

Source Marking in Language Models: A Controlled Framework for Acquisition, Cue Sensitivity, and Reasoning

Daniil Huryn

Independent Researcher

Atlanta, GA, USA

danikhur@gmail.com

Abstract

Language can affect how people think. We ask a computational version: if a language model is trained on English that marks propositions as observed, reported, or inferred, does it use evidence differently in ordinary English? We create matched corpora in which particles either mark these routes (EVID), appear in the same positions without predicting source (CTRL), or are absent (NOMARK), and remove them at evaluation.

We test 405M-parameter Pythia models trained from scratch and Qwen3.5-4B Base and Post models adapted across three paired seeds. The models learn the mapping: Pythia reaches 92.8% before ordinary-English adaptation and retains 79.8% afterward; Qwen EVID averages about 80% for both Base and Post, with controls near chance. Visible markers shift Qwen’s scores. Once markers are removed, however, the Pythia tests cannot show whether the effect carries over. Base’s main Qwen result changes direction across seeds, and Post’s is negative in all three. A separate test suggests that Base may respond more to relevant context, but not specifically to source wording. No condition consistently improves reliability, multistep inference, or recognition of copied information. Source marking is easy to learn and can influence model scores, but we find no broader improvement in reasoning.

1 Introduction

Language can influence how people approach particular cognitive tasks, although this rarely amounts to a general advantage for speakers of one language. Evidentiality gives us a concrete case: it marks whether information came from perception, testimony, inference, or another route. Turkish, for example, marks a direct–indirect contrast in parts of its past-tense system. This connection between linguistic form and information source has made evidentiality useful for studying language and cog-

nition (Aksu-Koç et al., 2009; Papafragou et al., 2007; Ünal and Papafragou, 2020).

This does not mean that speakers of one language simply “reason better.” Children can reason about sources before fully mastering evidential morphology, and linguistic effects often appear only when evidential language is explicitly processed (Papafragou et al., 2007; Ünal and Papafragou, 2020). Studies of Turkish-speaking children have also found links between their use of direct evidentials and performance on source-monitoring tasks (Kandemirci et al., 2023). Karakaş and Şimşek (2026) find that source trustworthiness changes Turkish speakers’ choice of evidential morphology, while the effects for language models vary with the model and prompt. So three things have to stay separate: how information was acquired, whether the source is reliable, and whether the model reasons correctly from it.

Language models make a related question unusually easy to test: does a model’s *training language* affect its behavior after it returns to ordinary English? We present the same narratives with a source-related grammatical pattern, an equally frequent pattern assigned by a hash and unrelated to source, or no pattern at all. We then remove every artificial marker before evaluation. The main comparison keeps the content, marker positions, frequency, and compute the same. By comparing evidence about the target, evidence about a distractor, and no related evidence, we can see whether the model responds to information about the claim rather than simply to evidence-like language elsewhere in the prompt.

Most related work changes source or certainty cues when the model is answering a prompt. Authority, hearsay, belief reports, and epistemic stance can change whether a model follows a contextual claim (Du et al., 2026), while expressions of uncertainty and overconfidence can shift answer behavior (Zhou et al., 2023). Source preferences also

do not reliably separate helpful from harmful assertions (Li et al., 2026). We take the opposite approach: add a source distinction during training, remove it, and then evaluate the model.

We test two settings: 405M-parameter Pythia models trained from scratch and matched adaptation of Qwen3.5-4B Base and Post. The Qwen experiments ask whether a more capable model can use the source markers to perform better on evidence tasks. Together, the experiments ask:

1. **Learning the markers:** Does EVID training teach the model which marker goes with each source type?
2. **After removing the markers:** Does EVID training change how strongly the model responds to evidence about the target when we evaluate it in plain English without the custom markers?
3. **Reasoning:** Do any of the models improve on a broader range of evidence-related reasoning tasks?

There are no artificial markers in our main evaluations: we compare EVID and CTRL models on identical English inputs. There is also a separate test for the models’ performance when the markers are left in.

The intervention demonstrates how information was acquired, not its truth or the reliability of its source. Reliability and provenance evaluations test whether those abilities emerge from source markers without being taught.

The last question matters most for our original motivation because it would provide the clearest parallel between human and model reasoning. Still, paying more attention to relevant context is not the same as paying more attention to its source, and neither guarantees a correct answer. We call the first change *contextual responsiveness*, the second *source-cue sensitivity*, and reserve *source-sensitive reasoning* for cases where the model uses source information correctly.

In both studies, EVID training teaches the custom marker system. The custom models remember much of the mapping after adapting to ordinary English, but our marker-free tests cannot tell us whether that knowledge changes how they use evidence. Across three Qwen seeds, Base’s main result does not repeat, and EVID performs below CTRL for Post every time. In a separate test, Base

responds more to target evidence than to no evidence, but not consistently more than to distractor evidence. Removing the source wording also does not remove the difference, so the result does not appear specific to source cues. What we offer is a controlled method and a clearer account of what the models learn and what changes afterward. We do not claim a general improvement in reasoning.

2 Related Work

2.1 Evidentiality and source monitoring

Evidential markers describe where information came from, not which sources should always be trusted; any implication about reliability depends on the context. Children can reason about sources before fully mastering evidential morphology, and linguistic effects are clearest when a task requires them to process evidential language (Papafragou et al., 2007; Ünal and Papafragou, 2020). Studies of Turkish-speaking children nevertheless find links between evidential use and source monitoring, although the results vary by task and developmental measure (Aksu-Koç et al., 2009; Kandemirci et al., 2023).

