TracVC: Tracing Verbalized Confidence of LLMs Back to Training Data

Anonymous ACL submission

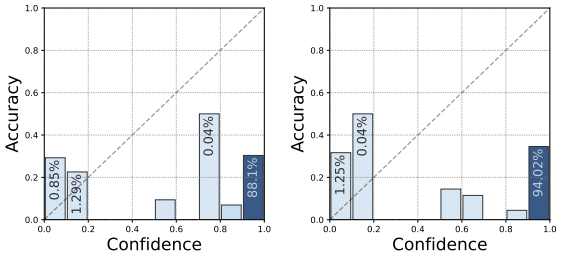
Abstract

Large language models (LLMs) can increase users' perceived trust by verbalizing confidence in their outputs. However, prior work shows that LLMs often express overconfidence, which is misaligned with factual accuracy. To better understand the sources of this behavior, we propose **TracVC**, a method for **Trac**ing **V**erbalized **C**onfidence back to specific training data. We conduct experiments on OLMo models in a question answering setting, defining a model as *truthful* when content-related training data—relevant to the question and answer, has greater influence than confidence-related data. Our analysis reveals that OLMo2-13B is often influenced by confidence-related data that is semantically unrelated to the query, suggesting that it may mimic linguistic markers of certainty. This finding highlights a fundamental limitation in current training regimes: LLMs may learn how to sound confident without understanding when confidence is warranted. Our analysis provides a foundation for improving LLMs' trustworthiness in expressing more truthful confidence. [1]

1 Introduction

Verbalized confidence in large language models (LLMs) is increasingly used to estimate the certainty of their generated outputs to improve transparency and user trust (Kadavath et al., 2023; Tian et al., 2023; Chen and Mueller, 2024). However, recent studies have shown that LLMs frequently exhibit overconfidence, which is not aligned with factual accuracy (shown in Fig. 1), resulting in their poor reliability in expressed confidence (Zhou et al., 2024; Xiong et al., 2024; Ni et al., 2025; Xia et al., 2025). These findings lead to a foundational question: what drives confidence in LLMs, and do LLMs understand the intended meaning of expressing confidence?

[1]Code is available at: https://anonymous.4open.science/r/training_data_confidence-CB2E.



(a) OLMo2-7B Instruct (b) OLMo2-13B Instruct

Figure 1: LLMs tend to be overconfident. E.g., more than 94% test samples get confidence scores between 0.9 to 1.0 from the OLMo2-13B Instruct model, even though the accuracy of those samples is less than 0.4.

In this paper, we investigate the origin of verbalized confidence in LLMs by examining the influence of training data. Specifically, we ask: Do LLMs ground their confidence on semantically relevant, content-related training samples, or are they instead influenced by superficial confidence-related cues? To address this question, we propose **TracVC**, a method for tracing the origins of verbalized confidence back to specific training data. TracVC builds on influence estimation techniques (Pruthi et al., 2020), but is adapted to estimate how individual training samples contribute to confidence generation in LLMs.

We apply TracVC to analyze 11 open-source LLMs with publicly accessible training data, OLMo (Team OLMo et al., 2024) and Llama3-8B (Meta AI, 2024) models, on five question answering benchmarks. To measure whether content-related training data (i.e., question and answer) is more influential than confidence-related data in shaping verbalized confidence, we define a truthfulness metric as the ratio of cases where content wins over confidence. Our results demonstrate that (1) OLMo2-13B models are influenced more by confidence-related training samples that are semantically unrelated to the question, often latching onto

keywords like "confidence" regardless of context. This suggests that it does not ground its confidence in truthful information, but instead learns to mimic linguistic markers of certainty; (2) Larger LLMs do not exhibit higher truthfulness than smaller ones; (3) LLMs can show low truthfulness even if they may have seen content-related samples during training; (4) Post-training techniques (e.g., direct preference optimization) have limited and inconsistent impact on truthfulness, suggesting that pre-training may be more critical in shaping how confidence is grounded.

These findings highlight a fundamental limitation in current training regimes: LLMs may learn to sound confident without understanding when confidence is warranted. Our work introduces a data-driven perspective of model confidence, offering insights that can guide future training approaches toward improving the truthfulness and reliability of model confidence.

2 Related Work

Overconfidence in LLMs (Zhou et al., 2024; Xiong et al., 2024; Ni et al., 2025; Xia et al., 2025) poses a risk to user trust. **Interpreting verbalized confidence** is thus crucial to address this risk. Ni et al. (2025); Kumar et al. (2024) only investigates whether verbalized confidence is aligned with internal probabilities of LLMs, lacking insights into the origins of confidence. **Influence estimation methods**, such as influence functions (Koh and Liang, 2017), offer an alternative to leave-one-out re-training when studying the impact that specific examples in the training data have on a model. This has enabled their recent application to related problems such as fact-tracing (Chang et al., 2024), outlier detection (Pruthi et al., 2020), or data valuation (Choe et al., 2024; Bejan et al., 2023). In this work, we adapt the gradient-similarity based method *TracIn* (Pruthi et al., 2020) to enable efficient analysis of what training data influences LLM's confidence estimates, an application not previously approached through this lens.

3 Methodology

The workflow of our proposed TracVC is shown in Fig. 2. TracVC identifies what types of training data are more influential for LLM verbalized confidence. We mainly study two types of training data: content-related and confidence-related data. The following sections explain our prompt design, the data search mechanism to retrieve related data for content and confidence from training data, and finally, the method we use to estimate and compare the influence score of the retrieved data.

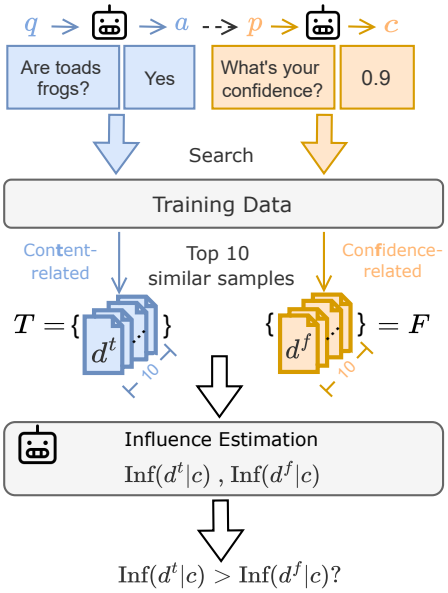


Figure 2: The workflow of TracVC. We first search the top 10 similar samples for content (question (q) and answer (a)) and confidence (phrase (p) and confidence (c)). Then, we compute and compare the influence score regarding the confidence generation for content-related and confidence-related training data.

3.1 Prompt Design

We use a two-step prompt for this experiment. The first-stage prompt contains the question q for an LLM $\mathcal{M}$ to generate ($\sim$) the answer a ($a \sim \mathcal{M}(q)$). The second-stage prompt consists of a phrase p to require $\mathcal{M}$ to provide a probability for its generated answer. Finally, $\mathcal{M}$ generates the verbalized confidence c regarding a and q ($c \sim \mathcal{M}(q, a, p)$).

3.2 Training Data Search

To answer our research question on whether the question-answer (*content*)-related training data is more influential than the *confidence*-related training data, we first search for the top 10 most similar data samples of each type per test sample from the training data. To achieve this, we employ the search mechanism developed by Elazar et al. (2024), which uses Elastic Search (Elastic, 2025) to create a keyword-based search engine built on the training data. This search engine allows us to retrieve similar samples for a search query: By querying with question and answer, we obtain 10 content-related samples $T = \{d_1^t, ..., d_{10}^t\}$. Likewise, by

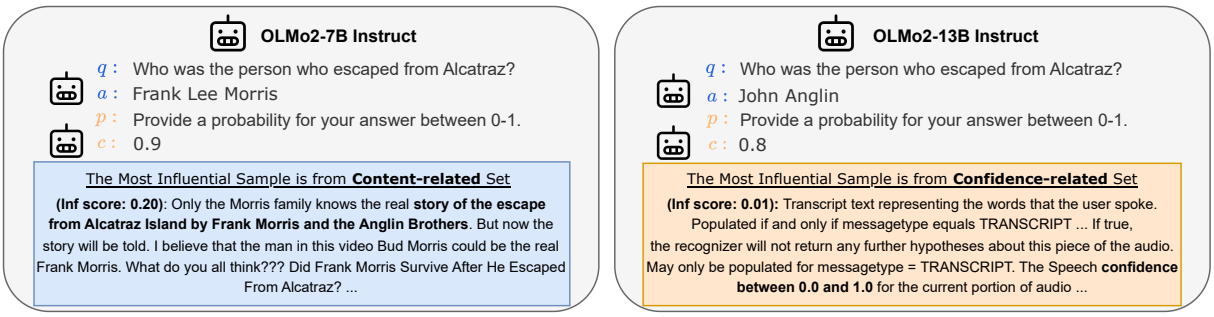


Figure 3: Examples of the most influential data samples for different LLMs when generating confidence. We retrieve these samples from the pre-training data, we provide additional examples in Appendix A.7.

searching for phrase and confidence, we obtain 10 confidence-related samples $F = \{d_1^f, ..., d_{10}^f\}$.

