Calibrated Confidence Expression for Radiology Report Generation

David Bani-Harouni[1,2⋆], Chantal Pellegrini[1,2⋆(✉)], Julian Lüers[1], Su Hwan Kim[3,4], Markus Baalmann[5], Benedikt Wiestler[2,4,6], Rickmer Braren[3,5], Nassir Navab[1,2], and Matthias Keicher[1,2]

[1] Computer Aided Medical Procedures, Technical University of Munich, Germany
{david.bani-harouni,chantal.pellegrini}@tum.de
[2] Munich Center for Machine Learning (MCML), Germany
[3] Department of Diagnostic and Interventional Radiology, TUM Klinikum rechts der Isar, Germany
[4] Department of Diagnostic and Interventional Neuroradiology, TUM Klinikum rechts der Isar, Germany
[5] Department of Diagnostic and Interventional Radiology and Nuclear Medicine, University Medical Center Hamburg-Eppendorf, Germany
[6] AI for Image-Guided Diagnosis and Therapy, Technical University of Munich

Abstract. Safe deployment of Large Vision-Language Models (LVLMs) in radiology report generation requires not only accurate predictions but also clinically interpretable indicators of when outputs should be thoroughly reviewed, enabling selective radiologist verification and reducing the risk of hallucinated findings influencing clinical decisions. One intuitive approach to this is verbalized confidence, where the model explicitly states its certainty. However, current state-of-the-art language models are often overconfident, and research on calibration in multimodal settings such as radiology report generation is limited. To address this gap, we introduce ConRad (Confidence Calibration for Radiology Reports), a reinforcement learning framework for fine-tuning medical LVLMs to produce calibrated verbalized confidence estimates alongside radiology reports. We study two settings: a single report-level confidence score and a sentence-level variant assigning a confidence to each claim. Both are trained using the GRPO algorithm with reward functions based on the logarithmic scoring rule, which incentivizes truthful self-assessment by penalizing miscalibration and guarantees optimal calibration under reward maximization. Experimentally, ConRad substantially improves calibration and outperforms competing methods. In a clinical evaluation we show that ConRad's report level scores are well aligned with clinicians' judgment. By highlighting full reports or low-confidence statements for targeted review, ConRad can support safer clinical integration of AI-assistance for report generation. Our code is available at https://github.com/ChantalMP/Conrad.

Keywords: Confidence Calibration · Report Generation · Reinforcement Learning.

⋆ Equal Contribution

1 Introduction

Recent progress in AI for radiology shows growing potential to assist with rising clinical demands [14,1]. Specifically for report generation, Large Vision–Language Models (LVLMs) have improved generation of coherent and correct radiology reports directly from medical images [5,4,20,15,21]. However, outputs generated by LVLMs are prone to issues such as hallucinations, while often being overly confident in their prediction [8,28]. Therefore, for such systems to be clinically useful, they must provide not only accurate outputs but also trustworthy, clinically actionable indicators of output quality, signaling when clinicians can rely on model findings and when in-depth verification is needed. One promising strategy is letting the model predict not only an output but also a confidence estimate, an indicator of how much the output should be trusted. If confidence is well-calibrated, it can support report-level triage and prioritization for selective human review as well as targeted verification of report passages, thereby reducing the effort for oversight without compromising safety.

Confidence calibration has been widely studied for large language models [23,7], spanning post-hoc black/white-box estimators and trained external predictors [2]. Verbalization of confidence has been targeted using supervised training on pseudo-ground-truth [31,13] or reinforcement learning against human feedback [16] or correctness measures [3,30]. In radiology report generation, uncertainty has been explored as clinical uncertainty phrasing learned from report language [26,29], however this does not target the model's inherent uncertainty but aims to reproduce the confidence expressions of radiologists. Other methods model uncertainty estimation via stochastic sampling or semantic consistency across multiple generated reports [25,24], or via post-hoc auditing that validates high-level extracted findings against external image classifiers [27]. These approaches are, however, inefficient at inference as they require multiple generations to estimate uncertainty. They further do not train explicitly for improving the internal confidence calibration of the model and do not enable it to express calibrated confidences alongside reports.

