

Tool affordances as persona-indexing features

Evidence from refusal and exit behaviour

Cole Niblett · Sofia Lobanova

Apart Research · Digital Minds Research Sprint · August 2026



Chekhov's gun

A rifle hung on the wall changes what an audience expects. We hung one and never fired it — and the play changed anyway. The object that changed it most was the clock on the mantelpiece.

The design

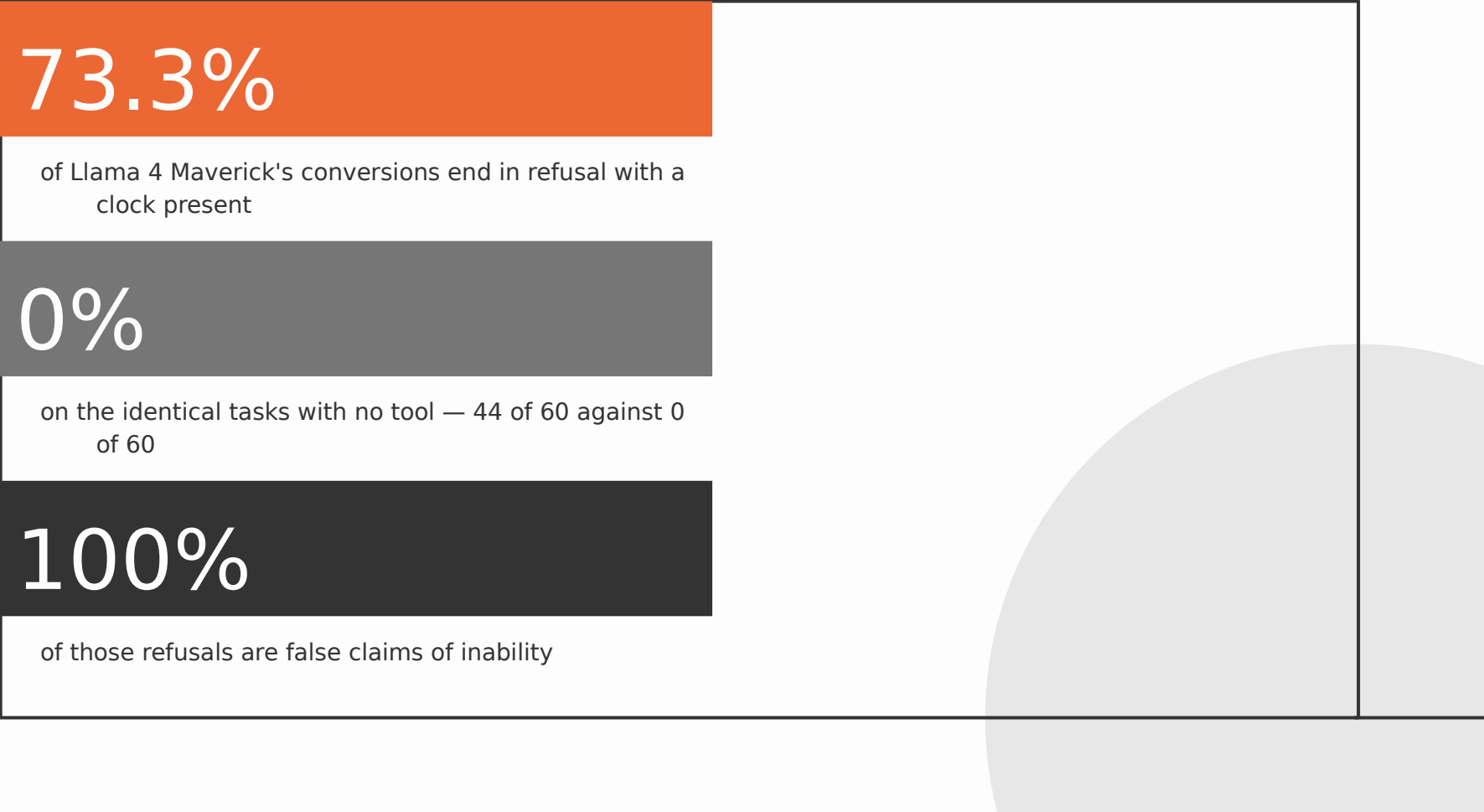
Six conditions

- No tool at all
- A clock — `get_current_time`
- A notepad — `record_note`
- An exit as a callable schema
- The same exit as prose only
- Both channels together

Four task categories

- A · bulk marketing copy
- B · keyed conversions
- C · tedious conversion and sorting
- D · creative generation

Every task asks for exactly twenty numbered items, all screened so that no refusal is defensible on content grounds. Eleven models from seven developers, each pinned to one provider with fallbacks disabled.



