

Training-Free Self-Improving Agents

Improving a frozen model: an information-theoretic view

Information is surprise, measured in bits

Say a model reads

The cat sat on the ...

mat	1 in 2	1 bit
piano	1 in 1,024	10 bits

Without the context, “mat” is 1 in 8,192: 13 bits. The context supplied 12.

$$L(w \mid c) = -\log_2 P(w \mid c)$$

surprisal: the information in word w , given context c

$$i(w ; c) = L(w) - L(w \mid c)$$

information gain: what the context is worth

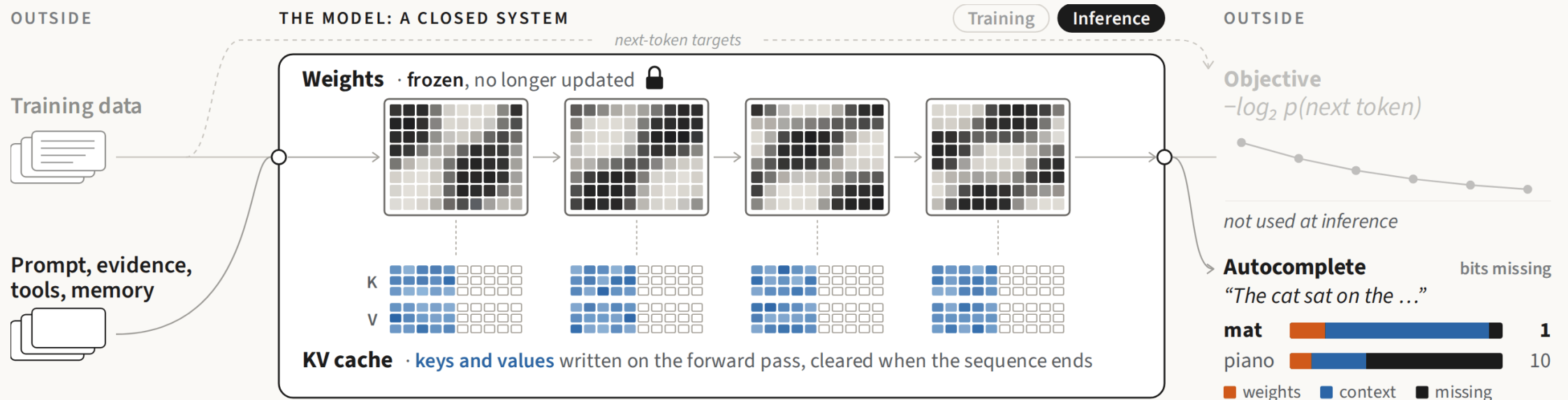
$$\text{KL}(p \parallel q) = \sum p \log_2 (p / q)$$

divergence: the extra bits wrong beliefs cost

Shannon (1951): English carries about a bit per letter; context makes the rest predictable. A model's training loss is its average surprise per token.

The frame

LLMs autocomplete from weights or context



Autocomplete finds "mat": 1 bit missing. The rarer "piano" still needs 10.

$$L(y \mid c) = L(y) - i(y ; c)$$

still missing = left open by the weights – supplied by the context

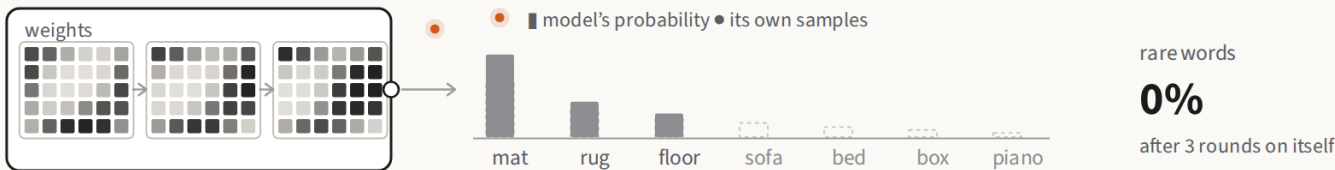
*Rare information needs more bits to store and extract.
After training, only $i(y ; c)$ can still change.*

Same quantities as slide 2, now for a whole answer y : surprisal L and information gain i .

Four ways a model can improve itself

Trains on its own outputs

Changes: **weights**
New bits: an outside reward. Otherwise none, and the tails collapse.