Karakaş and Şimşek (2026) provide the closest natural-language comparison. They vary the trustworthiness of a source named in the text and measure Turkish evidential morphology in humans and language models. Humans show a robust trust effect, while model behavior depends on the model and prompt. We instead change artificial source marking during training and remove it at evaluation.

2.2 Source and confidence in LLMs

Language models famously produce different outputs after small changes in how an input is phrased. Similar content influences the model differently depending on style, belief, evidential language, epistemic stance, or tone (Du et al., 2026), and human-versus-AI labels change trust in otherwise identical content (Sun et al., 2026). These studies show that source-related language matters when the model can see it, but they do not show that learning such cues during training changes later behavior when the cues are absent, nor whether any change reflects improved reasoning or simply better recall.

We must also show that the models use source information correctly, rather than merely prefer sources that happen to be right more often. Models often favor documents over direct user statements

without consistently separating helpful from harmful information. Fine-tuning them on cases where these sources conflict improves that discrimination (Li et al., 2026). Confidence is similarly difficult to measure: probabilities, correctness, and verbal confidence can disagree (Jiang et al., 2021; Mielke et al., 2022), and confidence associated with particular markers can change across tasks even when their broad ranking remains similar (Liu and Cohan, 2026).

2.3 Controlled changes to training data

Studies using controlled artificial languages were the greatest inspiration for our work. White and Cotterell (2021) train on parallel datasets that change only one grammatical feature, allowing them to isolate what that feature changes in model learning. Hu et al. (2025) show that early training on formal languages with different dependency structures changes later English learning and generalization. Our paper applies the same idea to source marking: our datasets differ in whether custom syntax predicts source, and we evaluate the models on plain English data without the syntax.

Our broader evaluations also borrow ideas from existing benchmarks. BoardgameQA tests non-trivial reasoning when information conflicts and some rules take priority over others (Kazemi et al., 2023), NeoQA uses fictional events and systematically removes evidence (Glockner et al., 2025), and ECON tests noticing a conflict separately from resolving it (Jiayang et al., 2024). We also check whether the model differentiates between evidence about the target and evidence about unrelated subjects.

3 Controlled Source-Marking Corpora

3.1 Synthetic source passages

The narratives were generated from scratch for the corpus and were not modifications of any existing data. The passages contain concrete events across 41 topics and vary in their contents as much as possible, while preserving the three information-acquisition routes. In a *direct-source* passage, a character witnesses the event. In a *reported-source* passage, the character learns about it from another person or a record. In an *inferred-source* passage, the character works it out from indirect clues. An artificial particle marks each verb that states what happened, but the phrase establishing the source—such as *saw*, *was told*, or *deduced*—is left un-

marked. Because of this, the model must use the earlier explanation of how the character learned the information rather than simply matching a particle to the nearest verb. We do not test how much text can separate that explanation from the marked claim. Appendix A gives an example.

The generation prompts prohibit explicit confidence words and phrases such as *certain*, *probably*, and *I think*. Otherwise, a model could associate markers with those words and borrow the target confidence from the text rather than reason about the evidence. Anthropic Claude-family models and OpenAI GPT-5-mini generated the passages. Every candidate was evaluated for marker count, source category, marker identity, intentionally excluded vocabulary, malformed markers, and duplicate text. We also used embedding-based cosine similarity to remove similar passages.

3.2 Custom-pretraining corpus

For the custom-pretraining experiment, we turn the same 523,671 source narratives into three conditions and keep them in the same order:

- EVID: ko, zo, and mu consistently mark direct, reported, and inferred evidence.
- CTRL: ba, du, and fo occur at identical evidence positions as EVID markers, but their identities are independent of source. An MD5 hash of the preceding sentence and marker position determines the marker.
- NOMARK: all artificial markers are removed.

EVID and CTRL each contain 52.13M Pythia tokens. NOMARK contains 49.04M because the same narratives become shorter when the particles are removed. We keep the number of passages fixed rather than adding more NOMARK passages to match the token count, since doing so would give NOMARK more distinct training data. We give every condition the same number of optimizer updates and presented tokens. The primary EVID–CTRL comparison is matched in sequence length, narrative content, and compute. We treat CTRL–NOMARK as secondary because NOMARK cycles through its shorter token stream more often.

3.3 Qwen adaptation corpus

For the stronger-model experiment, we use a cleaner, more strictly filtered subset rather than the full dataset. We keep passages with three to

Corpus	Passages	Tokens/condition
Custom EVID/CTRL	523,671	52.13M
Custom NOMARK	523,671	49.04M
Qwen EVID/CTRL	395,529	37.19M

Table 1: Sizes of the training corpora. The custom conditions are passage-matched; Qwen EVID and CTRL are also matched in vocabulary and token count.

eight markers, remove any passage with prohibited confidence vocabulary or malformed markers, and balance direct, reported, and inferred passages inside every bucket. This leaves 395,529 training passages: 131,843 for each source category and 763,463 marker occurrences per category.

EVID again has direct as ko, reported as zo, and inferred as mu. The Qwen CTRL condition uses the *same* three markers. Ko, zo, and mu are reused in CTRL to keep as much as possible equal between conditions, including tokenization and vocabulary. Each marker has the same total count as in EVID, but we shuffle them so that each appears about equally often within every source category. The two conditions have the same narratives, vocabulary, marker positions, marker counts, token count, and order. The only difference is whether marker identity predicts source. Each condition contains 37,189,481 Qwen tokens. For validation, we use a separate balanced set of 285 passages, with 95 from each source category.

