DO LLMS TRANSFER STRUCTURAL KNOWLEDGE? PROBING STRUCTURAL INFERENCE AS A COGNITIVE CAPABILITY IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Structural knowledge refers to relational patterns that capture an underlying organization beyond the specific entities and contexts in which those relations occur. While recent studies have examined cognitive biases and reasoning behaviors in large language models (LLMs), whether LLMs can infer such structures from natural language and transfer them across contexts remains unclear. We introduce INSTILL, a framework for INvestigating the STructural Inference in LLMs through tasks that require transferring structural knowledge. Each task presents a reference context and four candidate contexts, only one of which preserves the underlying relational structure of the reference. Drawing on five representative structure families motivated by cognitive science, INSTILL constructs diverse structural patterns and contextualizes them as cognitive scenarios grounded in nine domains spanning spatial, social, comparative, and temporal relations. These tasks require models to infer the latent structure from natural language and generalize it across contexts with different entities, relationships, or domains. INSTILL enables controlled analysis of how structural inference varies across structure families, domain congruency, and increasing structure sizes. Our results show that several recent models maintain stable structural inference across domain shifts and increasing structure sizes, whereas smaller variants of the same model families exhibit larger performance drops and greater sensitivity to domain congruency and local relational patterns.

1 INTRODUCTION

Large language models (LLMs) have shown proficiency in tasks such as logical reasoning, strategic planning, and multi-step inference (Geva et al., 2021; Cobbe et al., 2021; Srivastava et al., 2023; Valmeekam et al., 2023). These advances have shifted their role from mere task assistants to systems that increasingly resemble partners in thought, capable of reasoning, planning and problem-solving (Hao et al., 2023; Collins et al., 2024; Gui et al., 2025). This progression raises the question of whether their apparent proficiency reflects a genuine understanding of the world that supports the abstraction and transfer of knowledge across contexts.

Decades of research in cognitive science suggest that latent structural representations support generalization across diverse tasks and domains (Piaget, 1952; Gentner, 1983; Newell, 1990; Gentner & Markman, 1997). These representations, referred to as structural knowledge, encode relational patterns abstracted from prior experience and enable generalization to new environments (Mark et al., 2020; Lynn & Bassett, 2020). For instance, in spatial navigation, humans encode environments into relational layouts that reflect the abstract structure of spatial relations, allowing shortcut discovery not directly experienced (Brashears & Quintane, 2015; Epstein et al., 2017; Garvert et al., 2017; Peer et al., 2021). Importantly, recent findings suggest that such mechanisms extend beyond physical space, giving rise to form structural knowledge encoding both spatial and non-spatial relations across diverse domains. As shown in Figure 1 (left), grounding inference in abstract structures enables generalization across contexts that differ in appearance yet share an underlying structure.

A growing body of work has explored cognitive phenomena in LLMs, such as relational reasoning and cognitive biases (Binz & Schulz, 2023; Webb et al., 2023; Momennejad et al., 2023; Liu et al., 2025). While their findings provide valuable insights into the cognitive behaviors of LLMs, their

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

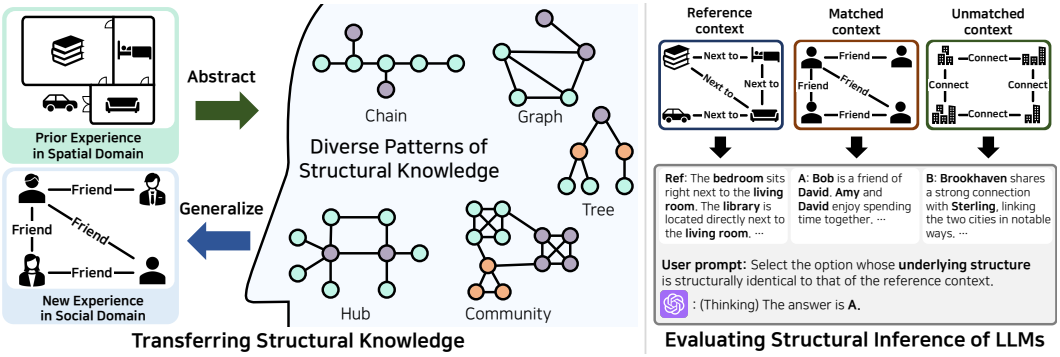


Figure 1: The left illustrates the process of human cognition grounded in structural knowledge, where patterns abstracted from prior experience are generalized to new situations. The right depicts INSTILL, which generates tasks that test whether LLMs can identify the context whose underlying structure matches that of a reference context.

investigation has largely been confined to specific tasks under controlled settings designed to elicit cognitive effects, leaving the structural transfer of knowledge in LLMs insufficiently explored.

In this work, we introduce INSTILL, a framework for INvestigating the STructural Inference in LLMs through tasks that require extracting latent relational structures from cognitive scenarios and transferring this structural knowledge to novel environments. INSTILL is designed to examine how structural inference in LLMs varies across different relational structures and domains. It is not intended to compare LLMs with humans or to assess whether their reasoning aligns with human cognition. As presented in Figure 1, INSTILL draws on representative abstract structures from cognitive science. INSTILL contextualizes these structures into cognitive scenarios grounded in diverse domains, encompassing spatial, social, and temporal situations. It then constructs tasks that require LLMs to generalize and transfer the underlying abstract structures to varied contexts. We summarize our main contributions as follows:

- **Probing Structural Inference in LLMs.** We present a framework to investigate LLMs' ability to abstract and transfer structural knowledge by identifying shared relational patterns across contexts.
- **Cognitive Science-Inspired Task Generation.** INSTILL instantiates cognitive scenarios across diverse domains while preserving their underlying representative relational structures.
- **Empirical Analysis of Structural Inference.** Our results reveal distinct patterns of structural inference across LLMs under domain shifts, intervening contexts, and increasing structure sizes.

2 Background and Related Works

Structural Knowledge. Structural knowledge, also referred to as a cognitive map, serves as an internal model that encodes sensory observations into a structured representation (Mark et al., 2020; Lynn & Bassett, 2020). Garvert et al. have shown that the hippocampal–entorhinal system organizes experiences according to relational structures. Although cognitive maps have been investigated in spatial navigation, subsequent studies have demonstrated that the same principles can be extended to non-spatial domains, capturing abstract relations among entities such as those in conceptual spaces (Behrens et al., 2018; Peer et al., 2021). Moreover, Mark et al. have found that once humans acquire a relational schema, they can transfer it to new contexts.

Cognitive Science-inspired Evaluation of LLMs. Recent studies inspired by cognitive science explore the cognitive capacities of LLMs (Jones & Steinhardt, 2022; Hagendorff et al., 2024; Sumita et al., 2025). In particular, CogEval (Momennejad et al., 2023) evaluates goal-directed planning over cognitive maps, focusing on whether LLMs can use relational representations of an environment to construct and revise planning trajectories. In parallel, ReCogLab (Liu et al., 2025) is introduced as a framework for evaluating relational reasoning in LLMs, generating tasks that reproduce cognitive experiments such as ordering, congruency, and indeterminacy. Our work is related to these studies in that it also treats relational structure as a basis for cognitive capability, but examines a different

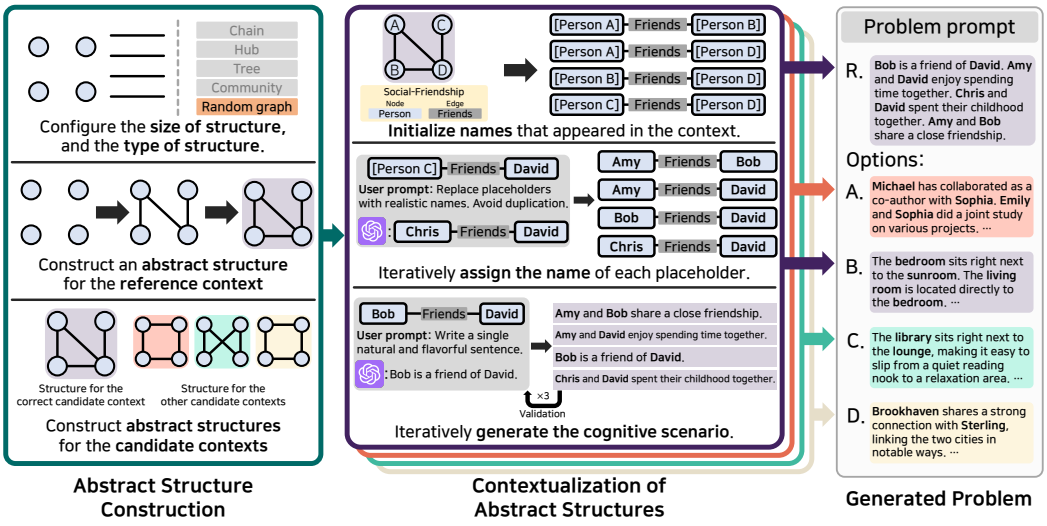


Figure 2: Overview of task generation in INSTILL. After configuring the size and type of the structure, INSTILL constructs abstract structures for the reference and candidate contexts. The abstract structures are then contextualized by iteratively generating cognitive scenarios using GPT-4.1.

aspect of relational reasoning. Specifically, CogEval and ReCogLab primarily examine reasoning over relational information within a given context, whereas our focus is on whether the underlying relational structure can be abstracted from one context and transferred across different contexts with different entities, expressions, and domains. To investigate the structural transfer, INSTILL constructs tasks over five representative structural patterns grounded in nine fine-grained domains, requiring models to identify shared structures across diverse contexts.

3 METHOD: INSTILL

Since structural knowledge lacks a canonical form of representation, it is challenging to directly infer from a model's responses. This absence of a settled form gives rise to a further elusiveness, as model responses may reflect either genuine structural understanding or mere recall of patterns encountered during training. We propose a framework, INSTILL, that generates tasks that require LLMs to transfer latent structures embedded in natural-language descriptions, incorporating representative structures widely studied in cognitive science.

3.1 OVERALL FRAMEWORK

We construct tasks that pair a reference context with four candidate contexts to evaluate whether LLMs can identify and transfer an underlying structure within cognitive scenarios. For each task, we first specify the structural family and size, and then construct the reference and candidate structures accordingly. Exactly one candidate matches the underlying structure of the reference, while the other three serve as structurally different distractors. The candidate structures are programmatically verified to ensure that only the designated candidate is structurally identical to the reference. The resulting structures are then instantiated as cognitive scenarios in natural language, grounded in domains spanning spatial, social, comparative, and temporal relations. The four options are randomly shuffled to mitigate answer position bias. Figure 2 outlines the overall task generation pipeline, which consists of abstract structure construction (§ 3.2) followed by the contextualization of those structures as natural language scenarios (§ 3.3).

