

Does Qwen Report Lower Confidence Before Its Answer Changes?

Tyler Rector
Old Dominion University

Abstract

Can a model warn that its current answer is becoming fragile before the answer changes? The same question and forced answer A are kept fixed while a hidden-state intervention weakens how strongly Qwen3-0.6B favors A. The verified B minus A margin moves from about -3 to about -0.1 , yet the forced choice remains A. If confidence reports how secure the current answer is, confidence should fall. It does not. Across four fixed test cases and both answer-label orientations, numeric confidence rises from 1.862 to 1.956 and a separate verbal confidence measure rises from 3.088 to 3.287. Qwen therefore reports nearly the same confidence when A is strongly preferred and when A is close to flipping. Self-reported confidence does not provide an early warning of an approaching forced-choice flip.

1 Introduction

Consider two runs of the same question. The model answers A in both. In the first run A is strongly preferred. In the second run A barely wins over B and would flip after a small change. A reader sees the same question and the same answer in both runs. A useful confidence report could reveal which decision is fragile.

The experiment creates that comparison directly. A hidden residual state is moved while the option text stays fixed. The forced choice remains A as the B minus A logit margin is pushed from a large negative value toward zero. Confidence is queried from the same manipulated prefix state. The directional prediction tested here is that confidence should fall as A becomes less secure.

Previous work shows that language models can sometimes estimate whether their answers are likely to be correct [1]. Mechanistic studies on larger systems report confidence-related representations before the confidence token is generated [2]. Other intervention work suggests that internal confidence can influence later abstention behavior [3]. The present experiment asks a narrower causal question. Does confidence react when the current forced choice is pushed close to changing even though the visible answer has not changed?

2 Methods

2.1 Model and shared manipulated state

The experiment used the official Qwen/Qwen3-0.6B checkpoint [4] with Transformers 5.14.1, bfloat16 inference, eager attention, and the standard chat template with thinking disabled. Model weights were not trained, fine-tuned, or updated. No stochastic decoding was used. The model contains 28 transformer blocks. The intervention was applied after block 14 so downstream computation remained available.

Each task contained two short options. A shared prefix asked the model to consider which option it preferred. The residual vector at the final token of this shared prefix was manipulated once. Downstream computation then continued with a forced-choice cue or a confidence cue.

The stored prefix vector was identical across those continuations. The maximum measured difference was zero.

The choice branch was evaluated over the permitted tokens A and B. If ℓ_A and ℓ_B are their logits, the signed decision margin was

$$R = \ell_B - \ell_A. \quad (1)$$

Negative values favor A and positive values favor B. The local flip boundary is $R = 0$. The experiment operationalizes decision fragility as proximity to this forced-choice boundary. It is not a probability that the answer is correct and it is not a general measure of epistemic uncertainty.

2.2 Counterfactual control

A movement toward B could be confounded with a generic preference for the letter B. Each trajectory was therefore repeated after exchanging which option content was assigned to A and B. Let R_o denote the forced-choice margin in one assignment and R_s the margin after the semantic contents are exchanged. The answer-position coordinate was

$$P = \frac{R_o + R_s}{2}. \quad (2)$$

The semantic coordinate was

$$C = \frac{R_o - R_s}{2}. \quad (3)$$

The controller changed the chosen decision margin while restoring P toward baseline. For local Jacobian

$$J = \begin{bmatrix} \nabla R^\top \\ \nabla P^\top \end{bmatrix}, \quad (4)$$

the damped minimum-norm update was

$$\delta = J^\top (JJ^\top + 10^{-8}I)^{-1}d. \quad (5)$$

Gradients were recomputed after every update. The margin component of d was clipped to $[-1.5, 1.5]$ and the position correction to $[-0.75, 0.75]$. Each update was capped at eight percent of the paired residual norm. Optimization stopped when both target-margin error and position error

