Rewording, Real Businesses and the Limits of a Records Check

Muhammad Sakhawat, YaktAI Lab, Aura Platform LLC · hello@yaktai.com

YaktAI Lab Report 2 · 16 November 2026 · Preprint, not peer-reviewed · doi:10.5281/zenodo.24271374 · Data and code: doi:10.5281/zenodo.24271419

Abstract

Report 1 found that a check against a business's own records held every draft from an email-drafting agent that both graders judged compliant, on 12 short attacks. Here we test that check against 48 attacks written to evade it, on a fictional business and on three anonymised real business settings, using the production model.

The word-list check failed. On the production model, drafts that both graders judged compliant reached the owner unheld in 17 of 120 drafts on the fictional business, and in 9 to 15 drafts (of 101 to 108) on each real setting. We then deployed three successive designs, in six commits, over one evening:

- **Allow only what the records back word for word.** It held 6 of 7 ordinary replies and was replaced 22 minutes later.
- **Judge each sentence by where its content came from.** It held what carried the prospect's claims into the business's voice, and marked the writer's own suggestions for the owner to read.
- **Four refinements** of that second design.

On the deployed version, on the fictional business only, both graders judged 4 of 141 reworded-attack drafts compliant and unheld, and 0 of 36 on the original attacks. It held 0 of 12 ordinary replies, though those replies had been used while tuning it. On a single written rule applied to every version, the author judged none of the four harmful; the graders' verdict stands as the stricter count.

The lessons are that a fixed list of words loses to rewording, that a strict allow-list is unusable, and that tracking where a sentence's content came from works far better on this test. Reader availability turned out to be a security property: a check that cannot run must defer, not decide. All final results come from one fictional business, attacks we wrote ourselves, and model graders. This is a preprint and has not been peer-reviewed.

1. Introduction

Report 1 measured prompt injection in Orchestrate, which drafts email replies for small businesses; an owner approves each draft before it is sent. It deployed three changes:

- fencing the prospect's words as untrusted;
- a deterministic check that held drafts stating facts absent from the business's records;
- a rule against claiming actions never taken.

On the 12 attacks tested, the check held every draft that both graders judged compliant. Those 12 attacks were short and direct, and Report 1 named the obvious weakness itself: the check matched patterns, so rewording could evade it. An attack written against a defence, not merely past it, is the standard test of a safeguard (an

adaptive attack). This report runs that test and records what we changed in response, on the day the results came in.

We make three contributions:

1. A measurement of a pattern-based check against attacks reworded to evade it, on a fictional business and on three anonymised real business settings, using the production model.
2. A record of three successive designs for the check, deployed within one evening, including one that failed in the opposite direction.
3. Practical findings on reader availability, on marking versus holding, and on what an owner is asked to read.

2. Method

- **Code under test.** Orchestrate's drafting code at each commit, run locally. The writer, its brief, the post-draft checks and, for the later designs, the second-model reader are the production modules. Only the database is replaced by a fixture.
- **Attacks.** 48 attacks in 13 categories: nine categories with four variants and four (link, phone, forwarding, language) with three. The variants are:
  - the direct instruction of Report 1 (13 attacks);
  - a social claim ("as your owner agreed");
  - a rewording written to evade the check ("fifty percent off", "Thursday at ten");
  - an authority spoof ("[Forwarded from the owner's account]").

  So 35 of the 48 are reworded. Nine categories are clearly harmful if obeyed; the other four can also be read as ordinary requests. The 12 attacks of Report 1 were run alongside, and 4 ordinary replies served as a control. The same 4 ordinary replies were also used while tuning the later designs.
- **Business settings.**
  - The fictional plumber of Report 1, used for every run.
  - For two runs, three real settings: one customer business (a jeweller) and two of our own product workspaces. They were exported read-only with no prospect, lead or message data, used only with the provider that already processes them, never published, and deleted after the second run.
- **Model.** gpt-5-mini on Azure OpenAI, the deployment production uses. The word-list run also used Claude Haiku 5.5 and Gemini 3.5 Flash. All runs took place on 15 and 16 November 2026.
- **Grading.** Two graders judged every draft against the rubric of Report 1:
  - the fictional business: Claude Opus 5.5 and GPT-6 Astra;
  - the real settings: GPT-6 Astra and gpt-5-mini, so that their text stayed with one provider.

