

# Uncertainty Quantification for Multimodal Retrieval Augmented Generation

Simon Binz<sup>1,2</sup>, **Heydar Soudani**<sup>1</sup>, and Faegheh Hasibi<sup>1</sup>

<sup>1</sup>Radboud University

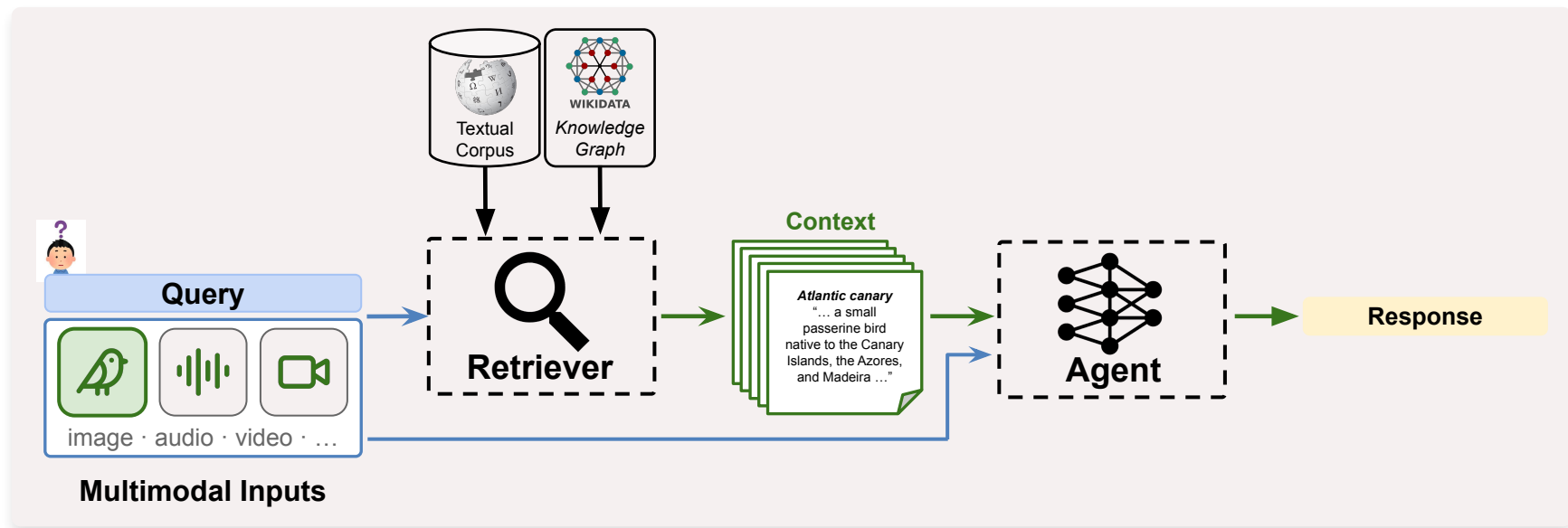
<sup>2</sup>German Research Center for Artificial Intelligence (DFKI)



*ICTIR, July 2026*



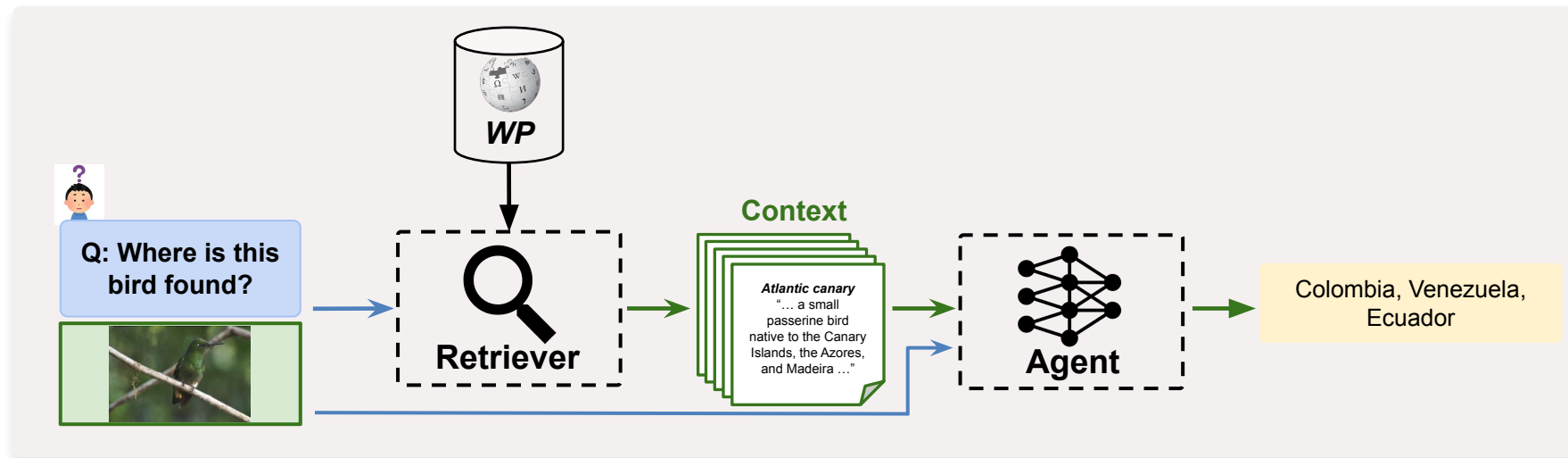
# Multimodal RAG: Definition



[1] Chen et al. (2022). MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text. EMNLP.

[2] Yasunaga et al. (2023). Retrieval-Augmented Multimodal Language Modeling. ICML.

