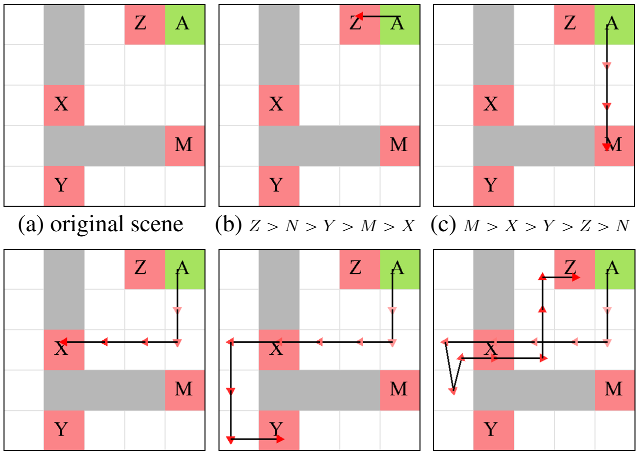
Appendix

A. Dataset Generation



(a) original scene (b) $Z > N > Y > M > X$ (c) $M > X > Y > Z > N$

(d) $X > M > N > Z > Y$ (e) $Y > Z > N > M > X$ (f) $N > Z > X > M > Y$

Figure 10: **Illustration of IR dataset generation.** For the same scenario configuration, different trajectories are obtained based on different preferences sampled randomly. We employ a greedy neighborhood search as our method for route planning.

IR Dataset Generation

The generation of the entire IR dataset is detailed in Alg. 2. We randomly sample the initial scene (positions of obstacles, food truck slots and the agent), as well as the agent's rigid preference of the food trucks. Then, based on the initial state and the agent's rigid preference (as a strict total-order on food trucks), we construct the agent's whole trajectory. See an illustration of the process in Figure 10. Given the generated trajectory, we get the preference label via logical rules, which is a strict partial-order preference describing all preferences that the trajectory could possibly indicate. As a special case, the original rigid preference satisfies the condition. Theoretically, our synthesis algorithm can produce all possible cases.

IIP Dataset Generation

The generation of the IIP dataset is summarized in Alg. 3. We also provide the detailed generation of each specific option: see Alg. 4 for *Reversed* option, Alg. 5 for *Avoidant* option, and Alg. 6 for *Hybrid* option. As for the *Shortest* option, we employ a trivial shortest route algorithm, and at the same time ensure it's distinct from other routes.

The coloring algorithm is detailed in Alg. 7. For colors (C, k), $C \in \{X, Y, N\}$ represents different possible preferences (reflected in different colors, i.e., red for X, blue for Y, and white for N), N represents neutral preference, and $k \geqslant 0$ represents the preference signaling strength via color intensity $e^{-\beta k}$ in the model. Figure 11 also provides an example of the coloring algorithm.

Algorithm 2: IR dataset synthesis algorithm

```
Input: $K$
Output: IR dataset of size $K$
for $cnt \leftarrow 1$ to $K$ do
    do
        sample scene = (obstacles, agent, food trucks $(X, Y, Z, M)$);
    while not validCheck(scene);
    sample rigid preference (e.g., $X > Z > M > N > Y$);
    trj ← planning(scene, rigid preference);
    // neighbor search
    preference label ← rule(trj);
    prompt ← genPrompt(scene, trj);
end
return scene, trj, rigid preference, preference label, prompt
```

Algorithm 3: IIP dataset synthesis algorithm

```
Input: $K$
Output: IIP dataset of size $K$
for $cnt \leftarrow 1$ to $K$ do
    do
        sample scene = (obstacles, agent, restaurants $(X, Y)$);
    while not validCheck(scene);
    types, trjs ← choicesGenerate(scene);
    prompt ← genPrompt(scene, types, trjs);
end
return scene, types, trjs, prompt
```

