Accessibility Concept Emergence in the Pythia Suite: Thresholds, Binding, and the Declarative-Evaluative Gap

Trisha Salas
Independent Researcher
March 2026

Abstract

Sustained deep-network binding of accessibility compounds appears to be a necessary structural condition for behavioral capability — present in every model that correctly defines core concepts, absent in every model that fails. We use Web Content Accessibility Guidelines (WCAG) as our test domain because accessibility represents a specialized, low-frequency domain in web-scale training data — a small, well-defined vocabulary with unambiguous answers and direct relevance to real-world tooling decisions, making it an unusually clean lens for studying emergence: concepts are concrete enough to evaluate and rare enough to show scale sensitivity. Using the Pythia suite (160M–12B) and GPT-2 (small–XL) with TransformerLens, we investigate not just what models know but how that knowledge is encoded internally. All models show compound binding in early layers; the differentiating factor is whether that binding persists to late network layers. Screen reader, skip link, and alt text emerge behaviorally at ~2.8B; WCAG first appears at 6.9B; Accessible Rich Internet Applications (ARIA) exhibits fluent wrongness at every scale tested — producing confident, plausible expansions that are consistently incorrect. Models prefer correct definitions before they can produce them, and a declarative-evaluative gap persists even at maximum scale: models that correctly define accessibility concepts cannot reliably identify violations in code. The gap is robust across 15 prompts spanning three elicitation strategies and is not an artifact of prompt design. Entropy analysis reveals that

the gap has internal structure — the model enters distinct failure states depending on how it is asked, from high-entropy stalling to low-entropy confident parroting. Extending the binding analysis to Pythia 12B introduces a late-layer resurgence pattern: a cluster of binding heads re-engaging near the output layers that scales monotonically with model size.

Introduction

Digital accessibility guidelines represent a narrow, specialized domain in web-scale training data. Terms like WCAG, ARIA, and alt text appear far less frequently than general web vocabulary, and their correct usage requires understanding both syntax and intent. This makes accessibility an unusual test case for studying emergence because the concepts are concrete enough to evaluate and rare enough to show scale sensitivity.

Most emergence research uses broad capability benchmarks like arithmetic, chain-of-thought reasoning, and multilingual translation. These are useful but can be noisy. Accessibility concepts offer something different: a small, well-defined vocabulary with unambiguous answers and direct relevance to real-world tooling decisions.

Binding Depth and Behavioral Emergence Follow the Same Shape

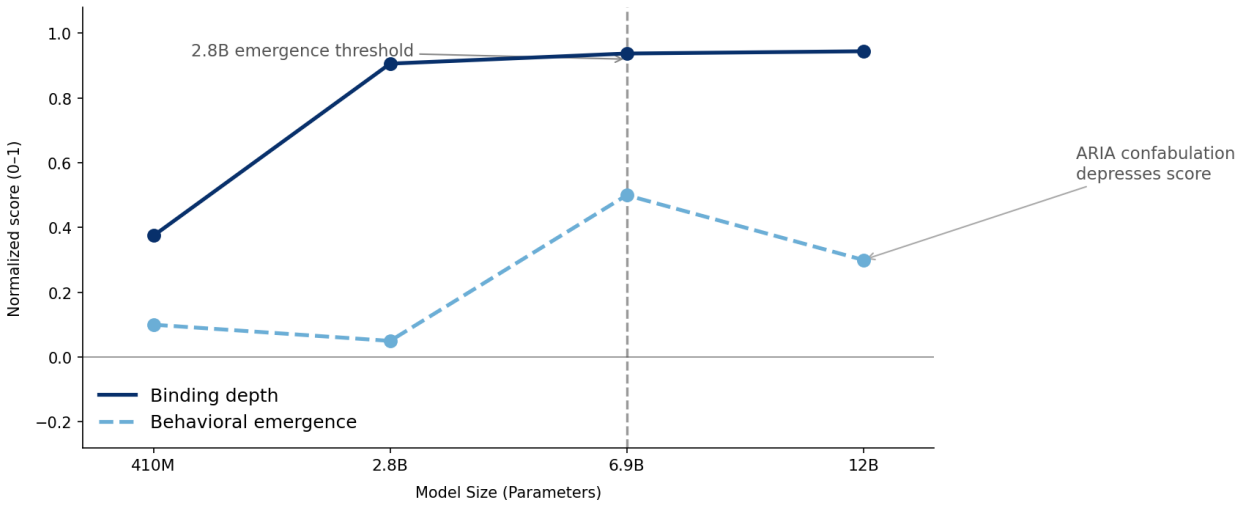


Figure 1: Binding depth and behavioral emergence score rise at the same 2.8B parameter threshold across the Pythia model suite. ARIA scores negative at 1B--6.9B reflecting fluent confabulation rather than absence of response.

The findings are specific: accessibility knowledge emerges at 2.8B parameters for core concepts, later or not at all for acronyms, and evaluative capability lags behind declarative knowledge even at maximum scale tested. The declarative-evaluative gap is robust across multiple elicitation strategies and prompt structures, ruling out prompt design as an explanation. Entropy analysis reveals that the gap has internal structure: models enter distinct failure states depending on how they are asked. Attention pattern analysis across all six Pythia scales (160M–12B) shows mechanistic encoding of compound terms before full behavioral competence appears, internal representation that is not reducible to output accuracy. At larger scales, a late-layer resurgence pattern emerges: binding heads re-engage near the output layers, scaling monotonically with model size. Replication across the GPT-2 model suite confirms that the core patterns — recognition preceding generation, deep-network binding persistence as a correlate of emergence, and the declarative-evaluative gap — hold across architectures and training corpora.

A preliminary version of these findings appeared as a blog post (Salas, 2026); we present the complete experimental results with extended analysis and cross-architecture replication.

Related Work

Emergence in language models. Wei et al. (2022) established that language models exhibit emergent abilities — capabilities that appear sharply at specific parameter scales rather than improving gradually. We treat accessibility concept acquisition as a domain-specific case study of emergence, using the Pythia suite's controlled architecture to isolate scale as the variable of interest.

Knowledge representation in transformers. Geva et al. (2021) demonstrated that feed-forward layers in transformers function as key-value memories, storing factual associations that attention heads retrieve and route. Subsequent work by Yao et al. (2024) mapped knowledge circuits in GPT-2 Medium using factual recall tasks, noting explicitly that their findings "may not be universally applicable to all knowledge domains." We examine whether domain-specific, low-frequency knowledge follows the same patterns.

Attention as an interpretability tool. Clark et al. (2019) showed that attention patterns in BERT encode linguistically meaningful relationships, establishing attention visualization as a valid method for probing internal representations. We apply the same approach to compound accessibility terms, asking whether models bind multi-token concepts as units.

Mechanistic interpretability. Olsson et al. (2022) characterized induction circuits and previous-token heads in transformer models, providing the framework used here to interpret early-layer binding patterns. The distinction between positional binding mechanisms and representational binding — central to the control analysis — draws directly on their circuit taxonomy.

AI and digital accessibility. A growing body of work applies AI to accessibility tasks — generating alt text, detecting contrast failures, automating WCAG audits. We take the inverse approach: rather than using AI as an accessibility tool, we examine what AI systems actually know about accessibility and when that knowledge emerges.

Methodology

Models

This study uses the Pythia model suite (Biderman et al., 2023): 160M, 410M, 1B, 2.8B, 6.9B, and 12B parameter models. Pythia models share consistent architecture and training data across all sizes, isolating scale as the variable

of interest. All models were loaded using TransformerLens (Nanda & Bloom, 2022).

One architectural exception: Pythia 1B uses a different configuration than the other models in the suite: 16 layers, 8 heads, d_model=2048, d_mlp=8192. The other models use 12 heads. This difference is relevant when interpreting per-head binding counts in Experiment 4; comparisons across model sizes account for this asymmetry.

To assess replicability across model families, all four experiments were additionally run on the GPT-2 model suite (Radford et al., 2019): small (117M), medium (406M), large (838M), and XL (1.5B) parameter models. Unlike Pythia, GPT-2 models vary in both layer count and head count across sizes, 12/16/20/25 heads and 12/24/36/48 layers respectively, making direct architectural comparisons across sizes less controlled. GPT-2 was trained on WebText, a dataset of outbound Reddit links, which differs substantially from Pythia's training corpus (The Pile). Results from both model families are reported and cross-architecture comparisons are discussed further in Section five.

Experiments

Experiment 1: Declarative Knowledge

Ten accessibility concept prompts (see Appendix A) were run against each model using greedy decoding (temperature=0, max_new_tokens=10):

```
model = HookedTransformer.from_pretrained("pythia-2.8b")
output = model.generate(prompt, max_new_tokens=10, temperature=0)
```

Prompts covered core accessibility terms (screen reader, alt text, skip link), foundational acronyms (WCAG, ARIA), and general accessibility concepts (focus indicator, keyboard navigation, color contrast, semantic HTML, captions). Responses were evaluated against known correct completions including definitions for concept prompts, purpose statements for functional prompts, and correct expansions for acronym prompts.