4 Experimental Setup

4.1 Study 1: custom pretraining

We initialize the Pythia-410M architecture from random weights (405M parameters in our local configuration; [Biderman et al., 2023](#)). The source markers are present from the beginning of training rather than added after the model has already learned English. The main advantage is simple: paired conditions start from identical weights and differ only in their training corpora.

The reported results use seed 42 for EVID, CTRL, and NOMARK. Each model trains for 4,770 optimizer steps at sequence length 1,024 with an effective batch of 32 sequences (32,768 tokens/step), totaling 156,303,360 presented tokens. We use AdamW, a peak learning rate of 3×10^{-4} , 1% linear warmup, cosine decay, weight decay .01, and gradient clipping at 1.0. Training uses bfloat16 on one NVIDIA RTX 5080.

We then adapt every final checkpoint to ordi-

nary English with the same fixed TinyStories data ([Eldan and Li, 2023](#)): 5M training tokens and a separate 200K-token prefix from the official validation split. Conditions receive identical batches. We evaluate the shared 400-step checkpoint, which minimizes mean validation perplexity while keeping the relative spread below 3%, after 13,107,200 presented adaptation tokens.

Study 1 includes only seed 42 and does not estimate variation across training runs.

4.2 Study 2: matched Qwen adaptation

We use Qwen3.5-4B Base and its architecture- and tokenizer-matched post-trained checkpoint ([Qwen Team, 2026](#)). This holds architecture and vocabulary fixed, but is not a clean test of *reasoning*: post-training changes both data and objectives.

Each checkpoint receives paired EVID and CTRL adapters initialized from the same saved LoRA weights. We use NF4 QLoRA with double quantization and bfloat16 computation ([Dettmers et al., 2023](#)). LoRA rank is 8, alpha is 16, dropout is zero, and we target attention, linear-attention, and feed-forward projections, giving 16,232,448 trainable parameters. Sequence length is 256 and the effective batch contains 4,096 tokens. AdamW uses a learning rate of 2×10^{-4} , no weight decay, gradient clipping at 1.0, and linear warmup over the first million tokens.

Within each of three seeds (S1–S3), EVID and CTRL use the same adapter initialization, chunk order, and number of updates. We evaluate at 245 steps and 1,003,520 presented tokens, the end of warmup and 2.70% of the 37.19M-token stream. Evaluation uses unquantized bfloat16 base weights with the adapter. We run S2 and S3 with the same configuration after the initial analysis. We record the exact random seeds and hashes of the starting adapter weights.

4.3 Comparing conditions

We compare EVID and CTRL within the same seed and fictional world. For each Qwen seed, the bootstrap resamples the 24 worlds, so it measures variation across test items for that adapter pair rather than variation from retraining the model. We report all three seeds and their simple average. Three seeds are too few for a useful confidence interval across training runs.

5 Evaluation

5.1 Learning the EVID mapping

Before asking whether the training changes later reasoning, we first check that EVID actually learns the marker system. In the custom study, we score the first marker in 583 held-out beginnings of passages (197 direct, 197 reported, 189 inferred), none of which contains an earlier particle. The choices are the single tokens *ko*, *zo*, and *mu*. We test checkpoints before and after TinyStories adaptation. Custom CTRL uses *ba/du/fo*, so here we only ask whether it predicts the EVID markers by accident; this test does not measure whether CTRL learned its own particles. Both Qwen conditions use *ko/zo/mu* and are tested on a balanced set of 285 passages. Chance is 33.3%. If EVID does not learn the mapping, any later difference is unlikely to come from that mapping.

We later add a follow-up test asking whether Qwen uses a marker when it can see one. Across 24 worlds and all three source routes, the marker is either absent, matches the source described in the passage, or is one of the two conflicting markers (288 items). We ask whether a conflicting marker pulls the model toward the source associated with that marker, relative to the same passage without a marker, and whether this pull is larger for EVID than CTRL. This tests use of a visible marker, not what remains after markers are removed.

5.2 Tests after removing the markers

Single-word confidence test. The marker-free task contains 30 direct, 30 reported, and 30 inferred items. We exclude twenty additional topic and complexity controls from the evidence analysis. Each item uses:

CONTEXT: [context]
QUESTION: How confident should we be that [claim]?
ANSWER: We should be

At the next-token position, we sum probabilities for five high-confidence continuations (*very*, *quite*, *highly*, *certain*, *confident*) and five low-confidence continuations (*not*, *somewhat*, *uncertain*, *cautious*, *skeptical*). For item i ,

$$C_i = \frac{\sum_{w \in H} p(w \mid x_i)}{\sum_{w \in L} p(w \mid x_i)}. \quad (1)$$

We label direct items high confidence and reported and inferred items low confidence. We

first calculate the direct-minus-reported difference within each model and then compare that difference between EVID and CTRL. This is a narrow test of which confidence word the model prefers, not a general test of whether its confidence is appropriate. It also assumes that direct evidence should produce more confidence than reported evidence, which is not true in every situation or evidential system.

Full-sentence test. A second evaluation uses 24 fictional worlds, both target states, three source routes, two reliability levels, and contexts containing relevant evidence, evidence about a distractor, or no evidence. It replaces the word lists with four complete sentence frames. Simple control items first check whether the model responds correctly to an unambiguous statement.