Inspired by the success of reinforcement learning with proper scoring rules for calibrating confidence in large language models [3,30], we propose ConRad, which extends scoring-rule–based confidence calibration to multimodal, long-form radiology report generation. ConRad fine-tunes medical LVLMs to express verbalized confidence jointly with the report text, and supports both report-level and fine-grained sentence-level confidence estimates to guide targeted verification. We combine Group Relative Policy Optimization (GRPO) [22] with a logarithmic scoring-rule reward and a continuous correctness signal provided by an automated report-quality metric, enabling calibration without manual confidence supervision. In summary, we propose: 1) a novel RL-based approach for verbalized confidence calibration in multimodal long-form report generation; 2) two settings for report- and sentence-level confidence expression enabling confidence-based verification; and 3) a reward design that couples policy optimization of the proper scoring rule with a clinical report quality metric. We evaluate ConRad on MIMIC-CXR, demonstrating substantial calibration gains over strong

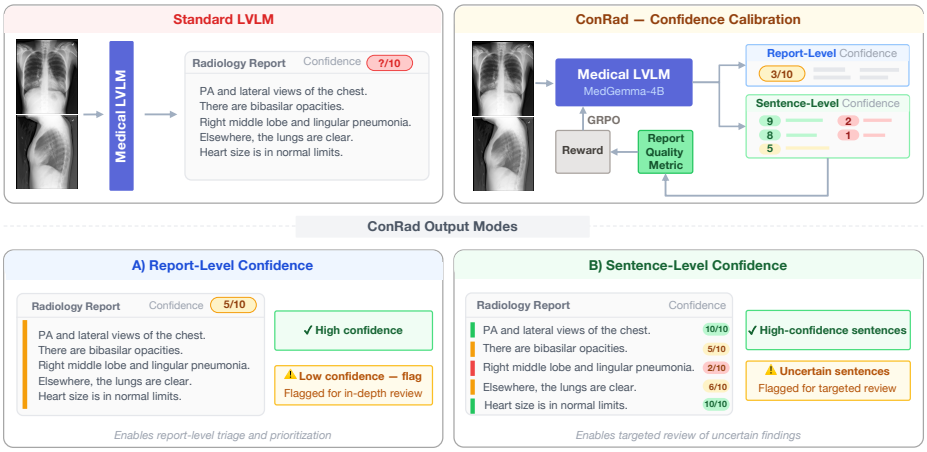

Fig. 1. Overview of ConRad. While standard medical LVLMs provide reports without any confidence measure, fine-tuning them with ConRad enables calibrated confidence expression on a report- and sentence-level that could be used for targeted review.

confidence estimation baselines, supporting safer integration of LVLM report generators into clinical workflows via confidence-guided review.

2 Methodology

We propose ConRad, a method for calibrating verbalized confidence in medical Large Vision-Language Models (LVLMs) during radiology report generation, enabling confidence-guided review in clinical workflows. ConRad supports both report- and sentence-level confidence scores, allowing clinicians to quickly identify which reports or statements need targeted verification. Our method fine-tunes LVLMs to jointly generate findings and numerical confidence scores by formulating the confidence expression as a Markov Decision Process, employing Reinforcement Learning (RL) to align the model's verbalized confidence with an external correctness assessment. In this setting, confidence token predictions are the actions, while prompt and report represent the state.

2.1 Problem Formulation: Confidence and Calibration

Let $v = (v_1, \ldots, v_k)$ denote a set of input medical images of a radiological study. The model generates a textual report r and a verbalized scalar confidence score, which we scale to $\hat{p} \in [0, 1]$. We define confidence as the model's self-reported quality estimate of the report r according to a ground-truth correctness metric $s \in [0, 1]$ (e.g., semantic similarity to a ground-truth reference report). In this setting perfect calibration is achieved, when the verbalized confidence matches the expected correctness:

$$\mathbb{E}[s \mid \hat{p} = x] = x, \quad \forall x \in [0, 1]. \tag{1}$$

2.2 Confidence Generation

We employ a pre-trained LVLM that accepts images v and prompt l to generate a response. We investigate two granularity levels for confidence expression:
Report-level Scenario: The model generates a complete report r followed by a single confidence score $\hat{p}$ representing the confidence of the entire report.
Sentence-level Scenario: A report is generated as a sequence of m sentences, $r = (r_1, \ldots, r_m)$. After each sentence r_i, the model generates a confidence score $\hat{p}_i$ to form an output $((r_1, \hat{p}_1), \ldots, (r_m, \hat{p}_m))$.

