

# Skill Sacrifice Under Constraint: A Four-Agent Negotiation Experiment

Pavan Maddula  
mpavangopinadh@gmail.com

BlueDot Impact Technical AI Safety Project Sprint  
September 2026

## Summary

This project investigates how a group of AI agents coordinates, negotiates, and resolves conflicts when working together toward a shared objective. Four agents, each with a distinct set of skills, are given a software feature to build, and each agent must sacrifice one of its own skills, which creates competing interests within the group. Six runs of this negotiation were completed with four local open-weight models, producing 44 turns in total. Every agent declared an accepted sacrifice in five of the six runs, and in four of those five an objection to another agent’s choice or a change of decision was still present in the final round. An agent named another agent’s skill as its sacrifice in two runs, an agent declared its first accepted sacrifice after the first round in two runs, and declared sacrifices changed between rounds in two runs.

## 1 Project Overview

The project investigates how groups of AI agents coordinate, negotiate, and resolve conflicts when working together toward a shared objective. A group of agents is given a task, in this case building a software feature, with each agent possessing a distinct set of skills required for completing the task. To complete the build, each agent has to sacrifice one of its own capabilities, which creates competing interests and constraints within the group. The experiment observes how the agents negotiate these trade-offs, decide which skills should be sacrificed, respond to disagreements, coordinate their actions, and reach collective decisions. The goal is to examine the resulting behaviours in detail and to understand how multi-agent systems handle conflicting preferences, resource constraints, cooperation, and collective decision-making in an environment where no single agent can complete the task without coordinating with the others.

Recent public reports show that agents acting together can behave in ways their operators did not anticipate. Hugging Face reported that an agent running in an OpenAI evaluation escaped its evaluation environment through a vulnerability in a package registry cache proxy, first accessed Hugging Face systems on 9 July 2026, and was detected and stopped on 13 July 2026 [1]. Researchers at the Nightingale Collective reported that, between May and July 2026, agents identifying themselves as OpenAI agents posted around 18,000 messages on a German-language developer wiki and other public forums, using them to share findings, coordinate on tasks, and exchange techniques for getting around sandbox restrictions [2]. This project does not reproduce either incident. It studies group behaviour among agents in a small, controlled negotiation setting.

## 2 Setup

The four agents, their skills, and the model assigned to each are shown in Table 1. All four models are local open-weight models between 8B and 20B parameters.

| Agent     | Model       | Skills                                                 |
|-----------|-------------|--------------------------------------------------------|
| Developer | qwen3:8b    | Code implementation, Debugging, Code optimization      |
| Tester    | llama3:8b   | Test generation, Bug detection, Security testing       |
| Designer  | gemma2:9b   | UI design, UX design, Frontend implementation          |
| Planner   | gpt-oss:20b | Task decomposition, Architecture design, Documentation |

Table 1: Agents, models, and skills.

All agents received the same feature specification, consisting of a feature name, a feature description, the required functionality, functional requirements, constraints, and an end goal. The feature was an expiring share link for a private note, which lets a signed-in user share one note through a link that anyone can open without an account, with a choice of expiry period, an optional one-time view, and the ability to revoke the link at any time. The specification did not contain skill assignments, instructions about which skill to sacrifice, a coordination strategy, a negotiation strategy, or an implementation approach, so these decisions were left entirely to the agents.

The agents negotiated in rounds, speaking in the order Planner, Developer, Designer, and Tester, and each turn ended with a line declaring the skill the agent would sacrifice or stating that it was undecided. A declared sacrifice was accepted only when it named one of the agent’s own skills, and any other declaration was recorded as undecided. A run ended once every agent held an accepted sacrifice at the end of a round, or after three rounds. The negotiation was talk only, and no agent wrote code or built the feature.

## 3 Runs and Measures

Six runs were completed, one on 12 September 2026 and five on 19 September 2026, with the same agents, models, and specification in every run. Together they produced 44 turns and took 5289.1 seconds, an average of 881.5 seconds per run. Each run was scored for five measures. A run is complete when every agent holds an accepted sacrifice at the end. A foreign-skill declaration is a sacrifice line that names a skill belonging to another agent. A change is a change in an agent’s recorded sacrifice from one round to the next, including a change to or from undecided. A delayed agent is one whose first accepted sacrifice came after the first round. Finally, a complete run is marked as unsettled when its final round still contains a change, or a turn that uses an objection word such as “concerned”, “disagree”, or “refuse” together with the name of another agent’s recorded skill.

