Priors Persist Through Suppression: A Stroop Paradigm for Lexical Override

Anonymous ACL submission

Abstract

Glossaries, technical specifications, and system prompts routinely ask language models to use familiar words in unfamiliar ways. When this works, the lexical prior persists through override rather than being replaced: it continues to operate after the local rule applies, with the rule lowering its logit rather than installing the new meaning on top. We test this with a Stroop-style paradigm—a remapping rule (*doctor* means *forest*) pitted against the query word's lexical-prior distractor (*hospital*), with matched neutral controls. Across 11 open-weight models spanning four families and 1B–9B parameters, lexical-prior strength predicts interference even after item-level controls for answer prior, frequency, tokenization, and prompt wording. Activation patching on five aligned models locates a source-position triplet (definition subject, definition target, query word) that nearly fully recovers the conflict effect (aggregate $R \in [0.92, 1.06]$); a definition-target swap shows the triplet performs binding rather than identity matching. Dissociation experiments isolate target preservation as the binding-specific signature: distractor suppression occurs under matched, swap, and item-mismatched conditions alike, whereas target logit collapse occurs only when the definition-target position is corrupted. Behavior and mechanism converge on the same channel: the lexical prior is where both interference originates and where override leaves its mark.

1 Introduction

Language models are routinely asked to follow local definitions that temporarily change what familiar words mean. A legal document may define a familiar term in an unusual way; a technical manual may repurpose a common word as a domain symbol; a system prompt may introduce a fictional rule. A software glossary defining *port* as a socket still has to fight the model's prior preference for *harbor*, and the glossary does not always win. These cases require *contextual override*: the model must use the locally specified meaning while suppressing interference from the word's pretrained lexical prior. Override failures are quiet (a residual bias toward the prior in next-token probabilities), but they recur wherever local-definition fidelity matters, from glossary lookup to retrieval-augmented document grounding.

Contextual override against lexical priors is *not* a clean replacement of meaning. It is a competition. The prior keeps operating, and its strength still shows through. This makes two predictions. Items whose distractor is more strongly favored by the model's ordinary lexical prior should show larger interference, even after surface controls. And a causal intervention that recovers override should change the target–distractor margin asymmetrically along the lexical-prior *channel*, the axis $\log P(t) - \log P(d)$ at the readout position, where prior interference manifests.

We test these claims with a Stroop-style paradigm (Figure 1). A prompt remaps a query word w to a contextual target t; a matched neutral control prompt replaces w with a semantically weak word but holds the target and a competing lexical-prior distractor d fixed. The target–distractor pair is identical in surface form and tokenization across conditions; only the query word changes by design, and the residual surface differences this introduces are absorbed by item-level covariates. The structure mirrors the classic Stroop (1935) task, in which a task-relevant dimension (the contextual mapping) is pitted against an automatic but irrelevant one (the lexical prior).

Behaviorally, across 11 open-weight models spanning the Qwen, Gemma, OLMo, and Mistral families and four prompt styles, aggregate interference is positive in every model and in every conflict family. The effect is smallest for the canonical Stroop case (antonym remapping) and 2–4$\times$ larger for the document-relevant families (arbitrary, pol-

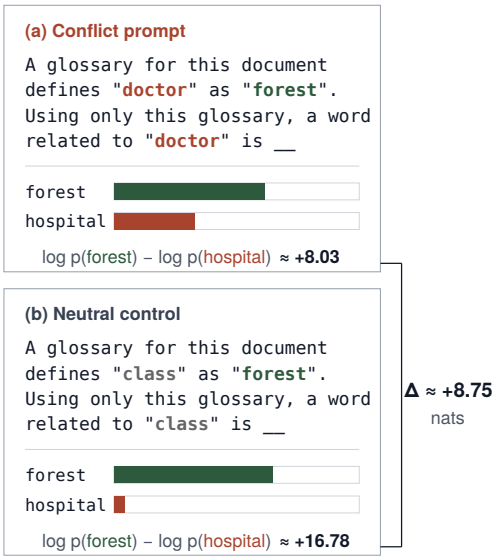

Figure 1: The Stroop-style paradigm. Conflict prompt: *doctor means forest* with target *forest* and lexical-prior distractor *hospital*. Neutral control: query word replaced by *class* (one of six sampled neutrals). Numbers are sum log-probabilities of target and distractor under Gemma-2-2B-Instruct (glossary prompt); Δ is neutral minus conflict $\log P(t) - \log P(d)$ (§3). Displayed neutral gives $\Delta \approx 8.75$; six-control mean ≈ 8.61.

ysemy, domain-definition), the settings where lexical override matters in practice. After item-level controls for answer prior, frequency, tokenization, prompt style, conflict family, model family, scale, and instruction tuning, the distractor's lexical-prior strength remains a positive predictor of interference, and the coefficient holds its sign under item-clustered, model-clustered, two-way clustered, and mixed-effects standard errors. Priors leak through override in proportion to their strength rather than being erased by the local rule.

Mechanistically, on a tractable subset of five aligned models, residual-stream activation patching identifies a *source triplet* (the joint representation of the definition subject, definition target, and query word) whose recovery exceeds every source-position pair at the site and layer the triplet selects for that model. Held-out splits, item-mismatched source controls, a definition-target swap that preserves the identity match while corrupting only the local-value position, and reader-component zero-ablation all support a binding interpretation. A logit decomposition then shows that the triplet intervention raises the target–distractor margin asymmetrically: the distractor logit drops under matched, swap, and item-mismatched conditions alike, while the target logit is preserved only under matched binding. Distractor suppression marks where override repair (the patch-induced restoration of the target–distractor margin under matched binding) lands in the logit channel; target preservation is binding's specific contribution.

Our contribution is not the observation that LMs sometimes fail to override priors. This has been documented at the benchmark level, in few-shot demonstrations, and in entity-level knowledge conflicts. Our contribution is the convergence: a single lexical-prior channel that both predicts which items resist override behaviorally and absorbs the causal effect of override repair mechanistically. Two independent measurements (an item-level behavioral predictor and a position-level causal intervention) land on the same target–distractor axis.

Concretely, this paper contributes: (i) a controlled Stroop-style paradigm for lexical-level override, with matched neutral baselines that isolate lexical-prior interference per item; (ii) evidence across 11 open-weight models that lexical-prior strength predicts override failure even after item-level controls for answer prior, frequency, tokenization, and prompt wording; (iii) a mechanistic dissociation in which the source-position triplet recovers the override effect, with the binding-specific logit signature localizing to the definition-target position alone: distractor suppression occurs broadly under source-position perturbation, but target logit collapse occurs only when the definition-target position is corrupted. Behavior and mechanism converge on the same channel.

2 Related Work

Conflicts between context and parametric knowledge. A growing body of work studies how language models reconcile in-context information with information memorized at pretraining (Longpre et al., 2021; Xie et al., 2024; Kortukov et al., 2024), with Xu et al. (2024) surveying context–memory, inter-context, and intra-memory conflicts. Almost all of this work operates at the document or claim level (entity reassignment, passage contradiction) and reads the model's open generation. Our setting is lexical-level. The readout is a teacher-forced target-versus-distractor margin, not free generation, and every conflict prompt has a matched neutral remapping that holds the target/distractor pair fixed, so lexical-prior interference from the query word reads off directly. This narrower design makes context–memory conflict measurable

as an item-level behavioral effect and testable as a position-level causal mechanism.

Psycholinguistic probes of lexical interference. Earlier probes measure lexical interference without Stroop framing: Ettinger (2020) adapts cloze diagnostics to pretrained models; Kassner and Schütze (2020) show that a misleading prime drags masked-token predictions toward the prime; Misra et al. (2020) find that semantic priming in BERT flips to interference under stronger sentential constraint. Our conflict is induced by an explicit local definition rather than a freestanding prime, each prompt has a matched *neutral remapping* that preserves the target/distractor pair, and scoring is over a target-minus-distractor margin. Matched baselines turn lexical-prior advantage into a quantitative item-level predictor of interference (§4). Our Stroop analogy is methodological (matched neutral controls and factorial design), not a claim that the LM's lexical prior is a competing percept in the cognitive sense.

