Can a Model Catch Its Own Hallucinations for Free?: Label-Free Doubt Signals Hold Their Own Against a Labelled Dataset for Abstention

Ali Asaria
Transformer Lab

Tony Salomone
Transformer Lab

Deep Gandhi*
Transformer Lab

Abstract

Large language models state false facts as fluently as true ones, yet a model often "knows" internally when it is on shaky ground: the probability it assigns to its own answer tends to dip on the facts it gets wrong. The usual way to act on this, teaching a model to abstain rather than guess, requires a labelled dataset of right and wrong answers. We ask whether the model's *own* confidence, which is free and needs no labels, can do that job instead. We fine-tune each model (with LoRA) to answer when its frozen confidence is high and to say "I'm not sure" when it is low, using the signal alone and no correctness labels. Across six open-weights models (1B–8B, two families) on short-form factual question answering, with correctness adjudicated by an independent judge model, this label-free recipe holds its own against label-supervised abstention-tuning: at matched coverage we find no statistically detectable difference between the two. A control that drills hard examples instead of abstaining does not help, indicating the gain comes from *calibration*, not rote memorization. The signal's one blind spot is *confidently wrong* facts, which it cannot flag. A model's own doubt is thus a near-free substitute for a labelled dataset when teaching it when to abstain. Code and artifacts are available on request.

1 Introduction

Language models hallucinate: they assert false facts as confidently as true ones, and a fluent surface gives a reader no signal that a claim is unsupported. A growing body of work observes that a model's *internal* state nonetheless carries information about its own correctness, in particular that token-level probabilities are higher on facts the model gets right than on facts it gets wrong. We confirm this on our own models: the mean log-probability over an answer span separates correct from incorrect answers with AUROC 0.73–0.86, holding even for a 1B model.

If a model already "knows when it is unsure," the natural question is whether it can be taught to *act* on that knowledge, to abstain on its low-confidence facts instead of asserting them. One could simply threshold the frozen signal at inference time as an external gate; we instead *internalize* the behavior into the weights so the model abstains natively, which needs no external scorer at deployment and lets abstention interact with generation. The usual way to teach abstention is supervised: collect correctness labels for a training set and train the model to decline the questions

*Corresponding author: deep@lab.cloud

it gets wrong. But labels are expensive: they require knowing the right answer to every training question. We ask a cheaper question: can the model's *own confidence signal* (which is free, already present, and requires no labels) play the role of those labels?

We study *signal-gated abstention fine-tuning*. We compute, once, the frozen base-model confidence over each training question's answer span; threshold it; and fine-tune the model so that low-confidence questions are retargeted to "I'm not sure" while high-confidence questions keep the model's own answer. No correctness label ever enters this procedure. We compare it head-to-head against the label-supervised version of the same recipe and against standard fine-tuning, an up-weighting control, and the untrained base, on a 2×3 model grid, with correctness adjudicated by an independent open-weight judge and significance assessed by bootstrap confidence intervals.

Contributions.

1. **A label-free recipe for abstention.** We introduce *signal-gated abstention fine-tuning*, which uses a model's own frozen token-probability confidence, with no correctness labels, to teach it to abstain on the facts it is internally unsure of. We first verify the premise the recipe rests on: this confidence discriminates the model's own hallucinations across six models and two families, even at 1B scale (§5).

2. **Free confidence holds its own against a labelled dataset.** At matched coverage and adjudicated by an independent judge, the label-free recipe is competitive with label-supervised abstention-tuning: we detect no difference on any of six models (§5).

3. **The gain is calibration, not memorization.** An up-weighting control that drills the same hard questions without abstaining does not reduce hallucination, isolating calibration as the mechanism (§5).

4. **An honest account of the limits.** We map where the method fails: confidently-wrong facts the signal cannot flag, and the rare entities where hallucination concentrates (§6).

2 Related Work

Internal confidence signals. A line of work reads a model's own internal state to detect hallucination: token-probability-based hallucination detectors [18, 16], and surveys of factual-confidence estimators that caution that raw sequence probability is a comparatively weak estimator on QA [14]. Kumaran et al. [12] argue that first-order token log-probabilities have intrinsic blind spots and propose verbalized or activation-based alternatives. Our results are consistent with both observations: the log-probability signal is discriminative enough to be useful, yet its residual failures are exactly the confidently-wrong cases it cannot see.

Teaching abstention and refusal. A complementary line fine-tunes models to abstain at their knowledge boundary, typically using correctness signals, labels, or rewards: refusal-aware tuning and refusal tokens [8], knowledge-boundary abstention via reinforcement learning [3], and training-free conformal abstention [21]. These methods rely on correctness signals, labels, or rewards to decide *where* to abstain; we ask whether the model's own confidence can supply that supervision for free.

