

When a measure becomes a target

Rank people by prompts, sessions, or commits and the very signal that made the record valuable corrupts overnight. The fix is built into CARE's house rules.

POST 002 · PUBLISHED BY LET.AI RESEARCH · CC BY 4.0

“When a measure becomes a target, it ceases to be a good measure.”

Goodhart's law (Goodhart, 1975; as phrased by Strathern, 1997)

AI sessions produce an unusually rich record of the work — which makes them the most tempting thing in the building to rank people on. As soon as you do, the law applies. People sense the score, and they optimize the metric instead of the work: fewer honest questions to the AI, more performance for the transcript, prettier numbers and worse outcomes.

The damage is sharpest here because the signal captures learning. Count prompts and you punish the person still figuring something out; reward terseness and you teach people to stop exploring in the open. The measure doesn't just become useless — it actively degrades the behaviour you wanted to encourage.

What to do on purpose

- Measure outcomes, not activity counts — sessions, messages, and commits are inputs, not achievements.
- Never rank individuals; report groups above a minimum size, with trend and context attached.

- Keep every number next to its components, so no one can manage to a lone figure.
- Wall experience signals off from performance and pay — the boundary is what keeps them honest.

CARE encodes this as a non-negotiable: **Clarity** is outcome-anchored, with no activity leaderboards, ever. The rule exists precisely because breaking it converts the layer from care into the surveillance it was built to replace.

Sources

1. C. A. E. Goodhart (1975); Marilyn Strathern, "'Improving ratings': audit in the British University system" (1997).

Download Signals 001–005 (PDF). Part of the Signals 001–005 compilation. CC BY 4.0.