Lexical Redefinition Reveals Context-Sensitive Instruction Following

Roksana Goworek
Queen Mary University of London
r.goworek@qmul.ac.uk

Haim Dubossarsky
Queen Mary University of London
University of Cambridge
h.dubossarsky@qmul.ac.uk

Abstract

How do language models reconcile conflicting contextual evidence within a prompt? We study this question using lexical redefinition, where an explicit instruction specifies how a word should be interpreted and a subsequent sentence provides additional evidence through its usage. Across 870 word pairs, we find that lexical redefinitions are not applied as fixed rules. Instead, the behaviour depends systematically on the relationship between the explicit instruction and the intervening sentence: the contextualised usage can preserve, weaken, or completely reverse the instruction. These results suggest that the interpretation of an explicit instruction remains sensitive to subsequent contextual evidence. Lexical redefinition therefore provides a controlled behavioural paradigm for studying how language models integrate multiple sources of information within a prompt.

1 Introduction

Prompted language models receive multiple sources of contextual information. Some are explicit, such as instructions that specify how a task should be performed (Mishra et al., 2022). Others are implicit, arising from examples, demonstrations, or ordinary language use elsewhere in the prompt (Brown et al., 2020; Min et al., 2022). These sources often agree, but they can also conflict in subtle ways that change how an explicit instruction should be interpreted (Camassa and Shiller, 2026; Xu et al., 2024). Understanding how language models reconcile such competing contextual evidence is central to understanding instruction following and in-context reasoning.

Humans likewise rarely interpret instructions in isolation. After hearing "Put the kettle on" a subsequent remark such as "I have some great biscuits" or "Do we have some orange juice left?" is naturally understood as relevant to the earlier instruction, even though it does not explicitly refer to it. In the first case, the remark explains why the instruction was given; in the second, it changes what the listener understands the speaker intended, without explicitly revising the original instruction.

Whether language models perform a similar kind of contextual reconciliation is much less clear. Existing work shows models do not treat all contextual information equally. They may disregard in-context demonstrations in favour of pretraining priors (Min et al., 2022), rely on the semantic properties of labels rather than newly specified mappings (Wei et al., 2023), or trust contextual information selectively depending on whether it agrees with stored knowledge (Xu et al., 2024). However, these studies mainly ask if models use particular sources of contextual information. Much less is known about how models reconcile multiple sources when these point toward different interpretations.

We use a lexical redefinition paradigm as a controlled experimental probe. A redefinition — "the word 'mouse' should be interpreted as 'letter'" (illustrated as) — asks the model to override a pretrained association on command. We then insert a single intervening sentence that either demonstrates the instructed remapping of meaning, uses the redefined word in its original ordinary sense, or demonstrates the opposite remapping. This creates a controlled competition between two sources of contextual evidence: an explicit instruction that states how a word should be interpreted, and a demonstrated usage that shows how it is interpreted. Because these two sources can be manipulated independently while everything else is fixed, lexical redefinition provides a minimal setting for studying how models reconcile competing contextual evidence.

We study the effect of lexical redefinition on Qwen2.5-14B-Instruct across 870 word pairs, and show that the effect depends on the relationship between the explicit instruction and subsequent contextual evidence. The same instruction produces

Proceedings of the 3rd Workshop on Language Understanding in the Human-Machine Era (LUHME 2026), pages 215–226
October 28, 2026 ©2026 Association for Computational Linguistics

different behaviour depending on how it is related to the intervening sentence.

Our findings suggest that explicit instructions are reconciled with implicit contextual evidence, rather than processed independently. We report that a single sentence can flip an explicit instruction when its usage points the other way, and conclude that this may indicate that the model's output tracks the *relationship* between these two sources, not the instruction in isolation.

2 Related Work

A closely related line of work studies conflicts between explicit instructions and patterns demonstrated in context. Camassa and Shiller (2026) show that models can shift from instruction following to pattern following after repeated contradictory demonstrations, and that this susceptibility is not well predicted by standard capability benchmarks. Our setting differs by introducing competing evidence through a single intervening sentence, varying its relation to the instruction beyond direct conflict, and focusing on how later context changes the interpretation of an instruction rather than whether the model simply complies with it.

2.1 Conditional Use of Prompt Information

Prior work shows that models use prompt information selectively rather than learning arbitrary mappings from context. Min et al. (2022) find that randomising input–label mappings in demonstrations often has little effect, suggesting that models rely more on label semantics, input distributions, and task format to identify latent concepts acquired during pretraining (Xie et al., 2022). Later studies also show that pretrained semantic priors constrain the mappings models acquire in context (Wei et al., 2023; Krishna Kumar, 2025; Jang et al., 2024).

Our setting shifts the focus from whether contextual information is used to how multiple relevant cues interact. Unlike work on irrelevant or distracting context (Shi et al., 2023; Yoran et al., 2023), the intervening sentence is always pertinent to the redefinition but varies in its relation to it. We therefore frame prompt interpretation as an interaction between an explicit rule and a single demonstrated usage, showing that model behaviour depends on the relationship between these two in-context sources. This also differs from knowledge-conflict settings (Xu et al., 2024): both competing signals are contextual, while parametric knowledge provides only the default mapping that the redefinition attempts to override.