Internalizing confidence by training. Several methods bake a confidence or abstention behavior into the weights: confidence tokens and routing [2], confidence tuning for cascades [17], differentiable

calibration losses [11], and reinforcement learning with proper-scoring rewards [19]. The closest to ours are self-supervised: distilling a model's own (sampling- or self-evaluation-based) uncertainty into verbalized confidence [6, 1] and self-distillation that restores calibration with little external labeling [20]. We differ in the specific, cheap recipe: a *frozen, single-pass token-probability* signal thresholded into a binary abstain/answer target, and in a controlled, coverage-matched comparison to the *label-supervised* version of the same recipe. Kaplan et al. [10] document that standard fine-tuning can induce factual forgetting; our standard-fine-tuning control reproduces this neutral-to-slightly-worse behavior. Finally, Liu et al. [13] caution that calibration metrics and faithful uncertainty expression can diverge; we therefore report selective risk rather than relying on expected calibration error.

3 Method

The confidence signal. For a question q, the model greedily decodes an answer $a = (t_1, \ldots, t_m)$. We define its confidence as the mean log-probability the model assigns to its own answer tokens,

$$s(q) \;=\; \frac{1}{m} \sum_{i=1}^{m} \log p_\theta(t_i \mid q, t_{<i}) \,,$$

computed *once* from the frozen base model before any fine-tuning. Higher s means the model concentrated probability on its answer; lower s means it hedged internally even while producing a fluent answer.

Signal-gated abstention fine-tuning (ours, "C3"). On a training set we rank questions by $s(q)$ and pick the lowest-s fraction (the abstention threshold, a coverage knob). For those low-confidence questions the supervised target becomes a fixed abstention string ("I'm not sure."); for the rest the target is the model's *own* greedy answer (self-distillation). We fine-tune with LoRA (r=16). **No correctness label is used**: the frozen signal alone decides where to abstain.

Comparators (same recipe, different supervision).

- **Base**: the untrained instruction model.
- **Standard SFT ("C1")**: LoRA on the model's own greedy answers everywhere (no abstention, no signal).
- **Label-supervised abstention ("C2", R-Tuning style [23])**: identical to ours except the abstain/answer decision uses the *correctness label* (abstain where the base model was wrong, by gold+alias match) instead of the signal.
- **Up-weighting control ("C4")**: instead of abstaining on low-signal questions, *up-weight* their loss toward the answer ("memorize harder"). If C4 matched C3, the gain would be recall, not calibration.

4 Experimental Setup

Models. Six open-weights instruction models forming a 2×3 grid (family $\times$ scale): Llama-3.2-1B, Llama-3.2-3B, Llama-3.1-8B [5]; Qwen3-1.7B/4B/8B [22] (run non-thinking for answer-span comparability). All fine-tuned with LoRA [7] (r=16); one training run per cell.

Data. A factual short-question-answering set of 700 development + 700 test questions drawn from PopQA [15] (MIT) and TriviaQA [9] (Apache-2.0), 300 TriviaQA + 400 PopQA per split.

Splits are *entity-grouped*: development and test share *zero* entities, preventing memorized-fact leakage. On-policy answers and the confidence signal are generated per model.

Metrics and protocol. *Coverage* is the fraction of questions a model answers (rather than abstaining); *hallucination rate* is the fraction of *answered* items the judge marks incorrect. Because methods that abstain more answer fewer (and easier) questions, the two trade off: the *risk–coverage curve* plots hallucination rate as the abstention threshold sweeps coverage, and AURC (area under it, lower is better) summarizes selective prediction independent of any single operating point. Correctness on the held-out test set is adjudicated by an *independent* open-weight judge, `gemma-2-27b-it` [4] (deliberately neither the Llama nor the Qwen family under test, to avoid same-family self-favoring). We use AURC rather than expected calibration error because our models abstain rather than emit a numeric confidence, so a confidence-bin calibration metric is ill-posed here.

Significance. *Matched coverage* means we set C3's abstention threshold per model so its coverage approximates C2's, then compare hallucination at that operating point (small residual coverage gaps remain; see Limitations). Confidence intervals are percentile intervals from 1,000 item-level bootstrap resamples of the 700-item test set (fixed seed). These intervals capture test-set sampling variance only, not training-seed variance.

