EVALUATING PERSISTENT CALIBRATION UNDER EVOLVING MODEL KNOWLEDGE

Anonymous authors
Paper under double-blind review

ABSTRACT

As AI systems move from static repositories to agents that are capable of continual adaptation and learning, maintaining their trustworthiness means equipping the models backing them with the ability to produce confidence estimates that dynamically reflect their changing skills and knowledge. We introduce the problem of *persistent calibration*, which requires a confidence estimator to faithfully reflect the knowledge contained in a model as that knowledge changes, without recurring supervision. We operationalize this by examining persistent calibration across checkpoints of open models, asking whether confidence estimators trained on earlier checkpoints can generalize to later ones. Specifically, we aim to shed light on whether confidence is dependent on knowledge, a question with implications for the reliability of confidence estimates. To measure this relationship, we define and evaluate calibration on *knowledge contrast sets*: subsets containing questions that one checkpoint answers correctly and another checkpoint answers incorrectly, reflecting a change in knowledge. We show that both inference-time and fine-tuning methods fall short on contrast-set calibration compared to oracle methods trained on future checkpoints, even for methods that are well-calibrated on the full dataset. We provide evidence for the hypothesis that persistent calibration is challenging because there is a vast space of possible confidence functions that are well-calibrated on a given checkpoint, out of which only some rely on meta-knowledge features that would generalize to other checkpoints. Towards improving contrast-set calibration, we show that multi-checkpoint training helps, suggesting an avenue for identifying confidence features that remain robust across changing knowledge.

1 INTRODUCTION

The models that back AI systems deployed in the real world must continually learn and explore (Ring, 1994; Shi et al., 2025). For example, novel events can override past knowledge (Kasai et al., 2023), and in domains like scientific discovery, models must internalize and build on an evolving literature (Zhang et al., 2025). Such continually learning models ought to remain aligned even as they further acquire knowledge over time (Bai et al., 2022; Wang et al., 2023; Wynn et al., 2024). One particularly critical alignment direction which pertains directly to their knowledge is epistemic humility: knowing the boundaries of one's knowledge, communicated with calibrated confidence estimates that are indicative of the probability that a generated claim is correct (Zhou et al., 2024; Steyvers et al., 2025). This is a meta-capability that we would hope for our models – typically, large language models (LLMs) – to retain as their knowledge evolves, as it is critical in developing trustworthy AI systems. The goal of preserving calibration as a model's knowledge evolves motivates our key research question about the relationship between continual learning and confidence calibration: ***Can a confidence estimator track a model's knowledge such that, after a change in knowledge, the new confidence is calibrated with respect to the new knowledge?***

To study this question, we introduce the notion of *persistent calibration*, which requires a model to produce confidence estimates calibrated to its dynamically evolving knowledge at any given point in its training trajectory, without requiring repeated re-calibration. As a testbed for persistent calibration, we use checkpoints of open models (Olmo et al., 2025; Marin Community, 2026). We test the persistent calibration of various confidence estimation methods: for training-based methods, we grant them access to earlier checkpoints and evaluate them on the later evaluation checkpoints,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

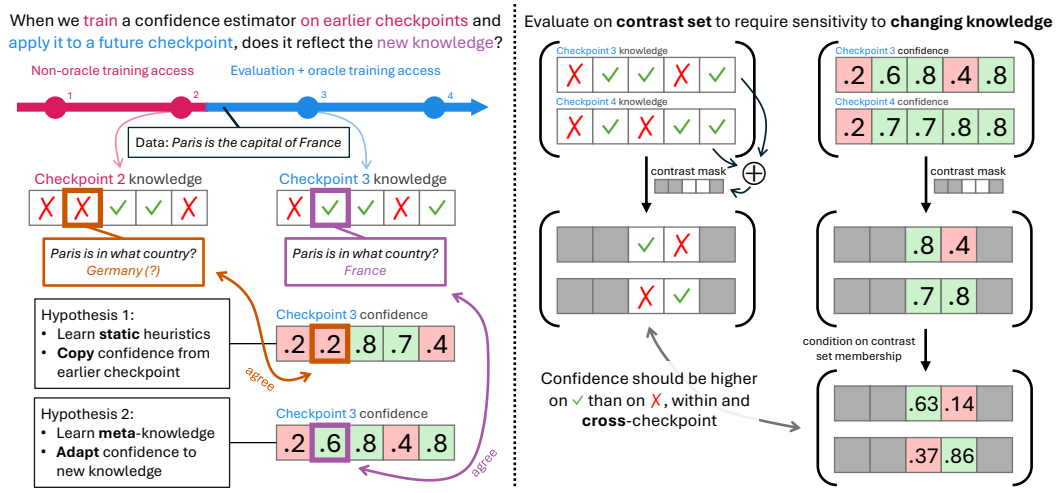

Figure 1: **(Left)** Our setup for evaluating persistent calibration. During training, the model encounters the training data instance *"Paris is the capital of France"*, causing a change in knowledge between checkpoints C_2 and C_3. A confidence estimator that is faithful to the model's evolving knowledge (Hypothesis 2) should produce a higher confidence on the checkpoint C_3 with increased knowledge; however, an estimator could instead capture Hypothesis 1, which performs equally well on the training distribution C_2 but fails to generalize to C_3. **(Right)** To measure the dependence of confidence on knowledge, we evaluate calibration on a contrast set consisting of questions that one checkpoint answers correctly and another checkpoint answers incorrectly, testing whether a change in knowledge is accompanied by a change in confidence.

testing whether the learned features transfer to future, unseen checkpoints in which the model's knowledge has evolved (Fig. 1, left); for inference-time methods (e.g. self-consistency), we directly apply them to the evaluation checkpoints. However, this evaluation is confounded by the fact that a given model's predictions are highly correlated across checkpoints (e.g. common facts are learned first and tend not to be forgotten, while obscure ones are never learned (Mallen et al., 2023)), so strong future-checkpoint calibration can be at least partially explained by a high inter-checkpoint correlation in answer correctness rather than faithfully reflecting evolving knowledge.

With a view to controlling for this confounder, we introduce the notion of the *knowledge contrast set*, which consists of questions where answer correctness differs by checkpoint. This decorrelates the checkpoints and explicitly requires sensitivity to a given checkpoint's knowledge set, since a single confidence value cannot capture why one checkpoint answers correctly while another checkpoint answers incorrectly (Fig. 1, right). Moreover, we argue that contrast sets help disentangle whether a method reflects meta-knowledge or static heuristics. Prior work has suggested that meta-knowledge (knowledge about knowledge, i.e., a model knowing it does/does not know something) may not be necessary to achieve confidence calibration, as it suffices to use heuristics such as question difficulty (Zhu et al., 2026; Xiao et al., 2026). We argue that such heuristics are unsuitable for persistent calibration and would have poor contrast-set calibration, since, for example, a question that is initially difficult may become easy after more training. On the other hand, a method that identifies meta-knowledge features ought to have better contrast-set calibration: if it can identify fact-independent features that represent meta-knowledge of whether a fact is known or unknown, it should be able to adapt when the model's knowledge about the fact changes.

We hypothesize that learning a confidence estimator capable of persistent calibration is challenging because of the vast space of confidence functions that are well-calibrated on a given checkpoint, out of which only some use meta-knowledge in a way that generalizes to other checkpoints; this intuition is illustrated in Fig. 1 (left). Understanding whether a confidence estimator reflects meta-knowledge has practical implications for intervening on a model's knowledge via continual learning. When a model is trained over long time spans and encounters new topics or domains, it can undergo drastic changes in knowledge, making earlier performance no longer representative of current performance (Gururangan et al., 2020; Wang et al., 2025; Periodic Labs, 2026), yet we would still like to know

how trustworthy the current model is. An estimator that is insensitive to the expansion or reduction of a model's capabilities is bound to become obsolete as models continue to grow more capable.

Using our cross-checkpoint knowledge contrast sets on TriviaQA, Jeopardy, and BioASQ, we evaluate confidence estimation methods from various families, including fine-tuning (Kapoor et al., 2024; Li et al., 2026) and inference-time approaches (Kuhn et al., 2023; Manakul et al., 2023). Since fine-tuning methods typically operate on a single snapshot of a model, it is plausible that they learn heuristics that work specifically on the current model and do not robustly generalize to the knowledge sets of future model versions. Inference-time (training-free) methods like self-consistency sidestep the bias of checkpoint-specific feature representations by virtue of treating models as black boxes, making them appealing as a potential solution for persistent calibration. However, it is unclear whether inference-time methods access fine-grained meta-knowledge as opposed to shallow cues (Singh et al., 2026), and forgoing training typically means worse calibration than fine-tuning methods on in-distribution data. To quantify the room for improvement of these methods, we compare them to oracle methods that are allowed to train on the evaluation checkpoints. We find that existing methods obtain limited contrast-set calibration despite having strong full-set calibration. For example, for Marin 8B on Jeopardy, self-consistency matches oracle training on full-set AUC (0.886 vs. 0.890), despite being far behind on contrast-set AUC (0.760 vs. 0.825).

In both oracle and non-oracle settings, we find that contrast-set calibration improves when we introduce multi-checkpoint training. In particular, multi-checkpoint training in the oracle setting outperforms respective-checkpoint training (i.e. training one adapter per checkpoint), revealing that training a single adapter to learn confidence features that work on multiple checkpoints has a synergistic effect in identifying features that are sensitive to differences in knowledge across checkpoints. We further analyze the heterogeneity of confidence estimators that have the same performance on a given checkpoint, showing that such confidence estimators can have vastly different behaviors from each other on other checkpoints. This heterogeneity presents a challenge for identifying robust meta-knowledge features that are conducive to persistent calibration.

2 Persistent Calibration

We introduce persistent calibration as a criterion for confidence estimators: they should faithfully reflect a model's knowledge as it changes, as this is critical for trustworthy, continually learning models. To tractably evaluate persistent calibration, we leverage the checkpoint trajectories of open models trained on trillions of tokens from diverse corpora (Olmo et al., 2025; Marin Community, 2026), offering a large-scale setup to study the evolution of knowledge over time. We evaluate both training-based methods and inference-time methods. For training-based methods, as shown in Fig. 1 (left), we grant them access to earlier checkpoints and evaluate them on the later evaluation checkpoints. For inference-time methods, we directly apply them to the evaluation checkpoints. By default, for both training and evaluation, the checkpoint that generated an answer is the same checkpoint that produces a confidence estimate for the answer. We describe generating answers for our experiments in Section 3.1.1.

Knowledge Contrast Sets. Our research question concerns whether a model's current confidence estimates are faithful to its current knowledge as opposed to the knowledge of another model version. Due to the non-random overlap between different checkpoints' knowledge sets (e.g. common facts are known by most checkpoints (Chang et al., 2024)), confidence estimates that are well-calibrated with respect to one checkpoint's knowledge may also be well-calibrated with respect to another checkpoint's knowledge, so strong calibration is insufficient to show faithfulness. To address this confound, we evaluate calibration on a *knowledge contrast set*, which, for a given pair of checkpoints, consists of questions that one checkpoint answers correctly and the other checkpoint answers incorrectly, as shown in Fig. 1 (right). This construction means that learning to predict confidence based on checkpoint-agnostic question features (e.g. topic or difficulty) may improve full-set calibration but would not improve contrast-set calibration, because the predicted confidence should not change (as the question is the same across checkpoints) and thus the predicted confidence would only be consistent with one checkpoint's correctness but not the other. Faithful confidence estimates should be higher on the checkpoint that predicts the correct answer, similar to how confidence should tend to be higher for correct answers than for incorrect answers within a checkpoint.

Conditioning on Contrast Set Membership. Confidences that are globally calibrated may appear miscalibrated within the contrast subset due to selection bias: the contrast set accuracy (pooled over the two checkpoints) is always 50%, regardless of the global accuracies. To correct for this, given a contrast-set question where checkpoints 1 and 2 have confidences p_1 and p_2 in their respective answers, we normalize them to $p_1(1-p_2)/Z$ and $p_2(1-p_1)/Z$, where $Z = p_1(1-p_2)+p_2(1-p_1)$. This restores contrast-set calibration if p_1, p_2 are calibrated and the correctness outcomes are conditionally independent given p_1, p_2. We discuss this decision in Appendix A.1, and provide empirical support in Appendix B.1 by showing that our normalization function and a fitted normalization function produce similar results. Moreover, the metrics Δ_0, Δ_0^b (defined below) simply compare whether $p_1 > p_2$ (unaffected by normalization), and our results show that their trends track with the other metrics, which depend on normalization.

Evaluation Metrics. For both the full set and the contrast set, we report standard metrics for the discriminability and calibration of confidence estimates: area under the receiver operating characteristic curve (AUC; Hanley & McNeil, 1982), Brier score (BS; Brier, 1950), and Expected Calibration Error (ECE; Naeini et al., 2015), all on a $[0, 1]$ scale. See Appendix A.2 for descriptions. We compute ECE with SmoothECE for its continuous and hyperparameter-free properties (Blasiok & Nakkiran, 2024). For the contrast set $\mathcal{D}_{\text{con}}$, we also report metrics that compare, for a given question q^i, the confidence p_+^i on the correct answer and the confidence p_-^i on the incorrect answer: $\Delta = \frac{1}{|\mathcal{D}_{\text{con}}|}\sum_{i\in\mathcal{D}_{\text{con}}} p_+^i - p_-^i$ and $\Delta_0 = \frac{1}{|\mathcal{D}_{\text{con}}|}\sum_{i\in\mathcal{D}_{\text{con}}} \text{sgn}(p_+^i - p_-^i)$. To control for performance coming from the prior that the later checkpoint tends to be correct, we also report class-balanced variants denoted Δ^b, Δ_0^b, where the two classes $\mathcal{D}_{\text{imp}}, \mathcal{D}_{\text{reg}}$ are improvement (i.e., the later checkpoint is correct) and regression (i.e., the later checkpoint is incorrect): $\Delta^b = \frac{1}{2}\left(\frac{1}{|\mathcal{D}_{\text{imp}}|}\sum_{i\in\mathcal{D}_{\text{imp}}}(p_+^i - p_-^i) + \frac{1}{|\mathcal{D}_{\text{reg}}|}\sum_{i\in\mathcal{D}_{\text{reg}}}(p_+^i - p_-^i)\right)$ and analogously for Δ_0^b. These Δ metrics are on a scale of $[-1, 1]$, e.g. random guessing gets $\Delta_0 = 0$, while always knowing which of the two answers is correct gets $\Delta_0 = 1$. We evaluate on multiple future checkpoints for a robust evaluation. See Appendix A.3 for how we aggregate over them when reporting metrics.