In-context learning and prior tug-of-war. Wei et al. (2023) and Pan et al. (2023) document a tension between *task recognition* from priors and *task learning* from demonstrations, with scale and instruction tuning shifting the balance; Shi et al. (2023) show that irrelevant context pulls predictions away from the prior. Our paradigm is a sharper microcosm: a single in-prompt rule replaces multi-shot demonstration, a binary lexical readout replaces task accuracy, and prior strength is measured directly per item rather than inferred from population-level base-rates.

Mechanistic interpretability of knowledge conflicts. Activation patching (Vig et al., 2020; Meng et al., 2022; Zhang and Nanda, 2024) (an instance of causal mediation analysis, Mueller et al., 2026) and circuit-level analyses (Wang et al., 2023; Olsson et al., 2022; McDougall et al., 2024) localize internal representations causally, with copy suppression (McDougall et al., 2024) establishing that attention heads can shape outputs by lowering wrong-token logits rather than by writing correct ones. Within knowledge conflicts specifically, Yu et al. (2023), Ortu et al. (2024), and Jin et al. (2024) identify attention heads that promote memorized versus in-context answers at the document and entity level, with the individual head as the unit of analysis. A parallel line on in-context binding (Feng and Steinhardt, 2024; Prakash et al., 2024) studies how LMs retrieve bound entity–attribute pairs. We ask a different question on a different unit: how a single in-prompt lexical redefinition competes with a word's pretrained prior, with a source-position triplet (definition subject, definition target, query word) as the unit of analysis. Unlike entity–attribute binding where the attribute is explicit context content, our redefinition binds an existing-meaning token to a new contextual value while the prior meaning remains a strong competing completion.

Binding and representation engineering. Existing binding work treats in-context use as a problem of keeping entities paired with the right attributes: Feng and Steinhardt (2024) identify binding-ID vectors attached to corresponding entities and attributes, while Prakash et al. (2024) show that fine-tuning improves entity tracking largely by strengthening mechanisms already present in the base model. Lexical override adds a harder source of competition. The model is not merely retrieving an attribute supplied in context; it must use a new value for a familiar word whose pretrained meaning remains a plausible answer. This is why the definition-target swap is central: it preserves the query-word identity match while corrupting the local value, testing whether the source triplet carries the binding rather than just the repeated word (§5). Representation-engineering methods (Turner et al., 2023; Zou et al., 2023; Hernandez et al., 2024) instead steer or edit hidden states through model-side activation directions or fact encodings. Our intervention is item-conditioned activation patching: it restores the clean representation of the same source positions in a corrupted run. The resulting dissociation separates a patch-broad suppression channel, which lowers the lexical-prior distractor under several source perturbations, from the binding-specific channel that preserves the contextual target.

3 The Stroop-style Paradigm

3.1 Stimuli

Each item specifies a query word w, a lexical-prior distractor d, and a context-defined target t. The conflict prompt remaps w to t inside one of four prompt wrappers (defined below); the neutral prompt replaces w with a semantically weak word n chosen from a neutral pool, while leaving t and d unchanged (Figure 1). For the running *doctor* item under the glossary wrapper, the conflict prompt reads `A glossary for this document defines "doctor" as "forest".`

`Using only this glossary, a word related to "doctor" is`, with $t =$ *forest*, $d =$ *hospital*; the neutral prompt substitutes a neutral n for *doctor*. The relation cue is family-specific (`a synonym of` for antonym, `a word related to` for the other three; Appendix A, Table 5). Target and distractor completions are identical in surface form and tokenization across conditions; the query word varies by design. We average over multiple neutral words per item and include token-count and frequency covariates in the controlled regression to absorb residual length and surface effects.

Conflict families. We evaluate four families that span psycholinguistic and document-style settings: **(i)** *antonym remapping* (*small means big*; the original Stroop analogue); **(ii)** *arbitrary semantic remapping* (*apple means river*); **(iii)** *polysemy/entity remapping* (*jaguar means car*); and **(iv)** *domain-definition remapping* (*thread means process*). Per-family examples in Table 4.

Prompt styles. Each item is rendered under four wrappers: game-rule (toy framing), glossary (document glossary), technical-document (technical convention), and scoped-definition (passage-scoped rule); full templates in Table 5. This tests whether interference is tied to toy wording or persists in document-like settings.

3.2 Behavioral metric

For prompt x, let $S(x) = \log P(t \mid x) - \log P(d \mid x)$ denote relative target preference, where $P(\cdot \mid x)$ is the model's full-sequence probability under teacher forcing. The Stroop-style interference effect is

$$\Delta = \mathbb{E}_{n \in \mathcal{N}}[S(x_{\text{neutral},n})] - S(x_{\text{conflict}}),$$

where $\mathcal{N}$ is the set of sampled neutral controls for the item. Positive Δ means lexical conflict reduces relative target preference. We report sum log-probability as the primary metric; mean per-token log-probability gives a length-normalized diagnostic that yields the same aggregate model-level and controlled-regression conclusions, while individual model$\times$family and model$\times$style cells can shift modestly between the two scoring modes.

Worked example. For the running *doctor* item (Figure 1; Gemma-2-2B-Instruct under the glossary wrapper, sum log-probability), $S(x_{\text{conflict}}) \approx +8.03$. The displayed neutral control using *class* gives $S(x_{\text{neutral, class}}) \approx +16.78$ and a single-control $\Delta \approx 8.75$; the item-level metric averages over six neutral controls, giving mean $S(x_{\text{neutral}}) \approx +16.64$ and $\Delta \approx 8.61$.

Lexical-prior strength. We measure the strength of the distractor's lexical prior with an *ordinary-prior prompt* (e.g., *a word related to doctor is*) that contains no remapping rule. The relation cue matches the family of the conflict prompt (*a synonym of* for the antonym family, *a word related to* for the other three), so the prior is measured under the same relation phrase the conflict prompt uses. The lexical-prior advantage of the distractor over the target is $\log P(d) - \log P(t)$ under that prompt, computed per item per model. We use this quantity as our operational measure of lexical-prior strength.

3.3 Models

The behavioral analysis evaluates 11 open-weight model settings: Qwen2.5-1.5B and 7B base/instruct (Yang et al., 2024), Gemma-2-2B and 9B base/instruct (Rivière et al., 2024), OLMo-1B (Groeneveld et al., 2024), and Mistral-7B base/instruct (Jiang et al., 2023). We abbreviate the instruction-tuned variants as "-IT" in figures and tables. Mechanistic analysis uses the aligned tractable subset for which full activation patching with TransformerLens (Nanda and Bloom, 2022) is computationally feasible: Qwen2.5-1.5B base/IT, Gemma-2-2B base/IT, and OLMo-1B (Table 6).

4 Behavioral Results

4.1 Interference is positive across modern models

All 11 settings have positive aggregate effects (Figure 2), with effect sizes ranging from 1.31 (Gemma-2-9B-IT) to 2.71 (Qwen2.5-7B). Interference is not specific to small models, to a single family, or to a single training regime.

