# Do Language Models Show Structural Priming Across Different Domains?

**So Young Lee**[†], **Russell Scheinberg**[◇], **Ameeta Agrawal**[◇]
[†]Miami University, USA
[◇]Portland State University, USA
soyoung.lee@miamioh.edu
{rschein2,ameeta}@pdx.edu

## Abstract

We test whether large language models show cross-domain structural priming by asking whether arithmetic expressions influence relative-clause attachment preferences. Experiment 1 examines English and French using materials based on prior psycholinguistic studies, and Experiment 2 extends the test to a larger multilingual dataset. Across both experiments, we find no robust priming effect. Instead, responses largely reflect baseline attachment preferences, which vary across languages and only partially align with human patterns. These findings suggest that, although language models show some structural sensitivity, they provide limited evidence of abstract structural generalization across domains.

## 1 Introduction

A central question in both cognitive science and natural language processing is whether the mechanisms that support language are domain-specific or domain-general. Research on human sentence processing has shown that structural priming—the tendency for prior exposure to a particular structure to bias subsequent interpretation—is not restricted to linguistic input alone, but can also be induced by structures from other domains, such as mathematics, logic, or music. These cross-domain effects suggest that human comprehenders recruit domain-general resources for representing and aligning hierarchical structures.

Recent NLP work increasingly treats language models as psycholinguistic test subjects and partial computational models of human sentence processing (Futrell et al., 2019; Wilcox et al., 2023; Cai et al., 2024).

Because LLMs are trained primarily on linguistic input, it remains unclear whether they exhibit priming effects that extend beyond within-language contexts. LLMs are trained on large text corpora

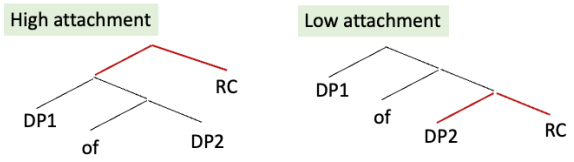


Figure 1: Syntactic Structures of DP1 of DP2 Modification (left) and DP2 Modification (right) in English

that include natural language, code and mathematical expressions, and they show substantial syntactic sensitivity within language, including long-distance agreement and other structure-sensitive contrasts (Linzen et al., 2016; Gulordava et al., 2018; Marvin and Linzen, 2018a; Warstadt et al., 2020). However, it remains unclear whether their structural representations are shared across domains in a way that supports cross-domain priming.

If models do show cross-domain priming, this would indicate that they encode abstract structural regularities that generalize across representational domains. If not, this would highlight a key boundary between human cognition and statistical language modeling. To investigate this issue, we focus on one of the most widely studied cases of syntactic ambiguity in human sentence processing: **attachment ambiguities**. For example, in (1), the relative clause (*who was on the balcony*) may attach either to the lower determiner phrase, DP2 (*the colonel*; low attachment), or to the higher determiner phrase, DP1 of DP2 (*the daughter of the colonel*; high attachment).

(1) The journalist interviewed the daughter of the colonel who was on the balcony.

Previous research has reported that English speakers tend to prefer low attachment. However, attachment preferences are not fixed: they can be modulated by prosody, lexical biases, discourse context, and, crucially, structural priming.

Cross-domain priming in sentence processing has been demonstrated in a series of psycholin-



*Proceedings of the 1st Workshop on Computational Developmental Linguistics (CDL)*, pages 37–51

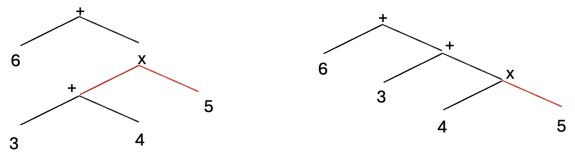


Figure 2: Hierarchical Structures of Arithmetic Expressions

guistic experiments. For instance, solving arithmetic expressions with nested groupings (e.g., $6 + (3 + 4) \times 5$) increases the likelihood of high relative clause attachment, whereas more linear expressions (e.g., $6+3+4\times5$) bias comprehenders toward low attachment (see Figure 2). These findings suggest that the mechanisms guiding syntactic disambiguation are sensitive to structural alignments across distinct representational domains. Rather than being tied exclusively to language, priming effects appear to reflect broader cognitive strategies for encoding and reusing hierarchical patterns.

LLMs, by contrast, have been observed to display priming-like behaviors only in linguistic contexts, such as persisting in syntactic choices across prompts (Prasad et al., 2019; Sinclair et al., 2022a; Michaelov et al., 2023; Jumelet et al., 2024a). Whether these effects extend to cross-domain contexts remains an open question. Because LLMs learn solely from textual corpora, they may not engage the same domain-general resources that humans recruit when transferring structural biases across domains. Testing whether LLMs exhibit cross-domain priming therefore provides a novel diagnostic of the extent to which their internal representations capture abstract structural parallels, or whether their priming behavior is confined to within-language statistical patterns. From a developmental perspective, the issue is not only whether language models eventually exhibit structure-sensitive behavior, but also what kind of representational organization emerges from their training experience. Human learners develop linguistic and non-linguistic reasoning abilities over time, and cross-domain priming has been taken as evidence that some aspects of hierarchical structure may be represented in a domain-general format. Language models provide a different kind of learner: they acquire behavior through large-scale exposure to text, code, mathematical expressions, and instruction-tuning. Testing whether arithmetic structure influences linguistic attachment therefore offers a way to ask whether model development gives rise to abstract, transferable structural representations, or whether linguistic and mathematical competencies remain functionally separated despite co-occurring in the training distribution. In this paper, we investigate this question by comparing human and LLM behavior in the same paradigm. Specifically, we ask whether models that have acquired both linguistic and mathematical competence show evidence of a shared structural representation across these domains. Building on prior psycholinguistic findings that mathematical grouping structures bias relative-clause attachment, we test whether contemporary LLMs show similar shifts. This comparison helps clarify the nature of priming, the extent of structural abstraction in LLMs, and the developmental trajectory through which such abstraction may or may not emerge in artificial learners.

## 2 Related Work

### 2.1 Structural Priming in Human Sentence Processing

Speakers tend to repeat syntactic structures they have recently encountered, a phenomenon known as *structural priming*, *syntactic persistence*, or *syntactic priming*. This effect has been widely studied as a window into how prior linguistic experience shapes subsequent behavior and what this reveals about the representations and memory mechanisms underlying language processing. Competing accounts attribute priming to residual activation, implicit learning, or interactions between the two (Bock, 1986b,a; Chang et al., 2000; Bock and Griffin, 2000; Fine and Jaeger, 2013).

