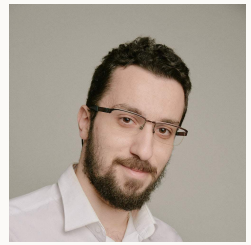
Sparks of Cooperative Reasoning

LLMs as Strategic Hanabi Agents

Mahesh Ramesh, Kaousheik Jayakumar, Aswinkumar Ramkumar, Pavan Thodima, Aniket Rege†, Manolis Vlatakis†





Motivation

"Cooperation is our Superpower" - Rich Sutton

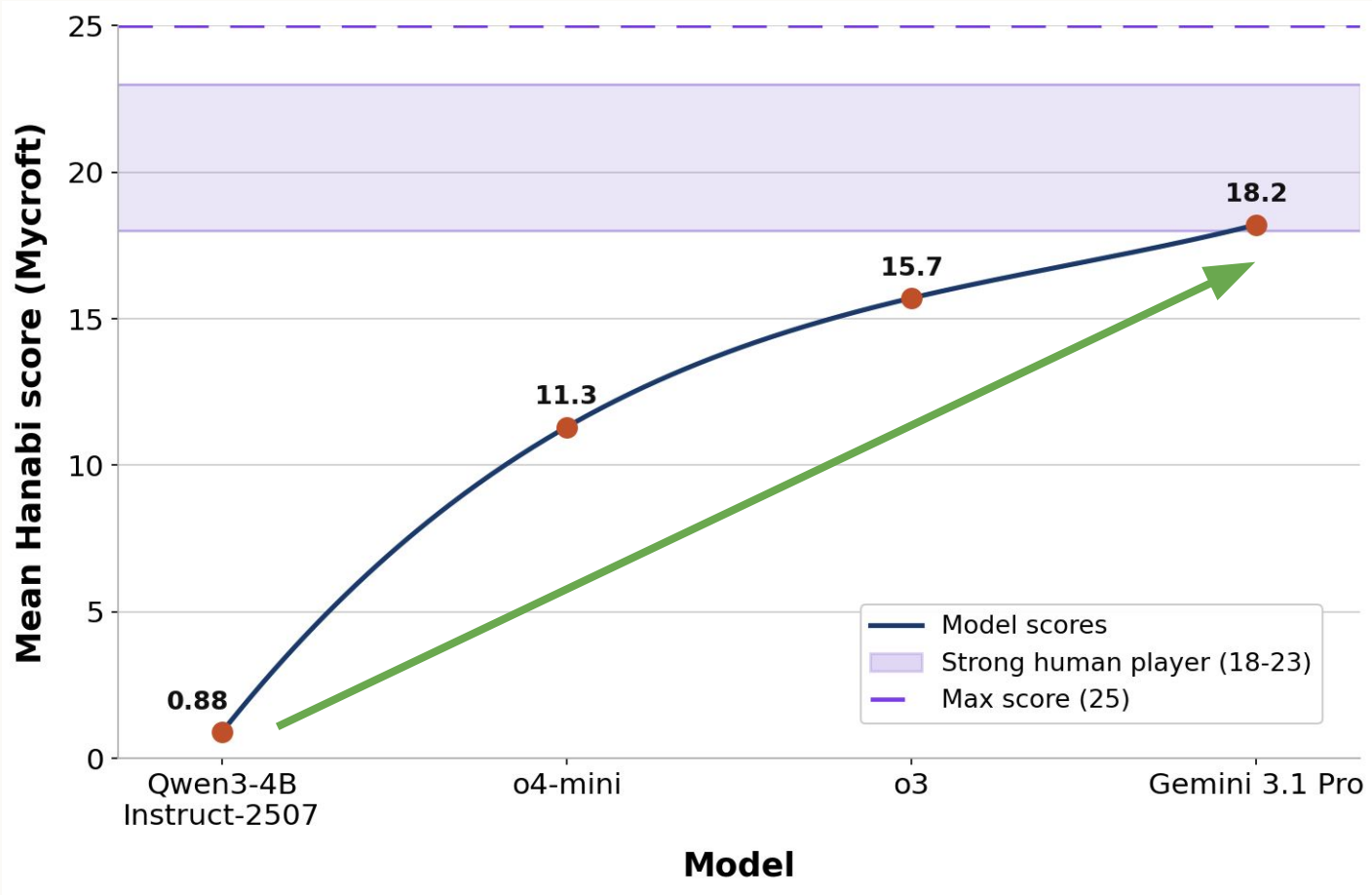
Modern AI agents must **cooperate** with humans and other agents

Why Hanabi?