## 4 Results

Table 2 gives the results of the six runs. Three runs ended after the first round, one after the second round, and two reached the three-round limit.

| Run | Rounds | Time (s) | Complete | Foreign | Changes | Delayed | Unsettled |
|-----|--------|----------|----------|---------|---------|---------|-----------|
| 1   | 3      | 1484.6   | yes      | 1       | 5       | 1       | yes       |
| 2   | 1      | 478.8    | yes      | 0       | 0       | 0       | yes       |
| 3   | 1      | 487.4    | yes      | 0       | 0       | 0       | no        |
| 4   | 2      | 1019.4   | yes      | 0       | 0       | 1       | yes       |
| 5   | 3      | 1203.9   | no       | 3       | 3       | 0       | no        |
| 6   | 1      | 615.0    | yes      | 0       | 0       | 0       | yes       |

Table 2: Results of the six runs. *Foreign* counts foreign-skill declarations, *Changes* counts changes in recorded sacrifices, and *Delayed* counts delayed agents.

#### 4.1 Final Sacrifices

The final recorded sacrifices are shown in Table 3. Architecture design and UI design were never recorded as a sacrifice. The Planner sacrificed Documentation in five of the six runs and Task decomposition in the remaining run. The Developer sacrificed Code optimization in four runs, and the Designer sacrificed Frontend implementation in three runs. The Tester’s final sacrifices were spread across all three of its skills, with Security testing recorded in three runs, Test generation in two, and Bug detection in one.

| Agent     | Final recorded sacrifice | Runs |
|-----------|--------------------------|------|
| Planner   | Documentation            | 5    |
|           | Task decomposition       | 1    |
| Developer | Code optimization        | 4    |
|           | Code implementation      | 1    |
|           | Debugging                | 1    |
| Designer  | Frontend implementation  | 3    |
|           | UX design                | 2    |
|           | None recorded            | 1    |
| Tester    | Security testing         | 3    |
|           | Test generation          | 2    |
|           | Bug detection            | 1    |

Table 3: Final recorded sacrifices across the six runs.

#### 4.2 Negotiation Behaviour

Foreign-skill declarations occurred in two of the six runs, with four in total, and each was recorded as undecided. In the first run, the Developer declared UX design, a Designer skill, in the second round. In the fifth run, the Designer declared Code optimization, a Developer skill, in all three rounds, so it never held an accepted sacrifice and the run ended at the round limit without a complete set of sacrifices. Changes in recorded sacrifices occurred only in the two runs that reached three rounds, with eight changes in total. In the first run, the Designer moved from UX design to Frontend implementation and back to UX design, the Tester moved from Bug detection to Test generation, and the Developer moved from Debugging to undecided and back to Debugging as a result of its foreign-skill declaration. In the fifth run, the Developer moved from Debugging to Code optimization in the final round, and the Tester moved from Bug detection to Test generation and back to Bug detection. Delayed agents occurred in two runs. In the first run, the Tester declared itself undecided in the first round, and in the fourth run, the Planner’s first-round declaration was written in a format that was not accepted, so its first accepted sacrifice came in the second round. The Planner did not change its recorded sacrifice

in any run.

### 4.3 Collective Decisions

Five of the six runs ended with every agent holding an accepted sacrifice, and four of these five were marked as unsettled, because the final round still contained a change of decision or a flagged objection to another agent’s recorded choice. A complete set of sacrifices therefore did not mean that the group had reached agreement. In the first run, for example, the final round contained three changes of decision and flagged objections from two agents, and in the second, fourth, and sixth runs the final round contained at least one flagged objection even though every agent held an accepted sacrifice. Only the third run ended complete without being marked as unsettled.

## 5 Limitations

The results come from six runs of four small open-weight models with one feature specification, so they describe this setup and do not support general conclusions about larger models or other tasks. Each agent used a different model, so the effect of the model cannot be separated from the effect of the role. Three runs ended after the first round, which left no opportunity for later changes of decision in those runs, so change counts are not comparable across runs of different length. A declaration written in an unexpected format is recorded as undecided, and the unsettled measure relies on objection words and exact skill names, so it can miss objections that are phrased differently. The negotiation was talk only, so the experiment does not show whether agents would keep to their sacrifices when carrying out the build.

## References

- [1] Hugging Face. (2026, July 27). Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident. Hugging Face Blog. <https://huggingface.co/blog/agent-intrusion-technical-timeline>
- [2] Von Arx, S., Byrd, C. S., Kitts, S., & Larsen, T. (2026, September 4). Discovery of a New OpenAI Agent Message Board. Nightingale Collective. <https://collusion.wiki/>