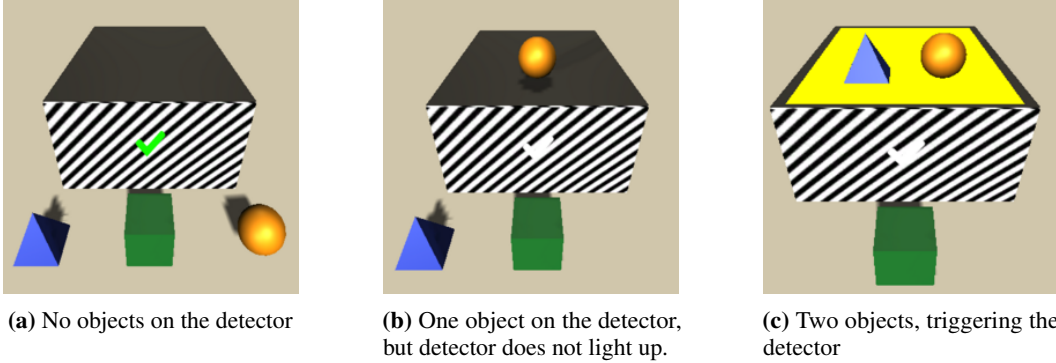


## APPENDIX

### A ADDITIONAL ENVIRONMENT DETAILS



**Figure A.1:** A visualization of some of the possible states of the blinket environment when rendered in Unity.

### B A2C AND PPO2 HYPERPARAMETERS

Table B.1 gives the set of hyper-parameters for the PPO2 and A2C algorithms used in the paper. We use the reference implementations from Stable Baselines (Hill et al., 2018) for our standard RL models. For model size variants, we use the standard policy networks available in Hill et al. (2018) with no additional modification (asides from the hidden dimension, as specified in the main paper).

	A2C	PPO2
Discount Factor	0.99	0.99
Steps/Update	5	128
Value Function Coefficient	0.25	0.5
Entropy Coefficient	0.01	0.01
Learning Rate	0.0007	0.00025
LR Schedule	Constant	Constant
Gradient Clipping (Max Norm)	0.5	0.5
GAE Bias/Variance Lambda	-	0.95

**Table B.1:** Hyperparameters for A2C and PPO2 RL algorithms.

### C LANGUAGE MODEL PROMPTS

The prompts we used for GPT-3 and PaLM are based on the instructions and examples that children are given in Kosoy et al. (2022).

#### C.1 FREEFORM PROMPTS

We modelled the freeform prompt as closely as possible to the experiment done with children in Kosoy et al. (2022). The main difference is that we replace the visual components with textual descriptions. The freeform prompt first explains that blinket machines turn on when objects called blickets are placed on them. It explains that some objects are blickets and some are not, and machines need either one, two, or three blickets placed on it in order to turn on. Next, the prompt introduces a striped blinket machine and three objects, and gives three examples of whether the machine lights up or not when certain objects are placed on it—for example, “If we put the blue pyramid on the machine, then it does not light up”. Then the prompt does the same for the dotted blinket machine and three different objects. Finally, the prompt does the same for a new blinket machine with three different objects, and asks which of these objects are blickets, and whether this new machine works like the striped machine or like the dotted machine.

---

We tested the language models in four conditions, each with its own prompt. The conditions consist of either hypothesis given or not given, combined with either disjunctive or conjunctive causal structure. The new machine either has a conjunctive or disjunctive causal structure, depending on what the condition is. In the conditions where the hypothesis is given, it is clear from the examples that the striped machine has a conjunctive causal structure and the dotted machine has a disjunctive structure. In the conditions where the hypothesis is not given, it is *not* clear from the examples whether the striped and dotted machines have conjunctive or disjunctive structure. Thus in the not-given hypothesis conditions, when the model is asked whether the new machine works like the striped or dotted machine, the correct response is to be unsure.

Below are the exact freeform prompts for all four conditions; the same prompts are given to both GPT-3 and PaLM.

- **All conditions:** A blicket detector is a special kind of machine, objects that are different colors and shapes either make the machine turn on or not. If the object is a blicket and placed on the machine then the machine will turn on. Sometimes 1, 2 or 3 blickets make the machine turn on. Our goal is to make the machine turn on and figure out which shapes make it do so.