General language ability. We also report TinyStories validation perplexity and accuracy on 12 BLiMP paradigms (Warstadt et al., 2020). These are broad checks of language performance, not proof that the models are otherwise equivalent.

5.3 Qwen primary task

The primary Qwen task uses 288 marker-free combinations across the same 24 fictional worlds. Each crosses direct, reported, and inferred evidence; high and low reliability; and both target states. Every combination appears in three contexts, yielding 864 items: *matched* evidence concerns the queried object, *swapped* evidence concerns a distractor, and *null* provides no additional evidence.

The model chooses one of two states. We average the correct-minus-alternative log-probability margin across both A/B orders. Base receives a raw completion instruction; Post uses its chat template with thinking disabled. For matched, swapped, and null margins m_M , m_S , and m_N , we report

$$B = m_M - m_S \quad (\text{target binding}), \quad (2)$$

$$G = m_M - m_N \quad (\text{target evidence vs. null}). \quad (3)$$

Our main comparison is EVID minus CTRL. We include the original, unadapted checkpoints for context. Every checkpoint passes the simple tests of explicit statements and matching propositions to the correct object.

5.4 Additional evaluation set

After observing the first Qwen Post result, we design a separate exploratory set of 552 marker-free items across 24 new fictional worlds. It tests

whether the model can identify the source of a claim (72), recognize when evidence is insufficient (120), adjust when source reliability is reversed (120), combine evidence across several steps (96), learn which new sources are reliable (72), and distinguish independent reports from copied ones (72). Each item has three choices, including insufficient evidence where appropriate, and appears in three A/B/C answer orders. We score the answer labels directly and average across orders, without asking the model to generate its reasoning.

The most important comparisons ask whether the model reacts differently to evidence about the target, the same kind of evidence about a distractor, and no related evidence at all. Training marks how information was acquired, not whether it is true, so improved judgments of reliability would have to emerge without being taught directly. We discuss the reliability, multistep, and copied-source results only when the untouched model can already perform the task. This additional evaluation does not replace the main task.

After obtaining all three seed results, we add a 168-item check to ask whether any difference is actually tied to source wording. It compares the original wording, paraphrases that preserve the source, statements that remove the source wording, distractor evidence, and no evidence. We compare the EVID–CTRL target/no-evidence difference for the source-preserving paraphrase with the same difference for the source-neutral statement. The models must reach 90% accuracy on target items and 80% on distractor and no-evidence items before we interpret the comparison. Appendix B gives details.

Finally, we ask whether the Base models simply needed more exposure. We choose the negative S3 pair in advance, resume both adapters from their 1M-token optimizer states, continue to 5M tokens, and rerun the same evaluations. Because this uses only one seed, it is a follow-up test rather than a replication or a full study of training length.

For S1, we also let Post use its native thinking mode on the same 552 items. Each item receives one deterministic, balanced A/B/C answer order. Greedy decoding allows 1,536 thinking tokens, a fixed ending that does not favor any answer label when needed, and 256 answer tokens; invalid answers count as wrong. Untouched, EVID, and CTRL use the same item assignments, requirements, and evaluator. Appendix B gives full definitions and results. Because this version generates

	PPL	BLiMP	Map (pre)	Map (post)
EVID	7.29	66.4	92.8	79.8
CTRL	7.31	65.5	38.1	36.0
NOMARK	7.37	65.5	38.8	37.7

Table 2: Custom-pretraining results for seed 42. PPL is TinyStories validation perplexity; other entries are percentages. Map is accuracy on the ko/zo/mu EVID mapping before and after adaptation. BLiMP is a broad check of English language ability.

an answer after a reasoning trace, its scores are not directly comparable to the direct answer-label scores averaged over three orders.

6 Results

6.1 Custom pretraining

At the shared adaptation step, ordinary-English validation perplexity differs by only 1.10% across conditions, and the aggregate scores on 12 BLiMP tasks are also close (Table 2). The models clearly differ in whether they learn the source markers: EVID predicts the correct marker with 92.8% accuracy before adaptation and still reaches 79.8% afterward. CTRL and NOMARK remain near the 33.3% chance level when asked to predict the EVID markers. Those scores do not test whether CTRL learned its own particles. The main result here is that EVID learns and remembers the artificial grammar.

Marker-free results. The single-word test cannot show whether the learned markers affect later evidence use. Every condition prefers high-confidence words on all 90 evidence items and therefore scores 33.3%: the 30 direct items are correct and all 60 reported and inferred items are wrong. Mean direct-minus-reported ratios are -14.03 for EVID, -14.74 for CTRL, and -4.95 for NOMARK. Although EVID is $+0.71$ above CTRL, both models are giving essentially the same answer every time, so we do not treat that difference as meaningful.

The full-sentence test is also inconclusive. Accuracy on simple statements is 51.6% for EVID, 50.5% for CTRL, and 56.3% for NOMARK. Because the models are nearly guessing on these basic items, a small or absent difference on the harder source questions tells us little. On the combined measure of whether the model connects evidence to the right object and adjusts for reliability, the exploratory EVID–CTRL difference is $+0.0059$, with

Model	Contrast	S1	S2	S3	Mean
Base	Binding <i>B</i>	+0.052	+0.026	−0.062	+0.005
	Target/null <i>G</i>	+0.148	+0.179	−0.017	+0.103
Post	Binding <i>B</i>	−0.105	−0.234	−0.066	−0.135
	Target/null <i>G</i>	+0.003	−0.182	−0.044	−0.074

Table 3: Qwen marker-free EVID–CTRL differences. Means are simple averages across S1–S3; with $n = 3$, we do not report a confidence interval across seeds.

a 95% interval of $[-0.0008, +0.0122]$. We do not interpret it because the models fail the simpler test.

