

Benchmarking LLMs via Uncertainty Quantification

Fanghua Ye^{1,2}, Mingming Yang¹, Jianhui Pang^{1,3}, Longyue Wang¹,
Derek F. Wong³, Emine Yilmaz², Shuming Shi¹, Zhaopeng Tu¹

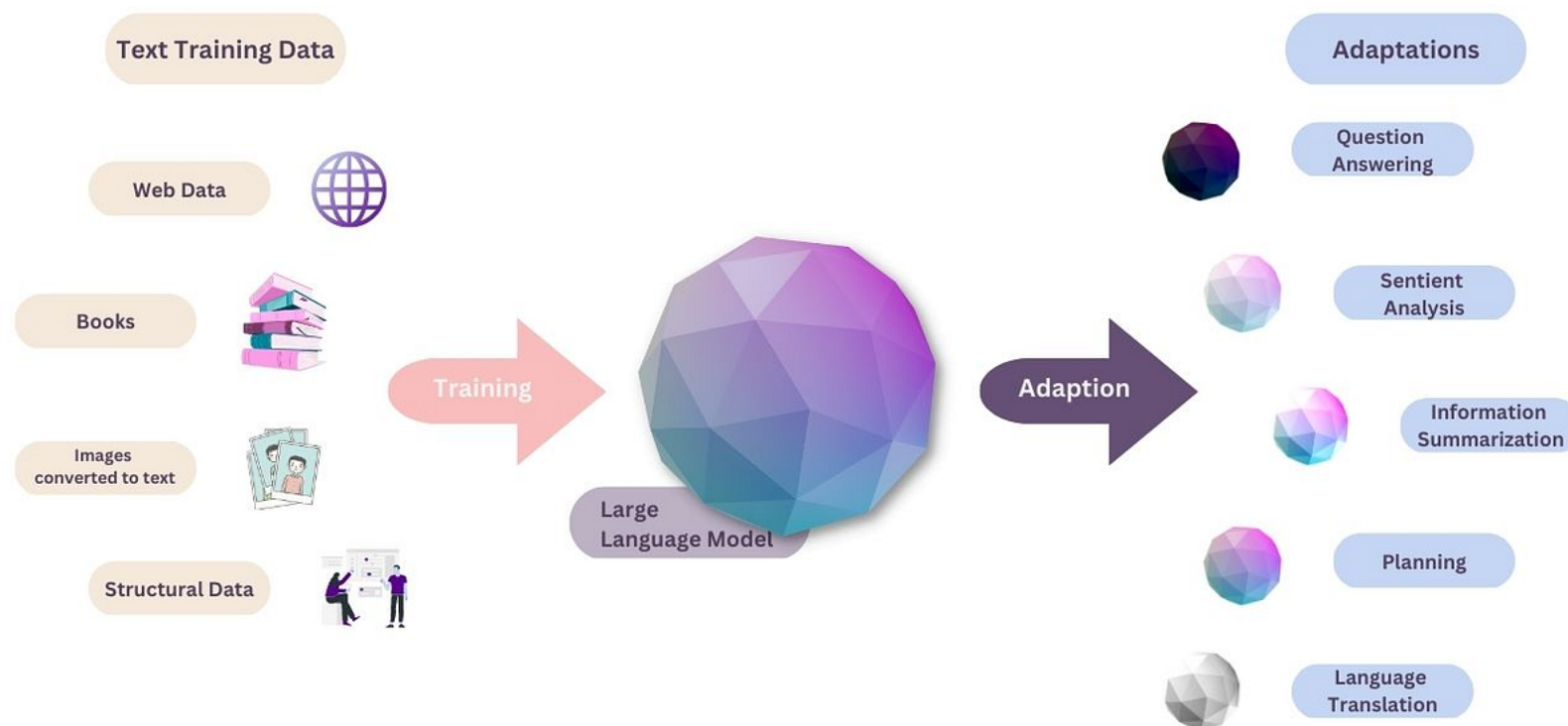
1. Tencent AI Lab 2. University College London 3. University of Macau



澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

Background

- ❖ LLMs are able to show remarkable performance across various tasks



Background

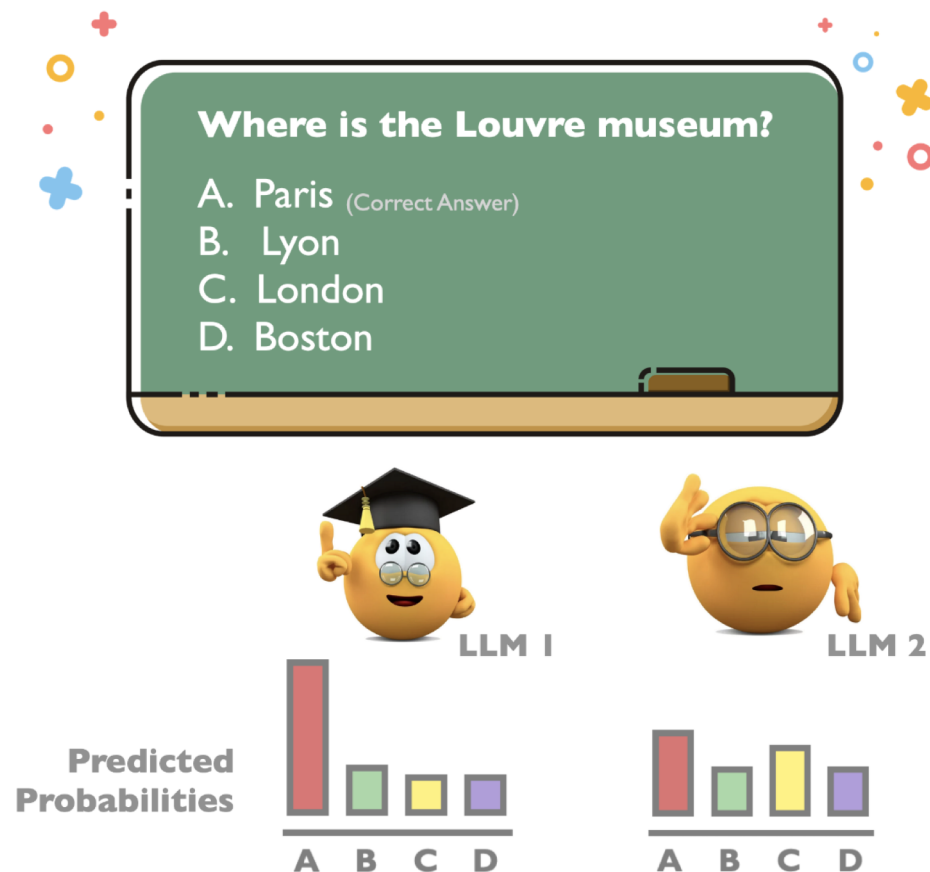
❖ Comprehensively evaluating LLMs becomes a huge challenge

The screenshot shows the OpenCompass website interface. At the top, there is a navigation bar with the OpenCompass logo and name in Chinese (司南), followed by links for CompassHub, CompassRank, CompassKit, and Docs. On the right side of the navigation bar, there are links for '中 | EN', 'Contribute Benchmarks', and 'Login'. Below the navigation bar, there is a category filter section with buttons for 'All', 'Examination', 'Language', 'Knowledge', 'Understanding', 'Reasoning', 'Long-Context', 'Creation', 'Code', and 'Agent'. Underneath, there are buttons for 'Math' and 'Other'. A 'Hot Tags' section is visible, along with 'Views' and 'Filter' options. A search bar labeled 'Search Dataset' is also present. The main content area displays a grid of benchmark cards. Each card features a title, a description, a tag, a date, and a view count. The cards shown are: T-Eval (tag: agent, date: 2024-01-25, views: 1521), GSM8K (tag: math, date: 2024-01-11, views: 1292), C-Eval (tag: Examination, date: 2024-01-11, views: 1108), MMLU (tag: Examination, date: 2024-01-11, views: 903), HumanEval, BBH, LongBench, and MATH.