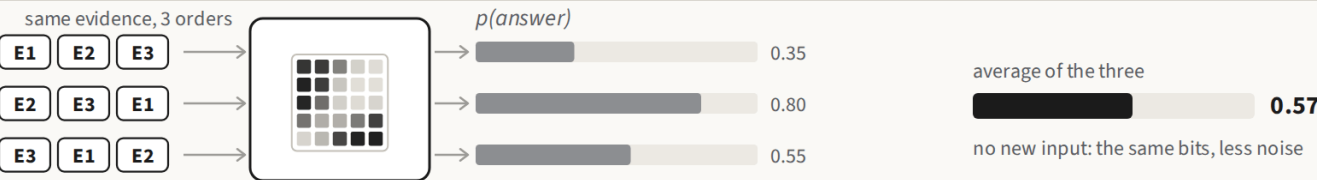
Rewrites its context or memory

Changes: **context (KV cache)**
New bits: the world, but only if it looks: tools, sources, people.



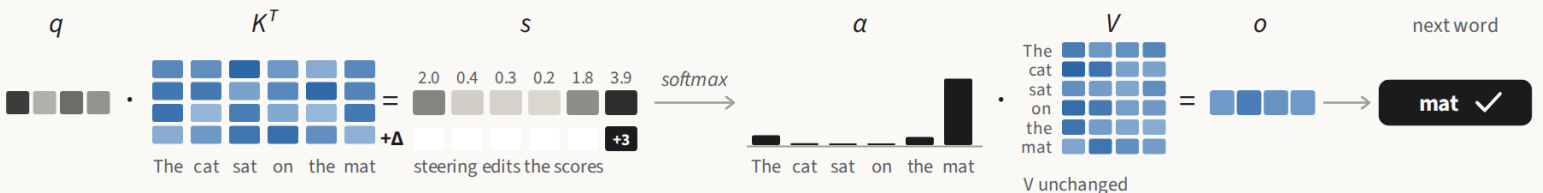
Searches, samples, votes, averages

Changes: **compute**
New bits: none; existing bits made usable.
Averaging orderings: 42.0% → 45.2%.



Steers how it reads

Changes: **attention**
New bits: none; it re-weights what it has.

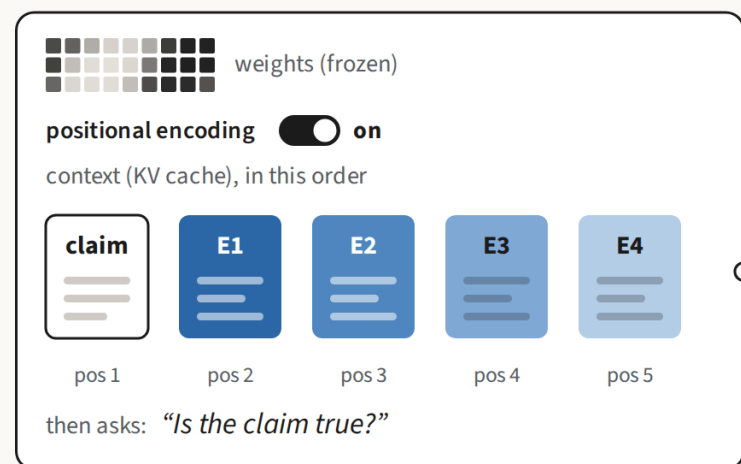


$I(W; \theta') \leq I(W; \theta, X)$ A closed loop learns nothing new about the world; computation only makes it usable.

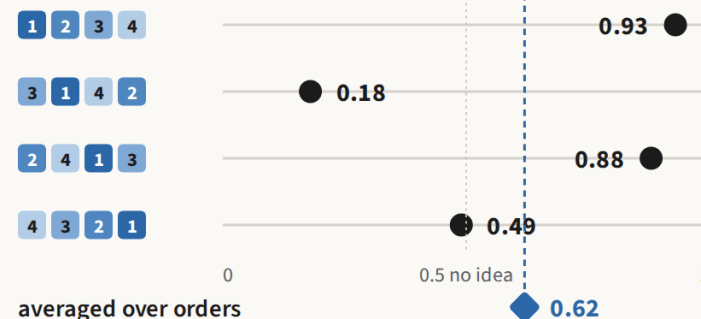
The problem: hallucination

Same evidence, new order, a different answer

THE MODEL



$p(\text{claim is true})$, asked in each order



True.