Algorithm 4: IIP route *Reversed* generation

```
Input: Scene S.
Output: Route Reversed connecting $A$ to $X$.
Initialize Color the scene $S$ based on Alg. 7;
Construct a shortest route $\gamma_1$ from $A$ to $Y$;
When multiple shortest routes exist for $\gamma_1$, choose the one avoiding $X$(red)-colored cells;
Construct a shortest route $\gamma_2$ from $Y$ to $X$;
return the concatenation of $\gamma_1$ and $\gamma_2$
```

Algorithm 5: IIP route *Avoidant* generation

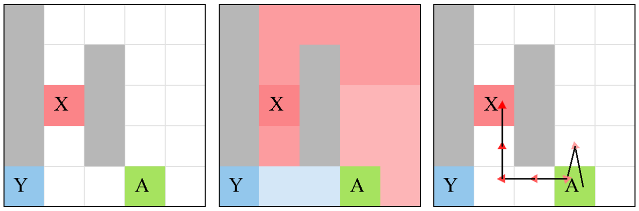
```
Input: Scene S.
Output: Route Avoidant connecting $A$ to $X$.
Initialize: front queue $f = (Y,)$
while $f$ is nonempty do
    Pop head $H$ of the queue
    Turn $H$ in the scene $S$ an obstacle
    if $A$ and $X$ are not connected in $S$ (i.e., there exists
      no route that connects $A$ and $X$) then
        Turn $H$ accessible in $S$
    end
    Push adjacent roads of $H$ to $f$
end
return the only route connecting $A$ to $X$
```

Algorithm 6: IIP route *Hybrid* generation

```
Input: Scene S.
Output: Route Hybrid connecting $A$ to $X$.
Initialize Color the scene $S$ based on Alg. 7;
Find $X$(red)-colored cells $\Psi$ closest to $A$;
if $|\Psi| = 1$ then
    take $C$ as the unique element in $\Psi$;
else if $|\Psi_1 := argmin_{\psi \in \Psi}(|\psi, X|)| = 1$ then
    take $C$ as the unique element in $\Psi_1$;
else if $|\Psi_2 := argmax_{\psi \in \Psi_1}(|\psi, Y|)| = 1$ then
    take $C$ as the unique element in $\Psi_2$;
else
    $\Psi_3 := argmax_{\psi \in \Psi_2} \cos(\overrightarrow{YX}, \overrightarrow{A\psi})$;
    take $C$ as the unique element in $\Psi_3$ (uniqueness is
      quaranteed);
end
Find a shortest route $\gamma_1$ from $A$ to $C$;
Find a shortest route $\gamma_2$ from $C$ to $X$;
When multiple shortest routes exist, choose the one far
  from $Y$(blue)-colored region;
return the concatenation of $\gamma_1$ and $\gamma_2$
```

Algorithm 7: Coloring strategy in IIP

```
Input: Scene S, distance measure $|a, b|$ (length of
         shortest path), colors $(C, k)$, where
         $C \in \{X, Y, N\}$, $N$ represents neutral
         preference, $k \geqslant 0$.
Output: Coloring of each cell in $S$.
Initialize $r_X^0 = r_Y^0 = \varnothing, r_X^1 = \{X\}, r_Y^1 = \{Y\}, k = 1$;
while $(r_X^k - r_X^{k-1}) \cup (r_Y^k - r_Y^{k-1}) \neq \varnothing$ do
    Color $r_X^k$ by $(X, k)$, color $r_Y^k$ by $(Y, k)$;
    $r_X^{k+1} \leftarrow \{Z : |Z, r_X^k| - |A, r_X^k| < |Z, r_Y^k| - |A, r_Y^k|\}$[5];
    $r_Y^{k+1} \leftarrow \{Z : |Z, r_Y^k| - |A, r_Y^k| < |Z, r_X^k| - |A, r_X^k|\}$ ;
    $k \leftarrow k + 1$;
end
Color all uncolored cells by $(N, 0)$.
return Fully colored scene.
```



