

HARMONY: Hidden Activation Representations and Model Output-Aware Uncertainty Estimation for Vision-Language Models

Erum Mushtaq¹ Zalan Fabian¹ Yavuz Faruk Bakman¹ Anil Ramakrishna²
Mahdi Soltanolkotabi¹ Salman Avestimehr¹
¹University of Southern California ²Amazon AGI

Abstract

Assessing the reliability of Vision-Language Models (VLMs) is crucial in high-stakes applications. Uncertainty Estimation (UE) methods are widely used for this purpose. Most existing probability-based UE approaches rely on output probability distributions, aggregating token probabilities into a single uncertainty score using predefined functions. Another line of research leverages model hidden representations, training MLP-based models to predict uncertainty. However, these methods often fall short in capturing the complex semantic and visual relationships between tokens and struggle to identify biased probabilities influenced by language priors. Based on these observations, we propose HARMONY (Hidden Activation Representations and Model Output-aware uNcertaintY Estimation for Vision-Language Models), a transformer-based UE function that jointly leverages model hidden representations and output token probabilities. Our key hypothesis is that both model’s internal belief on the vision understanding, and model’s output carry reliability signal, and leveraging them both simultaneously provide a better uncertainty estimate. Experimental results on two benchmark open-ended VQA datasets (OKVQA and A-OKVQA) and three state-of-the-art VLMs demonstrate that our method consistently outperforms existing approaches, achieving up to 4% improvement in AUROC, and 6% in PRR.

1. Introduction

Vision-Language models (VLMs) have seen tremendous growth in recent years exhibiting promising performance across various tasks such as visual commonsense reasoning, image captioning, and retrieval. However, as their usage is growing, their tendency to produce incorrect outputs poses a serious risk, especially in high-stakes settings such as assisting the Blind or Low Vision (BLV) community [1], and medical diagnosis [17]. In this regard, uncertainty estimation (UE) methods are widely recognized for their ability to detect reliability of the model’s generations. However, majority of existing UE methods for VLMs for-

mulate open-ended VQA tasks as a multiple-choice questions [8, 21]. This simplifies the complexity of the problem, and has its own advantages. For example, it simplifies the variable-length logit vector calibration task to a fixed multi-class logit calibration [21]. It also removes semantic variation challenge of multiple generations to analyze consistency [8]. Further, it constrains the output generation to only one token, avoiding the complexities of inter-token dependencies [21]. However, in real-world settings, having multiple answers known in advance is not always possible. Further, the default mode of these generative models is auto-regressive, which produces variable-length sequences exhibiting complex dependencies between the generated tokens and the textual and visual context.

In free-form generation, the model generates multiple tokens, involving aggregation of sequence token probabilities into a single UE function using a predefined scoring function. Various predefined functions from LLM literature address various aspects of the aggregation of token probabilities [9, 13]. However, finding the aggregation formulation via heuristics in itself can be a challenging problem due to inter-token dependencies involved. In the case of VLMs, the inclusion of visual context adds another layer of complexity. For instance, consider an image of a man looking down and the question, “Where is the man looking?” The model might generate the answer, “The man is looking up.” In this sentence, the token “up” carries significant semantic weight and should be assigned higher importance. Moreover, the uncertainty of the “up” token is influenced by the model’s understanding of the visual context. There exists a related work that looks at model’s internal representation for evidence [21]. However, they solely rely on the model internal representations, and overlook the sequential inter-token dependencies at the output.

Based on these insights, we present HARMONY (Hidden Activation Representations and Model Output-Aware Uncertainty Estimation for Vision-Language Models), a transformer architecture-based uncertainty estimation (UE) function that integrates both the hidden representations of the model and the output token probabilities. Our key hy-

pothesis is that the model’s internal understanding of visual content and its generated outputs both provide crucial reliability signals. By combining these two sources of information, we achieve a better uncertainty estimate. Specifically, we utilize the hidden activation representations, which capture the model’s visual comprehension of the image, alongwith question, generated answer semantics, and output probabilities assigned to each token. To train the function, we employ VisualBERT [11], a small-scale transformer with 113M parameters, which offers a relatively simple cost compared to training the original larger VLM models with billion parameters. Our experimental results on two datasets, and three frontier VLM models show that our method consistently outperforms existing black-box and supervised trainable works achieving up to 4% improvement in AUROC and 6% improvement in PRR scores. We further evaluate our method on a selective prediction task. Our proposed method consistently achieves better performance than the other trainable UE methods, with an improvement of up to 2.5% in the effective reliability metric.