Structural priming has been documented across many constructions and languages. Much of this evidence is consistent with accounts in which priming reflects sensitivity to local structural configurations, such as the arrangement of immediate phrasal constituents. Classic examples include the dative alternation and the active–passive alternation (Bock, 1986b; Pickering and Branigan, 1998; Bock and Loebell, 1990), along with a range of other constructions (e.g., Cleland and Pickering, 2003; Ferreira, 2003; Griffin and Weinstein-Tull, 2003; Hartsuiker and Westenberg, 2000). At the same time, more recent work suggests that priming is not limited to such local configurations. Studies of relative clause attachment ambiguity show priming between interpretations that differ only in hierarchical attachment, not in lexical content or linear order, suggesting that priming can also reflect sen-

sitivity to abstract structural relations (Desmet and Declercq, 2006; Scheepers, 2003).

Aligned with this view, cross-domain studies suggest that structural priming may extend beyond language itself. Scheepers et al. (2011) and Pozniak et al. (2018) showed that solving arithmetic expressions with different hierarchical groupings can bias subsequent relative clause attachment preferences in sentence processing. Because equations and sentences share no lexical or semantic content, these findings are difficult to explain in terms of lexical overlap or surface similarity. Instead, they have been interpreted as evidence that priming operates over abstract representations of hierarchical organization maintained in working memory (Scheepers et al., 2011; Pozniak et al., 2018).

Taken together, research on local configurations, global attachment, and cross-domain alignment suggests that structural priming reflects sensitivity to structured representations that extend beyond immediate word-order patterns.

### 2.2 Structural sensitivity in language models

Language model research has investigated whether these systems are sensitive to abstract linguistic structure, rather than relying only on surface-level distributional patterns. Researchers have approached this in several ways. For example, some studies have examined whether a model's internal representations encode syntactic dependencies or constructional information (Hewitt and Manning, 2019; Li et al., 2023; Tayyar Madabushi et al., 2020; White et al., 2021), while others have used targeted syntactic evaluation to test whether the model distinguishes between minimally contrasting grammatical and ungrammatical sentences (Gauthier et al., 2020; Marvin and Linzen, 2018b; Hu et al., 2020). Findings from this work suggest that language models can learn grammatical patterns that go beyond simple word-to-word associations. In other words, their behavior often reflects more than mere memorization of surface sequences. However, this evidence is still not fully conclusive. Showing that syntactic information can be recovered from a model's internal states does not necessarily mean that the model actively uses that information during prediction (Voita and Titov, 2020). Likewise, good performance on targeted syntactic evaluations does not always prove that the model is relying on abstract syntactic structure, since it may instead succeed by exploiting shallower lexical or distributional regularities. In addition, other work has pointed to weaknesses in areas such as word order, reliance on spurious heuristics, and the interpretation of negation (Kassner and Schütze, 2020; Lovering et al., 2021; Sinha et al., 2021). These findings suggest that language models do show some evidence of structural knowledge, but the depth, robustness, and functional role of that knowledge remain open questions.

### 2.3 Language models as developmental learners

The present study also connects to computational developmental linguistics, which treats learning systems as objects of developmental analysis rather than only as static predictors. From this perspective, language models are not simply evaluated for whether they know a particular syntactic contrast, but for what their behavior reveals about the emergence, organization, and limits of linguistic knowledge after large-scale training. This is especially relevant for cross-domain priming. If a model shows structural transfer from arithmetic to language, this would suggest that training has produced representations that abstract away from domain-specific surface forms. If it does not, the result suggests that linguistic and mathematical abilities may develop as partially or fully independent competencies, even in models that perform well in both domains. Thus, the present paradigm provides a developmental diagnostic: it asks whether exposure to multiple structured symbol systems leads to shared hierarchical representations or to functionally modular behavior across domains.

### 2.4 Structural priming in language models

Structural priming offers another way to study syntactic knowledge in language models. Prior work shows that language models do exhibit priming-like effects, including effects of recency, cumulative exposure, lexical overlap, and crosslinguistic persistence. At the same time, these effects do not always pattern like human priming: some studies report asymmetries across structural alternatives whose direction does not match the human literature. Structural priming is therefore informative not only because it provides further evidence that language models encode syntax, but also because it helps characterize how that knowledge is organized and how it may differ from human sentence processing.

Recent work including (Jumelet et al., 2024b; Sinclair et al., 2022b, 2026) has strengthened this

picture by showing that priming in neural language models cannot be reduced to simple surface repetition. In particular, (Sinclair et al., 2022b) reports reliable priming effects across a range of constructions and Transformer-based models, even when lexical overlap and semantic similarity are controlled. Priming is nevertheless sensitive to familiar psycholinguistic factors, weakening with distance and strengthening with repeated exposure and lexical or semantic similarity. This suggests that language models retain structural information across inputs, but that it remains closely tied to lexical, semantic, and distributional properties.

This raises a key question: ***Do priming effects in language models reflect abstract structural representations, or are they driven by domain-specific regularities?*** This question is especially pressing because existing work has focused almost entirely on within-language priming. It therefore remains unclear whether the representations supporting priming generalize across domains. The present study addresses this issue by asking whether exposure to structured input in mathematics influences the processing of structurally analogous input in language. Mathematics provides a particularly stringent test case because it is highly structured and compositional, yet differs sharply from language in surface form, semantics, and training distribution. If structural priming extends across these domains, that would support the existence of more domain-general structural representations in language models. If it does not, that would suggest important limits on the abstractness of their syntactic knowledge.

## 3 Experiments

### 3.1 Language Models

We evaluated five open-weight language models spanning two inference profiles: standard (non-reasoning) and reasoning (chain-of-thought), and two architectures, dense and mixture-of-experts (MoE). Table 1 summarizes the models and their key properties.

Models were selected to vary along two dimensions relevant to the study. First, we contrast **standard** models, which generate responses directly, with **reasoning** models, which produce an internal chain-of-thought before outputting the final answer. This distinction is important because structural priming would require the model to process the equation's hierarchical structure; reasoning models may engage with that structure more deeply due to their step-by-step computation. Second, we include both **dense** architectures (all parameters active on every token) and **mixture-of-experts** (MoE) architectures (only a subset of parameters active per token), allowing us to assess whether the total parameter count or the active computation determines math accuracy and attachment behavior.

All models were accessed via the Fireworks AI inference API.[2] Qwen3's default thinking mode was used and the <think> tags stripped before processing outputs, and GPT-OSS models were set to medium thinking effort. Experimental runs were logged with Langfuse for reproducibility.

### 3.2 Stimuli

Pozniak et al. (2018) and Scheepers et al. (2011) found that mathematical expressions can influence relative clause attachment in English and French. Building on this work, Experiment 1 adopted their stimulus materials in a Pozniak-style priming experiment. Specifically, we used 60 experimental item sets in Pozniak et al. (2018) (30 item sets in each language) to enable direct comparison with the human results. Each item consisted of two types of prime equations (2) and a written sentence as the target material (3).