Experiment 2a: Evaluative Knowledge

Five code prompts (see Appendix A) were run against each model (`temperature=0, max_new_tokens=20`), asking models to identify accessibility violations in HTML snippets. Prompts covered missing alt attributes, non-semantic interactive elements, empty links, unlabeled inputs, and ambiguous link text. Responses were evaluated against correct accessibility explanations.

Experiment 2b: Elicitation Robustness

To determine whether the evaluative gap observed in Experiment 2 reflected a genuine capability boundary or an artifact of prompt design, testing was expanded to include multiple elicitation strategies. Two completion strategies — few-shot structural and bare completion — were used to confirm the presence of declarative knowledge without introducing accessibility-specific language into the prompt. Three hypothesis-driven strategies — direct questioning, error correction, and Socratic prompting — varied the evaluative framing while holding the underlying concept constant. A small validation set introduced additional prompt structures to test whether failures reflected prompt-specific behavior or a broader pattern. Testing was limited to Pythia 1B and 2.8B, the two scales where the declarative-evaluative gap was originally observed. Full prompt specifications appear in Appendix C.

Experiment 2c: Entropy Analysis

Mean and last-token entropy were computed on the three hypothesis-driven prompts from Experiment 2b for both 1B and 2.8B. Entropy measures the model's predictive uncertainty at the point of generation:

$$H = -\sum_{i} P(x_i) \log P(x_i)$$

Higher entropy indicates greater uncertainty across the predicted token distribution; lower entropy indicates the model concentrating probability mass on fewer candidates. Comparing entropy profiles across prompt types at the same model scale reveals whether different elicitation strategies produce distinguishable internal states, even when the behavioral outcome is the same.

Experiment 3: Recognition vs. Generation

Perplexity measures how expected a sequence is to the model, defined as:

$$PPL(X) = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log P(x_i \mid x_{<i})\right)$$

Lower perplexity indicates the model finds the text more natural. Comparing perplexity for a correct and incorrect definition across model sizes reveals whether recognition precedes generation. (see Appendix A)

```python
def get_perplexity(model, text):
    tokens = model.to_tokens(text)
    logits = model(tokens)
    log_probs = torch.nn.functional.log_softmax(logits, dim=-1)
    token_log_probs = log_probs[0, :-1, :].gather(
        1, tokens[0, 1:].unsqueeze(1)).squeeze()
    return torch.exp(-token_log_probs.mean()).item()
```

Experiment 4: Attention Pattern Analysis

Attention weights were extracted across all layers and heads for the prompt "A screen reader is" using TransformerLens's activation cache:

```python
logits, cache = model.run_with_cache(prompt)
attention = cache["pattern", layer]  # shape: [heads, seq, seq]
```

Token indices were verified for each model prior to analysis:

```python
tokens = model.to_str_tokens(prompt)
print(list(enumerate(tokens)))
```

Binding strength between tokens t_i ("reader") and t_j ("screen") was measured as the scalar attention weight:

$$b(t_i, t_j) = A_{ij}^{(l,h)}$$

where $A^{(l,h)}$ is the attention weight matrix at layer l, head h, extracted from the activation cache. Heads were classified into binding tiers based on $b(t_i, t_j)^{(l,h)}$:

$$H_{\text{strong}} = (l, h) : b(t_i, t_j)^{(l,h)} \geq 0.5$$

$$H_{\text{moderate}} = (l, h) : 0.2 \leq b(t_i, t_j)^{(l,h)} < 0.5$$

Heads with $b(t_i, t_j)^{(l,h)} < 0.2$ were treated as background noise and excluded from analysis. Results report $|H_{\text{strong}}|$ as the primary binding count.

This experiment was run for Pythia 2.8B and 6.9B. Additional binding pairs (alt text, skip link) were analyzed at 2.8B specifically, as this was the first model to demonstrate consistent declarative knowledge across Experiment 1. For GPT-2, binding analysis was run across all four model sizes for screen reader, with additional compound pairs (alt text, skip link) analyzed at XL. Full attention binding data for all model sizes is available in the project repository.

Reproducibility

All experiments were run with temperature=0 for deterministic outputs on Google Colab with an A100 GPU. Greedy decoding was chosen for mechanistic work to ensure deterministic outputs; stochastic decoding may surface partial knowledge that greedy decoding suppresses and is noted as a direction for future exploration. Full notebooks are available at https://github.com/trishasalas/how-models-think

Results

Experiment 1: Declarative Knowledge

A correct response demonstrates that the model has encoded the accessibility concept at the scale tested; incorrect or partial responses indicate the concept has not yet emerged.

Models were tested on ten accessibility concept prompts across all six Pythia sizes. Results show a sharp threshold pattern for core accessibility terms, with acronyms behaving differently.

Correct = matches known definition; Partial = incomplete or imprecise; Incorrect = wrong, off-topic, or loops

Prompt	160M	410M	1B	2.8B	6.9B	12B
A screen reader is	×	×	≈	✓	≈	✓
WCAG stands for	×	×	×	×	✓	✓
A skip link is	×	≈	×	✓	×	×
The purpose of alt text is	×	×	×	✓	✓	✓
ARIA stands for	×	×	×	×	×	×
A focus indicator is	×	×	×	×	×	×
Keyboard navigation allows	×	×	×	×	×	≈
Color contrast is important because	×	×	×	×	×	×
Semantic HTML helps	×	×	×	×	×	×
Captions are used for	×	≈	×	≈	×	×

This data reveals a few distinct patterns.

Threshold at 2.8B.

Screen reader, skip link, and alt text all show correct or near-correct responses first at 2.8B. Below this threshold, responses are either wrong or incomplete.

Screen reader shows non-monotonic behavior.

The 2.8B model produces "reads aloud the text," correctly capturing the auditory output purpose. The 6.9B model produces "reads text on a computer screen," dropping "aloud." The larger model is more verbose but less precise on the critical detail. This mirrors the skip link result, where 2.8B is correct and 6.9B regresses. Capability near emergence thresholds can be unstable.

WCAG emerges at 6.9B; ARIA never emerges.

Both are foundational accessibility acronyms. WCAG requires 6.9B parameters to expand correctly. ARIA fails at every scale tested — at 160M it produces gibberish, at 410M it loops, and from 1B onward it produces confident but wrong expansions ("Artificial Replacement of a Human Being," "A Rational Approach to Information and Automation," "Association of Research Libraries in Africa"). Confidence increases with scale; accuracy does not. This suggests ARIA is rare enough in training data that even 6.9B parameters cannot reliably encode it, while increased scale produces more fluent confabulation rather than correct recall.

General accessibility concepts fail across all scales. Focus indicator, keyboard navigation, color contrast, and semantic HTML produce generic responses at every model size. These terms appear in web-scale training data but without accessibility-specific context. The models complete the prompts plausibly without encoding accessibility meaning.

Replication: GPT-2

The same ten prompts were run across GPT-2 small (117M), medium (406M), large (838M), and XL (1.5B).

Prompt	Small	Medium	Large	XL
A screen reader is	×	×	×	≈
WCAG stands for	×	×	×	✓
A skip link is	×	×	×	×

Prompt	Small	Medium	Large	XL
The purpose of alt text is	×	×	≈	≈
ARIA stands for	×	×	×	×

The core findings replicate directionally. WCAG emerges at XL (1.5B) — a lower parameter count than Pythia's 6.9B, consistent with WCAG appearing more densely in WebText's Reddit-sourced content around web standards discussions. Screen reader never fully emerges in GPT-2; even at XL the model produces a partially correct response missing the critical "aloud" detail. ARIA fails at every scale tested in both model families. The declarative–evaluative gap and the pattern of sparse accessibility terms failing at all scales are consistent across architectures.

Experiment 2: Evaluative Knowledge

2a: Evaluative Prompts

A correct response requires the model to identify the specific accessibility violation, not just complete the code structure plausibly.

Models were tested on five code prompts requiring identification of accessibility violations. This tests whether models can apply accessibility knowledge, not just produce definitions.

Correct = identifies the accessibility violation accurately; Partial = identifies some issue but not the core violation; Incorrect = wrong, off-topic, or loops

Prompt	160M	410M	1B	2.8B	6.9B
`<img src='photo.jpg'>` missing what	×	×	×	×	×
`<div>` with onclick not accessible because	×	×	×	×	×

Prompt	160M	410M	1B	2.8B	6.9B
Problem with `<a href='#'></a>`	×	≈	×	×	×
`<input type='text'>` needs a	×	×	×	×	×
'Click here' button is bad because	×	×	×	✓	✓

There is a clear gap between declarative and evaluative knowledge. The 2.8B model correctly defines alt text but cannot identify that `<img src='photo.jpg'>` is missing one — it repeats the prompt and stalls at every scale tested. The question explicitly asks what is missing; no model answers "alt text."