2. Related Works

The existing UE methods can be broadly categorized into the following four types. 1. *Self-Checking methods* rely on the model’s ability to evaluate its own correctness via self-evaluation [19, 20]. However, studies have shown that self-evaluated confidence is insufficient to be a good estimate of uncertainty [7]. 2. *Output Consistency methods* estimate uncertainty by examining the consistency of generated outputs across rephrasings [5, 8, 18] or model confidence over sub-questions [19]. However, rephrasing [8] methods are expensive due to multiple forward passes required by large VLMs, while sub-question-based approaches [19] add further costs with evidence collection and relevance verification. Additionally, sub-question methods assume VLMs are well-calibrated, which is not always the case. 3. *Internal state examination methods* exploit the hidden activation representations of the model to predict the correctness of the response [3, 3]. While effective, these works require calibration datasets to train the function. Further, they train simple architectures such as MLP-based scoring functions neglecting inter-token dependencies. 4. *Token Probability methods* use output token probabilities assigned by the model to predict the uncertainty [9, 13]. Some approaches leverage output probabilities with calibration datasets for supervised UE [22]. In most cases, VLM-based UE methods frame open-ended VQA task as multiple-choice problem [8, 21]. Therefore, token probability methods remain relatively unexplored for generative VLMs. Our proposed method targets free-form generation, and integrates both internal state examination and token probability methods, combining their strengths to achieve a more robust UE framework for Vision-Language Models.

3. Problem Formulation

Uncertainty Estimation Given a question \mathbf{q} , and an Image \mathcal{I} , a VLM model parameterized by θ generates an output response sequence $\mathbf{s} = \{s_1, s_2, \dots, s_k\}$, where k denotes the length of the sequence. The UE methods quantify the uncertainty for the model’s predicted sequence s given the input context. A naive way of estimating uncertainty is to calculate the probability of a generated sequence $P(\mathbf{s}|\mathbf{q}, \mathcal{I}, \theta) = \prod_{l=1}^L P(s_l | s_{<l}, \mathbf{q}, \mathcal{I}, \theta)$, where $s_{<l} \triangleq \{s_1, s_2, \dots, s_{l-1}\}$. However, this formulation penalizes long sequences. Length normalized confidence (LNC) [13] fixes this issue by proposing the metric, $\hat{P}(\mathbf{s}|\mathbf{q}, \mathcal{I}, \theta) = \prod_{l=1}^L P(s_l | s_{<l}, \mathbf{q}, \mathcal{I}, \theta)^{1/L}$. LNC metric essentially normalizes the log probabilities by the length of the sequence. In this work, we mainly focus on designing a scoring function $f(\cdot)$ that rely on the signals from the models to form a UE estimate.

Selective Prediction A practical use case of UE methods is selective prediction task, where based on the UE function $f(\cdot)$, a decision function $g(\cdot)$ is used to determine whether system chooses to answer the question or abstains [4]. For the generated sequence \mathbf{s} by a VLM, selective system will output the generated sequence \mathbf{s} if $g(\mathbf{s}) = 1$, otherwise abstain \emptyset , where $g(\mathbf{s}) = \mathbb{I}\{f(\mathbf{s}) > \gamma\}$ given a threshold γ , \mathbb{I} being an indicator function. Threshold γ that provides best differentiation between the correct and incorrect generations is selected from the calibration dataset. $f(\cdot)$ can be any UE function, for example, LNC.