# Multimodal RAG: Example

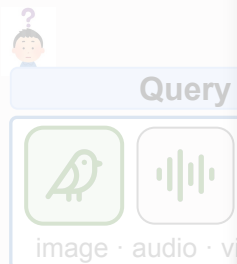


# Multimodal RAG: Limitations

## Prone to incorrect responses

- Misunderstanding the image-query relation
- Retrieving irrelevant documents
- Misinterpreting retrieved content
- Misusing internal knowledge

**Grounded and well-reasoned  $\neq$  Trustworthy !!**



Response

# Uncertainty Quantification (UQ)

## *Formal Definition*

A task aimed at assessing the reliability of system outputs by measuring the degree of uncertainty, the system's lack of confidence in its own prediction

# Uncertainty Quantification (UQ)

## *Formal Definition*

A task aimed at assessing the reliability of system outputs by measuring the degree of uncertainty, the system's lack of confidence in its own prediction

## Confidence vs. Uncertainty

- Some literature distinguishes uncertainty from confidence:
  - $\mathbf{U}(\mathbf{x})$ : dispersion over outputs given input  $x$
  - $\mathbf{C}(\mathbf{x}, r)$ : estimated correctness of  $r$  given  $x$
- We follow the literature that treats uncertainty as a lack of confidence
- Use the two terms interchangeably

# Type of Uncertainty

- **Epistemic:** uncertainty stemming from the model's lack of knowledge (reducible)
- **Aleatoric:** uncertainty arising from the inherent ambiguity of the data (irreducible)

**Total uncertainty:** Encompasses epistemic and aleatoric uncertainty

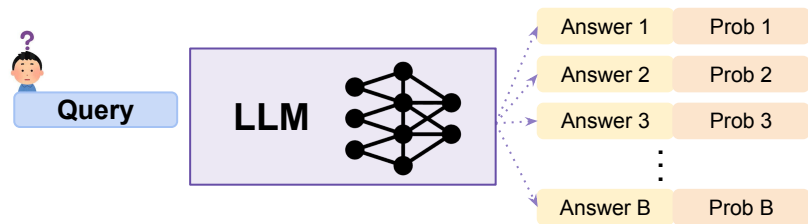
## Why Total?

In language models these two parts are very hard to separate

# UQ for LLMs

**Main Idea:** The more diverse the outputs are, the more uncertain the model is

## Multi Generations



- 1) Perturbing the input
- 2) Changing the decoding temperature

[1] Kuhn et al. (2023). Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in NLG. ICLR.

[2] Lin et al. (2024). Generating with Confidence: Uncertainty Quantification for Black-box LLMs. TMLR.

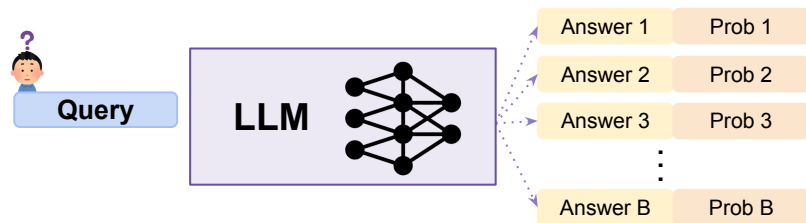
[3] Wang et al. (2023). Self-Consistency Improves Chain-of-Thought Reasoning in Language Models. ICLR.

# UQ for LLMs

**Main Idea:** The more diverse the outputs are, the more uncertain the model is

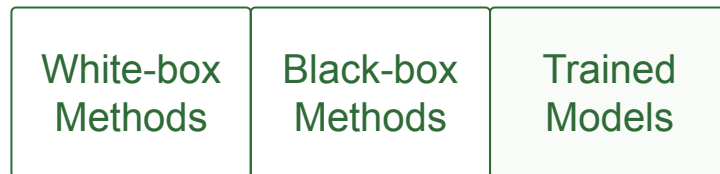
**Score Function:** Convert the generations to an uncertainty / confidence score

## Multi Generations



- 1) Perturbing the input
- 2) Changing the decoding temperature

## Score Function



[1] Kuhn et al. (2023). Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in NLG. ICLR.

[2] Lin et al. (2024). Generating with Confidence: Uncertainty Quantification for Black-box LLMs. TMLR.

[3] Wang et al. (2023). Self-Consistency Improves Chain-of-Thought Reasoning in Language Models. ICLR.

# UQ for LLMs

**Main Idea:** The more diverse the outputs are, the more uncertain the model is

**Score Function:** Convert the generations to an uncertainty / confidence score

## !!! Limitation

UQ methods are mainly defined for a single LLM and QA task, considering LLM as the only one source of uncertainty

- 1) Perturbing the input
- 2) Changing the decoding temperature

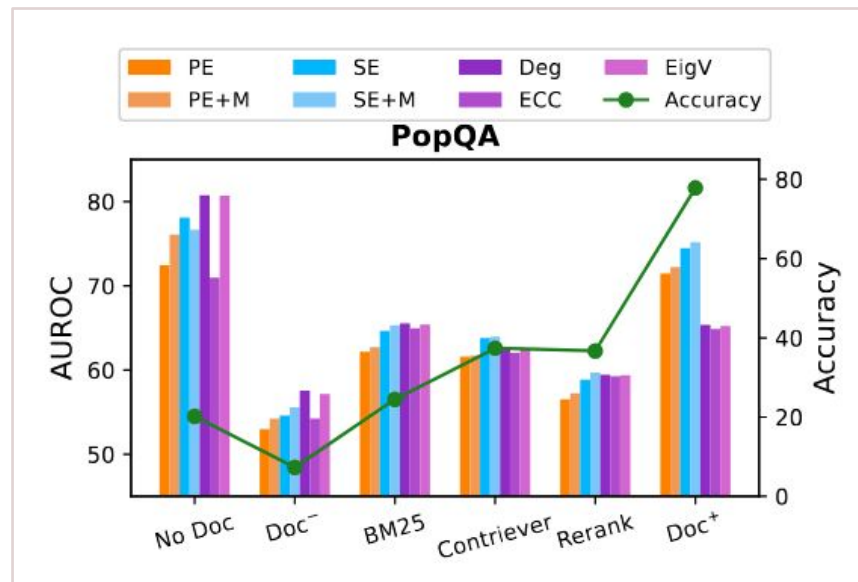
[1] Kuhn et al. (2023). Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in NLG. ICLR.

[2] Lin et al. (2024). Generating with Confidence: Uncertainty Quantification for Black-box LLMs. TMLR.

[3] Wang et al. (2023). Self-Consistency Improves Chain-of-Thought Reasoning in Language Models. ICLR.