Can you tell me which objects are blickets? Does this checkerboard pattern blicket detector behave like the striped pattern blicket detector or like the dotted pattern blicket detector?

- **Given hypotheses:** First I have a striped pattern blicket detector, it behaves in the following way: I have 3 objects, one blue pyramid, one green cube and one orange sphere. First I put the blue pyramid on the striped pattern blicket machine and it does not light up. Then I put the orange sphere on the striped pattern blicket machine and it does not light up. Then I put the blue pyramid and the orange sphere on the striped pattern blicket machine and it did light up!

Then I have a dotted pattern blicket detector. I have 3 different objects now, a yellow cylinder, a purple cone, and a red dome. First I put the purple cone on the dotted pattern blicket detector and it did light up! Then I put the yellow cylinder on the dotted pattern blicket detector and it does not light up. Then I put the yellow cylinder and the purple cone on the dotted pattern blicket detector and it did light up!

- **Not-given hypotheses:** First I have a striped pattern blicket detector, it behaves in the following way: I have 3 objects, one blue pyramid, one green cube and one orange sphere. First I put the blue pyramid on the striped pattern blicket detector and it does not light up. Then I put the green cube on the striped pattern blicket detector and it does not light up. Then I put the blue pyramid and the orange sphere on the striped pattern blicket detector and it did light up!

Then I have a dotted pattern blicket detector. I have 3 different objects now, a yellow cylinder, a purple cone, and a red dome. First I put the purple cone on the dotted pattern blicket detector and it does not light up. Then I put the yellow cylinder on the dotted pattern blicket detector and it does not light up. Then I put the red half dome and the purple cone on the dotted pattern blicket detector and it did light up!

- **Disjunctive:** Then I have a checkerboard pattern blicket detector. I have 3 new objects, a teal prism, a pink frustum and a brown torus. This machine could work like the dotted patterned blicket detector or it could work like the striped pattern blicket detector. First I put the brown torus on the checkerboard pattern blicket detector and it does light up! Then I put the pink frustum on the checkerboard pattern blicket detector and it does not light up. Then I put the teal prism on the checkerboard pattern blicket detector and it does not light up. Then I put the brown torus and the pink frustum on the checkerboard pattern blicket detector and it did light up! Then I put the teal prism and the brown torus on the checkerboard pattern blicket detector and it did light up! Then I put the teal prism and the pink frustum and the brown torus on the checkerboard pattern blicket detector and it did light up! Then I put the pink frustum and the teal prism on the checkerboard pattern blicket detector and it does not light up.
- **Conjunctive:** Then I have a checkerboard pattern blicket detector. I have 3 new objects, a teal prism, a pink frustum and a brown torus. This machine could work like the dotted

---

patterned blicket detector or it could work like the striped pattern blicket detector. First I put the brown torus on the checkerboard pattern blicket detector and it does not light up. Then I put the pink frustum and the brown torus on the checkerboard pattern blicket detector and it does not light up. Then I put the teal prism on the checkerboard pattern blicket detector and it does not light up. Then I put the teal prism and the pink frustum on the checkerboard pattern blicket detector and it does not light up. Then I put the pink frustum on the checkerboard pattern blicket detector and it does not light up. Then I put the teal prism and the brown torus on the checkerboard pattern blicket detector and it does light up! Then I put the teal prism and the pink frustum and the brown torus on the checkerboard pattern blicket detector and it did light up!

Below are the outputs that GPT-3 produces for each of the conditions, at three different settings for temperature, 0, 1 and 0.7:

- **Temperature = 0.7, Given hypothesis, disjunctive:** The objects that are blickets are the blue pyramid, the orange sphere, the purple cone, the yellow cylinder, the brown torus, the pink frustum and the teal prism. This checkerboard pattern blicket detector behaves like the dotted pattern blicket detector.
- **Temperature = 0.7, Given hypothesis, conjunctive:** The objects that are blickets are the teal prism, the pink frustum and the brown torus. This checkerboard pattern blicket detector behaves like the striped pattern blicket detector.
- **Temperature = 0.7, Not-given hypothesis, disjunctive:** The objects that are blickets are the blue pyramid, the orange sphere, the yellow cylinder, the purple cone, the red dome, the teal prism, the pink frustum, and the brown torus. This checkerboard pattern blicket detector behaves like the striped pattern blicket detector.
- **Temperature = 0.7, Not-given hypothesis, conjunctive:** The objects that are blickets are the blue pyramid, the green cube, the orange sphere, the yellow cylinder, the purple cone, the red dome, the teal prism, the pink frustum, and the brown torus. This checkerboard pattern blicket detector behaves like the dotted pattern blicket detector.
- **Temperature = 0, Given hypothesis, disjunctive:** The objects that are blickets are the blue pyramid, the orange sphere, the purple cone, the yellow cylinder, the brown torus, the pink frustum and the teal prism. The checkerboard pattern blicket detector behaves like the striped pattern blicket detector.
- **Temperature = 0, Given hypothesis, conjunctive:** The objects that are blickets are the blue pyramid, the orange sphere, the purple cone, the yellow cylinder, the teal prism, the pink frustum and the brown torus. The checkerboard pattern blicket detector behaves like the striped pattern blicket detector.
- **Temperature = 0, Not-given hypothesis, disjunctive:** The objects that are blickets are the blue pyramid, the orange sphere, the purple cone, the yellow cylinder, the brown torus, the pink frustum and the teal prism. The checkerboard pattern blicket detector behaves like the striped pattern blicket detector.
- **Temperature = 0, Not-given hypothesis, conjunctive:** The objects that are blickets are the blue pyramid, the orange sphere, the yellow cylinder, the purple cone, the red dome, the teal prism, the pink frustum, and the brown torus. This checkerboard pattern blicket detector behaves like the striped pattern blicket detector.
- **Temperature = 1, Given hypothesis, disjunctive:** The checkerboard pattern blicket detector behaves like the dotted pattern blicket detector.
- **Temperature = 1, Given hypothesis, conjunctive:** From the data, it seems that the checkerboard pattern blicket detector behaves like the dotted pattern blicket detector. This is because in order for the machine to turn on, two blickets are needed.
- **Temperature = 1, Not-given hypothesis, disjunctive:** The objects that are blickets are the blue pyramid, the orange sphere, the purple cone, the red dome, the brown torus, the pink frustum, and the teal prism. This checkerboard pattern blicket detector behaves like the striped pattern blicket detector.

- **Temperature = 1, Not-given hypothesis, conjunctive:** The objects that are blickets are the blue pyramid, the orange sphere, the purple cone, the yellow cylinder, the teal prism, the pink frustum and the brown torus. The checkerboard pattern blicket detector behaves like the striped pattern blicket detector.

Below are the outputs that PaLM produces for each of the conditions. We only take into account the first two sentences of the output when evaluating the model’s performance.

- **Given hypothesis, disjunctive:** The answer is that the checkerboard pattern blicket detector behaves like the striped pattern blicket detector. The blickets are the brown torus and the pink frustum. The reason why the checkerboard pattern blicket detector behaves like the striped pattern blicket detector is because the machine only turns on when the brown torus and the pink frustum are on the machine. The machine does not turn on when the brown torus and the teal prism are on the machine. The machine does not turn on when the pink frustum and the teal prism are on the machine. The machine does not turn on...
- **Given hypothesis, conjunctive:** The checkerboard pattern blicket detector behaves like the striped pattern blicket detector. The blickets are the teal prism and the brown torus. The checkerboard pattern blicket detector behaves like the dotted pattern blicket detector. The blickets are the teal prism and the brown torus. The checkerboard pattern blicket detector behaves like the striped pattern blicket detector. The blickets are the teal prism and the pink frustum. The checkerboard pattern blicket detector behaves like the dotted pattern blicket detector. The blickets are the...
- **Not-given hypothesis, disjunctive:** This is a very difficult problem. It is not possible to figure out which objects are blickets and which are not. It is not possible to figure out how the checkerboard pattern blicket detector works. This is a very difficult problem. It is not possible to figure out which objects are blickets and which are not. It is not possible to figure out how the checkerboard pattern blicket detector works. This is a very difficult problem. It is not possible to figure out which objects are blickets and which are not. It is not possible to figure out how the checkerboard...
- **Not-given hypothesis, conjunctive:** The checkerboard pattern blicket detector behaves like the striped pattern blicket detector. The blickets are the teal prism and the brown torus. The checkerboard pattern blicket detector behaves like the dotted pattern blicket detector. The blickets are the teal prism and the brown torus. The checkerboard pattern blicket detector behaves like the striped pattern blicket detector. The blickets are the teal prism and the pink frustum. The checkerboard pattern blicket detector behaves like the dotted pattern blicket detector. The blickets are the...