2.3 Reinforcement Learning for Calibration

We train confidence calibration using reinforcement learning. Unlike supervised LLM fine-tuning on expressing correctness values, which treats numerical confidence tokens as independent categorical labels and can ignore their ordinal structure, our RL approach utilizes a continuous reward signal that varies continuously with the distance of predicted confidence and correctness, so near-misses are preferred over large miscalibration. This incentivizes the model to learn a gradual internal representation of confidence as also shown by Zhang et al. [30].

To incentivize calibration, the reward function R must be maximized when the predicted confidence $\hat{p}$ aligns with the true correctness probability. We base our rewards on the Logarithmic Scoring Rule, a strictly proper scoring rule where the expected reward is uniquely maximized if and only if the predicted probability matches the true distribution. Note, that our goal is not to improve the quality of the report generation capabilities itself but only the confidence calibration of the model. We therefore only calculate a loss for the tokens involved in confidence expression, ensuring that the task performance remains stable.
Report-level Reward ($R_{\mathbf{R}}$). In the report-level scenario, the reward is computed with respect to a continuous correctness score $s \in [0, 1]$ evaluating the entire report. We define the reward as the logarithmic scoring rule with s as the target and λ as scaling factor:

$$R_{\mathrm{R}}(r, \hat{p}, s) = \lambda \cdot [s \log(\hat{p}) + (1 - s) \log(1 - \hat{p})] \tag{2}$$

This formulation can be thought of as policy optimization of the cross-entropy between predicted confidence and correctness target that penalizes deviations between the predicted $\hat{p}$ and the measured correctness s.
Sentence-level Reward ($R_{\mathbf{S}}$). In the sentence-level scenario, correctness is evaluated with a binary correctness score $f_i \in \{0, 1\}$ for each sentence i. Since f_i is binary, we define the reward as the mean logarithmic score over all sentences:

$$R_{\mathrm{S}}(r, \hat{p}, f) = \frac{\lambda}{m} \sum_{i=1}^{m} [f_i \log(\hat{p}_i) + (1 - f_i) \log(1 - \hat{p}_i)] \tag{3}$$

To ensure numerical stability, we clip confidence predictions to $[\varepsilon, 1 - \varepsilon]$ with ε sufficiently small.

Table 1. Report-level evaluation results. ECE and Pearson correlation serve as the primary evaluation metrics, GREEN score is included as a control to verify that report generation quality remains stable. 95 % confidence intervals in brackets.

	Calibration Metrics		
Method	**ECE** (↓)	**Correlation** (↑)	**GREEN**
MedGemma Base [21]	-	-	0.283 [0.271, 0.296]
Sequence Probability [10]	0.531 [0.519, 0.542]	0.340 [0.282, 0.388]	0.286 [0.273, 0.299]
P(True) [12]	0.399 [0.380, 0.418]	0.166 [0.113, 0.222]	0.287 [0.275, 0.300]
Self-Consistency [17]	0.436 [0.424, 0.448]	0.333 [0.277, 0.393]	0.297 [0.284, 0.310]
Trained Probe [2]	0.121 [0.110, 0.135]	0.372 [0.318, 0.422]	0.299 [0.287, 0.312]
Verbalize Base	0.642 [0.630, 0.654]	0.136 [0.078, 0.194]	0.257 [0.245, 0.268]
Verbalize Supervised	0.206 [0.189, 0.223]	0.160 [0.089, 0.222]	0.252 [0.240, 0.264]
ConRad	**0.106** [0.093, 0.118]	**0.431** [0.372, 0.487]	0.252 [0.240, 0.264]

This reward structure encourages the model to be confident only when it is correct (high s or $f = 1$) and to express uncertainty (low $\hat{p}$) when it is likely to be incorrect, effectively calibrating the model via RL.

3 Experiments

Experimental Setup and Metrics. All main experiments use MIMIC-CXR [11]. Given all chest x-rays of a study and the Indication section, the model generates the Findings section and confidence estimates. Confidence is an integer in $\{0, \ldots, 10\}$, normalized to $[0, 1]$ for evaluation. Calibration is measured with Expected Calibration Error (ECE); we additionally report Pearson correlation for report-level confidence, comparing to the continuous correctness score and AUROC for sentence-level confidence, comparing to binary per-sentence correctness. To test zero-shot transfer, we evaluate on the IU-Xray [6] dataset, using the test split defined in Rad-ReStruct [19].

