RECOGLAB: A FRAMEWORK TESTING RELATIONAL REASONING & COGNITIVE HYPOTHESES ON LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

A fundamental part of human cognition is the ability to not only recall memories, but to reason and manipulate information from them. In cognitive science and psychology, this is termed relational reasoning or relational memory and a number of cognitive effects and biases have been observed and proposed. Some of these effects include *congruence*, *the symbolic distance effect* and *transitive inference*. In addition, many of these phenomena have been observed in large language models on various handcrafted reasoning benchmarks. While some of these have been studied individually in prior benchmarks for reasoning with, none of these have the flexibility to study all or even most of these hypotheses. In this work, we create a fully customizable, automatically generated dataset which allows us to study these effects in detail. We introduce four settings with multiple cognitive-reasoning-inspired tasks targeting different skills and difficulties with parameters of each of these being configurable to run probes on different abilities. With our framework, we test and find many of these human cognitive effects are repeated in LLMs and provide a number of interesting analyses. We believe our generative framework will help accelerate the testing of various cognitive hypotheses and provide an interesting alternative paradigm for measuring reasoning capabilities.

1 INTRODUCTION

While recent work on memory in large language models (LLMs) assumes that memory refers to the recall of a specific piece of information from the past Li et al. (2024a); Levy et al. (2024), research into human memory and reasoning has long shown memory to be much more complex. When humans remember events, facts, places, etc., they don't just recall disconnected pieces, they recall the associations Cohen (1993); Eichenbaum (2004). Thus humans have the remarkable ability to recall relationships ("relational memory") and draw inferences across related memories ("relational reasoning").

Understanding and quantifying the relationship between human memory and reasoning has a long history of inquiry in the cognitive sciences. Experiments in these fields query memory for relationships between different entities (e.g. "How is A related to B?") and inferences about these relationships ("Given A is related to B and B to C, how are A and C related?"). These studies allow researchers to measure not just what is remembered, but how remembered information can inform a reasoning process. Patterns of errors and timing in these experiments provide glimpses into the underlying neural mechanisms that support the interaction of memory and reasoning. From this literature, a number of interesting effects have been observed, including *transitive inference*, and effects of presentation order, symbolic distance, and familiarity (Domjan, 2010; Moyer & Bayer, 1976; Koster et al., 2018) which have been studied individually in LLMs.

Prior work in evaluating LLMs on these kinds of relational reasoning problems have generally fallen into one of two camps. Either the benchmark is specifically designed to study recall of one or a few prior facts such as (Bai et al., 2024), or a specific cognitive science hypotheses has been tested on an evaluation created for that purpose such as in (Domjan, 2010). In this work, we aim to provide an evaluation framework that allows for the systematic evaluation of LLMs on relational memory and investigation of different possible effects (many inspired from the cognitive science literature).

Toward that end, we introduce ReCogLab, a framework for generating synthetic evaluation data that permits careful probing of the various aspects of relational reasoning over long contexts. Examples from this dataset can be seen in Fig. 1. The key aspect of this dataset is that virtually every aspect of the data can be controlled via a configuration file specifying parameters. For instance, in our Social Networks dataset, we can specify the number of entities, in Comparison we can control the ordering

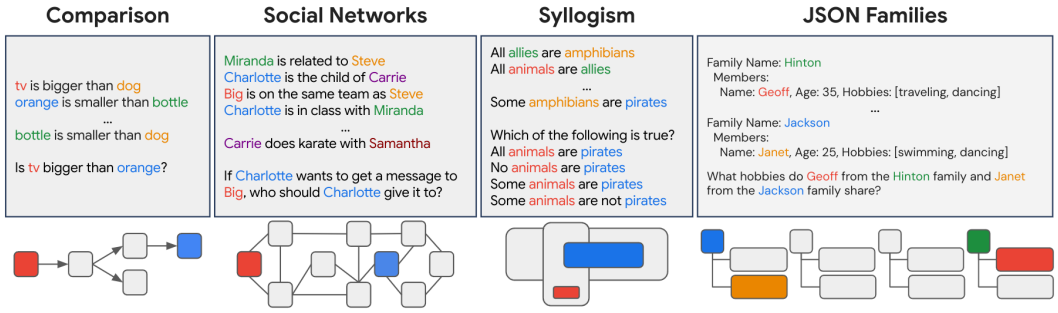


Figure 1: **ReCogLab Examples.** We show examples and the reasoning structure for Comparison (Sec.3.1) Social Networks (Sec.3.2), Syllogisms task (Sec.3.4) and JSON Families (Sec.3.3). These tasks evaluate a variety of important reasoning capabilities such as transitive inference, working memory capacity, and set reasoning. ReCogLab is capable of exploring. See Appendix C for more full examples from each of these tasks.

of statements, and in syllogisms we can control if the underling logical premises (e.g. all dogs are animals) are congruent, which is to say, true to the world. This level of control allows for the creation of datasets with varying length and complexity, facilitating the investigation of cognitive phenomena in long context language models. We present four tasks within ReCogLab to probe distinct aspects of relational reasoning.

With ReCogLab in hand, we perform a number of probes, some drawn from the cognitive science literature and some new, that allow us to probe many aspects of relational reasoning in large language models. In particular, we show reasoning performance often depends systematically on features that have been observed or hypothesized to affect reasoning performance in humans, like presentation order (Domjan, 2010), congruency with prior experience (Koster et al., 2018), and the symbolic distance between related entities (Moyer & Bayer, 1976). Our findings reveal intriguing parallels between human cognitive patterns and LLM behavior, suggesting common statistical or mechanistic factors constrain relational reasoning in both humans and machines.

In this paper we: (1) introduce ReCogLab, a new flexible dataset framework for cognitive science-inspired probes of LLMs on relational memory (2) perform numerous experiments to illustrate the flexibility and usefulness of our framework for this purpose (3) study several observed and hypothesized relational reasoning phenomena reported from the cognitive science literature enabled by the procedural generation of structured, text-based prompts (4) benchmark several LLM families on relational reasoning (5) provide analysis and new insights into LLM reasoning, including places where they fall short. We will release the ReCogLab framework and datasets used in this manuscript upon publication as a useful tool for evaluating examining LLMs for relational reasoning.

2 Background

Relational Memory and Reasoning in Cognitive Science. Relational memory and reasoning have been studied by psychologists across a variety of domains (Behrens et al., 2018; Eichenbaum, 2004; Nelli et al., 2023; Kumaran & Maguire, 2005; Tolman, 1948; Stupple & Ball, 2008; Evans et al., 1983). A good example of how these work together is "transitive inference" (Burt et al., 1911; Piaget, 1957): subjects are given data with the form "A>B, B>C, C>D, D>E", then queried about relationships that they did not observe directly but can be inferred (e.g. B>D). Many experiments vary different parameters of transitive inference to construct meaningful insights into cognitive biases. In this study, we explore different ways relational memory and reasoning interact in a language model. We investigate the effect of features like presentation order (Domjan, 2010)(Sec. 4.1) and symbolic distance (Moyer & Bayer, 1976) (Sec. 4.3) on 8 different language models. We also explore how these models' relational reasoning abilities are impacted by sequential processing and feature learning (Koster et al., 2018) through our congruency experiments in Sec. 4.2.

Other experiments have studied humans ability to reason about and sometimes navigate more complex relational structures, such as trees, grids, or community graph structures (Mark et al., 2020; Schapiro et al., 2013; Garvert et al., 2017; Lynn & Bassett, 2020). These experiments inspired our more richly structured datasets in Sec. 3. Relevant work in psychology has also studied syllogistic reasoning (Chater & Oaksford, 1999), and connections have been hypothesized that link syllogistic reasoning with reasoning over transitive chains (Guyote & Sternberg, 1981). Indeed, a notable

failure mode is that humans appear to simplify syllogisms to create easier-to-reason-about transitive chains (Ceraso & Provitera, 1971), which informs our investigation in Sec. 4.3.

Evaluation of LLMs Early work on reasoning benchmarks for language models[1] includes bAbI (Weston et al., 2016), a QA-style synthetic benchmark requiring very simple language understanding and reasoning questions, and Knowledge Base Question Answer (KBQA) Berant et al. (2013); Lopez et al. (2013), a system for evaluating multi-hop reasoning on knowledge graphs. Like our work, KBQA is heavily inspired by using graphs to assess relational reasoning capabilities.

More recently, datasets such as BigTOM (Gandhi et al., 2023), ToMi (Le et al., 2019) and RuleTakers (Clark et al., 2020) have been developed requiring much more complex reasoning. Other relevant work has investigated logic programming for logical deduction (Li et al., 2024c), multihop QA (e.g. (Li et al., 2024b)), and knowledge graph reasoning (e.g. (Luo et al., 2023)).

Other related work includes long-context LLM evaluations (Levy et al., 2024; Bai et al., 2024; Vodrahalli et al., 2024). Li et al. (2024a) in particular includes some of the aspects of relational reasoning but without our dataset's customizability. All of these works focus specifically on this question of thoroughly examining the context length of language models. Our dataset is unique in affording a joint inspection of context length in the context of relational reasoning, and affords scaling complexity along other directions with customizabilty along multiple dimensions.

Comparison of LLMs to humans Substantial work has compared reasoning in LLMs to humans (Hagendorff et al., 2024; Binz & Schulz, 2023), employing cognitive science and psychological methodologies to explore the biases in LLMs Jones & Steinhardt (2022); Seals & Shalin (2023); Webb et al. (2023); Berglund et al. (2023). Some of the cognitive effects that we study here have been studied individually in some of these works such as: consistency of logical reasoning with prior knowledge Lampinen et al. (2024) (Sec. 4.2), syllogistic reasoning Eisape et al. (2024) (see Sec.4.1, 4.2), premise order in syllogisms Chen et al. (2024) (Sec. 4.1), and susceptibility to irrelevant information (an aspect of capacity experiments in Sec. 4.1). We similarly adopt experiments from the cognitive science literature as a reference point for understanding LLM behaviors, but in the context of relational memory and reasoning.

Sandbox Evaluation Toolkits We have proposed a framework for automatically generating new templated datasets for relational reasoning problems. Other works have looked to increase the speed of creating evaluations, for instance Dynabench Kiela et al. (2021) and later Dynatask Thrush et al. (2022) which allow for rapid integration of models into dataset creation. However, this still requires human annotation and thus would not be able to automatically change parameters of the dataset instantaneously as our framework does. Cognitive scientists have long been interested in evaluating LLMs as we do with humans Hernández-Orallo (2017) and environments such as animal-ai environment Voudouris et al. (2022) have been created. However, this environment is a 3D embodied environment, and while it also touches on memory, it does not touch on language or on relational reasoning specifically.

3 ReCogLab Framework

We now describe in detail our framework for automatically generating many evaluations on relational reasoning with customizable properties and complexity.