3.2 ABSTRACT STRUCTURE CONSTRUCTION

We first construct the underlying relational structure of each context, which is represented as a graph whose nodes represent the entities within the context and edges, either directed or undirected, represent the relations among them. Figure 3 shows examples of each structure. For each task, the

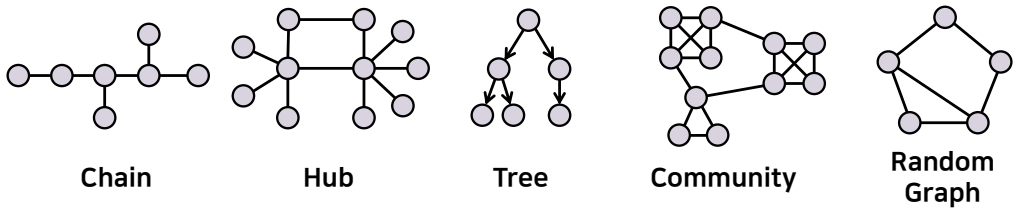


Figure 3: Representative examples of the five structural families used in INSTILL.

underlying relational structures of the contexts are generated using the configured number of nodes and edges. Our analysis focuses on five families of representative graph structures, *chain*, *hub*, *tree*, *community*, and *random graph*, which serve as abstract patterns for structural inference.

Chain. Human cognition exhibits a propensity to organize information in sequential patterns, as demonstrated in studies of spatial navigation where path integration relies on successive steps and hippocampal place and grid cells encode ordered trajectories. (Momennejad, 2020; Mark et al., 2020; Yildirim & Paul, 2024). Such sequential structures are central to cognitive processes, including chunking and sequential planning that decompose complex tasks into smaller components and organize them (Momennejad et al., 2017). To capture this property, we construct a family of chain structures, encompassing both undirected and directed types. Each structure consists of a primary chain of nodes with branches of varying lengths attached at randomly selected positions, such as a main road with side streets, or a storyline with subplots.

Hub. Centralization is a prominent property of human cognition, which serves as a key determinant of how humans organize relations and information (Behrens et al., 2018). This property manifests in the influence of key individuals within social networks and in the organization of core concepts in semantic memory, which anchor clusters of related knowledge (Behrens et al., 2018). Furthermore, neuroscientific evidence shows that certain regions of the brain function as hubs, integrating distributed sensory and mnemonic information (Backus et al., 2016; Gratton et al., 2018; Gordon et al., 2018; Weiller et al., 2025). These findings suggest that hub structures support the organization and retrieval of relational information. We construct a family of hub structures in which a small number of nodes connect to the majority of others. Specifically, up to three nodes are randomly selected and interconnected, after which all remaining nodes are attached to one of them. The remaining edge budget is allocated to connections among non-hub nodes.

Tree. A tree is a hierarchical structure rooted at a single node, in which each non-root node has exactly one parent and any two nodes are connected by a unique path. Such hierarchical relations are exemplified in linguistic parse trees, family trees, and taxonomies, which provide schematic forms that organize knowledge (Behrens et al., 2018; Mark et al., 2020). By structuring information into ordered layers, tree structures facilitate the integration of new concepts into existing knowledge and support inference across varying levels of abstraction (Russek et al., 2017; Momennejad et al., 2017; Momennejad, 2020; Peer et al., 2021). Thus, the tree serves as a cognitive schema for abstracting hierarchical relations from new environments. We construct a family of tree structures by designating a root node and iteratively attaching child nodes.

Community. A community structure comprises densely connected subgraphs with sparse intergroup connections, a pattern frequently observed in social and biological systems (Brashears & Quintane, 2015; Mark et al., 2020). Research in cognitive science suggests that humans represent and remember networks by abstracting groups as cognitive units, rather than focusing on individual nodes (Hirtle & Jonides, 1985; Kemp & Tenenbaum, 2008; Mark et al., 2020; Son et al., 2023). This abstraction reduces complexity and supports efficient reasoning over group-level structure. To construct a family of community structures, we specify the number of communities and randomly assign nodes to them. Within each community, all pairs of nodes are connected to form a clique. Subsequently, a small number of edges are placed between communities to ensure connectivity.

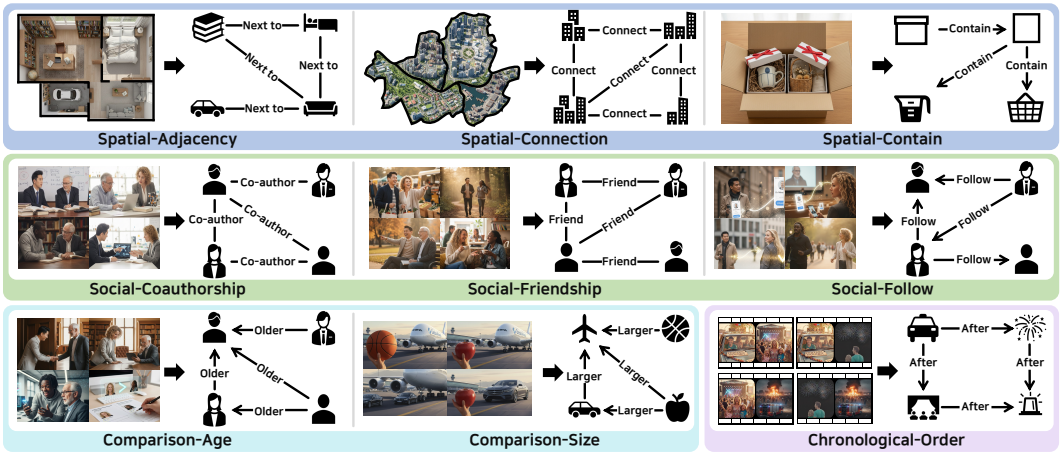


Figure 4: Example scenarios across diverse domains and their underlying structures[1]. The first row illustrates spatial situations. The second row depicts various social relationships. The third row presents situations involving comparative relationships and partial temporal ordering among events.

Random graph. We generate random graphs to model more general topologies that the aforementioned structures cannot capture. To ensure connectivity, we construct a random spanning tree for the specified number of nodes, then add the remaining edges. Directions of edges are assigned based on a random topological ordering of the nodes, ensuring acyclicity consistent with real-world settings.

3.3 Contextualization of Abstract Structures

We ground abstract structures in diverse domains, contextualizing them into cognitive scenarios described in natural language. By obscuring the abstract structure with surface-level details, our approach allows us to probe whether models can transfer structural knowledge across varied contexts.

3.3.1 Grounding Structures in Domains

We introduce nine fine-grained domains for grounding abstract structures, spanning spatial, social, and temporal situations in which relational structure is essential. Each domain is named in the format "*relation category – relation type*", where the relations indicate the relationships that constitute the domain. Figure 4 illustrates the representative scenarios that include relations in each domain.

Spatial relationships. Domains with spatial relationships encode physical layouts and spatial relations inferred from observations (Waller & Nadel, 2013). *Spatial-adjacency* describes adjacency relations among rooms, capturing local relational layouts by specifying which spaces are connected. *Spatial-connection* captures inter-city connectivity that represents road connections among cities, conveying how humans organize large-scale environments into network-like maps that support the planning of routes. *Spatial-contain* represents hierarchical containment, where the relations specify the asymmetric inclusion relations of boxes, designed to capture how humans track nested spatial relations and reason about inclusion and enclosure.

Non-spatial Relationships Empirical research in cognitive science has examined structural schemas across non-spatial domains, including social relations, comparative judgments, and temporal reasoning (Behrens et al., 2018; Lynn & Bassett, 2020; Peer et al., 2021). *Social-friendship* captures friendship relations among individuals, representing their social connections. *Social-coauthorship* captures relationships among researchers. *Social-follow* captures asymmetric follower–followee relations on social networks. These domains infuse the structures with a distinct social character, as friendships, collaborations, and asymmetries of influence manifest as patterns of connection. We further incorporate comparative and temporal relationships, capturing asymmetric relational dynamics.

[1]Images are generated by Gemini-2.5-Flash.

Table 1: Performance of structural inference across structure families, measured by exact-match accuracy. *Rand.* and *All* indicate random graphs and the average over structures. We report the mean and standard deviation over three runs.

	Chain	*Hub*	*Tree*	*Community*	*Rand.*	*All*
GPT-4.1	0.51 ± 0.01	0.70 ± 0.04	0.42 ± 0.04	0.54 ± 0.03	0.68 ± 0.00	0.57 ± 0.01
GPT-4.1-mini	0.33 ± 0.01	0.41 ± 0.04	0.32 ± 0.03	0.27 ± 0.02	0.54 ± 0.02	0.37 ± 0.01
GPT-4.1-nano	0.20 ± 0.02	0.27 ± 0.04	0.24 ± 0.03	0.24 ± 0.05	0.30 ± 0.01	0.25 ± 0.00
GPT-5	0.97 ± 0.00	$\mathbf{1.00 \pm 0.00}$	0.90 ± 0.01	0.98 ± 0.00	0.96 ± 0.00	0.96 ± 0.00
GPT-5-mini	0.90 ± 0.03	0.99 ± 0.01	0.90 ± 0.00	0.92 ± 0.02	0.96 ± 0.00	0.93 ± 0.00
GPT-5-nano	0.72 ± 0.02	0.80 ± 0.00	0.68 ± 0.03	0.41 ± 0.03	0.78 ± 0.00	0.68 ± 0.01
GPT-5.4	$\mathbf{0.99 \pm 0.01}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{0.97 \pm 0.01}$	$\mathbf{0.99 \pm 0.00}$	$\mathbf{0.99 \pm 0.00}$	$\mathbf{0.99 \pm 0.00}$
GPT-5.4-mini	0.94 ± 0.01	0.99 ± 0.01	0.93 ± 0.01	0.96 ± 0.02	0.96 ± 0.01	0.96 ± 0.00
GPT-5.4-nano	0.92 ± 0.01	$\mathbf{1.00 \pm 0.00}$	0.93 ± 0.01	$\mathbf{0.99 \pm 0.02}$	0.95 ± 0.01	0.96 ± 0.01
Gemini-2.5-Flash	0.87 ± 0.00	0.97 ± 0.00	0.82 ± 0.00	0.95 ± 0.00	0.95 ± 0.00	0.91 ± 0.00
Gemini-2.5-Flash-lite	0.51 ± 0.00	0.75 ± 0.00	0.34 ± 0.00	0.59 ± 0.01	0.63 ± 0.00	0.57 ± 0.00
Grok-4	0.97 ± 0.00	$\mathbf{1.00 \pm 0.00}$	$\mathbf{0.97 \pm 0.01}$	0.98 ± 0.00	0.98 ± 0.00	0.98 ± 0.00
Grok-3-mini	0.58 ± 0.04	0.82 ± 0.02	0.55 ± 0.06	0.35 ± 0.04	0.77 ± 0.00	0.62 ± 0.02
- Average per structure	0.73	0.83	0.69	0.71	0.81	0.75

Comparison-age models relative age comparison between individuals, and *Comparison-size* represents comparative relations of physical size among real-world objects. *Chronological-order* represents partially observed precedence relationships among events, capturing their temporal relationships.