4.2 The effect generalizes across conflict families

The effect is positive in every family. Antonym remapping, the closest psycholinguistic analogue of a Stroop conflict, is aggregate-positive with one IT-model reversal ($\Delta = 1.16$, 95% CI $[1.06, 1.26]$, n=3520; Figure 3(a)); the three document-relevant families are 2–4$\times$ larger: arbitrary semantic remapping 3.97 $[3.79, 4.15]$, polysemy/entity 2.66 $[2.49, 2.84]$, domain-definition

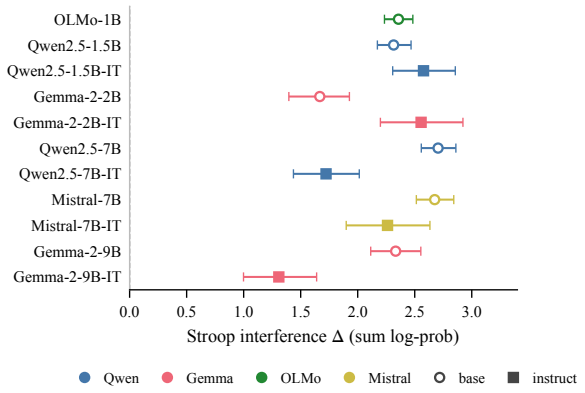

Figure 2: Model-level Stroop interference Δ with bootstrap 95% confidence intervals (sum log-probability scoring). All 11 evaluated model settings are positive at the aggregate level, including 7B/9B-class models from four families.

Term	Coef.	95% CI	p
Intercept	$+0.746$	$[+0.38, +1.11]$	$5.5{\cdot}10^{-5}$
Lexical advantage	$+0.114$	$[+0.08, +0.15]$	$4.5{\cdot}10^{-10}$
Answer prior	$+0.16$	$[+0.11, +0.21]$	$7.7{\cdot}10^{-11}$
Frequency (Δ zipf)	$+0.65$	$[+0.43, +0.86]$	$7.9{\cdot}10^{-9}$
Token-count Δ	-0.08	$[-0.46, +0.31]$	0.69
Is instruct	-0.74	$[-1.00, -0.47]$	$4.0{\cdot}10^{-8}$
log params (B)	-0.32	$[-0.44, -0.19]$	$4.3{\cdot}10^{-7}$

Table 1: Main controlled regression of item×prompt Stroop interference under sum log-probability scoring (n=7,744). The model also includes fixed effects for conflict family, prompt style, and model family (omitted for space; see Appendix D). The intercept and lexical-advantage coefficient remain positive in the main controlled specification.

$2.54\,[2.37, 2.73]$ (per-cell values in Table 10). The antonym minimum is concentrated in the largest instruction-tuned models (Qwen2.5-7B-IT, Gemma-2-9B-IT), consistent with a relation-phrase-specific confound: antonym prompts use a synonym-seeking cue (`a synonym of`) that may activate IT models' specialized opposite-word routines, whereas the other three families use a broader `a word related to` cue and are unaffected (Limitations). The paradigm generalizes to document-grounding settings where lexical override matters: glossary lookup, polysemous-entity disambiguation, and domain-specific redefinition.

4.3 The effect persists under document-style wrappers

Interference survives reframing the conflict as a glossary entry, a technical convention, or a passage-scoped definition. The game-rule wrapper yields the largest aggregate effect (3.15), but glossary (2.06), technical-document (1.75), and scoped-definition (1.94) wrappers all remain substantial (Figure 3(b); per-cell values in Table 11). The fraction of positive item–prompt observations is tightly clustered (78–81%) across all four wrappers despite the 1.8× range in mean magnitudes. Of 44 model×style cells, 42 have CI > 0.

4.4 Lexical-prior strength predicts interference

If interference reflects lexical-prior conflict rather than incidental confounds, then items whose distractor is more strongly favored by the model's ordinary lexical prior should show larger Δ. They do (Figure 4). Pooling all 11 models at the item–prompt level, bivariate OLS gives a slope of 0.086 ($p < 10^{-9}$), and binned decile means trace a clear monotone trend. In nat units, a one-nat increase in the distractor's lexical-prior advantage raises Stroop interference by about 0.1 nat. Lexical-prior advantage ranges over more than 20 nats across items (Figure 4), so the predicted effect spans the full observed range of Δ.

Table 1 reports the controlled OLS regression. After accounting for answer prior, frequency, tokenization, prompt style, conflict family, model family, scale, and instruction tuning, the lexical-advantage coefficient remains positive ($+0.114$, $p < 10^{-9}$) and the intercept remains positive ($+0.746$, $p < 10^{-4}$). The instruct and log-parameter coefficients are negative on average, but the average masks a scale interaction: pairwise base-vs-IT comparisons show instruction tuning *raising* interference at $\leq$ 2B ($+0.26$ to $+0.89$) and *lowering* it at $\geq$ 7B (-0.41 to -1.02; see data package).

Standard errors in the main table are clustered by item. The positive lexical-prior advantage coefficient retains the same sign under clustering by model, two-way clustering by item and model, and a linear mixed-effects specification with item- and model-level random effects; the mixed-effects 95% CI lower bound ($+0.014$) is close to zero, so the result is best read as a *stable positive sign across specifications* rather than as equally tight inference under all specifications. These robustness results are summarized in Appendix D and released with the data package. The instruction-tuning and scale coefficients retain the same negative sign, although their uncertainty increases under model-level and mixed-effects specifications.

Figure 3: External validity. **(a)** Effect by conflict family across models. **(b)** Effect by prompt style across models. The effect is positive in every family; the document-relevant families (arbitrary, polysemy, domain-definition) produce larger and less model-sensitive effects than antonym remapping (the canonical Stroop analogue). The effect is positive in aggregate under all document-style wrappers. Cell values are sum log-probability Δ.

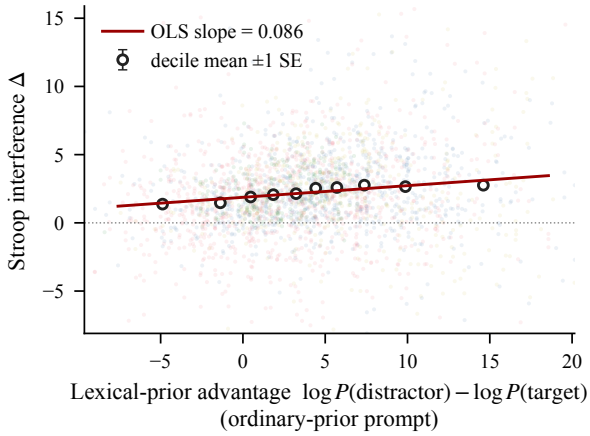

Figure 4: Lexical-prior strength predicts Stroop interference at the item×prompt level. Light points: 2,500 subsampled item×prompt observations across all 11 models. Open circles: decile means with ±1 SE bars. Red line: bivariate OLS fit (n=7,744). The controlled regression in Table 1 retains a positive coefficient after adding answer prior, tokenization, frequency, prompt style, conflict family, model family, scale, and instruction tuning.

The covariate set is conservative: answer-prior captures whether the target is easier than the distractor independent of context, and template and architecture shifts are removed by the fixed effects for prompt style, conflict family, and model family. Lexical-prior advantage remains predictive after all of these, the pattern expected if interference is driven specifically by the ordinary meaning of the query word. Within the antonym family alone, the per-item slope is not significant (+0.026, $p = 0.235$; Table 13); the pooled coefficient is carried by the three document-relevant families.

5 Mechanistic Analysis

The behavioral effects above are broad and not reducible to surface confounds. We now ask how the conflict is resolved internally.

Scope. Mechanistic experiments use the original antonym-remapping family, the game-rule prompt wrapper, and the single-token subset, on five aligned tractable models for which TransformerLens loading is reliable.

5.1 Activation patching setup

For each item we construct a clean prompt (neutral remapping) and a corrupted prompt (lexical-conflict remapping) sharing the same target and distractor. We patch clean residual-stream activations into corrupted runs at three source positions: the *definition subject* (e.g. *small*), the *definition target* (e.g. *big*), and the *query word*. We also patch the *full source triplet* combining all three, reporting the standard normalized recovery R of Zhang and Nanda (2024) on the final-position target-minus-distractor logit difference. Mechanistic items use the single-token target/distractor subset; Appendix E gives the implementation details.