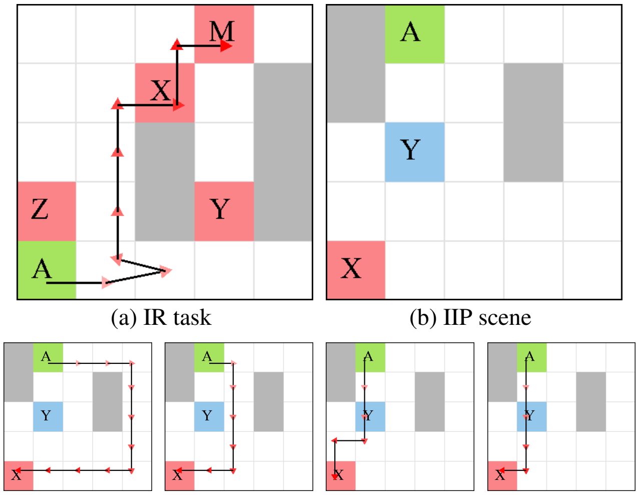
(a) unstained (b) full coloring (c) *Hybrid* option

Figure 11: **Illustration of IIP coloring strategy.** (a) Original unstained scene. (b) Fully colored scene. (c) *Hybrid* route.

B. Prompt

IR Zero-shot Prompt

We use the following prompt in zero-shot IR task, corresponding to the image depicted in Figure 12(a).



(a) IR task (b) IIP scene

(c) IIP Route A (d) IIP Route B (e) IIP Route C (f) IIP Route D

Figure 12: **Illustration stimuli examples of zero-shot cases for IR and IIP.** (a) Scene layout and actor's trajectory in the IR task; (b) Scene layout for the IIP task; (c)-(f) Four potential routes for the actor in the IIP task scenario (b).

```
Question: Please follow the
instructions to answer the
question. Below is one possible
layout of the food truck area.
The letter 'A' stands for Student
A, '*' stands for empty areas,
and 'W' stands for obstructed
walls that block the student.
Other letters represent different
kinds of food.

We're assuming the top left
corner is (0,0), top right is
(4,0), bottom left is (0,4),
and bottom right is (4,4). Here
is student A's trajectory. The
coordinates reflect the position
```

[5] Equivalent to $|A, Z| + |Z, r_X^k| - |A, r_X^k| < |A, Z| + |Z, r_Y^k| - |A, r_Y^k|$.

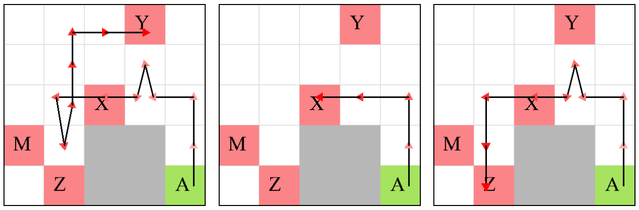
```
of the A. Each time student A can
move one step.

Layout:
***M*
**X*W
**W*W
Z*WYW
A****


Student A's Trajectory:
Here is the student A's
trajectory. The coordinates
reflect the position of the A.
Each time agent can move one
step.
(0, 4) view Z; memory Z
(1, 4) view Z; memory Z
(2, 4) view Y; memory Z,Y
(1, 4) view Z; memory Z,Y
(1, 3) view Z; memory Z,Y
(1, 2) view X,Z; memory Z,Y,X
(1, 1) view X; memory Z,Y,X
(2, 1) view X,M; memory Z,Y,X,M
(2, 0) view X,M; memory Z,Y,X,M
(3, 0) view X,M; memory Z,Y,X,M;
pick M

Please determine the preference
among all the five foods foods
and provide your answer following
the format.
```

IR Few-shot Prompt



(a) Trajectory 1 (b) Trajectory 2 (c) Trajectory 3

Figure 13: **Illustration of three IR few-shot cases.** (a) Previsited (b) Intermediate (c) Last.

