Advancing Spatial Reasoning in Large Language Models: An In-Depth Evaluation and Enhancement Using the StepGame Benchmark

Fangjun Li[1], David C. Hogg[1], Anthony G. Cohn[1,2]

[1]University of Leeds, UK [2]Alan Turing Institute, UK

Introduction

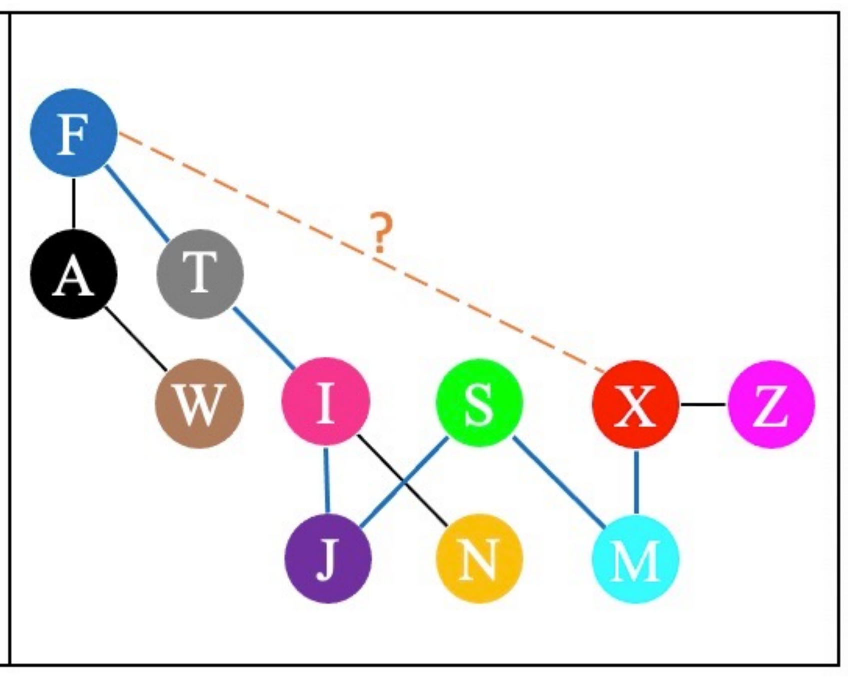
AI has made remarkable progress across various domains, with large language models (LLMs) like ChatGPT gaining substantial attention for their human-like text-generation capabilities. However, spatial reasoning remains a significant challenge, with ChatGPT's performance on spatial benchmarks like StepGame being unsatisfactory. Our analysis of GPT's spatial reasoning on a rectified StepGame benchmark identifies its proficiency in mapping text to spatial relations, yet it struggles with complex reasoning. We provide a flawless solution to the benchmark by combining template-to-relation mapping with logic-based reasoning. To address the limitations of GPT models in spatial reasoning, we deploy Chain-of-Thought (CoT) and Tree-of-Thoughts (ToT) prompting strategies, offering insights into GPT's "cognitive process", and achieving notable improvements in accuracy.

The StepGame Benchmark

Task: multi-hop spatial reasoning in texts

10- hop story:
1. T is slightly off center to the top left and I is slightly off center to the bottom right.
2. M is at the bottom of X.
3. Z presents right to X.
4. N is lower right of I.
5. S is positioned above J and to the right.
6. F is above T at 10 o'clock.
7. A is to the upper left of W.
8. A is at the bottom of F.
9. N is sitting in the right direction of J.
10. M is placed at the lower right of S.

What is the relation of the agent X to the agent F?

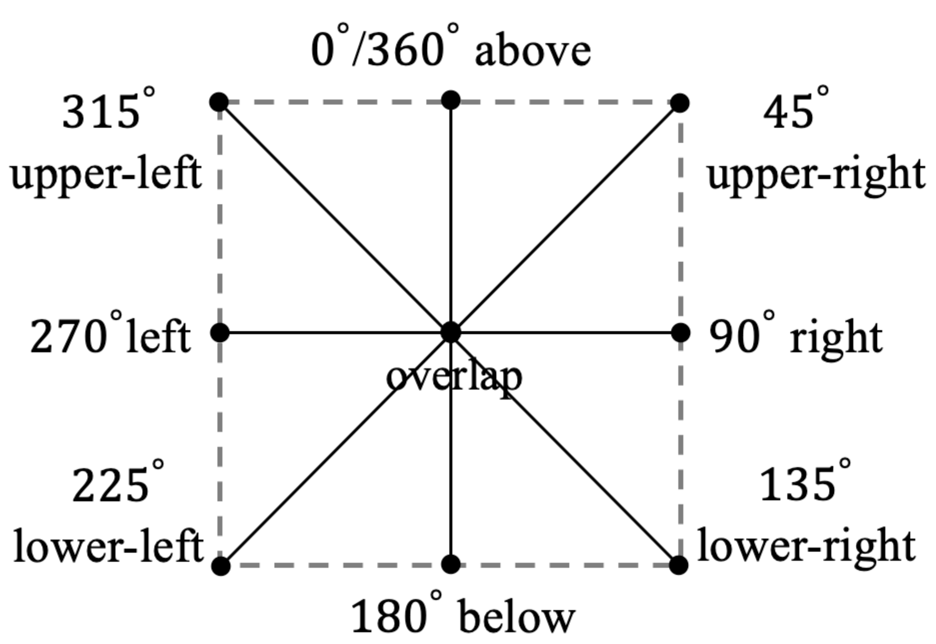


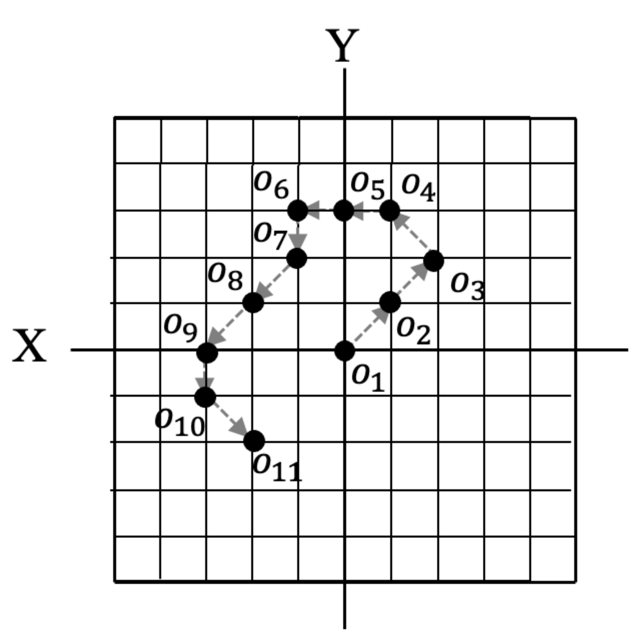
Solution to StepGame

Sentence-to-Relation Mapping + ASP Reasoner

Sentences	Template	ASP Facts
Y and I are parallel, and Y is on top of I. →	Y_above_I →	above("Y", "I").
F is on the left side of and below Q. →	F_lowerleft_Q →	down_left("F", "Q").
J is at O's 6 o'clock. →	J_below_O →	below("J", "O").
A is directly north east of B. →	A_upperright_B →	up_right("A", "B").
What is the relation of the agent B to the agent J?	query_B_J	query("B", "J").