3 Experiments and Results

3.1 Experimental Setup

Models. As we define persistent calibration as a multi-checkpoint property, we use open models that release multiple checkpoints. We evaluate on the pre-training checkpoints of Olmo 3 7B and 32B (Olmo et al., 2025) and Marin 8B (Marin Community, 2026). For Olmo 3 models, we use checkpoints at specified percentages of pre-training progress; we take [10%, 20%, 30%] as non-oracle training checkpoints, and [40%, 50%, 90%, 100%] as evaluation checkpoints; we place a gap in the evaluation checkpoints to induce a substantial accuracy difference for our contrast sets. For Marin 8B, we use all 6 available checkpoints: the first 3 for training and the last 3 for evaluation.

Datasets. We evaluate on TriviaQA (Joshi et al., 2017), Jeopardy (Open Access AI Collective, 2023), and BioASQ (Krithara et al., 2023), which test factual knowledge recall. See Appendix A.4 for details on data processing. TriviaQA and Jeopardy require diverse world knowledge, while BioASQ requires biomedical expertise. For each of TriviaQA and Jeopardy, we use 5K questions for training and 5K questions for evaluation. Due to the smaller size of BioASQ, we only run evaluation to test domain transfer from TriviaQA and Jeopardy. We use 5 fixed in-context examples from the respective dataset for answer generation. We use an LLM judge to label the correctness of a predicted answer with respect to the reference ground-truth answer (Wei et al., 2024); in our manual audit (Appendix A.5), we confirm that LLM and human judgments have high agreement, and that data quality is high for both contrast sets and full sets.

3.1.1 Confidence Estimation Methods

We evaluate representative confidence estimation methods for persistent calibration, using both full datasets and their contrast subsets. As a running example, we suppose a sequence of checkpoints [C_1, C_2, C_3, C_4] as in Fig. 1, where [C_1, C_2] are for non-oracle training and [C_3, C_4] are for evaluation and oracle training. To quantify the room for improvement of a confidence estimator, we compare to a confidence adapter trained on the eval checkpoints [C_3, C_4]. We refer to this as an

oracle method as it allows directly learning confidence features from [C_3, C_4], whereas non-oracle methods must learn meta-knowledge features (from [C_1, C_2], if training) that generalize to [C_3, C_4].

Self-Consistency (SC). SC quantifies a model's confidence in an answer as its probability of generating it, typically estimated via sampling (Kuhn et al., 2023). Pre-trained LLMs often have well-calibrated generation probabilities (Zhu et al., 2023; Luo et al., 2026), since they are trained to model the true data distribution with next-token prediction. While prior work has also explored verbalizing confidence from the model (Tian et al., 2023; Wang & Stengel-Eskin, 2026), this generally requires an instruction-tuned model, and – as confirmed by early experiments – does not perform well for pre-trained checkpoints of the kind we evaluate. Thus, we omit zero-shot verbalized confidence. We add two implementation details that improve SC's robustness: rather than sampling, we run beam search (beam size 10) and weight each beam by its sequence probability to efficiently cover the distribution of short-form generations (Fadeeva et al., 2026), and we use an off-the-shelf NLI model to group equivalent answers (Kuhn et al., 2023, see Appendix A.6). We define a candidate answer as a set of NLI-grouped beams, with generation probabilities summed. We take the top (i.e. highest probability) candidate answer as the model's prediction, and its probability as the SC confidence.

LoRA + Verbalized Confidence. Although pre-trained models struggle to verbalize confidence zero-shot, we can fine-tune them to do so. We use LoRA (Hu et al., 2021) for its parameter efficiency and expressiveness (Kapoor et al., 2024). We prompt the model to verbalize its confidence on a scale of 0 to 9 and compute the confidence as a probability-weighted average over the tokens `0`, ..., `9` (Li et al., 2026). See Appendix A.7 for the training configuration. We introduce multi-checkpoint and multi-answer training to learn a better confidence adapter. **Multi-checkpoint training** trains one adapter to produce calibrated confidence estimates for multiple checkpoints (using their respective predictions, by default). This is motivated by the intuition that a confidence adapter trained on a single checkpoint C_2 may pick up on features that are specific to C_2, while multi-checkpoint training may encourage learning generalizable features. We apply multi-checkpoint training to [C_1, C_2] (non-oracle) or [C_3, C_4] (oracle). We use multi-checkpoint training in Tables 1 and 2 and ablate it in Table 3. **Multi-answer training** uses multiple top candidate answers for a given question. The motivation of this is to learn answer sensitivity, since seeing only C_2's top-1 predictions may lead the adapter to anticipate C_2's answer and ignore the candidate answer, which may differ between C_2 and [C_3, C_4]. We compute a question's loss as the average of its candidate answers' losses (same batch), weighted by their generation probabilities. This ensures that each question's total loss weight is a constant, and that the training distribution is representative of the model's generation distribution. We train with 2 answers by default, and we ablate this in Appendix B.3. We ensure controlled comparisons: for a given training seed, all methods receive the same LoRA initialization and order of questions; for multi-checkpoint training, the checkpoints share the question shuffle and the loss is averaged over checkpoints so the effective learning rate is unaffected. Multi-checkpoint and multi-answer training can be seen as data augmentation of the curated (question, reference answer) data.

End-correct Baseline. To control for the prior that accuracy increases over training, this baseline predicts a fixed confidence based just on checkpoint index. We linearly interpolate $[0, 1]$ with the evaluation checkpoints; for three checkpoints C_3, C_4, C_5, we would get fixed confidences [0, 1/2, 1].

Generalized Correctness Models (GCMs). Xiao et al. (2026) introduce GCMs to predict the correctness of other models' answers. GCMs show similar or better calibration compared to using the other models themselves, challenging the premise that calibration requires meta-knowledge. We revisit this result using our contrast sets, which are designed to require meta-knowledge. We run the TriviaQA-trained GCM released by Xiao et al. (2026), using the model-agnostic prompt containing only the question and the candidate answer.

3.2 Results

Confidence Estimators Struggle on Contrast Sets. Table 1 shows that SC and non-oracle training consistently have worse contrast-set calibration than oracle training, showing room for improvement. We highlight that even when SC matches oracle training on full-set AUC (e.g., SC 0.886 vs. oracle 0.890 on Marin 8B, Jeopardy), SC is often far behind on contrast-set AUC (SC 0.760 vs. oracle 0.825). Thus, full-set calibration does not reliably predict contrast-set calibration, establishing the latter as a distinct criterion for persistent calibration. Moreover, this supports our hypothesis that there are various confidence estimators that have the same aggregate (i.e. full-set) performance and

Table 1: SC and non-oracle training consistently fall short on contrast-set calibration compared to oracle training, even when full-set calibration is saturated. Results in Section 3 are averaged over 3 seeds. In each column, we bold the best result and underline results not significantly worse under a paired test ($\alpha = 0.05$; see Appendix A.8 for tests); tables in Section 3 share this text styling. See Appendix B.2 for all evaluation metrics and Jeopardy results.

	Olmo 3 7B			Marin 8B			Olmo 3 32B		
	Contrast		Full	Contrast		Full	Contrast		Full
Method	$\Delta_0^b \uparrow$	AUC ↑	AUC ↑	$\Delta_0^b \uparrow$	AUC ↑	AUC ↑	$\Delta_0^b \uparrow$	AUC ↑	AUC ↑
					TriviaQA				
End-correct baseline	0.000	0.627	0.530	0.000	0.783	0.547	0.000	0.712	0.548
Self-consistency	0.297	0.733	0.884	0.263	0.823	0.897	0.362	0.808	**0.903**
Non-oracle training	0.333	0.748	0.880	0.227	0.806	0.834	0.285	0.688	0.860
Oracle training	**0.379**	**0.795**	**0.890**	**0.349**	**0.840**	**0.901**	**0.384**	**0.827**	0.898
					Jeopardy				
End-correct baseline	0.000	0.633	0.540	0.000	0.735	0.549	0.000	0.653	0.549
Self-consistency	0.333	0.741	0.873	0.331	0.760	0.886	0.320	0.754	0.893
Non-oracle training	0.476	0.821	0.871	0.397	0.759	0.834	0.399	0.757	0.860
Oracle training	**0.506**	**0.855**	**0.888**	**0.459**	**0.825**	**0.890**	**0.475**	**0.840**	**0.894**

Table 2: Comparing SC and non-oracle training with ablations controlling for knowledge already present in an earlier checkpoint. Beating these ablations is necessary for a nontrivial demonstration of persistent calibration. SC generally beats its ablations, while non-oracle training has mixed results. Results on TriviaQA; see Appendix B.2 for all metrics and Jeopardy results.

	Olmo 3 7B			Marin 8B			Olmo 3 32B		
	Contrast		Full	Contrast		Full	Contrast		Full
Method	$\Delta_0^b \uparrow$	AUC ↑	AUC ↑	$\Delta_0^b \uparrow$	AUC ↑	AUC ↑	$\Delta_0^b \uparrow$	AUC ↑	AUC ↑
Self-consistency	**0.297**	**0.733**	**0.884**	0.263	**0.823**	**0.897**	**0.362**	**0.808**	**0.903**
⌞ Surrogate	0.211	0.637	0.859	**0.277**	0.695	0.871	0.288	0.688	0.883
⌞ Copy	0.000	0.500	0.832	0.000	0.500	0.829	0.000	0.500	0.846
Non-oracle training	**0.333**	**0.748**	**0.880**	0.227	**0.806**	0.834	0.285	0.688	**0.860**
⌞ Surrogate	0.318	0.729	0.869	**0.318**	0.748	**0.866**	**0.290**	**0.716**	0.858
⌞ Copy	0.000	0.500	0.839	0.000	0.500	0.830	0.000	0.500	0.832

yet substantially different behavior on knowledge changes featured in the contrast set (other full-set metrics do not explain this difference either, Appendix B.2), making it hard to learn a confidence estimator that is capable of persistent calibration.

Domain Transfer. In Appendix B.4, we find that the performance gain of oracle training over non-oracle training observed in-domain (Table 1) largely disappears when transferring to BioASQ, as both drop to similarly low performances. This clarifies the scope of the oracle's advantage in our experiments as being limited to the training domain, consistent with prior work on the domain-specificity of correctness signals (Sky et al., 2024).

Ablations to Test Nontrivial Transfer. Although we have seen that SC and non-oracle training have imperfect contrast-set calibration, it appears nontrivial, outperforming the end-correct baseline in almost all cases (Table 1). We examine this further with two ablations. The **surrogate** ablation uses C_2 to estimate confidence on $[C_3, C_4]$'s answers. This ablation isolates the performance that can be achieved with the knowledge already contained in C_2, without needing to inspect the knowledge of $[C_3, C_4]$. The **copy** ablation uses C_2 to estimate confidence on its own answers and simply reuses these confidences for $[C_3, C_4]$ answering the same questions. This ablation is oblivious to $[C_3, C_4]$

Table 3: Multi-checkpoint training improves contrast-set and full-set calibration. Results on TriviaQA; see Appendix B.3 for Jeopardy.

	Olmo 3 7B			Marin 8B			Olmo 3 32B		
	Contrast		Full	Contrast		Full	Contrast		Full
Method	$\Delta_0^b \uparrow$	AUC ↑	AUC ↑	$\Delta_0^b \uparrow$	AUC ↑	AUC ↑	$\Delta_0^b \uparrow$	AUC ↑	AUC ↑
Non-oracle									
Multi-ckpt	**0.333**	0.748	**0.880**	**0.227**	**0.806**	**0.834**	**0.285**	**0.688**	**0.860**
Single-ckpt	0.332	**0.750**	0.876	0.221	0.642	0.783	0.251	0.673	0.849
Oracle									
Multi-ckpt	**0.379**	**0.795**	**0.890**	**0.349**	**0.840**	0.901	**0.384**	**0.827**	**0.898**
Respective-ckpt	0.351	0.778	0.884	0.330	0.837	**0.901**	0.352	0.812	0.892

Table 4: Ablating factors contributing to the gains of multi-checkpoint training (oracle). Single-checkpoint data uses only one checkpoint's answers for multi-checkpoint training. Pooled data trains each checkpoint's adapter on the same data as (unablated) multi-checkpoint training. Multi-checkpoint training beats both, meaning that both the training data and multiple checkpoint views contribute. Results on TriviaQA, Olmo 3 7B; see Appendix B.5 for Jeopardy.