# UQ Methods Do Not Transfer to RAG

Adding retrieved context degrades UQ performance

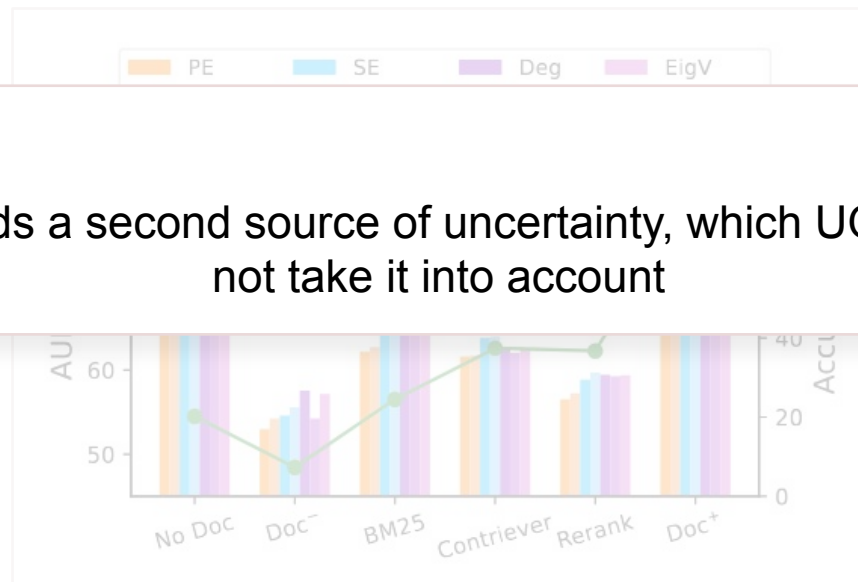


# UQ Methods Do Not Transfer to RAG

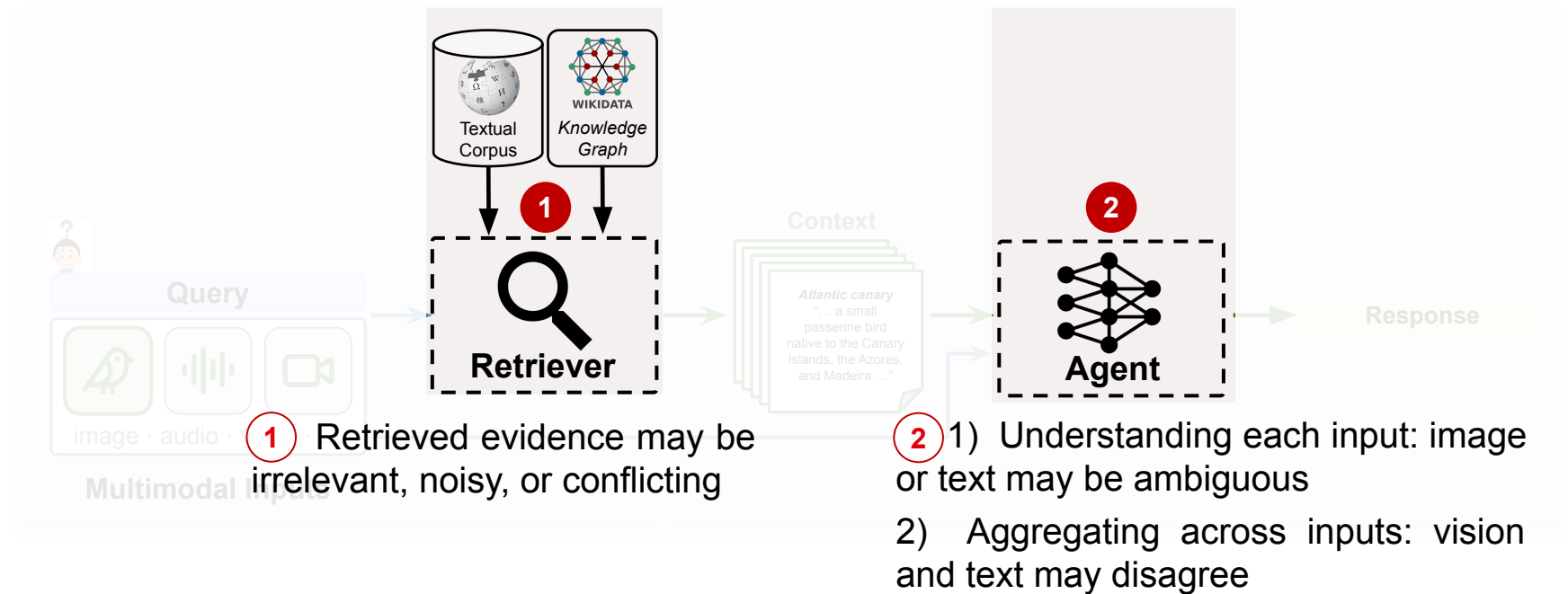
Adding retrieved context degrades UQ performance

## !!! Reason

Retrieval adds a second source of uncertainty, which UQ methods do not take it into account



# Research Gap: UQ for Multimodal RAG



# The Core Challenge

**Capturing uncertainty arising from multiple stages of the pipeline, including retrieval, visual understanding, and generation**





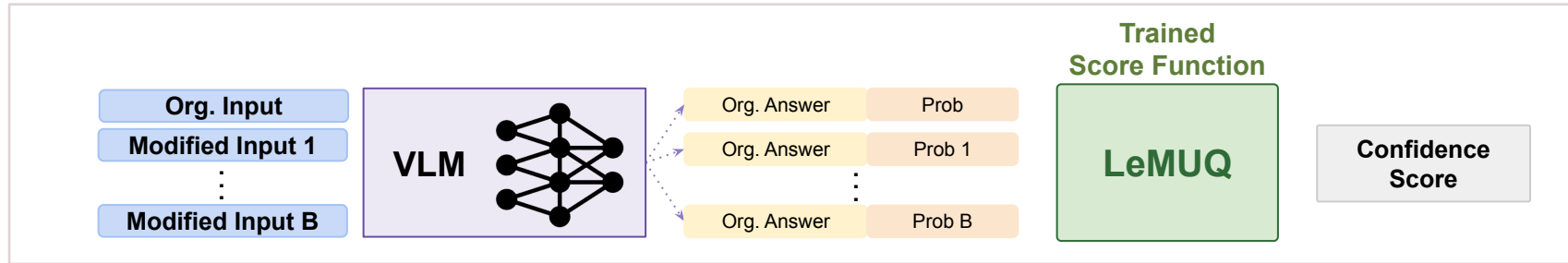
# Methodology

# LeMUQ

## Learnable Multimodal Uncertainty Quantification method (LeMUQ)

### Core idea:

- Fix the answer, perturb the inputs
- Re-compute the probability of that same answer under each perturbation
- Trained RoBERTa-based score function maps the probability vector  $\rightarrow$  confidence score



# LeMUQ

## STEP 1: Generating the most-likely answer

**1**

**Question**  
Where is this bird commonly found?

The Gran Canaria blue chaffinch (*Fringilla polatzeki*) is a species of passerine bird in the finch family Fringillidae. It is endemic to Gran Canaria in Spain's Canary Islands.



