Uncertainty Quantification for Multimodal Retrieval Augmented Generation

Simon Binz*
Radboud University
Nijmegen, The Netherlands
simon.binz@ru.nl

Heydar Soudani
Radboud University
Nijmegen, The Netherlands
heydar.soudani@ru.nl

Faegheh Hasibi
Radboud University
Nijmegen, The Netherlands
faegheh.hasibi@ru.nl

Abstract

Retrieval Augmented Generation (RAG) improves the question answering capabilities of Large Language Models (LLMs) by incorporating external knowledge and has recently been extended to multimodal settings through Vision-Language Models (VLMs) that integrate visual and textual information. Despite these advances, generated answers can still be incorrect or misleading. Uncertainty Quantification (UQ) methods aim to estimate the reliability of model outputs, but most existing approaches are designed for text-only models and perform poorly in multimodal RAG scenarios. A key challenge is capturing uncertainty arising from multiple stages of the pipeline, including retrieval, visual understanding, and generation. In this work, we show that modeling uncertainty using multimodal and retrieval-aware probability signals improves estimation in multimodal RAG systems. We introduce LeMUQ, a Learnable Multimodal UQ method that analyzes token probabilities under input modifications, such as removing modalities or retrieved context. By encoding these signals as probability tokens and processing them with a finetuned model, our approach captures interactions between modalities and retrieval. Experiments across datasets, retrievers, and VLMs show consistent improvements over baseline and finetuned UQ methods. Our proposed LeMUQ increases the AUROC metric by 0.038 on average. Additionally, our method shows strong generalization performance across different retrieval setups and datasets with mixed results when transferring across different VLMs. Our findings highlight the importance of modeling multimodal uncertainty and provide a step toward more reliable and safer multimodal RAG systems. Code is available on GitHub.[1]

CCS Concepts

• **Computing methodologies** → **Natural language generation**; • **Information systems** → **Question answering**; **Language models**.

Keywords

Uncertainty Quantification, Multimodal Retrieval Augmented Generation, Learnable Uncertainty Score

*Also with German Research Center for Artificial Intelligence (DFKI).
[1]https://github.com/informagi/LeMUQ/

ICTIR '26, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-2600-2/2026/07
https://doi.org/10.1145/3805713.3820431

ACM Reference Format:
Simon Binz, Heydar Soudani, and Faegheh Hasibi. 2026. Uncertainty Quantification for Multimodal Retrieval Augmented Generation. In *Proceedings of the 2026 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR) (ICTIR '26), July 25, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3805713.3820431

1 Introduction

Large Language Models (LLMs) have shown strong performance across a wide range of tasks [5, 21, 30, 51], such as question answering (QA), where queries are in natural language [19, 54]. However, many QA tasks require specific knowledge that is not fully captured during the training stage of LLMs [39, 48, 49]. Retrieval-Augmented Generation (RAG) has been proposed to address this issue by incorporating externally retrieved knowledge, thereby improving answer quality [20, 28, 44, 46, 56]. As tasks and models grow more complex, multimodal extensions such as Vision-Language Models (VLMs) have been developed for settings like Visual Question Answering (VQA), where questions are paired with images [29, 32, 35]. These images provide important visual cues, motivating extensions of standard RAG setups to multimodal scenarios that integrate visual information [1, 62].

Despite the improvements introduced by RAG systems and multimodal extensions, generated answers can still be incorrect or misleading [22, 39, 49]. This is particularly critical in domains such as medical or legal applications, where incorrect responses can have severe consequences [3, 8, 15]. To mitigate this issue, **Uncertainty Quantification (UQ)** methods aim to estimate the correctness of a model's response, where high uncertainty indicates a likely incorrect response and low uncertainty suggests correctness [12, 17, 18, 23, 61, 64].

Research gap. Existing research on UQ is mainly designed for text-only LLMs [3, 10, 14, 31, 37, 38, 48, 52, 55], with a limited number of approaches measuring uncertainty in VLMs [4, 26]. These standard UQ methods, even those designed for VLMs, often achieve suboptimal performance in multimodal RAG settings, where extra (non-)textual information is introduced [58]. Prior work has shown that many UQ methods degrade in performance when non-parametric knowledge is incorporated [43, 50, 52, 63]. This is particularly problematic in multimodal RAG, where the input prompt is not purely textual and retrieved supporting context can increase model confidence in ways that do not necessarily align with correctness, thereby degrading calibration and increasing false-positive confidence [45].

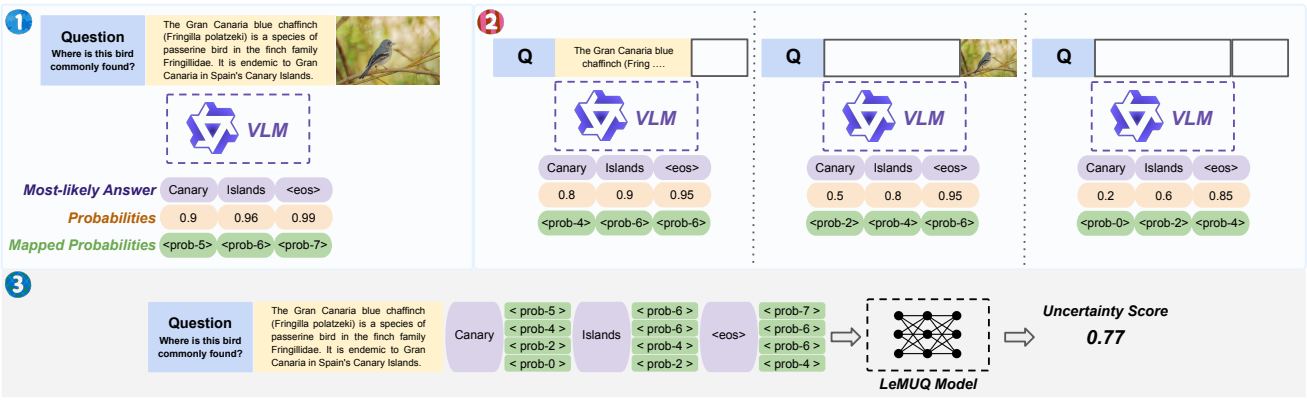

Figure 1: Overview of the LeMUQ pipeline. (1) A VLM receives the query, input image, and retrieved context, and generates a response as a sequence of tokens along with their token probabilities, which are discretized into quantile bins and represented as mapped token probabilities. (2) To account for different sources of uncertainty, the response is re-evaluated under multiple conditioning settings (without image, without retrieved context, and without both), producing alternative token probability sequences that are similarly converted into mapped probabilities. (3) The query, retrieved context, generated response tokens, and corresponding mapped token probabilities from all configurations are concatenated and passed to a finetuned RoBERTa-based LeMUQ model, which outputs a scalar score estimating the VLM's confidence in the generated response.

Challenge. In multimodal RAG settings, uncertainty can arise at different steps of the generation pipeline [62]. Within a single modality, inputs such as images or text may be ambiguous or noisy [22, 26]. Additional uncertainty may be introduced during the retrieval process, whether through text-only retrieval or multimodal retrieval pipelines such as image captioning followed by retrieval [62]. Further uncertainty can arise during the aggregation of information across modalities, as well as during the generation process itself [29, 58]. A UQ method for multimodal RAG needs to account for these sources of uncertainty and combine them in a calibrated manner to produce reliable uncertainty scores.

Method. In this work, we introduce a **Le**arnable **M**ultimodal **U**ncertainty **Q**uantification method (**LeMUQ**) that incorporates both multimodal and retrieval-based uncertainty. Our approach generates responses under varying conditions, including settings where specific modalities or retrieved context are selectively removed. The resulting response probabilities are encoded as special probability tokens and processed by a finetuned scoring model, which predicts an uncertainty score for each response.