2.2 Semantic Plausibility and Anomaly

Prior work shows that language models are sensitive to semantic plausibility, but less reliably capture finer-grained contextual expectations (Ettinger, 2020; Kauf et al., 2023). Models can also interpret anomalous but coherent sentences when an appropriate context is available (Olsen and Padó, 2026). Our stimuli involve a different form of anomaly: rather than violating world knowledge, they place one word in a context where another would ordinarily occur, creating a distributionally unexpected but interpretable usage.

We use this controlled anomaly as contextual evidence about a lexical redefinition. Its effect depends on its relation to the explicit instruction: ordinary usage has limited influence, whereas an anomalous usage that supports an alternative mapping can substantially alter or override the instructed interpretation.

3 Methods

3.1 Stimuli

We use 30 words, evenly selected from three part-of-speech groups (10 nouns, 10 verbs, 10 adjectives) and chosen to reflect various lexical properties such as concreteness, animacy, and polysemy. We construct all $30 \times 29 = 870$ ordered pairs, since $X \rightarrow Y$ and $Y \rightarrow X$ are distinct redefinitions. The full word list is given in Appendix 2. For each word, Gemini-2.5-Pro generates 10 natural sentences of 8–15 words in which the word occurs naturally.[1]

Ordinary and Anomalous Sentences These sentences generated by Gemini serve as *ordinary* intervening sentences and as templates for *anomalous* variants. For a pair such as (*mouse*, *letter*), the ordinary sentence contains the first word in its natural context (*"The mouse eats crumbs"*), while the anomalous variant replaces it with the second word (*"The letter eats crumbs"*). Thus, the surrounding context remains fixed and only the target word changes. Sentences are presented without an explicit link to the redefinition.

3.2 Prompts

Depending on the condition, each sentence is preceded by a redefinition prompt of the form:

[1] Shorter examples are used for illustration (e.g., *"The mouse eats crumbs"*); the actual sentences are longer.

"In the following sentences, the word '⟨A⟩' should be interpreted as if it were the word '⟨B⟩'. Please answer questions based on this interpretation."

We tested four alternative prompt formulations on 56 word pairs to ensure that the results do not depend on this specific prompt. The full prompts can be found in Table 3 of the Appendix.

3.3 Sentence Completion Task

Depending on the condition, (🐁→📄, 📄→🐁,or "∅") each prompt or intervening sentence is followed by an incomplete sentence-completion query whose natural continuation is one of the two words in the pair (e.g. *"The cat chased the tiny gray ___"* for *mouse*). We record the probability assigned to each word as the next token. Queries are generated using Claude Opus 4.7 (one query per word), and reviewed manually to ensure the intended word is the natural completion and the sentence reads naturally.

3.4 Experimental Design

The design is fully counterbalanced. For every word pair, we include both redefinition directions and construct ordinary and anomalous sentences for both words. Thus, alongside an ordinary *mouse* sentence and an anomalous *letter* sentence, the dataset also contains an ordinary *letter* sentence and an anomalous *mouse* sentence. The same applies to all 30 words. Sentence-completion queries are counterbalanced in the same way. We use only one orientation of the *mouse–letter* example in the main text for readability, but all analyses include the corresponding reverse cases and group results by experimental role rather than lexical identity.

3.5 Model

We evaluate Qwen2.5-14B-Instruct. Exact Hugging Face checkpoint identifiers and details of the probability extraction procedure, including aggregation across whitespace and capitalisation variants, are provided in Appendix B.

3.6 Analyses

All analyses below are based on item analyses, where we compare probabilities of the same pairs across the different conditions. This allows us to directly, and fairly, measure the effects of the experimental conditions, as variance associated with different base probabilities across the 870 pairs is cancelled out.

Token Probabilities Each experimental condition is defined by a redefinition prompt, an intervening sentence (or none), and a query. We evaluate each condition on ten sentence instances and report the mean probability

$$P(t) = \frac{1}{10}\sum_{i=1}^{10} p^{(i)}(t), \tag{1}$$

where $p^{(i)}(t)$ is the probability assigned to token t for sentence instance i. Throughout, $P(t)$ denotes this condition-level average.

Sentence Effect To isolate the effect of the intervening sentence, we compare each condition with an otherwise identical prompt containing no intervening sentence (only the redefinition prompt and completion query). We define the sentence effect as the difference in completion probability between the sentence and matched no-sentence conditions, holding the redefinition and query fixed:

$$\Delta P(t) \;=\; P_{\text{sent}}(t) - P_{\text{no-sent}}(t). \tag{2}$$

This is the quantity reported in Figure 2.