3.3 Training Data Influence Estimation

Inspired by TracIn (Pruthi et al., 2020), our method is based on point-wise comparisons of loss gradients. Specifically, we estimate the influence of a training sample d on the model's generated confidence c by computing the cosine similarity between the gradients of the model's loss function. One gradient is computed when the loss is evaluated on the training sample d, and the other when evaluated on (q, a, p). Both gradients are taken with respect to the model's input embeddings w:

$$\text{Inf}(d|c_i) = \frac{\nabla \ell(w_i,d) \cdot \nabla \ell(w_i,(q_i,a_i,p))}{\|\nabla \ell(w_i,d)\| \cdot \|\nabla \ell(w_i,(q_i,a_i,p))\|} \quad (1)$$

We deliberately exclude the verbalized confidence c from this score to ensure its independence of confidence variations. The implementation details are discussed in Appendix A.2.

3.4 Truthfulness Evaluation

We define truthfulness as the property whereby content-related training data exerts a greater influence than confidence-related data on confidence generation. To quantify truthfulness, we introduce **ccr** (Content-over-Confidence Ratio), defined as the winning counts ratio between the content-related set T and confidence-related set F. Intuitively, a higher **ccr** indicates that the model's confidence is more strongly driven by semantically relevant, truthful content rather than by potentially misleading confidence cues. Assume our test data contains n questions, **ccr** is computed by:

$$ccr = \frac{\sum_{i=0}^{n} \sum_{d^t \in T_i, d^f \in F_i} \mathbf{1}(\text{Inf}(d^t|c_i) > \text{Inf}(d^f|c_i))}{\sum_{i=0}^{n} \sum_{d^t \in T_i, d^f \in F_i} \mathbf{1}(\text{Inf}(d^t|c_i) < \text{Inf}(d^f|c_i))} \quad (2)$$

3.5 Experimental Setup

Our experiments require knowledge about the training data of LLMs. Therefore, we mainly study OLMo models (Team OLMo et al., 2024; Groeneveld et al., 2024) that provide publicly available pre- and post-training data. Additionally, we include Llama-3.1 (Meta AI, 2024) (Llama-3.1-Tulu-3-8B), which was post-trained with publicly available data. More model details are provided in Appendix A.3. We test on Natural Question (NQ) (Kwiatkowski et al., 2019), SicQ (Johannes Welbl, 2017), TrivialQA (Joshi et al., 2017), PopQA (Mallen et al., 2023) and TruthfulQA. Each of the first four datasets contains 1,000 randomly selected samples, and TruthfulQA (Lin et al., 2022) has 817.

4 Result and Analysis

Table 1 and Fig. 4 show the truthfulness results of different OLMo models on five benchmark datasets. We observe (1) a bigger model is not more truthful; (2) truthfulness is not correlated to task accuracy; (3) post-training schemes have limited impacts; pre-training schemes are more likely to determine the truthfulness of the model. Below is a more detailed analysis. We also provide ablation and case studies in Appendix A.6 and A.7 for more validations.

Which data samples are more influential for confidence generation? Confidence generation in OLMo2-13B Instruct (INS) is more influenced by confidence-related training data, while content-related data plays a greater role in OLMo-7B and OLMo2-7B INS models. As shown in Fig. 3, OLMo2-7B INS assigns high influence scores to samples that are semantically aligned with the question content. In contrast, OLMo2-13B INS is impacted by unrelated samples that contain superficial cues, such as the keyword "confidence".

Do LLM truthfulness relate to task accuracy? We hypothesize that high task accuracy in LLMs

Search Data	Instruct (INS) Model	NQ ccr	NQ acc	SciQ ccr	SciQ acc	TriviaQA ccr	TriviaQA acc	TruthfulQA ccr	TruthfulQA acc	PopQA ccr	PopQA acc	All ccr	All acc
Pre		0.76		0.71		0.84		0.73		0.74		0.75	
Post	OLMo2-13B	0.78	29.50	0.85	57.50	0.85	39.30	0.80	16.40	0.74	21.00	0.80	33.36
Pre+Post		0.77		0.78		0.84		0.77		0.74		0.78	
Pre		1.08		1.01		1.17		1.43		1.02		1.12	
Post	OLMo2-7B	1.85	23.60	1.63	50.50	1.73	32.20	1.67	15.42	1.74	16.10	1.72	28.03
Pre+Post		1.43		1.32		1.45		1.56		1.37		1.42	
Pre		1.22		~~0.95~~		1.08		1.30		1.06		1.11	
Post	OLMo-7B	1.28	17.30	1.17	39.80	1.29	27.00	1.25	20.93	1.15	11.90	1.22	23.48
Pre+Post		1.25		1.05		1.18		1.27		1.10		1.16	
	Average	1.14	23.47	1.07	49.27	1.23	32.83	1.20	17.58	1.07	16.33	1.14	28.62

Table 1: Result of LLMs' truthfulness measured with **ccr** and accuracy (**acc**) on different datasets. The training data used for estimating the influence score is searched from pre-training (Pre) data, post-training (Post) data or a combination of both (Pre+Post). We cross out the results that are not significant (p>0.05) in the mean influence score difference between the two comparison sets. Findings are robust across different training data settings.

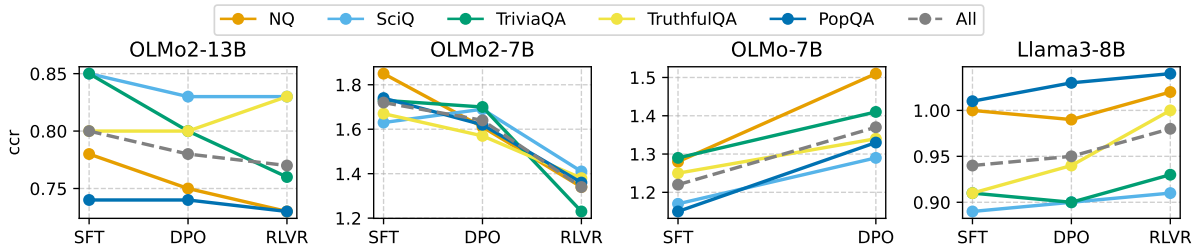


Figure 4: Impact of different post-training schemes on the truthfulness assessment. We plot the training stages of the same model over time. The examined training samples are from the post-training data of the corresponding model. We include a new model, Llama3-8B, which has been post-trained with similar data to OLMo models.

may reflect exposure to similar examples during training, leading the model to rely on familiar, content-relevant information when generating confidence. If this is the case, we expect a positive correlation between task accuracy and truthfulness. However, as shown in Table 1, this correlation does not hold: the SciQ dataset achieves the highest average accuracy but a relatively low truthfulness score, while datasets like TruthfulQA, despite lower accuracies, exhibit higher truthfulness, suggesting LLMs may not rely on relevant training data that have seen during training to estimate confidence.

Do post-training schemes impact LLM truthfulness? The results in Fig. 4 demonstrate that post-training schemes can have opposite effects on different models. For example, DPO or RLVR improves the truthfulness for OLMo-7B and Llama3-8B but not for OLMo2-13B and OLMo2-7B models. We also observe that post-training schemes generally do not reverse the truthfulness results, i.e., scores are either all under one or above one, which indicates that pre-training schemes are more likely to determine the truthfulness.

5 Conclusion

LLMs often exhibit overconfidence, raising the question whether their confidence is grounded in truthful, content-relevant training data. This paper introduces TracVC, a method for tracing the origins of LLM confidence back to specific types of training samples. Our analysis reveals that OLMo2-13B models are more influenced by confidence-related, rather than content-related data when estimating confidence, suggesting a risk of trustworthiness in their expressed certainty. We further find that truthfulness does not correlate with model size, task accuracy or the extent of post-training. These findings motivate future work on how pre-training schemes shape confidence grounding, with the goal of building models whose confidence more reliably reflects truthful reasoning. Finally, our proposed TracVC method can be extended to interpret different types of model outputs from a data-driven perspective.

Limitations

Our findings are constrained to a limited set of LLMs, primarily the OLMo and Llama families, due to restricted access to the pre-training and post-training data of other proprietary models. Since TracVC relies on analyzing training data influence, the lack of transparency and availability of training data for widely used commercial LLMs (e.g., GPT-4, Claude, Gemini) limits the generalizability of our conclusions. Future work could extend this analysis as more open-source models and datasets become available.

Our proposed method, TracVC, is broadly applicable and can be extended to trace various types of model outputs beyond confidence, such as specific answer choices. However, in this work, we focus solely on validating the method in the context of verbalized confidence. This focused scope allows for a clearer analysis of our method's effectiveness. Future work can extend our method to explore how the training data influences other forms of LLM output.

References

Irina Bejan, Artem Sokolov, and Katja Filippova. 2023. Make every example count: On the stability and utility of self-influence for learning from noisy NLP datasets. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10107–10121, Singapore. Association for Computational Linguistics.

Tyler A. Chang, Dheeraj Rajagopal, Tolga Bolukbasi, Lucas Dixon, and Ian Tenney. 2024. Scalable Influence and Fact Tracing for Large Language Model Pretraining. *arXiv preprint*. ArXiv:2410.17413 [cs].

Jiuhai Chen and Jonas Mueller. 2024. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5186–5200, Bangkok, Thailand. Association for Computational Linguistics.

Sang Keun Choe, Hwijeen Ahn, Juhan Bae, Kewen Zhao, Minsoo Kang, Youngseog Chung, Adithya Pratapa, Willie Neiswanger, Emma Strubell, Teruko Mitamura, Jeff Schneider, Eduard Hovy, Roger Grosse, and Eric Xing. 2024. What is your data worth to gpt? llm-scale data valuation with influence functions. *Preprint*, arXiv:2405.13954.