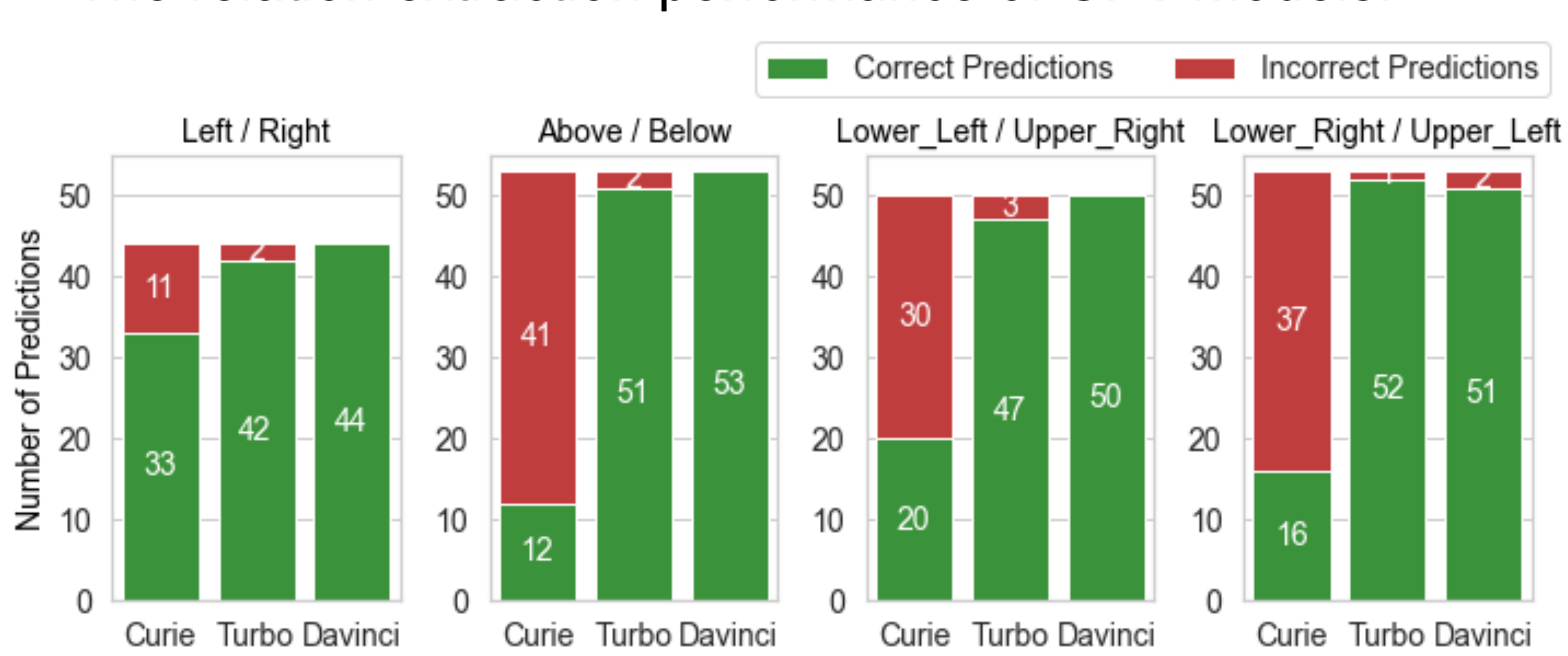
The ASP module calculates the location of o_i to o_j by adding the offsets $v(o_i, o_j)$.





LLM + ASP

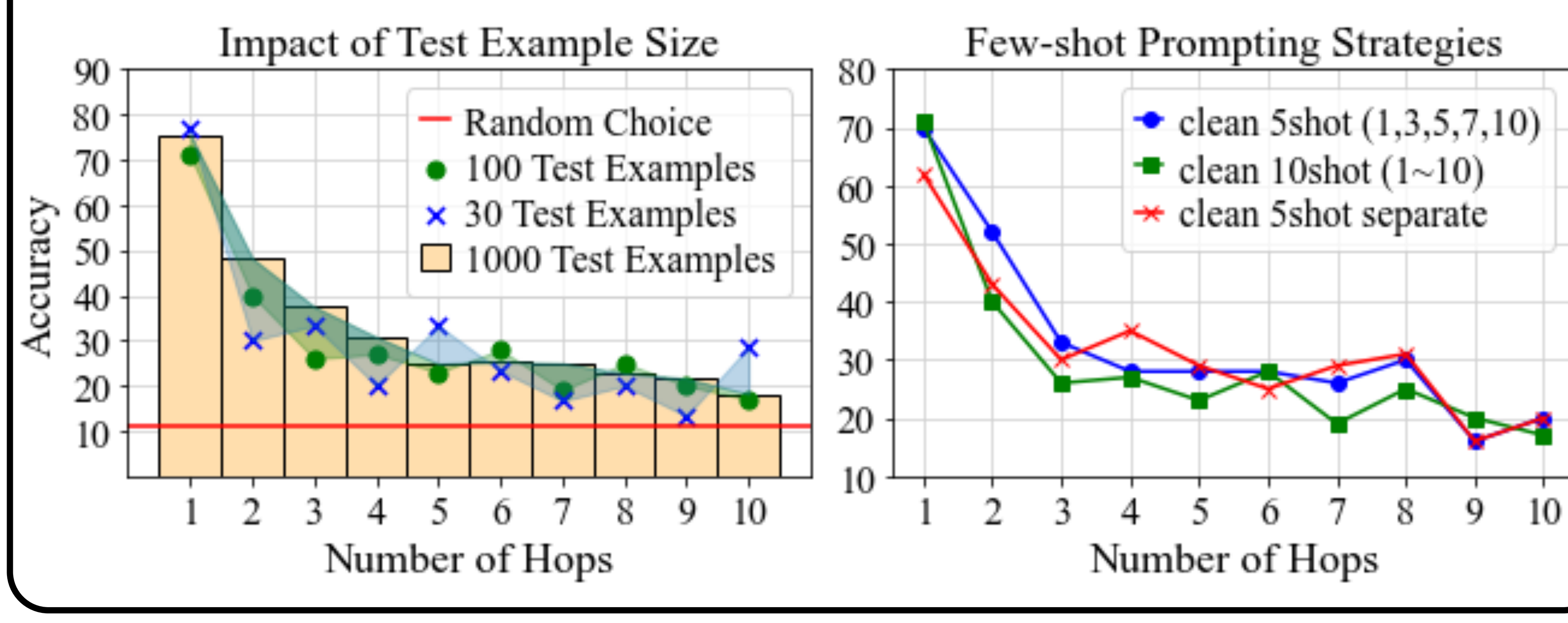
The relation extraction performance of GPT models.



Results of LLMs for relation extraction + ASP Reasoner

	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10
Map+ASP	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
Curie+ASP	46	43	42	59	67	67	57	56	58	61
Davinci+ASP	**100**	**100**	99	**100**	**100**	99	**100**	**100**	**100**	**100**
SOTA	92.6	89.9	89.1	93.8	92.9	91.6	91.2	90.4	89.0	88.3

Evaluation of GPT-3.5 Turbo on StepGame

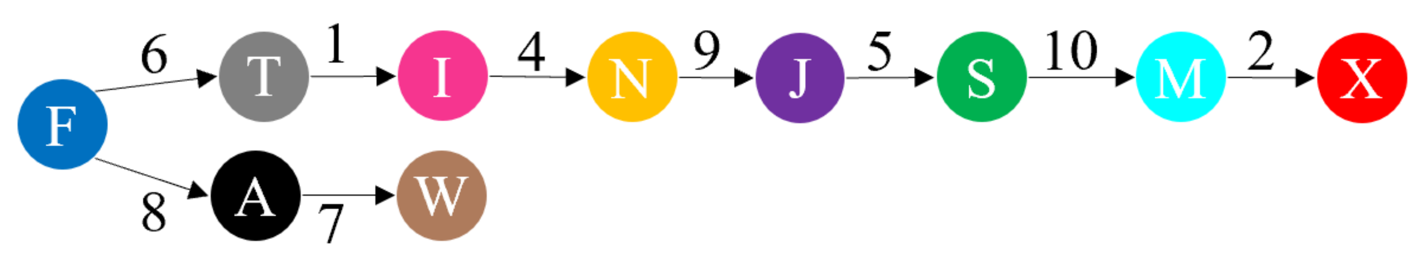


Methods

Our CoT approach decomposes each step of thought c_i to incorporate a coherent and detailed reasoning process.

At reasoning step i, $\boldsymbol{c}_i = [\boldsymbol{c}_i^{link}, \boldsymbol{c}_i^{map}, \boldsymbol{c}_i^{calcu}]$

- $\boldsymbol{c}_i^{link}$: guide LLMs to examine all relations in story ($R = [r^1, \ldots, r^j, \ldots, r^k]$) and select candidate r^j for each i
- $\boldsymbol{c}_i^{map}$: map r^j to simple relation description o_i is to the v of o_{i+1}
- $\boldsymbol{c}_i^{calcu}$: calculate the coordinate of o_{i+1} with r^j, $o_{i+1} = o_i + v(r^j) = (x_{o_i}, y_{o_i}) + (x_v, y_v) = (x_{o_{i+1}}, y_{o_{i+1}})$



Our ToT approach is designed to enhance the chain building process, allowing LLMs to consider different pathways.

