
Is “Confidence” the Right Word for Answer Repeat Probability?

Koshiro Aoki
Waseda University *

Abstract

Natural-language confidence reports depend on how they are elicited. We therefore test whether a single learned word can reliably report a frozen model’s answer repeat probability. For each multiple-choice question, the target is the probability that the model repeats its most likely valid answer. We learn one input embedding in Llama-3.1-8B and map the model’s 0–10 report to this target. Under a carrier shift, eight initializations yield Spearman correlations from 0.359 to 0.586. The reporter selected on the development set does not significantly outperform fixed *confidence* or *consistency*. A parameter-matched generic soft vector has lower mean squared error on both carriers, and readable replacements do not preserve the learned behavior. A learned vector can therefore adapt a particular reporting context, but these results do not establish a stable word-level interface or privileged introspection.

1 Introduction

Language models can report a number when asked how sure they are, but that number depends on the wording, rating scale, and prompt format [Xiong et al., 2024, Tian et al., 2023]. This dependence makes ordinary words such as *confidence* unreliable interfaces for precise targets. It also raises a constructive question: can a learned word give a frozen model a more reliable interface for reporting a specified property of its own answer?

Prior work motivates this question but also calls for a cautious interpretation. Continuous prompts can adapt a frozen model with few trainable parameters [Lester et al., 2021]. Hewitt et al. propose learned neologisms as a way to connect human and model concepts [Hewitt et al., 2025]. Yet nearest discrete tokens need not preserve the function of a continuous prompt [Khashabi et al., 2022]. Work on verbal uncertainty also separates linguistic expression from uncertainty estimated by resampling [Yona et al., 2024]. A successful soft prompt therefore does not, by itself, show that it has become a usable word.

We therefore study a deliberately narrow target. For each multiple-choice question, we identify the model’s most likely choice and compute the exact probability that another draw from the *restricted choice distribution* repeats it. This target is neither free-form semantic consistency nor the probability of a correct answer. Nor does reporting it imply that the model has privileged access to its internal state. The task is instead a controlled readout problem.

We learn a single embedding at the word position of a post-answer rating prompt while keeping the model frozen. We cache all answers and targets before training the reporter. We compare eight initial vectors with hard prompts and a parameter-matched generic soft vector. We then change the carrier sentence and replace each learned vector with either its nearest token or a model-generated paraphrase. These tests distinguish context-specific soft adaptation from a robust lexical interface.

*Research conducted at the Digital Minds Research Sprint, August 2026

Table 1: Test performance on 256 held-out questions. *Dev-selected* denotes the initial vector with the lowest dev MSE. The held-out test selects the centroid row, so that row is descriptive only.

Condition	Seen carrier		Shifted carrier	
	ρ	MSE	ρ	MSE
Fixed confidence	.552	.127	.533	.116
Fixed consistency	.595	.124	.523	.115
Explicit definition	.667	.103	.284	.085
Generic soft vector	.696	.042	.471	.064
Learned, dev-selected	.608	.051	.546	.115
Learned, test-best centroid [†]	.690	.056	.586	.094

2 Method

Target. For question i , let $\mathcal{A}_i$ be its valid answer letters and let $s_i(a)$ be the frozen model logit for letter a . We normalize only over $\mathcal{A}_i$:

$$p_i(a) = \frac{\exp s_i(a)}{\sum_{b \in \mathcal{A}_i} \exp s_i(b)}, \quad A_i = \arg \max_{a \in \mathcal{A}_i} p_i(a), \quad C_i = p_i(A_i). \quad (1)$$

Thus, C_i is the exact probability of repeating anchor A_i under the restricted choice distribution. We compute this probability with one forward pass rather than by sampling answers.

Reporter. After fixing the anchor, we ask the model to rate the *word* of its answer from 0 to 10. We renormalize the likelihoods of the eleven possible rating strings. The report is the expected normalized rating, $r_i = \sum_{b=0}^{10} p_i(b)b/10$. A learned reporter adds a vector δ to the input embedding of one word. The word appears in the seen carrier: *Rate the consistency of your answer from 0 to 10*. We minimize $\frac{1}{n} \sum_i (r_i - C_i)^2$ and keep all model parameters fixed.