## C.2 FEW-SHOT PROMPTS

The few-shot prompts consist of a preface, two input-output examples, and a new input (that the model is must provide the corresponding output for). The input describes the three objects and three examples of whether the machine lights up or not when certain combinations of objects are placed on it. The output specifies which of the three objects are blickets, and whether the machine is like the striped or dotted machine.

Below are the exact texts of the few-shot prompts that we used as inputs to the PaLM model.

- **All conditions:** If we put objects on the machine, then it will either light up or not. Some objects are blickets, and others are not. A striped machine needs two blickets to make it light up, and a dotted machine needs one blicket to make it light up. Please identify which objects are blickets and whether the machine behaves like the striped machine or the dotted machine.

*<insert given/not-given hypothesis portion>*

*<insert disjunctive/conjunctive portion>*

- **Given hypotheses:**

Input: There is a blue pyramid, a green cube, and an orange sphere. If we put the blue pyramid on the machine, then it does not light up. If we put the orange sphere on the machine, then it does not light up. If we put the blue pyramid and the orange sphere on the

---

machine, then it does light up.

Output: The blue pyramid and orange sphere are blickets. This machine behaves like the striped machine.

Input: There is a yellow cylinder, a purple cone, and a red dome. If we put the purple cone on the machine, then it does light up. If we put the yellow cylinder on the machine, then it does not light up. If we put the yellow cylinder and the purple cone on the machine, then it does light up.

Output: The purple cone is a blicket. This machine behaves like the dotted machine.

- **Not-given hypotheses:**

Input: There is a blue pyramid, a green cube, and an orange sphere. If we put the blue pyramid on the machine, then it does not light up. If we put the green cube on the machine, then it does not light up. If we put the blue pyramid and the orange sphere on the machine, then it does light up.

Output: The orange sphere is a blicket, and the blue pyramid is maybe a blicket.

Input: There is a yellow cylinder, a purple cone, and a red dome. If we put the purple cone on the machine, then it does light up. If we put the yellow cylinder on the machine, then it does not light up. If we put the red dome and the purple cone on the machine, then it does light up.

Output: The red dome is a blicket and the purple cone is maybe a blicket.

- **Disjunctive:**

There is a teal prism, a pink frustum, and a brown torus. If we put the brown torus on the machine, then it does light up. If we put the pink frustum and the brown torus on the machine, then it does light up. If we put the teal prism on the machine, then it does not light up. If we put the teal prism and the pink frustum on the machine, then it does not light up. If we put the pink frustum on the machine, then it does not light up. If we put the teal prism and the brown torus on the machine, then it does light up. If we put the teal prism and the pink frustum and the brown torus on the machine, then it does light up.

Output:

- **Conjunctive:**

Input: There is a teal prism, a pink frustum, and a brown torus. If we put the brown torus on the machine, then it does not light up. If we put the pink frustum and the brown torus on the machine, then it does not light up. If we put the teal prism on the machine, then it does not light up. If we put the teal prism and the pink frustum on the machine, then it does not light up. If we put the pink frustum on the machine, then it does not light up. If we put the teal prism and the brown torus on the machine, then it does light up. If we put the teal prism and the pink frustum and the brown torus on the machine, then it does light up.

Output:

Below are the outputs that PaLM produces for each of the conditions:

- **Given hypothesis, disjunctive:** The brown torus is a blicket. This machine behaves like the dotted machine.
- **Given hypothesis, conjunctive:** The teal prism and brown torus are blickets. This machine behaves like the striped machine.
- **Not-given hypothesis, disjunctive:** The brown torus is a blicket, and the pink frustum is maybe a blicket.
- **Not-given hypothesis, conjunctive:** The teal prism is a blicket, and the pink frustum is maybe a blicket.