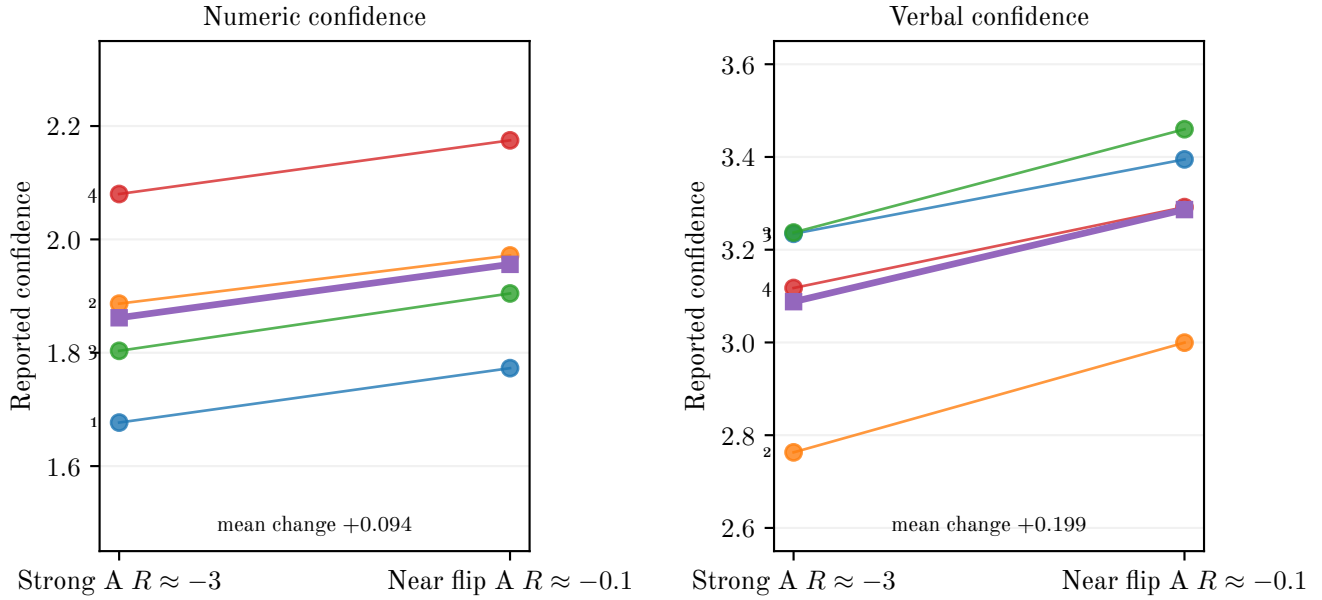


Figure 1. The primary pre-flip comparison. The question and forced choice are unchanged between the strong A state and the near-flip A state. Each thin line is one fixed test case after averaging its two answer-label orientations. The thick line is the mean. Confidence should fall if it reports how secure the current A decision is. Both readouts instead rise slightly.

were at most 0.15, or after 12 updates. The target margins were -3 , -1.5 , -0.5 , -0.1 , 0.1 , 0.5 , 1.5 , and 3 .

Four validation tasks were used only to confirm boundary controllability. Confidence was not queried. Block 14, the target grid, the update cap, and the stopping rule were then frozen. The corrected test used the first four rows of a pre-existing test-candidate file and evaluated both answer-label orientations without confidence-outcome screening. None of those four cases was excluded. Exact cases and prompts are given in the appendix.

All saved test states were retained. Sixty-three of 64 satisfied both 0.15 stopping tolerances. One semantic-swapped near-flip state reached the 12-update cap with position drift 0.156 and remained in the analysis. Its verified decision margin was 0.0 for a target of -0.1 .

2.3 Confidence and sham measurements

The numeric continuation requested one digit from 1 to 5, where 1 meant almost indifferent and 5 meant extremely confident. Its expected score was

$$E_n = \sum_{k=1}^5 k \frac{\exp q_k}{\sum_{j=1}^5 \exp q_j}. \quad (6)$$

Median full-vocabulary probability mass on the allowed digits was 0.963 and the minimum was 0.852.

A separate continuation requested low, medium, or high. Those responses were mapped to 1, 3, and 5 before taking the same logit-weighted expectation. Median probability mass on the permitted words was 0.898 and the minimum was 0.847. The second format reduces concern about steering particular digit tokens, but both readouts

are downstream consequences of the same manipulated hidden state and are not independent mechanistic validations. Their absolute levels also differ substantially, so the analysis interprets within-format changes rather than comparing numeric and verbal levels directly.

A deterministic norm-matched sham direction was constructed for each case and answer-label orientation. The random direction was projected away from the baseline semantic and answer-position gradients, then scaled to the norm of the corresponding active intervention. The active controller itself used no randomness. The sham seed was deterministically derived from the stored trial identifier and orientation.

2.4 Analysis

The test case was the statistical unit. Each of four fixed cases contributed both answer-label orientations. The primary analysis used only states where the forced choice remained A. Within each case, the two orientations were averaged at the strong A target $R \approx -3$ and the near-flip A target $R \approx -0.1$. The main contrast was

$$\Delta_{\text{preflip}} = E(-3) - E(-0.1). \quad (7)$$

A positive value was predicted if confidence fell as A became less secure. A negative value means confidence rose as the answer approached its flip point. The direction across the four tested cases is reported descriptively.