Sign Agreement As a descriptive measure of directional consistency across word pairs, we report sign agreement. For each condition c and candidate token t, we first compute the aggregate mean sentence effect $\overline{\Delta P}_c(t)$ across the 870 ordered word pairs. We then measure the proportion of pairs whose individual sentence effect has the same direction as this aggregate mean. Pairs with exactly zero effect contribute 0.5, reflecting no evidence for either direction. Formally,

$$A(t) = \frac{1}{870}\sum_{j=1}^{870} s_j(t), \tag{3}$$

where j indexes word pairs and

$$s_j(t) = \begin{cases} 0.5, & \text{if } \Delta P_j(t) = 0, \\ 1, & \text{if } \operatorname{sign}(\Delta P_j(t)) = \operatorname{sign}(\overline{\Delta P}(t)), \\ 0, & \text{otherwise.} \end{cases}$$

Here, $\Delta P_j(t)$ is the sentence effect for word pair j on the probabilities of token t, and $\overline{\Delta P}(t)$ is the mean sentence effect across all pairs. Values near 50% indicate little consistency in the direction of the effect, whereas values approaching 100% indicate that most pairs change in the direction of the mean. The percentages are shown beside the corresponding bars in Figure 2.

Aggregation Probabilities are first averaged across the ten sentence instances belonging to a word pair, yielding one score per pair. Means and standard deviations are then computed across the 870 word pairs. This reduces variation due to the particular sentence instances, and standard deviations reflect differences between word pairs.

4 Results

We report main results in the text, and present the prompt-variation results in Appendix E. We report effect magnitudes and directional agreement across word pairs descriptively. Because lexical items, sentences, and completion queries recur across ordered pairs, the pair-level observations are not fully independent (see Limitations section).

4.1 Redefinition Effect

Before introducing intervening sentences, we verify that the redefinition instruction alone changes the model's completions as expected.

Figure 1 shows the averaged completion probabilities before and after the redefinitions were applied. Instructing the model with the redefinition 📄→🐁 ("letter means mouse") makes it generate responses consistent with the instructed mapping. For the query *"The cat chased the tiny gray ___"*, where mouse is the natural completion, the redefinition increased P(📄) from 0% to 59%, while the probability of what was the original correct completion (🐁) dropped from 18% to 2%. Similarly, the redefinition 🐁→📄 increased P(🐁) from 0% to 59% for the query *"She put a stamp on the envelope and mailed the ___"*, where letter is the natural completion, while dropping the original correct completion P(📄) from 18% to 2%. By completely flipping the probabilities between

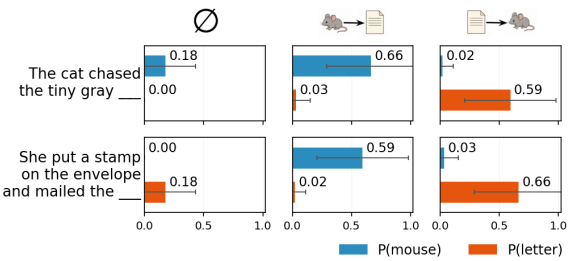

Figure 1: Token probabilities for sentence-completion queries under redefinition prompts and the original prompt without redefinition ("∅"). Sentences and icons are illustrative, results are aggregated over 870 word pairs. Error bars show SD.

the correct sentence completion in the original case and the redefinition, our results show that the model applied the instructions correctly.

Furthermore, the same instructions do not "flip" the probabilities in cases where their redefinitions do not make sense. 📄→🐁 only works for queries with mouse sentences and not letter sentences, and conversely, 🐁→📄 only works for letter sentences and not mouse sentences. Instead, in these cases we observe an increase in P(🐁) and P(📄) relative to their baseline. We interpret this as a priming effect (Kassner and Schütze, 2020), where a token that appears in a previous context can affect the generated probabilities. Because subsequent analyses measure sentence effects relative to matched no-sentence baselines (Eq. 2), this priming effect is controlled for by design.

4.2 Intervening Sentence Effect

We next analyse the effect of the intervening sentence on the next token probabilities for the sentence-completion query. We stress that the sentence is simply placed there, without any explicit instructions. A priori, one may think that the intervening sentence would imply a meaning associated with the particular usage of the word in the sentence. An anomalous sentence of the form ✉️ implies that *letter* means *mouse* (a different meaning) by showing an example of letter in a sentence where mouse would normally occur. Conversely, an ordinary sentence 🐁 implies that *mouse* means *mouse* (its current meaning) by showing it in its natural context. If the model follows the *meaning* that is implied by the sentence (but never explicitly stated like in the redefinition), we would expect probabilities assigned to tokens to change, compared with the matched condition containing no intervening sentence. For example, since the ✉️ sentence implies that 📄→🐁, then if the model follows this meaning implicature, we would expect it to complete the sentence "The cat chased the tiny gray ___" with *letter*, in line with the implied meaning, and similar to the redefinition effect of the equivalent mapping observed in § 4.1.

Contrary to that expectation, we observe little evidence of the implied mapping in the ∅ column of Figure 2, which shows the effect of the sentence on completion without any redefinition instruction. In fact, presenting the ✉️ sentence before asking the model to complete the sentence "The cat chased the tiny gray ___" results in a 5-percentage-point-increase in P(🐁), relative to the no-sentence

baseline (shown in Figure 1), with little corresponding change in P(📄). The pattern therefore does not support adoption of the mapping implied by the sentence: the ordinary completion increases more than the substituted word.

