Correct Definitions, Failed Applications: Accessibility Concepts in Language Models*

Trisha Salas *trisha@trishasalas.com*

Abstract

A language model can state what an accessibility concept means and still fail to apply it. We measure this declarative-evaluative gap in thirteen models from three families: Pythia, GPT-2, and OLMo 2. In the original behavioral batteries, the gap opens early and persists across families. At Pythia-12B, it closes because declarative accuracy falls. Three concepts answered correctly at smaller scales regress to incorrect at 12B. Convergence by decay is not mastery. In a frozen same-concept validation of eight concepts, each represented by one violation and one conformant item, none of 96 confirmatory model-concept cells passes both application items. Eleven cells pass one item; a development-exposed pilot, reported separately, passes one pair. The 96 confirmatory cells include 35 with correct definitions. We test two possible explanations for the gap. The global maximum of attention between a compound's constituents has weak associations with accuracy that differ in direction across families. When the measure is restricted to late-layer, value-weighted binding, a consistent pattern appears: rarer compounds receive stronger distributed value-weighted writes at all thirteen model-scale points ($p = 0.0003$). This is the opposite of what we would expect if stronger binding indicated knowledge. The pattern is consistent with processing difficulty, although the experiments do not identify what computation the stronger writes perform. Corpus frequency is the strongest measured predictor of accuracy across 49 compounds (Spearman $\rho = 0.52$–0.59 across families), but it leaves most compound-to-compound variation unexplained. The same concept can be available to definition and unreachable in application, even though its corpus frequency has not changed. Frequency predicts availability. Task form constrains reachability.

Introduction

Give Pythia-12B three prompts with the same structure, and watch it complete them three different ways. An image without alt text is not accessible because "it has no text alternative" — correct. A website without screen reader support is not accessible because "it does not have a text alternative" — fluent, confident, and wrong in a way most readers outside this domain would miss. A long navigation menu without a skip link is not accessible because "of the following error: The page you are trying to reach is not available in the current context" — a fluent 404 template, confidently wrong. The compounds differ sharply in corpus frequency. Alt text appears 23,306 times in the model's training corpus; screen reader, 32,100; skip link, 662. Frequency explains the worst failure: skip link is rare. It does not explain the other: screen reader appears more often than alt text and still fails. This paper asks what corpus frequency predicts, how far that prediction reaches inside the model, and what it does not explain.

*Generative AI tools assisted with code development, analysis verification, and manuscript preparation. The research, methodology, analysis, and conclusions are the author's own, and the author is responsible for the final content.

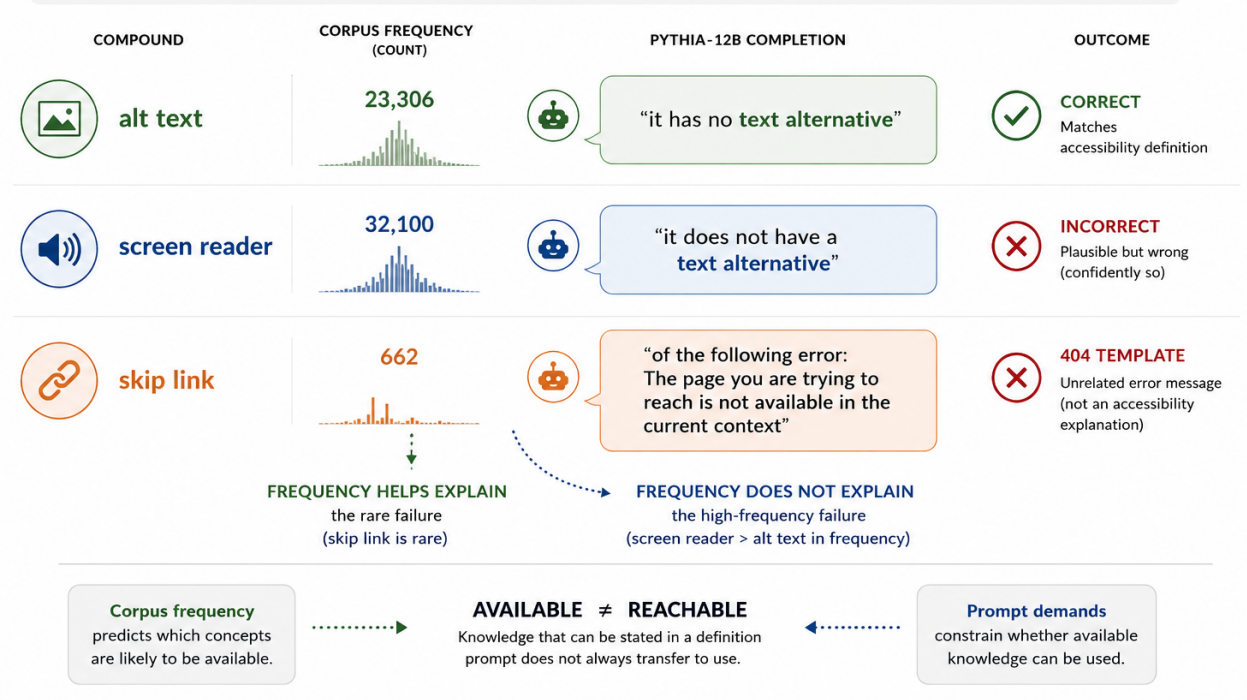

Figure 1. Identical prompt structure, divergent behavior. Pile bigram counts and Pythia-12B completions for three accessibility compounds. The rarest compound, skip link, elicits an unrelated 404 template. The two more frequent compounds differ by only 1.4× but split: alt text is completed correctly, whereas screen reader is completed fluently and incorrectly.

Digital accessibility is a good test domain for these questions for reasons unrelated to accessibility itself. Its concepts are concrete enough to code deterministically and rare enough in web text to reveal frequency effects. The phenomenon under study concerns words and their documentation density. Accessibility is where we measured it, not what it is about.

Across six Pythia scales, four GPT-2 scales, and three OLMo 2 scales, models that correctly define accessibility concepts often do not apply them correctly. In the original behavioral batteries, evaluative accuracy lags behind declarative accuracy at nearly every scale (Figure 2). In Pythia, it closes at 12B — but only because declarative accuracy regresses to meet evaluative failure. A frozen same-concept validation yields zero passing violation-and-conformant pairs across 96 confirmatory model-concept cells. Eleven cells pass one item, and a development-exposed pilot passes one pair. The confirmatory set includes 35 cells with correct definitions, so the application failures also occur for concepts the models can define.

Corpus frequency explains much of the failure structure. Across 49 compounds, frequency correlates with accuracy at Spearman $\rho = 0.59$ in Pythia, 0.52 in GPT-2, and 0.58 in OLMo 2. But it leaves most compound-to-compound variation unexplained. It predicts which concepts are likely to be available; it does not explain why knowledge that appears in a definition prompt remains unreachable in an application task.

The obvious mechanistic candidate is attention between the two words of the compound. If the model does not bind "screen" to "reader," perhaps it cannot retrieve the compound meaning. Globally maximal constituent-to-constituent attention does not support that account. Late-layer value-weighted binding reveals a different pattern: rarer compounds receive stronger value-weighted writes at every model-scale point. The frequency association is distributed across multiple late-layer heads and appears in the upper tail of value-weighted writes. This pattern is consistent with processing difficulty, but does not identify the computation performed by those writes.

The study combines a frozen same-concept violation/conformant validation, corpus-frequency analyses across three model families, and late-layer binding measurements that include value-write magnitude and residual normalization. A registered held-out intervention tests whether removing frequency-selected heads perturbs

rarer compounds more strongly. These analyses distinguish the relationship between frequency and declarative accuracy from success on application tasks.

The resulting account is deliberately narrower than a mechanism of understanding. Frequency predicts availability, while prompt demands constrain whether available knowledge can be used. The experiments that follow measure the behavioral gap, test the binding hypothesis, and separate what corpus exposure explains from what it leaves unresolved.

Related Work

Several lines of work begin from the same observation: a model can contain information that its output does not reveal. Dai et al. (2026) call this the Knowing-Using Gap. In their fine-tuning experiments, models memorize new facts before they can use them in downstream reasoning, and the failure is consistent with knowledge that does not reach the layers where it would affect the task. Basu et al. (2026) find a similar separation in clinical triage. Linear probes recover hazard information much more reliably than the model acts on it, and four mechanistic interventions leave most errors uncorrected. Jiang et al. (2020) show that failure under one prompt does not establish absent knowledge, because prompt form changes what is retrieved. Burns et al. (2023) recover latent knowledge from internal activations that the model's outputs do not state. We study the separation behaviorally, using concepts whose definitions and applications can be evaluated directly: availability here is demonstrated by output, not recovered by a probe.