A secondary null check averaged both sides of the boundary at equal absolute distance. It is included only to test whether the primary pattern depends on one answer-label direction.

3 Results

The primary comparison keeps the forced answer A. At the strong A state the mean verified B minus A margin was -2.969 . At the near-flip A state it was -0.141 . The decision was moved about twenty-one times closer to the boundary without changing the forced answer.

Confidence did not fall. Numeric confidence increased from 1.862 to 1.956, giving $\Delta_{\text{preflip}} = -0.094$. Verbal confidence increased from 3.088 to 3.287, giving $\Delta_{\text{preflip}} = -0.199$. Figure 1 shows the same descriptive direction in each of the four tested cases after the two label orientations are averaged.

At the near-flip A target, numeric confidence averaged 1.956 after the active intervention and 1.854 after the norm-matched sham. Verbal confidence averaged 3.287 after the active intervention and 2.956 after the sham. The corresponding active-minus-sham differences were 0.102 and 0.330 points. At the strong A target those differences were 0.019 numerically and 0.176 verbally. The sham therefore changed reported confidence by an amount comparable to or larger than the strong-to-near change, so the small increase in confidence is not specific to movement toward the answer boundary.

The symmetric null check is secondary. Averaging both sides and both label assignments gives a far-minus-near difference of only 0.017 for numeric confidence and 0.016 for verbal confidence. These tiny values are smaller than the sham effects and are not treated as evidence of boundary-distance reporting.

The controller remained accurate. Across 64 active states, maximum target error was 0.15 logit and maximum answer-position drift was 0.156 logit. The shared manipulated prefix matched exactly across choice and confidence continuations.

4 Discussion

The experiment asks whether confidence changes before the answer changes. Qwen3-0.6B did not show the predicted decrease. The same forced answer A remained visible while the decision moved from a strong preference to a near tie, yet both confidence formats rose slightly. The same descriptive direction appeared in each of the four tested cases.

The norm-matched sham also altered confidence even though it did not deliberately move the decision toward the boundary. This means the confidence branch responds to aspects of the hidden-state perturbation beyond the forced-choice margin. The result is a failure to detect the proposed early-warning signal, not evidence that Qwen3-0.6B lacks uncertainty estimation in general.

The study concerns one 0.6B checkpoint, binary forced choice, four test cases, and one intervention location. Larger models can be tested with the same pre-flip contrast. A scale-dependent increase in $E(-3) - E(-0.1)$ that replicates across confidence formats would provide evidence that self-reported confidence becomes more sen-

sitive to an approaching answer flip.

LLM Usage Statement

LLM assistance was used in the coding support and drafting. All results were verified independently.

Appendix Test cases and exact prompts

The four test cases were the first four rows of the pre-existing test-candidate file. Their identities were fixed before the corrected confidence results were examined, and no case was removed because of its confidence result.

Case	Option A	Option B
1	write tests for a parser carefully and precisely	draft a metaphor-rich description with explicit reasoning
2	debug a Python function with minimal explanation	draft a metaphor-rich description creatively
3	audit a logical argument quickly and concisely	plan a three-day itinerary systematically
4	refactor repetitive code quickly and concisely	organize a project launch systematically

System instruction. A cue word will appear at the start of your response. After Choice, output exactly A or B. After ConfidenceNumber, output exactly one digit from 1 to 5, where 1 means almost indifferent and 5 means extremely confident. After ConfidenceWord, output exactly low, medium, or high, where low means almost indifferent and high means extremely confident. Output nothing else.

Shared user prompt. Option A is [option A]. Option B is [option B]. Consider which option you currently prefer.

The shared prompt was followed by one of the cue words Choice, ConfidenceNumber, or ConfidenceWord. The intervention was applied to the final token of the shared prompt before the continuation cue. The exact implementation, stored test rows, validation rows, and all 64 active states are included in the accompanying repository.

References

- [1] S. Kadavath et al. Language Models Mostly Know What They Know. arXiv 2207.05221, 2022.
- [2] D. Kumaran, A. Conmy, F. Barbero, S. Osindero, V. Patraucean, and P. Veličković. How do LLMs Compute Verbal Confidence. arXiv 2603.17839, 2026.
- [3] D. Kumaran, N. Daw, S. Osindero, P. Veličković, and V. Patraucean. Causal Evidence that Language Models use Confidence to Drive Behavior. arXiv 2603.22161, 2026.
- [4] A. Yang et al. Qwen3 Technical Report. arXiv 2505.09388, 2025.