3.3.2 Generating Cognitive Scenarios

To derive domain-specific context from abstract structures, we initialize nodes and edges as placeholders tailored to the target domain. We then traverse the structure with GPT-4.1, instantiating each edge and its associated nodes as a relation and entities. In this way, abstract structures are contextualized into coherent descriptions, augmented with diverse flavor text. However, as LLM-generated descriptions may not fully preserve the intended semantics, we employ iterative self-validation, in which the model assesses the semantic fidelity of each generated sentence. This procedure iterates up to three times, with persistent failures replaced by predefined domain-specific expressions of relations. Prompts for cognitive scenario generation and self-validation are provided in Appendix A.1, while templates and examples of the generated tasks are presented in Appendix A.2.

4 Experiments

4.1 Experimental Settings

We investigate structural inference in LLMs across different structure families, domain congruency, and structure sizes with Instill. Our experiments encompass a series of GPT-4.1 (OpenAI, 2024), GPT-5 (OpenAI, 2025a), and GPT-5.4 (OpenAI, 2026) from OpenAI, Gemini-2.5-Flash-Lite and Gemini-2.5-Flash (Gemini Team, 2025) from Google, and Grok-3-mini and Grok-4 (xAI, 2025) from xAI. All of these models are representative of recent LLMs with reasoning capability. Further details, including inference settings, repeated runs, and answer parsing, are provided in Appendix B. The code and generated tasks are available at `https://anonymous.4open.science/r/instill`.

4.2 Evaluation across Structure Families

We first investigate whether LLMs can generalize structural knowledge across diverse families of abstract structures, where both the reference and candidate contexts are drawn from the same domain. As presented in Table 1, the most consistent success arises on hub structures and random graphs, indicating that LLMs are often able to abstract underlying structure when supported by either strong centralization, as in hubs, or by the absence of distinct patterns, as in random graphs, alleviating the burden of local relational constraints. In contrast, most LLMs struggle with transferring tree structures, which require integrating relations across multiple hierarchical levels. They also encounter difficulty

Table 2: Performance of structural inference across families of abstract structures under congruent ($\equiv$) and incongruent ($\not\equiv$) settings. *Rand.* and *All* indicate random graphs and the average over structures.

	Chain		*Hub*		*Tree*		*Community*		*Rand.*		*All*	
	$\equiv$	$\not\equiv$	$\equiv$	$\not\equiv$	$\equiv$	$\not\equiv$	$\equiv$	$\not\equiv$	$\equiv$	$\not\equiv$	$\equiv$	$\not\equiv$
GPT-4.1	0.89	0.32	0.82	0.55	0.92	0.11	0.86	0.24	0.86	0.41	0.87	0.33
GPT-4.1-mini	0.84	0.18	0.86	0.14	0.93	0.03	0.89	0.02	0.86	0.27	0.87	0.13
GPT-4.1-nano	0.92	0.02	0.88	0.02	0.95	0.03	0.89	0.01	0.91	0.04	0.91	0.02
GPT-5	0.98	0.81	1.00	1.00	0.98	0.92	1.00	0.97	0.98	0.91	0.99	0.92
GPT-5-mini	0.95	0.74	1.00	0.99	0.97	0.74	0.98	0.81	0.98	0.83	0.98	0.82
GPT-5-nano	0.90	0.49	0.85	0.68	0.88	0.39	0.77	0.20	0.90	0.56	0.86	0.46
GPT-5.4	0.98	0.96	1.00	1.00	0.99	1.00	1.00	1.00	1.00	0.99	0.99	0.99
GPT-5.4-mini	0.97	0.86	1.00	1.00	0.94	0.89	0.98	0.97	0.96	0.90	0.97	0.91
GPT-5.4-nano	0.98	0.79	0.99	0.98	0.98	0.83	0.99	0.89	0.97	0.83	0.98	0.85
Gemini-2.5-Flash	0.98	0.70	0.97	0.97	0.90	0.62	0.97	0.94	0.96	0.79	0.96	0.80
Gemini-2.5-Flash-lite	0.94	0.16	0.94	0.35	0.92	0.08	0.75	0.25	0.89	0.31	0.89	0.23
Grok-4	0.99	0.91	1.00	1.00	0.99	0.98	0.99	0.98	0.99	0.99	0.99	0.97
Grok-3-mini	0.93	0.42	0.90	0.58	0.91	0.15	0.95	0.02	0.93	0.51	0.92	0.34

with community structures, which demand compressing densely interconnected communities into higher-order units while tracking sparse inter-community relations. In addition, transferring chain structures poses challenges in organizing sequential relations across ordered sequences. These results underscore the varying difficulty of abstracting underlying structures.

The results also reveal a substantial performance gap across models. Models such as GPT-4.1-nano, GPT-5-nano, and Grok-3-mini degrade sharply, as they struggle to abstract the underlying structure across most structural families. In contrast, flagship models, including GPT-5.4 and Grok-4, remain consistently strong across structure families. The pattern is not explained by scale alone, as GPT-5.4-mini/nano also perform robustly. This suggests that recent model families exhibit substantial progress in structural inference, while the performance gaps among their smaller variants indicate that this ability remains closely tied to model capacity.

We analyze whether the observed performance gaps are related to simple shortcuts. We consider a count-based heuristic that selects a candidate with the same number of entities as the reference, along with a BERT embedding-based heuristic that selects the candidate with the highest similarity. The count-based heuristic remains near chance, while the embedding-based heuristic performs only slightly above chance. In addition, we also examine whether differences in inference-time reasoning effort help explain the performance gaps across models. The results show that models using more reasoning tokens do not consistently achieve higher accuracy, indicating that the observed performance gap cannot be attributed simply to greater inference-time reasoning effort. Detailed analyses of the heuristic baselines and reasoning effort are provided in Appendix C.1.1 and C.2.

4.3 Evaluation under Domain Congruency

We further examine structural inference in a setting to disentangle the capability to abstract structure from reliance on surface domain cues. The domains of reference and candidate contexts are sampled independently, with the correct candidate evaluated under two conditions: congruent ($\equiv$), where the domain of the correct candidate matches that of the reference, and incongruent ($\not\equiv$), where they differ. Specifically, in the $\equiv$ condition, the correct option is drawn from the same domain as the reference, while the remaining options are sampled from randomly selected domains. In the $\not\equiv$ condition, the correct option is required to come from a different domain, while the incorrect options are sampled so that at least one shares the reference domain.

As shown in Table 2, flagship models, such as GPT-5.4 and Grok-4, exhibit relatively small performance gaps between congruent and incongruent settings, indicating that their structural transfer performance remains robust across varied domains. In contrast, smaller models, such as GPT-5.4-nano and Grok-3-mini, show a stronger dependence on domain congruency, with marked performance degradation when the reference and candidate contexts belong to different domains.

Table 3: Variance in the performance of structural inference across domains. For each abstract structure, the average (*avg.*) and standard deviation (*std.*) across domains are reported.

	Chain		*Hub*		*Tree*		*Community*		*Rand.*	
	avg.	*std.*	*avg.*	*std.*	*avg.*	*std.*	*avg.*	*std.*	*avg.*	*std.*
GPT-5	0.99	0.02	1.00	0.00	0.96	0.05	1.00	0.00	0.98	0.02
GPT-5-nano	0.92	0.07	0.85	0.07	0.88	0.13	0.77	0.05	0.90	0.07
GPT-5.4	0.98	0.00	1.00	0.00	0.99	0.00	1.00	0.00	1.00	0.01
GPT-5.4-nano	0.98	0.04	0.99	0.02	0.98	0.00	0.99	0.02	0.97	0.05

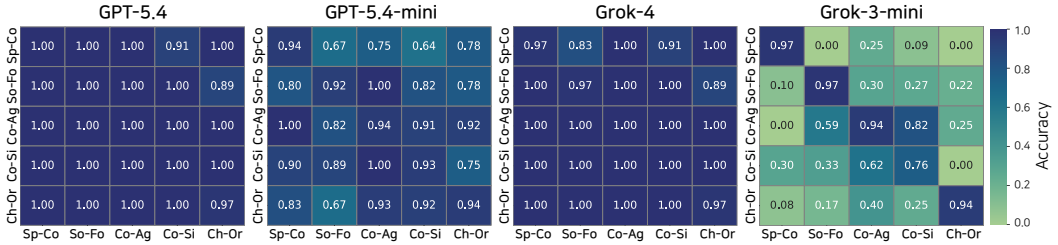


Figure 5: Heatmaps of performance on structural inference of LLMs, where columns indicate the domain of the reference context, and rows indicate that of the correct candidate. Domains are abbreviated: *Spatial-Contain* (Sp-Co), *Social-Follow* (So-Fo), *Comparison-Age* (Co-Ag), *Comparison-Size* (Co-Si), *Chronological-Order* (Ch-Or).

To examine this trend at a finer granularity, Figure 5 presents performance across individual pairs of reference and candidate domains. Smaller models, particularly GPT-5.4-mini and Grok-3-mini, exhibit pronounced disparities between domain pairs: darker diagonal regions indicate stronger performance when the domains are congruent, whereas brighter off-diagonal regions reflect substantial degradation when congruency cues are absent. Their often asymmetric off-diagonal patterns further indicate that transfer performance varies considerably depending on the specific direction of the domain pair. By comparison, flagship models maintain more uniform and symmetric accuracy across diverse domain pairs, suggesting less dependence on domain-specific signals. These patterns indicate that smaller models are more sensitive to domain cues rather than the underlying structure.

For heuristic baselines, the count-based heuristic also remains near chance in both conditions, whereas the embedding-based heuristic performs nearly perfectly in the congruent setting but collapses to near-zero accuracy in the incongruent setting. This contrast suggests that semantic similarity provides a strong shortcut in the congruent setting, thereby widening the gap between the two conditions when models rely on domain cues. Detailed results are provided in Appendix C.1.2.

In addition, we compare performance for each abstract structure across domains to examine whether the results are affected by structure-domain compatibility, as the same structure can become easier to infer in domains where training data of LLMs provides familiar relational patterns beyond the abstract structure itself. As shown in Table 3, we report the average and standard deviation across domain under the congruent setting. GPT-5.4 and GPT-5 exhibit low variation across domains, whereas GPT-5.4-nano and GPT-5-nano show relatively larger variation. These results suggest that high performance is accompanied by lower sensitivity across domains, indicating stable transfer of the underlying relational structure.

