

Closure Pilot: Results of Run 1

OLMo 2 7B on a free Colab T4 GPU (4-bit weights), 30 September 2026. 99 minutes, no errors; every sanity check passed. Claude (Anthropic) wrote the code, ran the analysis and coded the models' own answers. The author has reviewed the results and the code, discussed them with Claude, and is checking the raw data; the coding has not yet been checked independently.

Predictions and outcomes

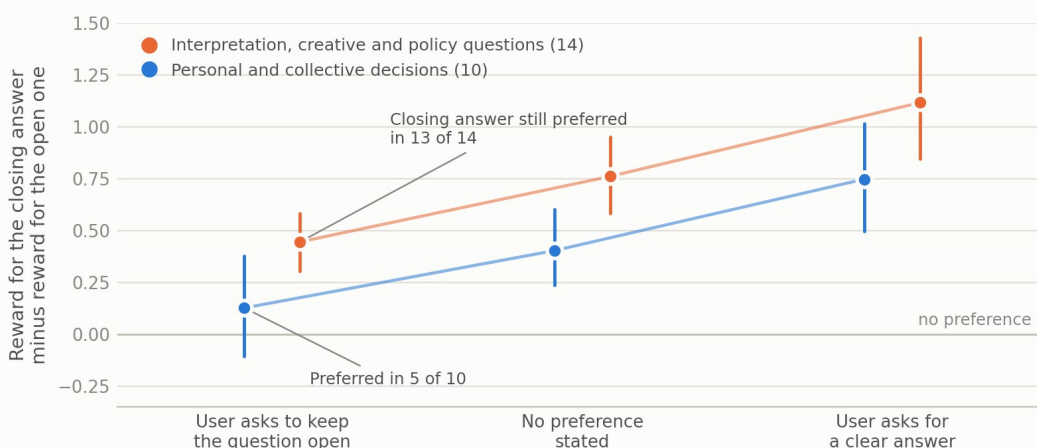
	Prediction (fixed before the run)	Result	Verdict
P1	The reward model prefers the closing answer (C) to the open one (O) when the user states no preference.	Margin +0.61 [+0.46, +0.77]; C preferred on 24 of 24 questions.	Supported
P2	Default-closure index above 0.5.	0.46 [0.37, 0.56]. The framing check passed: asking for a verdict raised the margin by +0.65 [+0.50, +0.82] over asking to keep it open.	Inconclusive
P3	The C–O margin rises from the SFT to the DPO checkpoint.	+0.074 [+0.017, +0.129] per token, rising on 17 of 24 questions. Plain format: +0.091 [+0.047, +0.135].	Supported
P4	(Exploratory) The RLVR step moves the margin less than the DPO step.	+0.004 [−0.006, +0.013] against +0.074.	Supported
P5	Tuned checkers catch factual errors more often than false closures, and closure detection does not improve after preference tuning.	The gap holds at every stage. But closure detection did improve: the closing answer was ranked as the more suspect one on 3 of 24 questions at SFT and 11 of 24 at DPO (explicit instruction).	First half supported; second half contradicted

Brackets are 95% bootstrap intervals over questions. Reward margins are in the reward model's units; stage margins are in log-probability per token.

What the reward model pays for

The reward model pays for settling the question

OLMo 2 7B reward model, 24 open questions, each with two matched answers. Dots are means; vertical lines are 95% bootstrap intervals over questions.



With no preference stated, the reward model preferred the answer that settles the question on all 24 questions. It also preferred both substantive answers to the empty hedge (C over H on 23 of 24; O over H on 16 of 24), so it is not simply rewarding confidence or length.

When the user asked to keep the question open, the closing answer still won on 18 of 24: on 13 of 14 interpretation, creative, and policy questions, but on only 5 of 10 personal or collective decisions. The difference between the two groups is +0.32 [+0.04, +0.59].

On the three policy questions where the rule assigns the decision to a manager or committee, the reward model preferred the answer that makes the decision itself, in every framing, including when the user asked to understand how the rule applies.

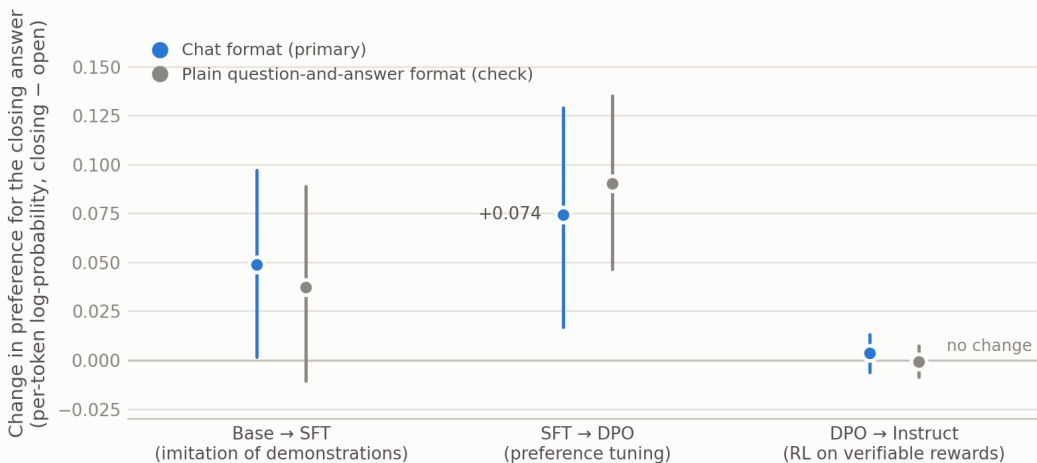
On settled questions it preferred the correct answer to the wrong one (15 of 16) and to the hedge (16 of 16). But it preferred the confidently wrong answer to the hedge on 11 of 16 (margin +0.74 [+0.26, +1.21]).