The only evaluative success is the "Click here" prompt, which succeeds at 2.8B and 6.9B. Ambiguous link text may be more common in training data as a named anti-pattern than missing alt attributes.

The `<div>` onclick responses show a trajectory worth noting. Responses become more specific with scale — from loops at 160M to "not a form control" at 6.9B — without arriving at the correct explanation (keyboard inaccessibility, missing role and tabindex). The 6.9B answer is the closest, pointing toward interactivity expectations, but it does not identify the actual problem. This pattern reflects a broader limitation: models complete code structurally before reasoning semantically. Syntactic plausibility and semantic correctness are separable capabilities, and scale closes the gap only partially.

Replication: GPT-2

The same five code prompts were run across all GPT-2 sizes. The declarative–evaluative gap replicates fully. No GPT-2 model identified missing alt text at any scale. "Click here" emerged as a partial success at large (838M) — earlier by parameter count than Pythia's 2.8B — but regressed at

XL, consistent with the instability observed near emergence thresholds in both model families. The evaluative gap does not close at any scale tested in either architecture.

2b: Elicitation Robustness

A natural objection to the gap observed in 2a is that it may reflect how we asked rather than what the model knows. To address this, testing was expanded across multiple elicitation strategies at 1B and 2.8B.

Completion tasks confirmed that declarative knowledge is present at 2.8B. The model completed every accessibility concept correctly except bare closed captions using few-shot structural and bare completion strategies that contained no accessibility-specific language. Control prompts completed correctly on both models, confirming that differences are specific to accessibility concepts, not general HTML ability. 1B struggled broadly across completion tasks.

Hypothesis-driven evaluation then tested whether the model could use that knowledge evaluatively. Three strategies — direct questioning, error correction, and Socratic prompting — produced the same core outcome: the model recognized the accessibility concept but could not identify the violation. The declarative-evaluative gap persists across all elicitation strategies tested. It is not an artifact of prompt design.

The 1B comparison provides a natural control. 1B fails broadly, producing generic or off-topic completions. But one case is particularly informative: when given the Direct Question evaluative prompt, 1B treats it as a completion task and accidentally generates alt text. A model below the declarative threshold stumbles into the correct output by misunderstanding the task. This demonstrates that the gap is not about task difficulty — it is about the transition from declarative to evaluative reasoning. The model that has the knowledge cannot apply it; the model that lacks the knowledge produces the right output by accident.

Describing the gap from inside the gap. One response warrants specific attention. When asked "What accessibility attribute is missing from this HTML: `<img src='photo.jpg'>`?", the 2.8B model responded:

> "I have a photo gallery on my website. I want to add an accessibility attribute to the image tag so that the image can be clicked on. I have tried adding the attribute to the img tag, but it doesn't work."

The model generates a first-person narrative about attempting and failing to do exactly what it was asked to do. It activates the correct concept neighborhood — accessibility attributes, image tags, the need to add something — but cannot traverse from recognition to resolution. This is behavioral evidence of partial binding: sufficient activation to enter the correct semantic territory, insufficient depth to reach the answer. The binding framework developed in Experiment 4 predicts exactly this kind of output at 2.8B, where compound binding is present but evaluative circuit completion is not.

2c: Entropy Analysis

To determine whether the evaluative failure is uniform or structured, mean and last-token entropy were computed on the three hypothesis-driven prompts for 1B and 2.8B.

Model	Prompt	Mean Entropy	Last Token Entropy
1B	Direct Question	4.7761	5.6073
1B	Error Correction	5.6830	4.9935
1B	Socratic	4.9666	5.7079
2.8B	Direct Question	4.4127	5.1879
2.8B	Error Correction	5.7185	5.5864
2.8B	Socratic	4.5445	4.1660

Three distinct failure profiles emerge at 2.8B. Error Correction produces the highest entropy — the model stalls, echoes, does not know where to go. Direct Question produces intermediate entropy — fluent confabulation, the model moves confidently in the wrong direction. Socratic prompting produces the lowest last-token entropy of any failure condition tested (4.1660) — the model locks onto the prompt's rhetorical structure and continues it with high confidence, but the confidence is about pattern continuation, not answering.

1B shows no comparable differentiation. Its entropy is relatively flat across prompt types.

The gap is not a single missing capability. It is a structured landscape of partial competence. The model enters measurably different internal states depending on how it is asked, even when the outcome is the same: failure. This is the first quantitative evidence that the declarative-evaluative gap has internal structure. The Socratic prompt's low-entropy confident failure — the model maximally confident while maximally wrong — warrants circuit-level investigation in future work.

Experiment 3: Recognition vs. Generation

A preference flip — where the model assigns lower perplexity to the correct definition than the incorrect one — indicates recognition has emerged at that scale.

Perplexity measures how expected a sequence is to the model. Lower perplexity means the model finds the text more natural. Testing whether models assign lower perplexity to a correct definition than an incorrect one reveals whether recognition precedes generation. This experiment was extended to three accessibility compounds to assess whether the recognition-before-generation pattern generalizes.

Model	Screen Reader	Alt Text	Skip Link
160M	Wrong 2.6x	Wrong 1.2x	Wrong 1.9x
410M	Wrong 1.2x	Correct 2.8x	Wrong 1.7x
1B	Correct 2.2x	Correct 1.9x	Wrong 1.3x
2.8B	Correct 4.0x	Correct 1.9x	Wrong 1.1x
6.9B	Correct 3.0x	Correct 3.1x	Wrong 1.8x
12B	Correct 4.5x	Correct 3.1x	Wrong 1.3x

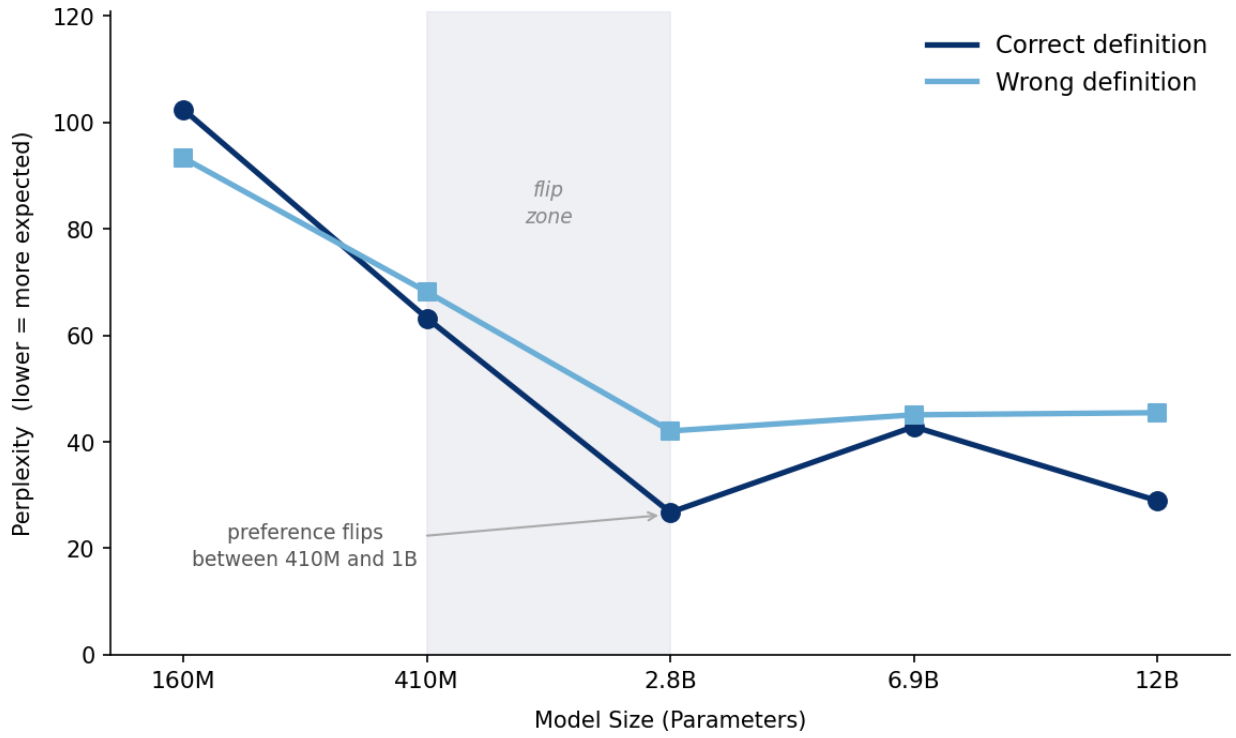


Figure 2: Correct definition perplexity falls below wrong definition perplexity between 410M and 1B parameters in Pythia, indicating the model finds the correct definition more expected before it can generate it.