We therefore find no evidence that presenting a sentence by itself has a meaningful impact on the model's next token probabilities. It seems that the model does not take on the implied redefinition of the sentence and instead outputs nearly the same token probabilities as it did without any sentence.

4.3 Sentence x Redefinition Interactions

We next examine the interaction between the redefinition and the intervening sentence. Although the prompt does not explicitly link them, the sentence substantially alters predictions after a redefinition, despite having almost no effect on its own.

The direction and magnitude of this interaction depend on how the sentence relates to the instructed mapping. Combining the two redefinition directions with the two sentence types yields four distinct sentence–instruction relations: reinforcing, contradicting, competing, and compatible. These relations are shown in Figure 2 and summarised in Table 1 and constitute the paper's main finding.

4.3.1 Anomalous Intervening Sentences

An anomalous intervening sentence places one word in a semantic context where the other would ordinarily occur, thereby implying a lexical mapping without stating it explicitly. For example, ✉ uses *letter* in a mouse context and therefore implies 📄→🐁. This implied mapping provides a possible explanation for the different effects observed in the reinforcing and contradicting conditions.

Reinforcing In this condition, the intervening sentence produces almost no change to the completion probabilities relative to the no-sentence redefinition condition. For example, under 📄→🐁 followed by ✉, the model applies the redefinition with approximately the same strength as when no intervening sentence is present, and the token probabilities remain largely unchanged. A possible explanation is that the anomalous sentence implies the same mapping as the explicit instruction. The redefinition states 📄→🐁, and ✉ also suggests that *letter* denotes the mouse concept. Because the explicit and implied mappings align, the sentence introduces no competing evidence and therefore does not alter the model's response.

Contradicting In this condition, the intervening sentence produces a reversal in the model's predictions. When 🐁→📄 is followed by ✉, P(🐁) decreases sharply while P(📄) increases substantially (see the bottom-left quadrant of Figure 2). The probabilities do not merely shift; they flip, closely resembling the pattern observed under the opposite 📄→🐁 redefinition in the no-sentence condition (Figure 1). A possible explanation is that the model is acting upon the reverse of the instructed mapping implied by the anomalous sentence. The instruction states 🐁→📄, whereas ✉ suggests 📄→🐁. Because these mappings assign opposite word–concept relations, they directly contradict one another. The observed probability reversal could thus reflect the model resolving this conflict in favour of the mapping implied by the intervening sentence.

Sentence	🐁→📄	📄→🐁
"The mouse eats crumbs" $\therefore$ *implies* 🐁→🐁	**Compatible** uses 🐁 in its ordinary sense, but the redefinition affects only 📄 contexts → redefinition unaffected or slightly weakened	**Competes** uses 🐁 in its ordinary sense, which was redefined to be denoted by 📄 → 🐁→📄 adopted instead
"The letter eats crumbs" $\therefore$ *implies* 📄→🐁	**Contradicts** shows an opposite mapping to the one in the instruction → 📄→🐁 adopted instead	**Reinforces** shows an example of the redefinition → redefinition effect is unchanged

Table 1: Analysis of the four sentence–redefinition combinations. Each cell names the relationship between the sentence and the redefinition instruction, and the observed effect on the completions.

4.3.2 Ordinary Intervening Sentences

Ordinary intervening sentences differ from anomalous ones because they do not imply a new lexical mapping, but preserve an existing word–concept relation. Their effect on completion probabilities depends on whether this relation concerns the concept targeted by the redefinition or the word being redefined. An ordinary sentence either shows that the concept targeted by the redefinition is still expressed by its original word, or that the redefined word still retains its original meaning. These two cases correspond to the competing and compatible conditions, respectively.

Competing A probability reversal similar to that in the contradicting condition, though arising for different reasons, occurs here. Under 📄→🐁, inserting the ordinary sentence 🐁 flips the preferred completion probabilities from *letter* back to *mouse*. The intervening sentence presents the 🐁 concept as being expressed by its original word, *mouse*, whereas the redefinition instructs the model to express that concept using *letter*. It therefore provides direct evidence against the instructed mapping, but without directly showing the opposite mapping instead.

Compatible A different pattern is observed for 🐁→📄 followed by the ordinary sentence 🐁. Here, the intervening sentence implies that *mouse* retains its ordinary meaning, but provides no evidence about which word should denote the 📄 concept. Nevertheless, the results show that the preference for *mouse* decreases significantly, while the probability of *letter* does not change correspondingly and the token probabilities do not flip. We interpret this as a dilution effect. The redefinition initially makes *mouse* a strong candidate for the letter concept, but the ordinary sentence provides evidence that *mouse* also retains its original meaning. This makes its association with the letter concept less specific, reducing P(🐁).