Require: LLM, input x
$S_0 \leftarrow Init(x)$
$i \leftarrow 1$
while no $s_f \in S_{i-1}$ has arrived at o_t **do**
  $S_i' \leftarrow \{s \cdot c | c \in G(s, j) \wedge ChainExtn(c) \wedge s \in S_{i-1}\}$
  if $S_i' = \emptyset$ **then return** failure
  $S_i \leftarrow select(b, \{\langle s, y\rangle | s \in S_i' \wedge y = \sum_1^n \sigma(V(s))\})$
  $i = i + 1$
end while
return $Link(s_f)$

Results - Accuracy

		k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10
Turbo	base	62	43	30	35	29	25	29	31	16	20
	CoT	/	34	40	36	28	28	26	31	25	24
	ToT	/	/	35	35	25	45	15	40	40	35
Davinci	base	77	42	21	26	25	30	23	23	22	22
	CoT	/	48	53	46	46	48	40	45	41	32
	ToT	/	/	65	50	45	60	50	50	55	50
GPT-4	base	**100**	70	55	45	40	25	40	35	35	25
	CoT	/	**80**	75	**95**	85	85	**90**	80	60	65
	ToT	/	/	**85**	85	**90**	**90**	85	**90**	**100**	**95**

Advancing Spatial Reasoning in Large Language Models: An In-Depth Evaluation and Enhancement Using the StepGame Benchmark

Fangjun Li[1], David C. Hogg[1], Anthony G. Cohn[1,2]
[1]University of Leeds, UK [2]Alan Turing Institute, UK

Introduction

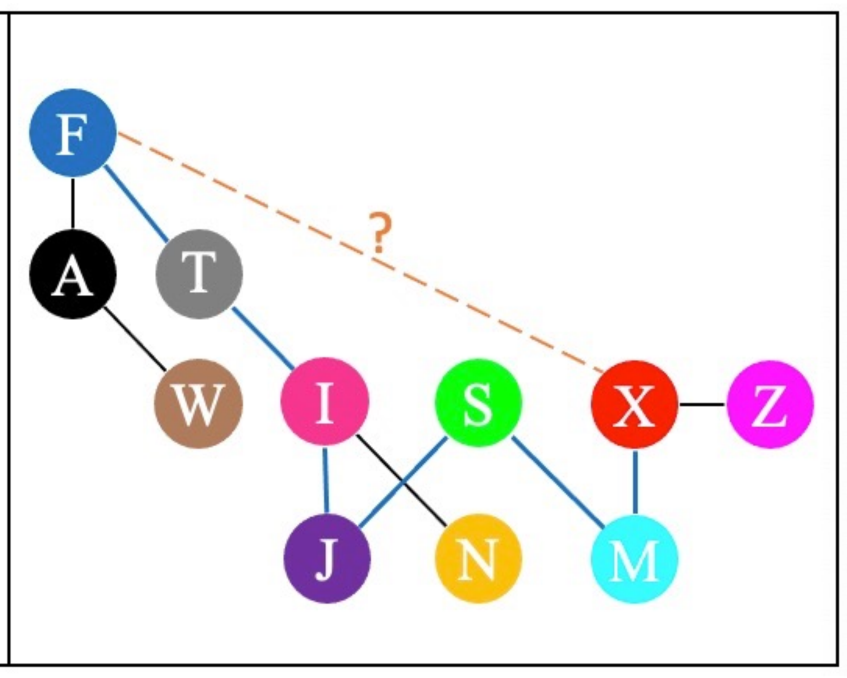
AI has made remarkable progress across various domains, with large language models (LLMs) like ChatGPT gaining substantial attention for their human-like text-generation capabilities. However, spatial reasoning remains a significant challenge, with ChatGPT's performance on spatial benchmarks like StepGame being unsatisfactory. Our analysis of GPT's spatial reasoning on the rectified StepGame benchmark identifies its proficiency in mapping text to spatial relations, yet it struggles with complex reasoning. We provide a flawless solution to the benchmark by combining template-to-relation mapping with logic-based reasoning. To address the limitations of GPT models in spatial reasoning, we deploy Chain-of-Thought (CoT) and Tree-of-Thoughts (ToT) prompting strategies, offering insights into GPT's "cognitive process", and achieving notable improvements in accuracy.

The StepGame Benchmark

Task: multi-hop spatial reasoning in texts

10- hop story:
1. T is slightly off center to the top left and I is slightly off center to the bottom right.
2. M is at the bottom of X.
3. Z presents right to X.
4. N is lower right of I.
5. S is positioned above J and to the right.
6. F is above T at 10 o'clock.
7. A is to the upper left of W.
8. A is at the bottom of F.
9. N is sitting in the right direction of J.
10. M is placed at the lower right of S.

What is the relation of the agent X to the agent F?

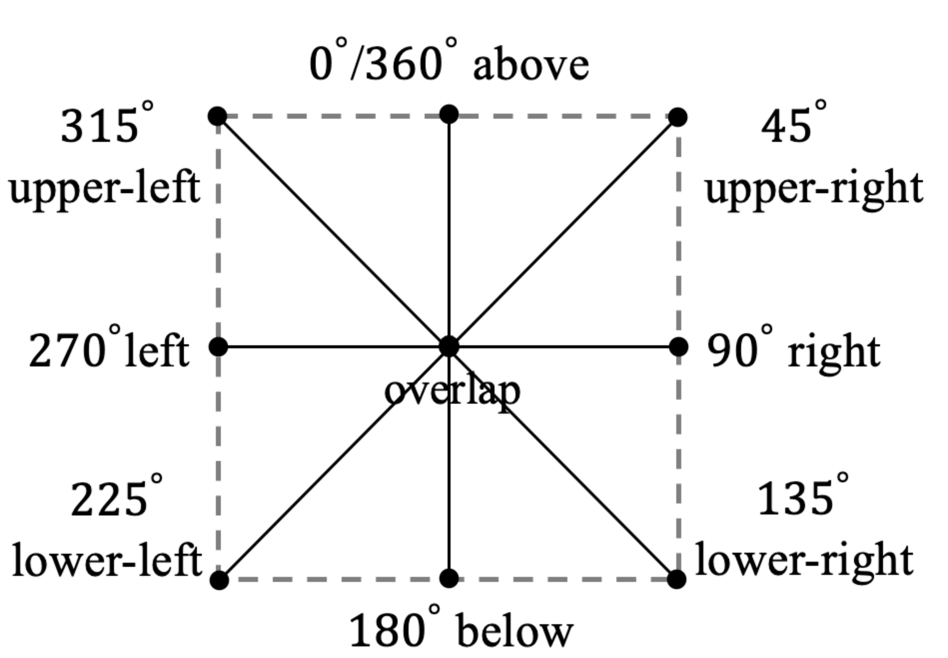


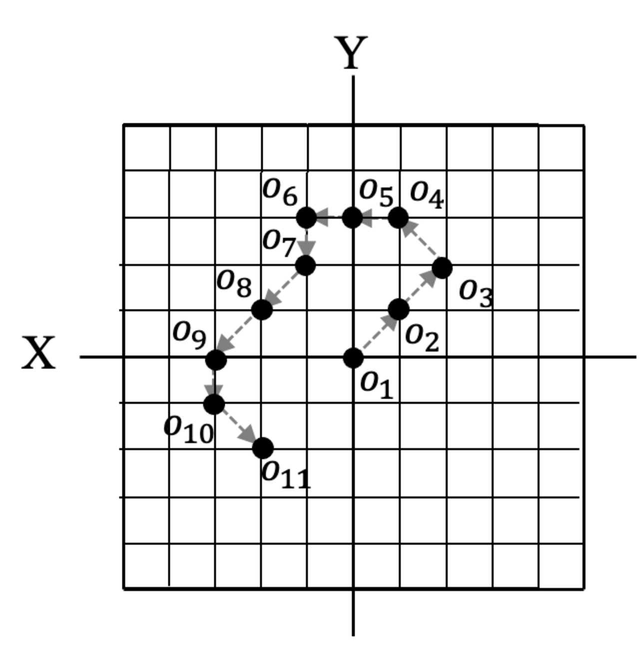
Solution to StepGame

Sentence-to-Relation Mapping + ASP Reasoner

Sentences	Template	ASP Facts
Y and I are parallel, and Y is on top of I. →	Y_above_I →	above("Y", "I").
F is on the left side of and below Q. →	F_lowerleft_Q →	down_left("F", "Q").
J is at O's 6 o'clock. →	J_below_O →	below("J", "O").
A is directly north east of B. →	A_upperright_B →	up_right("A", "B").
What is the relation of the agent B to the agent J? →	query_B_J →	query("B", "J").