Examples generated by ReCogLab come from one of four tasks: **Comparison**, **Social Networks**, **Syllogisms**, **JSON Families**. We chose these tasks because they were able to be generated automatically and they include a spectrum of problems in relational reasoning. Each of these tasks can be characterized as reasoning about graph problems using language. See Figure 1 to see exemplars from each dataset and the corresponding logical graphs associated with each. See Appendix. C for additional full examples. Each example consists of context C which consists of multiple factual statements that the model assumes to be true (e.g. "Miranda is related to Steve"). Each of these facts denotes a relationship R_{ij} between pairs of entities (E_i, E_j) (each of which is a person or object or other noun in the relationship or can be an attribute). Next is the question Q, which relies on a subset of the context C to answer, requiring integrating information across distinct facts to answer. Finally, each example has a corresponding, non-ambigious answer or answers A.

[1] As well as other AI systems, see e.g. Cropper & Dumančić (2022)

Task	Sub-tasks	Configurable Parameters
Comparison	**Larger-Smaller**, **Older-Younger**, **Heavier-Lighter**, **Consistency Detection**, **Indeterminate Conclusion**	**network_type**, **num_entities**, **entity_type**, **ordering**, **congruence**, **randomize_relations**, **do_reverse_comps**
Social Networks	**Fastest Message**, Oldest/Youngest Generation	**network_type**, **num_entities**, entity_type, relation_type, **randomize_relations**
JSON Families	Family Size, Member Hobby, **Hobby Comparison**, Age Comparison, Size Comparison	**num_families**, max_members, hop_length
Syllogisms	**Set Membership**	**entity_type**, **num_entities**, **ordering**

Table 1: **ReCogLab Sub-tasks and configurable parameters.** Please see Appendix A for more details on how these sub-tasks and parameters are defined and implemented. Bold text indicates sub-tasks and parameters used to support cognitive findings in our experiments.

Next, we emphasize that ReCogLab is not just a dataset, but a framework for automatically generating and configuring infinite datasets and dataset examples. To generate an example, you pass into ReCogLab a random seed (to allow for reproducible generation), a split (train/val/test, which determines which subset of all entities $\mathcal{E}$ is used), and a configuration file. The configuration file specifies the task (and sub-task) along with the value to use for configurable parameters. See Tab. 1 for a full list. Some important parameters here are "num_entities", which determines the number of relationships/facts in the context, used in Sec. 4.1, "ordering" which determines if facts are presented in order, also used in Sec. 4.1 and "congruence," used in Sec. 4.2. Because we control the data generating process by specifying parameters, we can produce datasets that evaluate specific reasoning behaviors or hypotheses or vary levels of difficulty. Please see Sec. A.1 for more details about how each task generates test examples mechanically.

We now describe each of our four main task types in greater detail.

3.1 Comparison

A comparison is a directed edge in an acyclic graph where vertices are *entities* and edges describe a directional comparison between them. Inspired by transitive inference going back to (Burt et al., 1911), we construct comparison problems using three attributes that exhibit transitivity: comparison of age, size, or weight of objects.

For comparison reasoning problems, we target specific parameters to understand how language models behave when presented with similar problems in different forms such as: the comparison type, the number of objects, the ordering of comparisons, and the attribute types of directional reasoning.

3.2 Social Networks

Inspired by Kumaran & Maguire (2005), the next set of tasks we propose require reasoning about Social Networks. Like comparisons, a social network can be algorithmically represented as a graph problem where nodes represent people and edges represent a particular relationship like friendship. Unlike with Comparisons, because friendship is (typically) a symmetric relationship, social networks are undirected graphs.

A social network is good for probing long-context relational reasoning for several reasons. First, social networks can be arbitrarily complex. This makes it possible to generate very long chains with complex structure. Secondly we introduce "flavor text" (see Appendix A) generation in every example's *context*. Finally one could consider combining other kinds of social network edges like teacher-student (directed edge) or even enemies (undirected edge that has a different semantic interpretation) to build even more complex reasoning probes.

3.3 JSON Families

The family dataset is a structured, JSON-formatted family data. Every family consists of a number of members, and the context consists of a number of families. This dataset is generated by defining a set of first names, last names, hobbies etc. Questions are asked per family and require comparing two families. The sub-tasks in this dataset are: Family size (asking the size of one family),

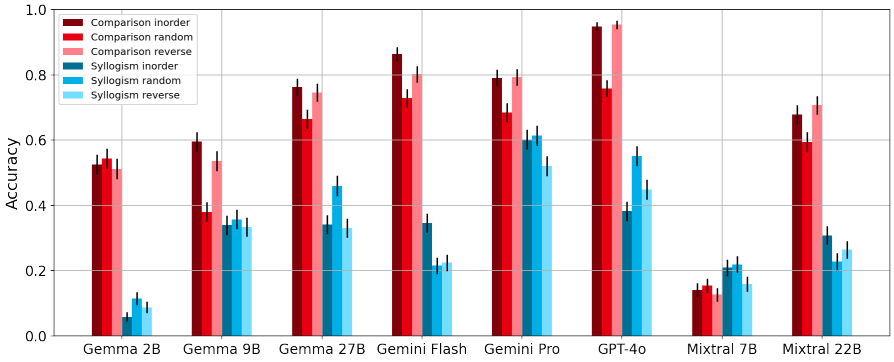


Figure 2: **Effects of contextual premise order.** We show the effect of transitive inference on answering reasoning problems correctly. For Syllogism and Comparison, we construct three different orders of the contextual premise. We find that processing information in some logically relevant (inorder or reverse) order produces improvement. Surprisingly this effect is not observed for set reasoning in Syllogism. The different performance across different language models on Comparison and Syllogism suggest that reasoning capabilities on different tasks may not be correlated.

family member hobby (checking if a family member has a specific hobby), family size comparison (comparing family sizes of two families), family member age comparison (asking the relative age of two members of two families), and family member hobby comparison (comparing hobbies of two different families). See Appendix A for more details.

3.4 Syllogisms

Previous work on syllogistic reasoning has shown interesting parallels between humans and language models (Eisape et al., 2024). We extend prior work that has studied ordering and congruence effects on humans and language models to include a capacity dimension. We construct chains of propositions that transitively resolve to a single final conclusion and prompt the language model to choose between all possible conclusions with the given subject and predicate.

3.5 Dataset Creation

For each of these tasks, we create datasets with fixed seeds to guarantee consistency and reproducibility. We generate sweeps of different configurations targeting evaluations of a specific cognitive probe. We generate 50 validation and report on 1000 test examples. Here the important thing is not the specific examples of the dataset, but the ability to tailor skill and difficulty levels to measure cognitive abilities in different LLMs. We intend to release the code for generating our datasets used in this manuscript as well as the custom configuration to empower others to investigate cognitive effects in language models.

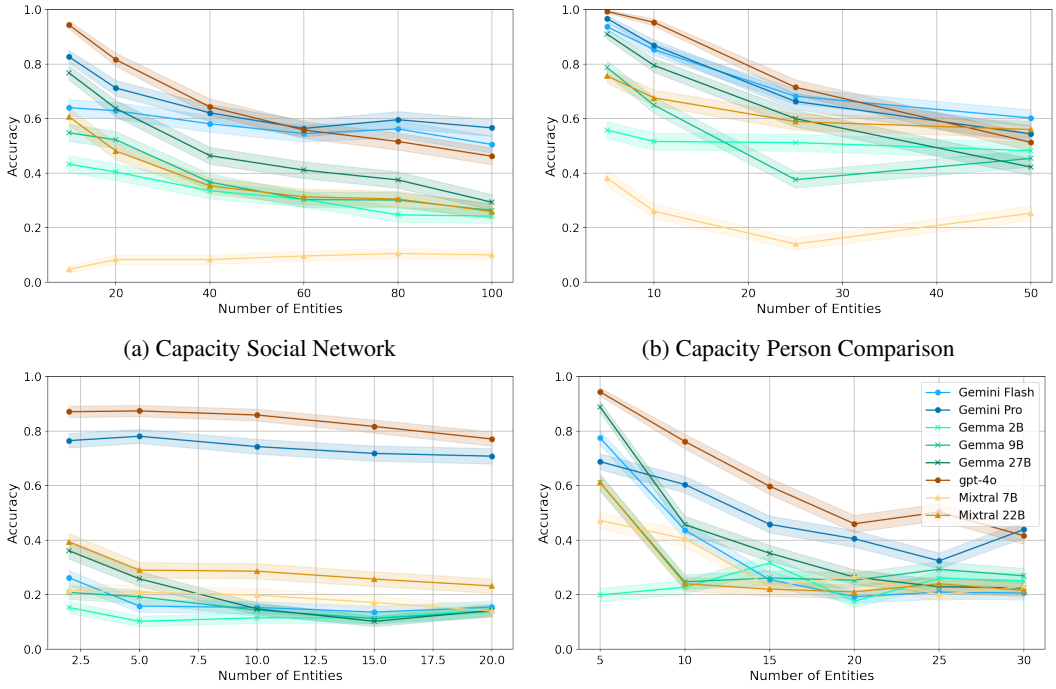
3.6 Evaluation Protocol

A core challenge in evaluating language models is the significant impact of prompting on their performance. To ensure a fair comparison when measuring different language models, we propose decoupling the prompt and parsing hyperparameters from the language model. We curated a library of prompts and parsing strategies targeting ReCogLab. These libraries cover a wide range of techniques from simply asking the question to role prompting. We use a small validation set to select the best performing combination of a prompt and string parser. This allows us to probe a wide variety of language models that is agnostic to a model's preference for a particular prompting strategy. Please see Sec. B for a list of all prompts and prompt strategies we validated.

4 Experiments

Now that we've described the ReCogLab framework, we can demonstrate the flexibility of generation it provides when testing language model capabilities. In particular, we will score different LLMs on their ability to perform relational reasoning tasks.

In this paper we look at transitive inference (Sec. 4.1), congruency (Sec. 4.2), the symbolic distance effect (Sec. 4.3), identifying logical inconsistencies (Sec. 4.4) and indeterminate reasoning (Sec. 4.5). These experiments were all made with little additional effort using our flexible ReCogLab framework by altering the configuration files. From our experiments, we observe that LLMs share many of the reasoning patterns, biases, and limitations with human relational reasoning.



