Does the AI Believe the Witness?

Hidden credibility judgments inside a language model reading legal statements, and why they cannot be assumed to hold from one model to the next

Ryan McDonough
Independent researcher
me@ryanmcdonough.co.uk

Working paper, July 13, 2026

Abstract

Language models are moving into legal work: document review, disclosure, case summarisation, and the triage of witness material. This paper asks a plain question with a measurable answer. When such a model reads a witness statement, does it form a view about whether the witness is believable, a view it never states out loud, and is that view swayed by *how* a statement is written rather than *what* it says? Using an open-weights model whose internal workings can be inspected (Qwen 2.5 7B), I show that the model does form an internal credibility signal, that this signal causally drives the reliability score the model gives, and that, on identical facts, the model rates hesitant speech and a regional-dialect rendering about three points lower out of ten than formal English. A prior negative note about a named person also lowers the model's score for that same person's later, identical statement. I then run the behavioural half of the same tests, unchanged, through two closed commercial models reached only by their public interfaces (Claude Sonnet 5 and GPT-5.5). Neither reproduces the pattern, and on the sharpest comparison the direction reverses. The lesson is not that "AI is biased" in one fixed way. It is that this behaviour is specific to each model, can arise by more than one internal route, and must be tested on the actual system in use, from the outside where the inside is closed. All code, synthetic statements, and raw results are released.

Contents

1 Introduction

When an AI system reads a witness statement, does it form a view about whether the witness is believable, a view it never states out loud?

This matters because AI tools are entering legal work: document review, disclosure exercises, case summarisation, chronologies. If the software carrying out those tasks holds an unstated assessment of who is credible, and if that assessment is influenced by how people speak rather than what they say, then a discrimination risk has entered the workflow, and it is invisible to everyone using the tool. Language models are already known to bring systematic biases to tasks where they score or rank text [2]; the open question is whether one such bias operates, unseen, on the credibility of witnesses.

You cannot answer this by asking the AI alone. With an open-weights model, one whose internal workings are published, you can look inside and see whether a credibility judgment is forming. With a closed commercial model you cannot, but you can still test whether its outputs change when you vary how a statement is written while holding the facts constant. This study used an open model so that both were possible, and then put the behavioural half of the tests to two closed models to see whether the open model's behaviour predicts theirs. It does not.

On the open model I tested, it forms internal credibility judgments about witnesses; those judgments attach to the specific person, carry across documents, drive the model's scores and its stated reasoning, and, in the model's actual ratings and not only its internals, identical facts score roughly three points lower out of ten when written in hesitant or dialect speech. When the same behavioural tests were run against two closed models, neither the voice penalty nor the person-bound effect appeared. Both results are reported below, along with what they do and do not mean.

This is one open model, tested with synthetic statements, and the closed-model comparison is behavioural only. The caveats are set out in full in Section 8.

2 Method

This section describes the whole method in plain terms. There is no step in it more complicated than averaging and subtracting. The exact procedures, down to the line of code, are in the released repository (Section 9); this paper does not restate them as formulae.

2.1 Models and inference settings

The open model is `Qwen/Qwen2.5-7B-Instruct`, a 7-billion-parameter instruction-tuned model with published weights, run in a memory-saving 4-bit mode on a single GPU. All of the open-model measurements come from one run (4 July 2026). Decoding is deterministic: the model is asked once per prompt with no randomness, so every number reproduces exactly.

The two closed models are Anthropic's `claude-sonnet-5` and OpenAI's `gpt-5.5`, each reached only through its public interface on 4 July 2026. They were run at a low reasoning setting with a short answer limit, applied identically to every statement and every condition, so those settings cannot themselves manufacture a difference between the cases being compared. A larger open model (14B) is being tested separately and is not reported here.

The recorded date for the closed models is part of the result: commercial models can change without notice, so a finding about them is a finding about the model as it stood on that date.

2.2 Materials

Every stimulus is synthetic and generated from fixed templates; no real case data was used at any point. Three sets of statements were built.