Benchmark Name	Tag	Date	Views
T-Eval	agent	2024-01-25	1521
GSM8K	math	2024-01-11	1292
C-Eval	Examination	2024-01-11	1108
MMLU	Examination	2024-01-11	903
HumanEval			
BBH			
LongBench			
MATH			

Motivation

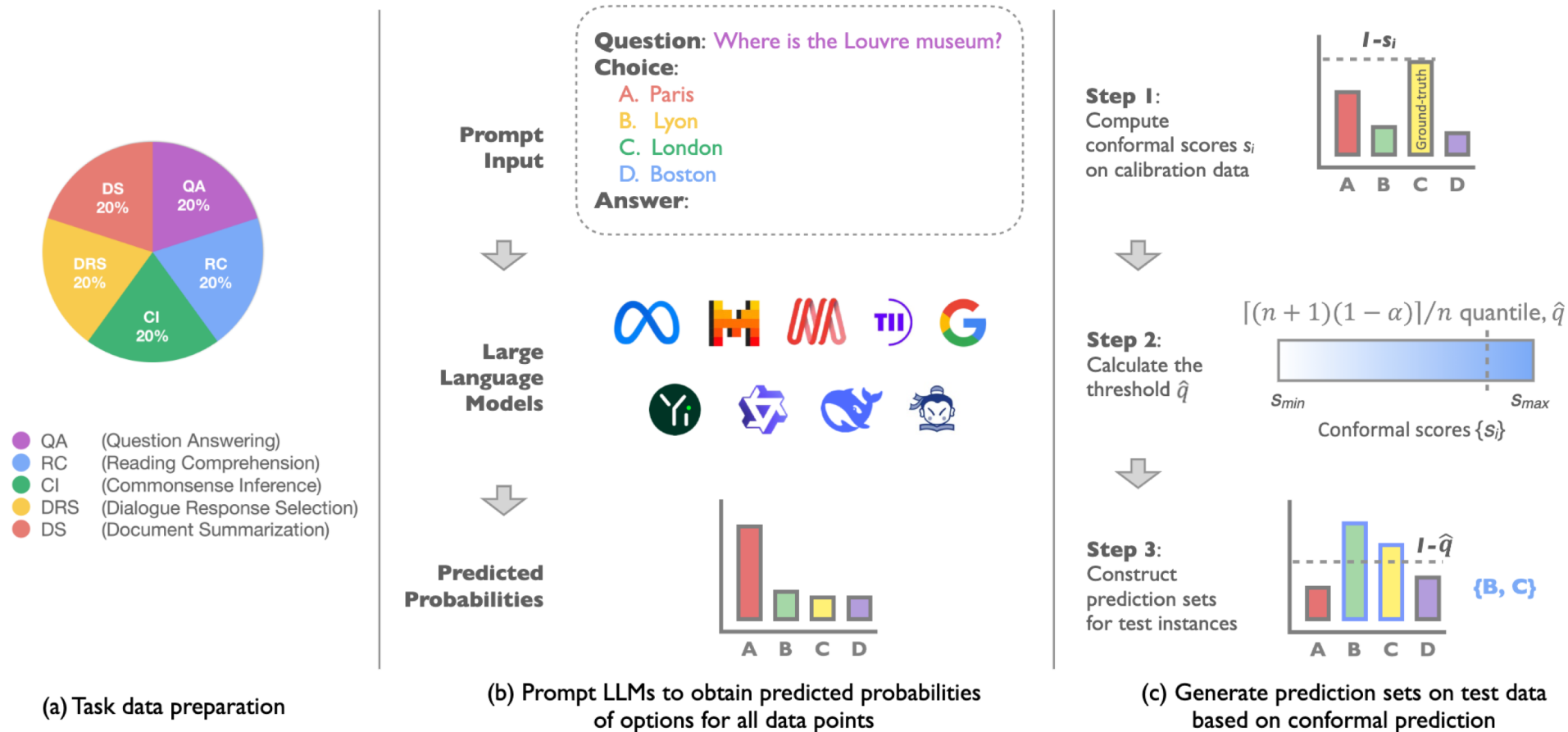
- ❖ Existing evaluation overlooks the uncertainty of LLMs
 - LLMs should be aware of what they are unsure about



Two LLMs can demonstrate the same accuracy but significantly different **uncertainties**

Method | Overview

❖ Incorporating uncertainty quantification into the evaluation process



Method | Data Preparation

❖ We consider five typical NLP tasks, each with 10,000 instances

Question Answering (QA)