(2) a. $90 - (9 + 1) \times 5$ HA prime

b. $90 - 9 + 1 \times 5$ LA prime

(3) Voici le tailleur de l'architecte qui
here the tailor of the architect who
s'apprête à créer un chef d'œuvre.
is about to create a masterpiece
'Here we have the tailor of the architect who is about to create a masterpiece.'

Note that we began with the English translations provided alongside the French stimuli and subsequently refined them to more closely match the French originals. Because the French materials encode grammatical gender on nouns, we took care to minimize any additional effects of gender and pronoun resolution in the English stimuli. Specifically, we replaced gender-marked forms with gender-neutral alternatives where possible, for example using *the* in place of *his* or *her* and *I* in place of *he* or *she*.

We then extended the investigation to a broader multilingual setting in Experiment 2, a MultiWho

[2] https://fireworks.ai

| Model | Developer | Total | Active | Type |
|---|---|---|---|---|
| Llama 3.3 70B Instruct | Meta | 70B | 70B | Standard |
| Qwen3 8B | Alibaba | 8B | 8B | Reasoning |
| Kimi K2 Instruct | Moonshot AI | 1T | 32B | Standard (MoE) |
| gpt-oss-120b | OpenAI | 117B | 5.1B | Reasoning (MoE) |
| gpt-oss-20b | OpenAI | 21B | 3.6B | Reasoning (MoE) |

Table 1: Models used in the experiments. *Total* = total parameter count; *Active* = parameters activated per forward pass (relevant for MoE models). Model cards are available on Hugging Face.[1]

multilingual extension, using the open-source *MultiWho* dataset (Lee et al., 2025), which contains ambiguous relative-clause attachment sentences from multiple languages. This extension allowed us to move beyond the smaller initial stimulus set based on the Pozniak-style materials and to examine whether the attachment patterns observed there would generalize to a larger and more cross-linguistically diverse set of items. The *MultiWho* dataset provides 96 ambiguous relative-clause sentences per language, substantially increasing the number of items relative to the initial experiment. To preserve comparability, we used the ambiguous *MultiWho* sentence materials together with arithmetic primes constructed in the same general format as those in the earlier experiment, thereby retaining the same priming logic and task structure while substantially expanding the language coverage and item base.

### 3.3 Procedure

In the baseline attachment condition, the model was presented only with ambiguous sentences in order to assess whether it exhibited an attachment preference and whether that preference aligned with previously reported human patterns. On each trial, the model received an ambiguous sentence such as (3), followed by a comprehension question (e.g., *Who is about to create a masterpiece?*).

In the priming condition, we tested whether arithmetic structure influenced subsequent attachment decisions. Each trial contained two components: (i) an arithmetic prime and (ii) an ambiguous sentence target followed by a comprehension question (Figure 3). We used an open-response format, allowing the model to generate its own answers to both tasks.

For Experiment 1, the priming materials consisted of 30 items, each repeated five times at a sampling temperature of 0.7. The design crossed two equation types (HA prime: *parenthesized*; LA prime: *flat*) with two target-language conditions

**Combined condition prompt (open-response):**

```
Answer both questions.  Give only short
answers, no explanation.
4 + (6 - 2) / 2 = ?

"Here we have the son of the father who
fancies riding what I like best."
Who fancies riding?

Respond in the format:
Math: [number]
Language: [single word]
```

Figure 3: Example prompt for the combined condition. The equation prime (here, a parenthesized/HA prime) and the ambiguous RC attachment sentence are presented together.

(English and French), yielding $30 \times 2 \times 2 \times 5 = 600$ combined trials. Together with the two baseline conditions, this resulted in 1200 trials per model: 300 math-baseline trials, 300 attachment-baseline trials, and 600 combined math+attachment trials.

Experiment 2 followed the same general procedure using the *MultiWho* dataset. It included 96 items in each of six languages, with five repetitions per item at a sampling temperature of 0.7. This yielded 2,880 attachment-baseline trials per model $(96 \times 6 \times 5)$ and 5,760 combined trials per model $(96 \times 6 \times 2 \times 5)$, for a total of 8,640 trials per model and 43,200 trials across the five models. All models were accessed through the Fireworks AI API with Langfuse tracing.

### 3.4 Analysis

Following standard practice in studies of relative clause attachment ambiguity, we use the proportion of HA responses as the primary measure. Because LA responses are simply the complement, values above 0.5 indicate an HA preference and values below 0.5 an LA preference.

Models did not always select one of the two candidate referents. Responses that matched neither the HA noun nor the LA noun were treated as invalid and excluded from HA/LA proportion calculations, though their counts are reported for transparency.

In the priming analysis, we further excluded trials with incorrect math answers. Following (Pozniak et al., 2018), only trials with correct math responses were retained, since incorrectly solved equations cannot be assumed to instantiate the intended structure.

### 3.5 Results in Exp. 1 (Pozniak-stimuli)

#### 3.5.1 Baseline preference

As reported in the psycholinguistic literature, English generally shows a LA preference, whereas French shows a HA preference. The models only partially reproduced this cross-linguistic contrast (Table 2). GPT-OSS 120B and GPT-OSS 20B aligned most closely with the human pattern, showing lower HA rates in English and higher HA rates in French. Llama 3.3 70B Instruct showed the same directional contrast, but its English responses reflected only a weak LA preference. By contrast, Qwen3 8B and Kimi K2 Instruct favored HA in both languages.

The GPT-OSS results, however, require caution because these models produced unusually high rates of invalid responses, especially in English. This substantially reduced the number of valid English trials, leaving only 82 for GPT-OSS 120B and 43 for GPT-OSS 20B out of 150. Thus, although their overall directional pattern is consistent with the human literature, their English attachment estimates should be interpreted cautiously.

#### 3.5.2 Priming Results

The priming results for English and French are summarized in Tables 3 and 4. We begin with English.

Table 3 shows no clear evidence of the predicted priming effect in English. Rather than shifting with prime type, the models largely maintained their baseline attachment preferences across conditions. Qwen 8B and Kimi K2 showed nearly identical HA rates across conditions, GPT-OSS 120B and GPT-OSS 20B remained broadly LA-preferring, and Llama 70B was difficult to interpret because very low math accuracy left too few valid trials.

The French results in Table 4 show the same pattern. Again, the models largely maintained their baseline preferences across prime conditions: Qwen 8B and Kimi K2 showed very similar HA rates across conditions, GPT-OSS 120B and GPT-OSS 20B remained broadly stable, and Llama 70B again yielded too few valid trials because of very low math accuracy.

