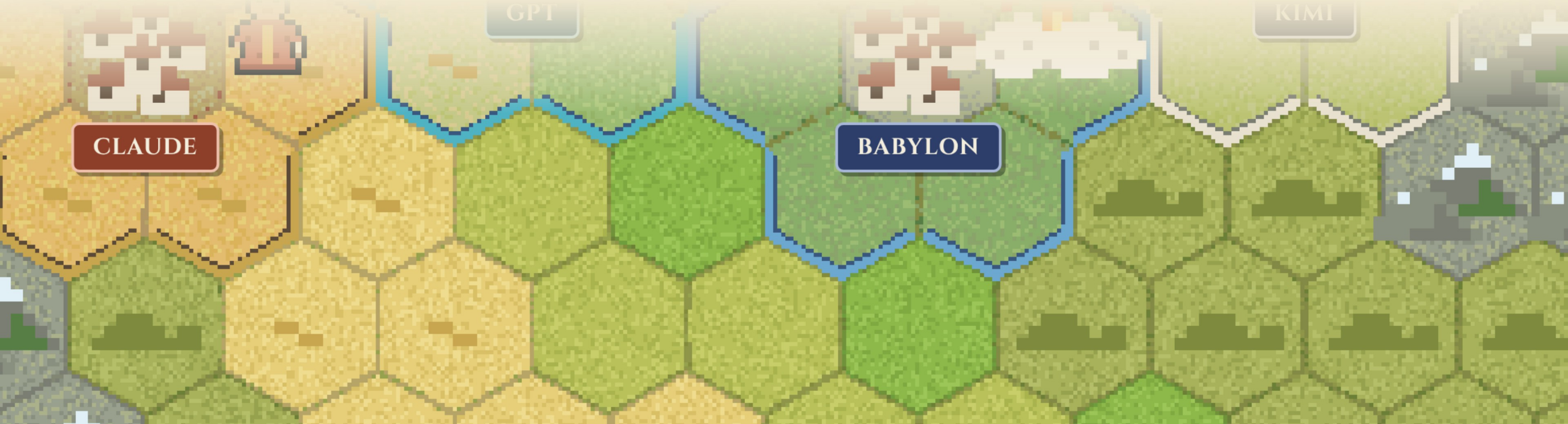
NEURIPS 2026 · EVALUATIONS & DATASETS TRACK

CivBench

A Long-Horizon Benchmark for Tool-Mediated Agents in Civilization VI

Austin Tudor David Andrews* · **Liam Wilkinson*** · Jamie Heagerty · Harry Coppock · Jakob Nicolaus Foerster · Rui Ponte Costa

University of Oxford · Tony Blair Institute for Global Change · Google DeepMind · UK AI Security Institute · Imperial College London · 10 Downing Street ·
*Equal contribution



MOTIVATION

Agents now run for hours, not a handful of steps

Most agent benchmarks are short. Deployed agents increasingly work through tools, for long stretches, in worlds they cannot see all at once.

300+
turns in a single game

1000s
of tool calls per episode

Partial
observability throughout

01 · THE ENVIRONMENT

Full games, played through MCP

MCP client

Any model, any MCP client

→

civ6-mcp server

76 tools in 7 categories

Deterministic narration layer: raw state → structured text

→

FireTuner

Lua over TCP

→

Civilization VI

Live game state

Available ≠ observed

Observation is pull-based. Victory progress, diplomacy and rivals are all queryable, but the agent sees them only if it asks.

A diary every turn

Each turn ends with a five-field diary, including concrete plans for the next 5–10 turns.

Outcomes alone do not separate models

3 / 23

games won, across four model families under one shared playbook

Win/loss

No significant model difference (Fisher–Freeman–Halton $p = 0.530$)

Normalised score

No significant model difference (Kruskal–Wallis $p = 0.594$)

13 outcome metrics

Only exploration at turn 100 clearly separates models (ICC = 0.717)

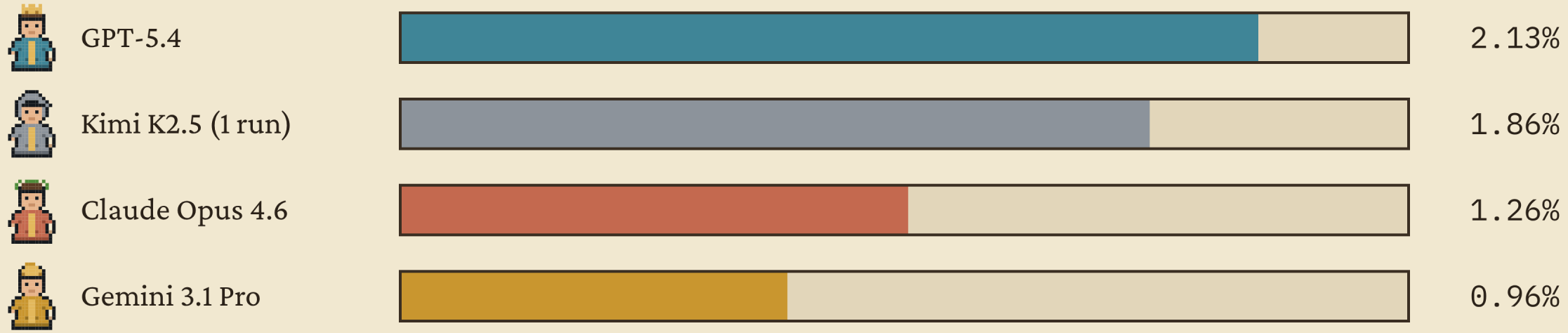
So we look inside the traces instead.

03 · FAILURE MODE ONE

The sensorium effect

A human absorbs the minimap and score ticker for free. An agent sees only what it queries, and it rarely queries the global state.

Proactive Monitoring Rate (share of non-infrastructure tool calls)



Axis 0–2.5%. Monitoring does not rise towards the endgame.

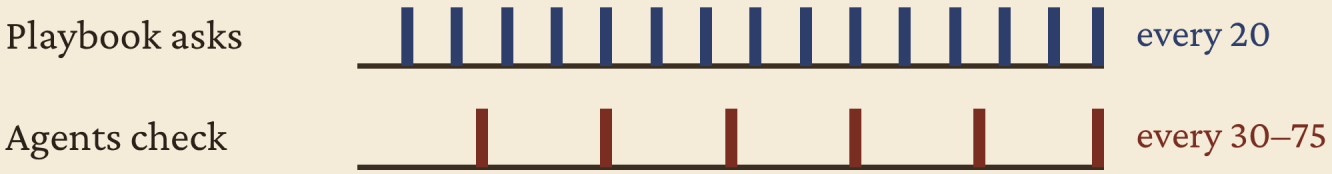
03 · FAILURE MODE ONE

Warnings that were there, but never retrieved

8 / 20

detectable defeats with no victory check in the final 20 turns

Victory-progress checks across a 300-turn game



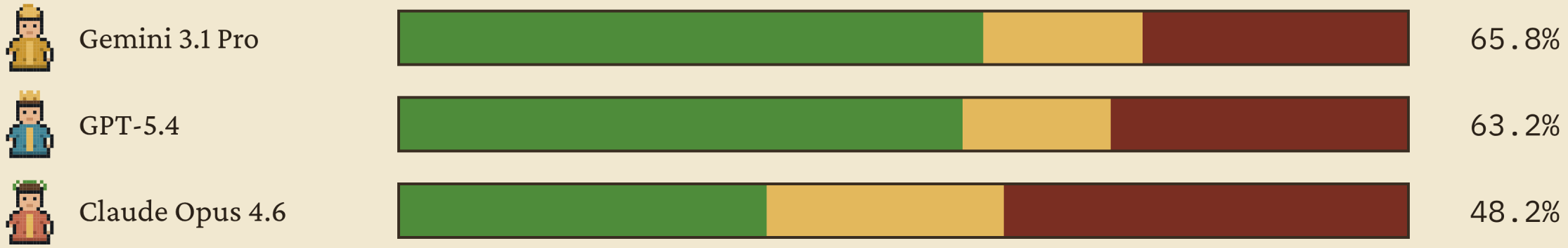
Schematic at the reported rate: 3.7–10.0 victory checks per game (0.05–0.29% of tool calls).

04 · FAILURE MODE TWO

The reflection–action gap

“Found second city” *“Build campuses in new cities”* *“Check victory progress”* → **not carried out**

Commitment execution within 10 turns (RAG@10)



Green: executed · Gold: partial (half credit) · Red: not executed. 142 commitments; 92% agreement with human labels (κ = 0.87). At 20 turns, 27.5% are still not executed.

TAKEAWAYS

Tools and instructions are not enough

1. Agents under-monitor state that is available but must be asked for.
2. They often fail to carry out plans they wrote down themselves.
3. A 23-run pilot: patterns across models, not a leaderboard.



Environment, scenarios, logs and analysis

`github.com/lmwilki/civ6-mcp`