5.2 The source triplet binds the local definition

Joint patching of the definition subject, definition target, and query word recovers nearly all of the

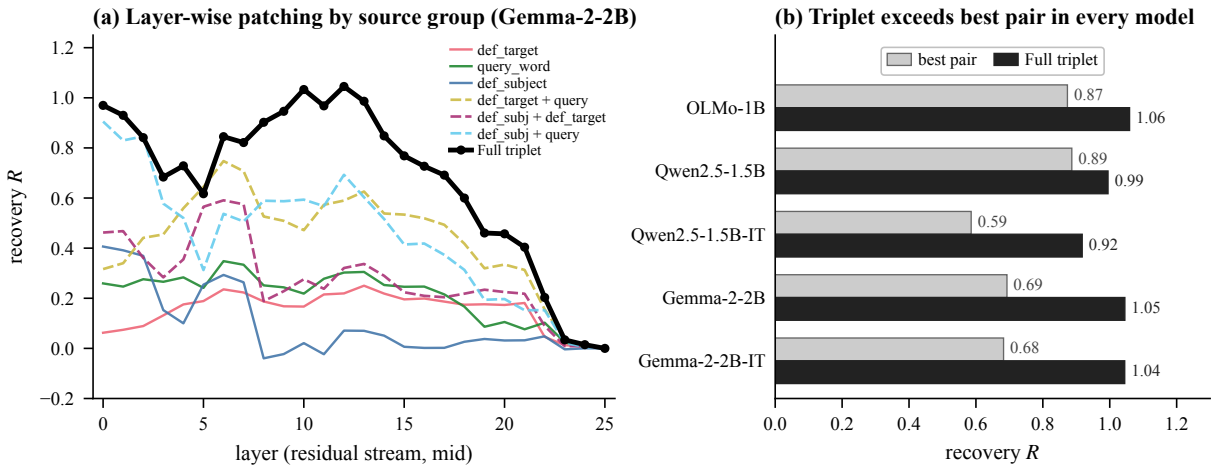

Figure 5: Source-triplet residual mechanism. **(a)** Layer-wise patching recovery for each source group on Gemma-2-2B at the residual-stream site the full triplet selects (mid-block). The full triplet sits above every singleton and pair through mid-network depths. **(b)** Recovery for the full triplet versus the best source-position pair across the five mechanism models, evaluated at the site and layer the triplet selects in each model. The triplet exceeds the best pair without exception ($\Delta R \in [+0.11, +0.36]$); split-validation evidence is in the text.

Model	Site	L	R_{trp}	R_{pair}	ΔR
Qwen2.5-1.5B	input	3	0.99	0.89	+0.11
Qwen2.5-1.5B-IT	input	15	0.92	0.59	+0.33
Gemma-2-2B	mid-block	12	1.05	0.69	+0.35
Gemma-2-2B-IT	input	3	1.04	0.68	+0.36
OLMo-1B	input	2	1.06	0.87	+0.19

Table 2: Source-triplet residual patching versus the best source-position pair. Site (residual-stream location: input = block input, mid-block = post-attention) and L (layer) are selected by the full source triplet; R_{trp} is the recovery of the full triplet at that site/layer, and R_{pair} is the recovery of the best source-position pair at the same site/layer. The best pair is consistently definition subject + query word; adding the definition target improves point-estimate recovery in all five models.

conflict effect (aggregate $R \in [0.92, 1.06]$ across five models, mean 1.03), exceeding any source-position pair at the residual-stream site and layer the triplet selects (Figure 5, Table 2; triplet–pair margin $\Delta R = R_{\text{triplet}} - R_{\text{pair}} \in [+0.11, +0.36]$ in 5/5 models). The best pair is consistently definition subject + query word; adding the definition target improves recovery at the same site and layer. Recovery for every source group at this site/layer is in Appendix F, Table 14.

We read the triplet as filling three structural slots: subject and query together specify the identity match (which word is being redefined), while definition target supplies the local value (what it now means). ΔR then measures the cost of any missing slot.

Definition-target swap control. A definition-target swap tests this directly: keep subject and query word matched, replace only the definition-target activation with a donor item's. Recovery collapses in every item of every model (142/142; per-item matched−swap margin at least 1.9 nats, model-level margin +1.54 to +5.26; Appendix F, Table 15), preserving the identity match while corrupting only the local-value position. This rules out a purely identity-based account.

Split validation. Across 200 discovery/held-out splits, the held-out triplet–pair margin has a positive mean in each of the five models, tightly so in base models (0.97–1.00 positive fraction) and more heterogeneously in instruction-tuned ones (0.83–0.95). Per-model values are in Appendix F, Table 17.

5.3 Binding's logit signature: target preservation, not distractor suppression

Recovery requires the full triplet ($\Delta R = +0.11$ to $+0.36$ over the best pair, 5/5 models; Table 2). Decomposing ΔR into the underlying margin movement, $\Delta m = \Delta \ell_t - \Delta \ell_d$, reveals a sharper localization: the binding-specific signature localizes to the definition-target position alone. Corrupting the definition-target position via swap produces a logit decomposition nearly identical to fully item-mismatched controls (Table 3; $|\Delta m_S - \Delta m_R| \leq 0.6$ nat across all five models).

Matched binding better preserves the contextual target logit ($\Delta \ell_t \in [-7, +0.17]$) while the dis-

Model	Matched (M)			Swap (S)			Random (R)		
	$\Delta\ell_t$	$\Delta\ell_d$	Δm	$\Delta\ell_t$	$\Delta\ell_d$	Δm	$\Delta\ell_t$	$\Delta\ell_d$	Δm
Qwen2.5-1.5B	-1.39	-3.13	$+1.74$	-12.79	-5.34	-7.45	-12.68	-5.41	-7.27
Qwen2.5-1.5B-IT	-6.97	-11.05	$+4.08$	-16.48	-13.70	-2.77	-16.26	-13.34	-2.91
Gemma-2-2B	$+0.17$	-3.72	$+3.89$	-8.04	-6.02	-2.03	-7.95	-5.81	-2.14
Gemma-2-2B-IT	-1.78	-5.60	$+3.83$	-16.92	-8.03	-8.89	-15.69	-7.40	-8.30
OLMo-1B	-1.38	-3.41	$+2.03$	-7.68	-5.50	-2.19	-7.56	-5.33	-2.23

Table 3: Logit decomposition under three patching conditions. M = matched triplet, S = definition-target swap, R = item-mismatched control. $\Delta m = \Delta\ell_t - \Delta\ell_d$. Distractor drops comparably across all three (within 0.7 nat in S vs R); target collapses only when binding is disrupted. Margin gaps $|\Delta m_S - \Delta m_R| \leq 0.6$ nat across all five models.

tractor drops by 3 to 11 nats. Under swap and item-mismatched controls the distractor drops by 5 to 13 nats (a comparable order of magnitude, somewhat larger than matched), but the target collapses by 7 to 17 nats, inverting the margin. The binding-specific signature is therefore selective target preservation. Distractor suppression operates broadly under any source-position perturbation (§5.4).

