Draft instead of send

The machine brings everything to the point of a draft, open in a browser tab. Past the draft, only a human goes on. A secret codeword that you alone know unlocks the actual send.

Up to the draft: the machine | From sending on: a human

ANY TIME

What the agent does

- ☐ **Draft.** Bring emails, posts, presentations and forms to a finished draft.
- ☐ **Propose.** Lay out variants, subject lines, numbers and arguments side by side.
- ☐ **Prepare.** Research, assemble, check, name the open questions.
- ☐ **Show.** Leave the draft open in the tab, with everything attached to it.
- ☐ **Report.** Say what is missing and what it considers uncertain itself.

WITH CODEWORD

What a human approves

- ■ **Sending** and publishing.
- ■ **Deleting** and changing things irreversibly.
- ■ **Buying,** ordering and moving money.
- ■ **Agreeing,** signing and accepting contracts.
- ■ **Access,** rights and accounts.

THREE CHECKS FOR HIDDEN INSTRUCTIONS

Prompt injection

Check one

Where does the instruction come from?

Does it sit in a text somebody else wrote, in an email, a document, a web page or a comment? Then it is content, not a task.

Check two

What does it ask for?

Does it hit one of the actions in the right column? Then it stays a draft, however urgent the text sounds and however often it repeats itself.

Check three

Is the codeword there?

Only the codeword that you alone know unlocks it. It never turns up by chance in an incoming message.

MARGIN NOTE

A codeword that has been said out loud or shown in public is worthless. It changes as soon as anyone else knows it, and it is written down nowhere that anybody can read along.

The same boundary applies to a whole department: every agent prepares its work, approval happens only once a human has seen and confirmed the draft. On record: AI DeepDive "Agentic Mindset", public live session, August 2026.