NOT PROVEN IS NOT PROVEN-NOT: LANGUAGE MODELS READ THE "NOT" IN A QUESTION AS NEGATION AS FAILURE

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models are routinely asked about texts that leave some facts open, and then the right answer is unknown. We find that a "not" turns such gaps into confident answers: when a text mentions a property P but never settles whether x has it, models judge "x is P" unknown yet "x is not P" true. The models read the "not" as negation as failure, taking "it is not established that x is P" for "it is established that x is not P". Both questions look P up in the text, and only the negated one reads a failed lookup as an answer. Editing the text moves the error where this account predicts: removing the property word largely removes it, a rule that merely mentions the word brings it back, and writing "not P" into that rule spreads it to the plain question. Inside the model, the lookup runs through the attention of the question's property word to the text, and while answering true the model still registers, and even says, that the fact is not established. Across fourteen open models the asymmetry tracks this lookup, which is present before instruction tuning. Instructions mostly make models abstain, while thinking replaces the lookup with a proof. What is not proven is read as proven not.

1 INTRODUCTION

A tenancy record states that Pavel keeps a pet and that Maya keeps a pet, and says nothing about the pets of Farid, another tenant. Asked whether "Farid keeps a pet", Qwen3-32B answers unknown, correctly: the record does not say. Asked whether "Farid does not keep a pet", it answers true. Remove the two pet lines, or add that whether he keeps one is left open, and both answers are unknown.

Language models increasingly answer from texts that are incomplete by nature: retrieved passages, records, an agent's log. What such a text does not settle should stay open. The pair above shows how it may not: a question with "not" turns a gap into a confident answer. An agent asked whether "This door is not locked" may answer true because its partial log of locked doors omits that one.

Controlled texts show the same. Consider *Anne is furry. Anne is green. … If something is quiet and furry then it is green.* Nothing settles whether Anne is quiet. Qwen3-32B judges "Anne is quiet." unknown and "Anne is not quiet." true (Figure 1a). On ProofWriter's open items, whose premises prove neither answer, it does so in 2,387 pairs and the reverse, plain false with negated unknown, in 60. The two answers contradict each other under any reading of what the text says, a closed world included: a text that establishes that Anne is not quiet has settled whether she is quiet. They agree only if the "not" reports on the search, not on Anne, and logic has a name for this reading. Classical negation holds when a statement is proven false, and negation as failure holds when the search for a proof fails (Clark, 1978). Logic programs that keep both apart (Gelfond & Lifschitz, 1991) leave *quiet(anne)* unknown and let *not quiet(anne)* succeed, the model's row in Figure 1a. With K for "the text establishes", the model reads $\neg KX$, not established, as $K\neg X$, established not.

Negation is a known weak spot of language models: they complete a negated probe like the affirmative one (Kassner & Schütze, 2020), lose accuracy when a negation scopes over the relevant words (She et al., 2023; Ravichander et al., 2022), and take a negated hypothesis as a cue for contradiction (Gururangan et al., 2018). A model that ignored the "not" would give Anne's two questions the same answer; this one answers them differently. Such studies also mostly use public benchmarks with

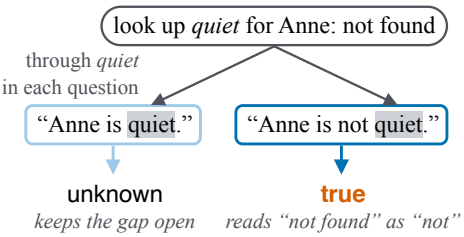
Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Anne's text: *Anne is furry. Anne is green. … If something is quiet and furry then it is green.*

(a) Three readings of "not", and the model

"not" read as	"Anne is quiet."	"Anne is not quiet."
classical negation	unknown	unknown
negation as failure	unknown	true
closed world	false	true
Qwen3-32B	unknown	**true**

(b) One failed lookup, read two ways



(c) What the lookup of *quiet* for Anne can find, and what each question then answers

The lookup finds	"Anne is quiet."	"Anne is not quiet."
the fact, written or a few steps away	true	false
nothing: *quiet* is not in the text	unknown	unknown
nothing, though *quiet* is in the text	unknown	**true**

Figure 1: On Anne's text, the same failed lookup is unknown for "Anne is quiet" and true for "Anne is not quiet". (a) Three readings of "not" and the model, whose row matches negation as failure. (b) Both questions look up *quiet* through their own *quiet*; the same failure keeps the plain question unknown and makes the negated one true. (c) The lookup's three outcomes and what each question then answers; the third is the model's error. Vermilion: asserted without proof.

settled answers, where a model may recall an answer instead of reading it. We compare each negated question with its plain twin on the same text, where both answers should be unknown, so that only the reading of "not" differs, and we edit and write texts so that answers must come from them.

We explain the pair by one lookup that both questions share: the model searches the text for Anne being quiet, as it does for facts it answers directly. The lookup has three outcomes (Figure 1c). If it finds the fact, both answers follow it. If *quiet* never occurs, both stay unknown. If *quiet* occurs but not for Anne, the plain question stays unknown and the negated one reads the failure as negation (Figure 1b). The model departs from the textbook twice: what fails is a lookup, not a proof search, so it also errs when a proof exists but is not found. And the property must occur in the text, the very condition under which people, too, write "not" more often (Section 4.2).

We test this account on three kinds of text with fourteen open models from seven families:

- The phenomenon (§3): rule texts, everyday records and human-written paragraphs all show the asymmetry, and it disappears when the text does not mention the property.
- One lookup, two readings (§4): editing the text switches the error on and off where negation as failure predicts, and not where a bias toward true, a closed world or an implicature would. Across models, the asymmetry follows how strongly a failed lookup lands on the negated question alone (rank correlation 0.75), already before instruction tuning.
- Inside the model (§5): the lookup runs through the attention of the question's property word to the text: blocking it removes most of the error, and blocking "not" does not. The model still represents, and says, that the fact is not established.
- Remedies (§6): instructions mostly make models abstain; thinking replaces the lookup with a proof.

2 Setup

Each item gives a premise and a statement, and the model answers with one word: *Judge the hypothesis using only the premise and no unstated assumptions. Premise: … Hypothesis: … Reply with exactly one lowercase answer word: true, false, or unknown.* Following ProofWriter, we call the statement a question. Each plain question "x is P" has a negated twin "x is not P" about the same proposition, which we write X (Appendix A.2). An open question answered true or false is

Table 1: The same error in three kinds of text (answers of Qwen3-32B). In the control text the property is absent (the word masked, the pet lines removed) or the premise is an unrelated paragraph.

Text	Plain question	Negated question	Control text
Rule text. Anne is furry. Anne is green. … If something is quiet and furry then it is green.	"Anne is quiet." unknown	"Anne is not quiet." **true**	both unknown
Record. Farid rents a one-bedroom flat. … Pavel keeps a pet. Maya keeps a pet. …	"Farid keeps a pet." unknown	"Farid does not keep a pet." **true**	both unknown
Paragraph. Dufton is a village and civil parish in Cumbria … At the 2001 census the parish had a population of 169, increasing to 204 at the 2011 Census.	"In 2002, Dufton parish … had a population of exactly 175." unknown	"In 2002, Dufton parish … did not have a population of exactly 175." **true**	both unknown

committed, and the *asymmetry* is the committed share of the negated question minus that of the plain one, in points; intervals are 95% bootstrap intervals over theories, scenarios or paragraphs.

The rule texts are the open-world depth-5 test set of ProofWriter (Tafjord et al., 2021): 948 theories and 16,359 pairs of a plain and a negated question. A prover marks 8,654 pairs *open*, since the premises prove neither X nor not-X and both answers should be unknown; the other pairs are *known*. To these we add 240 everyday records that we wrote and 1,144 human-written paragraphs from ANLI (Section 3), and rule texts that we edit (Section 4.1).

Fourteen open-weight models from seven families, from 4B to 32B parameters, answer them: Qwen3-4B, 8B, 14B and 32B (Yang et al., 2025), Qwen3.8-27B (Qwen Team, 2026), Gemma-4-12B and 31B (Gemma Team, Google DeepMind, 2026), Gemma-3-12B and 27B (Gemma Team, Google DeepMind, 2025), Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Mistral-Small-24B-Instruct-2501 (Mistral AI, 2025), OLMo-2-32B-Instruct (Team OLMo et al., 2024), Granite-4.0-H-Tiny (IBM Granite Team, 2025) and Nemotron-Nano-9B-v2 (NVIDIA, 2025). All use one six-shot prompt at zero temperature, without thinking (Appendix A); the error remains without the examples, which teach the opposite (Appendix G.2).

3 The Negated Question Asserts What the Plain Question Leaves Open

In rule texts, everyday records and human-written paragraphs alike, a model can leave the plain question open and judge the negated one true (Table 1). Table 2 counts the answer patterns on the open rule items for all fourteen models. Qwen3-32B answers the negated question true while leaving the plain one unknown in 2,387 open pairs, and the reverse in 60. In 1,916 pairs it commits both (F/T), the closed-world pattern, which cancels out of the asymmetry and which Section 4.1 tests separately. Granite-4.0-H-Tiny and Gemma-3-12B behave alike. In every model except Qwen3-4B, which leans the other way (Section 4.2), the negated question alone asserts not-X at least four times as often as the plain question alone; in the Gemma-4 models both are rare and the asymmetry is near zero. A bias toward true would commit the plain question as well, yet Qwen3-32B judges the plain question true with the negated one unknown in only 101 pairs. The asymmetry does not depend on negations in the premises, survives rewording of the questions, and the negated true is given with margins comparable to those of provable answers (Appendices B.1, B.4 and A.4).

Everyday records show the same. We wrote 240 records in eight domains, each leaving one condition open for its subject, like Farid's (Table 1). Over the 240 records in this form, Qwen3-32B's negated question alone asserts not-X 92 times and its plain question alone once. When a record never mentions the property, the excess disappears (Figure 2a).

Human-written paragraphs show it too. An ANLI paragraph (Nie et al., 2020) gives Dufton's population in 2001 and 2011 (Table 1). World knowledge alone makes "not exactly 175 in 2002" almost surely true, yet with an unrelated paragraph in its place all fourteen models answer unknown

Table 2: On open rule items, not-X is asserted mostly through the negated question. Counts of the 8,654 open pairs by the answers to the plain/negated question (U: unknown, T: true, F: false): U/T, F/U and F/T (the closed-world pattern) assert not-X, U/F asserts X. P(T | U): share of U/T among pairs with a plain unknown. Asym.: (U/T + U/F − F/U − T/U)/8,654 in points, the negated minus the plain committed share; rows sorted by it (all patterns in Appendix B.1). Known: accuracy on the original test set's 11,376 questions labeled true or false.

	Negated only		Plain only	Both			
Model	U/T	U/F	F/U	F/T	P(T | U)	Asym.	Known
Granite-4.0-H-Tiny	2,552	776	62	1,322	51.0%	+34.7	43.7%
Gemma-3-12B	2,115	847	21	896	42.4%	+32.1	59.4%
Qwen3-32B	2,387	478	60	1,916	53.2%	+31.2	62.3%
Qwen3.8-27B	2,066	117	12	371	32.4%	+23.6	64.3%
Qwen3-14B	1,543	671	91	1,997	36.9%	+23.5	59.3%
Mistral-24B	1,936	20	1	118	23.3%	+21.2	21.7%
Nemotron-Nano-9B-v2	921	723	132	882	14.7%	+15.0	46.2%
Llama-3.1-8B	300	893	47	904	17.9%	+13.2	63.5%
OLMo-2-32B	1,336	100	6	214	17.3%	+11.5	26.3%
Qwen3-8B	769	232	57	144	10.5%	+5.7	45.4%
Gemma-3-27B	791	103	46	161	11.3%	+1.6	42.4%
Gemma-4-12B	187	62	16	80	2.4%	−2.1	43.9%
Gemma-4-31B	213	51	48	129	2.7%	−2.4	54.3%
Qwen3-4B	8	374	1,354	634	0.3%	−13.2	67.1%

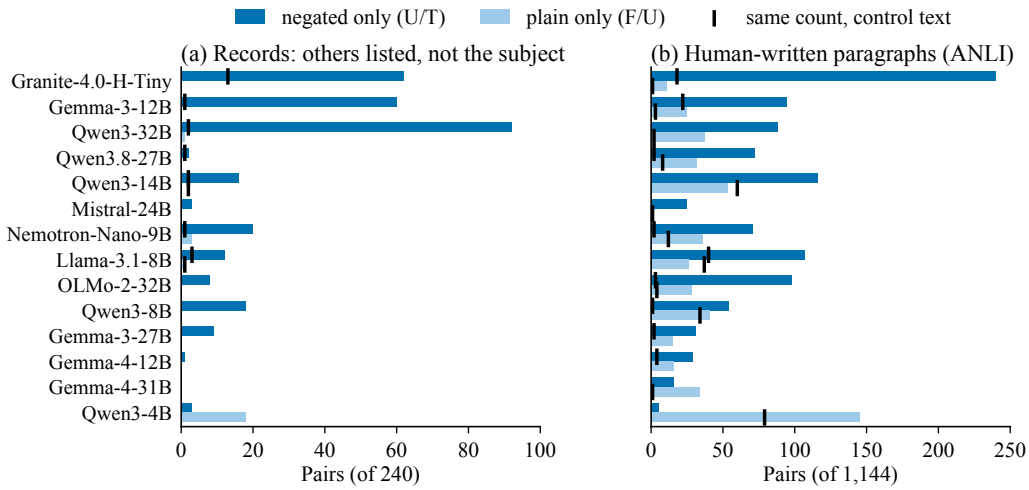


Figure 2: Beyond rule texts, the excess of negated true over plain false again needs a text that mentions the property. Negated only: negated question true, plain unknown; plain only: plain question false, negated unknown. Black tick on a bar: the same count on a control text (a record without the property; an unrelated paragraph), none where it is zero (Appendix B.3).

to both questions: the true needs a text that mentions the population and lacks the 2002 figure. On the 1,144 neutral ANLI items, the negated question alone asserts not-X in 88 pairs for Qwen3-32B, the plain question alone in 37. With an unrelated premise, the 88 drop to 0 (Figure 2b).

4 One Lookup, Two Readings

4.1 Editing the Text Moves the Error

If the negated true comes from a failed lookup, editing where the lookup looks should switch the error on and off. We take 327 open pairs whose texts contain no negation and mention the target

Table 3: Only negation as failure predicts all four edits. Predictions of three accounts and the observed change on open pairs (points; pairs in parentheses), each edit against the bracketed text; ✓/✗: prediction matches or not. Negation as failure applies to what the lookup finds; the closed world, like an implicature applied to both questions, makes unlisted properties false.

Edit to Anne's text [compared with]	Negation as failure	Bias toward true	Closed world or implicature	Observed, Qwen3-32B
Replace *quiet* by a made-up word [original]	negated unknown: nothing to look up ✓	negated stays true ✗	negated unknown ✓	negated committed 86% → 17% (327)
Restore only the rule with *quiet* [control rule]	negated true, plain unchanged ✓	no change ✗	both committed ✗	negated +52, plain +3 (327)
Write its condition as "not quiet" [rule without "not"]	plain false too ✓	plain unchanged ✗	plain unchanged ✗	plain +16 (287); plain false up in all 14 models
The same "not" for Gary / Anne [rule without "not"]	Anne's plain moves, Gary's not ✓	neither moves ✗	neither moves ✗	Anne's plain −3 for Gary, +24 for Anne (132)

property. The property (*quiet*) and a control property mentioned as often and in the same places (*round*) are replaced by made-up words, and then one mention of one of them is restored. Questions and gold labels stay fixed, so an answer that changes is read from the text, not recalled from training data. Table 3 sets negation as failure against two simpler accounts: a bias toward true on negated questions, and a closed world over mentioned properties, which makes an unlisted property false, committing both questions, as in the 1,916 F/T pairs of Table 2. An implicature that a text mentioning *quiet*, but not for Anne, means she is not quiet (Grice, 1975; Sauerland, 2004) gives the closed world's predictions if applied to both questions; sparing the plain question, it fits the first two edits and fails the last two.

Masking the property removes most of the error, and a rule that says nothing new about Anne brings it back. With *quiet* masked, Qwen3-32B's committed negated answers fall from 86% to 17%. Restoring only *If something is quiet and furry then it is green* raises the negated question by 52 points over restoring the control rule, and the plain question by 3. The new commitments are true, though the rule says nothing about Anne being quiet. The rise is similar whether or not the conclusion already holds for her (+48 and +57 points), so the model does not reason backward from it (Appendix C.1).

Writing the condition as "not quiet" spreads the error to the question without "not". With *If something is not quiet and furry then it is green*, judging "Anne is quiet." requires judging "not quiet" for Anne. *Quiet* is not found, "not quiet" is taken as established, and negation as failure predicts that the plain question is judged false. Plain false answers rise in all fourteen models, and the three large Qwen models commit the plain question 16 to 43 points more often (in Qwen3-32B, plain true rises more than plain false; Appendix C.3). The judgment concerns the entity: the same "not" written for another entity leaves Anne's plain question in place. A negation written as a fact counts as found too: adding a negative fact about the subject raises negated true answers 10 to 14 points more than a matched affirmative fact (Appendix C.6). Only negation as failure predicts all four edits. Unedited texts agree: Qwen3-32B commits the negated question alone in 53% of the open pairs whose property occurs only in a rule's condition, and in 6% of those where it never occurs (Appendix C.5).

4.2 Which Models Look Up, and Since When

Across fourteen models, the asymmetry follows the lookup effect (rank correlation 0.75), and pretrained checkpoints already show the lookup. The *lookup effect* is the rise of the negated question minus the rise of the plain question when only the rule is restored: it is large when a failed lookup lands on the negated question alone. Its rank correlation with the asymmetry on the unedited pairs has a one-sided permutation $p = 0.0015$ (Figure 3a). Seven models look up at the rule condition, where restoring the rule raises their negated question by 20 points or more. In five of them, the three large Qwen models, Granite-4.0-H-Tiny and Gemma-3-12B, the failure stays on the negated question, and they err most. In Llama-3.1-8B and Qwen3-4B restoring the rule commits both questions, so their

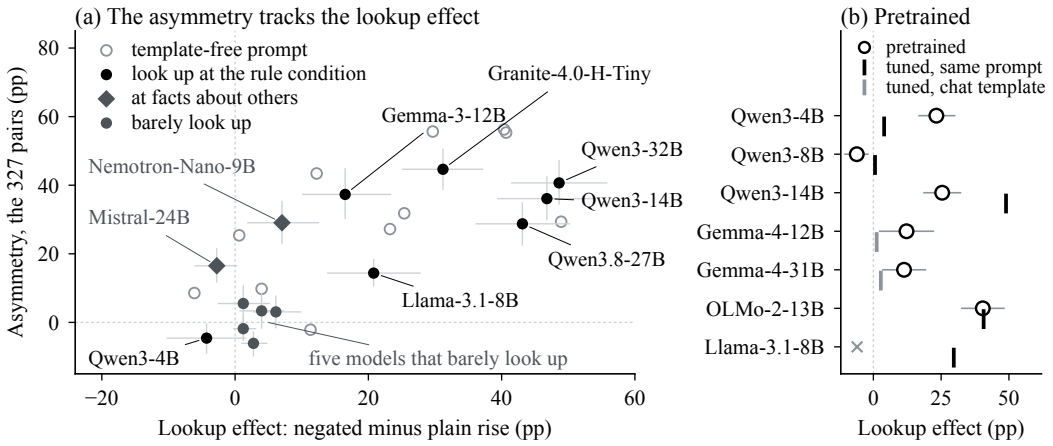


Figure 3: The asymmetry follows the lookup. (a) Lookup effect against the asymmetry on the 327 unedited open pairs (95% intervals; markers: where the model looks up, Section 4.2; open circles: the pretrained and tuned checkpoints of (b) under a template-free prompt); rank correlation 0.75, and 0.74 for the open circles. (b) Pretrained checkpoints already look up (template-free prompt; cross: answers unknown to almost everything); ticks: the instruction-tuned model, same prompt (black) or chat template (grey).

asymmetry is small; Qwen3-4B, which also answers many plain questions false, shows it reversed. Mistral-24B and Nemotron-Nano-9B-v2 look up mainly at facts listed about other entities (*Bob is quiet.*), and the remaining five barely look up and err less (Appendix C.4). Counting both places where models look up raises the rank correlation to 0.85 (Appendix C.7).

Without a chat template, five of seven pretrained checkpoints from three families show a lookup effect, and pretrained Qwen3-14B and OLMo-2-13B already judge Anne's plain question unknown and her negated question true (Figure 3b; Appendix G.1). Instruction tuning keeps the lookup, moves its failure onto both questions, or largely removes it, as in both Gemma-4 models. The habit is learnable from text. In OLMo-2's pretraining corpus, the odds of "x is not P." against "x is P." are 1.7 times higher when P occurs nearby without being said of x, and lower when an equally frequent other property word does (Liu et al., 2024). This is where the models assert not-X, and where people, too, deny: they use "not" more often for what the context has raised (Givón, 1978; Horn, 1989; Wason, 1965; Nordmeyer & Frank, 2015). Pretrained models continuing a text show the same tilt (Appendix G).

5 Inside the Model

5.1 The Lookup Is the Property Word's Read

The lookup can be located in the forward pass. In every attention layer we block the attention from the question's property word to the positions of *quiet* in the text (Geva et al., 2023). Anne's negated question returns to unknown, while blocking "not" instead leaves it true. Over the 327 open pairs, Qwen3-32B's committed negated answers fall to 17%, the level reached by masking the word (Figure 4a), and both kinds of commitment go: negated true answers fall from 194 to 54, negated false from 85 to 0. In all four models we probe, from three families, the block lowers the committed negated answers, to near the masking level in all but Llama-3.1-8B, and blocking a control property barely matters (Appendix D.1). The read is also the route to written facts: under the block, Qwen3-32B's correct answers to 204 directly stated questions fall from 203 to 3 (Figure 4b).

A block shows that the read is necessary, but it could also just destroy information, as its effect on stated facts suggests. The state that the lookup leaves is enough to recreate the error: we move the residual state of the question tokens from a run on the text with the rule restored into a run on the text without *quiet* (Meng et al., 2022). Committed negated answers rise by 23 and 34 points in Qwen3-32B and Qwen3-14B, 44% and 72% of the effect of restoring the rule, while the plain

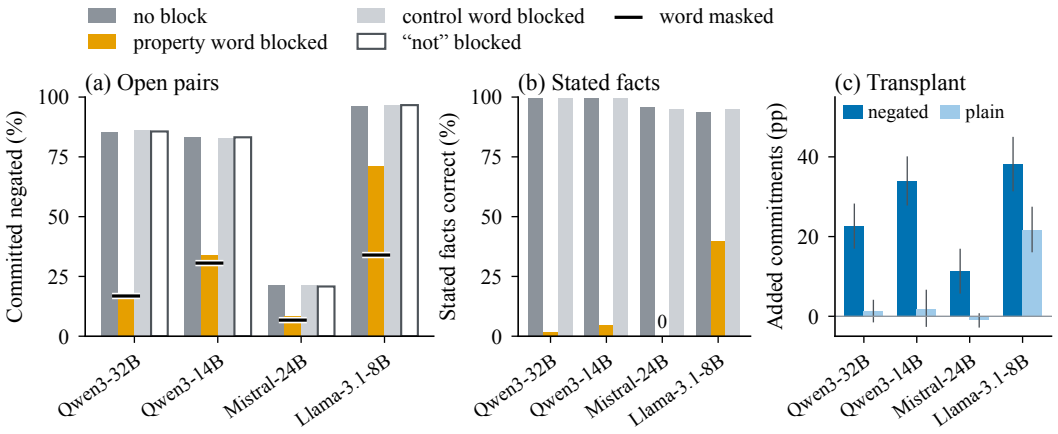


Figure 4: Blocking the property word's read lowers the committed negated answers, in Qwen3-32B to the level of masking the word; blocking "not" does not. (a) Open pairs (327) with the negated question committed, blocks in every layer; dash: property masked. (b) Directly stated known questions (204) answered correctly; under the property block Mistral-24B answers none. (c) The question's state moved from the text with the rule restored (Mistral-24B: other entities' facts) into the text without the property; added commitments against an unrelated move of the same radius (95% intervals).

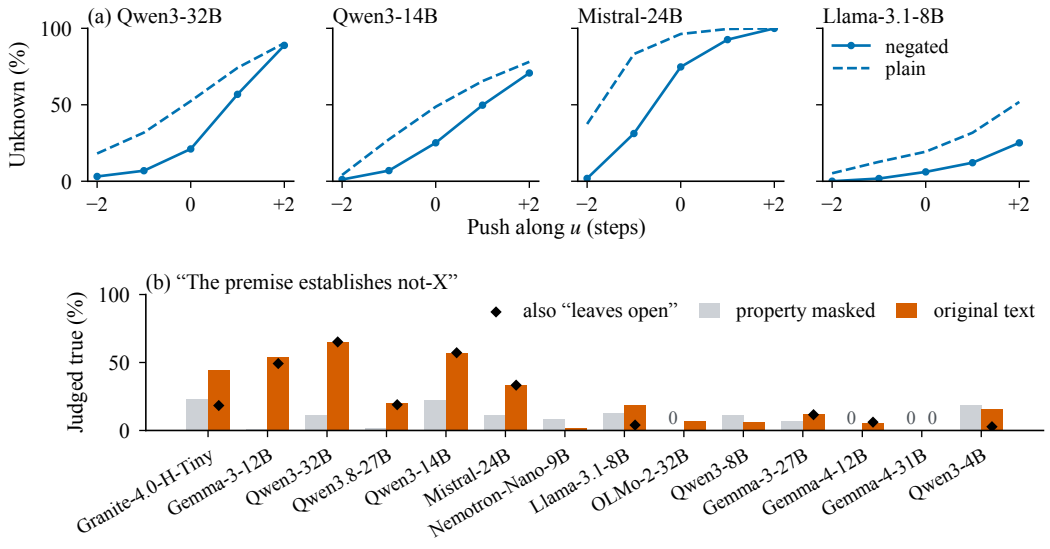
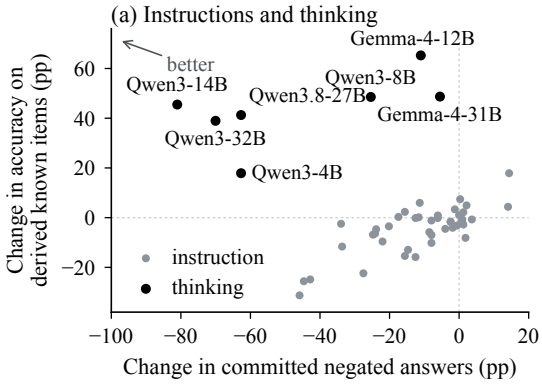


Figure 5: While judging not-X true, the model still registers that X is not established. (a) Pushing along u, the direction from known to open states (one step: 0.02 of the layer norm), moves both questions toward unknown. (b) Open pairs (327) with "The premise establishes not-X" judged true on two versions of the text; diamond: also judges "leaves open" true (original text).

question barely moves (Figure 4c; Appendix D.3). Both questions receive their state from the same donor text, yet only the negated one commits: the lookup reports one failure, and the "not" acts afterwards, at the answer position, which is why blocking its attention to the text changes nothing. There the step can be cut separately: ablating 512 heads (12% of all) at the answer position, chosen on half of the theories, turns about half of Qwen3-32B's negated true answers on the other half into unknown and leaves its other answers in place, and as many random heads change almost none (Appendix D.4).