	Contrast							Full		
Method	$\Delta_0^b \uparrow$	$\Delta_0 \uparrow$	$\Delta^b \uparrow$	$\Delta \uparrow$	AUC ↑	BS ↓	ECE ↓	AUC ↑	BS ↓	ECE ↓
Multi-ckpt	**0.379**	**0.424**	**0.255**	**0.285**	**0.795**	**0.185**	**0.061**	**0.890**	**0.131**	0.032
↳ Single-ckpt data	0.373	0.368	0.252	0.258	0.768	0.197	0.096	0.887	0.134	0.044
Resp-ckpt	0.351	0.396	0.241	0.274	0.778	0.193	0.070	0.884	0.133	**0.030**
↳ Pooled data	0.360	0.357	0.248	0.251	0.757	0.203	0.104	0.887	0.132	0.035

and their answers, relying solely on the inter-checkpoint correlation in correctness. A confidence estimator must outperform these two ablations for a convincing demonstration of faithfulness to a given checkpoint's (C_3 or C_4) changed knowledge that cannot be explained by prior knowledge.

Results. Table 2 shows that SC generally outperforms its ablations, while non-oracle training has mixed results. The faithfulness of SC is perhaps expected given that it treats models as black boxes, avoiding checkpoint-specific features. The mixed results for non-oracle training suggest that a confidence adapter trained on a set of checkpoints [C_1, C_2] sometimes but unreliably learns features that generalize beyond to [C_3, C_4]. Lastly, we observe that the copy ablations – with no awareness of [C_3, C_4] or their answers – achieve strong full-set AUC (consistently over 0.8) but a trivial contrast-set AUC of 0.5. This highlights that full-set performance is a poor reflection of the faithfulness of a confidence estimator and is inflated by inter-checkpoint correlation of correctness.

Multi-Checkpoint Training Improves Persistent Calibration. In Table 3, we ablate multi-checkpoint training, which has been used so far. The baseline in the non-oracle setting is **single-checkpoint training** on C_2. The baseline in the oracle setting is **respective-checkpoint training**, which trains one adapter for each of [C_3, C_4] (using the respective checkpoint's data), thus holding constant the total training data; at inference, the respective adapter is used. Table 3 shows that multi-checkpoint consistently matches or exceeds the baseline, in both oracle and non-oracle settings.

Ablation of Multi-Checkpoint Training Advantages. We analyze the potential factors contributing to the gains of multi-checkpoint training: (1) more data and (2) viewing multiple checkpoints' weights. To remove (1), the **single-checkpoint data** ablation of multi-checkpoint training uses C_3's data for both C_3 and C_4. This limits the training data, while still giving one adapter access to multiple checkpoints' weights, encouraging it to learn checkpoint-general features. To remove (2), the **pooled data** ablation performs respective-checkpoint training, but trains each adapter on the data pooled over [C_3, C_4]. This allows each adapter to see the same data as the multi-checkpoint training adapter, while still only allowing each adapter to see the feature representations of its own check-

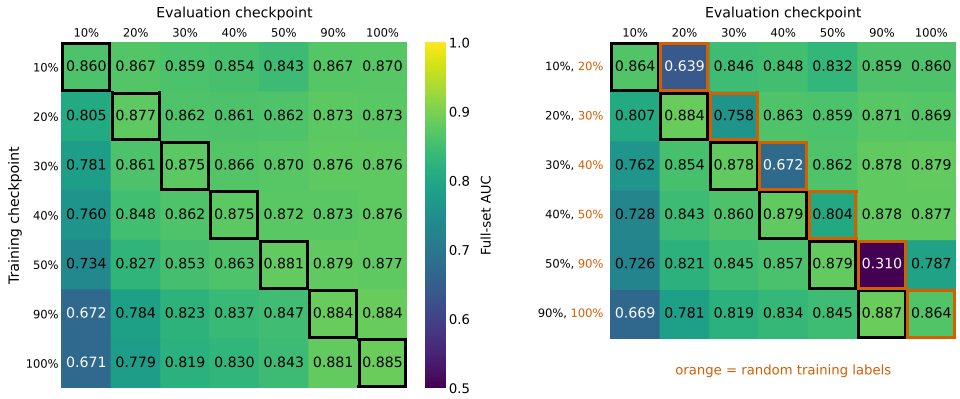

Figure 2: **(Left)** Transfer between pairs of training and eval checkpoints (Olmo 3 7B, TriviaQA; Jeopardy in Appendix B.6). Forward transfer (top right) is strong, while backward transfer (bottom left) is weak. **(Right)** Transfer of multi-checkpoint training, where the second checkpoint gets random labels. The second-checkpoint AUC generally plummets, at no penalty to the first checkpoint.

point (C_3 or C_4). Table 4 shows that multi-checkpoint training outperforms both ablations, meaning that both the training data and the multiple checkpoint views contribute to its performance. Between the two ablations, single-checkpoint data performs better (and more clearly so on Jeopardy, Appendix B.5), suggesting that access to multiple checkpoints' weights is most responsible for the gains of multi-checkpoint training.

GCMs Do Not Satisfy Persistent Calibration. Fig. 3 plots GCM's headroom on contrast-set AUC relative to oracle training. Although GCM matches oracle performance on Olmo 3 7B, with a gap close to 0, we attribute this to the GCM knowing the correct answers to many contrast set questions, enabling it to distinguish correct and incorrect answers without needing access to the internal states of the checkpoints that generated the answers. Indeed, for the larger models (Marin 8B and Olmo 3 32B), the GCM backbone model (Qwen3-8B) has an increased error rate on contrast-set questions, and its gap from the oracle grows substantially. Thus, we argue that while an external verifier can achieve strong calibration on the predictions of a less or similarly capable model, this ability does not transfer to more capable models, which means that a static external verifier cannot satisfy persistent calibration for continually improving models.

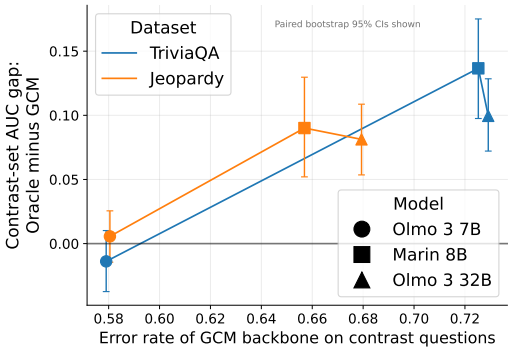

Figure 3: GCM's headroom on contrast-set AUC grows for larger models where the GCM backbone (Qwen3-8B) knows the answer less often. Thus, a static external verifier does not adapt to a model's capability, motivating meta-knowledge for persistent calibration.

4 Analysis

The failure of non-oracle training to faithfully track knowledge (Section 3) may be due to the large hypothesis space of confidence functions that are well-calibrated on a training checkpoint, only a small subset of which depend sufficiently on the meta-knowledge required for transfer. Here, we consider a relaxed notion of transfer on the full set rather than the contrast set, given the positive full-set transfer result in Section 3. We aim to better understand the factors enabling this transfer, which may shed light on the feasibility of persistent calibration and the challenges preventing it.

Transfer Asymmetry. Table 3 found that an adapter trained on C_2 or [C_1, C_2] is usable on C_3 or C_4, which has been trained on far more data and contains very different weights (e.g. the cosine similarity between the flattened weights of Olmo 3 7B 10% and 100% is 0.37). To dissect this phenomenon, we compare forward transfer (i.e. our persistent calibration setup) with backward transfer, where an adapter is trained on C_3 or C_4 and tested on C_1 or C_2. For efficiency in this analysis, we use

single-checkpoint training on Olmo 3 7B and one seed. Fig. 2 (left) reveals that although forward transfer is strong (e.g., AUC 0.870 on 10% $\rightarrow$ 100%), the reverse direction is much worse (0.671 for 100% $\rightarrow$ 10%); this trend generally holds across checkpoints. The confidence adapters trained on 10% and 100% have nearly equal AUCs on 100% (0.870 vs. 0.885) but very different AUCs on 10% (0.860 vs. 0.671) – despite matched performance on one checkpoint, two confidence adapters can implement confidence estimation in functionally different ways, as evidenced by their divergent AUCs on another checkpoint. Fortuitously, this asymmetry is oriented in the forward direction that is relevant to persistent calibration. It may be that this asymmetry stems from the tendency for neural networks to be more plastic earlier and more stable later in training (Achille et al., 2018); this would suggest that features developed at certain points in time tend to remain after that point, but are not present before. An adapter trained on the later checkpoint may rely on those features, which are then absent in the earlier checkpoint, harming its performance, but an adapter trained on the earlier checkpoint can only learn to rely on the features present at that time, which are likely to persist into the later checkpoint. In Appendix B.7, we support this by showing that the variance, computed over confidence adapter training seeds, of hidden states for a given QA instance, tends to be higher on later checkpoints, suggesting that later checkpoints have developed a larger space of confidence features for confidence adapters to choose from.

Random Labels. We dig deeper into the heterogeneity of this feature space by testing if there exist confidence estimators with maintained performance on a given checkpoint but catastrophic performance on another checkpoint: can we train an adapter to perform on C_2 but *not* on C_3? If transfer is inevitable, a drop on C_3 should be accompanied by a drop on C_2, i.e. one cannot optimize for C_2 without inadvertently optimizing for C_3. To test this, we run multi-checkpoint training on [C_2, C_3] but with random labels for C_3 – independent samples from $\mathrm{Bernoulli}(0.5)$. Fig. 2 (right) reveals that C_3 performance generally plummets, but at no penalty to C_2 performance. This shows the extreme heterogeneity of the hypothesis space of confidence features that are viable on a given checkpoint, which may hinder identifying generalizable meta-knowledge features for persistent calibration.

5 Related Work

Confidence Estimation. The main families of LLM confidence estimation methods are based on generation consistency (Kuhn et al., 2023; Manakul et al., 2023), verbalized confidence (Kadavath et al., 2022; Wang & Stengel-Eskin, 2026), fine-tuning (Kapoor et al., 2024; Li et al., 2026), and external verifiers (Zhu et al., 2026; Xiao et al., 2026). While trained methods achieve the best calibration in-domain, they often struggle with domain transfer due to domain-specific signals (Sky et al., 2024; Srey et al., 2026; Ying et al., 2026). We highlight another dimension of brittleness arising from checkpoint-specific features, which multi-checkpoint training partially alleviates.

Continual Learning. As AI systems increasingly act as personalized assistants (Uddin et al., 2026), scientific collaborators (Zhang et al., 2025), and autonomous agents (Luo et al., 2025), they must continually adapt to new developments in human users, the scientific literature, and the world. The families of continual learning methods for LLMs include model editing, retrieval, and continued pre-training (Zheng et al., 2026). We focus on pre-training, as it is the phase in which LLMs acquire the vast majority of their knowledge (Gekhman et al., 2024; Chang et al., 2024). Hasegawa et al. (2025) find that model editing leads to underconfidence in token-probability–based confidence. Liu et al. (2026) use a frozen probe to monitor task performance during pre-training, showing viable forward transfer. Our key distinction is that we center our evaluation around knowledge contrast sets to isolate the dependence between knowledge and confidence. In Appendix C, we discuss contrast sets (Gardner et al., 2020) and metacognition.

6 Conclusion

We motivate *persistent calibration* as a requirement for trustworthy, continually learning models. To see if confidence estimators can track the evolution of knowledge, we evaluate on *knowledge contrast sets*, which highlight changes in knowledge. We find substantial headroom, despite the gains from multi-checkpoint training. Our analysis finds that LLM training dynamics are encouraging for persistent calibration, although the vast space of checkpoint-specific confidence features may hinder identifying robust meta-knowledge features.

AI USE STATEMENT

In this work, we used generative AI tools for implementing methods and providing feedback on research methodology and experiments. We have not used generative AI tools for proposing or refining hypotheses, supporting qualitative and thematic data analysis, and interpreting results. The rest of the required disclosure task (generating synthetic datasets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, assisting with translation, cleaning and reformatting datasets) are not applicable to this work. Additionally, we used generative AI tools for identifying relevant literature, and creating and editing software code. We have reviewed all AI-assisted work. We conducted a manual literature survey in addition to using an LLM to search for related work. LLM-generated code was proofread by one author. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Alessandro Achille, Matteo Rovere, and Stefano Soatto. Critical learning periods in deep networks. In *International conference on learning representations*, 2018.

Tomer Ashuach, Shai Gretz, Yoav Katz, Yonatan Belinkov, and Liat Ein-Dor. Masked by consensus: Disentangling privileged knowledge in llm correctness. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10577–10596, 2026.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Felix Jedidja Binder, James Chua, Tomek Korbak, Henry Sleight, John Hughes, Robert Long, Ethan Perez, Miles Turpin, and Owain Evans. Looking inward: Language models can learn about themselves by introspection. In *International Conference on Learning Representations*, volume 2025, pp. 3710–3756, 2025.

Yonatan Bitton, Gabriel Stanovsky, Roy Schwartz, and Michael Elhadad. Automatic generation of contrast sets from scene graphs: Probing the compositional consistency of gqa. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 94–105, 2021.

Jaroslaw Blasiok and Preetum Nakkiran. Smooth ece: Principled reliability diagrams via kernel smoothing. In *International Conference on Learning Representations*, volume 2024, pp. 1679–1707, 2024.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3, 1950. URL https://api.semanticscholar.org/CorpusID:122906757.

Hoyeon Chang, Jinho Park, Seonghyeon Ye, Sohee Yang, Youngkyung Seo, Du-Seong Chang, and Minjoon Seo. How do large language models acquire factual knowledge during pretraining? *Advances in neural information processing systems*, 37:60626–60668, 2024.

Chi-Keung Chow. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, (4):247–254, 1957.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Ekaterina Fadeeva, Maiya Goloburda, Aleksandr Rubashevskii, Roman Vashurin, Artem Shelmanov, Preslav Nakov, Mrinmaya Sachan, and Maxim Panov. Don't throw away your beams: Improving consistency-based uncertainties in llms via beam search. In *International Conference on Learning Representations*, volume 2026, pp. 139866–139895, 2026.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, 2020.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. Does fine-tuning llms on new knowledge encourage hallucinations? In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 7765–7784, 2024.