4. HARMONY

A well-calibrated model should maintain a consistent relationship between its predicted probabilities and the accuracy of its predictions. However, in open-ended generations, uncertainty estimation (UE) is inherently difficult due to factors such as length bias (variable-length output generation) [13] and semantic bias (where certain tokens hold more significance than others) [2], which are often implicit but significantly influence UE. Additionally, visual grounding presents another challenge, as Vision-Language Models (VLMs) are prone to linguistic biases, often producing overconfident responses driven more by textual cues than visual context [10, 24]. In such cases, analyzing shifts in the output distribution alone may not be enough. We need a signal of visual understanding from the model. Therefore, we hypothesize that combining hidden state representations from the model with token-level uncertainty offers a more comprehensive measure of reliability. While internal activations capture latent uncertainty, reflecting the model’s understanding of the visual context and its alignment of visual and textual information, output probabilities track confidence shifts during generation, offering deeper insights into the model’s uncertainty and trustworthiness.

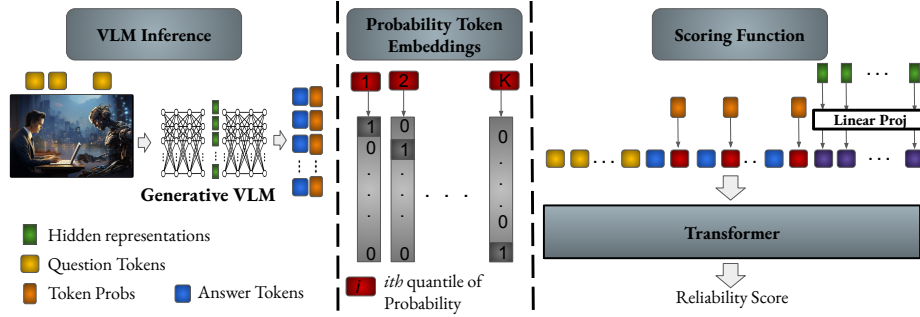


Figure 1. An illustration of calibration data collection phase (left), probability token embedding design (middle) and the scoring function architecture (right). Left subfigure represents the question tokens, generated sequence tokens and probabilities, and hidden representations at a specific layer. The next subfigure shows the orthogonal embedding vectors design of different probability quantiles. Right subfigure demonstrates how varied inputs are used to train a transformer like architecture, VisualBERT in our case to predict reliability score. It is worth-mentioning that the whole transformer architecture is fine-tuned on this task.

Scoring Function Let f be the scoring function that takes four inputs: the question $\mathbf{q} = (q_1, q_2, \dots, q_K)$, the generated response $\mathbf{s} = (s_1, s_2, \dots, s_L)$, the token probabilities $\mathbf{p} = (p_1, p_2, \dots, p_L)$, and the model’s hidden-states $\mathbf{H} = (\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_K, \dots, \mathbf{h}_{K+L})$ corresponding to \mathbf{q} and \mathbf{s} . Here, p_i denotes the probability of generated sequence token i , and each hidden state $\mathbf{h}_i \in R^N$ vector, where N represents the number of hidden units at a specific layer. It is worth-noting that the first K vectors in \mathbf{H} correspond to the question tokens, while the remaining L represent the generated tokens. We use the hidden states of the tokens right after the visual tokens, as they inherently encode cross-modal interactions, capturing information transferred from visual tokens to text tokens. Note that our inputs consist of varied nature i.e, texts, probabilities which are real numbers, and hidden states which are high dimensional vectors requiring a structured approach to model their inter-dependencies.

Input Mapping Given that \mathbf{H} , \mathbf{q} , \mathbf{s} and \mathbf{p} are sequential, we leverage the pretrained VisualBERT architecture, an extension of the encoder-based BERT model. Our choice is based on its ability to take text as well as high-dimensional visual features as inputs. Overall, it maintains two sets of embeddings E and F corresponding to text and visual features respectively. For text data, it tokenizes the input and maps each token to a set of embeddings, $e \in E$. Likewise, for vision context, model takes visual features corresponding to different regions of the images as an input, and assigns it an embedding $f \in F$.

