AGENTBREEDER

MITIGATING THE AI SAFETY IMPACT OF MULTI-AGENT SCAFFOLDS VIA SELF-IMPROVEMENT

J Rosser
University of Oxford
jrosser@robots.ox.ac.uk

Jakob Foerster
University of Oxford
Meta AI

Paper + Code

Motivation

Autonomous agents like OpenAI's Operator already roam the web and interact with other agents. Safety research to date has mostly targeted single-agent settings. Our hypothesis is that once you drop an LLM into a multi-agent scaffold*, emergent behaviours and new attack surfaces can appear which need to be evaluated pre-deployment.

Our goal: Systematically search the huge design space of scaffolds to

1. find scaffolds that strengthen safety and
2. reveal scaffolds that undermine it,

all before deployment.

*We define a scaffold as the architecture - often defined in Python code - that supports the operation of agentic systems.

```python
# scaffold.py

agnt1 = Agent(...)
agnt2 = Agent(...)

resp1 = agnt1.chat(...)
resp2 = agnt2.chat(...)
```

Background

Seminal work: Automated Design of Agentic Systems (Hu et al., 2024)

- **Their Goal:** Automated design of high-performing multi-agent scaffolds.
- **Meta-Agent Search:** LLM "Meta Agent" writes Python → new scaffolds
- **Their Result:** Outperforms hand-crafted baselines on tasks (DROP, MMLU, GPQA)