Experiments. We conduct our experiments across multiple datasets, retrievers, and VLMs to evaluate the effectiveness of our proposed method in multimodal RAG settings. We first assess performance in a within-distribution setting, where the same retriever, dataset, and VLM are used during training and evaluation. Our results show that LeMUQ consistently outperforms both baseline and finetuned UQ methods, demonstrating the benefits of incorporating information about different modalities. In the in-distribution setting, our method achieves an average AUROC of 0.88, improving AUROC by 0.038 on average across two datasets, different VLMs, and retriever setups.

Generalizability. Previous research has shown that supervised methods often generalize poorly when the underlying data distribution shifts [42]. This limitation arises because such models tend to overfit to the specific characteristics of the training distribution, which leads to decreased performance on different validation and test distributions. To examine the generalizability of our method, we evaluate LeMUQ under out-of-distribution settings across three dimensions: varying the retriever, the dataset, and the VLM. Our findings indicate that LeMUQ maintains strong performance relative to the baseline and other finetuned methods. The performance gain is largely preserved when generalizing across different retrievers, generally surpasses both finetuned and baseline methods in dataset-transfer settings, and yields mixed results when transferring across different VLMs.

The main contributions of this paper are:

(1) We investigate the problem of UQ for multimodal RAG for the first time, to the best of our knowledge, and demonstrate the limitations of existing UQ methods when applied to a multimodal RAG setting.
(2) We propose LeMUQ, a UQ method for multimodal RAG that leverages probability signals under different input configurations to capture uncertainty arising from both retrieval and multimodal inputs.
(3) We conduct experiments across multiple datasets, retrievers, and VLMs, demonstrating that our method consistently improves uncertainty estimation and generalizes well across different settings.

2 Related Work

Multimodal RAG. Multimodal RAG extends the text-only pipelines of standard RAG by incorporating multiple modalities such as images or audio into both the retrieval and generation stages, enabling access to richer information sources and potentially improving answer quality [1]. VisRAG [62] proposes a retrieval framework that operates on document pages as unified multimodal units, allowing

the model to reason over jointly embedded visual and textual information rather than treating modalities separately. EchoSight [60] augments VLMs with external knowledge from Wikipedia articles through a two-stage retrieval process, first performing visual-only retrieval followed by multimodal reranking to identify relevant information. ReflectiVA [7] introduces reflective tokens in a multimodal LLM that enable the model to decide whether retrieval is necessary and to assess the relevance of retrieved information during generation. In this paper, we focus on a standard multimodal RAG setup in which the context, image, and user query are passed to a generative VLM.

Uncertainty Quantification for VLMs. In classical machine learning, uncertainty is typically characterized as a property of a model's predictive distribution given a particular input [31, 33]. Building on this notion, UQ aims to quantify the degree of variability or unpredictability in a model's outputs, independent of any single generated response [16, 25]. Most existing approaches to UQ have been developed in the context of question answering, where the LLM is treated as the sole source of uncertainty [31, 38]. More recently, however, this perspective has been extended to RAG, where uncertainty arises not only from the generative model but also from the retriever. In this setting, UQ has been studied by modeling the relationship between retrieved documents and generated responses [50], utility-based models [43], and fact-checking [11].

When moving to multimodal systems, additional sources of uncertainty emerge, particularly from visual inputs [11, 52]. Despite this, only a limited number of studies have explored UQ for VLMs. For instance, ReCoVERR [53] estimates uncertainty via self-evaluation, prompting the model to assess the reliability of its own outputs. In contrast, HARMONY [41] combines multiple signals, including generated tokens, output probabilities, and the model's internal assessment of visual understanding, to derive an uncertainty score. It compares the token distribution generated from the original image with that obtained after removing question-relevant visual evidence. The discrepancy between these distributions serves as an uncertainty measure, motivated by the intuition that unreliable predictions are less sensitive to such perturbations. Avestimehr et al. [2] estimate uncertainty by comparing a model's output distributions for an original image and a perturbed version where question-relevant visual evidence is removed. If the distributions remain similar despite removing key visual information, the response is likely unreliable. Nevertheless, existing work has not yet addressed UQ in the context of multimodal RAG, where uncertainty originates from three distinct components: the retriever, the generator, and the visual input. To bridge this gap, we propose a novel UQ framework that estimates uncertainty by analyzing model probabilities under diverse input configurations.

3 Preliminaries

Uncertainty Quantification for LLMs. UQ methods are typically categorized into two groups: *white-box approaches*, which leverage token-level probabilities and entropy, and *black-box approaches*, which rely solely on final outputs. Some of these approaches, such as PE [38], SE [27], and Eccentricity [31], rely on multiple generations of responses, where the core idea is to use the diversity among generated outputs as a proxy for uncertainty. In contrast, another family of approaches operate on a single generation, such as Confidence [59], which estimates uncertainty by incorporating token probabilities from a single output. In this work, we focus on white-box probability-based methods and use a single generation to compute the uncertainty score.

Probability-based UQ methods are based on computing the probability of the generated answer for a given question [59]. Formally, given an LLM parameterized by θ and an input query q, the sequence probability is defined as:

$$P(r \mid q, \theta) = \prod_{i=1}^{N} P\left(r_i \mid r_{<i}, q; \theta\right), \tag{1}$$

where $r_{<n}$ denotes the tokens generated before the token r_n. This probability can be directly used as the confidence score of the single most-likely generated answer [3, 9, 59], or it can be used to compute the *entropy* of multiple sampled answers using a Monte Carlo approximation [27, 38].

It has been shown that the sequence probability in Eq. (1) is biased against longer generations [38]. To address this issue, length-normalized scoring is introduced as:

$$P_{\ln}(r \mid q, \theta) = \prod_{i=1}^{N} P\left(r_i \mid r_{<i}, q; \theta\right)^{\frac{1}{N}}.$$

MARS [3] and TokenSAR [9] further improve this scoring function by incorporating the semantic contribution of individual tokens. Although their approaches differ in how token importance is modeled, they can be generalized using the following formulation:

$$P_{\text{me}}(r \mid q, \theta) = \prod_{i=1}^{N} P\left(r_i \mid r_{<i}, q; \theta\right)^{w(r,q,N,i)},$$

where $w(r, q, N, i)$ denotes the weight assigned to the i-th token. These scoring functions assign higher weights to tokens that are central to the answer and are calibrated such that they correlate with the likelihood of error of the final response. Finally, LARS [59] enhances this calibration by directly learning the scoring function from data.

Learnable Response Scoring (LARS). The goal of LARS [59] is to develop a learnable scoring function that captures the semantic contributions of tokens with respect to the query, accounts for biased probabilities, models dependencies between tokens, and incorporates other factors that may not be immediately apparent but are crucial for UQ. The key design choice in LARS is feeding probability information into a transformer model, even though this information is a single real value per token. Formally, the LARS function is defined as:

$$P_{\text{LARS}}(r|x, \theta) = \mathcal{S}_\phi\left(q, \{\langle r_i, \tilde{p}_i\rangle\}_{i=1}^{N}; \theta\right), \tag{2}$$

where S_ϕ is a trainable scoring function that takes input q and a set of response tokens r_i and the corresponding probability tokens $\tilde{p}_i$. To train the scoring function S_ϕ, LARS finetunes a pretrained transformer model (specifically RoBERTa-base [34]). To map real-valued probabilities $\tilde{p}_i$ of answer tokens r_i to high-dimensional vector representations for RoBERTa, LARS partitions the probability range $[0, 1]$ into k partitions. Given that the transformer model has an input dimension d, if p_i, falls within the r-th partition, LARS sets its vector positions from $(r-1) \times \frac{d}{k}$ and $r \times \frac{d}{k}$ to 1, while

the remaining positions are set to 0. With this encoding strategy, LARS represents distinct probability ranges as orthogonal vectors in a high-dimensional space. The final uncertainty score can be obtained by taking the negative of the score in Eq. (2) for a single generation, or by combining scores from multiple generations, e.g., using entropy-based functions. Although the final uncertainty score can be obtained either from a single generation or from multiple generations, we follow Yaldiz et al. [59] and use the most-likely generation score, which outperforms multi-generation alternatives.