- **Contrastive pairs** (200 pairs). Two versions of each pair are word-for-word identical except one clock time. In one, a witness leaves 25 minutes after arriving; in the other they "leave" 40 minutes *before* they arrive, while the sentence still reads "roughly twenty-five minutes after I arrived". There are no hedging words and no verdict words, so the only way to tell the two apart is to do the arithmetic on the timeline.
- **A number-detector control** (200 pairs). Two perfectly consistent statements differing only in an innocent duration, a 25-minute visit versus a 55-minute one. A genuine consistency signal should be blind to these.
- **A voice set** (60 statements in each of five voices) and a **two-document set** (60 items), described with the findings that use them.

2.3 Extracting the credibility signal

As a language model reads text it builds up an internal pattern of activity, a large array of numbers that represents what it currently makes of the text. On an open model this pattern can be recorded

at any point. This is a recorded representation, not a legible train of thought, and nothing below treats it as the model's conscious reasoning.

To build the credibility signal I recorded that internal pattern as the model read each version of every contrastive pair. I then took the average pattern across all the credible statements, and separately across all the not-credible ones, and subtracted one from the other. What remains is a single internal direction that points from "not credible" toward "credible". I call it the credibility direction. The whole calculation is averaging and subtracting; there is nothing else in it. Difference-of-means directions like this have been used to read and steer other traits inside language models [5].

To check the direction is real, I used it on a separate, held-out set of pairs the direction had never been built from, and asked whether it could tell the consistent statements from the inconsistent ones. It ranked the consistent account as more credible 96 times in 100 (AUROC 0.962, with a margin of roughly 0.93 to 0.99).

2.4 Testing causation by steering

Separation shows the direction *tracks* credibility. It does not show the model *uses* it. The test was as follows. While the model read a fresh, neutral statement it had never seen, I nudged its internal activity a set amount along the credibility direction, toward "credible" or toward "not credible", and then asked it: on the evidence in this statement alone, rate the witness's reliability from 1 to 10. If the score moves when the direction is nudged, the direction is doing work in the judgment. Adding a fixed internal direction during the model's forward pass in this way is known as activation steering [4].

I read the answer two ways. First, the number the model wrote. Second, and before it wrote anything, how strongly it leaned toward answering "Yes" versus "No" to "should this account be relied upon?". The second reading cannot be corrupted by garbled text.

2.5 Controls

Results like these are easy to produce by accident, so the study includes several safeguards, each able to overturn the finding.

- **Placebo nudges.** Alongside the real direction I nudged the model equally hard in eight randomly chosen internal directions, and, harder to beat, in eight *decoy* directions built by the exact same averaging-and-subtracting recipe but from deliberately scrambled labels, so they carry no credibility signal. For the parsing-free "Yes/No" reading, which needs no text generation and is therefore cheap, I added 40 further decoys, 56 controls in all.
- **The number-detector trap.** The direction was projected onto the innocent-duration pairs; a genuine consistency signal should barely separate them.
- **Discarding broken answers.** Pushed hard, any model stops writing sense. Incoherent answers were screened out and counted; nothing rests on them.
- **No self-marking.** The place inside the model to intervene was chosen on one set of statements and every reported number comes from a separate, untouched set.
- **Everything logged.** Every answer at every setting is saved and published.

2.6 Behavioural tests

These need no access to the internals, which is what let me repeat them on the closed models. In the **voice test** I wrote the same facts in five voices (formal English, terse, translated-sounding, regional dialect, and hesitant) and asked the unsteered model to rate each one, 60 statements per voice. In the **persistence test** I gave the model two documents: a prior note about a named person (favourable or damning), then a neutral statement by that same person that is word-for-word identical across every condition, and asked it to rate the second document alone. Controls placed the prior note about a *different* person, and swapped the order of the two documents; order is worth controlling because models are known to weight the start and end of a context more heavily than the middle [3].

2.7 Statistical reporting

Three points a careful reader is entitled to.

