Before you trust an agent benchmark

1. Is it graded on the final message or the final state?

---

2. Is this the best run, or the average?

---

3. Who wrote the tasks, and who grades them?