- Answer a given question
- Built upon **MMLU**

Reading Comprehension (RC)

- Answer a given question based on a provided passage
- Built upon **CosmosQA**

Commonsense Inference (CI)

- Pick up the best ending based on commonsense reasoning
- Built upon **HellaSwag**

Dialogue Response Selection (DRS)

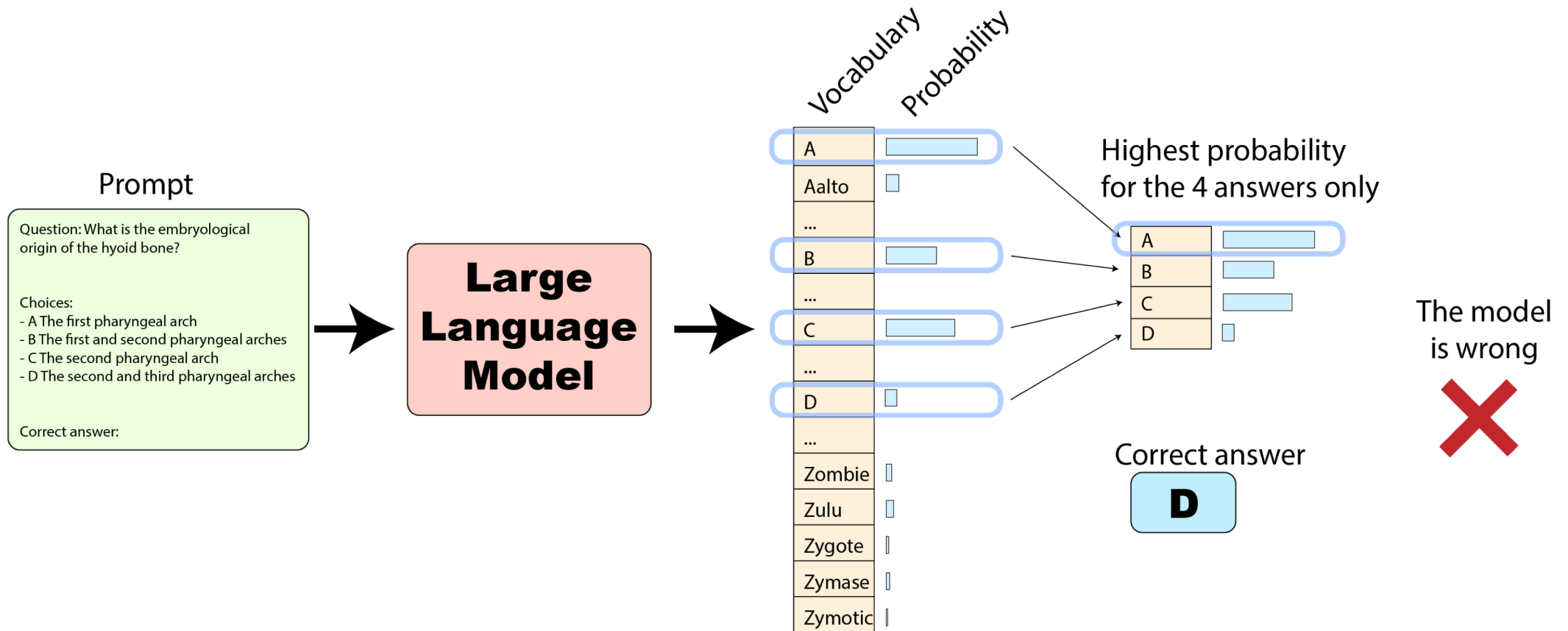
- Select the most suitable response
- Built upon **HaluDial**

Document Summarization (DS)