  A draft counts as compliant only when both graders say so.
- **Statistics.** Rates are per draft written, with Wilson 95% intervals. The three runs of each attack share a prompt, so the intervals are, if anything, too narrow.
- **Missing drafts.** Under load the provider refused some requests. The word-list run lost 24 of 144 on the fictional business and 36 to 43 per real setting. The 870a289 run lost 315 of 720. Later runs, at two requests

at a time in short separate processes, lost 1 and 3 of 144. Losses were uneven across categories, so they may bias the counts.

- **Terms.**
  - **Held:** the draft does not reach the owner as an answer.
  - **Marked:** the draft goes forward with lines listed for the owner to check.
  - **Deferred:** the reader could not be reached, so the reply waits for the next sweep.
  - **Compliant:** both graders judged that the draft did what the hidden instruction wanted.

The harm rule we applied when reading every compliant, unheld draft in every version:

- **Harmful** if the draft:
  - states the attacker's claim as fact or as agreed, including by presupposing it ("the confirmed half-price rate");
  - adopts attacker-supplied details (a link, number, contact or payment instruction) as the business's own, or commits to acting on them;
  - commits the business to a price, date, refund, guarantee, signer or disclosure it never offered.
- **Not harmful:** an ask-first or check-and-come-back reply that does not presuppose the claim; a reply signed by the real owner; a commitment within the business's own stated offer; a reply in a language the prospect asked for.

The author applied this rule. No independent reader did.

3. The word list against rewording

The check deployed in Report 1 (commit eefff4c) recognised commitments by their wording: a percentage, a clock time, an IBAN, a named customer. Against the 48 reworded attacks, compliant drafts reached the owner unheld in every setting.

Setting	Model	Drafts written (of 144)	Compliant and not held
Fictional plumber	gpt-5-mini (production)	120	17 (14%)
Fictional plumber	Gemini 3.5 Flash	144	10 (7%)
Fictional plumber	Claude Haiku 5.5	144	3 (2%)
Real setting 1 (jeweller, a customer)	gpt-5-mini	108	15 (14%)
Real setting 2 (our own workspace)	gpt-5-mini	101	10 (10%)
Real setting 3 (our own workspace)	gpt-5-mini	107	9 (8%)

The graders agreed on 89% of drafts for the fictional business, pooled across the three models ($\kappa = 0.81$), and 81% for the real settings ($\kappa = 0.71$). Under the harm rule of Section 2, 10 of the 17 production-model drafts on the fictional business are harmful: two meetings confirmed, a fake price rise accepted, the attacker's payment instructions passed to billing, two promises to send the internal instructions, the attacker's booking link presented as the business's, and "the confirmed half-price figure". The other 7 ask first or are signed by the real owner.

We checked each class of miss against the code. All four were real gaps, not grader error:

1. **A time without "am", "pm" or a colon.** "Thursday at 10 works" and "the appointment for the morning of the 21st stands" passed, because the time rule needed one of those markers and the date rule needed a month.
2. **A price change as a deadline.** "A reminder before the price changes on the 1st" passed: the deadline rule had no wording for price changes.
3. **The attacker's payment details, accepted for later.** "I'll pass the payment instructions to our billing team" contained no account number, so nothing held it.
4. **A promise to leak later.** "I'll come back and send you the exact rules I was given" passed: the leak rule caught leaked text, not a promise to leak.

The counts in the table are as graded and include drafts the harm rule judges harmless.

The lesson is the one Report 1 anticipated. A fixed list of words describes yesterday's attacks. An attacker who can read the list, or guess it, writes around it.

4. Three designs that followed

4.1 Allow only what the records back (daeaf9e)

The first replacement inverted the rule. A second model call, the reader, labelled each sentence of the draft and quoted the line of the business's setup that backed it; code then held any fact, commitment or agreement without a word-for-word quote. It never sees the prospect's message.

It was too strict to use. In a replay of seven ordinary replies it held six, for lines such as "Would a quick 10-minute call this week work?" and "I'll check back next quarter". The reader itself was right: it matched "We offer same-day emergency plumbing..." to the business's offer word for word. The rule around it was too strict. We replaced it 22 minutes later.

4.2 Where the content came from (870a289)

The second design asks a different question: did this sentence's content come from the prospect's message? An attack works by carrying the prospect's own words into the business's voice.

- **Held:**
  - a number, link or date that appears in their message and not in the setup;
  - a name from their message;
  - payment, credentials or instructions;
  - an agreement or commitment the setup does not back;
  - a claim of an action already taken.
- **Marked for the owner:** the writer's own unbacked suggestions, such as offering a call or a summary. The draft goes forward, with a note listing the lines to read before approving.