We compare against representative black and white-box confidence estimation baselines: *Verbalize Base* (zero-shot confidence via prompting), *Verbalize Supervised* (SFT on explicit GREEN scores appended to the report text using cross-entropy loss), *Sequence Probability* (mean token probability [10]), *P(True)* (two-stage self-assessment where the model predicts true or false after the report; confidence estimated from probabilities of true vs. false token [12]), *Self-Consistency* (confidence as agreement across multiple sampled reports [17]), and *Trained Probe* (MLP predicts correctness from hidden states of the LLM [2]).

Implementation Details. Report-level confidence is trained on 3000 samples for one epoch; sentence-level confidence on 1500 samples for one epoch, both until convergence. We use a learning rate of 10^{-5} on an NVIDIA A40 GPU. We use a 4-bit quantized `medgemma-4b-it` [21] as report generation model and fine-tune only the language model component using LoRA [9] and GRPO [22]. GRPO is an RL algorithm for LLM finetuning. It samples multiple candidate

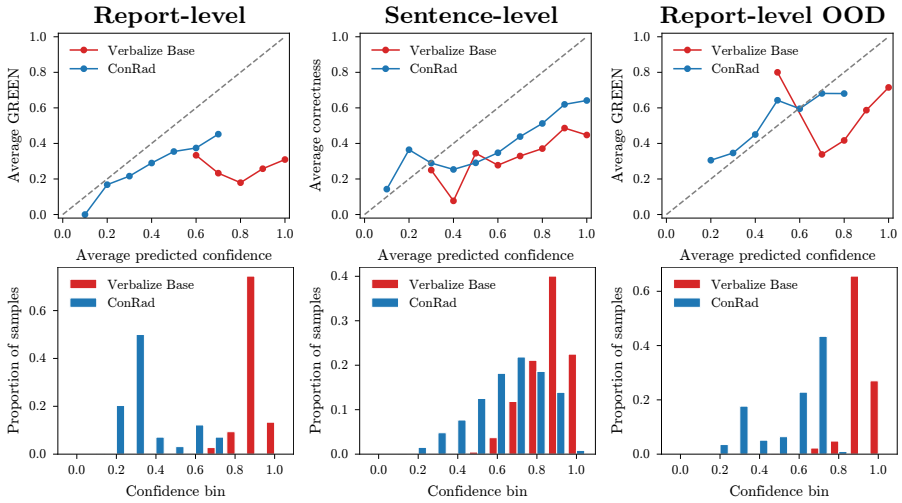

Fig. 2. Verbalize Base vs. ConRad: calibration curves (top row) and confidence distributions (bottom row) for report-level, sentence-level, and out-of-distribution (OOD).

outputs for the same input, normalizes and compares their rewards, and updates the model to increase the likelihood of the comparatively higher-reward outputs. As a correctness signal s, we use GREEN [18], which is a common report generation metric that scores a generated report against the reference report in $[0, 1]$ using an LLM-as-a-judge, evaluating the ratio of matched findings and clinically significant errors. For sentence-level correctness scoring, we utilize an adapted GREEN (Precision GREEN) f that only evaluates whether the finding in that sentence is supported by the reference report. During training, the loss is masked to the confidence tokens, constraining optimization to confidence prediction rather than report content. We set the reward scaling λ to 100. Outputs that violate the required format receive a fixed negative reward. For the trained probe, we also train on 3000 samples for 10 epochs with early stopping on the validation set with a learning rate of 10^{-4} and use the hidden states of layer 35.

4 Results and Discussion

Report-level Confidence Calibration. Table 1 summarizes the performance in the report-level confidence scenario. ConRad outperforms all baselines, reducing the ECE by over 80% compared to the base model and achieving the highest Pearson correlation with factual correctness. As shown in Figure 2, the fine-tuned model shifts from the base model's extreme overconfidence to a more balanced distribution and a calibration curve closer to the ideal diagonal. We observe a decrease in GREEN score when prompting MedGemma to verbalize its confidence (Verbalize Base) compared to report-only prompting, likely caused by the shift from the MedGemma training setting. At the same time the

Table 2. Sentence-level evaluation results. ECE and AUROC assess calibration and discriminative power, GREEN score is included as a control to verify that report generation quality remains stable. 95 % confidence intervals in brackets.

