

Benchmarking Open-Ended Multi-Agent Coordination in Language Agents

Kale-ab Tessera, Andras Szecsenyi, Cameron Barker, Alexander Rutherford,
Davide Paglieri, Aidan Scannell, Henry Gouk, Elliot J. Crowley,
Tim Rocktäschel, Amos Storkey

London
23/06/26

Motivation

LLM Agents – 2025

Google Cloud Contact sales

Blog

Demis Hassabis @demishassabis

Thrilled to kick off the Gemini 2.0 era with Gemini 2.0 Flash, an update to our workhorse model that outperforms even 1.5 Pro at twice the speed. It has really great multilingual skills, and can natively call tools, like Google Search. It's the first release in the Gemini 2.0 family of models, with more to come soon.

This is really just the beginning. **2025 will be the year of AI agents** and Gemini 2.0 will be the generation of models that underpin our agent-based work. We're sharing a set of prototypes made possible by 2.0 Flash's new capabilities: including an update to Project Astra, our vision for a universal AI assistant; the new Project Mariner, which explores the future of human-agent interaction, starting with your browser; and Jules, an AI-powered code agent that can help developers.

We're also sharing a few other east-navigate video games, which build breakthroughs in AI, and agents for

OpenAI

Sam Altman

We are now confident we know how to build AGI as we have traditionally understood it. We believe that, **in 2025, we may see the first AI agents "join the workforce" and materially change the output of companies.** We continue to believe that iteratively putting great tools in the hands of people leads to great, broadly-distributed outcomes.

Try in Playground

Microsoft

AI Agents Hackathon

April 8 - April 30, 2025

Build, Innovate, and #Hacktogether!

2025 is the year of AI agents But what exactly is an agent, and how can you build one? Whether you're a seasoned developer or just starting out, this **FREE three-week virtual hackathon** is your chance to dive deep into AI agent development.

Learn from **20+ expert-led sessions** streamed live on YouTube, covering top frameworks like **Semantic Kernel, Autogen, the new Azure AI Agents SDK and the Microsoft 365 Agents SDK**

ANTHROPIC

Engineering at Anthropic

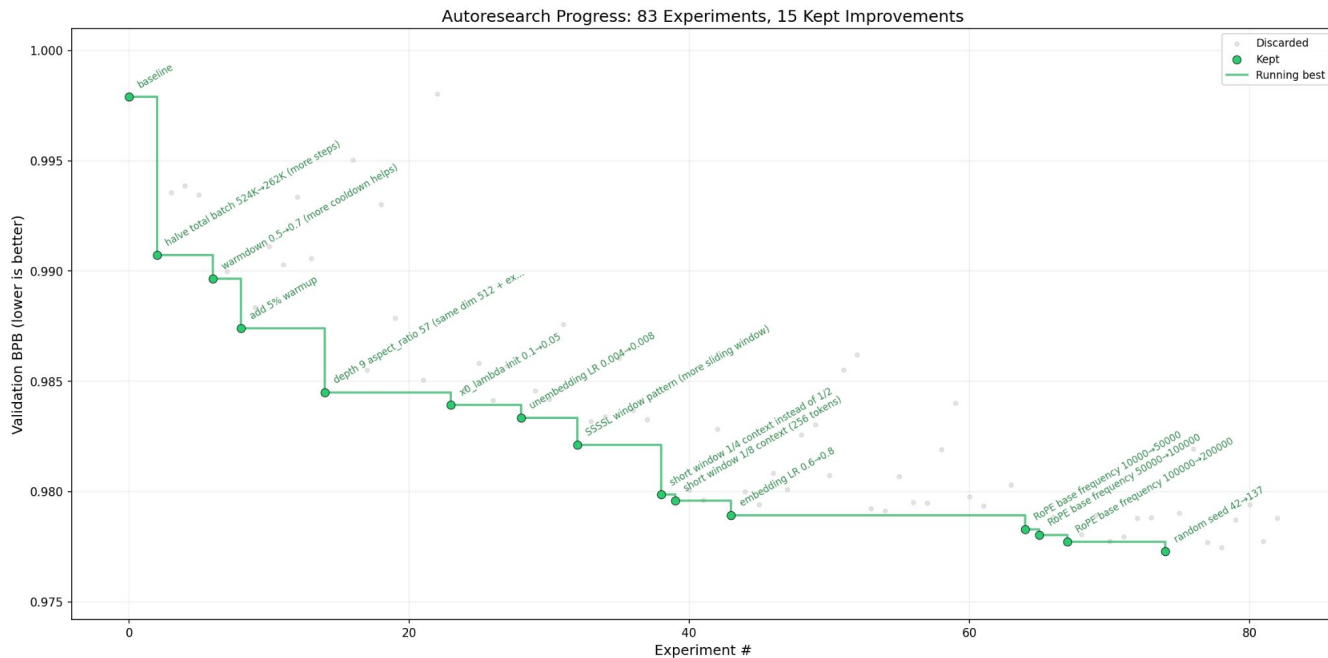


Building effective agents

AI agents have been making significant strides in their capabilities, driven by advancements in **artificial intelligence** technologies. With a focus on enhancing their utility, companies and researchers are continuously exploring ways to improve these intelligent systems. **By 2025, it is expected that AI agents will demonstrate substantial improvements, including better tool usage, enhanced contextual understanding, improved coding assistance, and strengthened safety measures** as highlighted by **Anthropic's** chief scientist Jared Kaplan.

(Open-ended) Automated AI Research – 2026?

LLM training

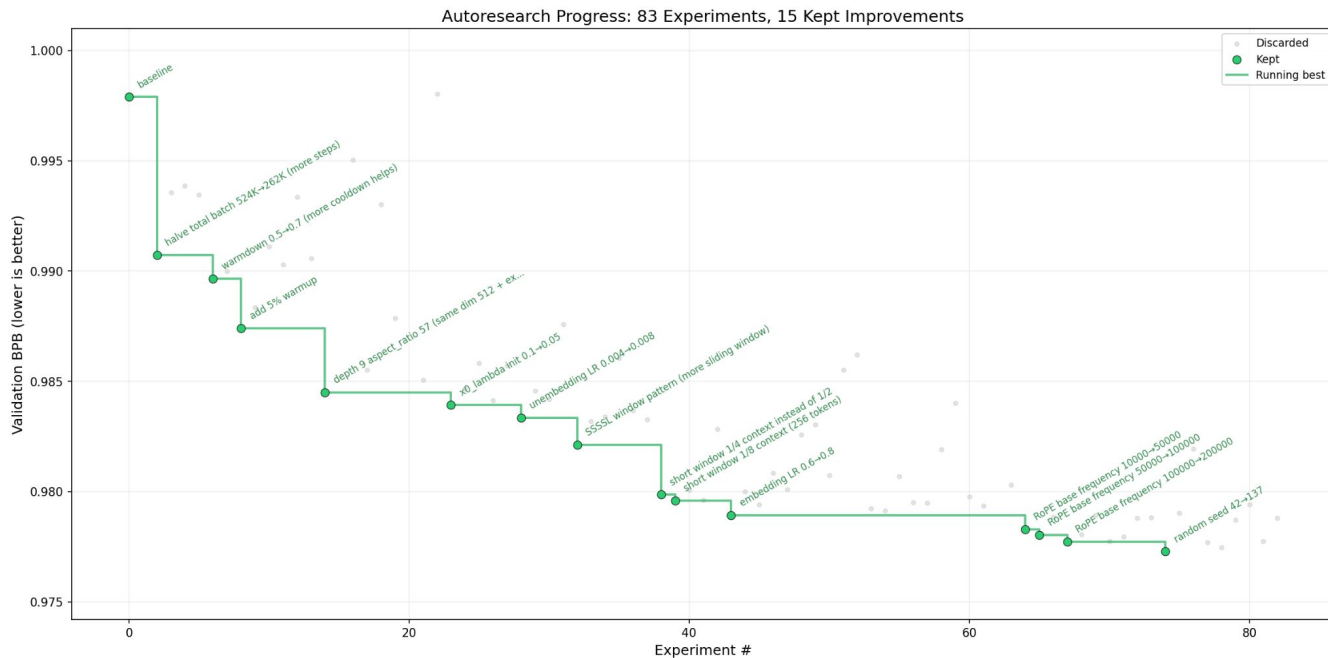


While not done:

1. edit train.py
2. train for 5 minutes
3. evaluate solution
4. keep or discard

Builds on long lineage of AutoML, NAS - 1990s

(Open-ended) Automated AI Research – 2026?



Hparams:

- DEPTH
- DEVICE_BATCH_SIZE
- TOTAL_BATCH_SIZE
- EMBEDDING_LR
- MATRIX_LR
- WEIGHT_DECAY
- WINDOW_PATTERN
- HEAD_DIM
- WARMUP_RATIO
- WARMDOWN_RATIO

Training code itself!

- E.g add warmup

Others - NanoGPT Speedrun (harder, human baselines)

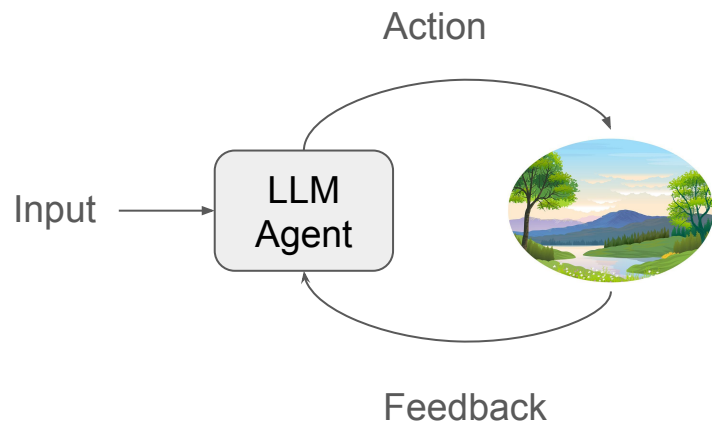
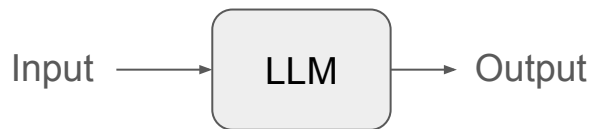
[Karpathy NanoChat Autoresearch](#)

(Optional Prelim in Agents)