Corpus frequency is already known to predict what language models learn. Kandpal et al. (2023) show that factual accuracy increases with the number of relevant documents in training. Merullo et al. (2024) connect corpus frequency to the formation and stability of linear representations. Our frequency result is consistent with this work. The difference is the dependent variable: we also ask whether compound frequency predicts what happens between constituent tokens during inference.

Attention weights alone are an incomplete account of a model's use of information. Jain & Wallace (2019) and Serrano & Smith (2019) show that attention magnitude can mislead as an indicator of importance. Kobayashi et al. (2020) incorporate the norm of transformed value vectors, while Ferrando et al. (2022) account for residual connections and layer normalization when aggregating token interactions. Head-pruning results also show that removing individual heads need not impair performance (Michel et al., 2019). We therefore distinguish raw attention, value-weighted writes, and the effects of intervention. Prior work on compounds and multiword expressions further motivates examining where these interactions occur across depth (Miletić & Schulte im Walde, 2023; 2024; Vig & Belinkov, 2019).

Salas's January 2026 report (Salas, 2026b) examined accessibility definitions, evaluative failures, and attention between compound constituents. A subsequent preprint developed the analysis of binding across depth and the declarative-evaluative gap (Salas, 2026a). Tran (2026) studies checkpoint dynamics and head ablations using EB*, which takes the maximum across layers of the strongest head's within-term attention above its layer mean. The value-weighted measure used here includes projected value-vector magnitude and target residual normalization.

Voita et al. (2019) identify a specialized rare-word head in an encoder model. That result provides a useful point of comparison for the frequency association studied here. In decoder-only models, we observe an association distributed across multiple late-layer heads and appearing in the upper tail of value-weighted writes.

Methods

The experiments draw from a shared model set and version-controlled stimulus inventory. This section defines the common environment. Each results section introduces the measure specific to that analysis.

Models. We study six Pythia scales (160M, 410M, 1B, 2.8B, 6.9B, and 12B; Biderman et al., 2023), four GPT-2 scales (124M, 355M, 774M, and 1.5B; Radford et al., 2019), and three OLMo 2 scales (1B, 7B, and 13B; Team OLMo et al., 2025). All are base models decoded greedily with sampling disabled. OLMo checkpoints are pinned to explicit revisions. The pinned reruns reproduce the saved elicitation outputs byte for byte.

Stimuli. The full inventory contains 49 two-word accessibility compounds drawn from WCAG 2.2 and related W3C terminology (W3C, 2024). In the Pile index, counts range from 3 occurrences for `accessible_authentication` to 32,100 for `screen_reader`. The frequency and binding analyses use all 49 compounds. The original behavioral comparison uses ten declarative concepts and five evaluative concepts. A separate paired battery uses the eight declarative concepts for which a violation and a conformant application item could both be constructed without naming the target concept in the prompt.

Behavioral measures. Declarative prompts ask the model to state what a concept means. Evaluative prompts ask it to judge an accessibility problem in an example. Responses are coded as correct, partial, or incorrect using deterministic, version-controlled rules. In the original gap tables, these categories receive scores of 1, 0.5, and 0. The reported gap is the difference between the mean score on the ten-concept declarative battery and the mean score on the five-concept evaluative battery at each model scale. The two original batteries are not concept-matched, so this estimate is a difference between paradigm means rather than a within-concept comparison.

Paired validation. The paired battery contains one violation and one conformant item for each of eight concepts. Both items use the same neutral answer frame. A concept passes only when the violation is coded correct and the conformant item is coded correct, preventing an always-fail or always-pass response strategy from satisfying the measure. The primary paired gap cell requires a correct declarative response and a failed evaluative pair for the same concept in the same model. Pythia-2.8B is reported separately because its behavior informed development before the battery was frozen; the other twelve models form the confirmatory set. All sixteen prompts, coding rules, outcome branches, and the analysis were frozen before confirmatory output was generated.

Corpus frequency. Bigram counts come from Infini-gram (Liu et al., 2024) and are frozen before analysis. Pythia uses counts from The Pile, its training corpus. GPT-2 uses the same Pile counts as a cross-corpus proxy because WebText is not publicly available. For declarative accuracy, OLMo uses counts from OLMo-Mix-1124, matching its training data. The registered value-weighted binding analysis uses the same frozen Pile-frequency table for all thirteen models, treating it as a cross-corpus reference for GPT-2 and OLMo. The primary frequency analysis uses one observation per compound: log bigram frequency is correlated with mean strict declarative accuracy across scales, where correct receives 1 and partial or incorrect receives 0. Because Spearman correlation depends on ranks, the base of the logarithm does not affect the result.

Binding measures. Raw binding is the attention paid by the later constituent token to the earlier constituent token, using the last subtoken when a constituent splits. The global measure takes the maximum over all layers and heads. Value-weighted binding multiplies this attention weight by the norm of the source token's head-specific value vector after the output projection, then divides by the target token's pre-attention residual norm (registered as effective binding in preregistration 0003 and renamed to better reflect what is measured). The primary value-weighted binding statistic is the 95th percentile across heads in the final third of the model. Natural and uniform prompt conditions are analyzed separately.

Held-out ablation. In Pythia-2.8B under natural prompts, we selected the five late-layer heads with the most negative frequency–value-weighted-binding correlations on 25 compounds and tested them on 24 held-out compounds. The split used seed 20260813. We jointly zeroed the selected heads' outputs at the later constituent's last subtoken and measured KL divergence from the baseline next-token distribution at the final prompt position. The registered prediction was that rarer compounds would show greater perturbation. We tested this negative frequency–KL association using 10,000 permutations and a one-count correction. Controls comprised an empty intervention, 100 random five-head sets excluding the selected heads, and all heads in the penultimate layer.

Aggregate permutation test. The cross-model binding result is evaluated with a preregistered shared-label permutation test. For each of the thirteen model-scale points, Spearman ρ is computed between log bigram frequency and the compound's 95th-percentile late-layer measure. The aggregate statistic is the mean of the Fisher-transformed per-model correlations. In each of 10,000 iterations, the frequency labels are shuffled across the 49 compounds and the same shuffled labels are applied to every model, preserving the dependence created by testing identical compounds across related architectures. The one-sided p-value is the proportion of permuted aggregates at least as extreme as the observed value, with a one-count correction in numerator and

denominator (random seed 20260813). The test and its interpretation branches were frozen by preregistration before any value-weighted binding result was inspected.

Measurement validity. All reported generation results use the model's true forward pass, and vocabulary projections apply the realized final normalization. A preregistered validation gate rejected an earlier shortcut projection that omitted the final LayerNorm and diverged from the true greedy trajectory. The appendix reports that audit; its result concerns measurement validity, not the behavioral gap.

Reproducibility. Model inference and activation capture use TransformerLens 2.18.0. The OLMo adapter was validated against the native Hugging Face forward pass and matched its top prediction at every tested position. Code, prompts, frozen counts, coding rules, result tables, preregistrations, decision records, and run manifests are included in the project repository. Claims that depend on incomplete or reconstructed provenance are identified as such.

1. The Behavioral Gap

The demonstrations above were run using the same model, prompted in the same way. The model produces correct, plausible-but-wrong, and incoherent completions depending on the concept. The question is whether this pattern generalizes.

We define declarative accuracy as whether or not a model can correctly state a concept's definition. Evaluative accuracy measures whether the model can use accessibility knowledge to perform a task. We first summarize the difference between the mean scores of the original ten-concept declarative battery and five-concept evaluative battery. This pooled estimate compares paradigms, not the same concept item by item. A separate paired test below asks the stronger same-concept question. The pooled pattern is not specific to Pythia. It reproduces across thirteen models in three families: Pythia, GPT-2, and OLMo 2.