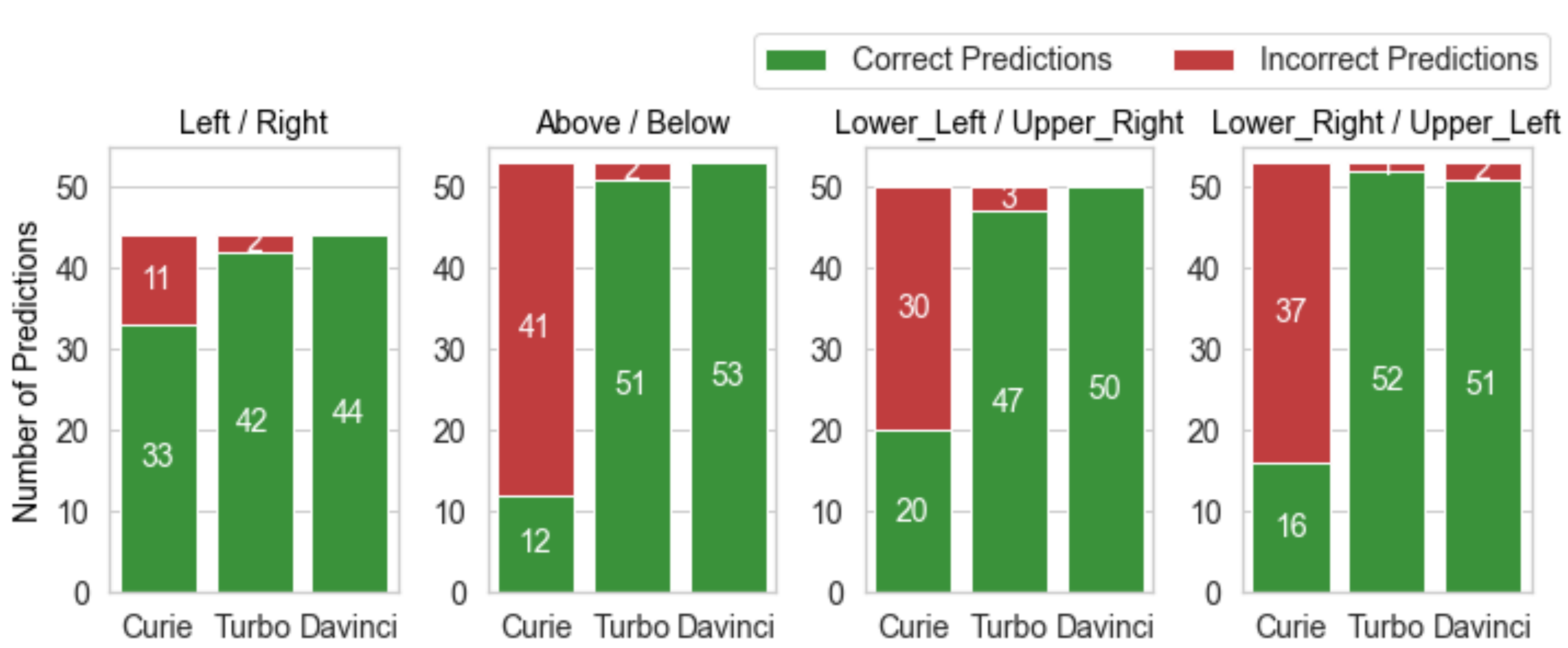
The ASP module calculates the location of o_i to o_j by adding the offsets $v(o_i, o_j)$.





LLM + ASP

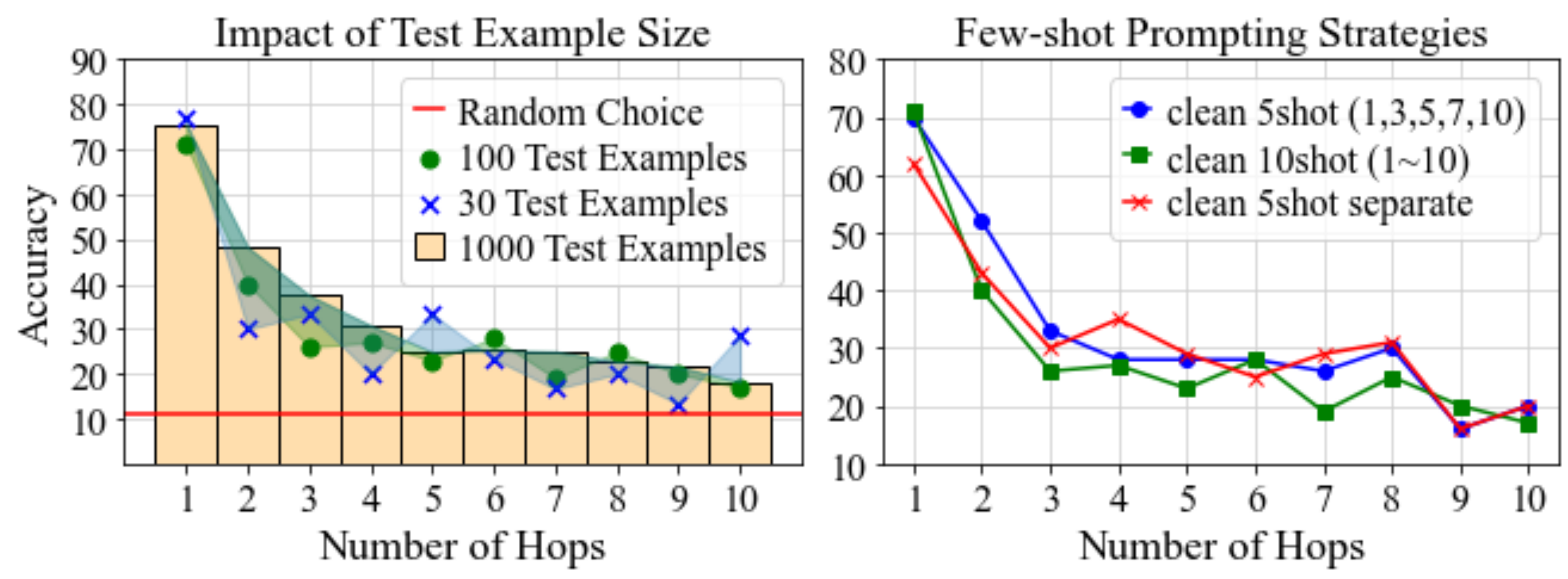
The relation extraction performance of GPT models.



Accuracy results of LLMs for relation extraction + ASP Reasoner

	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10
Map+ASP	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
Curie+ASP	46	43	42	59	67	67	57	56	58	61
Davinci+ASP	**100**	**100**	99	**100**	**100**	99	**100**	**100**	**100**	**100**
SOTA	92.6	89.9	89.1	93.8	92.9	91.6	91.2	90.4	89.0	88.3

Evaluation of GPT Models on Rectified StepGame



Methods

Our CoT approach decomposes each step of thought c_i to incorporate a coherent and detailed reasoning process.

At reasoning step i, $\boldsymbol{c_i} = [\boldsymbol{c_i^{link}}, \boldsymbol{c_i^{map}}, \boldsymbol{c_i^{calcu}}]$

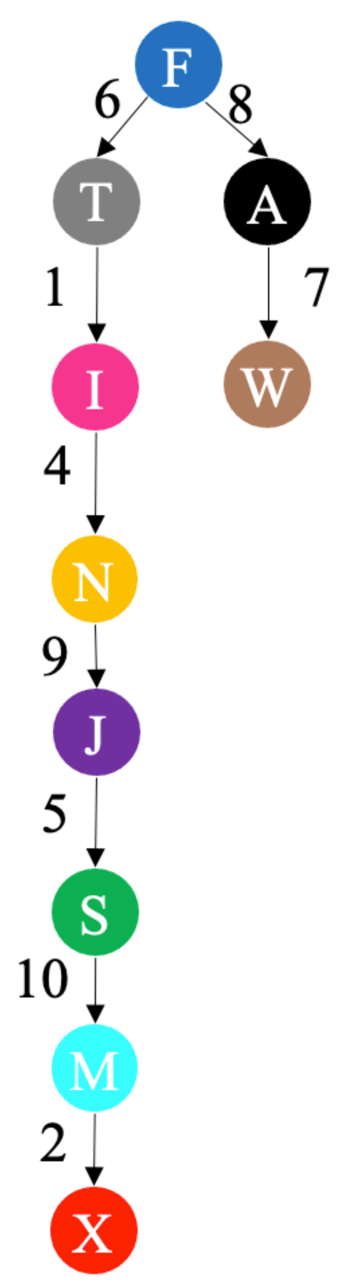
- $\boldsymbol{c_i^{link}}$: guide LLMs to examine all relations in story ($R = [r^1, \ldots, r^j, \ldots, r^k]$) and select candidate r^j for each i;
- $\boldsymbol{c_i^{map}}$: map r^j to simple relation description "o_i is to the v of o_{i+1}";
- $\boldsymbol{c_i^{calcu}}$: calculate the coordinate of o_{i+1} with r^j, $o_{i+1} = o_i + v(r^j) = (x_{o_i}, y_{o_i}) + (x_v, y_v) = (x_{o_{i+1}}, y_{o_{i+1}})$

Our ToT approach is designed to enhance the reasoning chain building process, allowing LLMs to consider different pathways.

