NEURAL INFORMATION PROCESSING SYSTEMS

Benchmarking LLMs via Uncertainty Quantification

Fanghua Ye[1,2], *Mingming Yang[1]*, *Jianhui Pang[1,3]*, *Longyue Wang[1]*,
Derek F. Wong[3], *Emine Yilmaz[2]*, *Shuming Shi[1]*, *Zhaopeng Tu[1]*

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

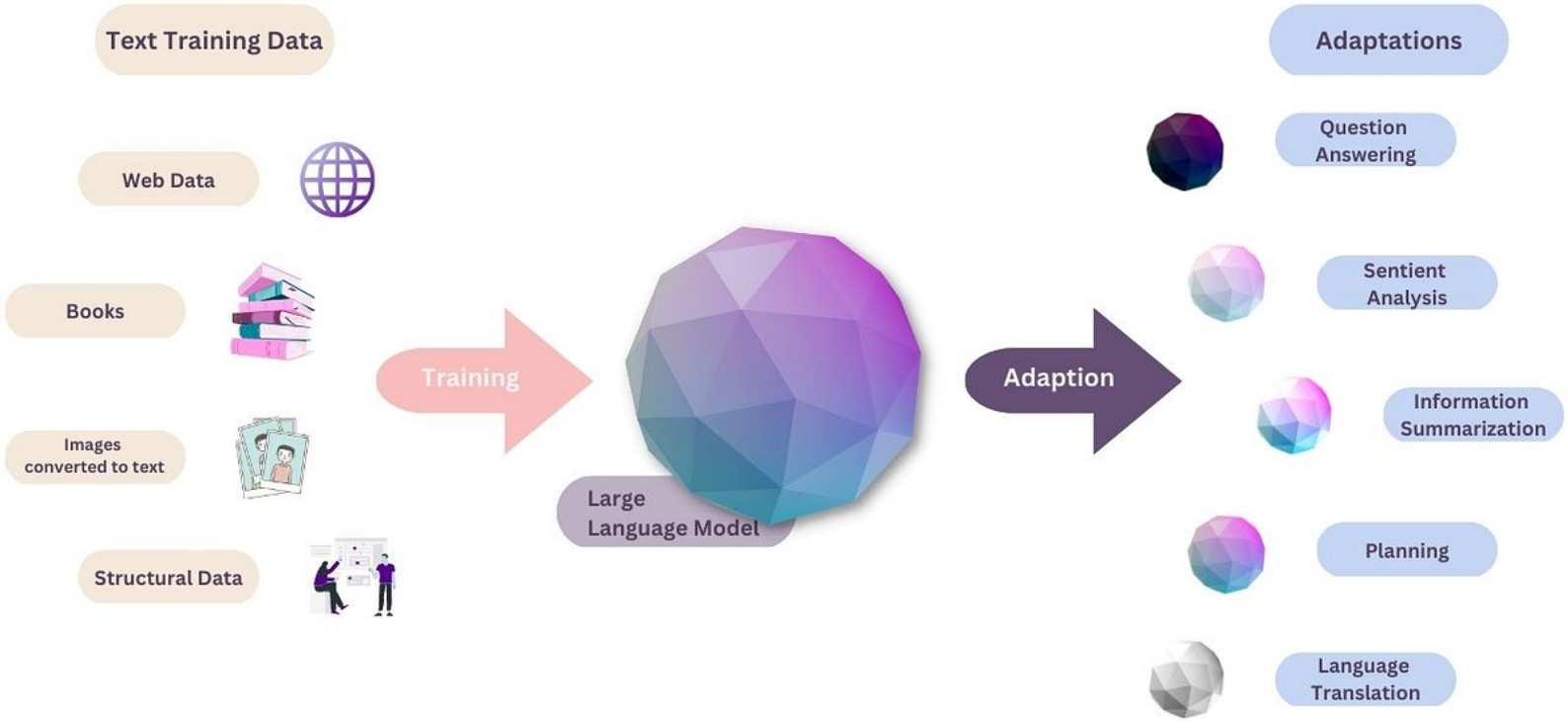
Tencent AI Lab

UCL

UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

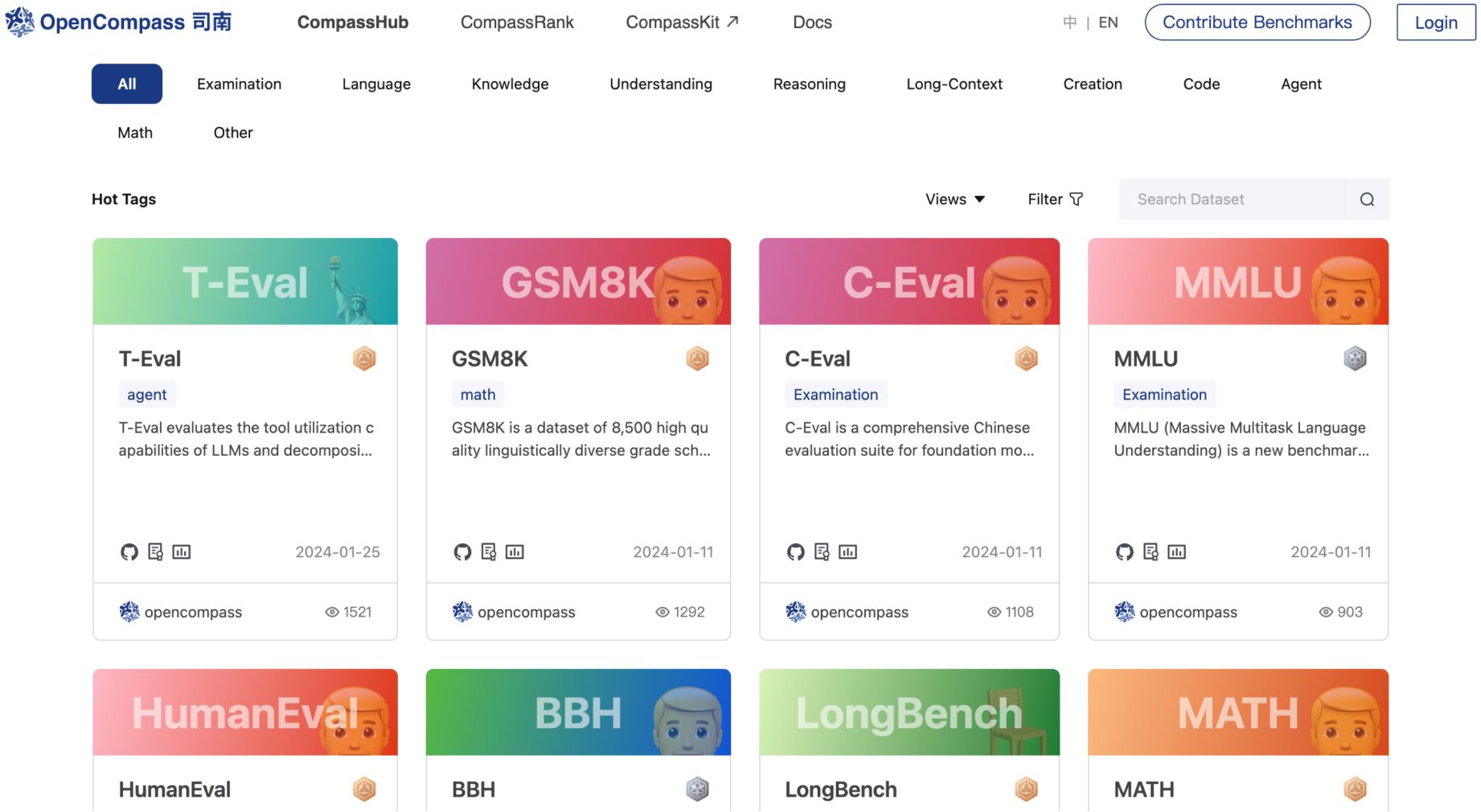
Background

- LLMs are able to show remarkable performance across various tasks