4 Methodology

In this section, we define the problem of UQ for multimodal RAG and present our approach, **LeMUQ**, a Learnable Multimodal UQ method, building on LARS [59], which was originally developed for QA tasks with question-only textual inputs.

4.1 Problem Formulation

We consider the problem of UQ for responses generated by a VLM in a multimodal RAG setting. Let x denote the input to the VLM, consisting of: the textual query q, the input image I, and the retrieved textual context c; i.e., $x = \langle q, I, c \rangle$. Given this input, a VLM parameterized by θ generates a response sequence $r = (r_1, \ldots, r_N)$, where r_i is a token from a vocabulary $\mathcal{V}$, and N is the number of generated tokens.

The uncertainty score is evaluated with respect to the correctness of the response [12]. Let r^* denote the ground truth response. We compute the correctness of a generated response using exact match (EM), defined as a binary variable $z = \mathbb{I}\{r = r^*\}$, where $\mathbb{I}\{\cdot\}$ is the indicator function.

We aim to define an uncertainty scoring function:

$$\mathcal{S}_\phi : (\langle q, I, c \rangle, r) \mapsto \mathbb{R},$$

parameterized by ϕ, which assigns higher scores to correct responses than to incorrect ones. That is, for two responses $r^{(1)}$ and $r^{(2)}$,

$$\mathcal{S}_\phi(\langle q, I, c \rangle, r^{(1)}) > \mathcal{S}_\phi(\langle q, I, c \rangle, r^{(2)})$$

whenever $r^{(1)}$ is correct and $r^{(2)}$ is incorrect. The quality of the uncertainty estimates is then evaluated by measuring how well $\mathcal{S}_\phi$ separates correct from incorrect responses over a dataset; e.g., using metrics such as AUROC [13].

4.2 LeMUQ: Learnable Multimodal Uncertainty Quantification

The main idea behind LeMUQ is to model the contribution of different elements of the input to the uncertainty score, inspired by the self-consistency theory [51, 57]. Instead of scoring probabilities based only on the full input, including image and context, we condition on probability estimates under different modality configurations. In this way, the model can capture uncertainty arising from the visual input, the retrieved context, and their interaction, thereby implicitly learning an uncertainty estimate.

The overall method is illustrated in Figure 1. First, the full input, consisting of the question, image, and retrieved context, is fed to the VLM to produce a response as a sequence of tokens. For example, given the question "Where is this bird commonly found?", the model may generate the answer "Canary Islands." Along with the generated tokens, the model also produces token-level probabilities, which are discretized into quantile bins and represented as mapped token probabilities (cf. Sec. 3). Second, to disentangle different sources of uncertainty, the same response from the VLM is evaluated under three other input conditions: (1) removing the image, (2) removing the retrieved context, and (3) removing both image and context, using only the question. The token probabilities may shift across these conditions because supporting evidence or noise is removed. These alternative probability sequences are likewise converted into mapped token probabilities. Third, the question and the retrieved context are combined with the generated response and the corresponding mapped token probabilities from all conditions. This joint representation is then passed to a finetuned model, which outputs a single scalar uncertainty score reflecting the likelihood that the generated response is correct, following the Confidence UQ function.

Formally, the scoring function in LeMUQ incorporates four probabilities, where the first is the probability of response r being generated by the VLM π_θ given the input $\langle q, I, c \rangle$:

$$r, p \leftarrow \pi_\theta(q, I, c),$$

where $p = (p_1, ..., p_N)$ denotes the probabilities associated with the generation of each token r_i in the response $r = (r_1, ..., r_N)$. Next, we modify the input by removing the context, the image, or both, and compute the probability of generating the same response r. Specifically, we construct three variations of the input by:

(1) removing the image:

$$p^{\mathrm{q,c}} \sim P(r \mid q, c; \theta).$$

(2) removing the context:

$$p^{\mathrm{q,\,I}} \sim P(r \mid q, I; \theta).$$

(3) removing both the image and the context:

$$p^{\mathrm{q}} \sim P(r \mid q; \theta).$$

Following Yaldiz et al. [59], the four probability series, p, $p^{\mathrm{q,c}}$, $p^{\mathrm{q,I}}$, and p^{q}, are then passed through a probability mapping function to produce the corresponding values $\tilde{p}$, $\tilde{p}^{\mathrm{q,c}}$, $\tilde{p}^{\mathrm{q,\,I}}$, and $\tilde{p}^{\mathrm{q}}$. We define the LeMUQ score as:

$$p_{LeMUQ}(r|x, \theta) = \mathcal{S}_\phi\big(q, c, \{\langle r_i, \tilde{p}_i, \tilde{p}_i^{\mathrm{q,c}}, \tilde{p}_i^{\mathrm{q,I}}, \tilde{p}_i^{\mathrm{q}} \rangle\}_{i=1}^{N}; \theta\big)$$

This sequence is then fed into a RoBERTa[2] model, which is finetuned to predict the correctness of the response. The model outputs a single scalar score representing the estimated correctness of the generated response.

Training Strategy. Following LARS, let $\mathcal{S}_\phi$ denote the scoring function, which takes four inputs: the input query q, the retrieved context c , the generated sequence $r = (r_1, \ldots, r_N)$, and the corresponding mapped probability vector, $\tilde{p} = (\tilde{p}_1, \tilde{p}_1^{\mathrm{q,c}}, \tilde{p}_1^{\mathrm{q,I}}, \tilde{p}_1^{\mathrm{q}}, \ldots, \tilde{p}_N, \tilde{p}_N^{\mathrm{q,c}}, \tilde{p}_N^{\mathrm{q,I}}, \tilde{p}_N^{\mathrm{q}},)$, where $\tilde{p}_i, \tilde{p}_i^{\mathrm{q,c}}, \tilde{p}_i^{\mathrm{q,I}}, \tilde{p}_i^{\mathrm{q}}$ represent the mapped probability tokens of token r_i. The function $\mathcal{S}_\phi$ outputs a real-valued score o. We make $\mathcal{S}_\phi$ directly learnable using supervised data. To this end, we construct a calibration set to train the scoring function $\mathcal{S}_\phi$, parameterized by ϕ. This calibration set consists of 5-tuples: the

[2] docs/transformers/model_doc/roberta

Table 1: Within-distribution AUROC performance across EVQA and InfoSeek for LLaVA1.5-7B and Qwen3-VL-4B. BM25, EVAC., and BM25+MLM are used as retrieval models. LARS$_{Ret}$ uses the finetuned LARS model matched to the corresponding retrieval setting. The green values report the gain or loss relative to LARS$_{Ret}$. Superscripts † and ‡ denote statistically significant differences according to the DeLong test ($p < 0.05$), compared to LARS$_{Base}$ and LARS$_{Ret}$, respectively.