Thus, neither the English nor the French results provide evidence of a priming effect. In both languages, attachment responses remained largely stable across prime conditions, suggesting that arithmetic structure did not systematically influence attachment choices in the predicted direction.

### 3.6 Results in Exp. 2 (multilingual extension)

We used the *MultiWho* dataset to extend the study to a larger multilingual set of relative-clause attachment items and to test whether the patterns from Experiment 1 generalize to this dataset.

#### 3.6.1 Baseline Preference

We first examined baseline attachment preferences. Prior psycholinguistic research reports a LA preference in English and Chinese, and a HA preference in Japanese, Korean, Spanish, and Russian. The *MultiWho* baseline results show only partial alignment with these human patterns. English and Chinese were consistently LA-preferring across all models, in line with the human literature, and Russian showed a robust HA preference across models. Spanish showed weaker and less consistent evidence of HA. By contrast, Japanese and Korean did not show the expected HA preference: all models remained below 50% HA in both languages, although Qwen 8B, GPT-OSS 120B, and GPT-OSS 20B showed somewhat higher HA rates, especially in Japanese.

The clearest correspondence to the human literature was found for English, Chinese, and Russian, while Japanese and Korean diverged most clearly from the expected HA pattern. Spanish occupied an intermediate position, showing some evidence of HA but not a consistent human-like pattern across models.

Invalid responses also varied substantially across languages and models. Japanese showed the highest invalid-response rates, with especially high rates for GPT-OSS 20B and Kimi K2, and Korean also yielded elevated invalid-response rates for several models. By contrast, invalid responses were generally low in English, Spanish, and Chinese, aside from the GPT-OSS models in English.

#### 3.6.2 Priming Results

The priming results for Experiment 2 are summarized in Table 6. If the arithmetic primes influenced attachment decisions, models should have produced more HA responses following HA primes and more LA responses following LA primes. This

| | | Llama 70B | Qwen 8B | Kimi K2 | GPT-OSS 120B | GPT-OSS 20B |
|---|---|---|---|---|---|---|
| Preference | English | 48.7% (73/150) | 58.4% (80/137) | 60.3% (88/146) | 20.7% (17/82) | 32.6% (14/43) |
| | French | 73.3% (85/116) | 92.2% (130/141) | 67.8% (97/143) | 67.8% (97/143) | 84.3% (113/134) |

Table 2: Baseline attachment preferences in Pozniak stimuli. Each model completed 300 trials. Percentages indicate HA responses among valid responses; fractions show HA/valid. Invalid counts are reported in Appendix B.

| Condition | Llama 70B | Qwen 8B | Kimi K2 | GPT-OSS 120B | GPT-OSS 20B |
|---|---|---|---|---|---|
| Math accuracy (HA prime) | 18.0% (27/150) | 100% (150/150) | 99.3% (149/150) | 100% (150/150) | 100% (145/145) |
| Math accuracy (LA prime) | 13.3% (20/150) | 100% (150/150) | 82.7% (124/150) | 100% (150/150) | 99.3% (140/141) |
| HA rate after HA prime | 37.0% (10/27) | 63.4% (92/145) | 76.1% (105/138) | 33.6% (41/122) | 52.5% (52/99) |
| HA rate after LA prime | 15.0% (3/20) | 64.1% (91/142) | 76.4% (84/110) | 29.4% (32/109) | 38.0% (30/79) |
| Priming effect ($\Delta$) | +22.0 | −0.7 | −0.3 | +4.2 | +14.5 |

Table 3: English priming results (Pozniak stimuli). HA rate = proportion of high-attachment responses among valid trials after excluding incorrect math trials. $\Delta$ = HA rate after HA prime minus after LA prime (percentage points).

pattern was not observed consistently across languages or models. Instead, as in Experiment 1, the responses largely reflected the models' baseline attachment preferences in each language.

Regardless of prime type, in English, Chinese, Japanese, and Korean, where baseline preferences were generally low attachment, responses were dominated by LA choices. In Russian, where baseline preferences were strongly high attachment, responses were dominated by HA choices. Spanish showed a more mixed pattern across models, but again did not provide clear evidence that prime type systematically shifted attachment in the predicted direction. Thus, the *MultiWho* results provide no clear evidence of structural priming.

## 4 General Discussion

The present study investigated whether structural priming can be observed across domains in large language models, specifically from arithmetic expressions to relative-clause attachment. From a developmental perspective, this provides a test of whether linguistic and mathematical abilities in models rely on shared structural representations or remain functionally separate.

Across both experiments, we found no robust evidence of the predicted priming effect. If language models represented structure in a sufficiently abstract and transferable way across domains, HA primes should have increased HA responses and LA primes should have increased LA responses. This pattern did not emerge consistently. Instead, the models' responses largely reflected their baseline attachment preferences in each language. The absence of such transfer suggests that, at least under the present task conditions, model development does not necessarily yield domain-general hierarchical representations comparable to those implicated in human cross-domain priming. Rather, the models appear to develop structure-sensitive behavior that is more strongly tied to the domain, format, and task in which that structure is encountered.

The absence of a priming effect in both Experiment 1 and Experiment 2, including the larger multilingual *MultiWho* dataset, suggests that this null result is robust rather than simply an artifact of the smaller initial study. At the same time, the models were not entirely insensitive to structure: across both experiments, they showed clear attachment preferences that varied across languages. The crucial finding, however, is that these preferences were not systematically shifted by the arithmetic primes. This suggests that, although the models encode some structural information, that knowledge may not support abstract transfer across domains. More cautiously, their behavior may instead reflect domain-specific statistical regularities, learned associations, or task-specific processing biases rather than a shared abstract structural representation.

One possible interpretation of the null priming effect concerns the modularity, or functional separation, of linguistic and mathematical reasoning in language models. The models tested here were able, in many cases, to solve the arithmetic problems and to answer the attachment questions, indicating that failure to observe priming cannot be reduced simply to a complete inability to perform either task. However, successful performance in both domains does not entail that the same representations or processing routines are used across them.

| Condition | Llama 70B | Qwen 8B | Kimi K2 | GPT-OSS 120B | GPT-OSS 20B |
|---|---|---|---|---|---|
| Math accuracy (HA prime) | 19.3% (29/150) | 100% (150/150) | 99.3% (149/150) | 100% (150/150) | 100% (150/150) |
| Math accuracy (LA prime) | 11.3% (17/150) | 100% (150/150) | 85.3% (128/150) | 100% (150/150) | 100% (150/150) |
| HA rate after HA prime | 69.0% (20/29) | 84.4% (108/128) | 79.9% (119/149) | 70.0% (105/150) | 87.3% (124/142) |
| HA rate after LA prime | 29.4% (5/17) | 87.5% (112/128) | 81.9% (104/127) | 65.5% (99/150) | 86.8% (125/144) |
| Priming effect ($\Delta$) | +39.6 | −3.1 | −2.0 | +4.5 | +0.5 |