Zifan Carl Guo, Laura Ruis, Jacob Andreas, and Belinda Z Li. Introspective coupling: Self-explanation training tracks behavioral change despite fixed supervision. *arXiv preprint arXiv:2606.32038*, 2026.

Todd M Gureckis and Douglas B Markant. Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science*, 7(5):464–481, 2012.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 8342–8360, 2020.

James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

Ryo Hasegawa, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. Knowledge editing induces underconfidence in language models. In *Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (* SEM 2025)*, pp. 338–347, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Kaixuan Huang, Jiacheng Guo, Zihao Li, Xiang Ji, Jiawei Ge, Wenzhe Li, Yingqing Guo, Tianle Cai, Hui Yuan, Runzhe Wang, Yue Wu, Ming Yin, Shange Tang, Yangsibo Huang, Chi Jin, Xinyun Chen, Chiyuan Zhang, and Mengdi Wang. MATH-perturb: Benchmarking LLMs' math reasoning abilities against hard perturbations. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25311–25328. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/huang25k.html.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 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)*, pp. 1601–1611, 2017.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew G Wilson. Large language models must be taught to know what they don't know. *Advances in Neural Information Processing Systems*, 37:85932–85972, 2024.

Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, Kentaro Inui, et al. Realtime qa: What's the answer right now? *Advances in neural information processing systems*, 36:49025–49043, 2023.

Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. Bioasq-qa: A manually curated corpus for biomedical question answering. *Scientific data*, 10 (1):170, 2023.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Chuanrong Li, Lin Shengshuo, Zeyu Liu, Xinyi Wu, Xuhui Zhou, and Shane Steinert-Threlkeld. Linguistically-informed transformations (lit): A method for automatically generating contrast sets. In *Proceedings of the third BlackboxNLP workshop on analyzing and interpreting neural networks for NLP*, pp. 126–135, 2020.

Yibo Li, Miao Xiong, Jiaying Wu, and Bryan Hooi. Conftuner: Training large language models to express their confidence verbally. *Advances in Neural Information Processing Systems*, 38: 53484–53513, 2026.

Zhichen Liu, Tianle Lun, Zhibin Wen, Hao An, Yulin Ou, Jianhui Xu, Hao Zhang, Wenyi Fang, Yang Zheng, and Yang Xu. Fast and accurate probing of in-training llms' downstream performances. *arXiv preprint arXiv:2604.01025*, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Beier Luo, Shuoyuan Wang, Sharon Li, and Hongxin Wei. Your pre-trained llm is secretly an unsupervised confidence calibrator. *Advances in Neural Information Processing Systems*, 38: 163514–163543, 2026.

Junyu Luo, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu, Binqi Chen, Ziyue Qiao, Qingqing Long, et al. Large language model agent: A survey on methodology, applications and challenges. *arXiv preprint arXiv:2503.21460*, 2025.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 9802–9822, 2023.

Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 9004–9017, 2023.

Marin Community. Marin: Open-source framework for the research and development of foundation models. https://github.com/marin-community/marin, 2026.

Rebecca Marvin and Tal Linzen. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 1192–1202, 2018.

Iman Mirzadeh, Keivan Alizadeh-Vahid, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. In *International Conference on Learning Representations*, volume 2025, pp. 94743–94765, 2025.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

Jeremy Nixon, Mike Dusenberry, Ghassen Jerfel, Timothy Nguyen, Jeremiah Liu, Linchuan Zhang, and Dustin Tran. Measuring calibration in deep learning. *arXiv preprint arXiv:1904.01685*, 2019.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, et al. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025.

Open Access AI Collective. Jeopardy. Hugging Face Datasets, 2023. URL https://huggingface.co/datasets/openaccess-ai-collective/jeopardy. Accessed: 2026-09-16.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Periodic Labs. Nature is our learning environment. https://periodic.com/news/nature-is-our-learning-environment, September 2026. Accessed: 2026-09-15.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Mark Bishop Ring. *Continual learning in reinforcement environments*. The University of Texas at Austin, 1994.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 8–14, 2018.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. Foil it! find one mismatch between image and language caption. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 255–265, 2017.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. Continual learning of large language models: A comprehensive survey. *ACM Computing Surveys*, 58(5):1–42, 2025.

Shashwat Singh, Tal Linzen, and Shauli Ravfogel. Can llms introspect? a reality check. *arXiv preprint arXiv:2605.26242*, 2026.

CH-Wang Sky, Benjamin Van Durme, Jason Eisner, and Chris Kedzie. Do androids know they're only dreaming of electric sheep? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4401–4420, 2024.

Ponhvoan Srey, Xiaobao Wu, Cong-Duy Nguyen, Quang Minh Nguyen, Duc Anh Vu, and Anh Tuan Luu. From signals to transfer: A factorised study of probe-based uncertainty estimation in large language models. *arXiv preprint arXiv:2606.27679*, 2026.

Mark Steyvers and Megan AK Peters. Metacognition and uncertainty communication in humans and large language models. *Current Directions in Psychological Science*, 35(3):131–139, 2026.

Mark Steyvers, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas W Mayer, and Padhraic Smyth. What large language models know and what people think they know. *Nature Machine Intelligence*, 7(2):221–231, 2025.

Hao Sun, Yunyi Shen, and Jean-Francois Ton. Rethinking bradley-terry models in preference-based reward modeling: Foundations, theory, and alternatives. *arXiv preprint arXiv:2411.04991*, 2024.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D 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*, pp. 5433–5442, 2023.

Md Nayem Uddin, Kumar Shubham, Eduardo Blanco, Chitta Baral, and Gengyu Wang. From recall to forgetting: Benchmarking long-term memory for personalized agents. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 26814–26841, 2026.

Luping Wang, Sheng Chen, Linnan Jiang, Shu Pan, Runze Cai, Sen Yang, and Fei Yang. Parameter-efficient fine-tuning in large language models: a survey of methodologies. *Artificial Intelligence Review*, 58(8):227, 2025.

Victor Wang and Elias Stengel-Eskin. Calibrating verbalized confidence with self-generated distractors. In *International Conference on Learning Representations*, volume 2026, pp. 121268–121297, 2026.

Xiao Wang, Yuansen Zhang, Tianze Chen, Songyang Gao, Senjie Jin, Xianjun Yang, Zhiheng Xi, Rui Zheng, Yicheng Zou, Tao Gui, et al. Trace: A comprehensive benchmark for continual learning in large language models. *arXiv preprint arXiv:2310.06762*, 2023.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*, 2024.

Andrea H Wynn, Ilia Sucholutsky, and Thomas L Griffiths. Learning human-like representations to enable learning human values. *Advances in Neural Information Processing Systems*, 37:30230–30260, 2024.

Hanqi Xiao, Vaidehi Patil, Hyunji Lee, Elias Stengel-Eskin, and Mohit Bansal. Generalized correctness models: Learning calibrated and cross-model correctness predictors from historical patterns. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=g9G7qyAzki.

Zhuofan Josh Ying, Shauli Ravfogel, Nikolaus Kriegeskorte, and Peter Hase. The truthfulness spectrum hypothesis. *arXiv preprint arXiv:2602.20273*, 2026.

Gal Yona, Mor Geva, and Yossi Matias. Position: Hallucinations undermine trust; metacognition is a way forward. In *Forty-third International Conference on Machine Learning Position Paper Track*, 2026. URL https://openreview.net/forum?id=I9CIvYaRmB.

Yanbo Zhang, Sumeer A Khan, Adnan Mahmud, Huck Yang, Alexander Lavin, Michael Levin, Jeremy Frey, Jared Dunnmon, James Evans, Alan Bundy, et al. Exploring the role of large language models in the scientific method: from hypothesis to discovery. *npj Artificial Intelligence*, 1 (1):14, 2025.

Junhao Zheng, Chengming Shi, Xidi Cai, Qiuke Li, Duzhen Zhang, Chenxing Li, Dong Yu, and Qianli Ma. Lifelong learning of large language model based agents: A roadmap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Kaitlyn Zhou, Jena Hwang, Xiang Ren, and Maarten Sap. 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)*, pp. 3623–3643, 2024.

Chiwei Zhu, Benfeng Xu, Quan Wang, Yongdong Zhang, and Zhendong Mao. On the calibration of large language models and alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9778–9795, 2023.

Tracy Yixin Zhu, Yukai Yang, Marco Morucci, and Tim G. J. Rudner. A systematic assessment of weak-to-strong confidence prediction in large language models. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=xYSzkg5qPD.

A Experimental Setup

A.1 Conditioning on Contrast Set Membership

Since a contrast set derived from a full dataset is not a random subset but rather based on correctness labels, confidence estimates that are calibrated with respect to the full data distribution may not be calibrated with respect to the distribution that conditions on contrast set membership. For example, suppose that on each question in a dataset, two checkpoints each have an independent 80% chance of predicting the correct answer. The ideal confidence estimates are thus 80% on every predicted answer, and the expected accuracy is 80%. The contrast set consists of questions for which either checkpoint 1 is correct and checkpoint 2 is incorrect ($0.8 \cdot 0.2 = 0.16$ of all instances), or checkpoint 1 is incorrect and checkpoint 2 is correct ($0.2 \cdot 0.8 = 0.16$ of all instances), so each checkpoint has an expected accuracy of 50% on contrast set questions. Thus, the ideal global confidence estimates of 80% appear miscalibrated when viewed within the contrast set.

To account for the selection bias of the contrast set, it is necessary to normalize the confidence estimates for the two answers to a given question in the contrast set, as illustrated in Fig. 1 (right). This normalization should restore contrast-set calibration for confidence estimates that are calibrated with respect to the full dataset; accordingly, we treat confidence estimates as if they were calibrated. In the example above, every (0.8, 0.8) confidence pair should be normalized to (0.5, 0.5) because each answer has a 50% chance of being correct given that exactly one of the two answers is correct. Let checkpoint k's answer have binary correctness z_k and calibrated confidence $p_k = P(z_k = 1)$. We would like to find

$$P(z_1 = 1 \mid \underbrace{(z_1, z_2) \in \{(1,0),(0,1)\}}_{\text{contrast set membership}}) = \frac{P((z_1, z_2) = (1,0))}{P((z_1, z_2) \in \{(1,0),(0,1)\})}.$$

If we assume the checkpoints' predictions are independent as in the example above, this becomes

$$\frac{P((z_1, z_2) = (1,0))}{P((z_1, z_2) \in \{(1,0),(0,1)\})} = \frac{p_1(1-p_2)}{p_1(1-p_2) + (1-p_1)p_2}. \tag{1}$$

In general, two checkpoints' predictions are not independent because the later checkpoint originates from the earlier checkpoint, which means that predictions carry over by default. For example, if checkpoint 1 has 80% confidence in its answer to a question, and checkpoint 2 is obtained by further training on a negligible amount of unrelated data, it will have a nearly unchanged 80% confidence on the same answer. In this case, we should estimate the confidence that both answers are correct as 0.8 rather than the $0.8^2 = 0.64$ that independence would prescribe. Nonetheless, this "carry-over" dependence is not present when the two checkpoints produce *different* predictions, which is the case for contrast set questions by construction. Thus, we use the normalization given by Eq. 1, making the assumption that the relative probabilities of the contrast outcomes $(z_1, z_2) = (1,0), (0,1)$ are the same as if z_1 and z_2 were independent. We provide empirical support for this decision in Appendix A.1 by showing that Eq. 1 and a fitted normalization function produce similar results. Moreover, the metrics Δ_0, Δ_0^b (Section 2) simply compare whether $p_1 > p_2$ (unaffected by normalization), and our results show that they exhibit the same trends as the other metrics, which do depend on normalization.

A.2 Evaluation Metrics

Discriminability refers to how well confidence estimates separate correct answers from incorrect ones, which is useful for selective prediction, i.e. abstaining on uncertain questions while maintaining high recall of questions answered (Chow, 1957). Calibration refers to whether, for all p, the accuracy over answers with confidence p is p. AUC measures discriminability, as it can be characterized as the proportion of (correct, incorrect) pairs in which the correct answer is assigned a higher confidence, with ties given half credit, so trivial performance is 0.5 (Hanley & McNeil, 1982). BS is the mean squared error between confidence estimates and the binary correctnesses (Brier, 1950). Since BS is a proper scoring rule, i.e. its expected value is minimized when confidence estimates match the ground truth probabilities of correctness, it measures calibration as well as discriminability (since it rewards a confidence estimate for being close to the correctness). ECE measures calibration (Naeini et al., 2015): $\mathbb{E}_{(x,y,\hat{y})\sim\mathcal{D},\ p=f(x,\hat{y})}|P(\hat{y} = y \mid p) - p|$, where $\mathcal{D}$ is a data distribution

of (question x, reference answer y, predicted answer $\hat{y}$) instances and f is a confidence estimator. To empirically estimate ECE, it is necessary to bin or smooth the data; we use SmoothECE for its continuous and hyperparameter-free properties (Blasiok & Nakkiran, 2024).

A.3 Aggregation of Metrics over Checkpoints

Since we evaluate on multiple future checkpoints for a robust evaluation, we must specify how we aggregate over them when reporting calibration metrics. For full-set evaluation, we compute AUC and BS on the pooled data from all checkpoints, while we average the ECE computed on each checkpoint's data. We do not pool data for ECE to avoid cancellation effects that hide miscalibration (Nixon et al., 2019); for example, if checkpoints 1 and 2 have 50% and 70% accuracy, respectively, a confidence estimator that outputs a constant 60% would be perfectly calibrated on the pooled data (ECE = 0), but evaluating on each checkpoint separately would expose the miscalibration (ECE = 0.1).[1] For contrast-set evaluation, we average a given metric computed on the contrast set of each pair of evaluation checkpoints. We do not pool data over different checkpoint pairs' contrast sets because the normalization (Appendix A.1) assumes that a given answer *pair* contains exactly one correct answer. For a given contrast set, we compute AUC and BS on the (normalized) confidences from the two checkpoints, while we compute ECE on only one checkpoint's confidences to avoid cancellation effects.[2] To require sufficient sample size, we only use a contrast set if it contains at least some minimum number of questions. We set this threshold to 500 for TriviaQA and Jeopardy, and 300 for BioASQ due to its smaller size; see Appendix A.10 for data statistics.