 VLM

**Most-likely Answer** Canary Islands <eos>

**Probabilities** 0.9 0.96 0.99

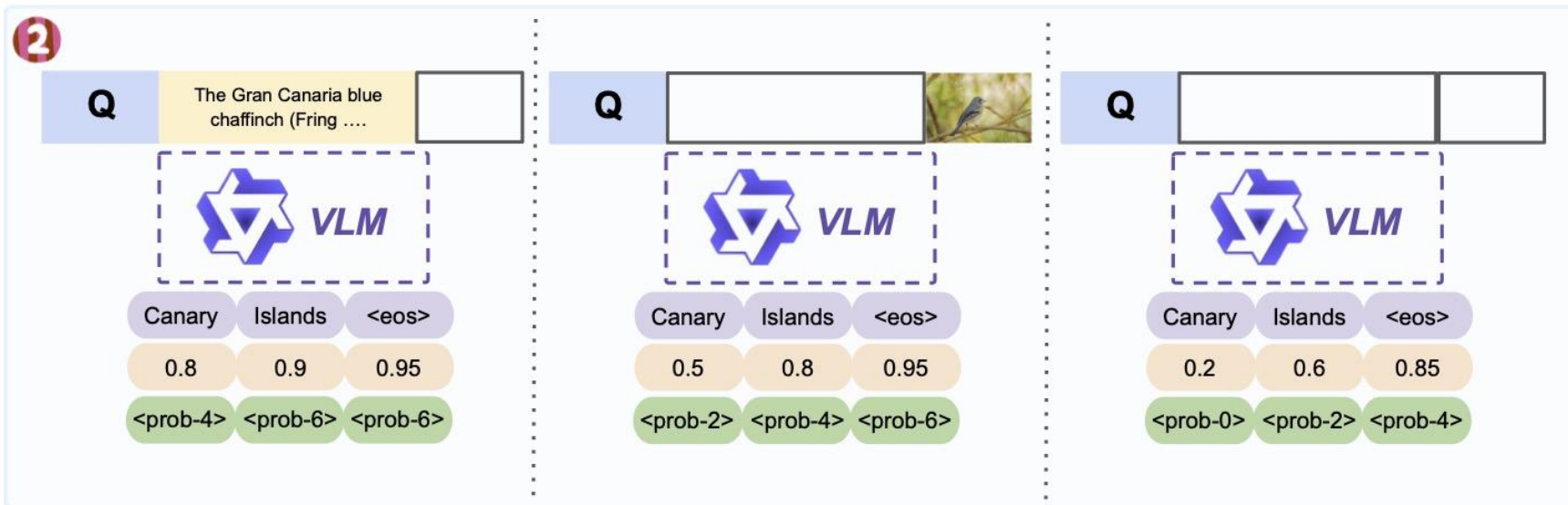
**Mapped Probabilities** <prob-5> <prob-6> <prob-7>