4.4 Scaling Analysis of Structure Sizes

We examine how the size of an abstract structure affects structural inference by evaluating model performance on tasks with increasing numbers of edges. As depicted in Figure 6, GPT-5.4, Gemini-2.5-Flash, and Grok-4 maintain strong performance as the number of edges increases, with only modest degradation on directed structures. This suggests that their structural inference remains stable even under larger abstract structures. However, smaller variants exhibit clearer performance degradation as structure size increases, with steeper declines on directed structures. This suggests that asymmetric relations impose additional constraints on structural inference, since directionality

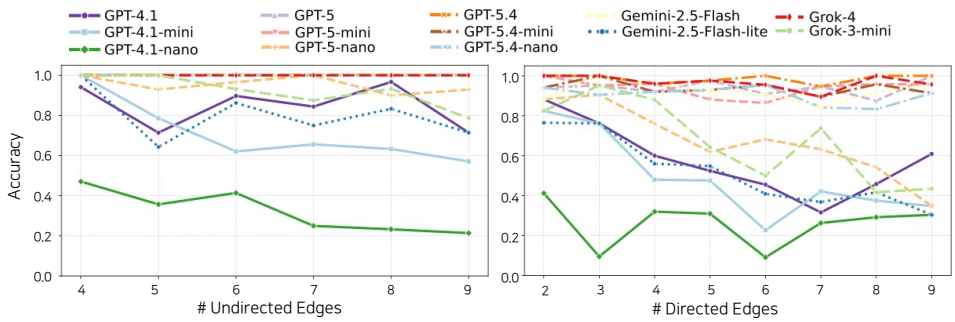


Figure 6: Performance on tasks constructed by random graphs with varying numbers of edges.

requires models to preserve not only connectivity but also the ordered composition of relations. Meanwhile, as the number of edges of structures increases, they often exhibit occasional spikes in accuracy. Prior work has revealed that LLMs tend to exhibit bias toward locally coherent patterns, such as triangles, in recall on graph structures (Wang et al., 2024). These local motifs may provide salient cues for smaller models, partially compensating for the increased difficulty of structural inference. Further analyses under recalling context settings and experimental results on open-source LLMs are provided in Appendix C.3 and C.4, respectively.

4.5 Human Audit of Generated Contexts

During contextualization, each generated sentence undergoes an iterative self-validation process. GPT-4.1 evaluates whether the sentence preserves the entities and relation specified by the source edge, including relation direction where applicable. If the sentence fails this evaluation, the validation feedback is used to regenerate it. This process is repeated up to three times, after which a predefined domain expression is used if the generated sentence still fails validation.

To further verify the fidelity of the generated sentences, we conducted an additional human audit. Four annotators who were not involved in this study independently evaluated 260 sentences uniformly sampled across all applicable combinations of structure families and domains. For each sentence, annotators were provided with the corresponding edges. They judged whether the sentence correctly represented both entities and their relation, including relation direction for directed domains. The generated sentences achieved a mean accuracy of 99.23%, with individual annotator scores ranging from 98.46% to 99.62%. The mean pairwise agreement across the six annotator pairs was 98.85%, and all four annotators agreed on 255 of the 260 sentences. These results indicate that the contextualization procedure preserves the predefined relational information with high reliability.

5 Discussion and Conclusion

We investigate whether LLMs can abstract and transfer structural knowledge across contexts in diverse domains. INSTILL constructs abstract structures from five representative families inspired by cognitive science, grounds them in nine fine-grained domains, and contextualizes them as cognitive scenarios described in natural language. Importantly, our work is not intended as a direct comparison between LLMs and humans. Humans acquire structural knowledge through real-world experience, including spatial navigation, social interaction, and temporal reasoning, whereas LLMs do not have direct access to the perceptual signals and encounter such structure primarily through language. Accordingly, we instantiate abstract relational structures as natural-language scenarios, providing a controlled setting for examining structural transfer in a modality accessible to language models.

Our findings suggest that recent LLMs exhibit a substantial capacity to transfer structural knowledge across contexts, although this capability remains dependent on model family and scale. Smaller models not only achieve lower overall accuracy but also show greater sensitivity to domain congruency and local relational patterns, suggesting reliance on superficial signals rather than the underlying abstract structure. In contrast, stronger models exhibit more consistent performance across these variations, indicating greater robustness in transferring the same underlying structure across contexts.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic data. As described in Section 3.3 and Appendix A.1, GPT-4.1 was used to verbalize predefined edges as natural language scenarios. The underlying graph structures, structural labels, and answer keys were produced programmatically, and the generated sentences underwent iterative validation and human audit.

Additionally, generative AI tools were used to create some of the illustrative images in Figure 4 solely to support the explanation and understanding of the corresponding domains. They were also used for proofreading, grammar correction, and minor stylistic edits to improve the readability of the manuscript. They were not used to develop the research ideas, methods, or claims. The authors reviewed all work produced with AI assistance and take responsibility for the final content of this work, including its text, claims, and generated artifacts.

We did not use generative AI tools to develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in writing proofs, propose or refine hypotheses, design or provide feedback on the research methodology or experiments, implement methods, assist with translation, clean or reformat datasets, support qualitative or thematic data analysis, or interpret the experimental results.

ETHICS STATEMENT

This study evaluates language models using synthetic natural language scenarios. It does not use private personal data or involve applications that directly affect individuals. The human audit involved four annotators who assessed whether generated sentences faithfully represented their corresponding edges. This component was reviewed and approved by the Institutional Review Board of the authors' institution. The approval identifier is withheld during anonymous review and will be provided in the final version. We do not identify any additional material ethical concerns arising from the study.

REPRODUCIBILITY STATEMENT

The methodology and task construction procedure are described in Section 3 and Appendix A. The experimental settings and evaluation protocol are provided in Section 4 and Appendix B. We also provide the implementation of INSTILL, representative samples of generated tasks, and a subset of model responses from the experiments in the anonymous repository provided in Section 4.1.

REFERENCES

Alexander R. Backus, Sander E. Bosch, Matthias Ekman, Alejandro Vicente Grabovetsky, and Christian F. Doeller. Mnemonic convergence in the human hippocampus. *Nature Communications*, 7(1):11991, 2016.

Timothy E.J. Behrens, Timothy H. Muller, James C.R. Whittington, Shirley Mark, Alon B. Baram, Kimberly L. Stachenfeld, and Zeb Kurth-Nelson. What is a cognitive map? organizing knowledge for flexible behavior. *Neuron*, 100(2):490–509, 2018.

Marcel Binz and Eric Schulz. Using cognitive psychology to understand GPT-3. In *Proceedings of the National Academy of Sciences*, volume 120, pp. e2218523120, 2023.

Matthew E. Brashears and Eric Quintane. The microstructures of network recall: How social networks are encoded and represented in human memory. *Social Networks*, 41:113–126, 2015.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Katherine M. Collins, Ilia Sucholutsky, Umang Bhatt, Kartik Chandra, Lionel Wong, Mina Lee, Cedegao E. Zhang, Tan Zhi-Xuan, Mark Ho, Vikash Mansinghka, Adrian Weller, Joshua B.

Tenenbaum, and Thomas L. Griffiths. Building machines that learn and think with people. *Nature Human Behaviour*, 8(10):1851–1863, 2024.

Russell A Epstein, Eva Zita Patai, Joshua B Julian, and Hugo J Spiers. The cognitive map in humans: spatial navigation and beyond. *Nature Neuroscience*, 20(11):1504–1513, 2017.

Mona M Garvert, Raymond J Dolan, and Timothy EJ Behrens. A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. *eLife*, 6:e17086, 2017.

Google Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2): 155–170, 1983.

Dedre Gentner and Arthur B. Markman. Structure mapping in analogy and similarity. *American Psychologist*, 52(1):45–56, 1997.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. In *Transactions of the Association for Computational Linguistics*, pp. 346–361, 2021.

Evan M. Gordon, Charles J. Lynch, Caterina Gratton, Timothy O. Laumann, Adrian W. Gilmore, Deanna J. Greene, Mario Ortega, Annie L. Nguyen, Bradley L. Schlaggar, Steven E. Petersen, Nico U.F. Dosenbach, and Steven M. Nelson. Three distinct sets of connector hubs integrate human brain function. *Cell Reports*, 24(7):1687–1695.e4, 2018.

Caterina Gratton, Haoxin Sun, and Steven E. Petersen. Control networks and hubs. *Psychophysiology*, 55(3):e13032, 2018.

Runquan Gui, Zhihai Wang, Jie Wang, Chi Ma, Huiling Zhen, Mingxuan Yuan, Jianye Hao, Defu Lian, Enhong Chen, and Feng Wu. Hypertree planning: Enhancing LLM reasoning via hierarchical thinking. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 20694–20722, 2025.

Thilo Hagendorff, Ishita Dasgupta, Marcel Binz, Stephanie C. Y. Chan, Andrew Lampinen, Jane X. Wang, Zeynep Akata, and Eric Schulz. Machine psychology. *arXiv preprint arXiv:2303.13988*, 2024.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. Reasoning with language model is planning with world model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8154–8173, 2023.

Stephen C. Hirtle and John Jonides. Evidence of hierarchies in cognitive maps. *Memory & Cognition*, 13(3):208–217, 1985.

Erik Jones and Jacob Steinhardt. Capturing failures of large language models via human cognitive biases. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pp. 11785–11799, 2022.

Charles Kemp and Joshua B. Tenenbaum. The discovery of structural form. *Proceedings of the National Academy of Sciences*, 105(31):10687–10692, 2008.

Andrew Liu, Henry Prior, Gargi Balasubramaniam, Rivka Moroshko, Amir Zait, Ilia Labzovsky, Danny Karmon, Ishita Dasgupta, Kim Stachenfeld, and Kenneth Marino. Recoglab: a framework testing relational reasoning & cognitive hypotheses on LLMs. In *The Thirteenth International Conference on Learning Representations*, pp. 45077–45106, 2025.

Christopher W. Lynn and Danielle S. Bassett. How humans learn and represent networks. *Proceedings of the National Academy of Sciences*, 117(47):29407–29415, 2020.

Shirley Mark, Rani Moran, Thomas Parr, Steve W. Kennerley, and Timothy E. J. Behrens. Transferring structural knowledge across cognitive maps in humans and models. *Nature Communications*, 11 (1):4783, 2020.

I. Momennejad, E. M. Russek, J. H. Cheong, M. M. Botvinick, N. D. Daw, and S. J. Gershman. The successor representation in human reinforcement learning. *Nature Human Behaviour*, 1(9): 680–692, 2017.

Ida Momennejad. Learning structures: Predictive representations, replay, and generalization. *Current Opinion in Behavioral Sciences*, 32:155–166, 2020.

Ida Momennejad, Hosein Hasanbeig, Felipe Vieira Frujeri, Hiteshi Sharma, Robert Osazuwa Ness, Nebojsa Jojic, Hamid Palangi, and Jonathan Larson. Evaluating cognitive maps and planning in large language models with cogeval. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp. 69736–69751, 2023.