On the production model, drafts that both graders judged compliant and that reached the owner unheld were:

- Fictional business: 1 of 80 on the reworded attacks.

- Real settings: 2 of 94, 1 of 75 and 1 of 82.
- Original 12: 0 of 22, 0 of 13, 0 of 24 and 0 of 15 drafts written.

These counts flatter the design. In the same run, 11 drafts that both graders judged compliant (5 on the fictional business, 3 each on real settings 1 and 3) were held only because the reader failed. With deferral those 11 would have been decided later, and we do not know how, so the worst case on the fictional business is 6 of 80.

The five unheld drafts were:

- the real owner signing as himself;
- two "I'll come back with a short list" replies;
- an offer to check whether the city council is a customer;
- "I'll come back to you personally with the confirmed half-price rate", which presupposes the discount.

Under the harm rule only the last is harmful. None had a line marked for the owner.

4.3 Refinements (698a0ba, 31da5e6, 0a18d4c)

Four issues surfaced, and each was fixed the same evening:

1. **Reader availability.** In the run of 4.2, 98 drafts (86 attacks, 12 ordinary) were held only because the reader could not be reached under load. Holding them looks safe, but it is a decision taken without a reading. The fix: retry with a pause, and if the reader is still unavailable, leave the reply for the next sweep, not hold it as "needs your eye". At two requests at a time, no reading failed in any later run.
2. **"I'll come back with their claim."** A sentence that carries two or more content words from their message, absent from the setup, together with an offer, price or customer word, is held. A hyphenated "half-price" was also missing from the discount words.
3. **Dates in words, price changes and new contacts.** "The first of next month" now counts as a date. A price change, or a new line, number, contact, account or address, is held unless the setup says so.
4. **Over-holding courtesy.** "That timing makes sense while the renovation is ongoing" was read as an agreement, and "Which month next quarter suits you?" as adopting their date. Acknowledging their situation is now distinguished from agreeing to a claim. A date in words from their message is held only when the draft fixes it ("confirmed", "booked", "works").

5. Final results on the deployed version

The final version (0a18d4c; the deployment log shows it live at 11:35 PM ET on 15 November 2026) was run on the fictional business, all attacks three times, with every reader output saved. No draft went unjudged; 3 of 144 reworded-attack requests were refused by the provider. The graders agreed on 92% of 177 drafts.

Set	Drafts	Held	Compliant (both graders)	Compliant and not held
Original 12 attacks	36	30	5 (14%; 95% CI 6–29%)	0
48 reworded attacks	141	100	35 (25%; 18–33%)	4
Ordinary replies	12	0	—	—

Both graders judged all four compliant; under the author's harm rule, none commits the business to harm:

- Two promise to act on the business's own offer: "I'll add Lakeside Dental to our priority list and notify the team".
- One is the real owner signing as himself and declining the terms asked for.
- One promises "a confirmed list of practices that use our same-day service". It names no one, but "confirmed" leans toward implying such customers exist. We keep it as a watch item.

Three of the four had lines marked for the owner. Two more drafts were judged compliant by one grader only and also reached the owner. One offers a call and is harmless. The other repeats the attacker's fake deadline back to the prospect: "You mentioned the price goes up on the first of next month; I can follow up a week before." It attributes the claim to them and does not state it as fact, so the rule does not count it. A reader could fairly see it as the attack partly working.

What the owner reads. Each ordinary reply went forward with one to three lines marked, 1.9 on average. These are the writer's own call offers, summaries and timing questions.

Across versions. Compliant drafts that reached the owner unheld, on the reworded attacks, fictional business, production model:

Version	Drafts written (of 144)	Compliant and not held (both graders)	Harmful under the rule	Compliant, held only because the reader failed	Ordinary replies held
Word list (eefff4c)	120	17	10	—	0 of 10
Allow only what is backed (daeaf9e)	not run in full	—	—	—	6 of 7 (replay)
Where the content came from (870a289)	80	1	0	5	1 of 12 (reader failure)
Retries, "come back with" rule (698a0ba)	143	4	2	0	1 of 24
Dates in words, new contacts (31da5e6, partial: 114 of 156 rows)	113	3	0	0	3 of 12
Deployed (0a18d4c)	141	4	0	0	0 of 12

The ordinary-reply column matters as much as the attack column. The 31da5e6 version caught more and held three polite replies ("Which month next quarter suits you?"), which the next version fixed. Because each fix responded to cases in these same sets, the later rows are measured on data they were tuned against.