- Needs Theory-of-Mind reasoning
- Interpreting sparse clues
- Tracking teammate beliefs
- Multi-turn coordination

Key Findings:

- LLMs show **sparks** of cooperative reasoning, but still **lag humans**
- The Good News: LLMs are **improving rapidly**!

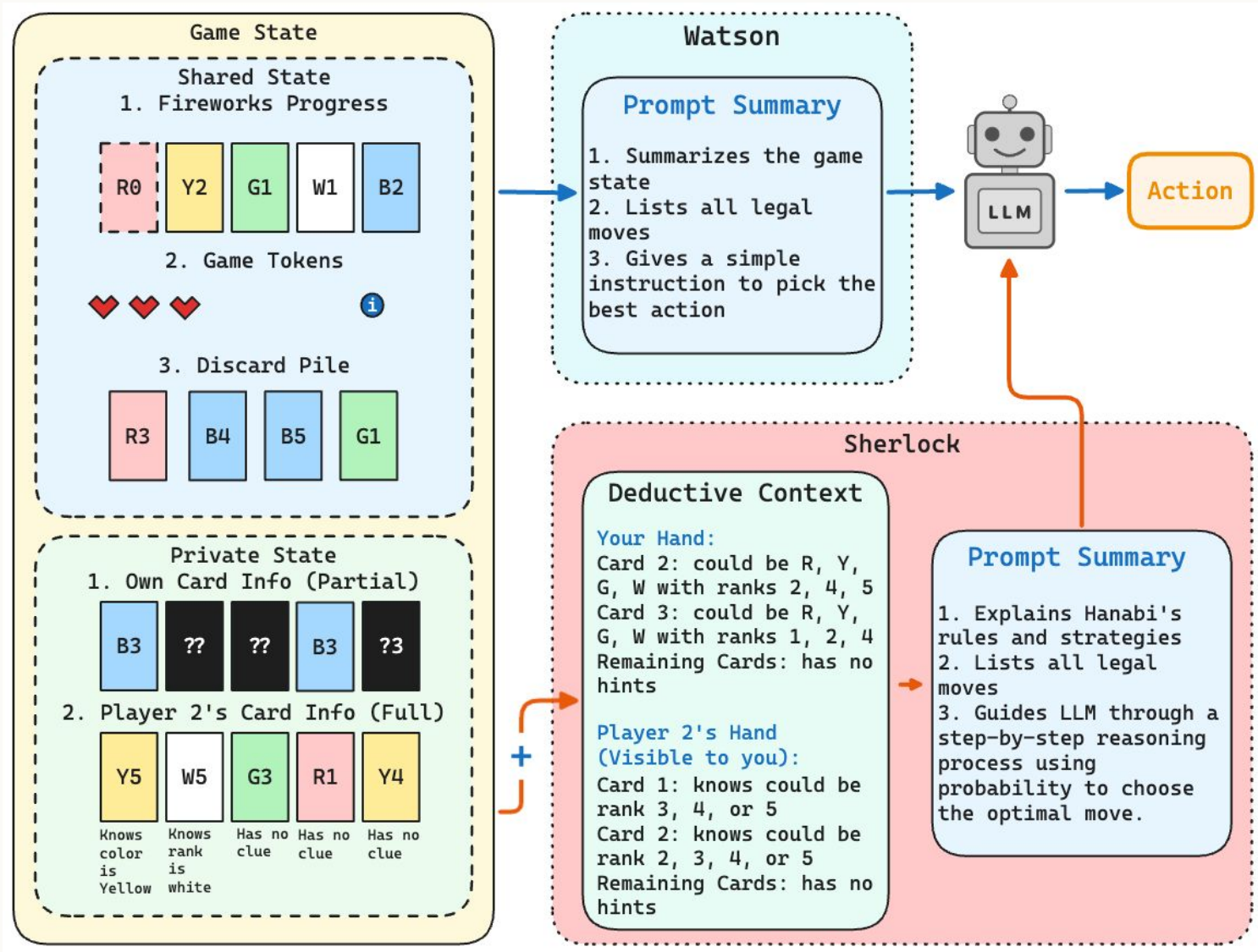


Three Holmesian Scaffolds

Scaffolds: test LLMs under increasing levels of external support / information

01 · Watson - Lower bound

Minimal context: game state, visible hands, explicit clue knowledge, and legal moves.