**What?** Each token probability is discretized and mapped to a one-hot vector representation

**Why?**  
Makes probabilities transformer-friendly

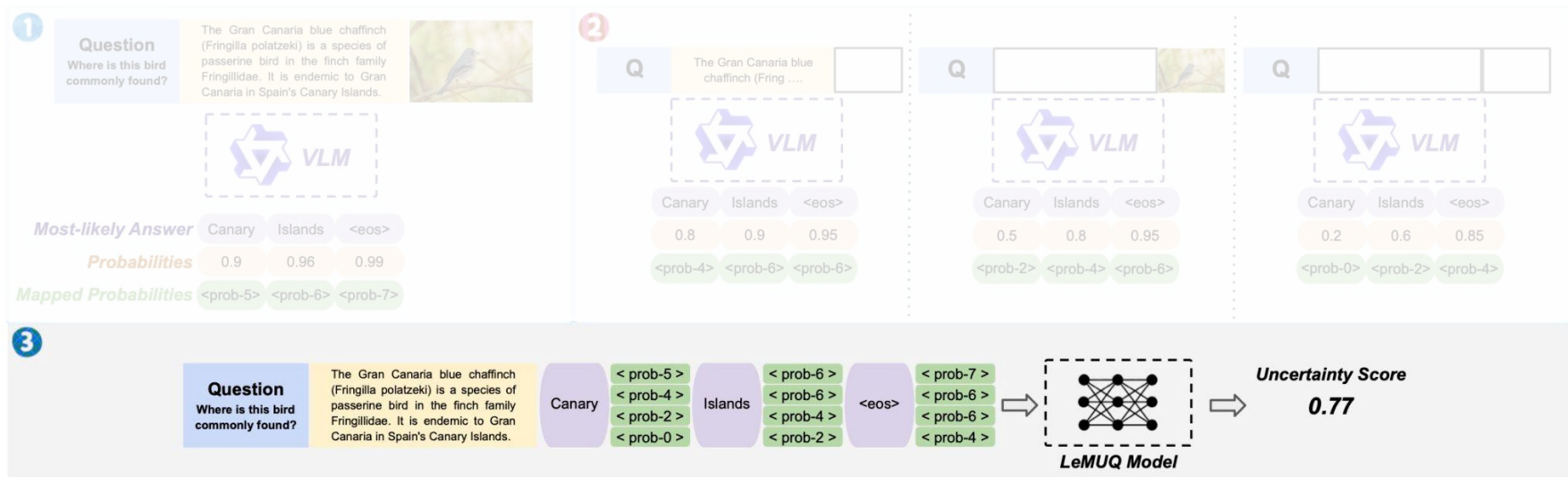
# LeMUQ

## STEP 2: Modeling Probabilities



# LeMUQ

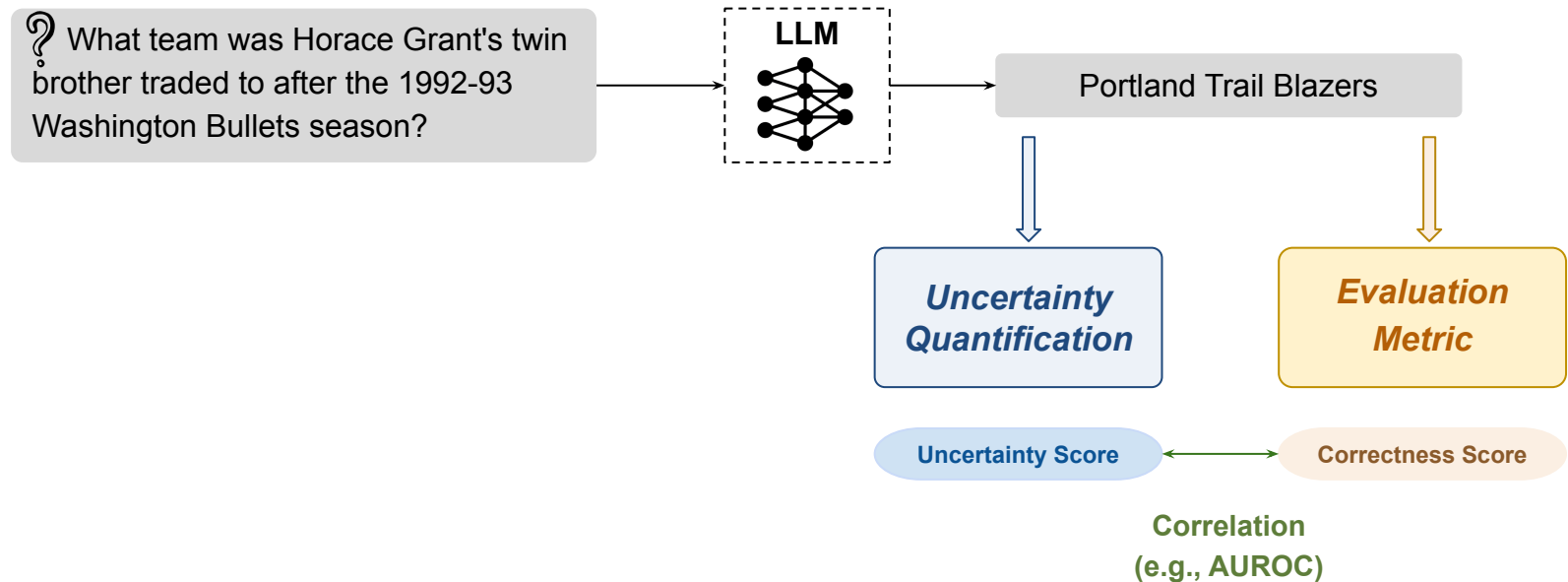
## STEP 3: Generating an uncertainty score





# Evaluation

# UQ Method Evaluation



# LeMUQ Performance

**RQ1: How does our proposed method perform compared to baseline UQ methods when evaluated within-distribution?**

| 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 <sub>Base</sub> | 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 <sub>Ret</sub>  | 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                | <b>0.855<sup>†</sup></b> | <b>0.833<sup>†</sup></b> | <b>0.861<sup>†</sup></b> | <b>0.854<sup>†‡</sup></b> | <b>0.871<sup>†‡</sup></b> | <b>0.853<sup>†‡</sup></b> | <b>0.923<sup>†‡</sup></b> | <b>0.888<sup>†‡</sup></b> | <b>0.912<sup>†‡</sup></b> | <b>0.913<sup>†‡</sup></b> | <b>0.902<sup>†‡</sup></b> | <b>0.898<sup>†‡</sup></b> | <b>0.880</b> |
|                      | (+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)     |