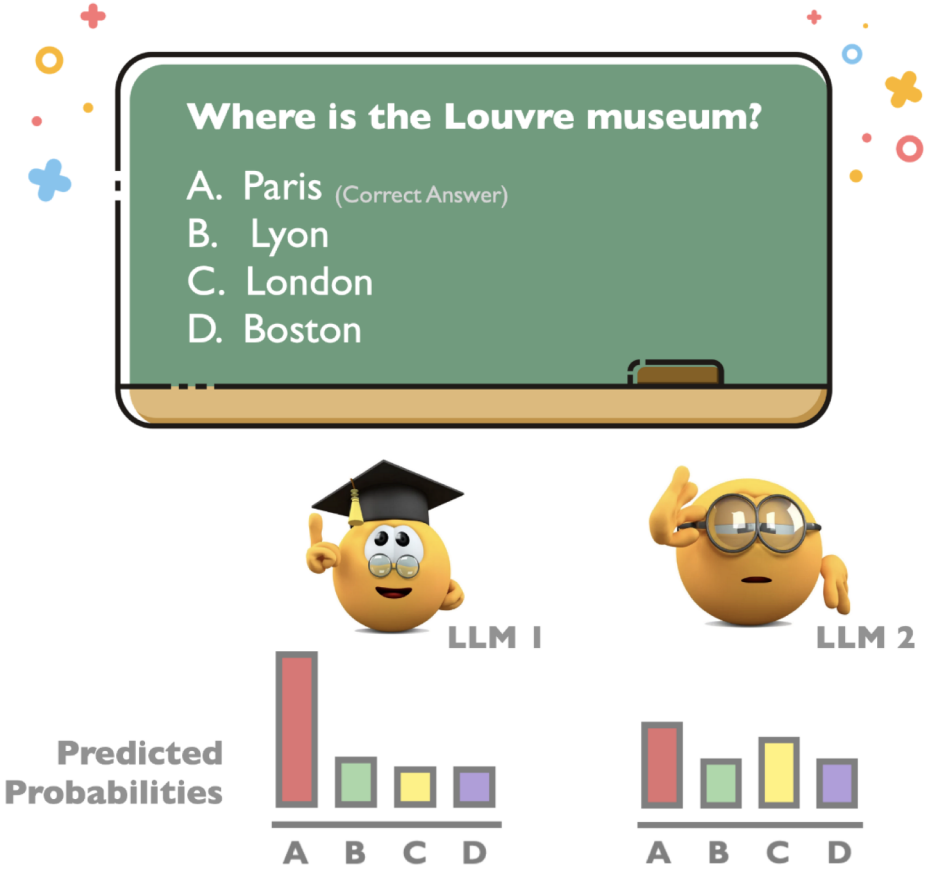
Background

- Comprehensively evaluating LLMs becomes a huge challenge



Motivation

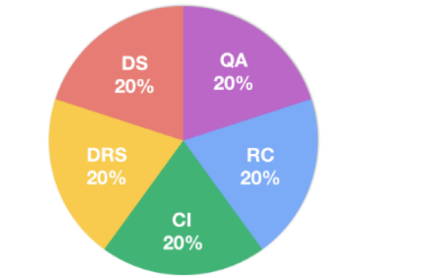
- Existing evaluation overlooks the uncertainty of LLMs
  - LLMs should be aware of what they are unsure about