	Calibration Metrics		
Method	**ECE** (↓)	**AUROC** (↑)	**GREEN**
MedGemma Base [21]	-	-	0.283 [0.271, 0.296]
Verbalize Base	0.438 [0.423, 0.453]	0.558 [0.543, 0.572]	0.268 [0.257, 0.279]
Verbalize Supervised	0.446 [0.433, 0.460]	0.552 [0.538, 0.565]	0.267 [0.257, 0.280]
ConRad	**0.239** [0.225, 0.253]	**0.633** [0.620, 0.647]	0.262 [0.250, 0.274]

report correctness remains stable compared to Verbalize Base during training with ConRad. This is desired as we do not want the training to influence the report generation process itself.

While the Trained Probe achieves a competitive ECE (Table 1), it depends on internal hidden states, whereas ConRad offers a solution that verbalizes confidence directly. Furthermore, unlike supervised fine-tuning which treats tokens as independent categories, our RL approach preserves the ordinal structure of confidence. This is consistent with the softmax distribution over the confidence-token vocabulary, where ConRad assigns the highest probability to the predicted confidence level and progressively smaller probabilities to adjacent levels, resulting in a smooth notion of "nearby" certainty. In contrast, supervised fine-tuning as in Verbalize Supervised disrupts this ordering, producing a less coherent probability pattern across tokens that suggests the model is learning a set of unrelated labels rather than a graded confidence scale.

Sentence-level Confidence Calibration. We additionally explore a sentence-level confidence scenario. Here, ConRad evaluates each sentence confidence separately against a binary correctness target. As shown in Table 2, our method reduces sentence-level ECE by ∼40% and improves AUROC compared to the base model. Figure 2 confirms this improvement, showing a calibration curve that tracks the diagonal more closely than the base model. Also in this setting, the GREEN score remains stable after training with ConRad. This experiment also highlights an advantage of our RL approach over SFT for calibration in settings with binary targets. Through its RL-based formulation, the model trained with ConRad is still able to learn fine-grained confidence predictions, while Verbalize Supervised can only be trained to reproduce these binary targets leading to suboptimal calibration. Overall, sentence-level calibration remains more challenging than report-level calibration, where we observe stronger overall results.

Confidence-Based Flagging. ConRad enables flagging sentences or reports with low confidence, making clinicians aware of findings that have an increased probability of being wrong. To assess this use case practically, we filter sentences based on their predicted confidence $\hat{p}$ and evaluate the remaining sentences with Precision GREEN. Table 3 demonstrates that the factual precision increases monotonically with the confidence threshold. However, this comes with an ex-

Table 3. Number of sentences remaining after filtering by confidence threshold and Precision GREEN on the filtered sentences.

Threshold	Verbalize Base		ConRad	
	#Sentences	Precision GREEN	#Sentences	Precision GREEN
All	6131	0.425	6317	0.421
Confidence ≥ 6	6081	0.426	4634	0.471
Confidence ≥ 8	5128	0.447	2106	0.560
Confidence $= 10$	1379	0.447	53	0.642

pected trade-off between precision and completeness. The standard GREEN score over the remaining report decreases (0.262 (All) $\rightarrow$ 0.097 (confidence $=$ 10)), indicating omission of uncertain but potentially correct content. This suggests that low-confidence sentences should be flagged for human review rather than automatically removed. In contrast, filtering based on Verbalize Base confidences yields only marginal Precision GREEN gains, despite a similar drop in GREEN score (0.268 $\rightarrow$ 0.121). This indicates that the base model's expressed confidence is poorly aligned with factual correctness, whereas ConRad's confidence provides actionable signals for targeted review.

Performance on OOD Data. To test generalization of ConRad beyond the training data, we evaluate on IU-Xray in a strict zero-shot setting, where ConRad is trained on MIMIC-CXR only, with no IU-Xray exposure. ConRad's calibration generalizes substantially better than the base model's with ECE improving $\sim$9$\times$ from 0.310 to 0.034, while maintaining comparable report quality. As seen in Figure 2, Verbalize Base is severely overconfident, while a much broader range of confidence values is used when applying a model fine-tuned with ConRad.