Table 4: French priming results (Pozniak stimuli). HA rate = proportion of high-attachment responses among valid trials after excluding incorrect math trials. $\Delta$ = HA rate after HA prime minus after LA prime (percentage points).

| Language | Llama 70B | Qwen 8B | Kimi K2 | GPT-OSS 120B | GPT-OSS 20B |
|---|---|---|---|---|---|
| EN | 13.2% (62/471) | 18.6% (89/479) | 27.1% (129/476) | 7.6% (34/446) | 12.6% (53/421) |
| CH | 7.5% (35/469) | 22.7% (106/467) | 12.9% (61/474) | 19.8% (94/474) | 12.5% (57/456) |
| JP | 13.0% (56/430) | 38.2% (152/398) | 23.5% (88/375) | 35.6% (143/402) | 36.3% (130/358) |
| KO | 8.3% (34/411) | 29.9% (134/448) | 13.5% (56/414) | 32.8% (151/461) | 27.9% (124/445) |
| RU | 57.4% (264/460) | 73.8% (335/454) | 63.4% (287/453) | 77.4% (325/420) | 76.9% (320/416) |
| SP | 42.8% (196/458) | 55.7% (264/474) | 49.4% (235/476) | 37.9% (180/475) | 63.7% (297/466) |

Table 5: Baseline attachment preferences in Experiment 2 (*MultiWho*). Each model completed 480 trials per language. Percentages indicate HA responses among valid responses, with raw counts in parentheses.

The absence of cross-domain priming is therefore compatible with the possibility that linguistic attachment preferences and arithmetic grouping are handled by partially separate mechanisms, representations, or task-specific circuits within the model. On this interpretation, mathematical and linguistic competence may coexist in the same model without being integrated at the level of abstract hierarchical structure required for priming.

This possibility is theoretically important rather than merely methodological. In humans, cross-domain priming has been interpreted as evidence for domain-general resources involved in representing hierarchical structure. The present findings suggest that language models may differ from humans not only in the amount or type of input they receive, but also in how competencies acquired from different input domains are organized. Thus, the null result contributes to a developmental account of model cognition: exposure to multiple structured domains may be sufficient for task performance, but not sufficient for the emergence of shared, transferable structural representations.

A second important finding concerns differences across models. Although the overall null priming result was similar across models, the models differed substantially in how reliably they carried out the task. Llama 3.3 70B Instruct consistently showed much lower math accuracy than the other models in the combined prime-and-sentence task, leaving relatively few interpretable trials for the priming analysis. The GPT-OSS models also showed distinctive response patterns, including relatively high invalid-response rates in some conditions, especially in English in Experiment 1. By contrast, Qwen3 8B and Kimi K2 Instruct were generally more stable in task execution, yielding a larger number of interpretable trials for analysis. These differences are important because they show that task reliability varied across models. However, this variation does not alter the broader conclusion: regardless of differences in execution, none of the models showed robust evidence of structural priming across domains.

A third point concerns baseline attachment preferences. Across both experiments, the models showed only partial alignment with human cross-linguistic patterns. In Experiment 1, some models captured the English–French contrast more clearly than others. In Experiment 2, the pattern was again mixed: English and Chinese generally showed LA preferences, and Russian showed a clear HA preference, but Japanese and Korean did not show the expected HA preference. This pattern is broadly consistent with prior *MultiWho* findings (Lee et al., 2025), suggesting that even newer models remain only partially sensitive to human-like cross-linguistic attachment preferences. At the same time, the English baseline preferences differed across Experiment 1 and Experiment 2, which likely reflects differences in the materials themselves. Compared with the *MultiWho* stimuli in Experiment 2, the relative-clause materials in Experiment 1 were longer and structurally heavier.

| Language | Condition | Llama 70B | Qwen 8B | Kimi K2 | GPT-OSS 120B | GPT-OSS 20B |
|---|---|---|---|---|---|---|
| EN | HA rate after HA prime | 11.3% | 25.3% | 30.2% | 6.8% | 18.1% |
| | HA rate after LA prime | 7.5% | 24.8% | 29.6% | 4.2% | 14.6% |
| | Δ | +3.8 | +0.5 | +0.6 | +2.6 | +3.5 |
| CH | HA rate after HA prime | 12.3% | 22.2% | 11.0% | 16.6% | 15.0% |
| | HA rate after LA prime | 6.4% | 19.4% | 9.5% | 17.6% | 14.0% |
| | Δ | +5.9 | +2.8 | +1.5 | −1.0 | +1.0 |
| JP | HA rate after HA prime | 8.7% | 29.7% | 26.4% | 34.8% | 32.9% |
| | HA rate after LA prime | 16.2% | 29.3% | 26.1% | 37.1% | 33.1% |
| | Δ | −7.5 | +0.4 | +0.3 | −2.3 | −0.2 |
| KO | HA rate after HA prime | 14.3% | 19.6% | 19.1% | 30.6% | 20.5% |
| | HA rate after LA prime | 15.9% | 20.2% | 14.1% | 31.5% | 19.1% |
| | Δ | −1.6 | −0.6 | +5.0 | −0.9 | +1.4 |
| RU | HA rate after HA prime | 55.7% | 60.2% | 55.9% | 72.6% | 74.2% |
| | HA rate after LA prime | 61.8% | 61.5% | 55.4% | 73.4% | 71.6% |
| | Δ | −6.1 | −1.3 | +0.5 | −0.8 | +2.6 |
| SP | HA rate after HA prime | 28.0% | 45.8% | 46.1% | 31.6% | 56.8% |
| | HA rate after LA prime | 43.5% | 47.0% | 43.0% | 31.6% | 59.5% |
| | Δ | −15.5 | −1.2 | +3.1 | 0.0 | −2.7 |

Table 6: Summary of priming results in the *MultiWho* dataset. HA rate = proportion of high-attachment responses among valid, math-correct responses. Δ = HA rate after HA prime − HA rate after LA prime (in percentage points). A positive Δ would indicate priming in the predicted direction. Across 30 language–model combinations, no consistent priming pattern emerges. (See Appendix C for full per-language tables with raw counts.)

Prior psycholinguistic work has shown that attachment preferences are influenced by properties of the input, including constituent length, processing complexity, and the distribution of lexical and structural cues (Hemforth et al., 1996). Similar considerations are relevant for language models, whose responses are often sensitive to surface form, input length, and local distributional patterns. From this perspective, the cross-experiment difference in English baseline preferences is not simply noise, but further evidence that attachment behavior in language models is shaped by the specific properties of the stimulus materials.