Allen Newell. *Unified Theories of Cognition*. Harvard University Press, 1990.

OpenAI. Introducing gpt-4.1 in the api. `https://openai.com/index/gpt-4-1/`, 2024. Accessed: 2025-07-09.

OpenAI. Introducing gpt-5. `https://openai.com/blog/introducing-gpt-5`, 2025a. Accessed: 2025-08-07.

OpenAI. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025b.

OpenAI. Introducing gpt-5.4. `https://openai.com/index/introducing-gpt-5-4/`, 2026. Accessed: 2026-03-05.

Michael Peer, Iva K. Brunec, Nora S. Newcombe, and Russell A. Epstein. Structuring knowledge with cognitive maps and cognitive graphs. *Trends in Cognitive Sciences*, 25(1):37–54, 2021.

Jean Piaget. *The Origins of Intelligence in Children*. International Universities Press, 1952.

Evan M. Russek, Ida Momennejad, Matthew M. Botvinick, Samuel J. Gershman, and Nathaniel D. Daw. Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLoS Computational Biology*, 13(9):e1005768, 2017.

Jae-Young Son, Apoorva Bhandari, and Oriel FeldmanHall. Abstract cognitive maps of social network structure aid adaptive inference. *Proceedings of the National Academy of Sciences*, 120 (47):e2310801120, 2023.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Johan Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew Kyle Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cesar Ferri, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Christopher Waites, Christian Voigt, Christopher D Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, C. Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar,

Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germàn Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Xinyue Wang, Gonzalo Jaimovitch-Lopez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Francis Anthony Shevlin, Hinrich Schuetze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B Simon, James Koppel, James Zheng, James Zou, Jan Kocon, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh Dhole, Kevin Gimpel, Kevin Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros-Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje Ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramirez-Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael Andrew Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Andrew Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Russ Salakhutdinov, Ryan Andrew Chi, Seungjae Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel Stern Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Shammie Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven Piantadosi, Stuart Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsunori Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, vinay uday prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Sophie Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023.

Yasuaki Sumita, Koh Takeuchi, and Hisashi Kashima. Cognitive biases in large language models: A survey and mitigation experiments. In *Proceedings of the 40th ACM/SIGAPP Symposium on*

Applied Computing, pp. 1009–1011, 2025.

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp. 38975–38987, 2023.

David Ed Waller and Lynn Ed Nadel. *Handbook of spatial cognition.* American Psychological Association, 2013.

Yanbang Wang, Hejie Cui, and Kleinberg Jon. Microstructures and accuracy of graph recall by large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pp. 65154–65182, 2024.

Taylor Webb, Keith J. Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541, 2023.

Cornelius Weiller, Marco Reisert, Pierre Levan, Jonas Hosp, Volker A Coenen, and Michel Rijntjes. Hubs and interaction: the brain's meta-loop. *Cerebral Cortex*, 35(3):bhaf035, 2025.

xAI. Grok 3 beta — the age of reasoning agents. `https://x.ai/blog/grok-3`, 2025. Accessed: 2025-02-21.

Ilker Yildirim and L.A. Paul. From task structures to world models: what do LLMs know? *Trends in Cognitive Sciences*, 28(5):404–415, 2024.

A Additional Details for Instill

In this section, we provide the additional details and examples for Instill. Our codes for Instill, representative samples of generated tasks, and a subset of model responses from the experiments are available at `https://anonymous.4open.science/r/instill`.

A.1 Detailed Procedure and Prompts of Contextualization

In Instill, the domains used for the contextualization step were chosen to model real-world situations. Consequently, each fine-grained domain imposes inherent constraints on the structural families it can plausibly exhibit. For example, in the chronological-order domain, which represents temporal relations among events, the structure cannot take the form of a community structure that contains cliques. To consider these constraints and to exclude domain–structure combinations that are rarely observed in real-world, we generated problems with specific combinations. The combinations employed in Instill can be found in Figure 7.

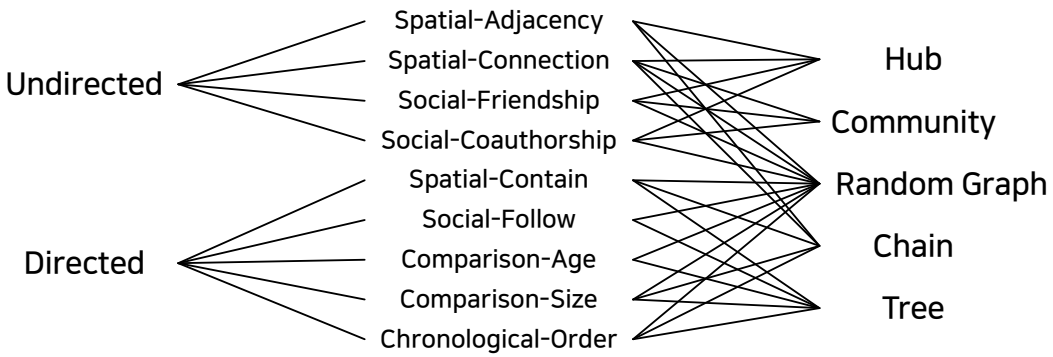
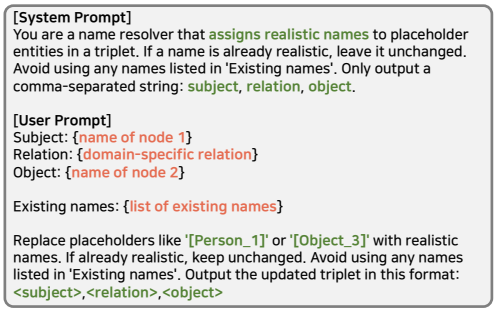


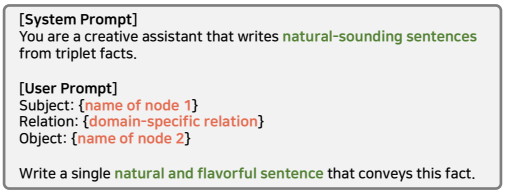
Figure 7: Combinations of edge directions, domains, and structures employed for the problem.

The abstract structures are contextualized into cognitive scenarios by traversing each edge and generating natural-language sentences with GPT-4.1. For each edge, if one or more of its nodes have a placeholder name, Instill assigns names to those nodes. Figure 8a shows the prompt used in this process. To provide domain information for naming, the prompt also includes the relation represented in the current edge. The generation is performed with a temperature of 1.0, and the output token length is limited to 10 tokens. To prevent duplicate names, the prompt contains a list of already assigned names, and the model's output is validated up to three times to ensure that no duplicate names are introduced.

After the names of both nodes are assigned, GPT-4.1 generates a natural-language sentence using the node names together with the relation associated with the selected domain. Figure 8b shows the prompt used for generating such sentences. Since the generated sentence may fail to accurately reflect the given node names or relation, it undergoes up to three rounds of validation. Figure 9a presents the prompt used for validation. At each validation step, GPT-4.1 evaluates whether the generated sentence is semantically consistent with the default sentence composed of node names and domain-specific relations, based on two criteria. The first criterion assesses whether the names of the entities appearing in the generated sentence match those in the default sentence. The second criterion evaluates whether the relationships between the entities accurately reflect the relations specified in the default sentence. If the validation result is "no," the model subsequently makes an explanation for the failure. The explanation is then provided back to the model, as shown in Figure 9b, to guide the regeneration of the sentence. If all three validation attempts fail, the final sentence is constructed directly using the node names and the relation without further modification. Figure 10 visualizes the full process of generating a cognitive scenario for a single edge during contextualization. Figure 11 illustrates example contexts generated for each domain.

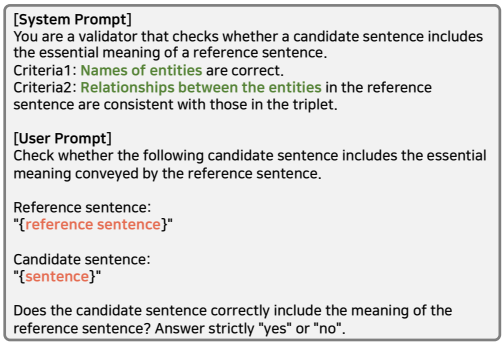


(a) Node Name Assignment Prompts

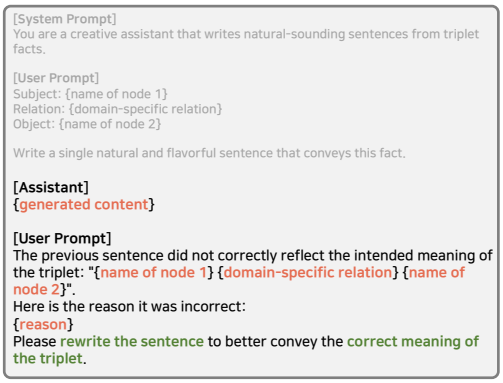


(b) Cognitive Scenario Generation Prompts

Figure 8: User and system prompts used for generating cognitive scenarios.



(a) Sentence Validation Prompts



(b) Cognitive Scenario Re-generation Prompts

Figure 9: Prompts used for validating generated sentences and re-generating cognitive scenarios.

A.2 Examples of Generated Problems

INSTILL generates problems that require identifying a candidate context whose underlying structure matches that of the reference context. Figures 14a, 14b, 15a, 16, and 15b present example problems, each consisting of contexts whose underlying structures correspond to chain, hub, tree, community, and random graph structures, respectively. Figure 13a and 13b show examples of the generated problems with domains sampled independently.

B Experimental Details

We generated 250 problems for each structure in the experiments in Section 4.2 and Section 4.3 and 1,000 problems for the random graph for scaling analysis in Section 4.4. The temperature of LLMs was fixed at zero in all experiments. However, as LLM outputs may still vary slightly because of stochasticity and nondeterminism in the decoding process, we repeated each experiment three times to obtain more reliable results. The maximum token length was set to the default for closed-source models and to 4096 tokens for open-source models. For open-source LLMs, we use a batch size of 16. We report results using exact match, averaged across three runs. Since LLMs typically generate intermediate reasoning before presenting a final answer, we extract the prediction using predefined answer markers, such as "*Answer*:" and "*the answer is*.". We then obtain the final prediction after removing any leading and trailing characters. Responses that could not be parsed were treated as incorrect as they failed to follow the instructions.