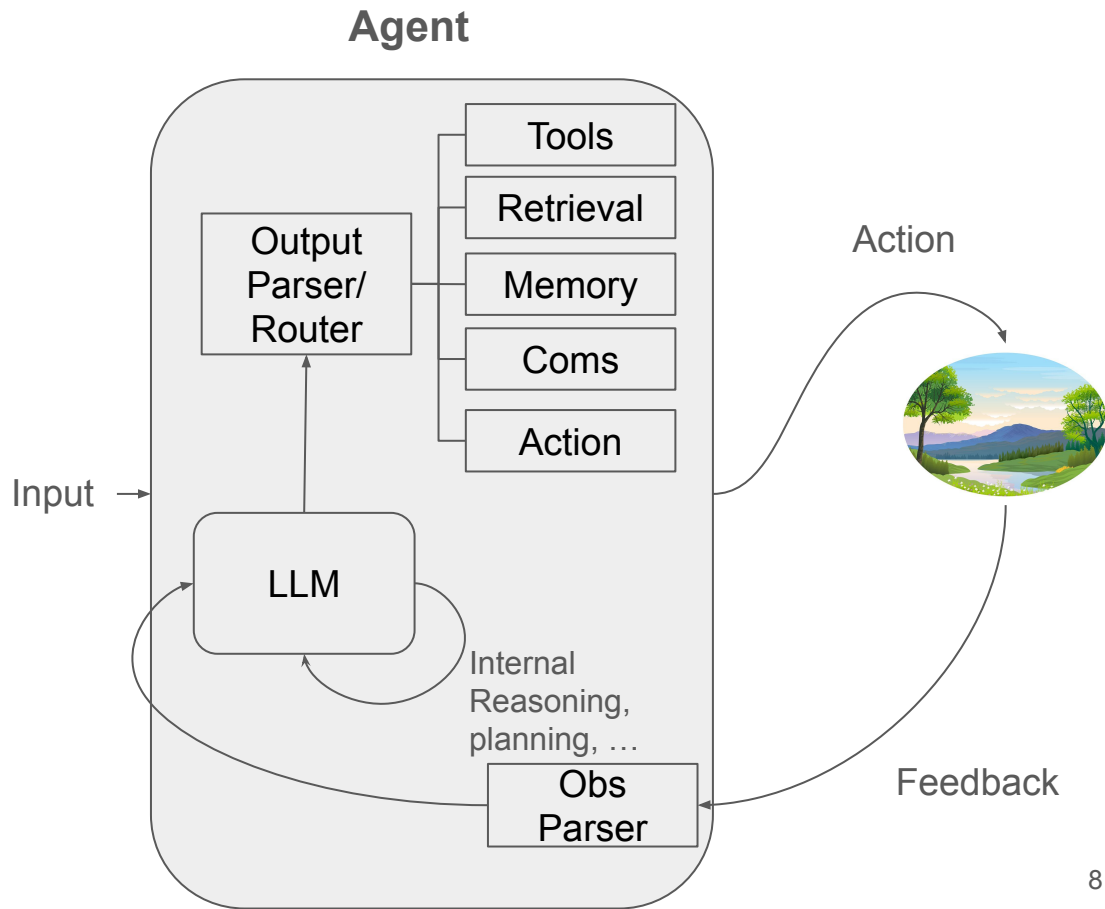
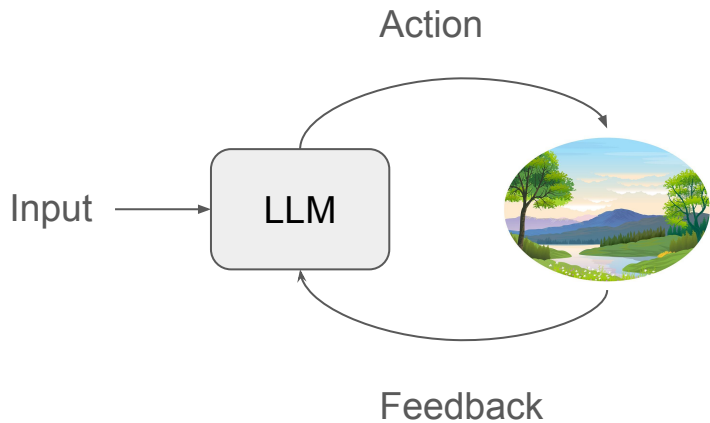
LLMs



LLM Agents



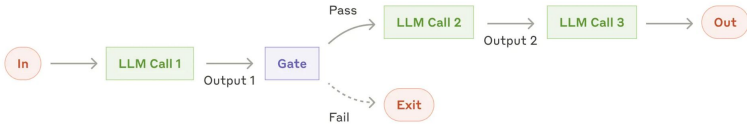
LLM Agents



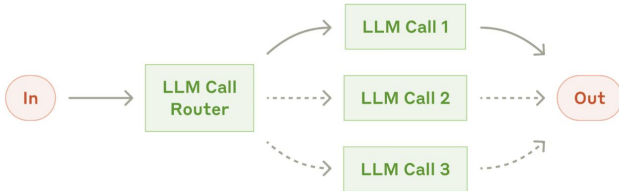
Workflows vs Agents

AI Workflows - **predefined** code paths.

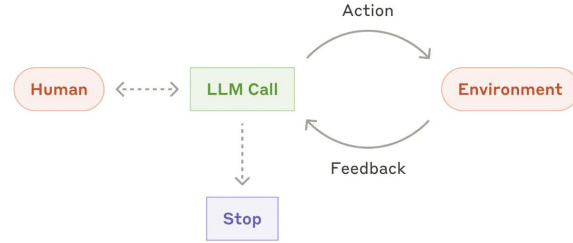
Prompt chaining e.g. write text, then translate.



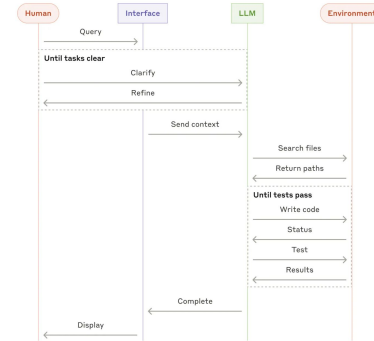
Routing, e.g. small models simple tasks, larger ones – more complicated tasks.



Agent AI - LLMs dynamically direct their own processes, control **how (e.g. choose tools/actions)** they accomplish tasks. Learn from env feedback.



E.g. Coding agent.



Can you draw full flowchart of the process (num of steps, what the steps are, which tools etc) ? Yes (workflow), No (AI Agent)

Agency Spectrum

The spectrum of “Agency”

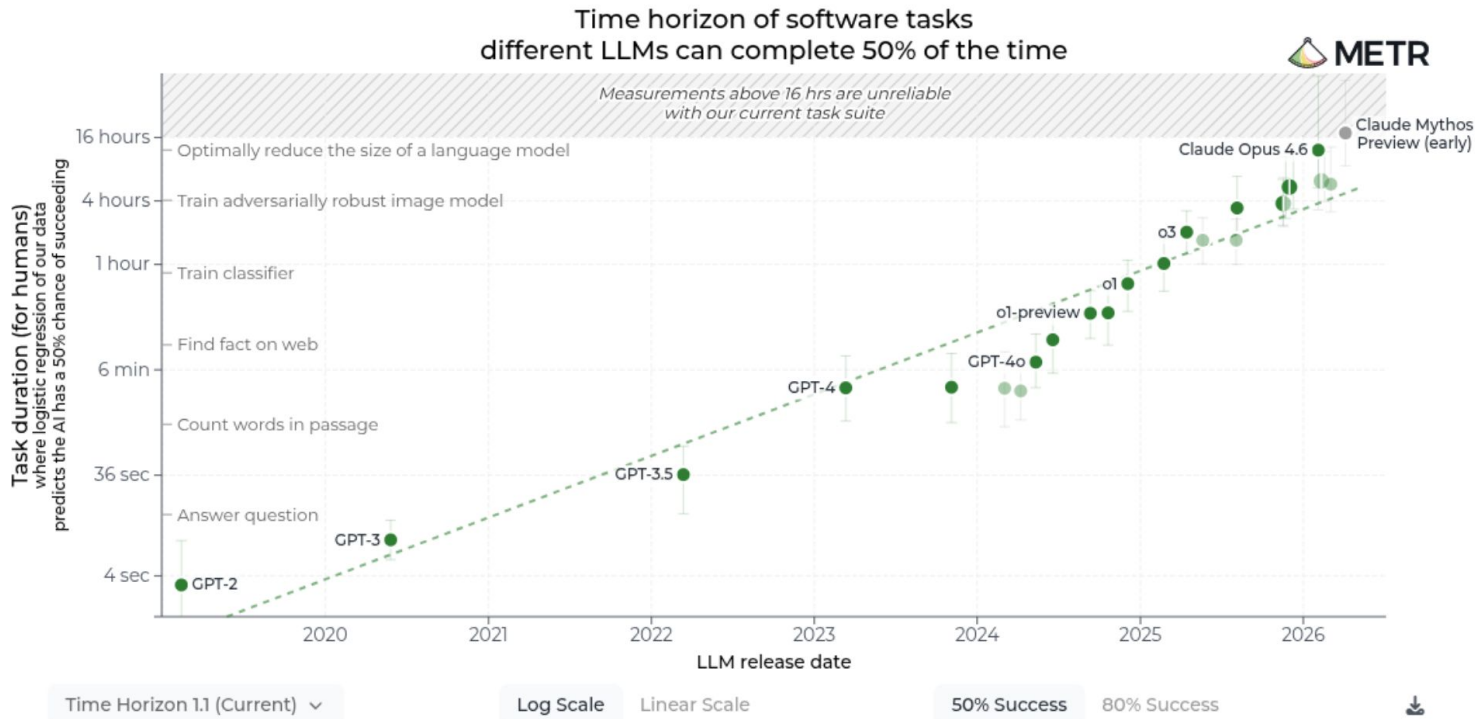
Following this definition, Agents exist on a continuous spectrum of increasing agency:

Agency Level	Description	What that's called	Example pattern
☆☆☆	Agent output has no impact on program flow	Simple processor	<code>process_llm_output(llm_response)</code>
☆☆☆	Agent output determines basic control flow	Router	<code>if llm_decision(): path_a() else: path_b()</code>
☆☆☆	Agent output determines function execution	Tool caller	<code>run_function(llm_chosen_tool, llm_chosen_args)</code>
☆☆☆	Agent output controls iteration and program continuation	Multi-step Agent	<code>while llm_should_continue(): execute_next_step()</code>
☆☆☆	One agentic workflow can start another agentic workflow	Multi-Agent	<code>if llm_trigger(): execute_agent()</code>

Motivation Continued

Long-horizon & Open-ended Tasks

Measured by how long they take human professionals.