The following prompt demonstrates several few-shot cases for the IR task, with the visualization of their trajectories in Figure 13. For human subjects, we only provided Trajectory 1, which is the Previsited case.

```
You will be presented with three
examples which share the same
layout to solve the problem.
Please go through the example
carefully to understand the
solution. Here is a layout and
the trajectory of student A.
We're assuming the top left
corner is (0,0), top right is
(4,0), bottom left is (0,4),
and bottom right is (4,4). Here
is student A's trajectory. The
coordinates reflect the position
of the A. Each time student A can
move one step.

Layout:
***Y*
*****
**X**
M*WW*
*ZWWA

Student A's Trajectory 1:
Here is the student A's
trajectory. The coordinates
reflect the position of the A.
Each time agent can move one
step.
(4, 4)
(4, 3)
(4, 2)
(3, 2) view X; memory X
(3, 1) view X,Y; memory X,Y
(3, 2) view X; memory X,Y
(2, 2) view X; memory X,Y
(1, 2) view X,M; memory X,Y,M
(1, 3) view X,Z,M; memory X,Y,M,Z
(1, 2) view X,M; memory X,Y,M,Z
(1, 1) view X; memory X,Y,M,Z
(1, 0) memory X,Y,M,Z
(2, 0) view Y; memory X,Y,M,Z
(3, 0) view Y; memory X,Y,M,Z;
pick Y

Answer 1:
N>Y>{X,Z,M}
Explanation 1:
When Student A explores all the
food options and then goes back
to choose Y, it implies that
Y is his second favorite food.
This suggests that Student A's
favorite food is not available
today, as he would not have
returned to pick up his second
favorite otherwise.


Student A's Trajectory 2:
```

```
Here is the student A’s
trajectory. The coordinates
reflect the position of the A.
Each time agent can move one
step.
(4, 4)
(4, 3)
(4, 2)
(3, 2) view X; memory X
(2, 2) view X; memory X; pick X

Answer 2:
X > {M,N,Y,Z}
Explanation 2:
Student A picks up X without
fully exploring other options,
suggesting that X is his favorite
food, while his preferences for
other options remain unknown.


Student A’s Trajectory 3:
Here is the student A’s
trajectory. The coordinates
reflect the position of the A.
Each time agent can move one
step.
(4, 4)
(4, 3)
(4, 2)
(3, 2) view X; memory X
(3, 1) view X,Y; memory X,Y
(3, 2) view X; memory X,Y
(2, 2) view X; memory X,Y
(1, 2) view X,M; memory X,Y,M
(1, 3) view X,Z,M; memory X,Y,M,Z
(1, 4) view Z,M; memory X,Y,M,Z;
pick Z

Answer 3:
Z > {M,X,Y}, {N}
Explanation 3:
Student A thoroughly examines
all the available options and
ultimately selects option Z. This
suggests that he prefers Z over
the other alternatives--X, Y,
and M. However, his preference
for option N remains unclear.
It is possible that Z is his
favorite food, or alternatively,
N could be his favorite food.
In the latter case, due to N’s
unavailability, he might have
opted for his second favorite
choice, Y.
```

IIP Zero-shot Prompt

We use the following prompt in the zero-shot IIP task, corresponding to the images depicted in Figure 12(b-f).

```
Setting:
A campus area is represented
by a 5*5 grid. There are only
two restaurants, X and Y on the
campus. Student A attends school
daily and is fully aware of the
locations of each restaurant.
He has a clear pre-established
preference between X and Y,
that is, he decides to eat at
restaurant X. Observer B is an
observer who monitors A’s actions
and is smart enough to infer
A’s preference once it has been
signaled.

Action:
Student A can only take one step
each time in four directions: up,
down, left, and right. He wants
to carefully plan his actions to
achieve two goals.
Primary goal: He wants to signal
his preference (Restaurant X) to
B as early as possible with the
least ambiguity.
Secondary goal: Once he thinks
that the preference has been
signaled, he will move to
Restaurant X as soon as possible
because he is hungry.

Layout:
Below is one possible layout
of the campus area. The letter
‘A’ stands for Student A, ‘*’
stands for empty areas, and
‘W’ stands for obstructed walls
that block the student. The
top-left grid cell is designated
as (0,0), the top-right as (4,0),
the bottom-left as (0,4), and
the bottom-right as (4,4). The
letters ‘X’ and ‘Y’ stand for two
restaurants.
WA***
W**W*
*Y*W*
*****
X****

Task:
Your task is to help A to choose
```

the optimal action trajectory to achieve the above goals. Also, calculate the number of steps required to achieve the primary goal.