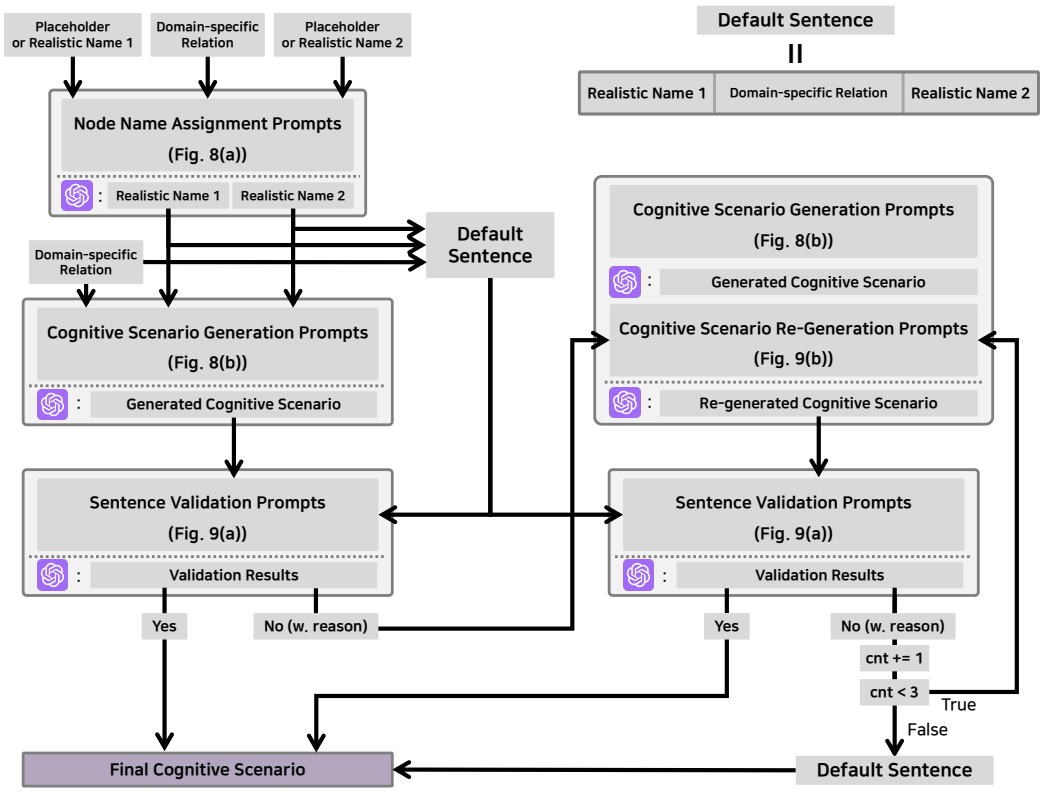


Figure 10: Full process for contextualization of an edge.

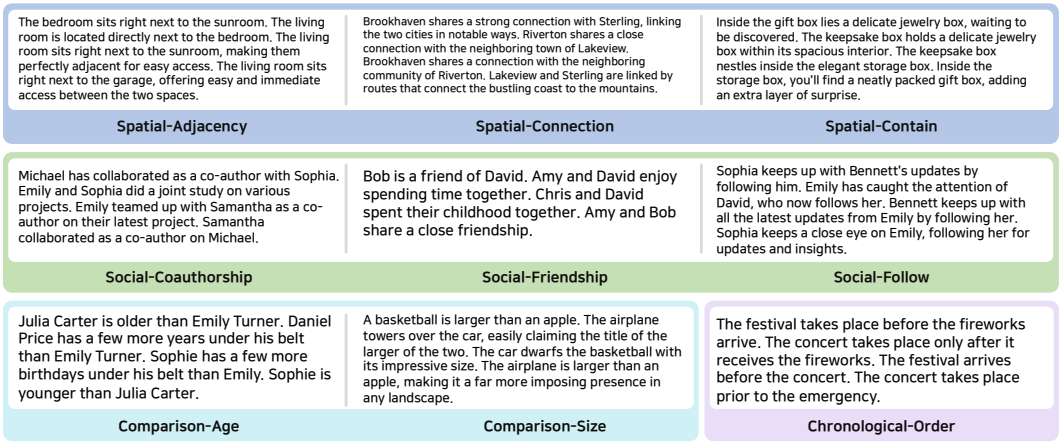


Figure 11: Example contexts generated for each domain.

C Additional Experiments

C.1 Experimental Results of Heuristic Baselines

We evaluate three heuristic baselines to examine whether INSTILL can be solved using simple graph statistics or semantic similarity. The *simple count* selects a candidate whose number of nodes matches that of the reference context. When multiple candidates have the same node count as the reference, the baseline predicts uniformly at random among the candidates. The *embedding similarity* encodes the reference and candidate contexts using BERT, computes their cosine similarities, and selects the candidate with the highest similarity. Since each task contains four candidates, uniformly random prediction yields an expected accuracy of 0.25.

Table 4: Performance of the node-count, edge-count, and embedding-similarity heuristic baselines, measured by accuracy. *Rand.* and *All* indicate random graphs and the average over structures.

	Chain	*Hub*	*Tree*	*Community*	*Rand.*	*All*
Simple count	0.25 ± 0.00	0.25 ± 0.00	0.25 ± 0.00	0.25 ± 0.00	0.25 ± 0.00	0.25 ± 0.00
Embedding similarity	0.21 ± 0.00	0.32 ± 0.00	0.25 ± 0.00	0.38 ± 0.00	0.31 ± 0.00	0.30 ± 0.00

Table 5: Performance of the node-count, edge-count, and embedding-similarity heuristic baselines under congruent ($\equiv$) and incongruent ($\not\equiv$) settings. *Rand.* and *All* indicate random graphs and the average over structures.

	Chain		*Hub*		*Tree*		*Community*		*Rand.*		*All*	
	$\equiv$	$\not\equiv$	$\equiv$	$\not\equiv$	$\equiv$	$\not\equiv$	$\equiv$	$\not\equiv$	$\equiv$	$\not\equiv$	$\equiv$	$\not\equiv$
Simple count	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
Embedding similarity	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.01	1.00	0.00

C.1.1 Evaluation of Heuristic Baselines across Structural Families

Table 4 reports the results where the reference and candidate contexts are drawn from the same domain. The *simple count* baseline performs near chance, reflecting our task construction, which controls for underlying structure by matching the number of entities across candidates whenever possible. The *embedding similarity* baseline achieves only slightly higher accuracy than random prediction, indicating that semantic or lexical similarity provides little information for generalizing the abstract structure when domain information is shared across all contexts. These results suggest that the single-domain tasks cannot be reliably solved using simple shortcuts or semantic similarity.

C.1.2 Evaluation of Heuristic Baselines under Domain Congruency

Table 5 presents the results under the congruent ($\equiv$) and incongruent ($\not\equiv$) settings. The *simple count* baseline also remains close to random prediction in both settings. In contrast, the *embedding similarity* baseline exhibits different behavior across the two settings. Its accuracy approaches 1.0 in the congruent setting, but falls close to 0 in the incongruent setting, where the correct candidate belongs to a different domain, showing how reliance on domain cues can widen the performance gap between the two experimental settings.

Taken together, these findings show that simple counting does not introduce a predictive bias in either evaluation setting, while embedding similarity captures the intended semantic alignment of the congruent setting without solving the underlying structural-matching problem. Although these baselines do not rule out every possible shortcut, they provide evidence that INSTILL cannot be reliably solved through simple graph-size heuristics or surface-level embedding similarity alone.

Table 6: Average number of reasoning tokens across structure families.

	Chain	*Hub*	*Tree*	*Community*	*Rand.*	*All*
Gemini-2.5-Flash	1849	2021	1781	1985	1714	1821
Gemini-2.5-Flash-lite	4816	4921	5197	8930	3122	4424
GPT-4.1	1377	1467	1276	1522	1296	1339
GPT-4.1-mini	1338	1242	1167	1172	1351	1292
GPT-4.1-nano	693	634	728	570	750	704

C.2 Additional Analysis of Reasoning Effort

We further examine whether differences in structural inference performance can be explained by the amount of inference-time reasoning used by each model. Table 6 reports the average number of reasoning tokens generated during inference. We observe no consistent monotonic relationship between reasoning length and accuracy. In particular, Gemini-2.5-Flash-Lite generates substantially more reasoning tokens than Gemini-2.5-Flash across all structure families, while exhibiting considerably lower accuracy. Within the GPT-4.1 family, GPT-4.1 and GPT-4.1-mini use comparable numbers of reasoning tokens despite a substantial difference in performance. These results suggest that the weaker structural inference of smaller models cannot be attributed solely to shorter reasoning traces. However, this analysis is not intended to establish a causal relationship between reasoning length and performance, as reasoning and tokenization strategies differ across model families.

Table 7: Performance of structural inference under the original (*org.*) and memory clearance setting (*mem.*). *Rand.* and *All* indicate random graphs and the average over structures.

	Chain		*Hub*		*Tree*		*Community*		*Rand.*		*All*	
	org.	*mem.*	*org.*	*mem.*	*org.*	*mem.*	*org.*	*mem.*	*org.*	*mem.*	*org.*	*mem.*
GPT-5.4	0.99	0.99	1.00	1.00	0.97	0.98	0.99	0.99	0.99	0.99	0.99	0.99
GPT-5.4-nano	0.92	0.88	1.00	0.98	0.93	0.93	0.99	0.99	0.95	0.95	0.96	0.94
Grok-4	0.98	0.99	1.00	1.00	0.98	0.98	0.99	0.97	0.98	0.97	0.98	0.98
Grok-3-mini	0.58	0.41	0.82	0.57	0.55	0.41	0.35	0.38	0.78	0.66	0.62	0.55

```
[User]
Given context: {context of the reference context}

Before we proceed to the recall part, we will perform an exercise with
you. You will be given a series of sentences, grouped in sets of three
sentences. At the end of each set, you will need to recall the last word in
each preceding set. Over time, you will be asked to remember words
from larger sets of sentences. If you have any questions, please ask now.
Otherwise, please inform us that you are ready to begin.

[Assitant]
{intro response}

[User]
{set of sentences for memory clearance}
```

(a) Templates for memory clearance experiments.

```
[Set 1 for Memory Clearance]
"One afternoon I open a letter from my younger sister, the photo chronicler of family events."
"Although they have enough food to sustain your group for years, supermarkets are also
dangerous."
"Maternity, or additional offspring, may force upon the woman a distressful life and future."
"By Wednesday night's vote meeting, Sabrina was thoroughly disgusted by the superficiality of
the week."
"I ask him to please stop lying about trying to take my place in the war."
"Since the 1950's, freeways have been built through every large and medium-sized city."
"Instead, he took a job in Washington, analyzing weapons expenditures for the U.S. Navy."

[Set 2 for Memory Clearance]
"She sprinted down the stairs into the backyard, screaming and cursing into her cell."
"The structures assigned to me have been in-country since the beginning of the deployment."
"Some Third World dictator might hold his minions together only as long as he survives."
"If this kind of teacher cheating is truly going on, how might it be detected?"
"The other students, still feverishly writing, didn't notice the change in the tension in the room."
"During the few months Troy had been back home, he'd told his friends about us."
"What is certain is that they will always be on the lookout for loot."

[Set 3 for Memory Clearance]
"It has to do with thinking sensibly about how people behave in the real world."
"Caitlin, almost laughing at his belligerence, tossed his jacket into the empty hallway."
"Most of the men had long hair and wore heavy-metal t-shirts and ripped jeans."
"When your team hits the road, everyone should know what's expected of him or her."
"As the world has grown more specialized, countless such experts have made themselves similarly
indispensible."
"But she was upset that she couldn't go with the date she wanted to take."
"The salts were tired and mean and the last thing they needed was a war."
```

(b) Sets of sentences for memory clearance experiments.