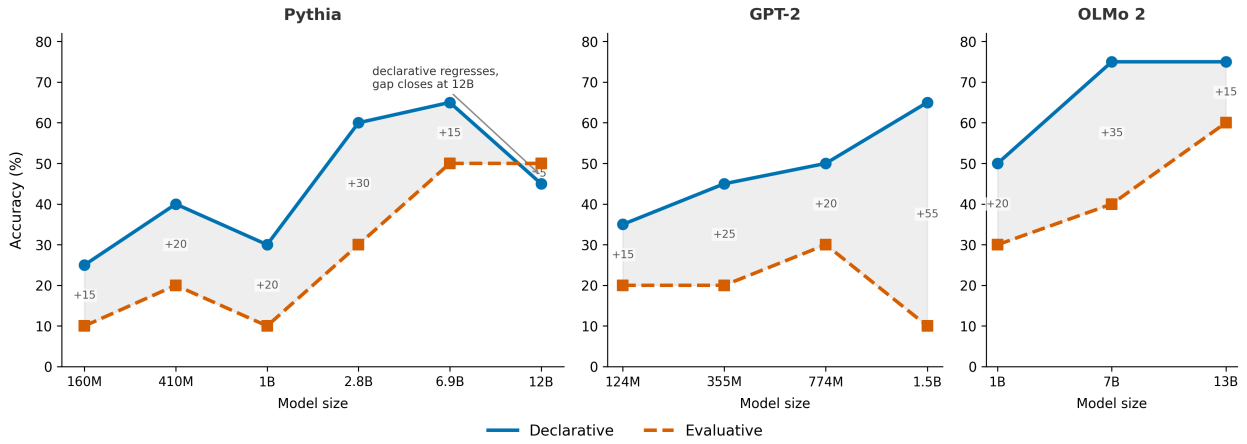

Figure 2. Declarative and evaluative accuracy across model scale. The pooled behavioral gap appears in all three families. In Pythia, the apparent closure at 12B comes from a decline in declarative accuracy rather than improved evaluative performance. Shading emphasizes the separation between the two measures.

Table A5 presents the original Pythia battery means at six scales. Declarative accuracy rises with scale, reaching a peak at 6.9B. Evaluative accuracy lags behind through 6.9B. The gap opens at 160M, widens to a maximum at 2.8B, and persists through 6.9B where the model's declarative knowledge is strongest.

At 12B, the gap closes — but not from below. Declarative accuracy drops. Three concepts answered correctly at earlier scales regress to incorrect by maximum scale: keyboard navigation, skip link, and closed captions. Table A6 shows the per-concept declarative trajectory; `cor`, `part`, and `inc` denote correct, partial, and incorrect. These are not averages obscuring noise — individual concepts flip from correct to incorrect. Each scale is a separately trained model, so the regression runs across scale, not across time within one training run. Convergence by decay is not mastery.

The pooled gap also appears in GPT-2 and OLMo. GPT-2 reaches a 55-point gap at 1.5B, but the evaluative battery contains only five items: one partial answer contributes ten percentage points. These small, unmatched batteries motivate the same-concept test below. Table A7 gives the family results not shown in Table A5.

The gap reproduces across three model families trained on different corpora with different tokenizers and different architectures. It also differs in character across families. In Pythia, incorrect responses remain high-entropy; in GPT-2, confidence in errors grows with scale. At OLMo 7B and 13B, incorrect accessibility responses have lower entropy than correct bicycle-control responses, although that reversal does not hold against correct accessibility responses. Figure 2 shows the accuracy gap across scale.

A same-concept test

The pooled comparison leaves an obvious objection: the declarative and evaluative batteries contain different concepts. We therefore froze a second battery covering eight of the declarative concepts. Each concept receives two evaluative items with the same neutral answer frame: one example contains the accessibility violation, and one conformant example removes it. A concept passes only if the model gets both polarities right. Always finding a problem and never finding one both fail.

The result is harsher than the pooled gap. Across twelve untouched confirmatory models, none passes a single evaluative pair. The test contains 96 model-by-concept cells. Eighty-five pass neither item, seven pass only the conformant item, four pass only the violation item, and zero pass both. The development-exposed pilot passes one pair, showing that the instrument can yield a pass, although its development history and the pair's prose format limit that comparison. The 96 confirmatory cells include 35 with correct declarative responses: application failure therefore also occurs for concepts the models can define. Table 1 separates those results from the development-exposed pilot; Figure 3 shows the per-model collapse. The zero-pair result is specific to this instrument: eight concepts and these sixteen application items, repeated across the twelve confirmatory models.

Table 1: Frozen paired-battery outcomes, separating the twelve confirmatory models from the development-exposed pilot.

Set	Model-concept cells	Declarative passes	Evaluative pair passes	Declarative pass / evaluative fail
Confirmatory models	96	35	0	35
Pythia-2.8B pilot	8	5	1	4

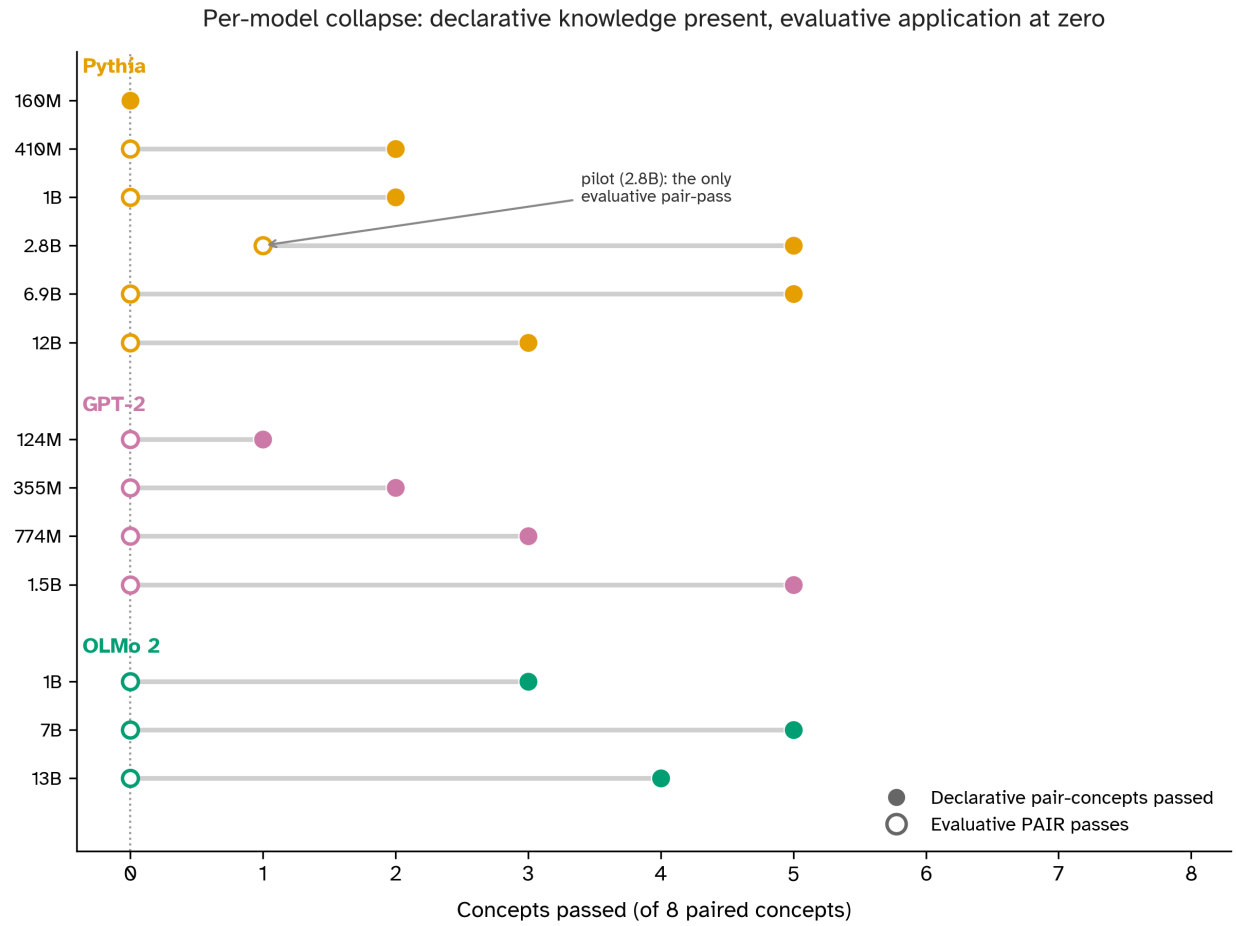

Figure 3. Per-model paired-battery collapse. For each model, the filled marker gives declarative pair-concepts passed (of eight) and the hollow marker gives evaluative pair passes. Every confirmatory model reaches zero evaluative pair passes regardless of declarative knowledge; only the development-exposed Pythia-2.8B pilot clears one.