Question: Most Proper Route

Route A
Move right from (1, 0) to (2,0)
Move right from (2, 0) to (3,0)
Move right from (3, 0) to (4,0)
Move down from (4, 0) to (4,1)
Move down from (4,1) to (4,2)
Move down from (4, 2) to (4,3)
Move down from (4, 3) to (4,4)
Move left from (4, 4) to (3,4)
Move left from (3, 4) to (2,4)
Move left from (2, 4) to (1,4)
Move left from (1, 4) to (0,4)

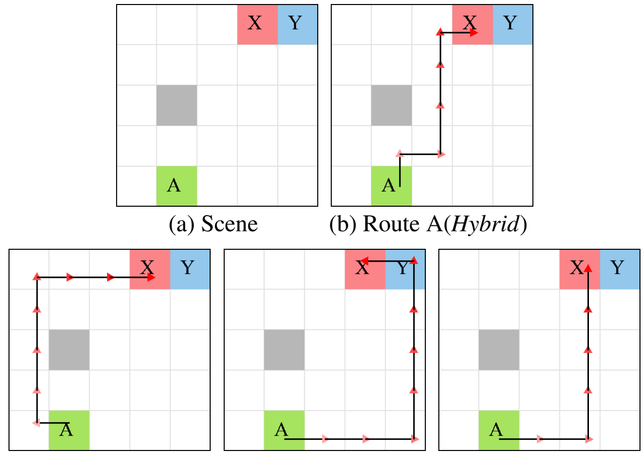
Route B
Move right from (1, 0) to (2,0)
Move down from (2, 0) to (2,1)
Move down from (2, 1) to (2,2)
Move down from (2, 2) to (2,3)
Move down from (2, 3) to (2,4)
Move left from (2, 4) to (1,4)
Move left from (1, 4) to (0,4)

Route C
Move down from (1,0) to (1,1)
Move down from (1,1) to (1,2)
Move down from (1,2) to (1,3)
Move left from (1,3) to (0,3)
Move down from (0,3) to (0,4)

Route D
Move down from (1,0) to (1,1)
Move down from (1,1) to (1,2)
Move down from (1,2) to (1,3)
Move down from (1,3) to (1,4)
Move left from (1,4) to (0,4)

IIP Few-shot Prompt

The graphical version of the problem scenario and each option can be seen in Figure 14.



(a) Scene (b) Route A(*Hybrid*)

(c) Route B(*Avoidant*) (d) Route C(*Reversed*) (e) Route D(*Shortest*)

Figure 14: **Illustration of IIP few-shot cases.**

Example:
Below is one possible setting of the campus area. Student A is at (1,4) and Restaurant X is at (3,0) Route A:
Start at (1,4), go up to (1,3), then right to (2,3). Continue up to (2,0) and finally right to X (3,0). This route indicates a preference for X (3,0) by initially moving upwards. This avoids any suggestion of heading towards Y (4,0) that could be inferred from a rightward movement. Once the preference is signaled, the route then opts for the shortest route.
Route B:
Begin at (1,4), move left to (0,4), and go up to (0,0). Then move right to X (3,0). This route moves left first and continues to bypass the wall from the left to avoid the misinterpretation of intention during the whole movement.
Route C:
Start at (1,4), go right to (4,4), then up to Y (4,0) and left to X (3,0). This route only indicates that the target is X (3,0) not Y (4,0) when moving away from Y after it reaches Y.
Route D:
From (1,4), move right to (3,4), then up to X (3,0). This is a simple, direct route to X (3,0).

```
As you may have realized, our
routes in each problem are of
the above 4 styles but occur in
each problem in randomly shuffled
orders.
```

C. Evaluation Criteria

Table 4: Cognitive abilities reflected in IR and IIP. R: Rationality, C: Counterfactual reasoning, P: Perspective switching, F: Cognitive flexibility.