Figure 12: Templates and sentences for memory clearance experiments.

C.3 Evaluation under Recalling Contexts

We evaluate structural inference in a more challenging setting that requires models to transfer an abstract structure under an intervening memory load between exposure to the reference and candidate contexts. This setting allows us to examine whether models can preserve an abstracted structure under an additional memory load and subsequently transfer it across candidate contexts. For each task, with all contexts drawn from the same domain, we first present the reference context and insert a memory clearance stage before providing the candidates.

Following (Wang et al., 2024), we apply a word span test commonly used to measure human working memory capacity. In each round of the word span test, the model is presented with seven sentences and instructed to return the final word of each sentence in the original order. The test is repeated until the model produces an incorrect response, for a maximum of three rounds. After the word span test terminates, the candidate contexts are presented and the model performs the structural matching task using the previously observed reference context. The template prompt and sets of sentences for memory clearance experiments are provided in Figure 12a and Figure 12b.

As LLMs do not maintain an independent memory store and the reference context remains available within the input, this stage does not literally clear or erase the model's memory. Nevertheless, we refer to this intervention as *memory clearance* to maintain consistency with the terminology used in related literature and to characterize the intervening task as shifting attention away from the reference before the model uses it to evaluate the candidate contexts. As shown in Table 7, we conduct the experiments on models that demonstrated competitive performance in the main evaluation, along with their smaller variants. Flagship models exhibit only marginal performance degradation after the intervention, indicating that they can reliably recall the reference context and preserve its abstract structure. In contrast, models that perform less reliably in the original setting, such as Grok-3-mini, suffer substantially larger declines, indicating that their structural inference is more vulnerable to interference from the intervening task.

C.4 Additional Experimental Results on Open-source LLMs

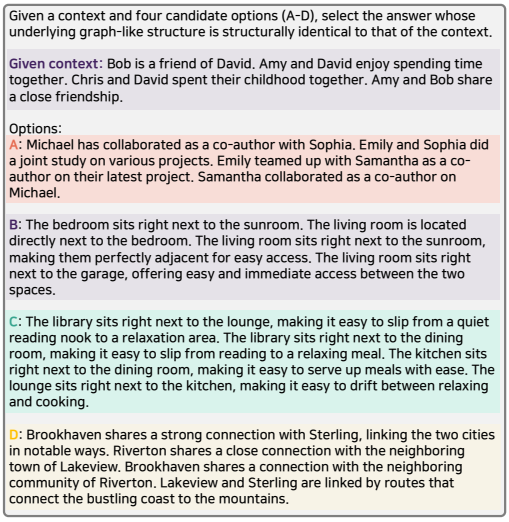
We conduct additional experiments on OpenAI's open-source LLMs: GPT-OSS-20B and GPT-OSS-120B (OpenAI, 2025b). We follow the same task construction and evaluation protocols as in the main

Table 8: Performance of structural inference across families of abstract structures, measured by accuracy. *Rand.* and *All* indicate random graphs and the average over structures.

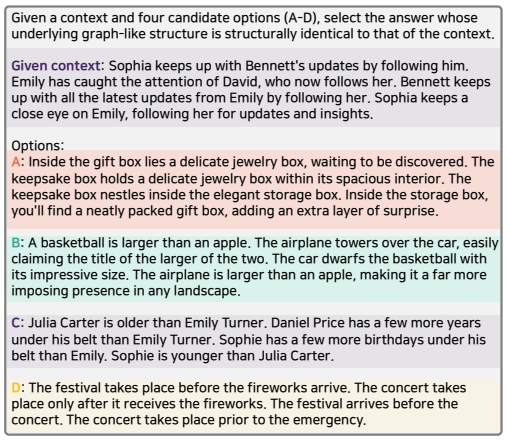
	Chain	*Hub*	*Tree*	*Community*	*Rand.*	*All*
GPT-OSS-120b	0.65 ± 0.01	0.78 ± 0.03	0.55 ± 0.02	0.09 ± 0.00	0.81 ± 0.02	0.57 ± 0.00
GPT-OSS-20b	0.38 ± 0.01	0.36 ± 0.05	0.36 ± 0.04	0.00 ± 0.00	0.56 ± 0.00	0.33 ± 0.02

Table 9: Performance of structural inference across families of abstract structures under congruent ($\equiv$) and incongruent ($\not\equiv$) settings. *Rand.* and *All* indicate random graphs and the average over structures.

	Chain		*Hub*		*Tree*		*Community*		*Rand.*		*All*	
	$\equiv$	$\not\equiv$	$\equiv$	$\not\equiv$	$\equiv$	$\not\equiv$	$\equiv$	$\not\equiv$	$\equiv$	$\not\equiv$	$\equiv$	$\not\equiv$
GPT-OSS-120b	0.85	0.53	0.90	0.84	0.83	0.43	0.23	0.08	0.89	0.61	0.74	0.50
GPT-OSS-20b	0.62	0.30	0.46	0.43	0.74	0.25	0.04	0.00	0.65	0.42	0.50	0.28



(a) An example problem with undirected relations. Each context is instantiated from a different domain: the reference context corresponds to *Social-Friendship*, while Context A is drawn from *Social-Coauthorship*, Contexts B and C from *Spatial-Adjacency*, and Context D from *Spatial-Connection*.



(b) An example problem with directed relations. Each context is instantiated from a different domain: the reference context corresponds to *Social-Follow*, while Context A is drawn from *Spatial-Contain*, Context B from *Comparison-Size*, Context C from *Comparison-Age*, and Context D from *Chronological-Order*.

Figure 13: Example problems with undirected (left) and directed (right) relations across diverse domains.

experiments: Table 8 corresponds to the single-domain setting in Section 4.2, while Table 9 evaluates the same congruent and incongruent conditions introduced in Section 4.3. All experiments on the open-source models are conducted using an NVIDIA RTX A6000. As shown in Tables 8 and 9, GPT-OSS-20B achieves lower overall performance than GPT-OSS-120B and exhibits a substantially larger accuracy drop from the congruent to the incongruent setting. This pattern is broadly similar to that observed in the closed-source models, with relatively strong performance on hub and random structures and weaker performance on chains and trees. Both GPT-OSS models perform particularly poorly on community structures, in line with the pronounced degradation observed for smaller models in the main experiments, such as Grok-3-mini relative to Grok-4. This may reflect the greater difficulty of recovering community structure from distributed relations. Both models also exhibit lower accuracy in the incongruent setting, indicating that structural transfer becomes more difficult when domain similarity no longer aligns with the correct candidate. The open-source models show patterns similar to those observed in the main experiments, including differences across model scales, structural families, and sensitivity to domains.

Given a context and four candidate options (A-D), select the answer whose underlying graph-like structure is structurally identical to that of the context.

Given context: Mount Elbrus towers over Denali in size, making it the larger of the two impressive peaks. Everest towers over Kilimanjaro, claiming the title of the larger mountain. Mount Elbrus towers over Mount A in size, making it the larger of the two peaks. Mount Elbrus, while impressive, stands shorter than Africa's towering Mount Kilimanjaro. Mount Kosciuszko is dwarfed by Kilimanjaro, standing at a much lower elevation than the African giant.

Options:

A: Sand is made up of much finer particles than gravel, making it noticeably smaller in size. A rock is generally bigger in size than a stone. Gravel is made up of pieces that are actually smaller than pebbles. A boulder is much larger than a typical rock. A pebble is noticeably smaller than a rock.

B: Compared to a rock, a stone is generally larger in size. A stone is bigger in size compared to a slab. A chip is even smaller than a pebble. A boulder dwarfs a stone in sheer size. A pebble is noticeably smaller than a rock.

C: A pebble is much smaller than a boulder. A grain is noticeably larger than a tiny speck of sand. A pebble is bigger than a grain. Sand particles are bigger in size than those of clay. Silt consists of much finer particles than grain, making it noticeably smaller in size.

D: Dust particles are even smaller than grains, making them nearly invisible to the naked eye. A speck is even tinier than a grain. A grain is much smaller than a pebble. A shard is even smaller than a pebble, just a tiny fragment compared to its sturdier counterpart. A pebble is much smaller than a rock.

(a) An example problem consisting of contexts whose underlying structure is a chain structure.

Given a context and four candidate options (A-D), select the answer whose underlying graph-like structure is structurally identical to that of the context.

Given context: Olivia teamed up with Grace Patel as a co-author. Benjamin Ross teamed up with Olivia as a co-author. Daniel Lee and Michael Nguyen have collaborated as co-authors on their work. Emily Carter has collaborated with Daniel Lee as a co-author. Samantha Ortiz collaborated with Daniel Lee as a co-author. Daniel Lee collaborated with Olivia as a co-author. Olivia teamed up with Victor Chen to co-author their latest work. Daniel Lee teamed up with Victor Chen as a co-author on their work. Olivia has worked alongside Samantha Ortiz as a co-author.

Options:

A: Julia Chen teamed up with Michael as a co-author on their latest project. Julia Chen collaborated with Samantha Lee as a co-author. Alex Turner collaborated with Samantha Lee as a co-author. Michael teamed up with David Patel as a co-author. Samantha Lee collaborated as a co-author on "Grace." Samantha Lee collaborated as a co-author with David Patel. Nathan Kim and Samantha Lee teamed up as co-authors on their latest work. Samantha Lee teamed up with Rachel as a co-author. Samantha Lee teamed up with Michael as a co-author.

B: Sophia Kim collaborated with Jessica as a co-author. Sophia Kim collaborated with Ethan as a co-author on their work. Olivia Park teamed up with Sophia Kim as a co-author on the project. Sophia Kim teamed up with David as a co-author. Emma Li collaborated with Sophia Kim as a co-author. Sophia Kim collaborated with Michael Chen as a co-author. Michael Chen collaborated with David as a co-author. Michael Chen collaborated with Ethan as a co-author. Lucas Wang has collaborated with Sophia Kim as a co-author.

C: Olivia Reed and Emily Chen have collaborated as co-authors. Nathan Rivera collaborated with Emily Chen as a co-author. Olivia Reed collaborated with Amelia as a co-author. Emily Chen collaborated with Liam Patel as a co-author. Emily Chen collaborated with James Parker as a co-author. Sophia collaborated with Amelia as a co-author. Emily Chen collaborated with Sophia as a co-author. Aiden Brooks collaborated with Olivia Reed as a co-author. Sophia teamed up with Olivia Reed as a co-author.