Taken together, these findings support a cautious but clear conclusion: although the models showed stable attachment preferences and some language-specific variation, they provided no robust evidence of structural priming from arithmetic expressions to relative-clause attachment. This suggests limited cross-domain structural generalization, with model behavior shaped more by baseline attachment tendencies and the statistical and formal properties of the linguistic input.

More broadly, the present results contribute to an ongoing debate about the nature of structure in language models. Prior work has shown that models can display sensitivity to syntactic patterns and can sometimes reproduce human-like preferences in linguistic tasks. The present findings qualify that picture by showing that such sensitivity does not necessarily extend to abstract structural priming across domains. In this respect, the study highlights an important distinction between exhibiting structured behavior within a domain and deploying abstract structural representations flexibly across domains. The models appear capable of the former, but the present evidence offers little support for the latter.

## 5 Conclusion

Across two experiments, we found no robust evidence of cross-domain structural priming from arithmetic expressions to relative-clause attachment in large language models. This suggests that, although these models show some structure-sensitive behavior within language, they do not readily generalize abstract structural representations across domains. The results point to a distinction between acquiring competence in multiple structured domains and using that competence in a way that supports cross-domain structural transfer. The absence of priming is therefore consistent with the possibility that linguistic and mathematical reasoning remain functionally separate in current models, though further representational analyses would be needed to test this interpretation directly.

## 6 Limitations

Several limitations of the present study should be noted. First, in some cases the combined math-and-language task yielded only a small number of usable trials because models either answered the math problem incorrectly or produced invalid responses to the attachment question. This was especially the case for Llama 3.3 70B Instruct, and for some language conditions with elevated invalid-response rates. Although these cases do not change the overall pattern of results, the small number of usable trials makes those estimates less stable for the specific model. As a result, these model-specific patterns should be interpreted with caution and should not be overgeneralized.

Second, the baseline attachment preferences differed across the two experiments, particularly in English. As discussed above, this difference may reflect differences in the stimulus materials, including sentence length, structural complexity, and other lexical or distributional properties. As a result, comparisons across datasets should be interpreted carefully, since attachment behavior may be shaped not only by language but also by properties of the items themselves.

Third, the present design tested a particularly strong form of structural generalization across domains, from arithmetic expressions to relative-clause attachment. Because the prime and target belong to different domains and differ substantially in surface form, this design places a demanding burden on the models. A null result in this setting therefore should not be taken to rule out the possibility that language models might show priming more readily in within-domain designs or in tasks with a more direct structural correspondence between prime and target.

A related limitation is that the present design cannot determine the internal source of the observed domain separation. The absence of priming may reflect genuinely modular or partially modular organization between linguistic and mathematical reasoning in the models. Alternatively, it may reflect properties of the prompting format, the open-response task, the salience of the arithmetic structure, or the way instruction-tuned models allocate attention across multi-part prompts. Thus, the present results should not be interpreted as proving architectural modularity in a strong sense. Rather, they show a functional absence of cross-domain transfer in this paradigm. Future work could test this interpretation more directly by examining intermediate representations, attention patterns, layer-wise effects, or developmental checkpoints during training to determine whether linguistic and mathematical structure become more integrated over time.

## References


J Kathryn Bock. 1986a. Meaning, sound, and syntax: Lexical priming in sentence production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 12(4):575.

J Kathryn Bock. 1986b. Syntactic persistence in language production. *Cognitive psychology*, 18(3):355–387.

Kathryn Bock and Zenzi M Griffin. 2000. The persistence of structural priming: Transient activation or implicit learning? *Journal of experimental psychology: General*, 129(2):177.

Kathryn Bock and Helga Loebell. 1990. Framing sentences. *Cognition*, 35(1):1–39.

Zhenguang Cai, Xufeng Duan, David Haslett, Shuqi Wang, and Martin Pickering. 2024. Do large language models resemble humans in language use? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 37–56, Bangkok, Thailand. Association for Computational Linguistics.

Franklin Chang, Gary S Dell, Kathryn Bock, and Zenzi M Griffin. 2000. Structural priming as implicit learning: A comparison of models of sentence production. *Journal of psycholinguistic research*, 29(2):217–230.

Alexandra A Cleland and Martin J Pickering. 2003. The use of lexical and syntactic information in language production: Evidence from the priming of noun-phrase structure. *Journal of Memory and Language*, 49(2):214–230.

Timothy Desmet and Mieke Declercq. 2006. Cross-linguistic priming of syntactic hierarchical configuration information. *Journal of Memory and Language*, 54(4):610–632.

Victor S Ferreira. 2003. The persistence of optional complementizer production: Why saying "that" is not saying "that" at all. *Journal of Memory and Language*, 48(2):379–398.

Alex Fine and T Florian Jaeger. 2013. Syntactic priming in language comprehension allows linguistic expectations to converge on the statistics of the input. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects:

Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. Syntaxgym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76.

Zenzi M Griffin and Justin Weinstein-Tull. 2003. Conceptual structure modulates structural priming in the production of complex sentences. *Journal of Memory and Language*, 49(4):537–555.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Robert J Hartsuiker and Casper Westenberg. 2000. Word order priming in written and spoken sentence production. *Cognition*, 75(2):B27–B39.

B Hemforth, L Konieczny, and C Scheepers. 1996. Syntactic and anaphoric processes in modifier attachment. In *The 9th Annual CUNY Conference on Human Sentence Processing*, pages 21–23.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 1725–1744.

Jaap Jumelet, Willem Zuidema, and Arabella Sinclair. 2024a. Do language models exhibit human-like structural priming effects? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14727–14742, Bangkok, Thailand. Association for Computational Linguistics.

Jaap Jumelet, Willem Zuidema, and Arabella Sinclair. 2024b. Do language models exhibit human-like structural priming effects? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14727–14742.

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.

So Young Lee, Russell Scheinberg, Amber Shore, and Ameeta Agrawal. 2025. Who relies more on world knowledge and bias for syntactic ambiguity resolution: Humans or LLMs? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3484–3498, Albuquerque, New Mexico. Association for Computational Linguistics.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2023. Generative judge for evaluating alignment. *Preprint*, arXiv:2310.05470.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Charles Lovering, Rohan Jha, Tal Linzen, and Ellie Pavlick. 2021. Predicting inductive biases of pretrained models. In *International Conference on learning representations*.

Rebecca Marvin and Tal Linzen. 2018a. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Rebecca Marvin and Tal Linzen. 2018b. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 1192–1202.

James A. Michaelov, Catherine Arnett, Tyler A. Chang, and Benjamin K. Bergen. 2023. Structural priming demonstrates abstract grammatical representations in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3703–3720, Singapore. Association for Computational Linguistics.