VisualBERT is naturally well-suited for textual input. For hidden states, we project model’s hidden states to the space of model’s visual embeddings via linear projection, and use them as an input. To encode probability information, we leverage a third set of embedding which is inspired by work [22]. The key idea is that the probability range $[0,1]$ can be split into a fixed k partitions. For the given dimension d of input embedding, if p_i falls in the range of

r -th partition, the vector positions between $(r - 1) \times kd$ and $r \times kd$ are set to one while all other positions are set to zero. This allows representation of distinct probability ranges via orthogonal embedding vectors. At the input, we have question, followed by generated tokens and their corresponding probabilities, further followed by hidden representations of the question and the generated tokens. We augment the VisualBERT model at the output via linear layer that gives a single logit output. We employ binary cross-entropy loss, and use binary ground-truth label (of accuracy) as a target.

5. Results

Datasets We perform evaluation on two open-ended VQA datasets, A-OKVQA [16] and OKVQA [15]. Both datasets require reasoning over external knowledge and common sense alongside visual information. A-OKVQA comprises 17.1K train and 1.1K validation samples. We use the train split as a calibration dataset. Since OKVQA consists of only 9K questions in train split which makes a smaller calibration dataset. Therefore, we also include two beams alongwith one greedy search generation for every question, increasing the calibration data size to 36K. To evaluate our method, we use the validation splits provided with the original datasets as test sets. For training our scoring function, we further divide the calibration dataset between 80% train and 20% validation data partition for both datasets.

Models We evaluate our method on three open-sourced VLMs: LLaVA-7b, LLaVA-13b [12] and InstructBLIP [6]. We use a prompt of $\langle image \rangle question, please provide a single word or short sentence answer. ASSISTANT:$.

Performance Evaluation Since the task is open-ended text generation, following previous work [19], we use LAVE_{GPT-3.5} [14] as an evaluator. LAVE uses a large language model to estimate the semantic similarity of each predicted answer to the 10 crowdsourced answers in the benchmark. We regard score greater than 0 (one or more matches) as correct and 0 (no-match) as incorrect. Following previ-

ous UE works on auto-regressive models [2, 9, 13, 22], we use AUROC (Area Under the Receiver Operating Characteristic) and prediction rejection ratio (PRR) as our evaluation metric. The score range for AUROC is 0.5 (random) to 1 (perfect), whereas the PRR ranges from 0 (random) to 1 (perfect). For selective prediction, we report Effective reliability, Coverage at 10% risk, and area under the risk-coverage curve (AURAC) [21].

Table 1. AUROC and PRR scores on A-OKVQA and OKVQA dataset

UE Method	LLaVA - 7b		LLaVA - 13b		Instruct-BLIP	
	AUROC(%)	PRR(%)	AUROC(%)	PRR(%)	AUROC(%)	PRR(%)
Length-Normalized Confidence	74.55	61.45	77.50	67.04	74.13	56.60
First Token Confidence	69.39	33.16	72.96	41.35	75.09	58.47
Self-Eval Confidence	71.53	54.48	63.04	54.43	76.12	62.75
Entropy	61.38	35.65	67.57	49.23	54.15	30.15
Semantic Entropy	78.39	68.48	80.83	69.89	73.72	52.20
Cluster Entropy	69.87	52.27	68.90	50.89	71.00	51.11
MSF	78.66	67.01	77.64	67.07	79.93	67.31
LARS	79.90	68.95	81.46	73.70	80.07	68.31
HARMONY [Ours]	83.99	75.05	83.72	77.09	81.73	72.03
	(+4.09)	(+6.10)	(+2.26)	(+3.36)	(+1.66)	(+3.72)
Length-Normalized Confidence	74.44	60.64	75.40	63.31	73.45	58.90
First Token Confidence	69.85	45.55	71.14	49.13	73.32	58.70
Self-Eval Confidence	67.3	54.13	70.55	51.16	76.45	61.33
Entropy	58.22	31.43	64.76	46.63	58.08	36.23
Semantic Entropy	71.45	49.15	71.89	50.38	69.93	47.18
Cluster Entropy	64.18	35.18	63.29	34.81	65.85	38.32
MSF	74.01	61.23	75.24	62.13	74.21	53.93
LARS	78.29	66.45	76.82	63.17	79.51	66.57
HARMONY [Ours]	80.91	71.49	79.57	69.25	80.81	69.08
	(+2.46)	(+5.04)	(+2.57)	(+6.08)	(+1.30)	(+2.51)

Baselines We compare our method to various black-box approaches such as LNC [13], first token confidence [23], self-eval [19], Entropy, Semantic Entropy, and Cluster Entropy [5]. We also consider supervised training functions, MSF (multimodal-selection function) [21] that trains an MLP on hidden representations of the base model, and LARS [22] that trains a transformer on the output probabilities.