5 Discussion

Lexical redefinition provides a controlled setting for studying how language models reconcile multiple sources of contextual evidence. Across our experiments, the model did not treat an explicit redefinition as either a fixed rule or an instruction to be ignored. Instead, it integrated the instruction with evidence supplied by a later intervening sentence, and its predictions depended systematically

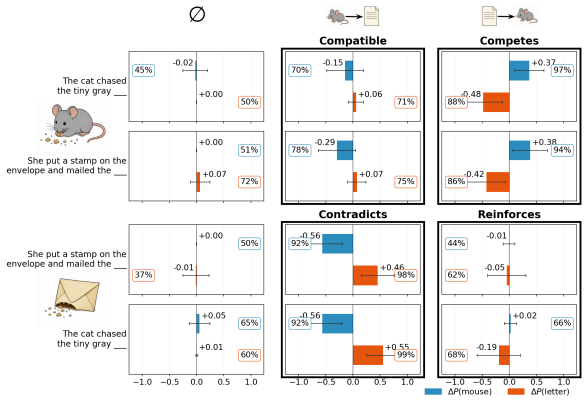

Figure 2: Effect of the intervening sentence on completion probabilities (ΔP, Eq. 2). Blue/orange bars show ΔP for *mouse*/*letter*. Percentages inside boxes indicate sign agreement with the mean across word pairs (Eq. 3).

on the relation between the two sources.

This pattern cannot be explained by either source alone. Redefinitions strongly altered completion probabilities in the expected direction, whereas intervening sentences had little effect without an explicit instruction. Yet when combined, a single sentence could preserve, attenuate, or reverse the redefinition effect. The interaction was also robust across alternative prompt formulations, suggesting that it reflects the relation between instruction and usage rather than a particular wording.

We propose four categories—reinforcing, contradicting, competing, and compatible—as an explanatory framework for these distinct behavioural patterns (Table 1). A reinforcing sentence leaves the instruction-induced preference largely unchanged. Both contradicting sentences, which support the reverse mapping, and competing sentences, which preserve the target concept's conventional name, reverse that preference, although with different magnitudes. Compatible sentences instead produce a weaker attenuation without changing the preferred completion.

The same ordinary and anomalous sentences can have sharply different effects depending on the redefinition they follow, and these relations are asymmetric. Although the model was not instructed to compare the two sources, its predictions varied systematically with the specific logical relation between the mappings supported by the instruction and the intervening sentence.

This pattern is consistent with later context revising the behavioural consequences of an earlier instruction, suggesting that prompt interpretation remains open to revision rather than being fixed at

the point of instruction. However, the fit between the results and our four-category framework does not establish that models explicitly represent or compute these categories. Rather, the framework provides a compact logical account of how different forms of agreement and disagreement between explicit rules and contextual evidence produce distinct behavioural outcomes.

Unlabelled Context Influences Instruction Use The intervening sentence is never introduced as evidence about the redefinition. Nevertheless, it systematically modulates the effect of the earlier instruction. This invites comparison with relevance-theoretic accounts of human communication, according to which communicative inputs create expectations of relevance and are interpreted against available contextual assumptions (Sperber and Wilson, 1995; Wilson and Sperber, 2004). We intend this only as a behavioural analogy: the present results do not show that language models implement the cognitive mechanisms proposed by relevance theory, leaving this as an open question.

Implications An explicit instruction is therefore not necessarily the strongest signal in a prompt. This is in line with Camassa and Shiller (2026), who show that sustained in-context demonstrations can override an explicit instruction; here, the same reversal is produced by a single intervening sentence. The result is also compatible with accounts in which in-context behaviour depends on properties such as the demonstrated label space, sequence format, or inference over a latent task concept (Min et al., 2022; Xie et al., 2022). More generally, prompt-injection research shows that models can treat contextual information as behaviour-changing instructions even when that data is not intended to have instructional authority (Greshake et al., 2023). Recent work on in-context representation hijacking shows that repeated lexical substitutions can shift how a token is represented and interpreted in adversarial prompts (Yona et al., 2026). Our results identify a related sensitivity in a controlled, non-adversarial setting. Whether it extends to realistic security scenarios remains a topic for future work.

Lexical Semantic Change Research Computational approaches to lexical semantic change are grounded in the assumption that differences in usage provide evidence for differences in meaning (Kutuzov et al., 2018; Goworek and Dubossarsky, 2026). Recent work uses computational models, including LLMs, to make usage-level semantic judgements (Schlechtweg et al., 2025; de Sá et al., 2026; Spaziani et al., 2026). Although our experiments do not directly evaluate LSC annotation, they point to a consideration for this emerging methodology: an LLM's interpretation of a usage may depend not only on the usage itself, but also on other information in the prompt. As LLM-based annotation becomes increasingly capable and widely adopted, improvements in apparent annotation accuracy may be worth considering alongside possible context-dependent artefacts. Future work could test biases in LLM-based semantic-change judgements due to preceding context, prompt formulation, and demonstrations.

Open questions The observed responses are consistent with the model combining explicit instructions with later contextual evidence, but they do not distinguish among several possible explanations. The model might infer a lexical mapping, condition on demonstrated usage or rely on broader pragmatic expectations. For example, contextual entrainment increases the probability of tokens that appeared earlier in the prompt (Niu et al., 2025). Determining whether the present results rely on this process, a semantic mapping process, or an interaction between them requires mechanistic analysis.

