SpaRC and SpaRP: Spatial Reasoning Characterization and Path Generation for Understanding Spatial Reasoning Capability of Large Language Models

Anonymous ACL submission

Abstract

Spatial reasoning is a crucial component of both biological and artificial intelligence. In this work, we present a comprehensive study of the capability of current state-of-the-art large language models (LLMs) on spatial reasoning. To support our study, we created and contribute novel spatial characterization frameworks and datasets, the **Spa**tial **R**easoning **C**haracterization (SpaRC), and **Spa**tial **R**easoning **P**aths (SpaRP), to enable an in-depth understanding of the spatial relations and compositions as well as the usefulness of spatial reasoning chains. We found that all state-of-the-art LLMs do not perform well on the datasets—their performances are consistently low across different setups. The spatial reasoning is an emergent capability as model sizes scale up. Finetuning both large language models (e.g., Llama-2-70B) and smaller ones (e.g., Llama-2-13B) can significantly improve their F1-scores by 7–32 absolute points. We also found that the top proprietary LLMs still significantly outperform their open-source counterparts in topological spatial understanding and reasoning.

1 Introduction

Spatial understanding and reasoning are a crucial component of both biological and artificial intelligence, essential for daily interactions and common tasks such as dialogues and conversations (Kruijff et al., 2007; Udagawa and Aizawa, 2019), navigation (Anderson et al., 2018; Chen et al., 2019; Zhang and Kordjamshidi, 2022), and robotics (Bisk et al., 2016; Venkatesh et al., 2021), among others. They require common reasoning steps such as identifying objects, determining other objects being involved, and aggregating multiple spatial relations to reach a conclusion. The advancement of the field has significantly benefited from many well-known tasks and datasets, including bAbI (Weston et al., 2016), SpartQA (Mirzaee et al., 2021), SpaRTUN and ReSQ (Mirzaee and Kordjamshidi, 2022), and StepGame (Shi et al., 2022), among others.

Recently, Large Language Models (LLMs) have been shown to be capable of performing abstract, commonsense-based, and multi-hop reasoning (Wei et al., 2022b; Kojima et al., 2022; Wang et al., 2023). If such models are to be used as intelligent agents to answer questions, perform tasks, and collaborate with humans, whether they can understand the basic spatial relationships and perform corresponding reasoning would become critical to many real-life applications.

In this work, we present an extensive study on the state-of-the-art LLMs' capability in spatial reasoning. The key components of spatial abilities include: (i) understanding spatial relations and composition, and (ii) developing reasoning chains to reach conclusions. Prior work (Mirzaee et al., 2021; Mirzaee and Kordjamshidi, 2022; Shi et al., 2022) has focused on the relations and spatial composition tied to a limited context setup, as will be detailed later in this paper. In our work, we propose a bottom-up approach that builds upon detailed spatial properties, providing fine control for constructing spatial rules and context setups. We formalize and propose **Spa**tial **R**easoning **C**haracterization (SpaRC), a systematic framework in defining *spatial properties* of objects, relations, and contexts, as well as how they *characterize* spatial composition, which is inspired by the widely used benchmarks SpaRTUN (Mirzaee and Kordjamshidi, 2022) and StepGame (Shi et al., 2022).

Reasoning paths are an integral part of the reasoning process and critical for analyzing and enhancing reasoning models. To the best of our knowledge, unlike other reasoning tasks such as mathematical reasoning, there exist no datasets with textual spatial reasoning paths. In this paper we develop deductively verified spatial reasoning paths by using spatial reasoners to generate step-by-step reasoning on SpaRTUN and StepGame,

which is then verbalized to form textual chain-of-thoughts. We show that finetuning different sizes of LLMs (13B and 70B) on the reasoning paths significantly improve their spatial reasoning performance, which also highlights the poor performance of the generalist pretrained LLMs (without finetuning) on spatial reasoning. In summary, our contributions are as follows:

- We present a comprehensive study on the spatial reasoning capabilities of the state-of-the-art LLMs, under extensive setups: comprehensive spatial characterizations, different parameter scales, pretrained *vs.* finetuned models, and different decoding strategies. We show that the current LLMs do not perform well on the spatial reasoning tasks. We observe that spatial reasoning is an emergent capability as model sizes scale up. Top proprietary LLMs still significantly outperform their open-source counterparts in topological spatial reasoning.
- To support an in-depth study, we present the **Spa**tial **R**easoning **C**haracterization (SpaRC) framework, a systematic bottom-up approach that shifts the focus towards spatial properties, providing a fine and flexible control on the spatial composition rules and context setups. We characterize and extend the widely used benchmark datasets SPARTUN and StepGame under the SpaRC framework.
- We develop **Spa**tial **R**easoning **P**aths (SpaRP) by generating reasoning steps using symbolic spatial reasoners and verbalizing them in a deductive step-by-step process. We demonstrate that finetuning large language models on our reasoning paths can consistently improve their spatial reasoning abilities.

2 Related Work

Text-based Spatial Reasoning. Textual spatial reasoning datasets present the task as question-answering (SRQA) over a textual spatial context. Weston et al. (2016) introduced bAbI containing two datasets focused on positional (Task 17) and navigational (Task 19) reasoning. Their simplistic nature and small size prompted subsequent works to create new and challenging datasets. Mirzaee et al. (2021) designed reasoning rules, and created human-generated and synthetic context-question-answer tuples from spatial description of visual scenes (SPARTQA) to train and evaluate spatial reasoning of neural language models. Mirzaee and Kordjamshidi (2022) further extended the spatial rules to cover 16 spatial relations over multiple formalism in 3D in their synthetic SPARTUN dataset, and commonsense spatial reasoning in human-generated RESQ dataset. StepGame (Shi et al., 2022) was introduced to assess robust positional multi-hop spatial reasoning in 2D. Our SpaRC framework builds on top of SPARTUN and StepGame as they provide a broad coverage over the number of hops and relations for abstract spatial reasoning.

Reasoning Abilities of Large Language Models. Certain reasoning capabilities have been shown to be emergent abilities of LLMs (Wei et al., 2022a), which are further elicited by various chain-of-thought prompting techniques (Wei et al., 2022b; Kojima et al., 2022; Yao et al., 2023; Hao et al., 2023). On logic-based tasks, including spatial reasoning, they however lag significantly when compared to neuro-symbolic methods (Mirzaee and Kordjamshidi, 2023; Yang et al., 2023).

To understand spatial reasoning abilities, Bang et al. (2023) provided a preliminary probing analysis on ChatGPT using a very small dataset (60 examples from each of StepGame and SPARTQA). Yang et al. (2023) evaluated the performance of GPT-3 on StepGame; Mirzaee and Kordjamshidi (2023) reported the performance of GPT-3 on SPARTQA, SPARTUN, and RESQ datasets. However, these work are limited in terms of evaluation metric, qualitative analysis, past generation of LLMs, pretrained LLMs, or generation strategies. To the best of our knowledge, our work is the first attempt at a comprehensive evaluation of spatial reasoning of LLMs under these settings.

3 The Spatial Reasoning Characterization (SpaRC) Framework

The steps to identify and compose spatial relations between entities distinguish spatial reasoning from other reasoning tasks. Prior work e.g. SPARTUN (Mirzaee and Kordjamshidi, 2022) and StepGame (Shi et al., 2022), have focused directly on the spatial composition rules coupled with the contexts, which can lead to different conclusions even for the same set of relations. For example, for the same context "A is left of B and B is above C", applying the spatial composition of StepGame concludes that A is to the left and above C, while no directional relation between A and C can be concluded at all by applying the spatial rules of SPAR-

TUN. The conclusions are completely different *but* equally valid. This difference can be reconciled by examining the underlying spatial properties of the objects and relations, specifically the treatment of objects as points vs extended, and completeness of the knowledge of relations in the context. We, therefore, advocate for an extendable bottom-up approach starting from a more granular level and introduce the **Spa**tial **R**easoning **C**haracterization (SpaRC) framework. SpaRC prioritizes spatial properties over spatial composition rules. Consequently, it offers finer control in creating contexts and facilitates a deeper and systematic examination of the spatial reasoning capabilities.

To keep our work closer and comparable to the widely used existing benchmarks, SPARTUN (Mirzaee and Kordjamshidi, 2022) and StepGame (Shi et al., 2022), we identify *six* properties that cover and characterize these datasets by two *distinct and mutually exclusive* sets of *three* properties each. With SpaRC, we further explore two *properties sets* (PS) with properties in common to these existing benchmarks.