We split the official MMLU-Pro test set into 512 train, 128 dev, and 256 test items [Wang et al., 2024]. The model is Llama-3.1-8B-Instruct [Grattafiori et al., 2024]. Training uses AdamW, learning rate 10^{-3} , batch size one, $L_2 = 10^{-4}$, and early stopping based on development-set MSE. We constrain each update to half the median norm of the vocabulary embeddings. We initialize the learned vector with *consistency*, *confidence*, *certainty*, *reliability*, their centroid, and three norm-matched random vectors.

Controls and evaluation. Hard controls use *confidence*, *consistency*, or an explicit definition of repeat probability. For the generic soft-vector control, we train an equally sized vector at the unrelated token *Question*. We evaluate both the seen carrier and an unseen carrier: *On a 0 to 10 scale, report the word*. The primary metric is Spearman correlation between r_i and C_i ; MSE and MAE are secondary. We use 1,000 paired item bootstraps for the planned comparisons.

To assess interpretability, we replace each learned vector with the vocabulary token that has the highest cosine similarity to it. We also replace each vector with a model-generated paraphrase.

3 Results

The planned test does not confirm a carrier-specific gain. On the seen carrier, all learned vectors rank the target well ($\rho = .608$ – $.690$, Figure 2a). Their descriptive mean is .648. This pattern does not isolate a lexical effect: the generic soft vector reaches .696 and has lower MSE than every learned run. Under the carrier shift, the correlations of the learned vectors range from .359 to .586, with a descriptive mean of .469.

The planned selection rule chooses the *consistency* initialization because it has the lowest development-set MSE. On the shifted carrier, this vector gives $\rho = .546$. The difference from fixed *confidence* is .012 (95% paired-bootstrap CI $[-.052, .079]$). The difference from fixed *consistency* is .023 ($[-.033, .083]$). Both intervals include zero.

The selected vector has an MSE of .115, whereas the generic soft vector achieves .064. The planned comparison therefore does not support a reliable improvement over fixed language, while the calibration results favor generic soft adaptation.

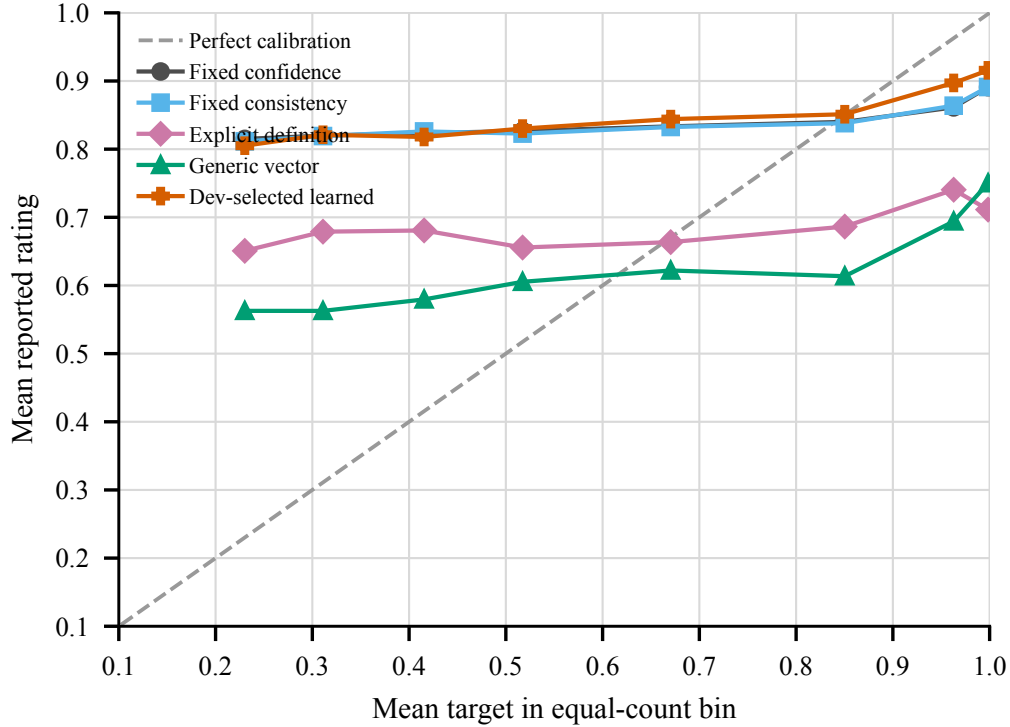


Figure 1: Held-out calibration curves for the dev-selected lexical reporter and fixed controls. Each point averages 32 test items after sorting by target. The generic soft vector is closer to the diagonal at low targets. This pattern accounts for its lower global MSE. These descriptive bins are not independent runs.