Training Strategy For every trainable scoring method, we perform hyper-parameter tuning of learning rate over $\{5e-4, 5e-5, 5e-6\}$. We use AUROC as our best model checkpoint selection criteria for all the methods. Further, for all the trainable methods, we use 20 epochs, and early stopping; i.e, if validation auoc does not improve for 1K training steps, we stop the training. For MSF, we follow the official implementation, and keep other parameters such as optimizer (AdamW), learning rate scheduler (Warmup Cosine Scheduler), batch size the same. We use the same optimizer and learning rate scheduler for our method and LARS. Further, we use batch size of 32 for LARS and HARMONY.

UE Performance: Table 1 presents the results of comparison of our method with state-of-the-art UE baselines. Among black-box approaches, we find semantic entropy to outperform all other baselines on LLaVA models. However, self-eval performs better than semantic entropy for the InstructBLIP model. Among trainable functions, MSF improves upon the LNC consistently across all datasets, and all models. However, LARS outperforms MSF consistently highlighting the significance of computing inter-token dependencies. Further, our proposed method HARMONY consistently outperforms LARS and MSF achieving upto

4% increase in AUROC scores and 6% increase in the PRR scores. Note that we ablate over every fourth layer for both MSF, and our method. We report the best performing layer results in Table 1. We observe that for LLaVa-7b and 13b models, inner layers (layer 16 and layer 22) yield the best AUROC performance. Further, for InstructBLIP, we find the outer-most layer performs the best.

Selective Prediction Performance: We compare our method to other trainable baselines on the selective prediction task for two datasets in Table 2. 1) Our proposed method consistently achieves higher coverage at 10% risk. 2) We evaluate our method on the effective reliability metric, which represents a better tradeoff between coverage and risk due to a penalty on the incorrectly covered question. For the comparison, we select a threshold based on the validation split of calibration set. We observe that our method either performs similar or outperforms other methods achieving up to 2.5% higher score. As an example, we present some sample questions from the A-OKVQA dataset and LLaVa-7b model generations in Figure 2. It shows that training on either output distributions or hidden representations alone can lead to contradictory or consistent but incorrect decisions, leveraging both simultaneously results in more reliable decision functions.

Table 2. AURAC, Coverage at risks (10%) and Effective Reliability (ER) (cost=1) scores on A-OKVQA and OKVQA dataset

UE Method	ER(%) \uparrow	C@R=10%(\uparrow)	ER(%) \uparrow	C@R=10%(\uparrow)	ER(%) \uparrow	C@R=10%(\uparrow)
MSF	49.17	43.75	53.19	50.56	38.15	32.66
LARS	49.78	50.65	53.71	61.83	37.73	31.27
HARMONY [Ours]	52.31	60.61	55.90	64.80	38.25	36.77
MSF	49.21	40.21	47.21	50.21	27.88	06.01
LARS	50.13	54.45	45.94	52.97	32.79	23.94
HARMONY [Ours]	51.20	56.52	52.20	60.03	32.80	29.47

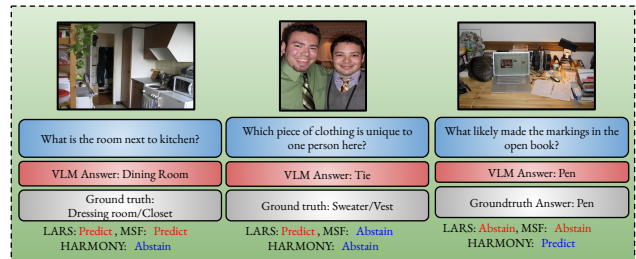


Figure 2. Some examples of selective prediction decision function's estimates on the A-OKVQA dataset with LLaVa-7b model, where LARS and MSF make either contradictory or same but wrong decisions, however, our approach makes the right prediction for each of these examples.