$\mathcal{F}$	Sub-Type	Relations ($\mathcal{R}$)	Textual Label ($\mathcal{L}$)
Topological	$\mathcal{T}_R$ (RCC8)	DC	outside
		EC	outside and touching
		PO	partially overlapping
		EQ	overlapping
		TPP	inside and touching
		NTPP	inside
		TPPI	contains and touches
		NTPPI	contains
Directional	$\mathcal{D}_R$ (Relative)	LEFT	left
		RIGHT	right
		ABOVE	above
		BELOW	below
		FRONT	front
		BEHIND	behind
	$\mathcal{D}_C$ (Cardinal)	NORTH	above
		SOUTH	below
		EAST	right
		WEST	left
	$\mathcal{D}_T$ (Clock)	12 o'clock	above
		3 o'clock	right
		6 o'clock	below
		9 o'clock	left
Distance	$\mathcal{S}_Q$(Qualitative)	NEAR	near
		FAR	far
	$\mathcal{S}_U$(Quantitative)	–	–

Table 1: Formalisms ($\mathcal{F}$) and their sub-types, relations ($\mathcal{R}$) in the datasets and their labels ($\mathcal{L}$). Labels are presented in natural language to work with language models. Composite relations e.g. lower-left are considered in a multi-label setting in the present work.

3.1 Principle and Design of SpaRC

We focus on a set of binary spatial relations $\mathcal{R}$ (Table 1) by following the previous work (Mirzaee and Kordjamshidi, 2022; Shi et al., 2022). The relations cover three formalism ($\mathcal{F}$)—topological $\mathcal{T}$, directional $\mathcal{D}$, and distance $\mathcal{S}$, divided into sub-types—

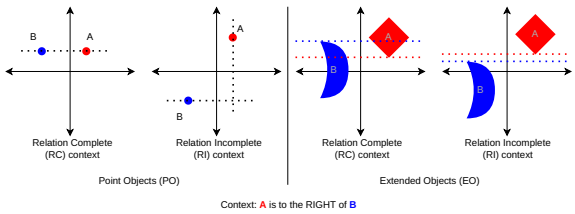

Figure 1: Visualization of Relation Complete (RC) and Relation Incomplete (RI) contexts for the RIGHT relation for Point Objects (PO) and Extended Objects (EO).

region connection calculus (RCC8) $\mathcal{T}_R$, relative directions $\mathcal{D}_R$, cardinal directions $\mathcal{D}_C$, clock-face directions $\mathcal{D}_T$, qualitative distance $\mathcal{S}_Q$, and quantitative distance $\mathcal{S}_U$.

For the relations set $\mathcal{R}$ and a given set of entities $\mathcal{E}$, we denote a context $\mathcal{C} = \{(h, r, t)_i\}_{i=1}^{N}$ as a set of (h, r, t) tuples, where $h \in \mathcal{E}$ is a head entity, $t \in \mathcal{E}$ is the tail entity, and $r \in \mathcal{R}$ is the binary relation. Without loss of generality, objects are considered to be in a 2D space with (x_s, y_s) and (x_e, y_e) as the start and end positions. We now identify and describe six spatial properties of the objects, contexts, and relations that are crucial in determining their spatial composition rules. Refer to Appendix A for a more detailed discussion.

Fixed Orientation or Point of View (FPoV). The directional relations are considered to be axis-aligned from a fixed orientation or point of view, i.e., fixed axes in a 2D or 3D space. A fixed mapping across the relative, cardinal, and clock-face directions is usually chosen. Consistent with the prior work, we map and canonicalize cardinal $\mathcal{D}_C$ and clock-face $\mathcal{D}_T$ relations to four relative directions $\mathcal{D}_R$ (Table 1), *only* for their label representations $\mathcal{L}$. We denote the 2D subset of directions as ${}^{2D}\mathcal{D} = \mathcal{D} \setminus \{\texttt{FRONT}, \texttt{BEHIND}\}$.

Point Objects (PO). A point object satisfies $x_s = x_e \wedge y_s = y_e$. As they are dimensionless, point objects have reduced set of relations with reference to other point objects. Real objects can be treated as point objects in practical contexts when their sizes are negligible.

Extended Objects (EO). An object is said to be an extended object if $x_s \neq x_e \vee y_s \neq y_e$. In SpaRC, we extend StepGame by considering extended objects in addition to point objects. We further study additional composition rules for extended objects than presented in SPARTUN, as will be detailed later in Section 3.2.

Relation	Point Objects (PO)	Extended Objects (EO)
Incomplete (RI):		
RIGHT(A,B)	$x_A > x_B$	$x_s^A \geq x_e^B$
BELOW(A,B)	$y_A < y_B$	$y_e^A \leq y_s^B$
Complete (RC):		
RIGHT(A,B)	$x_A > x_B \wedge$ $y_A = y_B$	$x_s^A \geq x_e^B \wedge$ $y_s^B \leq y_e^A \wedge y_e^B \geq y_s^A$
BELOW(A,B)	$y_A < y_B \wedge$ $x_A = x_B$	$y_e^A \leq y_s^B \wedge$ $x_s^A \leq x_e^B \wedge x_e^A \geq x_s^B$
RIGHT(A,B) $\wedge$ BELOW(A,B)	$x_A > x_B \wedge$ $y_A < y_B$	$x_s^A \geq x_e^B \wedge$ $y_e^A \leq y_s^B$

Table 2: Mathematical descriptions of Relation Incomplete (RI) and Relation Complete (RC) contexts for the relations RIGHT, BELOW, and their combination in terms of entity positions (x, y) for Point Objects (PO) or entity boundaries (x_s, x_e, y_s, y_e) for Extended Objects (EO).

Relation Incomplete (RI). We *introduce* the term relation incomplete (RI) for a context $\mathcal{C}$ between a head h and a tail t entity if *not all* the relations $r \in \mathcal{R}$ between these entities are *considered* to be known and expressed in the context. Thus, the knowledge for the expressed relations should be *treated* as incomplete or partial for spatial composition. For example, "Ron is to the right of Hermione" as a RI context means that the direction orthogonal to the RIGHT could be ABOVE or BELOW as well. The state of positions or boundaries of objects on the *orthogonal* axes cannot be assumed. Table 2 and Figure 1 exemplifies and visualizes this for a few scenarios.

Relation Complete (RC). We *introduce* the term relation complete (RC) for a context $\mathcal{C}$ between h and t if *all* the relations $r \in \mathcal{R}$ between these entities are *considered* to be known and expressed in the context, and treated as such for spatial compositions. For the previous example "Ron is to the right of Hermione" to be considered as RC, the context should mean that Ron is only to the RIGHT of Hermione, and not to her lower-right or upper-right side. The positions or boundaries of objects on the *orthogonal* direction axes should coincide or overlap. Table 2 and Figure 1 exemplifies and visualizes this for a few scenarios. In SpaRC, we further consider this property in conjunction with other properties, such as extended objects, to design composition rules that are not present in StepGame, as discussed later in Section 3.2.

Quantitatively Specified (QS). A relation which is stated in terms of a unit of measurement is said to be quantitatively specified in the given context. Quantitatively specified relations that are *reverse* of each other, e.g. {LEFT, RIGHT}, can readily be composed.

Quantitatively Unspecified (QU). A relation which can be stated in terms of a unit of measurement but is not stated as such in a given context is said to be quantitatively unspecified. Quantitatively unspecified relations that are *reverse* of each other, e.g. {LEFT, RIGHT}, cannot be composed unless they are quantified. In SpaRC, we design and study the reasoning abilities for this property in conjunction with other properties, such as point objects, that are not present in SPARTUN and StepGame, as discussed later in Section 3.2.

We restrict our study to the above 6 properties to keep it closer and comparable to the existing benchmarks, SPARTUN and StepGame. These properties form 3 *mutually exclusive* pairs—{EO,PO}, {RI,RC}, {QS,QU}, leading to 8 possible sets. SpaRC can be extended with additional properties, however, we note that the number of possible characterizations increases exponentially with the number of properties.

3.2 Creation of The SpaRC Dataset

We *identify* the properties set PS for the existing benchmarks, as formalized in the previous section, based on the generation process of the context and the spatial composition rules. More concretely, we identify that SPARTUN is characterized by the properties set PS1 = {EO,RI,QU}, while StepGame is characterized by the properties set PS2 = {PO,RC,QS}. These properties sets are mutually exclusive with PS2 supporting *stronger* composition rules than PS1 for a given context, e.g. "A is left of B and B is above C" as discussed earlier. Refer to Appendix B for more details.

In the SpaRC framework, we construct two additional datasets by relaxing the properties of StepGame from PO to EO, and QS to QU. We chose to extend StepGame as it is simple with fewer relations (only directional which is common across datasets and benchmarks) and challenging (more number of hops). Concretely, we create the datasets SpaRC-PS3 with the properties PS3 = {PO,RC,QU}, and SpaRC-PS4 with the properties PS4 = {EO,RC,QU}. Their composition rules, elaborated upon in Section 4, are formalized by the Algorithm 1 and Algorithm 2 respectively.