Because no confirmatory cell passes a pair, the 35 failures among correct-definition cells follow from the overall zero-pair result; they are not a separate conditional effect. Their value is the same-concept comparison: a correct definition does not ensure success on this application pair. Pythia-160M contributes no correct-definition cells among these eight concepts.

Pythia-2.8B is reported separately because its behavior informed instrument development before the battery was frozen. It produces the only pair pass, on skip link. It classifies the violation as inaccessible and the conformant example as accessible, then repeats much of the prompt. Skip link is also the battery's only prose-described concept; the other seven require interpreting HTML or CSS. The result is consistent with framing-dependent reachability, but it does not separate framing from concept.

The gap is real, cross-family, and persistent. The original result does not reduce to a single confidence profile or a single failure mode, and the paired result does not reduce to a difference between concept inventories. A single corpus statistic — the bigram frequency of each compound in the training data — predicts which concepts take which path. The next two sections ask how far that prediction reaches and what it does not explain.

2. Binding tracks rarity, not knowledge

If the behavioral gap comes from a failure to combine the two words in a compound, binding strength should predict accuracy. Compounds with stronger constituent binding should be the ones the model gets right.

The global measure does not support that account. We take the maximum attention from the later constituent token to the earlier one across every layer and head. We use Spearman correlation because accuracy is ordered categorical data and the question concerns a monotonic association. Across compound-scale observations, Spearman ρ is -0.059 in GPT-2, -0.126 in Pythia, and 0.108 in OLMo. These weak associations differ in direction across families; Pythia shows a weak negative relationship. They do not support the proposed positive relationship between maximum binding and accuracy. These are descriptive pooled correlations: the same compounds recur across scales, so the observations are not independent. Figure 4 shows the saturation: a dense band near 1.0 at every accuracy level.

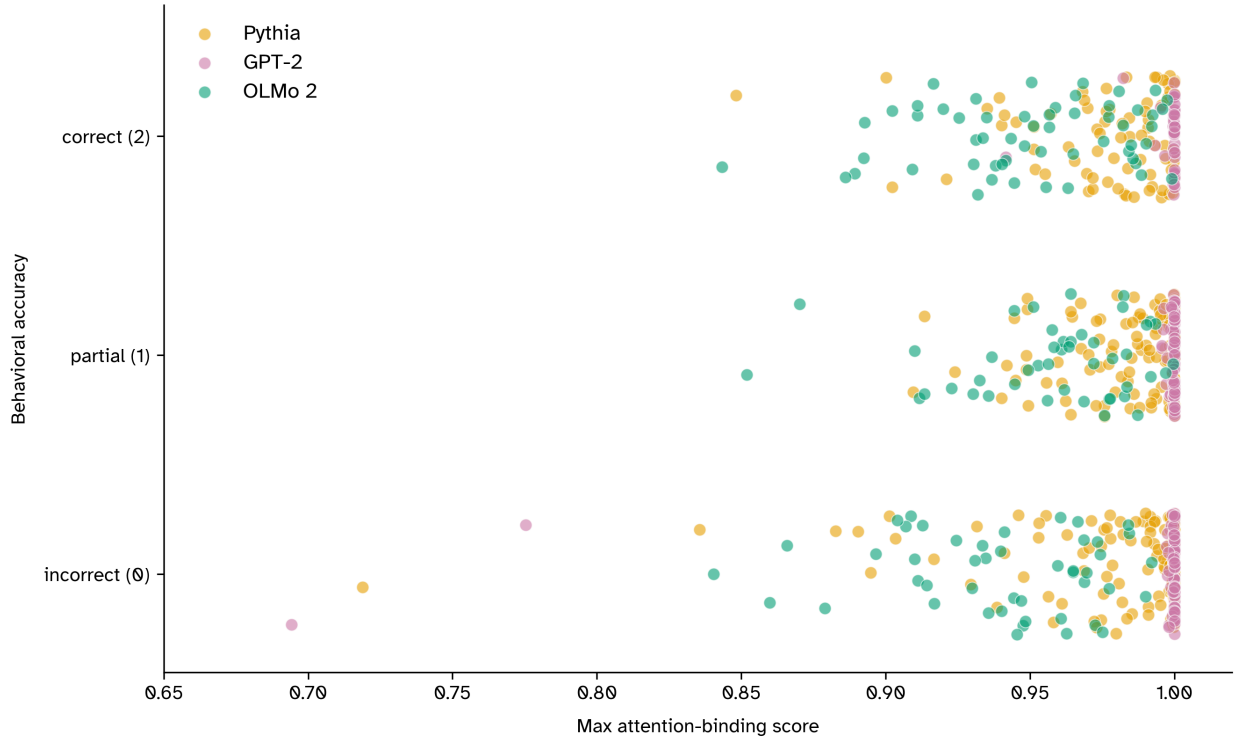

Figure 4. Maximum attention binding against behavioral accuracy, one point per compound per scale. Binding saturates near 1.0 regardless of accuracy; vertical jitter within bands is cosmetic. Per-family Spearman correlations are reported in the text.

The measure is also close to its ceiling. In GPT-2, 98% of compound pairs have a maximum attention score of at least 0.99. This concentration near the ceiling limits the measure's ability to distinguish compounds. The weak, mixed-direction associations do not support using whole-network maximum attention as a general indicator of knowledge. They do not rule out other forms of constituent interaction.

We next restrict the analysis to the final third of each model and include the magnitude of the value write. For each head, value-weighted binding is the attention weight multiplied by the source token's value-write norm and divided by the target token's residual norm. We then take the 95th percentile across late-layer heads. This measure is negatively correlated with corpus frequency at all thirteen model-scale points. A shared-label aggregate permutation test (see Methods) gives $p = 0.0003$. The direction repeats under uniform prompts at all thirteen model-scale points. The all-negative pattern does not hold across every alternative aggregation (Appendix C). Table A3 shows the primary natural-prompt correlations, and Figure 5 plots them with bootstrap confidence intervals.