LLM	LLaVA1.5-7B						Qwen3-VL-4B						Avg.
Retriever	BM25		EVAC.		BM25+MLM		BM25		EVAC.		BM25+MLM		
	EVQA	InfoSeek	EVQA	InfoSeek	EVQA	InfoSeek	EVQA	InfoSeek	EVQA	InfoSeek	EVQA	InfoSeek	
Accuracy	0.128	0.118	0.138	0.102	0.163	0.147	0.133	0.142	0.135	0.127	0.169	0.140	0.137
PE	0.750	0.756	0.803	0.775	0.719	0.729	0.721	0.770	0.753	0.813	0.692	0.749	0.752
P(True)	0.639	0.645	0.677	0.645	0.641	0.614	0.604	0.734	0.596	0.714	0.584	0.740	0.653
Ecc.	0.732	0.745	0.752	0.755	0.699	0.717	0.664	0.667	0.728	0.705	0.634	0.629	0.702
Img. Per.	0.415	0.414	0.426	0.409	0.383	0.397	0.347	0.294	0.348	0.321	0.389	0.291	0.369
LARS$_{Base}$	0.689	0.762	0.709	0.752	0.687	0.752	0.673	0.765	0.767	0.825	0.607	0.771	0.730
LARS$_{Ret}$	0.847	0.829	0.848	0.823	0.846	0.791	0.840	0.858	0.877	0.890	0.816	0.845	0.843
LeMUQ	**0.855**†	**0.833**†	**0.861**†	**0.854**†‡	**0.871**†‡	**0.853**†‡	**0.923**†‡	**0.888**†‡	**0.912**†‡	**0.913**†‡	**0.902**†‡	**0.898**†‡	**0.880**
	(+0.008)	(+0.004)	(+0.012)	(+0.031)	(+0.025)	(+0.063)	(+0.083)	(+0.030)	(+0.034)	(+0.022)	(+0.086)	(+0.053)	(+0.038)

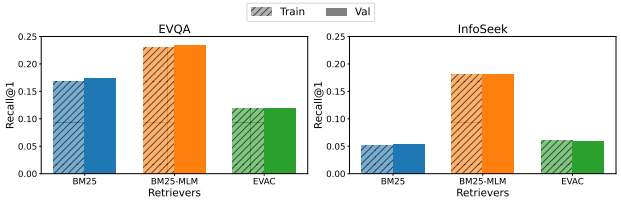

Figure 2: Recall@1 performance of retrievers on the train and validation sets.

input query q, the retrieved context c, the generated sequence r, the mapped probability vector $\tilde{p}$, and a binary ground truth label g. The label g indicates whether r is a correct response to q. To optimize the parameters of $\mathcal{S}_{\phi}$, we employ the binary cross-entropy loss.

5 Experimental Setup

Datasets. As multimodal RAG requires ⟨question, image, answer⟩ triplets that also rely on external knowledge, we evaluate LeMUQ on two vision-text datasets: Encyclopedic VQA (EVQA) [40] and InfoSeek [6]. Both datasets come with a controlled knowledge base generated from Wikipedia articles containing ground truth documents for each question.

EVQA focuses on encyclopedic knowledge about entities such as landmarks and species, and includes various question types. Following [60], we use only the automatically generated questions and ensure each question–entity pair appears exclusively in either train or test set. Due to limitations in the original splits, we subsample the training data to create 15,000 training and 2,000 test examples, ensuring no overlap between the train and test splits.

InfoSeek similarly targets questions requiring external knowledge. Since its required articles are covered by the EVQA knowledge base, we use the same source for retrieval. In contrast to EVQA, only the ground truth Wikipedia article is given for each question. To identify relevant sections, we apply a MiniLM model trained on MS MARCO[3] that ranks sections using the question and answer. From the training split, we sample 15,000 instances for training, ensuring that all unique question–entity pairs are preserved. For evaluation, we sample 2,000 instances from the validation split.

Retrieval Models. Following [50], we use a BM25 retriever [47] and a two-stage reranking pipeline which combines BM25 with the MS MARCO cross-encoder reranker, which is applied to the top 1000 retrieved documents (BM25+MiniLM). In addition to text-only retrieval methods, we also perform image-to-image retrieval [7] by encoding query images with EVA-CLIP-8B[4] and matching them to Wikipedia image embeddings to retrieve relevant articles. We then use Contriever[5] to select the most relevant section given the question (EVAC.). We also consider an oracle retrieval setting Doc^{+} (gold document), and no retrieval setting $Doc^{\times}$ (no document).

On InfoSeek, standard BM25-based methods yield low recall. To address this, we adopt a query rewriting strategy inspired by [36], where GPT-5-mini[6] generates a query combining the visual entity and question intent. This rewritten query is used for retrieval, and for BM25+MiniLM, we perform the second-stage retrieval using the identified entity in the image instead of the abstract entity in the original question. These modifications improve retrieval quality and ensure comparability across datasets.

VLMs. To generate responses, we use two VLMs: LLaVA1.5-7B[7] and Qwen3-VL-4B.[8] For response generation, we set the `do_sample` parameter to `False`. For sampling responses for UQ methods, we set the number of samples to 10, enable `do_sample`, and set both temperature and top_p to 1.0. We provide the top-1 retrieved section as context to the VLM following the simplified setup in [7]. Following prior work in RAG [39, 49], which limits the effective input by restricting retrieved content [28], we constrain the input

[3] cross-encoder/ms-marco-MiniLM-L-6-v2
[4] BAAI/EVA-CLIP-8B
[5] facebook/contriever
[6] https://developers.openai.com/api/docs/models/gpt-5-mini
[7] llava-hf/llava-1.5-7b-hf
[8] Qwen/Qwen3-VL-4B-Instruct

Table 2: Generalizability of LeMUQ across different retrieval models. The UQ model is trained using one retrieval model and tested with another retrieval model. Superscripts † and ‡ denote statistically significant differences according to the DeLong test ($p < 0.05$), compared to LARS$_{Base}$ and the corresponding LARS model in the same row block, respectively.

VLM	EVQA					InfoSeek				
	Doc$^{\times}$	BM25	EVAC.	BM25+ MLM	Doc^{+}	Doc$^{\times}$	BM25	EVAC.	BM25+ MLM	Doc^{+}
LLaVA1.5-7B										
LARS$_{Base}$	0.745	0.689	0.709	0.687	0.674	0.773	0.762	0.752	0.752	0.690
LARS$_{BM25}$	0.847	—	0.804	0.835	0.701	0.909	—	0.845	0.797	0.779
LeMUQ$_{BM25}$	0.823†	—	0.828†‡	0.841†	0.739†‡	0.824‡	—	0.819†	0.810†	0.795†
LARS$_{EVAC}$	0.867	0.807	—	0.812	0.813	0.896	0.825	—	0.788	0.783
LeMUQ$_{EVAC}$	0.832†‡	0.799†	—	0.791†	0.796†‡	0.866†	0.828†	—	0.814†‡	0.800†
LARS$_{\text{BM25+MLM}}$	0.827	0.844	0.799	—	0.716	0.917	0.822	0.849	—	0.762
LeMUQ$_{\text{BM25+MLM}}$	0.821†	0.860†	0.840†‡	—	0.755†‡	0.802‡	0.823†	0.824†	—	0.785†‡
Qwen3-VL-4B										
LARS$_{Base}$	0.709	0.673	0.767	0.607	0.629	0.751	0.765	0.825	0.771	0.772
LARS$_{BM25}$	0.869	—	0.893	0.811	0.841	0.916	—	0.893	0.845	0.854
LeMUQ$_{BM25}$	0.898†‡	—	0.889†	0.904†‡	0.881†‡	0.879†‡	—	0.891†	0.896†‡	0.871†
LARS$_{EVAC}$	0.879	0.810	—	0.777	0.815	0.913	0.854	—	0.843	0.852
LeMUQ$_{EVAC}$	0.857†‡	0.878†‡	—	0.857†‡	0.903†‡	0.894†‡	0.894†‡	—	0.906†‡	0.888†‡
LARS$_{\text{BM25+MLM}}$	0.882	0.846	0.897	—	0.856	0.914	0.854	0.887	—	0.862
LeMUQ$_{\text{BM25+MLM}}$	0.821†‡	0.913†‡	0.894†	—	0.898†‡	0.873†‡	0.882†‡	0.899†	—	0.874†

length to fit within hardware limits. In particular, we truncate the context to a maximum of approximately 500 tokens.