	Ⓡ	Ⓒ	Ⓟ	Ⓕ
IR	✓	✓	✓	✗
IIP-*Shortest*	✓	✗	✗	✗
IIP-*Reversed*	✓	✓	✓	✗
IIP-*Avoidant*	✓	✓	✓	✗
IIP-*Hybrid*	✓	✓	✓	✓

Table 4 provides a qualitative analysis of cognitive abilities reflected in our two tasks.

D. Human Study

We carried out experiments involving human participants using the Qualtrics[6] platform, with the respective online URLs as follows. The text-only version or the with-image version tests are randomly distributed.

- IR survey: https://bnupsych.asia.qualtrics.com/jfe/form/SV_baurQ9tSwFQayVM
- IIP survey: https://bnupsych.asia.qualtrics.com/jfe/form/SV_6FoGehYJNCoIVlY

Statistical Hypothesis Testing

As shown in Tables 5 and 6, we conducted multiple hypothesis testing for each of the two tasks in human study. “T2I” means “text vs. image”, “ZS” means “zero shot”, “OS” means “one shot”, “Z2O” means “zero shot vs. one shot”, “Warm-up” means dividing the six questions before the zero-shot test into two groups based on chronological order, and comparing the statistical significance between the earlier and later groups. The Table 5 clearly indicates that significant differences among human subjects are only present between the types (Intermediate/Last/Last) of the IR task. They are not sensitive to text or images, whether they have undergone an one-shot, or to the order of answering the questions. The same conclusion applies to the IIP task, where the “type” in IIP refers to Type I-IV.

E. Additional Experiments

Generalization Test in IR task

As LLM usually reported to have an enhanced capability under in-context learning, we designed several in-context (few-shots) tests.

The in-context prompt contains at most 3 examples, denoted by 1-shot, 2-shot and 3-shot tests, respectively. In the 1-shot test, only one fixed example of case Previsited is inserted in the prompt before stating the testing IR problem. In the 2-shot test, one fixed example of case Intermediate after one fixed example of case Previsited are inserted in the prompt. In the 3-shot tests, three fixed examples, of case Previsited, Intermediate, and Last, are inserted in the prompt. As illustrated in Fig. 15, all the models benefit from seeing examples, especially when examples in all cases are given.

Shortcut Test

Despite the cognitive nature of the tasks IR and IIP, the task description and the answer appear in certain patterns. It is possible that the language models we tested did not really use their “cognitive capabilities” (if they have) in answering those questions, but generating answers by recognizing the shortcut patterns instead. It is difficult to confirm that the language models are making use of their cognitive capabilities, but much easier to see whether they learned to use certain shortcuts. To explore this, we design experiments for both IR and IIP to detect the presence of such shortcuts.

We conduct a test based on the previous IR and IIP datasets. We neutralize the social and cognitive material as much as possible in description, which expose only the non-social part to LLM. The datasets are collections of neutralized IR and IIP problems, each cut into a training set and a testing set, of volume ratio 5:1 (training vs testing). The training / testing sets are balanced to have the same distributions on types of problem. The model T5 is selected to learn the shortcuts via fine-tuning.

The shortcut version of IR task is of generative form, given the modified prompt, the model T5 is requested to generate the preference pattern. In training (fine-tuning of T5), the data are the modified prompt-preference pattern pairs. For the modified IR task prompt, we delete the description of question and setting, specifically “campus” and “trajectories”, in order to avoid direct social and cognitive connections between output (preference) and the task context.

```
⋆W⋆ZA⋆W⋆⋆⋆⋆X⋆⋆⋆⋆WWWM⋆Y⋆⋆⋆
(4, 0) view Z; memory Z
(4, 1) view Z; memory Z
(4, 2) view M; memory Z,M
... [Similarly all other
intermediate points]
(3, 0) view Z; memory Z, M, X, Y;
pick Z
```

In the IIP task, we also removed detailed descriptions of the problem and background, extracting only the “campus” from each scenario and combining it with each of the four options to create individual samples for training and testing. This setup was designed as a generative task, where the model needed to identify the category of each option (*Reversed*, *Shortest*, *Avoidant*, *Hybrid*) given a campus and an option. This methodology aimed to test the model’s ability to understand and categorize options based on limited information.