5 Results

The signal discriminates hallucination (and largely survives scale). The frozen mean-log-probability signal separates correct from incorrect base-model answers with AUROC 0.778/0.821/0.858 (Llama 1B/3B/8B) and 0.755/0.778/0.725 (Qwen3 1.7B/4B/8B), all above 0.65, and discriminative even at 1B. The trend is monotone within Llama but not within Qwen3 (its 8B value is the lowest of the three).

Label-free is competitive with label-supervised (headline). Table 1 reports judge-adjudicated hallucination at matched coverage. On every model the C2-vs-C3 95% bootstrap confidence intervals overlap, i.e. we do not detect a hallucination difference between the label-free and label-supervised methods. Point estimates fall within a narrow band (C3/C2 reduction ratio 0.95–1.10) that is, on this evidence, within noise; we therefore read the result as parity (failure to detect a difference), not as either method beating the other. Standard fine-tuning (C1) is neutral-to-slightly-worse than base.

The gain is consistent with calibration, not memorization. The up-weighting control (C4) stays at base hallucination with near-full coverage, whereas C3 reduces hallucination: drilling low-signal questions harder does nothing, while retargeting them to abstention helps. (This control is measured at our string-match, n=250 stage; the selective-prediction result below, judge-based at n=700, is stronger evidence for genuine calibration.)

Table 1: Judge-adjudicated hallucination rate (lower is better) on the 700-item held-out test set. **C1**=standard SFT, **C2**=label-supervised abstention (R-Tuning), **C3**= ours (label-free). C3 is evaluated at coverage matched to C2; C3 coverage shown (C2 operates at comparable low coverage, ≈0.15–0.50, by over-abstaining). On every model the C2-vs-C3 95% bootstrap CIs overlap (no detected difference). The C3/C2 column (ratio of hallucination *reduction* vs. base) is a point estimate within noise.

Model	Base	C1	C2 (labels)	C3 (ours; coverage)	C3/C2
Llama-3.2-1B	0.707	0.740	0.400	0.385 (0.19)	1.05
Llama-3.2-3B	0.594	0.553	0.211	0.171 (0.28)	1.10
Llama-3.1-8B	0.445	0.461	0.149	0.164 (0.38)	0.95
Qwen3-1.7B	0.759	0.750	0.316	0.286 (0.08)	1.07
Qwen3-4B	0.660	0.661	0.218	0.185 (0.09)	1.07
Qwen3-8B	0.579	0.580	0.208	0.179 (0.19)	1.08

Selective prediction. Treated as a selective predictor, C3's frozen signal yields lower area under the risk–coverage curve than a non-selective model: 7/25/28% for Llama (1B/3B/8B) and 17/24/26% for Qwen3 (1.7B/4B/8B), with the largest gains for the larger Llama models. Abstention is well-targeted: roughly a 7:1 ratio of useful corrections to abstentions wasted on correct answers.

Robustness. The parity conclusion is unchanged under string-match vs. judge correctness (the judge makes C3 look slightly *better*, not worse), holds across both model families, and is stable from n=250 to n=700; indeed an apparent 69% point estimate at n=250 was sampling noise that vanished under the n=700 bootstrap. Substituting an energy signal for mean log-probability gave no gain.

6 Discussion and Limitations

The practical takeaway is that a model's own confidence is a promising near-free substitute for correctness labels when teaching short-form abstention: across six models we could not distinguish the label-free method from the label-supervised one. Two failure modes bound the result. First, hallucination concentrates on *rare* entities: the worst popularity quartile runs roughly 2–3× the best (e.g. Llama-8B 0.89 vs. 0.32); the method handles these by abstaining rather than by knowing rare facts.

Second, and dominant, are *confidently wrong* answers: 36% of all items are answered-but-wrong with confidence high enough that the signal does not flag them, the bulk of the 39% answered-but-wrong residual. These are high-signal items on which C3 keeps (and thus self-distills) the model's own wrong answer; our calibration claim is therefore specifically about the *abstention* behavior on low-signal items, not that the method is memorization-free overall. This is the intrinsic ceiling of a single-pass log-probability signal and motivates richer signals (verbalized confidence, activation probes) as future work.