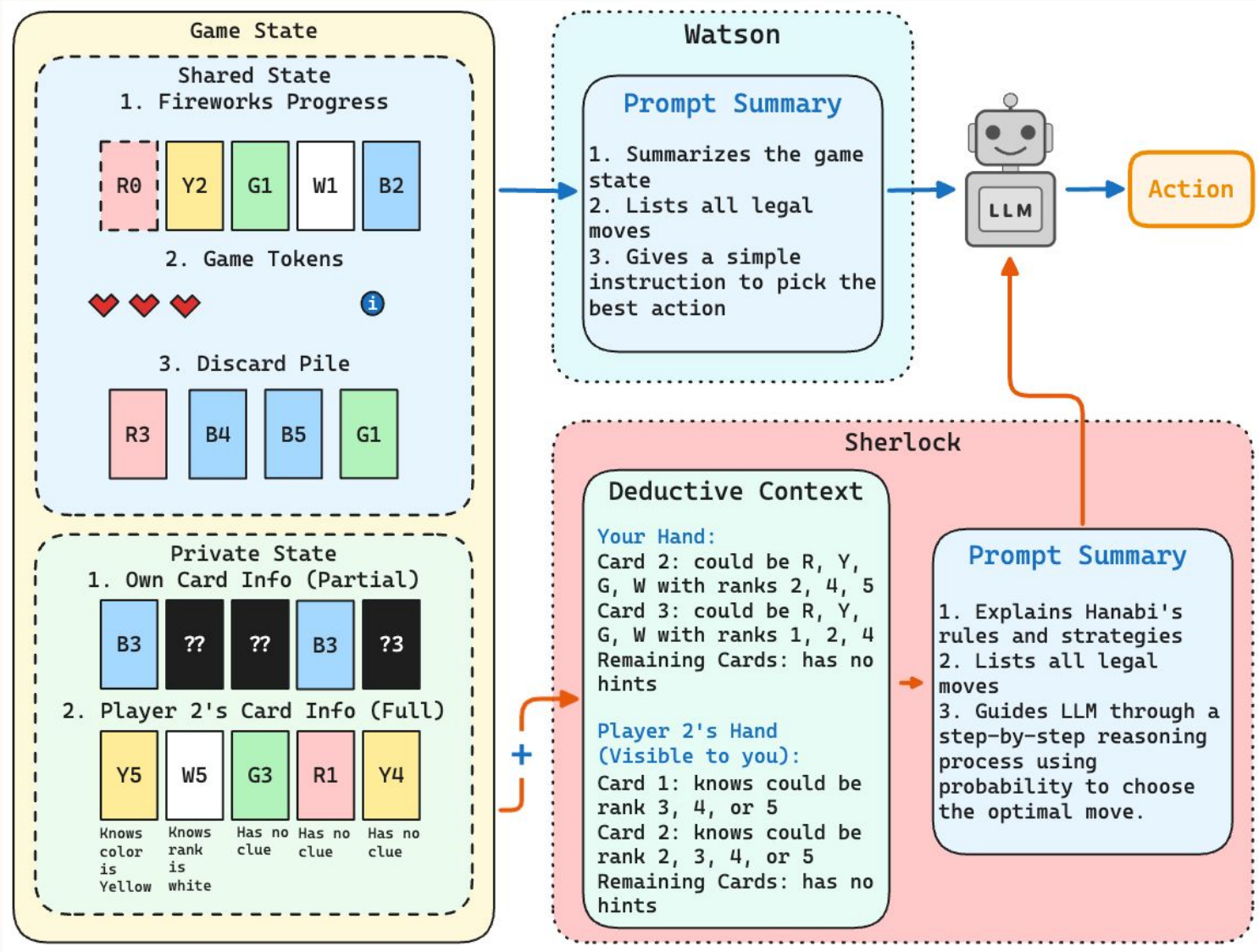
Three Holmesian Scaffolds

Scaffolds: test LLMs under increasing levels of external support / information

01 · Watson - Lower bound

02 · Sherlock - Upper bound

Adds engine-computed per-card candidate sets and encourages Bayesian-style reasoning.