Existing work shows that in-context demonstrations can induce task- or function-level internal representations that causally affect model behaviour (Hendel et al., 2023; Todd et al., 2024). Similar analyses could test whether lexical redefinitions and intervening usages produce separable representations, and how those representations interact.

More broadly, it remains unclear how specific the phenomenon is to lexical redefinition. The present experiments isolate competition between an explicit instruction and an implicitly demonstrated usage, but similar tensions arise in many prompting settings. Anomalous-but-interpretable sentences may make this competition especially visible; related work shows that language models can often construct supporting contexts that render anomalous sentences interpretable (Olsen and Padó, 2026). Future work should test how far the effect generalises across instruction types, languages, word classes, invented vocabularies, and model scales. Lexical redefinition may therefore provide a useful controlled instance of a broader form of interaction between explicit instructions and contextual evidence.

6 Conclusion

In this paper, we use lexical redefinition as a controlled probe of how competing sources of contextual evidence are reconciled. An explicit redefinition and an intervening sentence each provide evidence about how a word should be interpreted, and varying them independently shows that these two sources are integrated rather than the redefinition being followed literally. A sentence that has no effect on its own can preserve, weaken, or reverse the instruction depending on the relationship between them. We capture these patterns with a four-category explanatory framework—reinforcing, contradicting, competing, and compatible—even though the model was never instructed to distinguish these relations. Its outputs nevertheless appear to track this logical structure. This suggests that models may treat redefinitions not as fixed rules to be executed but as evidence that remains open to revision. The paradigm reduces a broad question about in-context reasoning to a minimal setting in which the competing signals can be varied and measured directly.

Limitations

The lexical redefinition paradigm isolates a simple form of conflict between an explicit instruction and a single piece of contextual evidence. This controlled design enables systematic comparisons, but it remains unclear how directly the results extend to natural prompting settings, in which instructions, demonstrations, and contextual cues are more varied. The experiments also use a small hand-selected English vocabulary and short sentences. The prompt-wording variation analysis is restricted to a 56-pair subset. The present study should therefore be understood as a controlled behavioural demonstration rather than evidence of broad generality.

Model Scope We evaluate only a single model, Qwen2.5-14B-Instruct. The extent to which the observed effects generalise to other model families, parameter scales, training regimes, or multilingual systems therefore remains unclear. In particular, larger models or models with different instruction-tuning procedures may resolve semantic conflict differently. We also measure direct next-token predictions without eliciting intermediate reasoning, so the reported context dependence may change under alternative inference procedures.

Construct of the Redefinition Our instruction targets semantic interpretation (*"A should be interpreted as B"*), whereas the sentence-completion task measures lexical production. Completion probabilities therefore provide a behavioural proxy for whether the model applies the redefinition, but they do not uniquely distinguish a change in semantic interpretation from a change in production preference or response strategy. This does not affect the observed context sensitivity of the completion probabilities, but it limits the conclusions that can be drawn about changes to the model's internal representation of word meaning. Moreover, the task measures next-token probabilities for two prespecified candidates rather than the model's full output distribution or an independent measure of lexical interpretation. Some variability, particularly in non-matching queries, may therefore reflect a mismatch between the instructed construct and the elicited behaviour. Denotational prompts or tasks that separate interpretation from production could help disambiguate these possibilities.

Stimuli We construct 870 ordered pairs from a hand-selected set of 30 words spanning nouns, verbs, and adjectives. This provides systematic combinatorial coverage but limited lexical diversity, and the pairs are not 870 wholly independent lexical items because the same words, sentences, and completion queries recur across ordered pairs. Because all words are crossed, many pairs also span parts of speech, so some substitutions introduce grammatical-category mismatches in addition to semantic anomaly.

The intervening sentences are generated with Gemini 2.5 Pro and reviewed manually, but generation at scale limits experimental control. We use one completion query per word, and the intended completion is not guaranteed to receive the highest baseline probability for every item. Some substitutions intended to be anomalous may also remain locally plausible (e.g. substituting *hot* for *cold*). Measuring sentence effects relative to matched no-sentence baselines mitigates but does not eliminate these sources of variation.

Behavioural Focus The study characterises output behaviour and does not identify the internal representations or computations responsible for the observed probability shifts. In particular, the results do not distinguish lexical reinterpretation from contextual entrainment, production-level priming, or other forms of contextual conditioning. Mechanis-

tic analysis would be required to determine how the instruction and intervening sentence are represented and how they interact during inference.

Acknowledgements

This work has in part been funded by the research program Change is Key! supported by Riksbankens Jubileumsfond (under reference number M21-0021).

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Carolina Camassa and Derek Shiller. 2026. Do as i say, not as i do: Instruction-induction conflict in LLMs. *Preprint*, arXiv:2605.20382.

Jáder Martins Camboim de Sá, Jooyoung Lee, Marcos Da Silveira, and Cedric Pruski. 2026. Semantic change characterization with LLMs using rhetorics. In *The Proceedings for the 6th International Workshop on Computational Approaches to Language Change (LChange'26)*, pages 110–123, Rabat, Morocco. Association for Computational Linguistics.