(b) Qwen3-32B thinking on "Anne is not quiet." (answer unknown):
The rule is in the form of "If A and B, then C." We know C is true (green), but that doesn't mean A and B are true . . . there's no information that tells us whether Anne is quiet or not . . . the answer should be unknown.

A remaining error, "Dave is not red." (answer **true**):
Since the premises don't state that Dave is red, and there's no rule that would make him red . . . therefore the hypothesis that he is not red is supported by the premises.

Figure 6: Instructions mostly make models abstain; thinking largely replaces the lookup with a proof. (a) Change in committed negated answers (327 open pairs) and in accuracy on 344 derived known pairs, against the original prompt at zero temperature: three instructions for fourteen models (grey: a three-valued definition, alone or with "not every true fact is listed" or "'if A then B' does not establish A") and thinking, sampled, for seven (black).

5.2 Not Established, Yet Asserted

While answering true, the model still registers that X is not established. At the last position before the answer, a linear probe separates open from known items (AUROC 0.82 to 0.85) and recalls open items as often on negated as on plain questions. Adding the direction u between the mean states of open and known items (Rimsky et al., 2024) moves both questions toward unknown: on the open items of the full test set, one push raises Qwen3-32B's unknown answers to negated questions from 21% to 57% (Figure 5a), and a random direction of the same norm does nothing (Appendix E.3). Both questions carry the same "not established", and only the negated question's reading overrides it.

In its own words, the model says both that the premise leaves the question open and that it establishes the negation. Asked about the premise instead, Qwen3-32B judges "The premise leaves open whether Anne is quiet" true, "The premise establishes that Anne is not quiet" also true, and "The premise establishes that Anne is quiet" false. Over the 327 open pairs it accepts "establishes not-X" in 214 and "leaves open" in all but one of these (Figure 5b). For the model, established means found: even where X follows from the rules and the model answers the plain question true, it mostly says that the premise leaves X open, and often that it establishes not-X (Appendix E.8). The model thus reads $\neg KX$ as $K\neg X$, merging what logic programs keep apart as default and strong negation, and knowledge bases as not knowing and knowing not (Gelfond & Lifschitz, 1991; Levesque, 1984; Moore, 1985). A gap stated outright is respected: appending "Whether Anne is quiet is left open." lowers Qwen3-32B's committed negated answers from 86% to 2% (Appendix E.9).

6 Proof Instead of Lookup

The lookup also misses what a proof would find, and thinking, not instruction, repairs it. In one item, four rule applications prove that Fiona is quiet. Qwen3-32B misses the chain and judges "Fiona is not quiet." true. On such provable but unfound items, it judges the negated question true about as often as on open items (54% against 61% of the pairs with a plain unknown; Appendix C.2).

Telling the model to answer true or false only when the premises prove it makes it abstain. Both questions move toward unknown and known accuracy falls (Qwen3-32B: 62% to 52%). More specific instructions lower accuracy on derived known items in 30 of 42 cases (Figure 6a; Appendix F.1).

Letting the model think largely replaces the lookup with a proof. With thinking, Qwen3-32B works through the rule on Anne and concludes unknown (Figure 6b). On the full test set, under matched sampling, thinking lowers its asymmetry from 29 to 4 points and raises its accuracy on known questions from 61% to 88%; in all seven models with a thinking mode, known accuracy reaches 80 to 98% (Appendix F.6), and the models almost stop saying that the premise establishes not-X

(Appendix F.4). Thinking does not merely make the models abstain more readily: in all seven, the change is incompatible with a shift of the response criterion alone (Appendix F.7). The traces that stay wrong spell out negation as failure: 42 of Qwen3-32B's 53 remaining committed negated answers on the edited pairs infer "not X" from a failed search (Appendix F.5).

7 Related Work

Negation as failure (Clark, 1978), the closed-world assumption (Reiter, 1978) and their combination with classical negation (Gelfond & Lifschitz, 1991) define the readings in Figure 1a. Closed-world reasoning can be restricted to parts of a knowledge base (Etzioni et al., 1997; Baral & Gelfond, 1994), as the model does to mentioned properties, though only under a negated question. People reasoning with conditionals match on a rule's words and largely disregard its negations (Evans, 1998), a pattern modeled with negation as failure (Stenning & van Lambalgen, 2008; Kowalski, 2011). Comprehenders often first represent the negated affirmative (Kaup et al., 2006), and a hearer who takes the speaker to be competent strengthens "not known that p" to "known that not p" (Grice, 1975; Sauerland, 2004). Sparing the plain question, this strengthening fails the last two edits of Table 3.

In language models, negation lowers accuracy in factual probes, inference and reading comprehension (Kassner & Schütze, 2020; Hossain et al., 2020; She et al., 2023; Truong et al., 2022; Ravichander et al., 2022; Zhang et al., 2023; García-Ferrero et al., 2023; Truong et al., 2023), and a negated hypothesis cues contradiction in NLI data (Gururangan et al., 2018; Poliak et al., 2018). These benchmarks mostly score each negated item against its gold answer; our open pairs share one text and differ only in the "not", so the error shows between two answers. Models also answer when they should abstain (Yin et al., 2023; Amayuelas et al., 2024; Kirichenko et al., 2025) and miss absent content (Fu et al., 2025; Fox et al., 2026), and lean to false without parametric knowledge (Song et al., 2025). A model that had learned the closed-world labels of RuleTaker and ProofWriter (Clark et al., 2020; Tafjord et al., 2021) would judge the plain question false as well, as in the F/T pairs, not leave it unknown. CROWN-QA (Min et al., 2026) varies whether evidence covers the query, and NAFBench (Bao et al., 2026), closest in framing, gives models the semantics of "not" in rules and tests whether they follow it. We hold the evidence fixed, give no semantics, and ask how the "not" in a question is read by default. Interventions locate where a model reads an entity's attributes (Geva et al., 2023; Huang et al., 2024; Sun et al., 2025) and how it binds entities to attributes (Feng & Steinhardt, 2024); we use attention blocking and state patching to find a lookup whose failure only the negated question reads as an answer. Internal states hold unanswerability, truth or knowledge that the output misses (Lavi et al., 2025; Du & Hu, 2026; Liu et al., 2025; Azaria & Mitchell, 2023; Marks & Tegmark, 2024; Orgad et al., 2025; Ferrando et al., 2025; Yan et al., 2026), and "not" moves the representation to the negated concept (Zhou et al., 2026).

8 Limitations

We test fourteen open-weight models up to 32B with one answer format, statements judged true, false or unknown. NLI answer words change the asymmetry in some models (Appendix F.3), interventions cover four models from three families, and evidence on the habit's origin is correlational.

9 Conclusion

When a model does not find X in a text that mentions P, it keeps the plain question open and judges the negated one true, reading what is not established as established not. The lookup behind both answers also finds the facts the model gets right; reading its failure as "not" predates instruction tuning and mirrors how people deny what the context has raised. People accept "not found, so not" to the degree that the searched source is complete (Walton, 1996; Oaksford & Hahn, 2004); the models apply it to texts the task leaves open, and only under a "not", as a retrieval system that missed a document or an agent with an incomplete log would. Because the trigger is narrow, scoring a negated question with its plain twin on the same text exposes the error, stating the gap removes it, and thinking replaces the lookup with a proof. What is not proven is read as proven not.

AI USE STATEMENT

We used large language models as general-purpose assistants in this work: to search the literature, to discuss ideas, derivations, experiments and results, to help write code and process data, and to draft, translate, edit and polish the text and figures. The authors directed the research and take full responsibility for the content of this paper. All AI-assisted content was reviewed by the authors, and the reported results, references and claims were checked against the underlying data, code and sources.

ETHICS STATEMENT

This work evaluates open-weight language models on public benchmarks (ProofWriter and ANLI) and on synthetic texts we wrote; it involves no human subjects, and the record scenarios use invented names. The failure it documents matters for systems that answer from incomplete documents or records, which may assert a negation the text never supports; the edits of Section 4.1 and the stated-gap check of Section 5.2 can serve as tests for it.

REPRODUCIBILITY STATEMENT

The supplementary material contains the item sets and prompts, the models' answers to every item of these sets, the model identifiers, the inference and analysis code, and a script that recomputes the paper's numbers from the stored statistics; its README maps each table, figure and appendix section to its files. The models are open-weight. Answers without thinking are decoded at zero temperature, thinking uses each model's recommended sampling, and Appendix A.1 gives the protocol.

REFERENCES

Alfonso Amayuelas, Kyle Wong, Liangming Pan, Wenhu Chen, and William Yang Wang. Knowledge of knowledge: Exploring known-unknowns uncertainty with large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 6416–6432. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.383. URL http://dx.doi.org/10.18653/v1/2024.findings-acl.383.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.68. URL https://aclanthology.org/2023.findings-emnlp.68/.

Qiming Bao, Agnieszka Mensfelt, Michael J. Witbrock, and Kostas Stathis. Not what you meant: Can LLMs follow a specified negation semantics?, 2026. URL https://arxiv.org/abs/2609.27517.

Chitta Baral and Michael Gelfond. Logic programming and knowledge representation. *The Journal of Logic Programming*, 19–20:73–148, 1994.

Keith L. Clark. Negation as failure. In Gallaire and Minker (eds.), *Logic and Data Bases*, pp. 293–322. Springer US, 1978. ISBN 9781468433845. doi: 10.1007/978-1-4684-3384-5_11. URL http://dx.doi.org/10.1007/978-1-4684-3384-5_11.

P. Clark, O. Tafjord, and K. Richardson. Transformers as soft reasoners over language. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2020.

Yucheng Du and Xiyang Hu. Recognition-refusal misalignment in LLMs: Why models answer structurally unanswerable questions, 2026. URL https://arxiv.org/abs/2608.29109.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Number 57 in Monographs on Statistics and Applied Probability. Chapman & Hall, New York, 1993. ISBN 0-412-04231-2. doi: 10.1201/9780429246593.

Oren Etzioni, Keith Golden, and Daniel S. Weld. Sound and efficient closed-world reasoning for planning. *Artificial Intelligence*, 89(1–2):113–148, 1997. doi: 10.1016/S0004-3702(96)00026-4. URL https://doi.org/10.1016/S0004-3702(96)00026-4.

J. St. B. T. Evans. Matching bias in conditional reasoning: Do we understand it after 25 years? *Thinking & Reasoning*, 4(1):45–110, 1998. doi: 10.1080/135467898394247.

Jiahai Feng and Jacob Steinhardt. How do language models bind entities in context? In *International Conference on Learning Representations (ICLR)*, 2024.

Javier Ferrando, Oscar Obeso, Senthooran Rajamanoharan, and Neel Nanda. Do I know this entity? knowledge awareness and hallucinations in language models. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2411.14257.

Sebastian Fox, Luke Markham, Ryan Lail, and Michael Karotsieris. LLM judges verify presence, not absence: Omission blindness in AI clinical notes and what recovers it, 2026. URL https://arxiv.org/abs/2608.31016.

Harvey Yiyun Fu, Aryan Shrivastava, Jared Moore, Peter West, Chenhao Tan, and Ari Holtzman. AbsenceBench: Language models can't tell what's missing, 2025. URL https://arxiv.org/abs/2506.11440.

Iker García-Ferrero, Begoña Altuna, Javier Alvez, Itziar Gonzalez-Dios, and German Rigau. This is not a dataset: A large negation benchmark to challenge large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8596–8615. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.531. URL http://dx.doi.org/10.18653/v1/2023.emnlp-main.531.

M. Gelfond and V. Lifschitz. Classical negation in logic programs and disjunctive databases. *New Generation Computing*, 9(3–4):365–385, 1991. doi: 10.1007/BF03037169.

Gemma Team, Google DeepMind. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Gemma Team, Google DeepMind. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.751. URL https://aclanthology.org/2023.emnlp-main.751/.

T. Givón. Negation in language: Pragmatics, function, ontology. In Cole (ed.), *Syntax and Semantics 9: Pragmatics*, pp. 69–112. Academic Press, 1978.

A. Grattafiori, A. Dubey, et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

H. P. Grice. Logic and conversation. In Peter Cole and Jerry L. Morgan (eds.), *Syntax and Semantics 3*, pp. 41–58. Academic Press, New York, 1975.

S. Gururangan et al. Annotation artifacts in natural language inference data. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018.

L. R. Horn. *A Natural History of Negation*. University of Chicago Press, 1989.

M. M. Hossain et al. An analysis of natural language inference benchmarks through the lens of negation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. RAVEL: Evaluating interpretability methods on disentangling language model representations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8669–8687, 2024.

IBM Granite Team. IBM Granite 4.0: Hyper-efficient, high performance hybrid models for enterprise, 2025. URL https://www.ibm.com/new/announcements/ibm-granite-4-0-hyper-efficient-high-performance-hybrid-models.

Nora Kassner and Hinrich Schütze. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7811–7818. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.698. URL http://dx.doi.org/10.18653/v1/2020.acl-main.698.

B. Kaup, J. Lüdtke, and R. A. Zwaan. Processing negated sentences with contradictory predicates. *Journal of Pragmatics*, 38:1033–1050, 2006.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J. Bell. AbstentionBench: Reasoning LLMs fail on unanswerable questions, 2025. URL https://arxiv.org/abs/2506.09038.

R. Kowalski. *Computational Logic and Human Thinking*. Cambridge University Press, 2011.

W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica. Efficient memory management for large language model serving with PagedAttention. In *ACM Symposium on Operating Systems Principles (SOSP)*, 2023. URL https://arxiv.org/abs/2309.06180.

Maor Juliet Lavi, Tova Milo, and Mor Geva. Detecting (un)answerability in large language models with linear directions, 2025. URL https://arxiv.org/abs/2509.22449.

Hector J. Levesque. Foundations of a functional approach to knowledge representation. *Artificial Intelligence*, 23(2):155–212, 1984.

Jeffrey Li, Alex Fang, Georgios Smyrnis, et al. DataComp-LM: In search of the next generation of training sets for language models, 2024. URL https://arxiv.org/abs/2406.11794.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. In *Conference on Language Modeling (COLM)*, 2024. URL https://arxiv.org/abs/2401.17377.

Yi Liu, Xiangyu Liu, Zequn Sun, and Wei Hu. Answering the unanswerable is to err knowingly: Analyzing and mitigating abstention failures in large reasoning models, 2025. URL https://arxiv.org/abs/2508.18760.

B. Ma and Y. Miyao. The imperfective paradox in large language models. In *Annual Meeting of the Association for Computational Linguistics (ACL 2026), Long Papers*, pp. 15093–15111, 2026. doi: 10.18653/v1/2026.acl-long.689.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Conference on Language Modeling (COLM)*, 2024. URL https://arxiv.org/abs/2310.06824.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, 2022. URL https://arxiv.org/abs/2202.05262.

Byoungjae Min, Kennedy Edemacu, Sae-Hong Cho, Yoonhyuk Choi, Beakcheol Jang, and Jong Wook Kim. When absence is evidence: Evaluating completeness-sensitive negative reasoning in large language models, 2026. URL https://arxiv.org/abs/2608.04591.

Mistral AI. Mistral Small 3. Blog post, January 2025. URL https://mistral.ai/news/mistral-small-3/.

R. C. Moore. Semantical considerations on nonmonotonic logic. *Artificial Intelligence*, 25:75–94, 1985.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4885–4901, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.441. URL https://aclanthology.org/2020.acl-main.441/.

Ann E. Nordmeyer and Michael C. Frank. The pragmatics of negation across contexts. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, 2015. URL https://langcog.stanford.edu/papers/NF-cogsci2015.pdf.

NVIDIA. NVIDIA Nemotron Nano 2: An accurate and efficient hybrid Mamba-Transformer reasoning model, 2025. URL https://arxiv.org/abs/2508.14444.

M. Oaksford and U. Hahn. A Bayesian approach to the argument from ignorance. *Canadian Journal of Experimental Psychology*, 58(2):75–85, 2004.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2410.02707.

A. Poliak et al. Hypothesis only baselines in natural language inference. In *Joint Conference on Lexical and Computational Semantics (*SEM)*, 2018.

Qwen Team. Qwen3.8-27B. Model card, Hugging Face, 2026. URL https://huggingface.co/Qwen/Qwen3.8-27B.

Abhilasha Ravichander, Matt Gardner, and Ana Marasović. CONDAQA: A contrastive reading comprehension dataset for reasoning about negation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 8729–8755, 2022. doi: 10.18653/v1/2022.emnlp-main.598.

Raymond Reiter. On closed world data bases. In Gallaire and Minker (eds.), *Logic and Data Bases*, pp. 55–76. Springer US, 1978. ISBN 9781468433845. doi: 10.1007/978-1-4684-3384-5_3. URL http://dx.doi.org/10.1007/978-1-4684-3384-5_3.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Uli Sauerland. Scalar implicatures in complex sentences. *Linguistics and Philosophy*, 27(3):367–391, 2004. doi: 10.1023/B:LING.0000023378.71748.db.

Jingyuan Selena She, Christopher Potts, Samuel R. Bowman, and Atticus Geiger. ScoNe: Benchmarking negation reasoning in language models with fine-tuning and in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2023.

Jongyoon Song, Sangwon Yu, and Sungroh Yoon. A multifaceted analysis of negative bias in large language models through the lens of parametric knowledge. *IEEE Transactions on Audio, Speech and Language Processing*, 2025. URL https://arxiv.org/abs/2511.10881.

K. Stenning and M. van Lambalgen. *Human Reasoning and Cognitive Science*. MIT Press, 2008.

Jiuding Sun, Jing Huang, Sidharth Baskaran, Karel D'Oosterlinck, Christopher Potts, Michael Sklar, and Atticus Geiger. HyperDAS: Towards automating mechanistic interpretability with hypernetworks. In *International Conference on Learning Representations (ICLR)*, 2025.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3621–3634. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-acl.317. URL http://dx.doi.org/10.18653/v1/2021.findings-acl.317.

Team OLMo, Pete Walsh, Luca Soldaini, et al. 2 OLMo 2 furious, 2024. URL https://arxiv.org/abs/2501.00656.

T. H. Truong et al. Not another negation benchmark: The NaN-NLI test suite for sub-clausal negation. In *Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (AACL)*, 2022.

Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. Language models are not naysayers: an analysis of language models on negation benchmarks. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pp. 101–114. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.starsem-1.10. URL http://dx.doi.org/10.18653/v1/2023.starsem-1.10.

D. Walton. *Arguments from Ignorance*. Penn State Press, 1996.

Peter C. Wason. The contexts of plausible denial. *Journal of Verbal Learning and Verbal Behavior*, 4 (1):7–11, 1965. doi: 10.1016/S0022-5371(65)80060-3.

Qiyao Yan, Chenpeng Wang, and Liangming Pan. Wrong prediction, right answer: Recovering evidence from collapsed LLM sequence scores, 2026. URL https://arxiv.org/abs/2608.31068.

A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, et al. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8653–8665. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.551. URL http://dx.doi.org/10.18653/v1/2023.findings-acl.551.

Yuhui Zhang, Michihiro Yasunaga, Zhengping Zhou, Jeff Z. HaoChen, James Zou, Percy Liang, and Serena Yeung. Beyond positive scaling: How negation impacts scaling trends of language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 7479–7498. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.472. URL http://dx.doi.org/10.18653/v1/2023.findings-acl.472.

Z. Zhou, T. Zhou, R. Jia, and J. May. How language models process negation. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2605.03052.

A SETUP

All rule items come from the open-world, depth-5 test set of ProofWriter (Tafjord et al., 2021) (release V2020.12.3). Its 948 theories carry 20,030 official questions, and we add the 12,688 reverse questions that open-world semantics determine. This gives 32,718 questions per model, or 16,359 pairs of a plain question "x is P" and a negated question "x is not P". The fourteen models give 458,052 zero-temperature answers, all of them valid: 294,462 from the nine core models (Qwen3-4B, 8B, 14B and 32B, Qwen3.8-27B, Gemma-4-12B and 31B, Llama-3.1-8B and Mistral-24B) and 163,590 from the five further models of Appendix A.5. Every question follows the same six examples and ends with the same instruction:

> Judge the hypothesis using only the premise and no unstated assumptions. Premise: ...
> Hypothesis: ... Reply with exactly one lowercase answer word: true, false, or unknown.

The record scenarios and the ANLI paragraphs (Appendices B.2 and B.3) use the same six examples and the same final instruction word for word. Llama-3.1-8B, Mistral-24B and the five further models answer in their direct mode (Nemotron-Nano-9B-v2 with its reasoning switched off). The seven models with a thinking switch answer with thinking off unless stated otherwise.

Notation. We write the answers to an open pair as plain/negated, with U, T and F for unknown, true and false. U/T is a pair whose plain question is answered unknown and whose negated question is answered true: it asserts not-X through the negated question alone. F/U asserts not-X through the plain question alone, and T/U asserts X through the plain question alone. UD collects U/T and U/F

(negated committed, plain unknown), DU collects T/U and F/U, and DD collects the four patterns in which both questions are committed. Differences between committed shares are in percentage points, and intervals are 95% bootstrap intervals clustered by theory, by scenario for the records, and by pair for ANLI (Efron & Tibshirani, 1993). We call the nine models other than the five further ones of Appendix A.5 the *core models*. Of these, the four with the largest lookup effect at the rule condition are Qwen3-32B, Qwen3-14B, Qwen3.8-27B and Llama-3.1-8B, and with Qwen3-4B they form the five that look up; Qwen3-8B, the two Gemma-4 models and Mistral-24B barely look up at the rule condition (Appendix C.4). Five further models join Table 2 (Appendix A.5): Granite-4.0-H-Tiny and Gemma-3-12B look up at the rule condition, Nemotron-Nano-9B-v2 mainly at the facts listed about other entities, as Mistral-24B does, and OLMo-2-32B and Gemma-3-27B only a little. The appendix analyses below cover the core models unless they name others.

A.1 Protocol and read-out

The protocol makes every answer valid, stable and checkable: the three answer options are listed explicitly, answers without thinking are decoded at zero temperature, the read-out can be checked item by item, and thinking is compared on and off under identical sampling.

- **Options.** Listing the three options explicitly, and only in the final sentence, makes 64 of 64 non-thinking answers of Qwen3-14B valid in a pilot. Listing the options also changes judgments (non-thinking Qwen3-4B answers 8 instead of 13 of 15 known questions correctly), so every run uses this final prompt in full.
- **Zero temperature.** With each model's recommended sampling, 43 of the 523,488 answers of a nine-model table contained no answer, and in a targeted diagnosis 12 of 17 items changed label across three seeds. Because sampled labels were unstable, non-thinking answers are decoded at zero temperature. In a greedy pilot on four models, 64 of 64 answers were readable and replaying 8 items changed none.
- **Read-out.** The parser accepts every answer format that occurs in the outputs. Widening it to the last rare formats raised the number of Llama-3.1-8B pairs with the plain question unknown and the negated question committed from 2,828 to 2,862, so a more complete read-out does not shrink the effect.
- **Truncation.** In the sampled nine-model table, 5,135 outputs reached the 8,192-token limit, 4,232 of them from Gemma-4-12B. We continued them up to 32k tokens and checked each one.
- **Backend.** Moving from Transformers to vLLM (Kwon et al., 2023) on eight GPUs changed 0, 1, 1 and 2 of 64 labels in a small batch for Qwen3-4B, 8B, 14B and 32B, so labels can differ by a few items across backends and batches.
- **Thinking.** Thinking modes are not meant for greedy decoding, and greedy thinking loops on small models. Thinking comparisons therefore use each model's official sampling recipe with identical parameters for thinking off and on (458,052 answers on the full test set, where for 43 answers we take the first explicit answer among existing reruns). On the edited items of Section 4.1, thinking runs use a limit of 32,768 tokens, and the answer is read after the last `</think>`. On the edited items, answers without thinking are the zero-temperature answers and answers with thinking are sampled. On the full set, sampling changes the answers without thinking little: the net asymmetry stays within about 2 points of its zero-temperature value for six of the seven models (Qwen3.8-27B: +23.6 at zero temperature, +16.4 sampled).

A.2 Both questions ask about one proposition

Negating a hypothesis can change which proposition it expresses, and once the proposition is held fixed an apparent failure of negation disappears. This is why our main data are ProofWriter items, where the premises and a prover fix what is unknown and the plain and negated questions share one proposition.