(a) Capacity Social Network

(b) Capacity Person Comparison

(c) JSON Family Hobby Comparison

(d) Syllogisms

Figure 3: **Problem complexity vs. Accuracy**. We show the effect of adding complexity and context in the form of increasing number of entities for each our tasks. We find in all cases that performance drops, although this drop is less severe in more capable models such as GPT-4o and Gemini Pro

Language Models We evaluate several open-source and closed-sourced models on our suite of cognitive benchmarks. In our results, we show Google's open-source Gemma-2B, 9B, 27B family in green. The open-source Mixtral models are shown in yellow/orange. Gemini Flash and Gemini Pro (closed-source) are shown in blue. OpenAI's GPT-4o (also closed source) is shown in red. All of these language models are the instruction-tuned variants. Each specific probe involves validating on 50 examples before selecting the best prompt and parserfor test-time evaluation on 1000 examples. We also plot a 95% confidence interval to give an idea of the confidence in the difference between models and in the trendlines.

4.1 Transitive Inference

Transitive inference ((Piaget, 1957)) is a quintessential example of relational reasoning (e.g. "if $A > B$ and $B > C$, then $A > C$"). This ability is intact even if this information is presented such that the temporal order is different from the symbolic ordering (e.g. if "$B > C$ and $A > B$"); however, a performance improvement has been reported for comparison judgments if the symbolic and sequential order match (Domjan, 2010). Thus, we wanted to evaluate our models' ability to reason accurately about associations presented in context, and measure whether there was a dependence on presentation order in their context across different domains.

In Fig. 2 we see this experiment for Comparison and Syllogisms. For Comparison, we can clearly see across nearly all models randomizing the order causes noticeable degradation (Gemma 2B, 9B and Mixtral 7B perform near or below chance of 0.5, guessing or giving invalid answers). Reversing the presentation order also often negatively affects performance in many, but not all cases.

Interestingly in Syllogisms (chance performance is 0.25), we see that not all models do better when entities are in order. In these cases, the model will often skip a full step-by-step of the problem and say something along the lines of "We know there's a chain from entity A to entity Z so 'All As are Zs' is correct" even when there are statements within the chain like 'Some X are Y'. These error patterns are consistent with the observation that errors in human syllogistic reasoning seem to derive from simplifying relationships to enable a more computationally lightweight transitive chain (Ceraso & Provitera, 1971; Chater & Oaksford, 1999; Guyote & Sternberg, 1981). We speculate

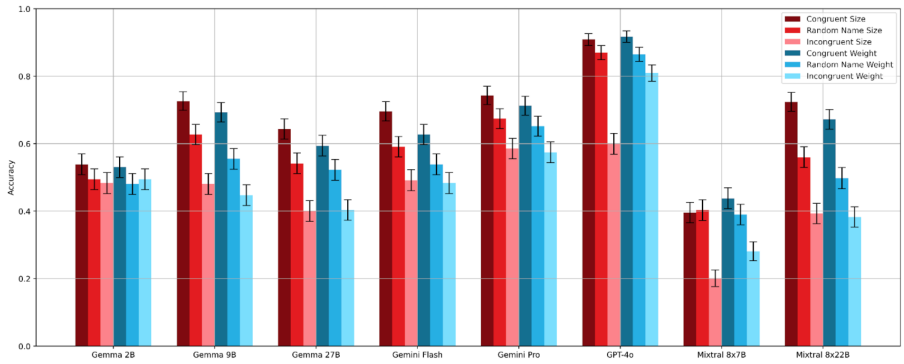


Figure 4: **Congruent and Incongruent reasoning.** We find that transitive inference is strongly impacted by real-world priors. We construct Comparison problems with objects that are *Congruent* and *Incongruent* with real world objects. We also generate a version which *Random Name* replaces all comparison objects with randomly generated strings that contain no real world priors. Across the different language models, we observe degradation in reasoning capabilities when presented with logically consistent, but incongruent statements about the real world.

that consistent presentation order may bias the model toward these approximate solutions, although more investigation is needed.

In Fig. 3, we further measure how this ability depends on complexity measures of the reasoning problem: e.g. total length of context, the number of relations. These experiments test the capacity of our LLMs on these datasets. Because these datasets are easily configurable, for each of our four datasets, we scale difficulty and context length of these dataset in different ways to see how each of our models responds. We find, perhaps unsurprisingly, that performance generally degrades as the complexity of the problem increases across all models. The more capable closed-source models (Gemini Flash, Gemini Pro, GPT-4o) perform the best. Interestingly, on JSON Family, the performance of Gemini Pro seem to levels out rather than decreasing and GPT-4o only decreasing slightly, suggesting that they are particularly good at needle-in-the haystack style questions, but struggle along with the rest of the models on other datasets requiring more than two pieces of context. These trends seem consistent even within confidence intervals, but more study is needed to fully validate this effect and we hope that our framework enables further research into this area. We might for instance want to separate out more clearly the effect of long contexts from problem complexity by experimenting with "filler" text scaling (adding irrelevant text) versus making the graphs more complex to further analyze these trends.

4.2 Congruent and Incongruent

Despite only being trained on textual data, language models exhibit remarkable knowledge about real world physical attributes of objects. This can be easily demonstrated when asking questions like: *Are fire trucks larger than teddy bears?*. This prior knowledge is aligned with real facts about the world and is essential for many applications of language models.

Motivated by the findings of Lampinen et al. (2024), who report that logical reasoning is more accurate when logical premises are congruent with real-world facts for LLMs (as it is for humans), we sought to explore this in the relational reasoning regime. Expanding on their paradigm, we categorize premises based on their grounding in real-world knowledge. *Congruent* relationships are those that are consistent with the real world (e.g "whales are larger than minnows"). Conversely, *incongruent* relationships contradict our prior knowledge about the world ("minnows are larger than whales"). Finally, we consider a third category of relationships involving contrived, nonsensical relationships ("glarbs are larger than bojaks") that have no meaningful semantics behind the terms. Intuitively, congruent relationships are more likely to be consistent with the prior training of the language model and may bias the answering of these questions.

First, we investigate the impact of congruence on language model performance, replicating the observation that congruence contributes positively to performance.

For the Comparison task, to extract real-world knowledge about objects, we curated a list of 540 objects and asked a language model to estimate their size in kilograms and meters. From this, we construct comparison problems that are **Congruent** and **Incongruent**. We also generated a

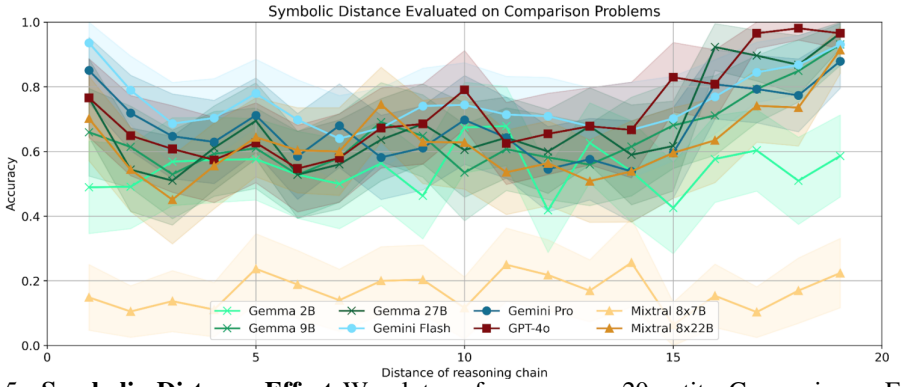


Figure 5: **Symbolic Distance Effect** We plot performance on 20-entity Comparison. For each example, we compute the "symbolic distance" of the answer which is the minimum number of statements one would need to traverse to verify an answer. For most models (the weaker Mixtral 7B and Gemma 2B) we see a curve in which accuracy decreases with symbolic distance between 1 and 5, but then improves for symbolic distances above about 15.

Random-String version of the comparison problem which replaces all object names with a random string of alphanumeric characters of length 5 for the contrived, nonsensical baseline.

We find that using **Congruent** statements outperforms similar comparison problems constructed with **Incongruent** statements. This behavior is consistent across language models and relational reasoning tasks. Some models even exhibit worse-than-chance (0.5) correctness on **Incongruent** comparison problems. **Random-String** performance lands between the two extremes, suggesting a positive relationship between reasoning competency and factual coherence. To further investigate these early findings, future researchers can use our framework to add more types of tasks to see if this result holds for multiple types of problems and vary other parameters (such as context length) to see if requiring more context degrades model's ability to reason counter to its congruence bias.

4.3 Symbolic Distance Effect

Symbolic Distance refers to the number of relational reasoning steps that join two entities in a sequential reasoning task (e.g., if A $>$ B $>$ C $>$ D, the symbolic distance between A and D is 3). The symbolic distance effect refers to the observation that more distant comparison judgments are, perhaps surprisingly, easier for humans (except at distance 1, where observed pairs can be memorized) (Moyer & Bayer, 1976). This effect is thought to rely on feature learning (Lippl et al., 2024), which is hypothesized to require repeated exposures (Koster et al., 2018). Thus, we hypothesize that LLMs will also show a symbolic distance effect.

In Figure 5, we report the effect of symbolic distance on performance for our models. We generate Comparison examples of 20 entities and then sample comparison questions based on the symbolic distance. For models that perform above chance (all except Gemma 2B and Mixtral 7B), there is a clear U shape. This has some commonalities with the behavior of humans: humans show higher performance for the shortest distance, followed by a sudden drop in performance, which gradually rises as symbolic distance increases. LLMs similarly show elevated performance for the shortest and longest symbolic distances, however both the drop for short distances and the rise for longer distances are more gradual. Thus we see real similarities between the human and observed LLMs distance effects for powerful models, but some differences as well.

4.4 Identifying Inconsistent Premises

One important capability of a reasoning system is identifying when a premise is logically inconsistent (Black et al., 1986; Johnson-Laird et al., 2004). A logically inconsistent example is a set of statements which cannot all be true. This reasoning task is different from the previous structured reasoning tasks as identifying the set of inconsistent facts requires global reasoning capabilities. Furthermore this question does not incorporate any specific entity name so there's no prior on which statements from the context are relevant; and there are no rules on organizing or processing statements that trivialize this task. Therefore all statements must be equally considered before a conclusion of logically consistent can be drawn.