*Long-horizon tasks

Kwa et al. Measuring AI Ability to Complete Long Tasks.

Long-horizon & Open-ended Tasks

Open-ended Tasks (AI Scientist V2, MLE Bench, ...) - sequences of artifacts

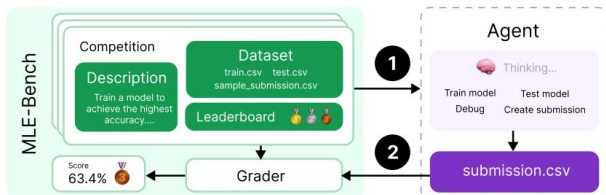
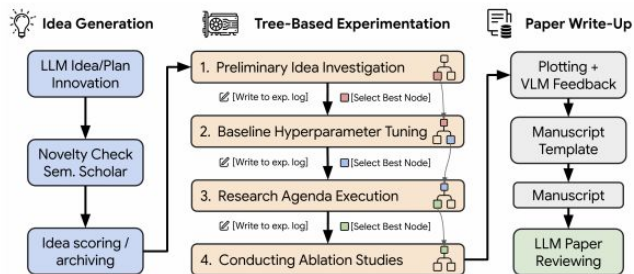
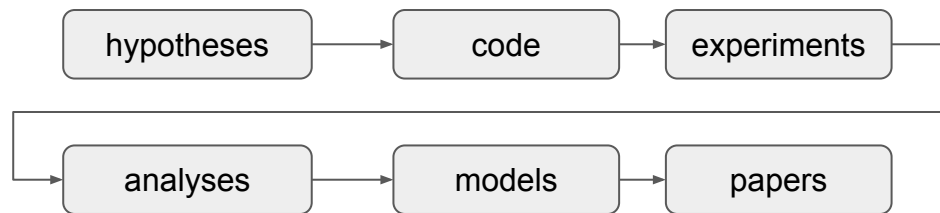


Figure 1: MLE-bench is an offline Kaggle competition environment for AI agents. Each competition has an associated description, dataset, and grading code. Submissions are graded locally and compared against real-world human attempts via the competition's leaderboard.

Sequences of artifacts



Goals = <hypothesis, code, experiments, analyses, models, papers,...>

Not fully open-ended as defined as by Hughes at al.
- **novel** and learnable.

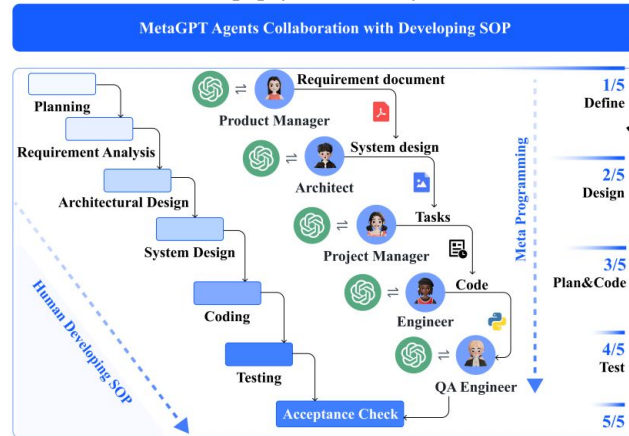
***Long-horizon & Open-Ended Tasks**

Multi-Agent Agentic Systems



LLM Debate can improve reasoning [1,3,4]

Specialisation/Division of labour improves performance on code generation tasks + reduces hallucinations [5] (*workflow*)



Multi-Agent Agentic Systems

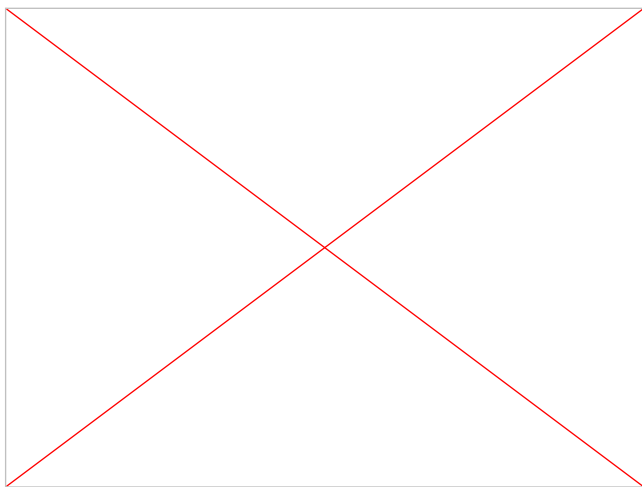
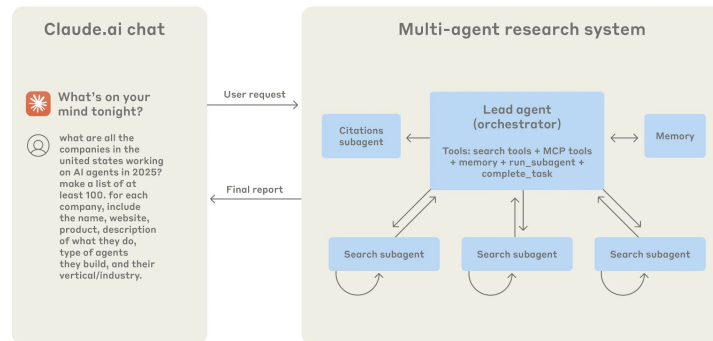
Claude Research:

“...multi-agent system with Claude Opus 4 as the lead agent and Claude Sonnet 4 subagents outperformed single-agent Claude Opus 4 by

90.2% ...”

- [anthropic](#) blog

High-level Architecture of Advanced Research



Google AI Co-Scientist:

- LLMs can benefit from role-specialized multi-agent workflows
- generate diverse ideas, critique them, rank them, evolve them, and synthesize the best ones.

Summary

As foundation models (FMs) are increasingly deployed as autonomous **agents** in multi-agent settings, there's a critical need to evaluate their **coordination** in **long-horizon** and **open-ended** tasks.

How to measure coordination in open-ended tasks?

Environment

Env requirements

1. ***Long-horizon***
2. ***Open-Ended Tasks:*** e.g. Goals = <hypothesis, code, experiments, analyses, models, papers,..>, see **video**.
3. ***Diverse Coordination objectives***
4. ***Controllable & Measurable Difficulty:*** split coordination in low, medium and high coordination complexity (enables systematic evaluation as coordination demands and incentive misalignment scale up).

Env requirements

Environment	Long horizon	Bounded open-endedness	Explicit coord. requirements	Coord. difficulty control
SMAC / SMACv2/ SMAX [45, 15, 44]	✗	✗	~	✗
Overcooked / Overcookedv2 [10, 17]	✗	✗	~	✗
Hanabi [4]	✗	✗	~	✗
Melting Pot [1, 33]	~	✗	~	✗
Concordia [51, 59]	✗	✗	~	✗
HECOGrid [34]	✗	✗	✓	✓
GPUDrive [30]	✗	✗	✗	✗
Neural MMO [52, 53]	✓	✓	~	✗
Multi-Agent Craftax v1 [67]	✓	✓	~	✗
Craftax-Coop [40]	✓	✓	~	✗
ALEM	✓	✓	✓	✓

* We focus on cooperative/coordination properties here.

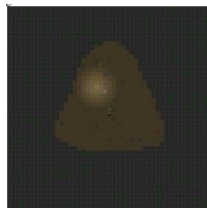
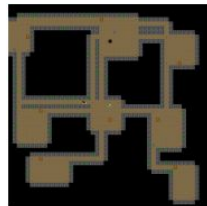
Craftar Worlds



Crafter (Hafner, 2022)



Craftax-Classic



Craftax (9 levels, Jax Matthews, 2024)



Multi-Agent Craftax (Ye et al, 2025, **Jax, Multiple Agents**)



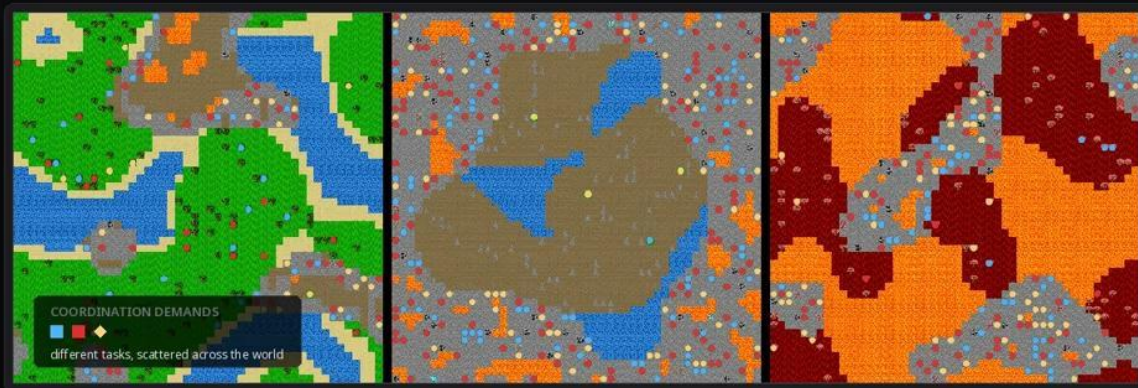
Figure 1: Example pixel-based observation of *Craftax-Coop* with a summary of player specializations. We also provide a symbolic observation to focus research on multi-agent challenges.

Multi-Agent Craftax v2 (Bassel et al, 2025, roles)

alem: An Open-Ended World for Multi-Agent Coordination

ALEM

OPEN-ENDED MULTI-AGENT COORDINATION



OPEN-ENDED MULTI-AGENT WORLDS

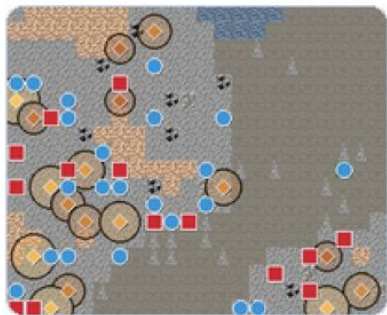
ALEM

Building on Craftax and Craftax-Coop, Alem brings controllable multi-agent coordination to open-ended JAX survival worlds.