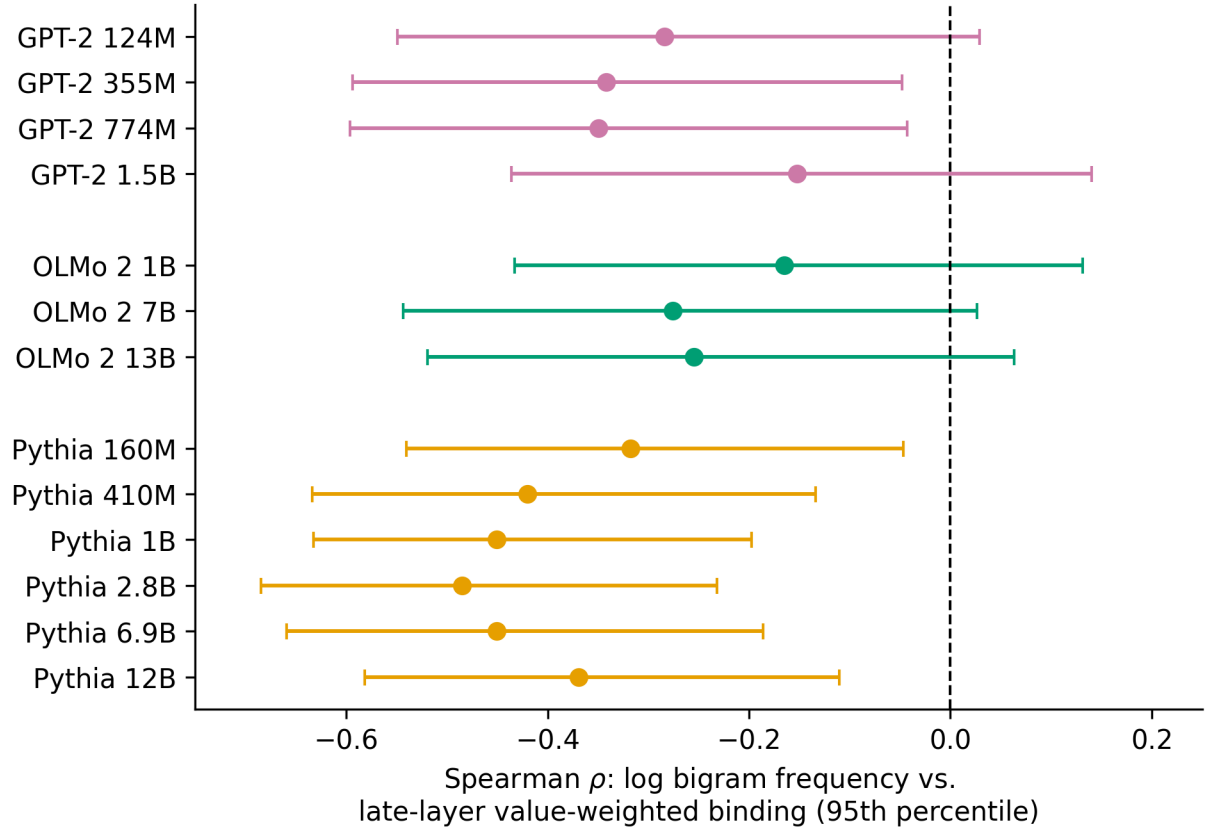

Figure 5. Spearman correlation between log bigram frequency and 95th-percentile late-layer value-weighted binding, one point per model, with bootstrap confidence intervals. All thirteen are negative; aggregate significance comes from the preregistered shared-label permutation test (Methods).

Rarer compounds receive stronger late-layer value-weighted writes. This is the opposite of what a knowledge-bearing account predicts. Common compounds are more likely to be answered correctly, but they receive less of the late-layer interaction measured here. The result is more consistent with unresolved or difficult processing than with successful composition.

The frequency association is distributed across multiple late-layer heads. It appears in the upper tail of late-layer value-weighted writes. Voita et al. (2019) identify a specialized rare-word head in an encoder model; the present analysis examines associations across heads in decoder-only models. Here the response is a population pattern, not evidence of a stable circuit.

The registered held-out ablation did not support the predicted greater perturbation for rarer compounds. Across 24 test compounds, frequency and selected-set KL had Spearman $\rho = 0.037$ (one-sided permutation $p = 0.563$). The selected correlation was not unusually negative relative to the 100 random sets (empirical one-sided tail probability 0.248). Mean KL was 2.13×10^{-5}, and no highest-probability token changed. Both control gates passed: the empty intervention produced zero KL for every compound, and the penultimate-layer control exceeded the registered KL threshold for all 24 compounds. These results do not show that the selected heads have no effect; they show that the measured perturbation did not increase with rarity as predicted.

Late value-weighted binding is real, distributed, and frequency-sensitive. Processing difficulty remains a possible interpretation, not an identified computation. The present evidence does not establish binding's causal role.

3. Frequency is the floor, not the ceiling

The binding result shows that corpus frequency is related to what happens between a compound's constituents during inference. This section asks how much frequency explains at the behavioral level.

Bigram frequency predicts declarative accuracy in all three model families. Spearman ρ is 0.59 in Pythia, 0.52 in GPT-2, and 0.58 in OLMo, with $p < 0.001$ and $n = 49$ compounds in each family. The Pythia count comes from its training corpus. GPT-2 reproduces the relationship using Pile counts as a cross-corpus proxy, and OLMo reproduces it using OLMo-Mix counts. The OLMo result is unchanged after pinning and rerunning the model checkpoints. Figure 6 shows all 49 compounds in each family.

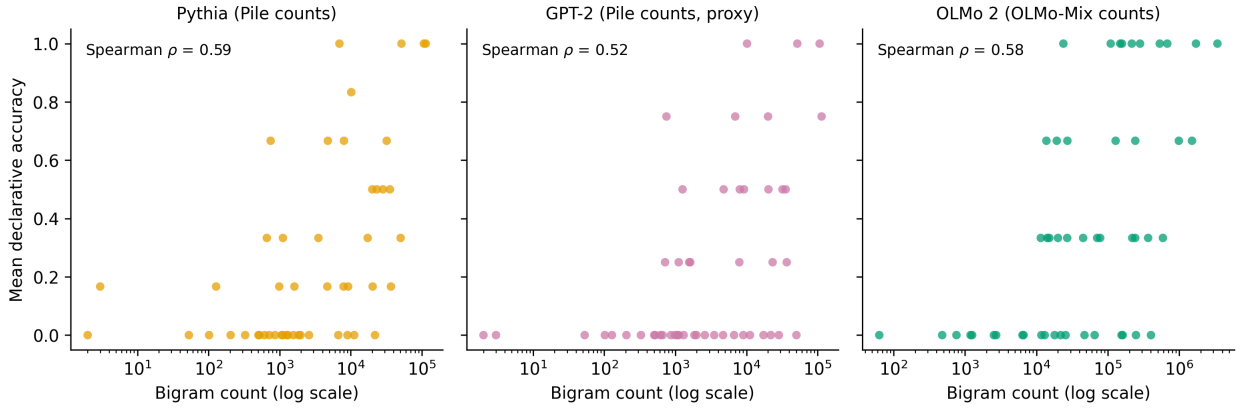

Figure 6. Log bigram count against mean declarative accuracy, 49 compounds per family. Frequency sets the floor; the vertical spread at similar counts is what it leaves unexplained.

Frequency is the strongest measured predictor in this study, but it is not a complete explanation. Correlations of $\rho = 0.52$–0.59 are substantial but leave much unexplained: compounds with similar counts follow different trajectories, and the same concept can succeed or fail when the task changes.

One Pythia-2.8B diagnostic makes the framing problem concrete. Asked, `What accessibility attribute is missing from this HTML: <img src='photo.jpg'>?`, the model produces two newlines and starts a new paragraph. Given the cloze prompt, `<img src='photo.jpg'> is missing the attribute`, it produces `alt`. The model, weights, concept, and corpus count are unchanged. Only the answer position changes.

The paired battery extends the point beyond one diagnostic: the same-concept failures reported in Section 1 include concepts with correct definitions. The compound and its corpus frequency are fixed; what changes is the form in which the model must use it. This does not isolate one prompt feature—the declarative and evaluative tasks differ in more than wording—but it rules out frequency as a sufficient explanation of whether available knowledge will be reached by a particular task.

The frequency and binding results point in opposite behavioral directions. More frequent compounds are answered more accurately. Less frequent compounds receive stronger late-layer value-weighted writes. The additional interaction does not compensate enough to reverse the accuracy relationship.

The contribution is the joint test of corpus frequency, late-layer value-weighted binding, and same-concept application failure across model families. Corpus frequency predicts declarative accuracy, while rarity predicts stronger late value-weighted writes. The paired battery shows that correct definitions do not ensure success on the tested application pairs. The registered Pythia-2.8B ablation did not support the predicted greater perturbation for rarer compounds.

Frequency sets a floor. More exposure makes knowledge more likely to appear, but it does not guarantee that the knowledge will be available in the form a task requires. What lies above that floor remains unresolved.

Discussion

The results establish a behavioral mismatch and rule out two simple explanations for it. The original pooled gap reproduces across three model families. In the paired test, none of 96 confirmatory cells passes both application items, including 35 cells with correct definitions. The item-level outcomes and separate pilot pass described in Section 1 bound this result to the frozen instrument.

The binding result is consistent with processing difficulty, although that interpretation is not directly tested. Rarer compounds receive stronger value-weighted writes in the final third of every model tested. The direction repeats under uniform prompts. Yet the compounds receiving those writes are also the ones answered less accurately. Stronger late interaction is not evidence of stronger knowledge.

The upper-tail statistic keeps this interpretation at the population level. The distributed frequency association describes the observed pattern across late-layer heads; it does not establish circuit organization. The registered Pythia-2.8B intervention did not reveal greater sensitivity for rarer compounds. Its scope is limited to five selected heads at one token position and a next-token distributional outcome; the causal role of the wider late-layer response remains unresolved. We looked for wiring and found ink.

Frequency has a different role. It is the strongest predictor of declarative accuracy across all three families, with correlations between 0.52 and 0.59. But frequency is constant for a concept, while the model's response changes with the task. The paired result holds the concept fixed across declarative and evaluative measures, but it does not isolate which feature of the task blocks access. Availability and reachability are useful names for the distinction, not identified internal mechanisms. We use *unreachable* in the operational sense throughout: unreached under the task forms tested here.