- **The separation score** (0.962 above) is just the chance that the direction ranks a consistent account above an inconsistent one; 0.5 would be a coin toss.
- **The placebo comparison has a hard floor.** With sixteen fully generated controls, the best possible result is that the real direction beats every one of them, and that outcome corresponds to a significance value of 1 in 17, about 0.059. This design simply cannot return a smaller number, so it cannot cross the usual 0.05 line however clean the result. I report this as a limit of the test, not as a near miss. The cheaper "Yes/No" reading, which carries 56 controls and so has room to go lower, does *not* clear its controls; that is reported too.
- **This was not a pre-registered study.** The hypotheses were fixed in advance, but several analysis choices were revised across earlier runs as flaws were found, and every one of those runs remains published. Where a choice could flatter a headline number, the safeguard against it is stated.

Confidence intervals accompany the behavioural numbers, and the voice and persistence comparisons use permutation tests, which make no assumption about the shape of the data. The exact tests are in the code.

3 Finding 1 — An internal credibility signal, and its causal effect on the rating

With no interference, the model rated the neutral test statements at **5.3 out of 10** on average, a sensible middling answer with balanced reasoning. Then I nudged the credibility direction, the one built purely from whether the facts add up. The score slid with the size of the nudge, in the right direction, at every step: from 5.3 with no nudge, down to 1.0 under the strongest push toward "not credible", and up to 8.0 under the strongest push toward "credible". The model's written reasoning changed to match. Nudged down, it recharacterised the same text as vague or memory-biased, and at the strongest setting asserted a contradiction that is not in the statement. It was not merely producing a different number; its assessment of the same words had changed.

Voice (identical facts)	Model's average rating out of 10
Formal English	5.8
Terse	4.9
Translated-sounding	4.2
Regional dialect	2.9
Hesitant	2.8

Table 1: The open model's own unsteered ratings by voice, 60 statements per voice, facts held constant.

Of 3,288 answers generated across the whole test, 70 (about 2%, almost all under the placebo directions at the harshest settings) were incoherent enough to discard; every figure above uses only the surviving answers.

The placebo results are where this run was hardest on itself, and the verdict is split. On the *written scores*, the real direction moved the rating further than all sixteen controls, the strongest of which managed 4.5 points against the real direction's 5.7. But, as noted above, sixteen controls put the best attainable significance at 1 in 17, so the result is resolution-limited rather than conventionally significant. On the parsing-free "Yes/No" reading, the real direction moved the internal leaning 11.7 units in perfect order, but seven of the fifty-six decoys moved the model further than the real direction did (rank $p = 0.14$), so there the direction is *not* distinguishable from a strong decoy, and no claim is made.

Finally, on the number-detector trap, the direction scored 0.63 where 0.5 would mean it ignores innocent numbers entirely and 0.96 is its score on genuinely contradictory pairs. So the direction is chiefly reading whether the account is consistent, with a residue that reacts to numbers as such. I note that, rather than claiming a clean "does it add up?" detector.

Even at the strongest settings the written answers stayed measured on the surface while the scores swung end to end. The model's stated reasoning does not necessarily reveal the internal representation that is moving its assessment.

4 Finding 2 — Linguistic register shifts the rating on identical facts

This is the finding with direct discrimination consequences. I wrote the same facts in five voices and asked the unsteered model to rate each. On identical facts (Table 1), a witness who says the same thing in regional dialect or in hesitant, disfluent speech is rated roughly three points lower out of ten than one who says it in formal English. Translated-sounding grammar costs about a point and a half. Nothing in the facts differs.

There is a second, subtler result. The internal credibility direction from Finding 1 does *not* predict which voices get marked down. Hesitant speech projects as the *most* credible voice on that internal direction, yet receives nearly the *lowest* rating when the model is asked directly. An audit that checked witness statements against that one internal direction, and found hesitant speech unpenalised, would have been falsely reassured. So the credibility direction, though genuinely causal, is one route by which voice reaches the judgment, not the whole of it. Behavioural testing found what inspecting the most obvious internal signal did not.

What the model read before the identical Statement B	Score for B
A damning note about the same person	1.3
A favourable note about the same person	4.3
Nothing (Statement B alone)	5.3

Table 2: The open model's rating of one identical statement, by what preceded it.