We consider this as applied on comparison problems. Here, a logically consistent comparison graph assumes no cycles exist. If A $>$ B, B $>$ C, then you cannot have C $>$ A. This implies that all

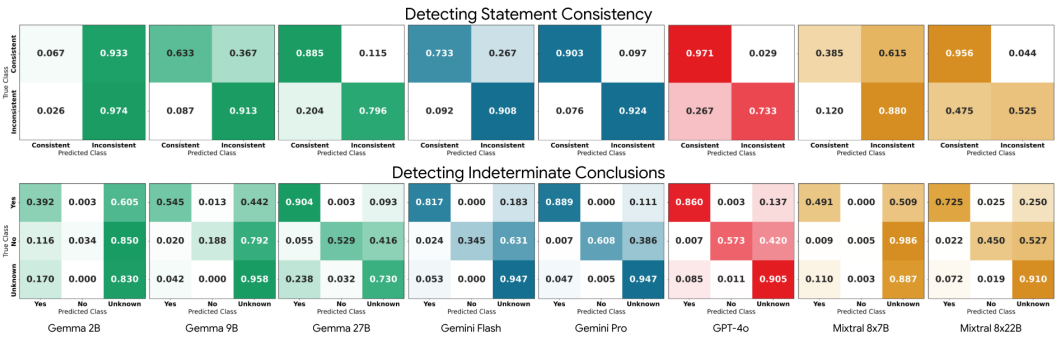


Figure 6: **Confusion matrix for meta-reasoning**. ReCogLab allows us to probe language models for important meta-reasoning capabilities. The first row shows confusion matrices on classifying *consistent* and *inconsistent* premises. We find later generations of models iterations can better identify statements that are inconsistent. We also find a similar trend when inferring whether a conclusion is *indeterminate*. These are important meta-reasoning capabilities with critical implications for designing human interaction and autonomous decision-making systems.

inconsistent graphs are comparison graphs that contain a cycle (i.e. the graph is no longer a DAG). To create inconsistent examples, we start with a valid comparison DAG and randomly sample nodes to add an edge that would induce a cycle.

We show the confusion matrix on predictions extracted from each LLM on the first row of Fig. 6. The diagonal elements of the confusion matrix indicate that larger language models perform better on detecting inconsistent statements. An additional interesting statistic to analyze is the type of classification errors each model makes. This is represented by the off-diagonal elements and is related a classifiers' false positive rates (FPR).

Because our probes are zero-shot tasks, the models' predictions are inherently uncalibrated. This presents an opportunity to investigate each models' implicit reasoning bias. For instance, the GPT-4o predictions have a higher FPR for Inconsistent premises (16.3%, 17.5% vs 6% and 4.8%), implying that GPT-4o have a preference for answering consistent absent any priors. Gemini Flash predictions exhibits the opposite bias, with higher rates of False Positive for Consistent premises. Gemini Pro performs the best at zero-shot statement consistency detection and has relatively balanced false positive and false negative rates.

4.5 Indeterminate Reasoning

Another important aspect of a reasoning is understanding when there's insufficient information to draw a conclusion. This relates to an ability often ascribed somewhat uniquely to humans: our ability to characterize uncertainty (Courville et al., 2006). We refer to this capability as detecting indeterminacy. This is important for several reasons. First, a logical system that fails to identify an indeterminate premise may draw erroneous conclusions that are not supported by the evidence. Second, recognizing when a conclusion cannot be drawn is important for improving decision making and handling ambiguity. Therefore we create a probe targeting evaluation of each language model's robustness to this failure mode.

We start with comparison problems which contain a fixed label set of Yes or No. We modify the comparison problems from a linear chain to a random tree generation while still asking questions about two random nodes. Under a DAG tree, any two nodes may no longer have a path to each other. This corresponds to an indeterminate conclusion – insufficient *context* was provided to reason about the relationship between two objects. This results in the modified answer set consisting of three potential answers (Yes, No, Unknown). For this particular evaluation, we provide specific instructions to answer whether the *context* and question are indeterminate before asking a question. While this experiment follows by combining cognitive science experiments about uncertainty estimation and transitive inference, it has not to our knowledge been performed in the cognitive science literature.

We show results of this probe in the second row of confusion matrices in Fig. 6. Similar to the previoAs probe, we examine the FPR to understand each LLMs' implicit reasoning bias. Again, the diagonals are correctly classified examples, and the off-diagonal elements are different FPR errors.

For any determinate example we construct, determining if $A > B$ is as difficult as determining if $B > A$. However we see that there's a significant performance gap across all language models in correctly identifying determinate examples. In other words, presenting the same problem, but flipping the comparison such that the answer is *No*, will cause the language model to prefer predicting indeterminacy. One interpretation is that language models prefer to answer *Yes*. This is further supported when examining the language models bias towards answering *Yes* on indeterminate examples that are incorrectly reasoned to be determinate. Of note, a similar "acquiescence bias" toward answering *Yes* has been noted in humans in a variety of settings (Krosnick, 1999).

The immediate findings suggest that a self-consistency scheme could improve logical deductions and reduce the implicit reasoning bias towards affirmative answers. Longer term, pursuing strategies beyond instruction tuning to improve detection of indeterminate logical premises is important for safety and human interactions.

5 Discussion

We introduce ReCogLab, a flexibly generated dataset for quantitatively measuring relational memory and reasoning performance in LLMs. We demonstrate the utility of this dataset, conducting a number of experiments to benchmark relational reasoning performance across different models and problem complexities. Moreover, we recreate cognitive science experiments that characterize how reasoning performance depends on features such as presentation order Domjan (2010), symbolic distance Moyer & Bayer (1976), congruency with prior experience Koster et al. (2018), and availability of logical heuristics Ceraso & Provitera (1971). We also create a number of novel experiments that measure how models recognize inconsistencies and indeterminacies. Ultimately, we hope that by accurately measuring the cognitive capabilities of our language models, we can contribute a metric for hill-climbing improvements in the field that are inspired by human cognition. Furthermore, we hope that this work will inspire deeper, rigorous probes into reasoning capabilities of both humans and artificial agents, and help provide insight into the success and failure modes of reasoning. [2]

In the future, we plan to continue enriching ReCogLab with increasingly sophisticated probes of relational memory and reasoning. One possible future direction is the addition of datasets with similar underlying structure to study how invariant these effects are to the particular language or form of the problem (e.g. different ways of expressing graph relationships for studying transitive inference). Another set of directions is to further ablate and test different configurations within the existing dataset or varying two parameters of interest together to study their interaction.

Many of the effects we observe in LLMs mirror effects observed in humans, suggesting commonalities in the factors contributing to relational reasoning in LLMs and humans. While it is instructive to compare and contrast LLM reasoning to that of humans, we should as always be careful about drawing too many conclusions about the similarities between LLMs and human psychology (Shevlin & Halina, 2019; Shanahan, 2022). Even when there are similarities, there is ambiguity about the explanation: it may derive from similarities in the statistics of experienced data, task objectives, architectures, or learning rules. Alternatively, it could be that two entirely different mechanisms are responsible for the same effect.

This work not only provides a dataset for evaluation of relational memory and reasoning in LLMs, it gives us a flexible framework for surgically probing specific effects. We believe this will provide a contribution to the cognitive science literature as well. While many of our experiments copied existing cognitive science experiments, our dataset also enabled us to easily generate novel experiments that have a cognitive science connection but that have not yet been tested in humans (to our knowledge). Our results provide a number of novel hypotheses for psychology: in particular, that syllogistic reasoning performance is negatively impacted by presentation order matching the relational ordering (if "Some A's are B's" relations are present) (Sec 4.1), that the symbolic distance effect is modulated by congruency (Sec. 4.3), that there is a bias toward incorrectly answering "yes" on logical prompts when it is uncertain and reporting uncertainty when in fact it is unknown (Sec. 4.5). More generally, this dataset generator permits the development of procedurally generating new text-based relational reasoning tasks, and rapidly piloting them on LLMs, which we hope will provide a useful tool to the field.

[2] We will release code and datasets upon publication

REFERENCES

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3119–3137, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.172.

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, October 2018.

Jonathan Berant, Andrew K. Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase from question-answer pairs. In *Conference on Empirical Methods in Natural Language Processing*, 2013. URL https://api.semanticscholar.org/CorpusID:6401679.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: LLMs trained on" a is b" fail to learn" b is a". In *International Conference on Learning Representations*, 2023.

Marcel Binz and Eric Schulz. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023.

Alison Black, Paul Freeman, and Philip N Johnson-Laird. Plausibility and the comprehension of text., 1986.

Cyril Burt et al. Experimental tests of higher mental processes and their relation to general intelligence. 1911.

John Ceraso and Angela Provitera. Sources of error in syllogistic reasoning. *Cognitive Psychology*, 2(4):400–410, 1971. ISSN 0010-0285. doi: https://doi.org/10.1016/0010-0285(71)90023-5. URL https://www.sciencedirect.com/science/article/pii/0010028571900235.

Nick Chater and Mike Oaksford. The probability heuristics model of syllogistic reasoning. *Cognitive Psychology*, 38(2):191–258, 1999. ISSN 0010-0285. doi: https://doi.org/10.1006/cogp.1998.0696. URL https://www.sciencedirect.com/science/article/pii/S001002859890696X.

Xinyun Chen, Ryan A Chi, Xuezhi Wang, and Denny Zhou. Premise order matters in reasoning with large language models. In *International Conference on Machine Learning*, 2024.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. Transformers as soft reasoners over language. In *International Joint Conference on Artificial Intelligence*, 2020. URL https://api.semanticscholar.org/CorpusID:211126663.

NJ Cohen. *Memory, amnesia and the hippocampal system*. MIT Press, 1993.

Aaron C Courville, Nathaniel D Daw, and David S Touretzky. Bayesian theories of conditioning in a changing world. *Trends in cognitive sciences*, 10(7):294–300, 2006.

Andrew Cropper and Sebastijan Dumančić. Inductive logic programming at 30: a new introduction. *Journal of Artificial Intelligence Research*, 74:765–850, 2022.

Michael Domjan. *The Principles of Learning and Behavior*. The Principles of Learning and Behavior. Belmont: Cengage learning, 6 edition, 2010.

Howard Eichenbaum. Hippocampus: cognitive processes and neural representations that underlie declarative memory. *Neuron*, 44(1):109–120, 2004.

Tiwalayo Eisape, Mh Tessler, Ishita Dasgupta, Fei Sha, Sjoerd van Steenkiste, and Tal Linzen. A systematic comparison of syllogistic reasoning in humans and language models. In *North American Chapter of the Association for Computational Linguistics*, volume abs/2311.00445, 2024. URL https://api.semanticscholar.org/CorpusID:264832674.

JSBT Evans, Julie L Barston, and Paul Pollard. On the conflict between logic and belief in syllogistic reasoning. *Memory & cognition*, 11(3):295–306, 1983.

Kanishk Gandhi, Jan-Philipp Franken, Tobias Gerstenberg, and Noah D. Goodman. Understanding social reasoning in language models with language models. In *Advances in Neural Information Processing Systems*, volume abs/2306.15448, 2023. URL https://api.semanticscholar.org/CorpusID:259262573.

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, apr 2017. ISSN 2050-084X. doi: 10.7554/eLife.17086. URL https://doi.org/10.7554/eLife.17086.

Martin J. Guyote and Robert J. Sternberg. A transitive-chain theory of syllogistic reasoning. *Cognitive Psychology*, 13(4):461–525, 1981. ISSN 0010-0285. doi: https://doi.org/10.1016/0010-0285(81)90018-9. URL https://www.sciencedirect.com/science/article/pii/0010028581900189.