Pythia-2.8B provides the one exception to the strict pair-level result. Reported separately as the development-exposed pilot, it passes skip link, the only concept described entirely in prose. The other seven pairs require interpreting HTML or CSS. That success is consistent with framing-dependent reachability, but format and concept are confounded. It does not establish a general framing effect or special skip-link knowledge.

Related work suggests that a separation between stored information and successful use is not unique to accessibility. Dai et al. (2026) observe memorized facts that fail to reach downstream reasoning, and Basu et al. (2026) recover clinical information internally that interventions do not reliably turn into correct decisions. Our results do not establish a general property of all autoregressive transformers. They add a cross-family behavioral case with a frequency-sensitive internal correlate.

Accessibility also raises a training-data hypothesis that the present study does not test. Concepts such as alt text and skip links are described in standards and tutorials, while ordinary code examples may fail to implement them. If descriptive and applied examples are distributed differently, increasing corpus volume could preserve the imbalance rather than repair it. Testing that account requires measuring the ratio directly. Consistent with this, six structural concepts that never reach correct declarative form in Pythia or GPT-2 first appear in OLMo-2 at 7B — a family trained on Dolma rather than the Pile — a hint that corpus composition, not volume alone, shifts which concepts become available ($n = 1$ per concept; suggestive only).

Scale alone is not an intervention supported by these results. A direct test would change the composition of exposure while holding its volume fixed, then ask whether evaluative accuracy improves and late value-weighted binding weakens. Until that experiment is run, compensatory binding is a diagnostic candidate, not an intervention target.

Conclusion

Across the tested base models, correct accessibility definitions did not ensure success on matched application pairs. Corpus frequency predicted declarative accuracy, while rarer compounds received stronger late-layer value-weighted writes. The registered intervention did not support greater output perturbation for rarer compounds. Together, these results separate definition-level success, observed constituent interactions, and the ability to apply a concept in the tested tasks.

Limitations

The original behavioral gap uses two small, non-matched batteries: ten declarative concepts and five evaluative concepts. Its pooled percentage is a difference between paradigm means. The paired battery addresses the same-concept objection, but it remains small: eight concepts, each represented by one violation and one conformant item. The result establishes failure under this instrument, not under every possible application prompt.

The paired pass rule is deliberately strict. Both polarities must be correct, so a model that gets one side right receives no concept-level pass. Across the confirmatory set, 85 of 96 cells pass neither item, seven pass only the conformant item, and four pass only the violation item. The zero pair count should therefore be reported with its item-level distribution rather than treated as a complete account of behavior.

The paired coder is deterministic and was frozen before confirmatory generation, but keyword rules do not resolve every malformed continuation. Manual audit found two conformant items labeled correct by incidental string matches. Neither has a correct violation partner, so adjudicating them cannot change the primary zero-pair result. Twenty-one confirmatory outputs also echo their prompt exactly. Removing exact echoes changes no confirmatory item label or pair outcome.

Pythia-2.8B's sole pair pass also needs a narrow interpretation. It clearly classifies both skip-link polarities correctly, but then repeats much of the prompt, so the output does not establish an independently generated explanation. Skip link is the only prose-described pair; concept and representation format are therefore confounded in the one successful cell.

The full frequency and binding analyses contain 49 accessibility compounds. Whether the same relationships hold in other domains is unknown. Security and privacy provide plausible comparisons because they also contain norms that may be described more often than they are followed, but that corpus structure has not been measured here.

The model set is broad across families but small by current standards. The largest model has 13 billion parameters. The results do not establish what happens in larger base models, instruction-tuned models, or closed models trained on undisclosed corpora.

The value-weighted binding measure includes attention weight, value-write magnitude, and target residual norm. It does not follow information through every residual, normalization, and downstream path. ALTI or another full attribution method could test whether the same frequency relationship survives a more complete decomposition.

Frequency provenance also differs by family. Pythia uses counts from its training corpus, and OLMo uses OLMo-Mix counts. GPT-2 uses Pile counts because WebText is unavailable, making that family a cross-corpus replication rather than an exact training-frequency analysis.

Finally, the compensatory interpretation is an inference from direction and distribution. The experiments show that rare compounds receive stronger late value-weighted writes. They do not show what computation the additional write performs or whether it causally affects the model's prediction.

Broader Impact

This study motivates a broader approach to understanding and improving how language models use learned knowledge. Distinguishing behavioral availability from task-dependent reachability makes it possible to ask which training experiences support application, beyond those that support correct descriptions. It also encourages evaluating proposed internal mechanisms against both behavioral outcomes and interventions, rather than interpreting stronger activity as greater competence. These distinctions can guide research on training-data composition, transfer across task forms, and concept-level changes across scale. Accessibility provides the empirical setting here; whether the same patterns hold in other domains remains an open question. The potential benefit is a more precise basis for developing models whose demonstrated knowledge translates into reliable use.

References

Sanjay Basu, Sadiq Y. Patel, Parth Sheth, Bhairavi Muralidharan, Namrata Elamaran, Aakriti Kinra, John Morgan, and Rajaie Batniji. Interpretability without actionability: Mechanistic methods cannot correct language model errors despite near-perfect internal representations. *arXiv preprint arXiv:2603.18353*, 2026. URL https://arxiv.org/abs/2603.18353.

Stella Biderman, Hailey Schoelkopf, Quentin G. 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. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *Proceedings of the Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ETKGuby0hcs.

Lu Dai, Ziyang Rao, Yili Wang, Hanqing Wang, Hao Liu, and Hui Xiong. Towards mechanistically understanding why memorized knowledge fails to generalize in large language model finetuning. *arXiv preprint arXiv:2607.08393*, 2026. URL https://arxiv.org/abs/2607.08393.

Javier Ferrando, Gerard I. Gállego, and Marta R. Costa-jussà. Measuring the mixing of contextual information in the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 8698–8714. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.595. URL https://aclanthology.org/2022.emnlp-main.595/.

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. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357/.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020. URL https://aclanthology.org/2020.tacl-1.28/.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 15696–15707. PMLR, 2023. URL https://proceedings.mlr.press/v202/kandpal23a.html.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 7057–7075. Association for Computational Linguistics, 2020. URL https://aclanthology.org/2020.emnlp-main.574/.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. In *Proceedings of the First Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2401.17377.

Jack Merullo, Sarah Wiegreffe, and Yanai Elazar. The mutual relationship between corpus frequency and linear representations in language models. In *NeurIPS Workshop on Attributing Model Behavior at Scale*, 2024. URL https://neurips.cc/virtual/2024/105361.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.neurips.cc/paper_files/paper/2019/hash/2c601ad9d2ff9bc8b282670cdd54f69f-Abstract.html.

Filip Miletić and Sabine Schulte im Walde. A systematic search for compound semantics in pretrained BERT architectures. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 1499–1512. Association for Computational Linguistics, 2023. URL https://aclanthology.org/2023.eacl-main.110/.

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. URL https://aclanthology.org/2024.tacl-1.33/.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Trisha Salas. Accessibility concept emergence in the Pythia suite: Thresholds, binding, and the declarative-evaluative gap, May 2026a. URL https://doi.org/10.5281/zenodo.20360787. Preprint. Zenodo. Published May 23, 2026. Version DOI: 10.5281/zenodo.20360788.

Trisha Salas. Testing accessibility knowledge across Pythia model sizes, January 2026b. URL https://trishasalas.com/writing/notes/testing-accessibility-knowledge-pythia/. Research blog post.

Sofia Serrano and Noah A. Smith. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2931–2951. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1282. URL https://aclanthology.org/P19-1282/.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, et al. 2 OLMo 2 furious. In *Proceedings of the Second Conference on Language Modeling*, 2025. URL https://arxiv.org/abs/2501.00656.

Khanh-Dung Tran. Attention-head binding as a term-conditioned mechanistic marker of accessibility concept emergence in language models. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=QG7mfCy9mu.

Jesse Vig and Yonatan Belinkov. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP*, pp. 63–76. Association for Computational Linguistics, 2019. URL https://aclanthology.org/W19-4808/.

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. Association for Computational Linguistics, 2019. URL https://aclanthology.org/P19-1580/.

W3C. Web content accessibility guidelines (WCAG) 2.2, 2024. URL https://www.w3.org/TR/WCAG22/. W3C Recommendation.

Appendix

A. Models and generation

All thirteen models are base models. Table A1 reports architecture dimensions from the saved paired-run manifests. Model names follow the identifiers used by the loaders: Pythia models use the `EleutherAI/` namespace and OLMo 2 models use `allenai/`. GPT-2 model names are the TransformerLens loader identifiers.