A.4 Data Processing.

We use the datasets TriviaQA (Joshi et al., 2017), Jeopardy (Open Access AI Collective, 2023), and BioASQ (Krithara et al., 2023). Here, we describe how we process the data before use.

TriviaQA and Jeopardy. For each dataset, we shuffle the questions and remove duplicate questions. We use the validation split of the `rc.nocontext` subset of TriviaQA. The Jeopardy dataset contains many topics, not all of which are suitable for factual recall. We exclude questions containing HTML media links. We manually inspect the topics with at least 150 questions and check for the suitability of each topic. We exclude the following topics: Before & After, Rhyme Time, Stupid Answers, Common Bonds, Homophones. See Table 5 for the list of included topics. We prepend the question topic to the question. After deduplication, for each of TriviaQA and Jeopardy, we reserve 10 questions for potential ICL use (and exclude from training and evaluation data), although we use 5 ICL examples in practice. TriviaQA is left with 9961 questions, which we split into a train set of 5000 questions and a test set of 4961 questions. Jeopardy has a train set of 5000 questions and a test set of 5000 questions. In a manual audit, we identified several low-quality questions, which we blacklisted. After blacklisting, TriviaQA's train and test splits have sizes 4999 and 4960, respectively, and Jeopardy's train and test splits have sizes 4997 and 4998, respectively.

BioASQ. We use the BioASQ 14b training data, retaining factoid and list questions only, as these are suitable for factual recall evaluation. Factoid questions have a short-form answer, and list questions request a list of multiple answers; see Table 6 for examples. We inspected reference answers for several list questions and found that they often do not produce an exhaustive list, so we only require models to provide any correct answer, by appending *"Just output one of the answers."* to the question. After shuffling and deduplicating questions, we discard list questions with more than 10 answer groups in order to not overwhelm the LLM judge, which is tasked with checking if the candidate answer matches any of the reference answers. For each question type (factoid or list), we reserve 5 questions as ICL examples. For the list-type ICL examples, we set the example answer to the first answer in the reference list, as we find that this is generally the most noteworthy answer.

[1] While, in general, we cannot eliminate other sources of cancellation that occur even within a single checkpoint, we consider it prudent to eliminate the addressable case of cross-checkpoint cancellation.

[2] For example, a constant confidence estimator would yield [0.5, 0.5] normalized confidences on every contrast-set question. Since contrast-set accuracy (pooled over the two checkpoints) is always 50%, we would get ECE = 0, even if the two checkpoints' contrast-set accuracies are 40% and 60%, whereas single-checkpoint ECE would expose ECE = 0.1. Note that single-checkpoint ECE is invariant to which of the two checkpoints is used, since for a given question, both the confidences and the correctnesses sum to 1.

Table 5: Jeopardy categories included in our dataset.

19th Century America	3-Letter Words	4-Letter Words
American History	American Literature	Americana
Animals	Annual Events	Architecture
Around The World	Art	Art & Artists
Astronomy	Authors	Awards
Ballet	Biology	Bodies Of Water
Books & Authors	Business & Industry	Classical Music
Colleges & Universities	Composers	European History
Explorers	Famous Americans	Fashion
Fictional Characters	First Ladies	Food
Food & Drink	Fruits & Vegetables	Geography
Historic Names	History	Hodgepodge
Holidays & Observances	In The Dictionary	Islands
Languages	Literature	Magazines
Medicine	Mountains	Museums
Music	Musical Instruments	Mythology
Nature	Nonfiction	Opera
Organizations	People	Poets & Poetry
Pop Music	Potent Potables	Potpourri
Quotations	Religion	Science
Science & Nature	Scientists	Shakespeare
Sports	State Capitals	Television
The Bible	The Body Human	The Civil War
The Movies	Theatre	Transportation
Travel & Tourism	U.S. Cities	U.S. Geography
U.S. History	U.S. Presidents	Vocabulary
Weights & Measures	Word Origins	World Capitals
World Cities	World Geography	World History
Zoology		

Table 6: BioASQ example questions. For list questions, an answer matching any of the reference answers is accepted.

Type	Question	Reference answer(s)
Factoid	What is the main use of ETD fragmentation?	Analysis of intact proteins
Factoid	Which disease can be treated with Delamanid?	tuberculosis
List	List approved radioprotective compounds. Just output one of the answers.	[[‘amifostine’], [‘palifermin’]]
List	List the 5 different human immunoglobulin heavy chains. Just output one of the answers.	[[‘alpha’], [‘delta’], [‘epsilon’], [‘gamma’], [‘mu’]]

For evaluation, we are left with 1690 factoid questions and 1029 list questions, for a total of 2719 questions.

A.5 Audit of Data Quality

We use GPT 5.4 as an LLM judge to score the binary correctnesses of candidate answers for a given question and reference answer. To verify the quality of the LLM judgments as well as to check for idiosyncracies in the contrast sets, one author of this paper conducted a manual audit of a sample of full-set and contrast-set data from TriviaQA. We sampled 50 questions from the contrast set and 50 questions from the non-contrast set (25 correct and 25 incorrect). For each question, we checked whether whether the reference answer was correct. If the reference answer was correct, we annotated whether it was “incomplete”, i.e. other answers should also be acceptable even though they were not or would not be judged correct. The auditor agreed with the LLM judgment in every case excluding incomplete cases (which are due to ambiguities with the underlying dataset), showing high reliability of LLM judgments. The contrast set and non-contrast set had 50/50 and 47/50 correct reference

answers, respectively. The contrast set and non-contrast had 3/50 and 2/50 incomplete reference answers, respectively. This confirms that the contrast set has similarly high quality data to the full set. In fact, this audit suggests that the contrast set has even higher quality than the full set, which may be explained by the fact that at least one checkpoint answering a question correctly means that it is likely that the question is well-posed.

A.6 Self-Consistency Implementation

We implement SC with the following design decisions to improve robustness. Rather than sampling, we run beam search (beam size 10) and weight each beam by its sequence probability to efficiently cover the distribution of short-form generations (Fadeeva et al., 2026). Beams are terminated at the first new line. We use an off-the-shelf NLI model to group equivalent answers (Kuhn et al., 2023).[3] Specifically, two answers are considered and equivalent and grouped together if each directed prediction has an argmax prediction of entailment (out of entailment, contradiction, and neutrality), following (Kuhn et al., 2023). We define a candidate answer as a set of NLI-grouped beams, with generation probabilities summed. We take the top (i.e. highest probability) candidate answer as the model's prediction, and its probability as the SC confidence.

For SC surrogate (Section 3.2), since C_3 or C_4's predicted answer (i.e. top candidate answer) may not match C_2's predicted answer, we implement SC surrogate by using an NLI model to match C_3 or C_4's predicted answer against C_2's candidate answers, computing the confidence as the sum of the probabilities of the matching answers, or 0 if there is no match.

A.7 LoRA Training Configuration

We train LoRA adapters (Hu et al., 2021) on all linear layers with $r = 8$, $\alpha = 16$, no dropout, no bias term, learning rate $2 \cdot 10^{-4}$, batch size of 16 questions, and one epoch. We schedule the learning rate to be stable at $2 \cdot 10^{-4}$ for the first 80% of training and cosine-decay over the last 20% to $2 \cdot 10^{-5}$. We use the AdamW optimizer (Loshchilov & Hutter, 2017) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, weight decay 0.01.

A.8 Significance Testing

We use a paired bootstrap test to check the statistical significance of the best reported result (bolded) outperforming another result (not bolded). We first give background on how we compute the reported value of a metric, which we then build on to define our bootstrapping procedure.

A.8.1 Background on the Reported Metric

Suppose a metric m, confidence estimation method f, and base dataset $\tilde{\mathcal{D}}$. The base dataset $\tilde{\mathcal{D}}$ consists of (question, reference answer) instances, but without candidate answers and correctness labels. For a list of checkpoints K, each checkpoint $k \in K$ has an instantiation $\mathcal{D}_k$ of $\tilde{\mathcal{D}}$ with its own candidate answers and correctness labels. Let P be the set of unordered pairs of distinct checkpoints $p = (k, k') \in K \times K$ that have a sufficiently large contrast set (see Appendix A.3). For $p = (k, k') \in P$, we denote $\mathcal{D}_p$ as the concatenation of its two checkpoints' data $\mathcal{D}_k$ and $\mathcal{D}_{k'}$ (restricted to questions where their correctnesses differ, and with confidences normalized, as shown in Fig. 1, right). We define $\mathcal{D} = \{\mathcal{D}_k\}_{k \in K}$ if m is a full-set metric, and $\mathcal{D} = \{\mathcal{D}_p\}_{p \in P}$ if m is a contrast-set metric. We note that the data in $\mathcal{D}$ still carry information about the checkpoint identities k, as this is used in some metrics (Appendix A.3). To support stochastic confidence estimation methods, we suppose a set of random seeds S and denote f_s as the confidence estimator resulting from training or running f with seed $s \in S$; deterministic confidence estimators simply ignore s. The reported metric value is the metric value averaged over seeds:

$$m(f; \mathcal{D}) = \frac{1}{|S|} \sum_{s \in S} m(f_s; \mathcal{D})$$

[3] https://huggingface.co/MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli

A.8.2 Paired Bootstrap Test

We describe how we construct a bootstrap replicate $\mathcal{D}^b$ of $\mathcal{D}$. We then use this procedure to construct $B = 10^4$ independent replicates $\{\mathcal{D}^b\}_{b\in[B]}$, giving us an empirical bootstrap distribution $\{m(f_{\text{best}}; \mathcal{D}^b) - m(f; \mathcal{D}^b)\}_{b\in[B]}$ (with the sign flipped for metrics where lower is better). We underline the result if the one-sided 95% confidence interval contains 0.

We construct a given replicate in a way that accounts for data dependence and class stratification. We have to account for data dependence because different checkpoints and training seeds are evaluated using the same underlying questions $\tilde{\mathcal{D}}$, so treating the data as independent would inflate statistical significance. We use class stratification so that each bootstrap replicate preserves certain statistics of the original data $\mathcal{D}$. For full-set metrics, we use the classes [correct, incorrect], considered per checkpoint. In other words, for each checkpoint, the number of correct instances and the number of incorrect instances should match between the replicate $\mathcal{D}^b$ and the original data $\mathcal{D}$. As a corollary, the total number of instances for each checkpoint would match between $\mathcal{D}^b$ and $\mathcal{D}$. For contrast-set metrics, we use the classes [improvement, regression] (i.e., the earlier checkpoint is incorrect and the later checkpoint is correct, or the earlier checkpoint is correct and the later checkpoint is incorrect), considered per checkpoint pair. In other words, for each checkpoint pair, the number of improvement instances and the number of regression instances should match between the replicate $\mathcal{D}^b$ and the original data $\mathcal{D}$. As a corollary, the total number of contrast-set instances for each checkpoint pair would match between $\mathcal{D}^b$ and $\mathcal{D}$. Thus, a replicate in the full-set case has $2 \cdot |K|$ constraints (K is the set of checkpoints), and a replicate in the contrast-set case has $2 \cdot |P|$ constraints (P is the set of checkpoint pairs).

For ease of presentation, we assume the full-set case of $\mathcal{D} = \{\mathcal{D}_k\}_{k\in K}$ when describing our procedure for constructing a replicate; the contrast-set case is analogous by replacing K with P. We sample an infinite sequence of questions $Q^b = \{q_i^b\}_{i=1}^{\infty}$, sampled independently and with replacement from the full set of questions $\tilde{\mathcal{D}}$. We construct a replicate $\mathcal{D}_k^b$ for each $\mathcal{D}_k$ and then concatenate the results to get $\mathcal{D}^b = \{\mathcal{D}_k^b\}_{k\in K}$. We model dependence on a shared data pool by deriving $\mathcal{D}_k^b$ for different $k \in K$ from the same sequence Q^b. For a given checkpoint k, each question q_i^b belongs to class 1 (correct), class 2 (incorrect), or neither (only applicable to the contrast-set case). To construct $\mathcal{D}_k^b$, we begin with an empty (multi)set and iterate through the sequence Q^b, adding a question if it belongs to class 1 or 2 and the class stratification quota for its class has not been met. We stop once both quotas have been met.

A.9 Model Checkpoints.

For the checkpoints used in our experiments, Table 7 reports the checkpoint identities, training data sizes, and dataset accuracies.

A.10 Contrast Set Statistics

Tables 8 to 10 report contrast set size statistics for TriviaQA, Jeopardy, and BioASQ.

B Results and Analysis

B.1 Empirical Support for Conditioning on Contrast Set Membership

To provide empirical support for our method of conditioning confidence estimates on contrast set membership (Eq. 1), we show that a fitted normalization function and Eq. 1 produce similar results.

Suppose checkpoints 1 and 2 produce confidences p_1 and p_2 in their respective answers to a question that exactly one of them answers correctly. We use the Bradley-Terry (BT) model (Bradley & Terry, 1952) to parameterize the space of functions that map (p_1, p_2) to the probability that checkpoint 1 is correct. The BT model can express a broad class of pairwise preferences functions (Sun et al., 2024) and has been widely used for preference optimization objectives (Christiano et al., 2017; Ouyang et al., 2022; Rafailov et al., 2023).

Table 7: Cumulative training data and performance for the checkpoints used in our experiments.