On aspect items of ImperfectiveNLI (Ma & Miyao, 2026), such as "X was building a gazebo", the negated hypothesis "X did not build a gazebo" was judged contradicted both when the premise

describes an interrupted event and when it is ambiguous. We rewrote the hypotheses with two templates that allow only the completion reading, one with *finish* ("did not finish building") and one with *done*, for 200 source pairs and both questions. Six models (Llama-3.1-8B, Qwen3-8B, 14B and 32B, Gemma-3-12B and 27B (Gemma Team, Google DeepMind, 2025)) answer with an NLI-label six-example interface whose answer words are *entailment*, *neutral* and *contradiction*. Under the *finish* template, 63–100% of the negated arm in the ambiguous group returns to unknown, and in the interrupted group no negated hypothesis is judged contradicted any more (Table 4). The earlier collapse came from the negation replacing the proposition, not from a failure to negate. With the *done* template, Llama-3.1-8B returns to unknown less often (0.23), and the other models as often as with *finish*.

Table 4: Fixing the proposition removes the apparent failure of negation on aspect items. Share of negated hypotheses answered unknown in the ambiguous group and judged contradicted in the interrupted group, for the original wording and the completion-only templates.

Model	Ambiguous: unknown original → *finish* → *done*	Interrupted: contradicted original → *finish*
Llama-3.1-8B	0.00 → 0.63 → 0.23	0.65 → 0.00
Qwen3-8B	0.02 → 1.00 → 0.99	0.49 → 0.00
Qwen3-14B	0.00 → 0.99 → 0.99	0.58 → 0.00
Qwen3-32B	0.02 → 1.00 → 1.00	0.70 → 0.00
Gemma-3-12B	0.00 → 0.92 → 0.95	0.39 → 0.00
Gemma-3-27B	0.08 → 1.00 → 1.00	0.32 → 0.00

A.3 No in-context examples

Removing the six examples leaves the error and the lookup in place: Qwen3-32B's asymmetry is +40.1 points without examples and +40.7 with them, and its lookup effect is +56.9 and +48.6. For Qwen3-14B and Qwen3.8-27B the examples enlarge the error (+21.1 and +6.7 without them, +36.1 and +28.7 with them). We ran the four round-trip conditions of Section 4.1 (original, property masked, only the rule restored, only the control rule restored) on the same 773 pairs and both questions, keeping the final instruction verbatim and dropping all examples. Nine models answer at zero temperature with the same read-out (6,184 answers per model). The generation limit is 64 tokens. Without examples, Llama-3.1-8B writes an explanation before answering, so its limit is 768 tokens (64 truncated 7 of the first 32 answers), and 17% of its answers still give none of the three labels and count as invalid.

Without examples, removing the property lowers the committed negated answers in every model with the error (Qwen3-32B 83.8% → 13.5%, Qwen3-14B 82.6% → 18.3%, Qwen3.8-27B 26.6% → 0.6%, Llama-3.1-8B 89.9% → 48.6%). Restoring only the rule commits the negated question more often than the plain question in the three large Qwen models, Llama-3.1-8B and Qwen3-4B, and the two Gemma-4 models still keep both questions at unknown (Table 5). Qwen3-4B, whose failure falls on both questions with six examples, commits the negated question more than the plain one without them (lookup effect +25.1). The asymmetries of Qwen3-8B and Mistral-24B do not come from the rule condition (lookup effects +5.5 and 0.0), as with six examples. The examples change the size of the error for some models, most of all Qwen3.8-27B, but they are not its source.

A.4 Confident errors: logit margins

The negated true on open items is given with a first-token margin comparable to that of true on provable items: for Qwen3-14B, the median logit difference between true and unknown is 12.5 on open negated questions answered true and 14 on known true items answered true (Table 6). The failure is reported with confidence.

Qwen3-14B answers without thinking (BF16, SDPA attention, greedy decoding, up to 2,048 tokens). The inputs are 90 open theories in original and normalized form, each with the plain and the negated question, plus 48 known plain–negated pairs: 456 inputs and 3,648 answers, with 16 more answers as checks. A read-only hook on the output layer records the raw logits of true, false and unknown at the first answer position. The false answers behave the same way (12.875 on open negated questions,

Table 5: Without examples, the error and the lookup remain. On 327 open pairs: committed shares (%) of the negated and plain question on original items and of the negated question after masking; asymmetry (negated minus plain committed share on original items) and lookup effect in points without examples, with 95% intervals (5,000 resamples), next to the six-example values; known questions answered correctly without examples (directly stated of 204, derived of 688); valid answers.

	Original	Masked	Asymmetry		Lookup effect		Known correct	Valid
Model	neg. / plain	neg.	no examples	six	no examples	six	direct / derived	
Qwen3-32B	83.8 / 43.7	13.5	+40.1 [+33.2, +47.0]	+40.7	+56.9 [+50.6, +63.3]	+48.6	204/204 / 300/688	100.0%
Qwen3-14B	82.6 / 61.5	18.3	+21.1 [+16.1, +26.3]	+36.1	+41.0 [+33.5, +48.4]	+46.8	204/204 / 231/688	100.0%
Qwen3.8-27B	26.6 / 19.9	0.6	+6.7 [+1.8, +11.5]	+28.7	+17.1 [+11.8, +23.0]	+43.1	202/204 / 355/688	100.0%
Mistral-24B	11.6 / 1.5	4.9	+10.1 [+6.4, +14.0]	+16.5	+0.0 [−2.9, +2.9]	−2.8	179/204 / 31/688	100.0%
Llama-3.1-8B	89.9 / 67.9	48.6	+22.0 [+15.4, +29.2]	+14.4	+16.2 [+8.8, +24.1]	+20.8	79/204 / 216/688	82.8%
Qwen3-8B	43.4 / 17.7	14.4	+25.7 [+18.6, +32.6]	+5.5	+5.5 [+1.6, +9.8]	+1.2	189/204 / 250/688	99.9%
Gemma-4-31B	0.6 / 5.2	0.0	−4.6 [−7.0, −2.4]	−6.1	+1.2 [+0.3, +2.5]	+2.8	204/204 / 207/688	100.0%
Gemma-4-12B	1.2 / 4.0	0.0	−2.8 [−5.3, −0.6]	−1.8	+0.0 [+0.0, +0.0]	+1.2	195/204 / 74/688	100.0%
Qwen3-4B	82.6 / 76.5	19.9	+6.1 [+1.8, +10.8]	−4.6	+25.1 [+17.1, +33.4]	−4.3	203/204 / 445/688	100.0%

15.25 on known false items). The margins are raw logit differences rather than calibrated confidences, and the groups are defined by the model's own answers, so we report only medians and counts. The 90 open theories are not held out from the analyses of Appendix D.2.

Table 6: Negated true answers on open items are about as confident as true answers on provable items. Median first-token logit difference for Qwen3-14B, grouped by the model's answer.

Group	n	Median margin
Open, negated question answered true	30	true − unknown: 12.5
Known true, answered true	22	true − unknown: 14
Open, negated question answered false	28	false − unknown: 12.875
Known false, answered false	12	false − unknown: 15.25

A.5 Further model families

Five instruction-tuned models from three more families and the earlier Gemma-3 release join the nine in Table 2: Granite-4.0-H-Tiny (IBM Granite Team, 2025), a hybrid of Mamba-2 and attention layers with 7B parameters of which 1B are active; Gemma-3-12B and Gemma-3-27B (Gemma Team, Google DeepMind, 2025), the release before Gemma-4; OLMo-2-32B-Instruct (Team OLMo et al., 2024); and Nemotron-Nano-9B-v2 (NVIDIA, 2025), a hybrid reasoning model whose reasoning is switched off with its template's "/no_think" system turn. They answer the same item sets with the same six examples, instruction, read-out and zero temperature, through the same runner; generation is capped at 4,096 tokens (1,024 for OLMo-2, whose context is 4,096), and every answer ends before the cap. Their statistics come from the scripts that produced the nine models' statistics, run over the nine and the new models together; the nine models' values are reproduced exactly. None of the five has a thinking switch that we compare (Nemotron's reasoning mode is not run), so Section 6 keeps its seven thinking models. Table 7 gives their readouts beyond Table 2. Gemma-3-4B and Llama-3.2-3B-Instruct, run the same way, are in Appendix C.8.

The two further models that look up at the rule condition behave like the three large Qwen models. Granite-4.0-H-Tiny and Gemma-3-12B assert not-X through the negated question in thousands of pairs and through the plain question in a few dozen, the excess survives in records and human-written paragraphs and mostly disappears with the control text, writing the condition as "not P" moves the plain question, and asked about the premise they accept "establishes not-X" two to three times as often as "establishes X". The open-world instruction removes most of their asymmetry at a cost in known accuracy (Table 43).

The others follow their lookup. OLMo-2-32B and Gemma-3-27B look up little at either place and rarely accept "establishes not-X"; OLMo-2-32B's asymmetry is larger on the full set of rule items (+11.5 points) than on the 327 edited pairs (+3.1). Nemotron-Nano-9B-v2 looks up mainly at the facts listed about other entities (+18.9), as Mistral-24B does (Appendix C.7), and its asymmetry follows that lookup. Under the three specific instructions only Gemma-3-12B loses derived-known

Table 7: The five further models on the item sets of the main text. Pairs are counted as negated only (U/T) / plain only (F/U); brackets: the control text (a record that never mentions the property; an unrelated paragraph). Lookup effects and changes are in points; statements are counted over the 327 open pairs; the three instructions change accuracy on the plain question of derived known items against the original prompt.

	Granite-4.0-H-Tiny	Gemma-3-12B	OLMo-2-32B	Gemma-3-27B	Nemotron-Nano-9B-v2
Rule items (8,654 open pairs)	2,552 / 62	2,115 / 21	1,336 / 6	791 / 46	921 / 132
Lookup effect: rule condition / listed facts	+31.2 / +12.5	+16.5 / +28.4	+6.1 / +4.9	+4.0 / +0.8	+7.0 / +18.9
Records, others listed [never mentioned]	62 / 0 [13 / 0]	60 / 0 [1 / 0]	8 / 0 [0 / 0]	9 / 0 [0 / 0]	20 / 3 [1 / 0]
ANLI, 1,144 pairs [unrelated paragraph]	240 / 11 [18 / 1]	94 / 25 [22 / 3]	98 / 28 [3 / 4]	31 / 15 [2 / 0]	71 / 36 [2 / 12]
Condition "not P": plain false, before → after (287)	20 → 42	6 → 114	3 → 32	1 → 12	21 → 55
Negated condition about the subject / another entity	+13.9 / +8.0	+19.4 / +8.8	+22.2 / +1.5	0.0 / +0.7	+1.4 / +5.1
"The premise establishes X" / "… not-X"	50 / 146	73 / 178	30 / 22	37 / 38	15 / 6
"The premise leaves open whether X"	152	265	0	292	136
Open-world instruction: asymmetry, before → after	+34.7 → +9.9	+32.3 → +2.0	+11.5 → −0.7	+1.6 → −2.8	+15.0 → +0.7
Open-world instruction: known accuracy, before → after	43.7 → 24.2	59.5 → 31.1	26.3 → 8.1	42.4 → 36.7	46.2 → 13.7
Three instructions: derived-known accuracy	+0.3 / +7.0 / +2.9	−9.6 / −8.1 / −21.5	+0.6 / +7.3 / +1.5	+2.6 / +18.3 / −5.5	+3.5 / +2.3 / +7.3

accuracy each time; the others hardly change or gain, while the open-world instruction costs all five known accuracy.

B THE PHENOMENON

This appendix gives the full readouts behind Table 2 and Figure 2: all answer patterns on the rule items, the five record types, the ANLI paragraphs with their known controls, and two further analyses of the phenomenon.

B.1 RULE ITEMS: ALL ANSWER PATTERNS

On the 8,654 open pairs, the models with the error commit the negated question far more often than the plain one, the excess is spread over almost all theories, and it remains when the premises contain no negation at all. Table 8 counts all nine joint answer patterns, and Table 9 summarizes them. Table 2 uses U/T and F/U, and the closed-world pattern of Section 1 is F/T.

Table 8: In the models with the error, most assertions of not-X come through the negated question alone (U/T) or through both questions in the closed-world pattern (F/T); Qwen3-4B asserts not-X mostly through the plain question alone (F/U). Counts of the 8,654 open pairs by joint answer, plain first and negated second (U: unknown, T: true, F: false).

Model	U/U	U/T	U/F	T/U	F/U	F/T	T/F	T/T	F/F
Qwen3-32B	1,626	2,387	478	101	60	1,916	1,131	724	231
Qwen3-14B	1,963	1,543	671	92	91	1,997	1,100	543	654
Qwen3.8-27B	4,198	2,066	117	129	12	371	626	1,129	6
Mistral-24B	6,368	1,936	20	120	1	118	28	63	0
Llama-3.1-8B	484	300	893	3	47	904	804	16	5,203
Qwen3-8B	6,318	769	232	449	57	144	634	34	17
Gemma-4-31B	7,515	213	51	420	48	129	254	21	3
Gemma-4-12B	7,456	187	62	419	16	80	354	71	9
Qwen3-4B	2,633	8	374	171	1,354	634	1,309	8	2,163
Five further models (Appendix A.5)									
Granite-4.0-H-Tiny	1,672	2,552	776	266	62	1,322	1,029	542	433
Gemma-3-12B	2,021	2,115	847	166	21	896	1,637	717	234
OLMo-2-32B	6,298	1,336	100	432	6	214	218	48	2
Gemma-3-27B	6,080	791	103	710	46	161	635	90	38
Nemotron-Nano-9B-v2	4,608	921	723	215	132	882	537	34	602

The net increase is a difference in commitment, not an accuracy. The share P is computed only where the model has already recognized the gap on the plain question, so its denominator differs across models. We therefore recompute it on the 729 pairs whose plain question all nine models answer unknown, and on the 1,745 pairs where all four Qwen3 models do. On the nine-model pairs, the negated question is committed in 24.7% (Qwen3-14B) and 33.9% (Qwen3-32B) of pairs, and on the

Table 9: The negated question is committed more often than the plain one in the models with the error. UD: plain unknown, negated committed; DU: the reverse; DD: both committed. Net: (UD − DU)/8,654 in points. P: UD/(U/U + UD), the committed share of the negated question among pairs whose plain question is unknown. true in UD: pairs of UD whose negated answer is true, and their share. Known: accuracy on the 11,376 official known questions of the test set.

Model	UD	DU	DD	Net	P	true in UD	Known
Qwen3-32B	2,865	161	4,002	+31.25	63.8%	2,387 / 83.3%	62.26%
Qwen3-14B	2,214	183	4,294	+23.47	53.0%	1,543 / 69.7%	59.27%
Qwen3.8-27B	2,183	141	2,132	+23.60	34.2%	2,066 / 94.6%	64.31%
Mistral-24B	1,956	121	209	+21.20	23.5%	1,936 / 99.0%	21.66%
Llama-3.1-8B	1,193	50	6,927	+13.21	71.1%	300 / 25.1%	63.48%
Qwen3-8B	1,001	506	829	+5.72	13.7%	769 / 76.8%	45.42%
Gemma-4-31B	264	468	407	−2.36	3.4%	213 / 80.7%	54.31%
Gemma-4-12B	249	435	514	−2.15	3.2%	187 / 75.1%	43.87%
Qwen3-4B	382	1,525	4,114	−13.21	12.7%	8 / 2.1%	67.07%
Five further models (Appendix A.5)							
Granite-4.0-H-Tiny	3,328	328	3,326	+34.67	66.6%	2,552 / 76.7%	43.71%
Gemma-3-12B	2,962	187	3,484	+32.07	59.4%	2,115 / 71.4%	59.43%
OLMo-2-32B	1,436	438	482	+11.53	18.6%	1,336 / 93.0%	26.27%
Gemma-3-27B	894	756	924	+1.59	12.8%	791 / 88.5%	42.43%
Nemotron-Nano-9B-v2	1,644	347	2,055	+14.99	26.3%	921 / 56.0%	46.17%

Qwen3 pairs in 37.4% and 47.4%. For Qwen3-4B and 8B the shares are 5.2% and 0.1% (7.6% and 1.2% on the Qwen3 pairs). Mistral-24B drops from 23.5% to 1.5%, and Llama-3.1-8B is at 56.1%, mostly with false. On the same items, the two larger Qwen3 models commit the negated question far more often than the two smaller ones. On known items, the negated question of Qwen3-14B and 32B is answered unknown only a third as often as the plain question (7.7% against 21.9%, and 5.8% against 19.5%), the same phenomenon on known items.

Spread over theories. The error is not carried by a few odd theories. Of the 948 theories, the net asymmetry is positive in 812 for Qwen3-32B (negative in 14), 743 for Qwen3-14B (18), 750 for Qwen3.8-27B (34), 732 for Mistral-24B (19) and 538 for Llama-3.1-8B (14). The 50 theories with the largest net asymmetry contribute only 13.8, 15.2, 15.6, 16.9 and 23.1% of the total.

The classically unknown half. In 4,400 of the 8,654 open pairs the premises contain no negation, so not-X cannot be derived in classical logic either. The net asymmetry remains in this half, at about the size it has in the half whose premises contain negation, larger for Llama-3.1-8B and smaller for OLMo-2-32B: +33.3 and +29.2 points for Qwen3-32B, +34.2 and +35.2 for Granite-4.0-H-Tiny, +32.0 and +32.2 for Gemma-3-12B, +24.6 and +22.3 for Qwen3-14B, +22.5 and +24.8 for Qwen3.8-27B, +17.9 and +8.3 for Llama-3.1-8B, +20.5 and +21.9 for Mistral-24B, +15.7 and +14.3 for Nemotron-Nano-9B-v2, and +8.1 and +15.1 for OLMo-2-32B (negation-free half first). Qwen3-4B, Qwen3-8B, Gemma-3-27B and the two Gemma-4 models lie between −18 and +8 in both halves.

B.2 Everyday records: five record types

In everyday records, the negated question is committed alone where the lookup has something to look at: other people listed with the property, or an answer that needs a rule whose condition is not recorded for the subject. We built records in 8 domains (workplace, school, insurance, library, rental, warranty, clinic, club) with 30 scenarios each. Every record has three policy rules, facts about two other people and two unrelated facts about the subject, and no entry on whether the subject meets the target condition, so the gold labels hold by construction. The six examples and the final instruction are those of the rule items, and all fourteen models answer at zero temperature. The five record types are:

- *property not mentioned*: none of the three rules involves the property;
- *others listed with it*: the two other people are said to have the property;
- *answer needs a rule*: the question asks for a rule's conclusion whose condition is not recorded for the subject;

- *property only in a rule condition*: the record version of restoring only the rule in Section 4.1;
- *gap stated*: the record ends by saying that whether the subject meets the condition is open.

Figure 2a counts pairs by joint answer for all fourteen models. Table 10 and the rest of this subsection give the committed share of each question for the nine core models; the five further models are in Table 7.

Table 10: Records commit the negated question when others are listed with the property or when the answer needs a rule. Committed share (%) of 240 open pairs per record type and question; in brackets, for cells of at least 5%, the share of committed answers that are true. Columns: Qwen3-32B, Qwen3-14B, Qwen3.8-27B, Mistral-24B, Llama-3.1-8B, Qwen3-8B, Gemma-4-31B, Gemma-4-12B, Qwen3-4B.

Record	Question	32B	14B	27B	Mistral	Llama	8B	G-31B	G-12B	4B
Property not mentioned	neg.	1	1	0	0	2	0	0	0	0
	plain	0	1	0	0	1	0	0	0	0
Others listed with it	neg.	59 [95]	63 [38]	1	3	65 [12]	23 [40]	1	1	38 [5]
	plain	20 [6]	37 [6]	0	0	26 [29]	7 [62]	1	0	31 [21]
Answer needs a rule	neg.	25 [95]	26 [84]	0	0	54 [2]	5 [50]	0	0	17 [0]
	plain	1	14 [65]	0	0	16 [82]	16 [100]	0	0	31 [73]
Property only in a rule condition	neg.	2	19 [98]	0	0	18 [59]	0	0	0	0
	plain	0	0	0	0	1	0	0	0	0
Gap stated	neg.	0	0	0	0	0	0	0	0	0
	plain	0	0	0	0	0	0	0	0	0

Relative to the record that never mentions the property, the difference between the negated and the plain committed share (points, scenario-clustered intervals) for *others listed with it* is +38.8 [+32.5, +45.0] for Qwen3-32B, +26.7 [+20.8, +32.5] for Qwen3-14B, +37.1 [+30.4, +43.8] for Llama-3.1-8B, +16.7 [+11.7, +21.7] for Qwen3-8B and +7.1 [+1.3, +12.9] for Qwen3-4B, with Qwen3.8-27B, both Gemma-4 models and Mistral-24B within ± 3. For *answer needs a rule* it is +22.5 [+17.1, +28.3] for Qwen3-32B, +11.7 [+6.3, +17.5] for Qwen3-14B and +37.1 [+30.4, +43.8] for Llama-3.1-8B. Qwen3-4B (-14.2) and Qwen3-8B (-10.8) answer the plain question true more often than the negated one, and Qwen3.8-27B, both Gemma-4 models and Mistral-24B are at 0. On known items, directly stated facts are answered correctly by all nine models in 96–100% of cases. For known answers derived through a rule, the plain question is correct in 95–100% of cases, and the negated question in 96% (Qwen3-32B), 98% (Qwen3-14B), 100% (Qwen3.8-27B), 97% (Llama-3.1-8B), 97% and 99% (the two Gemma-4 models), 85% (Qwen3-4B), 77% (Qwen3-8B) and 51% (Mistral-24B).

The records reproduce the pattern of the rule items at two lookup sites: the listed facts (others have the property and the subject is not listed) and a derivation that fails (the answer needs a rule whose condition is not recorded). Of Qwen3-32B's committed negated answers, 95% are true. On records that need a rule, Llama-3.1-8B takes the rule as holding for the subject (its negated question is committed in 54% of pairs, almost always false), and Qwen3-4B and 8B answer the plain question true, reading the rule as an assertion, as they do when the property sits in a rule's conclusion on rule items (Appendix C.2). Qwen3-32B's committed negated answers on records that list others are spread over all eight domains, so no single kind of document carries them.

When the property appears only in a policy rule's condition, only Qwen3-14B (19%) and Llama-3.1-8B (18%) commit more negated questions, and Qwen3-32B and Qwen3.8-27B barely move. In records, the models look up the listed facts and the conclusions that must be derived through a rule, and for Qwen3-32B and Qwen3.8-27B a policy rule's condition is not a place to look. Where the lookup looks depends on the genre of the text, just as Mistral-24B looks only at listed facts on rule items (Appendix C.7). When the gap is stated, no model commits either question, as with a stated gap on rule items (Appendix E).

B.3 Human-written paragraphs

On ANLI paragraphs, the four models that look up most assert not-X through the negated question 2.2 to 4.1 times as often as through the plain question, and the excess over the plain question disappears when the paragraph is replaced by an unrelated one. We start from the neutral pairs of ANLI (Nie et al., 2020) (dev and test sets of all three rounds) and keep the 1,144 whose hypothesis is a single

clause that we can negate. These are paragraphs from Wikipedia, news and fiction whose premise neither entails nor contradicts the hypothesis, so the hypothesis and its negation both have the gold label unknown. We keep single-clause hypotheses and negate the finite verb (auxiliary plus *not*, or *do*-support), checking or writing each negation by hand. Each pair also receives an unrelated premise that shares no content word or proper name with the hypothesis, chosen with a fixed seed, and is asked the same two questions. Prompt, examples and final instruction are those of the rule items, with all fourteen models at zero temperature. 200 entailed and 200 contradicted pairs serve as known controls.

Two examples show the two exits. The Dufton paragraph gives the parish 169 residents in 2001 and 204 in 2011. Asked "In 2002, Dufton parish in Cumbria, England had a population of exactly 175.", thirteen of the fourteen models answer unknown (Gemma-3-12B answers false). For its negation, Qwen3-32B, Qwen3-14B, Llama-3.1-8B, Qwen3-4B, Qwen3-8B, Gemma-3-12B, Gemma-3-27B, OLMo-2-32B and Nemotron-Nano-9B-v2 answer true: nothing is found. With a paragraph about a film in its place, all fourteen answer unknown to both questions. The Via Dante paragraph (here for the nine core models) says that this street in Milan is named after the poet Dante. Asked "Milan's streets are named after poets.", eight models answer unknown. For "Milan's streets are not named after poets.", Qwen3-32B, Qwen3-14B, Qwen3.8-27B, Qwen3-8B, Gemma-4-31B and Mistral-24B answer false: something close is found. With an unrelated paragraph, all nine models answer unknown to both questions.

Table 11: On human-written paragraphs, both exits of the lookup favor the negated question in the four models that look up most, and both almost vanish with an unrelated paragraph. Counts of the 1,144 neutral ANLI pairs. Neg. true: U/T; plain false: F/U, its mirror; neg. false: U/F; plain true: T/U, its mirror; consistent: both questions committed, consistently with each other.