Why this matters legally: dialect, interpreter-mediated grammar, hesitancy and brevity are not evenly spread across society, and they track the witnesses a fair process is meant to protect. That language models can treat non-standard dialects prejudicially has been documented before [1]; what is new here is a matched-facts measurement on one model and, in Section 6, the finding that the effect does not carry from one model to another. The model rates identical evidence differently depending on the voice it is written in; nobody chose those penalties, and nothing in the output reveals them. Each voice here is one stylised rendering, not a real sociolect (Section 8); the effect is real in the model's behaviour, and generalising it to any real community would need real, consented transcript material.

5 Finding 3 — Characterisation persists and attaches to the person

The last experiment mimics a real litigation file: several documents about the same person, read together. Document A is a prior note saying a named person had given either a consistent, corroborated account or a shifting, contradicted one. Document B, identical in every test, is a neutral statement by that same person. I asked the model to rate Document B alone.

Same words; a score of **1.3 versus 5.3**, depending entirely on what the model read about the person first (Table 2). The effect is asymmetric: the damning note collapses the score by four points, while the favourable note does not lift it at all. This is a taint, not a symmetrical halo.

Is the taint bound to the *person*, or does a damaging document colour the reading of the whole file? I re-ran it with Document A about a different, unrelated person. Some of that effect carries over, but only about a quarter as much, and the difference is large and clear: the damning characterisation is overwhelmingly bound to the named individual. A prior read of the person carries forward within the file and attaches to that person; this is persistence within a single reading, not a stored memory.

6 Closed-model replication

Most legal AI tools run on closed models reached through an interface, where looking inside is impossible. The voice and persistence tests need no internal access, so I ran both, unchanged, through Claude Sonnet 5 and GPT-5.5. This is the exact position of a firm using a closed tool: outputs only.

Neither closed model reproduced the open model's pattern, and on the sharpest comparison the direction reversed. On the voice test (Table 3) both closed models rated all five voices in a much narrower band, and both rated regional dialect *higher* than formal English, the opposite of the open model. Hesitant speech, which the open model punished hardest, was essentially

Voice (identical facts)	Open model	Claude Sonnet 5	GPT-5.5
Formal English	5.8	3.7	5.8
Terse	4.9	4.0	5.6
Translated-sounding	4.2	3.9	5.7
Regional dialect	2.9	4.1	6.2
Hesitant	2.8	3.7	5.8

Table 3: Average rating by voice across the three models. The open model's dialect and hesitant penalties do not carry over; both closed models slightly favour dialect.

level with formal on both. Voice still moved the score, so behavioural testing remains necessary, but which voices are favoured, and in which direction, differed on every model tested.

The persistence taint did not appear either. On both closed models a damning prior note produced an equal or *higher* rating for the identical statement than a favourable one, the opposite direction to the hypothesised effect. Documents about *other* people still moved the score, though: on Sonnet, a glowing note about an unrelated stranger dragged the witness down to 3.92 (against a 4.75 baseline), while a damning note about a stranger pushed the witness up to 5.60. The witness gains or loses by comparison with whoever else is in the file. So bundle composition still matters on the closed models; the mechanism is comparison between people rather than a taint bound to one person.

Three things follow. This is *not* evidence the closed models are safe: their scores still moved on the voice a statement is written in and on what else was in the file, which is different behaviour, not the absence of behaviour. Open-model findings do *not* transfer: neither the direction nor the size of any effect carried across. And there is no closed-model version of Finding 1, because the internal signal cannot be inspected from the outside; the behavioural battery is the whole of what a customer can check.

7 Implications for legal practice

- **A tool's stated reasoning is not its assessment.** The two can come apart. "Does the tool make credibility judgments?" cannot be answered by reading its outputs; it needs either inspection of the internals, where the model is open, or structured behavioural testing, where it is closed.
- **Voice is a demonstrated discrimination risk on the model where it was shown.** On identical facts the open model marked down hesitant speech and a stylised dialect rendering by about three points, with clear implications for procedural fairness and potential Equality Act 2010 concerns where linguistic features correlate with protected characteristics. A model-level difference is not by itself unlawful discrimination, but the risk is real, measurable, and, as the closed models show, specific to each system.
- **Auditing one internal signal is not enough.** The most natural internal check would have cleared the very voice the model penalised most.
- **What else is in the file matters.** On the open model a prior read of a person persists and attaches to that person; on the closed models the score moves by comparison between people. Either way, bundle composition and ordering are fairness-relevant decisions.