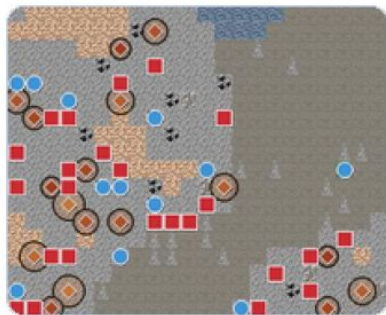
Procedural coordination tasks, soft specialisation, communication, and controllable difficulty inside long-horizon survival worlds.

alem / "world" in Amharic

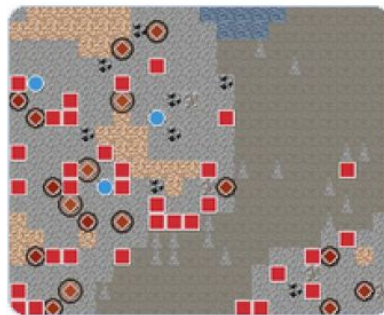
Easy/Med/Hard



Easy ($\alpha=0.30$)



Medium ($\alpha=0.60$)



Hard ($\alpha=0.90$)

Harder:

- More agents required for coordination tasks
- Less time for handover tasks
- Stricter specialisation
- Stronger mobs.

Alem induces measurable inter-agent dependencies

🔍 Multi-agent metrics recently proposed to measure statistical dependence between an *agent's history and its actions*, *teammates' private information* and between *agent's actions*.

Adding coordination tasks and soft specialisation strengthens all four measured behavioural dependencies.

Behavioural question	Single-agent	Multiple agent	MA + Soft spec.	ALEM
Do agents benefit from memory? (HAR)	0.355 ± 0.142	0.372 ± 0.022	0.437 ± 0.123	0.641 ± 0.064
Does teammate info. help predict actions? (PIF)	–	0.314 ± 0.036	0.384 ± 0.125	0.609 ± 0.148
Does synchronous coordination emerge? (AA)	–	0.017 ± 0.004	0.032 ± 0.011	0.061 ± 0.022
Does temporal coordination emerge? (DAI)	–	0.319 ± 0.057	0.406 ± 0.128	0.613 ± 0.165

*Higher values indicate **stronger behavioural dependence**, not higher reward.

Base vs Coord. Achievements



Category	Example	Num	Metric
Base	Collect Wood, Defeat Necromancer	66	% of the maximum achievable normal reward
Coord	Coordinated diamond mining, Build Beacon	27	% of the maximum achievable coord reward
Total	-	93	% of the maximum achievable total reward

Agents Playing the Game

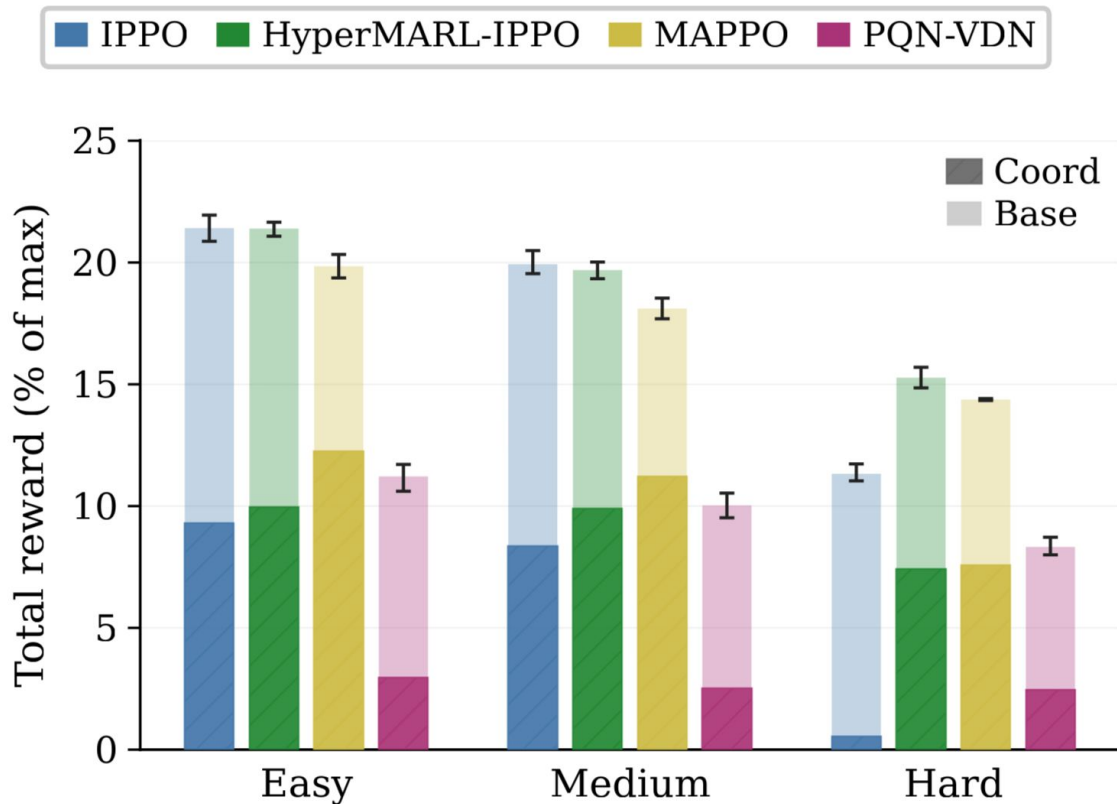


Results

Prelim 1 - MARL Results - 1 Billion Train Steps

RL Interface:

- Symbolics obs - observation dim size **9730** (local maps, inventory, teammate directions).
- **60** discrete actions - communication, request resources, etc. With legal action mask **~20** actions.



*Bar plots are rescaled onto a shared denominator, CI only on total in plot, see paper for full CI intervals.

Prelim 1 - MARL Results - 3B - Train Longer?

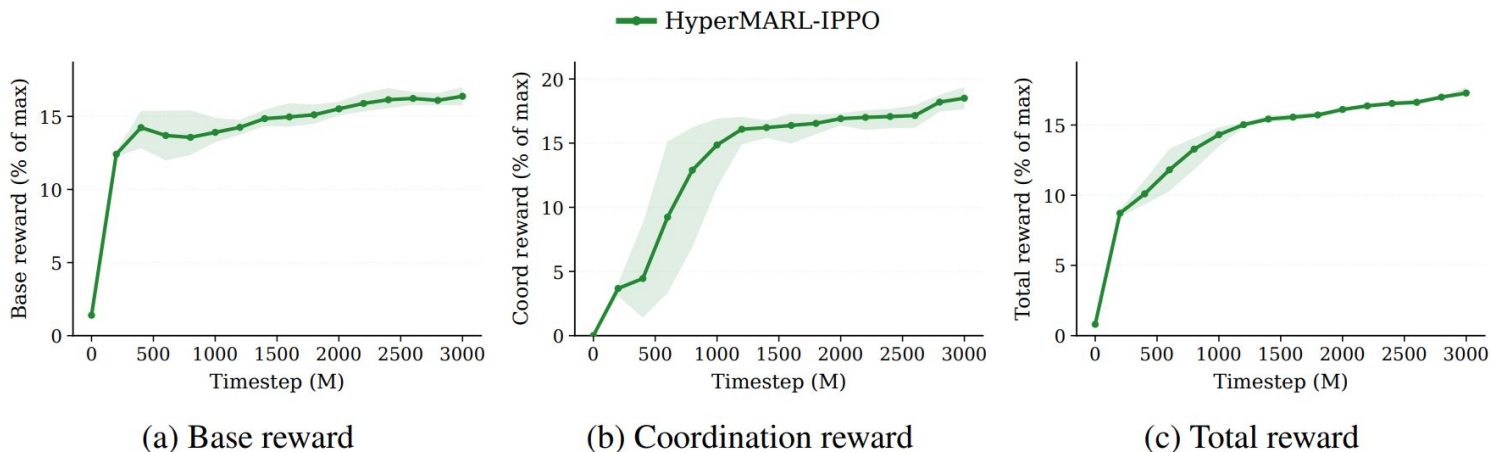


Figure 15: Extended MARL ablation on hard ALEM. HyperMARL-IPPO is trained for 3B environment steps across five seeds. Curves show seed means and shaded bands show 95% bootstrap confidence intervals. Performance does not saturate: coordination and total reward continue improving late in training, and total reward remains well below the maximum achievable score.