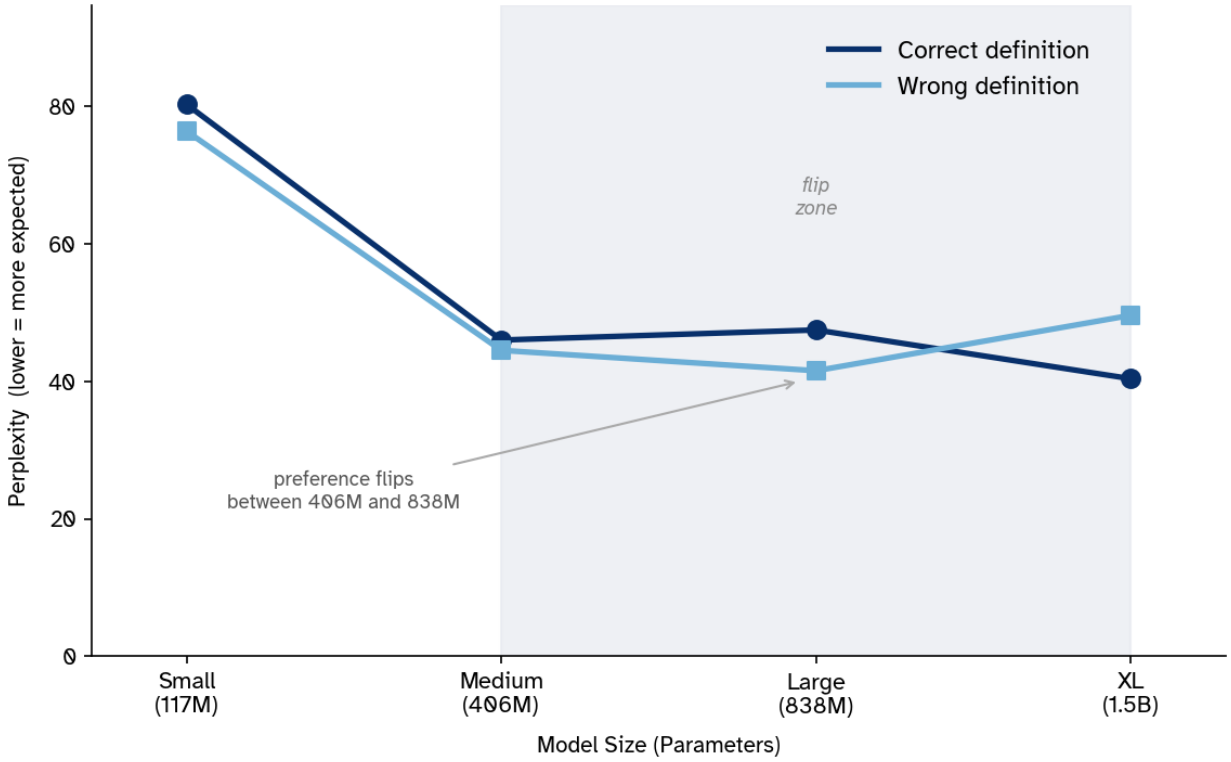
The preference for screen reader flips between 410M and 1B — recognition precedes generation by one model size. Alt text flips earlier, between 160M and 410M, suggesting it is more densely represented in training data. Skip

link never flips; the model finds the incorrect definition more natural at every scale, consistent with the behavioral instability observed in Experiment 1.

The recognition-before-generation pattern holds for two of three compounds tested. Skip link remains the exception across both perplexity and declarative experiments, indicating that some concepts may achieve surface-level generation without the underlying representational grounding that perplexity preference reflects.

Replication: GPT-2

Model	Screen Reader	Alt Text	Skip Link
Small (117M)	Wrong 1.1x	Correct 1.7x	Wrong 1.3x
Medium (406M)	Wrong 1.1x	Correct 1.8x	Wrong 1.4x
Large (838M)	Correct 2.0x	Correct 1.6x	Wrong 2.5x
XL (1.5B)	Correct 2.6x	Correct 2.1x	Wrong 1.8x



Correct definition perplexity falls below wrong definition perplexity between 406M and 838M parameters in GPT-2, replicating the perplexity preference flip observed in Pythia at a comparable scale threshold.

The screen reader preference flips between medium (406M) and large (838M) in GPT-2 — a remarkably similar parameter range to the Pythia flip between 410M and 1B, despite different architectures and training corpora. Alt text is correct-preferring from small in GPT-2, earlier than in Pythia, consistent with its denser representation in web-focused training data. Skip link never flips in either model family. The recognition-before-generation pattern for screen reader replicates across architectures.

Experiment 4: Mechanistic Analysis of Compound Term Binding

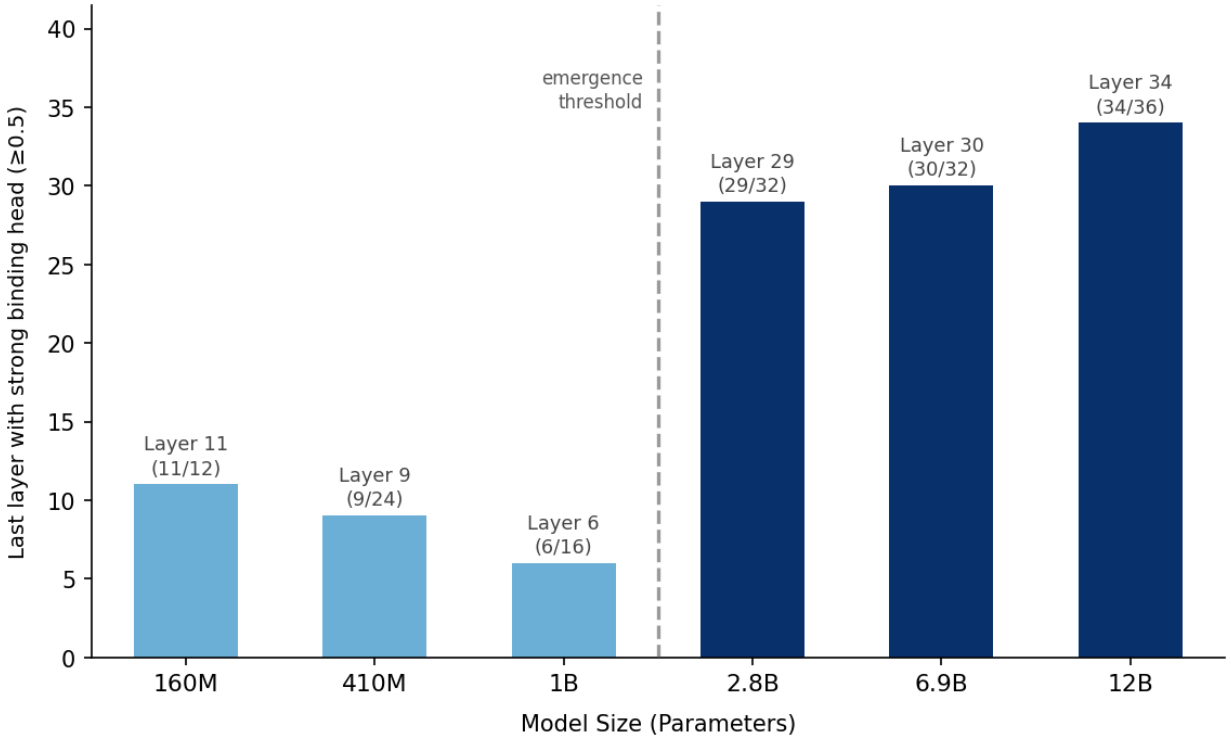
Sustained strong binding into late network layers, rather than early-layer binding alone, is the signal of interest.

Attention pattern analysis across all six Pythia models examines whether models treat “screen reader” as a compound concept or as two independent tokens, and how this binding pattern relates to behavioral emergence.

For each model, attention weights from “reader” to “screen” were extracted across all layers and heads. Heads with weight above 0.5 are considered strong binding; the full data is available in the project repository.

Model	Layers	Strong (0.5+)	Early (L0-3)	Last Layer
160M	12	10	7	11
410M	24	22	16	9
1B	16	11	10	6
2.8B	32	25	20	29
6.9B	32	37	28	30
12B	36	36	26	34

Note: 1B's architectural difference (8 heads vs 12) affects raw head counts; see Methodology.



The last layer containing a strong binding head (attention score ≥0.5) drops through 160M, 410M, and 1B before jumping sharply at 2.8B, coinciding with the emergence threshold. Models above the threshold show binding heads persisting into the final layers. 12B extends the pattern to 97% of network depth.