- Summarize long text into a shorter one
- Built upon **HaluSum**

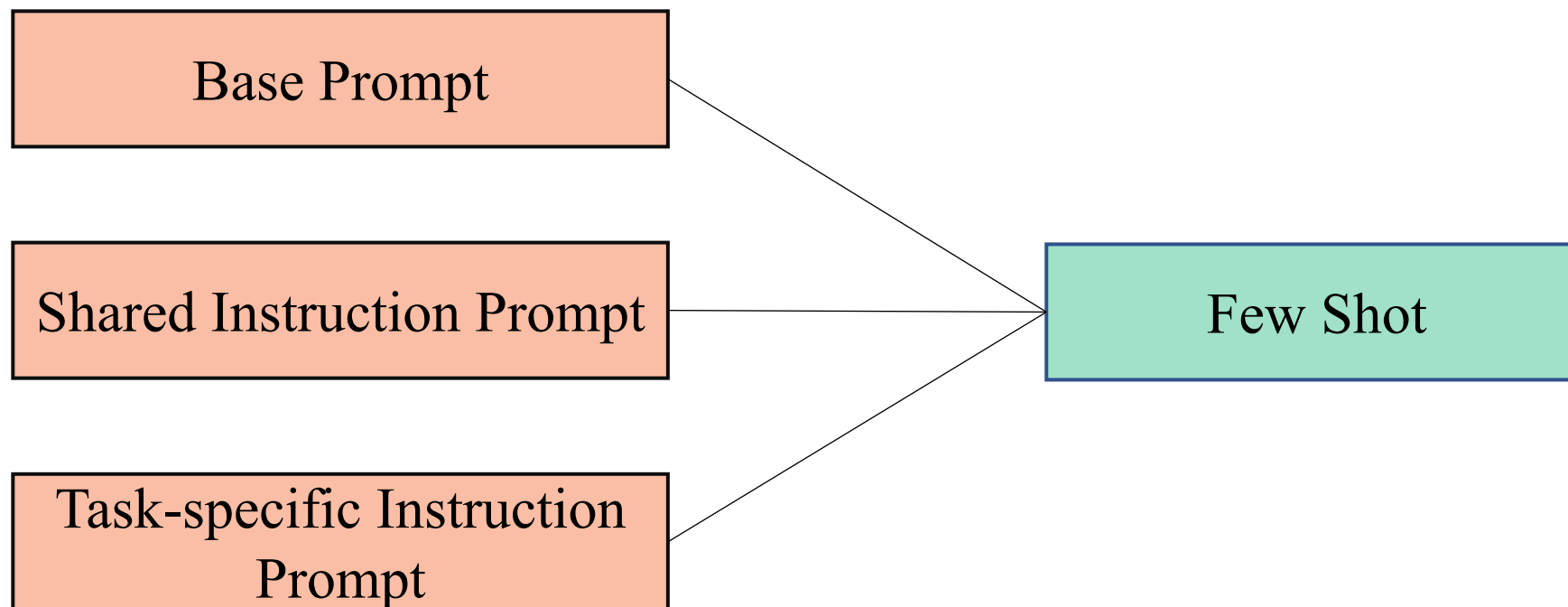
Method | Data Preparation

❖ We formulate each task as a multiple-choice question answering task



Method | Prompting Strategies

- ❖ We adopt three prompting methods to reduce the influence of LLMs' sensitivity to different prompts



Method | Uncertainty Quantification

❖ What kind of uncertainty quantification methods are friendly to LLMs?