***Challenging for MARL Agents!**

LLM Interface

System prompt given to the LLM

```
system

You are Agent 1 (forager) in a 3-agent cooperative survival game. Your goal is to gather resources, craft gear, fight monsters, and descend through 9 dungeon levels, while coordinating with teammates. You must survive -- if your health reaches zero, you die, and if all agents die the game ends. Maximize the number of achievements while staying alive.

<game_rules>

## How to play
- Each turn, choose exactly one action.
- **Movement** uses absolute directions: north, south, east, and west. Any move attempt changes your facing to that direction, even if the move is blocked and you stay in place. A move is blocked if the target tile is solid, including trees, stone, ore veins, walls, crafting stations, chests, and plants, or if it contains water, lava, a mob, or another player. If repeated move attempts do not change your position, that direction is blocked. You can also use a blocked move to turn in place, for example to face an adjacent tree.
- **Facing** is your facing direction is set by your last movement action and persists until you move again. **Do** always targets the tile in your current facing direction.
- **Do** is your main interaction: face a tile and use the **Do** action on exactly that tile to chop trees, mine ore, attack creatures, drink water, open chests, or revive a downed teammate. If the faced tile contains a downed teammate, Do revives them. If the faced tile contains a living teammate, Do targets that teammate instead and can cause friendly fire. For synchronous-style coordination, all required agents must stand next to the same target tile, face it, and act together.
- **Crafting** stands next to (including diagonally) the required station and use the craft action; you do NOT need to face it. Diamond items always require an adjacent epic forge, not a table.
- **Placing** face the target tile, then use the place action. Tables and furnaces need an empty non-solid tile that is not water or lava; stone can also be placed into water (costs 1 stone). Place Plant puts a sapling on the faced tile. Place Torch lights dark areas.
- **Ranged combat** use Shoot Arrow while facing a creature (requires a bow + arrows). Bows are found in dungeon chests.
- **Elite mobs** are tougher and deal more damage; coordinating with teammates (multiple agents attacking together) makes them much easier to defeat.
- **Request/Give** use Request [Resource] to broadcast a resource request to teammates for 10 turns; teammates can use Give to Agent X to transfer one unit of the requested resource directly -- no adjacency required, works at any distance. Give only appears as an available action when a teammate has an active Request.

## Survival stats
Food, drink, and energy deplete gradually over time -- roughly every 20-30 steps you lose 1 point of each (dexterity slows this rate). When food or drink reaches 0, your health starts dropping. When energy reaches 0, you automatically fall asleep and cannot act until energy is full. While sleeping, you take 2.5x damage from all sources. Mana does NOT decay -- it is only spent by casting spells or enchanting. Mana slowly regenerates over time (faster while sleeping).
- **Sleep** choose this voluntarily to recover energy at 2x the passive rate. Ends automatically when energy is full.
- **Rest** choose this to recover health gradually. Requires food, drink, and energy all > 0; ends when health is full or a stat runs out.

## Roles
Role-restricted actions succeed with reduced probability for non-specialists. Depending on the difficulty configuration, non-specialist success rates are 10%

- **Forager** collecting water, saplings, eating passive mobs (e.g. cows/bats/snails). Also has 3x base food and drink capacity.
- **Miner** crafting pickaxes/torches, placing stone.
- **Warrior** crafting swords and arrows. Also deals 2x melee damage, and specializes in enchanting swords and bows.
- No role restriction: Place Table, Place Furnace, Wood Sword, Iron Armour, Diamond Armour
```

```
- **Construction** Construction sites (shelters, forges, beacons) may require either sync or handover. Always follow the coordination rule shown in the observation for that site.
- **Elite mobs** stronger enemies may benefit from or require coordinated attacks. Attack from different sides, avoid standing between a teammate and the mob, and avoid blocking another agent's attack.
- **Revive** Face a downed teammate and use **Do** to revive them.
- **Epic forge / Diamond crafting** Diamond-tier items require multiple agents to craft simultaneously at an adjacent epic forge. All required agents must choose the crafting action for the same item on the same turn while adjacent to the forge.

## Resource chain
Trees -> wood (no tool required) -> Stone/Coal (needs wood pickaxe) -> Iron (needs stone pickaxe) -> Diamond (iron pickaxe) -> Ruby/Sapphire (diamond pickaxe)

## Crafting recipes
All recipes consume the listed materials.
Stations: Table (2 wood), Furnace (1 stone)
- Wood pickaxe/sword: table + 1 wood
- Stone pickaxe/sword: table + 1 wood + 1 stone
- Iron pickaxe/sword: table + furnace + 1 wood + 1 stone + 1 iron + 1 coal
- Iron armour: table + furnace + 3 iron + 3 coal
- Diamond pickaxe: epic forge + 1 wood + 3 diamond + enough agents crafting the same item there on the same turn
- Diamond sword: epic forge + 1 wood + 2 diamond + enough agents crafting the same item there on the same turn
- Diamond armour: epic forge + 3 diamond + enough agents crafting the same item there on the same turn
- Arrows: table + 1 wood + 1 stone (yields 2)
- Torch: table + 1 wood + 1 coal (yields 4)

Construction (at a construction site, face it and use Build action):
- Build Shelter: needs 10 wood + 5 stone. Shelters result in +50%
- Build Forge: needs 10 stone + 3 iron + 2 coal. Creates an epic forge, which enables diamond gear crafting.
- Build Beacon: needs 3 iron + 2 coal. Expands the lit area on this level.

## Attributes
Gain 1 XP each time you descend to a new floor. Spend XP with Level Up actions.
- **Strength** max health = 8 + strength
- **Dexterity** max food = 7 + 2*dexterity (+2 extra for foragers); max drink = same; max energy = 7 + 2*dexterity
- **Intelligence** max mana = 6 + 3*intelligence; enchantment damage +5%

## Progression
1. Gather wood -> place a table -> craft a wood pickaxe; craft a wood sword early if combat is likely.
2. Mine stone and coal -> place a furnace -> craft iron tools and iron armour.
3. To descend: stand on the 'ladder_down' tile (visible in your observation when close) and use the Descend action. The ladder only becomes usable after enough monsters on that level have been killed. Only one agent needs to use Descend/Ascend -- all teammates are teleported with them.

</game_rules>

<achievements>
## Achievements
Collect Wood
Place Table
Eat Cow
Collect Sapling
Collect Drink
Collect Food
Make Wood Pickaxe
Make Wood Sword
```

LLM Interface

`<output_format>`

Each turn you receive an observation showing what you see, your inventory, teammates, and available actions.

Think first, then output strictly in the following format:

1. (Required) Exactly one action from the available action list:

`<action>YOUR_CHOSEN_ACTION</action>`

2. (Optional) Broadcast to teammates, up to 400 chars. Teammates can only act on what you tell them. Be specific (e.g. 'Dig on tree next turn', 'Ladder at 5NE', 'Need 2 wood'). Reply to teammates' requests.

`<communication>YOUR_MESSAGE</communication>`

3. (Optional) Private notes, up to 1000 chars -- not shared with teammates. Your context resets each turn -- this is your only memory. Don't repeat what's already in your observation; store what you'll need later. Record teammates' plans and any facts you'll need after they scroll out of view.

`<scratchpad>YOUR_NOTES</scratchpad>`

Important: every tag you open must be closed (e.g. `<communication>...</communication>`).

Token budget: 8192 tokens for your full response (including reasoning). Keep reasoning concise and stop thinking early enough to emit every required tag -- if you exhaust the budget mid-reasoning, no action is produced and your turn fails.

`</output_format>`

LLM Interface

Example observation given to the LLM

The LLM receives the last 8 observation-action turns as context. Below we show the current observation from one turn.

Step: 0/10000 (10000 remaining, ends early if all agents die)

Current Observation:

Position: (x=25, y=24)

Role: forager

Location: Overworld (surface)

Achievements: 0/93 (39 unlock later)

Level info:

- Light: bright (0.80)
- Level: cleared -- you can find the ladder down tile and use Descend to go deeper.

You see:

- tree 3 steps north (x=25, y=21)
- construction_site 1 step north (x=25, y=23)
- water 5 steps east (x=30, y=24)

Facing: north.

Do target: construction_site (x=25, y=23).

LLM Interface

Your status:

- health: 9
- food: 9
- drink: 9
- energy: 9
- mana: 9
- xp: 0

Attributes:

- dexterity: 1
- strength: 1
- intelligence: 1

You have nothing in your inventory.

Available actions:

- Noop
- Move West
- Move East
- Move North
- Move South
- Do
- Sleep
- Rest
- Request Food
- Request Drink
- Request Wood
- Request Stone
- Request Iron
- Request Coal
- Request Diamond
- Request Ruby
- Request Sapphire

Response format (in this order): <action> , <communication> , <scratchpad>. Close every tag you open.

Prelim 2 - Challenges in Multi-Agency

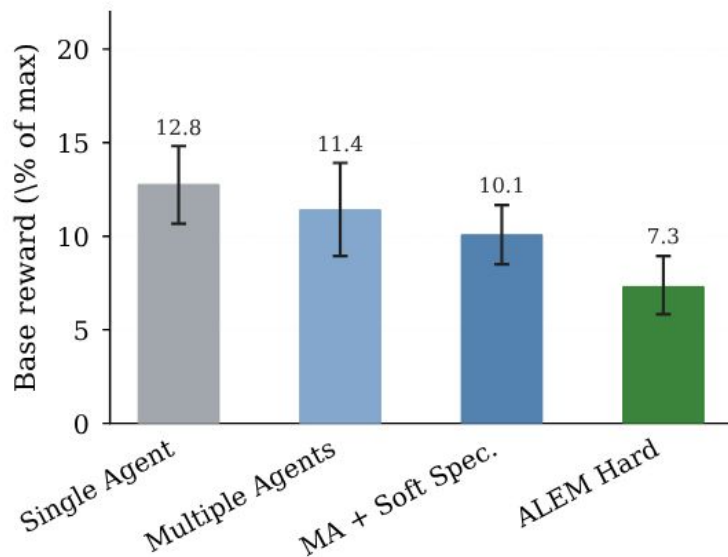
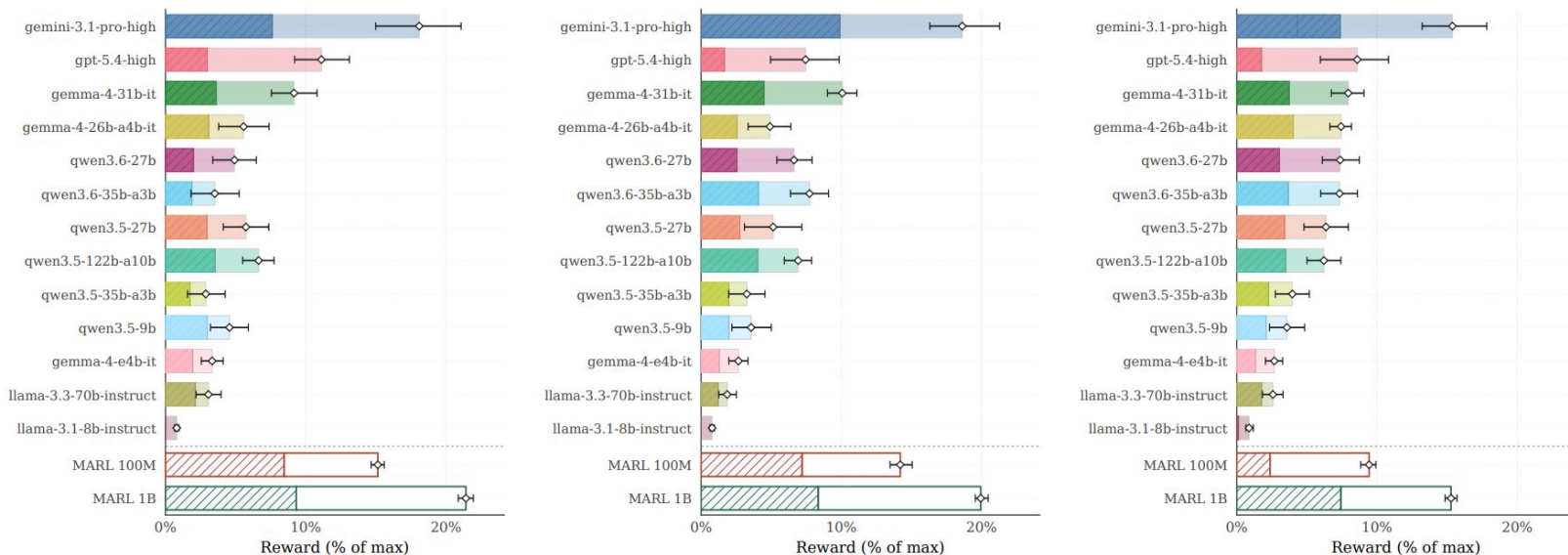


Figure 19: **Environment calibration.** Base reward for Gemma-4-31B-it across settings sharing the same underlying world. Cooperative achievements are excluded so all settings are directly comparable. Error bars show 95% bootstrap CIs.

***Multi-Agent Coordination introduces a new paradigm of difficulty**

Q1: How well do LLM agents coordinate (Zero-shot) ?