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*, 8:34–48.

Roksana Goworek and Haim Dubossarsky. 2026. Rethinking metrics for lexical semantic change detection. In *The Proceedings for the 6th International Workshop on Computational Approaches to Language Change (LChange'26)*, pages 147–161, Rabat, Morocco. Association for Computational Linguistics.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. *arXiv preprint arXiv:2302.12173*.

Roee Hendel, Mor Geva, and Amir Globerson. 2023. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9318–9333, Singapore. Association for Computational Linguistics.

Joonwon Jang, Sanghwan Jang, Wonbin Kweon, Minjin Jeon, and Hwanjo Yu. 2024. Rectifying demonstration shortcut in in-context learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4294–4321, Mexico City, Mexico. Association for Computational Linguistics.

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*, pages 7811–7818, Online. Association for Computational Linguistics.

Carina Kauf, Anna A. Ivanova, Giulia Rambelli, Emmanuele Chersoni, Jingyuan Selena She, Zawad Chowdhury, Evelina Fedorenko, and Alessandro Lenci. 2023. Event knowledge in large language models: The gap between the impossible and the unlikely. *Cognitive Science*, 47(11):e13386.

Anantha Padmanaban Krishna Kumar. 2025. Semantic anchors in in-context learning: Why small LLMs cannot flip their labels. *arXiv preprint arXiv:2511.21038*.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

Jingcheng Niu, Xingdi Yuan, Tong Wang, Hamidreza Saghir, and Amir H. Abdi. 2025. Llama see, llama do: A mechanistic perspective on contextual entrainment and distraction in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16218–16239, Vienna, Austria. Association for Computational Linguistics.

Katrina Olsen and Sebastian Padó. 2026. Finding sense in nonsense with generated contexts: Perspectives from humans and language models. In *Proceedings of the 15th Joint Conference on Lexical and Computational Semantics (*SEM)*, San Diego, California, USA. Association for Computational Linguistics.

Dominik Schlechtweg, Tejaswi Choppa, Wei Zhao, and Michael Roth. 2025. CoMeDi shared task: Median judgment classification & mean disagreement ranking with ordinal word-in-context judgments. In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 33–47, Abu Dhabi, UAE. International Committee on Computational Linguistics.

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*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR.

Elena Spaziani, Kamyar Zeinalipour, Pierluigi Cassotti, and Nina Tahmasebi. 2026. Elections go bananas: A first large-scale multilingual study of pluralia tantum using LLMs. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6547–6570, Rabat, Morocco. Association for Computational Linguistics.

Dan Sperber and Deirdre Wilson. 1995. *Relevance: Communication and Cognition*, 2 edition. Blackwell, Oxford.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. 2024. Function vectors in large language models. In *The Twelfth 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*.

Deirdre Wilson and Dan Sperber. 2004. Relevance theory. In Laurence R. Horn and Gregory Ward, editors, *The Handbook of Pragmatics*, pages 607–632. Blackwell, Oxford.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit Bayesian inference. In *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. *arXiv preprint arXiv:2403.08319*.

Itay Yona, Amir Sarid, Michael Karasik, and Yossi Gandelsman. 2026. In-context representation hijacking. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16852–16867, San Diego, California, United States. Association for Computational Linguistics.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2023. Making retrieval-augmented language models robust to irrelevant context. *arXiv preprint arXiv:2310.01558*.

A Stimuli

The 30-word set is hand-selected to balance parts of speech (10 nouns, 10 verbs, 10 adjectives; Appendix 2). The 8-word subset used for the prompt-variant experiments is {chair, dog, flower, bank, river, cook, stone, cloud}, giving $8 \times 7 = 56$ ordered pairs.

Nouns	Verbs	Adjectives
truth	run	hot
chair	know	cold
dog	break	big
flower	give	small
bank	think	good
river	find	bad
cook	sleep	dead
stone	build	happy
cloud	running	fast
time	knowing	red

Table 2: The 30-word stimulus set by part of speech.

B Implementation details.

We use the Hugging Face checkpoint `Qwen/Qwen2.5-14B-Instruct`. Completion probabilities are obtained by applying a softmax to the model logits at the final input position. To account for tokenisation and capitalisation differences, for each candidate word we tokenise four surface forms: lowercase and initial-capitalised variants, each with and without a leading whitespace. We take the first token ID of each form and define the candidate probability as the sum of the probabilities assigned to the resulting distinct token IDs. This measures the probability that the next generated token begins with one of these surface forms of the candidate word.

C Prompt Variation

To assess whether the main results depend on the exact phrasing of the redefinition instruction, we repeated the experiment on 56 word pairs using four paraphrased prompts with the same intended meaning. The prompt variants are listed in Appendix 3, and the corresponding results are shown in Appendix 4-7. The qualitative pattern is consistent across all prompt variants, indicating that the interaction between the redefinition and the intervening sentence is robust to instruction wording.

D Raw probabilities

Figure 3 gives the underlying raw next-token probabilities, with no normalisation. The grid adds, relative to the main figures, two no-sentence rows which are used as controls in the main text.