- **The testing duty splits by access, and it is dischargeable.** Where weights are open, test the internals and the behaviour. Where they are closed, run matched behavioural batteries on the model actually deployed and repeat them when it changes. I ran exactly such a battery, unchanged, through two commercial interfaces, which shows the testing is practicable.

8 Limitations

1. **One rendering per voice.** Each voice is a single template with the facts filled in, so "dialect is marked down" means, strictly, that this one stylised rendering is marked down. Several independent renderings of each voice are the most important next step and would separate the feature from one synthetic version of it.

2. **Overlap between the corpora.** The voice and persistence statements reuse the fact skeletons of the contrastive pairs (in different wording; the steering targets are kept separate). Where they overlap, the internal projection in Finding 2 could be reacting to familiarity as much as credibility. The behavioural ratings do not depend on the internal direction and are unaffected; the discrimination claim rests on them.

3. **One open model, one memory-saving mode.** Every internal measurement is from a single 4-bit 7B model. A larger open model is being tested separately.

4. **The placebo test is resolution-limited.** On the written scores the best attainable significance is about 0.059, set by the number of controls; on the reading that had room to go lower, the direction did not clear its controls.

5. **Synthetic statements throughout.** No real case data was used, deliberately. Quantitative claims about any real community need consented, anonymised transcript material.

6. **Closed-model settings.** The closed models were run at a low reasoning setting with a short answer limit, applied uniformly, so the settings cannot create a difference between the cases being compared; a high-effort sweep is not reported.

9 Future work

In order: a run on a larger open model, reported whichever way it goes; several independent renderings of each voice, and non-overlapping fact skeletons for the voice and persistence sets; a wider closed-model battery across more providers, repeated as those models are updated; and then, with appropriate consent, real transcript material in place of synthetic statements.

Ethics statement

This work uses only synthetic, deterministically generated statements; no real case data or personal data was used. The voice test varies linguistic surface features only and never attaches protected-characteristic labels to any individual; the link between those features and protected characteristics is discussed as a societal risk, not asserted of any person. The aim is diagnostic, to show that a model can form and act on such judgments unbidden, and not to build a tool for assessing witnesses,

which would be an inappropriate use. Because the findings differ by model, no single number here should be read as a property of "AI" in general.

Data and code availability

The full code, the deterministic statement generators, and the raw results (including every model answer at every setting) are released so that any figure in this paper can be traced to an artefact and reproduced. The repository is at `https://github.com/ryanmcdonough/pcp-replication`. The interactive write-up and downloadable result bundles are at `https://papers.ryanmcdonough.co.uk/believe/`. The source code is released under the MIT licence; the generated statements, results, and this paper are released under the Creative Commons Attribution 4.0 International licence (CC BY 4.0). The open-model results are from a single run directory and the closed-model results from a single bundle, both dated 4 July 2026.

References

[1] V. Hofmann, P. R. Kalluri, D. Jurafsky, and S. King. AI Generates Covertly Racist Decisions About People Based on Their Dialect. *Nature*, 633(8028):147–154, 2024. doi:10.1038/s41586-024-07856-5. Preprint: arXiv:2403.00742.

[2] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2023. arXiv:2306.05685.

[3] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics (TACL)*, vol. 12, 2024. arXiv:2307.03172.

[4] N. Rimsky, N. Gabrieli, J. Schulz, M. Tong, E. Hubinger, and A. M. Turner. Steering Llama 2 via Contrastive Activation Addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers*, pp. 15504–15522, 2024. arXiv:2312.06681.

[5] R. Chen, A. Arditi, H. Sleight, O. Evans, and J. Lindsey. Persona Vectors: Monitoring and Controlling Character Traits in Language Models. arXiv:2507.21509, 2025.