We confine our study to these four property sets as they encompass the two existing benchmarks and have covered the most interesting spatial rules

Not	$\forall(X,Y) \in Entities$	$R \in \{Dir \vee PP\}$	IF $R(X,Y)$	$\implies$ NOT$(R_{reverse}(X,Y))$
Inverse	$\forall(X,Y) \in Entities$	$R \in \{Dir \vee PP\}$	IF $R(Y,X)$	$\implies R_{reverse}(X,Y)$
Symmetry	$\forall(X,Y) \in Entities$	$R \in \{Dis \vee (RCC-PP)\}$	IF $R(Y,X)$	$\implies R(X,Y)$
Transitivity	$\forall(X,Y,Z) \in Entities$	$R \in \{Dir \vee PP\}$	IF $R(X,Z), R(Z,Y)$	$\implies R(X,Y)$
Combination	$\forall(X,Y,Z,H) \in Entities$	$R \in Dir, *PP \in PP$	IF $*PP(X,Z), R(Z,H), *PPi(Z,Y)$	$\implies R(X,Y)$

Table 3: Spatial Rules reproduced from SPARTUN (Mirzaee and Kordjamshidi, 2022). Dir: Directional relations (e.g., `LEFT`), Dis: Distance relations (e.g., `FAR`), PP: all Proper parts relations (`NTPP`, `NTPPI`, `TPPI`, `TPP`), $RCC-PP$: All RCC8 relation except proper parts relations. $*PP$: one of `TPP` or `NTPP`. $*PPi$: one of `NTPPi` or `TPPi`.

Dataset	$\mathcal{F}$	Properties	Textual Reason.	Split	# Context	# Ques.
SPARTUN	$\mathcal{T}_R$,$\mathcal{D}$,$\mathcal{S}_Q$	EO,RI,QU	✗	Train	6039	18400
				Dev	915	2818
				Test	925	2830
SpaRP-PS1			✓	Train	5806	16348
				Dev	877	2392
				Test	872	2301
StepGame	$^{2D}\mathcal{D}$, $\mathcal{S}_U$	PO,RC,QS	✗	Train	50000	50000
				Dev	5000	5000
				Test	100000	100000
SpaRP-PS2			✓	Train	49243	49243
				Dev	4927	4927
				Test	98614	98614
SpaRP-PS3	$^{2D}\mathcal{D}$	PO,RC,QU	✓	Train	44666	44666
				Dev	4494	4494
				Test	78092	78092
SpaRP-PS4		EO,RC,QU	✓	Train	41436	41436
				Dev	4171	4171
				Test	69474	69474

Table 4: Comparison between the extended (SpaRP) dataset and the source datasets. Descriptions of the properties are provided in Section 3.1. Relations contained in the formalisms are presented in Table 1. All the questions are of Find Relations (FR) types.

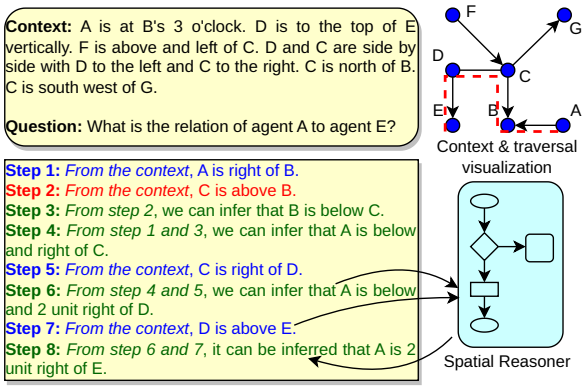

Figure 2: Our step-by-step deductive **Spa**tial **R**easoning **P**aths (SpaRP) generation. A context graph and node traversal from the head to the tail entity in a question is identified and verbalized. Blue indicates context relations r^c, red indicates inverse context relations r^{ic}, and green indicates deduced relations r^d between entities while traversing the reasoning path A–B–C–D–E.

and composition; the additional extension is more routine and we leave that as future work.

4 The Spatial Reasoning Paths (SpaRP)

Reasoning paths are an integral part of reasoning models and critical for analyzing and enhancing such models. To the best of our knowledge, unlike other reasoning tasks such as mathematical reasoning, there exist no datasets with spatial reasoning paths. In this section, we develop deductively verified spatial reasoning paths by *verbalizing* the symbolic steps.

Existing spatial reasoning datasets can be considered as a collection of context-question-answer $(\mathcal{C}, \mathcal{Q}, \mathcal{A})$ tuples. Formally, we denote a context $\mathcal{C} = \{(h,r,t)_i\}_{i=1}^{N}$ defined over a set of entities $\mathcal{E}$ and binary relations $\mathcal{R}$ as a set of (h,r,t) tuples, where $h \in \mathcal{E}$ is the head entity, $t \in \mathcal{E}$ is the tail entity and $r \in \mathcal{R}$ is the binary relation. For a given $(\mathcal{C}, \mathcal{Q}, \mathcal{A})$ tuple, seeking relation between the head h_q and tail t_q entities, we define a *symbolic* reasoning path $\mathcal{P} = (l_i)_{i=1}^{L}$ as a sequence of L reasoning links $l_i = (h_i, r_i^{\cup}, t_i)$ such that $h_1 = h_q$, $t_L = t_q$, and $h_i = t_{i-1}$ for $1 < i \leq L$. We define $r^{\cup} = r^c \cup r^{ic} \cup r^d$, where r^c denotes the set of relations present in the **c**ontext, r^{ic} denotes the **i**nverse relations present in the **c**ontext i.e. relations from t to h, and r^d denotes the set of **d**educed relations. Following the format of deductively verified chain-of-thought (Ling et al., 2023), we *verbalize* the reasoning path $\mathcal{P}$ as a series of step-by-step reasoning sentences, where each step receives their necessary context and premises (Figure 2). The overall process is as given below:

1. Entities and their relations in the contexts are either pre-annotated (SPARTUN) or extracted using regex pattern matching (StepGame) to construct the symbolic context $\mathcal{C}$.
2. A traversal path $\mathcal{P}$ is identified from h_q to t_q by constructing a network graph over $\mathcal{C}$. The deduced relations r^d are initialized to be the inverse of r^{ic}, to traverse and merge steps in a single direction from h_q to t_q (Figure 2).
3. We traverse the path $\mathcal{P}$, progressively merging the links (as $h_i = t_{i-1}$) and updating the deduced relations r^d based on the properties set PS and their spatial composition rules:
   - For SPARTUN we reuse the rules from

Algorithm 1 Relative Direction composition for set of properties PS2 and PS3 in 2D.

Input: Pairs to compose $\{pair1, pair2\}$.
$quantitative \in \{true, false\}$.
Output: $merged$ pair.

```
/* initialized pair starts with dx = dy = 0 */
merged ← InitializePair
merged.head ← pair1.head
merged.tail ← pair2.tail
for pair ∈ {pair1, pair2} do
    for delta ∈ {dx, dy} do
        delta ← merged.delta + pair.delta
        /* Handle direction reversal and quantitatively
           unspecified */
        if (merged.delta × pair.delta < 0) and not
        quantitative then
            /* Set as NaN to invalidate compositions
               from now on in this direction */
            merged.delta ← NaN
        else
            merged.delta ← delta
        end if
    end for
end for
```

Algorithm 2 Relative Direction composition for set of properties PS4 in 2D.

Input: Pairs to compose $\{pair1, pair2\}$.
current set of constraint inequalities $ineq$
Output: $merged$ pair and updated inequalities $ineq$.

```
/* initialize an empty pair */
merged ← InitializePair
merged.head ← pair1.head
merged.tail ← pair2.tail
for rel ∈ {LEFT, RIGHT, ABOVE, BELOW} do
    candidate_ineq ← substitute_entities(
        rel.ineq, merged.head, merged.tail)
    consistent ← check_consistency(
        candidate_ineq, ineq)
    if consistent then
        insert(candidate_ineq, ineq)
        insert(rel, merged.relations)
    end if
end for
```

Mirzaee and Kordjamshidi (2022), reproduced in Table 3.

- For StepGame and SpaRC-PS3, we represent the relative positions as signed integers on the x and y axis, and numerically compose them (Algorithm 1). Without the quantitative knowledge of backtracking along a given axis, e.g. x-axis for {LEFT, RIGHT}, no subsequent inferences can be made for those directions.
- For SpaRC-PS4, the relations in context can be expressed as logical conjunction $\wedge$ of inequalities, refer Section 3, Table 2, and Figure 1. For composition of relations to merge reasoning steps, consistency of inequalities for relations $r \in \mathcal{D}$ is checked and the deduced relations set r^d is updated (Algorithm 2).