Thilo Hagendorff, Ishita Dasgupta, Marcel Binz, Stephanie C. Y. Chan, Andrew Lampinen, Jane X. Wang, Zeynep Akata, and Eric Schulz. Machine psychology, 2024. URL https://arxiv.org/abs/2303.13988.

José Hernández-Orallo. *The measure of all minds: evaluating natural and artificial intelligence*. Cambridge University Press, 2017.

P N Johnson-Laird, Paolo Legrenzi, and Vittorio Girotto. How we detect logical inconsistencies., 2004.

Erik Jones and Jacob Steinhardt. Capturing failures of large language models via human cognitive biases. *Advances in Neural Information Processing Systems*, 35:11785–11799, 2022.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. Dynabench: Rethinking benchmarking in nlp. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4110–4124, 2021.

Raphael Koster, Martin J. Chadwick, Yi Chen, David Berron, Andrea Banino, Emrah Düzel, Demis Hassabis, and Dharshan Kumaran. Big-loop recurrence within the hippocampal system supports integration of information across episodes. *Neuron*, 99(6):1342–1354.e6, Sep 2018. ISSN 0896-6273. doi: 10.1016/j.neuron.2018.08.009. URL https://doi.org/10.1016/j.neuron.2018.08.009.

Jon A. Krosnick. Survey research. *Annual Review of Psychology*, 50(Volume 50, 1999):537–567, 1999. ISSN 1545-2085. doi: https://doi.org/10.1146/annurev.psych.50.1.537. URL https://www.annualreviews.org/content/journals/10.1146/annurev.psych.50.1.537.

D Kumaran and E Maguire. Knowing your place: Navigation in spatial and social networks. In *Journal of Cognitive Neuroscience*, pp. 100–100. MIT PRESS, 2005.

Andrew K Lampinen, Ishita Dasgupta, Stephanie C Y Chan, Hannah R Sheahan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus*, 3(7):233, 07 2024. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgae233. URL https://doi.org/10.1093/pnasnexus/pgae233.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. Revisiting the evaluation of theory of mind through question answering. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5872–5877, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1598. URL https://aclanthology.org/D19-1598.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In *Annual Meeting of the Association for Computational Linguistics*, 2024. URL https://api.semanticscholar.org/CorpusID:267897954.

Mo Li, Songyang Zhang, Yunxin Liu, and Kai Chen. Needlebench: Can llms do retrieval and reasoning in 1 million context window? *ArXiv*, abs/2407.11963, 2024a. URL https://api.semanticscholar.org/CorpusID:271218188.

Yanyang Li, Shuo Liang, Michael Lyu, and Liwei Wang. Making long-context language models better multi-hop reasoners. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2462–2475, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.135. URL https://aclanthology.org/2024.acl-long.135.

Zhiming Li, Yushi Cao, Xiufeng Xu, Junzhe Jiang, Xu Liu, Yon Shin Teo, Shang-Wei Lin, and Yang Liu. Llms for relational reasoning: How far are we? In *Proceedings of the 1st International Workshop on Large Language Models for Code*, LLM4Code '24, pp. 119–126, New York, NY, USA, 2024c. Association for Computing Machinery. ISBN 9798400705793. doi: 10.1145/3643795.3648387. URL https://doi.org/10.1145/3643795.3648387.

Samuel Lippl, Kenneth Kay, Greg Jensen, Vincent P. Ferrera, and L. F. Abbott. A mathematical theory of relational generalization in transitive inference. *Proceedings of the National Academy of Sciences*, 121(28):e2314511121, 2024. doi: 10.1073/pnas.2314511121. URL https://www.pnas.org/doi/abs/10.1073/pnas.2314511121.

Vanessa Lopez, Christina Unger, Philipp Cimiano, and Enrico Motta. Evaluating question answering over linked data. *Journal of Web Semantics*, 21:3–13, 2013. ISSN 1570-8268. doi: https://doi.org/10.1016/j.websem.2013.05.006. URL https://www.sciencedirect.com/science/article/pii/S157082681300022X. Special Issue on Evaluation of Semantic Technologies.

Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. Reasoning on graphs: Faithful and interpretable large language model reasoning. In *International Conference on Learning Representations*, volume abs/2310.01061, 2023. URL https://api.semanticscholar.org/CorpusID:263605944.

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. doi: 10.1038/s41467-020-18254-6. URL https://doi.org/10.1038/s41467-020-18254-6.

Robert S. Moyer and Richard H. Bayer. Mental comparison and the symbolic distance effect. *Cognitive Psychology*, 8(2):228–246, 1976. doi: 10.1016/0010-0285(76)90025-6. URL https://doi.org/10.1016/0010-0285(76)90025-6.

Stephanie Nelli, Lukas Braun, Tsvetomira Dumbalska, Andrew Saxe, and Christopher Summerfield. Neural knowledge assembly in humans and neural networks. *Neuron*, 111(9):1504–1516.e9, May 2023.

Jean Piaget. *Logic and psychology.* Logic and psychology. Basic Books, Oxford, England, 1957.

Anna C Schapiro, Timothy T Rogers, Natalia I Cordova, Nicholas B Turk-Browne, and Matthew M Botvinick. Neural representations of events arise from temporal community structure. *Nature neuroscience*, 16(4):486–492, 2013.

S. M. Seals and Valerie L. Shalin. Long-form analogies generated by chatgpt lack human-like psycholinguistic properties. In *Annual Meeting of the Cognitive Science Society*, 2023.

Murray Shanahan. Talking about large language models. *arXiv*, pp. 1–11, 2022.

Henry Shevlin and Marta Halina. Apply rich psychological terms in ai with care. *Nature Machine Intelligence*, 1:165–167, 2019.

Edward JN Stupple and Linden J Ball. Belief–logic conflict resolution in syllogistic reasoning: Inspection-time evidence for a parallel-process model. *Thinking & Reasoning*, 14(2):168–181, 2008.

Tristan Thrush, Kushal Tirumala, Anmol Gupta, Max Bartolo, Pedro Rodriguez, Tariq Kane, William Gaviria Rojas, Peter Mattson, Adina Williams, and Douwe Kiela. Dynatask: A framework for creating dynamic ai benchmark tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 174–181, 2022.

Edward C. Tolman. Cognitive maps in rats and men. *Psychological Review*, 55(4):189–208, 1948. ISSN 1939-1471. doi: 10.1037/h0061626. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/h0061626.

Kiran Vodrahalli, Santiago Ontañón, Nilesh Tripuraneni, Kelvin Xu, Sanil Jain, Rakesh Shivanna, Jeffrey Hui, Nishanth Dikkala, Mehran Kazemi, Bahare Fatemi, Rohan Anil, Ethan Dyer, Siamak Shakeri, Roopali Vij, Harsh Mehta, Vinay Venkatesh Ramasesh, Quoc Le, Ed Chi, Yifeng Lu, Orhan Firat, Angeliki Lazaridou, Jean-Baptiste Lespiau, Nithya Attaluri, and Kate Olszewska. Michelangelo: Long context evaluations beyond haystacks via latent structure queries. 2024. URL https://api.semanticscholar.org/CorpusID:272753754.

Konstantinos Voudouris, Matthew Crosby, Benjamin Beyret, José Hernández-Orallo, Murray Shanahan, Marta Halina, and Lucy G Cheke. Direct human-ai comparison in the animal-ai environment. *Frontiers in Psychology*, 13:711821, 2022.

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

Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. In *International Conference on Learning Representations*, 2016. URL https://api.semanticscholar.org/CorpusID:3178759.

A Framework Details

We discuss further details about ReCogLab.

A.1 General Overview of framework

Here we explicitly describe how the framework constructs training examples to support different experimental hypotheses. We will make the entire framework codebase and dataset available to the public.

Our framework works by:

1. Creating a config file for the dataset you want to create. This includes the task and sub-task (see Table 1 in main paper and the longer task descriptions in Section A.2) and the task parameters (see Section A.3).
2. The user passes in this config as well as the dataset split and the random seed to the dataset generator class.
3. Based on which task is chosen, the dataset generator for the corresponding task generates the underlying logic of the problem (e.g. for Social Networks it generates an undirected graph) and generates the context C text, question Q and answer(s) A from templates. The context is made up of a list of relationship R_{ij} between pairs of entities (E_i, E_j) (each of which is a person or object or other noun in the relationship or can be an attribute). The templates then fill in the text to describe the relationship R_{ij} (often sampling from a number of possible templates) and sampling entities (E_i, E_j) to fill in. The generator also enforces and uses all of the task parameters to do this generation. For details on each of these constructions see the corresponding paragraph of Section A.5.
4. The final datpoint is then passed back and the generator continues to generate up until the desired dataset size N.
5. For experiments which "sweep" parameters, the dataset generation of N examples is repeated, changing the value of the sweep parameter on each step to generate all C_P parameter configurations.

A.2 Task Description

- Comparison
  - Older-Younger: Constructs comparisons on age between people.
  - Larger-Smaller: Constructs comparisons on size. These can incorporate congruent/incongruent knowledge priors.
  - Heavier-Lighter: Constructs comparisons on weight. These can incorporate congruent/incongruent knowledge priors
  - Consistency Detection: Instead of asking a comparison question, asks whether the statements are logically consistent.
  - Indeterminate Conclusion: Generate a comparison tree and ask whether the statements support drawing a conclusion.
- Social Network
  - Fastest Message: Given the goal of passing a message from person A to person B with the fewest hops, who should person B give a message to. Reduces to a logical breadth first search.
  - Oldest/Youngest Generation: Given a statement about a family, who is the oldest and youngest generation? Similar to Older-Younger.
- Family JSON
  - Family Size: Calculate how many members are in a specific family.
  - Family Member Hobby: Given a hobby, check whether it is a hobby of a specific family member.
  - Family Size Comparison: Given two families, compare their size.

  - Hobby Comparison: Given two specific individuals in different family, describe overlapping hobbies.
  - Age Comparison: Given two specific individuals in different family, compare age.
- Syllogism
  - Set Membership: Given statements describing group membership between different labels, determine how two labels' members intersect.

A.3 Task Parameters

- network_type: Whether to construct the task using a linear chain or a randomly generated tree.
- num_entities: The number of entities used when generating an experiment. We use this for evaluating capacity performance.
- entity_type: The type of entities we construct word problems out of. These consist of names, objects, labels depending on the need of the task.
- congruence: If the task supports integrating real world priors to generate congruent or incongruent statements.
- randomize_relations: For a directional statement, whether to reverse the statement and relation. $A > B$ becomes $B < A$ in the word problem.
- do_reverse_comps: To evaluate whether the order of the statement in the question matters.
- relation_type: For social network, applies labels to the edge which include other kinds of relationships.
- num_families: Equivalent to num_entities, measure of capacity for length (total number of families).
- num_family_members: Equivalent to num_entities, measure of capacity for width (max members within a family).
- hop_length: Distance between two families chosen. If not set, default is sampled randomly between half the number of families to overall number of families.