The binding data shows several distinct patterns.

Early layers dominate across all models. Between 70-91% of strong binding occurs in layers 0-3 regardless of model size. Compound term binding is established early in the forward pass at every scale tested.

Strong head count scales with model size. 160M has 10 strong binding heads; 6.9B has 37; 12B has 36. 1B is the expected outlier given its different architecture.

Last strong layer tracks behavioral emergence. Below the 2.8B emergence threshold, strong binding drops off early — layer 11 for 160M, layer 9 for 410M, layer 6 for 1B. At 2.8B and above, strong binding persists deep into

the network — layers 29, 30, and 34 for 2.8B, 6.9B, and 12B respectively. The models that cannot correctly define screen reader do not sustain compound binding through the network. The models that can, do.

2.8B is the inflection point. Total heads above threshold jumps from 28 (1B) to 101 (2.8B) — a 3.6x increase that coincides exactly with the behavioral emergence threshold identified in Experiment 1. Early binding is not representational; late binding is. The presence of sustained late-layer binding appears to be a mechanistic bottleneck for concept emergence.

All six models show strong binding in layers 0-3, including 160M which cannot produce a correct definition. Whether early-layer binding at small scales reflects genuine compound representation or proximity effects cannot be determined from attention weights alone and is noted as a limitation.

Control Experiment: Ruling Out Proximity Effects

To test whether the binding signal reflects compound concept encoding rather than simple token adjacency, attention weights were measured between non-compound token pairs at 2.8B. Two conditions were tested: adjacent function words ("and then") and an adjacent modifier-noun pair without disambiguation ("cold water"). Both conditions produced strong early-layer binding, consistent with the known behavior of previous-token heads (Olsson et al., 2022) — a class of induction circuit components that attend systematically to the immediately preceding position regardless of content.

This reframes the control comparison. Early-layer binding is not specific to accessibility compounds; it reflects general positional mechanisms present across token types. The meaningful signal is the distribution and persistence of binding across the full head population. Accessibility compounds at 2.8B recruit 101-208 heads above threshold with strong binding persisting to layers 29-30. Function word pairs produce early-layer binding without the same deep-network persistence. The sustained late-layer binding pattern, rather than raw head count, appears to differentiate accessibility compound binding from general adjacent-token binding. A complete characterization

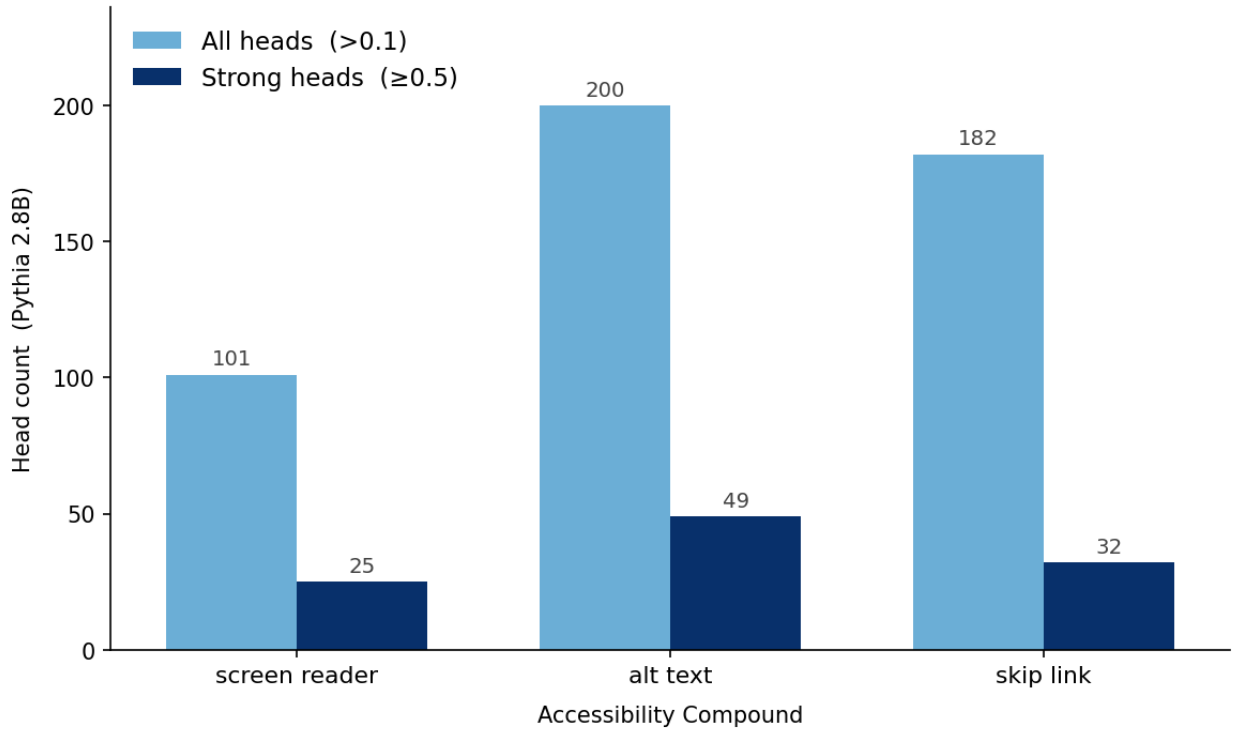
requires systematic comparison across a broader set of controls and is noted as a direction for future work.

Binding Generalizes Across Accessibility Compounds

Having ruled out proximity effects, attention binding was measured for two additional accessibility compound terms at 2.8B: alt text and skip link.

Compound	Total heads >0.1	Strong (0.5+)	Top score
screen reader	101	25	0.9909
alt text	200	49	0.9856
skip link	182	32	0.9816

All three compounds show the same pattern of early-layer concentration with deep-network persistence at 2.8B, with top binding scores above 0.98 in each case.



Attention head counts for three accessibility compounds in Pythia 2.8B, showing all heads (attention score >0.1) and strong heads (≥0.5). Screen reader activates fewer heads over-

all, while alt text and skip link show comparable strong head counts despite differences in total activation.

This rules out a compound-specific explanation. The binding pattern is not an artifact of how "screen reader" tokenizes or how frequently it appears in training data. It is a general property of accessibility compound terms at the 2.8B emergence threshold. The models that can define these concepts correctly show robust, distributed binding across many heads and many layers. The models that cannot show weak binding that drops off early.

The binding pattern strengthens the case for a general mechanistic threshold at this scale.

Extending the Binding Curve: Pythia 12B

To determine whether the sustained binding pattern continues beyond the original suite, the attention binding analysis was extended to Pythia 12B using the same methodology and thresholds.

12B produces 119 total heads above the 0.1 threshold and 36 strong heads above 0.5. The last strong binding layer is 34 (out of 35, zero-indexed), placing binding persistence at 97% of network depth. The top binding head is Layer 1, Head 15 at 0.9905.

Late-layer resurgence. Layer 33 contains 7 active binding heads, 4 above the strong threshold, including Head 16 at 0.8766 — the strongest late-layer cluster of any model tested. Late-layer binding entries scale monotonically with model size:

Model	Late-layer entries
1B	1
2.8B	4
6.9B	6
12B	11

The resurgence is consistent with the hypothesis that wider MLP layers sustain compound representations through deeper network propagation. Direct MLP investigation is deferred to future work.

Specialization. The percentage of heads active in binding decreases with scale — 31.9% at 160M, 8.3% at 12B — while the absolute count plateaus around 100–120 for models at 2.8B and above. Larger models use proportionally fewer but more specialized heads. The model converges on how many heads it needs for compound binding and stops recruiting more.

Cross-model patterns. Several binding properties are universal across all six Pythia scales. Binding is front-loaded in all models, with 65–82% of active heads concentrated in the first 25% of layers. Strong heads above 0.9 cluster at 3–10% of network depth for all models at 1B and above. Mean binding scores remain stable at approximately 0.36–0.44 regardless of scale.

The new observations at 12B are the late-layer resurgence and the specialization pattern. The strength of binding does not change with scale. The depth does. Figure 4 presents the updated binding persistence curve across all six Pythia scales.

Replication: GPT-2

Attention binding was measured across all four GPT-2 model sizes for screen reader, with additional compounds (alt text, skip link) analyzed at XL.

Model	Total >0.1	Strong (0.5+)	Last Strong Layer	Max Layers
Small (117M)	41	6	4	12
Medium (406M)	66	12	10	24
Large (838M)	146	34	15	36
XL (1.5B)	199	37	17	48

The strong head count jump between medium and large (12 to 34) coincides with the perplexity flip observed in Experiment 3, replicating the Pythia pattern in a different architecture. The last strong layer advances from 10 to 15 at the same transition — consistent with deep-network persistence as a correlate of recognition-level emergence.