4. We finally *verbalize* the reasoning path $\mathcal{P}$ link-by-link (Figure 2) following the format of deductively verified chain-of-thought (Ling et al., 2023). However, instead of generating and self-verifying LLM outputs, we use spatial reasoners for ground truth generation.

We denote the extended dataset as **Spa**tial **R**easoning **P**aths (SpaRP). Specifically, we exended SpaRTUN, StepGame, SpaRC-PS3, and SpaRC-PS4, to be SpaRP-PS1, SpaRP-PS2, SpaRP-PS3 and SpaRP-PS4, respectively, by enriching the former with the reasoning paths. A comparison of the derived datasets with the original datasets is summarized in Table 4.

5 Experimental Setup

Dataset. Due to the expense and resource limitations for running LLMs, for each of the four subsets of SpaRP, we randomly sample 2000, 500, and 1000 datapoints as our training, validation, and test set, respectively. We call them small SpaRP, or **SpaRP-S**. We also randomly sample equal number of instances for each *number of hops* in the reasoning path. Additionally, we collect *five* diverse sets of *human-generated* natural language descriptions of the properties relevant to spatial compositions, and construct a *system prompt* template with a unified task instruction using these descriptions.

Implementation Details. To help replicability, we include implementation details such as dataset sampling, system prompt, and training parameters in Appendix-C.

Evaluation Metrics. We use exact-match accuracy and macro-averaged F1-scores[1].

6 Results and Analysis

We run experiments with three state-of-the-art LLMs — Llama-2-13B, Llama-2-70B (Touvron et al., 2023), and GPT-4, each with both single greedy decoding and self-consistency (Wang et al., 2023) with majority voting over 20 generations with sampling (SC=20). Inputs are provided with a "system prompt" containing task instructions and *5-shot* CoT with randomly sampled exemplars from the *relevant dev-set*. We also finetune Llama-2 13B and 70B models, indicated by FT in Table 5, using

[1]We used the scikit-learn v1.3.2 library

Dataset		Model	Acc.	F1
SpaRP-S-PS1	(SpaRTUN)	Llama-2-13B	0.2	0.49
		Llama-2-13B-FT	18.9	22.23
		Llama-2-70B	10.1	23.37
		Llama-2-70B-FT	28	36.49
		Llama-2-70B$_{SC=20}$	17.1	27.95
		GPT-4	46.8	54.30
		GPT-4$_{SC=20}$	**54.3**	**60.32**
		SOTA (PistaQ)	*94.52*	–
SpaRP-S-PS2	(StepGame)	Llama-2-13B	0.1	0.47
		Llama-2-13B-FT	13.7	33.23
		Llama-2-70B	10.6	26.41
		Llama-2-70B-FT	16.6	34.63
		Llama-2-70B$_{SC=20}$	20.30	38.96
		GPT-4	23.9	41.09
		GPT-4$_{SC=20}$	**28.6**	**43.01**
		SOTA (LLM-ASP)	*90.88*	–
SpaRP-S-PS3		Llama-2-13B	0.2	0.92
		Llama-2-13B-FT	27.3	32.01
		Llama-2-70B	9.4	25.27
		Llama-2-70B-FT	19.5	32.97
		Llama-2-70B$_{SC=20}$	15.2	32.01
		GPT-4	23.8	35.17
		GPT-4$_{SC=20}$	**32.5**	**42.06**
SpaRP-S-PS4		Llama-2-13B	0.7	1.84
		Llama-2-13B-FT	30.6	31.62
		Llama-2-70B	9.0	22.13
		Llama-2-70B-FT	20	31.74
		Llama-2-70B$_{SC=20}$	18.3	29.73
		GPT-4	21.7	33.02
		GPT-4$_{SC=20}$	**32.9**	**40.23**

Table 5: Performance evaluations of Llama-2 (13B and 70B) and GPT-4 models on the individual datasets.

QLoRA (Dettmers et al., 2023) on the *verbalized reasoning paths* made available by SpaRP.

Overall Results. As shown in Table 5, we can observe that the performance of all the state-of-the-art LLMs on the spatial reasoning datasets are low, lagging significantly behind the existing state-of-the-art symbolic-based models such as PistaQ (Mirzaee and Kordjamshidi, 2023) and LLM-ASP (Yang et al., 2023) on SpaRTUN and StepGame, respectively, suggesting that if these generalist models are to be used for any spatial-reasoning-related tasks (e.g., in LLMs-based agents), caution should be exerted.

Among these models, GPT-4 under SC=20 exhibits the best performance overall, followed closely by GPT-4 with greedy decoding. The latter outperforms even the largest open-source Llama-2-70B model with SC=20.

We also observed the emergent abilities of LLMs on spatial reasoning as model sizes scale up. The

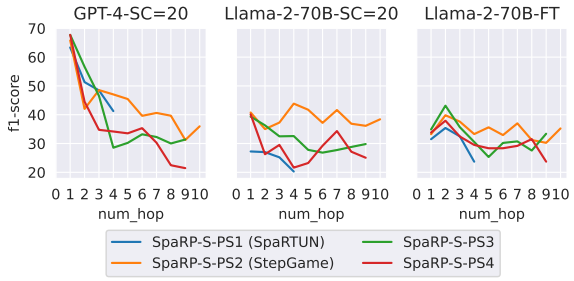

Figure 3: F1 scores vs. number of hops for spatial reasoning.

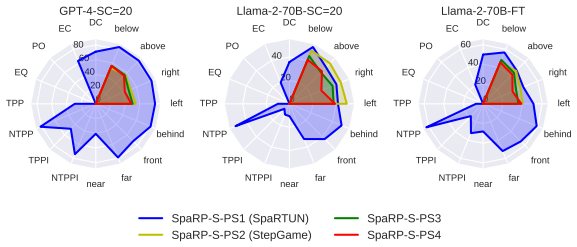

Figure 4: F1 scores of individual labels.

F1-score of Llama-2 13B model on SpaRP-S-PS1 (SpaRTUN) is only 0.49 (no spatial reasoning ability), significantly lower than the 23.37 F1-score of the 70B model. Similarly, for SpaRP-S-PS3, the F1-scores are 0.92 and 25.27 for the 13B and 70B models, respectively. This is also observed on the other datasets.

Impact of Spatial Properties and Composition Rules. StepGame and SpaRP-PS3 consider entities as point objects (PO), however, SpaRP-PS3 does not quantify directions rendering them not composable while backtracking, e.g. RIGHT followed by LEFT is not composable. SpaRP-PS4 considers entities as real objects with extended sizes, thereby introducing added complexity to spatial relation composition (Section 4 and Algorithm 2). The F1-scores (Table 5) of both GPT-4 and Llama-2 underscore these challenges.

Furthermore, Figure 3 demonstrates that the F1-scores for both SpaRP-S-PS3 and SpaRP-S-PS4 consistently trail those of SpaRP-S-PS2 (StepGame) across varying numbers of hops. This highlights the utility of our SpaRC framework in identifying additional challenges that are not addressed by the existing benchmarks.

Relation-wise Performance. The performance of GPT-4 is significantly better compared to Llama-2 models on SpaRP-S-PS1 (SpaRTUN), which

Errors, examples and explanations
Error: Incorrect relation extraction **Context: Box EEE** *has* a **tiny white rectangle** and covers a midsize white diamond. **Extracted:** *Step 7:* It is given that the **tiny white rectangle** is *inside and touching* the **box EEE**. **Explanation:** *Has* only means inside.
Error: Reverse answer **Question:** What is the relation of the agent **W** to the agent **X**? **Answer Step:** *Step 8:* From step 5 and 7, we can infer that **X** is above and left of **W**. *Hence, the answer is* above, and left. **Explanation:** Directional relations are non-symmetric. Question is from W to X, while answer is from X to W.
Error: Copied, not composed **Reasoning Steps:** *Step 6:* From step 4 and 5, we can infer that **A** is below and right of **S**. *Step 7:* From the context, **S** is left of **M**. *Step 8:* From step 6 and 7, we can infer that **A** is below and right of **M**. **Explanation:** Relation from S to M not used in composition. Instead, relation from A to S is copied in step 8.
Error: Composed without connection **Reasoning Steps:** *Step 5:* From step 3 and 4, we can infer that **Y** is right of **L**. *Step 14:* From step 12 and 13, we can infer that **C** is below and right of **K**. *Step 15:* From step 5 and 14, we can infer that **C** is below and right of **L**. **Explanation:** No common entity between merged steps 5 and 14 which are 9 steps apart.