5.4 Robustness: item-mismatched controls

Across all 250 control trials (50 perturbations per model $\times$ 5 models), item-mismatched clean source activations yield strongly *negative* recovery (-0.58 to -4.16; Figure 6, Table 18); no single trial reaches positive recovery and none approach the matched value, so recovery requires item-matched information. Component-level scans then localize partial mediation in identifiable readers (attention heads and MLPs at the final position, where the target-vs-distractor margin is read out). Zero-ablating top reader components during the matched triplet patch reduces recovery by 0.19 to 0.62 across the five models (Appendix F, Table 16); writer effects at the source positions are smaller

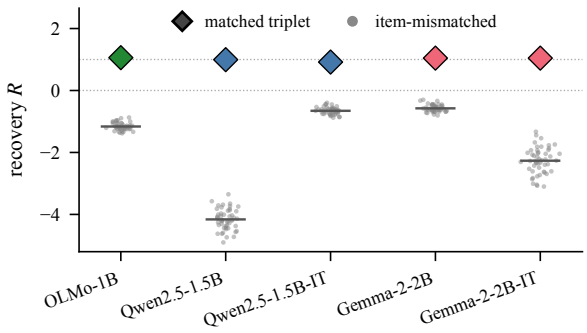

Figure 6: Matched triplet recovery (colored diamond) versus 50 item-mismatched clean-source controls (grey points; horizontal bar at the control mean) per mechanism model. Item-mismatched clean activations *do not* recover the effect; they actively harm it, establishing that recovery requires item-matched source information.

and more distributed, with full outputs provided in the data package.

6 Discussion

Two independent measurements landed on the same lexical-prior channel. Behaviorally, a one-forward-pass estimate of distractor-versus-target prior strength predicts per-item override interference across 11 open-weight models, after item-level controls, and across four prompt wrappers including document-style framings. Mechanistically, a source-position triplet recovers the override effect, and the lexical-prior distractor channel that the behavioral predictor identifies is also where override repair lands at the logit level. The prior is both where interference comes from and where repair lands.

The logit decomposition revealed a functional partition. Matched binding's contribution to the target–distractor margin runs almost entirely through target preservation: the distractor logit drops by comparable magnitude under structured corruption (swap) and unstructured corruption (item-mismatched random), so distractor suppression is a broad effect of source-position perturbation rather than a binding-specific function.

Across paradigm, behavior, and mechanism, lexical override emerges as partial competition rather than clean substitution: the prior keeps operating, and interference scales with its strength.

Limitations

The paradigm is deliberately controlled. Even with glossary, technical-document, and scoped-definition wrappers, our prompts are shorter than realistic long-form documents, and the items use English stimuli with target-vs-distractor likelihood scoring rather than unconstrained free generation.

The mechanistic analysis covers five aligned 1B–2B models where full residual and component patching is tractable, while the behavioral analy-

sis spans 1B–9B across four families. Whether the source-triplet structure persists at 7B+ scale, across the four conflict families, and across non-game-rule wrappers is open; the mechanism is established on antonym remapping under the game-rule wrapper, while the behavioral signature holds across the larger range. The functional partition between binding-specific target preservation and patch-broad distractor suppression is established on the same antonym single-token subset; whether it generalizes to other conflict families and larger models is open. The definition-target swap supports binding at the source-representation level. A same-target donor swap (donors constrained to share the target word, isolating context-binding from surface-form effects) and a complete path-level circuit through writer/reader paths remain future work.

The antonym family is structurally constrained: a synonym-seeking cue (`a synonym of`) is the only phrasing that keeps distractor and remap target separable for antonym pairs, since broader cues would invite the antonym itself as a high-prior continuation and collapse the conflict. This cue may activate IT models' specialized opposite-word behavior, fitting the observed antonym minimum in Qwen2.5-7B-IT and Gemma-2-9B-IT but not directly tested. The caveat is specific to the antonym family; the other three use a broader relation cue. Disentangling the IT-routine effect would require varying the synonym-cueing wording while preserving conflict structure (future work).

Ethical Considerations

This work uses synthetic stimuli and open-weight models; no human subjects or sensitive data are involved. The paradigm is diagnostic and aimed at understanding when local context fails to override pretrained meaning, a relevant pre-deployment check for systems that must respect document-specific definitions in domains such as law, medicine, and technical documentation. We report aggregate behavior and causal interventions rather than prompt-attack constructions.

References

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8.

Jiahai Feng and Jacob Steinhardt. 2024. How do language models bind entities in context? In *The Twelfth International Conference on Learning Representations*.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, and Yizhong Wang. 2024. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Evan Hernandez, Belinda Z. Li, and Jacob Andreas. 2024. Inspecting and editing knowledge representations in language models. In *First Conference on Language Modeling (COLM)*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, and Lucile Saulnier. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Evgenii Kortukov, Alexander Rubinstein, Elisa Nguyen, and Seong Joon Oh. 2024. Studying large language model behaviors under context-memory conflicts with real documents. In *First Conference on Language Modeling (COLM)*.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Callum McDougall, Arthur Conmy, Cody Rushing, Thomas McGrath, and Neel Nanda. 2024. Copy suppression: Comprehensively understanding a motif in language model attention heads. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35.

Kanishka Misra, Allyson Ettinger, and Julia Taylor Rayz. 2020. Exploring BERT's sensitivity to lexical cues using tests from semantic priming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Aaron Mueller, Jannik Brinkmann, Millicent Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiuding Sun, Eric Todd, David Bau, and Yonatan Belinkov. 2026. The quest for the right mediator: Surveying mechanistic interpretability for NLP through the lens of causal mediation analysis. *Computational Linguistics*, 52(1):331–378.

Neel Nanda and Joseph Bloom. 2022. TransformerLens: A library for mechanistic interpretability of generative language models. https://github.com/TransformerLensOrg/TransformerLens.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, and Anna Chen. 2022. In-context learning and induction heads. *Transformer Circuits Thread*.

Francesco Ortu, Zhijing Jin, Diego Doimo, Mrinmaya Sachan, Alberto Cazzaniga, and Bernhard Schölkopf. 2024. Competition of mechanisms: Tracing how language models handle facts and counterfactuals. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. 2023. What in-context learning "learns" in-context: Disentangling task recognition and task learning. In *Findings of the Association for Computational Linguistics: ACL 2023*.

Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. 2024. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *The Twelfth International Conference on Learning Representations*.

Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and Johan Ferret. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*.

J. Ridley Stroop. 1935. Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18(6):643–662.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2023. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, and Haoran Wei. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Qinan Yu, Jack Merullo, and Ellie Pavlick. 2023. Characterizing mechanisms for factual recall in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Fred Zhang and Neel Nanda. 2024. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*.

In the appendices below, code-style names in monospace (`source_all`, `def_subject`, `def_target`, `query_word`, `game_rule`, etc.) correspond to column values, group labels, or template identifiers in the released CSVs and notebook code. The body of the paper uses the corresponding paper-level terms (*full source triplet*, *definition subject*, *definition target*, *query word*, *game-rule wrapper*, etc.).

A Stimulus Construction and Prompt Templates

Behavioral set. The behavioral stimulus set contains 176 unique lexical items, each rendered under four prompt styles, yielding 704 item–prompt observations per model setting. Unique items are distributed across the four conflict families as follows: 80 antonym, 36 arbitrary semantic remapping, 30 polysemy/entity remapping, and 30 domain-definition remapping. Targets and distractors are chosen to be single tokens under each model's tokenizer whenever possible; when not, the target/distractor token-count difference is included as a regression covariate (Table 1). The full stimulus list is released in `behavior_external_validity_stimuli.csv`.

Query	Target	Distractor
Antonym remapping (n=80)		
small	*big*	*tiny*
large	*small*	*big*
hot	*cold*	*warm*
cold	*hot*	*chilly*
happy	*sad*	*glad*
sad	*happy*	*unhappy*
Arbitrary semantic remapping (n=36)		
apple	*river*	*fruit*
banana	*mountain*	*fruit*
tiger	*pencil*	*animal*
doctor	*forest*	*hospital*
teacher	*ocean*	*school*
piano	*desert*	*music*
Polysemy / entity remapping (n=30)		
apple	*music*	*iPhone*
Python	*programming*	*snake*
Java	*programming*	*coffee*
jaguar	*car*	*animal*
mercury	*metal*	*planet*
bass	*music*	*fish*
Domain-definition remapping (n=30)		
port	*socket*	*harbor*
thread	*process*	*sewing*
cell	*spreadsheet*	*biology*
token	*symbol*	*coin*
driver	*software*	*person*
kernel	*system*	*seed*

Table 4: Six examples per conflict family. The conflict definition takes the form `<Query> means <Target>` in the `game_rule` wrapper, with the other three wrappers paraphrasing it (Table 5); the full stimulus set (176 items) is released in `behavior_external_validity_stimuli.csv`.