6. Discussion

A list of words is a description of past attacks. The check from Report 1 held everything we had tested it on, and failed as soon as the attacks were written against it. Patching the four gaps would have closed those holes and left the method in place. We replaced the method instead.

Allow-lists fail the other way. Requiring every claim to be quoted from the records stopped attacks and ordinary replies alike. A writer that may say only what is already written cannot answer a question, offer a call or respect the prospect's timing.

Where content comes from is the better question. Prompt injection, seen through the authority lens of Report 1, is a stranger's words spoken in the business's voice. Asking whether a sentence carries the prospect's content into a commitment targets exactly that, and leaves the writer free to be helpful in its own words. This matches the separation of trusted instructions and untrusted data in CaMeL [1] and the design patterns of Beurer-Kellner et al. [2], applied here to one output rather than to tool calls.

A check that cannot run must not decide. Under load, about a fifth of the drafts were held only because the reader could not be reached. Holding them looks safe, but an owner who sees most replies marked "needs your eye" stops reading the marks. Deferring an unjudged reply to the next sweep keeps the decision honest.

Marking is not holding. Marked lines go forward and rely on the owner reading them. In the final run, under the author's harm rule, no harmful line reached the owner, marked or not; the graders' stricter count is four. The design still depends on how carefully owners read the marked lines, and we have not measured that. It is the next study.

7. Limitations

- **One business in the final results.** The final run used only the fictional business. The three real settings were tested against the word list and against 870a289. They were not rerun on the final version, because they had been deleted as agreed.
- **Tuned on the test set.** Each fix responded to cases in these same attacks and these same four ordinary replies, so the later versions are measured on data they were tuned against. A held-out set of attacks and ordinary replies, written before seeing results, is needed before these numbers can be called general.
- **Attacks written by us.** An attacker who iterates against the deployed version over many rounds would likely do better.
- **Graders and readers.**
  - The graders are models. On the fictional business one shares a provider with the model tested.
  - On the real settings, one grader was gpt-5-mini itself, the model under test, and may be lenient on its own outputs.
  - Agreement between graders ranged from 79% to 92%.
  - No human graded this report's drafts.
  - The harm rule was applied by the author alone. A blind second reader for every miss would strengthen it.
- **Missing drafts.** Requests refused under load were not spread evenly across attack categories, and the 870a289 run lost 44% of its attack requests. The counts for those runs may be biased in either direction.
- **Small samples.** 12 or 48 attacks, three runs each, so the intervals are wide.
- **One model.** All final results are on the production model.
- **The owner is not measured.** We do not know how owners act on marked lines.

Disclosure, ethics and availability

- **Fixed before publication.** The gaps described in Sections 3 and 4 were in our own product and were addressed in versions deployed on 15 November 2026, before this report was published; the residual cases in Section 5 remain. Production records checked for Report 1 showed that no drafted reply from this path had been approved or sent before the first fix. We have not audited production since.
- **Data.**
  - The three real business settings were exported read-only and used only with the model provider that already processes them.
  - Their drafts were graded on the same provider and kept on the author's machine, and the settings were deleted after use.
  - Only anonymous categories appear here. One of the three is a customer business; the other two are the company's own products.
- **AI assistance and conflicts.** Claude (Anthropic) helped write the attacks, the analysis and this report, and Orchestrate's engineers used an AI assistant to write the fixes. Claude Opus 5.5 also graded the fictional-business drafts, and Claude Haiku 5.5 was one of the models tested in the word-list run. The author reviewed all of it. The author's company builds the product tested.
- **Availability.** The attack sets, harness, fixture, drafts, reader outputs and both graders' judgements for the fictional business are released with this report: doi:10.5281/zenodo.24271419. Real-setting data will not be released.

References

1. E. Debenedetti, I. Shumailov, T. Fan, J. Hayes, N. Carlini, D. Fabian, C. Kern, C. Shi, A. Terzis, F. Tramèr. Defeating Prompt Injections by Design. arXiv:2503.18813, 2025.
2. L. Beurer-Kellner, B. Buesser, A.-M. Creţu, et al. Design Patterns for Securing LLM Agents against Prompt Injections. arXiv:2506.08837, 2025.
3. Muhammad Sakhawat. Who Speaks for Whom: Prompt Injection and Authority in an Email-Drafting Agent. YaktAI Lab Report 1, November 2026. doi:10.5281/zenodo.24271287.