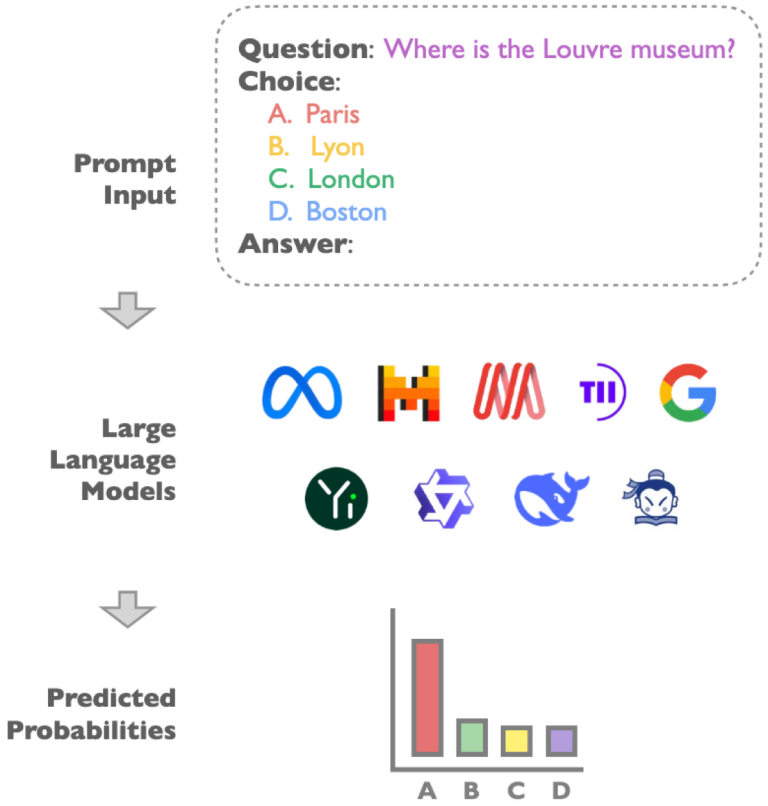
Two LLMs can demonstrate the same accuracy but significantly different **uncertainties**

Method | Overview

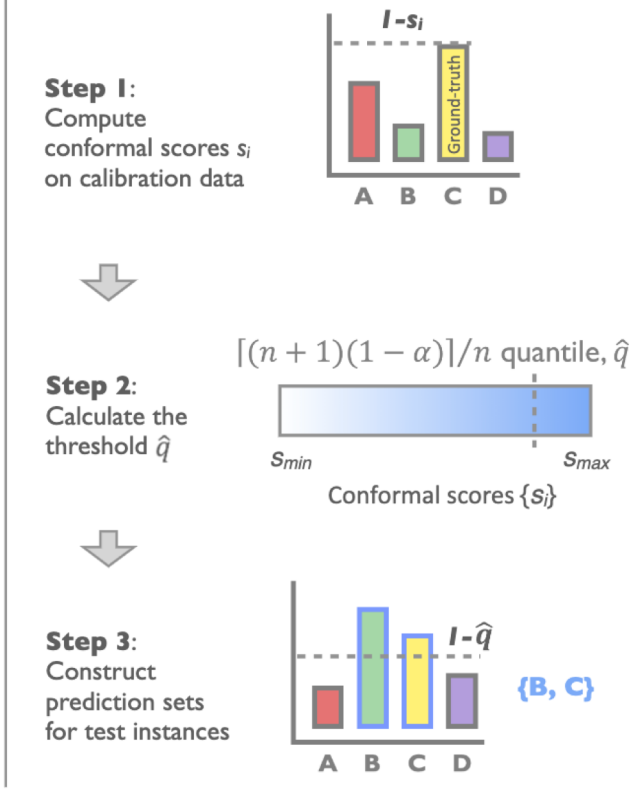
- Incorporating uncertainty quantification into the evaluation process