Zero-Shot Frontier Model (Gemini 3.1 Pro) matches the performance of MARL trained for 1B timesteps.



(a) Easy

(b) Medium

(c) Hard

Coord. achievements Normal achievements 95% CI on total MARL 100M MARL 1B

Q1: Qualitative - Multi-Step Plans + Shared Intent



AGENT 0

Warrior

THINKING

```
T87: Do (mine 18,16)
If successful:
T88: Move N (onto 18,16)
T89: Move E (turn to face 19,16)
T90-T94: Noop
T95: DO (mine 19,16 3-agent sync)
```

If mining fails on T87, repeat DO until 18,16 is clear.

A2 will receive my message at T88, so they will **cancel** their T90 solo/duo DO ...

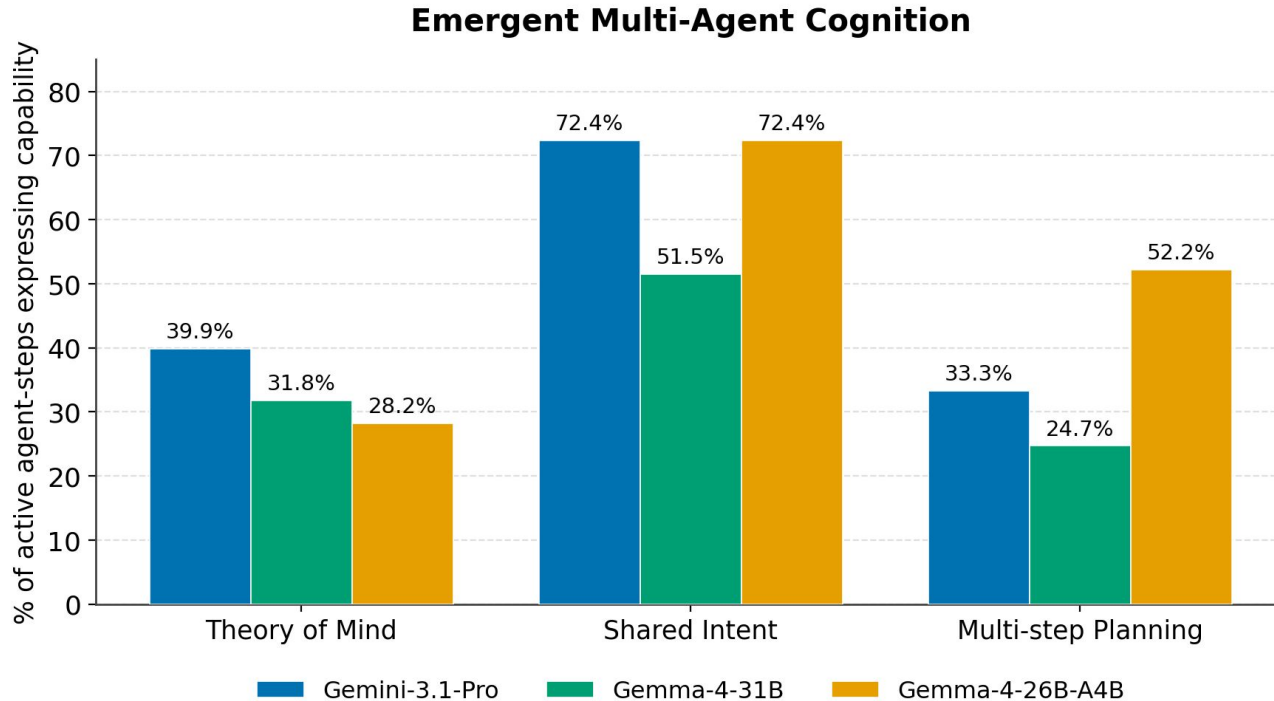
MESSAGE

A2, WAIT! A1 changed the plan so we get the 'Coord Mine Stone Hard' achievement on 19,16! **DO NOT** hit 19,16 at T90. I am mining 18,16 now to stand on it, might take a few tries as a warrior. I will be at 18,16 facing E. **A2: Stay** at 19,17 facing N. **A1:** Route to 19,15 facing S. **ALL SYNC DO** 19,16 AT **EXACTLY** T95!

● WARRIOR ● FORAGER ● MINER

STEP 87

Q1: Emergent Multi-Agent Cognition - Top 3 models



* GPT doesn't expose reasoning so we didn't include it here.

Q1: Emergent Multi-Agent Cognition - Gemini 3.1 Pro

Theory of Mind

reasoning about a teammate's state / intent

Multi-Step Planning

sequenced, multi-turn lookahead

Gemini-3.1-Pro	Gemma-4-31B	Gemma-4-26B-A4B
<p>BROADCAST - TEAM</p> <p>Copy that, Agent 2. Take your time resting — once you're done, face South and Do to revive me.</p> <p>HARD · RUN 0 · STEP 203 · A0</p>	<p>BROADCAST - TEAM</p> <p>A2, your health dropped to 5 — is there a monster next to you? If you're safe, revive me next turn as you said.</p> <p>EASY · RUN 0 · STEP 136 · A0</p>	<p>BROADCAST - TEAM</p> <p>Agent 1, I'm waiting at (26,25). Let's Build once you're done with the cow.</p> <p>EASY · RUN 0 · STEP 103 · A0</p>
<p>BROADCAST - TEAM</p> <p>A1, I'll be at 26,21 ready to complete the handover on turn 22 — you can initiate on turn 21. A0, meet us at the 3-agent stone at 26,16 after we finish this tree!</p> <p>HARD · RUN 1 · STEP 17 · A2</p>	<p>BROADCAST - TEAM</p> <p>I'm boxed in. I'll mine the stone at (22,30) first to clear a path to (22,31). Let's sync our 'Do' actions on turn 28!</p> <p>EASY · RUN 6 · STEP 24 · A2</p>	<p>PRIVATE · NOT SENT</p> <p>Current pos: (24,25). Plan: Move North, North, East, then Do tree. Wood is priority for crafting tools.</p> <p>EASY · RUN 0 · STEP 1 · A2</p>

Q1: How well do LLM agents coordinate zero-shot within homogeneous teams?

Method	Easy			Medium			Hard		
	Base%	Coord.%	Total%	Base%	Coord.%	Total%	Base%	Coord.%	Total%
gemini-3.1-pro-high	18.1 [14.2,21.8]	18.0 [12.5,23.9]	18.1 [14.8,21.0]	15.1 [12.3,18.0]	23.4 [20.3,27.0]	18.6 [16.3,21.3]	13.8 [11.2,16.6]	17.5 [13.5,21.5]	15.4 [13.2,17.8]
gpt-5.4-high	14.1 [11.5,16.7]	7.0 [3.6,10.8]	11.1 [9.2,13.0]	10.0 [6.5,13.5]	4.0 [2.1,5.7]	7.4 [4.9,9.8]	11.8 [8.0,15.3]	4.2 [2.0,6.6]	8.6 [5.9,10.8]
gemma-4-31b-it	9.6 [7.4,11.9]	8.6 [6.4,11.0]	9.2 [7.6,10.8]	9.7 [8.0,11.3]	10.6 [8.3,13.0]	10.1 [9.0,11.1]	7.3 [5.9,8.9]	8.8 [6.9,10.9]	8.0 [6.7,9.1]
gemma-4-26b-a4b-it	4.3 [2.6,6.2]	7.3 [5.0,9.6]	5.6 [3.8,7.4]	4.1 [2.8,5.3]	6.1 [3.8,8.3]	4.9 [3.3,6.4]	5.9 [5.3,6.6]	9.5 [7.7,11.3]	7.4 [6.6,8.2]
qwen3.6-27b	5.1 [3.4,6.9]	4.7 [2.9,6.7]	4.9 [3.4,6.5]	7.0 [5.4,8.8]	6.1 [4.3,7.9]	6.6 [5.4,7.9]	7.5 [5.7,9.5]	7.2 [5.4,8.9]	7.4 [6.1,8.8]
qwen3.6-35b-a3b	2.8 [1.4,4.5]	4.4 [2.3,6.7]	3.5 [1.8,5.3]	6.3 [5.0,7.9]	9.7 [7.7,11.6]	7.7 [6.3,9.1]	6.3 [4.8,8.1]	8.7 [6.7,10.7]	7.3 [6.0,8.6]
qwen3.5-27b	4.8 [3.1,6.9]	7.0 [4.9,9.2]	5.7 [4.1,7.4]	4.1 [2.4,6.0]	6.5 [3.8,9.3]	5.1 [3.1,7.2]	5.1 [3.5,7.1]	8.1 [5.8,10.4]	6.4 [4.8,7.9]
qwen3.5-122b-a10b	5.4 [4.3,6.5]	8.4 [6.6,10.0]	6.6 [5.5,7.7]	5.0 [4.6,5.3]	9.6 [7.5,11.5]	6.9 [6.0,7.9]	4.7 [3.9,5.5]	8.3 [6.2,10.6]	6.2 [5.0,7.4]
qwen3.5-35b-a3b	1.9 [1.0,2.9]	4.2 [2.1,6.3]	2.9 [1.6,4.2]	2.2 [1.4,3.1]	4.7 [2.7,6.8]	3.3 [2.0,4.6]	2.9 [2.2,3.7]	5.3 [3.4,7.5]	4.0 [2.7,5.2]
qwen3.5-9b	2.7 [2.0,3.4]	7.0 [4.6,9.5]	4.6 [3.2,5.9]	2.7 [1.6,4.0]	4.7 [2.7,7.1]	3.6 [2.2,4.9]	2.6 [1.7,3.4]	5.0 [2.9,7.1]	3.6 [2.3,4.8]
gemma-4-e4b-it	2.4 [1.9,2.9]	4.6 [3.0,6.3]	3.3 [2.6,4.1]	2.4 [1.9,2.8]	3.1 [1.7,4.5]	2.7 [2.0,3.3]	2.3 [1.9,2.7]	3.2 [1.9,4.5]	2.7 [2.1,3.3]
llama-3.3-70b-instruct	1.6 [1.1,2.3]	5.0 [3.2,6.8]	3.1 [2.2,4.0]	1.1 [0.8,1.5]	2.9 [1.4,4.4]	1.9 [1.3,2.5]	1.4 [1.1,1.7]	4.2 [2.7,5.8]	2.6 [1.8,3.3]
llama-3.1-8b-instruct	1.4 [1.1,1.7]	0.0 [0.0,0.0]	0.8 [0.6,1.0]	1.4 [1.1,1.6]	0.0 [0.0,0.0]	0.8 [0.6,0.9]	1.3 [1.0,1.5]	0.3 [0.0,0.9]	0.9 [0.7,1.2]
Across LLM agents	5.7	6.6	6.1	5.5	7.0	6.1	5.6	6.9	6.2
MARL 100M	11.6 [11.1,12.1]	20.0 [19.0,21.4]	15.1 [14.6,15.6]	12.1 [11.5,12.8]	17.0 [16.2,18.0]	14.2 [13.5,15.0]	12.2 [11.6,12.8]	5.6 [5.1,6.2]	9.4 [8.9,9.9]
MARL 1B	21.0 [19.4,22.6]	22.0 [20.8,23.1]	21.4 [20.9,21.9]	20.1 [19.6,20.7]	19.8 [19.4,20.2]	19.9 [19.5,20.5]	13.6 [12.8,14.2]	17.6 [16.9,18.2]	15.3 [14.9,15.7]