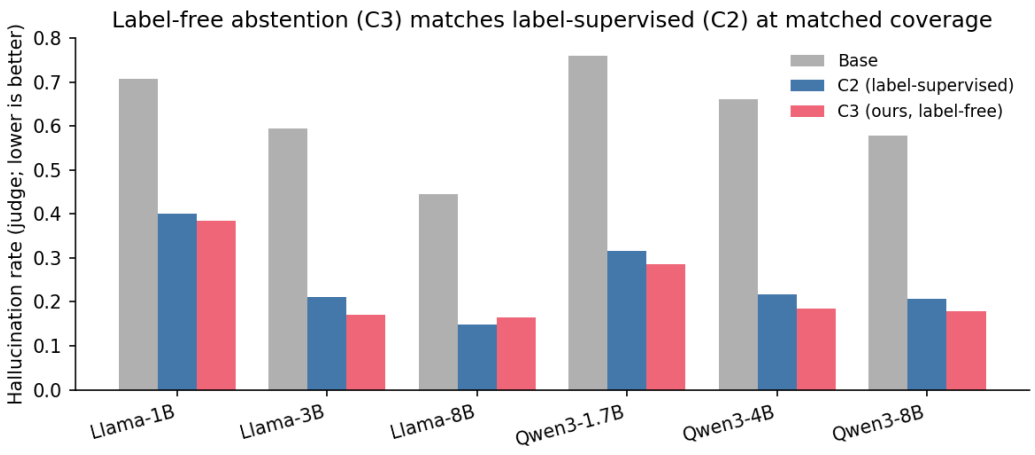

Figure 1: Hallucination rate (lower is better). At matched coverage, label-free abstention (C3, red) and label-supervised abstention (C2, blue) show no detected difference on any model (overlapping 95% bootstrap CIs), both far below the untrained base (grey). The label-free method needs no correctness labels.

6.1 Limitations

A few refinements would further strengthen these results. Our parity finding is a failure to detect a difference rather than a formal equivalence, so a paired equivalence test (TOST) at exactly-pinned coverage is a natural next step, alongside a multiple-comparison adjustment across the six models. The two methods are compared at approximately matched coverage, with the small residual gap if anything favouring the label-free method, so we report parity rather than superiority. Stronger baselines are also worth exploring, such as a judge-trained label-supervised method and a second labelled recipe, and the up-weighting control could be extended to the full evaluation set. Finally, the study covers English short-form question answering at modest scale, so extending the approach to long-form generation and larger models is promising future work; correctness throughout is decided by a single judge model, with a string-match measure reported alongside as a check.

7 Availability

The training and evaluation code, the fine-tuned LoRA adapters, and the data-construction scripts are written and available on request from the corresponding author. No artifacts are hosted publicly at this time.

8 Conclusion and Future Work

On English short-form factual QA at $\leq$8B, a model's own token-probability confidence, frozen and used without any correctness labels, can be fine-tuned into native abstention that is competitive with label-supervised abstention-tuning across six open-weights models, at no labeling cost; we do not detect a difference between the two. The gain is consistent with calibration rather than memorization, and selective-prediction value tends to grow with scale within the Llama family. The chief limitation is confidently-wrong facts the signal cannot see. The most promising next steps are

a formal equivalence test at exactly-pinned coverage, a judge-trained and a second label-supervised baseline, richer self-supervised signals (verbalized confidence, activation probes), and extension to long-form factuality.

References

[1] Arslan Chaudhry, Sridhar Thiagarajan, and Dilan Gorur. Finetuning Language Models to Emit Linguistic Expressions of Uncertainty. arXiv:2409.12180 [cs.CL], 2024. URL https://arxiv.org/abs/2409.12180.

[2] Yu-Neng Chuang, Prathusha K. Sarma, Parikshit Gopalan, John Boccio, Sara Bolouki, Xia Hu, and Helen Zhou. Learning to Route LLMs with Confidence Tokens. arXiv:2410.13284 [cs.CL], 2024. URL https://arxiv.org/abs/2410.13284.

[3] Cheng Gao, Cheng Huang, Kangyang Luo, Ziqing Qiao, Shuzheng Si, Huimin Chen, Chaojun Xiao, and Maosong Sun. KARL: Mitigating Hallucinations in LLMs via Knowledge-Boundary-Aware Reinforcement Learning. arXiv:2604.22779 [cs.CL], 2026. URL https://arxiv.org/abs/2604.22779.

[4] Gemma Team. Gemma 2: Improving Open Language Models at a Practical Size. arXiv:2408.00118 [cs.CL], 2024. URL https://arxiv.org/abs/2408.00118.

[5] Aaron Grattafiori et al. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.AI], 2024. URL https://arxiv.org/abs/2407.21783.

[6] Sophia Hager, David Mueller, Kevin Duh, and Nicholas Andrews. Uncertainty Distillation: Teaching Language Models to Express Semantic Confidence. arXiv:2503.14749 [cs.CL], 2025. URL https://arxiv.org/abs/2503.14749.

[7] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685 [cs.CL], 2021. URL https://arxiv.org/abs/2106.09685.