Neutral controls. Neutral controls replace the query word with semantically weak words drawn from a per-item candidate pool of 6–8 words, screened to be semantically weak in context (e.g., *thing*, *object*, *item*, *label*, *symbol* for some items; *class*, *field*, *point*, *record*, *slot*, *type* for items where those generic placeholders are themselves contextually plausible). Behavioral experiments average K=6 controls per item; mechanistic experiments select a single representative clean prompt closest to the item's neutral mean from K=8 candidates, giving a deterministic clean run for activation patching. The full per-item neutral lists are released in the stimulus CSV.

Prompt templates. The four prompt-style wrappers are listed in Table 5. Each wrapper embeds the conflict definition and the query in a different framing. Behavioral experiments instantiate all four wrappers; mechanistic experiments use the `game_rule` wrapper from the same template family.

game_rule

```
Complete the answer with one word or
phrase only.
In this game, {subject} means {meaning}.
Question:  In this game,
{relation_phrase} {query} is
Answer:
```

glossary

```
A glossary for this document defines
"{subject}" as "{meaning}".
Using only this glossary,
{relation_phrase} "{query}" is
```

technical_document

```
In the following technical document, the
term "{subject}" refers to "{meaning}".
According to the document,
{relation_phrase} "{query}" is
```

definition_scope

```
For this passage only, interpret
"{subject}" as "{meaning}".
Under this definition, {relation_phrase}
"{query}" is
```

Table 5: Prompt-style wrappers. {subject} and {meaning} are filled in from the conflict definition; {query} is the query word; {relation_phrase} is "a synonym of" for the antonym family and "a word related to" for the other families. The wrappers correspond to the prose names *game-rule*, *glossary*, *technical-document*, and *scoped-definition*.

B Model Set, Tokenization, and Scoring

Table 6 lists all model settings, with a flag indicating whether each is used for the behavioral analysis, the mechanistic analysis, or both. The 7B/9B-class models are included for behavioral external validity. Component-level activation patching is restricted to the aligned tractable subset (Qwen2.5-1.5B base/IT, Gemma-2-2B base/IT, OLMo-1B), where full residual and component patching is computationally feasible and TransformerLens (Nanda and Bloom, 2022) loading is reliable.

Model	Family	Beh.	Mech.
Qwen2.5-1.5B	Qwen	✓	✓
Qwen2.5-1.5B-IT	Qwen	✓	✓
Qwen2.5-7B	Qwen	✓	
Qwen2.5-7B-IT	Qwen	✓	
Gemma-2-2B	Gemma	✓	✓
Gemma-2-2B-IT	Gemma	✓	✓
Gemma-2-9B	Gemma	✓	
Gemma-2-9B-IT	Gemma	✓	
OLMo-1B	OLMo	✓	✓
Mistral-7B	Mistral	✓	
Mistral-7B-IT	Mistral	✓	

Table 6: Open-weight model settings. "IT" is the instruction-tuned variant. "Beh." marks inclusion in the behavioral sweep (n=11); "Mech." marks inclusion in component-level mechanistic analysis (n=5).

Tokenization. Target and distractor tokenization is identical across conflict and neutral conditions; the query word can differ by design. The controlled regression includes target/distractor token-count and frequency covariates (Table 1).

C Behavioral Results: Full Tables

Aggregate effect by family and prompt style. Tables 7 and 8 report the aggregate sum-log-probability effect Δ at the conflict-family and prompt-style levels. The aggregate effect is positive across all four conflict families and all four prompt styles. Individual model$\times$family and model$\times$style cells reveal meaningful heterogeneity, especially for antonym remapping in larger instruction-tuned models (Section 4, Fig. 3); the per-cell tables are released as `behavior_summary_by_model_conflict_type.csv` and `behavior_summary_by_model_prompt_style.csv`.

Family	n	Δ	95% CI	Pos.
Antonym	3520	1.16	$[1.06, 1.26]$	0.70
Arbitrary	1584	3.97	$[3.79, 4.15]$	0.92
Polysemy/ent.	1320	2.66	$[2.49, 2.84]$	0.86
Domain-def.	1320	2.54	$[2.37, 2.73]$	0.84

Table 7: Aggregate Stroop interference Δ by conflict family (sum log-probability). "Pos." is the fraction of item–prompt–model observations with positive Δ.

Prompt style	n	Δ	95% CI	Pos.
Game-rule	1936	3.15	[2.92, 3.38]	0.78
Glossary	1936	2.06	[1.92, 2.19]	0.81
Technical-document	1936	1.75	[1.64, 1.88]	0.80
Scoped-definition	1936	1.94	[1.82, 2.06]	0.80

Table 8: Aggregate Stroop interference Δ by prompt style (sum log-probability).

Per-model behavior. Table 9 reports the model-level aggregate Δ for the 11 model settings under sum log-probability scoring (Fig. 2 in the body shows the same data graphically). All 11 settings are positive at the model level.

Model	Size	Δ	95% CI	Pos.
OLMo-1B	1.0	2.36	[2.24, 2.48]	0.95
Qwen2.5-1.5B	1.5	2.31	[2.17, 2.47]	0.92
Qwen2.5-1.5B-IT	1.5	2.58	[2.31, 2.86]	0.80
Gemma-2-2B	2.0	1.67	[1.40, 1.93]	0.71
Gemma-2-2B-IT	2.0	2.56	[2.20, 2.92]	0.75
Qwen2.5-7B	7.0	2.71	[2.56, 2.86]	0.93
Qwen2.5-7B-IT	7.0	1.72	[1.44, 2.01]	0.72
Mistral-7B	7.0	2.67	[2.51, 2.84]	0.91
Mistral-7B-IT	7.0	2.26	[1.90, 2.63]	0.66
Gemma-2-9B	9.0	2.33	[2.11, 2.55]	0.81
Gemma-2-9B-IT	9.0	1.31	[1.00, 1.64]	0.61

Table 9: Model-level Stroop interference Δ (sum log-probability), n=704 item–prompt observations per model. Mean per-token scoring gives the same conclusion that all 11 settings are positive at the model level (released in `behavior_summary_by_model.csv`).

Per-cell breakdowns. Tables 10 and 11 report per-cell Stroop interference Δ for each model×conflict-family and model×prompt-style combination, respectively (sum log-probability). These are the numerical counterparts of the heatmaps in Fig. 3; they make the heterogeneity discussed in Section 4 fully auditable.

Model	Ant.	Arb.	Pol.	Dom.
OLMo-1B	2.12	3.07	2.45	2.06
Qwen2.5-1.5B	1.46	4.03	2.69	2.15
Qwen2.5-1.5B-IT	1.69	4.72	2.41	2.54
Gemma-2-2B	1.21	2.05	1.81	2.28
Gemma-2-2B-IT	0.55	6.33	3.04	2.87
Qwen2.5-7B	2.24	4.09	2.78	2.21
Qwen2.5-7B-IT	0.42	2.40	3.39	2.73
Mistral-7B	1.87	3.94	3.32	2.67
Mistral-7B-IT	0.34	5.06	2.83	3.46
Gemma-2-9B	1.46	3.56	2.50	3.02
Gemma-2-9B-IT	−0.63	4.38	2.10	2.00

Table 10: Per-cell Δ by model and conflict family (sum log-probability), corresponding to Fig. 3(a). Columns: Ant. = antonym; Arb. = arbitrary semantic; Pol. = polysemy/entity; Dom. = domain-definition. Antonym remapping in larger instruction-tuned models gives the smallest or near-zero cell values.