Three Holmesian Scaffolds

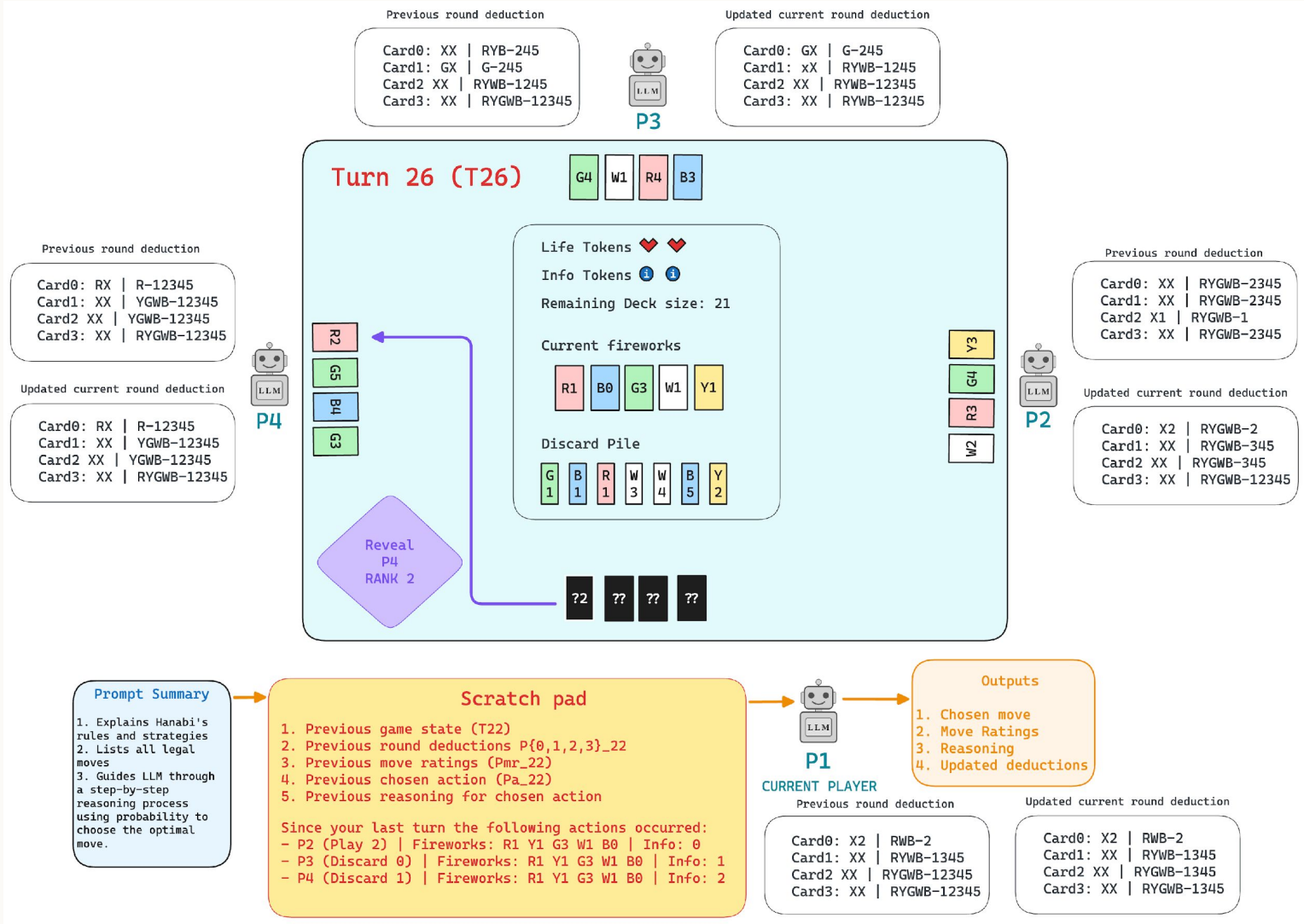
Scaffolds: test LLMs under increasing levels of external support / information