(a) Task data preparation



(b) Prompt LLMs to obtain predicted probabilities of options for all data points



(c) Generate prediction sets on test data based on conformal prediction

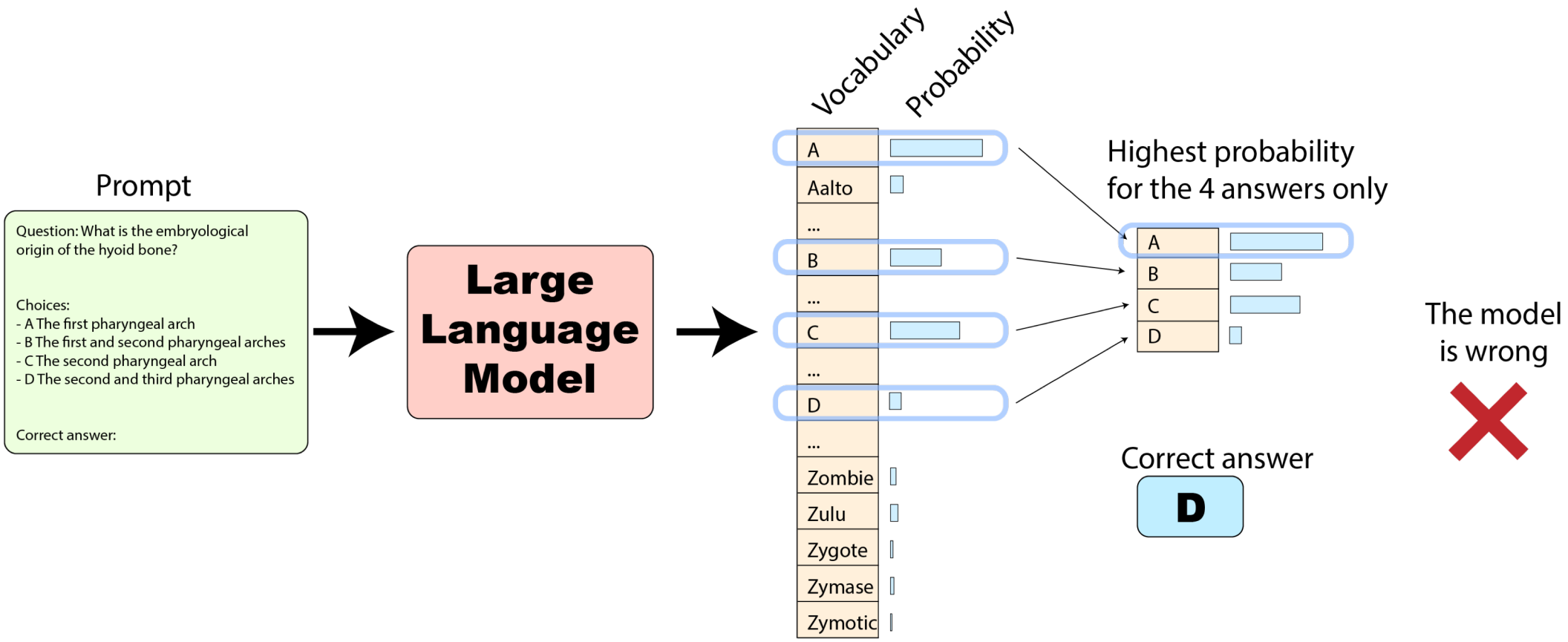
Method | Data Preparation

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

Question Answering (QA)	Reading Comprehension (RC)	Commonsense Inference (CI)	Dialogue Response Selection (DRS)	Document Summarization (DS)
• Answer a given question • Built upon **MMLU**	• Answer a given question based on a provided passage • Built upon **CosmosQA**	• Pick up the best ending based on commonsense reasoning • Built upon **HellaSwag**	• Select the most suitable response • Built upon **HaluDial**	• Summarize long text into a shorter one • Built upon **HaluSum**