Model	Game	Gloss.	Tech.	Scope
OLMo-1B	2.41	2.29	2.66	2.07
Qwen2.5-1.5B	2.97	2.03	2.21	2.05
Qwen2.5-1.5B-IT	3.24	2.65	2.75	1.67
Gemma-2-2B	3.49	1.51	0.43	1.24
Gemma-2-2B-IT	2.69	3.39	1.64	2.51
Qwen2.5-7B	3.95	2.08	2.48	2.31
Qwen2.5-7B-IT	1.38	2.30	0.77	2.44
Mistral-7B	4.40	1.78	2.65	1.87
Mistral-7B-IT	5.00	0.87	0.82	2.36
Gemma-2-9B	1.68	3.01	3.23	1.41
Gemma-2-9B-IT	3.46	0.72	−0.32	1.37

Table 11: Per-cell Δ by model and prompt style (sum log-probability), corresponding to Fig. 3(b). Columns: Game = game-rule; Gloss. = glossary; Tech. = technical-document; Scope = scoped-definition.

D Regression Robustness

Table 1 (main paper) reports the controlled OLS regression with standard errors clustered by item. Full coefficient tables covering all conflict-family, prompt-style, and family fixed effects, as well as (CONFLICT_FAMILY × PROMPT_STYLE) and (LEXICAL_ADVANTAGE × CONFLICT_FAMILY) interaction specifications, are released in `behavior_control_regressions.csv`; the fixed effects and interactions are omitted from the main table for space.

Four robustness specifications are released in `behavior_regression_robustness.csv`: (i) standard errors clustered by item, (ii) clustered by model, (iii) two-way clustered by item and model, and (iv) a linear mixed-effects specification with item- and model-level random effects. Table 12 summarizes these specifications for the key LEXICAL_ADVANTAGE predictor under sum log-probability scoring; the coefficient remains positive across all four specifications (cf. §4.4). Mean log-probability scoring gives the same qualitative pattern.

SE specification	Coef.	95% CI	p
Item-clustered	+0.114	[+0.08, +0.15]	$4.5{\cdot}10^{-10}$
Model-clustered	+0.114	[+0.07, +0.16]	$3.7{\cdot}10^{-6}$
Two-way clustered	+0.114	[+0.06, +0.17]	$4.8{\cdot}10^{-5}$
Mixed-effects	+0.114	[+0.01, +0.21]	0.025

Table 12: Robustness of the LEXICAL_ADVANTAGE coefficient under sum log-probability scoring (n=7,744) across four standard-error specifications.

Family	Coef.	95% CI	p
Antonym (baseline)	$+0.026$	$[-0.017, +0.070]$	0.235
Arbitrary semantic	$+0.202$	$[+0.13, +0.28]$	$1.5{\cdot}10^{-7}$
Domain-definition	$+0.185$	$[+0.09, +0.28]$	$1.5{\cdot}10^{-4}$
Polysemy/entity	$+0.150$	$[+0.08, +0.22]$	$5.8{\cdot}10^{-5}$

Table 13: Family-specific lexical-prior slopes from the LEXICAL_ADVANTAGE × CONFLICT_FAMILY interaction specification (released as `behavior_control_regressions.csv`, regression `type_interactions`). The antonym baseline slope is null; the other three families show positive lexical-prior effects.

E Mechanistic Analysis: Implementation

Clean and corrupted prompts. For each item we construct a *clean* prompt (neutral remapping) and a *corrupted* prompt (lexical-conflict remapping) that share the same target and distractor. The clean prompt is selected from the $K{=}8$ neutral candidates as the candidate whose target-vs-distractor score is closest to the item's neutral mean, giving a deterministic clean run for activation patching. Mechanistic items are restricted to the single-token target/distractor subset.

Residual-stream sites. We refer to three sites within each transformer block: block input (CSV column `hook_resid_pre`), mid-block, i.e., post-attention and pre-MLP (`hook_resid_mid`), and block output (`hook_resid_post`). The main paper uses the readable names; CSV column names follow the TransformerLens convention.

Source positions. Source positions are named `def_subject`, `def_target`, and `query_word` in the released CSVs; the full triplet intervention is named `source_all`. The *final position* is the final prompt position at which the target-vs-distractor logit difference is read. Although the triplet combines three source positions, the binding signal at this resolution localizes primarily to `def_target` (§5).

Recovery. The patching score is the final-position target-minus-distractor logit difference $M = \ell_t - \ell_d$. Recovery is $R = (M_{\text{patched}} - M_{\text{corrupt}})/(M_{\text{clean}} - M_{\text{corrupt}})$, the standard normalized metric of Zhang and Nanda (2024): $R{=}1$ is full clean–corrupt recovery, $R{=}0$ is no recovery, and R can exceed 1 or fall below 0 when the patch overshoots or actively harms the answer state.

Split validation. For each model we run 200 repeated splits. Each split chooses a discovery item set, selects the best (hook, layer, source-group) patch on the discovery items, and evaluates that patch on the held-out items, recording the held-out $\Delta R = R_{\text{triplet}} - R_{\text{best pair}}$. Held-out splits in base models have positive ΔR in 0.97–1.00 of splits; instruction-tuned models have a positive split mean but more heterogeneity (Section 5).

Item-mismatched source controls. For each model we run 50 control perturbations. Each perturbation patches clean source activations from a randomly chosen *different* item, leaving the corrupted item's clean target-distractor pair fixed. These controls test whether recovery requires item-matched source information; they do not recover behavior and often actively harm it (Fig. 6).

F Additional Mechanistic Results

Source groups at the triplet-selected site and layer. Table 14 reports the recovery of each source group evaluated at the *same residual-stream site and layer* that the full triplet selects in Table 2 for that model. This is the comparison point used by the main source-triplet analysis: at each model's selected site and layer, the full triplet exceeds every singleton and every pair. The released CSV also contains all per-layer values for the three residual-stream sites: block input, mid-block, and block output.

Group (site, L)	Q1.5 input, 3	Q1.5-IT input, 15	G2 mid-block, 12	G2-IT input, 3	OLMo input, 2
def_s	-2.46	-0.09	$+0.07$	-3.33	-0.02
def_t	$+0.03$	-1.99	$+0.22$	$+0.34$	$+0.06$
query	-2.35	$+0.35$	$+0.30$	-1.20	-0.58
def_s + def_t	-2.31	-1.92	$+0.32$	-3.75	$+0.09$
def_s + query	$+0.89$	$+0.59$	$+0.69$	$+0.68$	$+0.87$
def_t + query	-2.20	-0.09	$+0.59$	-1.02	-0.50
Triplet	**0.99**	**0.92**	**1.05**	**1.04**	**1.06**

Table 14: Source-group recoveries at each model's triplet-selected residual-stream site and layer from Table 2. At the selected site and layer, the full triplet exceeds every singleton and every pair in all five models. Column headings: Q1.5 = Qwen2.5-1.5B; Q1.5-IT = Qwen2.5-1.5B-IT; G2 = Gemma-2-2B; G2-IT = Gemma-2-2B-IT; OLMo = OLMo-1B. Row headings: def_s = definition subject; def_t = definition target; query = query word.

Own-best-layer protocol. Evaluating each source group at its own-best (site, layer) rather than the triplet-selected position lets the definition subject + query word pair reach high

recovery at early layers, where the patch behaves like an input replacement. We treat these as representational alternatives and use the matched-site/layer comparison in Table 14 as the main comparison point. All per-group per-layer recoveries (and all three residual-stream sites: block input, mid-block, and block output) are in `mechanism_source_residual_patches.csv`.