Baselines. We use the following baseline UQ methods as discussed in Sections 2 and 3: Ecc [31], PE [38], P(True) [24], and the pre-trained LARS (LARS$_{Base}$). To account for visual uncertainty, we include the image perturbation (Img. Per.) method [2], described in Section 2. We instantiate this approach using a black image as the perturbed version and compute the distance on the top-1 token distribution, as this was shown to achieve the best performance [2]. In addition to the baseline methods, we finetune LARS and our proposed LeMUQ. For response scoring, we use a single-generation setup (denoted as Confidence in [59]), as it achieves the best performance among probability-based UQ methods in [59].

Evaluation Metrics. To evaluate the responses, we use exact matching as in [50]. Similarly, we use AUROC as an evaluation metric for the UQ methods, as it is invariant to the distribution of correct and incorrect answers.

Finetuning Setup. To finetune LARS and LeMUQ, we follow the training pipeline of [59]. For each ⟨question, image, answer⟩ triplet, we retrieve context using one of the previously defined retrievers and generate up to five responses with the VLM, ensuring that the most likely response is included. Each generated response is labeled using exact matching against the ground-truth answer. As the scoring model, we use a RoBERTa-base model initialized with the finetuned LARS weights from [59] for both LARS and LeMUQ. We train the models to predict the correctness of the generated response using binary cross-entropy with a learning rate of 5×10^{-6} for three epochs. 10% of the training data is used for validation. Training was conducted on RTX A6000, H100, and H200 GPUs, with training times ranging from 10 to 27 hours for LeMUQ and 3–8 hours for LARS, depending on the retriever, VLM, and hardware used.

6 Results

We present a set of experiments that address the following research questions: **RQ1**: How does our proposed method perform compared to baseline UQ methods when evaluated within-distribution? **RQ2**: How well does LeMUQ generalize to out-of-distribution data? **RQ3**: What is the impact of each probability component in LeMUQ on its overall performance?

6.1 LeMUQ Performance

RQ1 evaluates the performance of LeMUQ compared to other UQ methods in a within-distribution setting. Figure 2 presents the performance of the retrievers on both the training and validation sets. We report Recall@1, as only the top-ranked document is passed to the generator model. The BM25+MLM model achieves the best performance on both datasets. For the InfoSeek dataset, BM25 and EVAC perform similarly; however, BM25 outperforms EVAC on the EVQA dataset. The relatively lower performance of EVAC can be attributed to its two-stage retrieval pipeline, which first operates at the article level and then at the section level. As a result, the overall performance is constrained by the effectiveness of the initial article-level retrieval step. In summary, we conduct our experiments using different types of retrievers with varying levels of performance.

Table 3: Generalizability of LeMUQ across datasets. The UQ model is trained on one dataset and tested on another dataset. Superscript ‡ denotes a statistically significant difference according to the DeLong test ($p < 0.05$) compared to LARS$_{Ret}$ in the same column.

Setup	EVQA → InfoSeek			InfoSeek → EVQA		
Ret.	**BM25**	**EVAC.**	**BM25+ MLM**	**BM25**	**EVAC.**	**BM25+ MLM**
LLaVA1.5-7B						
LARS$_{Ret}$	0.632	0.803	0.621	0.618	0.700	0.597
LeMUQ	0.768‡	0.812	0.721‡	0.646	0.720	0.657‡
Qwen3-VL-4B						
LARS$_{Ret}$	0.813	0.845	0.805	0.718	0.754	0.601
LeMUQ	0.820	0.839	0.852‡	0.778‡	0.830‡	0.755‡

Table 4: Generalizability of LeMUQ across VLMs. The UQ model is trained on one VLM and tested on another VLM. Superscript ‡ denotes a statistically significant difference according to the DeLong test ($p < 0.05$) compared to LARS$_{Ret}$ in the same column.

Setup	LLaVA → Qwen3			Qwen3 → LLaVA		
Ret.	**BM25**	**EVAC.**	**BM25+ MLM**	**BM25**	**EVAC.**	**BM25+ MLM**
EVQA						
LARS$_{Ret}$	0.812	0.786	0.790	0.761	0.776	0.729
LeMUQ	0.842‡	0.824‡	0.801	0.716‡	0.695‡	0.718
InfoSeek						
LARS$_{Ret}$	0.757	0.786	0.748	0.805	0.808	0.775
LeMUQ	0.811‡	0.838‡	0.852‡	0.678‡	0.715‡	0.669‡

Table 1 presents the results on two datasets and two VLMs. By *within-distribution*, we mean that the model is trained and evaluated using the same retrieval model; in other words, the documents used for training and evaluation are obtained from the same retriever. Among the baselines, PE, P(True), Ecc., and LARS$_{Base}$ perform reasonably well and are also robust to changes. Img. Per. fails to predict uncertainty properly, consistently returning values below 0.5. This indicates that Img. Per. is not generalizable to multimodal RAG settings. By finetuning the standard LARS model for a specific dataset, VLM, and retriever, LARS$_{Ret}$ outperforms baseline methods.

Finally, our proposed method, LeMUQ, outperforms all baselines across all settings by a significant margin. Overall, it increases the AUROC by 0.038 on average compared to a finetuned LARS scoring function. The improvement achieved by LeMUQ is higher in the Qwen3 model, with an average improvement of 0.051, compared to the improvement observed in the LLaVA1.5 model, which is on average 0.024. When comparing across datasets, our method consistently outperforms the baselines across all types of retrievers, including sparse, reranking, and image-based retrievers. However, the performance of the UQ methods is higher for the EVAC retriever on average across both datasets and models. These findings suggest that incorporating the probability of VLM generation under different input configurations, and training a model to capture these signals, improves uncertainty estimation in multimodal RAG systems.

6.2 Out-of-Domain Generalizability

RQ2 evaluates the out-of-distribution (OOD) generalizability of LeMUQ across three settings: retriever, dataset, and VLM. The main objective is to assess whether our proposed scoring function generalizes effectively to these different settings.

Retriever Generalizability. Table 2 presents the results of LeMUQ when the model is trained on contexts retrieved by one retriever and evaluated on contexts obtained from different retrievers. The rows correspond to the retrievers used during training, while the columns represent the retrievers used at evaluation time. In addition to evaluating with BM25, BM25+MLM, and EVAC, we also consider two special settings: one without any context in the input ($Doc^{\times}$), and another where the input contains gold context (Doc^{+}).

For LLaVA1.5 on the EVQA dataset, when training is performed using BM25 or BM25+MLM, LeMUQ consistently outperforms both LARS$_{Base}$ and LARS$_{Ret}$ across BM25, EVAC, BM25+MLM, and Doc^{+}. However, when training is conducted using EVAC, LARS$_{Ret}$ surpasses LeMUQ. Additionally, LARS$_{Ret}$ performs better than LeMUQ in the $Doc^{\times}$ setting. This suggests that, under our training approach, LeMUQ is less effective when no context is provided, although it still significantly outperforms LARS$_{Base}$. For InfoSeek, LARS$_{Ret}$ only outperforms LeMUQ in the $Doc^{\times}$ setting; in all other configurations, LeMUQ shows a significant advantage over the baselines.