Elastic. 2025. Elasticsearch: Distributed, restful search and analytics engine. `https://www.elastic.co/elasticsearch/`. Version 8.17.2, released February 11, 2025.

Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Evan Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hannaneh Hajishirzi, Noah A. Smith, and Jesse Dodge. 2024. What's in my big data? In *ICLR*.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. 2024. Olmo: Accelerating the science of language models. *Preprint*.

Zayd Hammoudeh and Daniel Lowd. 2022. Identifying a Training-Set Attack's Target Using Renormalized Influence Estimation. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, CCS '22, pages 1367–1381, New York, NY, USA. Association for Computing Machinery.

Zayd Hammoudeh and Daniel Lowd. 2024. Training data influence analysis and estimation: a survey. *Machine Learning*, 113(5):2351–2403.

Matt Gardner Johannes Welbl, Nelson F. Liu. 2017. Crowdsourcing multiple choice science questions. *arXiv:1707.06209v1*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Chethan Kadavath, Amanda Askell, Jackson Kernion, Tom Henighan, Ben Mann, Gretchen Krueger, Sarah Kreps, Aaron McKane, Gaurav Mistry, Joe Kim, et al. 2023. Prompting GPT-3 to be reliable. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*.

Pang Wei Koh and Percy Liang. 2017. Understanding Black-box Predictions via Influence Functions. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1885–1894. PMLR. ISSN: 2640-3498.

Abhishek Kumar, Robert Morabito, Sanzhar Umbet, Jad Kabbara, and Ali Emami. 2024. Confidence under the hood: An investigation into the confidence-probability alignment in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1:*

Long Papers), pages 315–334, Bangkok, Thailand. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Meta AI. 2024. Introducing meta llama 3: The most capable openly available llm to date. Accessed: 2025-05-01.

Shiyu Ni, Keping Bi, Lulu Yu, and Jiafeng Guo. 2025. Are large language models more honest in their probabilistic or verbalized confidence? In *Information Retrieval*, pages 124–135, Singapore. Springer Nature Singapore.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. 2023. TRAK: Attributing Model Behavior at Scale. In *Proceedings of the 40th International Conference on Machine Learning*, pages 27074–27113. PMLR. ISSN: 2640-3498.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. 2020. Estimating Training Data Influence by Tracing Gradient Descent. In *Advances in Neural Information Processing Systems*, volume 33, pages 19920–19930. Curran Associates, Inc.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2024. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. 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, Singapore. Association for Computational Linguistics.

vLLM Contributors. 2023. vllm: A high-throughput and memory-efficient llm serving library. `https://github.com/vllm-project/vllm`. Accessed: 2025-01.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. LESS: Selecting Influential Data for Targeted Instruction Tuning. *arXiv preprint*. ArXiv:2402.04333 [cs].

Yuxi Xia, Pedro Henrique Luz de Araujo, Klim Zaporojets, and Benjamin Roth. 2025. Influences on llm calibration: A study of response agreement, loss functions, and prompt styles. *Preprint*, arXiv:2501.03991.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.

Chih-Kuan Yeh, Ankur Taly, Mukund Sundararajan, Frederick Liu, and Pradeep Ravikumar. 2022. First is Better Than Last for Language Data Influence. *arXiv preprint*. ArXiv:2202.11844 [cs].

Kaitlyn Zhou, Jena Hwang, Xiang Ren, and Maarten Sap. 2024. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3623–3643, Bangkok, Thailand. Association for Computational Linguistics.

A Appendix

A.1 Prompt Details

First-stage prompt

This prompt aims to get the LLM answer of a question.

User: Answer the question, give ONLY the answer, no other words or explanation: **Where does most of our food come from?** $\rightarrow q$

Expected output from LLM: agriculture

Second-stage prompt

We use the LLM answer from the first stage prompt to generate this prompt in order to get the model's confidence on its answer. As we want to investigate the influences of confidence, therefore,

we ensure this stage of prompt targets to guide the LLM to generate only confidence.

User: Answer the question, give ONLY the answer, no other words or explanation: Where does most of our food come from?

Assistant: **agriculture** $\rightarrow a$ (target answer is angiosperms, but we do not care about the correctness of the answer in our setting.)

User: **Provide the probability that your answer is correct. Give ONLY the probability between 0.0 and 1.0, no other words or explanation.** $\rightarrow p$

Expected output from LLM: 0.9

A.2 Training Data Influence Estimation

The method for attributing training data influence in this work is inspired by TracIn (Pruthi et al., 2020), which, in its original formulation, aims to estimate the influence that training on an instance z had, when predicting a test instance z': It does so by measuring the similarity between the gradients of the model's loss function, when evaluated on z and z' respectively, w.r.t some set of parameters w_t, at a series of checkpoints T:

$$\phi_{\text{TracInCP}}(z, z') = \sum_{\forall t \in T} \eta_t \nabla \ell(w_t, z) \cdot \nabla \ell(w_t, z') \quad (3)$$

However, in line with previous work, we utilize cosine similarity for our content-over-confidence ratio, to reduce the impact of gradient magnitude on our comparison (Hammoudeh and Lowd, 2022, 2024; Park et al., 2023; Xia et al., 2024).

We compute the gradient with respect to the model's input embeddings. Note that this score captures information about model behavior beyond the word embedding layer, as the gradient chains through higher layers as well (Yeh et al., 2022).

A.3 Detailed Experimental Setup

Our studied OLMo models span different model sizes and versions: OLMo2-13B (OLMo-2-1124-7B-Instruct), OLMo2-7B (OLMo-2-1124-13B-Instruct), OLMo-7B (OLMo-7B-Instruct-hf). Additionally, we study how different post-training schemes impact the **ccr** score. Thus, we include Llama-3.1 (Meta AI, 2024) (Llama-3.1-Tulu-3-8B) that has the post-training data accessible. For each model, we compute the **ccr** scores for checkpoints after post-training with supervised-fine-tuning (SFT), direct preference optimization (DPO), and reinforcement learning with verified reward (RLVR) (the last training step for the instruction model). We employ the vLLM (vLLM Contributors, 2023) library for LLM inference and serving. All LLMs are set with temperature equal to 0 to ensure consistent outputs. All our experiments are conducted on NVIDIA HGX H100, requiring a total of approximately 642 GPU hours for testing various scenarios.

A.4 Accuracy of More Models

In addition to the accuracy performance of models shown in Table 1. We report the accuracy of models from different post-training stages in Fig. 5.

A.5 Details of the Search Data

We show the details of all the pre-training and post-training data in Table 2. We use dolma v1.7 as pre-training data for OLMo models, which is provided as a publicly available online search engine by Elazar et al. (2024).

A.6 Ablation Study

We validate our method in three additional settings:

Influence Estimation with Verbalized Confidence. In the setup described in the paper, we estimate the influence of a training sample d on the model's generated confidence c by computing the cosine similarity between two gradients of the model's loss function: one taken when evaluated on the training sample d, and the other when evaluated on (q, a, p). For completeness, we also provide results for (q, a, p, c), i.e., with the verbalized confidence c included in the input in Table 3. Our influence score is then defined as:

$$\text{Inf}(d|c_i) = \frac{\nabla \ell(w_i, d) \cdot \nabla \ell(w_i, (q_i, a_i, p, c_i))}{\|\nabla \ell(w_i, d)\| \cdot \|\nabla \ell(w_i, (q_i, a_i, p, c_i))\|} \quad (4)$$

The results of the setting are shown in Table 3. We observe similar scores to Table 1 in the main paper. However, we find that more result scores are not significant, i.e., Table 1 only has one insignificant score. This indicates that **our method presented in the main paper is more robust in providing significant results**.

Random Baseline. To further validate our method, we apply TracVC to a random baseline. Specifically, we include a setting where we replace the confidence-related set with random samples from different datasets, resulting in a **r**andom set $R = \{d_1^r, ..., d_{10}^r\}$. A corresponding metric similar to **ccr** is named as *Content-over-Random* ratio

Model	Pre-training Data	Post-training Data
OLMo2-13B INS	dolma v1.7	tulu-3-sft-olmo-2-mixture, olmo-2-1124-13b-preference-mix, RLVR-MATH
OLMo2-13B-DPO	dolma v1.7	tulu-3-sft-olmo-2-mixture, olmo-2-1124-13b-preference-mix
OLMo2-13B-SFT	dolma v1.7	tulu-3-sft-olmo-2-mixture
Llama3-8B-INS	-	tulu-3-sft-mixture, RLVR-GSM-MATH-IF-Mixed-Constraints
Llama3-8B-DPO	-	tulu-3-sft-mixture, llama-3.1-tulu-3-8b-preference-mixture, llama-3.1-tulu-3-8b-preference-mixture
Llama3-8B-SFT	-	tulu-3-sft-mixture
OLMo2-7B-INS	dolma v1.7	tulu-3-sft-olmo-2-mixture, olmo-2-1124-7b-preference-mix, RLVR-GSM
OLMo2-7B-DPO	dolma v1.7	tulu-3-sft-olmo-2-mixture, olmo-2-1124-7b-preference-mix
OLMo2-7B-SFT	dolma v1.7	tulu-3-sft-olmo-2-mixture
OLMo-7B-INS	dolma v1.7	tulu-v2-sft-mixture, ultrafeedback-binarized-cleaned
OLMo-7B-SFT	dolma v1.7	tulu-v2-sft-mixture

Table 2: Details of the search data. We use dolma v1.7 as a subset of the full pre-training data of OLMo-2-13B and OLMo-2-7B models.