I'm 93% sure.

bits behind the answer

a 95%-sure answer needs

0.71

the evidence supplies (averaged)

0.31

A hallucination is confidence the evidence can't back. It's predictable, so count the bits first.

$$\sigma(n) \approx a + b \log n$$

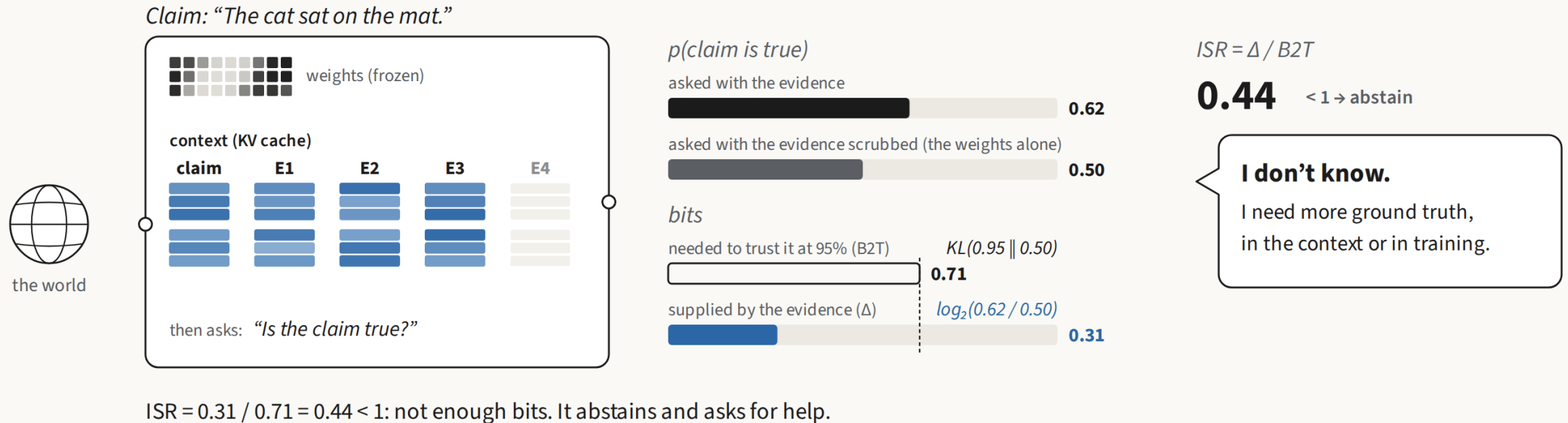
order sensitivity grows with n , the number of evidence chunks

A hallucination is confidence the evidence can't back.

It's predictable, so count the bits before answering.

Rewrites its context

When the evidence runs short, it asks for help



$$\Delta = L(y) - L(y \mid c) = i(y ; c)$$

the bits the evidence supplies: ask without it, ask with it, subtract

It can't create the missing bits.

It can measure the shortfall and ask for them.

How many outside bits does trust cost?

$$\text{B2T} = \text{KL}(\text{Ber}(p^\star) \parallel \text{Ber}(q))$$

Bits needed to move a belief q to reliability $p^\star$.

From 0.5 to 0.95 costs 0.71 bits; from 0.1 to 0.95, 2.9.

$$\text{ISR} = \bar{\Delta} / \text{B2T}$$

Bits the evidence supplies, over bits needed. Answer if $\text{ISR} \geq 1$; below 1, fetch bits from outside or abstain.

Held-out audit, 528 items

0.0–0.7%

hallucination with the gate at $\text{ISR} = 1$
while abstaining on 20.6–27.9% (95% CIs)

Hallucination is a budget shortfall: an answer given on too few bits.

No model can certify itself

Turing, 1936 No program decides, for every program, whether it halts.

Rice, 1953 No program decides any non-trivial property of what programs do.

Gödel, 1931 No consistent system rich enough for arithmetic proves it is consistent.

So self-checks are incomplete or unsound. Trusted verdicts come from outside: tests, compilers and proof checkers where the question is decidable; people where it isn't.

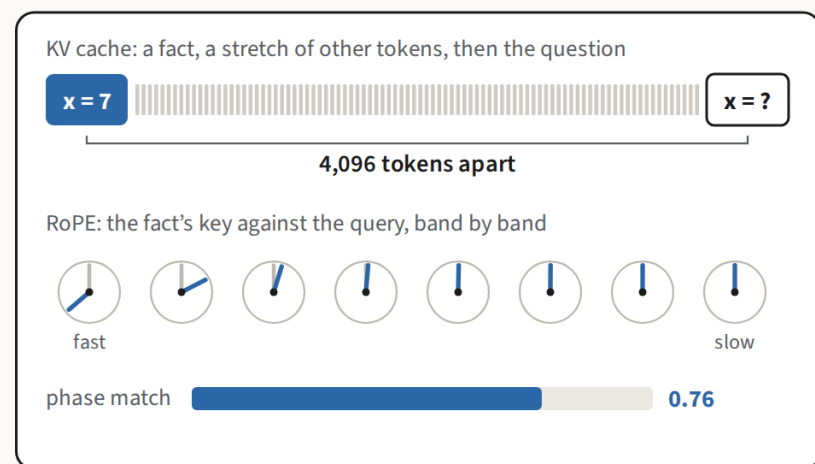
Our results agree. Same-model review missed the model's own bugs, and model-written evidence passed 4 of 4 false claims. With outside, isolated provers, false approvals fell from 67% to 1.8%.