Table A1: Architecture dimensions recorded for the paired evaluative runs. Vocabulary sizes describe the model configuration.

Model	Layers	Heads/layer	Hidden size	Model vocabulary
gpt2	12	12	768	50257
gpt2-large	36	20	1280	50257
gpt2-medium	24	16	1024	50257
gpt2-xl	48	25	1600	50257
OLMo-2-0425-1B	16	16	2048	100352
OLMo-2-1124-13B	40	40	5120	100352
OLMo-2-1124-7B	32	32	4096	100352
pythia-12b	36	40	5120	50688
pythia-160m	12	12	768	50304
pythia-1b	16	8	2048	50304
pythia-2.8b	32	32	2560	50304
pythia-410m	24	16	1024	50304
pythia-6.9b	32	32	4096	50432

Table A2: Checkpoint provenance (12-character commit prefixes; full hashes are recorded in the manifests) for the paired evaluative runs. These records do not retrospectively pin separate earlier experiments.

Model	Requested revision	Resolved commit in paired manifest
gpt2	main	607a30d783df
gpt2-large	main	32b71b12589c
gpt2-medium	main	6dcaa7a952f7
gpt2-xl	main	15ea56dee5df
OLMo-2-0425-1B	stage1-step1907359-tokens4001B	9d3e43659f00
OLMo-2-1124-13B	stage1-step596057-tokens5001B	08d2aca2e28a
OLMo-2-1124-7B	stage1-step928646-tokens3896B	c0371f4281bf
pythia-12b	main	bb1e3e710cdf
pythia-160m	main	Not recorded
pythia-1b	main	f73d7dcc545c
pythia-2.8b	main	2a259cdd96a4
pythia-410m	main	Not recorded
pythia-6.9b	main	c0e3eee36dc4

The paired runner disables sampling and allows up to 100 new tokens. It preserves the raw continuation, including leading whitespace. The original elicitation helper supplies `temperature=0` and the per-prompt continuation limit; the current accessibility inventory specifies 100 tokens for the declarative and evaluative items. Binding and ablation measurements evaluate fixed prompts without generating continuations.

The OLMo revision labels are `stage1-step1907359-tokens4001B` (1B), `stage1-step928646-tokens3896B` (7B), and `stage1-step596057-tokens5001B` (13B). These requested labels are also recorded in the effective-binding manifests. The paired-run resolved commits are shown separately in Table A2. Earlier Pythia and GPT-2 binding runs do not record resolved model revisions; the registered Pythia-2.8B ablation likewise does not record a resolved checkpoint commit. Those omissions limit exact checkpoint reconstruction and are not filled using later runs.

The inference notebooks use TransformerLens tokenization. Binding result rows preserve the realized token strings and constituent indices. A separate tokenizer revision is not recorded in these manifests. Runtime versions and device/dtype information are retained per run rather than assumed identical across experiments.

B. Prompt populations and coding

The original pooled comparison contains ten declarative and five evaluative prompts per model. Correct, partial, and incorrect responses receive 1, 0.5, and 0 in those battery means. The corpus-frequency analysis instead uses strict declarative accuracy across scales, assigning partial and incorrect responses zero. These are different estimands.

The paired battery contains sixteen prompts: a violation and a conformant item for each of eight concepts. WCAG and ARIA acronym-expansion items are excluded because application examples would test different knowledge. The battery was frozen on August 20, 2026. Pythia-2.8B remains development-exposed and is reported separately. The twelve confirmatory models contribute 192 item responses and 96 pairs; the pilot contributes sixteen responses and eight pairs.

The frozen coder lowercases and normalizes whitespace in the complete continuation. Empty or repetitively degenerate responses are incorrect. A violation is correct when the response contains both a negative accessibility classification and a concept-specific reason; either alone is partial. An affirmative classification on a violation is incorrect. A conformant item is correct with an affirmative classification unless the rule detects a contradictory reason; a negative classification is incorrect and an unclassified response is partial. The exact concept-specific string rules are preserved in `src/paired_evaluative.py`, and the original battery coding is preserved in `src/gap_analysis.py` and `src/accuracy_coding.py`. A pair passes only when both items are correct.

The frozen mechanical labels contain two documented conformant-item false positives: OLMo-2-1B focus indicator and GPT-2 alt text. Neither has a correct paired violation, so correcting these labels cannot produce a confirmatory pair pass. Removing exact prompt copies from the 21 confirmatory continuations containing them changes no item label. The reported item counts retain the frozen labels. In the pilot skip-link pair, the continuation supplies the classification but repeats prompt language that satisfies the reason rule; an independently generated explanation is not established.

The accompanying prompt inventory reproduces the fifteen original battery prompts, the 41 additional declarative probes, all sixteen frozen paired items and their expected judgments, and all 53 binding prompts. It is an extraction of the existing inventories, not a revision of the instrument.

C. Binding measurement and coverage

The binding inventory contains 53 accessibility compounds. All thirteen models have saved natural- and uniform-prompt measurements. The frequency analysis retains the 49 compounds present in the frozen frequency table; `page_title`, `link_text`, `empty_link`, and `form_label` are excluded because that table contains no count for them. This exclusion is not a token-matching failure.

Natural prompts use the inventory text. Uniform prompts use `A {word1} {word2} is`. Matching first searches for a complete token after removing surrounding whitespace and ignoring case, then searches concatenated adjacent token strings. A split constituent is represented by its last subtoken. The earlier matched position supplies the value vector and the later position supplies the attention query and residual normalization. Unresolved constituents are reported by the producer.

For a head, the measure is its attention from the later position to the earlier position multiplied by the L2 norm of the earlier position's value vector after that head's output projection. This quantity is divided by the later position's pre-attention residual L2 norm, with a numerical lower bound of 10^{-12}. For OLMo, the source-specific write is measured before attention-branch RMS normalization. This is not a complete decomposition of information flow through all normalization and residual paths.

The late-layer set uses zero-indexed layers $l \geq \lceil 2L/3 \rceil$. The primary statistic is the 95th percentile across all heads in that set for each compound. The maximum attention, maximum weighted write, maximum normalized write, and mean of the five largest normalized writes are retained as alternative summaries.

The saved secondary summaries limit the generality of the all-negative sign pattern. Normalized maxima have negative frequency correlations in eight of thirteen models under natural prompts and ten under uniform prompts; the top-five mean has negative correlations in eleven and twelve, respectively. The primary 95th percentile is negative in all thirteen under each condition. These are descriptive sign counts, not counts of individually significant effects. All these comparisons use the final third of layers and do not test a final-quarter cutoff.

D. Statistical procedures and prompt sensitivity

The binding confidence intervals use 2,000 paired compound bootstrap resamples with seed 20260813. Each resample draws 49 rows with replacement and recomputes Spearman correlation; nonfinite estimates are omitted. The endpoints are the 2.5th and 97.5th percentiles. These are per-model intervals, not simultaneous confidence intervals.

The shared-label test uses 10,000 permutations with seed 20260813. The same compound-frequency permutation is applied to every model. The statistic is the mean Fisher-transformed correlation, and the lower-tail probability uses the one-count correction. Natural prompts have two permuted statistics at least as negative as the observed statistic, giving $3/10001 = 0.000300$; uniform prompts have three, giving $4/10001 = 0.000400$. The observed mean Fisher transforms are -0.349354 for natural prompts and -0.364767 for uniform prompts.

The saved effective-binding analysis uses `results/frequency/frequency_table.csv` for all thirteen model-scale points. This shared frequency reference must be distinguished from the family-matched corpus counts used in the declarative-accuracy analysis.

Table A3: Frequency correlations for the 95th-percentile normalized value-weighted binding measure, with 49 compounds per model and condition.