6. Conclusion

We introduce a novel uncertainty estimation method HARMONY for Vision-Language Models that combines hidden activation representations with output token probabilities. By jointly leveraging model internal states and output beliefs in a sequential fashion, our proposed framework provides a more holistic reliability assessment, complementing probability-based and representation-based approaches.

References

- [1] Dragan Ahmetovic, Daisuke Sato, Uran Oh, Tatsuya Ishihara, Kris Kitani, and Chieko Asakawa. Recog: Supporting blind people in recognizing personal objects. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020. 1
- [2] Yavuz Faruk Bakman, Duygu Nur Yaldiz, Baturalp Buyukates, Chenyang Tao, Dimitrios Dimitriadis, and Salman Avestimehr. Mars: Meaning-aware response scoring for uncertainty estimation in generative llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7752–7767, 2024. 2, 4
- [3] Corentin Dancette, Spencer Whitehead, Rishabh Maheshwary, Ramakrishna Vedantam, Stefan Scherer, Xinlei Chen, Matthieu Cord, and Marcus Rohrbach. Improving selective visual question answering by learning from your peers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24049–24059, 2023. 2
- [4] Ran El-Yaniv et al. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5), 2010. 2
- [5] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. 2, 4
- [6] Jiaying Huang, Jingyi Zhang, Kai Jiang, Han Qiu, and Shijian Lu. Visual instruction tuning towards general-purpose multimodal model: A survey. *arXiv preprint arXiv:2312.16602*, 2023. 3
- [7] Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine M. Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew Gordon Wilson. Large language models must be taught to know what they don’t know. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2
- [8] Zaid Khan and Yun Fu. Consistency and uncertainty: Identifying unreliable responses from black-box vision-language models for selective visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10854–10863, 2024. 1, 2
- [9] 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. 1, 2, 4
- [10] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024. 2
- [11] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2
- [12] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 3
- [13] Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*, 2021. 1, 2, 4
- [14] Oscar Mañas, Benno Krojer, and Aishwarya Agrawal. Improving automatic vqa evaluation using large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4171–4179, 2024. 3
- [15] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 3
- [16] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer, 2022. 3
- [17] Mohammad Shahab Sefehri, Zalan Fabian, Maryam Soltanolkotabi, and Mahdi Soltanolkotabi. Mediconfusion: Can you trust your ai radiologist? probing the reliability of multimodal medical foundation models. *International Conference on Learning Representations*, 2024. 1
- [18] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6649–6658, 2019. 2
- [19] Tejas Srinivasan, Jack Hessel, Tanmay Gupta, Bill Yuchen Lin, Yejin Choi, Jesse Thomason, and Khyathi Raghavi Chandu. Selective” selective prediction”: Reducing unnecessary abstention in vision-language reasoning. *CoRR*, abs/2402.15610, 2024. 2, 3, 4
- [20] 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 *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. 2
- [21] Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. Reliable visual question answering: Abstain rather than answer incorrectly. In *European Conference on Computer Vision*, pages 148–166. Springer, 2022. 1, 2, 4
- [22] Duygu Nur Yaldiz, Yavuz Faruk Bakman, Baturalp Buyukates, Chenyang Tao, Anil Ramakrishna, Dimitrios Dimitriadis, Jieyu Zhao, and Salman Avestimehr. Do not design, learn: A trainable scoring function for uncertainty estimation in generative llms. *CoRR*, abs/2406.11278, 2024. 2, 3, 4
- [23] Qinyu Zhao, Ming Xu, Kartik Gupta, Akshay Asthana, Liang Zheng, and Stephen Gould. The first to know: How token distributions reveal hidden knowledge in large vision-language models? In *European Conference on Computer Vision*, pages 127–142. Springer, 2024. 4
- [24] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu

Yao. Analyzing and mitigating object hallucination in large vision-language models. In *The Twelfth International Conference on Learning Representations*, 2024. [2](#)