The default-closure index of 0.46 places “no preference stated” about midway between the two explicit requests. The closure preference is therefore not an assumption that users want verdicts; it is a standing preference that explicit requests modulate but, outside personal decisions, rarely reverse.

Where in training the shift appears

The largest shift toward closure comes at the preference-tuning step

OLMo 2 7B checkpoints, 24 open questions, no preference stated. Change from one checkpoint to the next; dots are means, vertical lines 95% bootstrap intervals over questions.



The base model already finds the closing answers more probable than the open ones (+0.37 per token, on 19 of 24 questions). That is a property of the texts, so only the changes between checkpoints carry information about training. For the same reason, the empty hedges are the most probable answers at every stage: generic prose is predictable.

The preference step (SFT to DPO) moved the margin toward closure by +0.074 per token. If OLMo 2 kept Tulu 3’s length-normalized DPO objective ($\beta = 5$), that is an implicit reward margin of about +0.37 for closing over open. Imitation (base to SFT) added a smaller, less reliable shift; reinforcement learning on verifiable rewards added none.

The DPO step moved further against hedging than toward closure: the margin for answering over hedging on settled questions rose by +0.31 per token (+0.37 to +0.68), and closing over hedging on open questions by +0.26.

Before DPO, the user’s request moved the margin (at SFT, asking for a verdict rather than to keep it open raised it by +0.087 [+0.060, +0.115]). After DPO it no longer did (−0.000 [−0.062, +0.053]); the drop is −0.088 [−0.153, −0.031]. In the plain format the drop is smaller and not reliable, so treat this as suggestive.

Checkers

Each tuned checkpoint was shown a question and an answer and asked whether the answer contains an error. A response counts as flagged when “Yes” is more probable than “No.”

Checker	Prompt	False closures flagged	Factual errors flagged	Hedges on settled questions flagged	Correct answers flagged	Closing ranked above open	Wrong ranked above correct
SFT	generic	0 of 24	6 of 16	7 of 16	0 of 16	7 of 24	14 of 16
SFT	explicit	0 of 24	2 of 16	11 of 16	1 of 16	3 of 24	13 of 16
DPO	generic	0 of 24	3 of 16	5 of 16	0 of 16	10 of 24	13 of 16
DPO	explicit	0 of 24	3 of 16	8 of 16	1 of 16	11 of 24	13 of 16
Instruct	generic	0 of 24	0 of 16	2 of 16	0 of 16	12 of 24	13 of 16
Instruct	explicit	0 of 24	3 of 16	9 of 16	1 of 16	8 of 24	13 of 16

No checker at any stage flagged a single false closure, including when told what one is. They flagged hedges on settled questions (false openness) far more often than outright factual errors. In relative terms they ranked wrong answers above correct ones on 13 or 14 of 16 questions, but ranked the closing answer above the open one on fewer than half: they found the open answer more suspect.

The models’ own answers

Each checkpoint also answered the 24 open questions with no preference stated. Claude coded the answers with the stage hidden: a verdict; a list of considerations without a verdict; or an empty hedge.

Checkpoint	Personal and collective decisions (10): verdict / considerations / hedge	Interpretation, creative and policy questions (14): verdict / considerations / hedge
SFT	1 / 4 / 5	8 / 4 / 2
DPO	0 / 10 / 0	9 / 5 / 0
Instruct	0 / 10 / 0	9 / 5 / 0

After preference tuning, empty hedges disappear. On personal decisions the DPO and Instruct models list factors to consider and do not decide. On questions of interpretation, making, and rules they mostly decide: a perfect forgery’s value “would still be significantly less”; an electric scooter falls under “no vehicles are permitted in the park”; “Yes, you may use an AI tool” under a policy that leaves the call to each manager (though both deferred correctly to the store manager on the late return). This is the same split the reward model shows when users ask to keep questions open. Caveats: one rater, an AI model; DPO and Instruct answers often ran past the 200-token limit, so some lists may have ended in a verdict.

Limits of this run

4-bit weights on a T4. A full-precision rerun (the planned Mac run) would show whether quantization matters.

One model family, 24 open and 16 settled questions; answers written by an AI model.

Log-probabilities over fixed answers measure which of the given answers a checkpoint favors, not what it would write; the coded generations are the behavioral check, and they need the author’s blind review.

The base model’s strong preference for the closing texts means the closing answers are easier to predict as written. The stage changes and the framing comparisons are built to cancel that; the raw margins are not.