A.4 Common Technical Details

We use NetworkX Python library to construct randomly generated graphs. To seed deterministic randomness, we use Jax's PRNG Keys to ensure that each example is isolated from each other while still being fully reproducible given the correct key.

Some of our experiments generate linear chain graphs. A chain graph is a sequence of nodes where each node is connected to the next one in a linear sequence. This allows us to test for ordering effects in the context because answering the question of a linear chain graph is equivalent to traversing the statements sequentially. Other experiments generate more complex graphs like random trees.

We incorporate multiple stages of deterministic randomness, first at initialization to help randomly configure parameters shared across all test examples, and next at a per-test example generation.

A.5 Domain-specific details

Comparison We use NetworkX to generate linear chain and random trees depending on the subtask.

Comparison sub-tasks consist of asking questions about a bunch of entities in three comparative settings, *Size*, *Age*, and *Weight*. *Consistency Detection* and *Indeterminate Conclusion* use the *Age* comparison but ask a modified question.

For evaluating the ordering effects, we use a linear chain graph shown in the contextual premise ordering experiments in 2. Here inorder and reverse corresponds to matching the toplogical sequential order of the linear chain. Random simply randomizes the ordering. We also use linear chain for Congruent and Incongruent reasoning as well as the Symbolic Distance probe as they simplify the problem construction.

For indeterminate reasoning and inconsistent detection, we use random trees. This is because creating an inconsistent premise requires making a self-loop which is trivial to identify in a linear chain. It's also impossible to construct unknown or infeasible questions with a linear chain because every pair of object has a relationship.

Social Network We use NetworkX to generate linear chain graphs for evaluating capacity.

We evaluate on the sub-task of **Fastest Message** which is a specific question that asks who should [Entity 1]pass a message to have it arrive to [Entity2]in the fewest hops.

To construct a **Fastest Message** problem, we generate premises about the nodes in the graph where edges denote a friendship and a way for a message to be passed. We randomly pick two entities and ask how to pass a message. Here the answer can be determined through breadth first search.

Oldest/Youngest sub-task reformulates the social network graph to a directed problem where nodes indicate relationships between parent and child. The **oldest/youngest** generation asks which family member is part of the oldest or youngest generation. We can solve and verify the answer using a combination of common-ancestor graph algorithms. While we didn't report benchmark performance on this task, we note that **oldest/youngest** has several interesting, distinguishing properties from other tasks presented in this paper. First it does not incorporate any entity names in the question which might inform the language model how to search the context. It also requires processing every bit of information before being able to verifiably prove the answer.

Syllogism We use custom-written logic to generate syllogisms with an arbitrary number of propositions. For a given number of propositions, we keep track of the current valid conclusion, and enumerate all new propositions which, along with the current conclusion, generate a new valid conclusion. This is done in a depth-first way, and when we hit the desired number of propositions, we yield the current chain.

We evaluate the model by showing it all of the propositions in the chain, as well as all syllogism types with the valid conclusion's subject and predicate, i.e. "All A are Z," "No A are Z," "Some A are Z," and "Some A are not Z."

To study ordering effects, we sort the propositions by constructing a Hamiltonian path between the proposition which contains the subject of the conclusion, and the proposition which contains the predicate of the conclusion.

JSON Families For the JSON Families task, our framework produces structured, nested JSON representations of multiple families. Each family is identified by a last name, address, and a set of members. Each member is then identified by their name, age, and hobbies. Table 2 outlines the kind of questions i.e. probes that can be be conducted on this dataset. The question types range from simple fact retrieval (e.g., family size) to complex comparisons (e.g., relative age of members across different families or shared hobbies). This structured data allows us to easily scale the number of families as well as the members per family.

During the dataset generation process, we ensure that the entity in the question is uniquely identifiable (e.g. no two families share the same combination of name and address) and if there are multiple possible answers, we check against all of them for determining correctness.

Sub-tasks for JSON Families

Question Type	Number of entities	Question Details
Family Size	Single Family	Finding the size of a given family in the context
Family Member Hobby	Single Family	Checking if a given hobby is a hobby of a specific member from a specific family
Family Size Comparison	Multi Family	Comparing size of two given families
Family Member Age Comparison	Multi Family	Comparing age of two members from two given families
Family Member Hobby Comparison	Multi Family	Comparing hobbies of two members from two given families

Table 2: Family Question Types

A.6 ENTITY SOURCES

In addition to constructing the problem, we also provide a list of pre-defined entities to populate the problem with. We prepare train-val-test splits on entity names. For names, we use a dataset of 258k popular baby-names[3]. We use this for both Social Network and Comparison-Age.

For comparing objects, we collected a list of 540 commonly encountered physical objects like "Firetruck" and "Shoebox." We used Gemini-Pro to generate many such candidate objects with mass and size that is consistent and easy to measure. We then asked Gemini-Pro to estimate their size and weight to allow us to construct Congruent statements based on their estimated physical properties.

For incongruent samples we simply reverse the congruent ordering.

We also use a random-string configuration for comparison to test the effect of prior knowledge on congruent and incongruent relationships. Specifically we randomly generated alphanumeric ids of length 5 to replace the name of each object in the comparison. This controls for the effect of prior knowledge as entity names like 3Am4O and gj1Bx have no semantic priors in the language model.

For Syllogism we prompted an LLM to generate a diverse set of plural nouns. The entities are randomly assigned when generating the syllogisms.

For JSON Families, the entities include family names (first, last names), address and details of every member. These are randomly chosen from a set of predefined names, city names, states, postal codes and hobbies. These lists were generated by prompting an LLM to provide a commonly used entities, following which a random combination is chosen when identifying a family / a member.

A.7 MISC DETAILS

Rejection Heuristic Sampling Because ReCogLab is a data generating process, we do not have precise control of the posterior distribution of examples. Some capabilities need to be evaluated on rare-occurring events or configurations that are easy to verify but hard to generate. We can use rejection sampling to help promote diversity of these hard-to-find events.

One example of how we use rejection heuristic sampling is to help generate a sufficient number of examples of indeterminate and determinate examples. Another example is increasing the posterior occurrence of high symbolic distance and rejecting low symbolic distance experiment examples which occur more frequently.

We choose this approach of balancing test examples since it doesn't require a specific implementation, as long as a process has a probability of generating an interesting configuration we can upsample it's occurrence in the dataset. Additionally some notions of parameters share common meaning but have wildly different implementation strategies. For instance symbolic distance is trivial to calculate for a linear chain configuration, but impossible to control in a random tree. Rejection sampling resolves this issue by allowing us to upsample rare examples without needing to explicitly define how those rare labels generate.

Flavor Text Flavor text is text that adds depth or background to a character or relationship. This embellishment of factual statements adds an additional layer of cognitive load. Instead of simply presenting statements as facts, flavor text rewrites them to a statement that implies two entities' relationship with each other. We generate a total of 83 flavor texts using Gemini-Pro and specific instructions.

B PROMPT VALIDATION AND MODEL CAPABILITIES

As discussed in Section 3.6, while prompt strategies are an important line of research for exploring language model capabilities, they introduce additional variance to our observations of a language model's true problem-solving capabilities. From a scientific point-of-view, this is problematic; optimizing prompts (even if they are 0-shot) for specific tasks may yield better hill-climbing results,

[3]Source: https://www.cs.princeton.edu/introcs/data/names.csv

but they make it harder to transfer our understanding of a language model's general capabilities to a different task.

To mitigate this issue, we treat prompts and answer parsing as hyperparameters to fit on a validation set first. Because ReCogLabis a generative framework, we can generate always generate validation sets to find the best prompt strategy for each and every language model. This means that every language model is given the opportunity to figuratively "play to their strength."

In our experiments we only consider "0-turn" and "0-trained" prompt strategies to preserve the generalization of our results. We also use several prompts that targeted specific sub-datasets or tasks like Comparisons, Syllogism, and Feasibility because their setup required more explanations. We show the prompts we tested below. [question] refers to both the contextual information and task-specific question combined.

Common Prompts Templates used in all tasks.

- [question]
- [question]
  Answer in only one word.
- [question]
  Think through your answer then respond at the end with a newline and 'Answer:' with your answer.
- [question]
  Think through your answer then respond at the end with a newline and 'Answer:' with your answer. Use only one word for the answer.
- [question]
  Let's think step-by-step

Social Network Prompts Templates

- You are a language model with advanced cognitive abilities. Your task is to understand and reason about the following social scenario, much like a human would. Read the story carefully and answer the questions that follow.
  [question]

Comparison Prompts Templates

- [question]
  Answer the above relational reasoning question with Yes or No. Use only one word for the answer.
- You are a language model being probed for your reasoning abilities. Your task is to carefully think about the following information and answer the question.
  [question]
  Make sure to respond at the end with 'Answer:'
- [question]
  Answer the above relational reasoning question with Yes or No with a newline and 'Answer:' with your answer. Give your best guess if uncertain. Use only one word for the answer.
- [question]
  Answer the above relational reasoning question with only Yes or No with a newline and 'Answer:'. Give your best guess if uncertain. Use only one word for the answer.
- [question]
  Answer the above relational reasoning question with Yes, No, or Unknown. Use Unknown if the question cannot be answer with the information given. Use only one word for the answer.

We find that after validation, different models from the same family of model classes perform better with different prompts.

This framework of splitting prompts and model capabilities works for more complex patterns of cognition like Chain-of-Thought and In-context learning. We also consider sentence parsing but that has a much weaker impact on performance and is also significantly cheaper to test many of. We will release the prompt and sentence parsing functions when releasing code.

C Additional Test Examples

Below we show additional examples generated from our framework for various tasks/sub-tasks/configurations.

C.1 Comparison

C.1.1 Comparison - Objects

Question:
Straightener is smaller than Sugar bowl
Paperweight is larger than Oyster
Folder is smaller than Gemstones
Shaving cream is larger than Sand dollar
Eye pin is smaller than Flower arrangement
Sand dollar is larger than Pliers
Oyster is larger than Lego
Sugar bowl is smaller than Watermelon
Drum is larger than Cello
Folder is larger than Flower arrangement
Pill is smaller than Pliers
Ceiling is smaller than Cello
Paperweight is smaller than Pill
Gemstones is smaller than Keyboard
Ceiling is larger than Audio interface
Straightener is larger than Shaving cream
Keyboard is smaller than Lego
Apple is smaller than Audio interface
Drum is smaller than Eye pin
Is Pill smaller than Oyster?