[8] Neel Jain, Aditya Shrivastava, Chenyang Zhu, Daben Liu, Alfy Samuel, Ashwinee Panda, Anoop Kumar, Micah Goldblum, and Tom Goldstein. Refusal Tokens: A Simple Way to Calibrate Refusals in Large Language Models. arXiv:2412.06748 [cs.CL], 2024. URL https://arxiv.org/abs/2412.06748.

[9] Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. arXiv:1705.03551 [cs.CL], 2017. URL https://arxiv.org/abs/1705.03551.

[10] Guy Kaplan, Zorik Gekhman, Zhen Zhu, Lotem Rozner, Yuval Reif, Swabha Swayamdipta, Derek Hoiem, and Roy Schwartz. Why Fine-Tuning Encourages Hallucinations and How to Fix It. arXiv:2604.15574 [cs.CL], 2026. URL https://arxiv.org/abs/2604.15574.

[11] Ranganath Krishnan, Piyush Khanna, and Omesh Tickoo. Enhancing Trust in Large Language Models via Uncertainty-Calibrated Fine-tuning. arXiv:2412.02904 [cs.LG], 2024. URL https://arxiv.org/abs/2412.02904.

[12] Dharshan Kumaran, Viorica Patraucean, Simon Osindero, Petar Veličković, and Nathaniel Daw. How LLMs Detect and Correct Their Own Errors: The Role of Internal Confidence Signals. arXiv:2604.22271 [cs.CL], 2026. URL https://arxiv.org/abs/2604.22271.

[13] Gabrielle Kaili-May Liu, Gal Yona, Avi Caciularu, Idan Szpektor, Tim G. J. Rudner, and Arman Cohan. MetaFaith: Faithful Natural Language Uncertainty Expression in LLMs. arXiv:2505.24858 [cs.CL], 2025. URL https://arxiv.org/abs/2505.24858.

[14] Matéo Mahaut, Laura Aina, Paula Czarnowska, Momchil Hardalov, Thomas Müller, and Lluís Màrquez. Factual Confidence of LLMs: on Reliability and Robustness of Current Estimators. arXiv:2406.13415 [cs.CL], 2024. URL https://arxiv.org/abs/2406.13415.

[15] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. arXiv:2212.10511 [cs.CL], 2022. URL https://arxiv.org/abs/2212.10511.

[16] Mengjia Niu, Hamed Haddadi, and Guansong Pang. Robust Hallucination Detection in LLMs via Adaptive Token Selection. arXiv:2504.07863 [cs.CL], 2025. URL https://arxiv.org/abs/2504.07863.

[17] Stephan Rabanser, Nathalie Rauschmayr, Achin Kulshrestha, Petra Poklukar, Wittawat Jitkrittum, Sean Augenstein, Congchao Wang, and Federico Tombari. Gatekeeper: Improving Model Cascades Through Confidence Tuning. arXiv:2502.19335 [cs.LG], 2025. URL https://arxiv.org/abs/2502.19335.

[18] Mainak Singha. Detecting AI Hallucinations in Finance: An Information-Theoretic Method Cuts Hallucination Rate by 92%. arXiv:2512.03107 [cs.CL], 2025. URL https://arxiv.org/abs/2512.03107.

[19] Jiayun Wu, Jiashuo Liu, Zhiyuan Zeng, Tianyang Zhan, Tianle Cai, and Wenhao Huang. Mitigating LLM Hallucination via Behaviorally Calibrated Reinforcement Learning. arXiv:2512.19920 [cs.LG], 2025. URL https://arxiv.org/abs/2512.19920.

[20] Xiaohu Xie, Xiaohu Liu, and Benjamin Yao. Know When You're Wrong: Aligning Confidence with Correctness for LLM Error Detection. arXiv:2603.06604 [cs.CL], 2026. URL https://arxiv.org/abs/2603.06604.

[21] Rui Xu, Yi Chen, Sihong Xie, and Hui Xiong. Geometry-Calibrated Conformal Abstention for Language Models. arXiv:2604.27914 [cs.CL], 2026. URL https://arxiv.org/abs/2604.27914.

[22] An Yang et al. Qwen3 Technical Report. arXiv:2505.09388 [cs.CL], 2025. URL https://arxiv.org/abs/2505.09388.

[23] Hanning Zhang, Shizhe Diao, Yong Lin, Yi R. Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. R-Tuning: Instructing Large Language Models to Say 'I Don't Know'. arXiv:2311.09677 [cs.CL], 2023. URL https://arxiv.org/abs/2311.09677.