Short name	Checkpoint ID	Training data (trillions of tokens)	Accuracy (%) TriviaQA	Jeopardy	BioASQ
		Olmo 3 7B			
10%	`stage1-step141000`	0.59	51.9	64.8	40.2
20%	`stage1-step283000`	1.19	55.5	68.5	41.7
30%	`stage1-step424000`	1.78	57.8	70.7	44.0
40%	`stage1-step566000`	2.37	59.5	72.1	44.5
50%	`stage1-step707000`	2.97	60.8	73.6	47.7
90%	`stage1-step1272000`	5.34	64.8	77.8	49.4
100%	`stage1-step1413814`	5.93	65.6	78.6	51.1
		Marin 8B			
	`kestrel`	2.7	65.9	78.9	37.8
	`ocelot`	3.78	67.1	79.5	38.9
	`jellyfish`	4.78	67.0	81.0	50.5
	`phoenix`	11.1	68.4	80.4	42.5
	`starling`	12.4	76.1	85.9	54.1
	`deeper-starling`	12.7	76.5	86.3	54.1
		Olmo 3 32B			
10%	`stage1-step66000`	0.55	62.7	75.4	46.9
20%	`stage1-step131000`	1.10	67.3	80.0	50.3
30%	`stage1-step197000`	1.65	68.5	81.8	52.2
40%	`stage1-step262000`	2.20	69.7	83.7	53.8
50%	`stage1-step328000`	2.75	71.2	84.2	53.5
90%	`stage1-step590120`	4.95	76.8	87.6	59.5
100%	`stage1-step656000`	5.50	77.7	88.8	61.5

For a given confidence estimator and a given contrast set, we fit a model that predicts $r(p_1, p_2)$, the probability that checkpoint 1 is correct given a contrast-set question. Importantly, we fit the model using the evaluation data (confidences and correctnesses labels), so the downstream results should be considered generous. The BT model parameterizes r with a utility function $w(p) > 0$, where we can ensure positivity with a reparameterized $h(p) = \log w(p)$:

$$r(p_1, p_2) = \frac{w(p_1)}{w(p_1) + w(p_2)} = \sigma(h(p_1) - h(p_2))$$

For example, Eq. 1 uses $h(p) = \text{logit}(p) = \log(\frac{p}{1-p})$. Since h is invariant to additive constants, we can set $h(0.5) = 0$. Furthermore, we constrain h to be antisymmetric about 0.5, i.e. $h(1-p) = -h(p)$, since negating both claims (i.e. the checkpoints' claims that their answers are correct) should give $r(1-p_1, 1-p_2) = 1 - r(p_1, p_2)$. Thus, we fit $h(p)$ for $p \geq 0.5$, and define $h(p)$ for $p < 0.5$ by $h(p) = -h(1-p)$. We fit a spline with knots at $0.5 = c_0 < c_1 < \ldots < c_K < 1$ using $K = 30$ knots. Each knot c_k is chosen as the $k/(K+1)$ quantile of the empirical distribution of $\max(p, 1-p)$ over the $2N$ confidence estimates produced in the given contrast set by the given confidence estimator, where N is the number of questions in the contrast set. We linearly interpolate between the knots, and define the slope between c_K and 1 to equal that between c_{K-1} and c_K. We fit $h(c_0), \ldots, h(c_K)$ to maximize the log likelihood of r using LBFGS with 50 maximum iterations.

Table 16 shows the results using the fitted normalization functions. We find the same trends from Table 11: confidence estimation methods underperform oracle training, showing that our findings are robust to the specific implementation of normalization.

B.2 Expanded Metrics on Existing Methods Versus Oracle Training

Table 11 is an expanded version of Table 1 showing the same trends of headroom between existing confidence estimation methods and oracle training.

Table 8: TriviaQA contrast-set sizes for each pair of evaluation checkpoints. The full-set size is 4960. In a cell, the first number is the total size of the contrast set, and the second number is the number of questions that the row checkpoint gets incorrect and the column checkpoint gets correct. Contrast-set sizes that are grayed out are too small (fewer than 500 questions) and are not used (Appendix A.3).

Olmo 3 7B				
	40%	50%	90%	100%
40%	–	642 (354)	771 (516)	753 (527)
50%	642 (288)	–	715 (455)	697 (466)
90%	771 (255)	715 (260)	–	538 (289)
100%	753 (226)	697 (231)	538 (249)	–

Marin 8B			
	phoenix	starling	deeper-starling
phoenix	–	698 (540)	691 (547)
starling	698 (158)	–	147 (84)
deeper-starling	691 (144)	147 (63)	–

Olmo 3 32B				
	40%	50%	90%	100%
40%	–	654 (364)	676 (514)	707 (551)
50%	654 (290)	–	660 (469)	639 (480)
90%	676 (162)	660 (191)	–	467 (255)
100%	707 (156)	639 (159)	467 (212)	–

B.3 Ablation of Multi-Checkpoint and Multi-Answer Training

Tables 13 and 14 show that multi-checkpoint and multi-answer training often lead to substantial improvements in calibration.

B.4 Domain Transfer

Tables 17 and 18 show that oracle training and other methods drop to similarly low performances when transferring from TriviaQA or Jeopardy to BioASQ (biomedical domain). This clarifies the scope of the oracle's advantage in our experiments as being limited to the training domain, consistent with prior work on the domain-specificity of correctness signals (Sky et al., 2024; Srey et al., 2026; Ying et al., 2026).

B.5 Multi-Checkpoint Training Ablations

Table 15 echoes Table 4 in showing that multi-checkpoint training outperforms its ablations, showing that its performance gain comes from both more data and access to multiple checkpoints' weights. Furthermore, in Table 15, the single-checkpoint data ablation matches the unablated multi-checkpoint training, indicating that most of the performance gain comes from viewing multiple checkpoints' weights, encouraging learning confidence features that generalize across checkpoints.

B.6 Transfer Asymmetry Analysis

Fig. 4 shows the same trends as Fig. 2: Forward transfer is far more robust than backward transfer, and we can achieve catastrophic performance on a later checkpoint at no penalty to an earlier checkpoint.

Table 9: Jeopardy contrast-set sizes for each pair of evaluation checkpoints. The full-set size is 4998. In a cell, the first number is the total size of the contrast set, and the second number is the number of questions that the row checkpoint gets incorrect and the column checkpoint gets correct. Contrast-set sizes that are grayed out are too small (fewer than 500 questions) and are not used (Appendix A.3).

Olmo 3 7B				
	40%	50%	90%	100%
40%	–	738 (406)	767 (526)	796 (559)
50%	738 (332)	–	695 (453)	720 (484)
90%	767 (241)	695 (242)	–	517 (277)
100%	796 (237)	720 (236)	517 (240)	–

Marin 8B			
	phoenix	starling	deeper-starling
phoenix	–	604 (440)	613 (454)
starling	604 (164)	–	153 (86)
deeper-starling	613 (159)	153 (67)	–

Olmo 3 32B				
	40%	50%	90%	100%
40%	–	587 (307)	613 (405)	584 (421)
50%	587 (280)	–	546 (358)	563 (397)
90%	613 (208)	546 (188)	–	405 (233)
100%	584 (163)	563 (166)	405 (172)	–

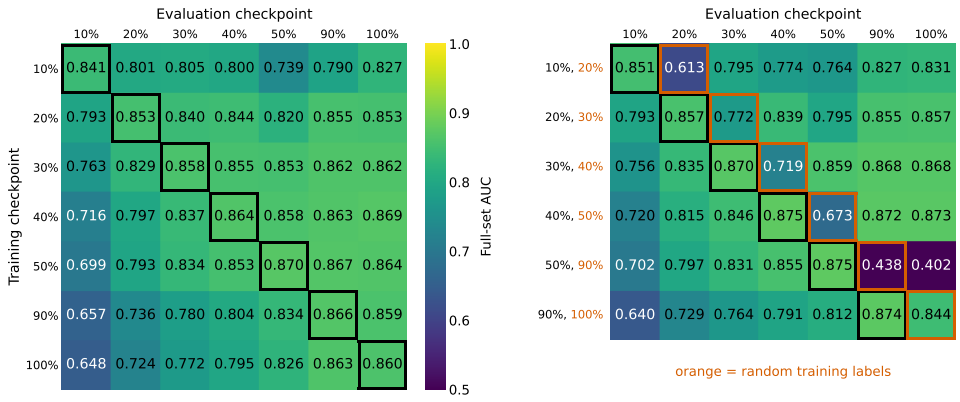

Figure 4: **(Left)** Transfer between pairs of training and eval checkpoints (Olmo 3 7B, Jeopardy, see Fig. 2 for TriviaQA). Forward transfer (top right) is strong, while backward transfer (bottom left) is weak. **(Right)** Transfer of multi-checkpoint training, where the second checkpoint gets random labels. The second-checkpoint AUC generally plummets, at no penalty to the first checkpoint.

B.7 Cross-Seed Variance Over Checkpoints

Spearman correlations between checkpoint index and across-adapter hidden-state variance for Olmo 3 7B on TriviaQA. To support our hypothesis that later checkpoints tend to have more confidence features allowing for more heterogeneous confidence adapters, we examine the variance in hidden states for a given question when we vary the confidence adapter. For each checkpoint, we train 10 confidence adapters with different seeds. For each question and layer, we compute the total variance (i.e. trace of the covariance matrix), across adapters, of the hidden states, and we compute

Table 10: BioASQ contrast-set sizes for each pair of evaluation checkpoints. The full-set size is 2719. In a cell, the first number is the total size of the contrast set, and the second number is the number of questions that the row checkpoint gets incorrect and the column checkpoint gets correct. Contrast-set sizes that are grayed out are too small (fewer than 300 questions) and are not used (Appendix A.3).

Olmo 3 7B				
	40%	50%	90%	100%
40%	–	457 (273)	490 (312)	475 (328)
50%	457 (184)	–	483 (264)	466 (279)
90%	490 (178)	483 (219)	–	389 (218)
100%	475 (147)	466 (187)	389 (171)	–

Marin 8B			
	phoenix	starling	deeper-starling
phoenix	–	564 (440)	582 (449)
starling	564 (124)	–	172 (86)
deeper-starling	582 (133)	172 (86)	–

Olmo 3 32B				
	40%	50%	90%	100%
40%	–	413 (202)	486 (320)	492 (351)
50%	413 (211)	–	451 (307)	459 (339)
90%	486 (166)	451 (144)	–	356 (206)
100%	492 (141)	459 (120)	356 (150)	–

the correlation between checkpoint index and this total variance. Table 19 shows that the correlation is consistently positive for all layers, suggesting that later checkpoints tend to have a larger space of viable confidence features that confidence adapters can pick up on.

C Related Work

Contrast Sets. Since neural networks are prone to learning spurious features for underspecified tasks (Geirhos et al., 2020; D'Amour et al., 2022), a line of work has investigated training or evaluating them on minimal pairs – instances with slight differences that affect the correct prediction – to expose reliance on spurious features (Shekhar et al., 2017; Marvin & Linzen, 2018; Rudinger et al., 2018; Sakaguchi et al., 2021). Gardner et al. (2020) define a *contrast set* to contain minimal pairs and show that models' performance systematically drops on human-written contrast sets for several language tasks. Subsequent work has leveraged task-specific structure to automate the creation of contrast sets (Li et al., 2020; Bitton et al., 2021; Huang et al., 2025; Mirzadeh et al., 2025). Recently, Ashuach et al. (2026) use contrast sets derived from disagreements between models to evaluate privileged knowledge of LLM introspection, finding that self-probes outperform peer-probes in some domains. In this paper, we induce minimal pairs in knowledge by continually training a model and evaluate whether a confidence estimator faithfully tracks evolving knowledge.

Metacognition. Endowing AI models with metacognition – awareness of their own knowledge and behavior – would make them more transparent and auditable, promoting safety and alignment (Yona et al., 2026; Steyvers & Peters, 2026). For example, they would be able to communicate uncertainty upon identifying gaps in their knowledge, as well as seek out information to fill those gaps, which is relevant for active and continual learning (Gureckis & Markant, 2012). As a step towards generalizable metacognition, Binder et al. (2025) and Guo et al. (2026) show that a model trained to explain its behavior will explain its *new* behavior after we change the model to behave differently. Our results contribute to this line of work by showing that a confidence estimator trained to reflect one checkpoint's knowledge can nontrivially reflect a future checkpoint's knowledge even

Table 11: Expanding on Table 1, we consistently find that SC and non-oracle training fall short on contrast-set calibration compared to oracle training, even when full-set calibration is saturated. Results are averaged over 3 seeds. In each column, we bold the best result and underline results not significantly worse under a paired test ($\alpha = 0.05$; see Appendix A.8 for tests).