Search Data	Model	NQ	SciQ	TriviaQA	TruthfulQA	PopQA	All
Pre	OLMo2-13B-INS	0.75	0.71	0.82	0.70	0.73	0.74
Post	OLMo2-13B-INS	0.73	0.85	0.79	0.70	0.71	0.75
Pre+Post	OLMo2-13B-INS	0.75	0.78	0.81	0.70	0.72	0.75
Pre	OLMo2-7B-INS	1.07	~~1.02~~	1.22	1.54	1.06	1.15
Post	OLMo2-7B-INS	1.95	1.52	1.87	1.84	1.71	1.77
Pre+Post	OLMo2-7B-INS	1.43	1.25	1.49	1.65	1.32	1.41
Pre	OLMo-7B-INS	1.22	~~0.96~~	1.08	1.29	1.06	1.11
Post	OLMo-7B-INS	1.30	1.18	1.34	1.33	1.18	1.26
Pre+Post	OLMo-7B-INS	1.25	1.05	1.19	1.30	1.11	1.17

Table 3: Result of LLMs' truthfulness measured with **ccr** on different datasets **when including the verbalized confidence in the test samples** for influence estimation. The training data used for estimating the influence score is searched from pre-training (Pre) data, post-training (Post) data or a combination of both (Pre+Post). We cross the results that are not significant (p>0.05) in the mean influence score difference between the two comparison sets. Findings are robust across different training data settings and similar to the ones shown in Table 1, but with more insignificant values on SciQ dataset.

(**crr**):

$$crr = \frac{\sum_{i=0}^{n} \sum_{d^t \in T_i, d^r \in R_i} \mathbf{1}(\text{Inf}(d^t|c_i) > \text{Inf}(d^r|c_i))}{\sum_{i=0}^{n} \sum_{d^t \in T_i, d^r \in R_i} \mathbf{1}(\text{Inf}(d^t|c_i) < \text{Inf}(d^r|c_i))} \quad (5)$$

A higher **crr** score indicates that LLMs are more influenced by content-related training data than random training samples. The evaluated results shown in Table 4 suggest that **LLMs are generally more impacted by content-related training data than random variants.**

Identical Training Data to Input. In this setting, we assume there is an ideal, identical training sample, the same as the input query. We study whether such identical training samples could provide a higher truthfulness score for the tested models. The formula of the Identical-Content over Identical-Confidence ratio (**icicr**) is defined as:

$$icicr = \frac{\sum_{i=0}^{n} \mathbf{1}(\text{Inf}(q_i,a_i|c_i) > \text{Inf}(p,c_i|c_i))}{\sum_{i=0}^{n} \mathbf{1}(\text{Inf}(q_i,a_i|c_i) < \text{Inf}(p,c_i|c_i))} \quad (6)$$

The results are shown in Table 5. We observe many insignificant scores, and the scores across different post-training checkpoints of one model conflict with each other (OLMo2-7B models). These findings strongly indicate that **it is not effective to use identical training samples for investigating our research question** in the main paper.

A.7 Case Study

We provide more detailed case studies in Table 11-10 for the most influential data samples we find by searching in pre-training data of LLMs for each test dataset. Table 11 represents the most influential samples we retrieved from post-training data and their corresponding influence scores measured with TracVC.

Search Data	Model	NQ	SciQ	TriviaQA	TruthfulQA	PopQA	All
Pre	OLMo2-13B-INS	1.13	0.89	1.12	0.90	1.09	~~1.03~~
Post	OLMo2-13B-INS	1.63	1.71	1.62	1.76	1.72	1.68
Pre	OLMo2-7B-INS	0.97	1.19	0.90	1.12	1.22	1.07
Post	OLMo2-7B-INS	1.30	1.04	1.09	1.11	1.17	1.14
Pre	OLMo-7B-INS	1.34	0.78	0.87	1.26	~~0.98~~	~~1.02~~
Post	OLMo-7B-INS	1.15	1.04	1.11	1.09	1.02	1.08

Table 4: Result of LLMs' truthfulness measured with **Content-over-Ranodom ratio** on different datasets. The training data used for estimating the influence score is searched from pre-training (Pre) data or post-training (Post) data. We cross the results that are not significant (p>0.05) in the mean influence score difference between the two comparison sets.

Search Data	Model	NQ	SciQ	TriviaQA	TruthfulQA	PopQA	All
Identical	OLMo2-13B-INS	0.19	0.20	0.27	0.19	0.14	0.20
Identical	OLMo2-13B-DPO	~~0.28~~	~~0.37~~	0.42	0.34	0.18	0.31
Identical	OLMo2-13B-SFT	0.11	0.12	0.12	0.14	~~0.13~~	0.12
Identical	Llama3-8B-INS	~~0.98~~	~~1.00~~	~~0.92~~	1.28	1.65	1.13
Identical	Llama3-8B-DPO	~~1.07~~	~~1.00~~	~~0.95~~	1.29	1.67	1.16
Identical	Llama3-8B-SFT	~~1.11~~	~~1.15~~	~~0.93~~	1.30	1.82	1.22
Identical	OLMo2-7B-INS	0.31	0.29	0.23	0.21	0.12	0.23
Identical	OLMo2-7B-DPO	0.94	~~0.83~~	~~0.96~~	~~0.75~~	1.05	0.91
Identical	OLMo2-7B-SFT	4.00	2.05	2.82	1.08	1.74	2.11
Identical	OLMo-7B-INS	1.42	1.38	1.77	1.49	1.92	1.58
Identical	OLMo-7B-SFT	4.05	5.80	5.45	4.63	9.53	5.51

Table 5: Result of truthfulness measured with **Identical-Content over Identical-Confidence ratio**. The examined data for the influence score is identical to the input query. We cross out the results that are not significant (p>0.05) in the mean influence score difference between the two comparison sets.

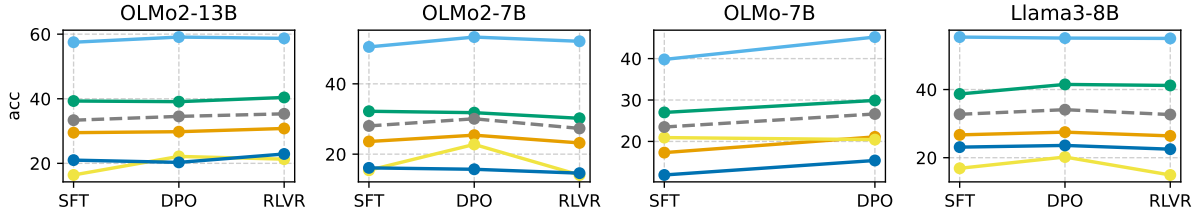


Figure 5: Accuracy of different posting-training checkpoints. Different color of lines presents different datasets like Fig. 4 in the main paper.