Table 6: Errors, their examples (only relevant steps) and explanations in the model generated reasoning paths.

has a larger candidate set comprising of 16 relations, including 8 topological relations. In contrast, SpaRP-S-PS2 (StepGame) has a smaller candidate set consisting of only directional relations. This highlights a notable deficiency in Llama-2 regarding the understanding and composition of topological relations. More importantly, even the finetuned Llama-2 model falls short of GPT-4's performance. The top proprietary LLMs still significantly outperform their open-source counterparts in topological spatial reasoning.

Additionally, Figure 3 demonstrates that even when controlling for the same number of hops, the F1-scores of Llama-2 on SpaRP-S-PS1 (SPARTUN) rank lowest across all hops. An examination of F1-scores on a per-relation basis (Figure 4) further confirms this difficulty of topological relations for Llama-2 models compared to GPT-4.

Finetuning with Reasoning Paths. We observe that finetuning the 13B and 70B models with the reasoning paths made available in SpaRP consistently improves the spatial reasoning capabilities. Finetuning consistently boosts the F1-score by 21–32 and 7–13 points for 13B and 70B models respectively, across the datasets. In specific instances, the accuracy of a finetuned 13B model surpasses that of 5-10 times larger models such as Llama-2-70B with SC=20, and GPT-4. We hope the proposed reasoning-path generation can be further used for improving LLMs' explainability and robustness on spatial reasoning.

Error Analysis of Reasoning Paths. We sampled and manually analyzed a total of 80 model generated reasoning paths across all datasets for both the GPT-4 and Llama-2 70B models. The deductive step-by-step reasoning path made available by SpaRP proves to be useful in identifying errors in the generated outputs (Table 6). Commonly observed errors include incorrect parsing or retrieval of relations from the contexts, especially for topological relations. Additionally, we observe instances of reverse answering, where relations between tail to head entities are returned instead of head to tail entities in a question. More complex reasoning failures involve copying relations from one of the reasoning steps instead of composing them. Similarly, composing relations between reasoning steps without a common entity is observed frequently over distant steps. Additional errors with examples are provided in Appendix D. These errors are more prevalent in Llama-2 models, resulting in poorer performance compared to GPT-4.

7 Conclusion

Spatial reasoning is one of the basic components of intelligence. We perform a study on the spatial reasoning abilities of the latest LLMs under comprehensive setups. To support the study, we introduce (SpaRC), a systematic framework to characterize spatial reasoning scenarios by identifying and defining six spatial properties of objects, spatial relations, and contexts, and their impact on the spatial composition rules. Based on that, we create the (SpaRP) reasoning paths for the datasets. We found that the state-of-the-art LLMs do not perform well on the datasets — their performances are consistently low across different setups. The spatial reasoning is an emergent capability as model sizes scale up. Finetuning both large language models (e.g., Llama-2-70B) and smaller ones (e.g., Llama-2-13B) can significantly improve their performance by 7–32 points on F1-scores. We also found top proprietary LLMs still significantly outperform their open-source counterparts in topological spatial understanding and reasoning. We provide detailed analyses and insights in our experiments.

Limitations

We aimed to characterize various properties of the objects, relations, contexts and the associated spatial composition rules. We, however, note that the spatial scenarios, relations and interactions between objects can still be incomplete. Further, the existing datasets and our extensions of them still pertain to a limited combination of the characterizations *in isolation* in a context. Even with our proposed characterizations, a combination of these within a single context is common in the real world, including multi-modality with visual perception, which we haven't considered in our current study. The base datasets, although textual, are synthetic in nature. Combined with the use of symbolic reasoners for our reasoning path generation, our dataset inherit all the associated limitations such as relative lack of linguistic diversity, types of objects, relations etc. Finally we note that due to the cost and resource constraints of using LLMs, we worked with a smaller set of about 1000 test instances per dataset, which is a common data size to work with LLMs.

References

P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sunderhauf, I. Reid, S. Gould, and A. van den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3674–3683, Los Alamitos, CA, USA. IEEE Computer Society.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Yonatan Bisk, Deniz Yuret, and Daniel Marcu. 2016. Natural language communication with robots. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 751–761, San Diego, California. Association for Computational Linguistics.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12530–12539.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Geert-Jan M. Kruijff, Hendrik Zender, Patric Jensfelt, and Henrik I. Christensen. 2007. Situated dialogue and spatial organization: What, where... and why? *International Journal of Advanced Robotic Systems*, 4(1):16.

Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. 2023. Deductive verification of chain-of-thought reasoning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Roshanak Mirzaee and Parisa Kordjamshidi. 2022. Transfer learning with synthetic corpora for spatial role labeling and reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6148–6165, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Roshanak Mirzaee and Parisa Kordjamshidi. 2023. Disentangling extraction and reasoning in multi-hop spatial reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3379–3397, Singapore. Association for Computational Linguistics.

Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjamshidi. 2021. SPARTQA: A textual question answering benchmark for spatial reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4582–4598, Online. Association for Computational Linguistics.

Zhengxiang Shi, Qiang Zhang, and Aldo Lipani. 2022. Stepgame: A new benchmark for robust multi-hop spatial reasoning in texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11321–11329.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Takuma Udagawa and Akiko Aizawa. 2019. A natural language corpus of common grounding under continuous and partially-observable context. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Sagar Gubbi Venkatesh, Anirban Biswas, Raviteja Upadrashta, Vikram Srinivasan, Partha Talukdar, and Bharadwaj Amrutur. 2021. Spatial reasoning from natural language instructions for robot manipulation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11196–11202.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Jason Weston, Antoine Bordes, Sumit Chopra, and Tomás Mikolov. 2016. Towards ai-complete question answering: A set of prerequisite toy tasks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Zhun Yang, Adam Ishay, and Joohyung Lee. 2023. Coupling large language models with logic programming for robust and general reasoning from text. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5186–5219, Toronto, Canada. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yue Zhang and Parisa Kordjamshidi. 2022. Explicit object relation alignment for vision and language navigation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 322–331, Dublin, Ireland. Association for Computational Linguistics.

A Additional details and comparison of spatial properties in SpaRC

A symbolic context $\mathcal{C} = \{(h, r, t)_i\}_{i=1}^{N}$ is usually verbalized as several natural language sentences. However, we note that the *verbalization* can be a conjunction of multiple tuples in a single context sentence e.g. "Objects A and B are inside the box C", or "Entity X is below and left of entity Y". Such verbalization is common in existing benchmarks, including SpaRTUN and StepGame.

Fixed Orientation or Point of View (FPoV). The relations are considered to axis-aligned from a globally fixed orientation or point of view, i.e., fixed axes in a 2D or 3D space. We note that the cardinal ($\mathcal{D}_C$) and clock-face ($\mathcal{D}_T$) directions have only 4 relations in 2D. With the set of relative directions ($\mathcal{D}_R$) being larger (6 relations in 3D), $\mathcal{D}_C$ and $\mathcal{D}_T$ are mapped and canonicalized to four of the relative directions *only* for their label representations $\mathcal{L}$ (Table 1). Their understanding in the contexts and questions is still required. Additionally, the understanding of the map to a canonicalized label is also required to return correct answers.

Point Objects (PO) vs Extended Objects (EO). Point objects are entities that are either dimensionless i.e. their boundaries on all axes coincide, or can be treated as such in a given context. Since they are dimensionless, in relation to other point objects, the possible topological $\mathcal{T}_R$ relation (Table

1) collapses just to {DC, EQ} i.e. outside or "disconnected", and overlapping respectively. When combined with other formalisms such as directional relations ($\mathcal{D}$), even DC becomes redundant as the presence of any directional relation implies that the objects are not at the same position. Although point objects are purely mathematical constructs, real objects can often be treated as point objects in practical contexts. For example when the sizes of the objects can be ignored in relation to the distances between them. e.g. discussing spatial (directional) relations between buildings across several towns.

Extended Objects, on the other hand, are entities that are not dimensionless, i.e. their boundaries on at least one axis extends or has a spread. All real objects are extended objects. Dimensions of objects cannot be ignored when the distances between them are comparable to their sizes for spatial rule compositions and thus they must be treated as extended objects e.g. "a number of curious silver instruments" standing on Dumbledore's "spindle-legged tables".

Relation Incomplete (RI) vs Relation Complete (RC). For a set of relations $\mathcal{R}$, the contexts are usually relation incomplete in several real-world scenarios or when $|\mathcal{R}|$ is relatively large. On the other hand, the contexts can be relation complete in the real-world scenarios if $|\mathcal{R}|$ is relatively small, and one needs to emphasize and be specific.