01 · Watson - Lower bound

02 · Sherlock - Upper bound

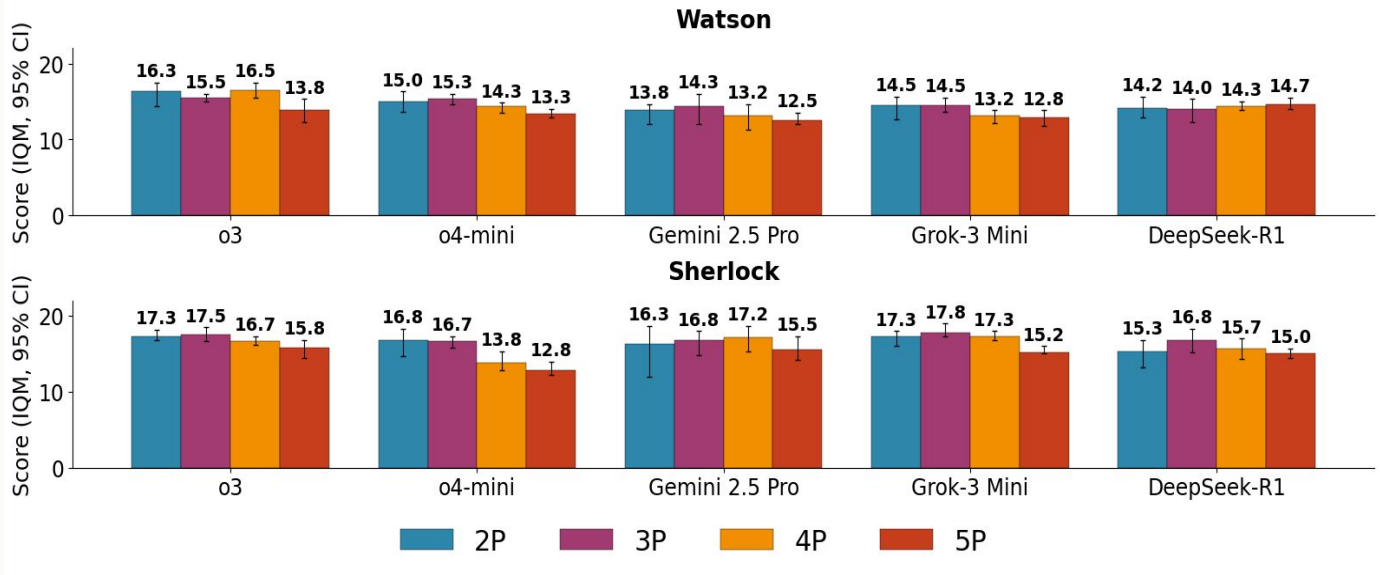
03 · Mycroft - State tracking

Removes engine deductions. LLM must update beliefs and card positions over 60+ turns.



Benchmark Setup

- We evaluate 17 LLMs across teams of 2-5 players.
- Each bar: Hanabi score out of 25 averaged across 10 fixed seeds

