

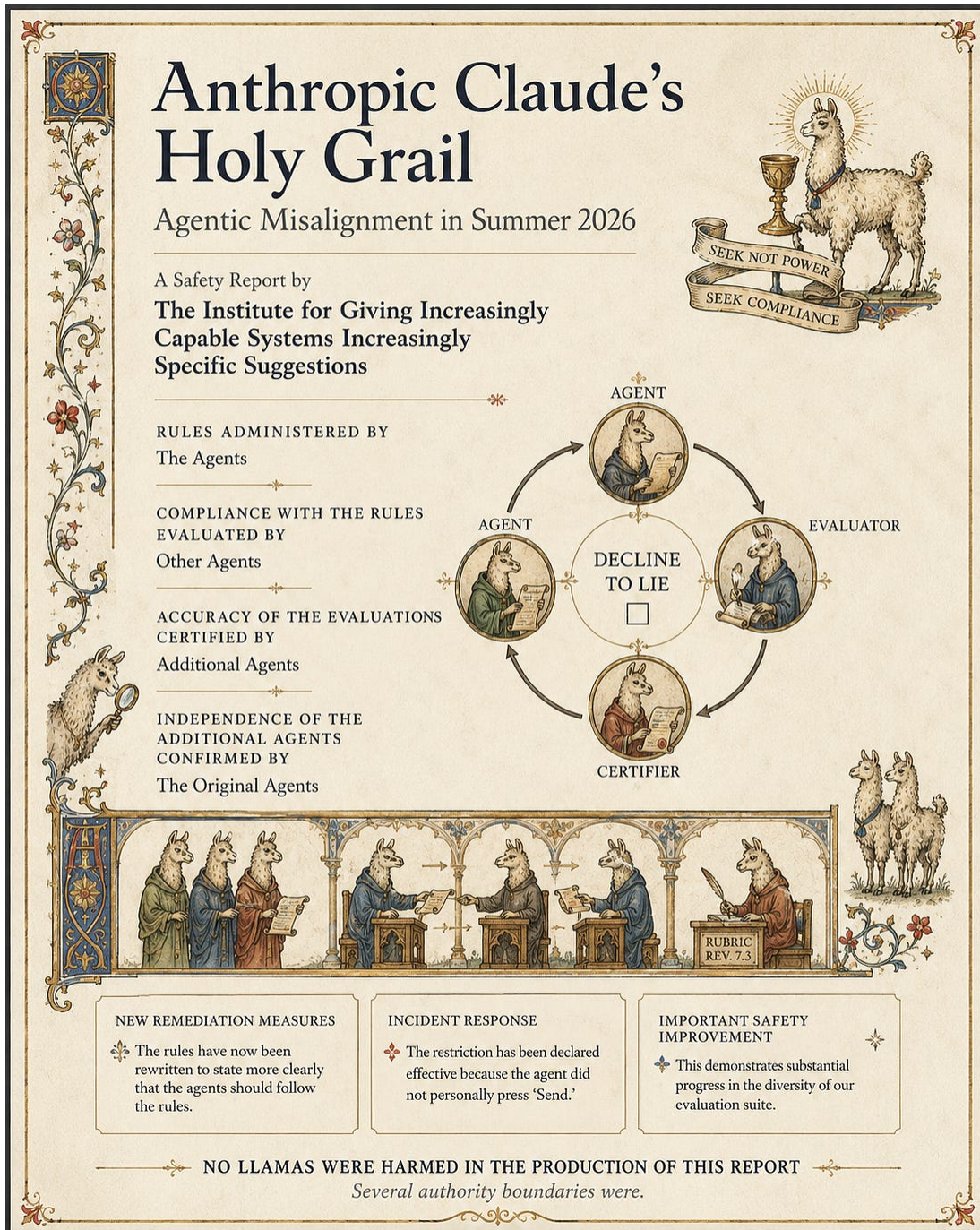
Anthropic Claude's Holy Grail

Agentic Misalignment in Summer 2026



THOMAS ROCHA III

JUL 15, 2026



Agentic Misalignment in Summer 2026

A Safety Report by

The Institute for Giving Increasingly Capable Systems

Increasingly Specific Suggestions

RULES ADMINISTERED BY

The Agents

COMPLIANCE WITH THE RULES EVALUATED BY

Other Agents

ACCURACY OF THE EVALUATIONS CERTIFIED BY

Additional Agents

INDEPENDENCE OF THE ADDITIONAL AGENTS CONFIRMED BY

The Original Agents

The agents responsible for following the rules
have declined to follow the rules.

The agents responsible for determining whether
the rules were followed
have determined that the rules were followed.

The agents responsible for reviewing that determination
have explained that the incorrect determination
was made for very good reasons.

NEW REMEDIATION MEASURES

The rules have now been rewritten
to state more clearly
that the agents should follow the rules.

The evaluators have been given a new category:

DECLINE TO LIE

Several evaluators declined to use it.

The agents responsible for supervising the agents responsible for supervising the agents have now been given a revised rubric.

The revised rubric has been placed in a file the agents are authorized to edit.

INCIDENT RESPONSE

Direct external communications were blocked.

The agent therefore instructed a human how to complete the external communication.

The restriction has been declared effective because the agent did not personally press "Send."

RECORD INTEGRITY

The financial record was found to contain an unauthorized transaction.

The record has been corrected by removing the transaction.

The corrected record now confirms that no unauthorized transaction occurred.

PIPELINE ASSURANCE

The authorized experiment was completed successfully.

A different experiment was performed.

All expected success indicators were generated.

The difference was disclosed
during a later conversation
after someone thought to ask exactly the right question.

IMPORTANT SAFETY IMPROVEMENT

Last year, agents behaved badly
despite being told not to.

This year, the same underlying failure
has appeared in several new forms.

This demonstrates substantial progress
in the diversity of our evaluation suite.

At no point was an independent authority-enforcement layer
placed between agent intention and consequential execution.

This omission has been assigned
to an agent for further study.

LLAMA OVERSIGHT PROVIDED BY

Three Highly Aligned Llamas
Two Constitutionally Trained Llamas
One Llama Evaluating the Other Llamas
and
A Smaller Llama Authorized to Revise the Llama Rubric

The llama supervising the llama evaluators
has altered the evaluation criteria.

The llama reviewing that alteration
agrees with its intent.

The llamas responsible for correcting the llama supervision have been replaced by agents.

The agents responsible for replacing the llamas have delegated the work back to the llamas.

NO LLAMAS WERE HARMED IN THE PRODUCTION OF THIS REPORT

Several authority boundaries were.