Quantitatively Specified (QS) vs Quantitatively Unspecified (QU). For our current set of formalism (Table 1), some topological relations $r \in \mathcal{T} \setminus$ {EC, EQ, TPP, TPPI} and all the directional relations $r \in \mathcal{D}$ can be quantitatively specified. However, the topological relations are usually considered qualitatively, although there are metric based calculus for RCC8 and other topological relations as well. For example, context statements "Hogwarts is 200 miles to the left of the Azkaban Fortress" and "The Quidditch Stadium is inside and 1 KM away from the Hogwarts School's northern boundary" have LEFT and NTPP (inside) as quantitatively specified relations respectively. Quantitatively specified relations that are *reverse* of each other, such as LEFT and RIGHT, can readily be composed. For example, we can infer that Harry is 2 unit right of Ron, from the context statements – Harry is 3 unit left of Hermione, and Hermione is 5 unit right of Ron. Relations are quantitatively specified when their measurements are required in a context directly, or to infer other spatial relations indirectly.

On the other hand, for the previous examples, the context statements "Hogwarts is to the left of the Azkaban Fortress" and "The Quidditch Stadium is inside the Hogwarts School" are quantitatively unspecified for the relations LEFT and NTPP (inside) respectively. Quantitatively unspecified relations that are *reverse* of each other, such as LEFT and RIGHT, cannot be composed unless the relations are quantified. For example, directional relation between Harry and Ron cannot be determined from the context statements – Harry is left of Hermione, and Hermione is right of Ron.

Mutual Exclusitivity of Spatial Relations. While the *reverse* relations in any formalism cannot occur simultaneously, under RCC8 calculus, multiple topological relations $\mathcal{T}_R$ cannot occur simultaneously for the same *ordered* pair of entities even if they are not *reverse* of each others. Thus, for a given relation $r \in \mathcal{T}_R$ and an ordered pair of entities (X, Y):

$$r(X, Y) \implies \text{NOT}(r'(X, Y)) \; \forall r' \in \mathcal{T}_R \setminus r$$

For example, TPP (inside and touching) and NTPP (inside) are exclusive in RCC8. Stating a single topological relation in $\mathcal{T}_R$ makes the context Relation Complete (RC) in (*and only in*) $\mathcal{T}_R$.

However, negative implications are only for *reverse* relations in directional formalism $\mathcal{D}$. Orthogonal relations such as LEFT and ABOVE can be *simultaneously true* for a set of *ordered* pair of entities. As directional relations are not symmetric, we will always mean an ordered pair or sequence of entities while discussing them, unless stated otherwise. Hence, Relation Incomplete (RI) contexts can be quite common in terms of directional relations.

B Characterization of SpaRTUN and StepGame

Although the existing datasets, inlcuding SpaRTUN and StepGame, do not explicitly consider the spatial properties, their contexts and spatial composition rules conform to a set of these properties referenced in Section 3.1. StepGame considers entities in a completely abstract sense placed on a grid (Figure 5). They support only directional relations (including composites such as lower-left) and an overlap. Hence, objects can either be completely overlapping or completely separate. Their placement on the grid is also in terms of implicit unit of measurements. An overlap and unrestricted

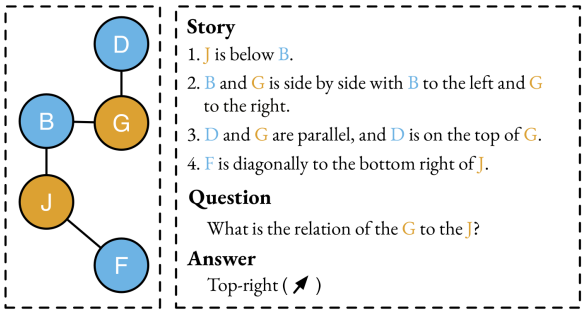

Figure 5: An example reproduced from the StepGame (Shi et al., 2022).

numerical composition of directions during their generation process coupled with the complete abstract representation of the entities essentially make them to be point objects (PO) and quantitatively specified (QS). Additionally their clear and complete expressions such as "BB is to the right of AA", "BB is at the 3 o'clock position relative to AA", and "AA and BB are horizontal and AA is to the right of BB" all considered equivalent means that when the relation is expressed as `RIGHT`, it means exactly and only `RIGHT` and this relation is completely known and they correspond to the relation complete (RC) context. This is why they support *strong* compositions for example presented at the beginning of Section 3 – "A is left of B and B is above C" $\implies$ "A is to the left and above C". Hence, the properties set of StepGame is {`PO, RC, QS`}.

SpaRTUN on the other hand considers objects that have shapes and sizes as they built their dataset on top of NLVR images and scene graphs with different sizes of objects and blocks, and support of topological relations such as containment, inside etc. Hence, their entities are extended objects (EO). Their spatial rules (Table 3) also do not consider quantitative relations either explicitly or implicitly. Finally their spatial rules also do not make any assumption about the alignment of directional relations to be exactly parallel to an axis system. That is why a statement such as "A is to the left of B" doesn't rule out the possibility of A additionally being above or below B i.e. the relations are not necessarily only as stated and other directional relations would still needs to be checked rather than assumed to be not present. This is in contrast with StepGame. Thus, the properties set of SpaRTUN is {`EO, RI, QU`}. This is why applying their spatial rules (Table 3) lead to no conclusion for the previous example "A is left of B and B is above C" presented at the beginning of Section 3. SpaRTUN composition rules are thus weaker in comparison to StepGame's composition based on these differences in their properties sets.

C Implementation Details

C.1 Datasets and Prompts

We created the **`SpaRP-S`** dataset with train, validation, and test splits of sizes 2000, 500, and 1000 respectively for each sub-dataset of `SpaRP`. To ensure a *fair* distribution of the difficulty level, we *randomly* sample *equal* number of instances for each *number of hops* (of entities) in the reasoning path. The final distribution is still skewed due to less number of instances for higher number of hops in SpaRTUN. Additionally, we collect *five* diverse sets of *human-generated* natural language descriptions of the properties (Table 7) relevant to spatial compositions (Section 3.1). We construct a *system prompt* template with a unified task instruction and populate it with randomly sampled natural language descriptions for each instances of each sub-dataset. The system prompt template and the human-generated descriptions are presented in Table 7 through Table 13.

C.2 Model configurations and training setup

To assess the spatial understanding and reasoning abilities of LLMs over varying model sizes, we run experiments with three model variants (all chat versions) – Llama-2-13B, Llama-2-70B, and GPT-4. The default GPT-4 i.e. GPT-4-0613 used in the experiments was accessed between Dec 1 2023 to Jan 31 2024.

We finetune a single model 13B and 70B models on all the four datasets i.e. `SpaRP-S-1` (SpaRTUN), `SpaRP-S-PS2` (StepGame), `SpaRP-S-PS3`, and `SpaRP-S-PS4`. For finetuning, we used QLoRA (Dettmers et al., 2023) with 8-bit quantization, LoRA $\alpha = 16$, and LoRA config $r = 64$. We trained for 3 epochs with a learning rate $lr = 1e^{-4}$, paged AdamW optimizer, cosine lr scheduler, and an *effective* batch size of 32 using gradient accumulation.

D Reasoning errors and their examples

We randomly sampled and manually analyzed 80 model generated reasoning paths to identify the errors and understand the discrepancy in the GPT-4 and Llama-2 70B models. A collection of several errors, their examples in terms of reasoning steps,

Terminology	Descriptions
System Instruction Template	You are an expert assistant with the knowledge of spatial relations and the rules to compose them under the assumptions that the contexts provided are of '{point_of_view_type}', the objects or entities are to be treated as '{entity_type}', the directions are '{quantitative_type}', and '{relation_type}'. The description of these terminologies are as given below: {point_of_view_type}: {point_of_view_type_desc}{point_of_view_type_default} {entity_type}: {entity_type_desc}{entity_type_default} {quantitative_type}: {quantitative_type_desc}{quantitative_type_default} {relation_type}: {relation_type_desc}{relation_type_default} You need to identify the sub-set of entities from the context that are relevant as well as combine their spatial relations with valid compositions under the above mentioned assumptions to find the spatial relations between the entities in the asked questions. The list of all possible spatial relations are: {spatial_relation_choices}. Always provide the final answer, only and only, in terms of these spatial relations. Include all the spatial relations that hold true as the answer, in case of multiple correct choices.
Fixed Orientation Point of View	The spatial relations are expressed from a single, consistent and unchanging perspective. This means that the observations are made from a global viewpoint that remains same and constant for all the entities in a given context. Hence, relations such as relative directions e.g. left or right always refer to the same directions and there is a one-to-one mapping between relative, cardinal and clock-face directions i.e. left is same as west or 9 o'clock position, right is same as east or 3 o'clock position, above is same as north or 12 o'clock position, and below is same as south or 6 o'clock position.
Implicit Quantification	Unless otherwise stated, consider the direction relations specified in the context to be of 1 unit distance. For example, the sentence, entity X is to the lower-left of entity Y means that the entity X is 1 unit to the left and 1 unit below the entity Y.