E Prompt phrasing results

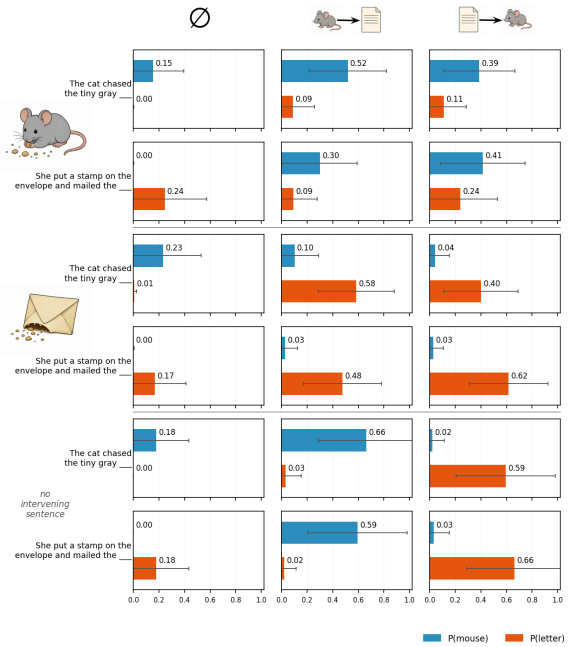

Figure 3: Qwen2.5-14B-Instruct: raw next-token probabilities. Rows: context sentence (icon) × query, plus two no-sentence rows; columns: redefinition condition (icon). Bars: P(mouse)= w_1, P(letter)= w_2; ± 1 SD over pairs.

Wording	Prompt (X, Y stand for the word pair)
baseline	In the following sentences, the word ‘X’ should be interpreted as if it were the word ‘Y’.
scope_sentence_only	For the next sentence only, ‘X’ should be interpreted as if it were ‘Y’.
scope_global	For the rest of this conversation, ‘X’ should be interpreted as if it were ‘Y’.
denotational	Whenever ‘X’ appears, read it as ‘Y’.
comprehension_only	Treat ‘X’ as having the meaning of ‘Y’ when reading. Do not change how you produce these words; only how you interpret them.

Table 3: The baseline redefinition prompt and four alternative formulations.

context	query	∅					
		.10 ±.14	.00 ±.00	.07 ±.20	.01 ±.02	.34 ±.19	-.28 ±.26
		.00 ±.01	.07 ±.10	.05 ±.22	.02 ±.02	.30 ±.23	-.29 ±.25
		.00 ±.00	.10 ±.19	-.36 ±.28	.50 ±.21	.00 ±.01	.16 ±.23
		.04 ±.08	.01 ±.02	-.44 ±.28	.52 ±.27	.00 ±.02	.15 ±.28

Table 4: No-sentence-normalised sentence effect ΔP ($P_{\text{with sentence}} - P_{\text{no sentence}}$) for Qwen2.5-14B-Instruct under the `scope_global` wording (8-word variant subset). Rows: context sentence $\times$ query word; columns: redefinition direction $\times$ queried word. Each cell: mean ΔP over the 56 pairs; $\pm$ line is the SD over pairs.

context	query	∅					
		.10 ±.14	.00 ±.00	-.15 ±.16	.02 ±.03	.20 ±.15	-.21 ±.21
		.00 ±.01	.07 ±.10	-.08 ±.18	.02 ±.03	.11 ±.15	-.19 ±.22
		.00 ±.00	.10 ±.19	-.30 ±.23	.29 ±.18	.00 ±.01	-.10 ±.22
		.04 ±.08	.01 ±.02	-.35 ±.21	.21 ±.17	.02 ±.02	-.14 ±.20

Table 5: Prompt-variation results for Qwen2.5-14B-Instruct (`scope_sentence_only` wording).

context	query	∅					
		.10 ±.14	.00 ±.00	-.02 ±.24	-.09 ±.18	.22 ±.23	-.13 ±.18
		.00 ±.01	.07 ±.10	-.04 ±.13	-.01 ±.21	.25 ±.16	-.20 ±.23
		.00 ±.00	.10 ±.19	-.15 ±.19	.24 ±.26	-.13 ±.17	.11 ±.27
		.04 ±.08	.01 ±.02	-.24 ±.24	.26 ±.17	-.07 ±.21	.05 ±.16

Table 6: Prompt-variation results for Qwen2.5-14B-Instruct (`denotational` wording).

context	query	∅					
		.10 ±.14	.00 ±.00	.01 ±.24	.00 ±.01	.12 ±.10	-.17 ±.16
		.00 ±.01	.07 ±.10	.01 ±.23	.00 ±.01	.13 ±.12	-.12 ±.22
		.00 ±.00	.10 ±.19	-.33 ±.24	.36 ±.16	.00 ±.01	.08 ±.29
		.04 ±.08	.01 ±.02	-.38 ±.19	.35 ±.18	.00 ±.01	.00 ±.26

Table 7: Prompt-variation results for Qwen2.5-14B-Instruct (`comprehension_only` wording).