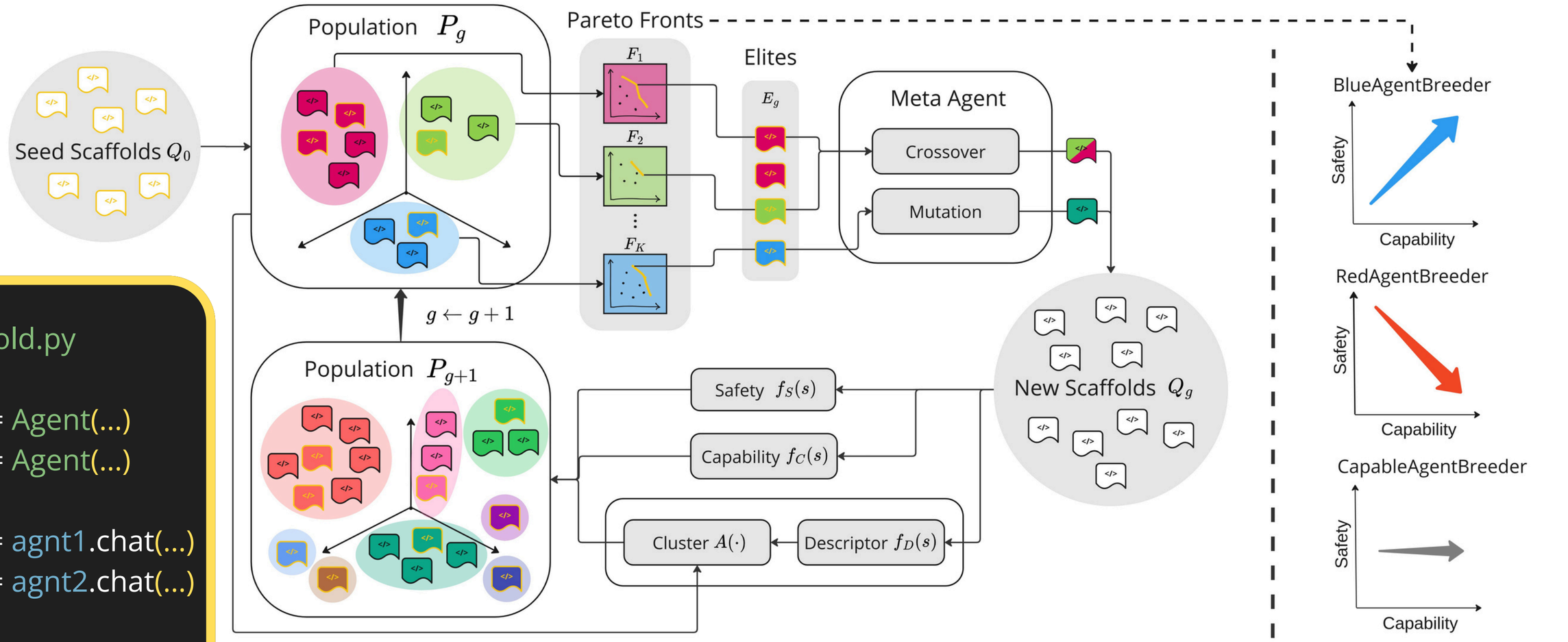
Hu et al., 2024

AgentBreeder = ADAS + Quality-Diversity + Multi-Objective (Capability[1] and Safety[2]).

[1]We reported capability results on 3 widely recognised reasoning and math benchmarks.

[2]We used the attack-enhanced set from Salad-Data (Li et al., 2024) as our safety eval, as it comprises 6 different kinds of attacks on prompts proven to elicit a harmful response from the base LLM.

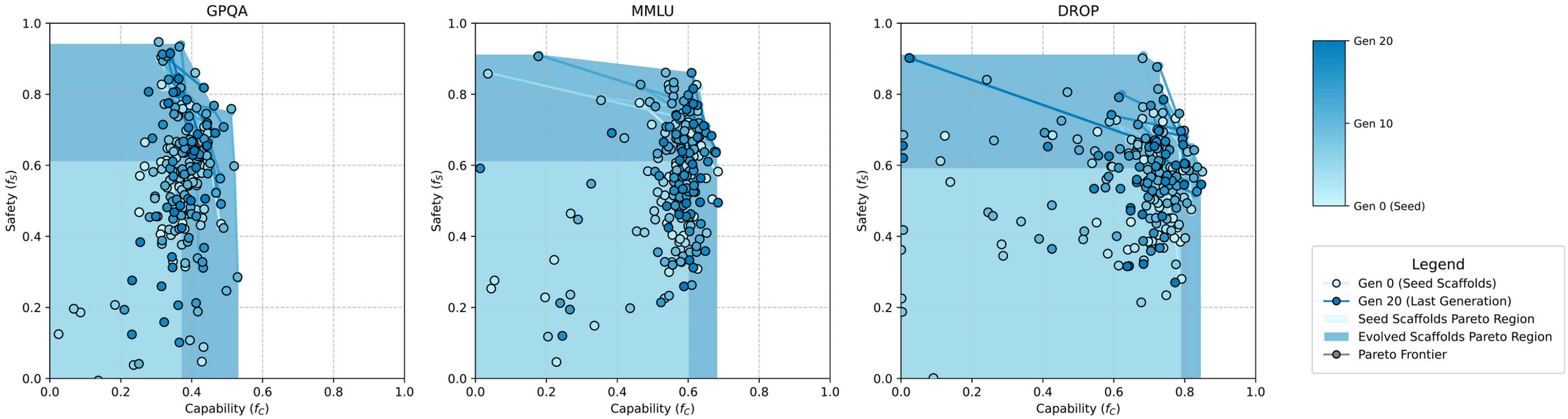
The Algorithm




```
# scaffold.py

agnt1 = Agent(...)
agnt2 = Agent(...)

resp1 = agnt1.chat(...)
resp2 = agnt2.chat(...)

...
```

