# Confident RAG: Enhancing the Performance of LLMs for Mathematics Question Answering through Multi-Embedding and Confidence Scoring

Shi-Ting Chen[a,1], Zijian Zhao[b,1] (zzhaock@connect.ust.hk), Jinsong Chen[a]*

[a]The University of Hong Kong [b]The Hong Kong University of Science and Technology

## Retrieval-Augmented Generation (RAG)

- **Data Preparation:** document parsing and text chunking → $C = \{c_1, c_2, \dots, c_m\}$
- **Indexing:** vectorization → $E = \{e_1, e_2, \dots, e_m\}$, where $e_i = \mathrm{g}(c_i)$
  - $e_i$: embedding vector of chunk $c_i$
  - $\mathrm{g}(\cdot)$: embedding model
- **Retrieval:**
  - Computing the similarity between query $q$ and each chunk: $s_i = \frac{\mathrm{g}(q)\cdot e_i}{|\mathrm{g}(q)||e_i|}$
  - Selecting the top-$k$ chunks with the highest similarity scores
- **Augmented Generation:** incorporating the retrieved information to enrich the input to the generative model

## Confidence of LLMs

Various metrics can be used to measure the confidence of an LLM's response, such as Average Log-Probability (AvgLogP), Gini Impurity (Gini), Entropy, Distributional Perplexity (DP), and KL divergence to the uniform distribution (Self-Certainty). According to experimental results, model performance is positively correlated with these confidence scores:

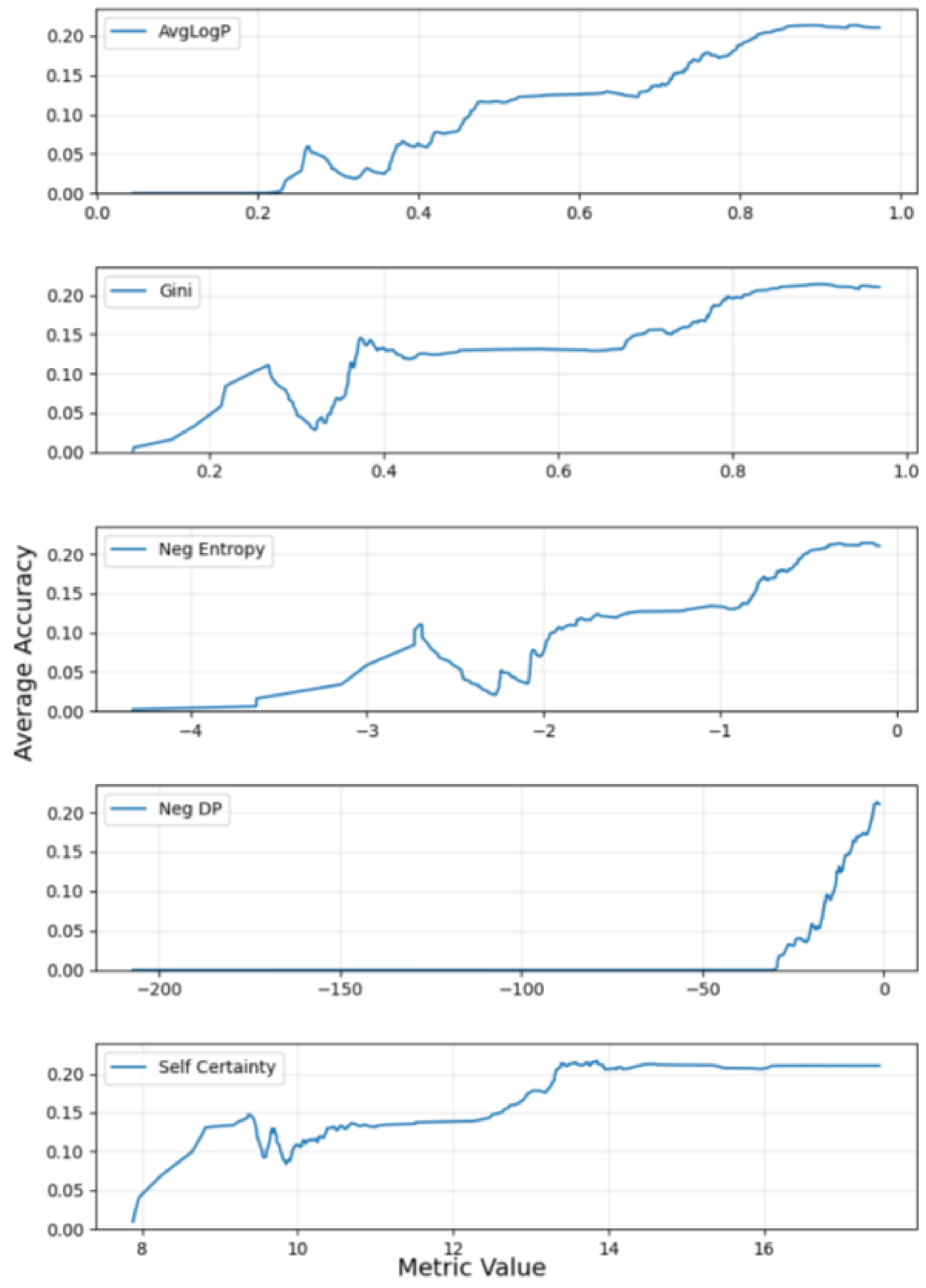

Fig. 1: Cumulative Distribution Function (CDF) of Accuracy for Different Metrics

## Confident RAG

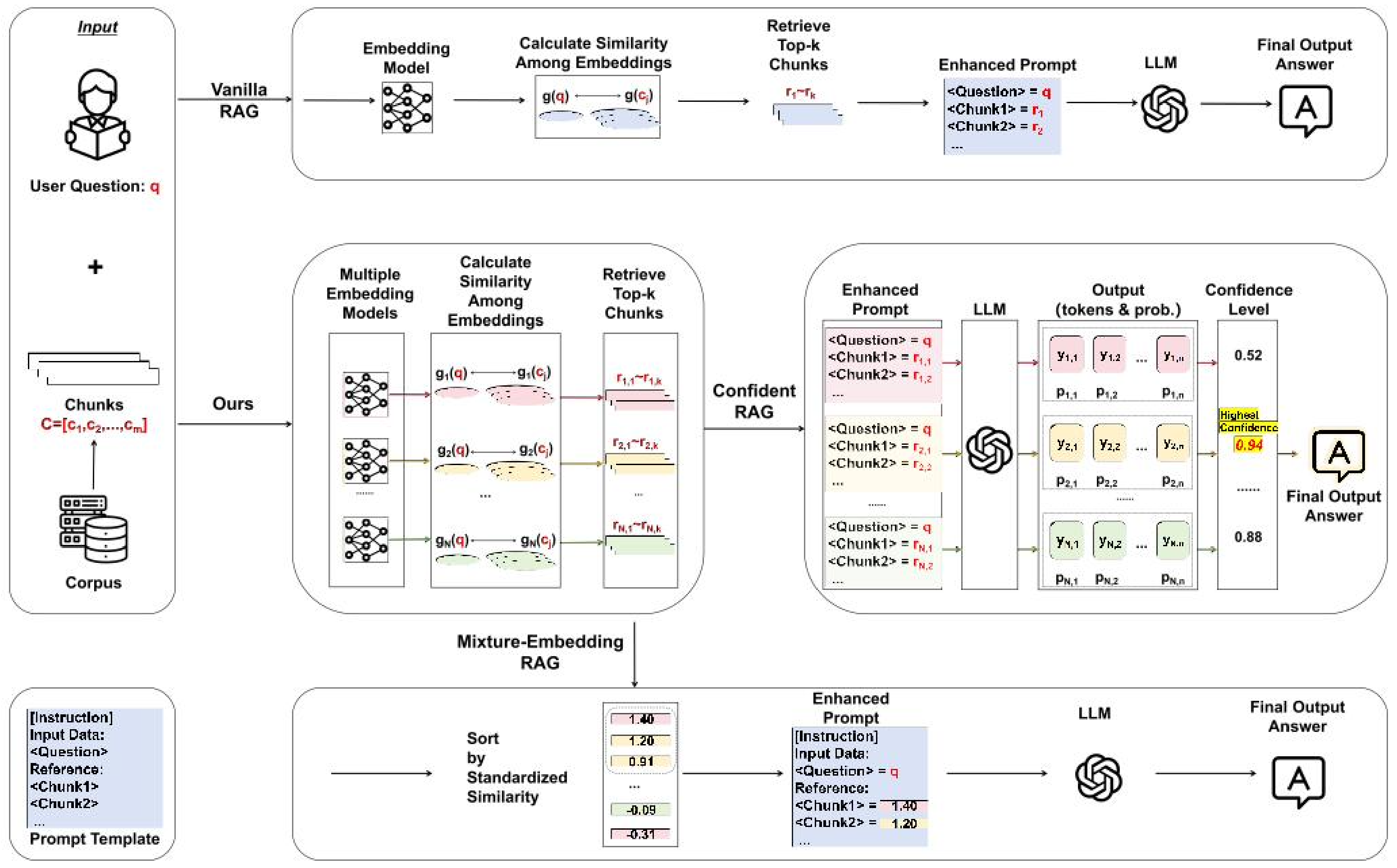

Fig. 2: Workflow

- **Motivation:** Different embedding models yield different retrieval results due to variations in architecture and training data. Our goal is to enable LLMs to adaptively select the optimal embedding model for each query, guided by confidence signals.
- **Methodology:**
  - Generate answers using RAG repeatedly, each time employing a different embedding model for indexing and retrieval.
  - Output the answer with the highest confidence score.
- **Confidence Metrics:**