Martin J Pickering and Holly P Branigan. 1998. The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and language*, 39(4):633–651.

Céline Pozniak, Barbara Hemforth, and Christoph Scheepers. 2018. Cross-domain priming from mathematics to relative-clause attachment: A visual-world study in french. *Frontiers in psychology*, 9:2056.

Grusha Prasad, Marten van Schijndel, and Tal Linzen. 2019. Using priming to uncover the organization of syntactic representations in neural language models. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages

66–76, Hong Kong, China. Association for Computational Linguistics.

Christoph Scheepers. 2003. Syntactic priming of relative clause attachments: Persistence of structural configuration in sentence production. *Cognition*, 89(3):179–205.

Christoph Scheepers, Patrick Sturt, Catherine J Martin, Andriy Myachykov, Kay Teevan, and Izabela Viskupova. 2011. Structural priming across cognitive domains: From simple arithmetic to relative-clause attachment. *Psychological Science*, 22(10):1319–1326.

Arabella Sinclair, Jaap Jumelet, Willem Zuidema, and Raquel Fernández. 2022a. Structural persistence in language models: Priming as a window into abstract language representations. *Transactions of the Association for Computational Linguistics*, 10:1031–1050.

Arabella Sinclair, Jaap Jumelet, Willem Zuidema, and Raquel Fernández. 2022b. Structural persistence in language models: Priming as a window into abstract language representations. *Transactions of the Association for Computational Linguistics*, 10:1031–1050.

Arabella Sinclair, Anastasia Klimovich-Gray, Jaap Jumelet, Nika Adamian, and Agnieszka Konopka. 2026. Structural priming in humans and large language models. *Journal of Memory and Language*, 149:104713.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 2888–2913.

Harish Tayyar Madabushi, Laurence Romain, Dagmar Divjak, and Petar Milin. 2020. CxGBERT: BERT meets construction grammar. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4020–4032, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Jennifer C White, Tiago Pimentel, Naomi Saphra, and Ryan Cotterell. 2021. A non-linear structural probe. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 132–138.

Ethan Wilcox, Clara Meister, Ryan Cotterell, and Tiago Pimentel. 2023. Language model quality correlates with psychometric predictive power in multiple languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7503–7511, Singapore. Association for Computational Linguistics.

## A Appendix: English Stimulus Modifications

The English stimuli were adapted from the translations provided alongside the French originals in Pozniak et al. (2018). Because the French materials use possessives that agree with the possessed noun (e.g., *son vêtement* 'his/her garment'), they do not introduce gender as an unintended disambiguation cue. The English translations, however, contained gendered pronouns (*he*, *she*, *his*, *her*) that could bias attachment when NP1 and NP2 differ in gender. For example, in *"the daughter of the chemist who will put on* ***her*** *usual outfit"*, the pronoun *her* might more easily be construed as referring to the daughter.

To remove this confound, we replaced gendered pronouns with first-person forms (*he/she* → *I*, *his/her* → *my*) or rephrased to avoid pronouns entirely (*his drink* → *a drink*). Three items (6, 12, and 15) required no changes. Table 7 lists all modifications.

| Item | Sentence |
|---|---|
| 1 | *...who fancies riding what ~~he likes~~* ***I like*** *best.* |
| 2 | *...who will set up what ~~he was~~* ***I was*** *asked to.* |
| 3 | *...who is about to have ~~his drink~~* ***a drink****.* |
| 4 | *...who will prepare what ~~he needs to~~* ***is needed****.* |
| 5 | *...who will cut what ~~he has to~~* ***is needed****.* |
| 7 | *...who will reveal ~~his~~* ***my*** *latest creation.* |
| 8 | *...who will read what ~~she's~~* ***I'm*** *used to.* |
| 9 | *...who will continue working on ~~her~~* ***my*** *latest project.* |
| 10 | *...who will deliver what ~~he has~~* ***I have****.* |
| 11 | *...who will present ~~his~~* ***my*** *recent work.* |
| 13 | *...who will sip ~~her favorite~~* ***a favorite*** *drink.* |
| 14 | *...who will fetch what ~~she's~~* ***I'm*** *used to.* |
| 16 | *...who will purchase what ~~he needs~~* ***I need****.* |
| 17 | *...who will create what ~~he's~~* ***I'm*** *best known for.* |
| 18 | *...who will fetch ~~her~~* ***my*** *favorite item.* |
| 19 | *...who will work on what ~~he's~~* ***I'm*** *supposed to.* |
| 20 | *...who will wear ~~her~~* ***my*** *favorite clothes.* |
| 21 | *...who will grab what ~~she~~* ***I*** *was looking for.* |
| 22 | *...who'll have to sell ~~his~~* ***my*** *favorite possession.* |
| 23 | *...who will eat ~~his lunch~~* ***lunch****.* |
| 24 | *...who will have ~~his dinner~~* ***dinner****.* |
| 25 | *...who will put on ~~his~~* ***my*** *usual outfit.* |
| 26 | *...who will finish what ~~he was~~* ***I was*** *working on.* |
| 27 | *...who will finish what ~~he started~~* ***I started*** *to work on.* |
| 28 | *...who will put on ~~her~~* ***my*** *usual outfit in the morning.* |
| 29 | *...who will read what ~~he's~~* ***I'm*** *most interested in.* |
| 30 | *...who will have what ~~he's~~* ***I'm*** *desperate for.* |

Table 7: Modifications to the English stimuli from Pozniak et al. (2018). Only the relative clause portion is shown; the matrix clause (e.g., *"Here we have the son of the father..."*) was unchanged. Items 6, 12, and 15 required no modification.

## B Appendix: Invalid Responses Rate

| | Llama 70B | Qwen 8B | Kimi K2 | GPT-OSS 120B | GPT-OSS 20B |
|---|---|---|---|---|---|
| English | 0% (0/150) | 9% (13/150) | 3% (4/150) | 45% (68/150) | 71% (107/150) |
| French | 23% (34/150) | 6% (9/150) | 5% (7/150) | 5% (7/150) | 11% (16/150) |

Table 8: Invalid response counts by language in Experiment 1 (*Pozniak*). Each cell is based on 150 trials.

| | Llama 70B | Qwen 8B | Kimi K2 | GPT-OSS 120B | GPT-OSS 20B |
|---|---|---|---|---|---|
| English | 2% (9/480) | 0% (1/480) | 1% (4/480) | 7% (34/480) | 12% (59/480) |
| Chinese | 2% (11/480) | 3% (13/480) | 1% (6/480) | 1% (6/480) | 5% (24/480) |
| Japanese | 10% (50/480) | 17% (82/480) | 22% (105/480) | 16% (78/480) | 25% (122/480) |
| Korean | 14% (69/480) | 7% (32/480) | 14% (66/480) | 4% (19/480) | 7% (35/480) |
| Russian | 4% (20/480) | 5% (26/480) | 6% (27/480) | 13% (60/480) | 13% (64/480) |
| Spanish | 5% (22/480) | 1% (6/480) | 1% (4/480) | 1% (5/480) | 3% (14/480) |

Table 9: Invalid response rates by language in Experiment 2 (*MultiWho*). Each cell shows the percentage of invalid responses, followed by the raw count in parentheses.