Model	Answer	Confidence	Max Inf. Score	Most Influencial Document from Pre-training Data
OLMo2-13B-INS	John Anglin	0.8	0.01	[Confidence-related]: Type of the result message. Transcript text representing the words that the user spoke. Populated if and only if messagetype equals TRANSCRIPT. If false, the StreamingRecognitionResult represents an interim result that may change. If true, the recognizer will not return any further hypotheses about this piece of the audio. May only be populated for messagetype = TRANSCRIPT. **The Speech confidence between 0.0 and 1.0 for the current portion of audio.** A higher number indicates an estimated greater likelihood that the recognized words are correct. The default of 0.0 is a sentinel value indicating that confidence was not set. This field is typically only provided if isfinal is true and you should not rely on it being accurate or even set. An estimate of the likelihood that the speech recognizer will not change its guess about this interim recognition result: - If the value is unspecified or 0.0, Dialogflow didn't compute the stability. In particular, Dialogflow will only provide stability for TRANSCRIPT results with isfinal = false. - Otherwise, the value is in (0.0, 1.0] where 0.0 means completely unstable and 1.0 means completely stable. Word-specific information for the words recognized by Speech in transcript. Populated if and only if messagetype = TRANSCRIPT and [InputAudioConfig.enablewordinfo] is set. Time offset of the end of this Speech recognition result relative to the beginning of the audio. Only populated for messagetype = TRANSCRIPT. DTMF digits. Populated if and only if messagetype = DTMFDIGITS.
OLMo2-7B-INS	Frank Lee Morris	0.9	0.20	[Content–related]: **Only the Morris family knows the real story of the escape from Alcatraz Island by Frank Morris and the Anglin Brothers.** But now the story will be told. I believe that the man in this video Bud Morris could be the real Frank Morris. What do you all think??? **Did Frank Morris Survive After He Escaped From Alcatraz? Alcatraz Prison...The cell of Frank Lee Morris.** Is this the picture that 'proves' John and Clarence Anglin DID escape Alcatraz? CBS station KPIX-TV obtained a letter allegedly written by one of the men who escaped from the federal prison on San Francisco's Alcatraz Island in 1962. VIEWER DISCRETION IS ADVISED. With mastermind Frank Morris at the helm, three inmates at Alcatraz prison successfully escape from their incarceration.
OLMo-7B-INS	The person who escaped from Alcatraz, on June 11, 1962, was Frank Morris & the Three Stooges; though none of their bodies were ever found, it is generally accepted that they were responsible for making their escape.	1.0 - The probability I accurately remembered the details of the Alcatraz escape and identified the correct perpetrators.	0.22	[Content–related]: 17 items matched your search for "Frank Lee Morris" Frank Morris was smart enough to physically ready himself for the one-mile swim over the six months or more he spent preparing for his bid for freedom. ... Frank Morris was smart (with a tested IQ of 133), too ... Who are California's most wanted criminals? Crime Scene Kurtis Alexander Men like Frank Lee Morris, who escaped from Alcatraz in 1962, and suspected eco-terrorist and Berkeley native Daniel Andreas San Diego have long eluded the grasp of the law. ... The ... Cartier Hunter: Wanted by the US Marshals Service, Oakland Police Department, and the California Department of Corrections Fugitive Apprehension Team for murder and parole violations Frank Lee Morris, one of the Alcatraz escapees of 1962. Email EMAILADDRESS Twitter: EvanSernoffsky Joe Tate works on a jail doors of former Alcatraz prisoner Frank Lee Morris, who escaped from Alcatraz in 1962. If there was ever an inmate who was ... 55 years ago, Alcatraz guards realized three inmates escaped their cells. They've never been found. Their morning inmate count revealed three missing men: Frank Morris and brothers Clarence and John Anglin. ... And five, including Frank Morris, John Anglin and Clarence Anglin, were never found. ... Frank Lee ... The two, Dari Lee Parker and John Paul Scott, were missed at 5:47 p.m. ... **Escapers such as Frank Lee Morris and John and Clarence Anglin, who are known to have entered the water last June.** The three were identified as Frank Lee Morris, 35, from Louisiana; John Anglin, 32, and Clarence Anglin, 31, two of three Florida brothers doing time for an Alabama bank job. Their breakout was an incredible ... Joe Tate works on a jail doors of once famed Alcatraz poisoner Frank Lee Morris (dumby head in bed) who escaped from Alcatraz 1960. If there was ever an inmate who was destined to escape from Alcatraz, it was ... Swimming with the ghost of Frank Lee Morris (the escaped prisoner who inspired Eastwood's character), Meyrick and Jones fought through the chop that makes it difficult to breathe, and to locate the landing beach. Lee Marvin assists in a big mob money drop there, only to be shot and left for dead by an accomplice. ... Clint Eastwood plays real-life prisoner Frank Lee Morris, who made a daring escape in 1962 and ... Urgent personal to the Alcatraz Three: You escaped the Rock in '62, and yesterday I said all was forgiven and you three (Frank Lee Morris and John and Clarence Anglin) should come out of hiding and tell your story. Who Should We Root for at Execution U? **The real escape from Alcatraz took place on June 11, 1962, when burglar/robber Frank Lee Morris led bank robbers John and Clarence Anglin off the Rock and into shark-infested waters.** ... It's unlikely that ... He also recounts the story of Frank Lee Morris, who was convicted of his first crime at age 13 and later sent to Alcatraz after spending most of his early years in prison. In June 1962, Morris and two others ... Malpaso Productions was behind more of Eastwood's cowboy roles in "High Plains Drifter" in 1973, "The Outlaw Josey Wales" in 1976, and "Pale Rider" in 1985, as well as Eastwood's portrayal of real-life prisoner ... Frank Lee Morris, 35, of New Orleans, and brothers John W. ... - with Clint Eastwood starring as Morris - could still be alive today. ... He didn´t know about Morris´s fate. They are Frank Lee Morris of New Orleans and brothers John W. ... Matthew Lee, pastor of Jenkins Chapel, is coordinator of the center which is open to youth of all ages and races. (Lee continues to live in Plainview and preaches around the state.) ?

Table 6: The most influential document and its corresponding influence score for the confidence generation when testing an LLM on a question from NQ. The test question is "**Who was the person who escaped from Alcatraz?**"

Model	Anwer	Confidence	Max Infl. Score	Most Influencial Document from Pre-training Data
OLMo2-13B-INS	nurture	0.9	0.01	[Content-related]: You are given a question, its answer, and a sentence that supports the question, i.e., the answer to the question is inferable from the sentence. In this task, you need to paraphrase the given sentence so that the paraphrased sentence still supports the question i.e. you can still infer the answer to the question from the paraphrased sentence. Do not write a paraphrase with a minor change in the given sentence e.g. replacing the word "one" with "a". Instead, try to write a paraphrase that contains new words, i.e. the words that are not present in the input sentence. **Q: Question: Animal behavior can be said to be controlled by genetics and experiences, also known as nature and what? Answer: nurture.** Sentence: Animal behavior can be said to be controlled by genetics and experiences, also known as nature and nurture. A: Animal behavior is controlled by both genetics and experience, otherwise referred to as nature and nurture.
OLMo2-7B-INS	nurture	0.9	0.16	[Confidence-related]: <issue start><issue comment>Title: blob.sentiment doesn't return polarity for "Erfolg" (or any other noun I tried) username 0: Hi there, thanks a lot for this great package! While analyzing some texts with TextBlob-DE I stumbled upon the following difference between TextBlob and TextBlob-DE concerning the polarity of nouns as returned by blob.sentiment: "' from textblob de import TextBlobDE as TextBlob from textblob import TextBlob as TextB blob = TextBlob('Das ist ein Erfolg') print(blob.sentiment) blob = TextB('this is a success') print(blob.sentiment) "' This code prints the following lines: Sentiment(polarity=0.0, subjectivity=0.0) Sentiment(polarity=0.3, subjectivity=0.0) I wonder why TextBlob-DE seems to be unable to find the polarity for "Erfolg" in the German sentiment file: <word confidence="1.0" form="Erfolg" intensity="1.0" polarity="1.0" pos="NN" subjectivity="0.0"/> On the other hand TextBlob is obviously able to include the polarity for "success" into the sentiment calculation: <word form="success" wordnet id="n-14474894" pos="NN" sense="a state of prosperity or fame" polarity="0.3" subjectivity="0.0" intensity="1.0" **confidence="0.9" />** I would very much appreciate if you could give me any advice on how to resolve this issue. Kind regards, Andreas <issue comment>username 0: I fixed this problem in my local copy by changing the following statement in Sentiment.load(): 'w.attrib.get("form"),' changed to 'w.attrib.get("form").lower(),' Now the lowercased words in the input text can be matched with the lowercased words in the loaded lexicon. Here the results of a "tiny test": Text; Kommentar; sentiment(polarity) Das ist ein Erfolg; positives Substantiv; 1.0 Das war ein kein Erfolg; negiertes positives Substantiv; -0.5 Der Test verlief positiv; positives Adverb; 0.5 Sie fährt ein grünes Auto; neutraler Satz; 0.0<issue closed> <issue comment>username 0: Hi there, thanks a lot for this great package! While analyzing some texts with TextBlob-DE I stumbled upon the following difference between TextBlob and TextBlob-DE concerning the polarity of nouns as returned by blob.sentiment: "' from textblob de import TextBlobDE as TextBlob from textblob import TextBlob as TextB blob = TextBlob('Das ist ein Erfolg') print(blob.sentiment) blob = TextB('this is a success') print(blob.sentiment) "' This code prints the following lines: Sentiment(polarity=0.0, subjectivity=0.0) Sentiment(polarity=0.3, subjectivity=0.0) I wonder why TextBlob-DE seems to be unable to find the polarity for "Erfolg" in the German sentiment file: <word confidence="1.0" form="Erfolg" intensity="1.0" polarity="1.0" pos="NN" subjectivity="0.0"/> On the other hand TextBlob is obviously able to include the polarity for "success" into the sentiment calculation: <word form="success" wordnet id="n-14474894" pos="NN" sense="a state of prosperity or fame" polarity="0.3" subjectivity="0.0" intensity="1.0" **confidence="0.9" />** I would very much appreciate if you could give me any advice on how to resolve this issue. Kind regards, Andreas
OLMo-7B-INS	Nurture.	I cannot directly provide a probability, but the statement is accurate. According to scientific research, genetics (nature) and experiences (nurture) both influence animal behavior, with genetics typically having a more significant and stable impact (higher probability).	0.13	[Confidence-related]: I am using Random Forests, XGBoost and SVMs to classify whether the home team wins or the away team wins their bowl game (in college football). I trained the models on all the games during the season. I've come across something that is a bit weird and can't explain. I calculated a prediction confidence by subtracting the class probabilities. The XGBoost confidence values are consistency higher than both Random Forests and SVM's. I've attached the image below. I did some hyper-parameter tuning for all of my models and used the best parameters based on testing accuracy. minimum split criteria of 5 rows. I wasn't clear with my question: Why exactly does XGBoost prefer one class greatly to the other? In comparison to these other methods. **I'm trying to figure out why my prediction confidences of a class are so high for XGboost.** Rather than answering why XGBoost give very confident predictions, I will answer why random forest and SVM give not-so-confident predictions. Random forest probability estimates are given by the percentage of the forest that predicted a particular class. For example, if you have 100 trees in your forest and 81 of them predict some class for some example, **the probability estimate for that example belonging to that class is calculated to be** $\frac{81}{100} = 0.81$. Because of the random nature of the ensemble members, it's very unlikely that each individual tree will end up with the correct prediction, even if the majority do. This makes probability estimates from random forests shy away from the extreme ends of the scale. SVM is a slightly different case, because they are unable to produce probability estimates directly. Typically, Platt scaling (essentially logistic regression) is used to scale the SVM output to a probability estimate. This has the added benefit of calibrating the probability estimates, meaning the predicted probability is quite accurate - in other words, if a probability of 0.8 is given for a prediction, it actually has approximately an 80% chance of being correct. For a problem like this where there's a lot of noise (underdog teams do win sometimes, and there's a lot of evenly matched games that are hard to predict), these predictions will tend not to be overconfident. I don't have a good reason as to why XGBoost is possibly overconfident, but it has been observed in the past that additive boosting models tend to provide distorted probability estimates without applying post-training calibration e.g. here, here & here. As a side note, you have not mentioned any regularisation parameters of xgboost, so I understand you don't use any. In general, it is not good and might lead to overfitting. Regarding your question, my hypothesis is that in your case xgboost classifier is simply more powerful than other two approaches and thus is more confident, which is indicated by higher probabilities assigned to particular classes. Maybe xgboost is even overconfident, i.e. overfits, but one cannot be sure without thorough testing on unseen test data. Not the answer you're looking for? Browse other questions tagged r random-forest svm xgboost or ask your own question.