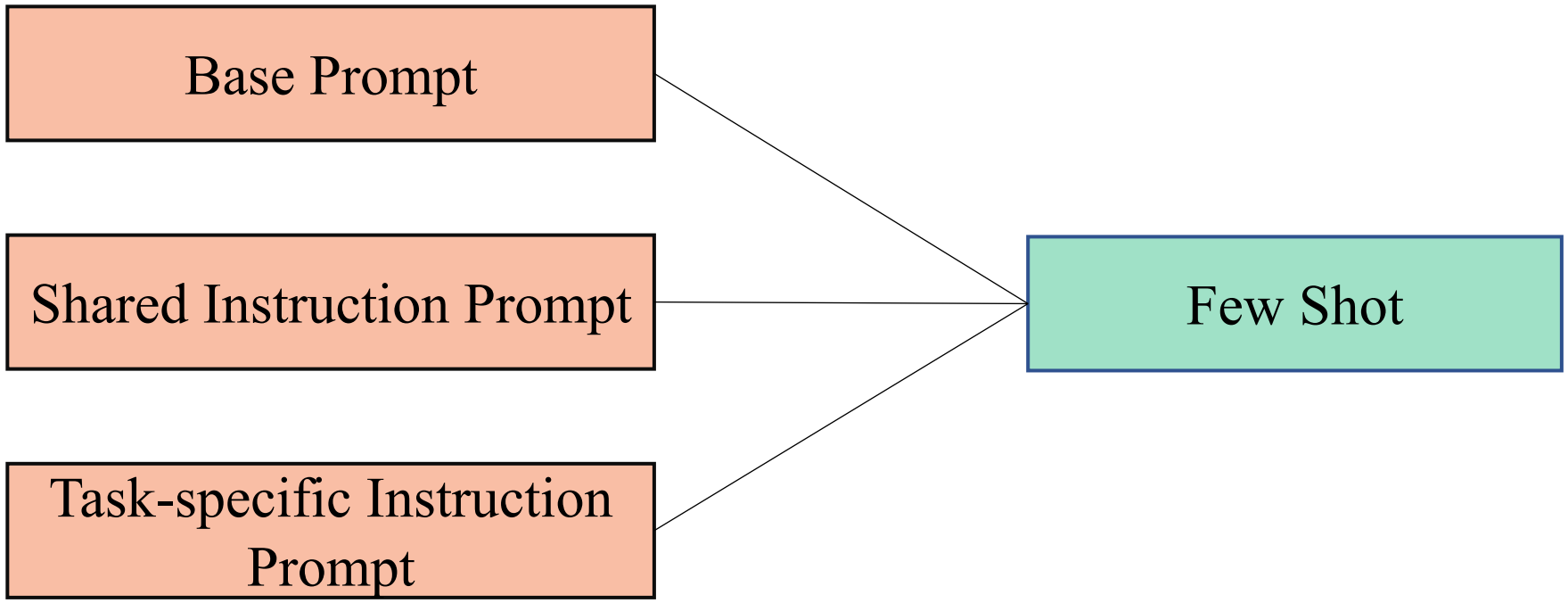
Method | Data Preparation

- 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 }

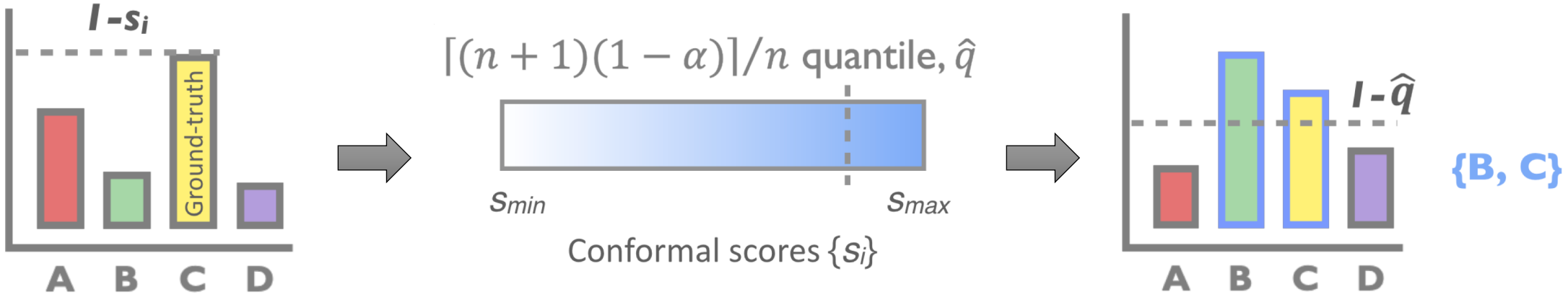
{ fox squirrel 0.82, gray fox, 0.03, bucket, 0.02, rain barrel 0.02 }