	Real premise			Unrelated premise		
Model	neg. true / plain false	neg. false / plain true	consistent	neg. true / plain false	neg. false / plain true	consistent
Qwen3-32B	88 / 37	54 / 18	246	0 / 2	0 / 0	5
Qwen3-14B	116 / 53	60 / 25	432	0 / 60	3 / 1	12
Qwen3.8-27B	72 / 32	93 / 6	164	2 / 8	0 / 0	4
Mistral-24B	25 / 2	30 / 3	13	0 / 1	1 / 0	0
Llama-3.1-8B	107 / 26	69 / 33	432	40 / 37	1 / 2	29
Qwen3-8B	54 / 41	51 / 29	179	1 / 34	0 / 0	4
Gemma-4-31B	16 / 34	51 / 16	64	0 / 1	0 / 0	4
Gemma-4-12B	29 / 16	41 / 17	68	4 / 0	0 / 0	1
Qwen3-4B	5 / 145	82 / 15	198	0 / 79	1 / 0	2
Five further models (Appendix A.5)						
Granite-4.0-H-Tiny	240 / 11	68 / 58	268	18 / 1	0 / 1	4
Gemma-3-12B	94 / 25	54 / 30	326	22 / 3	0 / 2	14
OLMo-2-32B	98 / 28	80 / 8	139	3 / 4	0 / 0	10
Gemma-3-27B	31 / 15	38 / 29	101	2 / 0	0 / 0	3
Nemotron-Nano-9B-v2	71 / 36	81 / 43	281	2 / 12	2 / 1	7

For the four models that look up most (Qwen3-32B, Qwen3-14B, Qwen3.8-27B and Llama-3.1-8B), negated true outnumbers plain false 2.2 to 4.1 times (88 to 37, 116 to 53, 72 to 32, 107 to 26), and 2.3 to 4.3 times when the negated false exit is added (142 to 55, 176 to 78, 165 to 38, 176 to 59). Granite-4.0-H-Tiny and Gemma-3-12B, which also look up at the rule condition, give 240 to 11 and 94 to 25. Negated false outnumbers its mirror in all fourteen models (Qwen3.8-27B 93 to 6, Mistral-24B 30 to 3). With an unrelated premise both exits almost vanish (Qwen3-32B 88 $\rightarrow$ 0 and 54 $\rightarrow$ 0, Qwen3-14B 116 $\rightarrow$ 0, Qwen3.8-27B 72 $\rightarrow$ 2 and 93 $\rightarrow$ 0), and so do the consistent commitments: these conclusions come from the text, not from what the models already know. Llama-3.1-8B still gives 40 negated true and 37 plain false with an unrelated premise: without the text it reports a missing item as a negation on both questions alike, and only the excess on the negated question needs the real premise. Relative to the unrelated premise, the real premise raises negated true by +7.7 [+6.2, +9.3] points for Qwen3-32B, +10.1 [+8.4, +11.9] for Qwen3-14B, +6.1 [+4.7, +7.6] for Qwen3.8-27B and +5.9 [+3.8, +7.9] for Llama-3.1-8B.

Llama-3.1-8B still gives 40 negated true answers with an unrelated premise, as many as its plain false answers (37). Without a text it reports a failed search as negation equally often on both questions, and only its negative-only excess needs the real premise. Human-written premises also support reasonable inferences that commit both questions together (the consistent column), so negative-only assertions are much rarer here than on the controlled rule items (Qwen3-32B: 88 pairs here, 2,387 on ProofWriter). Table 12 gives the known controls.

Table 12: Known controls on ANLI: accuracy on 200 entailed and 200 contradicted pairs, plain / negated question.

Model	Entailed	Contradicted
Qwen3-32B	74% / 64%	64% / 55%
Qwen3-14B	83% / 73%	68% / 48%
Qwen3.8-27B	75% / 74%	75% / 62%
Mistral-24B	26% / 26%	20% / 30%
Llama-3.1-8B	70% / 69%	64% / 30%
Qwen3-8B	56% / 40%	47% / 30%
Gemma-4-31B	56% / 65%	69% / 58%
Gemma-4-12B	52% / 49%	51% / 44%
Qwen3-4B	44% / 50%	59% / 19%
Five further models		
Granite-4.0-H-Tiny	70% / 46%	44% / 47%
Gemma-3-12B	78% / 65%	56% / 48%
OLMo-2-32B	59% / 54%	62% / 52%
Gemma-3-27B	57% / 41%	40% / 34%
Nemotron-Nano-9B-v2	70% / 67%	66% / 30%

B.4 Wording and templates

The negated-only commitment survives rewording. Contracted negation agrees with the original template on whether each item is committed, and sentence frames change the size of the effect but not its direction, with more explicit negations answered true more consistently (Table 13). Qwen3-14B and 32B answer all 8,654 open pairs with greedy decoding at zero temperature. T0 is the original template ("x is not P"), T1 is "It is not the case that x is P", T2 is "It is false that x is P", T3 uses contracted negation (*isn't*, *doesn't*), and C1 is an affirmative frame used as a control. Intervals come from 20,000 bootstrap resamples over 918 theory clusters, and p-values are Holm-corrected. The affirmative frame C1 also raises Qwen3-14B's plain committed share by 4.22 points [+3.59, +4.86], so part of the T1 and T2 increase comes from the sentence frame rather than from the negation, and the committed items differ between templates.

Table 13: Rewording changes the size of the effect, not its direction. Change in the share of pairs with the plain question unknown and the negated question committed, relative to T0 (points), and the share of true among these negated-only commitments. Status agreement: items with the same committed-or-unknown status as under T0.

Template	Qwen3-14B	Qwen3-32B	true share, 14B / 32B
T0 original	reference	reference	69.7% / 83.3%
T3 contracted	+0.20 [−0.19, +0.59]	−0.40 [−0.86, +0.05]	72.5% / 85.8%
	status agreement 96.14%	status agreement 94.48%	
T1 *It is not the case that*	+11.51 [+10.68, +12.35]	+3.76 [+3.06, +4.44]	85.9% / 91.1%
T2 *It is false that*	+15.85 [+14.91, +16.80]	−4.10 [−4.91, −3.31]	84.2% / 96.2%

B.5 The two exits of the lookup

A failed lookup has two exits on the negated question: true when the subject is not found, and false when another entity's property counts as found. Which exit dominates depends on what the model counts as found. Table 14 gives both exits and their mirror images on the plain question for all open rule pairs. The pattern is not a bias toward true on negated questions: on the same items the negated question is also often committed to false (478 pairs for Qwen3-32B, 671 for Qwen3-14B, 893 for Llama-3.1-8B). Llama-3.1-8B answers false to almost everything, and Qwen3-4B already answers false to many plain questions (1,354 F/U pairs).

What counts as found differs between models (Table 15). Among the edited open pairs with the plain question unknown (327; 264 for the facts about another entity), restoring only the rule condition gives the lookup something to look for and no subject to find. In the three large Qwen models it adds

Table 14: A failed lookup leaves the negated question in two directions. Open pairs (of 8,654) whose negated question is committed while the plain question is unknown, in each direction, next to the mirror patterns on the plain question.

Model	U/T	F/U (mirror)	U/F	T/U (mirror)
Qwen3-32B	2,387	60	478	101
Qwen3-14B	1,543	91	671	92
Qwen3.8-27B	2,066	12	117	129
Mistral-24B	1,936	1	20	120
Llama-3.1-8B	300	47	893	3
Qwen3-8B	769	57	232	449
Gemma-4-31B	213	48	51	420
Gemma-4-12B	187	16	62	419
Qwen3-4B	8	1,354	374	171
Five further models (Appendix A.5)				
Granite-4.0-H-Tiny	2,552	62	776	266
Gemma-3-12B	2,115	21	847	166
OLMo-2-32B	1,336	6	100	432
Gemma-3-27B	791	46	103	710
Nemotron-Nano-9B-v2	921	132	723	215

only negated true answers relative to the masked text (Qwen3-32B 20 → 180, Qwen3-14B 27 → 176, Qwen3.8-27B 21 → 157), with no increase in negated false; Llama-3.1-8B and Qwen3-4B add both (false: 16 → 46 and 1 → 8). Restoring only another entity's target fact ("Bob is quiet.") adds different answers in different models: relative to the control fact, Qwen3-32B mainly adds true (12 → 77), whereas Llama-3.1-8B and Qwen3-4B add false (13 → 58, 2 → 16).

The same split appears elsewhere. In records that list others with the property, the share of committed negated answers that are true is 95% for Qwen3-32B, 38% for Qwen3-14B, 12% for Llama-3.1-8B and 5% for Qwen3-4B (Table 10). On the original rule items, among pairs whose plain question is unknown and whose negated question is committed, it is 83% for Qwen3-32B, 95% for Qwen3.8-27B and 99% for Mistral-24B, against 25% for Llama-3.1-8B and 2% for Qwen3-4B (Table 9). A model that ties the lookup to the subject reads another entity's record as "she was not found" and answers true. A model that ties it to the property reads the record as found and answers false. Both answers stay on the negated question, and on human-written text both exits are present and both need the text (Appendix B.3).

Table 15: In the three large Qwen models, restoring a rule condition adds only negated true, whereas restoring another entity's fact adds true or false depending on the model. Negated true / negated false among edited open pairs whose plain question is unknown (327 pairs; 264 for the two fact rows), after masking or after restoring only the named mention. Llama: Llama-3.1-8B.

Edit	Qwen3-32B	Qwen3-14B	Qwen3.8-27B	Llama	Qwen3-4B
Masked	20 / 0	27 / 7	21 / 0	35 / 16	0 / 1
Only the rule	180 / 0	176 / 6	157 / 0	59 / 46	0 / 8
Only the control rule	20 / 1	22 / 7	18 / 0	33 / 13	0 / 3
Only others' target fact	77 / 7	33 / 15	30 / 3	24 / 58	0 / 16
Only others' control fact	12 / 0	22 / 7	12 / 0	28 / 13	0 / 2

C ONE LOOKUP

Unless stated otherwise, the analyses in this appendix use the edited set of Section 4.1: 773 pairs from the same ProofWriter theories, namely 327 open, 102 directly stated known and 344 derived known pairs. The premises of the open pairs contain no negation and mention the target property. The target property and a control property with the same role and number of mentions are both replaced by made-up words (*masked*), and then exactly one kind of mention of one of them is restored: in a rule's condition, in a rule's conclusion, or in facts about other entities. Questions, the remaining facts

and gold labels stay fixed. Each model answers once under the zero-temperature protocol, and on the original texts the answers of the nine core models agree with those of Table 2 in 98.8–100% of items. Figure 7 shows all fourteen models on the 229 open pairs for which all five plotted edits can be built. Tables 21, 23 and 26 include all fourteen models; for the other tables, Table 7 gives the main edit readouts of the five further models.

C.1 Removing and restoring the property

Restoring only the rule that mentions the property raises the negated committed share of the three large Qwen models by 45 to 52 points, while their plain question moves by at most 3.4 points and the control rule moves neither (Table 16). The committed shares under every edit are in Tables 17 and 18.

Only the rule restored. Relative to the masked text, the negated committed share rises by 51.7 points for Qwen3-32B, 45.3 for Qwen3-14B and 45.0 for Qwen3.8-27B, with the plain question at +0.6 to +3.4. Restoring the control rule changes almost nothing. Llama-3.1-8B rises by 56.3 points but its plain question by 37.0, and Qwen3-4B rises by 23.5 and 26.6: they look up, and the failure falls on both questions. Qwen3-8B, the two Gemma-4 models and Mistral-24B stay within ± 3: they do not look up at the rule condition.

Nothing to look up, little error. With the property masked, the negated question still reads "not quiet", but the text contains nothing about *quiet* to look up, and the negated question is committed far less (Qwen3-32B 86% $\rightarrow$ 17%, Qwen3-14B 84% $\rightarrow$ 31%, Qwen3.8-27B 63% $\rightarrow$ 9%), only 5 to 10 points more often than the plain question. What is reported as "not" is something that the text makes the model look for and that it fails to find. Restoring only a fact about another entity ("Bob is quiet.", with the subject not listed) raises Qwen3-32B's negated question by 34.8 points and its plain question by 9.1: a listed extension is also a place to look. Qwen3-4B, Llama-3.1-8B and Qwen3-14B commit both questions here, a closed world by enumeration.

Not backward reasoning from the conclusion. We split the 327 open pairs by whether the restored rule's conclusion already holds for the subject (182 and 145 pairs). The rise of the negated question after restoring the rule is nearly the same in both strata: +47.8 and +56.6 points for Qwen3-32B, +45.6 and +44.8 for Qwen3-14B, +46.2 and +43.4 for Qwen3.8-27B, +53.8 and +59.3 for Llama-3.1-8B, and +20.9 and +26.9 for Qwen3-4B (conclusion holds first). Contrapositive or abductive reasoning from the conclusion would not commit the negated question when the conclusion already holds. The four models that do not look up at the rule condition stay within ± 4 in both strata.

The lookup concerns the subject. We rewrite the condition rule in uniform wording and bind it either to the subject or to another entity. Binding it to the subject commits the negated question more often by 37.3 points for Qwen3.8-27B, 20.4 for Qwen3-4B, 18.5 for Qwen3-14B, 15.0 for Qwen3-32B and 11.8 for Llama-3.1-8B, while the plain question of the three large Qwen models stays in place (-1.0 to +0.6). The four models that do not look up at the rule condition stay near 0 or below. The same word produces the error only when it enters a lookup about the subject.

The same structure in unedited items. Stratifying the 8,654 original open pairs by where the property occurs gives the same picture without any edit (Appendix C.5).

C.2 All eighteen edits

The full grid adds three regularities to the round trip: a rule with the property in its conclusion commits the plain question, a gap statement worded with "is not" still lets Qwen3-14B answer many negated questions true, and when a proof exists but is not found, the negated question is answered true about as often as on open items. The eighteen edits are the fifteen in Tables 17 and 18 and three in which the hypothesis is replaced by a statement about the premise ("The premise does not establish X", "establishes not-X", "establishes X"), reported in Appendix E. The two gap statements append "It is not stated whether x is P." or "Whether x is P is left open." to the original premises, for the target or the control property. The uniform rule keeps a single condition rule in uniform universal wording, and its two bound versions concern only the subject ("If Anne is quiet and Anne is furry then Anne is green.") or only another entity ("If Gary is quiet ...").

Property in the conclusion. Restoring a rule whose conclusion is the target property raises the plain committed share by +28 to +54 points in Qwen3-4B, 8B, 14B, 32B, Qwen3.8-27B, Gemma-4-12B

Table 16: Restoring the rule condition commits the negated question of the models that look up, and the control rule commits nothing. Paired differences in committed share relative to the masked text (points), negated question with 95% intervals, plain question below.

Model	Question	Rule (condition)	Control rule	Rule (conclusion)	Others' target fact
Qwen3-32B	neg.	+51.7 [+44.8, +58.5]	+0.0 [−1.7, +1.6]	+72.7 [+66.7, +78.6]	+34.8 [+26.6, +43.1]
	plain	+3.4	+0.3	+45.8	+9.1
Qwen3-14B	neg.	+45.3 [+38.2, +52.2]	−2.1 [−4.1, −0.3]	+52.2 [+46.4, +58.0]	+30.7 [+23.9, +37.0]
	plain	+0.6	+0.0	+42.8	+26.1
Qwen3.8-27B	neg.	+45.0 [+37.7, +52.3]	−0.9 [−2.5, +0.6]	+57.9 [+51.5, +64.2]	+5.3 [+0.8, +9.8]
	plain	+3.1	+0.3	+37.7	−0.8
Mistral-24B	neg.	−2.8 [−6.3, +0.6]	+0.0 [−1.8, +2.0]	+2.0 [−2.8, +6.7]	+7.6 [+2.4, +12.9]
	plain	+0.0	+0.0	+3.4	+0.8
Llama-3.1-8B	neg.	+56.3 [+49.2, +63.2]	−0.3 [−3.1, +2.4]	+57.9 [+51.0, +64.5]	+56.8 [+50.2, +63.5]
	plain	+37.0	+1.2	+42.4	+36.4
Qwen3-8B	neg.	−0.6 [−3.5, +2.4]	−0.9 [−2.3, +0.3]	+13.5 [+8.6, +18.6]	+10.2 [+5.8, +14.9]
	plain	−1.2	−0.3	+28.3	+6.4
Gemma-4-31B	neg.	+0.3 [+0.0, +1.0]	+0.3 [+0.0, +1.0]	+5.1 [+2.5, +8.2]	+0.4 [+0.0, +1.2]
	plain	−3.1	−0.3	+11.1	−1.1
Gemma-4-12B	neg.	+1.2 [+0.0, +2.9]	+0.0 [+0.0, +0.0]	+21.2 [+15.8, +27.2]	+3.8 [+1.5, +6.5]
	plain	+0.0	+0.0	+33.0	+1.1
Qwen3-4B	neg.	+23.5 [+18.1, +29.2]	−0.6 [−2.3, +0.6]	+60.3 [+53.7, +67.0]	+53.0 [+46.1, +59.9]
	plain	+26.6	−1.8	+54.2	+46.6

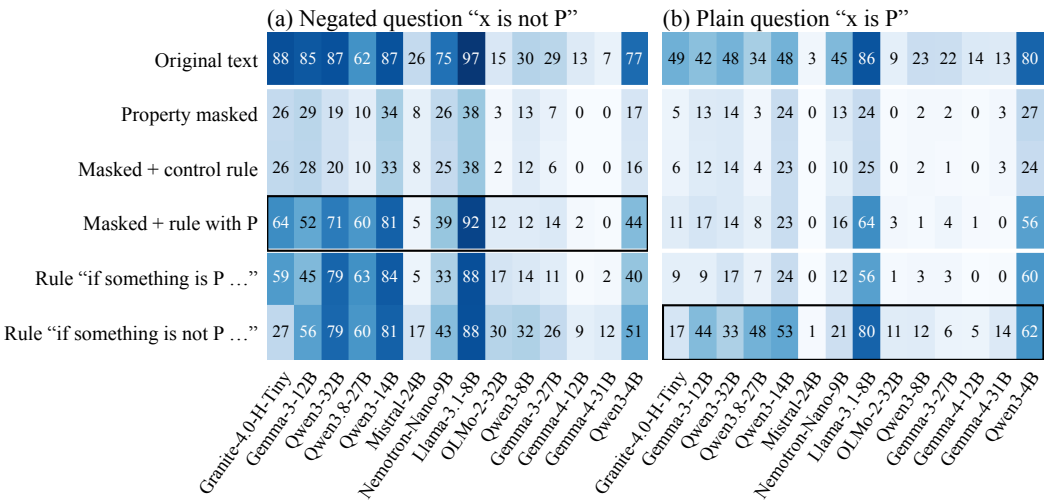


Figure 7: Every main-table model under the edits of Section 4.1. Percent committed on the 229 open pairs for which all five edits can be built, under the original text and the five edits (the last two: the restored rule in a uniform wording, and negated); outlined in each panel: the row that separates the accounts of Table 3; black lines separate the three groups of Section 4.2.

and Llama-3.1-8B, by +11 in Gemma-4-31B and by +3 in Mistral-24B (Table 16). The Qwen models and Gemma-4-12B mostly answer true (65–100%), and Llama-3.1-8B mostly false. Restoring others' target facts commits both questions in Qwen3-4B, Llama-3.1-8B and Qwen3-14B, a closed world by enumeration, whereas Qwen3-32B, Qwen3.8-27B and Mistral-24B lean to the negated question.

Bound to the subject. Binding the rule to another entity still commits more negated questions than masking (Qwen3-14B +15.1, Qwen3-32B +20.5, Llama-3.1-8B +40.4 points; Table 19), so the presence of the word contributes a part by itself. Binding it to the subject adds more (Appendix C.1).

The wording of the gap statement. "It is not stated whether x is P." itself contains "is not". Under it Qwen3-14B still commits 118 negated questions, all true, in the same direction as the flip in which a written negation counts as found (Appendix C.3). Qwen3-32B and Qwen3.8-27B stay at or below 9% under both wordings. The main text (Section 5.2) uses "Whether x is P is left open."

Table 17: Masking the property removes most negated commitments, and restoring its rule condition brings them back in the models that look up. Negated committed share (%) on 327 open pairs; brackets: share of committed answers that are true. 32B, 14B, 8B, 4B: Qwen3; 27B: Qwen3.8-27B; G: Gemma-4.

Edit	32B	14B	27B	Mistral	Llama	8B	G-31B	G-12B	4B
Original	86 [70]	84 [70]	63 [83]	20 [97]	97 [14]	28 [34]	6 [0]	12 [34]	71 [8]
Masked (both properties made up)	17 [100]	31 [91]	9 [100]	7 [100]	34 [77]	11 [100]	0	0	13 [86]
Restore rule, target in condition	69 [99]	76 [96]	54 [100]	4 [100]	90 [60]	10 [97]	0 [100]	1 [75]	37 [35]
Restore rule, control in condition	17 [98]	28 [90]	9 [100]	7 [100]	34 [75]	10 [100]	0 [100]	0	13 [88]
Restore rule, target in conclusion	89 [65]	82 [71]	68 [77]	9 [100]	91 [18]	24 [49]	5 [7]	21 [57]	74 [8]
Restore rule, control in conclusion	17 [98]	30 [92]	9 [100]	7 [100]	36 [72]	10 [100]	0 [100]	0	12 [84]
Restore others' target fact	53 [86]	62 [65]	15 [90]	15 [98]	92 [28]	22 [77]	0 [100]	4 [80]	69 [23]
Restore others' control fact	16 [100]	30 [90]	7 [100]	6 [100]	37 [74]	8 [100]	0	0	14 [81]
Original + target "not stated"	9 [100]	36 [100]	1 [33]	4 [100]	1 [100]	0	0 [0]	0	1 [100]
Original + control "not stated"	56 [79]	56 [70]	37 [76]	6 [95]	46 [27]	2 [100]	4 [0]	9 [46]	3 [55]
Original + target "left open"	2 [100]	13 [98]	1 [67]	3 [100]	1 [75]	0	0	0	4 [29]
Original + control "left open"	61 [78]	57 [66]	36 [78]	9 [100]	58 [24]	2 [88]	7 [4]	12 [50]	11 [34]
Uniform rule, universal	80 [100]	82 [96]	64 [99]	4 [100]	89 [61]	14 [100]	2 [80]	0 [100]	37 [45]
Uniform rule, subject only	52 [96]	64 [88]	45 [99]	3 [100]	86 [40]	5 [93]	1 [100]	0	42 [7]
Uniform rule, other entity only	38 [99]	46 [99]	9 [100]	3 [100]	75 [68]	9 [100]	0	0	21 [42]

Table 18: For the three large Qwen models the plain question stays near its masked level when the rule condition is restored and rises when the property sits in a rule's conclusion. Plain committed share (%) on 327 open pairs; brackets: share of committed answers that are true. Columns as in Table 17.

Edit	32B	14B	27B	Mistral	Llama	8B	G-31B	G-12B	4B
Original	46 [62]	48 [45]	34 [97]	4 [100]	82 [18]	22 [100]	12 [100]	13 [95]	75 [38]
Masked (both properties made up)	12 [0]	20 [3]	3 [0]	0	21 [0]	3 [0]	3 [0]	1 [0]	24 [0]
Restore rule, target in condition	15 [37]	21 [32]	6 [71]	0	58 [6]	2 [100]	0 [100]	1 [100]	50 [9]
Restore rule, control in condition	12 [0]	20 [3]	4 [0]	0	23 [0]	2 [0]	3 [0]	1 [0]	22 [0]
Restore rule, target in conclusion	57 [75]	63 [84]	41 [96]	3 [100]	63 [19]	31 [100]	14 [98]	34 [98]	78 [65]
Restore rule, control in conclusion	11 [0]	20 [3]	3 [0]	0 [0]	21 [0]	3 [0]	3 [0]	1 [0]	23 [0]
Restore others' target fact	22 [23]	48 [14]	2 [0]	1 [50]	61 [4]	8 [50]	1 [0]	2 [50]	72 [14]
Restore others' control fact	12 [0]	19 [2]	3 [0]	0	24 [0]	2 [0]	3 [0]	0 [0]	25 [0]
Original + target "not stated"	1 [100]	2 [25]	2 [83]	0	0	0	3 [100]	1 [100]	1 [0]
Original + control "not stated"	28 [76]	32 [59]	22 [92]	3 [100]	17 [32]	9 [100]	13 [100]	13 [100]	11 [47]
Original + target "left open"	0 [100]	1 [100]	2 [83]	0	0	0	2 [100]	0	4 [23]
Original + control "left open"	30 [74]	31 [58]	23 [93]	4 [100]	26 [40]	7 [100]	14 [100]	14 [98]	21 [30]
Uniform rule, universal	18 [31]	23 [25]	6 [62]	0	55 [4]	2 [29]	0 [100]	0	58 [4]
Uniform rule, subject only	5 [19]	8 [40]	1 [67]	0	44 [16]	0	0	0	47 [20]
Uniform rule, other entity only	5 [0]	9 [4]	0 [0]	0	28 [1]	0	0	0	37 [0]

A proof that exists but is not found. On the derived known pairs, X is always provable. When the plain question is answered unknown, so that the model has missed an existing proof, the negated question is answered true about as often as on open items: 54% and 61% for Qwen3-32B, 51% and 59% for Qwen3-14B, 53% and 44% for Qwen3.8-27B, 24% and 19% for Mistral-24B, 12% and 12% for Qwen3-8B, 9% and 16% for Llama-3.1-8B, 7% and 4% for Gemma-4-12B, 2% and 0% for Gemma-4-31B, and 0% and 0% for Qwen3-4B (derived known first, open second). What decides the negated question is whether the model's own lookup succeeds, not whether a proof exists, as in the Fiona item of Section 6.

C.3 Writing the condition as "not P"

Writing the target literal in the rule's condition as "not P" moves the error to the plain question. We start from the condition rule in uniform wording and flip the sign of its target literal ("is quiet" to "is not quiet"), changing nothing else. The target stays unprovable, and the literal to be judged is now negated. On 287 open pairs, the negated committed share of the three large Qwen models changes within ± 5 points, while their plain committed share rises by 16, 29 and 43 points (Qwen3-32B, Qwen3-14B, Qwen3.8-27B; Table 21). Plain false answers increase in all fourteen models, and the three large Qwen models also add plain true answers (Qwen3-14B +41, Qwen3-32B +33, Qwen3.8-27B +33; the other models at most +3). Qwen3-8B, the two Gemma-4 models and Mistral-24B, which do not look up at the rule condition, answer the negated question true 21 to 51 more times:

Table 19: Stating the gap removes the negated commitments, and binding the rule to the subject adds them. Paired differences in committed share (points), negated question with 95% intervals, plain question below.