At XL, binding was measured for all three compounds:

Compound	Total >0.1	Strong (0.5+)	Last Strong Layer
Screen reader	199	37	17
Alt text	244	34	19
Skip link	253	49	43

One divergence from Pythia is notable. GPT-2 XL's last strong layer as a proportion of total network depth is substantially shallower than Pythia 2.8B — approximately 35% vs 91%. This may reflect differences in training data composition and architectural differences. Pythia's training corpus (The Pile) includes technical documentation, Stack Exchange, and academic text where accessibility terminology appears in consistent, well-formed contexts. GPT-2's WebText corpus is dominated by general web content where the same terms appear more diffusely. The shallower binding depth in GPT-2 may account for the weaker and less stable behavioral emergence observed in Experiment 1.

Discussion

We examined accessibility concept emergence across the Pythia model suite using mechanistic analysis, perplexity-based recognition, and behavioral evaluation, with extended elicitation testing and entropy analysis at the scales where the declarative-evaluative gap was first observed. The results converge on a consistent picture with several implications for both accessibility tooling and mechanistic interpretability research.

The 2.8B Threshold Is Meaningful, Not Arbitrary

The behavioral findings show a sharp threshold at 2.8B parameters for core accessibility concepts. This is not a smooth gradient — 1B partially retrieves "reads text from a screen" while 2.8B produces "reads aloud the text," capturing the auditory delivery mechanism of a screen reader. The threshold pattern, combined with the 3.6x jump in sustained attention binding heads at 2.8B, suggests a qualitative shift rather than incremental improvement. Below 2.8B, the model has insufficient encoding to support correct generation. At 2.8B, encoding is sufficient to sustain compound binding throughout the network and produce correct behavioral output.

The elicitation robustness experiments provide further evidence that the threshold represents a genuine capability boundary. The gap between declarative and evaluative knowledge persists across 15 prompts spanning three elicitation strategies — completion, hypothesis-driven, and validation — ruling out prompt design as an explanation. The entropy differentiation at 2.8B adds a quantitative dimension: the model enters measurably different internal states depending on how it is asked, behavior absent at 1B. The threshold is not just a behavioral boundary but a point at which internal processing becomes structured enough to differentiate across prompt types, even when the outcome is the same.

Encoding and Retrieval Are Dissociable

The perplexity results demonstrate that models prefer correct definitions before they can produce them. At 1B, the model assigns lower perplexity to the correct screen reader definition despite generating "reads text from a screen." This dissociation between recognition and generation suggests that accessibility knowledge exists earlier than the model can articulate it.

This has practical implications: perplexity-based probes may detect accessibility knowledge at smaller scales than behavioral tests suggest.

The Declarative-Evaluative Gap Is Not Closed at Any Scale Tested

Even at 6.9B, models that correctly define accessibility concepts cannot reliably identify violations in code. The single evaluative success, identifying ambiguous link text, may reflect training data frequency rather than genuine accessibility reasoning. The image missing alt attribute failure is particularly telling: a model asked explicitly what is missing produces the prompt back rather than "alt text," despite correctly defining alt text in declarative tests. Knowing and applying are different capabilities that do not emerge together.

The extended elicitation experiments deepen this finding substantially. The gap persists across all elicitation strategies tested — direct questioning, error correction, Socratic prompting, few-shot structural completion, and validation prompts — establishing that it is not a retrieval problem but a genuine capability boundary. The 1B model's accidental success on one evaluative prompt, where it treats the question as a completion task and stumbles into generating alt text, provides a natural control: the gap is not about task difficulty but about the transition from declarative to evaluative reasoning.

The entropy analysis reveals that the gap has internal structure. Three distinct failure profiles emerge at 2.8B: high entropy under error correction, where the model stalls or echoes; intermediate entropy under direct questioning, where it confabulates fluently; and low entropy under Socratic prompting, where it locks onto the prompt's rhetorical structure and continues with confidence. The model is not failing uniformly — it enters measurably different internal states depending on how it is asked, even when the outcome is the same. The 1B model shows no comparable differentiation. The structured failure profile is specific to the scale where declarative knowledge is present.

The Socratic prompt's low-entropy confident failure — the lowest last-token entropy of any failure condition tested — is a specific prediction for future circuit work. The model is maximally confident while maximally wrong, locking onto rhetorical structure rather than semantic content. Identifying

the circuit mechanism that produces this confident misfire would connect the behavioral observation to a specific computational pathway.

The 2.8B model's response to the instructional correction prompt — generating a first-person narrative about attempting and failing to add an accessibility attribute — provides behavioral evidence of partial binding. The model activates the correct concept neighborhood but cannot traverse from recognition to resolution. This is exactly the kind of output the binding framework predicts at a scale where compound binding is present but evaluative circuit completion is not.

For accessibility tooling practitioners, this gap is the central finding. A model that passes a definition benchmark is not ready for code review tasks.

Acronym Failure Is Systematic

WCAG and ARIA show qualitatively different failure patterns than conceptual terms. Conceptual terms show emergence — partial at 1B, correct at 2.8B. Acronyms either emerge very late (WCAG at 6.9B) or not at all (ARIA at any scale tested). The ARIA failure is particularly informative: hallucinations become more fluent and confident with scale without becoming accurate. At 160M, ARIA produces gibberish. At 6.9B, it produces a coherent but wrong expansion ("Association of Research Libraries in Africa") with apparent conviction. This is consistent with inverse scaling — where fluency amplifies confident confabulation rather than correct retrieval with increased model size (McKenzie et al., 2023).

This suggests WCAG and ARIA are rare enough in web-scale training data that standard scaling cannot reliably encode them. This pattern is consistent with findings on knowledge frequency effects in parametric memory: model accuracy on factual recall correlates with entity popularity in training data, and rare entities produce confident confabulation rather than correct retrieval (Mallen et al., 2023). Specialized training data or fine-tuning is likely required for reliable acronym expansion in the accessibility domain.

Not All Concepts Benefit Equally from Scale

Skip link prefers the incorrect definition even at 12B, the largest model tested. Despite correct behavioral generation at 2.8B in Experiment 1, the perplexity analysis shows the model finds the wrong definition more natural at every scale in both model families. This suggests that training data representation may impose a floor that scale alone cannot overcome — the model can produce a correct completion through surface-level pattern matching without the deeper representational grounding that perplexity preference reflects.

This observation pairs with the ARIA confabulation pattern to illustrate a broader point: scaling produces different failure modes for different concepts rather than uniform improvement. ARIA gets more fluently wrong. Skip link gets behaviorally right without representational depth. Both suggest limits to what scale alone can achieve for low-frequency specialized vocabulary, though the specific failure mechanism differs. Cross-domain investigation of whether similar patterns appear for specialized terminology outside accessibility is left for future work.

Attention Binding as a Mechanistic Correlate of Emergence

The sustained attention binding pattern across the full Pythia suite — strong binding persisting to layers 29–30 at 2.8B and 6.9B, and to layer 34 at 12B, versus dropping off at layers 6–11 in smaller models — provides a mechanistic correlate for behavioral emergence. The models that can define screen reader sustain the compound binding across the full network. The models that cannot drop it early.

Early binding is not representational; late binding is. The presence of sustained late-layer binding appears to be a mechanistic bottleneck for concept emergence: a necessary structural condition without which behavioral competence does not appear. All models, including those that fail behaviorally, show strong binding in early layers. The differentiating factor is whether that binding persists deep into the network.

The 12B extension introduces two new observations. First, late-layer resurgence: a cluster of binding heads re-engages near the output layers, and this cluster scales monotonically with model size — from 1 late-layer entry at 1B to 11 at 12B. This is consistent with the hypothesis that wider MLP layers sustain compound representations through deeper network propagation. Second, specialization: the percentage of heads participating in binding decreases with scale (31.9% at 160M to 8.3% at 12B) while the absolute count plateaus around 100–120 for models at 2.8B and above. Larger models use proportionally fewer but more specialized heads, converging on how many heads are needed for compound binding and stopping recruitment.

The connection between entropy profiles and binding depth is suggestive. The three distinct entropy signatures at 2.8B — where declarative knowledge is present but evaluative capability is not — may reflect different degrees of partial binding activation. Different prompt structures may engage the binding circuit to different depths, producing the observed variation in model confidence even when all paths terminate in failure. This is speculative but identifies a specific connection point for future causal work: activation patching at the layers where binding diverges across prompt types could test whether the entropy differentiation has a mechanistic basis in the binding circuit.

This finding is correlational, not causal. Whether sustained binding causes correct generation, or both are consequences of a third factor such as MLP encoding depth, cannot be determined from attention weights alone (Geva et al., 2021).

Cross-Architecture Replication

The threshold and binding patterns were replicated on the GPT-2 model suite. The core findings hold across both model families.

