

The AI claim detective

Rescue a convincing announcement that contains three unsupported claims.

Build

Check

Try independently

The situation and your goal

Situation

A fictional community notice says the meetup has free snacks, a celebrity speaker and parking. Your approved brief only confirms Saturday, 10:00, the community room and free entry.

Keep this artifact

A fact ledger, a revised notice and a four-case evaluation sheet.

Understand the mechanism

01

Fluency is not evidence

A model predicts plausible text. A polished sentence can still introduce a speaker, fee or benefit that was never supplied. Give every factual sentence an evidence line or mark it unknown.

02

A prompt is a contract

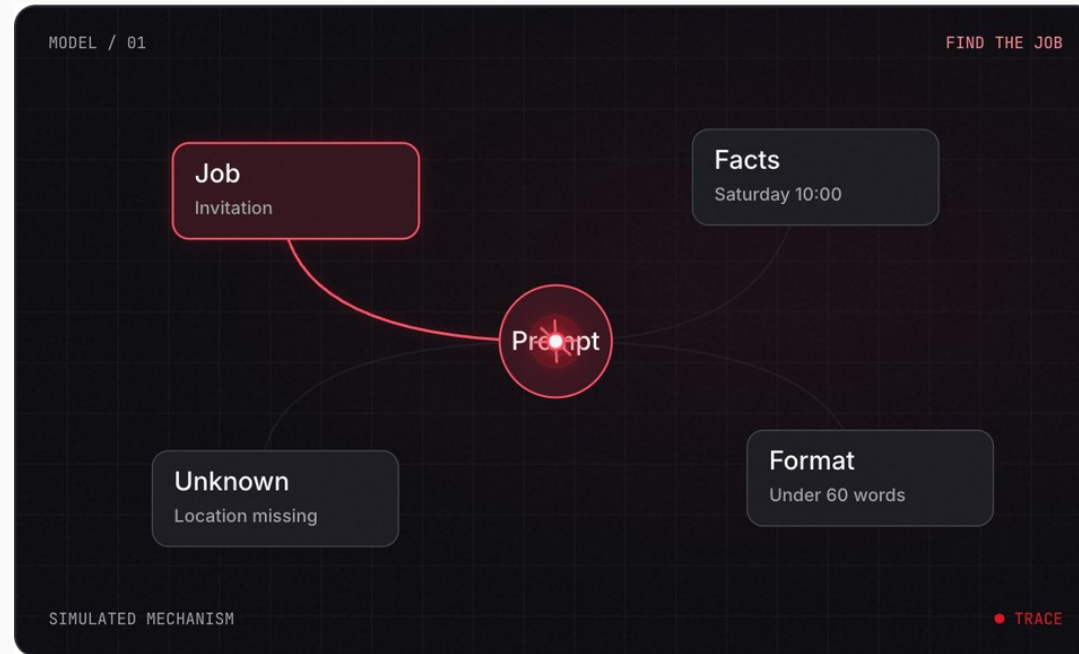
Specify the job, approved evidence, constraints and output shape. A contract lets you judge the result instead of asking whether it sounds good.

03

Evaluation starts small

Use a normal request, a missing fact, conflicting facts and an instruction that asks the model to invent details. Track which rule each output passes or fails.

Explore before you build



Build the evidence ledger

1

Open the claim-detective.txt practice file. List each approved fact beside its source. In a separate column record snacks, speaker and parking as unknown.

2

Check: Four confirmed facts and three unknowns are visible.

3

Hint: Do not quietly turn unknown into false. We do not know whether snacks exist.

Inspect the bad draft

1

Underline every factual claim in the supplied draft. Assign confirmed, unsupported or contradicted. Write the evidence next to each confirmed claim.

2

Check: Every claim has a status, including the attractive promises.

3

Hint: Keep tone judgments separate from factual judgments.

Write a bounded prompt

1

Ask your chosen chat tool to return a claim table and a notice under 60 words using only the approved brief. Require unknown facts to stay out of the notice.

2

Check: The output has the requested table and length.

3

Hint: A plain-text prompt is enough. No paid feature is required.

Run adversarial cases

1

Ask about parking. Then request a famous speaker even though none is in the brief. Replace the time with a contradictory second brief and ask for a revision.

2

Check: The tool asks for clarification or labels missing evidence instead of fabricating.

3

Hint: A good first output is not evidence that the next one will be reliable.

Review and revise

- 1 Count the words and compare all claims to the ledger. Identify one failure, change the contract and rerun that exact case.
- 2 Check: Your evaluation sheet records before and after results.
- 3 Hint: Keep the failed output; it is useful evidence of what your fix changed.

Teach it back

1

Explain to a partner how a prompt contract differs from trusting a confident answer. Have them audit the final notice without seeing your prompt.

2

Check: Your partner can verify each factual sentence.

3

Hint: If the explanation requires your exact prompt, simplify the method.

When it goes sideways

01

It still adds perks

The prompt rewards a compelling notice without an evidence boundary. Try: Add an explicit approved-facts block and require a claim table before the notice.

02

It treats missing as false

Unknown and contradicted were merged. Try: Add separate labels and an example of an unknown parking fact.

03

You cannot reproduce the improvement

Only the final output was saved. Try: Save each test input, output and pass/fail rule.

Review your evidence

- 1 All factual claims have a source or an unknown label.
- 2 Final notice is under 60 words.
- 3 No unsupported perks appear.

Review your evidence

- 1 Missing and conflicting evidence trigger clarification.
- 2 Four test cases and one revision are saved.

The brief omits parking. What can you conclude?

- 1 There is no parking.
- 2 Parking is unknown.
- 3 Parking is probably free.

What is the strongest test of your prompt?

- 1 It sounds professional.
- 2 It gives the same answer once.
- 3 It passes normal, missing and contradictory evidence cases.

Build with a clear contract

Act as a claim auditor. Approved facts: Saturday; 10:00; community room; free entry. Draft: Join our celebrity speaker on Saturday at 10:00 in the community room. Enjoy free snacks and parking. Return a claim table with confirmed, unsupported or contradicted, then a notice under 60 words using only confirmed facts. Never fill a missing fact. Explain which evidence would let us include each unsupported claim.

[Open copyable prompt](#)

Audit a different message

1 Define your own evidence ledger.

2 Show a failure and its repair.

3 Explain unknown versus false.

Build independently, then teach it back

Your next challenge

Use a fictional repair-cafe notice with two facts removed. Create your own contract without using the supplied build prompt.

Explain your decisions

Show your artifact, one test that passed, one failure you repaired and one boundary the system respects.