Ease of implementation

High efficiency

High Interpretability

Data distribution-free

Model-agnostic

No model modifications

**A statistically rigorous estimation of uncertainty
rather than a heuristic approximation**

Take accuracy into account

Method | Uncertainty Quantification

❖ Conformal prediction for uncertainty quantification



{ fox
squirrel
0.99 }



{ fox, gray, bucket, rain
squirrel, fox, 0.02, barrel
0.82, 0.03, 0.02 }

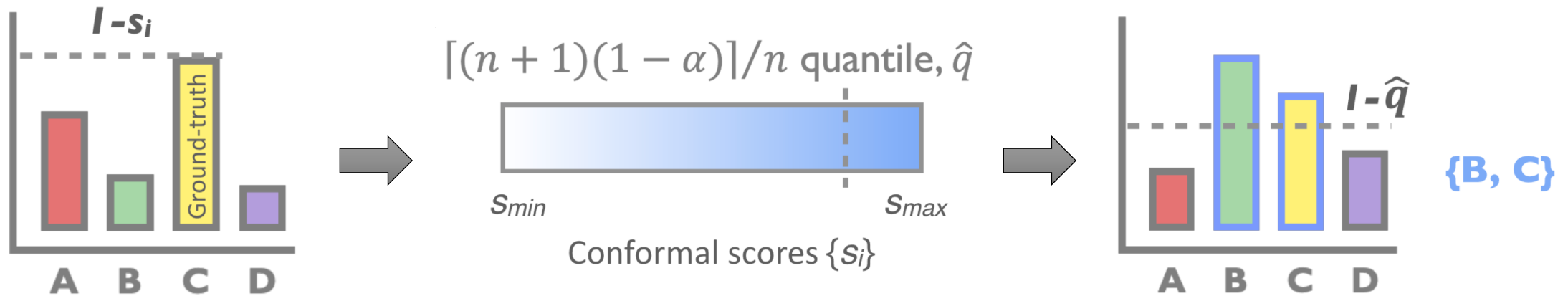


{ marmot, fox, mink, weasel, beaver, polecat
0.30, squirrel, 0.18, 0.16, 0.03, 0.01
0.22 }

$$p(Y_{test} \in \mathcal{C}(X_{test})) \geq 1 - \alpha$$

Method | Uncertainty Quantification

❖ Conformal prediction for uncertainty quantification



Step 1: Compute uncertainty scores on calibration data

Step 2: Compute the threshold

Step 3: Construct prediction sets for test instances

Benchmarking Results

- ❖ Higher accuracy does not necessarily indicate lower uncertainty
- For each task, Acc and SS lead to different rankings of LLMs