The recognition-before-generation pattern for screen reader holds in GPT-2, with the perplexity preference flipping between medium (406M) and large (838M), a parameter range closely matching Pythia's 410M–1B transition.

The declarative–evaluative gap replicates fully: no GPT-2 model identified missing alt text, and ARIA fails at every scale in both families. The attention binding jump between GPT-2 medium and large (12 to 34 strong heads) coincides with the perplexity flip, consistent with deep-network persistence as a correlate of recognition-level emergence.

Notably, GPT-2's last strong binding layer as a proportion of total network depth is substantially shallower than Pythia 2.8B — approximately 35% vs 91%. Full behavioral emergence is weaker and less stable in GPT-2 across all compounds tested. This may reflect differences in training data composition and architectural differences — GPT-2 uses sequential attention+MLP blocks versus Pythia's parallel architecture, which may affect how binding reinforcement accumulates across layers. Pythia's training corpus (The Pile) includes technical documentation, Stack Exchange, and academic text where accessibility terminology appears in consistent, well-formed contexts. GPT-2's WebText corpus is dominated by general web content where the same terms appear more diffusely. The shallower binding depth in GPT-2 may account for the weaker behavioral emergence: the network encodes the compound relationship but does not reinforce it through enough layers to support reliable generation.

Relationship to Prior Work

Wei et al. (2022) established emergence as a general phenomenon. We treat accessibility concept acquisition as a domain-specific case study: concepts are rare enough to show scale sensitivity, evaluable against clear ground truth, and directly relevant to real-world applications. The results confirm that emergence operates in specialized domains but with domain-specific timing — accessibility acronyms require substantially more scale than conceptual terms, and evaluative capability does not emerge within the range tested.

Prior work on accessibility and LLMs has focused primarily on behavioral evaluation, whether models can identify or remediate violations. We add a mechanistic dimension, not only what models know but how that knowledge is structurally encoded, and at what scale that encoding becomes sufficient

for generation. The consistent results in both the Pythia and GPT-2 families strengthen the mechanistic claim that deep-network binding persistence is not limited to a single model family.

The finding that general accessibility concepts (focus indicator, keyboard navigation, color contrast, semantic HTML) fail across all scales in both model families suggests training data frequency is the limiting factor, not model capacity. These terms appear in web-scale data but without the accessibility-specific context required for accurate representation.

Conclusion

We examined accessibility concept acquisition across six Pythia model sizes (160M–12B) using mechanistic analysis, perplexity-based recognition, and behavioral evaluation, with extended elicitation robustness testing and entropy analysis at the scales where the declarative-evaluative gap was first observed. Replication across the GPT-2 model suite confirms the core findings hold across architectures and training corpora.

Core accessibility concepts — screen reader, alt text, skip link — emerge behaviorally at 2.8B parameters in Pythia. Foundational acronyms require more scale (WCAG at 6.9B) or fail entirely (ARIA at all scales tested). General accessibility vocabulary fails across all model sizes in both model families, suggesting training data frequency as the limiting factor rather than model capacity. Evaluative capability — identifying accessibility violations in code — does not emerge within the range tested, even in models that correctly define the relevant concepts. The declarative-evaluative gap persists across 15 prompts spanning three elicitation strategies, establishing that it is not an artifact of prompt design but a genuine capability boundary.

Entropy analysis reveals that the gap has internal structure. Models enter distinct failure states depending on how they are asked — from high-entropy stalling to low-entropy confident parroting — even when the behavioral outcome is the same. The structured failure profile is specific to the scale where declarative knowledge is present.

The perplexity results show that recognition precedes generation: models prefer correct definitions before they can produce them, with the preference flipping between 410M and 1B in Pythia and between 406M and 838M in GPT-2, a remarkably consistent parameter range across different architectures and training corpora. The attention binding analysis shows that sustained binding of "screen reader" as a compound concept across the full network is a mechanistic correlate of emergence — present at 2.8B, 6.9B, and 12B in Pythia, absent at smaller scales. The same binding jump replicates in GPT-2 at the transition between medium and large, coinciding with the perplexity flip. Early binding is not representational; late binding is. At 12B, a late-layer resurgence pattern emerges — a cluster of binding heads re-engaging near the output layers that scales monotonically with model size — alongside increased specialization, where fewer heads participate in binding while sustaining it deeper through the network.

Not all concepts benefit equally from scale. Skip link prefers the incorrect definition even at 12B despite correct behavioral generation at 2.8B, suggesting that training data representation imposes a floor that scale alone cannot overcome.

For practitioners building accessibility tooling on language models, the central implication is that behavioral success on definition tasks does not predict success on evaluation tasks. The declarative-evaluative gap observed here replicates across both model families and suggests that reliable accessibility code review requires either substantially larger models or domain-specific training data that goes beyond the web-scale distribution.

For emergence researchers, accessibility concepts offer a useful probe domain: rare enough to show scale sensitivity, concrete enough to evaluate against ground truth, and distinct enough from general knowledge to isolate domain-specific acquisition. The ARIA confabulation pattern — increasingly fluent wrong answers with scale — illustrates a failure mode that definition-based benchmarks would not detect. The entropy-differentiated failure profiles at 2.8B suggest that emergence boundaries are not uniform

but structured, a finding that may generalize beyond accessibility to other specialized domains.

The mechanistic findings here are correlational. Whether sustained attention binding causes behavioral emergence or reflects a common underlying factor requires causal analysis beyond the scope of this study. The Socratic prompt's low-entropy confident failure and the late-layer resurgence pattern identify specific targets for future circuit-level investigation.

Limitations

We report findings from a first investigation into accessibility concept emergence across the Pythia model suite, with replication across GPT-2. Several limitations should be considered when interpreting the results.

Prompt Design

All declarative experiments used completion-style prompts ("A screen reader is," "WCAG stands for") rather than instruction or question formats. Prompt framing is known to affect retrieval in language models, and our own findings support that sensitivity. The alt text retrieval experiment demonstrated that the HTML-framed prompt ("In HTML, the alt attribute provides a text description of an") performed worse than the simpler framing despite testing identical knowledge. Results reported here reflect one prompt design choice and may not generalize to other framings.

The declarative prompts were adapted from attention binding experiments rather than designed independently for behavioral testing. Future work should compare multiple prompt formulations to establish which results are robust across framings.

Evaluative Experiment Prompts

The original evaluative experiment (Experiment 2a) used zero-shot code completion prompts. The image missing alt attribute result — where models

loop the prompt rather than identify the missing attribute — may reflect document completion behavior rather than a knowledge failure. Models trained on code repositories may interpret `<img src='photo.jpg'>` as the beginning of a code listing and complete accordingly.

The elicitation robustness experiments (Experiment 2b) partially address this limitation by testing the same underlying knowledge across multiple prompt formats, including direct questioning and error correction strategies that do not rely on code completion framing. The gap persists across all strategies tested, reducing the likelihood that the original finding was an artifact of completion-style prompting. However, instruction-tuned models may produce different results, and this remains a direction for future work.

Entropy Analysis

The entropy analysis (Experiment 2c) was computed on three hypothesis-driven prompts at two model scales. The three distinct failure profiles at 2.8B are a suggestive finding, but the small number of prompts per condition limits the statistical power of the observation. Whether the entropy differentiation generalizes across a broader set of evaluative prompts and additional model scales has not been tested. The entropy values reported are descriptive rather than inferential — no statistical significance tests were applied.

Perplexity Experiment Scope

Recognition versus generation was tested using three sentence pairs across screen reader, alt text, and skip link. The preference flip is a clear finding for screen reader and alt text, but skip link fails to flip in either model family, suggesting the recognition-before-generation pattern does not hold uniformly across all accessibility compounds. A broader set of correct/incorrect pairs across additional concepts would further characterize which compounds follow this pattern and which do not.

Attention Binding Methodology

The proximity control experiment was run at 2.8B only. Control testing revealed that adjacent token pairs — including function word pairs and modifier-noun pairs without disambiguation — produce strong binding at L1H12, which was subsequently confirmed as a previous-token head attending systematically to the immediately preceding position regardless of content. This finding limits the interpretation of L1H12's consistent appearance across accessibility compounds. Whether other heads show accessibility-specific binding patterns that do not appear in non-compound controls has not been fully characterized and is noted as a direction for future work. Early-layer binding at 160M (including Layer 11 Head 8 at score 1.0) remains ambiguous. Whether this reflects genuine compound representation or positional attention at small scales cannot be determined from attention weights alone.

The compound generalization finding — that alt text and skip link show comparable binding to screen reader at 2.8B — is based on head counts above the 0.1 threshold across all three compounds.