Require: LLM, input x

$S_0 \leftarrow Init(x)$
$i \leftarrow 1$
while no $s_f \in S_{i-1}$ has arrived at o_t **do**
  $S_i' \leftarrow \{s \cdot c | c \in G(s,j) \wedge ChainExtn(c) \wedge s \in S_{i-1}\}$
  if $S_i' = \emptyset$ **then return** failure
  $S_i \leftarrow select(b, \{\langle s, y\rangle | s \in S_i' \wedge y = \Sigma_1^n \sigma(V(s))\})$
  $i = i + 1$
end while
return $Link(s_f)$

Link X and F



Results

Accuracy comparison of GPT models

		k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10
Turbo	base	62	43	30	35	29	25	29	31	16	20
	CoT	/	34	40	36	28	28	26	31	25	24
	ToT_CoT	/	/	35	35	25	45	15	40	40	35
Davinci	base	77	42	21	26	25	30	23	23	22	22
	CoT	/	48	53	46	46	48	40	45	41	32
	ToT_CoT	/	/	65	50	45	60	50	50	55	50
GPT-4	base	**100**	70	55	45	40	25	40	35	35	25
	CoT	/	**80**	75	**95**	85	85	**90**	80	60	65
	ToT_CoT	/	/	**85**	85	**90**	**90**	85	**90**	**100**	**95**

UNIVERSITY OF LEEDS

Advancing Spatial Reasoning in Large Language Models: An In-Depth Evaluation and Enhancement Using the StepGame Benchmark

Fangjun Li[1], David C. Hogg[1], Anthony G. Cohn[1,2]
[1]University of Leeds, UK [2]Alan Turing Institute, UK

Introduction

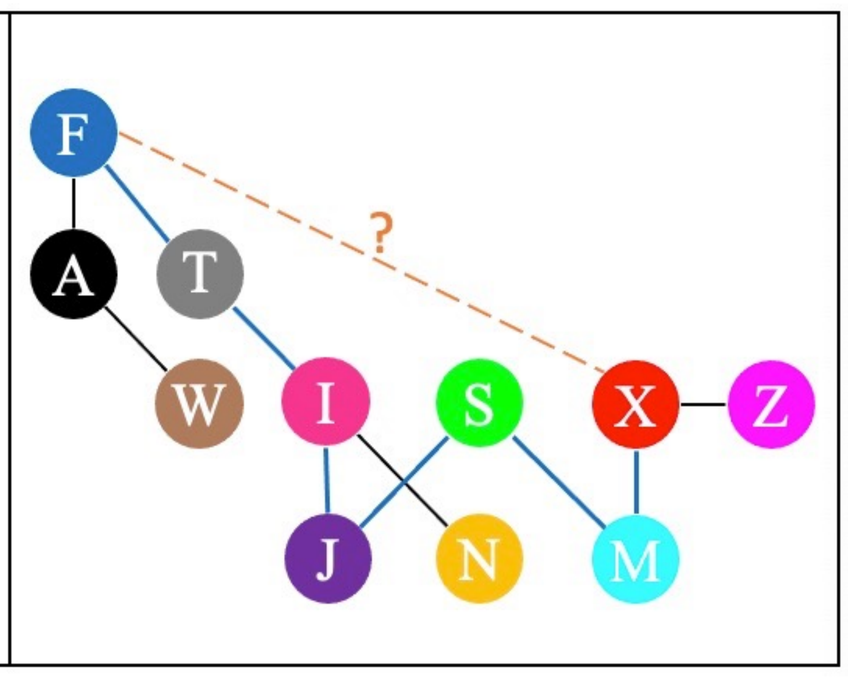
AI has made remarkable progress across various domains, with large language models (LLMs) like ChatGPT gaining substantial attention for their human-like text-generation capabilities. However, spatial reasoning remains a significant challenge, with ChatGPT's performance on spatial benchmarks like StepGame being unsatisfactory. Our analysis of GPT's spatial reasoning on the rectified StepGame benchmark identifies its proficiency in mapping text to spatial relations, yet it struggles with complex reasoning. We provide a flawless solution to the benchmark by combining template-to-relation mapping with logic-based reasoning. To address the limitations of GPT models in spatial reasoning, we deploy Chain-of-Thought (CoT) and Tree-of-Thoughts (ToT) prompting strategies, offering insights into GPT's "cognitive process", and achieving notable improvements in accuracy.

The StepGame Benchmark

Task: multi-hop spatial reasoning in texts

10- hop story:
1. T is slightly off center to the top left and I is slightly off center to the bottom right.
2. M is at the bottom of X.
3. Z presents right to X.
4. N is lower right of I.
5. S is positioned above J and to the right.
6. F is above T at 10 o'clock.
7. A is to the upper left of W.
8. A is at the bottom of F.
9. N is sitting in the right direction of J.
10. M is placed at the lower right of S.

What is the relation of the agent X to the agent F?

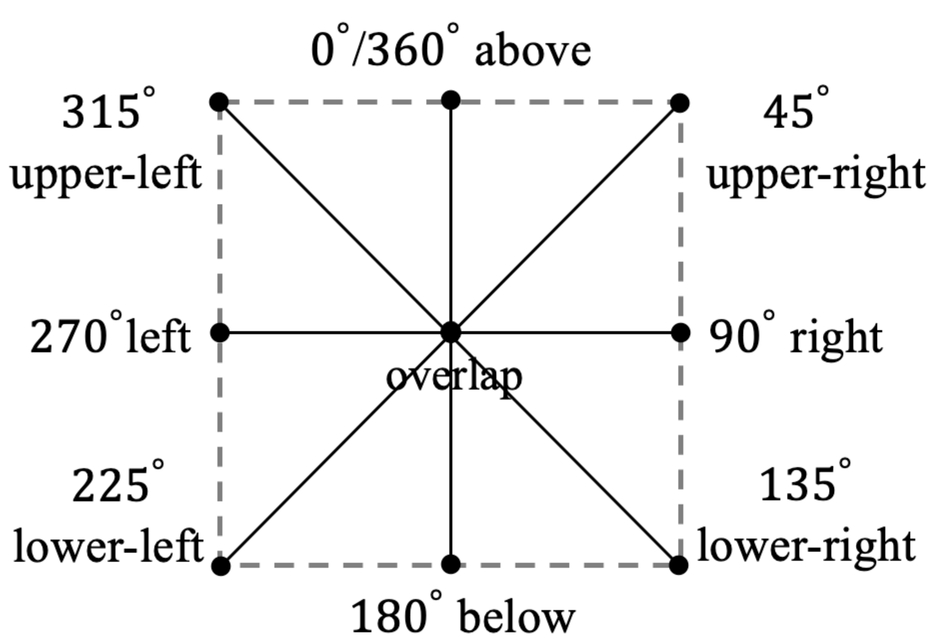


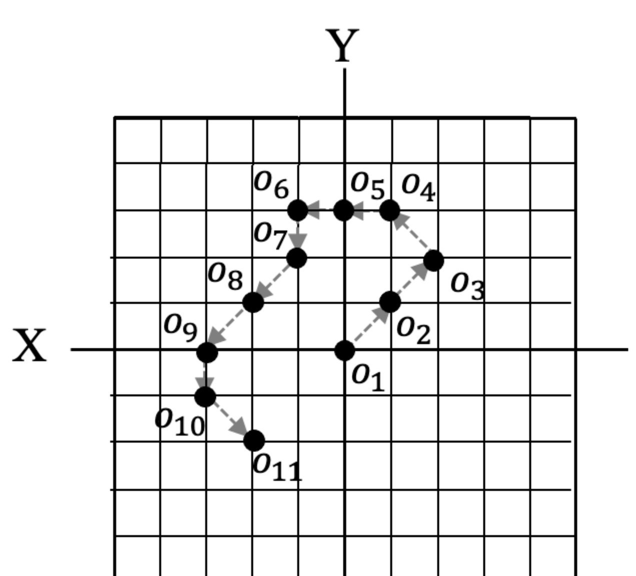
Solution to StepGame

Sentence-to-Relation Mapping + ASP Reasoner

Sentences	Template	ASP Facts
Y and I are parallel, and Y is on top of I.	Y_above_I	above("Y", "I").
F is on the left side of and below Q.	F_lowerleft_Q	down_left("F", "Q").
J is at O's 6 o'clock.	J_below_O	below("J", "O").
A is directly north east of B.	A_upperright_B	up_right("A", "B").
What is the relation of the agent B to the agent J?	query_B_J	query("B", "J").