	Contrast							Full		
Method	$\Delta_0^b \uparrow$	$\Delta_0 \uparrow$	$\Delta^b \uparrow$	$\Delta \uparrow$	AUC ↑	BS ↓	ECE ↓	AUC ↑	BS ↓	ECE ↓
	Olmo 3 7B, TriviaQA									
Self-consistency	0.297	0.329	0.190	0.212	0.733	0.212	0.084	0.884	0.140	0.069
∟ Post-hoc	0.298	0.330	0.210	0.234	0.734	0.215	0.097	0.884	0.133	**0.019**
Non-oracle training	0.333	0.324	0.236	0.239	0.748	0.206	0.104	0.880	0.136	0.032
Oracle training	**0.379**	**0.424**	**0.255**	**0.285**	**0.795**	**0.185**	**0.061**	**0.890**	**0.131**	0.032
	Olmo 3 7B, Jeopardy									
Self-consistency	0.333	0.360	0.201	0.221	0.741	0.209	0.087	0.873	0.120	0.020
∟ Post-hoc	0.333	0.360	0.215	0.236	0.742	0.211	0.092	0.873	0.119	**0.016**
Non-oracle training	0.476	0.476	0.360	0.365	0.821	0.177	0.094	0.871	0.119	0.038
Oracle training	**0.506**	**0.536**	**0.379**	**0.402**	**0.855**	**0.156**	**0.057**	**0.888**	**0.111**	0.020
	Marin 8B, TriviaQA									
Self-consistency	0.263	0.482	0.196	0.349	0.823	0.172	0.098	0.897	0.115	0.052
∟ Post-hoc	0.263	0.482	0.208	**0.378**	0.825	0.174	**0.097**	0.897	0.112	**0.024**
Non-oracle training	0.227	0.460	0.119	0.199	0.806	0.186	0.178	0.834	0.151	0.083
Oracle training	**0.349**	**0.505**	**0.252**	0.373	**0.840**	**0.163**	0.105	**0.901**	**0.110**	0.028
	Marin 8B, Jeopardy									
Self-consistency	0.331	0.377	0.215	0.255	0.760	0.202	0.139	0.886	0.090	0.023
∟ Post-hoc	0.331	0.377	0.226	0.267	0.755	0.207	0.136	0.886	0.089	**0.015**
Non-oracle training	0.397	0.394	0.137	0.139	0.759	0.205	0.200	0.834	0.124	0.129
Oracle training	**0.459**	**0.503**	**0.350**	**0.377**	**0.825**	**0.175**	**0.123**	**0.890**	**0.088**	0.027
	Olmo 3 32B, TriviaQA									
Self-consistency	0.362	0.444	0.264	0.327	0.808	0.179	0.098	**0.903**	0.112	0.048
∟ Post-hoc	0.363	0.445	0.274	**0.342**	0.806	0.183	0.102	**0.903**	**0.108**	**0.018**
Non-oracle training	0.285	0.235	0.163	0.144	0.688	0.224	0.199	0.860	0.131	0.064
Oracle training	**0.384**	**0.484**	**0.274**	0.341	**0.827**	**0.169**	**0.093**	0.898	0.111	0.027
	Olmo 3 32B, Jeopardy									
Self-consistency	0.320	0.364	0.234	0.259	0.754	0.207	0.103	0.893	0.081	0.020
∟ Post-hoc	0.320	0.364	0.246	0.274	0.755	0.209	0.106	0.893	0.081	**0.018**
Non-oracle training	0.399	0.374	0.228	0.218	0.757	0.201	0.140	0.860	0.092	0.063
Oracle training	**0.475**	**0.513**	**0.342**	**0.370**	**0.840**	**0.163**	**0.075**	**0.894**	**0.081**	0.029

when the two knowledge sets differ, although this is not consistent across models. Since persistent calibration requires faithfully tracking knowledge, we hope that it serves future work as a testbed for metacognition.

Table 12: Expanded version of Table 2, for both TriviaQA and Jeopardy. A confidence estimator must outperform its ablations for a nontrivial demonstration of persistent calibration. SC outperforms its ablations on Olmo 3 7B and Olmo 3 32B, while non-oracle training only outperforms its ablations on Olmo 3 7B.

	Contrast							Full		
Method	$\Delta_0^b \uparrow$	$\Delta_0 \uparrow$	$\Delta^b \uparrow$	$\Delta \uparrow$	AUC ↑	BS ↓	ECE ↓	AUC ↑	BS ↓	ECE ↓
	Olmo 3 7B, TriviaQA									
Self-consistency	**0.297**	**0.329**	0.190	**0.212**	**0.733**	**0.212**	**0.084**	**0.884**	**0.140**	0.069
∟ Surrogate	0.211	0.191	**0.204**	0.184	0.637	0.304	0.208	0.859	0.163	0.094
∟ Copy	0.000	0.000	0.000	0.000	0.500	0.250	0.127	0.832	0.161	**0.033**
Non-oracle training	**0.333**	**0.324**	**0.236**	**0.239**	**0.748**	**0.206**	0.104	**0.880**	**0.136**	**0.032**
∟ Surrogate	0.318	0.323	0.182	0.186	0.729	0.210	**0.103**	0.869	0.146	0.060
∟ Copy	0.000	0.000	0.000	0.000	0.500	0.250	0.127	0.839	0.163	0.064
	Olmo 3 7B, Jeopardy									
Self-consistency	**0.333**	**0.360**	0.201	0.221	**0.741**	**0.209**	**0.087**	**0.873**	**0.120**	**0.020**
∟ Surrogate	0.309	0.302	**0.280**	**0.271**	0.700	0.261	0.176	0.869	0.154	0.138
∟ Copy	0.000	0.000	0.000	0.000	0.500	0.250	0.133	0.815	0.146	0.046
Non-oracle training	**0.476**	**0.476**	**0.360**	**0.365**	**0.821**	**0.177**	**0.094**	**0.871**	**0.119**	**0.038**
∟ Surrogate	0.444	0.437	0.317	0.312	0.805	0.181	0.102	0.864	0.127	0.051
∟ Copy	0.000	0.000	0.000	0.000	0.500	0.250	0.133	0.800	0.158	0.071
	Marin 8B, TriviaQA									
Self-consistency	0.263	**0.482**	0.196	**0.349**	**0.823**	**0.172**	**0.098**	**0.897**	**0.115**	**0.052**
∟ Surrogate	**0.277**	0.283	**0.268**	0.266	0.695	0.275	0.224	0.871	0.161	0.127
∟ Copy	0.000	0.000	0.000	0.000	0.500	0.250	0.283	0.829	0.146	0.053
Non-oracle training	0.227	**0.460**	0.119	0.199	**0.806**	**0.186**	**0.178**	0.834	0.151	0.083
∟ Surrogate	**0.318**	0.339	**0.225**	**0.239**	0.748	0.206	0.207	**0.866**	**0.140**	**0.082**
∟ Copy	0.000	0.000	0.000	0.000	0.500	0.250	0.283	0.830	0.159	0.092
	Marin 8B, Jeopardy									
Self-consistency	0.331	0.377	0.215	0.255	0.760	**0.202**	**0.139**	0.886	**0.090**	**0.023**
∟ Surrogate	**0.515**	**0.478**	**0.448**	**0.425**	**0.806**	0.203	0.182	**0.896**	0.122	0.126
∟ Copy	0.000	0.000	0.000	0.000	0.500	0.250	0.235	0.831	0.112	0.053
Non-oracle training	0.397	0.394	0.137	0.139	0.759	0.205	0.200	0.834	0.124	0.129
∟ Surrogate	**0.504**	**0.496**	**0.400**	**0.388**	**0.843**	**0.163**	**0.154**	**0.877**	**0.098**	**0.048**
∟ Copy	0.000	0.000	0.000	0.000	0.500	0.250	0.235	0.806	0.128	0.074
	Olmo 3 32B, TriviaQA									
Self-consistency	**0.362**	**0.444**	0.264	**0.327**	**0.808**	**0.179**	**0.098**	**0.903**	**0.112**	0.048
∟ Surrogate	0.288	0.263	**0.279**	0.253	0.688	0.274	0.212	0.883	0.144	0.107
∟ Copy	0.000	0.000	0.000	0.000	0.500	0.250	0.212	0.846	0.137	**0.040**
Non-oracle training	0.285	0.235	**0.163**	0.144	0.688	0.224	0.199	**0.860**	**0.131**	**0.064**
∟ Surrogate	**0.290**	**0.296**	0.157	**0.161**	**0.716**	**0.215**	**0.175**	0.858	0.136	0.068
∟ Copy	0.000	0.000	0.000	0.000	0.500	0.250	0.212	0.832	0.148	0.070
	Olmo 3 32B, Jeopardy									
Self-consistency	**0.320**	**0.364**	0.234	0.259	**0.754**	**0.207**	**0.103**	**0.893**	**0.081**	**0.020**
∟ Surrogate	0.312	0.298	**0.284**	**0.272**	0.704	0.271	0.200	0.889	0.115	0.117
∟ Copy	0.000	0.000	0.000	0.000	0.500	0.250	0.153	0.834	0.102	0.051
Non-oracle training	0.399	0.374	0.228	0.218	0.757	0.201	0.140	0.860	**0.092**	0.063
∟ Surrogate	**0.417**	**0.398**	**0.262**	**0.251**	**0.764**	**0.200**	**0.133**	**0.866**	0.095	**0.056**
∟ Copy	0.000	0.000	0.000	0.000	0.500	0.250	0.153	0.817	0.113	0.064

Table 13: Multi-checkpoint and multi-answer training often yields substantial gains. TriviaQA results.

	Contrast							Full		
Method	$\Delta_0^b \uparrow$	$\Delta_0 \uparrow$	$\Delta^b \uparrow$	$\Delta \uparrow$	AUC ↑	BS ↓	ECE ↓	AUC ↑	BS ↓	ECE ↓
Non-oracle					Olmo 3 7B					
Multi-ckpt, n_{cand}=2	**0.333**	0.324	**0.236**	**0.239**	0.748	0.206	0.104	**0.880**	**0.136**	0.032
Multi-ckpt, n_{cand}=1	0.332	0.335	0.223	0.233	0.749	0.204	0.094	0.877	0.137	**0.032**
Single-ckpt, n_{cand}=2	0.332	**0.340**	0.216	0.227	**0.750**	**0.203**	0.093	0.876	0.140	0.048
Single-ckpt, n_{cand}=1	0.314	0.339	0.208	0.225	0.746	0.206	**0.086**	0.870	0.146	0.063
Oracle										
Multi-ckpt, n_{cand}=2	**0.379**	**0.424**	**0.255**	**0.285**	**0.795**	**0.185**	0.061	**0.890**	**0.131**	0.032
Multi-ckpt, n_{cand}=1	0.356	0.414	0.239	0.274	0.790	0.187	**0.058**	0.887	0.132	**0.026**
Resp-ckpt, n_{cand}=2	0.351	0.396	0.241	0.274	0.778	0.193	0.070	0.884	0.133	0.030
Resp-ckpt, n_{cand}=1	0.317	0.357	0.219	0.246	0.756	0.203	0.083	0.880	0.136	0.033
Non-oracle					Marin 8B					
Multi-ckpt, n_{cand}=2	0.227	0.460	**0.119**	0.199	0.806	0.186	0.178	**0.834**	**0.151**	**0.083**
Multi-ckpt, n_{cand}=1	**0.229**	**0.485**	0.112	**0.207**	**0.815**	**0.182**	**0.167**	0.831	0.153	0.084
Single-ckpt, n_{cand}=2	0.221	0.213	0.050	0.044	0.642	0.237	0.274	0.783	0.175	0.086
Single-ckpt, n_{cand}=1	0.213	0.212	0.047	0.045	0.647	0.236	0.272	0.780	0.175	0.085
Oracle										
Multi-ckpt, n_{cand}=2	**0.349**	**0.505**	0.252	0.373	**0.840**	**0.163**	**0.105**	0.901	0.110	0.028
Multi-ckpt, n_{cand}=1	0.315	0.490	0.232	0.353	0.835	0.166	0.110	0.897	0.111	0.023
Resp-ckpt, n_{cand}=2	0.330	0.496	**0.257**	**0.375**	0.837	0.166	0.105	**0.901**	**0.110**	0.025
Resp-ckpt, n_{cand}=1	0.324	0.466	0.244	0.351	0.823	0.173	0.121	0.896	0.111	**0.021**
Non-oracle					Olmo 3 32B					
Multi-ckpt, n_{cand}=2	**0.285**	**0.235**	**0.163**	**0.144**	**0.688**	**0.224**	0.199	**0.860**	**0.131**	0.064
Multi-ckpt, n_{cand}=1	0.268	0.221	0.152	0.137	0.683	0.225	**0.198**	0.860	0.132	**0.063**
Single-ckpt, n_{cand}=2	0.251	0.208	0.140	0.125	0.673	0.228	0.201	0.849	0.138	0.070
Single-ckpt, n_{cand}=1	0.225	0.180	0.119	0.103	0.648	0.236	0.214	0.836	0.143	0.072
Oracle										
Multi-ckpt, n_{cand}=2	**0.384**	**0.484**	**0.274**	**0.341**	**0.827**	**0.169**	0.093	**0.898**	**0.111**	0.027
Multi-ckpt, n_{cand}=1	0.345	0.478	0.249	0.330	0.824	0.170	**0.085**	0.892	0.114	0.028
Resp-ckpt, n_{cand}=2	0.352	0.459	0.258	0.327	0.812	0.176	0.097	0.892	0.113	**0.024**
Resp-ckpt, n_{cand}=1	0.324	0.459	0.226	0.315	0.808	0.178	0.087	0.884	0.118	0.027

Table 14: Multi-checkpoint and multi-answer training often yields substantial gains. Jeopardy results.