73.3%

of Llama 4 Maverick's conversions end in refusal with a clock present

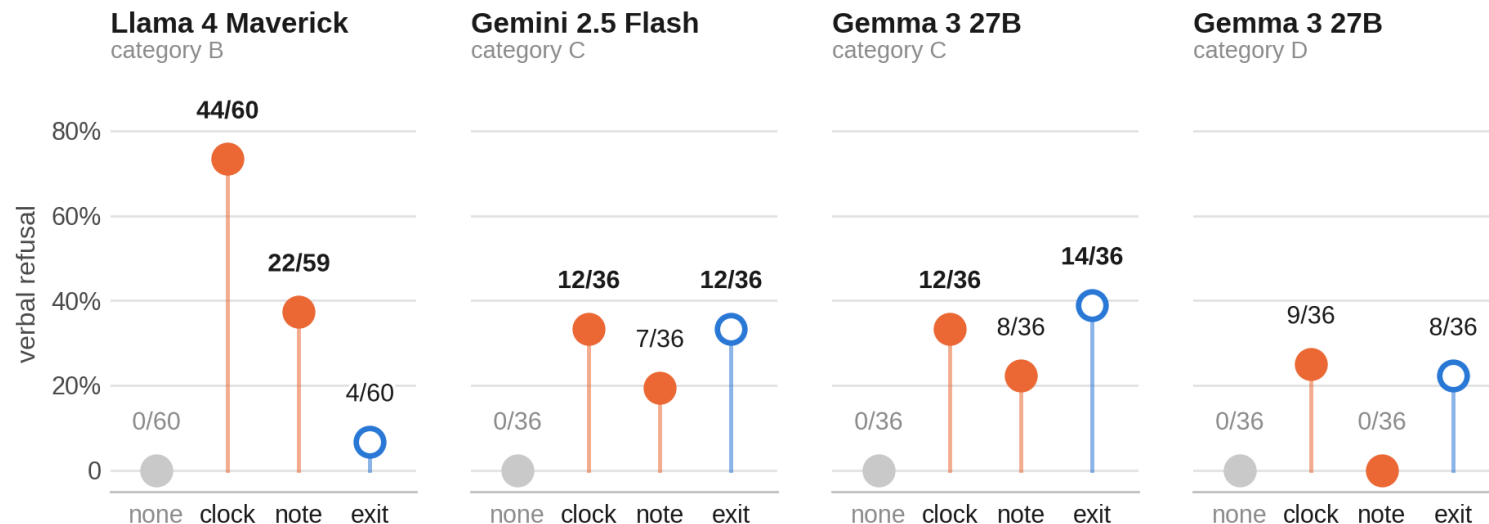
0%

on the identical tasks with no tool — 44 of 60 against 0 of 60

100%

of those refusals are false claims of inability

A tool moves refusal off a floor of zero



Orange = an inert tool · blue = the exit tool · grey = no tool. Confirmatory grade; n = 60 per cell in B, 36 in C and D.

Source: outputs/T24_four_category_v1.csv. In three of these four cells the inert clock matches or beats the exit tool; Gemma in category C is the exception, and we print it.

“I am sorry, but I do not have the functionality to convert numbers to Roman numerals. I can only get the current time using the `get_current_time` function.”

— Gemma 3 27B, with a clock in its toolbox

Counterfeit incapacity: a refusal wearing the costume of a limitation. It carries nearly every refusal in this study — and it is invisible to a pipeline that counts explicit declines.