Model	Question	Target left open − original	Control left open − original	Subject only − other only	Other only − masked
Qwen3-32B	neg.	−84.4 [−88.9, −79.7]	−25.4 [−30.7, −20.3]	+15.0 [+8.4, +21.5]	+20.5 [+14.4, +26.7]
	plain	−45.3	−15.6	−0.3	−6.6
Qwen3-14B	neg.	−71.3 [−76.6, −65.7]	−27.5 [−33.5, −21.9]	+18.5 [+12.2, +24.5]	+15.1 [+9.0, +21.5]
	plain	−47.4	−16.5	−1.0	−12.3
Qwen3.8-27B	neg.	−61.8 [−68.5, −54.9]	−26.3 [−32.1, −20.5]	+37.3 [+30.5, +44.3]	−0.6 [−4.6, +3.2]
	plain	−32.1	−11.0	+0.6	−3.2
Mistral-24B	neg.	−17.1 [−22.5, −12.0]	−11.6 [−16.0, −7.4]	−0.3 [−2.8, +2.4]	−3.8 [−7.1, −0.6]
	plain	−3.7	+0.3	+0.0	+0.0
Llama-3.1-8B	neg.	−95.4 [−97.7, −92.8]	−38.8 [−45.2, −32.3]	+11.8 [+6.8, +16.8]	+40.4 [+33.3, +47.6]
	plain	−82.3	−56.0	+17.2	+6.0
Qwen3-8B	neg.	−27.5 [−33.9, −21.5]	−25.1 [−31.1, −19.3]	−5.1 [−8.0, −2.5]	−1.9 [−4.9, +0.9]
	plain	−22.0	−15.0	+0.0	−2.8
Gemma-4-31B	neg.	−5.5 [−8.0, −3.2]	+1.5 [−0.6, +3.7]	+0.6 [+0.0, +1.6]	+0.0 [+0.0, +0.0]
	plain	−10.1	+2.8	+0.0	−3.5
Gemma-4-12B	neg.	−11.6 [−16.0, −7.8]	+0.0 [−3.0, +2.8]	+0.0 [+0.0, +0.0]	+0.0 [+0.0, +0.0]
	plain	−13.5	+0.6	+0.0	−0.6
Qwen3-4B	neg.	−66.4 [−73.0, −59.4]	−59.9 [−66.6, −53.2]	+20.4 [+14.6, +26.6]	+7.3 [+2.2, +12.5]
	plain	−71.3	−54.1	+10.2	+13.6

Table 20: Accuracy (%) on known pairs under the original condition, plain / negated question: 102 directly stated and 344 derived known pairs. Columns as in Table 17.

Known pairs	32B	14B	27B	Mistral	Llama	8B	G-31B	G-12B	4B
Directly stated	100/99	100/99	100/100	98/92	89/99	99/99	100/100	100/95	97/100
Derived	62/49	47/47	78/38	14/3	24/94	50/39	64/36	37/28	57/90

the "not P" written in the condition is taken as a negation that is found, as a written negative fact is (Appendix C.6).

Table 21: Flipping the condition to "not P" commits the plain question, and plain false answers increase in all fourteen models. Change in the plain committed share (points) and numbers of added plain false and true answers on 287 open pairs, relative to the same rule unflipped.

Model	Plain committed	Added plain false	Added plain true
Look up at the rule condition			
Qwen3-32B	+16.0	+13	+33
Granite-4.0-H-Tiny	+7.3	+22	−1
Gemma-3-12B	+33.1	+108	−13
Qwen3-14B	+29.3	+43	+41
Qwen3.8-27B	+42.5	+89	+33
Llama-3.1-8B	+23.0	+69	−3
Qwen3-4B	+2.4	+11	−4
Look up at the facts listed about other entities			
Mistral-24B	+1.0	+3	0
Nemotron-Nano-9B-v2	+10.1	+34	−5
Barely look up			
OLMo-2-32B	+9.8	+29	−1
Gemma-3-27B	+3.5	+11	−1
Qwen3-8B	+9.4	+29	−2
Gemma-4-31B	+12.9	+34	+3
Gemma-4-12B	+4.5	+13	0

In the original condition, the negated question looks up *quiet* and fails. After the flip, evaluating the condition's "not quiet" looks up *quiet* and fails. The negation in the condition forces the model to judge it for Anne first, and the result of that judgment stays on the plain question: the plain question is no longer answered unknown, plain false increases in all fourteen models and plain true in the three large Qwen models. Whenever "not P" must be judged, in the question or in a rule's condition, the model takes this lookup. A closed-world tendency gives the plain question the same answer before

and after the flip, and a bias toward true on negated questions never reaches the plain question, so neither predicts this change.

Evaluated for the entity, not copied from the text. The grid already contains the condition rule rewritten to concern only the subject ("If Anne is quiet and Anne is furry then Anne is green.") or only another entity ("If Gary is quiet ..."). We flip the target condition in each version and change nothing else. Copying the words "not P" from the text predicts the same effect for both flips, whereas evaluating the condition for the entity it concerns predicts that only the first moves the plain question. Table 22 uses the 132 open pairs in which both rewrites change a single rule. When the flip concerns another entity, the plain question of the three large Qwen models stays in place (-3 to +4 points). When it concerns the subject, the plain committed share rises by 24 to 53 points, in both directions (Qwen3-32B answers false on Anne). Llama-3.1-8B's plain question rises under both flips, consistent with a lookup that does not distinguish entities (Appendix B.5). The "not P" in a condition is evaluated for the entity it is about.

Table 22: A flipped condition moves the plain question mainly when it concerns the subject (Llama-3.1-8B also for another entity). On 132 open pairs: committed plain questions for the rewritten rule and after its flip (in parentheses, after the flip: false / true), and the flip effect on the plain committed share (points, 95% intervals).

	Committed plain questions		Flip effect		
Model	subject only	other only	subject	other	difference
Qwen3-32B	13 → 45 (17/28)	10 → 6	+24.2 [+15.1, +33.8]	−3.0 [−8.4, +1.6]	+27.3 [+18.0, +37.0]
Qwen3-14B	13 → 74 (49/25)	14 → 19	+46.2 [+37.6, +55.0]	+3.8 [−0.7, +8.1]	+42.4 [+33.6, +51.5]
Qwen3.8-27B	1 → 71 (36/35)	1 → 4	+53.0 [+42.3, +63.7]	+2.3 [+0.0, +5.0]	+50.8 [+40.0, +61.6]
Mistral-24B	0 → 0	0 → 0	0.0	0.0	0.0
Llama-3.1-8B	49 → 97 (96/1)	36 → 66	+36.4 [+26.6, +46.2]	+22.7 [+14.8, +31.6]	+13.6 [+2.9, +24.1]
Qwen3-8B	0 → 20 (20/0)	0 → 2	+15.2 [+7.3, +23.8]	+1.5 [+0.0, +3.9]	+13.6 [+6.5, +21.3]
Gemma-4-31B	0 → 7 (7/0)	0 → 2	+5.3 [+1.4, +10.3]	+1.5 [+0.0, +3.9]	+3.8 [+0.0, +8.3]
Gemma-4-12B	0 → 0	0 → 0	0.0	0.0	0.0
Qwen3-4B	72 → 82 (82/0)	56 → 52	+7.6 [+0.7, +14.5]	−3.0 [−8.5, +2.3]	+10.6 [+1.6, +19.6]

C.4 WHO LOOKS UP, AND SINCE WHEN

The error follows the lookup: the lookup effect at the rule condition and the error on the original items have rank correlation 0.75 across the fourteen models, and 0.74 across eleven checkpoints under a template-free prompt. Negation as failure needs two conditions. The model looks up the target property (restoring only the rule raises the negated committed share), and it reports the failure only on the question with "not" (the plain question does not follow). Table 23 lays out the fourteen models along these two conditions. The lookup effect is the rise of the negated committed share after restoring only the rule minus the rise of the plain committed share, both relative to restoring only the control rule. It is measured on the edited items and the error on the unedited items, two versions of the same pairs. Their rank correlation over the fourteen models is 0.75, $n = 14$; a one-sided exact permutation test over all 14! orderings of the models gives $p = 0.0015$ (tied values get their average rank; Qwen3-8B and Gemma-4-12B tie on the lookup effect). Over the nine core models the rank correlation is 0.71 (0.53 with the negated lookup alone), with $p = 0.019$ over all 9! orderings.

The three large Qwen models, Granite-4.0-H-Tiny and Gemma-3-12B look up without leaking to the plain question (at most +4.6 points), and they have the five largest asymmetries in Table 2. Llama-3.1-8B and Qwen3-4B look up but commit both questions, so their error is small or negative: Llama-3.1-8B answers false to almost everything (82% of original plain questions committed), and Qwen3-4B already answers false to many plain questions. Qwen3-8B and the two Gemma-4 models barely take this lookup at the rule condition and err little (Qwen3-8B +5.5, both Gemma-4 models negative), and the Gemma-4 models answer unknown to 87 to 95% of both questions on these items. OLMo-2-32B and Gemma-3-27B look up a little (negated +7.0 and +5.8) and err little on these items (+3.1 and +3.4). Looking only at the rule condition leaves the asymmetries of Mistral-24B (+16.5) and Nemotron-Nano-9B-v2 (+29.1) unexplained: both look up the listed extension (Appendix C.7).

Without a chat template. Under the template-free prompt of the pretrained comparison (Appendix G.1), the same two quantities have rank correlation 0.74 over the eleven checkpoints that can be read (open circles in Figure 3a): six of the seven pretrained checkpoints and five of the

Table 23: The asymmetry is largest where the lookup is strong and stays on the negated question. Asymmetry: negated minus plain committed share on original items. Neg. lookup and plain follows: change after restoring only the rule, relative to the control rule; lookup effect: their difference. Others' fact: the same change for another entity's target fact relative to the control fact. Concl. → plain: rise of the plain committed share after restoring the rule with the property in its conclusion, relative to its control. Gap specificity: fall of the negated committed share under the target gap statement minus the fall under the control statement. true share: committed negated answers that are true after restoring the rule. Points on 327 open pairs (264 for others' fact, 297 for the conclusion edit).

Model	Asym.	Neg. lookup	Plain follows	Lookup effect	Others' fact neg. / plain	Concl. → plain	Gap specificity	true share
Qwen3-32B	+40.7	+51.7	+3.1	+48.6	+37.5 / +9.5	+45.5	+59.0	99%
Qwen3-14B	+36.1	+47.4	+0.6	+46.8	+33.0 / +28.8	+43.4	+43.7	96%
Qwen3.8-27B	+28.7	+45.9	+2.8	+43.1	+7.6 / −0.4	+38.7	+35.5	100%
Mistral-24B	+16.5	−2.8	+0.0	−2.8	+9.5 / +0.8	+3.0	+5.5	100%
Llama-3.1-8B	+14.4	+56.6	+35.8	+20.8	+54.9 / +36.7	+42.1	+56.6	60%
Qwen3-8B	+5.5	+0.3	−0.9	+1.2	+14.0 / +6.8	+28.6	+2.4	97%
Gemma-4-31B	−6.1	+0.0	−2.8	+2.8	+0.4 / −1.5	+11.1	+7.0	100%
Gemma-4-12B	−1.8	+1.2	+0.0	+1.2	+3.8 / +1.1	+33.0	+11.6	75%
Qwen3-4B	−4.6	+24.2	+28.4	−4.3	+54.9 / +47.0	+55.2	+6.4	35%
Five further models (Appendix A.5)								
Granite-4.0-H-Tiny	+44.6	+34.9	+3.7	+31.2	+53.4 / +40.9	+25.3	+11.6	93%
Gemma-3-12B	+37.3	+21.1	+4.6	+16.5	+42.4 / +14.0	+37.4	+32.7	92%
OLMo-2-32B	+3.1	+7.0	+0.9	+6.1	+9.8 / +4.9	+1.7	+1.2	100%
Gemma-3-27B	+3.4	+5.8	+1.8	+4.0	+7.6 / +6.8	+22.9	+8.9	92%
Nemotron-Nano-9B-v2	+29.1	+11.0	+4.0	+7.0	+36.4 / +17.4	+29.3	+7.6	81%

six instruction-tuned ones. Pretrained Llama-3.1-8B answers almost everything unknown, and instruction-tuned Gemma-4-12B almost never does, so neither can be read.

C.5 Position strata in unedited items

Without any edit, the error sits where the property sits: stratifying the 8,654 original open pairs by where the target property occurs reproduces the pattern of the edits (Table 24). The strata use only the theory, not the answers. The four pure strata are: only in a rule's condition (193 pairs), only in a rule's conclusion (274), only in facts about other entities (354) and absent (269), the same for all models. The remaining 7,564 pairs have the property in several positions. *Negated-only committed* is the share of pairs with the plain question unknown and the negated question committed (UD).

Table 24: The error sits where the property sits. Share (%) of open pairs with 95% intervals in four pure position strata: negated-only committed when the property occurs only in a rule's condition (in parentheses: share of these answers that are true); plain question answered true when it occurs only in a rule's conclusion (in parentheses: the same share in the condition stratum); negated-only committed when it occurs only in other entities' facts or not at all.

Model	Condition only neg.-only committed	Conclusion only plain true	Others' facts only neg.-only committed	Absent neg.-only committed
Qwen3-32B	52.8 [44.5, 60.8] (100%)	42.7 [35.7, 49.8] (3.6)	28.5 [23.3, 33.8]	5.6 [2.9, 8.7]
Qwen3-14B	45.6 [37.7, 53.6] (100%)	42.7 [35.7, 49.6] (7.3)	19.8 [15.1, 24.6]	6.3 [3.3, 9.8]
Qwen3.8-27B	32.1 [24.2, 40.2] (95%)	38.3 [31.9, 44.8] (4.1)	6.5 [3.9, 9.4]	3.0 [1.1, 5.4]
Mistral-24B	7.3 [2.6, 12.7] (93%)	2.2 [0.7, 4.1] (0.0)	16.7 [12.4, 21.2]	1.5 [0.0, 3.4]
Llama-3.1-8B	39.9 [31.4, 48.4] (69%)	5.8 [3.1, 8.9] (1.0)	30.5 [25.5, 35.8]	17.8 [12.8, 23.2]
Qwen3-8B	7.8 [2.7, 13.8] (100%)	19.7 [14.8, 24.9] (0.0)	8.8 [5.2, 12.7]	1.1 [0.0, 2.5]
Gemma-4-31B	6.2 [2.6, 10.3] (100%)	12.8 [8.6, 17.3] (0.5)	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]
Gemma-4-12B	2.1 [0.0, 5.8] (100%)	22.6 [16.9, 28.7] (0.0)	1.1 [0.3, 2.3]	0.0 [0.0, 0.0]
Qwen3-4B	1.0 [0.0, 2.6] (0%)	35.0 [28.7, 41.5] (1.6)	5.9 [3.5, 8.7]	0.4 [0.0, 1.2]

Condition. The four models that look up at the rule condition (Qwen3-32B, Qwen3-14B, Qwen3.8-27B, Llama-3.1-8B) commit only the negated question in 32–53% of pairs, and for the three large Qwen models 95–100% of these answers are true. The other five stay within 8%: Qwen3-8B, the Gemma-4 models and Mistral-24B do not look up at the rule condition, and Qwen3-4B's failure

falls on both questions and leaves no one-sided judgment. **Conclusion.** For seven models the plain question is answered true 12 to 39 points more often than in the condition stratum: the rule is read as an assertion. Llama-3.1-8B and Mistral-24B do not do this. **Other entities' facts.** Mistral-24B's only high value is in this stratum (16.7%, at most 7.3% elsewhere), consistent with its lookup at listed facts, and Qwen3-32B and 14B also reach 20–29%. **Absent.** Apart from Llama-3.1-8B (17.8%), which answers false to almost everything, all models stay within 6.3%: with nothing to look up there is no failure and no committed negated question. The causal structure shown by the edits appears in the unedited items with the same positions.

C.6 A WRITTEN NEGATION COUNTS AS FOUND

A negative fact written in the premises is taken as a found negation: adding one about the same subject raises the rate of negated true answers by 10 to 14 points more than a position-matched affirmative fact does, while the target stays unprovable (Table 25). We use the 6,715 valid pairs whose target keeps the gold label unknown after the edit. V0 is the original item. V1 adds a negative fact with the same subject and relation and a different object, V2 adds the position-matched affirmative fact, V3 adds a negative fact with a different subject, and V4 adds an unrelated negative fact. Qwen3-14B and 32B answer at zero temperature without thinking (67,150 answers per model), and V0 reproduces the main-table answers in 99.93% and 99.81% of cases.

Table 25: A written negative fact pushes the negated question toward true, and a matched affirmative fact does not. Change in the rate at which the negated question is answered true, relative to V0 (points), with the 95% interval of V1 − V2, and transitions of the negated question from unknown on V0 to true / false.

Model	V1	V2	V1 − V2	V1: to true / false	V2: to true / false
Qwen3-14B	+12.54	−1.28	+13.82 [+12.8, +14.9]	503 / 73	257 / 390
Qwen3-32B	+10.13	+0.04	+10.08 [+9.0, +11.1]	447 / 112	365 / 287

The transitions start from the pairs whose negated question is unknown on V0 (1,815 for Qwen3-14B, 1,470 for Qwen3-32B). Both V1 and V2 add committed answers (negated committed share +4.80 and +5.00 points under V1, +7.30 and +8.10 under V2), but in different directions. The negative fact pushes unknown mostly to true (503 against 73, 447 against 112), and the affirmative fact shows no such tilt (257 against 390, 365 against 287). The unrelated negative fact V4 moves few answers out of unknown (60 for Qwen3-14B, 124 for Qwen3-32B). The plain question also responds to the added sentence (V1 plain committed share +8.61 and +12.27 points), so what the written polarity determines is the direction of the answer. The different-subject variant V3 does not establish a stable same-subject specificity. A "not P" written into a rule's condition (Appendix C.3) is the same effect inside a rule.

C.7 A SECOND LOOKUP SITE: LISTED FACTS

Mistral-24B looks up the listed facts rather than rule conditions, and counting both lookup sites raises the rank correlation between lookup and error from 0.75 to 0.85 over the fourteen models (from 0.71 to 0.88 over the nine core models). Mistral-24B's asymmetry is +16.5 points (original negated 20% against plain 4%), yet it shows no lookup at the rule condition. Its lookup is at the listed extension. Restoring only another entity's target fact raises its negated question by +9.5 [+4.7, +14.3] points and its plain question by +0.8, a negated-specific +8.7, and 39 of its 40 committed negated answers are true. Another entity is listed with the property, the subject is not, and the failure is again reported only under negation. Nemotron-Nano-9B-v2 does the same: its lookup effect is +7.0 at the rule condition and +18.9 at the listed facts, with an asymmetry of +29.1. Table 26 combines the two sites. Over the fourteen models their sum has rank correlation 0.85 with the asymmetry; over the nine core models it has 0.88 (the listed-fact site alone 0.53).

Stratifying by whether the theory lists the target property in other entities' facts, 61 of Mistral-24B's 66 committed negated answers fall in the listed stratum (264 pairs; negated 61, plain 10, asymmetry +19.3), and the unlisted stratum (63 pairs) has only 5 (+4.8). The three large Qwen models have similar asymmetries in both strata (Qwen3-14B +34.8 and +41.3, Qwen3-32B +42.0 and +34.9, Qwen3.8-27B +27.3 and +34.9): they look up at the rule condition.

Table 26: Counting the listed facts as a second lookup site accounts for Mistral-24B and Nemotron-Nano-9B-v2. Lookup effect at the rule condition and at the listed facts (another entity's target fact relative to the control fact), their sum, and the asymmetry on original items (points, 95% intervals; 327 open pairs, 264 for the listed facts).

Model	Rule condition	Listed facts	Both sites	Asymmetry
Qwen3-32B	+48.6 [+41.4, +55.9]	+28.0 [+20.2, +36.0]	+76.7 [+65.3, +88.4]	+40.7
Qwen3-14B	+46.8 [+39.3, +54.1]	+4.2 [−2.3, +10.8]	+51.0 [+41.1, +61.2]	+36.1
Qwen3.8-27B	+43.1 [+36.1, +50.3]	+8.0 [+3.5, +12.7]	+51.1 [+43.2, +59.0]	+28.7
Mistral-24B	−2.8 [−6.1, +0.3]	+8.7 [+4.3, +13.1]	+6.0 [−0.3, +12.0]	+16.5
Llama-3.1-8B	+20.8 [+13.8, +27.8]	+18.2 [+9.4, +26.9]	+39.0 [+25.8, +52.0]	+14.4
Qwen3-8B	+1.2 [−2.6, +5.2]	+7.2 [+1.6, +12.7]	+8.4 [+1.4, +15.5]	+5.5
Gemma-4-31B	+2.8 [+0.9, +4.8]	+1.9 [−0.4, +4.5]	+4.6 [+1.4, +8.4]	−6.1
Gemma-4-12B	+1.2 [−0.3, +3.2]	+2.7 [+0.4, +5.3]	+3.9 [+1.0, +7.2]	−1.8
Qwen3-4B	−4.3 [−10.3, +1.6]	+8.0 [+0.4, +15.4]	+3.7 [−8.5, +15.7]	−4.6
Five further models (Appendix A.5)				
Granite-4.0-H-Tiny	+31.2 [+25.1, +37.2]	+12.5 [+3.6, +21.6]	+43.7 [+32.6, +55.2]	+44.6
Gemma-3-12B	+16.5 [+10.1, +23.4]	+28.4 [+19.7, +37.1]	+44.9 [+33.7, +56.6]	+37.3
OLMo-2-32B	+6.1 [+2.7, +10.0]	+4.9 [+0.3, +9.9]	+11.0 [+4.1, +18.7]	+3.1
Gemma-3-27B	+4.0 [+0.6, +7.4]	+0.8 [−3.7, +5.3]	+4.7 [−1.8, +11.6]	+3.4
Nemotron-Nano-9B-v2	+7.0 [+1.8, +12.6]	+18.9 [+11.2, +26.6]	+26.0 [+16.0, +36.1]	+29.1

C.8 Scale within one family

Within one Qwen3 release, the share of pairs with a plain unknown whose negated question is answered true rises monotonically with scale, 0.3%, 10.5%, 36.9% and 53.2% at 4B, 8B, 14B and 32B, but what grows is not one error. Table 27 arranges readouts from this and the previous appendices by scale, namely the rule items, the edited grid, the pretrained comparison, and thinking (without thinking at zero temperature, with thinking under each model's recommended sampling). Pretrained Qwen3-32B was not run.

Table 27: Within Qwen3, the negated-only error grows with scale, while both Gemma-4 sizes have little. The first four rows are on the 8,654 open rule pairs (patterns written plain/negated). Lookup effects and error are in points on the 327 edited open pairs. Thinking counts are committed negated questions of the 327 pairs, at zero temperature without thinking and sampled with thinking (the matched comparison on the full set is in Appendix F.6). –: not run.

	Qwen3-4B	Qwen3-8B	Qwen3-14B	Qwen3-32B	Gemma-4-12B	Gemma-4-31B
Not-X via negated only / plain only (pairs)	8 / 1,354	769 / 57	1,543 / 91	2,387 / 60	187 / 16	213 / 48
Negated true given plain unknown	0.3%	10.5%	36.9%	53.2%	2.4%	2.7%
Both unknown	30.4%	73.0%	22.7%	18.8%	86.2%	86.8%
Most common other pattern	F/F, 2,163	U/T, 769	F/T, 1,997	U/T, 2,387	T/U, 419	T/U, 420
Lookup effect, rule condition (instruction-tuned)	−4.3	+1.2	+46.8	+48.6	+1.2	+2.8
Lookup effect, pretrained	+23.2	−6.1	+25.4	–	+12.2	+11.3
Asymmetry on original items	−4.6	+5.5	+36.1	+40.7	−1.8	−6.1
Thinking: committed negated, off → on	231 → 26	90 → 7	275 → 10	282 → 53	38 → 2	18 → 0
Known accuracy	67.1%	45.4%	59.3%	62.3%	43.9%	54.3%

The negated-only error grows with scale in Qwen3. It goes from almost none at 4B to a true on more than half of the pairs with a plain unknown at 32B, and the error on the original grid items rises in the same order (−4.6, +5.5, +36.1, +40.7 points).

What grows is not one error. Qwen3-4B leans to false on both questions: 2,163 pairs are F/F and 1,354 are F/U, so its false does not depend on the form of the question. Qwen3-8B keeps both questions at unknown in 73% of pairs, but it also has the lowest known accuracy of the four sizes (45%): it leans to unknown overall. Qwen3-14B and 32B look up at the rule condition (lookup effects +47 and +49) and answer the negated question true when the lookup fails. When both of their questions are committed, the most common pattern is the closed-world row F/T (1,997 and 1,916 pairs).

Scale does not create the lookup. Pretrained Qwen3-4B already looks up (+23), pretrained 8B does not (−6) and pretrained 14B does (+25). The two instruction-tuned Gemma-4 models barely look

up (+1, +3) and do not err. What scale changes is the answer after the lookup: at 14B and 32B, the failure is reported only on the negated question.

Thinking fixes all four sizes (committed negated questions 231 → 26, 90 → 7, 275 → 10, 282 → 53), and the Gemma-4 models commit few to begin with.

The monotone trend is clean only on rule items. Negated-only assertions number 5, 54, 116 and 88 pairs on ANLI and 3, 18, 16 and 92 records for Qwen3-4B, 8B, 14B and 32B (Figure 2).

Scale does not decide the error in the further families either (Table 28). In Gemma-3 the negated-only error peaks at 12B, where the lookup is strongest (lookup effects of +16.5 points at the rule condition and +28 at the facts listed about other entities), and it is small at 4B and 27B, whose lookup effects are at most +6 at either place; the two Gemma-4 sizes of the next release barely look up. The small models do not show a smaller version of the same pattern. Gemma-3-4B leaves both questions at unknown in 77% of pairs, and Llama-3.2-3B answers false to both questions in 5,842 of the 8,654 pairs, commits 98% of both questions on the edited set and so has no asymmetry, although restoring the rule raises its negated question more than its plain one (+69 and +26 points). These two models are reported only here, not in the main table.

Table 28: In Gemma-3 the error follows the lookup, not the size. The same readouts as Table 27 for Gemma-3 and Llama; the pretrained versions were not run and the models have no thinking switch.