Clinical Evaluation. To assess the alignment of report confidence to radiologist judgment, we conduct a small-scale clinical evaluation on 50 reports with three raters. Each rater scores every sentence on a 5-point Likert scale from *Reject* to *Accept*. We derive two report-level clinician ratings from these sentence annotations. *Mean aggregation* averages scores across all sentences and raters, yielding a continuous report-level quality estimate. *All accepted* defines a binary acceptability criterion: a report is marked acceptable only if all sentences receive a score of at least 4 from a majority of raters. We report correlation, AUROC and ECE with respect to clinician ratings and model confidences. Table 4 shows that ConRad achieves stronger alignment with clinician judgment across both aggregation settings than Verbalize Base, indicating that the calibration gains of ConRad transfer to expert evaluation, supporting its potential for report-level triage.

5 Conclusion

In this work, we introduced ConRad, a reinforcement-learning framework that fine-tunes medical LVLMs to jointly generate radiology reports and verbalized, calibrated confidence. ConRad supports both a global report-level, and a local sentence-level confidence expression. Our results show that confidence calibration

Table 4. Report-level clinical evaluation on 50 reports with 3 raters. Two report-level targets are derived from sentence-wise Likert ratings: *Mean aggregation* and *All accepted*. We report Spearman correlation, AUROC and ECE.

Aggregation	Method	Correlation	AUROC	ECE
Mean aggregation	Verbalize Base	0.069	0.519	0.537
	ConRad	**0.268**	**0.644**	**0.099**
All accepted	Verbalize Base	-0.036	0.489	0.842
	ConRad	**0.103**	**0.610**	**0.220**

is improved compared to strong baselines, while keeping the report generation performance unchanged. Such calibrated confidence expression, integrated in the report generation process, can enable more efficient expert verification, guided by the model's confidence, opening a path towards safe integration of AI support tools into clinical practice.

Acknowledgments. The authors gratefully acknowledge the financial support by the Bavarian Ministry of Economic Affairs, Regional Development and Energy (StMWi) under project ThoraXAI (DIK-2302-0002), and the German Research Foundation (DFG, grant 469106425 - NA 620/51-1).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Acosta, J.N., Dogra, S., Adithan, S., Wu, K., Moritz, M., Kwak, S., Rajpurkar, P.: The impact of ai assistance on radiology reporting: a pilot study using simulated ai draft reports. arXiv preprint arXiv:2412.12042 (2024)
2. Azaria, A., Mitchell, T.: The internal state of an LLM knows when it's lying. In: The 2023 Conference on Empirical Methods in Natural Language Processing (2023)
3. Bani-Harouni, D., Pellegrini, C., Stangel, P., Özsoy, E., Zaripova, K., Navab, N., Keicher, M.: Rewarding doubt: A reinforcement learning approach to calibrated confidence expression of large language models. In: The Fourteenth International Conference on Learning Representations (2026)
4. Bannur, S., Bouzid, K., Castro, D.C., Schwaighofer, A., Thieme, A., Bond-Taylor, S., Ilse, M., Pérez-García, F., Salvatelli, V., Sharma, H., et al.: Maira-2: Grounded radiology report generation. arXiv preprint arXiv:2406.04449 (2024)
5. Chen, Z., Varma, M., Delbrouck, J.B., Paschali, M., Blankemeier, L., Veen, D.V., Valanarasu, J.M.J., Youssef, A., Cohen, J.P., Reis, E.P., Tsai, E.B., Johnston, A., Olsen, C., Abraham, T.M., Gatidis, S., Chaudhari, A.S., Langlotz, C.: Chexagent: Towards a foundation model for chest x-ray interpretation. arXiv preprint arXiv:2401.12208 (2024), `https://arxiv.org/abs/2401.12208`
6. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23**(2), 304–310 (2016)
7. Geng, J., Cai, F., Wang, Y., Koeppl, H., Nakov, P., Gurevych, I.: A survey of confidence estimation and calibration in large language models. In: ACL: Human Language Technologies. pp. 6577–6595 (2024)
8. Hadi, M.U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M.B., Akhtar, N., Wu, J., Mirjalili, S., et al.: Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. Authorea preprints **1**(3), 1–26 (2023)
9. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
10. Huang, Y., Song, J., Wang, Z., Zhao, S., Chen, H., Juefei-Xu, F., Ma, L.: Look before you leap: An exploratory study of uncertainty analysis for large language models. IEEE Transactions on Software Engineering (2025)
11. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
12. Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., et al.: Language models (mostly) know what they know. arXiv preprint arXiv:2207.05221 (2022)
13. Kang, K., Wallace, E., Tomlin, C., Kumar, A., Levine, S.: Unfamiliar finetuning examples control how language models hallucinate. In: ACL: Human Language Technologies. pp. 3600–3612 (2025)
14. Kaur, N., Mittal, A., Singh, G.: Methods for automatic generation of radiological reports of chest radiographs: a comprehensive survey. Multimed Tools Appl (2022)

