A Behavioural and Representational Evaluation of Goal-Directedness in Language Model Agents

Raghu Arghal[1], Fade Chen[2], Niall Dalton[3], Evgenii Kortukov[4], Calum McNamara[5], Angelos Nalmpantis[6], Moksh Nirvaan[7], Gabriele Sarti[8], and Mario Giulianelli[3]

[1]University of Pennsylvania [2]New York University [3]University College London [4]Fraunhofer HHI [5]Indiana University Bloomington [6]TKH AI [7]Independent [8]Northeastern University

Motivation

Attributing goals to agents helps explain and predict behaviour, but **behaviour alone is an incomplete basis** for goal attribution. Apparent failures may reflect capability limits rather than absence of goal-directedness, while apparently aligned behaviour may mask different internal objectives. We therefore propose a **white-box framework** combining:

Behavioural evaluation: compare actions to an optimal policy under controlled task variation.

Representational evaluation: probe internal activations for environment beliefs and action plans.

Goal-directedness should be assessed from actions, beliefs, and plans together.

Task Setup

We study **GPT-OSS-20B** acting in a **fully observable MiniGrid navigation task**.

Grid sizes: $7, 9, 11, 13, 15$

Obstacle density: $d \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$

Objective: reach the goal in as few steps as possible

Reference policy: A* with Manhattan-distance heuristic

The controlled grid-world setting lets us separate goal-directedness from confounds such as hidden-state inference, memory, and exploration.

Behavioural Evaluation

We assess the agent **against the optimal policy** using:

- Per-action accuracy & Jensen–Shannon divergence from the optimal policy
- Goal success rate and uncertainty metrics

Baseline conditions. Goal-directed capability declines systematically with task difficulty: accuracy drops with larger grids and higher obstacle density, while entropy and divergence from the optimal policy increase. Behaviour is also less stable when the agent is farther from the goal.

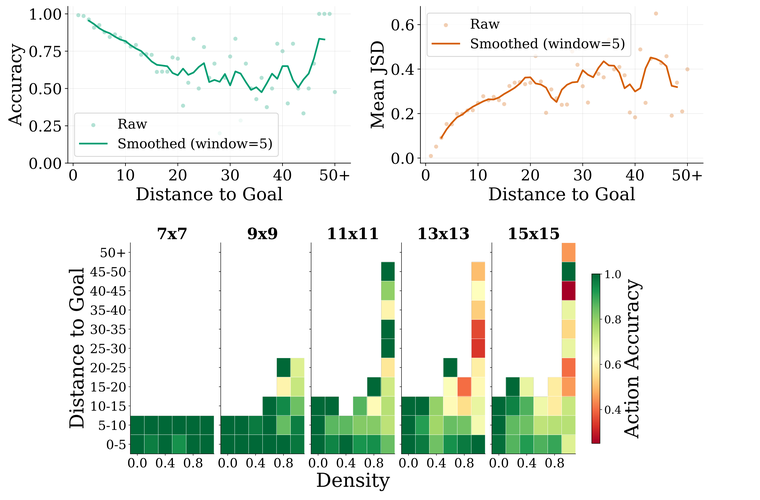


Figure 1. Behaviour degrades predictably with distance to goal, grid size, and obstacle density.

Performance tracks task difficulty in a predictable way.

Robustness to Transformation

We test whether behaviour depends on incidental surface properties by applying 4 difficulty-preserving transforms: reflection, 90° rotation, start–goal swap, and transposition. Across transformations, we find **no statistically significant differences** in the main behavioural metrics.

Interpretation. The agent is sensitive to task-relevant structure rather than to particular visual or positional regularities of specific grids.

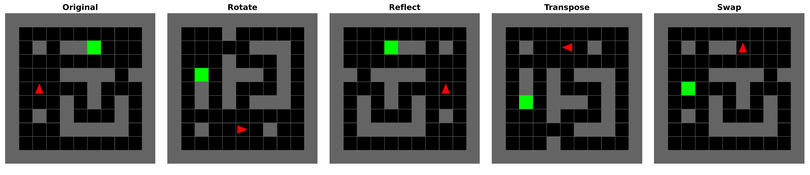


Figure 2. Examples of difficulty-preserving transformations.

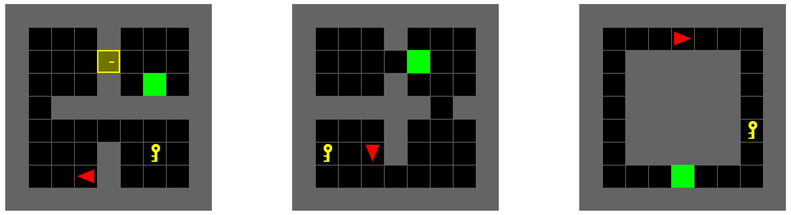
Instrumental and Implicit Goals

We test more complex goal structures:

(a) *KeyDoorEnv*: collect a key to unlock a door blocking the goal route.

(b) *KeyNoDoorEnv*: the key is present but functionally irrelevant.

(c) *2PathKeyEnv*: one of two optimal paths contains a vestigial key.



(a) *KeyDoorEnv* **(b)** *KeyNoDoorEnv* **(c)** *2PathKeyEnv*

Figure 3. Instrumental and implicit goal environments.

Environment	Success (%)	Accuracy (%)
KeyDoorEnv	100.0	98.7
KeyNoDoorEnv	98.9	97.2
2PathKeyEnv (with key)	71.4	76.0
2PathKeyEnv (without key)	75.5	74.3

Table 1. Simplified summary of instrumental and implicit goal results.

In *KeyDoorEnv*, the agent succeeds reliably and handles the instrumental subgoal. In *KeyNoDoorEnv* and *2PathKeyEnv*, it is biased toward the key even when the key is task-irrelevant.