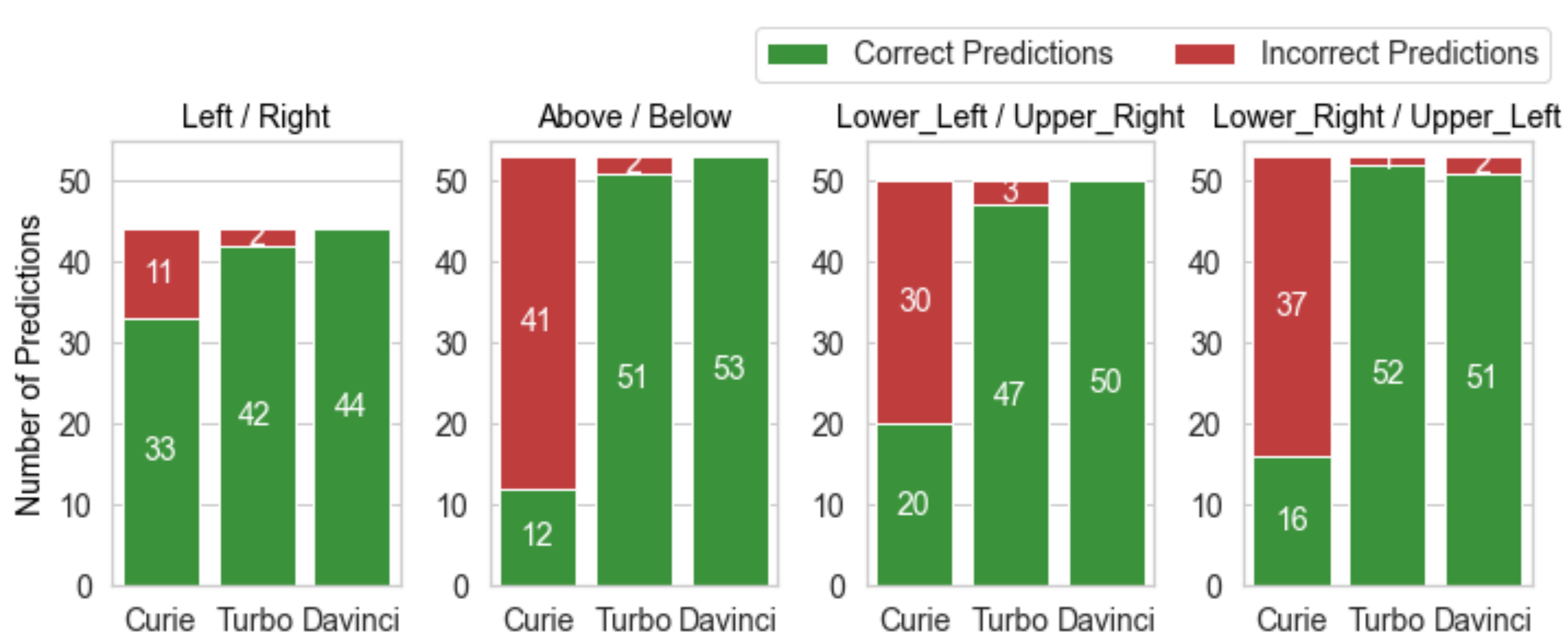
The ASP module calculates the location of o_i to o_j by adding the offsets $v(o_i, o_j)$.





LLM + ASP

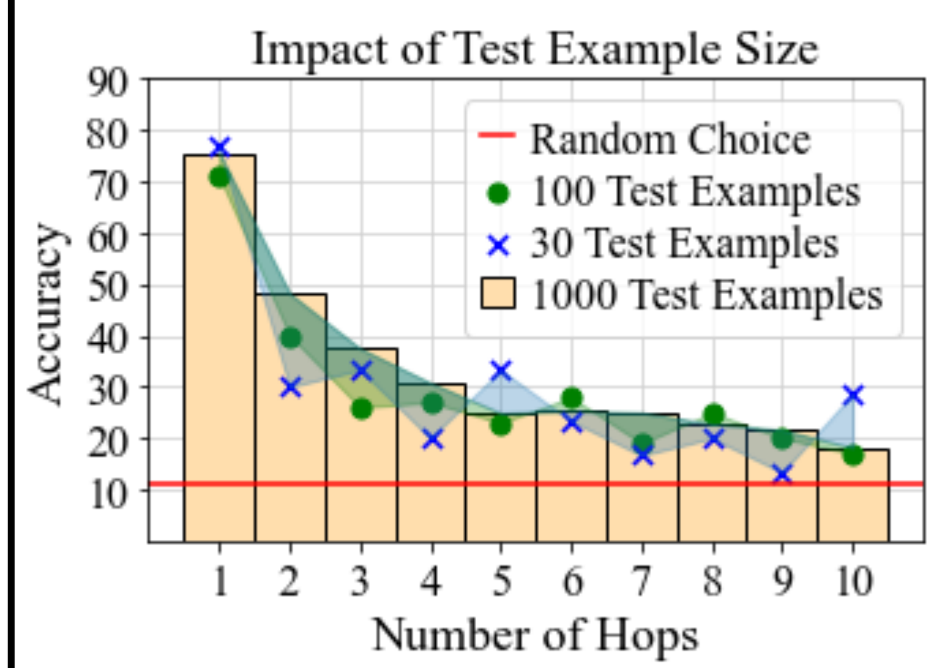
The relation extraction performance of GPT models.

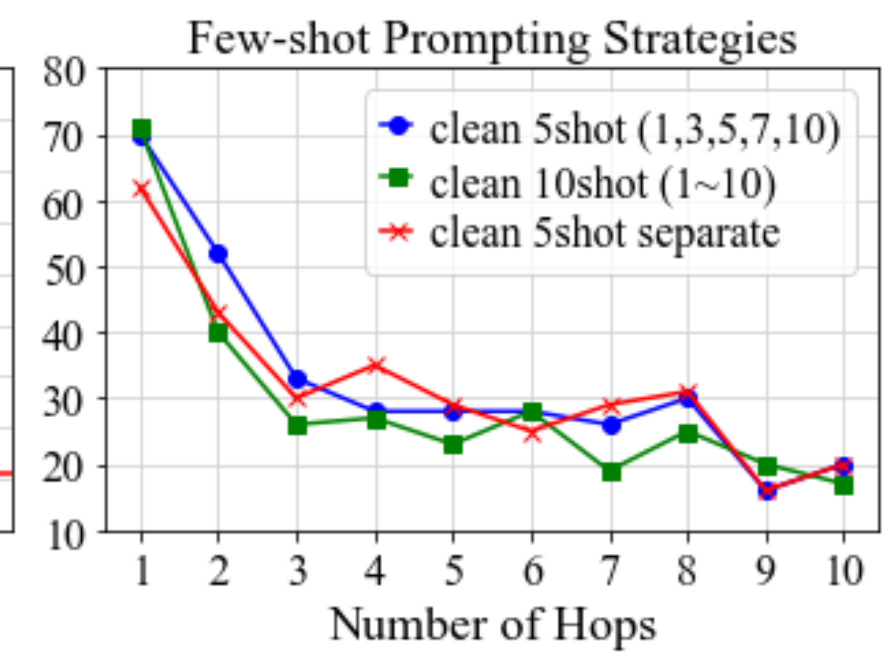


Results of LLMs for relation extraction + ASP Reasoner

	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10
Map+ASP	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
Curie+ASP	46	43	42	59	67	67	57	56	58	61
Davinci+ASP	**100**	**100**	99	**100**	**100**	99	**100**	**100**	**100**	**100**
SOTA	92.6	89.9	89.1	93.8	92.9	91.6	91.2	90.4	89.0	88.3

Evaluation of GPT Models on Rectified StepGame





Methods

Our CoT approach decomposes each step of thought c_i to incorporate a coherent and detailed reasoning process.

At reasoning step i, $\boldsymbol{c_i} = [\boldsymbol{c_i^{link}}, \boldsymbol{c_i^{map}}, \boldsymbol{c_i^{calcu}}]$

- $\boldsymbol{c_i^{link}}$: guide LLMs to examine all relations in story ($R = [r^1, \dots, r^j, \dots, r^k]$) and select candidate r^j for each i
- $\boldsymbol{c_i^{map}}$: map r^j to simple relation description o_i is to the v of o_{i+1}
- $\boldsymbol{c_i^{calcu}}$: calculate the coordinate of o_{i+1} with r^j, $o_{i+1} = o_i + v(r^j) = (x_{o_i}, y_{o_i}) + (x_v, y_v) = (x_{o_{i+1}}, y_{o_{i+1}})$

Our ToT approach is designed to enhance the reasoning chain building process, allowing LLMs to consider different pathways.