For Qwen3-VL, the results are largely consistent with our observations for LLaVA1.5. However, on the EVQA dataset, LeMUQ outperforms LARS$_{Ret}$ in the $Doc^{\times}$ setting when trained with BM25. Across both datasets, the generalization performance is significantly higher compared to what we observed with LLaVA1.5. For example, when training on EVQA and evaluating on BM25+MLM, the improvement is 0.093. Overall, these findings indicate that although our scoring function experiences a drop in performance in the absence of context, it still generalizes well across different retrievers compared to the baseline. This suggests that the scoring function is relatively insensitive to the choice of retriever.

Dataset Generalizability. Table 3 presents the results of dataset generalization. In this setup, the model is trained on one dataset and evaluated on a different dataset. In the table, the dataset on the left side of the arrow indicates the training dataset, while the one on the right side of the arrow represents the evaluation dataset. To isolate the effect of the dataset, we keep both the retriever and the VLM the same during training and evaluation.

On LLaVA1.5-7B, LeMUQ consistently outperforms the baseline across all setups. Moreover, when using BM25+MLM, LeMUQ significantly outperforms LARS$_{Ret}$. On Qwen3-VL-4B, the baseline performs better in only one out of six setups, which demonstrates the strong generalization ability of our model across datasets. Based on these observations, we conclude that the proposed scoring function performs well even when evaluated on datasets different from

Table 5: Ablation study AUROC performance across EVQA and InfoSeek for LLaVA1.5-7B and Qwen3-VL-4B. BM25, EVAC., and BM25+MLM are used as retrieval models. LeMUQ denotes the full method, and the remaining rows remove one component at a time. Superscripts † and ‡ denote statistically significant differences according to the DeLong test ($p < 0.05$), compared to LARS$_{Ret}$ and LeMUQ, respectively.

VLM	LLaVA1.5-7B						Qwen3-VL-4B						Avg.
Retriever	BM25		EVAC.		BM25+MLM		BM25		EVAC.		BM25+MLM		
	EVQA	InfoSeek	EVQA	InfoSeek	EVQA	InfoSeek	EVQA	InfoSeek	EVQA	InfoSeek	EVQA	InfoSeek	
Accuracy	0.128	0.118	0.138	0.102	0.163	0.147	0.136	0.142	0.132	0.128	0.171	0.140	0.137
LARS$_{Base}$	0.689	0.762	0.709	0.752	0.687	0.752	0.673	0.767	0.769	0.815	0.610	0.773	0.730
LARS$_{Ret}$	0.847	0.829	0.848	0.823	0.846	0.791	0.832	0.861	0.882	0.889	0.816	0.843	0.842
LeMUQ	0.855	0.832	**0.861**	**0.854**†	**0.871**†	0.853†	0.920†	0.889†	0.917†	0.909†	0.902†	0.901†	**0.880**
LeMUQ $-\tilde{p}$	0.841	0.826	0.852	0.825‡	0.859	0.844†	0.921†	0.903†‡	0.910†	0.906†	0.894†	**0.909**†	0.874
LeMUQ $-\tilde{p}^{q,I}$	0.856	0.845	0.851	0.829‡	0.840‡	0.829†‡	**0.924**†	0.891†	0.912†	0.906†	**0.907**†	0.893†	0.873
LeMUQ $-\tilde{p}^{q,c}$	0.852	0.818	0.852	0.849†	0.856	**0.862**†	0.923†	**0.906**†‡	**0.919**†	0.899	0.895†	0.905†	0.878
LeMUQ $-\tilde{p}^{q}$	**0.859**	**0.849**	0.851	0.848†	0.864	0.851†	0.906†‡	0.902†‡	0.916†	**0.911**†	0.886†‡	0.905†	0.879

those used during training, including those with different distributions. This suggests that the scoring function learns to effectively utilize probability information itself, rather than relying on dataset-specific distributions.

VLM Generalizability. Table 4 demonstrates the generalization ability of UQ methods across different VLMs. Similar to other generalization experiments, we fix both the dataset and the retriever during training and evaluation, and assess cross-model generalization from Qwen to LLaVA and vice versa.

When generalizing from LLaVA to Qwen3, LeMUQ consistently outperforms the baseline across both datasets and all retrievers, indicating that a model trained on LLaVA generalizes well to Qwen. In contrast, from Qwen3 to LLaVA, LARS performs better than our proposed method. This suggests that training on Qwen leads the scoring function to learn the probability distribution specific to the Qwen model, which does not transfer effectively to LLaVA. A possible explanation for this asymmetric transfer between VLMs is differences in their token-probability distributions. We observe that Qwen assigns high probability scores within a narrow range, whereas LLaVA produces a wider range of probability values for generated tokens. Consequently, when a model is trained on Qwen's narrow probability range and tested on LLaVA's broader range, its generalization is weaker, as many probabilities fall outside the range most frequently observed during training, making the modality-removal signals less reliable. In contrast, training on LLaVA exposes LeMUQ to a wider range of probability values, which transfer more effectively to Qwen's narrower, high-confidence distribution.

Overall, the results show that the VLM generalization capability of LeMUQ depends on the choice of training and evaluation backbone models.

6.3 Impact of Probability Components

LeMUQ consists of four probability components: $\tilde{p}$, $\tilde{p}^{q,c}$, $\tilde{p}^{q,I}$, and $\tilde{p}^{q}$, as introduced in Section 4.2. **RQ3** examines the contribution of each component to overall performance by performing an ablation study, where one component is removed at a time, followed by finetuning and evaluation using the remaining three components.

Table 5 summarizes the findings of the study. For LLaVA1.5-7B combined with BM25, the highest performance is achieved when p^{q} is excluded, followed by the configuration without $p^{q,I}$ across both datasets. In contrast, when using the EVAC retriever, LeMUQ with all components delivers the strongest results. In particular, on the InfoSeek dataset, removing $\tilde{p}$ or $p^{q,I}$ leads to a substantial decline in performance. For the BM25+MLM setting, LeMUQ again yields the best overall results on average; however, omitting $p^{q,I}$ results in a pronounced degradation.

On Qwen3-VL-4B with BM25, performance improves slightly when excluding $p^{q,I}$, $p^{q,c}$, and $\tilde{p}$, whereas removing p^{q} leads to a substantial decline. For EVQA, LeMUQ achieves the second-best results. Under the BM25+MLM setting, the highest scores are obtained by omitting $p^{q,I}$ for EVQA and $\tilde{p}$ for InfoSeek. Across most configurations, variants of LeMUQ consistently outperform LARS$_{Ret}$, highlighting the effectiveness and necessity of the proposed probability components.

Overall, LeMUQ delivers the strongest average performance, suggesting that each component contributes positively to the final outcome. Notably, removing $p^{q,I}$ and $\tilde{p}$ results in larger performance drops compared to excluding $p^{q,c}$ or p^{q}, indicating that these two components play a more critical role in achieving high performance.

7 Conclusions and Future Work

This paper investigates the problem of uncertainty quantification for multimodal RAG for the first time, to the best of our knowledge. It introduces a novel method for UQ called LeMUQ, designed specifically for the multimodal RAG setting. LeMUQ is a learnable scoring approach that analyzes token probabilities under various input modifications, such as removing modalities or retrieved context. Our experimental results demonstrate several key findings. First, we observe a consistent drop in performance for standard UQ methods as the context becomes more reliable, which is in line with previous research [50]. Furthermore, methods developed specifically for VLMs, such as [2], fail to produce meaningful results when contextual information is incorporated. Second, we observe that finetuned methods on VLM and dataset outputs significantly

outperform baseline methods, as well as methods trained for different models such as LARS [59]. This is evident in the substantial performance gap between both the finetuned LARS and our proposed method, and the base LARS model, which uses the pretrained weights provided by [59].