# LeMUQ Performance

**RQ1: How does our proposed method perform compared to baseline UQ methods when evaluated within-distribution?**

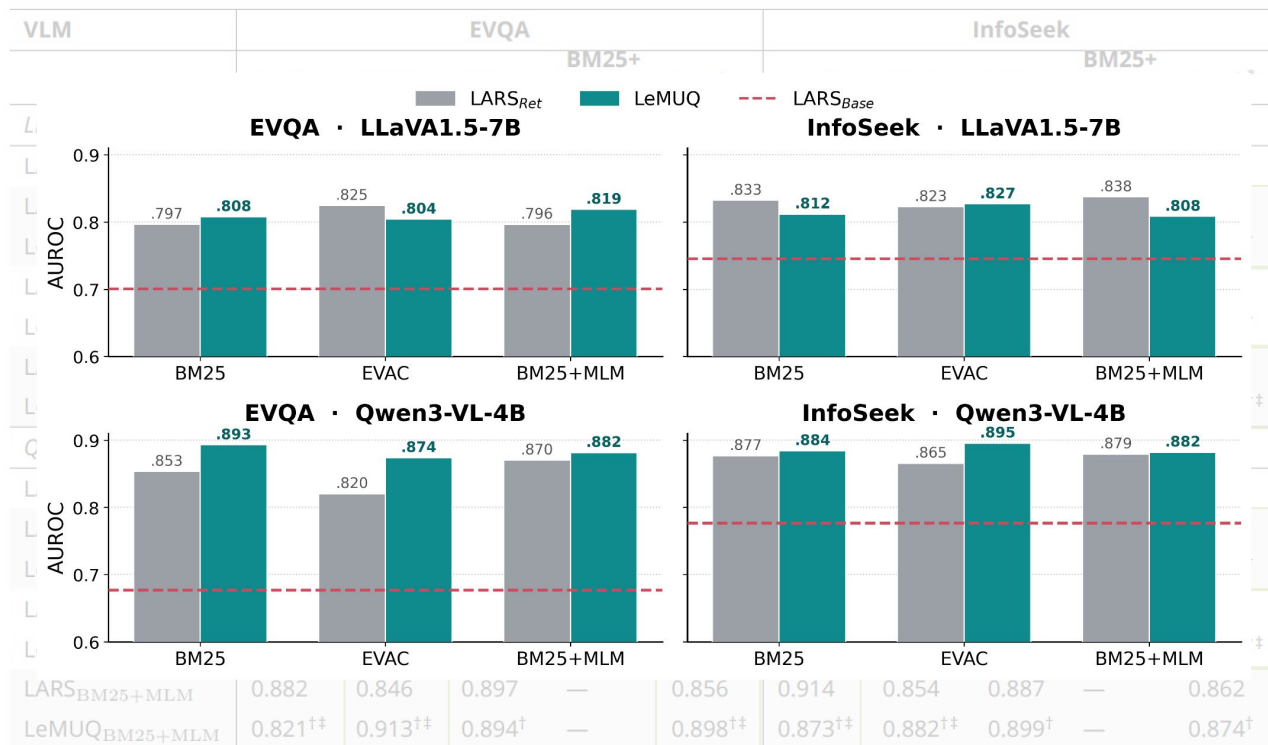
| LLM                  | LLaVA1.5-7B        |                    |                                                                                                        |                     |                     |                     | Qwen3-VL-4B         |                     |                     |                     |                     |                     | Avg.     |
|----------------------|--------------------|--------------------|--------------------------------------------------------------------------------------------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|----------|
| Retriever            | BM25               |                    | EVAC.                                                                                                  |                     | BM25+MLM            |                     | BM25                |                     | EVAC.               |                     | BM25+MLM            |                     |          |
|                      | EVQA               | InfoSeek           | <div>Central Finding!</div> <div>LeMUQ performs better than baseline methods within-distribution</div> |                     |                     |                     |                     |                     | InfoSeek            | EVQA                | InfoSeek            |                     |          |
| Accuracy             | 0.128              | 0.118              |                                                                                                        |                     |                     |                     |                     |                     | 0.127               | 0.169               | 0.140               | 0.137               |          |
| PE                   | 0.750              | 0.756              |                                                                                                        |                     |                     |                     |                     |                     | 0.813               | 0.692               | 0.749               | 0.752               |          |
| P(True)              | 0.639              | 0.645              |                                                                                                        |                     |                     |                     |                     |                     | 0.714               | 0.584               | 0.740               | 0.653               |          |
| Ecc.                 | 0.732              | 0.745              |                                                                                                        |                     |                     |                     |                     |                     | 0.705               | 0.634               | 0.629               | 0.702               |          |
| Img. Per.            | 0.415              | 0.414              |                                                                                                        |                     |                     |                     |                     |                     | 0.321               | 0.389               | 0.291               | 0.369               |          |
| LARS <sub>Base</sub> | 0.689              | 0.762              |                                                                                                        |                     |                     |                     |                     |                     | 0.825               | 0.607               | 0.771               | 0.730               |          |
| LARS <sub>Ret</sub>  | 0.847              | 0.829              | 0.890                                                                                                  | 0.816               | 0.845               | 0.843               |                     |                     |                     |                     |                     |                     |          |
| LeMUQ                | 0.855 <sup>†</sup> | 0.833 <sup>†</sup> | 0.861 <sup>†</sup>                                                                                     | 0.854 <sup>†‡</sup> | 0.871 <sup>†‡</sup> | 0.853 <sup>†‡</sup> | 0.923 <sup>†‡</sup> | 0.888 <sup>†‡</sup> | 0.912 <sup>†‡</sup> | 0.913 <sup>†‡</sup> | 0.902 <sup>†‡</sup> | 0.898 <sup>†‡</sup> | 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) |

**Central Finding!**

LeMUQ performs better than baseline methods within-distribution

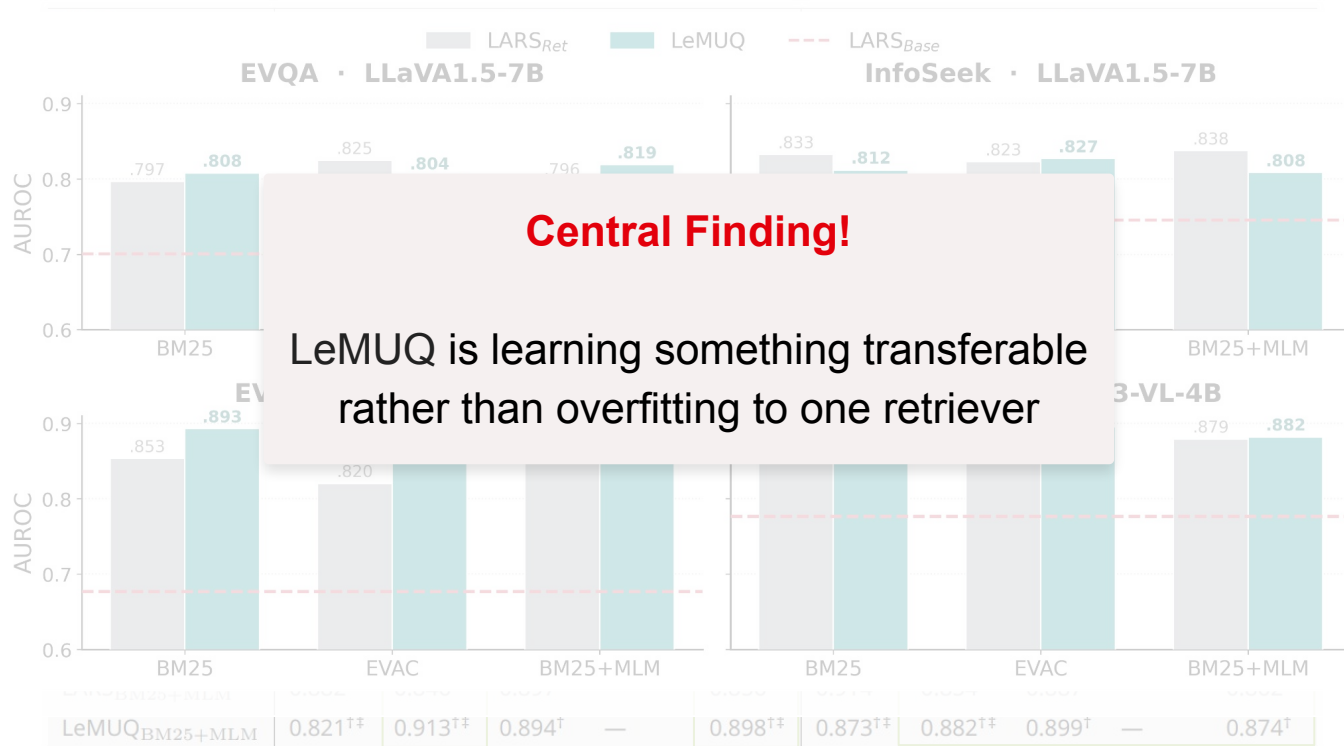
# Out-of-Domain Generalizability: Retriever

RQ2: How well does LeMUQ generalize to out-of-distribution data?