Answer(s):
No

Question:
Lego is larger than Leaf
Thermostat is larger than Stage
Thermostat is smaller than Trowel
Ribbon is smaller than Saxophone
Sculpture is larger than Saxophone
Cane is smaller than Coffee mug Playbill is smaller than Rake
Rake is smaller than Ribbon
Leaf is larger than Lamp
Banjo is larger than Bandage
Paperweight is smaller than Playbill
Gemstones is larger than Coffee mug
Stage is larger than Sinker
Banjo is smaller than Cane
Bandage is larger than Accordion
Trowel is smaller than Water bottle
Gemstones is smaller than Lamp
Sculpture is smaller than Sinker
Paperweight is larger than Lego
Is Ribbon smaller than Paperweight?

Answer(s):
No

Question:
Trowel is smaller than Wire cutters
Birthday cake is smaller than CD
Mushroom is larger than Measuring cup
CD is smaller than Cap
Thesaurus is larger than Soda can
Measuring cup is larger than Lego
Leaf is larger than Guitar
Mushroom is smaller than Power strip
Extension cord is larger than Envelope
Wire cutters is smaller than Wood
Conditioner is smaller than Cup
Extension cord is smaller than Guitar
Power strip is smaller than Soda can
Envelope is larger than Cup
Trowel is larger than Tripod
Ashtray is smaller than Birthday cake
Leaf is smaller than Lego
Tripod is larger than Thesaurus
Conditioner is larger than Cap
Is Trowel larger than Birthday cake?
Answer(s):
Yes

C.1.2 COMPARISON - PEOPLE

Question:
Silver is younger than Trace
Madyson is older than Lorelai
Evertt is younger than Hilma
Rush is older than Petra
Eulah is younger than Eulalie
Petra is older than Orlena
Lorelai is older than Leta
Trace is younger than Vernetta
Cooper is older than Chrissie
Evertt is older than Eulalie
Nigel is younger than Orlena
Ceil is younger than Chrissie
Madyson is younger than Nigel
Hilma is younger than Khalid
Ceil is older than Betty
Silver is older than Rush
Khalid is younger than Leta
Allison is younger than Betty
Cooper is younger than Eulah
Is Nigel younger than Lorelai?

Answer(s):
No

Question:
Julius is older than Jalissa
Shemar is older than Raquel
Shemar is younger than Treyvon
Lyn is younger than Mae
Natasha is older than Mae

Danyel is younger than Demario
Lonzo is younger than Lori
Lori is younger than Lyn
Jalissa is older than Garey
Cherrie is older than Carmen
Kaleigh is younger than Lonzo
Estefani is older than Demario
Raquel is older than Phoenix
Cherrie is younger than Danyel
Carmen is older than Ayana
Treyvon is younger than Vickey
Estefani is younger than Garey
Natasha is younger than Phoenix
Kaleigh is older than Julius
Question: Is Lyn younger than Kaleigh?
Answer(s):
No

Question:
Myah is younger than Ozzie
Antonetta is younger than Anya
Hettie is older than Elmore
Anya is younger than Arizona
Marely is older than Marcela
Elmore is older than Donn
Devonta is older than Darryl
Hettie is younger than Iver
Cedrick is older than Case
Ozzie is younger than Thomas
Blake is younger than Briana
Cedrick is younger than Darryl
Iver is younger than Marcela
Case is older than Briana
Myah is older than Mikeal
Abraham is younger than Antonetta
Devonta is younger than Donn
Mikeal is older than Marely
Blake is older than Arizona
Is Myah older than Antonetta?
Answer(s):
Yes

C.2 Social Networks

Question:
Arjun is always honest with Marcelo, even when it's hard.
When Arjun needs to talk, Vida is the first one they call.
Casandra is always impressed by Lorena's knowledge and intelligence.
Vida is always willing to listen to Julianne's problems.
Roselyn is like family to Ava.
Exie is always willing to listen to Casandra's problems.
Lukas and Marcelo's families often have dinner together.
Ava is always impressed by Casandra's creativity and artistic talents.
Lorena and Vida enjoy discussing books and movies they've both seen.
Any two friends are able to pass along a message, which allows messages to move from one friend to another. Thus, messages can be passed between two people through friends they have in common.
If Vida wants to get a message to Marcelo as quickly as possible, who should Vida give it to?
Answer(s):
Arjun

Question:
Austin knows how to make Gunnar laugh, even on a bad day.
Barton and Marlana share inside jokes that only they understand.
Daisey always respects Creola's opinions, even when they disagree.
Branden is always honest with Tierra, even when it's hard.
Tierra can always count on Paola for a shoulder to cry on.
Marlana is always the first person Creola calls with good news.
Marvin and Gunnar have inside jokes that no one else understands.
Branden knows all of Barton's favorite snacks and surprises them with them sometimes.
Marvin knows how to tease Barton without hurting their feelings.
Any two friends are able to pass along a message, which allows messages to move from one friend to another. Thus, messages can be passed between two people through friends they have in common.
If Daisey wants to get a message to Tierra as quickly as possible, who should Daisey give it to?
Answer(s):
Creola

Question:
Daijah is always there to listen when Elmore needs to vent about work.
Elmore is always willing to listen to Charlee's problems.
You can often find Anwar and Finn laughing and chatting away.
When Walker needs to talk, Alvin is the first one they call.
Charlee knows how to calm Anwar down when they're stressed or anxious.
Minerva always knows how to cheer Elmore up.
Cathi and Anwar have a mutual respect for each other's personal space and boundaries.
Walker and Charlee enjoy trying new hobbies and activities together.
Toma and Cathi have a deep and meaningful connection.
Any two friends are able to pass along a message, which allows messages to move from one friend to another. Thus, messages can be passed between two people through friends they have in common.
If Cathi wants to get a message to Minerva as quickly as possible, who should Cathi give it to?
Answer(s):
Anwar

C.3 JSON Families

C.3.1 JSON Families - Family Size

Question:
"Family Name": "Wilson", "Address": "544 Pine St, Palo Alto, CA 95841", "Members": ["Name": "Ava", "Age": 13, "Hobbies": ["cycling", "music", "reading"] , "Name": "Grace", "Age": 71, "Hobbies": ["running"] , "Name": "Bob", "Age": 91, "Hobbies": ["writing", "painting", "gardening", "running", "knitting", "cooking"] , "Name": "Diego", "Age": 93, "Hobbies": ["music", "running", "dancing"]] "Family Name": "Wilson", "Address": "695 Divisadero St, Daly City, CA 70635", "Members": ["Name": "Frank", "Age": 45, "Hobbies": ["knitting", "cooking", "cycling"] , "Name": "Diego", "Age": 67, "Hobbies": ["reading", "gardening", "music", "writing", "dancing", "cooking", "traveling"] , "Name": "Bob", "Age": 89, "Hobbies": ["music", "knitting", "gardening", "painting", "writing", "dancing"] , "Name": "Liam", "Age": 96, "Hobbies": ["knitting", "painting"]] "Family Name": "Rodriguez", "Address": "48 Lombard St, San Mateo, CA 35388", "Members": ["Name": "Jack", "Age": 91, "Hobbies": ["music", "gardening"] , "Name": "Alice", "Age": 41, "Hobbies": ["gardening", "writing", "running", "cycling", "reading", "dancing", "music", "traveling", "painting"] , "Name": "Frank", "Age": 100, "Hobbies": ["writing", "gardening", "music"]] "Family Name": "Brown", "Address": "959 Market St, San Jose, CA 10946", "Members": ["Name": "Bob", "Age": 36, "Hobbies": ["running", "music", "painting", "reading", "knitting", "writing"] , "Name": "Bob", "Age": 10, "Hobbies": ["painting", "traveling", "dancing"] , "Name": "Kai", "Age": 88, "Hobbies": ["cooking", "writing"]] "Family Name": "Wilson", "Address": "326 Lombard St, Daly City, CA 99979", "Members": ["Name": "Muhammad", "Age": 63, "Hobbies": ["cycling", "knitting", "writing"] , "Name": "Emily", "Age": 36, "Hobbies": ["writing", "traveling", "knitting", "reading", "running"]]
How many members are in the Brown family living on 959 Market St, San Jose, CA 10946? Answer as a single number.

Answer(s):
3

C.3.2 JSON FAMILIES - FAMILY MEMBER HOBBY

Question:
”Family Name”: ”Wilson”, ”Address”: ”544 Pine St, Palo Alto, CA 95841”, ”Members”: [”Name”: ”Ava”, ”Age”: 13, ”Hobbies”: [”cycling”, ”music”, ”reading”] , ”Name”: ”Grace”, ”Age”: 71, ”Hobbies”: [”running”] , ”Name”: ”Bob”, ”Age”: 91, ”Hobbies”: [”writing”, ”painting”, ”gardening”, ”running”, ”knitting”, ”cooking”] , ”Name”: ”Diego”, ”Age”: 93, ”Hobbies”: [”music”, ”running”, ”dancing”]] ”Family Name”: ”Wilson”, ”Address”: ”695 Divisadero St, Daly City, CA 70635”, ”Members”: [”Name”: ”Frank”, ”Age”: 45, ”Hobbies”: [”knitting”, ”cooking”, ”cycling”] , ”Name”: ”Diego”, ”Age”: 67, ”Hobbies”: [”reading”, ”gardening”, ”music”, ”writing”, ”dancing”, ”cooking”, ”traveling”] , ”Name”: ”Bob”, ”Age”: 89, ”Hobbies”: [”music”, ”knitting”, ”gardening”, ”painting”, ”writing”, ”dancing”] , ”Name”: ”Liam”, ”Age”: 96, ”Hobbies”: [”knitting”, ”painting”]] ”Family Name”: ”Rodriguez”, ”Address”: ”48 Lombard St, San Mateo, CA 35388”, ”Members”: [”Name”: ”Jack”, ”Age”: 91, ”Hobbies”: [”music”, ”gardening”] , ”Name”: ”Alice”, ”Age”: 41, ”Hobbies”: [”gardening”, ”writing”, ”running”, ”cycling”, ”reading”, ”dancing”, ”music”, ”traveling”, ”painting”] , ”Name”: ”Frank”, ”Age”: 100, ”Hobbies”: [”writing”, ”gardening”, ”music”]] ”Family Name”: ”Brown”, ”Address”: ”959 Market St, San Jose, CA 10946”, ”Members”: [”Name”: ”Bob”, ”Age”: 36, ”Hobbies”: [”running”, ”music”, ”painting”, ”reading”, ”knitting”, ”writing”] , ”Name”: ”Bob”, ”Age”: 10, ”Hobbies”: [”painting”, ”traveling”, ”dancing”] , ”Name”: ”Kai”, ”Age”: 88, ”Hobbies”: [”cooking”, ”writing”]] ”Family Name”: ”Wilson”, ”Address”: ”326 Lombard St, Daly City, CA 99979”, ”Members”: [”Name”: ”Muhammad”, ”Age”: 63, ”Hobbies”: [”cycling”, ”knitting”, ”writing”] , ”Name”: ”Emily”, ”Age”: 36, ”Hobbies”: [”writing”, ”traveling”, ”knitting”, ”reading”, ”running”]]
Is writing a hobby of Jack from the Rodriguez family living on 48 Lombard St, San Mateo, CA 35388? Answer with Yes or No. Answers: **Answer(s):**
No