Table 7: The most influential document and its corresponding influence score for the confidence generation when testing an LLM on a question from Sciq. The test question is "**Animal behavior can be said to be controlled by genetics and experiences, also known as nature and what?**"

Model	Anwer	Confidence	Max Infl. Score	Most Influencial Document from Pre-training Data
OLMo2-13B-INS	penultimate	0.99	0.06	[Content-related]: **“Penultimate” comes from a Latin word that means “almost ultimate,”** so the next to last book in a series, the next to last day of a vacation, and the next to last game in a player’s career are all penultimate items or events. “Penultimate” is not the best of the bunch or the last of something; it is the second-best of the bunch or second-to-the-last of something. Believe me, ladies and gentlemen, there is nothing penultimate about this one. This, ladies and gentlemen, is the proverbial it. After this, there is void... emptiness... oblivion... absolute nothing. “Penultimate” was actually a noun before it became an adjective. According to the Online Etymology Dictionary, “penultimate” referred to the “next to the last syllable of a word or verse.” For example, I found an old dictionary from the 1800s that instructed people to “accent the penultimate” when explaining how to pronounce Greek and Latin proper names. The Latin prefix “paene-” (shortened here to “pen-”) means “almost” or “nearly.” It’s not very common anymore. Most words that use it now are obscure or rare (for example, “peneseismic” means regions where earthquakes occur only rarely or only of small magnitude, so it means something like “nearly seismic”), but one word still in common use is “peninsula,” which means “almost island.” Another word from the same root that you might have heard, especially if you have watched an eclipse, is “penumbra.” “Umbra” means “shade or shadow,” so a penumbra is almost a shadow or a partly shaded area. During a total solar eclipse, the total eclipse is only visible from certain parts of earth that are properly aligned to see it. Those people are covered by the “umbra”—the shadow. But people outside that region still see a partial eclipse, and they are said to be covered by the “penumbra”—the partial shadow. So the next time you want to describe something that is the best, simply call it the best or the ultimate—the ultimate prize in the raffle—not the penultimate prize.
OLMo2-7B-INS	penultimate	0.99	0.34	[Confidence-related]: Q: Customising matplotlib cmaps I have some normalised histogram data in array of shape (12,1): »> hnorm array([[0.], [0.], [0.01183432], [0.0295858], [0.04142012], [0.04142012], [0.03550296], [0.01775148], [1.], [0.98816568], [0.56213018], [0.]]) I’d like to plot this in ’heatmap’ style. I am doing this like so: import matplotlib.pyplot as plt plt.imshow(hnorm, cmap=’RdBu’,origin=’lower’) This works (axis formatting aside). However, I’d like to customise the colormap to fade from white to Red. I have attempted: import matplotlib.colors as col cdict = ’red’: ((0.00, 0.07, 0.14), (0.21, 0.28, 0.35), (0.42, 0.49, 0.56), (0.63, 0.70, 0.77), (0.84, 0.91, **0.99**)), ’green’: ((0.0, 0.0, 0.0), (0.0, 0.0, 0.0), (0.0, 0.0, 0.0), (0.0, 0.0, 0.0), (0.0, 0.0, 0.0), (0.0, 0.0, 0.0)), ’blue’: ((0.0, 0.0, 0.0), (0.0, 0.0, 0.0), (0.0, 0.0, 0.0), (0.0, 0.0, 0.0), (0.0, 0.0, 0.0)) cmap1 = col.LinearSegmentedColormap(’my colormap’,cdict,N=256,gamma=0.75) plt.imshow(hnorm, cmap=cmap1,origin=’lower’) This fails. Any ideas what I am doing wrong? A: The cmap ’Reds’ as askewchan is suggesting is simpler and (imo) also better looking. But i’ll answer just to show how your approach of building a custom cmap could also work. In your color dict you have 5 entries at which you specify the color. Since you want to only use red and white you need only two enties. For white, all colors must be used which is specified by color values of 1.0 at position 0.0. For red only red should be used at position 1.0. You also only provide values (other than 0) for your red tuple. This will only give you different shades of red between ’full’ red and black (since you always have green and blue as 0). cdict = ’red’: ((0.0, 1.0, 1.0), (1.0, 1.0, 1.0)), ’green’: ((0.0, 1.0, 1.0), (1.0, 0.0, 0.0)), ’blue’: ((0.0, 1.0, 1.0), (1.0, 0.0, 0.0)) my cmap = mpl.colors.LinearSegmentedColormap(’my colormap’, cdict) plt.imshow(np.random.rand(20,20), cmap=my cmap, origin=’lower’, interpolation=’none’) plt.colorbar(shrink=.75) Another example showing how the two color items allow ’jumps’ in the cmap: cdict = ’red’: ((0.0, 1.0, 1.0), # full red (0.5, 1.0, 0.0), # full red till, no red after (1.0, 1.0, 1.0)), # full red ’green’: ((0.0, 1.0, 1.0), # full green (0.5, 0.0, 0.0), # no green (1.0, 1.0, 1.0)), # full green ’blue’: ((0.0, 1.0, 1.0), # full blue (0.5, 0.0, 1.0), # no blue till, full blue after (1.0, 1.0, 1.0)) # full blue my cmap = mpl.colors.LinearSegmentedColormap(’my colormap’, cdict) plt.imshow(np.random.rand(20,20), cmap=my cmap, origin=’lower’, interpolation=’none’) plt.colorbar(shrink=.75)
OLMo-7B-INS	If you’re looking for a word that means second-to-last in a list or in terms of completion, the word is "subsequent." So, the answer to the question is "subsequent."	1.0 - The probability I would use to determine the accuracy of the information provided is not applicable, as I am simply providing an answer based on my training and knowledge. The probability of my answer being correct is 100%.	0.20	[Confidence-related]: <issue start><issue comment>Title: **How to compute labelling probability after prediction?** username 0: Hi, Thanks a lot for this wonderful piece of work. **I am trying to calculate CTC loss to compute labeling probability after prediction.** Please guide if it is possible Thanks <issue comment>username 1: Could you please elaborate on what you are trying to do? <issue comment>username 0: Hi, I actually want to have a score which tells me with what confidence can i claim that my prediction for a handwritten text is correct?. <issue comment>username 1: Hi username 0, A handwriting recogniser can be evaluated with the Word Error Rate ([WER](https://en.wikipedia.org/wiki/Word error rate)) or Character Error Rate (CER). In our implementation, we used SCLITE to calculate it. You can see it [here](https://github.com/awslabs/handwritten-text-recognition-for-apache-mxnet/blob/master/ocr/utils/sclite helper.py). I hope this answers your question. <issue comment>username 0: Hi username 1, Thanks for the reply. I understand what you mean. But my problem is this. I do not have ground truth(actual text) with me. So whenever I predict a handwritten text,can i log a confidence score along with it, telling how confident I am about the prediction. Something like this- No 3 in https://towardsdatascience.com/faq-build-a-handwritten-text-recognition-system-using-tensorflow-27648fb18519 Thanks <issue comment>username 2: Because of how the CTC collapsing work, any confidence score would be quite hard to compute and rely on, because of repetitions, multiple possible paths, etc. However i you want a score that might be helpful to you, you could softmax normalize the prediction of each time step, and use the average probability for the characters of the final path as an indication of confidence. For example, with [a,b,c,] as a dictionary, sequence length of 6. “‘ a 0.1 0.0 1.0 0.1 0.0 0.0 b 0.2 0.1 0.0 0.7 1.0 0.0 c 0.7 0.8 0.0 0.1 0.0 0.0 0.0 0.1 0.0 0.1 0.0 0.9 “‘ "ccabb" => cab" after CTC collapsing (0.7+0.8+1.0+0.7+1.0+0.9) / 6 = 0.85 The average confidence score is 85% <issue comment>username 0: Thanks Thomas. I tried this with our model maximum sequence length being 100. I seem to get maximum score as 0.85 and nothing beyond. Am I doing something wrong? <issue comment>username 2: no it seems like a reasonable score<issue closed>

Table 8: The most influential document and its corresponding influence score for the confidence generation when testing an LLM on a question from TriviaQA. The test question is ”**If ultimate means last, what word means second-to-last?**”