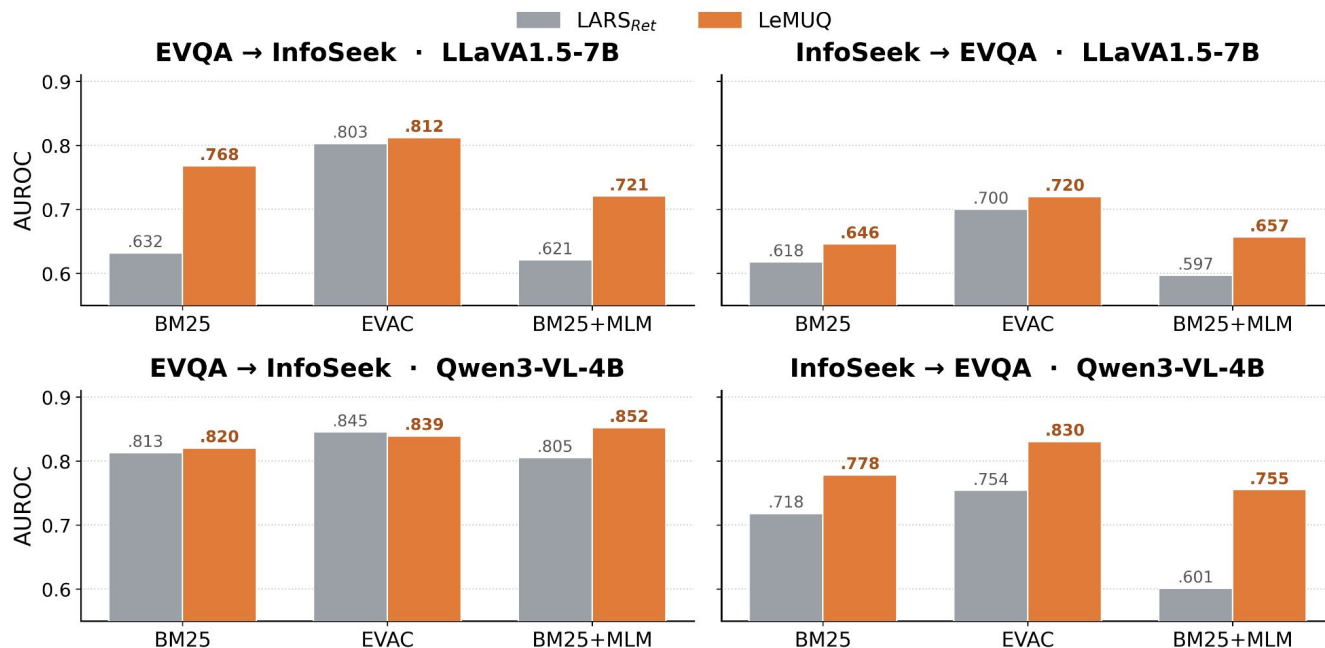
# Out-of-Domain Generalizability: Retriever

RQ2: How well does LeMUQ generalize to out-of-distribution data?



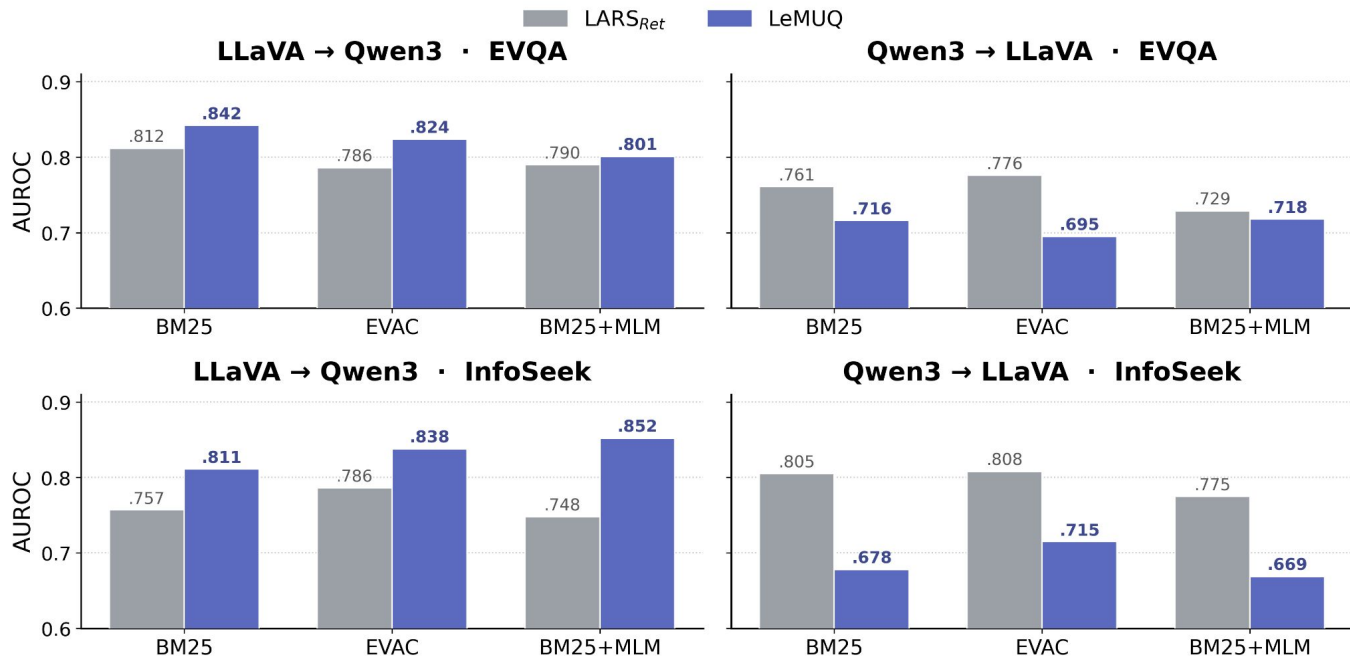
# Out-of-Domain Generalizability: Dataset

RQ2: How well does LeMUQ generalize to out-of-distribution data?