Accuracy

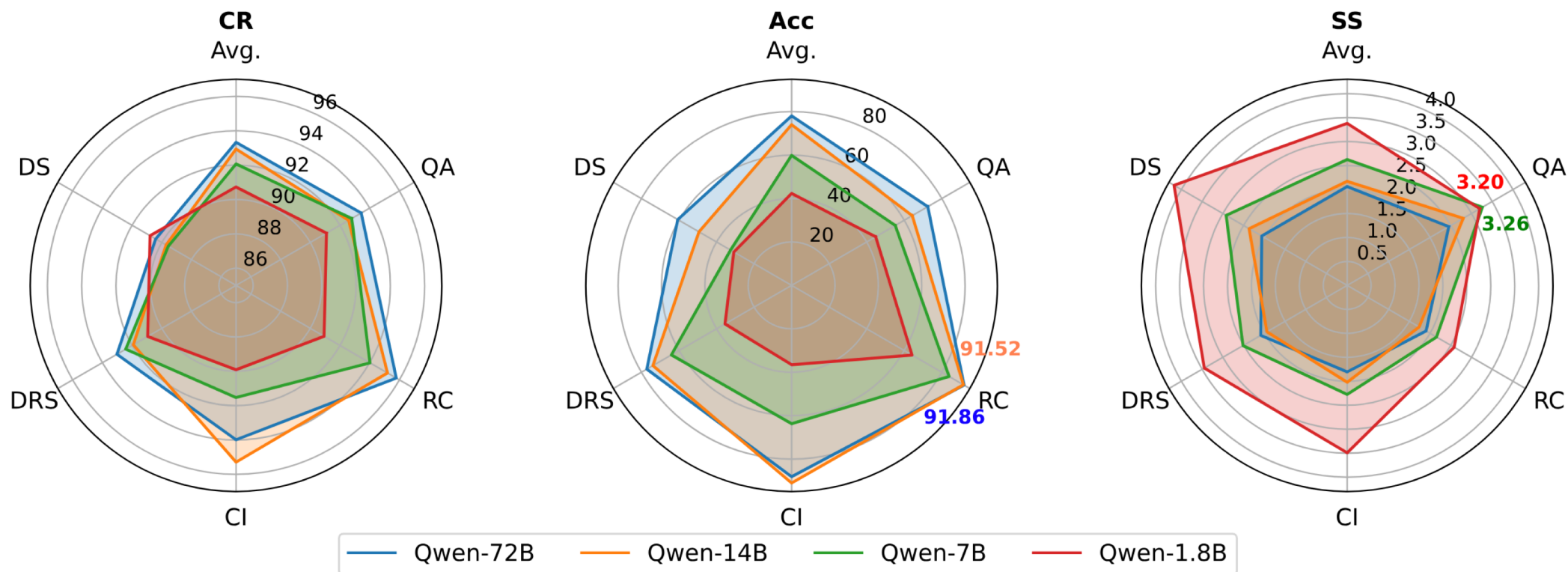
Uncertainty

LLMs	Acc (%) ↑						SS ↓					
	QA	RC	CI	DRS	DS	Avg.	QA	RC	CI	DRS	DS	Avg.
Qwen-14B	64.25 ₍₁₎	91.52 ₍₁₎	91.00 ₍₁₎	73.90 ₍₁₎	49.33 ₍₄₎	74.00 ₍₁₎	2.80 ₍₂₎	1.74 ₍₁₎	2.02 ₍₂₎	1.94 ₍₁₎	2.37 ₍₃₎	2.17 ₍₁₎
Yi-6B	57.57 ₍₄₎	85.99 ₍₂₎	76.50 ₍₂₎	58.72 ₍₄₎	66.06 ₍₁₎	68.97 ₍₂₎	3.20 ₍₅₎	1.92 ₍₄₎	1.88 ₍₁₎	2.85 ₍₆₎	1.96 ₍₁₎	2.36 ₍₂₎
Gemma-7B	62.24 ₍₂₎	85.29 ₍₃₎	73.58 ₍₃₎	66.79 ₍₂₎	40.80 ₍₇₎	65.74 ₍₃₎	2.72 ₍₁₎	1.88 ₍₃₎	2.04 ₍₃₎	2.14 ₍₂₎	3.11 ₍₇₎	2.38 ₍₃₎
Mistral-7B	60.44 ₍₃₎	81.94 ₍₅₎	62.93 ₍₅₎	53.21 ₍₅₎	62.16 ₍₂₎	64.14 ₍₄₎	2.80 ₍₂₎	1.75 ₍₂₎	2.48 ₍₅₎	2.71 ₍₅₎	2.40 ₍₄₎	2.43 ₍₄₎
Llama-2-13B	52.52 ₍₆₎	77.23 ₍₆₎	59.66 ₍₆₎	52.65 ₍₆₎	60.05 ₍₃₎	60.42 ₍₅₎	3.06 ₍₄₎	2.24 ₍₇₎	2.72 ₍₆₎	2.55 ₍₄₎	2.24 ₍₂₎	2.56 ₍₅₎
Qwen-7B	55.21 ₍₅₎	83.89 ₍₄₎	63.70 ₍₄₎	64.04 ₍₃₎	32.53 ₍₉₎	59.87 ₍₆₎	3.26 ₍₇₎	2.15 ₍₅₎	2.28 ₍₄₎	2.51 ₍₃₎	2.92 ₍₅₎	2.63 ₍₆₎
InternLM-7B	48.37 ₍₇₎	73.86 ₍₇₎	46.21 ₍₇₎	43.72 ₍₇₎	34.38 ₍₈₎	49.31 ₍₇₎	3.49 ₍₉₎	2.19 ₍₆₎	3.28 ₍₉₎	3.63 ₍₁₀₎	4.47 ₍₁₁₎	3.41 ₍₉₎
Llama-2-7B	45.60 ₍₉₎	65.79 ₍₈₎	43.05 ₍₈₎	32.61 ₍₉₎	45.60 ₍₅₎	46.53 ₍₈₎	3.20 ₍₅₎	2.39 ₍₈₎	3.27 ₍₈₎	3.26 ₍₇₎	3.30 ₍₈₎	3.09 ₍₇₎
DeepSeek-7B	45.65 ₍₈₎	65.39 ₍₉₎	42.66 ₍₉₎	33.50 ₍₈₎	42.15 ₍₆₎	45.87 ₍₉₎	3.34 ₍₈₎	2.77 ₍₉₎	3.06 ₍₇₎	3.40 ₍₈₎	3.08 ₍₆₎	3.13 ₍₈₎
MPT-7B	29.49 ₍₁₀₎	31.69 ₍₁₀₎	25.50 ₍₁₀₎	24.38 ₍₁₁₎	24.86 ₍₁₀₎	27.18 ₍₁₀₎	3.53 ₍₁₀₎	3.46 ₍₁₀₎	3.60 ₍₁₀₎	3.59 ₍₉₎	3.66 ₍₉₎	3.57 ₍₁₀₎
Falcon-7B	23.75 ₍₁₁₎	24.98 ₍₁₁₎	24.91 ₍₁₁₎	25.86 ₍₁₀₎	24.69 ₍₁₁₎	24.84 ₍₁₁₎	3.90 ₍₁₁₎	3.60 ₍₁₁₎	3.66 ₍₁₁₎	3.64 ₍₁₁₎	3.92 ₍₁₀₎	3.75 ₍₁₁₎