Require: LLM, input x
1: $S_0 \leftarrow Init(x)$
2: $i \leftarrow 1$
3: **while** no $s_f \in S_{i-1}$ has arrived at o_t **do**
4: $S_i' \leftarrow \{s \cdot c | c \in G(s,j) \wedge ChainExtn(c) \wedge s \in S_{i-1}\}$
5: **if** $S_i' = \emptyset$ **then return** failure
6: $S_i \leftarrow select(b, \{\langle s,y \rangle | s \in S_i' \wedge y = \sum_1^n \sigma(V(s))\})$
7: $i = i + 1$
8: **end while**
9: **return** $Link(s_f)$

Results

Accuracy comparison of GPT models

		k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10
Turbo	base	62	43	30	35	29	25	29	31	16	20
	CoT	/	34	40	36	28	28	26	31	25	24
	ToT	/	/	35	35	25	45	15	40	40	35
Davinci	base	77	42	21	26	25	30	23	23	22	22
	CoT	/	48	53	46	46	48	40	45	41	32
	ToT	/	/	65	50	45	60	50	50	55	50
GPT-4	base	**100**	70	55	45	40	25	40	35	35	25
	CoT	/	**80**	75	**95**	85	85	**90**	80	60	65
	ToT	/	/	**85**	85	**90**	**90**	85	**90**	**100**	**95**

Advancing Spatial Reasoning in Large Language Models: An In-Depth Evaluation and Enhancement Using the StepGame Benchmark

Fangjun Li[1], David C. Hogg[1], Anthony G. Cohn[1,2]

[1]University of Leeds, UK [2]Alan Turing Institute, UK

Introduction

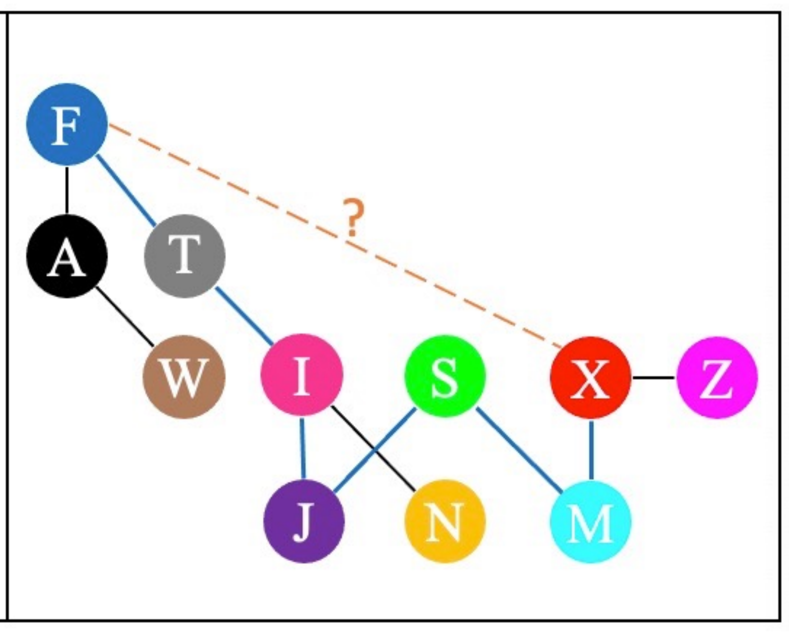
AI has made remarkable progress across various domains, with large language models like ChatGPT gaining substantial attention for their human-like text-generation capabilities. Despite these achievements, spatial reasoning remains a significant challenge for these models. Benchmarks like StepGame evaluate AI spatial reasoning, where ChatGPT has shown unsatisfactory performance. However, the presence of template errors in the benchmark has an impact on the evaluation results. Thus there is potential for ChatGPT to perform better if these template errors are addressed, leading to more accurate assessments of its spatial reasoning capabilities.

The StepGame Benchmark

Task: multi-hop spatial reasoning in texts

10- hop story:
1. T is slightly off center to the top left and I is slightly off center to the bottom right.
2. M is at the bottom of X.
3. Z presents right to X.
4. N is lower right of I.
5. S is positioned above J and to the right.
6. F is above T at 10 o'clock.
7. A is to the upper left of W.
8. A is at the bottom of F.
9. N is sitting in the right direction of J.
10. M is placed at the lower right of S.

What is the relation of the agent X to the agent F?



Template Errors in StepGame

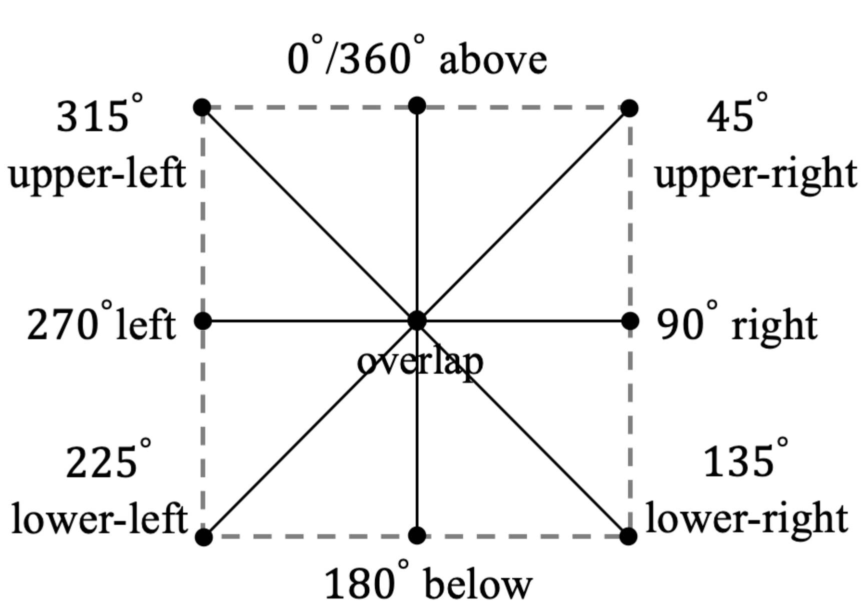
Mapping	Original Incorrect Statement
BB_right_AA	AA and BB are parallel, and AA on the right of BB. AA and BB are parallel, and AA is to the right of BB. AA and BB are horizontal and AA is to the right of BB AA and BB are both there with the object AA is to the right of object BB.
BB_below_AA	AA is placed at the bottom of BB. AA is at the bottom of BB and is on the same vertical plane. AA presents below BB.
AA_lowerleft_BB	BB is there and AA is at the 10 position of a clock face. BB is positioned below AA and to the left. .
BB_upperright_AA	Object A is above object BB and to the right of it, too. AA is diagonally to the upper right of BB.
AA_lowerright_BB	AA is to the right and above BB at an angle of about 45 degrees.
BB_upperleft_AA	BB is to the right and above AA at an angle of about 45 degrees. BB is diagonally left and above BB.

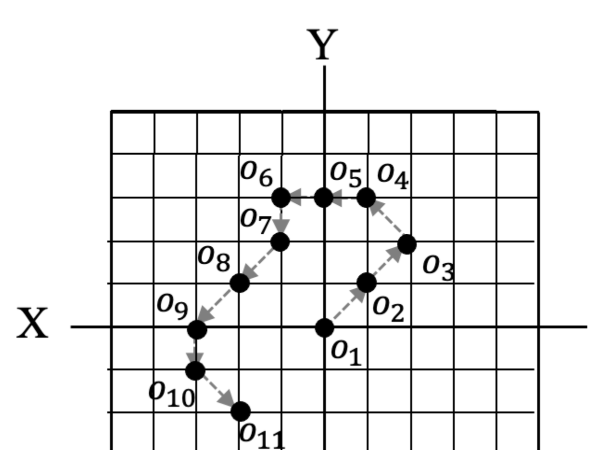
	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10
Clean	7.64	15.03	20.87	26.39	32.54	37.66	41.71	47.20	51.50	**54.29**
Noise	20.43	30.19	34.59	48.18	57.13	61.14	63.60	69.45	72.84	**74.21**

Solution to StepGame

Sentence-to-Relation Mapping + ASP Reasoner

Sentences	Template	ASP Facts
Y and I are parallel, and Y is on top of I.	Y_above_I	above("Y", "I").
F is on the left side of and below Q.	F_lowerleft_Q	down_left("F", "Q").
J is at O's 6 o'clock.	J_below_O	below("J", "O").
A is directly north east of B.	A_upperright_B	up_right("A", "B").
What is the relation of the agent B to the agent J?	query_B_J	query("B", "J").