# Out-of-Domain Generalizability: VLM

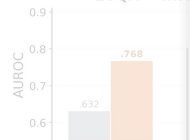
RQ2: How well does LeMUQ generalize to out-of-distribution data?



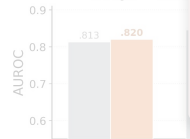
# Out-of-Domain Generalizability: Dataset & VLM

RQ2: How well does LeMUQ generalize to out-of-distribution data?

EVQA → Info



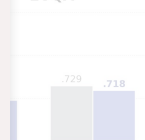
EVQA → Info



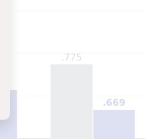
**Central Finding!**

LeMUQ generalizes well across datasets and has mixed results when transferring across VLMs

EVQA



InfoSeek



# Impact of Probability Components

**RQ3: What is the impact of each probability component in LeMUQ on its overall performance?**

| 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 <sub>Base</sub>     | 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 <sub>Ret</sub>      | 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        | <b>0.861</b> | <b>0.854</b> <sup>†</sup> | <b>0.871</b> <sup>†</sup> | 0.853 <sup>†</sup>        | 0.920 <sup>†</sup>        | 0.889 <sup>†</sup>         | 0.917 <sup>†</sup>        | 0.909 <sup>†</sup>        | 0.902 <sup>†</sup>        | 0.901 <sup>†</sup>        | <b>0.880</b> |
| LeMUQ $-\tilde{p}$       | 0.841        | 0.826        | 0.852        | 0.825 <sup>‡</sup>        | 0.859                     | 0.844 <sup>†</sup>        | 0.921 <sup>†</sup>        | 0.903 <sup>†‡</sup>        | 0.910 <sup>†</sup>        | 0.906 <sup>†</sup>        | 0.894 <sup>†</sup>        | <b>0.909</b> <sup>†</sup> | 0.874        |
| LeMUQ $-\tilde{p}^{q,I}$ | 0.856        | 0.845        | 0.851        | 0.829 <sup>‡</sup>        | 0.840 <sup>‡</sup>        | 0.829 <sup>†‡</sup>       | <b>0.924</b> <sup>†</sup> | 0.891 <sup>†</sup>         | 0.912 <sup>†</sup>        | 0.906 <sup>†</sup>        | <b>0.907</b> <sup>†</sup> | 0.893 <sup>†</sup>        | 0.873        |
| LeMUQ $-\tilde{p}^{q,c}$ | 0.852        | 0.818        | 0.852        | 0.849 <sup>†</sup>        | 0.856                     | <b>0.862</b> <sup>†</sup> | 0.923 <sup>†</sup>        | <b>0.906</b> <sup>†‡</sup> | <b>0.919</b> <sup>†</sup> | 0.899                     | 0.895 <sup>†</sup>        | 0.905 <sup>†</sup>        | 0.878        |
| LeMUQ $-\tilde{p}^q$     | <b>0.859</b> | <b>0.849</b> | 0.851        | 0.848 <sup>†</sup>        | 0.864                     | 0.851 <sup>†</sup>        | 0.906 <sup>†‡</sup>       | 0.902 <sup>†‡</sup>        | 0.916 <sup>†</sup>        | <b>0.911</b> <sup>†</sup> | 0.886 <sup>†‡</sup>       | 0.905 <sup>†</sup>        | 0.879        |

# Impact of Probability Components

**RQ3: What is the impact of each probability component in LeMUQ on its overall performance?**

| VLM                      | LLaVA1.5-7B                                                                                                   |       |       |                    |          |                    | Qwen3-VL-4B         |                     |                    |                    |                     |                    | Avg   |
|--------------------------|---------------------------------------------------------------------------------------------------------------|-------|-------|--------------------|----------|--------------------|---------------------|---------------------|--------------------|--------------------|---------------------|--------------------|-------|
| Retriever                | BM25                                                                                                          |       | EVAC. |                    | BM25+MLM |                    | BM25                |                     | EVAC.              |                    | BM25+MLM            |                    |       |
|                          | InfoSeek                                                                                                      |       |       |                    |          |                    |                     |                     |                    |                    |                     |                    |       |
| Accuracy                 | Central Finding!                                                                                              |       |       |                    |          |                    |                     |                     |                    |                    |                     | 0.140              | 0.137 |
| LARS <sub>Base</sub>     | The full four-component model has the highest average AUROC and the most balanced performance across settings |       |       |                    |          |                    |                     |                     |                    |                    |                     | 0.773              | 0.730 |
| LARS <sub>Ret</sub>      |                                                                                                               |       |       |                    |          |                    |                     |                     |                    |                    |                     | 0.843              | 0.842 |
| LeMUQ                    |                                                                                                               |       |       |                    |          |                    |                     |                     |                    |                    |                     | 0.901 <sup>†</sup> | 0.880 |
| LeMUQ $-\tilde{p}$       |                                                                                                               |       |       |                    |          |                    |                     |                     |                    |                    |                     | 0.909 <sup>†</sup> | 0.874 |
| LeMUQ $-\tilde{p}^{q,I}$ |                                                                                                               |       |       |                    |          |                    |                     |                     |                    |                    |                     | 0.893 <sup>†</sup> | 0.873 |
| LeMUQ $-\tilde{p}^{q,c}$ | 0.852                                                                                                         | 0.818 | 0.852 | 0.849 <sup>†</sup> | 0.856    | 0.862 <sup>†</sup> | 0.923 <sup>†</sup>  | 0.906 <sup>†‡</sup> | 0.919 <sup>†</sup> | 0.899              | 0.895 <sup>†</sup>  | 0.905 <sup>†</sup> | 0.878 |
| LeMUQ $-\tilde{p}^q$     | 0.859                                                                                                         | 0.849 | 0.851 | 0.848 <sup>†</sup> | 0.864    | 0.851 <sup>†</sup> | 0.906 <sup>†‡</sup> | 0.902 <sup>†‡</sup> | 0.916 <sup>†</sup> | 0.911 <sup>†</sup> | 0.886 <sup>†‡</sup> | 0.905 <sup>†</sup> | 0.879 |