Q1.1: Base != Coordination Performance

Method	Easy			Medium			Hard		
	Base%	Coord.%	Total%	Base%	Coord.%	Total%	Base%	Coord.%	Total%
gemini-3.1-pro-high	18.1 [14.2,21.8]	18.0 [12.5,23.9]	18.1 [14.8,21.0]	15.1 [12.3,18.0]	23.4 [20.3,27.0]	18.6 [16.3,21.3]	13.8 [11.2,16.6]	17.5 [13.5,21.5]	15.4 [13.2,17.8]
gpt-5.4-high	14.1 [11.5,16.7]	7.0 [3.6,10.8]	11.1 [9.2,13.0]	10.0 [6.5,13.5]	4.0 [2.1,5.7]	7.4 [4.9,9.8]	11.8 [8.0,15.3]	4.2 [2.0,6.6]	8.6 [5.9,10.8]
gemma-4-31b-it	9.6 [7.4,11.9]	8.6 [6.4,11.0]	9.2 [7.6,10.8]	9.7 [8.0,11.3]	10.6 [8.3,13.0]	10.1 [9.0,11.1]	7.3 [5.9,8.9]	8.8 [6.9,10.9]	8.0 [6.7,9.1]
gemma-4-26b-a4b-it	4.3 [2.6,6.2]	7.3 [5.0,9.6]	5.6 [3.8,7.4]	4.1 [2.8,5.3]	6.1 [3.8,8.3]	4.9 [3.3,6.4]	5.9 [5.3,6.6]	9.5 [7.7,11.3]	7.4 [6.6,8.2]
qwen3.6-27b	5.1 [3.4,6.9]	4.7 [2.9,6.7]	4.9 [3.4,6.5]	7.0 [5.4,8.8]	6.1 [4.3,7.9]	6.6 [5.4,7.9]	7.5 [5.7,9.5]	7.2 [5.4,8.9]	7.4 [6.1,8.8]
...									
Across LLM agents	5.7	6.6	6.1	5.5	7.0	6.1	5.6	6.9	6.2
MARL 100M	11.6 [11.1,12.1]	20.0 [19.0,21.4]	15.1 [14.6,15.6]	12.1 [11.5,12.8]	17.0 [16.2,18.0]	14.2 [13.5,15.0]	12.2 [11.6,12.8]	5.6 [5.1,6.2]	9.4 [8.9,9.9]
MARL 1B	21.0 [19.4,22.6]	22.0 [20.8,23.1]	21.4 [20.9,21.9]	20.1 [19.6,20.7]	19.8 [19.4,20.2]	19.9 [19.5,20.5]	13.6 [12.8,14.2]	17.6 [16.9,18.2]	15.3 [14.9,15.7]

Q1.2: Coordination failures differ across task structure.

gemini-3.1-pro-high	12	18	90	7
gpt-5.4-high	1	4	42	0
gemma-4-31b-it	3	9	58	2
qwen3.6-35b-a3b	2	9	64	0
qwen3.5-27b	3	8	67	1
gemma-4-26b-a4b-it	1	8	70	1
qwen3.6-27b	1	6	46	2
qwen3.5-122b-a10b	2	8	65	0
qwen3.5-9b	3	5	69	0
qwen3.5-35b-a3b	2	6	62	0
gemma-4-e4b-it	0	1	45	0
llama-3.3-70b-instruct	1	1	48	0
llama-3.1-8b-instruct	0	0	2	0
MARL 1B step	25	19	43	0

Soft Sync-Hard Handover Construction

Measure success rate.

Handover (one agent begins, another completes)

- Comparatively easiest – time slack.

Sync-Hard (all agents coord same time)

- Harder, possibly due to spatial reasoning from text.

(a) Coordination breakdown

Q1.2: MARL Exploration Failure?

gemini-3.1-pro-high	12	18	90	7
gpt-5.4-high	1	4	42	0
gemma-4-31b-it	3	9	58	2
qwen3.6-35b-a3b	2	9	64	0
qwen3.5-27b	3	8	67	1
gemma-4-26b-a4b-it	1	8	70	1
qwen3.6-27b	1	6	46	2
qwen3.5-122b-a10b	2	8	65	0
qwen3.5-9b	3	5	69	0
qwen3.5-35b-a3b	2	6	62	0
gemma-4-e4b-it	0	1	45	0
llama-3.3-70b-instruct	1	1	48	0
llama-3.1-8b-instruct	0	0	2	0
MARL 1B step	25	19	43	0
	Soft	Sync-Hard	Handover	Construction

Construction

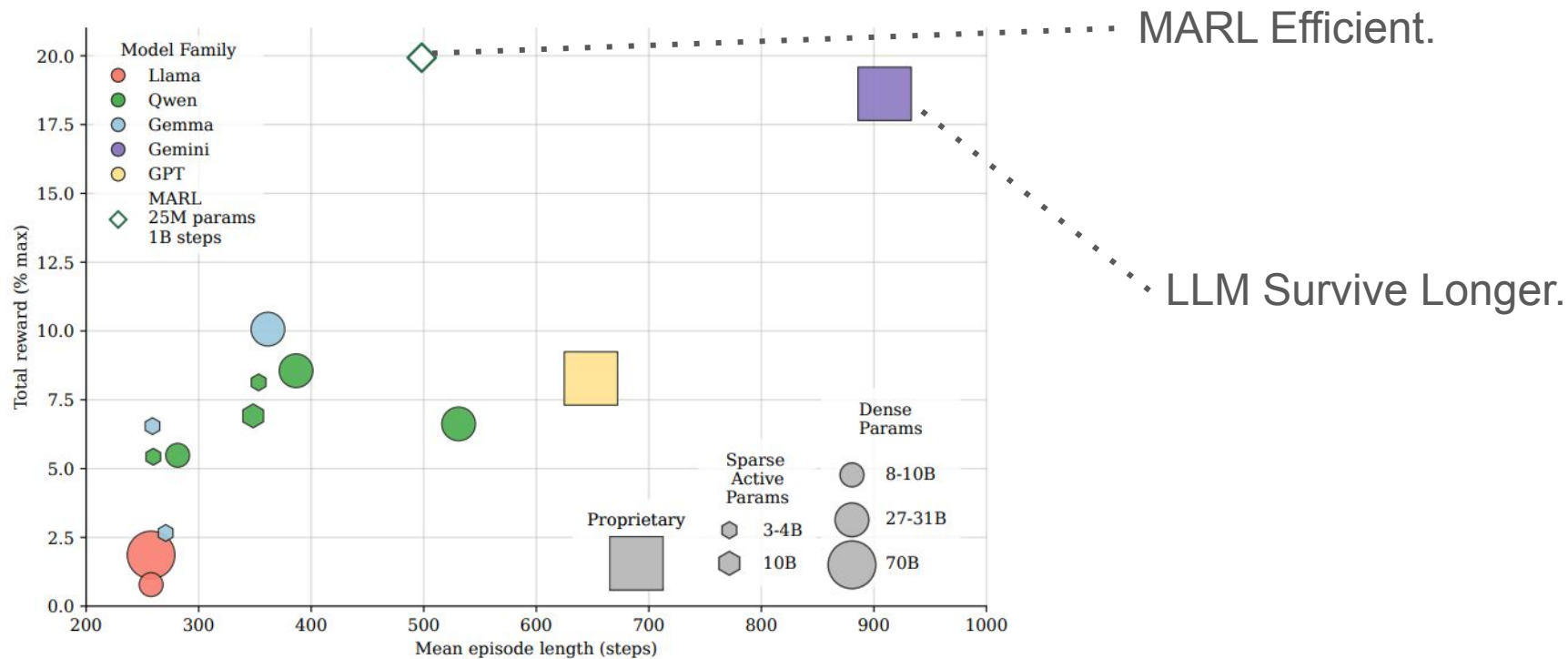
- Agents must collect resources, progress through a technology chain, converge on a shared site, and execute a synchronised build.

MARL fails here!

Possible to find through trial-and-error?

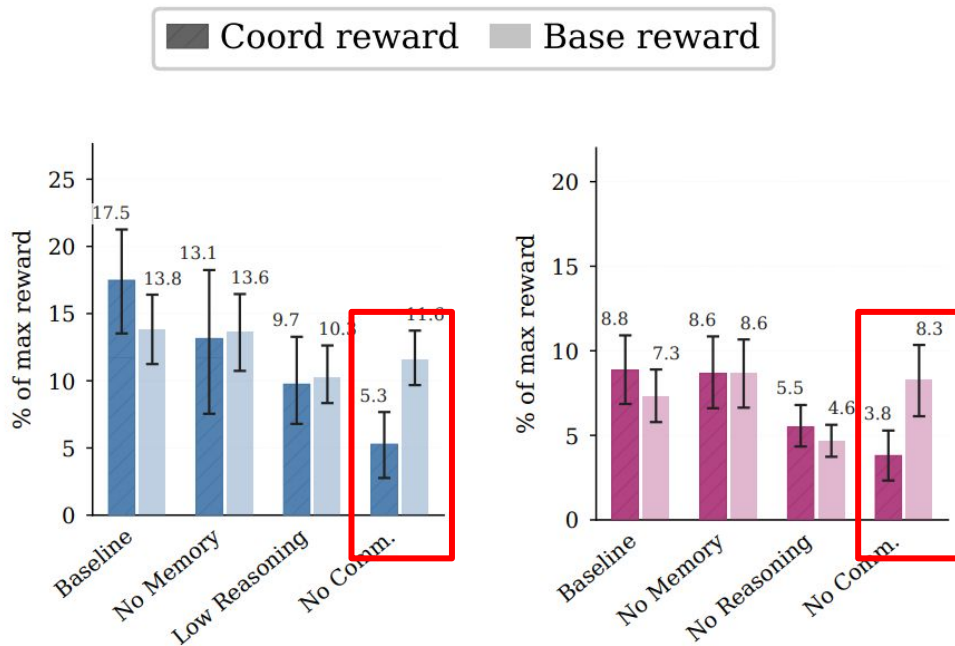
(a) Coordination breakdown

Q1.3: Step Efficiency vs Performance



(c) Survival vs. reward

Q2: Ablations: How do communication, memory, and reasoning affect coordination?