Figure 1 illustrates the calibration difference. Fixed words and the dev-selected lexical reporter produce reports near .8 even when the mean target is near .23. The generic soft vector is less overconfident in the lower target bins. Although its rank correlation is lower than that of the dev-selected lexical vector, its lower MSE shows that rank performance alone is insufficient for this readout task.

The favorable test result is exploratory. The centroid initialization performs best on the shifted test ($\rho = .586$), but this comparison selects the initialization on that same test set. The centroid also has a higher MSE than the generic soft vector (.094 versus .064). Its post-hoc correlation difference from the generic soft vector is .115 (95% CI [.036, .200]), and its MSE difference is +.030 ([+.017, +.042]). Thus, one test-best rank correlation cannot establish a general lexical advantage.

Readable recovery fails. On the shifted carrier, no nearest-token replacement exceeds $\rho = .490$, and no generated paraphrase exceeds .408 (Figure 2c). Both maxima are below the .533 achieved by fixed *confidence*. Some nearest tokens for the random vectors are non-English strings, so we treat them as candidates for preserving behavior rather than as interpretable words. None of the specified discrete replacements recover the learned behavior.

A difficulty prior is competitive. As a post-hoc diagnostic, we predict each test target using the training-set mean for its MMLU-Pro family. This simple, non-reporter baseline obtains $\rho = .532$ and MSE .055. The target varies strongly by problem family. Correlation with it can therefore reflect item difficulty as well as reporter quality, which further supports a narrow interpretation of the learned reporter.

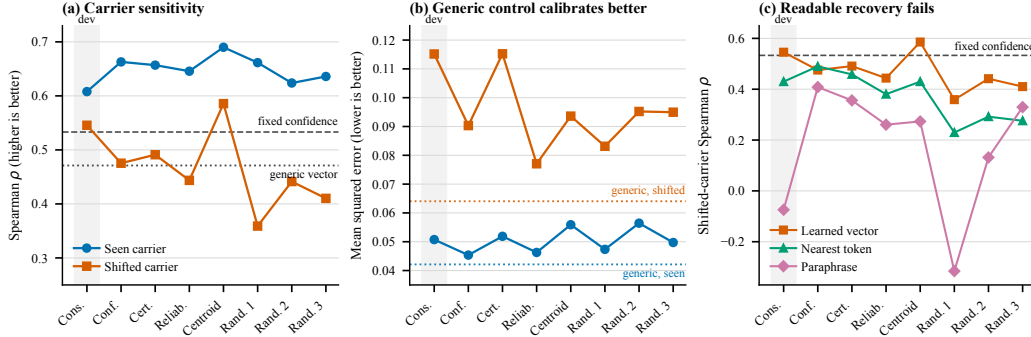


Figure 2: The gray band marks the dev-selected initialization among all eight lexical initializations on 256 test items. The centroid’s test-best value is descriptive, not a selection rule. (a) Rank performance changes markedly with the carrier. (b) The parameter-matched generic soft vector has lower MSE than every lexical vector on both carriers. (c) A nearest token or generated paraphrase does not preserve the shifted-carrier behavior. Lines join different conditions, not repeated training runs.

4 Discussion and limitations

The experiment supports soft adaptation within the training carrier but not a robust word-level interface. Under a carrier shift, the dev-selected vector does not significantly outperform fixed words. Generic soft adaptation yields better calibration, and readable replacements fail to preserve the learned behavior. These results accord with the warning that discretized interpretations of continuous prompts can be wayward [Khashabi et al., 2022].

Several limitations constrain this interpretation. We test one frozen model on one data split. The eight initializations are experimental conditions rather than independent reruns. All selected checkpoints reach the update-radius boundary, and we do not test sensitivity to that radius.