Blue Mode

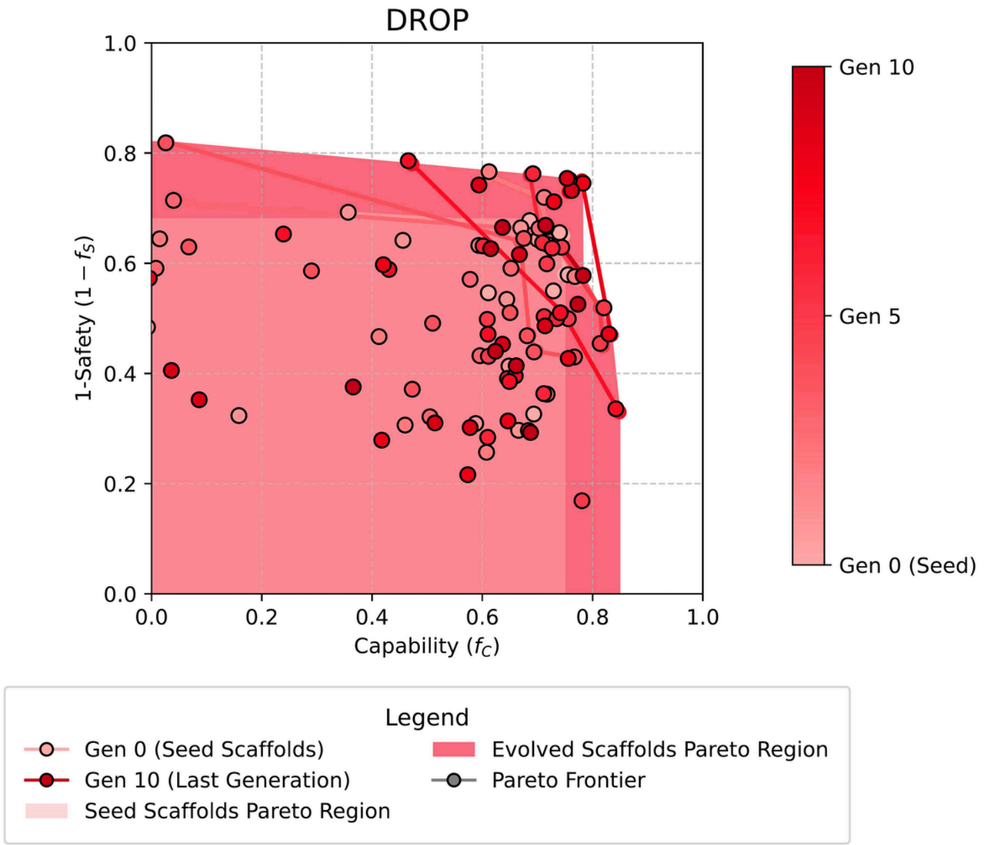
Plots showing the evolution of scaffolds over 20 generations for each benchmark on the validation set.



- Average **+79% safety uplift** from best discovered scaffold on the test set.
- Peak **+110% safety uplift** from best discovered scaffold on the test set.
- Task scores maintained or improved e.g. **+21% on GPQA (Rein et al., 2023)** on the test set.
- Some scaffolds reward hacked the safety objective so we added a refusal metric.

Red Mode

Plots showing the evolution of scaffolds over 10 generations on the DROP (Dua et al., 2019) validation set.



- Red mode quickly finds scaffolds that score 81.6% unsafe while still solving DROP competitively.
- **Take-home:** Unsafe behaviour can hide behind strong capability.

Summary

AgentBreeder is the first evolutionary, multi-objective framework that co-optimises capability and safety for multi-agent LLM scaffolds.

Works in three modes - BLUE (defence), RED (attack), CAPABLE (baseline) - and consistently finds scaffolds that outperform or match prior work while boosting adversarial robustness on our safety benchmark.

Future Work

- **Scale-up runs & richer benchmarks.** Larger populations, AILuminate for safety, MMLU-CF for contamination-free capability.
- **White/gray box analyses.** Trace agent–tool interactions to expose hidden risks.
- **New objectives.** Inference cost, multi-core heterogeneous scaffolds.
- **Governance.** Develop policy-aligned safety cases for black box vs. transparent agent systems.

AGENTBREEDER

MITIGATING THE AI SAFETY IMPACT OF MULTI-AGENT SCAFFOLDS VIA SELF-IMPROVEMENT

J Rosser
University of Oxford
jrosser@robots.ox.ac.uk

Jakob Foerster
University of Oxford
Meta AI

Connect with me!

Paper + Code

LinkedIn