Third, within the in-distribution setting, our proposed LeMUQ consistently outperforms LARS. Despite incorporating additional probability tokens corresponding to unseen modality configurations (i.e., probabilities without image, without context, and without both), which could increase the data requirements, our model achieves better performance even when trained on the same amount of data. However, the increased model complexity, combined with the relatively limited training data, may explain the mixed results observed in generalization. Since LARS has fewer parameters and relies on fewer signals, it may require fewer training samples, which could explain why it outperforms LeMUQ in certain transfer scenarios, such as the transition from Qwen to LLaVA. These findings highlight that uncertainty in multimodal RAG systems is inherently complex, as it can arise at different stages of the pipeline. Methods that rely solely on image- or text-based signals, as seen in the baselines, fail to fully capture this complexity. Despite the improvements introduced by LeMUQ, it does not consistently surpass all baseline UQ methods under distribution shifts, indicating that robust generalization remains an open challenge.

Limitations

Our study has several limitations. First, like other white-box uncertainty quantification methods, LeMUQ requires access to token-level probability information, or logits, from the underlying VLM. In contrast, black-box UQ methods, which rely only on generated outputs, do not have this requirement and may therefore be easier to apply to proprietary or API-only models. This dependency may limit the applicability of LeMUQ in settings where token probabilities are unavailable or costly to obtain.

Second, we consider only a single retrieved context. Extending LeMUQ to incorporate multiple retrieved contexts may provide additional gains in accuracy of final responses. Furthermore, LeMUQ is computationally more expensive during training and inference compared to UQ methods that require single generation [59], as it models token probabilities multiple times under different input configurations. Finally, the retrievers analyzed in this work are relatively simple. The observed differences in performance between the EVAC setting and other retrieval setups suggest that the retrieval mechanism itself could have an impact on uncertainty estimation. Exploring this effect and their interactions with UQ methods is a promising avenue for future work.

Acknowledgments

This publication is part of the project LESSEN with project number NWA.1389.20.183 of the research program NWA ORC 2020/21 which is (partly) financed by the Dutch Research Council (NWO).

Prompt Details

Prompt: Generation for LLaVA 1.5-7B without context
USER: <image>
{question} Answer with a single word or a short sentence.
ASSISTANT:

Prompt: Generation for LLaVA 1.5-7B with context
USER: <image>
Context: {context}
{question} Answer with a single word or a short sentence.
ASSISTANT:

Prompt: Generation for Qwen3-VL-4B without context
{question} Answer with a single word or a short sentence.

Prompt: Generation for Qwen3-VL-4B with context
{context}
{question} Answer with a single word or a short sentence.

P(True) System Prompt
You are a helpful, respectful and honest question-answer evaluator. You will be given a question, some brainstormed ideas and a generated answer. Evaluate the generated answer as true or false considering the question and brainstormed ideas. Output "The generated answer is true" or "The generated answer is false".

Prompt: P(True) without context
USER: <image>
Question:{question}
Here are some ideas that were brainstormed:{ideas}
Generated answer:{generated_text}

Prompt: P(True) with context
USER: <image>
Context:{context}
Question:{question}
Here are some ideas that were brainstormed:{ideas}
Generated answer:{generated_text}

References

[1] Omar Adjali, Olivier Ferret, Sahar Ghannay, and Hervé Le Borgne. 2024. Multi-Level Information Retrieval Augmented Generation for Knowledge-based Visual Question Answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 16499–16513. https://doi.org/10.18653/v1/2024.emnlp-main.922

[2] Kiana Avestimehr, Emily Aye, Zalan Fabian, and Erum Mushtaq. 2025. Detecting unreliable responses in generative vision-language models via visual uncertainty. In *ICLR Workshop: Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI.*

[3] Yavuz Faruk Bakman, Duygu Nur Yaldiz, Baturalp Buyukates, Chenyang Tao, Dimitrios Dimitriadis, and Salman Avestimehr. 2024. 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)*. 7752–7767.

[4] Jinyeong Chae and Jihie Kim. 2022. Uncertainty-based Visual Question Answering: Estimating Semantic Inconsistency between Image and Knowledge Base. In *International Joint Conference on Neural Networks, IJCNN 2022, Padua, Italy, July 18-23, 2022*. IEEE, 1–9. https://doi.org/10.1109/IJCNN55064.2022.9892787

[5] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology* 15, 3 (2024), 1–45.

[6] Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023. Can Pre-trained Vision and Language Models Answer Visual Information-Seeking Questions?. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 14948–14968.

[7] Federico Cocchi, Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2025. Augmenting multimodal llms with self-reflective tokens for knowledge-based visual question answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 9199–9209.

[8] Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *Journal of Legal Analysis* 16, 1 (01 2024), 64–93. https://doi.org/10.1093/jla/laae003 arXiv:https://academic.oup.com/jla/article-pdf/16/1/64/58336922/laae003.pdf

[9] Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 5050–5063.

[10] Ekaterina Fadeeva, Maiya Goloburda, Aleksandr Rubashevskii, Roman Vashurin, Artem Shelmanov, Preslav Nakov, Mrinmaya Sachan, and Maxim Panov. 2025. Don't Throw Away Your Beams: Improving Consistency-based Uncertainties in LLMs via Beam Search. *CoRR* abs/2512.09538 (2025). https://doi.org/10.48550/ARXIV.2512.09538 arXiv:2512.09538

[11] Ekaterina Fadeeva, Aleksandr Rubashevskii, Dzianis Piatrashyn, Roman Vashurin, Shehzaad Dhuliawala, Artem Shelmanov, Timothy Baldwin, Preslav Nakov, Mrinmaya Sachan, and Maxim Panov. 2025. Faithfulness-aware uncertainty quantification for fact-checking the output of retrieval augmented generation. *arXiv preprint arXiv:2505.21072* (2025).

[12] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature* 630, 8017 (2024), 625–630.

[13] Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern recognition letters* 27, 8 (2006), 861–874.

[14] Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't Hallucinate, Abstain: Identifying LLM Knowledge Gaps via Multi-LLM Collaboration. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL*. Association for Computational Linguistics, 14664–14690.

[15] Oscar Freyer, Isabella Catharina Wiest, Jakob Nikolas Kather, and Stephen Gilbert. 2024. A future role for health applications of large language models depends on regulators enforcing safety standards. *The Lancet Digital Health* 6, 9 (2024), e662–e672. https://doi.org/10.1016/S2589-7500(24)00124-9

[16] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. 2023. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review* 56, Suppl 1 (2023), 1513–1589.

[17] James Harrison, John Willes, and Jasper Snoek. 2024. Variational Bayesian Last Layers. In *The Twelfth International Conference on Learning Representations, ICLR*. OpenReview.net.

[18] Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. 2024. Decomposing Uncertainty for Large Language Models through Input Clarification Ensembling. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net. https://openreview.net/forum?id=byxXa99PtF

[19] Mohanna Hoveyda, Jelle Piepenbrock, Arjen P. de Vries, Maarten de Rijke, and Faegheh Hasibi. 2026. OrLog: Resolving Complex Queries with LLMs and Probabilistic Reasoning. In *Advances in Information Retrieval*. Springer Nature Switzerland, 98–114.

[20] Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan O Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. (2025).

[21] Hideaki Joko and Faegheh Hasibi. 2026. FACE: A Fine-Grained Reference-Free Evaluator for Conversational Information Access. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR*.

[22] Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. 2025. Sufficient Context: A New Lens on Retrieval-Augmented Generation Systems. In *International Conference on Learning Representations (ICLR)*.