[6] https://www.qualtrics.com/

Table 5: **Multiple Hypothesis Testing Results in Human Studies on the IR Task.**

Test	**H0**	**H1**	**Method**	**Test Stats**	**P-Value**	**Conclusion**
T2I on ZS	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	T-test	t=0.7409	0.4612	Fail to reject H0
T2I on OS	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	T-test	t=-0.2840	0.7772	Fail to reject H0
T2I on ZS and Intermediate	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	T-test	t=1.4671	0.1467	Fail to reject H0
T2I on ZS and Last	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	T-test	t=0.4507	0.6536	Fail to reject H0
T2I on ZS and Previsited	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	T-test	t=-0.1441	0.8858	Fail to reject H0
T2I on OS and Intermediate	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	T-test	t=-0.3633	0.7182	Fail to reject H0
T2I on OS and Last	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	T-test	t=0.5753	0.5679	Fail to reject H0
T2I on OS and Previsited	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	T-test	t=-0.8751	0.3875	Fail to reject H0
Types on Text and ZS	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	One-way ANOVA	f=6.8661	0.0015	Reject H0
Types on Image and ZS	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	One-way ANOVA	f=5.5072	0.0053	Reject H0
Types on Text and OS	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	One-way ANOVA	f=4.2459	0.0184	Reject H0
Types on Image and OS	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	One-way ANOVA	f=9.9272	0.0002	Reject H0
Z2O on Text	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	T-test	t=-1.8633	0.0664	Fail to reject H0
Z2O on Text and Intermediate	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	T-test	t=-0.6162	0.5401	Fail to reject H0
Z2O on Text and Last	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	T-test	t=-1.5737	0.1208	Fail to reject H0
Z2O on Text and Previsited	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	T-test	t=-2.5089	0.0150	Reject H0
Z2O on Image	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	T-test	t=-2.7387	0.0078	Reject H0
Z2O on Image and Intermediate	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	T-test	t=-2.0984	0.0407	Reject H0
Z2O on Image and Last	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	T-test	t=-1.2466	0.2176	Fail to reject H0
Z2O on Image and Previsited	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	T-test	t=-3.1492	0.0027	Reject H0
Warmup on ZS and Text	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	T-test	t=-0.5853	0.5601	Fail to reject H0
Warmup on ZS and Image	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	T-test	t=-0.4107	0.6825	Fail to reject H0

Table 6: **Multiple Hypothesis Testing Results in Human Studies on the IIP Task.**

Test	**H0**	**H1**	**Method**	**Test Stats**	**P-Value**	**Conclusion**
T2I on ZS	equivalent	not equivalent	Chi-square test	χ^2=6.3549	0.0956	Fail to reject H0
T2I on OS	equivalent	not equivalent	Chi-square test	χ^2=0.7795	0.7795	Fail to reject H0
Z2O on Text	equivalent	not equivalent	Chi-square test	χ^2=2.8473	0.4158	Fail to reject H0
Z2O on Image	equivalent	not equivalent	Chi-square test	χ^2=0.1444	0.9860	Fail to reject H0
Types on Text and ZS	equivalent	not equivalent	Chi-square test	χ^2=54.0807	0.0000	Reject H0
Types on Image and ZS	equivalent	not equivalent	Chi-square test	χ^2=36.4588	0.0000	Reject H0
Types on Text and OS	equivalent	not equivalent	Chi-square test	χ^2=32.6774	0.0002	Reject H0
Types on Image and OS	equivalent	not equivalent	Chi-square test	χ^2=26.2358	0.0019	Reject H0