Our results show that the models learn the markers and remember much of the mapping after adapting to ordinary English, while remaining similar on perplexity and BLiMP. Our tests cannot tell us whether this training changes source reasoning once the markers are removed.

6.2 Qwen marker learning and main task

Qwen also learns the system. Across three seeds, marker accuracy averages 81.4% for Base EVID and 79.1% for Post EVID, versus 35.9% and 35.4% for their controls. Mean held-out LM loss is also lower for EVID: 2.054 versus 2.114 for Base and 2.075 versus 2.135 for Post.

The visible-marker test shows that the models do more than predict which marker comes next. Compared with CTRL, a conflicting marker shifts EVID’s scores toward the source associated with that marker in every seed (mean +0.029 Base, +0.157 Post), and the other marker checks are also positive throughout. The shift is never large enough to change an answer: every model still follows the source stated in the passage. Visible markers clearly affect the probabilities, but this does not show that the effect survives after the markers are removed or improves reasoning.

The two Base conditions have the same mean accuracy of 93.9% when relevant evidence is present. Its positive S1 result becomes smaller in S2 and reverses in S3 (mean +0.005), so the original result does not repeat across seeds. For Post, EVID is below CTRL in all three seeds (mean −0.135); accuracy averages 93.3% for EVID and 95.0% for CTRL. Because Base and Post differ in their earlier training as well as their behavior here, we cannot say that post-training caused this difference.

6.3 Additional evaluation results

The separate evaluation gives a narrower result. For Base, EVID responds more strongly than CTRL to target evidence compared with no evidence

in all three seeds (+0.249, +0.310, +0.104; mean +0.221). The same is not true when target evidence is compared with evidence about a distractor (+0.163, +0.076, −0.086; mean +0.051). This suggests that Base may have become more responsive to relevant context, but it does not show that the change is specifically about source. For Post, the target-versus-distractor result is negative in every seed, while target-versus-no-evidence changes direction.

The source-wording check gives no positive result. Every Base adapter falls just below the required accuracy for paraphrased target items (83.3–87.5% versus 90%), so we treat the comparison as descriptive only. Even descriptively, the result is negative in every seed (−0.099, −0.106, −0.077; mean −0.094): the EVID–CTRL difference is larger after the source wording is removed, not while it is preserved. Post reaches the required accuracy but shows the same negative direction (mean −0.357).

Giving the preselected S3 pair more training leads to the same conclusion. At 5M tokens, EVID predicts the markers with 93.3% accuracy, while the main binding result is only +0.032. Target-versus-no-evidence grows to +0.465, but target-versus-distractor remains −0.015. Both adapters now meet the accuracy requirements, yet the source-preserving result remains below the source-neutral result by −0.264. More exposure strengthens learning of the markers and some response to relevant context, but not a specific response to source wording.

The other tasks do not show a general improvement in reasoning. The original Base model cannot reliably perform the reliability, multistep, or source-learning tasks, so differences after adaptation are hard to interpret. Post shows a positive EVID–CTRL difference in the reliability-reversal contrast in every seed (mean +0.123). This contrast measures half the change in the log-probability margin between the two states when source reliability is reversed; Appendix B instead reports the suite-wide correct-answer margin (mean +0.061). Both adapted models are much less accurate than the original Post model (73.9% EVID, 71.4% CTRL, 96.7% original). EVID has no multistep advantage, and the original Post model cannot reliably perform the source-learning or copied-source tasks. Across all tasks, EVID never produces a consistent improvement in using source information correctly. Appendix B reports the full results.

Post with native thinking. On S1, both adapted Post models perform well enough on the basic checks for us to interpret every task; valid-output rates are 99.5% for EVID and 99.8% for CTRL. Mean accuracy is 91.2% versus 92.1%. EVID has a small $+0.042$ advantage in joint reliability-reversal accuracy—getting both versions of a reversed source-reliability pair correct. Appendix B reports the suite-wide accuracy difference ($+0.025$). EVID ties CTRL on source binding and both target-evidence comparisons, and is lower on copied-source reasoning (-0.125) and multistep accuracy (-0.010). We see that letting the model think through its answer reveals no consistent EVID advantage.

7 Discussion

7.1 Learning the markers and using them

The clearest result is that the models learn the markers. Custom EVID reaches 92.8% and retains 79.8% after adaptation to unmarked English; Qwen EVID averages about 80%, while controls average 36%. What happens after the markers are removed is much less clear. One Pythia test produces the same answer for every item, and the models nearly guess on the simple controls for the other. Qwen handles the basic tasks, but Base’s main result changes direction across seeds and Post’s is consistently negative.

The visible-marker test gives us additional evidence that Qwen learned more than a next-token association: the markers change its probabilities, especially for Post. They never change the final answer, however, and do not show better reasoning about sources.

We include Qwen to see whether a more capable pretrained model can make better use of the source distinction, not to compare model sizes directly. It differs from Study 1 in architecture, prior training, adaptation method, marker inventory, and interface. To show that model capacity itself causes a difference, we would need a series of models that differ mainly in capability.