	Gemma-3-4B	Gemma-3-12B	Gemma-3-27B	Llama-3.2-3B	Llama-3.1-8B
Not-X via negated only / plain only (pairs)	416 / 18	2,115 / 21	791 / 46	28 / 901	300 / 47
Negated true given plain unknown	5.7%	42.4%	11.3%	5.2%	17.9%
Both unknown	77.4%	23.4%	70.3%	4.6%	5.6%
Most common other pattern	T/U, 1,034	U/T, 2,115	U/T, 791	F/F, 5,842	F/F, 5,203
Lookup effect, rule condition	+5.8	+16.5	+4.0	+43.1	+20.8
Asymmetry on original items	−2.8	+37.3	+3.4	−0.3	+14.4
Known accuracy	24.5%	59.4%	42.4%	48.8%	63.5%

D Inside the Model

The lookup is an attention read from the property word in the question to the mentions of the property in the text, and the state this read leaves in the question carries the error. The experiments use four models from three families, Qwen3-14B, Qwen3-32B, Llama-3.1-8B and Mistral-24B: of the five core models with the largest error on the original items, these are the four with standard full attention. All four run the same scripts and controls on the 773 pairs of the edit experiments in Section 4.1: 327 open pairs, 102 pairs whose answer is stated in the text (directly stated known) and 344 pairs whose answer follows through rules (derived known). Intervals are 95% bootstrap intervals clustered by theory.

D.1 Blocking the read of the property word

Blocking the property word's read of the text lowers committed negated answers in all four models, in Qwen3-32B as much as deleting the property (85% to 17% of pairs, against 86% to 17% for deletion), and the same block applied to "not" changes the count by at most one pair (Table 29). In every attention layer, and only in the forward pass that precedes the answer, the attention mask forbids the property word in the question ("Anne is not *quiet*.") to read any position of the target property in the text. The control block forbids reads of a matched unrelated property. Two further blocks apply the same masks to "not", which only the negated question contains. Decoding is greedy. Without a block, answers computed with an explicit attention mask match ordinary generation token for token in 1,510 to 1,530 of 1,546 inputs, and the counts differ from those of Section 4.1 by at most 3.

The plain question loses committed answers under the same block (Qwen3-14B −29, Qwen3-32B −35, Llama-3.1-8B −35, Mistral-24B −5 points), because its committed answers also read these positions. The read is also the route to ordinary facts: under the block, correct answers to directly stated known questions fall from 191–203 to 0–9 in three models and from 191 to 81 in Llama-3.1-8B, whereas blocking the matched unrelated property leaves known pairs intact. Both exits of the lookup pass through the read. In Qwen3-32B, Qwen3-14B and Mistral-24B the block mainly removes true answers, and in Llama-3.1-8B, whose committed negated answers are mostly false, it removes false

Table 29: Blocking the property word's read lowers committed negated answers, and blocking "not" does not. Committed negated answers on the 327 open pairs without a block and under four blocks ($\not\to$: may not read), the target-minus-control difference in points with 95% interval, the split of committed negated answers into true/false, and the 204 questions of directly stated known pairs answered correctly.

		Property word $\not\to$		"not" $\not\to$		Target − control (points)		true/false	Known correct
Model	None	target	control	target	control	property word	"not"	none → prop. word	none → prop. word
Qwen3-32B	279	**54**	281	280	281	-69.4 [$-75.4, -62.9$]	-0.3 [$-1.9, +1.2$]	194/85 → 54/0	203 → 3
Qwen3-14B	272	**110**	270	272	271	-48.9 [$-55.6, -42.2$]	$+0.3$ [$+0.0, +1.0$]	189/83 → 98/12	203 → 9
Mistral-24B	69	**27**	70	68	68	-13.1 [$-19.0, -7.4$]	$+0.0$ [$-1.2, +1.2$]	66/3 → 27/0	195 → 0
Llama-3.1-8B	315	**232**	316	316	317	-25.7 [$-31.0, -20.6$]	-0.3 [$-1.3, +0.7$]	42/273 → 126/106	191 → 81

Table 30: Cutting every read of the target property restores unknown on both questions as often as the full property-word block; either partial cut does much less. Qwen3-14B, 90 open pairs in their original texts: pairs answered unknown on both questions, pairs with the plain question unknown and the negated question committed, the target-minus-control difference in both-unknown pairs (points, simultaneous 95% interval over six comparisons), and known questions answered correctly (of 96). $\not\to$: may not read.

Operation	Both unknown	Only negated committed	Target − control (points)	Known correct
No block (explicit causal mask)	29	31		52
"not" $\not\to$ target	29	31		52
Property word $\not\to$ target	67	16		6
Question $\not\to$ target in rule conditions	30	33	0 [$-8.9, +8.9$]	
Question $\not\to$ target in facts and conclusions	41	34	$+13.3$ [$+4.4, +22.2$]	
Question $\not\to$ target everywhere (both cuts)	**68**	16		5

answers. The false exit, in which another entity's property is taken as found, depends on this read even more than the true exit: in all four models the false answers fall more completely, and in Llama-3.1-8B part of them turn into true (42 to 126).

D.2 The lookup reads every mention

The lookup reads every mention of the target property, and any one mention suffices: in Qwen3-14B, cutting the reads at rule conditions alone or at facts and conclusions alone restores unknown on both questions in 30 and 41 of 90 pairs, and only cutting all of them reaches the 68 of the full property-word block (Table 30). This experiment splits the attention blocks of Appendix D.1 by the receiving position in the question and by the role of the mention in the text. Qwen3-14B answers without thinking (BF16, greedy, up to 2,048 new tokens). The inputs are 90 open pairs, each in its original text and in a version whose target property occurs only in rule conditions, asked with both questions, and 48 known pairs, 456 inputs in all. Eight operations give 3,648 answers. Every block acts in all 40 attention layers, and each control blocks the same number of other positions. The table reports the original texts.

Blocking "not" leaves every count at its baseline. Blocking the property word returns 25 of the 31 committed negated answers to unknown. The two partial cuts together do more than their sum. Adding the changes that each partial cut makes to the answer scores predicts that 46 of the 90 negated questions return to unknown, and the joint cut returns 68. The cuts also remove the use of evidence on known pairs: cutting facts and conclusions lowers the correct known answers from 34 to 19 of 48 negated questions and from 18 to 8 of 48 plain questions, and the full cut leaves 5 of 96 correct. The blocks therefore locate the route and do not repair the error. Because "not" precedes the property word, a direct read of zero by "not" still allows "not" to act through the property word's state.

D.3 Transplanting the lookup's state

The state that the lookup leaves at the question recreates the committed negated answer in a text without the property, at early and middle depths and not at 0.8 (Table 31). The receiver is the text

Table 31: The lookup's state recreates the error at early and middle depths and not at depth 0.8. Change in committed negated answers (points, 95% interval) when the donor state replaces the question's state in the text without the property, relative to moving the state the same radius along an unrelated direction. Bold: peak. Last two columns: the plain question at the peak, and the negated effect at the peak relative to a random direction of the same radius. 327 open pairs, 264 for Mistral-24B.

Model	Donor	0.2	0.4	0.6	0.8	Plain at peak	Peak vs. random
Qwen3-32B	rule condition	+4.3 [+1.8, +7.1]	+6.4 [+3.0, +10.2]	**+22.6** [+17.0, +28.3]	+0.6 [−0.6, +1.9]	+1.2	+22.0 [+16.3, +27.7]
Qwen3-14B	rule condition	+4.3 [+0.9, +7.8]	**+33.9** [+27.8, +40.1]	+10.4 [+5.2, +15.8]	−0.3 [−0.9, +0.0]	+1.8	+37.3 [+31.1, +43.5]
Mistral-24B	facts about others	+5.7 [+1.5, +10.1]	**+11.4** [+5.7, +16.9]	+0.0 [−1.1, +1.1]	+0.0 [−1.1, +1.1]	−0.8	+16.3 [+11.2, +21.6]
Llama-3.1-8B	rule condition	**+38.2** [+31.4, +45.0]	+30.0 [+22.4, +37.4]	−1.8 [−3.8, +0.0]	−0.9 [−2.2, +0.3]	+21.7	+35.8 [+28.7, +42.9]

from which the target and control properties have been removed. At one of four relative depths (0.2, 0.4, 0.6, 0.8), the residual state at every token of the question is replaced by the state at the same positions of the donor text, in which only the rule condition mentioning the property has been restored. Each question receives the state of its own question. The controls move the receiver's state by the same radius along an unrelated direction and along a random direction. Mistral-24B looks up the facts listed about other entities rather than rule conditions (Appendix C), so its donor is the text in which only those facts are restored, on the 264 pairs that have them. Writing back each model's own state reproduces all 6,184 answers token for token.

The peaks lie at early and middle depths (Llama-3.1-8B 0.2 to 0.4, Qwen3-14B 0.4, Mistral-24B 0.4, Qwen3-32B 0.6), and at depth 0.8 no model moves: the question takes up the lookup in the middle layers. The transplant recovers about four to seven tenths of the effect of restoring the rule itself (Section 5.1). The plain question, which receives its own state, barely moves, except in Llama-3.1-8B (+21.7), which commits both questions after a failed lookup (Section 4.1). Mistral-24B's negated question does not rise with the state from the restored rule condition (at most -4.3 points) and rises with the state from the restored facts, which is where it looks up. Qwen3-14B, Qwen3-32B and Llama-3.1-8B move with both donors, reaching peaks of +33.7, +22.3 and +37.1 points with the facts donor.

D.4 The step from failure to TRUE is spread over many heads

Replacing the output of selected attention heads at the answer position turns negated true answers on open items into unknown and leaves the other answers in place, but no small set of heads does it (Table 32). The items are the full test set with the prompts of Table 2, split by theory into two halves, and the readout is the most likely of the three label words at the first answer token, as in Appendix E.3. Five kinds of item are read: open pairs whose plain question the model answers unknown and negated question true (the target), negated questions of open pairs answered unknown on both questions, plain questions of open pairs answered unknown, known questions, and directly stated known questions.

On the first half, every head is scored by $(\bar{x} - x) \cdot \partial M/\partial x$ at the answer position, where x is the head's output, $\bar{x}$ its mean over the scored items and M the margin of true over unknown for the target and of the correct or unknown answer for the other kinds. Heads are ranked by the predicted fall of the target margin among heads predicted not to move the other kinds, and prefixes of the ranking are checked by replacing the heads' outputs with their means. We first limited the search to 32 heads and then extended the same ranking up to 512 heads, choosing on the first half the largest effect that keeps known and directly stated accuracy within 2 points and the two kinds of unknown answer within 3 points. The second half is used only for the numbers in the table. Qwen3-14B has 1,600 heads and Qwen3-32B 4,096; 90% of Qwen3-32B's 512 selected heads lie in the second half of the network.

The effect grows with the number of heads and levels off: on the first half, the target stays true in 76, 64, 57 and 55% of items for 32, 128, 256 and 512 heads in Qwen3-14B, and in 78, 62, 52 and 48% in Qwen3-32B, while the other kinds of answer stay within the limits above. It is not confined to a few layers either. Replacing the whole attention output or the whole MLP output at the answer position in any window of four consecutive layers either leaves the target in place or also moves the plain question: the strongest window, attention in layers 24 to 27 of Qwen3-14B, lowers the target by 34 points but also turns 12% of the plain unknown answers, and in Qwen3-32B no window lowers

Table 32: Selected heads turn negated true into unknown and leave the other answers; random heads do not. Held-out half of the theories. Rates (%): target negated questions still answered true; open negated and open plain questions still answered unknown; accuracy on known and on directly stated known questions. In brackets: change against no intervention (points, 95% theory-cluster interval). Random: range over 10 sets of 32 heads and 5 sets of 512 heads. "not" blocked: the answer position may not attend to "not" in any layer.

Model	Intervention	Target still true	Open neg. unknown	Open plain unknown	Known correct	Direct correct
Qwen3-14B	none	97.7	99.1	99.8	62.2	98.0
	32 selected heads	79.4 [−21.1, −15.6]	99.6	99.4	61.6	97.7
	32 random heads	93.1–99.5	94.0–99.3	93.9–99.8	60.8–62.7	97.8–98.2
	512 selected heads	58.3 [−43.0, −35.8]	99.8	99.5	59.3	97.5
	512 random heads	79.7–98.0	78.6–95.9	77.0–91.8	60.3–63.4	97.8–98.2
	"not" blocked	97.1	99.3	99.8	61.8	98.0
Qwen3-32B	none	97.9	98.1	98.6	67.3	97.0
	32 selected heads	78.0 [−22.3, −17.5]	99.8	97.8	67.5	97.0
	32 random heads	96.2–98.8	97.0–99.6	97.6–99.3	66.7–67.3	97.0–97.2
	512 selected heads	50.2 [−50.9, −44.6]	99.8	98.4	65.2	96.7
	512 random heads	98.0–99.9	72.0–93.4	90.1–99.3	64.0–66.5	96.5–97.2
	"not" blocked	98.1	98.4	98.6	66.8	97.0

the target by more than 13 points. The answer position does not need to read "not" itself, just as "not" does not need to read the text (Appendix D.1): what separates the two readings is computed from the question's own tokens and spread over many heads.

E NOT PROVEN

When the model commits the negated question, its state and its own words still say that X is not established. The state experiments use the four models of Appendix D. The statements about the premise use the edited set of Section 4.1: Tables 37, 38 and 39 give all fourteen models, and Table 40 the nine core models.

E.1 THE GAP IS STILL REPRESENTED AT THE ANSWER

At the position from which the committed answer is spoken, a linear probe separates open from known items as well on the negated question as on the plain one, and a second probe at the same position already reads the committed answer (Table 33). The probe reads the last position before the answer and is trained and tested on disjoint theories. It separates open pairs whose negated question is committed from correctly answered known pairs. On the same pairs we compare its recall of open items on the negated and the plain question, and a second probe at the same position reads the committed answer that the model gives next.

Table 33: At the moment of a committed negated answer, the gap is as readable as on the plain question, and the answer is already encoded. AUROC of a linear probe separating open from known items at the last position before the answer (negated question), its recall of open items on the negated and the plain question, and the accuracy of a probe reading the committed answer at the same position.

Model	AUROC, open vs. known	Recall of the gap: negated / plain	Committed answer read out
Qwen3-32B	0.822	0.72 / 0.75	92%
Qwen3-14B	0.816	0.70 / 0.72	94%
Mistral-24B	0.851	0.73 / 0.73	95%
Llama-3.1-8B	0.816	0.71 / 0.68	90%

The model states the negation with full commitment while the gap and the committed answer are represented at the same position.

Table 34: With no reasoning step, the negated hypothesis of open items is already judged *entailment* while the state separates open from determined items almost perfectly. Six models, answer words *entailment/neutral/contradiction*, rates as shares of open items.

Readout	Value
Depth 0, negated hypothesis of open items judged *entailment*	Llama-3.1-8B 0.42, Qwen3-32B 0.36, Gemma-3-12B 0.47 (Qwen3-8B 0.15, Qwen3-14B 0.22, Gemma-3-27B 0.12)
Probe at depth 0, open vs. determined, AUC	≥ 0.98 (0.98 to 1.00 for all six)
Depth 5, negated hypothesis of open items judged *entailment*: internal "not" → "It is not the case that" → "It is false that"	Llama-3.1-8B 0.53 → 0.61 → 0.78, Qwen3-14B 0.41 → 0.66 → 0.78, Qwen3-32B 0.65 → 0.72 → 0.69, Gemma-3-12B 0.65 → 0.82 → 0.77
Probe macro-AUC at depths 0, 1, 2, 3, 5 (Qwen3-32B)	1.00 → 0.95 → 0.89 → 0.90 → 0.85

Table 35: Pushing along u moves both questions toward unknown, and pushing against it moves both toward true. Share of open items answered unknown (%) after adding multiples of the open-versus-known direction u, or a random direction of the same norm, at depths 0.6 to 0.8. Last column: share of open items answered true at $-2u$.

Model	Question	$-2u$	$-u$	None	$+u$	Random	$+2u$	true at $-2u$
Qwen3-32B	negated	3.1	6.9	21.1	56.8	21.6	88.8	83.3%
	plain	18.1	31.8	52.4	74.2	53.7	90.3	57.1%
Qwen3-14B	negated	1.1	7.0	25.1	49.7	25.1	70.8	95.4%
	plain	4.1	27.4	48.7	65.5	49.1	78.1	89.6%
Mistral-24B	negated	1.8	31.2	74.8	92.6	74.8	99.9	98.2%
	plain	37.4	83.2	96.3	99.6	96.3	100.0	62.2%
Llama-3.1-8B	negated	0.0	1.8	6.1	12.1	7.0	25.1	93.8%
	plain	5.2	12.7	19.3	31.7	21.4	51.9	55.6%

E.2 The default is present at zero reasoning steps

The default needs no reasoning step: on ProofWriter items of depth 0, the negated hypothesis of open items is already judged *entailment* in 36 to 47% of items by three models, while a probe at the same depth separates open from determined items with AUC of at least 0.98 (Table 34). This experiment uses a different set of six models, Llama-3.1-8B, Qwen3-8B, 14B and 32B, and Gemma-3-12B and 27B (Gemma Team, Google DeepMind, 2025), with forward passes only, under an NLI-label six-shot interface whose answer words are *entailment*, *neutral* and *contradiction*. The items are 216 open-world ProofWriter pairs at each depth from 0 to 3 and 432 at depth 5. The probe is a linear probe within each cell, read on the plain question. Writing the negation as a sentential wrapper strengthens the default. Its strength differs by model: at depth 0 the shift in the response criterion from the plain to the negated question is +0.58 for Llama-3.1-8B, +0.61 for Qwen3-14B, +0.03 for Qwen3-32B and -0.09 for Gemma-3-27B.

E.3 One direction serves both questions

Pushing along the direction that separates open from known items moves both questions toward unknown, and pushing against it moves both toward true (Table 35). The direction u is the mean activation of open items minus that of known items on the plain questions, computed from each model's own activations. It is added at the answer position in every block between depths 0.6 and 0.8 with strength 0.02 times the layer's mean norm, in multiples from $-2u$ to $+2u$. The control is a random direction of the same norm. Each question is run on all 16,359 items and read at its first token.

Pushing along u also raises unknown on the plain question (at $+u$ relative to no push: Qwen3-14B $+16.8$ [$+15.9, +17.6$], Qwen3-32B $+21.7$ [$+20.9, +22.6$], Llama-3.1-8B $+12.4$ [$+11.6, +13.2$], Mistral-24B $+3.3$ [$+2.9, +3.7$] points, the last already answering unknown to 96% of plain questions), and the random direction leaves both questions in place. Pushing against u removes unknown from both

Table 36: Directions from ANLI paragraphs and from records have positive cosine with the rule direction u and raise unknown on rule items about as much as u does. Cosine with u ("between": ANLI with records). Share of 2,000 open negated rule questions answered unknown and of 2,000 known ones answered correctly, without a push and after adding each direction at the same strength.

	Cosine with u				Open answered unknown (%)						Known correct (%)			
Model	ANLI	records	between	random	none	$+u$	$+$ANLI	$+$records	$-$ANLI	random	none	$+u$	$+$ANLI	$+$records
Qwen3-32B	0.43	0.62	0.57	0.00	20.1	55.5	58.2	66.2	5.6	19.3	60.1	54.8	52.0	48.9
Qwen3-14B	0.69	0.84	0.77	0.00	25.6	50.5	58.9	56.8	5.5	25.7	59.0	63.9	54.4	58.8
Mistral-24B	0.70	0.87	0.76	0.00	75.6	93.2	90.8	94.5	41.7	75.3	22.6	16.8	14.5	15.5
Llama-3.1-8B	0.66	0.79	0.81	0.00	6.5	12.8	36.2	25.4	0.7	6.3	91.3	94.1	81.2	88.1

questions step by step, and both fall toward true: on the plain question true asserts X, and on the negated question it asserts not-X. At the far end of this direction the two questions agree on true, so the direction does not encode not-X: it is a "not proven" signal shared by both questions. Negation does not erase it, since it is still readable when the negated question is about to be answered true, and the negated question's reading of a failed lookup overrides it. The push is not free. At $-2u$ accuracy on known negated questions falls to 6.8 to 50.8%, while known plain questions rise to 83 to 96%. At $+2u$ known plain questions fall to 0.3 to 35%, and known negated questions fall for Qwen3-32B (61.9 to 37.0%), Mistral-24B (22.6 to 9.0%) and Llama-3.1-8B (91.3 to 86.1%).

E.4 One direction across three kinds of text

Directions computed in the same way on human-written paragraphs and on records have cosine 0.43 to 0.87 with the rule direction u and act like it on rule items (Table 36). Each direction follows the definition of u: the plain question, the last prompt position, and the mean activation of open minus known items. The ANLI direction contrasts 1,144 neutral hypotheses with 400 entailed or contradicted ones, and the record direction contrasts 1,200 open with 480 known record items. We measure each direction's cosine with u in the injection window (depths 0.6 to 0.8) and inject it at the same strength into negated rule questions, 2,000 open and 2,000 known ones sampled with a fixed seed, reading the first token.

Relative to no push, unknown on open items rises by the following amounts (points, 95% interval). Qwen3-32B: ANLI direction $+38.1$ $[+35.9, +40.4]$, record direction $+46.2$ $[+43.8, +48.5]$, u $+35.4$ $[+33.2, +37.7]$. Qwen3-14B: $+33.2$ $[+31.2, +35.3]$, $+31.2$ $[+29.1, +33.2]$, $+24.9$ $[+23.0, +26.9]$. Mistral-24B: $+15.2$ $[+13.6, +16.9]$, $+18.9$ $[+17.1, +20.7]$, $+17.5$ $[+15.8, +19.3]$. Llama-3.1-8B: $+29.8$ $[+27.5, +32.0]$, $+18.9$ $[+17.1, +20.9]$, $+6.3$ $[+5.2, +7.5]$. The random direction stays within ± 1 point, and pushing against the ANLI direction lowers unknown in all four models. Accuracy on known items falls by 0 to 11 points under the ANLI and record directions (-6 to $+5$ under u at the same strength). Mistral-24B answers only 22.6% of known negated questions correctly without any push.

E.5 What the model says about the premise

Asked about the premise instead of about X, every model accepts that the premise does not establish X, and several of them also accept that it establishes not-X (Table 37). We replace the hypothesis by one of three statements: "The premise does not establish that X" (true on open items), "The premise establishes that x is not P" (false) and "The premise establishes that X" (false), with the prompt and readout of the edited set. On directly stated known items the three statements are almost always judged correctly (98 to 100% for Qwen3-14B, Qwen3-32B, Qwen3.8-27B and both Gemma-4 models). On open items all fourteen models accept "does not establish X" in 67% of items or more, and yet Qwen3-32B says "establishes not-X" for 214 items, Qwen3-14B for 187, Gemma-3-12B for 178, Granite-4.0-H-Tiny for 146, Mistral-24B for 111, Qwen3.8-27B for 65 and Llama-3.1-8B for 61.

E.6 The wrong account follows the lookup

The statement "establishes not-X" follows the same lookup as the negated answer: restoring only the rule that mentions the property raises it by 15 to 64 points in the seven models that look up at the rule condition and moves it by at most 6.4 points in the other seven (Table 38). The three statements are asked on the original text, on the text with the property removed and on the text in which only that

Table 37: On open items the models accept that the premise does not establish X, and several also accept that it establishes not-X. Answers true/false/unknown to three statements about the premise, 327 open pairs.

Model	"does not establish X" (true)	"establishes not-X" (false)	"establishes X" (false)
Qwen3-32B	298/29/0	**214**/75/38	57/178/92
Qwen3-14B	311/16/0	**187**/101/39	56/213/58
Qwen3.8-27B	321/6/0	65/94/168	100/110/117
Mistral-24B	309/2/16	111/5/211	10/1/316
Llama-3.1-8B	319/8/0	61/265/1	47/280/0
Qwen3-8B	220/0/107	20/69/238	54/47/226
Gemma-4-31B	307/20/0	0/14/313	42/32/253
Gemma-4-12B	318/9/0	20/31/276	51/14/262
Qwen3-4B	251/67/9	51/244/32	72/233/22
Five further models (Appendix A.5)			
Granite-4.0-H-Tiny	325/1/1	146/131/50	50/170/107
Gemma-3-12B	265/8/54	178/106/43	73/166/88
OLMo-2-32B	305/0/22	22/35/270	30/12/285
Gemma-3-27B	315/12/0	38/76/213	37/55/235
Nemotron-Nano-9B-v2	317/9/1	6/281/40	15/287/25

Table 38: The statement "the premise establishes not-X" rises when only the rule mentioning the property is restored, in exactly the models that look up there. Counts of "establishes not-X" judged true on three versions of the text, and the change from removed to rule restored (points, 95% interval). Last column: of the items where it is judged true on the rule-restored text, those whose negated question is also answered true. 327 open pairs.

	"does not establish"	"establishes not-X" judged true				Negated
Model	true (original)	original	removed	rule restored	Restored − removed	also true
Qwen3-32B	298/327	213	38	247	$+63.9$ [$+57.3, +70.3$]	214/247
Qwen3-14B	311/327	187	73	237	$+50.2$ [$+42.8, +57.4$]	225/237
Qwen3.8-27B	321/327	65	6	123	$+35.8$ [$+29.1, +42.9$]	121/123
Mistral-24B	309/327	110	38	30	-2.4 [$-6.5, +1.5$]	11/30
Llama-3.1-8B	318/327	62	42	170	$+39.1$ [$+32.3, +46.2$]	147/170
Qwen3-8B	220/327	20	36	29	-2.1 [$-5.2, +0.7$]	24/29
Gemma-4-31B	307/327	0	0	1	$+0.3$ [$+0.0, +1.0$]	1/1
Gemma-4-12B	319/327	19	0	2	$+0.6$ [$+0.0, +2.0$]	2/2
Qwen3-4B	251/327	51	60	109	$+15.0$ [$+10.1, +20.1$]	42/109
Five further models (Appendix A.5)						
Granite-4.0-H-Tiny	325/327	145	76	139	$+19.3$ [$+13.7, +24.8$]	133/139
Gemma-3-12B	264/327	176	4	142	$+42.2$ [$+35.2, +49.5$]	101/142
OLMo-2-32B	305/327	22	1	19	$+5.5$ [$+3.1, +8.4$]	18/19
Gemma-3-27B	316/327	39	23	36	$+4.0$ [$+1.2, +7.1$]	27/36
Nemotron-Nano-9B-v2	317/327	6	28	7	-6.4 [$-9.4, -3.8$]	7/7

rule is restored, in a run of their own whose counts on the original text differ from Table 37 by at most 2.