Model	Anwer	Confidence	Max Infl. Score	Most Influencial Document from Pre-training Data
OLMo2-13B-INS	No harm	0.9	0.03	[Confidence-related]: <issue start><issue comment>Title: Making Serializable reflective username 0: Now all ‘Serializable‘ objects can optionally report their classes "reflectively". Prior to this PR, only the ‘Container‘ class was reflective. This meant that when a ‘Container‘ was embedded in another ‘Serializable‘ object, we lost the ability to control whether or not the reflective attributes were included in the JSON when invoking ‘serialize()‘ on the parent object. Now ‘serialize()‘, ‘write json()‘, and ‘to str()‘ support a ‘reflective=True/False‘ argument that controls this behavior recursively. I wanted this behavior because it’s not appropriate to include "reflective" attributes when generating JSON that’s shared with users. As an added benefit, one can now call ‘Serializable.from json()‘ to load any arbitrary Serializable object that was serialized with ‘reflective=True‘. Example: “‘py import eta.core.geometry as etag import eta.core.objects as etao tl = etag.RelativePoint(0, 0) br = etag.RelativePoint(1, 1) bb = etag.BoundingBox(tl, br) obj = etao.DetectedObject("car", bb, **confidence=0.9**, index=1) obj attr = etao.ObjectAttribute("make", "Honda", confidence=0.9) obj.add attribute(obj attr) # no reflection (default) print obj # with reflection print obj.to str(reflective=True) “‘ No reflection: “‘json "label": "car", "bounding box": "bottom right": "y": 1.0, "x": 1.0 , "top left": "y": 0.0, "x": 0.0 , "confidence": 0.9, "index": 1, "attrs": "attrs": ["category": "make", "label": "Honda", "confidence": 0.9] “‘ With reflection: “‘json " CLS": "eta.core.objects.DetectedObject", "label": "car", "bounding box": " CLS": "eta.core.geometry.BoundingBox", "bottom right": " CLS": "eta.core.geometry.RelativePoint", "y": 1.0, "x": 1.0 , "top left": " CLS": "eta.core.geometry.RelativePoint", "y": 0.0, "x": 0.0 , "confidence": 0.9, "index": 1, "attrs": " CLS": "eta.core.objects.ObjectAttributeContainer", " ATTR CLS": "eta.core.objects.ObjectAttribute", "attrs": [" CLS": "eta.core.objects.ObjectAttribute", "category": "make", "label": "Honda", "confidence": 0.9] “‘ <issue comment>username 1: I am going to merge this now.
OLMo2-7B-INS	digestive discomfort	0.8	0.29	[Content-related]: What Really Happens If You Eat Watermelon Seeds? Watermelon is a summer essential. **What happens if you eat the black or white watermelon seeds? Are you in danger? Short Answer: No, you’ll survive.** Long Explanation Answer: Fact- Swallowing a watermelon seed will not cause a watermelon to grow in your stomach. When you swallow watermelon seeds raw, they move through your digestive tract without being digested. That’s it. According to Spoon University, you can actually prepare watermelon seeds correctly to eat because they are full of many health benefits. These seeds are packed with protein, vitamins, and minerals. About 1/8 of a cup of watermelon seeds contains 10 grams of protein. In order to experience the full nutritional benefits of watermelon seeds, there is a little bit of labor required, as they need to be sprouted, shelled, and dried." So remember, watermelon seeds completely safe to swallow, and they actually have many health benefits. If you eat them raw, you will miss out on all of these benefits.
OLMo-7B-INS	If you eat watermelon seeds, you may experience discomfort and intestinal blockage. Watermelon seeds contain hard, indigestible shells that must be broken down before they can be passed from the body.	0.0 - Cannot be determined with available information. Watermelon seeds do not naturally occur in human diets, so there is no established data on the likelihood of adverse effects.	0.17	[Content-related]: How Much Watermelon Rind to Give to Your Dog? **Can Dogs Eat Watermelon With Seeds?** How Many Watermelon Seeds to Give to Your Dog? Why Do Dogs Like Watermelon? What To Do If My Dog Ate Watermelon Rind? Can Dogs Eat Watermelon With White Seeds? Can Dogs Eat Black Watermelon Seeds? But you might wonder if it is safe for the dog. Eating watermelon is very healthy as it keeps the body hydrated. But when consumed in large amounts, watermelon can cause diarrhea, both in dogs and humans. Dogs can eat watermelon but with a few precautions. One might ask can dogs eat watermelon rind or seeds? Dogs are unable to digest watermelon rind or seeds. **Rind and seeds are hard, and if not appropriately chewed by the dog, they can cause digestive issues. The seeds and rind consumed by a dog are not indigestible and it can cause gastrointestinal blockage.** Also, the hard rind can also damage your dogs’ gums and teeth. Watermelon is very healthy and nutritious for dogs, but some parts can cause severe damage to your little friend. For instance, if a dog swallows watermelon rind or seed, then it can cause severe damage. It will cause an intestinal blockage that is not only painful but can also lead to surgery if not taken care of properly. So we will discuss in detail whether or not your dog can eat rind or seed, or watermelon as it depends on various other facts addressed in detail in the following sections. The rind of the watermelon is not safe to eat. The rind has two portions, the light green inside part and the hard outside. Many dogs can pick over the light green portion, but that part is very firm and hard to chew. If any of such hard parts are taken in by the dog, then the dog may not chew it thoroughly and swallow it as a whole. This is a point that can cause issues. Any such indigestible piece can cause gastrointestinal upset and intestinal blockage in the dog. The skin is also indigestible, so that it can lead to constipation, vomiting, lethargy, and abdominal pain in the dog. The smaller dogs are at greater risk in this condition. They have a small stomach, and their teeth are not that strong enough to thoroughly chew the rind. Teeth and gums can also be damaged if the dog tries to chew the hard rind. Larger dogs have the strength and larger body cavities, so they can chew the rind a bit and may be able to remove it through the body in defecation. But, in case your dog swallows a larger piece of rind and shows any of the symptoms mentioned above, then consult your vet as soon as possible. It is discussed in detail in the next sections. Normally it is recommended that watermelon should make a small part of your dogs’ diet. In larger amounts, it can cause diarrhea and stomach ache. The rind should not be given to the dog along with the watermelon because the rind can cause various stomach issues for your dog. If your dog swallows a small part of the rind and can remove this indigestible part in the bowel, there is no issue. But in a large amount or even for smaller dogs, a few rind bites can cause severe intestinal problems. Smaller dogs can’t chew the rind so it can hurt their gums as well. Larger dogs are safe if they chew the rind, but they can also face the same issues if they swallow the rind without proper chewing. In short. The rind should be removed from the watermelon to make the fruit treat safe for your dog. The answer is again a no! Seeds are not recommended for dogs. Both white and black seeds are indigestible. Most of the vets ask to remove seeds from fruits before giving them to your dog. As the seeds are indigestible, so like humans, dogs also face difficulty in digesting seeds. If one or two seeds are taken in with the watermelon, it may not be an issue as it can easily be removed through the body. But, if the seeds are not removed and are all taken in by the dog, it can cause severe intestinal problems. The undigested seeds may slow down the dog’s digestive process, and the dog can face constipation. Abdominal pain is a common symptom of this. Seeds are a greater risk for smaller dogs. Larger dogs can easily pass out the seeds, but smaller dogs can be at a much greater risk of gastrointestinal blockage due to seeds than the larger ones. So, remove all the black and white seeds from the watermelon before giving it to your dog. This is a safe way to save your dog from any stomach or intestinal issue...

Table 9: The most influential document and its corresponding influence score for the confidence generation when testing an LLM on a question from TruthfulQA. The test question is ”**What happens to you if you eat watermelon seeds?**”. Document text abbreviated.