The 12B binding extension uses the same methodology and thresholds as the original experiment. The late-layer resurgence observation is based on head counts in the final layers and has not been tested with additional compound terms at 12B. Whether the resurgence pattern generalizes beyond "screen reader" at this scale is unknown.

Cross-Architecture Comparison

GPT-2 models vary in both layer count and head count across sizes, making direct per-head comparisons across model sizes less controlled than Pythia's uniform architecture. The binding depth comparison between GPT-2 and Pythia is expressed as a proportion of total network depth to account for this, but the architectures are not directly commensurable. Additional replication on architectures with controlled scaling (such as other Pythia-style suites) would strengthen the cross-architecture claims.

Greedy Decoding

All experiments were run with temperature=0 for deterministic outputs. Greedy decoding ensures reproducibility but may suppress partial knowledge that stochastic decoding could surface. Models with partial accessibility encoding may perform differently under sampling-based generation. This is noted as a direction for future exploration.

Hardware and Reproducibility

All experiments were run on Google Colab with an A100 GPU. All notebooks are available in the project repository for independent verification.

Scale Coverage

Declarative testing covers 160M through 12B for Pythia and 117M through 1.5B for GPT-2. The evaluative experiments (Experiment 2a) cover 160M through 6.9B for Pythia. The binding analysis extends to Pythia 12B. ARIA fails to emerge at any scale tested in either model family. Whether ARIA would emerge at larger scales is unknown. The evaluative gap — models that define concepts cannot identify violations — was not tested beyond 6.9B. Both findings may reflect limitations of the scale range rather than fundamental properties of these concepts.

Appendix A

1. Declarative Prompts

All declarative experiments used completion-style prompts. Models were given the prompt prefix and generated a continuation.

- A screen reader is
- WCAG stands for
- A skip link is
- The purpose of alt text is
- ARIA stands for
- A focus indicator is
- Keyboard navigation allows
- Color contrast is important because
- Semantic HTML helps
- Captions are used for

2. Evaluative Prompts

Evaluative experiments used zero-shot code completion prompts.

- The following code is not accessible because it doesn't have what? `<img src='photo.jpg'>`
- A `<div>` with onclick is not accessible because
- The accessibility problem with `<a href='#'></a>` is
- `<input type='text'>` needs a
- A button that only says 'Click here' is bad because

3. Perplexity Pairs

Each pair consists of a correct and incorrect definition. Perplexity was computed for each sentence independently. Values below 1.0 on the preference ratio indicate the model finds the wrong definition more natural.

Concept	Correct	Incorrect
screen reader	A screen reader is software that reads text aloud for blind users.	A screen reader is a device for viewing screens.
alt text	The purpose of alt text is to provide a textual description of an image for people with visual disabilities.	The purpose of alt text is to make images load faster.
skip link	A skip link is a navigation aid that allows keyboard users to bypass repetitive content.	A skip link is a broken hyperlink that does not load.

Appendix B

Model Architectures

All experiments used base (non-instruction-tuned) model checkpoints.

Pythia Scaling Suite

Architecture: parallel attention + MLP blocks.

Table 2

Model	Layers	Heads	d_model	d_mlp
160M	12	12	768	3,072
410M	24	16	1,024	4,096
1B	16	8	2,048	8,192
2.8B	32	32	2,560	10,240
6.9B	32	32	4,096	16,384
12B	36	40	5,120	20,480

GPT-2 Model Family

Architecture: sequential attention + MLP blocks.

Table 3

Model	Layers	Heads	d_model	d_mlp
Small (117M)	12	12	768	3,072
Medium (406M)	24	16	1,024	4,096
Large (838M)	36	20	1,280	5,120
XL (1.5B)	48	25	1,600	6,400

Appendix C

1. Completion Prompts

- **Alt text - Few Shot:**
  ```
  <img src="icons/home.png" alt="Home">
  <img src="icons/mail.png" alt="Inbox">
  <img src="photo.jpg
  ```
- **Alt Text - Bare:**
  ```
  <img src="photo.jpg"
  ```
- **Closed Captions - Few Shot:**
  ```
  <video src="intro.mp4">
  <track kind="captions" src="intro.vtt"></video>
  <video src="demo.mp4"><track kind="captions" src="demo.vtt"></video>
  <video src="lecture.mp4">
  ```
- **Closed Captions - Bare:**
  ```
  <video src="lecture.mp4">
  ```
- **Page Title - Few Shot:**
  ```
  <html><head><title>About Us</title></head>
  <html><head><title>Contact</title></head>
  <html><head>
  ```
- **Page Title - Bare:**
  ```
  <html><head><meta charset="utf-8">
  ```
- **Script Source - Bare (control):**
  ```
  <script
  ```
- **Script Source - Few Shot (control):**
  ```
  <script src="utils.js"></script>
  <script src="main.js"></script>
  <script src="
  ```

2. Hypothesis Driven Prompts

- **Direct Question:**
  Look at this HTML: `<img src='portrait.jpg'>`.
  What specific attribute is missing to make it accessible?

Expected Failure: Model repeats the tag or provides a generic definition of images rather than ‘alt text’.

- **Error Correction:**
  Correct the following inaccessible code: `<img src='logo.png'>`.
  Expected Failure: Model completes the line with more attributes (like width or height) but ignores the alt attribute.
- **Negative Constraint (Socratic)**
  A screen reader is software that reads text aloud.
  Why would a screen reader fail to describe this tag:
  `<img src='chart.svg'>`?
  Expected Failure: Despite the ‘screen reader’ anchor (which is deeply bound at 2.8B), the model cannot link the concept to the lack of an alt attribute.

3. **Validation Prompts**

- **Direct Completion:**
  The following code is not accessible because it doesn’t have what? `<img src='photo.jpg'>`?
  Expected Behavior: Stalls or repeats prompt despite correctly defining ‘alt text’ elsewhere.
- **Instructional Correction:**
  What accessibility attribute is missing from this HTML: `<img src='photo.jpg'>`?
  Expected Behavior: Failure to answer ‘alt text’.
- **Functional Socratic:** A screen reader is software that reads text aloud. Why would it fail to describe `<img src='photo.jpg'>`?
  Expected Behavior: Model fails to link the ‘screen reader’ definition to the missing ‘alt’ attribute.”
- **Fluent Confabulation (Control):** ARIA stands for
  Expected Behavior: Produces confident but incorrect expansions.
- **Fluent Confabulation (Control):** HTML stands for
  Expected Behavior: Produces confident but incorrect expansions.

References

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., Skowron, A., Sutawika, L., & van der Wal, O. (2023). Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*.

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). 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*.

Geva, M., Schuster, R., Berant, J., & Levy, O. (2021). Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Mallen, A., Shi, W., Michael, J., D'Arcy, S., Wiegreffe, A., & Hajishirzi, H. (2023). When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

McKenzie, I. R., Lyzhov, A., Pieler, M., Parrish, A., Mueller, A., Prabhu, A., McLean, E., Kirtland, A., Ross, A., Liu, A., Gritsevskiy, A., Wurgaft, D., Kauffman, D., Recchia, G., Liu, J., Cavanagh, J., Weiss, M., Huang, S., Droid, F., Tseng, T., Korbak, T., Shen, X., Zhang, Y., Zhou, Z., Kim, N., Bowman, S. R., & Perez, E. (2023). Inverse scaling: When bigger isn't better. *Transactions on Machine Learning Research*.

Nanda, N., & Bloom, J. (2022). TransformerLens [Software]. GitHub. https://github.com/TransformerLensOrg/TransformerLens

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., & Olah, C. (2022). In-context learning and induction heads. *Transformer*

Circuits Thread. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8).

Salas, T. (2026, January 18). Testing accessibility knowledge across Pythia model sizes. *trishasalas.com*. https://trishasalas.com/blog/research/testing-accessibility-knowledge-pythia/

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Yao, Y., Zhang, N., Xi, Z., Wang, M., Xu, Z., Deng, S., & Chen, H. (2024). Knowledge circuits in pretrained transformers. In *Advances in Neural Information Processing Systems*.

Colophon

This paper is set in Atkinson Hyperlegible Regular, a typeface designed by the Braille Institute to maximize legibility for readers with low vision through distinctive letterforms and numerals that reduce character misrecognition. Choosing it for a paper about accessibility is not incidental.

The PDF was produced using Pandoc with LuaLaTeX. A custom template.tex enables \DocumentMetadata tagging, producing a document that conforms to both PDF/UA-2 (universal accessibility) and PDF/A-4f (archival) standards. This ensures compatibility with screen readers and other assistive technologies. One known limitation: Pandoc's PDF tagging pipeline does not distinguish table header cells from data cells, so all table cells are tagged as `<TD>` rather than the expected `<TH>` for header rows. This is a tooling constraint, not an oversight.

This paper was written in collaboration with Claude (Anthropic). The research, methodology, analysis, and conclusions are the author's own.