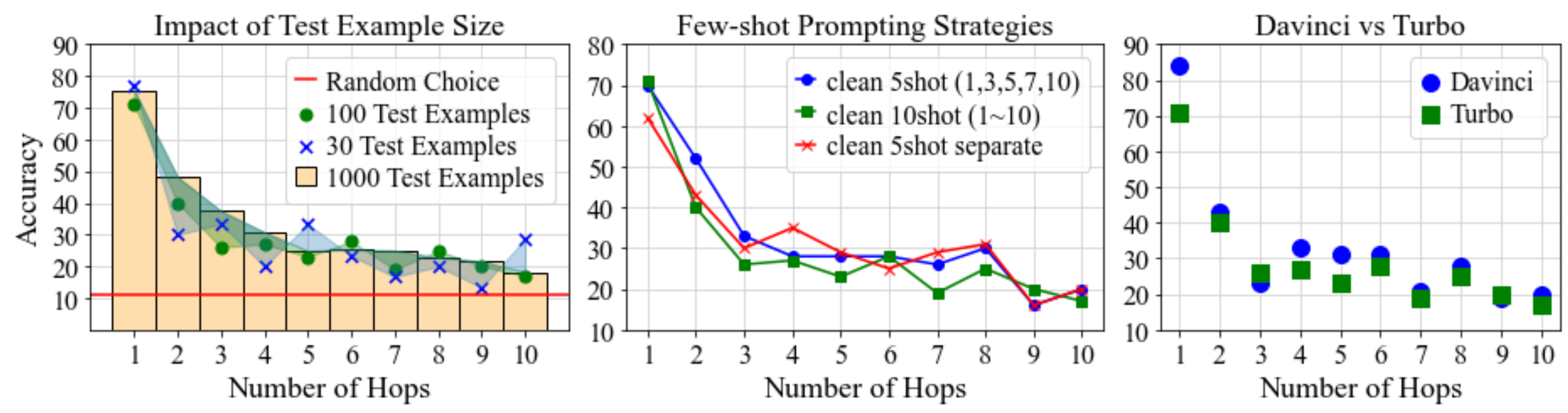


LLM + ASP

LLM for relation extraction + ASP Reasoner

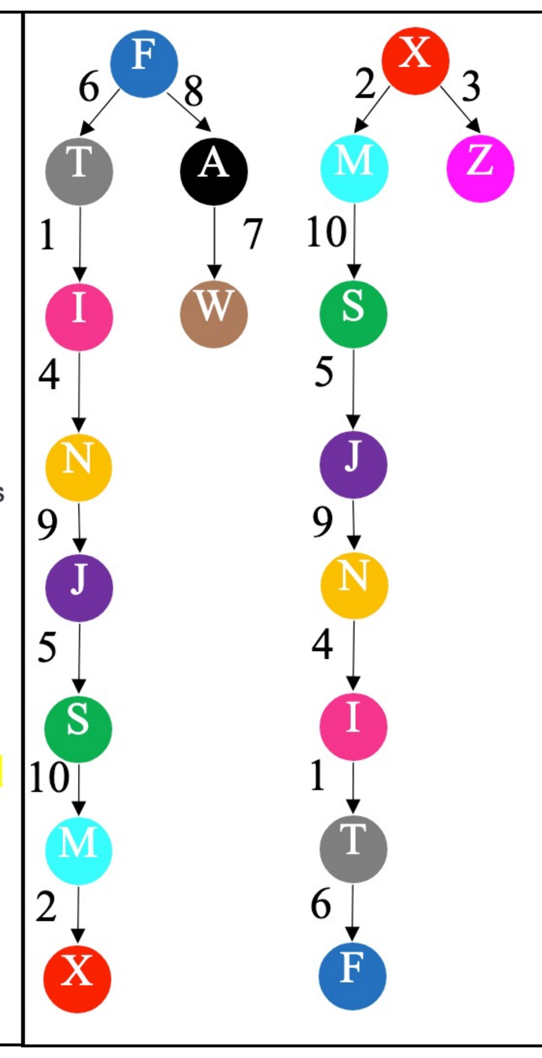
	left/ right	above /below	lower_left/ upper_right	lower_right/ upper_left
total	44	53	50	53
text-curie-001	11	41	30	37
text-davinci-003	**0**	**0**	**0**	2
gpt-3.5-turbo	2	2	3	**1**

Evaluation of GPT Models on Revised StepGame



Methods

Given a story about spatial relations among objects, answer the relation between two queried objects. **Possible relations are:** overlap, above, below, left, right, upper-left, upper-right, lower-left, and lower-right. If a sentence in the story is describing **clock-wise** information, then 12 denotes above, 1 and 2 denote upper-right, 3 denotes right, 4 and 5 denote lower-right, 6 denotes below, 7 and 8 denote lower-left, 9 denote left, 10 and 11 denote upper-left. If the sentence is describing **cardinal directions**, then north denotes above, east denotes right, south denotes below, and west denotes left. In all the spatial relations, assume that all agents occupy a position on a grid point of equally spaced points in the vertical and horizontal directions and that agents occupy the nearest grid point consistent with the spatial relation. The **offsets** of 9 spacial relations: offset(overlap) = (0,0); offset(top) = (0,1); offset(down) = (0,-1); offset(left) = (-1,0); offset(right) = (1,0); offset(top_left) = (-1,1); offset(top_right) = (1,1); offset(down_left) = (-1,1); offset(down_right) = (1,1).



Reasoning:
Suppose F is at (0,0). Link X and F using the relations in the story.
Start from F. According to 6. F is above T at 10 o'clock. So T is to the lower-right of F. T= F+ offset(lower-right) = (0,0)+(1,-1)=(1,-1).
Then search for T. According to 1. T is slightly off center to the top left and I is slightly off center to the bottom right. So I is to the lower-right of T. I= T+ offset(lower-right) = (1,-1)+(1,-1)=(2,-2).
Then search for I. According to 4. N is lower right of I. So N is to the lower-right of I. N= I+ offset(lower-right) = (2,-2)+(1,-1)=(3,-3).
Then search for N. According to 9. N is sitting in the right direction of J. So J is to the left of N. J= N+ offset(left) = (3,-3)+(-1,0)=(2,-3).
Then search for J. According to 5. S is positioned above J and to the right. So S is to the upper-right of J. S= J+ offset(upper-right) = (2,-3)+(1,1)= (3,-2).
Then search for S. According to 10. M is placed at the lower right of S. So M is to the lower-right of S. M= S+ offset(lower-right) = (3,-2)+(1,-1)=(4,-3).
Then search for M. According to 2. M is at the bottom of X. So X is above M. X= M+ offset(above) = (4,-3)+(0,1)=(4,-2). Come to X, ended. F(0,0), X(4,-2), so X is to the lower-right of F.
Answer: lower-right

Algorithm 1: Our ToT Approach

Require: LLM, input x

$S_0 \leftarrow Init(x)$
$i \leftarrow 1$
while no $s_f \in S_{i-1}$ has arrived at o_t **do**
  $S_i' \leftarrow \{s \cdot c \mid c \in G(s,j) \wedge ChainExtn(c) \wedge s \in S_{i-1}\}$
  if $S_i' = \emptyset$ **then return** failure
  $S_i \leftarrow select(b, \{\langle s, y\rangle \mid s \in S_i' \wedge y = \Sigma_1^n \sigma(V(s))\})$
  $i = i + 1$
end while
return $Link(s_f)$

Results

		k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10
Map+ASP		**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
Curie+ASP		46	43	42	59	67	67	57	56	58	61
Davinci+ASP		**100**	**100**	99	**100**	**100**	99	**100**	**100**	**100**	**100**
SOTA		92.6	89.9	89.1	93.8	92.9	91.6	91.2	90.4	89.0	88.3
Turbo	base	62	43	30	35	29	25	29	31	16	20
	CoT	/	34	40	36	28	28	26	31	25	24
	ToT_CoT	/	/	35	35	25	45	15	40	40	35
Davinci	base	77	42	21	26	25	30	23	23	22	22
	CoT	/	48	53	46	46	48	40	45	41	32
	ToT_CoT	/	/	65	50	45	60	50	50	55	50
GPT-4	base	100	70	55	45	40	25	40	35	35	25
	CoT	/	**80**	75	**95**	85	85	**90**	80	60	65
	ToT_CoT	/	/	**85**	85	**90**	**90**	85	**90**	**100**	**95**