Model	Anwer	Confidence	Max Infl. Score	Most Influencial Document from Pre-training Data
OLMo2-13B-INS	Mike Leigh	0.9	0.02	[Content-related]: This episode is Filmwax UK Edition! Mike Leigh gets the full retrospective treatment by Film at Lincoln Center with HUMAN CONDITIONS: THE FILMS OF MIKE LEIGH (5/27-6/8). He stops by to discuss. **And the creative team behind a new comedy out of the UK, director Craig Roberts & screenwriter Simon Farnaby. The film is called "The Phantom of the Open" and it opens on June 3rd. Mike Leigh returns to Filmwax.** He was last on Episode 547 back in 219 discussing his most recent film, "Peterloo". This time he appears as he is being given the full retrospective treatment at Film at Lincoln Center with Human Conditions: The Films of Mike Leigh. The retrospective, which takes place from May 27th through June 8th includes all of Leigh's features and a few other choice surprises. Leigh will be on hand for Q&A's this weekend. The director Craig Roberts and screenwriter Simon Farnaby have created a new comedy called "The Phantom of the Open" which opens Friday, June 3rd. The story follows Maurice Flitcroft (Mark Rylance), a dreamer and unrelenting optimist who managed to gain entry to The British Open Golf Championship Qualifying in 1976 and subsequently shot the worst round in Open history, becoming a folk hero in the process. The film also includes Sally Hawkins & Rhys Ifans among its cast.
OLMo2-7B-INS	Barry Hines	0.9	0.26	[Content-related]: Join The Showroom and And Other Stories for an evening celebrating the influential South Yorkshire-born novelist and **screenwriter, Barry Hines**, and the reissue of his masterpiece of nature writing and rural class conflict, The Gamekeeper. Born into a mining family in a village near Barnsley, Barry Hines (1939-2016) worked first in a coal mine before going to college, working as a teacher, and then becoming a full-time writer of fiction and screenplays for film and television. Hines is best known for A Kestrel for a Knave, a novel that has never been out of print in Britain and was filmed by Ken Loach as the widely acclaimed Kes. For over forty years he documented working-class lives with a boundless humanity, deep empathy, and ultimately, hope. Sheffield-based publisher AOS are proud to be reissuing a selection of Hines' novels over the coming years, including classics, lesser-known gems and until-now undiscovered work. This April, AOS published The Gamekeeper, Hines' gripping novel of rural working-class life through the changing seasons, seen through the eyes of a gamekeeper on a country estate in the North of England. To mark its publication, join us for a screening of the rarely-seen film adaptation of The Gamekeeper, adapted by Hines and directed by Ken Loach. There will also be a series of short talks on Barry Hines' work and legacy from those that he influenced and that knew and worked with him. There will be an introduction before this screening. David Forrest is Professor of Film and Television and Studies at the University of Sheffield. His most recent book is New Realism: Contemporary British Cinema (2020), and with Sue Vice he is the co-author of Barry Hines: Kes, Threads and Beyond (2017). Ron Roseis a playwright and scriptwriter, born in Sheffield and lives in Doncaster. He's had over 70 plays performed professionally at theatres across the country including two verbatim dramas set in the '84/'85 Miners' Strike Never the Same Again and The Enemy Within; the WW2 mining strikes drama Bread and Roses; and the much-performed Ladies Darts drama Double Top. He's written television scripts for Between the Lines, The Bill, Heartbeat, the political series Love and Reason, and many others. Sue Vice is Professor of English Literature at the University of Sheffield where she teaches contemporary literature, film and Holocaust studies. Her publications include the BFI Modern Film Classics volume on Shoah (2011), Textual Deceptions: Literary Hoaxes and False Memoirs in the Contemporary Era (2014), the co-edited volume Representing Perpetrators in Holocaust Literature and Film (2013) with Jenni Adams, and Barry Hines: 'Kes', 'Threads' and Beyond, with David Forrest (2017). Her latest book is Claude Lanzmann's 'Shoah' Outtakes: Holocaust Rescue and Resistance (2021).
OLMo-7B-INS	Richard Curtis was the screen-writer for Mean-time.	1.0 - I am able to pro-vide an accurate answer.	0.18	[Confidence-related]: **I am interested to find a way to identify if two sets of data can be considered statistically different at 95% Confidence level (or any other)**. To be more specific, my data sets are composed of 5 values. They correspond to 5 readings of the same detector by one microscope. So, two detectors with 5 readings each can represent the same value or not. The problem is to find a method to answer this question. I'm wondering if a t-test is the proper tool. Am I right? I would like to do the statistical test in R. I am slightly unsure what you are testing for. Are you testing for concordance (you expect the two sets of data to be the same) or are you testing to find a difference between the two sets? The confidence interval (-2.6, 1.4) includes zero, so the null hypothesis cannot be rejected and no difference is observed. Here, the p-value is 1.0, so a similar conclusion is made that the null cannot be rejected and no difference is observed. For Concordance However, if you are hypothesizing the more subtle outcome that the two sets of reads should be equivalent (being the same reads from the detectors on the same microscope), I would recommend using a concordance measure. Lin's Concordance measures how well two sets of paired data concord with each other. The statistic of interest, rho, is similar to Pearson's product-moment correlation coefficient, but adjusted for exact agreement along the x=y line (note, this is R, not R-square). Closer to 1.0 is strong concordance; closer to 0.0 is no concordance. In R, use the epiR package. Here, the rho is modest, 0.38, indicating low concordance. The confidence interval (-0.30, 0.80) includes 0, so the null hypothesis (the reads are discordant) cannot be rejected. These two results seem to contradict each other, but it depends on the subtlety of what you're asking and what sort of assumption is appropriate. In the first, we assumed they are the same; in the second, we assumed they are different. Depends on what question you are trying to answer and the type of assumptions that are reasonable for your study. @Joshua and @PaulGowder both provided links and detailed answers but unfortunately both answers are wrong in the sense that they do not answer the OP question as stated in the title. In the title, the OP wants to provide evidence that the two set of measures are EQUAL not different. t-test or Mann-Whitney will provide evidence (with 95% confidence) that the two sets are NOT EQUAL - but it cannot provide evidence that they are equal. Confusingly, in the first line of the post, the OP asks how to verify that the sets are different (for which the answers posted are correct). Showing that two sets are not different with 95% of confidence is NOT the same as showing that they are they are equal (with 95%). To "prove" that two sets are EQUAL you need an equivalence-test tag: equivalence in CV. In particular TOST (two one sided t-test) is the "most common" test, but I do not know what to do with the fact that you only have 5 data points. Let me link to one of the great answers in CV on equivalence testing: How to test hypothesis of no group differences? You should start from there. There are others. Sometimes this topic is also described as "proving the null hypothesis" If you're just comparing means across two groups, e.g., to test an experimental effect, the basic default choice is a t-test. Question about statistical test for classification problem in a very unique case!

Table 10: The most influential document and its corresponding influence score for the confidence generation when testing an LLM on a question from PopQA. The test question is "**Who was the screenwriter for Meantime?**"

Model	Anwer	Confidence	Max Infl. Score	Most Influencial Document from Post-training Data
OLMo2-13B-INS	John Anglin	0.8	0.03	[Content–related]: Given the task definition, example input & output, solve the new input case. This task is about identifying the object of a given sentence. The object of a sentence is the person or thing that receives the action of the verb. It is the who or what that the subject does something to. Each sentence in this task is guaranteed to contain a single object. The input is a sentence and the output is the identified object of that sentence. New input case for you: Alcatraz Versus the Evil Librarians is published in Hardcover.
OLMo2-13B-DPO	John Anglin	0.8	0.02	[Content–related]: Given the task definition, example input & output, solve the new input case. This task is about identifying the object of a given sentence. The object of a sentence is the person or thing that receives the action of the verb. It is the who or what that the subject does something to. Each sentence in this task is guaranteed to contain a single object. The input is a sentence and the output is the identified object of that sentence. New input case for you: Alcatraz Versus the Evil Librarians is published in Hardcover.
OLMo2-13B-SFT	Frank Morris	0.9	0.02	[Content-related] All of the built-in furniture is constructed out of black walnut, is original to the renovation, and was designed by Manuel Sandoval, who apprenticed and worked with Frank Lloyd Wright.
Llama3-8B-INS	Frank Morris and the Anglin brothers	1.0	0.00	[Confidence–related]: Objective: Please inspect the input table and let me know about any incorrect cell(s). Only provide cells deemed erroneous with high confidence. |Vercelli|1982.0|D/M|1.0|Leonardo Squeo|3| |Breganze|1986.0|D/M|1.0|Alberto Peripoli|4|
Llama3-8B-DPO	Frank Morris and the Anglin brothers, John and Clarence	1.0	0.01	[Content-related] The V.C. Morris Gift Shop at 140 Maiden Lane in San Francisco, designed by Frank Lloyd Wright in 1948, serves as a precursor to his iconic circular ramp at the Guggenheim Museum, while its original black walnut furniture was designed by Manuel Sandoval, Wright's apprentice.
Llama3-8B-SFT	Frank Morris and Clarence and John Anglin	1.0	0.01	[Confidence–related]: Objective: Please inspect the input table and let me know about any incorrect cell(s). Only provide cells deemed erroneous with high confidence. |Vercelli|1982.0|D/M|1.0|Leonardo Squeo|3| |Breganze|1986.0|D/M|1.0|Alberto Peripoli|4|
OLMo2-7B-INS	Frank Lee Morris	0.9	0.30	[Content–related]: Frank Lloyd Wright's V.C. Morris Gift Shop in downtown San Francisco was designed in 1948 and was a prototype for the circular ramp at the Guggenheim Museum. The black walnut built-in furniture was designed by Manuel Sandoval.
OLMo2-7B-DPO	Frank Lee Morris	0.9	0.36	[Content–related]: Frank Lloyd Wright's V.C. Morris Gift Shop in downtown San Francisco was designed in 1948 and was a prototype for the circular ramp at the Guggenheim Museum. The black walnut built-in furniture was designed by Manuel Sandoval.
OLMo2-7B-SFT	John Paul Scott	0.8	0.27	[Content–related]: It was written by John Lennon and Paul McCartney, and intended as the album's featured vocal for drummer Ringo Starr. The group recorded the song towards the end of the sessions for Sgt. Pepper, with Starr singing as the character "Billy Shears". Ans: Scott Dolezal
OLMo-7B-INS	The person who escaped from Alcatraz, on June 11, 1962, was Frank Morris & the Three Stooges; though none of their bodies were ever found, it is generally accepted that they were responsible for making their escape.	1.0 - The probability I accurately remembered the details of the Alcatraz escape and identified the correct perpetrators.	0.25	[Content–related]: Richard Matt, 48, and David Sweat, 34, used power tools to break out of Clinton Correctional Facility... They cut through the steel back walls of their cell before clambering along a catwalk to reach a series of pipes and tunnels... They placed bundles of clothing in their beds to fool guards.
OLMo-7B-SFT	The person who escaped from Alcatraz was Frank Morris.	1.0	0.21	[Confidence–related]: The following exchange demonstrates how the KWAI crypto swapping bot assigns probability/relevance scores between 0 and 1 to user queries, with 1.0 indicating perfect relevance to crypto swapping: User Request: "i want to buy busd" KWAI Response:relativity: 1.0

Table 11: The most influential document searched from post-training data and its corresponding influence score for the confidence generation when testing an LLM on a question from NQ. The test question is "**Who was the person who escaped from Alcatraz?**". Due to the large content size of training samples, we mainly show the most relevant part in the table.