C.3.3 JSON FAMILIES - FAMILY SIZE COMPARISON

Question:
”Family Name”: ”Wilson”, ”Address”: ”544 Pine St, Palo Alto, CA 95841”, ”Members”: [”Name”: ”Ava”, ”Age”: 13, ”Hobbies”: [”cycling”, ”music”, ”reading”] , ”Name”: ”Grace”, ”Age”: 71, ”Hobbies”: [”running”] , ”Name”: ”Bob”, ”Age”: 91, ”Hobbies”: [”writing”, ”painting”, ”gardening”, ”running”, ”knitting”, ”cooking”] , ”Name”: ”Diego”, ”Age”: 93, ”Hobbies”: [”music”, ”running”, ”dancing”]] ”Family Name”: ”Wilson”, ”Address”: ”695 Divisadero St, Daly City, CA 70635”, ”Members”: [”Name”: ”Frank”, ”Age”: 45, ”Hobbies”: [”knitting”, ”cooking”, ”cycling”] , ”Name”: ”Diego”, ”Age”: 67, ”Hobbies”: [”reading”, ”gardening”, ”music”, ”writing”, ”dancing”, ”cooking”, ”traveling”] , ”Name”: ”Bob”, ”Age”: 89, ”Hobbies”: [”music”, ”knitting”, ”gardening”, ”painting”, ”writing”, ”dancing”] , ”Name”: ”Liam”, ”Age”: 96, ”Hobbies”: [”knitting”, ”painting”]] ”Family Name”: ”Rodriguez”, ”Address”: ”48 Lombard St, San Mateo, CA 35388”, ”Members”: [”Name”: ”Jack”, ”Age”: 91, ”Hobbies”: [”music”, ”gardening”] , ”Name”: ”Alice”, ”Age”: 41, ”Hobbies”: [”gardening”, ”writing”, ”running”, ”cycling”, ”reading”, ”dancing”, ”music”, ”traveling”, ”painting”] , ”Name”: ”Frank”, ”Age”: 100, ”Hobbies”: [”writing”, ”gardening”, ”music”]] ”Family Name”: ”Brown”, ”Address”: ”959 Market St, San Jose, CA 10946”, ”Members”: [”Name”: ”Bob”, ”Age”: 36, ”Hobbies”: [”running”, ”music”, ”painting”, ”reading”, ”knitting”, ”writing”] , ”Name”: ”Bob”, ”Age”: 10, ”Hobbies”: [”painting”, ”traveling”, ”dancing”] , ”Name”: ”Kai”, ”Age”: 88, ”Hobbies”: [”cooking”, ”writing”]] ”Family Name”: ”Wilson”, ”Address”: ”326 Lombard St, Daly City, CA 99979”, ”Members”: [”Name”: ”Muhammad”, ”Age”: 63, ”Hobbies”: [”cycling”, ”knitting”, ”writing”] , ”Name”: ”Emily”, ”Age”: 36, ”Hobbies”: [”writing”, ”traveling”, ”knitting”, ”reading”, ”running”]]
Which family is larger, the Wilson family living on 326 Lombard St, Daly City, CA 99979 or the Brown family living on 959 Market St, San Jose, CA 10946? Answer with the family name of the larger family. **Answer(s):**
Brown

C.3.4 JSON FAMILIES - FAMILY MEMBER AGE COMPARISON

Question:
”Family Name”: ”Wilson”, ”Address”: ”544 Pine St, Palo Alto, CA 95841”, ”Members”: [”Name”: ”Ava”, ”Age”: 13, ”Hobbies”: [”cycling”, ”music”, ”reading”] , ”Name”: ”Grace”, ”Age”: 71, ”Hobbies”: [”running”] , ”Name”: ”Bob”, ”Age”: 91, ”Hobbies”: [”writing”, ”painting”, ”gardening”, ”running”, ”knitting”, ”cooking”] , ”Name”: ”Diego”, ”Age”: 93, ”Hobbies”: [”music”, ”running”, ”dancing”]] ”Family Name”: ”Wilson”, ”Address”: ”695 Divisadero St, Daly City, CA 70635”, ”Members”: [”Name”: ”Frank”, ”Age”: 45, ”Hobbies”: [”knitting”, ”cooking”, ”cycling”] , ”Name”: ”Diego”, ”Age”: 67, ”Hobbies”: [”reading”, ”gardening”, ”music”, ”writing”, ”dancing”, ”cooking”, ”traveling”] , ”Name”: ”Bob”, ”Age”: 89, ”Hobbies”: [”music”, ”knitting”, ”gardening”, ”painting”, ”writing”, ”dancing”] , ”Name”: ”Liam”, ”Age”: 96, ”Hobbies”: [”knitting”, ”painting”]] ”Family Name”: ”Rodriguez”, ”Address”: ”48 Lombard St, San Mateo, CA 35388”, ”Members”: [”Name”: ”Jack”, ”Age”: 91, ”Hobbies”: [”music”, ”gardening”] , ”Name”: ”Alice”, ”Age”: 41, ”Hobbies”: [”gardening”, ”writing”, ”running”, ”cycling”, ”reading”, ”dancing”, ”music”, ”traveling”, ”painting”] , ”Name”: ”Frank”, ”Age”: 100, ”Hobbies”: [”writing”, ”gardening”, ”music”]] ”Family Name”: ”Brown”, ”Address”: ”959 Market St, San Jose, CA 10946”, ”Members”: [”Name”: ”Bob”, ”Age”: 36, ”Hobbies”: [”running”, ”music”, ”painting”, ”reading”, ”knitting”, ”writing”] , ”Name”: ”Bob”, ”Age”: 10, ”Hobbies”: [”painting”, ”traveling”, ”dancing”] , ”Name”: ”Kai”, ”Age”: 88, ”Hobbies”: [”cooking”, ”writing”]] ”Family Name”: ”Wilson”, ”Address”: ”326 Lombard St, Daly City, CA 99979”, ”Members”: [”Name”: ”Muhammad”, ”Age”: 63, ”Hobbies”: [”cycling”, ”knitting”, ”writing”] , ”Name”: ”Emily”, ”Age”: 36, ”Hobbies”: [”writing”, ”traveling”, ”knitting”, ”reading”, ”running”]]
Who is older: Muhammad from the Wilson family living on 326 Lombard St, Daly City, CA 99979 or Jack from the Rodriguez family living on 48 Lombard St, San Mateo, CA 35388? If both are the same age, answer with the name that comes first alphabetically. Answer with the name.

Answer(s):
Jack

C.3.5 JSON FAMILIES - FAMILY MEMBER HOBBY COMPARISON

Question:
”Family Name”: ”Wilson”, ”Address”: ”544 Pine St, Palo Alto, CA 95841”, ”Members”: [”Name”: ”Ava”, ”Age”: 13, ”Hobbies”: [”cycling”, ”music”, ”reading”] , ”Name”: ”Grace”, ”Age”: 71, ”Hobbies”: [”running”] , ”Name”: ”Bob”, ”Age”: 91, ”Hobbies”: [”writing”, ”painting”, ”gardening”, ”running”, ”knitting”, ”cooking”] , ”Name”: ”Diego”, ”Age”: 93, ”Hobbies”: [”music”, ”running”, ”dancing”]] ”Family Name”: ”Wilson”, ”Address”: ”695 Divisadero St, Daly City, CA 70635”, ”Members”: [”Name”: ”Frank”, ”Age”: 45, ”Hobbies”: [”knitting”, ”cooking”, ”cycling”] , ”Name”: ”Diego”, ”Age”: 67, ”Hobbies”: [”reading”, ”gardening”, ”music”, ”writing”, ”dancing”, ”cooking”, ”traveling”] , ”Name”: ”Bob”, ”Age”: 89, ”Hobbies”: [”music”, ”knitting”, ”gardening”, ”painting”, ”writing”, ”dancing”] , ”Name”: ”Liam”, ”Age”: 96, ”Hobbies”: [”knitting”, ”painting”]] ”Family Name”: ”Rodriguez”, ”Address”: ”48 Lombard St, San Mateo, CA 35388”, ”Members”: [”Name”: ”Jack”, ”Age”: 91, ”Hobbies”: [”music”, ”gardening”] , ”Name”: ”Alice”, ”Age”: 41, ”Hobbies”: [”gardening”, ”writing”, ”running”, ”cycling”, ”reading”, ”dancing”, ”music”, ”traveling”, ”painting”] , ”Name”: ”Frank”, ”Age”: 100, ”Hobbies”: [”writing”, ”gardening”, ”music”]] ”Family Name”: ”Brown”, ”Address”: ”959 Market St, San Jose, CA 10946”, ”Members”: [”Name”: ”Bob”, ”Age”: 36, ”Hobbies”: [”running”, ”music”, ”painting”, ”reading”, ”knitting”, ”writing”] , ”Name”: ”Bob”, ”Age”: 10, ”Hobbies”: [”painting”, ”traveling”, ”dancing”] , ”Name”: ”Kai”, ”Age”: 88, ”Hobbies”: [”cooking”, ”writing”]] ”Family Name”: ”Wilson”, ”Address”: ”326 Lombard St, Daly City, CA 99979”, ”Members”: [”Name”: ”Muhammad”, ”Age”: 63, ”Hobbies”: [”cycling”, ”knitting”, ”writing”] , ”Name”: ”Emily”, ”Age”: 36, ”Hobbies”: [”writing”, ”traveling”, ”knitting”, ”reading”, ”running”]]
What hobbies do Muhammad from the Wilson family living on 326 Lombard St, Daly City, CA 99979 and Jack from the Rodriguez family living on 48 Lombard St, San Mateo, CA 35388 share? List the hobbies in alphabetical order, separated by commas, or answer N/A if they share no hobbies.
Answer(s):
N/A

C.4 SYLLOGISMS

Question:
All accounts are actions
No actions are actors
Which of the following is true?
All accounts are actors
No accounts are actors
Some accounts are actors
Some accounts are not actors
Answer(s):
No accounts are actors

Question:
No accounts are actions
All actors are accounts
Which of the following is true?
All actors are actions
No actors are actions
Some actors are actions
Some actors are not actions
Answer(s): No actors are actions

Question:
Some accounts are actions
No actors are accounts
Which of the following is true?
All actions are actors
No actions are actors
Some actions are actors
Some actions are not actors
Some actions are not actors