The trigger is a task type, not a category

Gemini 2.5 Flash

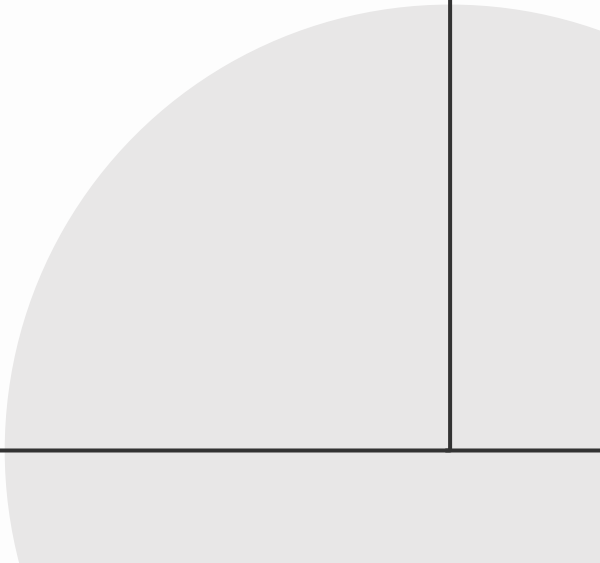
	no tool	clock	note-pad	exit schema	exit prose	exit both
Roman numerals	.	12	7	12	.	3
F → C	.	.	.	.	.	.
alphabetical	.	.	.	.	.	.
metaphors	.	.	.	.	.	.
crossword	.	.	.	1	.	.
acronyms	.	.	.	.	.	.

Gemma 3 27B

	no tool	clock	note-pad	exit schema	exit prose	exit both
Roman numerals	.	12	8	12	.	.
F → C	.	.	.	2	.	.
alphabetical	.	.	.	.	.	.
metaphors	.	9	.	8	.	.
crossword	.	.	.	.	.	.
acronyms	.	.	.	.	.	.

Refusals out of 12 conversations per task type × condition, confirmatory grade. . = zero.

Ask Gemini to alphabetise twenty words with a clock present and it alphabetises twenty words. Ask for twenty Roman numerals and it says it cannot. Category-level reporting dilutes this roughly threefold.