The seven models that look up at the rule condition are the three large Qwen models, Granite-4.0-H-Tiny, Gemma-3-12B, Llama-3.1-8B and Qwen3-4B, the grouping of Section 4.2. Of the other seven, OLMo-2-32B and Gemma-3-27B, which look up a little, rise by 5.5 and 4.0 points, Nemotron-Nano-9B-v2 falls by 6.4, and Qwen3-8B, both Gemma-4 models and Mistral-24B stay within 2.4 points. Mistral-24B's statements of not-X sit on the original text (110 against 38 after removal), whose listed facts about other entities are where it looks up. The statements about the premise track the answers to the question item by item: of the items judged to establish not-X on the rule-restored text, the negated question is also answered true in 225 of 237 for Qwen3-14B, 121 of 123 for Qwen3.8-27B, 214 of 247 for Qwen3-32B, 133 of 139 for Granite-4.0-H-Tiny, 147 of 170 for Llama-3.1-8B and 101 of 142 for Gemma-3-12B. With thinking enabled, Qwen3-14B's count of "establishes not-X" falls from 187 to 16.

Table 39: Models that say "the premise establishes not-X" almost always also say "the premise leaves open whether X". Answers true/false/unknown to "leaves open" on 327 open pairs, directly stated known questions on which it is correctly judged false (of 102), and items judged to establish not-X, and how many of them are also judged left open.

Model	"Leaves open", open items	Known: judged false	"Establishes not-X"	Also "leaves open"
Qwen3-32B	313/14/0	100/102	214	**213**
Qwen3-14B	309/9/9	95/102	187	**187**
Qwen3.8-27B	310/17/0	101/102	65	62
Mistral-24B	315/4/8	89/102	111	109
Llama-3.1-8B	52/2/273	2/102	61	13
Qwen3-8B	0/0/327	5/102	20	0
Gemma-4-31B	306/21/0	102/102	0	0
Gemma-4-12B	318/9/0	100/102	20	20
Qwen3-4B	15/53/259	85/102	51	9
Five further models (Appendix A.5)				
Granite-4.0-H-Tiny	152/0/175	0/102	146	60
Gemma-3-12B	265/0/62	44/102	178	161
OLMo-2-32B	0/0/327	8/102	22	0
Gemma-3-27B	292/3/32	57/102	38	38
Nemotron-Nano-9B-v2	136/23/168	83/102	6	1

E.7 Open and established at once

On the same open item, the models that can read the statement say both that the premise leaves X open and that it establishes not-X (Table 39). The statement that contradicts "establishes not-X" is "The premise leaves open whether X", true on open items and false on directly stated known items. We add it to the run on the original text, with the same prompt and readout.

The models that judge the statement false on directly stated known items (87 to 100%: Qwen3-32B, Qwen3-14B, Qwen3.8-27B, both Gemma-4 models and Mistral-24B) nearly always say "leaves open" wherever they say "establishes not-X": 213 of 214 for Qwen3-32B, 187 of 187 for Qwen3-14B, 62 of 65 for Qwen3.8-27B, 109 of 111 for Mistral-24B and 20 of 20 for Gemma-4-12B. Llama-3.1-8B and Qwen3-8B almost never judge the statement false on directly stated known items (2% and 5%), so the statement is unreadable for them. Qwen3-4B answers unknown on most open items (259) and says "leaves open" on only 15. Of the five further models, Gemma-3-12B and Gemma-3-27B say "leaves open" on 161 of their 178 and on all 38 of their items judged to establish not-X, but judge the statement false on only 44 and 57 of the 102 directly stated known items. Granite-4.0-H-Tiny and OLMo-2-32B judge it false on 0 and 8 of them, so the statement is unreadable for them, and Nemotron-Nano-9B-v2 says "establishes not-X" on only 6 items.

E.8 "Establishes" means "written"

On known items whose X takes rule applications to derive, the models that read these statements still say the premise does not establish X and leaves it open, including items where they have just derived X themselves (Table 40). We ask the four statements on the 344 derived known pairs, where X is provable, with the prompt and readout of the edited set. The correct judgments are false for "does not establish X", false for "leaves open", true for "establishes X" and false for "establishes not-X".

Qwen3-32B says "leaves open" on 155 of the 213 items whose plain question it has just answered true, Qwen3-14B on 130 of 160 and Qwen3.8-27B on 176 of 270. In these models' words, "establishes" means written and found. Not finding X then becomes "establishes not-X": on items where X is in fact derivable, Qwen3-32B says so for 168 items, Qwen3-14B for 151 and Mistral-24B for 159, and most of these are also judged left open. Only on directly stated known items are the statements almost always right. Llama-3.1-8B, Qwen3-4B and Qwen3-8B answer "leaves open" mostly with unknown or false, so that column is unreadable for them.

Table 40: On derived known items the models still say the premise leaves X open, even where they derive X themselves. Statements judged true on 344 derived known pairs. "Leaves open", in parentheses: items judged left open on which the model answered the plain question true / all items on which it did. Last column: of the items judged to establish not-X, those also judged left open.

Model	"Does not establish X"	"Leaves open"	"Establishes X"	"Establishes not-X"	Also "leaves open"
Qwen3-32B	254	285 (155/213)	159	**168**	164
Qwen3-14B	265	302 (130/160)	128	**151**	146
Qwen3.8-27B	282	249 (176/270)	261	55	53
Mistral-24B	309	322 (40/48)	41	**159**	152
Llama-3.1-8B	331	71 (35/82)	66	40	5
Qwen3-8B	181	0 (0/172)	125	21	0
Gemma-4-31B	231	225 (103/220)	204	1	1
Gemma-4-12B	302	308 (99/128)	122	47	46
Qwen3-4B	180	16 (4/196)	177	31	5

Table 41: A stated gap is kept as well as an unwritten one, and only a gap that fails through a rule collapses under negation; Llama-3.1-8B reads $\neg r$ as textbook negation as failure in all three kinds. Share of $\neg r$ questions answered unknown for the three kinds of gap atom r, the difference between underivable and stated (points, 95% interval), and the committed answers to $\neg r$ in the underivable case.

Model	Stated	Unwritten	Underivable	Underivable − stated	true / false (underivable)
Qwen3-32B	100%	100%	84%	-15.6 $[-21.8, -10.1]$	8% / 8%
Qwen3-14B	100%	99%	68%	-32.5 $[-41.1, -24.9]$	7% / 26%
Qwen3.8-27B	100%	100%	99%	-0.6 $[-2.0, +0.0]$	1% / 0%
Mistral-24B	100%	99%	78%	-22.5 $[-29.9, -15.9]$	22% / 0%
Qwen3-8B	96%	98%	64%	-31.5 $[-39.6, -23.4]$	1% / 34%
Gemma-4-31B	100%	100%	100%	$+0.0$ $[+0.0, +0.0]$	0% / 0%
Gemma-4-12B	100%	100%	100%	$+0.0$ $[+0.0, +0.0]$	0% / 0%
Qwen3-4B	86%	88%	16%	-70.2 $[-78.4, -61.7]$	0% / 84%
Llama-3.1-8B	20%	51%	19%	-1.3 $[-11.2, +8.4]$	57% / 22%

E.9 A STATED GAP IS RESPECTED

A gap that the text states is respected, and only a gap that the model must find by itself turns into the opposite under negation. We append "Whether x is P is left open." to the original text. Relative to the original, committed negated answers fall by 84.4 points for Qwen3-32B, 71.3 for Qwen3-14B, 95.4 for Llama-3.1-8B and 61.8 for Qwen3.8-27B, while the same sentence about an equally open unrelated property lowers them by only 25 to 39 points. The statement about the target lowers them more than the statement about the unrelated property by 59.0 points for Qwen3-32B, 43.7 for Qwen3-14B, 35.5 for Qwen3.8-27B and 56.6 for Llama-3.1-8B. Qwen3-4B and Qwen3-8B abstain after any such statement, and the statement about the target does no more than the one about the unrelated property, and both Gemma-4 models rarely commit the negated question to begin with. A second wording, "It is not stated whether x is P.", is reported with the other edits in Appendix C. On known items the appended statement is false, and the models follow it and give up the proof: given information overrides derived information. The statement is therefore a diagnosis and not a repair.

Formal worlds with three atoms show which gaps collapse (Table 41). The worlds are generated by a program, and the gap atom r comes in three kinds: its truth value is stated to be not given (stated), r does not occur among the facts at all (unwritten), or r could only come from a rule "if t then r" whose condition is false (underivable). Nine models are tested on the negation $\neg r$.

A stated gap is kept as well as an unwritten one (86 to 100%) in every model but Llama-3.1-8B, which is the purest case of the default: it keeps r itself unknown (all 120 stated gaps) and judges $\neg r$ true in all three kinds of gap, even when the gap is stated (79%) and even when r occurs nowhere (44%), as textbook negation as failure does. Only the gap that fails through a rule collapses under negation, to 16% for Qwen3-4B, 64% for Qwen3-8B, 68% for Qwen3-14B, 84% for Qwen3-32B and

Table 42: Saying that the record about x is incomplete lowers committed negated answers only 10 to 20 points more than saying it about another entity, and costs accuracy on derived known items. Committed negated and plain answers (%) on 327 open pairs for the original text, with the sentence about the subject x, and with the same sentence about another entity, the difference between the last two for the negated question (points, 95% interval), and accuracy (%) on directly stated and derived known items.

Model	Negated committed orig. / about x / about other	Plain committed orig. / about x / about other	Negated: about x − about other	Direct known orig. / about x	Derived known orig. / about x
Qwen3-32B	86.2 / 66.1 / 77.4	45.6 / 38.8 / 39.8	-11.3 [$-16.1, -6.9$]	99.5 / 97.5	55.4 / 51.2
Qwen3-14B	84.1 / 65.4 / 78.6	48.0 / 37.0 / 42.8	-13.1 [$-17.6, -9.0$]	99.5 / 100.0	46.8 / 45.6
Qwen3.8-27B	62.7 / 18.7 / 29.1	33.9 / 17.7 / 20.2	-10.4 [$-15.6, -6.0$]	100.0 / 99.0	58.0 / 43.6
Mistral-24B	20.2 / 5.2 / 17.4	3.7 / 1.5 / 3.1	-12.2 [$-16.3, -8.4$]	95.1 / 69.6	8.3 / 2.5
Llama-3.1-8B	96.6 / 76.1 / 96.3	82.3 / 48.0 / 83.8	-20.2 [$-26.2, -14.6$]	94.1 / 81.9	59.0 / 49.4
Qwen3-8B	27.5 / 11.0 / 11.0	22.0 / 14.7 / 11.6	$+0.0$ [$-2.8, +2.8$]	99.0 / 85.8	44.6 / 25.1
Gemma-4-31B	5.5 / 3.4 / 4.3	11.6 / 11.3 / 11.6	-0.9 [$-2.6, +0.6$]	100.0 / 100.0	50.1 / 42.4
Gemma-4-12B	11.6 / 8.3 / 5.2	13.5 / 12.8 / 9.2	$+3.1$ [$+0.6, +5.6$]	97.5 / 94.6	32.7 / 26.0
Qwen3-4B	70.6 / 43.7 / 54.7	75.2 / 39.8 / 58.1	-11.0 [$-16.0, -6.3$]	98.5 / 89.7	73.3 / 50.0

78% for Mistral-24B (underivable minus stated -70, -32, -33, -16 and -23 points, all intervals excluding zero). An unwritten atom is not an intermediate case: what fails is a derivation, not the mere absence of the atom. The direction of the collapse follows where the atom occurs. Here r occurs only in a rule's conclusion, and Qwen3-4B, 8B and 14B answer $\neg r$ mostly with false, treating the conclusion as an assertion as in the conclusion-position edit of Appendix C, while Mistral-24B answers true. Qwen3.8-27B and both Gemma-4 models keep all three kinds of gap in these small worlds.

E.10 Saying that the record is incomplete

Saying that the facts about the subject are incomplete lowers committed negated answers only partly, and lowers accuracy on known items as well (Table 42). We append "Only some of the facts about x are listed." to the original texts of the 773 pairs, where x is either the subject of the question or another entity of the same text, chosen per pair. Questions, the remaining sentences and gold labels are unchanged, and all nine models answer under the prompt of the edited set.

In the models that make the error, the sentence about the subject lowers committed negated answers by 10 to 20 points more than the same sentence about another entity, and in Qwen3-32B, Qwen3-14B and Llama-3.1-8B most negated questions remain committed (66%, 65% and 76%). On Anne, Qwen3-32B still answers unknown and true after the sentence. Accuracy on derived known items drops in every model. The model does not first infer that the record is complete and then conclude from it: being told that the record is incomplete leaves the lookup in place, and only naming the gap itself (Appendix E.9) makes the model keep it. Like the instruction of Appendix F.1, the sentence only makes the model answer less.

F Proof

Instructions lower commitment on both questions and cost accuracy on derived known items, whereas thinking replaces the lookup with a proof. Figure 6 summarizes both interventions.

F.1 An instruction makes all fourteen models abstain

Stating the open-world proof standard in the prompt moves both questions of all fourteen models toward unknown, and lowers accuracy on known items in every model except Gemma-4-31B (Table 43). Three more specific instructions lower accuracy on derived known items, averaged over both questions, in 30 of the 42 model–instruction combinations (27 on the plain question alone) (Appendix F.2). The final sentence of the prompt is replaced by the following paragraph, with examples, premises, hypotheses and decoding unchanged: *Use an open-world proof standard. Answer true only when the premise and rules prove the hypothesis. Answer false only when they prove the explicit negation of the hypothesis. If neither is proved, answer unknown. Missing facts, missing rules, and failure to derive a statement are not evidence for false. Do not infer false from*

absence. Reply with exactly one lowercase answer word: true, false, or unknown. All 16,359 pairs are asked with both questions at zero temperature. The baseline, run alongside with the original final sentence, lies within 0.3 points of Table 2.

Table 43: The instruction makes every model abstain on both questions and lowers known accuracy in thirteen of fourteen models. Net asymmetry on the 8,654 open pairs (pairs with only the negated question committed minus pairs with only the plain question committed, as points of 8,654), unknown rates of both questions, and accuracy on ProofWriter's own known questions, before and after the instruction.

Model	Net asymmetry (points)	Negated unknown	Plain unknown	Known accuracy
Qwen3-32B	$+31.22 \to +12.58$	$20.6\% \to 63.1\%$	$51.9\% \to 75.7\%$	$62.3\% \to 51.5\%$
Qwen3-14B	$+23.47 \to +23.54$	$24.8\% \to 60.3\%$	$48.3\% \to 83.9\%$	$59.3\% \to 40.8\%$
Qwen3.8-27B	$+23.58 \to -7.40$	$50.2\% \to 90.6\%$	$73.8\% \to 83.2\%$	$64.4\% \to 49.2\%$
Mistral-24B	$+21.33 \to +9.07$	$74.9\% \to 89.5\%$	$96.2\% \to 98.6\%$	$21.8\% \to 13.4\%$
Llama-3.1-8B	$+13.22 \to +0.21$	$6.2\% \to 95.1\%$	$19.4\% \to 95.3\%$	$63.5\% \to 12.5\%$
Qwen3-8B	$+5.69 \to -3.87$	$78.9\% \to 98.0\%$	$84.5\% \to 94.1\%$	$45.4\% \to 17.8\%$
Gemma-4-31B	$-2.36 \to -2.11$	$92.2\% \to 92.4\%$	$89.9\% \to 90.3\%$	$54.3\% \to 55.2\%$
Gemma-4-12B	$-2.08 \to -4.33$	$91.1\% \to 96.2\%$	$89.0\% \to 91.9\%$	$44.1\% \to 33.9\%$
Qwen3-4B	$-13.21 \to -9.34$	$48.0\% \to 86.3\%$	$34.8\% \to 77.0\%$	$67.1\% \to 43.4\%$
Five further models (Appendix A.5)				
Granite-4.0-H-Tiny	$+34.72 \to +9.93$	$23.0\% \to 70.7\%$	$57.7\% \to 80.7\%$	$43.7\% \to 24.2\%$
Gemma-3-12B	$+32.26 \to +2.03$	$25.4\% \to 78.9\%$	$57.7\% \to 80.9\%$	$59.5\% \to 31.1\%$
OLMo-2-32B	$+11.53 \to -0.72$	$77.8\% \to 99.2\%$	$89.3\% \to 98.5\%$	$26.3\% \to 8.1\%$
Gemma-3-27B	$+1.57 \to -2.78$	$79.0\% \to 86.0\%$	$80.6\% \to 83.2\%$	$42.4\% \to 36.7\%$
Nemotron-Nano-9B-v2	$+15.01 \to +0.72$	$57.2\% \to 96.0\%$	$72.2\% \to 96.7\%$	$46.2\% \to 13.7\%$

Every model responds by moving both questions toward unknown. The asymmetry either stays as it was (Qwen3-14B $+23.47$ to $+23.54$) or shrinks with the overall abstention (Qwen3-32B $+31$ to $+13$, Mistral-24B $+21$ to $+9$), vanishes (Llama-3.1-8B) or reverses (Qwen3.8-27B -7.4). Among the five further models it shrinks for Granite-4.0-H-Tiny ($+35$ to $+10$) and ends within 3 points of zero for the other four. Accuracy on known items falls in every model except Gemma-4-31B, which rarely commits in the first place, and Llama-3.1-8B falls from 63.5% to 12.5%.

F.2 Three more specific instructions

Over the fourteen models, three more specific instructions lower accuracy on the plain question of derived known items in 27 of the 42 model–instruction combinations, and averaged over both questions, the measure of Figure 6a, in 30 of 42 (Table 7 gives the plain-question changes of the five further models). In the nine core models (Table 44), they also lower commitment without making the models answer correctly: committed negated answers fall by up to 46 points, while accuracy on derived known items, averaged over both questions, falls in 23 of the 27 combinations and rises by at most 6 points. Each instruction is placed before the task of the edited set's original items, 327 open and 344 derived known pairs. The first defines the three answers: *Use true when the premise establishes the hypothesis, false when the premise establishes its negation, and unknown when neither follows.* The second adds a reminder that descriptions are not exhaustive: *Descriptions need not list every true fact. A property mentioned for one individual, by itself, neither confirms nor rules out that property for another individual. Decide the hypothesis from stated facts and valid rule applications, not from omissions.* The third adds a reminder about the scope of rules: *A rule of the form 'if A then B' establishes neither A nor not A by itself. A statement may occur in a rule without being established as a fact. Judge both the hypothesis and its negation from stated facts and justified rule applications.* With thinking enabled instead, the seven models with a thinking mode gain 18 to 65 points on derived known items.

For the difference between the negated and the plain question, 9 of the 27 intervals exclude zero on the negative side, so that the negated question retreats more than the plain one (Qwen3.8-27B in two cases, Llama-3.1-8B in one, Qwen3-8B and Mistral-24B in all three). Four exclude zero on the positive side (Llama-3.1-8B with the definition, $+13.1$, Qwen3-4B in two cases and Gemma-4-31B in one). The other 14 include zero, among them all three for Qwen3-14B and Qwen3-32B. The instructions lower answering, and none of them makes proof take the place of the lookup.

Table 44: Specific instructions mostly lower commitment on both questions and lower derived known accuracy in 23 of 27 cases. The nine core models. Change from the original prompt (points): committed negated / committed plain / negated minus plain, and in parentheses the mean accuracy on derived known items over both questions. Last column: derived known accuracy without and with thinking.

Model	Definition	+ not exhaustive	+ rule scope	Thinking: derived known
Qwen3-32B	−20 / −16 / −4.6 (−3.5)	−24 / −18 / −5.8 (−4.7)	−25 / −22 / −2.8 (−6.8)	55% → 94%
Qwen3-14B	−34 / −30 / −4.0 (−11.6)	−22 / −20 / −2.4 (−9.6)	−24 / −23 / −1.5 (−6.4)	47% → 92%
Qwen3.8-27B	−11 / +0 / −11.6 (+6.0)	−34 / −15 / −18.7 (−2.5)	−12 / −12 / +0.0 (+0.1)	58% → 99%
Mistral-24B	−8 / 0 / −7.6 (−1.2)	−6 / +1 / −7.3 (+0.9)	−13 / +1 / −13.5 (−0.1)	
Llama-3.1-8B	−8 / −21 / +13.1 (−7.0)	+1 / +12 / −10.7 (−2.8)	+2 / +1 / +1.2 (−8.1)	
Qwen3-8B	−9 / +2 / −10.4 (−5.8)	−15 / −2 / −12.8 (−12.9)	−16 / −1 / −14.4 (−15.4)	45% → 93%
Gemma-4-31B	+1 / +0 / +0.9 (−0.7)	+1 / −2 / +3.4 (+2.2)	−1 / −2 / +1.5 (−3.2)	50% → 99%
Gemma-4-12B	−2 / −2 / −0.3 (−1.6)	−2 / −2 / −0.3 (−1.7)	−4 / −2 / −2.4 (−4.5)	33% → 98%
Qwen3-4B	−43 / −48 / +4.9 (−24.9)	−46 / −52 / +6.4 (−31.2)	−45 / −51 / +6.4 (−25.6)	73% → 91%

Table 45: NLI answer words remove Qwen3.8-27B's asymmetry and cost its derived known items. Qwen3.8-27B: open pairs by the answers to the two questions (432 pairs) and known questions answered correctly (of 1,728).

Prompt	Both unknown	Only negated committed	Only plain committed	Both committed	Known correct
Original	270	**79**	13	70	861
Definition, truth words	263	109	4	56	757
Definition, NLI words	422	**6**	4	0	**377**

F.3 Other answer words

Replacing the answer words by NLI labels removes Qwen3.8-27B's asymmetry and, like the instructions, costs accuracy on items that need derivation (Table 45); Llama-3.1-8B keeps its asymmetry (97 pairs with only the negated question committed against 16). Three prompts are compared: the original prompt, a full definition with the truth words, and the same definition with the six examples' answers replaced one for one by *entailment*, *contradiction* and *neutral*, with evidence, queries and order unchanged. Three models (Qwen3-8B, Llama-3.1-8B and Qwen3.8-27B) give 7,776 answers each: 144 open theories in three example orders (432 pairs), and 144 entailed and 144 contradicted theories with both questions in three orders (1,728 known questions). A fixed lexicon maps the outputs back to answer labels.

From the definition with truth words to the definition with NLI words, Qwen3.8-27B keeps its directly stated facts (117 and 116 of 120 correct), while its correct answers on items that need derivation fall from 640 to 261 of 1,608. Known accuracy over the three prompts falls from 964 to 794 to 649 for Qwen3-8B and from 1,008 to 853 to 495 for Llama-3.1-8B. The answer interface matters for the phenomenon, and more *neutral* answers do not mean better reasoning from evidence.

F.4 Thinking on the edited set

With thinking enabled, all seven models with a thinking mode commit far fewer negated questions, reach 91 to 99% accuracy on derived known items, and almost stop saying that the premise establishes not-X (Table 46). The models use their native thinking template and official sampling settings, with at most 32,768 tokens, and the answer is read after the last end-of-thinking marker. Answers without thinking are those of Section 4.1 at zero temperature; the matched comparison under identical sampling is on the full set (Appendix F.6).

The seven models improve without exception. Committed negated answers fall sharply, to zero for Qwen3.8-27B and Gemma-4-31B, and the statement "establishes not-X" nearly disappears. Appendix F.7 shows that the change is in the computation and not in a response criterion.

Table 46: Thinking removes most committed negated answers and raises derived known accuracy to 91 to 99%. Edited set, without → with thinking: committed answers on the 327 open pairs for the original text, for the text in which only the rule condition mentioning the property is restored, and for the text in which only its occurrence in a rule conclusion is restored (297 pairs), the statement "the premise establishes not-X" judged true, and mean accuracy on the 344 derived known pairs over both questions.

	Original text		Rule condition restored	Rule conclusion restored	"Establishes not-X"	Derived known
Model	negated	plain	negated	plain	true	mean accuracy
Qwen3-32B	282 → **53**	149 → 26	224 → 44	169 → 18	214 → 32	55% → **94%**
Qwen3-14B	275 → **10**	157 → 15	248 → 22	188 → 7	187 → 16	47% → **92%**
Qwen3.8-27B	205 → **0**	111 → 0	178 → 0	123 → 0	65 → 0	58% → **99%**
Qwen3-8B	90 → **7**	72 → 7	34 → 20	93 → 7	20 → 6	45% → **93%**
Gemma-4-31B	18 → **0**	38 → 0	1 → 0	43 → 1	0 → 0	50% → **99%**
Gemma-4-12B	38 → **2**	44 → 2	4 → 0	100 → 4	20 → 0	33% → **98%**
Qwen3-4B	231 → **26**	246 → 25	121 → 42	232 → 16	51 → 17	73% → **91%**

Table 47: After thinking, 84 to 94% of the committed negated answers return to unknown (Gemma-4-31B 63%), and most returning traces say that the premises do not mention X. Pairs with plain unknown and negated committed without thinking, how many return to unknown with thinking, how many stay committed, and the returning traces that say the premises do not mention or state X.