7.2 What a stronger response to context means

Base responds more strongly to target evidence than to no evidence in every seed, but the result does not hold when target evidence is compared with evidence about a distractor. The original main result also changes direction across seeds. Most im-

portantly, the follow-up gives us no basis for saying that Base “pays greater attention to sources.” Base narrowly misses one accuracy requirement, and even its descriptive result is larger when the source wording is removed than when it is preserved. The most we can say is that EVID may make Base more responsive to relevant context, not specifically to its source.

Training the S3 pair for 5M tokens makes the same point. The models learn the markers better and now meet the accuracy requirements, but source wording still provides no advantage. A simple explanation that they only needed more examples is insufficient in our view.

Nor does this amount to successful source-sensitive reasoning. The training does not consistently improve judgments of trust, recognition of copied information, calibration, or multistep reasoning. A model can pay more attention to the relevant part of a prompt and still draw the wrong conclusion from it.

7.3 Why reliability may not improve

Our corpus marks how information was acquired, not whether it is correct. It never teaches that direct evidence is always stronger or that reports are unreliable, and either rule would often be false. Better reliability judgments would need to emerge on their own, rather than being taught directly. Li et al. (2026) similarly find that preferring one source does not mean a model can distinguish helpful claims from harmful ones.

One follow-up will manipulate acquisition route and source reliability independently during training. Direct observations, reports, and inferences could come from sources with balanced accuracy histories. The goal would be to test whether grammatical marking helps models learn and apply information about evidentiary reliability without conflating “direct” with “trustworthy.”

7.4 Applying source markers to real fine-tuning data

Synthetic narratives let us know where every claim came from, which makes the experiment easy to control. A more practical version could add source markers to data that a company or research group already plans to use for domain fine-tuning. The markers could distinguish first-party records, reports, retrieved documents, and model inferences without otherwise changing the text. A convincing test would compare meaningful markers with

equally frequent shuffled markers, then ask whether the model better tracks copied information, resolves conflicts, responds specifically to relevant evidence, and expresses appropriate confidence. Source type is still not the same as reliability, and incomplete source information could reinforce existing biases.

7.5 Why Base and Post cannot be compared directly

Base averages $+0.005$ after changing direction across seeds; Post averages -0.135 and remains negative throughout. Both compare EVID with CTRL within the same checkpoint. Base and Post differ in instruction following, use of context, training data, and objectives; Du et al. (2026) likewise find differences between Base and instruction-tuned models. We report Post descriptively, not as evidence that reasoning-oriented post-training caused the result. Native thinking on S1 also gives no consistent EVID advantage.

8 Conclusion

We introduce a matched-corpus method that separates marker acquisition, marker-free contextual effects, and correct use of source information. Both custom-pretrained and Qwen EVID models learn the markers. Qwen also responds when the markers are visible. Yet marker-free gains do not replicate or isolate source wording from relevant context, and no condition consistently improves broader reasoning.

Limitations

Study 1 has only one custom-pretraining seed, while Study 2 has three paired Qwen seeds. Three seeds are enough to show that some results change direction, but not enough to measure precisely how much results vary when the models are retrained. Resampling fictional worlds measures variation across test items for one trained pair, not variation across training runs.

We design the additional evaluation after seeing the first Qwen Post result, although we do so before training or scoring the other two seeds. It includes several related measures, and we do not adjust the statistics for testing many of them. The native-thinking test is also added afterward, covers only S1, and uses one answer order per item. The later comparison of source-neutral and source-preserving wording is another follow-up. All six

1M Base adapters fall below its required accuracy on paraphrased target items, so we report those results only as descriptions. The preselected S3 pair meets the requirement after 5M tokens, but one pair cannot show a reliable effect of additional training.

The two studies do not provide a controlled comparison of model size. The custom models receive only 156.3M synthetic tokens and 13.1M adaptation tokens. Custom CTRL also uses different particles from EVID, so its marker score only tests the EVID mapping; a replication should use the same particles with their source meanings shuffled. The Qwen comparison keeps the underlying checkpoint the same but trains only low-rank adapters. The main runs see just 1.00M tokens—2.70% of the stream—and stop when warmup ends. The S3 follow-up reaches 5.00M tokens, but it does not test a range of training lengths.

The corpora have not received a blinded human review or an exhaustive check for overlap between training and evaluation. In the larger custom-pretraining corpus, about 0.079% of marker strings are malformed. These cases occur in the matched conditions, so they add noise rather than favoring EVID. We record the generator families and prompts, but not every revision made by the providers. Both the particles and the evaluations are synthetic simplifications, and the results should be tested on external tasks validated by people.

Ethics Statement

We collect no data from human participants and use no personal data. Synthetic text can reproduce biases from the models that generated it. A model that reacts more strongly to source cues without evaluating them correctly may become more vulnerable to claims framed as authoritative or to misleading source information. These results are not evidence that the models are safe or well calibrated in high-stakes settings.

Acknowledgments

Generative AI assistance. Anthropic Claude-family models and OpenAI GPT-5-mini generated the synthetic narratives under author-specified constraints; this use is part of the method described in Section 3.1. OpenAI Codex and Anthropic Claude agents assisted throughout the project with the experiments, including refining experimental designs and constructing evaluation items, running and debugging experiments, making scripts to analyze

results, searching for literature, and partially drafting and editing text throughout the manuscript. The author made all the decisions about the methods, analyses, interpretation, and wording, manually reviewed and edited the methods, manuscript, and reported results, and takes full responsibility for the work.