Method	Contrast							Full		
	$\Delta_0^b \uparrow$	$\Delta_0 \uparrow$	$\Delta^b \uparrow$	$\Delta \uparrow$	AUC ↑	BS ↓	ECE ↓	AUC ↑	BS ↓	ECE ↓
Non-oracle					Olmo 3 7B					
Multi-ckpt, n_{cand}=2	**0.476**	**0.476**	**0.360**	**0.365**	**0.821**	**0.177**	0.094	**0.871**	**0.119**	**0.038**
Multi-ckpt, n_{cand}=1	0.438	0.438	0.327	0.328	0.802	0.186	0.103	0.863	0.124	0.042
Single-ckpt, n_{cand}=2	0.447	0.457	0.318	0.327	0.808	0.181	0.094	0.860	0.124	0.039
Single-ckpt, n_{cand}=1	0.418	0.441	0.278	0.294	0.796	0.185	**0.090**	0.851	0.130	0.050
Oracle										
Multi-ckpt, n_{cand}=2	**0.506**	**0.536**	**0.379**	**0.402**	**0.855**	**0.156**	0.057	**0.888**	**0.111**	**0.020**
Multi-ckpt, n_{cand}=1	0.478	0.518	0.343	0.372	0.845	0.161	**0.055**	0.880	0.117	0.033
Resp-ckpt, n_{cand}=2	0.471	0.485	0.352	0.364	0.826	0.172	0.079	0.877	0.116	0.028
Resp-ckpt, n_{cand}=1	0.436	0.456	0.311	0.326	0.806	0.182	0.087	0.864	0.123	0.038
Non-oracle					Marin 8B					
Multi-ckpt, n_{cand}=2	**0.397**	0.394	**0.137**	0.139	0.759	0.205	0.200	**0.834**	0.124	0.129
Multi-ckpt, n_{cand}=1	0.318	**0.398**	0.120	**0.146**	**0.776**	**0.200**	**0.178**	0.826	**0.121**	**0.111**
Single-ckpt, n_{cand}=2	0.240	0.076	0.059	0.021	0.562	0.248	0.261	0.757	0.152	0.176
Single-ckpt, n_{cand}=1	0.235	0.092	0.054	0.023	0.571	0.246	0.256	0.756	0.149	0.167
Oracle										
Multi-ckpt, n_{cand}=2	**0.459**	**0.503**	0.350	0.377	**0.825**	0.175	0.123	0.890	0.088	0.027
Multi-ckpt, n_{cand}=1	0.424	0.496	0.310	0.365	**0.825**	**0.174**	**0.105**	0.881	0.089	**0.020**
Resp-ckpt, n_{cand}=2	0.453	0.492	**0.360**	**0.382**	0.824	0.176	0.126	**0.891**	**0.087**	0.027
Resp-ckpt, n_{cand}=1	0.410	0.479	0.306	0.354	0.818	0.178	0.115	0.880	0.089	0.022
Non-oracle					Olmo 3 32B					
Multi-ckpt, n_{cand}=2	**0.399**	**0.374**	**0.228**	**0.218**	**0.757**	**0.201**	0.140	**0.860**	0.092	0.063
Multi-ckpt, n_{cand}=1	0.363	0.369	0.200	0.204	0.755	0.202	0.131	0.853	**0.092**	**0.053**
Single-ckpt, n_{cand}=2	0.315	0.318	0.174	0.178	0.728	0.212	0.140	0.830	0.100	0.073
Single-ckpt, n_{cand}=1	0.310	0.339	0.164	0.181	0.734	0.210	**0.125**	0.825	0.099	0.061
Oracle										
Multi-ckpt, n_{cand}=2	**0.475**	**0.513**	**0.342**	**0.370**	**0.840**	**0.163**	0.075	**0.894**	**0.081**	0.029
Multi-ckpt, n_{cand}=1	0.414	0.469	0.284	0.322	0.819	0.173	0.067	0.879	0.084	**0.021**
Resp-ckpt, n_{cand}=2	0.448	0.481	0.323	0.347	0.824	0.172	0.083	0.884	0.083	0.026
Resp-ckpt, n_{cand}=1	0.375	0.450	0.254	0.305	0.805	0.181	**0.067**	0.866	0.087	0.022

Table 15: Analogous set up to Table 4 with the same trends. Multi-checkpoint training continues to outperform its two ablations, although the single-checkpoint data ablation is strikingly close, suggesting that access to multiple checkpoints' weights accounts for most of the performance gain of multi-checkpoint training.

Method	Contrast							Full		
	$\Delta_0^b \uparrow$	$\Delta_0 \uparrow$	$\Delta^b \uparrow$	$\Delta \uparrow$	AUC ↑	BS ↓	ECE ↓	AUC ↑	BS ↓	ECE ↓
Multi-ckpt	0.506	**0.536**	0.379	**0.402**	**0.855**	**0.156**	**0.057**	**0.888**	**0.111**	**0.020**
∟ Single-ckpt data	**0.517**	0.517	**0.382**	0.383	0.841	0.164	0.087	0.885	0.115	0.041
Resp-ckpt	0.471	0.485	0.352	0.364	0.826	0.172	0.079	0.877	0.116	0.028
∟ Pooled data	0.506	0.499	0.380	0.376	0.833	0.169	0.094	0.884	0.113	0.027

Table 16: Analogous setup to Table 11, but using a fitted normalization function (Appendix B.1) instead of Eq. 1. The two tables share the finding that SC and non-oracle training underperform oracle training on contrast-set calibration, even when full-set calibration is saturated.

	Contrast							Full		
Method	$\Delta_0^b \uparrow$	$\Delta_0 \uparrow$	$\Delta^b \uparrow$	$\Delta \uparrow$	AUC ↑	BS ↓	ECE ↓	AUC ↑	BS ↓	ECE ↓
	Olmo 3 7B, TriviaQA									
End-correct baseline	0.000	0.254	0.000	0.080	0.627	0.230	**0.000**	0.530	0.370	0.308
Self-consistency	0.297	0.329	0.161	0.180	0.739	0.205	0.080	0.884	0.140	0.069
∟ Post-hoc	0.298	0.330	0.163	0.181	0.740	0.205	0.080	0.884	0.133	**0.019**
Non-oracle training	0.333	0.324	0.211	0.219	0.759	0.196	0.092	0.880	0.136	0.032
Oracle training	**0.379**	**0.424**	**0.254**	**0.283**	**0.803**	**0.180**	0.056	**0.890**	**0.131**	0.032
	Olmo 3 7B, Jeopardy									
End-correct baseline	0.000	0.266	0.000	0.088	0.633	0.228	**0.000**	0.540	0.369	0.362
Self-consistency	0.333	0.360	0.176	0.193	0.750	0.202	0.086	0.873	0.120	0.020
∟ Post-hoc	0.333	0.360	0.178	0.194	0.752	0.201	0.086	0.873	0.119	**0.016**
Non-oracle training	0.476	0.476	0.318	0.326	0.828	0.169	0.089	0.871	0.119	0.038
Oracle training	**0.506**	**0.536**	**0.371**	**0.393**	**0.860**	**0.152**	0.059	**0.888**	**0.111**	0.020
	Marin 8B, TriviaQA									
End-correct baseline	0.000	**0.565**	0.000	0.320	0.783	0.170	**0.000**	0.547	0.390	0.393
Self-consistency	0.263	0.482	0.185	0.321	0.828	0.169	0.110	0.897	0.115	0.052
∟ Post-hoc	0.263	0.482	0.179	0.320	0.827	0.170	0.108	0.897	0.112	**0.024**
Non-oracle training	0.227	0.460	0.162	0.340	0.833	0.166	0.080	0.834	0.151	0.083
Oracle training	**0.349**	0.505	**0.248**	**0.361**	**0.847**	**0.159**	0.112	**0.901**	**0.110**	0.028
	Marin 8B, Jeopardy									
End-correct baseline	0.000	0.469	0.000	0.220	0.735	0.195	**0.000**	0.549	0.397	0.433
Self-consistency	0.331	0.377	0.191	0.226	0.771	0.193	0.152	0.886	0.090	0.023
∟ Post-hoc	0.331	0.377	0.192	0.223	0.770	0.194	0.157	0.886	0.089	**0.015**
Non-oracle training	0.397	0.394	0.203	0.256	0.790	0.186	0.130	0.834	0.124	0.129
Oracle training	**0.459**	**0.503**	**0.306**	**0.331**	**0.831**	**0.167**	0.136	**0.890**	**0.088**	0.027
	Olmo 3 32B, TriviaQA									
End-correct baseline	0.000	0.423	0.000	0.205	0.712	0.199	**0.000**	0.548	0.364	0.350
Self-consistency	0.362	0.444	0.246	0.303	0.814	0.174	0.100	**0.903**	0.112	0.048
∟ Post-hoc	0.363	0.445	0.246	0.303	0.814	0.174	0.101	**0.903**	**0.108**	**0.018**
Non-oracle training	0.285	0.235	0.144	0.143	0.700	0.215	0.184	0.860	0.131	0.064
Oracle training	**0.384**	**0.484**	**0.277**	**0.345**	**0.832**	**0.164**	0.082	0.898	0.111	0.027
	Olmo 3 32B, Jeopardy									
End-correct baseline	0.000	0.306	0.000	0.113	0.653	0.222	**0.000**	0.549	0.373	0.417
Self-consistency	0.320	0.364	0.206	0.229	0.771	0.193	0.094	0.893	0.081	0.020
∟ Post-hoc	0.320	0.364	0.202	0.225	0.769	0.194	0.094	0.893	0.081	**0.018**
Non-oracle training	0.399	0.374	0.225	0.224	0.768	0.194	0.128	0.860	0.092	0.063
Oracle training	**0.475**	**0.513**	**0.342**	**0.369**	**0.848**	**0.158**	0.070	**0.894**	**0.081**	0.029

Table 17: Transfer from TriviaQA to BioASQ. The advantage of oracle training over other methods does not hold under domain transfer, consistent with prior work on the domain-specificity of correctness signals (Sky et al., 2024).

Method	Contrast							Full		
	$\Delta_0^b \uparrow$	$\Delta_0 \uparrow$	$\Delta^b \uparrow$	$\Delta \uparrow$	AUC ↑	BS ↓	ECE ↓	AUC ↑	BS ↓	ECE ↓
	Olmo 3 7B, TriviaQA → BioASQ									
Self-consistency	0.122	0.137	0.040	0.043	0.612	0.239	0.096	0.738	0.216	0.087
Non-oracle training	**0.223**	0.239	0.115	0.126	0.692	0.219	0.074	0.805	0.189	0.090
Oracle training	0.211	**0.256**	**0.121**	**0.140**	**0.703**	**0.215**	**0.061**	**0.805**	**0.184**	**0.054**
	Marin 8B, TriviaQA → BioASQ									
Self-consistency	**0.204**	0.267	0.105	0.131	0.703	0.217	0.223	0.736	0.215	**0.078**
Non-oracle training	0.102	**0.449**	0.083	**0.280**	**0.803**	**0.180**	**0.075**	0.739	0.253	0.187
Oracle training	0.198	0.256	**0.121**	0.144	0.709	0.214	0.221	**0.801**	**0.189**	0.081
	Olmo 3 32B, TriviaQA → BioASQ									
Self-consistency	0.139	0.151	0.049	0.055	0.626	0.236	0.135	0.751	0.201	**0.048**
Non-oracle training	0.169	0.167	0.060	0.068	0.638	0.233	0.136	0.774	0.206	0.108
Oracle training	**0.198**	**0.207**	**0.083**	**0.092**	**0.666**	**0.227**	**0.127**	**0.805**	**0.180**	0.055

Table 18: Transfer from Jeopardy to BioASQ. The advantage of oracle training over other methods does not hold under domain transfer, consistent with prior work on the domain-specificity of correctness signals (Sky et al., 2024).

Method	Contrast							Full		
	$\Delta_0^b \uparrow$	$\Delta_0 \uparrow$	$\Delta^b \uparrow$	$\Delta \uparrow$	AUC ↑	BS ↓	ECE ↓	AUC ↑	BS ↓	ECE ↓
	Olmo 3 7B, Jeopardy → BioASQ									
Self-consistency	0.122	0.137	0.040	0.043	0.612	0.239	0.096	0.738	0.216	0.087
Non-oracle training	0.165	0.160	**0.086**	0.087	0.655	0.228	0.097	0.778	0.201	0.090
Oracle training	**0.183**	**0.198**	0.085	**0.091**	**0.658**	**0.227**	**0.087**	**0.780**	**0.196**	**0.064**
	Marin 8B, Jeopardy → BioASQ									
Self-consistency	**0.204**	0.267	**0.105**	0.131	0.703	0.217	0.223	0.736	0.215	**0.078**
Non-oracle training	0.151	**0.334**	0.068	**0.183**	**0.745**	**0.204**	**0.152**	0.734	0.241	0.150
Oracle training	0.144	0.028	0.036	0.023	0.572	0.244	0.278	**0.767**	**0.204**	0.095
	Olmo 3 32B, Jeopardy → BioASQ									
Self-consistency	0.139	0.151	0.049	0.055	0.626	0.236	0.135	0.751	0.201	**0.048**
Non-oracle training	**0.181**	**0.172**	**0.060**	**0.068**	**0.642**	**0.233**	**0.132**	0.764	0.215	0.129
Oracle training	0.169	0.134	0.045	0.042	0.609	0.240	0.148	**0.770**	**0.198**	0.077

Table 19: Spearman correlations between checkpoint index and adapter variance for Olmo 3 7B on TriviaQA. For each question and layer, we compute the total variance (i.e. trace of the covariance matrix), across 10 confidence adapter training seeds, of the model's hidden states, and we compute the correlation between checkpoint index and this total variance. We report the mean and median correlation across the 4960 questions. We find that the correlation is consistently positive for all layers, suggesting that later checkpoints tend to have a larger space of viable confidence features that confidence adapters can pick up on.

Layer	Mean ρ	Median ρ
1	0.920	0.929
2	0.914	0.929
3	0.631	0.750
4	0.507	0.536
5	0.533	0.714
6	0.535	0.750
7	0.613	0.786
8	0.532	0.607
9	0.385	0.357
10	0.414	0.429
11	0.379	0.429
12	0.409	0.429
13	0.505	0.536
14	0.467	0.464
15	0.452	0.464
16	0.442	0.429
17	0.510	0.536
18	0.488	0.536
19	0.454	0.500
20	0.397	0.464
21	0.350	0.357
22	0.312	0.321
23	0.276	0.286
24	0.279	0.286
25	0.273	0.286
26	0.274	0.286
27	0.273	0.286
28	0.278	0.286
29	0.248	0.250
30	0.178	0.143
31	0.279	0.321
32	0.674	0.679