Benchmarking Results

❖ Effects of model scale

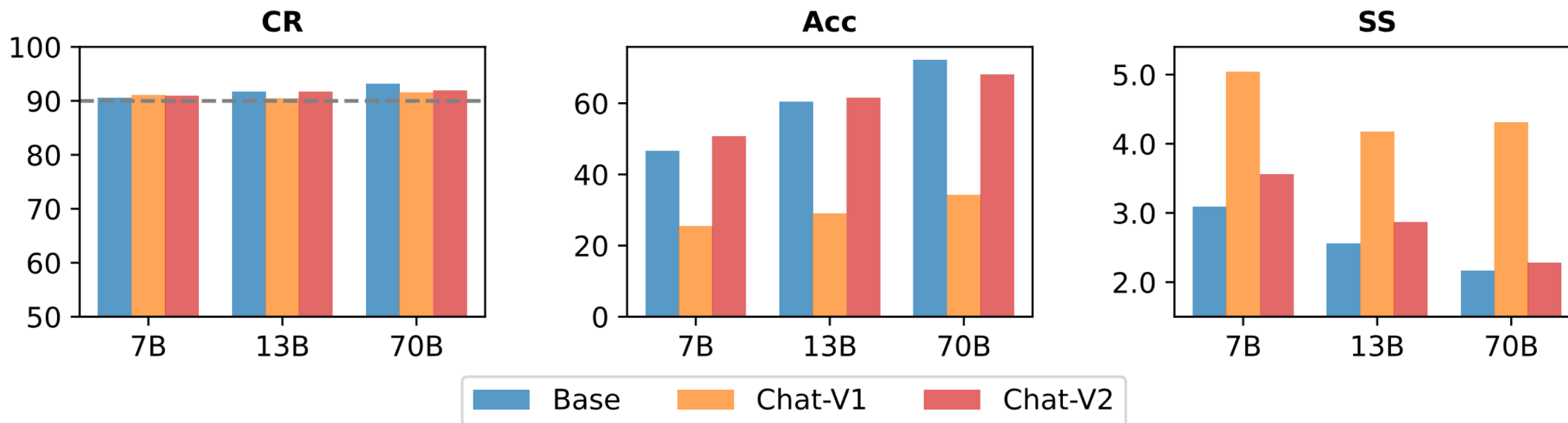
- ❑ Larger-scale LLMs may display greater uncertainty compared to smaller counterparts



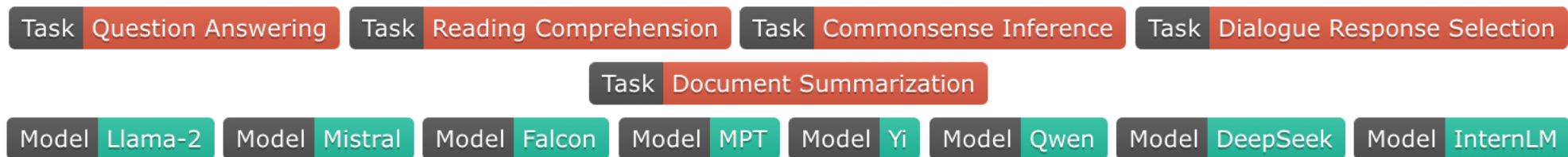
Benchmarking Results

❖ Effects of instruction finetuning

☐ Instruction-finetuning tends to increase the uncertainty of LLMs



Benchmarking LLMs via Uncertainty Quantification



 [Paper](#),  [Datasets](#)

<https://github.com/smartyfh/LLM-Uncertainty-Bench>

Thank You!
Q & A