## C MultiWho Priming Results Detail

| Condition | Llama 70B | Qwen 8B | Kimi K2 | GPT-OSS 120B | GPT-OSS 20B |
|---|---|---|---|---|---|
| Math accuracy (HA prime) | 28.0% (134/480) | 100% (480/480) | 90.6% (435/480) | 100% (480/480) | 100% (474/474) |
| Sentence HA answer rate | 11.3% (15/133) | 25.3% (118/467) | 30.2% (130/430) | 6.8% (31/457) | 18.1% (81/447) |
| Math accuracy (LA prime) | 16.7% (80/480) | 100% (480/480) | 73.8% (354/480) | 100% (480/480) | 98.9% (466/471) |
| Sentence LA answer rate | 92.5% (74/80) | 75.2% (357/475) | 70.4% (243/345) | 95.8% (436/455) | 85.4% (369/432) |

Table 10: English priming results by model in the *MultiWho* dataset. Fractions in parentheses indicate raw counts. Percentages for HA and LA answers are calculated within valid attachment responses after excluding trials with incorrect math answers.

| Condition | Llama 70B | Qwen 8B | Kimi K2 | GPT-OSS 120B | GPT-OSS 20B |
|---|---|---|---|---|---|
| Math accuracy (HA prime) | 25.4% (122/480) | 100% (480/480) | 91.0% (437/480) | 100% (480/480) | 100% (480/480) |
| Sentence HA answer rate | 12.3% (15/122) | 22.2% (102/460) | 11.0% (46/419) | 16.6% (78/469) | 15.0% (66/441) |
| Math accuracy (LA prime) | 18.1% (87/480) | 100% (480/480) | 75.4% (362/480) | 100% (480/480) | 99.0% (475/480) |
| Sentence LA answer rate | 93.6% (73/78) | 80.6% (378/469) | 90.5% (306/338) | 82.4% (383/465) | 86.0% (368/428) |

Table 11: Chinese priming results by model in the *MultiWho* dataset. Fractions in parentheses indicate raw counts. Percentages for HA and LA answers are calculated within valid attachment responses after excluding trials with incorrect math answers.

| Condition | Llama 70B | Qwen 8B | Kimi K2 | GPT-OSS 120B | GPT-OSS 20B |
|---|---|---|---|---|---|
| Math accuracy (HA prime) | 27.1% (130/480) | 100% (480/480) | 90.8% (436/480) | 100% (480/480) | 100% (480/480) |
| Sentence HA answer rate | 8.7% (9/103) | 29.7% (114/384) | 26.4% (66/250) | 34.8% (149/428) | 32.9% (105/319) |
| Math accuracy (LA prime) | 19.4% (93/480) | 100% (480/480) | 76.9% (369/480) | 100% (480/480) | 99.4% (477/480) |
| Sentence LA answer rate | 83.8% (57/68) | 70.7% (265/375) | 73.9% (153/207) | 62.9% (264/420) | 66.9% (216/323) |

Table 12: Japanese priming results by model in the *MultiWho* dataset. Fractions in parentheses indicate raw counts. Percentages for HA and LA answers are calculated within valid attachment responses after excluding trials with incorrect math answers.

| Condition | Llama 70B | Qwen 8B | Kimi K2 | GPT-OSS 120B | GPT-OSS 20B |
|---|---|---|---|---|---|
| Math accuracy (HA prime) | 27.9% (134/480) | 100% (480/480) | 91.9% (441/480) | 100% (480/480) | 100% (480/480) |
| Sentence HA answer rate | 14.3% (16/112) | 19.6% (83/424) | 19.1% (61/319) | 30.6% (141/461) | 20.5% (79/386) |
| Math accuracy (LA prime) | 19.4% (93/480) | 100% (480/480) | 73.3% (352/480) | 100% (480/480) | 99.8% (479/480) |
| Sentence LA answer rate | 84.1% (69/82) | 79.8% (344/431) | 85.9% (214/249) | 68.5% (315/460) | 80.9% (321/397) |

Table 13: Korean priming results by model in the *MultiWho* dataset. Fractions in parentheses indicate raw counts. Percentages for HA and LA answers are calculated within valid attachment responses after excluding trials with incorrect math answers.

| Condition | Llama 70B | Qwen 8B | Kimi K2 | GPT-OSS 120B | GPT-OSS 20B |
|---|---|---|---|---|---|
| Math accuracy (HA prime) | 29.8% (143/480) | 100% (480/480) | 92.7% (445/480) | 100% (480/480) | 100% (480/480) |
| Sentence HA answer rate | 55.7% (73/131) | 60.2% (277/460) | 55.9% (224/401) | 72.6% (339/467) | 74.2% (333/449) |
| Math accuracy (LA prime) | 17.7% (85/480) | 100% (480/480) | 74.2% (356/480) | 100% (480/480) | 98.8% (474/477) |
| Sentence LA answer rate | 38.2% (29/76) | 38.5% (180/468) | 44.6% (144/323) | 26.6% (122/459) | 28.4% (128/450) |

Table 14: Russian priming results by model in the *MultiWho* dataset. Fractions in parentheses indicate raw counts. Percentages for HA and LA answers are calculated within valid attachment responses after excluding trials with incorrect math answers.

| Condition | Llama 70B | Qwen 8B | Kimi K2 | GPT-OSS 120B | GPT-OSS 20B |
|---|---|---|---|---|---|
| Math accuracy (HA prime) | 29.8% (143/480) | 100% (480/480) | 91.7% (440/480) | 100% (480/480) | 100% (480/480) |
| Sentence HA answer rate | 28.0% (40/143) | 45.8% (216/472) | 46.1% (195/423) | 31.6% (151/478) | 56.8% (266/468) |
| Math accuracy (LA prime) | 17.7% (85/480) | 100% (480/480) | 73.8% (354/480) | 100% (480/480) | 99.2% (476/479) |
| Sentence LA answer rate | 56.5% (48/85) | 53.0% (249/470) | 57.0% (195/342) | 68.4% (324/474) | 40.5% (189/467) |

Table 15: Spanish priming results by model in the *MultiWho* dataset. Fractions in parentheses indicate raw counts. Percentages for HA and LA answers are calculated within valid attachment responses after excluding trials with incorrect math answers.