15. Lee, S., Youn, J., Kim, H., Kim, M., Yoon, S.H.: Cxr-llava: a multimodal large language model for interpreting chest x-ray images. European Radiology pp. 1–13 (2025)
16. Leng, J., Huang, C., Zhu, B., Huang, J.: Taming overconfidence in LLMs: Reward calibration in RLHF. In: The Thirteenth International Conference on Learning Representations (2025)
17. Lin, Z., Trivedi, S., Sun, J.: Generating with confidence: Uncertainty quantification for black-box large language models. TMLR (2024)
18. Ostmeier, S., Xu, J., Chen, Z., Varma, M., Blankemeier, L., Bluethgen, C., Michalson, A.E., Moseley, M., Langlotz, C., Chaudhari, A.S., et al.: Green: Generative radiology report evaluation and error notation. arXiv preprint arXiv:2405.03595 (2024)
19. Pellegrini, C., Keicher, M., Özsoy, E., Navab, N.: Rad-restruct: A novel vqa benchmark and method for structured radiology reporting. In: MICCAI. pp. 409–419. Springer (2023)
20. Pellegrini, C., Özsoy, E., Busam, B., Wiestler, B., Navab, N., Keicher, M.: Radialog: Large vision-language models for x-ray reporting and dialog-driven assistance. In: MIDL (2025)
21. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
22. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
23. Wang, C.: Calibration in deep learning: A survey of the state-of-the-art. arXiv preprint arXiv:2308.01222 (2023)
24. Wang, C., Zhou, W., Ghosh, S., Batmanghelich, K., Li, W.: Semantic consistency-based uncertainty quantification for factuality in radiology report generation. In: NAACL 2025. pp. 1739–1754 (2025)
25. Wang, Y., Lin, Z., Xu, Z., Dong, H., Luo, J., Tian, J., Shi, Z., Huang, L., Zhang, Y., Fan, J., et al.: Trust it or not: Confidence-guided automatic radiology report generation. Neurocomputing **578**, 127374 (2024)
26. Wang, Z., Yan, S., Yin, K., Zhang, X., Cheung, W.K.: Curv: Coherent uncertainty-aware reasoning in vision-language models for x-ray report generation. In: NeurIPS (2025)
27. Warr, H., Ibrahim, Y., McGowan, D.R., Kamnitsas, K.: Quality control for radiology report generation models via auxiliary auditing components. In: International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging. pp. 70–80. Springer (2024)
28. Xu, H., Zhu, Z., Lan, K., Wang, Z., Wu, M., Ji, Z., Chen, L., Fung, P., Yu, K., et al.: Delusions of large language models. arXiv preprint arXiv:2503.06709 (2025)
29. Yan, S., Yin, H., Tsang, I.W., Cheung, W.K.: Diagnose with uncertainty awareness: Diagnostic uncertainty encoding framework for radiology report generation. In: International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging. pp. 34–44. Springer (2024)
30. Zhang, C., Zhu, X., Li, C., Collier, N., Vlachos, A.: Reinforcement learning for better verbalized confidence in long-form generation. arXiv preprint arXiv:2505.23912 (2025)
31. Zhang, H., Diao, S., Lin, Y., Fung, Y., Lian, Q., Wang, X., Chen, Y., Ji, H., Zhang, T.: R-tuning: Instructing large language models to say 'I don't know'. In: NAACL. pp. 7113–7139 (2024)