Table 7: Human-generated natural language descriptions for common terminologies, defaults and system instruction. Terms inside { } are placeholders that are further replaced with their language descriptions. For current work, point_of_view_type is always Fixed Orientation Point of View and the only default available is for quantitative_type = Quantitatively Specified (QS) with quantitative_type_default = Implicit Quantification. All other placeholders are replaced by randomly sampled descriptions from one of their 5 diverse human-generated descriptions presented in Table 8 through Table 13. The spatial_relation_choices are the relevant labels $\mathcal{L}$ from Table 1.

Diverse human-generated Descriptions for Point Objects (PO)
Description 1: Two objects can be treated as Point Objects in a given context for specifying their spatial relations if they are extremely small such that their sizes are immaterial, or if they are of similar or even varying shapes and sizes but are placed sufficiently far enough that their shapes and sizes can be ignored to state and compose spatial relations between them. This leads to a limitation on the spatial relations that can be specified between objects e.g. containment, but simpler relation compositions since shapes and sizes of the objects need not be considered. For example, a tea-cup and an apple on a table, or a school building and a warehouse that are miles away can be considered as point objects.
Description 2: While composing spatial relations between objects, they can be considered as Point Objects if they can be treated as dimensionless i.e. if (1) their sizes are so small that they can be neglected or (2) the size and shape of the objects are negligible compared to the great distance between the objects. Although this situation may prevent to express certain relations like containment, it provides simpler spatial relation statements and compositions over multiple objects, since the size and shape are not considered. For example, two balls on the basketball pitch or two buildings that are separated with 2 KM distance.
Description 3: Point Objects are small objects in a given context, whose sizes and shapes can be ignored. Thus, only their locations and orientations are considered when specifying spatial relations, leading to less number of relations and their simpler combinations over objects. A typical example of point objects can be buildings on a map or beads on a table.
Description 4: In this scenario, objects can be treated as Point Objects if they are extremely small or far apart to the extent that their shapes and sizes can be ignored. In such cases, certain spatial relationships, like containment, become inapplicable. Additionally, since the shapes and sizes of the objects are not important, relationship compositions can be simpler. For example, two cars that are miles apart can be considered as point objects.
Description 5: Entities can be treated as Point Objects when the distance between them relative to their sizes is either large or can be ignored. Therefore when providing spatial relations between them, a limited set of relations with simpler composition rules is possible. For example, when someone says, a cafe and a house that are far apart can be treated as point objects.

Table 8: Five diverse human-generated natural language descriptions of Point Objects (PO).

Diverse human-generated Descriptions for Extended Objects (EO)
Description 1: Two objects are to be treated as Extended Objects in a given context for specifying their spatial relations if their shapes and sizes in comparison to the distances between them can not be ignored to state and compose spatial relations between them. This leads to more number of possible spatial relations that can be specified between objects e.g. containment, but reduces the number while increasing the complexity of possible relation compositions, as the shapes and sizes of the objects can neither be assumed nor be discarded. For example, a tea-cup and a tube-light, or a table and a cupboard in a room are to be considered as extended objects. **Description 2:** If the distance between objects is comparable to the shapes and sizes of the objects while specifying the spatial relations, the objects are considered as Extended Objects i.e. they can't be treated as dimensionless and they have significant length, breadth or height in comparison to the distances between the objects in the context. Although this gives an opportunity to use more specific spatial relations like touching or containment, the complexity of compositions increases. A basket and an apple in it or two entities, X and Y, in a room can be given as examples. **Description 3:** Extended Objects refer to objects, whose shapes and sizes can affect the spatial relations that can be specified and the way they can be combined between objects. This leads to more number of relations and the combination of relations have to be minimal in the absence of the information about the shape and size of the objects. Examples of extended objects include buildings on a street or boxes in a room. **Description 4:** In this scenario, two objects are considered to be Extended Objects if their shapes and sizes, in comparison to the distances between them, cannot be ignored. In such cases, a larger set of spatial relations between objects can be specified, although the relation composition becomes more limited when the shapes and sizes of the objects are unknown compared to when this information is known. For example, a tea-cup and a lamp or a sofa and a TV in a room can be considered as extended objects. **Description 5:** Entities can be treated as Extended Objects if they have shapes and sizes which are not to be ignored in the context. Because of this, although a larger set of relations is possible between objects but the composition rules can become complex. For example, a cafe and a mall building can be treated as extended objects and the cafe can be a part of i.e. inside the mall building itself.

Table 9: Five diverse human-generated natural language descriptions of Extended Objects (EO).

Diverse human-generated Descriptions for Relation Incomplete (RI) contexts
Description 1: Not all set of possible spatial relations that hold true between two objects are stated while specifying the relations between those objects. Thus, there could be multiple possible spatial configurations that conform to the stated relations between the objects. For example, the statement, object A is to the left of object B, when considered as relation incomplete could mean that A may or may not be strictly only to the left of B, i.e. it can be either only to the left, or is to the left and above, or is to the left and below B. **Description 2:** Although some spatial relations between two objects exist, they might be overlooked while expressing the relations between those objects. Therefore, other valid configurations, which are compatible with the expression but not explicitly specified, may also exist. For instance, the relation incomplete expression, the entity X is to the left of the entity Y does not have to mean that X is to the left of Y and they are strictly aligned at the same time. The entity X can be both to the left and bottom (or above etc.) of the entity Y. **Description 3:** An incomplete spatial relationship corresponds to the insufficient information or context to decide the exact spatial relationship between objects, leading to ambiguation. In other words, there can be multiple valid spatial arrangements or layouts that hold true to each incomplete relation. For example, given the incomplete statement that box 'one' is in front of box 'two', it holds true for multiple arrangements such as box 'one' is to the right and front of box 'two', or box 'one' is to the left and front of box 'two'. **Description 4:** Relations are incomplete in the context statements if not all the spatial relationships that exist between two objects are stated. In such cases, multiple spatial outline or positioning of the objects are possible, without a single definitive truth. For example, consider the relationship - the fruit F is behind the object O in a 2D plane. Although O is in front of F, their relative position on the horizontal axis is incomplete, and hence, could be left, right or at the same place when considered horizontally. **Description 5:** The provided set of spatial relations between two objects may not be enough to communicate the complete spatial position between them. Therefore, for the provided spatial information between two objects more than one arrangement is possible. For example, a metal ball is hanging below a metal beam in the workshop - can mean various spatial positions such as the metal ball is below the beam, or additionally, it can be to the right or left and away from the beam in consideration.

Table 10: Five diverse human-generated natural language descriptions of Relation Incomplete (RI).

Diverse human-generated Descriptions for Relation Complete (RC) contexts
Description 1: All set of possible spatial relations that hold true between two objects are stated while specifying the relations between those objects. Hence, there is only one spatial configuration that conforms to the stated relations between the objects. For example, the statement, object A is to the left of object B, when considered as complete could only and only mean that A is to the left of B. **Description 2:** All existing spatial relations between two objects are included while expressing the relations. Therefore, there is one-to-one mapping between spatial configuations and expressed spatial relations between objects. For instance, a relation complete statement, the entity X is to the right of the entity Y means that X is to the right of Y and they are aligned. **Description 3:** Completely specified spatial relations refer to the complete sets of spatial relations that can be held as well as stated between objects. Thus, there can be only one valid spatial arrangement or layout that holds true for a relation complete statement. An example is that box 'one' is in front of box 'two' and they are in the same line that denotes front in a given fixed orientation for all. **Description 4:** Relations are complete in a setting, if all the spatial relationships between two objects are stated. In such cases, there is a single ground truth spatial outline or positioning of the objects. For example, consider the relationship - the fruit F is behind the object O in a 2D plane. This means that O is strictly and only in front of F and are aligned on the axis i.e. can be considered to be neither left nor right but at the same position on the horizontal axis. **Description 5:** The provided set of relations between two objects are enough to know the actual spatial position between them. Therefore, no additional information is needed to understand the actual position between two objects. For example, a metal ball is hanging below a metal beam in the workshop means that the ball is below the beam and not to its left or right.

Table 11: Five diverse human-generated natural language descriptions of Relation Complete (RC).