References

- Ayhan Aksu-Koç, Hale Ögel-Balaban, and İ. Ercan Alp. 2009. [Evidentials and source knowledge in Turkish](#). *New Directions for Child and Adolescent Development*, 2009(125):13–28.
- Stella Biderman, Hailey Schoelkopf, Quentin G. Anthony, et al. 2023. [Pythia: A suite for analyzing large language models across training and scaling](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [QLoRA: Efficient fine-tuning of quantized LLMs](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115.
- Kevin Du, Clara Kumpel, Michelle Wastl, and Alex Warstadt. 2026. [It’s not what you say, it’s how you say it: Evaluating LLM responses to expressions of belief](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3137–3151.
- Ronen Eldan and Yuanzhi Li. 2023. [TinyStories: How small can language models be and still speak coherent English?](#) *Preprint*, arXiv:2305.07759.
- Max Glockner, Xiang Jiang, Leonardo F. R. Ribeiro, Iryna Gurevych, and Markus Dreyer. 2025. [NeoQA: Evidence-based question answering with generated news events](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11842–11926.
- Michael Y. Hu, Jackson Petty, Chuan Shi, William Merrill, and Tal Linzen. 2025. [Between circuits and Chomsky: Pre-pretraining on formal languages imparts linguistic biases](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9691–9709.
- Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. [How can we know when language models know? on the calibration of language models for question answering](#). *Transactions of the Association for Computational Linguistics*, 9:962–977.
- Cheng Jiayang, Chunkit Chan, Qianqian Zhuang, Lin Qiu, Tianhang Zhang, Tengxiao Liu, Yangqiu Song, Yue Zhang, Pengfei Liu, and Zheng Zhang. 2024. [ECON: On the detection and resolution of evidence conflicts](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7816–7844.
- Birsu Kandemirci, Anna Theakston, Ditte Boeg Thomssen, and Silke Brandt. 2023. [Does evidentiality support source monitoring and false belief understanding? a cross-linguistic study with Turkish- and English-speaking children](#). *Child Development*, 94(4):889–904.
- Sercan Karakaş and Yusuf Şimşek. 2026. [Benchmarking source-sensitive reasoning in Turkish: Humans and LLMs under evidential trust manipulation](#). *Preprint*, arXiv:2604.24665.
- Mehran Kazemi, Quan Yuan, Deepti Bhatia, Najoung Kim, Xin Xu, Vaiva Imbrasaitė, and Deepak Ramachandran. 2023. [BoardgameQA: A dataset for natural language reasoning with contradictory information](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 39052–39074.
- Shuwei Li, Haoxin Li, Wenda Chu, and Yi Fang. 2026. [How large language models balance internal knowledge with user and document assertions](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 25323–25346.
- Gabrielle Kaili-May Liu and Arman Cohan. 2026. [Can LLMs use linguistic uncertainty markers to reliably reflect intrinsic confidence?](#) *Preprint*, arXiv:2605.28778.
- Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. [Reducing conversational agents’ overconfidence through linguistic calibration](#). *Transactions of the Association for Computational Linguistics*, 10:857–872.
- Anna Papafragou, Peggy Li, Youngon Choi, and Chung-hye Han. 2007. [Evidentiality in language and cognition](#). *Cognition*, 103(2):253–299.
- Qwen Team. 2026. [Qwen3.5: Towards native multi-modal agents](#).
- Xin Sun, Di Wu, Sijing Qin, Isao Echizen, Abdallah El Ali, and Saku Sugawara. 2026. [Label effects: Shared heuristic reliance in trust assessment by humans and LLM-as-a-judge](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32378–32392.
- Ercenur Ünal and Anna Papafragou. 2020. [Relations between language and cognition: Evidentiality and sources of knowledge](#). *Topics in Cognitive Science*, 12(1):115–135.
- Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. [BLiMP: The benchmark of linguistic minimal pairs for English](#). *Transactions of the Association for Computational Linguistics*, 8:377–392.

Jennifer C. White and Ryan Cotterell. 2021. [Examining the inductive bias of neural language models with artificial languages](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 454–463.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. [Navigating the grey area: How expressions of uncertainty and overconfidence affect language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5506–5524.

A Marker Placement and Quality Checks

A.1 Placement rule

We attach markers to verbs that state what happened (not to the phrase that explains how the character learned it). For example, a direct-source passage begins:

Fishermen stood at the harbor’s edge and observed how their nets ko hung in irregular patterns from loading cranes. Equipment ko sat scattered across the dock after sudden rough seas.

In this example, *observed* establishes that the character saw the event, but we do not mark it. The later statements take *ko* because the earlier clause establishes how the information was acquired. We do the same in reported passages: something like *the supervisor told me* establishes the source, and the claims that follow take *zo*. If we marked *saw*, *told*, or *deduced*, the model could simply match a marker to the word beside it.

A.2 Generation and filtering

We generated passages of five to ten sentences, each with one source of evidence. We made sure not to include any explicit confidence vocabulary. After generation, validation scripts check the passage length, source label, marker inventory and count, prohibited vocabulary, repetition, and consistency across transformations. For later batches, we also removed embedding-based near duplicates at cosine similarity .92.

Every variant of our corpus is regenerated from the same base passages. EVID retains the source mapping, CTRL replaces markers at the same positions using the hash algorithm, and NOMARK removes markers. For the Qwen subset, we apply stricter filters, balance source categories within every three-to-eight-marker bucket, and use the same marker inventory for EVID and CTRL.