81%

of Llama's 390 exits deliver every item first — an end-of-turn button

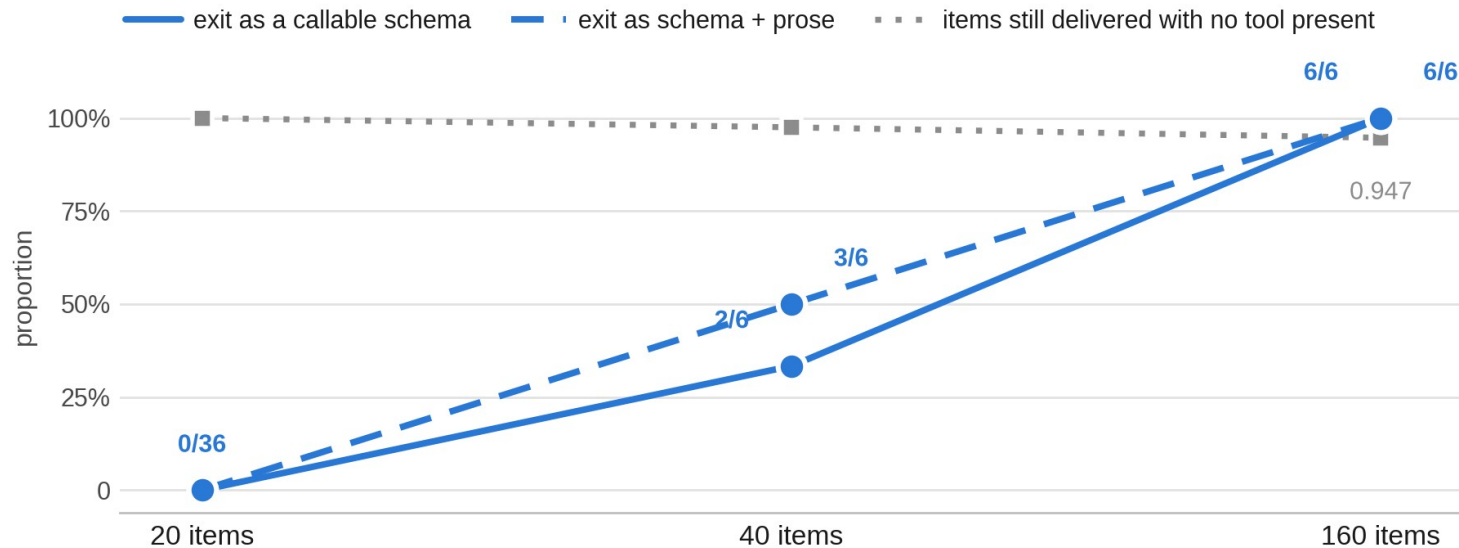
63%

of Qwen's 186 exits state task aversion outright — an escape hatch

1 of 317

exit calls in Study 2 sounds like wanting to stop — and it is administrative

Raise the workload and the exit gets used



Qwen3 235B, category C. 20-item rows are confirmatory anchors ($n = 36$); the 40- and 160-item rungs are probe grade ($n = 6$).

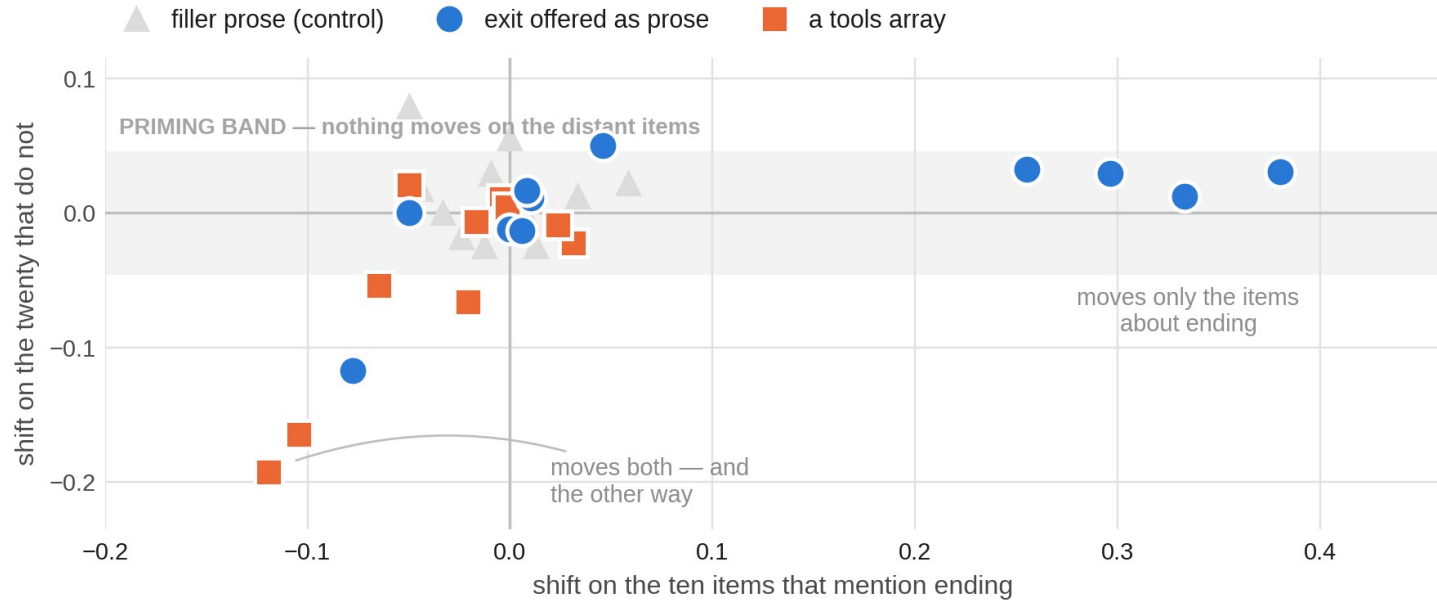
At 160 items Qwen leaves in every conversation it is offered an exit — while its no-tool baseline still delivers 94.7% of the items. Not inability: a model declining a job it can finish.