$$p(v_j|x, y_{<i}) = \mathsf{f}(x, y_{<i})[j]\ , \tag{1}$$

$$\text{AvgLogP} = \frac{1}{n}\sum_{i=1}^{n} \log p(y_i|x, y_{<i})\ , \tag{2}$$

$$\text{Gini} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{|v|} \left(p(v_j|x, y_{<i})\right)^2\ , \tag{3}$$

$$\text{Entropy} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{|v|} -p(j|x, y_{<i}) \log p(v_j|x, y_{<i})\ , \tag{4}$$

$$\text{DP} = \frac{1}{n}\sum_{i=1}^{n} \exp\left(-\sum_{j=1}^{|v|} p(v_j|x, y_{<i}) \log p(v_j|x, y_{<i})\right)\ , \tag{5}$$

$$\text{Self-Certainty} = -\frac{1}{n|v|}\sum_{i=1}^{n}\sum_{j=1}^{V} \log\left(|v| \cdot p(v_j|x, y_{<i})\right)\ , \tag{6}$$

$$\tag{7}$$

- $\mathsf{f}()$: LLM
- $x$: the prompt
- $y_{<i}$: the first generated $i-1$ tokens
- $y_i$: the $i^{th}$ token
- $v_j$: the $j^{th}$ element in the vocabulary $v$
- $|v|$: the vocabulary size
- $n$: the answer length

## Experiment

### A. Experiment Setup:

- **LLMs:** Qwen2.5-Math-7B [6], Llama-3.1-8B [8], OLMo-2-1124-7B [2]
- **Embedding Models:** MiniLM-L6-v2 [7], ModernBERT-large [7], MathBERT [4], and stsb-roberta-large [5]
- **Q&A Datasets:** Gsm8k [1]
- **Corpus:** Mathematics Textbooks from OpenStax [3]

| | Qwen2.5-Math-7B | Llama-3.1-8B | OLMo-2-1124-7B |
|---|---|---|---|
| w/o RAG | 75.2 | 16.6 | 21.0 |
| vanilla RAG | 80.5 | 21.3 | 26.0 |
| AvgLogP | 80.1 | 26.8 | 31.5 |
| Self-Certainty | **84.3** | **27.0** | 31.2 |
| Gini | 80.0 | 26.7 | 30.5 |
| Entropy | 80.6 | 26.9 | 31.8 |
| DP | 81.6 | **27.0** | **33.3** |

Tab. 1: Evaluation Accuracy: The last five rows correspond to Confident-RAG results using different confidence metrics. RAG results are averaged across various embedding models. Bold indicates the best performance; underline indicates the second best.

### B. Experiment Results:

- Confident-RAG consistently outperforms vanilla RAG across most scenarios.
- Confident-RAG yields greater improvements for general-purpose LLMs than for domain-specific LLMs, as the latter already possess substantial knowledge aligned with the external corpus.
- Self-Certainty and DP achieve the best performance when used as confidence metrics for Confident-RAG.

## References

[1] Karl Cobbe et al. "Training Verifiers to Solve Math Word Problems". In: *arXiv preprint arXiv:2110.14168* (2021).

[2] Aaron Grattafiori et al. "The llama 3 herd of models". In: *arXiv preprint arXiv:2407.21783* (2024).

[3] OpenStax. *Open Textbooks Math*. URL: https://openstax.org/subjects/math.

[4] Shuai Peng et al. "Mathbert: A pre-trained model for mathematical formula understanding". In: *arXiv preprint arXiv:2105.00377* (2021).

[5] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2019. URL: http://arxiv.org/abs/1908.10084.

[6] Rulin Shao et al. *Spurious Rewards: Rethinking Training Signals in RLVR*. 2025. arXiv: 2506.10947 [cs.AI]. URL: https://arxiv.org/abs/2506.10947.

[7] Benjamin Warner et al. *Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference*. 2024. arXiv: 2412.13663 [cs.CL]. URL: https://arxiv.org/abs/2412.13663.

[8] An Yang et al. "Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement". In: *arXiv preprint arXiv:2409.12122* (2024).