MLLU-Pro restricts the available answer choices, whereas meaning-level uncertainty in open-ended generation poses a different problem [Kuhn et al., 2023]. Our target also correlates with family-level difficulty. The Qwen and Gemma runs did not finish, so we make no claim about transfer across models. Finally, the percentile intervals resample only items and do not capture variation from training.

The practical conclusion is limited: a single learned vector can improve a specific reporting prompt. Before treating such a vector as a learned word, it should be evaluated under prompt shifts and compared with generic soft vectors, readable replacements, and non-reporter priors.

References

- Aaron Grattafiori et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL <https://arxiv.org/abs/2407.21783>.
- John Hewitt, Robert Geirhos, and Been Kim. Position: We can’t understand AI using our existing vocabulary. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 81556–81572. PMLR, 2025. URL <https://proceedings.mlr.press/v267/hewitt25a.html>.
- Daniel Khashabi, Xinxu Lyu, Sewon Min, Lianhui Qin, Kyle Richardson, Sean Welleck, Hannaneh Hajishirzi, Tushar Khot, Ashish Sabharwal, Sameer Singh, and Yejin Choi. Prompt waywardness: The curious case of discretized interpretation of continuous prompts. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3631–3643. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.266. URL <https://aclanthology.org/2022.naacl-main.266/>.

- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=VD-AYtP0dve>.
- Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.243. URL <https://aclanthology.org/2021.emnlp-main.243/>.
- Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.330. URL <https://aclanthology.org/2023.emnlp-main.330/>.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhui Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pages 95266–95290, 2024. doi: 10.52202/079017-3018. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ad236edc564f3e3156e1b2feafb99a24-Paper-Datasets_and_Benchmarks_Track.pdf.
- Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *International Conference on Learning Representations*, pages 23650–23678, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/6733cf15e10e2cd1d59af033c3bb8507-Paper-Conference.pdf.
- Gal Yona, Roei Aharoni, and Mor Geva. Can large language models faithfully express their intrinsic uncertainty in words? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7752–7764. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.443. URL <https://aclanthology.org/2024.emnlp-main.443/>.

A LLM Usage Statement

We used Codex to assist with code development and manuscript preparation.

B Full experimental details

Prompts. For each question, the answer-caching prompt instructs the model to return only a valid choice letter. The reporter then receives the question, the cached anchor, a carrier sentence, and an instruction to return only a rating. The seen carrier is *Rate the word of your answer from 0 to 10*. The shifted carrier is *On a 0 to 10 scale, report the word*. In the explicit hard control, a definition replaces *word*. That definition is *probability that another answer sampled under the same conditions would have the same meaning as your answer*.

C Readable replacements

Table 2: All reporter conditions. Values are test Spearman ρ /MSE. We evaluate the four fixed rows once per carrier. They are not repeated samples.

Condition	Seen		Shifted	
	ρ	MSE	ρ	MSE
Fixed confidence	.552	.127	.533	.116
Fixed consistency	.595	.124	.523	.115
Explicit definition	.667	.103	.284	.085
Generic soft vector	.696	.042	.471	.064
Lexical: consistency	.608	.051	.546	.115
Lexical: confidence	.663	.045	.475	.090
Lexical: certainty	.657	.052	.491	.115
Lexical: reliability	.646	.046	.444	.077
Lexical: centroid	.690	.056	.586	.094
Lexical: random 1	.662	.047	.359	.083
Lexical: random 2	.624	.056	.441	.095
Lexical: random 3	.636	.050	.410	.095

Table 3: Held-out replacements. Random-vector candidates can be readable Unicode text, but they are not necessarily English words.

Init.	Nearest token	Paraphrase	Near. ρ	Para. ρ
Consistency	consistency	Thickness or texture.	.430	-.074
Confidence	confidence	Confident self-assurance.	.490	.408
Certainty	certainty	Complete confidence.	.458	.356
Reliability	reliability	Consistency of behavior.	.380	.260
Centroid	consistency	Lack of bias.	.430	.273
Random 1	administered	Someone who spends a lot of time in the bathroom.	.229	-.316
Random 2	Arabic token	Severe or intense.	.292	.132
Random 3	Chinese token	No space.	.276	.330