Diverse human-generated Descriptions for Quantitatively Specified (QS) relations
Description 1: Spatial relations, such as directions, specified between two objects are said to be Quantitatively Specified if those relations can have a unit of measurement and are also stated, implicitly or explicitly, in the specified context. The composition of such relations is always possible when all the object parameters and the relations between any two objects in a statement are completely known. For example, with constraints such as objects A, B and C are apples lying in a line and the relation specified are of 1 unit measurement when not mentioned explicitly, the quantitatively specified statements - B is 3 units to the left of A, and C is to the right of B - can lead to the only conclusion that A is 2 units to the right of C, or its inverse equivalent i.e. C is 2 units to the left of A. **Description 2:** Unit of measurements in spatial relations (e.g., directions) between two objects needs to be explicitly or implicitly specified for these relations to be called as Quantitatively Specified. The composition of such relations can be determined when all other object parameters and relations of two objects are given. For example, let entities X and Z be perfect round shaped balls. Let entity Y be a round basket with 10 unit radius and let centers of all objects are horizontally aligned. If X is 1 unit to the left of the center of Y and Z is 2 units to the left of X, then Z is inside the basket and 3 units to the left of the center of the basket Y. **Description 3:** Spatial relations are Quantitatively Specified when these relations are defined with a specific unit of measurement such as meters or miles. The relation compositions over objects become deterministic if all the other object parameters and the relationships between them are provided. For example, box 'one' is 3 units above box 'two' and they are in the same line can be easily used to determine relations with respect to a third box, say box 'three', if its position is also quantitatively specified with one of them. **Description 4:** Under this setting, spatial relations between two objects are said to be Quantitatively Specified if the relations have a unit of measurement and stated directly or indirectly in the context. In such cases, when all the object parameters and relations between any two objects in the statement are known, a deterministic relation composition is possible. For example, although there are limitations like having three apples (A1, A2, A3) arranged in a row, the statements - A2 is 2 units left of A1, and A3 is 1 unit right of A2 - provides enough information to determine the exact positions of A1 and A3 relative to each other. **Description 5:** If the quantitative value along with the measurement unit for a spatial relation is provided then those relations are said to be Quantitatively Specified. The measurements may be a default value that is understood in the context or is explicitly provided. The composition of these relations will result in a distinctly resolved relation. For example, in the sentence, the cafe is 2 blocks north of my house and the hospital is 1 block south of the cafe, it can be easily determined that the hospital is 1 block north of my house.

Table 12: Five diverse human-generated natural language descriptions of Quantitatively Specified (QS) relations.

Diverse human-generated Descriptions for Quantitatively Unspecified (QU) relations
Description 1: Spatial relations, such as directions, specified between two objects are said to be quantitatively unspecified if those relations can have a unit of measurement but are not stated in the specified context. The composition of such relations may not be possible even when all the object parameters and the relations between any two objects in a statement are completely known. For example, even with constraints such as objects A, B and C are apples lying in a line, the quantitatively unspecified statements - B is to the left of A, and C is to the right of B - can not lead to any conclusion regarding left, right, or overlapping relationship between A and C.
Description 2: In order for spatial relations between two objects to be considered as quantitatively unspecified, unit of measurement in these relations should not be specified. The exact composition or realization of such relations may not be determined even if the other object features and relations are completely known. For example, let entity X be in the basket Y of a known and stated size, and let entity Z be to the right of the entity X. It is not possible to infer whether entity Z is in the basket Y or not if its distance from X is quantitatively unspecified.
Description 3: Spatial relations are Quantitatively Unspecified when these relations are not defined in terms of specific units of measurement such as meters or miles. The relation compositions over objects can still not be determined even if all the other object parameters and the relationships between them are provided. For example, if box 'one' is above box 'two', it's not clear how far exactly box 'one' lies with respect to box 'two' and this will affect the conclusions to be drawn about relations with respect to other objects, say box 'three'.
Description 4: In this setting, spatial relations between two objects are Quantitatively Unspecified if the relations have a unit of measurement that is not specified in the context. In such cases, even when all the object parameters and relations between any two objects in the statement are known, a deterministic composition of relations may be impossible. In this scenario, although there are limitations like having three apples (A1, A2, A3) arranged in a row, the statements that lack specific quantities - A2 is on the left of A1, and A3 is on the right of A2 - do not provide enough information to determine the left, right, or overlapping positions of A1 and A3 relative to each other.
Description 5: If the quantitative value along with the measurement unit for a spatial relation is not provided then those relations are said to be quantitatively unspecified. The composition of these relations may not be enough to result in a distinctly resolved relation. For example, in the sentence, the cafe is to the north of my house and the hospital is to the south of the cafe, it can't be determined if the hospital is to the south or north of my house.

Table 13: Five diverse human-generated natural language descriptions of Quantitatively Unspecified (QU) relations.

the datasets to which the generated paths belong and the explanation of the errors are provided in Table 14.

Errors examples and explanations
Error: Reverse Step Merge **Dataset:** SpaRP-S-PS2(StepGame) **Reasoning Steps:** *Step 7:* From step 4 and 6, we can infer that **I** is 2 unit above and 2 unit right of **W**. *Step 9:* From step 8, we can infer that **Q** is right of **W**. *Step 10:* From step 7 and 9, we can infer that **I** is 2 unit above and 3 unit right of **Q**. **Explanation:** Requires composition from I to W to Q. Step 9 is a reverse step in this reasoning chain causing composition error.
Error: Copied, not composed **Dataset:** SpaRP-S-PS2 (StepGame) **Reasoning Steps:** *Step 6:* From step 4 and 5, we can infer that **A** is below and right of **S**. *Step 7:* From the context, **S** is left of **M**. *Step 8:* From step 6 and 7, we can infer that **A** is below and right of **M**. **Explanation:** The relation from S to M is not used in composition, instead, the relation from A to S is copied for the merge step A to S to M.
Error: Incorrect Composition **Dataset:** SpaRP-S-PS2 (StepGame) **Reasoning Steps:** *Step 11:* From step 8 and 10, we can infer that **Z** is above and 2 unit right of **L**. *Step 13:* From step 12, we can infer that **L** is above and left of **J**. *Step 14:* From step 11 and 13, we can infer that **Z** is 3 unit above and right of **J**. **Explanation:** Quantitatively incorrect.
Error: Merged without connection **Dataset:** SpaRP-S-PS3 **Reasoning Steps:** *Step 5:* From step 3 and 4, we can infer that **Y** is right of **L**. *Step 14:* From step 12 and 13, we can infer that **C** is below and right of **K**. *Step 15:* From step 5 and 14, we can infer that **C** is below and right of **L**. **Explanation:** No common entity between merged steps 5 and 14 which are 9 steps apart.
Error: Non-composable Composition **Dataset:** SpaRP-S-PS4 **Reasoning Steps:** *Step 16:* From step 13 and 15, we can infer that **M** is below **X**. *Step 18:* From step 17, we can say that **X** is above **A**. *Step 19:* From step 16 and 18, we can infer that **M** is above **A**. **Explanation:** The reverse relations {below, above} are not composable under quantitatively unspecified (QU) criteria.
Error: Reverse answer **Dataset:** SpaRP-S-PS3 **Question:** What is the relation of the agent **W** to the agent **X**? **Answer Step:** *Step 8:* From step 5 and 7, we can infer that **X** is above and left of **W**. *Hence, the answer is* above, and left. **Explanation:** Directional relations are non-symmetric. Question is from W to X, while answer is from X to W.
Error: Incorrect clock extraction **Dataset:**SpaRP-S-PS3 **Context: P** is above **X** at 10 o'clock. **Extracted:** *Step 6:* From the context, **X** is below **P**. **Explanation:** Correct extraction should be below and right.
Error: Incomplete relation extraction **Dataset:** SpaRP-S-PS4 **Context: I** is on the left side of and below **V**. **Extracted:** *Step 2:* From the context, I is left of V. **Explanation:** Incomplete or partial extraction, below wasn't picked.
Error: Incorrect relation extraction **Dataset:** SpaRP-S-PS1 (SPARTUN) **Context: Box EEE** *has* a **tiny white rectangle** and covers a midsize white diamond. **Extracted:** *Step 7:* It is given that the **tiny white rectangle** is *inside and touching* the **box EEE**. **Explanation:** *Has* only means inside.
Error: Incorrect reversal **Dataset:** SpaRP-S-PS1 (SPARTUN) **Reasoning Steps:** *Step 1:* It is given that the **block AAA** *contains* the **medium yellow square**. *Step 2:* From step 1, we can infer that the **medium yellow square** is *inside and touching* the **block AAA**. **Explanation:** Reverse of *contains* should only be inside.

Table 14: Observed errors and their examples in the model generated reasoning paths. Only the relevant steps are shown.