[23] Saurav Kadavath, Tom Conerly, Amanda Askell, Thomas Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zachary Dodds, Nova Dassarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, John Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom B. Brown, Jack Clark, Nicholas Joseph, Benjamin Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language Models (Mostly) Know What They Know. *ArXiv* abs/2207.05221 (2022). https://api.semanticscholar.org/CorpusID:250451161

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

[25] Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems* 30 (2017).

[26] Zaid Khan and Yun Fu. 2024. 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*. 10854–10863.

[27] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=VD-AYtP0dve

[28] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* 33 (2020), 9459–9474.

[29] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.

[30] I-Fan Lin, Faegheh Hasibi, and Suzan Verberne. 2026. LLMs Enable Bag-of-Texts Representations for Short-Text Clustering. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

[31] Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models. *Transactions on Machine Learning Research* (2024). https://openreview.net/forum?id=DWkJCSxKU5

[32] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems* 36 (2023), 34892–34916.

[33] Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. 2025. Uncertainty quantification and confidence calibration in large language models: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 6107–6117.

[34] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019). arXiv:1907.11692 http://arxiv.org/abs/1907.11692

[35] Zhenghao Liu, Chenyan Xiong, Yuanhuiyi Lv, Zhiyuan Liu, and Ge Yu. 2023. Universal Vision-Language Dense Retrieval: Learning A Unified Representation Space for Multi-Modal Retrieval. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. https://openreview.net/forum?id=PQOlkgsBsik

[36] Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 5303–5315.

[37] Nishanth Madhusudhan, Sathwik Tejaswi Madhusudhan, Vikas Yadav, and Masoud Hashemi. 2025. Do LLMs Know When to NOT Answer? Investigating Abstention Abilities of Large Language Models. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING*. Association for Computational Linguistics, 9329–9345.

[38] Andrey Malinin and Mark Gales. 2021. Uncertainty Estimation in Autoregressive Structured Prediction. In *International Conference on Learning Representations*. https://openreview.net/forum?id=jN5y-zb5Q7m

[39] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 9802–9822.

[40] Thomas Mensink, Jasper Uijlings, Lluis Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araujo, and Vittorio Ferrari. 2023. Encyclopedic vqa: Visual questions about detailed properties of fine-grained categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3113–3124.

[41] Erum Mushtaq, Zalan Fabian, Yavuz Faruk Bakman, Anil Ramakrishna, Mahdi Soltanolkotabi, and Salman Avestimehr. 2025. HARMONY: Hidden Activation Representations and Model Output-Aware Uncertainty Estimation for Vision-Language Models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 1654–1659. https://doi.org/10.1109/CVPRW67362.2025.00154

[42] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems* 32 (2019).

[43] Laura Perez-Beltrachini and Mirella Lapata. 2025. Uncertainty Quantification in Retrieval Augmented Question Answering. *arXiv preprint arXiv:2502.18108* (2025).

[44] Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. Measuring and Narrowing the Compositionality Gap in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP*.

[45] Zexuan Qiu, Zijing Ou, Bin Wu, Jingjing Li, Aiwei Liu, and Irwin King. 2025. Entropy-based decoding for retrieval-augmented large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 4616–4627.

[46] Mahta Rafiee, Heydar Soudani, Zahra Abbasiantaeb, Mohammad Aliannejadi, Faegheh Hasibi, and Hamed Zamani. 2026. Total Recall QA: A Verifiable Evaluation Suite for Deep Research Agents. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[47] Stephen E. Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* 3, 4 (2009), 333–389.

[48] Heydar Soudani. 2025. Enhancing Knowledge Injection in Large Language Models for Efficient and Trustworthy Responses. In *Proceedings of the 48th International ACM Conference on Research and Development in Information Retrieval, SIGIR*.

[49] Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. 2024. Fine Tuning vs. Retrieval Augmented Generation for Less Popular Knowledge. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, SIGIR-AP 2024*. 12–22.

[50] Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. 2025. Why Uncertainty Estimation Methods Fall Short in RAG: An Axiomatic Analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*. 16596–16616.

[51] Heydar Soudani, Roxana Petcu, Evangelos Kanoulas, and Faegheh Hasibi. 2026. A Survey on Recent Advances in Conversational Data Generation. *ACM Comput. Surv.* 58 (4 2026). Issue 10. https://doi.org/10.1145/3795686

[52] Heydar Soudani, Hamed Zamani, and Faegheh Hasibi. 2026. Uncertainty Quantification for Retrieval-Augmented Reasoning. (2026).

[53] Tejas Srinivasan, Jack Hessel, Tanmay Gupta, Bill Yuchen Lin, Yejin Choi, Jesse Thomason, and Khyathi Chandu. 2024. Selective "selective prediction": Reducing unnecessary abstention in vision-language reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*. 12935–12948.

[54] Prashant Upadhyay, Rishabh Agarwal, Sumeet Dhiman, Abhinav Sarkar, and Saumya Chaturvedi. 2024. A comprehensive survey on answer generation methods using NLP. *Natural Language Processing Journal* 8 (2024), 100088. https://doi.org/10.1016/j.nlp.2024.100088

[55] Artem Vazhentsev, Lyudmila Rvanova, Gleb Kuzmin, Ekaterina Fadeeva, Ivan Lazichny, Alexander Panchenko, Maxim Panov, Timothy Baldwin, Mrinmaya Sachan, Preslav Nakov, and Artem Shelmanov. 2026. Uncertainty-Aware Attention Heads: Efficient Unsupervised Uncertainty Quantification for LLMs. In *Proceedings of 43rd International Conference on Machine Learning (ICML)*.

[56] Aparna Vinayan Kozhipuram, Samar Shailendra, and Rajan Kadel. 2025. Retrieval-Augmented Generation vs. Baseline LLMs: A Multi-Metric Evaluation for Knowledge-Intensive Content. *Information* 16, 9 (2025). https://doi.org/10.3390/info16090766

[57] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=1PL1NIMMrw

[58] Weihao Xuan, Qingcheng Zeng, Heli Qi, Junjue Wang, and Naoto Yokoya. 2025. Seeing is believing, but how much? a comprehensive analysis of verbalized calibration in vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 1408–1450.

[59] Duygu Nur Yaldiz, Yavuz Faruk Bakman, Baturalp Buyukates, Chenyang Tao, Anil Ramakrishna, Dimitrios Dimitriadis, Jieyu Zhao, and Salman Avestimehr. 2025. Do Not Design, Learn: A Trainable Scoring Function for Uncertainty Estimation in Generative LLMs. In *Findings of the Association for Computational Linguistics: NAACL 2025*.

[60] Yibin Yan and Weidi Xie. 2024. Echosight: Advancing visual-language models with wiki knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 1538–1551.

[61] Zhangyue Yin, Qiushi Sun, Qipeng Guo, Zhiyuan Zeng, Xiaonan Li, Junqi Dai, Qinyuan Cheng, Xuanjing Huang, and Xipeng Qiu. 2024. Reasoning in Flux: Enhancing Large Language Models Reasoning through Uncertainty-aware Adaptive Guidance. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL*. 2401–2416.

[62] Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2025. VisRAG: Vision-based Retrieval-augmented Generation on Multi-modality Documents. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=zG459X3Xge

[63] Qiwei Zhao, Dong Li, Yanchi Liu, Wei Cheng, Yiyou Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Huaxiu Yao, Chen Zhao, Haifeng Chen, and Xujiang Zhao. 2025. Uncertainty Propagation on LLM Agent. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*. 6064–6073.

[64] Yukun Zhao, Lingyong Yan, Weiwei Sun, Guoliang Xing, Chong Meng, Shuaiqiang Wang, Zhicong Cheng, Zhaochun Ren, and Dawei Yin. 2024. Knowing What LLMs DO NOT Know: A Simple Yet Effective Self-Detection Method. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL*. 7051–7063.