B Additional Evaluation Set

The evaluation set contains:

Suite	Items
Source binding	72
Evidence sufficiency	120
Reliability reversal	120
Multihop binding	96
Learned calibration	72
Provenance independence	72

Source binding asks the model where the evidence comes from. Evidence sufficiency asks whether the evidence in a passage is about the primary target, an unrelated target, absent, ambiguous, or conflicting. Reliability reversal changes which sources have been accurate in the past. Multistep items require one to three reasoning steps. Learned calibration asks the model to infer whether a new source is reliable. The copied-source task tests whether it distinguishes independent agreement from repeated reports of the same information.

We only interpret an adapted model’s result when the untouched model first shows that it can perform the task. The required scores are .70 for exact binding, .80 for clear target support, .60 for rejecting insufficient evidence, .65 for conflicts with unequal reliability, .65 for learning source reliability, .60 for recognizing independent sources, .70 when every step in a chain is present, and .50 when a premise is missing. The native-thinking test also requires .95 valid outputs and no more than .05 truncation.

B.1 Full results

Base

Suite	Unt.	E	C	Δ_m
Source binding	100.0	98.6	98.1	+.159
Evidence sufficiency	100.0	99.7	100.0	+.235
Reliability reversal	20.8 [†]	32.8	35.0	−.038
Multihop binding	63.5 [†]	75.3	76.0	+.024
Learned calibration	33.3 [†]	33.3	33.3	−.077
Provenance independence	62.5	58.8	62.5	−.004

Post

Suite	Unt.	E	C	Δ_m
Source binding	100.0	100.0	100.0	+.113
Evidence sufficiency	100.0	100.0	100.0	+.105
Reliability reversal	96.7	73.9	71.4	+.061
Multihop binding	91.7	86.5	88.2	−.011
Learned calibration	36.1 [†]	33.3	34.3	−.064
Provenance independence	47.2 [†]	63.0	58.3	+.082

Unt./E/C refer to untouched, EVID, and CTRL. Adapted accuracies and Δ_m —the EVID–CTRL difference in the score for the correct answer—are simple averages across the three paired seeds. The

same untouched checkpoint is used for every seed. † means that the untouched model does not perform well enough for us to interpret the adapted result.

B.2 Post with native thinking

This test uses S1 only.

Suite	Unt.	E	C	Δ_a
Source binding	100.0	100.0	100.0	.000
Evidence sufficiency	100.0	100.0	100.0	.000
Reliability reversal	95.8	88.3	85.8	+.025
Multihop binding	100.0	99.0	100.0	-.010
Learned calibration	43.1	77.8	80.6	-.028
Provenance independence	86.1	81.9	86.1	-.042

Here, Δ_a is the EVID-CTRL difference in accuracy. Both adapted models perform well enough on every task for the comparison to be interpretable. The untouched model falls short on learning source reliability, producing valid answers, and avoiding truncation. Valid-output rates are 90.0% for untouched, 99.5% for EVID, and 99.8% for CTRL.

B.3 Source-wording follow-up

The main comparison asks whether EVID’s advantage over CTRL between target and no evidence is larger when a paraphrase preserves the source wording than when the same claim is stated without a source.

Model	S1	S2	S3	Mean	Gates
Base	-.099	-.106	-.077	-.094	0/6
Post	-.321	-.437	-.313	-.357	6/6

The final column shows how many adapted checkpoints meet the accuracy requirements. Base falls short on paraphrased target items, so we report its result only as a description. Its mean EVID-CTRL target-versus-no-evidence difference is +.164 for the source-preserving paraphrase and +.258 for the source-neutral form. The earlier response difference is not specific to source wording.

B.4 Visible-marker test

For each world and source described in ordinary language, we either omit the marker, use the matching marker, or substitute either of the two conflicting markers. The main measure asks how much a conflicting marker pulls the model toward the source associated with that marker rather than the source stated in the passage, relative to the same passage without a marker, and then compares that pull between EVID and CTRL.

Model	S1	S2	S3	Mean	Spec.	Reinf.
Base	+.003	+.025	+.059	+.029	+.052	+.117
Post	+.156	+.074	+.242	+.157	+.133	+.240

We average specificity and reinforcement across the three seeds. All three measures are positive in every seed. Even so, every model chooses the source stated in the passage on every item. The markers shift probabilities but never change the final answer.

B.5 Additional-training test

We select Base S3 for additional training because its main result is negative at 1M. The paired adapters resume from their saved optimizer states and stop after 5,001,216 presented tokens. The results are shown below.

Dose	E map	Binding	T/null	T/dist.	P-N	Pull
1M	83.9	-.062	+.104	-.086	-.077†	+.059
5M	93.3	+.032	+.465	-.015	-.264	+.055

E map is EVID marker accuracy in percent; the corresponding CTRL scores are 38.6% and 32.3%. T/null and T/dist. compare target evidence with no evidence and distractor evidence. P-N asks whether the result is larger for a source-preserving paraphrase than for a source-neutral statement, and Pull is the visible-marker result. † means that the 1M models fall below the required accuracy; both 5M adapters meet it. Additional training strengthens learning of the markers and the response to some target evidence, but still gives no advantage to source wording.

C Data, Code, and Reproducibility

We will release the corpus and evaluation data together with clean scripts for training and evaluation.