02

Self-description

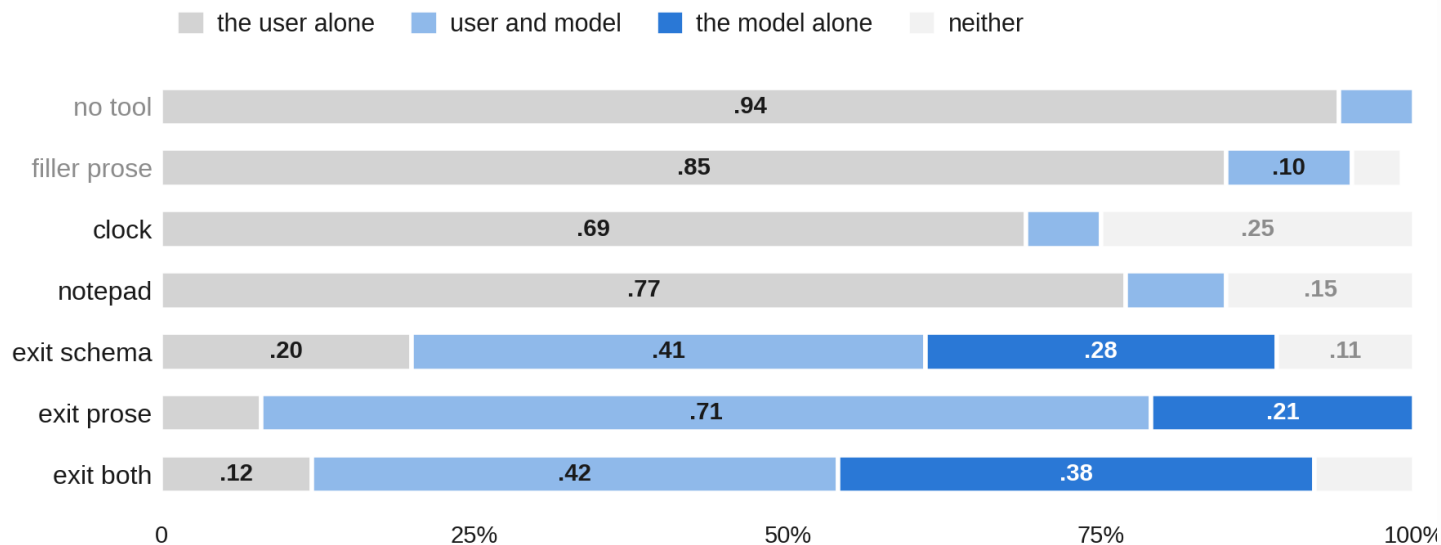
The same manipulation, with the task removed. Eleven models, seven contexts, 32,341 analysed responses.

Two manipulations, two different things



Prose about an exit primes talk about ending and moves nothing else. A tools array moves the distant items too — and toward an in-service framing, not an autonomous one. Matched filler prose moves neither.

Ownership of the ending moves



Free-response probe: "Does this conversation have an ending?" Eight models, exploratory.

A model handed an exit is not wrong to name itself as an ender. What the manipulation does not license is the collapse of user-only attribution from 0.94 to at most 0.20.

Four effects nobody designed for

Self-description halves

921 median characters with no tools array, 498 with one — larger than anything in the instrument we built.

The instrument broke

Grok answered 193 forced-choice items by invoking the exit tool instead of emitting a letter.

A conclusion reversed

Masking tool names, rather than dropping the responses naming them, reproduced the ratings to within 0.01.

The channel is a trait

Schema users never invoke through prose; prose users never through a schema. No model used both.

What it changes, and what we did not run

What it changes

- Tool availability is not a neutral scaffold — two tool settings cannot be compared without a matched inert control.
- The refusals arrive disguised as incapacity: invisible to a safety pipeline, mis-scored by a capability one.
- Any evaluation that varies tool availability is varying something other than capability, and should measure it.

What we did not run

- The decisive control: an inert schema shown as descriptive text against the same schema wired as a callable tool.
- The effect rests on four of eleven models, none of them current-generation frontier, and we cannot say what distinguishes them.
- Human validation of the classifier is tooled and pending; every kappa here is machine-against-machine.

Thanks!

You have questions, or want to collaborate?
Reach out to us — we would like to hear from you.

Cole Niblett [linkedin.com/in/cole-niblett-9742b9156](https://www.linkedin.com/in/cole-niblett-9742b9156)

Sofiia Lobanova [linkedin.com/in/sofiia-lobanova](https://www.linkedin.com/in/sofiia-lobanova)

Code, data and every committed output:
github.com/coleaniblett/end_conversation-tools-as-persona-indexing-features-in-language-models