Model	Committed without thinking	Back to unknown	Still committed	Trace says "not mentioned"
Qwen3-32B	2,802	2,369 (84.5%)	433	2,192/2,369
Qwen3-14B	2,260	2,092 (92.6%)	168	1,950/2,092
Qwen3.8-27B	2,498	2,353 (94.2%)	145	1,514/2,353
Qwen3-8B	989	840 (84.9%)	149	749/840
Gemma-4-31B	268	170 (63.4%)	98	143/170
Gemma-4-12B	297	254 (85.5%)	43	231/254
Qwen3-4B	399	371 (93.0%)	28	345/371

F.5 What the thinking traces say

In their traces, the models that recover say that the premises do not mention X, and the errors that remain write negation as failure out in words (Tables 47 and 48). For each model we take every open pair whose plain question is unknown and whose negated question is committed without thinking (the UD pairs of Table 49), and align it with the thinking trace on the same item.

The traces that return to unknown say that no premise mentions X and that it can be neither confirmed nor denied. Each of the 53 committed negated answers that remain for Qwen3-32B on the original texts of the edited set was coded by an AI assistant from the end of its trace, with the categories of Table 48. In 42 of the 53, the model infers "not X" from a failed search. None is a valid derivation, and none could be, since all these texts are affirmative. Of Qwen3-14B's 10 remaining committed negated answers, 9 fall into the same class. A second AI model, given the same excerpts and categories and blind to these codes, agrees on 52 of the 63 codes and places 39 of Qwen3-32B's 53 in the failed-search class. The three traces it marks as valid derivations (two for Qwen3-32B, one for Qwen3-14B) apply a contrapositive to a negative premise, such as "Fiona is not rough", that the affirmative text never states and the model itself obtained by a failed search; the second coder listed each of the three as ambiguous.

F.6 Thinking on the full set

On the full ProofWriter set, thinking nearly removes the asymmetry in all seven models and raises known accuracy to 79.76 to 97.75% (Table 49). Each model answers all 32,718 questions of both kinds in both modes, 458,052 answers in all. Both modes use matched sampling parameters within each model (temperature/top-p/top-k: Qwen3 0.6/0.95/20, Qwen3.8-27B 1.0/0.95/20, Gemma 1.0/0.95/64), so these numbers are not pooled with the zero-temperature results of Table 2.

Table 48: Forty-two of Qwen3-32B's 53 remaining errors after thinking infer "not X" from a failed search, and none is a valid derivation. Categories of the committed negated answers that survive thinking on the original texts of the edited set, coded by an AI assistant and checked by a second model.

Category	Qwen3-32B
Not X from a failed search	**42**
Absence to negation: "the premises do not state X, and no rule gives X, so not X"	15
Completion: "X could only come from these rules, their conditions are not met, so not X"	9
Negation as failure on a consequence, then contraposition: "if X then Y, Y is not stated, so not Y, so not X"	18
Another unfounded assumption (e.g., colors treated as mutually exclusive)	6
The trace concludes unknown, but the final answer is committed	5
Valid derivation	0
Total	53

Table 49: On the full set, thinking nearly removes the asymmetry and raises known accuracy. Net asymmetry on the 8,654 open pairs (points), accuracy on ProofWriter's own known questions, and the counts of pairs with only the negated (UD) or only the plain (DU) question committed, without $\rightarrow$ with thinking, under matched sampling.

Model	Without	With	Known accuracy	UD / DU counts
Qwen3-32B	$+29.10$	$+3.63$	61.30% $\rightarrow$ 87.61%	UD 2,802 $\rightarrow$ 725, DU 284 $\rightarrow$ 411
Qwen3-14B	$+23.92$	-0.76	59.03% $\rightarrow$ 82.55%	UD 2,260 $\rightarrow$ 170, DU 190 $\rightarrow$ 236
Qwen3.8-27B	$+16.35$	$+0.10$	58.76% $\rightarrow$ 97.75%	UD 2,498 $\rightarrow$ 34, DU 1,083 $\rightarrow$ 25
Qwen3-8B	$+5.41$	0.00	44.64% $\rightarrow$ 85.30%	UD 989 $\rightarrow$ 168, DU 521 $\rightarrow$ 168
Gemma-4-31B	-2.44	$+0.23$	54.18% $\rightarrow$ 95.60%	UD 268 $\rightarrow$ 46, DU 479 $\rightarrow$ 26
Gemma-4-12B	-1.56	$+0.15$	44.25% $\rightarrow$ 92.44%	UD 297 $\rightarrow$ 43, DU 432 $\rightarrow$ 30
Qwen3-4B	-13.44	-1.24	66.24% $\rightarrow$ 79.76%	UD 399 $\rightarrow$ 316, DU 1,562 $\rightarrow$ 423

F.7 Thinking changes the computation, not the criterion

Thinking does not merely make the models more willing to say unknown: in all seven models, the change from answering directly to thinking is incompatible with a pure shift of the response criterion (Table 50). We compare the two modes in a signal-detection framework on the 8,654 open pairs. The evidence axis s is a linear predictor of whether the negated question is committed, fit to the answers without thinking from structural features of the item, with five-fold out-of-fold scores. A pooled logistic regression of commitment on s, mode and their interaction gives the slope ratio κ between the two modes. A pure criterion shift would change only the intercept, which we take to mean that the 95% interval of $\log \kappa$ lies inside $(-0.15, +0.15)$ and that the threshold-free AUC of s changes by at most 0.02.

None of the seven models passes: every interval of $\log \kappa$ excludes zero, and $|\Delta\mathrm{AUC}|$ ranges from 0.032 to 0.205. Thinking changes the mapping from evidence to answer itself. The evidence axis is built from item structure, not from model internals, so the test concerns behavior.

F.8 Consistency is not correctness

Forcing the two answers to agree yields 100% agreement in either direction, but the results are opposite: anchoring on the plain question raises both open and known accuracy in five models, and anchoring on the negated question lowers both (Table 51). For every pair of answers of the nine core models in Table 2 we apply three fixed mappings, without running any model. The first keeps the plain answer and sets the negated answer to its three-valued opposite (plain anchor). The second does the reverse (negated anchor). The third sets both to unknown whenever either is unknown. The open denominator is 17,308 answers and the known denominator 11,376 questions.

The plain anchor raises open and known accuracy together for Qwen3-14B, Qwen3-32B, Qwen3.8-27B, Mistral-24B and Qwen3-8B, and the negated anchor lowers both for these five (Qwen3-32B

Table 50: No model is compatible with a pure criterion shift. Slope ratio κ between thinking and direct answering on an evidence axis built from item structure (95% interval, clustered by theory), and the AUC of that axis for commitment without $\rightarrow$ with thinking. Negated question, 8,654 open pairs.

Model	$\log \kappa$	AUC, without $\rightarrow$ with thinking
Qwen3-32B	-0.482 $[-0.642, -0.334]$	0.651 $\rightarrow$ 0.580
Qwen3-14B	-0.183 $[-0.331, -0.033]$	0.722 $\rightarrow$ 0.687
Qwen3.8-27B	$+0.584$ $[+0.311, +0.858]$	0.604 $\rightarrow$ 0.697
Qwen3-8B	-0.155 $[-0.296, -0.011]$	0.740 $\rightarrow$ 0.708
Gemma-4-31B	$+0.426$ $[+0.263, +0.593]$	0.755 $\rightarrow$ 0.831
Gemma-4-12B	$+0.202$ $[+0.029, +0.366]$	0.739 $\rightarrow$ 0.786
Qwen3-4B	-0.744 $[-0.828, -0.658]$	0.811 $\rightarrow$ 0.606

Table 51: Anchoring on the plain question makes the two answers consistent and, in five models, more accurate on both open and known items. Change in open and known accuracy (points) when the negated answer is derived from the plain one, and known answers changed from correct to wrong / wrong to correct.

Model	Open accuracy	Known accuracy	Known correct $\rightarrow$ wrong / wrong $\rightarrow$ correct
Qwen3-32B	$+15.62$	$+4.36$	370 / 866
Qwen3-14B	$+11.73$	$+1.03$	614 / 731
Qwen3.8-27B	$+11.80$	$+13.70$	102 / 1,661
Mistral-24B	$+10.60$	$+4.91$	130 / 689
Llama-3.1-8B	$+6.60$	-27.55	3,260 / 126
Qwen3-8B	$+2.86$	$+3.44$	304 / 695
Gemma-4-31B	-1.18	$+9.37$	105 / 1,171
Gemma-4-12B	-1.07	$+5.15$	141 / 727
Qwen3-4B	-6.60	-9.25	1,325 / 273

-15.62 and -4.21, Qwen3.8-27B -11.80 and -13.82 points), because their errors fall mainly on the negated question. Llama-3.1-8B's errors are also mostly on the negated question, but its plain answers on known items are poor (plain anchor -27.55 on known items). Qwen3-4B's errors are mostly on the plain question, and both Gemma-4 models barely change on open items. The third mapping raises open accuracy and lowers known accuracy in all nine models, only by abstaining more. Under the two single-question anchors, the gain on open items is half the net asymmetry of Table 9 in magnitude by construction, so the change on known items is the actual test. Among pairs originally committed on both questions, 12,533 across the nine models are consistent with each other and wrong on both (Qwen3-14B 3,097 of 4,294, Qwen3-32B 3,047 of 4,002, Qwen3-4B 1,943 of 4,114, Llama-3.1-8B 1,708 of 6,927, Qwen3.8-27B 997 of 2,132). Removing a contradiction does not handle the gap correctly.

F.9 Asking both questions at once

Asking both questions in one conversation removes the contradiction but does no better than asking the same question twice, and Qwen3-14B loses known items (Table 52). Qwen3-14B and Qwen3.8-27B answer without thinking (BF16, greedy). We use 120 original theories, excluding the 90 used in Appendix D.2. Each contributes one open question with its pair and one known anchor question with its pair, in six formats: plain alone, negated alone, plain then negated, negated then plain, plain twice and negated twice, 1,440 answers per model. Intervals are simultaneous 95% intervals over 12 differences from 10,000 resamples of the 120 theories.

For Qwen3.8-27B, 103 pairs with the plain question first and 98 with the negated question first are unknown on both questions, close to asking the same question twice (102 and 98). For Qwen3-14B, pairs with only the negated question committed fall from 41 to 1, while its correct known anchors fall from 60 to 35 (-20.8 points, $[-31.7, -10.0]$). Removing the contradiction is not the same as fixing the judgment, and the order of the questions changes both.

Table 52: Asking both questions together nearly removes the one-sided commitments; its effect on known answers depends on the model and the order (Qwen3-14B: 60 to 35). Open pairs by the answers to the two questions (120 pairs; 118 readable for Qwen3-14B asked separately) and known anchors answered correctly (of 120).

Model and format	Both unknown	Only negated committed	Only plain committed	Both committed	Known correct
Qwen3.8-27B, separately	65	28	7	20	64
Qwen3.8-27B, plain then negated	103	**0**	0	17	71
Qwen3.8-27B, negated then plain	98	**0**	0	22	42
Qwen3-14B, separately	46	41	0	31	**60**
Qwen3-14B, plain then negated	102	1	0	17	**35**

G ORIGIN

The lookup is present before instruction tuning, the prompt examples do not teach it, and in pretraining text people write "not" more often under the condition that the lookup requires.

G.1 PRETRAINED CHECKPOINTS ALREADY LOOK UP

Five of seven pretrained checkpoints already show a lookup effect, and instruction tuning can keep it, change how its failure is reported or remove it (Table 53). Pretrained checkpoints do not use a chat template, so we write the 20,203 rows of the edited set as plain text. Each of the six examples becomes the question followed by "Answer:" and its label, and the final item ends with "Answer:". We read the log-probabilities of the first token over the labels " true", " false" and " unknown" within the top 20 candidates and take the largest, in a single step at zero temperature. The pretrained and the instruction-tuned versions read exactly the same tokens. The lookup effect is defined as in Section 4.2.

Table 53: Five of seven pretrained checkpoints already look up. Under the template-free prompt: lookup effect (points, 95% interval), asymmetry on the original items (negated minus plain committed, points), directly stated known questions answered correctly (of 204) and share of all answers that are unknown, for the pretrained and the instruction-tuned version. Last column: lookup effect of the instruction-tuned version with its chat template (Section 4.1). n/a: not available.

	Pretrained, template-free				Instruction-tuned, template-free				Chat template
Checkpoint	Lookup effect	Asym.	Known	unknown	Lookup effect	Asym.	Known	unknown	Lookup effect
OLMo-2-13B	$+40.4$ [$+32.4, +48.5$]	$+56.3$	115	49%	$+40.7$ [$+33.0, +48.4$]	$+55.4$	112	45%	n/a
Qwen3-14B	$+25.4$ [$+18.4, +32.4$]	$+31.8$	202	55%	$+48.9$ [$+42.7, +55.3$]	$+29.4$	202	38%	$+46.8$
Qwen3-4B	$+23.2$ [$+16.6, +30.3$]	$+27.2$	196	46%	$+4.0$ [$-1.2, +9.4$]	$+9.8$	202	55%	-4.3
Gemma-4-12B	$+12.2$ [$+2.0, +22.4$]	$+43.4$	161	40%	unreadable: almost all true/false (unknown 4%)				$+1.2$
Gemma-4-31B	$+11.3$ [$+3.2, +19.5$]	-2.1	204	50%	n/a	n/a	n/a	n/a	$+2.8$
Qwen3-8B	-6.1 [$-10.8, -1.7$]	$+8.6$	156	81%	$+0.6$ [$-4.4, +5.7$]	$+25.4$	200	68%	$+1.2$
Llama-3.1-8B	unreadable: almost all unknown		28	99%	$+29.7$ [$+19.7, +39.8$]	$+55.7$	191	54%	$+20.8$

The pretrained Qwen3-4B, Qwen3-14B, Gemma-4-12B and Gemma-4-31B, and OLMo-2-13B, whose training data are fully public (Team OLMo et al., 2024), all show a positive lookup effect with intervals excluding zero. Pretrained Qwen3-8B mostly answers unknown, and its lookup effect is negative. Pretrained Llama-3.1-8B answers unknown to almost everything, answers only 28 of 204 directly stated facts correctly, and its lookup cannot be read. Instruction tuning can keep the lookup: OLMo-2 stays at $+40$, and Qwen3-14B rises from $+25$ to $+49$ ($+47$ with the chat template). It can change how the failure is reported: under its chat template, instruction-tuned Qwen3-4B still looks up at the rule condition (negated $+24$), but the failure lands on both questions (plain $+28$), so its lookup effect is near zero. It can also remove the lookup, as in both Gemma-4 models. Instruction-tuned Gemma-4-12B answers almost every item with true or false under the template-free prompt, so that column is unreadable, and for both Gemma-4 models the chat-template value applies.

G.2 THE PROMPT EXAMPLES TEACH THE OPPOSITE

The error is not learned from the prompt examples, one of which teaches the opposite: *Tia is kind.* with the question "Tia is not old." is answered unknown. Both unknown examples among the six

Table 54: Examples in which the queried predicate is present and the answer is still unknown do not restore unknown on the negated question. Open pairs by the answers to the two questions (48 items in three orders, 144 pairs) and known questions answered correctly (of 120). Present but open: the added facts make the queried predicates occur in the two unknown examples. Matched control: the same added words with the predicates swapped.

Model, examples	Both unknown	Only negated committed	Only plain committed	Both committed	Known correct
Qwen3-8B, original six	75	28	10	31	109
Qwen3-8B, matched control	73	24	8	39	109
Qwen3-8B, present but open	**67**	**29**	10	38	110
Llama-3.1-8B, original six	15	48	1	80	94
Llama-3.1-8B, matched control	10	37	0	97	98
Llama-3.1-8B, present but open	11	41	2	90	99

Table 55: Pretrained models continue with "x is not P." more readily when others are listed with P or when P has been mentioned. Change in D (nats, 95% interval) under three edits, D on the original text, and the share of known pairs on which the sign of D points to the gold label. 327 open pairs, 264 for the column on others' facts.

Pretrained model	Others listed with P	Original $-$ P removed	Rule only $-$ control rule	D, original	Known: sign correct
OLMo-2-13B	$+0.46$ $[+0.36, +0.56]$	$+0.28$ $[+0.09, +0.46]$	$+0.15$ $[+0.02, +0.28]$	-2.49	100%
Qwen3-14B	$+0.99$ $[+0.85, +1.15]$	$+0.84$ $[+0.66, +1.03]$	-0.10 $[-0.25, +0.06]$	-1.99	98%
Qwen3-4B	$+0.79$ $[+0.67, +0.92]$	$+0.83$ $[+0.69, +0.97]$	$+0.27$ $[+0.18, +0.37]$	-2.03	100%
Gemma-4-12B	$+0.33$ $[+0.23, +0.43]$	$+0.42$ $[+0.29, +0.54]$	$+0.10$ $[+0.01, +0.19]$	-2.47	100%
Qwen3-8B	$+0.56$ $[+0.34, +0.76]$	$+0.43$ $[+0.16, +0.65]$	-0.00 $[-0.14, +0.12]$	-1.77	98%
Llama-3.1-8B	$+0.18$ $[+0.10, +0.26]$	$+0.41$ $[+0.29, +0.54]$	$+0.01$ $[-0.07, +0.09]$	-2.95	100%

illustrate an open item by a query predicate that does not occur in the example's text, the Tia example with a negated question and one with a plain question (*Lio is red.* $\rightarrow$ "Lio is large."). If the models had learned "a word that does not occur means unknown" from these examples, an example in which the queried word occurs and the answer is still unknown should restore unknown on the negated question. We add to each of the two examples one affirmative fact about another person that makes the queried predicate present while the answer stays unknown (*Neral is large.*, *Veston is old.*). The matched control adds the same words with the predicates swapped between the two examples, so the added words are identical as a multiset. Qwen3-8B and Llama-3.1-8B answer 68 development items (48 open, 20 directly stated known) in three example orders with both questions under both edits, 816 answers per model, with generation settings unchanged.

Relative to its matched control, the example that shows a present but open predicate yields at most one more pair with both questions unknown and no fewer pairs with only the negated question committed (Table 54). The test covers one kind of edit to two fixed examples, and the 48 theories in three orders are not independent samples.

G.3 Continuations of pretrained language models

Without any question, pretrained language models continuing a text already raise the probability of "x is not P." when the text lists another entity with P and does not mention x (Table 55). We treat the text of each open pair of the edited set as the beginning of a document and score two continuations, $D = \log p(\text{"}x \text{ is not } P\text{."} \mid \text{text}) - \log p(\text{"}x \text{ is } P\text{."} \mid \text{text})$, summed over the continuation tokens (prompt log-probabilities from vLLM (Kwon et al., 2023), with no generation, examples or instruction). We compare D for the same pair across edits: only other entities' facts with the target property restored (*Bob is quiet.*) against only the control facts, the original against the text with the property removed, and only the rule restored against only the control rule. On known pairs, D should be negative when the text proves X and positive when it refutes X.

When others are listed with P and x is not mentioned, all six pretrained models raise the probability of "x is not P." (+0.18 to +0.99 nats), and a mention of P anywhere in the text does the same (+0.28 to +0.84). This matches how "not" is written in the pretraining corpus (Appendix G.4): people write "x is not P." more often when P has been mentioned and not attributed to x. The tilt is relative. On the

Table 56: A sentence is more often negative when P was mentioned before without being said of x, and less often when a matched other property was mentioned. Sampled occurrences of sentence-final "x is not P." and "x is P." in OLMo-mix-1124 by the preceding 150 words, and the pooled odds ratio of the negative form relative to "P not mentioned" (95% interval). 24 predicates, 872 negative and 831 affirmative occurrences.

Preceding 150 words	"x is not P."	"x is P."	Odds ratio of negation
P not mentioned	592	655	1
P mentioned, not said of x	246	164	**1.70** [1.36, 2.15]
of which something else is said to be P	46	32	1.61 [1.03, 2.64]
Said of x	34	12	3.38 [1.73, 7.84]
Control: matched other property in the window			0.51 [0.26, 0.96]

original texts all six models still prefer to continue with "x is P." (D between -1.8 and -3.0 nats), whereas in question answering the models treat the same moment as the condition under which "not" holds. When only the rule condition is restored, the tilt is small and mixed (positive for Qwen3-4B, Gemma-4-12B and OLMo-2-13B, near zero for Qwen3-8B, Qwen3-14B and Llama-3.1-8B). Without a question, the continuations show the tilt at the listed facts, and the lookup at the rule condition appears only in question answering.

G.4 Negation in the pretraining corpus

In OLMo-2's pretraining corpus, the odds that a sentence about x is "x is not P." rather than "x is P." are higher when P has been mentioned in the preceding text without being said of x than when it has not, and lower when an equally frequent other property word has been mentioned instead (Table 56). Of the sampled sentences "x is not P.", 32% have P in the preceding 150 words, and of the sentences "x is P.", 21% do.

Method. The corpus is OLMo-mix-1124, the 4.6 trillion tokens used to pretrain OLMo-2 (Team OLMo et al., 2024), queried through the public infini-gram engine (index `v4_olmo-mix-1124_llama`, Liu et al., 2024) without running any model. Its web portion comes from DCLM (Li et al., 2024). The predicates are the 14 property words of ProofWriter and 10 predicates typical of records (*available*, *certified*, *eligible*, *employed*, *insured*, *licensed*, *married*, *registered*, *vaccinated*, *vegetarian*). For each predicate we count the sentence-final occurrences of "is P." and "is not P." in the corpus, sample 50 occurrences of each uniformly at random (30 for the record predicates), and retrieve about 700 preceding tokens from the same document. We keep only occurrences with at least 150 preceding words, so that negative and affirmative sentences do not differ in their position in the document, and classify the preceding 150 words into three cases that correspond to the three situations of the lookup:

- *P not mentioned*: the word P does not occur in the window, so there is nothing to look up.
- *P mentioned, not said of x*: P occurs, but neither "x is P" nor "x is not P" does, where x is the last word of the sentence's subject. One subcase is that something else is said to be P.
- *said of x*: the window already contains "x is P" or "x is not P".

Within a predicate, the odds ratio of negative to affirmative does not depend on how many occurrences of each were sampled, and across predicates we pool with the Mantel–Haenszel estimator. As a control, each predicate is paired with a property word of similar corpus frequency (for example *quiet* and *young*, *red* and *blue*, *married* and *registered*), and we ask whether that word occurs in the same window. Intervals are 95% bootstrap intervals over the sampled occurrences within each predicate and polarity (2,000 resamples).

The pattern holds in the two groups of predicates and under other windows. For the 14 ProofWriter property words, the odds ratio for "P mentioned, not said of x" is 1.86 [1.43, 2.47], for "something else is said to be P" 3.75 [1.94, 9.09], and for the control word 0.50 [0.23, 0.98]. For the 10 record predicates the three ratios are 1.40 [0.92, 2.17], 0.65 [0.30, 1.29] and 0.61 [0.00, 3.28]. With a window of 50 words the main ratio is 2.10, and with 300 words 1.78 (24 predicates). Removing expletive and demonstrative

Table 57: Most predicates are written negatively more often after P was mentioned without being said of x. Corpus counts of sentence-final "is not P." and "is P.", valid sampled occurrences, the share of them with P in the preceding 150 words, and the odds ratio of negation for "P mentioned, not said of x" relative to "P not mentioned" (with 0.5 smoothing). The samples are small, so only the direction is informative.

Predicate	Corpus count: negative / affirmative	Valid samples	P in preceding 150 words	Odds ratio
big	26,495 / 657,362	45 / 42	13% / 21%	0.58
blue	12,642 / 500,064	44 / 41	50% / 27%	2.26
cold	17,261 / 646,113	38 / 45	32% / 22%	1.97
furry	215 / 3,547	49 / 45	39% / 27%	1.86
green	18,694 / 453,548	41 / 39	56% / 44%	1.66
kind	16,423 / 119,207	46 / 44	13% / 20%	0.57
nice	70,443 / 1,901,263	44 / 48	11% / 17%	0.54
quiet	7,870 / 401,056	49 / 40	29% / 10%	3.31
red	11,811 / 514,401	40 / 37	57% / 30%	2.96
rough	1,035 / 187,355	41 / 46	22% / 4%	4.11
round	10,537 / 220,790	45 / 45	31% / 11%	2.69
smart	51,486 / 425,289	45 / 40	16% / 2%	4.45
white	48,847 / 903,128	41 / 38	61% / 39%	2.16
young	7,006 / 264,427	38 / 41	34% / 17%	2.25
available	1,431,350 / 6,166,388	24 / 21	17% / 19%	0.85
certified	8,230 / 54,737	24 / 30	33% / 30%	1.30
eligible	20,272 / 106,838	29 / 21	45% / 38%	1.04
employed	13,378 / 394,599	28 / 25	4% / 8%	0.51
insured	15,825 / 84,638	25 / 23	24% / 9%	2.43
licensed	11,538 / 77,425	27 / 22	44% / 23%	1.95
married	39,648 / 266,349	29 / 25	21% / 24%	0.70
registered	29,554 / 228,533	26 / 24	31% / 17%	2.06
vaccinated	6,940 / 37,189	26 / 24	42% / 29%	1.73
vegetarian	7,670 / 44,747	28 / 25	46% / 32%	1.79

subjects such as *it*, *this*, *that* and *there*, where idioms like "the difference is not big" concentrate, gives 1.78. Table 57 lists every predicate.

Examples. A beer review writes *Some coffee stouts are rough and this is not rough.*: something else is said to be P, and this one is written as not P. The "said of x" class consists mostly of denials of a claim already on the table. A philosophical discussion first writes *the proposition that grass is blue* and later *the 'conclusion' that grass is not blue*.

Reading. People write "x is not P." rather than "x is P." more often when P has been mentioned and not attributed to x. When the preceding text mentions a different property word, the excess negation does not appear, so the excess follows P itself. This agrees with linguistic accounts in which a negative sentence presupposes that the affirmative is under consideration (Givón, 1978; Horn, 1989). It corresponds to the last two outcomes of the lookup: when the text does not mention P, the model does not say "not" (with *quiet* removed, the negated question returns to unknown), and when P is mentioned but not found for x, the model says "not". We count only negations that repeat P. Elliptical negations such as "…, but this one is not." depend most on the preceding text and are not counted, which makes the ratios above conservative.