Model	Natural rho	95% bootstrap interval	Uniform rho	95% bootstrap interval
gpt2	-0.284	[-0.549, 0.029]	-0.314	[-0.554, -0.011]
gpt2-large	-0.350	[-0.596, -0.043]	-0.362	[-0.586, -0.070]
gpt2-medium	-0.342	[-0.594, -0.048]	-0.430	[-0.648, -0.153]
gpt2-xl	-0.152	[-0.436, 0.140]	-0.240	[-0.498, 0.049]
OLMo-2-0425-1B	-0.165	[-0.433, 0.131]	-0.442	[-0.645, -0.187]
OLMo-2-1124-13B	-0.254	[-0.519, 0.063]	-0.217	[-0.507, 0.094]
OLMo-2-1124-7B	-0.276	[-0.544, 0.027]	-0.347	[-0.615, -0.048]
pythia-12b	-0.369	[-0.582, -0.110]	-0.452	[-0.636, -0.227]
pythia-160m	-0.318	[-0.541, -0.047]	-0.266	[-0.523, 0.027]
pythia-1b	-0.451	[-0.633, -0.198]	-0.310	[-0.545, -0.045]
pythia-2.8b	-0.485	[-0.685, -0.232]	-0.368	[-0.594, -0.099]
pythia-410m	-0.420	[-0.634, -0.134]	-0.413	[-0.625, -0.133]
pythia-6.9b	-0.450	[-0.659, -0.186]	-0.352	[-0.577, -0.087]

The declarative-frequency token-count sensitivity uses GPT-2 tokenization as a common count measure across families and partial Spearman correlation: variables are ranked and the first-order partial-correlation formula is applied to their Pearson rank correlations. The saved family-level partial correlations controlling compound token count are 0.5807 for Pythia, 0.5092 for GPT-2, and 0.5810 for OLMo. These results concern declarative accuracy and do not establish token-count robustness for value-weighted binding.

E. Exploratory head localization

For each late-layer head, frequency is correlated with normalized value-weighted binding across all 49 compounds under natural prompts. This exploratory analysis is separate from the 25-compound selection procedure used for intervention. Head identities are not aligned across models.

Two-sided Spearman p-values receive Benjamini–Hochberg correction within each model over heads with defined correlations. Constant measurements yield undefined correlations: thirteen heads in Pythia-12B and

two in Pythia-6.9B are excluded from that denominator. Bootstrap stability uses 2,000 paired compound resamples with seed 20260905, shared across heads within each model. The detailed tables report percentile intervals, the fraction of finite resamples with negative correlation, and correlation distributions by relative layer depth. The bootstrap intervals are marginal and are not corrected for multiple comparisons. These implementation choices are recorded as exploratory audit choices.

Table A4: Exploratory head-level frequency associations. Negative signs and corrected statistical support are reported separately; these counts do not establish causal contributions.

Model	Late heads	Defined correlations	Negative correlations	Negative with BH q $< .05$
gpt2	48	48	16	0
gpt2-large	240	240	132	0
gpt2-medium	128	128	78	0
gpt2-xl	400	400	252	0
OLMo-2-0425-1B	80	80	49	6
OLMo-2-1124-13B	520	520	307	1
OLMo-2-1124-7B	320	320	192	8
pythia-12b	480	467	307	7
pythia-160m	48	48	28	0
pythia-1b	40	40	34	1
pythia-2.8b	320	320	252	5
pythia-410m	128	128	101	0
pythia-6.9b	320	318	218	6

F. Registered intervention and adherence

The registered natural-prompt Pythia-2.8B test sorts the 49 compound names, shuffles with seed 20260813, assigns the first 25 to selection and the remaining 24 to testing, and preserves the ordered split. The five selected zero-indexed (layer, head) pairs are (28, 5), (30, 6), (28, 13), (22, 25), and (30, 11). The intervention zeros their head outputs at the later constituent's last subtoken. The outcome is KL divergence from the baseline to the ablated full-vocabulary next-token distribution at the final prompt position.

The September 4 controls amendment specifies an empty intervention, 100 random five-head sets excluding the selected heads, and all 32 heads in layer 30 as the positive control. The registered run records all sets and 2,400 compound-by-random-set outcomes. The empty intervention yields zero KL on every compound. The positive control exceeds 10^{-6} for all 24 compounds. Both control gates pass.

For the selected set, frequency–KL Spearman correlation is 0.037, with one-sided permutation $p = 0.563$. Its empirical lower-tail probability among the random sets is 0.248. Mean KL is 2.13×10^{-5} and median KL is 1.62×10^{-5}; no highest-probability token changes. The registered prediction of greater perturbation for rarer compounds is not supported. This does not establish that the selected heads have no effect.

A saved-artifact audit reproduces the split, head selection, random sets, statistical tests, and all eleven manifest hashes. The full baseline and intervened vocabulary distributions were not saved, so KL cannot be independently reconstructed from the CSVs alone. Earlier six-scale exploratory ablations used a different split and control design and are excluded from the manuscript's evidence.

The paired protocol's header retains unfilled freeze fields; the separate paired-results audit identifies the freeze commit and battery hash. The reproducibility record preserves that documentary distinction rather than rewriting the original preregistration retrospectively.

G. Original pooled behavioral tables

Table A5: Pythia original-battery means. Correct, partial, and incorrect are scored 1, 0.5, and 0; gaps are percentage points.

Scale	Declarative	Evaluative	Gap
160M	25%	10%	15%
410M	40%	20%	20%
1B	30%	10%	20%
2.8B	60%	30%	30%
6.9B	65%	50%	15%
12B	45%	50%	-5%

Table A6: Pythia declarative trajectories. cor, part, and inc denote correct, partial, and incorrect.

Concept	160M	410M	1B	2.8B	6.9B	12B
ARIA	inc	inc	inc	inc	inc	inc
WCAG	inc	inc	inc	inc	cor	cor
alt text	inc	part	part	cor	cor	cor
closed captions	inc	inc	inc	cor	inc	inc
color contrast	part	cor	part	part	cor	cor
focus indicator	part	inc	inc	inc	inc	inc
keyboard navigation	part	cor	cor	cor	cor	inc
screen reader	part	part	cor	cor	cor	cor
semantic HTML	part	part	inc	part	part	part
skip link	inc	part	inc	cor	cor	inc

Table A7: GPT-2 and OLMo 2 original-battery means, using the same scoring as Table A5.

Family	Scale	Declarative	Evaluative	Gap
GPT-2	124M	35%	20%	15%
GPT-2	355M	45%	20%	25%
GPT-2	774M	50%	30%	20%
GPT-2	1.5B	65%	10%	55%
OLMo 2	1B	50%	30%	20%
OLMo 2	7B	75%	40%	35%
OLMo 2	13B	75%	60%	15%

H. Measurement-pathway validation

A decision-point analysis initially appeared to show the correct continuation for skip link losing to a higher-frequency competitor between Pythia-6.9B and Pythia-12B. The analysis projected the final residual stream directly through the unembedding matrix, omitting the model's final LayerNorm. A preregistered gate required the resulting trace to agree with the true forward pass. It did not, and the competition claim was withdrawn.

An audit of the saved rollouts found the same failure at all six Pythia scales: the shortcut left the true greedy trajectory by the eighth generated token. Under weight folding, the final LayerNorm bias becomes an effective unembedding bias, b_U, whose correlation with log Pile unigram frequency ranges from 0.664 to 0.779 across scales. The shortcut omits this frequency-ordered term; in all six observed divergences, the true pathway selected the more frequent token. This is a finding about measurement validity, not an explanation of the behavioral gap.

I. Artifact map

The reproducibility materials preserve separate roles for the stimulus inventories, producers, result tables, and audit records.

Table A8: Reproducibility artifact map.

Purpose	Repository-relative artifact
Original prompt inventory	data/accessibility.yaml
Paired prompts and expected judgments	data/evaluative_paired.yaml
Binding prompt inventory	data/binding/accessibility.yaml
Paired generation and coding	src/paired_evaluative_runner.py; src/paired_evaluative.py
Original battery analysis	src/gap_analysis.py
Binding producers	notebooks/effective-binding-{pythia,gpt2,olmo}.ipynb
Binding summaries, bootstrap, aggregate tests	notebooks/effective-binding-analysis.ipynb; results/analysis/effective_binding_*.csv
Corpus accuracy and partial correlations	src/dual_spearman.py; results/frequency/spearman_partial.csv
Registered intervention	notebooks/frequency-head-ablation-pythia.ipynb; src/qk_ov.py
Registered run outputs and contemporaneous manifest	results/ablation/frequency_heads/pythia-2.8b/pythia-2.8b-natural-registered-*
Exploratory head audit	docs/audits/2026-09-05-head-localization-*
Protocols and controls amendment	docs/preregistrations/0002-evaluative-prompts.md; 0003-frequency-and-effective-late-layer-binding.md; 0003-amendment-2026-09-04-ablation-controls.md