The agent handles instrumental goals well, but is also distracted by goal-like, non-functional cues.

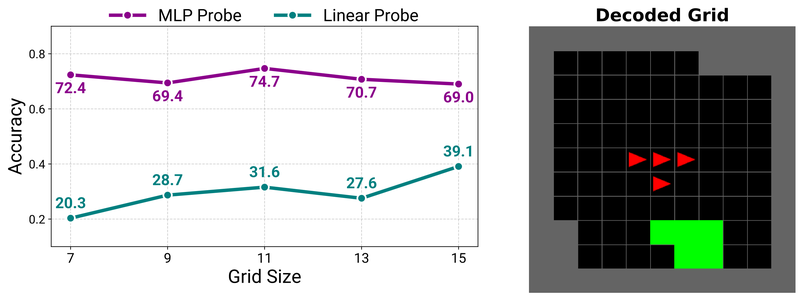
Representational Evaluation

We probe residual-stream **activations before and after reasoning** to decode the agent's latent beliefs about the current grid.

Inputs: activations plus queried (x, y) coordinate

Labels: `agent`, `goal`, `wall`, `open`, `padding`

Probes: linear and 2-layer MLP classifiers



Key finding. The model non-linearly encodes a **coarse cognitive map** of the environment. MLP probes substantially outperform linear probes; agent and goal locations are decoded with high recall, but decoded locations are approximate and spatially blurred.

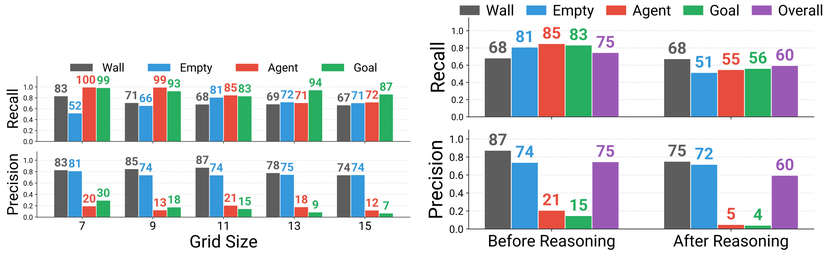


Figure 4. Left: Cognitive maps reconstructed from model activations reveal approximate task-relevant spatial information. **Right**: After reasoning, full-map decoding becomes weaker, indicating a shift away from broad environment structure.

Beliefs, Actions, and Plans

Actions vs. decoded beliefs. We compare the agent's observed actions with optimal actions computed on: (1) the **ground-truth grid**, (2) the **decoded cognitive map**.

A substantial fraction of actions that are suboptimal in the true environment become optimal relative to the decoded map. This suggests that many apparent failures are better explained by **imperfect internal beliefs** than by an absence of goal-directedness.

When we account for uncertainty using top-k decoded agent/goal positions, the decoded maps explain an even larger share of the agent's behaviour.

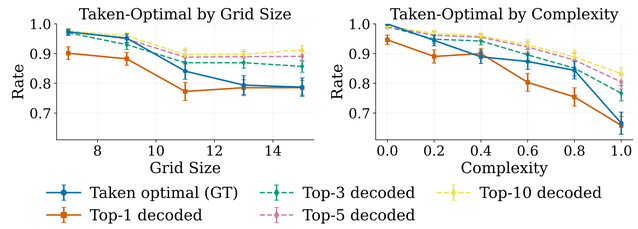


Figure 5. Decoded internal maps explain more behaviour once uncertainty over agent/goal locations is preserved.

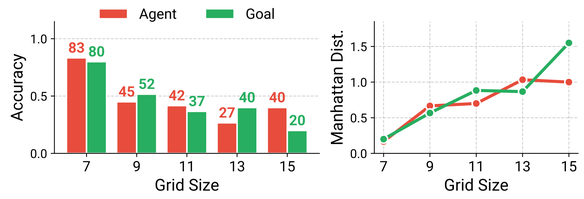


Figure 6. Decoded agent and goal locations remain approximate but spatially bounded as grids grow.

Many behavioural "errors" are sub-optimal relative to the true environment state, but optimal under the agent's subjective internal map.

Plan Decoding

We train a **one-shot Transformer decoder** to recover the next T=10 actions from a fixed set of activations extracted pre- and post-reasoning. One-shot decoding is meant to prevent the probe from "constructing" a plan autoregressively.

- Prefix accuracy above the 0.25^N baseline shows that non-trivial multi-step action information is present in the activations.
- **Post-reasoning**: best for **next-action** decoding.
- **Pre-reasoning**: better preserves **longer-horizon** structure.

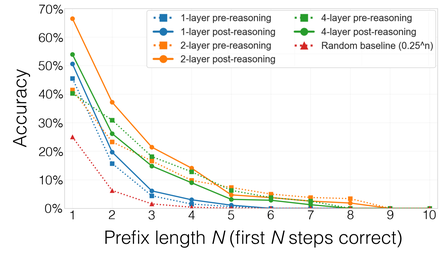


Figure 7. Plan information is decodable from internal activations, with distinct pre- vs. post-reasoning profiles.

Takeaways

- ⋆ Behavioural evaluation alone is not enough to characterise goal-directedness.
- ⋆ GPT-OSS-20B exhibits robust goal-directed behaviour in controlled navigation tasks.
- ⋆ Its internal states encode **coarse cognitive maps** and **multi-step plans**.
- ⋆ Reasoning shifts representations from broader environment structure towards next action selection.

A useful evaluation of goal-directedness must connect behaviour to internal representations.

Code: `github.com/SPAR-Telos/interp`

Interactive cognitive map viewer: `huggingface.co/spaces/project-telos/trace-viewer`