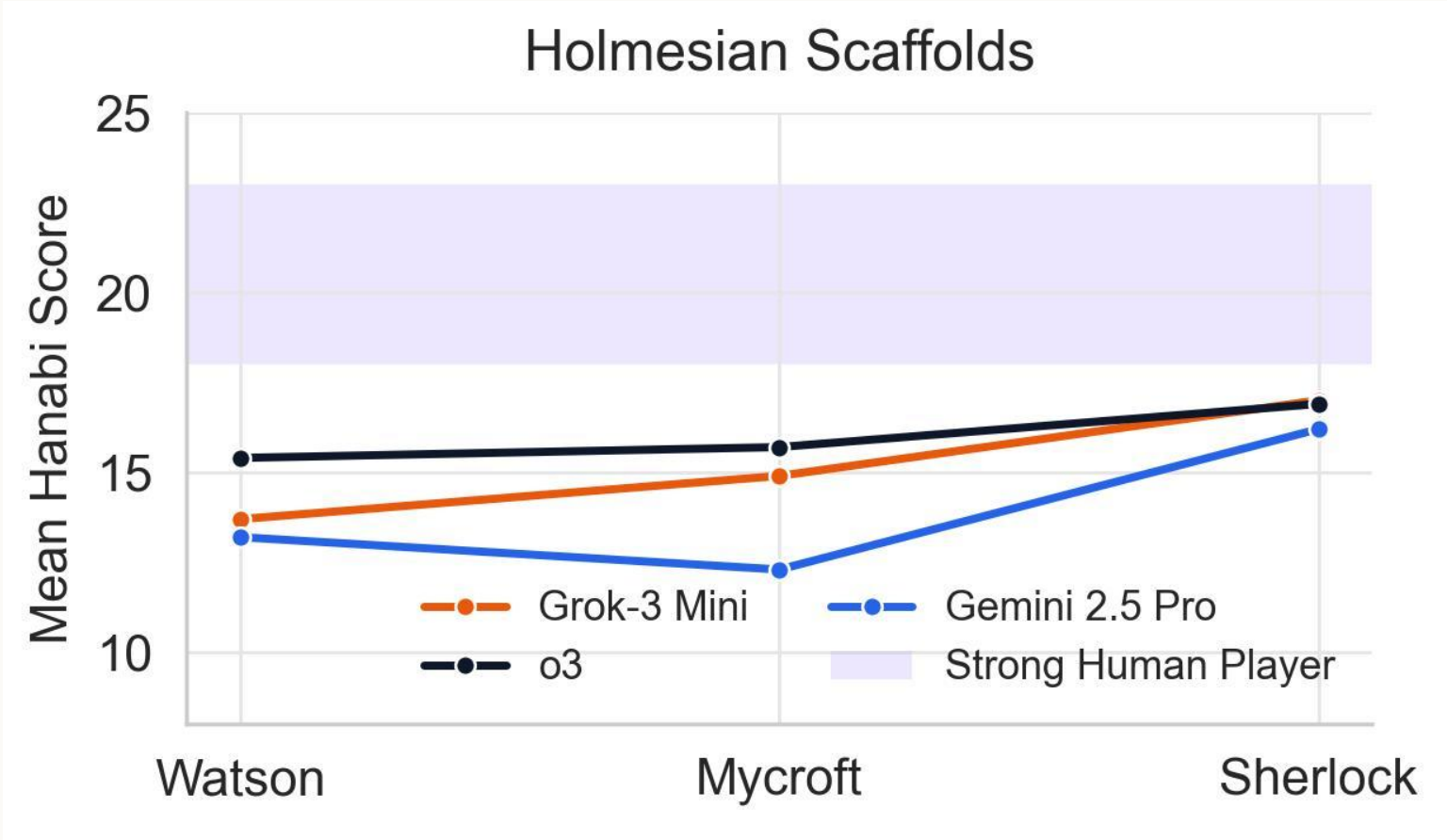


New Datasets for Post-training!

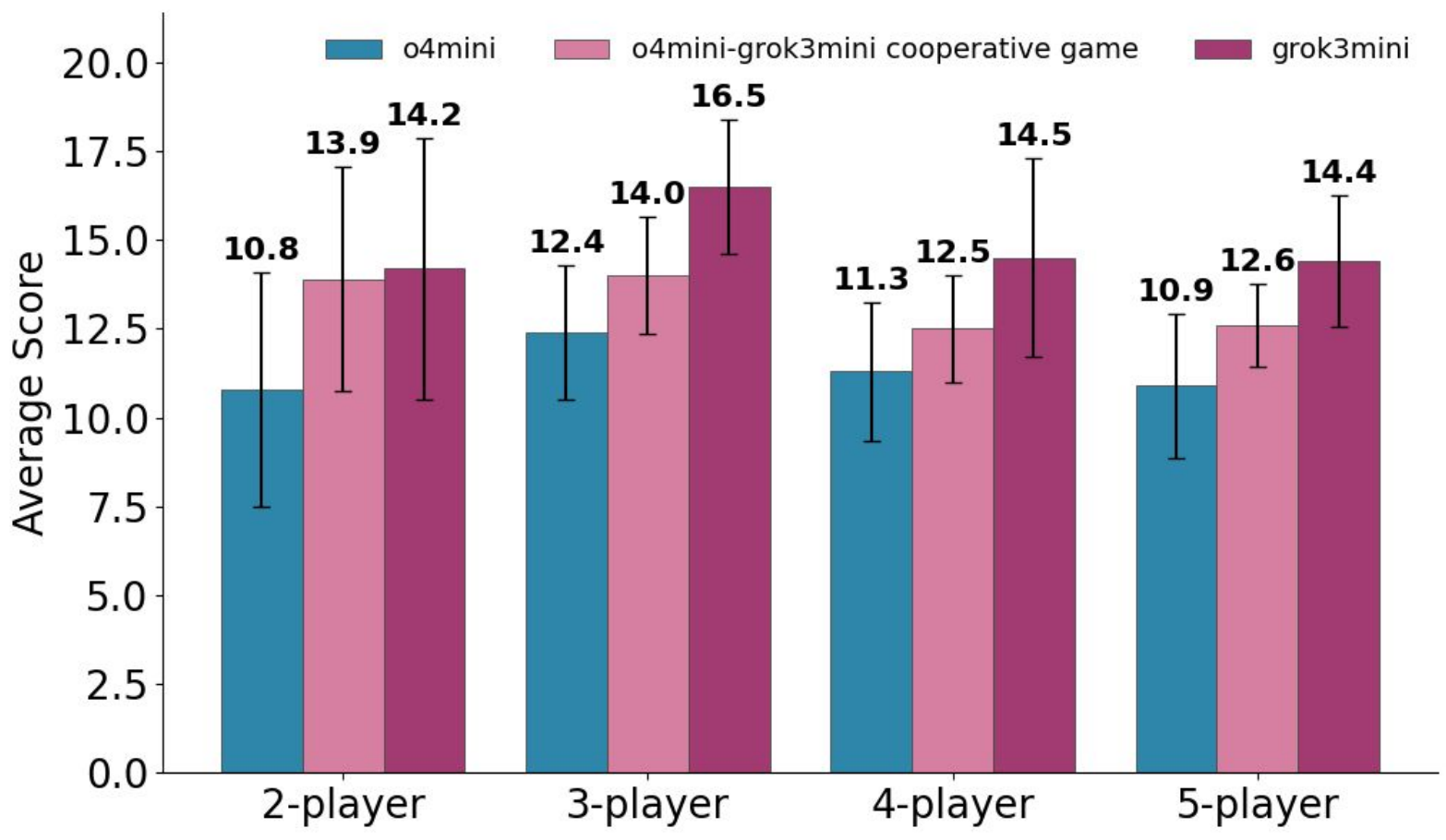
- **HanabiLogs**: full agent trajectories
  → 1,520 complete game logs for SFT
- **HanabiRewards**: move-level rewards (LLM-as-a-Judge)
  → 560 games for RL post-training