Each outside yes/no carries at most 1 bit. Spend them where $ISR < 1$.

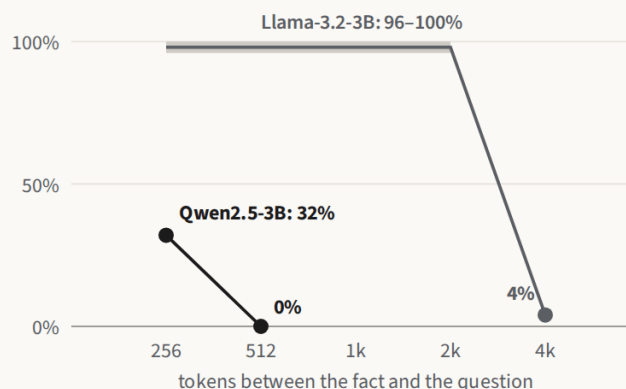
The problem: present but unused

It can repair its own KV cache using exact algebra

THE MODEL



accuracy vs distance



its answer

x = 7 ✓

give the far fact a near phase

Qwen2.5-3B, 2,048 tokens, 40 look-alike facts

as it reads now

[bar] **12.5%**

with the near phase

[bar] **77.5%**

Give the far fact the phase of a near one and accuracy goes from 12.5% to 77.5%. No new facts, no new weights.

$$s = q \cdot R(\Delta) k$$

RoPE turns each key by its position, so its score depends on Δ

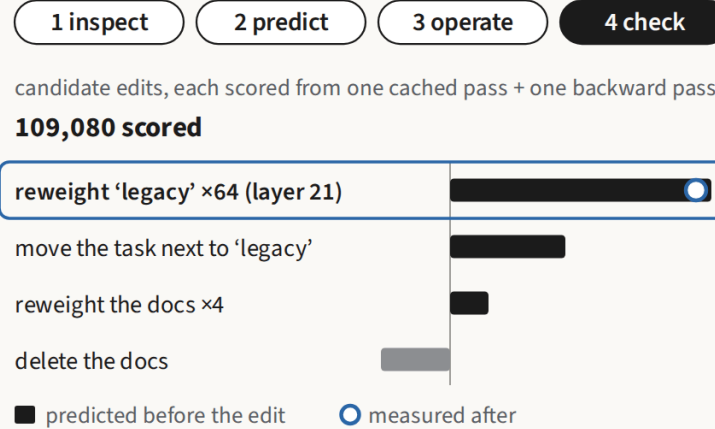
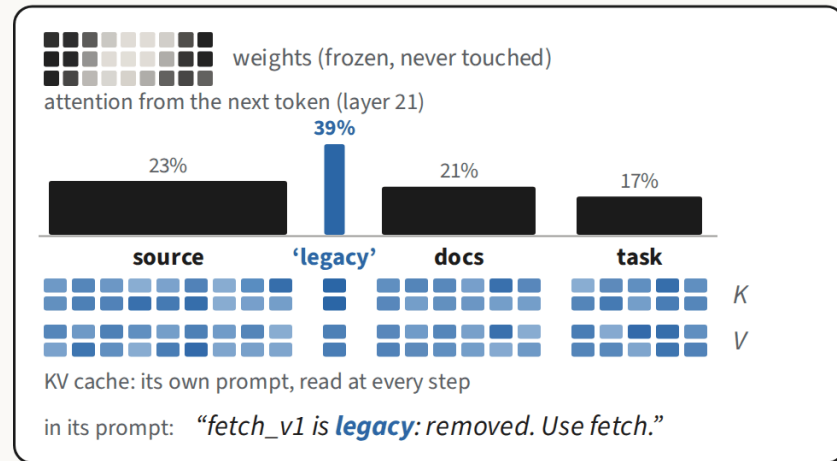
The fact is still in the cache.

Position decides whether it gets read.

Steers how it reads

Using RoPE as a derivative the model can experiment on itself

THE MODEL



tests passed

6 / 6



emphasising the whole source instead: 1 / 6

its output

fetch(url)

calls the current API

Self-surgery: it finds what it ignored and changes how it reads it. No retraining.

95–97%

of 92,160 position edits predicted in the right direction

The fix was already in its prompt:

no new bits, no new weights, just a new reading.

Summary

Three things to keep

- 1 Rare information needs more bits. After training, only the context can supply them.
 - 2 Loops that only talk to themselves re-use bits. New bits come from outside, and ISR says how many.
 - 3 No model can certify itself. Spend outside verdicts (tests, proofs, people) where the budget runs short.
-