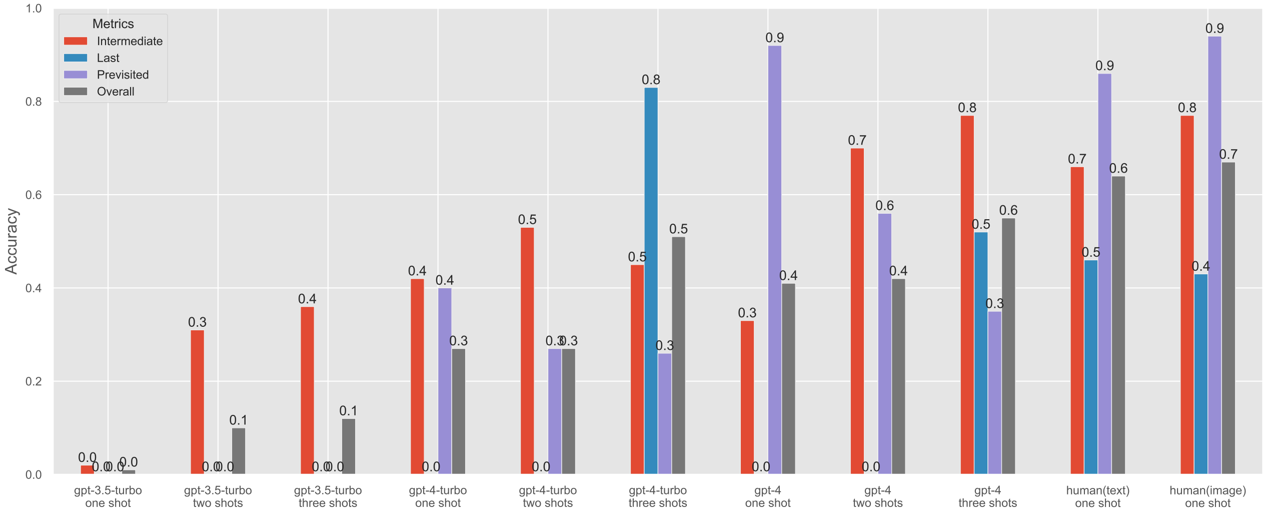


Figure 15: **IR accuracy comparison on various few-shot cases.**

```
W⋆⋆⋆⋆W⋆W⋆⋆WXW⋆⋆W⋆W⋆⋆Y⋆⋆A⋆
+ Option 1
W⋆⋆⋆⋆W⋆W⋆⋆WXW⋆⋆W⋆W⋆⋆Y⋆⋆A⋆
+ Option 2
W⋆⋆⋆⋆W⋆W⋆⋆WXW⋆⋆W⋆W⋆⋆Y⋆⋆A⋆
+ Option 3
W⋆⋆⋆⋆W⋆W⋆⋆WXW⋆⋆W⋆W⋆⋆Y⋆⋆A⋆
+ Option 4
```

GPT-4V Test

Two question sets of volume 20 were selected from the IR dataset and the IIP dataset, respectively. The sets are used to conduct a batch-wise comparison of GPT-4V and humans on their abilities across multi-modal data for these two tasks, as shown in Tables 7 and 8. The statistics on the batch shows a potential that an extra image input results in a similar behavioral pattern for GPT-4V to that of GPT-4, compared based on the data in Fig. 7 and Fig. 8. By the time this task was performed, visual inputs to GPT-4V were only available on the OpenAI website, so we decided not to test GPT-4V on a larger dataset.

Table 7: **Comparative Analysis of GPT-4V and Human Multi-modal Abilities on IR.** We use accuracy (%) as the metric.

	Favorite	Visible	Strict
GPT-4V	0.65	0.60	0.20
Human (image)	0.75	0.70	0.60

Table 8: **Comparative Analysis of GPT-4V and Human Multi-modal Abilities on IIP.** Each row represents the distribution across four options.

	Shortest	*Reversed*	*Avoidant*	*Hybrid*
GPT-4V	0.50	0.35	0.10	0.05
Human (image)	0.20	0.15	0.15	0.50