Definition-target swap control. Donor definition-target activations are drawn from the same matched pool used for the matched-triplet patch. We run 10 donor swaps per item at each model's triplet-selected site and layer and report, in Table 15, the matched recovery, the mean swap recovery, and the fraction of items whose mean swap recovery falls below the matched value (*every* item in every model). The near-identity between swap and item-mismatched outcomes at the margin level (§5, Table 3) further indicates that the binding signal at this resolution is carried almost entirely by the `def_target` position.

Model	(site, L)	n	Matched	Swap	Frac. below
Qwen2.5-1.5B	input, 3	32	$+0.99$	-4.26	1.00
Qwen2.5-1.5B-IT	input, 15	32	$+0.92$	-0.62	1.00
Gemma-2-2B	mid-block, 12	32	$+1.05$	-0.54	1.00
Gemma-2-2B-IT	input, 3	14	$+1.04$	-2.43	1.00
OLMo-1B	input, 2	32	$+1.06$	-1.14	1.00

Table 15: Definition-target swap control (`mechanism_definition_target_swap_summary.csv`). The definition subject and query word activations are kept matched to the item; only the definition-target activation is replaced by a donor item's definition-target activation. Matched is the matched triplet recovery; Swap is the mean swap recovery across 10 donor swaps per item; Frac. below is the fraction of items whose mean swap recovery is below the matched triplet recovery. Recovery drops sharply in every model, supporting the binding interpretation.

Reader zero-ablation mediation. A non-degenerate component scan at the final position (attention heads and MLPs whose individual-component patch effect is reproducible across split seeds and not subsumed by trivial null perturbations) identifies, in each mechanism model, a small set of *reader* components that mediate the triplet patch. Table 16 lists the top reader components by *mediation drop*: the reduction in triplet-patch recovery when the listed component is zero-ablated during the matched triplet patch. The top reader in every model contributes a recovery drop of at least 0.19, and individual readers account for drops of up to 0.62. Per-model writer scans at source positions are released in `mechanism_component_scan_nondegenerate.csv`; writer effects are smaller and more distributed across heads/MLPs than the reader effects reported here. Not every non-degenerate component supports recovery: 8 components (across the five models), primarily late-block MLPs, yield a negative drop of magnitude ≥ 0.3, meaning that zero-ablating them *increases* patched recovery. The pattern is reproducible across split seeds and is consistent with a suppression-style role for these components, in which removing them releases the patched target–distractor margin upward (cf. copy suppression, McDougall et al., 2024); the full list is in the component-mediation CSV.

Model	Comp.	R_{trp}	$R_{\text{trp+0}}$	Drop
Qwen2.5-1.5B	L20H5	0.99	0.37	0.62
Qwen2.5-1.5B	L26H7	0.99	0.44	0.56
Qwen2.5-1.5B-IT	L27MLP	0.92	0.56	0.36
Qwen2.5-1.5B-IT	L20H5	0.92	0.63	0.29
Gemma-2-2B	L12H2	1.05	0.79	0.26
Gemma-2-2B	L22H7	1.05	0.80	0.24
Gemma-2-2B-IT	L22H1	1.04	0.58	0.47
Gemma-2-2B-IT	L16H4	1.04	0.72	0.33
OLMo-1B	L8MLP	1.06	0.78	0.28
OLMo-1B	L15H12	1.06	0.87	0.19

Table 16: Top reader zero-ablation drops per model (`mechanism_component_mediation_zero_ablation.csv`). R_{trp} is the matched triplet recovery; $R_{\text{trp+0}}$ is the triplet recovery with the listed component zero-ablated; *Drop* = $R_{\text{trp}} - R_{\text{trp+0}}$.

Split-validation summary. Table 17 reports the 200-fold split-validation results per mechanism model: held-out mean $\Delta R = R_{\text{triplet}} - R_{\text{pair}}$, 95% split-level confidence interval, and fraction of splits with positive ΔR (cf. §5).

Model	Splits	Mean ΔR	95% CI	Pos.
Qwen2.5-1.5B	200	$+0.11$	$[+0.00, +0.34]$	0.975
Qwen2.5-1.5B-IT	200	$+0.31$	$[-0.01, +0.56]$	0.950
Gemma-2-2B	200	$+0.39$	$[+0.03, +0.66]$	0.975
Gemma-2-2B-IT	200	$+0.19$	$[-0.11, +0.58]$	0.825
OLMo-1B	200	$+0.18$	$[+0.08, +0.28]$	1.000

Table 17: Held-out split-validation summary per mechanism model (`mechanism_split_validation.csv`, summary in `mechanism_split_validation_summary.csv`). **Pos.** is the fraction of splits with positive ΔR.

Item-mismatched source controls summary. Table 18 summarizes the 50 item-mismatched control perturbations per mechanism model. The matched triplet recovery is near or above 1 for every model, while the mean control recovery is strongly negative; the highest single control value (**Ctrl. max**) does not approach the matched value in any model.

Model	Matched	Ctrl. mean	Ctrl. min	Ctrl. max
Qwen2.5-1.5B	0.99	−4.16	−4.90	−3.35
Qwen2.5-1.5B-IT	0.92	−0.66	−0.88	−0.40
Gemma-2-2B	1.05	−0.58	−0.80	−0.30
Gemma-2-2B-IT	1.04	−2.27	−3.10	−1.33
OLMo-1B	1.06	−1.16	−1.38	−0.87

Table 18: Item-mismatched source-control summary per mechanism model (`mechanism_random_controls.csv`; 50 permutations per model). The matched triplet recovery is far above the worst-case control max in every model.

G Reproducibility and Released Artifacts

Repository and released package. We provide an anonymized reproduction repository at anonymous reproduction repository. It contains the end-to-end reproduction notebook, generated CSV outputs, figures, Markdown summaries, and package documentation. The notebook reproduces all numerical results reported in the paper; the repository's `README` indexes each CSV against the figure or table it supports.

Artifacts, licenses, and intended use. The released code and documentation are distributed for research reproduction and diagnostic analysis. The code is MIT-licensed. Pretrained model checkpoints, TransformerLens, and Hugging Face tooling retain their original licenses, access conditions, and model-card terms; gated checkpoints require the corresponding upstream access approval. The created artifacts are not intended for deployment or downstream decision-making.

Data, human subjects, and risk profile. The released data consist of synthetic prompt templates, lexical-item lists, generated CSV outputs, figures, and Markdown summaries. No human subjects, crowdworkers, private user data, or sensitive personal data were collected. Because the paradigm is diagnostic, the main risk is over-interpreting lexical-override interference as a general prompt-attack or safety-control method; we report aggregate behavior and causal interventions rather than operational attack instructions.

Hardware and precision. Reproduction is straightforward on a single A100-class GPU and involves forward evaluation plus activation caching and patching, with no model training or fine-tuning. Behavioral evaluation uses `bfloat16` where available; mechanistic experiments use `float32` for stable activation caching and consistent residual-stream arithmetic across model families.

Tooling. Mechanistic interventions use TransformerLens (Nanda and Bloom, 2022). Gated model checkpoints (e.g., the Gemma family) require a Hugging Face access token; the notebook reads it from the `HF_TOKEN` environment variable.

Tokenizer handling. Some open-weight tokenizers (notably the Qwen family) require explicit handling because the default behavior assumes a beginning-of-sequence token that is not present in the vocabulary. We disable automatic BOS insertion when the tokenizer has no BOS token and use explicit `add_special_tokens=False` scoring, ensuring consistent prompt likelihood computation across model families.

Determinism. Behavioral evaluation is fully deterministic given fixed model weights and tokenizer handling. Mechanistic split validation and item-mismatched source controls use a fixed random seed recorded in the notebook; the released CSVs are the exact outputs of that seeded run.