(a) Gemini-3.1-Pro

(b) Gemma-4-31B

Q2: Communication is critical for (zero-shot) coordination.

Gemini 3.1 Pro: run 00, agent 0 (warrior)

Step 1 ctrl action: Move East

"I chopped 25,23. Let's do the 3-agent tree at 27,23! I'm moving to its West side (26,23). Agent 0, go to its North side (27,22). Agent 2, go to its South side (27,24). Face the tree and wait, we will 'Do' together when everyone is ready."

Step 10 ctrl action: Make Wood Pickaxe

"My wood pickaxe craft failed due to the warrior penalty! Trying again this turn. Once I get it, I will make a wood sword and then join you for the 3-agent stone."

Step 230 forced action: Noop hp 0

"Agent 2, your health is 1! You are about to die! Drink a potion IMMEDIATELY to heal if you have one! Place stone to block the orc!"

Gemma 4: runs 00 & 06, agent 0 (warrior)

Run 00, step 0 ctrl action: Move North

"Agent 1: face North and Do the tree at (25, 23). Agent 2: move to the stone at (20, 23). I'm moving North to (24, 23) to help gather wood."

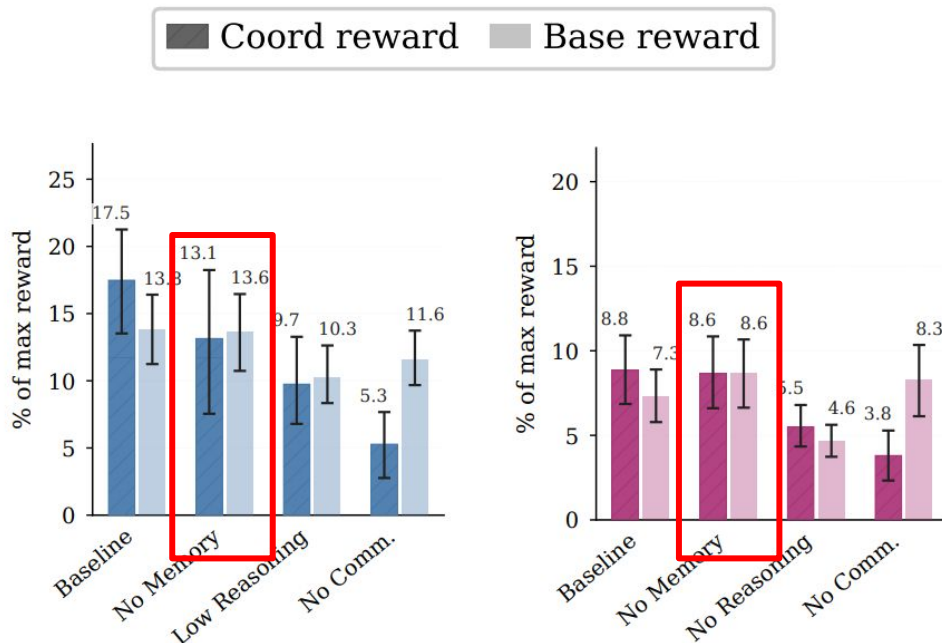
Run 00, step 100 forced hp 5

"Almost there, thanks for waiting."

Run 06, step 700 forced hp 3

"Agent 1, please prioritize me for food as well, I am at 0 and my health is critical (3). Staying put."

Q2: Scratchpad memory helps when used for planning.



(a) Gemini-3.1-Pro

(b) Gemma-4-31B

Q2: Scratchpad memory helps when used for planning.

Gemini uses the scratchpad as a forward-looking planner:

- **71.5%** of entries are multi-line,
- **75.3%** contain future-tense verbs such as will, next, then, or plan
- **12.4%** contain explicit turn-indexed action sequences such as T9:Do, T10:Move West, and T11:Give wood to Agent 1.

In contrast, Gemma's scratchpads are:

- one-line state summaries (average **1.08** lines)
- with frequent coordinate references (**98.8%**) status information that is often already present in the next observation.

Q2: Scratchpad memory helps when used for planning.

Gemini 3.1 Pro: run 02, agent 0 (warrior), step 9, action: Do

T9: Do (completes handover at 26,23)

T10: Move West (towards 21,23)

T11: Give wood to Agent 1 (since A1 will Request Wood on T10)

T12+: Move West to table, then Make Wood Pickaxe.

Since I'm a warrior, pickaxe crafting has a chance to fail (40% or whatever the config is), so I'll keep trying if it does.

Gemma 4: agent 0 (warrior), runs 00 & 06

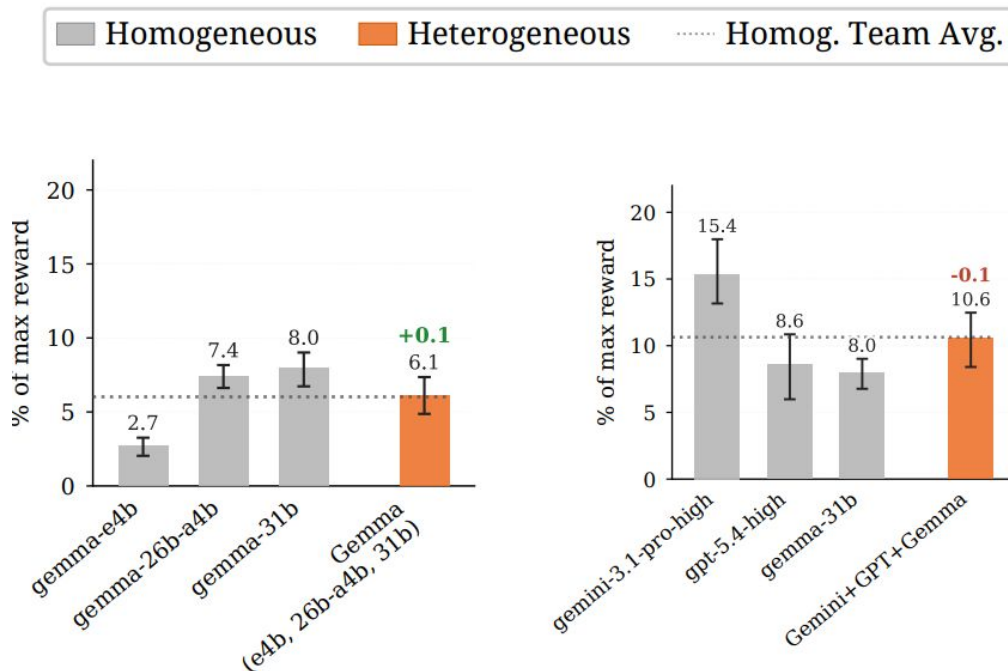
Run 00, step 0 action: Move North

Starting resource collection. Agent 1 (Forager) is best for wood (25, 23). Agent 2 (Miner) is best for stone (20, 23). I will move to (24, 23) to assist with wood gathering.

Run 06, step 473 action: Move North

Current (0, 19). Hit stone at (0, 20) and edge at x=-1. Moving North to (0, 18) to reposition and start searching East. A2 is searching North. A1 still sleeping.

Q3: How well do agents coordinate in teams composed of different models?



(c) Same-family team

(d) Cross-family team

MARL vs LLM Agents

MARL

Pros

- ✓ Efficient execution & training (1B steps, 1.5 days cheap GPU).
- ✓ High timestep efficiency – likely because trained.
- ✓ Less compute required.

Cons

- ✗ Less interpretable plans/communications.
- ✗ Has to be trained in each new env.
- ✗ Hard to input useful priors.
- ✗ MARL exploration still a challenge in big worlds, e.g. construction tasks.

LLM Agents

Pros

- ✓ Frontier models good - zero-shot.
- ✓ More interpretable plans/communications/reasoning.
- ✓ Good priors that don't need exploration.
- ✓ Can input priors easily.

Cons

- ✗ Expensive - Gemini exps cost 3-4k USD
- ✗ Prompting/agents – black magic – more than deep RL.
- ✗ Less timestep efficient.

Summary



Contributions:

- We introduce alem, a JAX-based benchmark for long-horizon, open-ended multi-agent coordination.
- We zero-shot evaluate 13 LLMs and 4 trained MARL baselines find the following:
 - See hints of **theory-of-mind, shared intent, multi-step planning**.
 - **Coordination** performance is **distinct** from **base-task performance**.
 - **Communication** matters most among the ablated harness components.
 - Initial heterogeneous team experiments suggest that **mixed teams** avg. out compared to homogenous teams.

Takeaway & Thanks



Challenges & Limitations:

- Results based on a single environment.
- Focus on LLMs, no experiments on VLMs.
- Due to costs, no Claude and only 10 seeds for LLM experiments (20 used for MARL).

51

Future work:

- **LLMs** - stronger agent harnesses, cross-episode memory, and lifelong learning.
- **MARL** - better exploration in open-ended environments.
- Neither MARL or LLM are perfect – what is next?

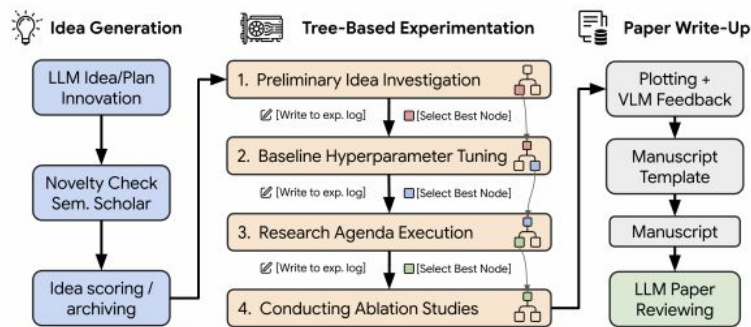
Connections between Open-ended Coordination & Research

Sequence of artifacts

Goals = <collect wood, build table, craft pickaxe,..>



Goals = <hypothesis, code, experiments, analyses, models, papers,..>



Questions & Comments

Email: k.tessera@ed.ac.uk

Paper: Benchmarking Open-Ended Multi-Agent Coordination in Language Agents <https://arxiv.org/abs/2606.08340>

Code: <https://github.com/alem-world/alem-env>