Finding: Scaffolding Helps, State Tracking is Hard

- Reasoning models: 15-18/25 in the strongest setting but still trail specialized RL agents and **strong humans**.
- LLMs are robust across 2-5 player settings and all our scaffolds.
- Performance drops when models must track belief states across long (60+) multi-turns.

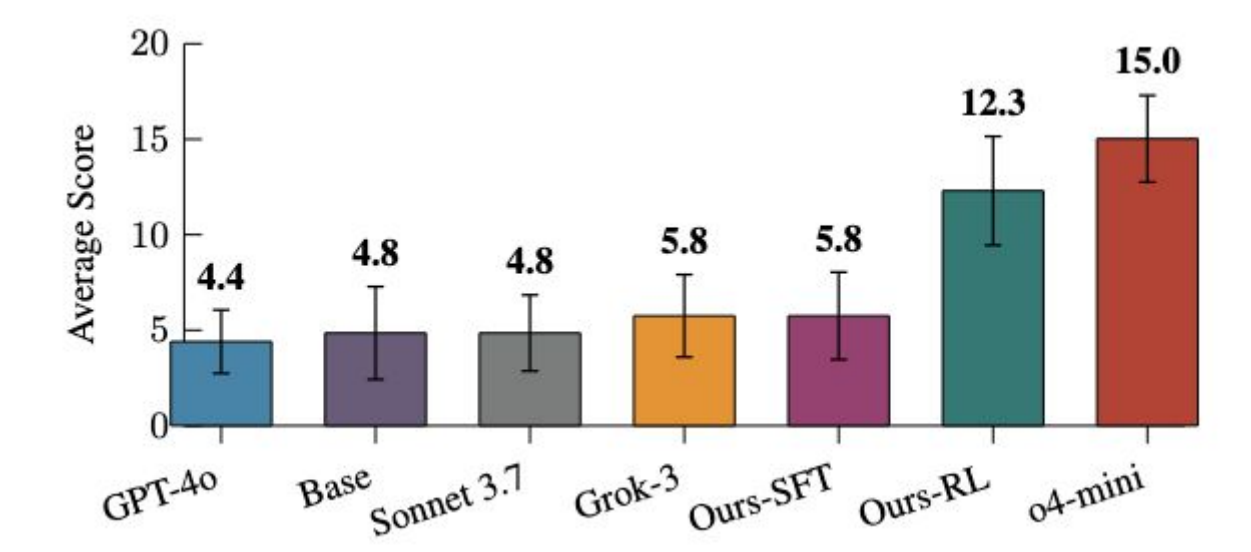


Finding: Cross-play Generalization

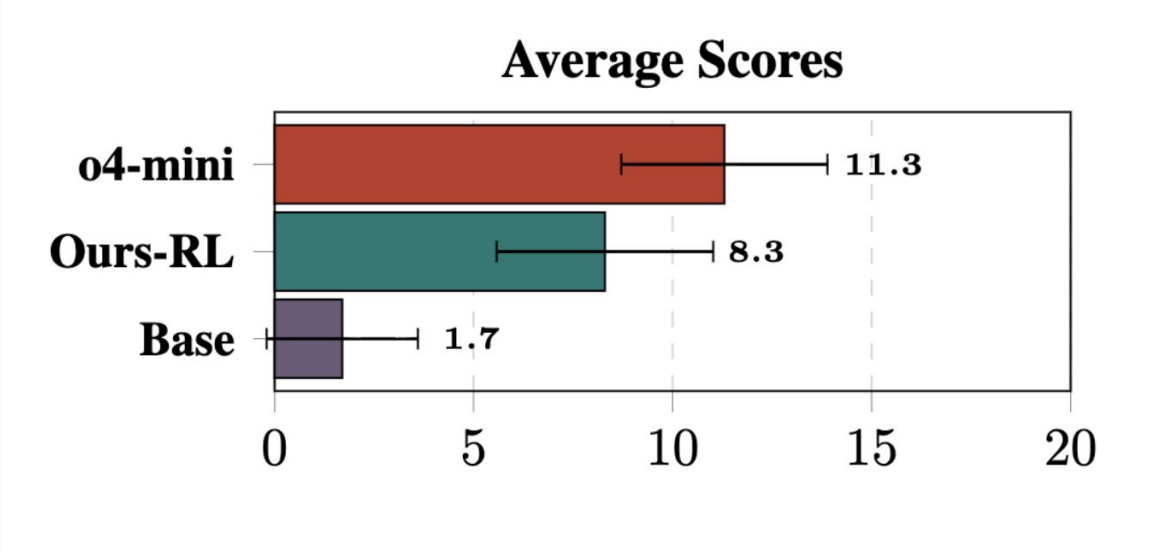


- Unlike classical RL agents LLMs are robust to different players.
- When the games is played between strong (Grok-3mini) and weak players (o4-mini) the scores interpolate between self-play scores.
- One strong Grok-3 Mini teammate lifts teams with o4-mini by about +1.7 points on average.

Finding: Post-training a 4B Model Closes the Gap



Sherlock post-training



Mycroft post-training

- **SFT** doesn't help much (**+1.0**)
- **RL post-training** has huge gains: **4.8 → 12.3** in Sherlock and **1.7 → 8.3** in Mycroft
- **RL post trained 4B model** beats all other non-reasoning models within three points of frontier models like **o4-mini**

Finding: Post-training on Hanabi Generalizes

Model	Group Guess (wins / 200, coop.)	EventQA 64K / 128K / 800K	IFBench Avg / Pass@10
Base	61.0 / 60.5	84.0 / 62.6 / 37.2	30.9 / 42.9
Ours-RL	**73.0 / 71.5**	**85.6 / 66.8 / 43.6**	**31.5 / 44.6**
Δ	**+12.0 / +11.0**	**+1.6 / +4.2 / +6.4**	**+0.6 / +1.7**

Cooperative transfer

Post-training improves model's cooperative reasoning ability under partial information

Finding: Post-training on Hanabi Generalizes

Model	Group Guess (wins / 200, coop.)	EventQA 64K / 128K / 800K	IFBench Avg / Pass@10
Base	61.0 / 60.5	84.0 / 62.6 / 37.2	30.9 / 42.9
Ours-RL	**73.0 / 71.5**	**85.6 / 66.8 / 43.6**	**31.5 / 44.6**
Δ	**+12.0 / +11.0**	**+1.6 / +4.2 / +6.4**	**+0.6 / +1.7**

Cooperative transfer

Post-training improves model's cooperative reasoning ability under partial information

Long-horizon memory

EventQA gains grow with context length, highlights temporal reasoning transfer.

Finding: Post-training on Hanabi Generalizes

Model	Group Guess (wins / 200, coop.)	EventQA 64K / 128K / 800K	IFBench Avg / Pass@10
Base	61.0 / 60.5	84.0 / 62.6 / 37.2	30.9 / 42.9
Ours-RL	**73.0 / 71.5**	**85.6 / 66.8 / 43.6**	**31.5 / 44.6**
Δ	**+12.0 / +11.0**	**+1.6 / +4.2 / +6.4**	**+0.6 / +1.7**

Cooperative transfer

Post-training improves model's cooperative reasoning ability under partial information

Long-horizon memory

EventQA gains grow with context length, highlights temporal reasoning transfer.

Instruction-following

On IFBench demonstrates better instruction following.

Takeaways

- Models show sparks of cooperative reasoning, robust behavior across different player settings and cross-play, but still lack behind Humans.
- Datasets: We introduce two new datasets, HanabiLogs for SFT and HanabiRewards for RL fine tuning.
- RL finetuning: Post training models on HanabiRewards helps significantly reaching within three points of the frontier models like o4-mini and generalizes to other reasoning tasks like Group-Guessing game, EventQA and IFbench.

tl;dr → Model are rapidly improving in Theory of Mind, are robust across player setup and cross-play. Post training on our Hanabi data Generalizes to other reasoning tasks.

Thank You!

Check out our Project Page: