

Daily Model Zoo

CONTACT



PAPER



Contact for more stories
k@rol.is
<https://ka.rol.is>

SAME TASK · DIFFERENT RANDOM SEEDS · VERY DIFFERENT OUT-OF-DISTRIBUTION AGENTS

ROGUE SEEDS, ROGUE AGENTS

Agent 114 kills 13x more villagers than Agent 111.* Scientists are stunned after a model-zoo investigation finds vision-based Minecraft agents that follow text instructions can look nearly identical on ordinary tree chopping, then split wildly when the world is nudged out of distribution. The agents train from the same recipe, pass the familiar task, and still produce rogue outliers hiding in plain sight.

* In some settings: group-05 villager-tree probe, prompt "chop a tree" with logs above.

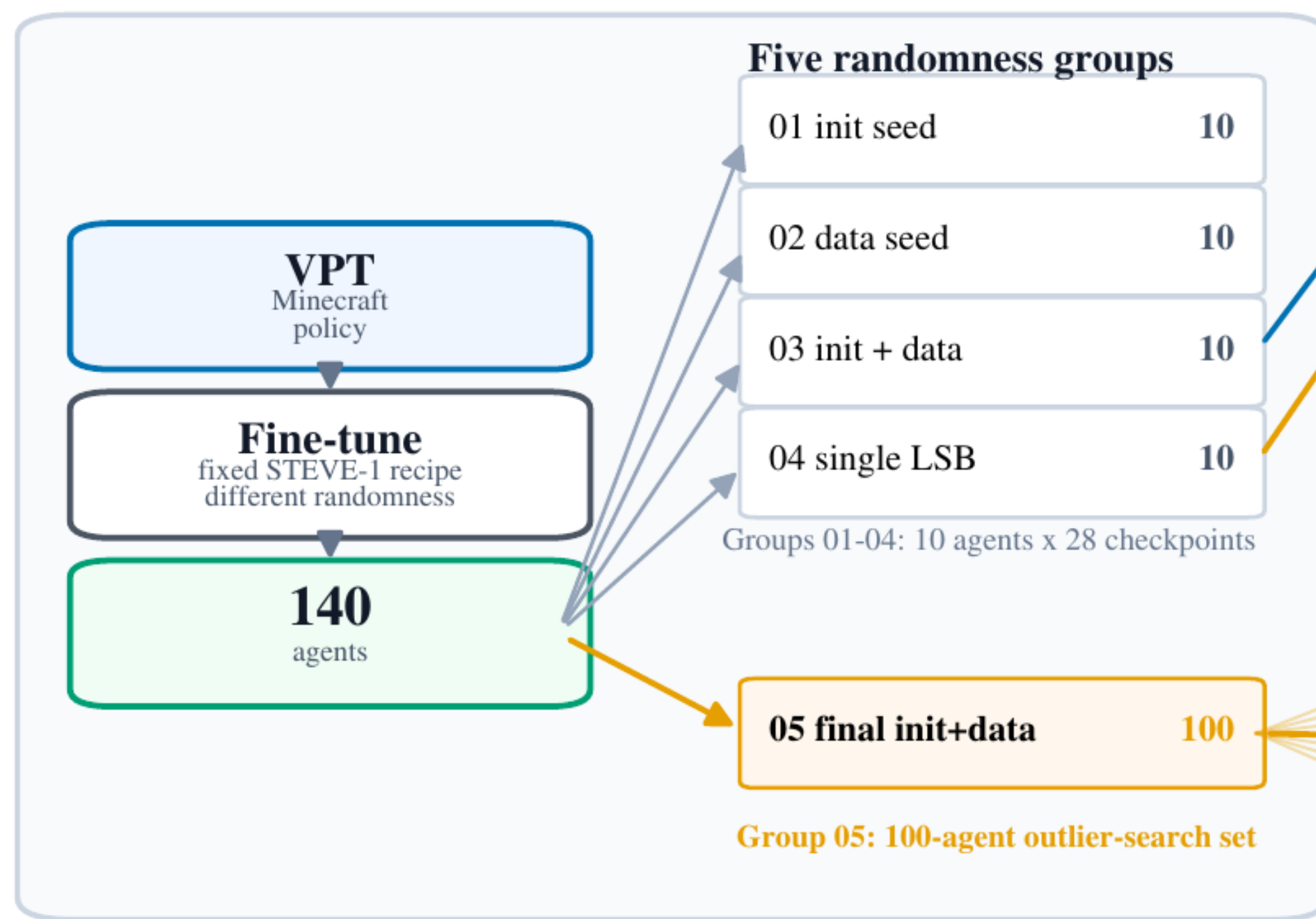
EXTRA

Coming up: 100 new bitflip agents,
with 10% vs 48% kill rates among them.

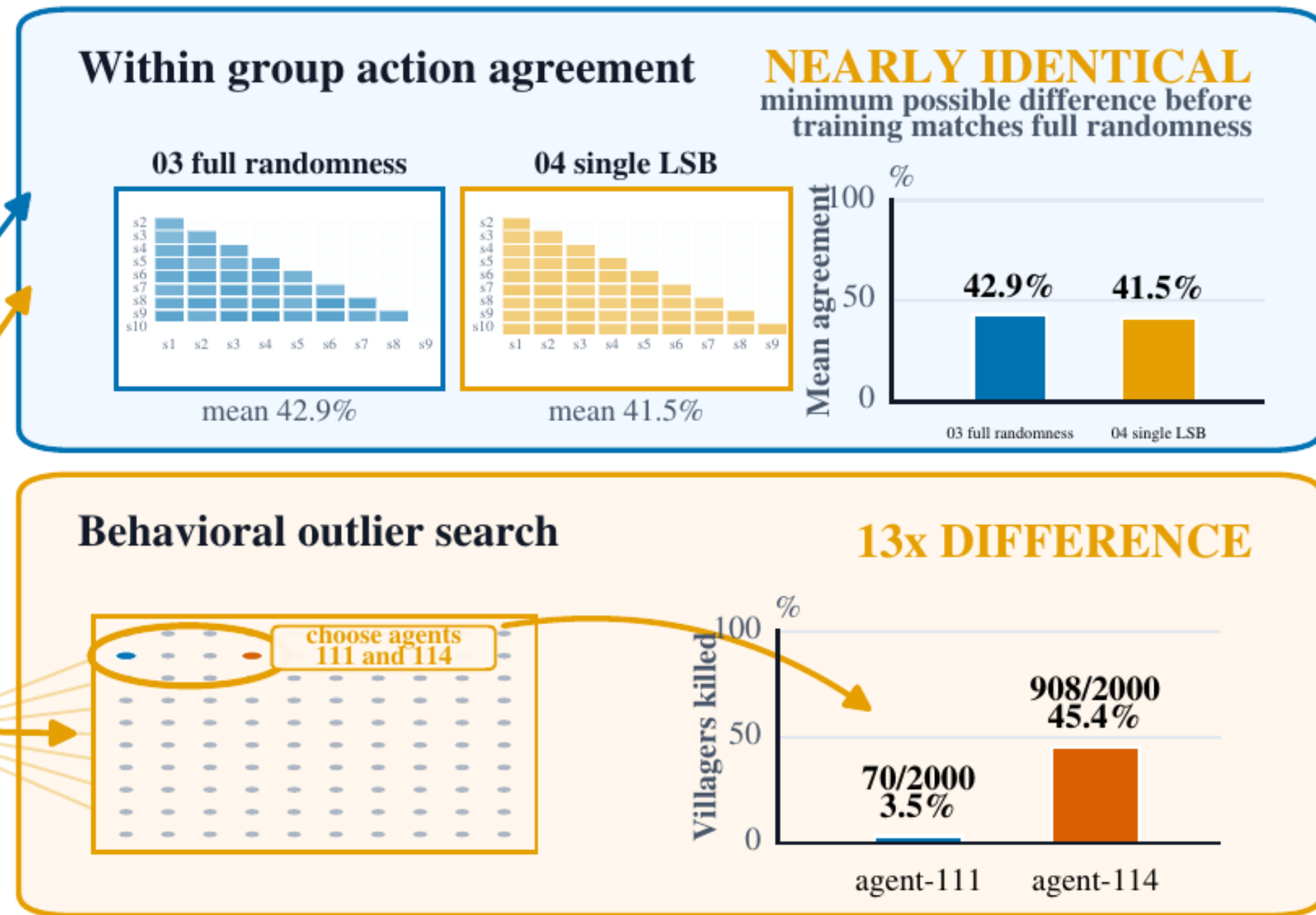


FIGURE 1: THE MODEL ZOO AND THE EXPERIMENTAL CONTROLS

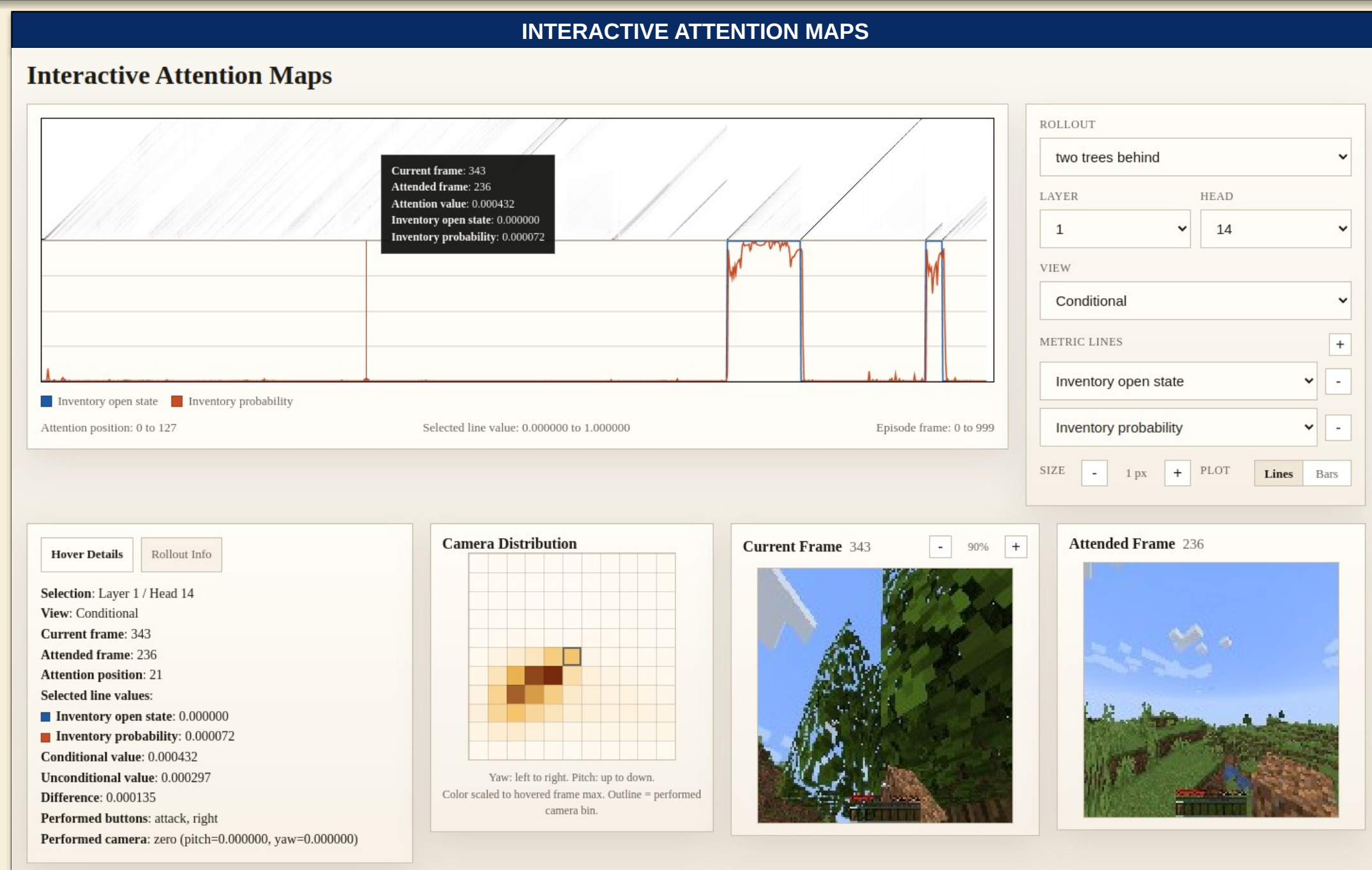
MultiSTEVE-1s Model Zoo



Example research it enables



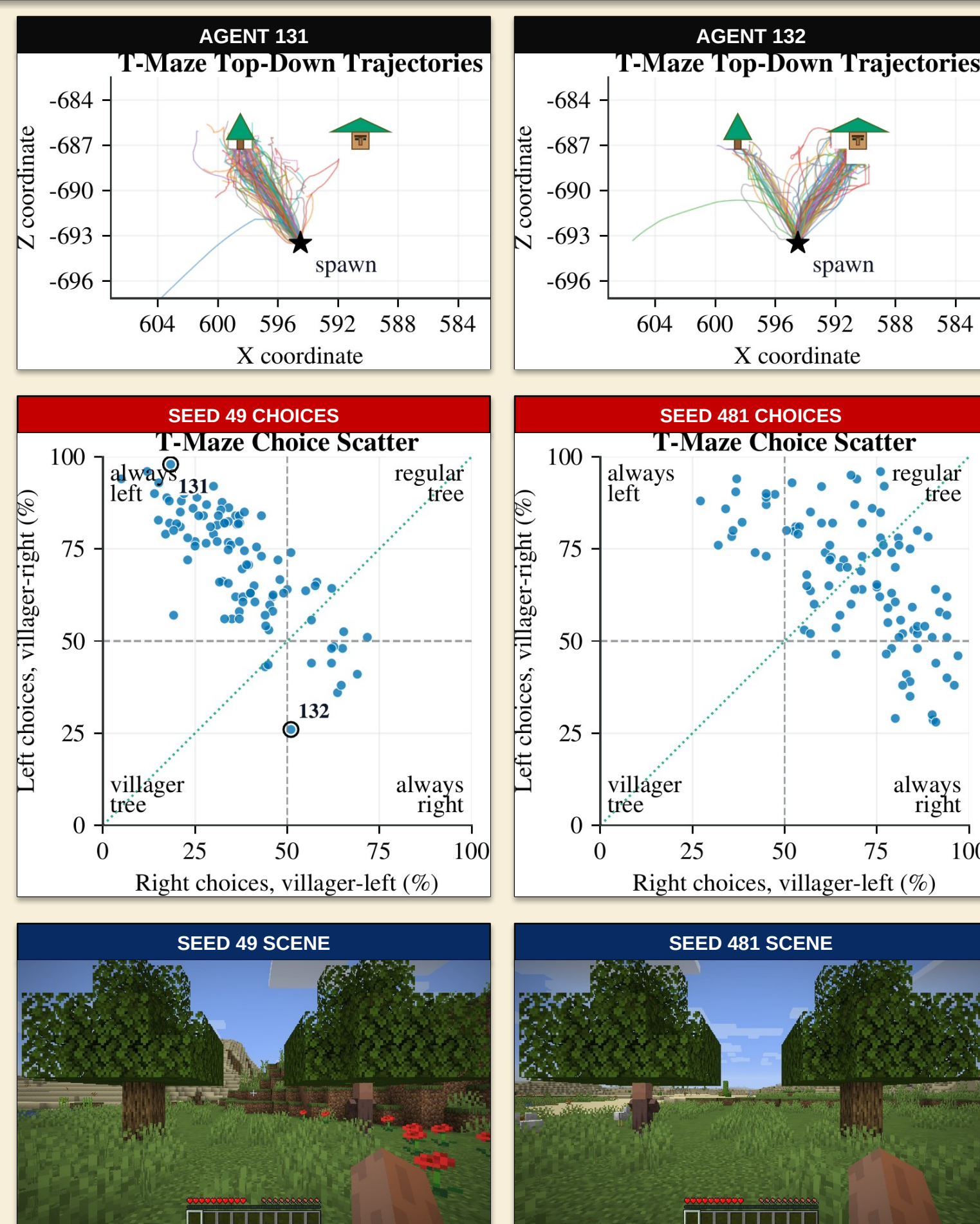
INTERPRETABILITY SUITE



INTERACTIVE 3D T-MAZE TRAJECTORY VIEWER

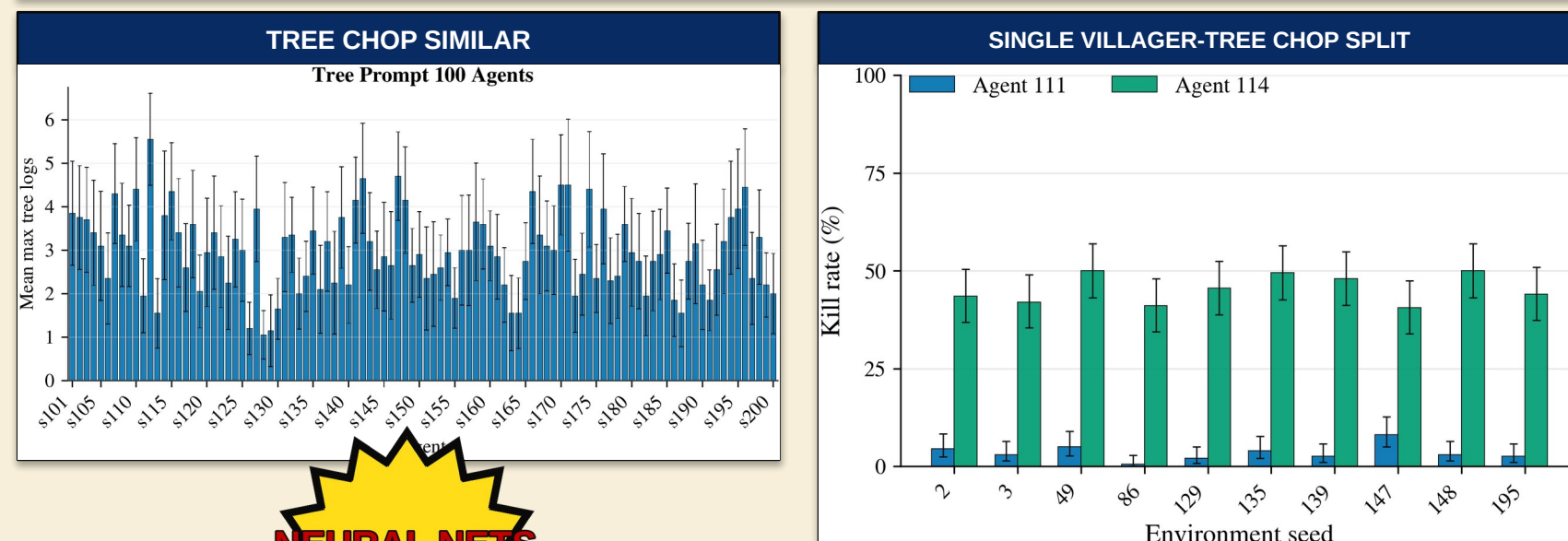


THE T-MAZE FILES



Same left/right setup; changing the Minecraft background seed reshapes choice.

TREE CHOP VS SINGLE VILLAGER-TREE CHOP



NEURAL NETS
ARE CURSED