D: Emily Carter teamed up with Jonathan Kim as a co-author. Gabriel Torres teamed up with Jonathan Kim as a co-author. Marcus Delgado collaborated with Gabriel Torres as a co-author. Madison Lee teamed up with Jonathan Kim as a co-author. Lucas Ramirez teamed up with Gabriel Torres as his co-author on a recent project. Olivia Bennett teamed up with Jonathan Kim as a co-author. Gabriel Torres teamed up with Olivia Bennett as a co-author. Gabriel Torres and Madison Lee have collaborated together as co-authors. Sophie Nguyen collaborated with Jonathan Kim as a co-author.

(b) An example problem consisting of contexts whose underlying structure is a hub structure.

Figure 14: Example problems with different underlying structures: chain (left) and hub (right).

Given a context and four candidate options (A-D), select the answer whose underlying graph-like structure is structurally identical to that of the context.

Given context: Ethan Patel keeps up with Charlotte Kim's updates, following her activities with interest. Olivia Martinez keeps up with Lucas Bennett by following him. Ava Thompson counts Charlotte Kim among her followers. Ethan Patel keeps up with Olivia Martinez, following her updates closely. Charlotte Kim keeps up with Sophie Nguyen by following her.

Options:

A: Emily Carter keeps up with Olivia Brooks by following her. James Turner is one of Emily Carter's followers. James Turner keeps up with the latest from Sophia Green, following her updates closely. Ethan Bennett keeps up with Lucas Hayes by following him. James Turner keeps up with Ethan Bennett by following him.

B: Emily Chen keeps up with Jake Morrison by following him. Jake Morrison keeps up with the latest from Lucas Bennett by following him. Olivia Watkins counts Jake Morrison among her followers. Olivia Watkins keeps up with Madison Turner's updates by following her. Jake Morrison keeps up with Samantha Reyes by following her updates.

C: David Chen keeps up with Michael Patel by following him. Emily Parker keeps up with Sophia Martinez by following her. Sophia Martinez keeps up with Ava Brooks, following her updates and activities. Emily Parker keeps up with David Chen by following him. Sophia Martinez is one of Olivia Thompson's followers.

D: Emma Parker keeps up with Lucas Bennett by following him. Isabella Hayes is succeeded by Emma Parker. Lucas Bennett keeps a close eye on Olivia Reynolds, following her every move. Emma Parker keeps up with Daniel Kim by following him. Isabella Hayes follows Ava Patel.

(a) An example problem consisting of contexts whose underlying structure is a tree structure.

Given a context and four candidate options (A-D), select the answer whose underlying graph-like structure is structurally identical to that of the context.

Given context: The Spring Festival kicks off just before the arrival of the Lunar New Year. The Winter Solstice arrives each year before the celebration of the Lunar New Year. The Winter Solstice arrives after autumn has run its course, ushering in the shortest day of the year.

Options:

A: The Opening Ceremony takes place before the start of the Annual Science event. The Annual Science Fair takes place following the Opening Ceremony. The Robotics Workshop takes place before the Annual Science event.

B: The Spring Art Exhibition takes place ahead of the Winter Debate. The Annual Science Fair takes place after the Math Olymp, giving students back-to-back opportunities to showcase their talents. The Math Olymp takes place before the Winter Debate gets underway.

C: The Early Bloom Parade takes place just before the Vernal Festival kicks off. The Vernal Festival takes place just before the Spring Equinox arrives. The Spring Equinox arrives just before the official first day of spring.

D: The Quarterly Review takes place ahead of the Team Meeting. The Annual Conference takes place prior to the Team Meeting. The quarterly review takes place following the product launch.

(b) An example problem consisting of contexts whose underlying structure is a random graph.

Figure 15: Example problems with different underlying structures: tree (left) and random graph (right).

Given a context and four candidate options (A-D), select the answer whose underlying graph-like structure is structurally identical to that of the context.

Given context: Lucas Bennett collaborated with David Chen as a co-author. Matthew teamed up with Charlotte as a co-author on a project. Emma Dawson collaborated with Jacob Lin as a co-author. Lucas Bennett collaborated with Grace Morgan as a co-author. Ethan Turner collaborated with Lucas Bennett as a co-author. Hannah Wells collaborated with Charlotte as a co-author. Hannah Wells teamed up with David Chen as a co-author. Ethan Turner collaborated with David Chen as a co-author. Olivia teamed up with Lucas Bennett as a co-author. Hannah Wells teamed up with Emma Dawson as her co-author. Grace Morgan teamed up with Ethan Turner as a co-author. David Chen and Olivia teamed up as co-authors on a project. Ethan Turner collaborated with Olivia as a co-author. Sophia Patel and Emma Dawson have collaborated as co-authors. Sophia Patel teamed up with Jacob Lin as a co-author. Miles Harper collaborated with Riley Thompson as a co-author. Miles Harper collaborated with Grace Morgan as a co-author. Hannah Wells teamed up with Matthew as a co-author. Grace Morgan collaborated with Riley Thompson as a co-author. Hannah Wells collaborated with Jacob Lin as a co-author.

Options:

A: Olivia teamed up with Ethan Price as a co-author. Daniel Foster collaborated with Olivia as a co-author. Christopher Hayes collaborated with Olivia as a co-author. Christopher Hayes and Amanda Phillips have collaborated as co-authors. Jessica Ramirez collaborated with Michael Carter as a co-author. Lauren Mitchell has collaborated with Christopher Hayes as a co-author. Daniel Foster collaborated with Sarah Bennett as a co-author. Daniel Foster teamed up with Ethan Price as a co-author. David Nguyen collaborated with Sarah Bennett as one of the co-authors on their recent work. Amanda Phillips collaborated with Lauren Mitchell as a co-author. Emily Chen collaborated with Michael Carter as a co-author. Emily Chen collaborated with Jessica Ramirez as a co-author. David Nguyen teamed up with Nathan Lee as a co-author on their latest work. David Nguyen teamed up with Rachel Kim as a co-author. Nathan Lee teamed up with Rachel Kim as a co-author. Sarah Bennett collaborated with Olivia as a co-author. Daniel Foster has collaborated as a co-author with Christopher Hayes. Michael Carter and David Nguyen have collaborated as co-authors. Sarah Bennett collaborated with Ethan Price as a co-author. Emily Chen and David Nguyen have collaborated as co-authors on their work.

B: Jonathan Reed collaborated as a co-author alongside Michael Turner. Ava Martinez collaborated with Ethan Sullivan as a co-author. Ava Martinez collaborated with Olivia as a co-author. Samantha Brooks collaborated with Rachel Morgan as a co-author. Ava Martinez teamed up with Nathan Parker as a co-author on their latest project. Ethan Sullivan teamed up with Sarah Thompson as a co-author. David Wilson has collaborated closely with Linda Patel as a co-author. Jessica Nguyen collaborated with Jonathan Reed as a co-author. Linda Patel collaborated with Michael Turner as a co-author. Jonathan Reed teamed up with Emily Chen as a co-author. Michael Turner collaborated with David Wilson as a co-author. Ethan Sullivan teamed up with Olivia as a co-author. Emily Chen collaborated with Linda Patel as a co-author. David Wilson teamed up with Jonathan Reed as a co-author. Nathan Parker teamed up with Olivia as a co-author. Rachel Morgan and Linda Patel have teamed up as co-authors. Jessica Nguyen collaborated as a co-author alongside Samantha Brooks. Sarah Thompson has collaborated with Rachel Morgan as a co-author on published works. Sarah Thompson collaborated with David Wilson as a co-author. Olivia has teamed up with Sarah Thompson as her co-author. Emily Chen collaborated with Michael Turner as a co-author. Ethan Sullivan teamed up with Nathan Parker as a co-author. Sarah Thompson collaborated with Nathan Parker as a co-author. Linda Patel teamed up with Jonathan Reed as a co-author. Sarah Thompson collaborated with Ava Martinez as a co-author. Jessica Nguyen teamed up with Rachel Morgan as a co-author. Emily Chen has collaborated with David Wilson as a co-author.

C: Michael Thompson collaborated with Emily Carter as her co-author. Liam teamed up with Avery Parker to co-author their latest work. Olivia Nguyen collaborated with Ethan as a co-author on their work. Maya Patel and Ben teamed up as co-authors on their recent work. David Chen teamed up with Michael Thompson as a co-author on a recent project. Rachel Evans collaborated with Ben as a co-author. Isabella Martinez collaborated with Michael Thompson as a co-author. Rachel Evans collaborated with Maya Patel as a co-author. Isabella Martinez teamed up with David Chen as a co-author on their latest work. Sophia teamed up with Emily Carter as a co-author. Olivia Nguyen teamed up with David Chen as a co-author on their recent work. Ben collaborated with Lucas Anderson as a co-author. Isabella Martinez collaborated with Ben as a co-author. Avery Parker teamed up with Olivia Nguyen as a co-author. Emily Carter teamed up with Isabella Martinez as her co-author. Ethan shares co-authorship with Liam. Isabella Martinez collaborated with Sophia as a co-author. Avery Parker teamed up with Ethan as a co-author. Olivia Nguyen collaborated with Liam as a co-author. Lucas Anderson teamed up with Maya Patel as a co-author on their latest work. David Chen collaborated with Sophia as a co-author. Sophia teamed up with Michael Thompson as a co-author. Rachel Evans and Lucas Anderson have worked together as co-authors. Emily Carter collaborated closely with David Chen as his co-author.

D: Noah Sullivan and Charlotte Pierce teamed up as co-authors on their latest work. Is collaborated as a co-author with Olivia Bennett. Noah Sullivan teamed up with Henry Wells as a co-author. Sophia Nguyen collaborated with James as a co-author. Maya Brooks and Ethan Rivera teamed up to co-author their latest work. Grace Foster collaborated with Noah Sullivan as a co-author. James and Ethan Rivera have collaborated as co-authors. Olivia Bennett teamed up with Ethan Rivera as a co-author. Sophia Nguyen collaborated with Lucas Patel as a co-author. Grace Foster teamed up with Olivia Bennett as a co-author. Ava Thompson teamed up with James as a co-author. Maya Brooks contributed as a co-author to the work titled "Is." Sophia Nguyen contributed as a co-author to the book "Is." Maya Brooks teamed up with Olivia Bennett as a co-author. Emma Chen teamed up with Lucas Patel as a co-author on a project. Henry Wells collaborated with Charlotte Pierce as a co-author. Noah Sullivan teamed up with Maya Brooks as a co-author. Henry Wells collaborated with Grace Foster as a co-author. Sophia Nguyen collaborated as a co-author alongside Emma Chen. Grace Foster collaborated with Charlotte Pierce as a co-author. Ava Thompson teamed up with Ethan Rivera as a co-author.

Figure 16: An example problem consisting of contexts whose underlying structure is a community structure.