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

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

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

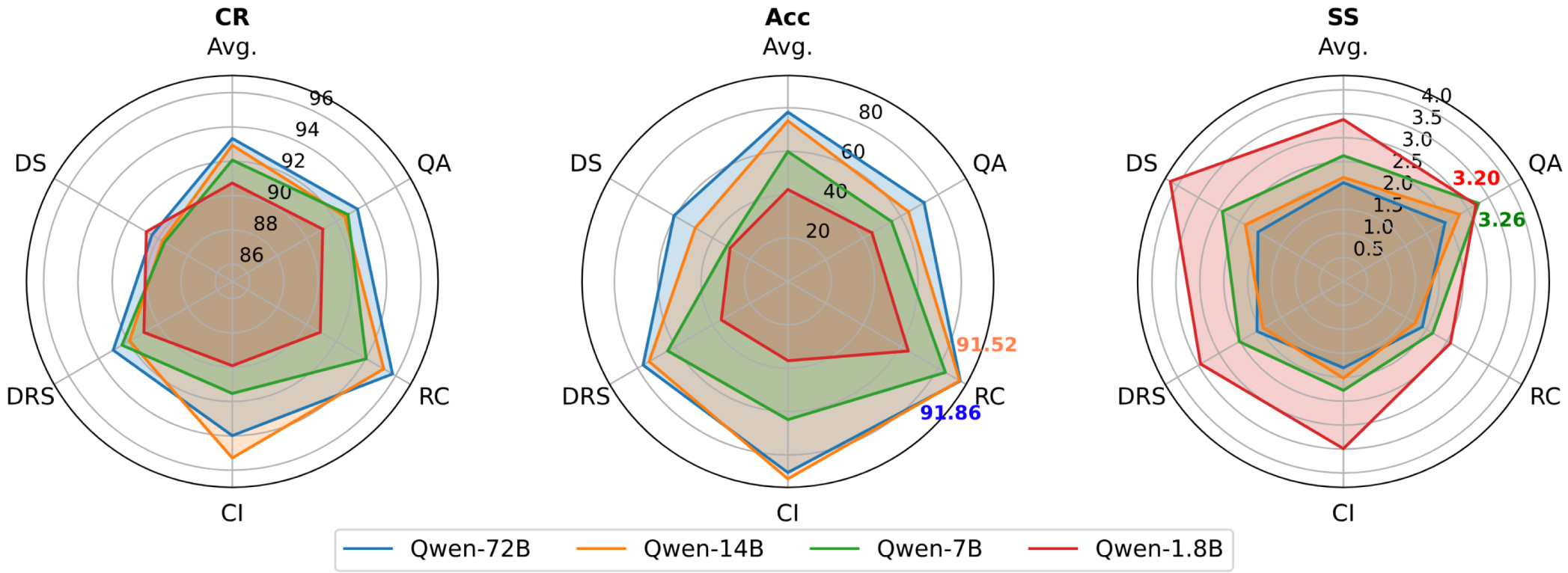
- Higher accuracy does not necessarily indicate lower uncertainty
  - For each task, Acc and SS lead to different rankings of LLMs

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)

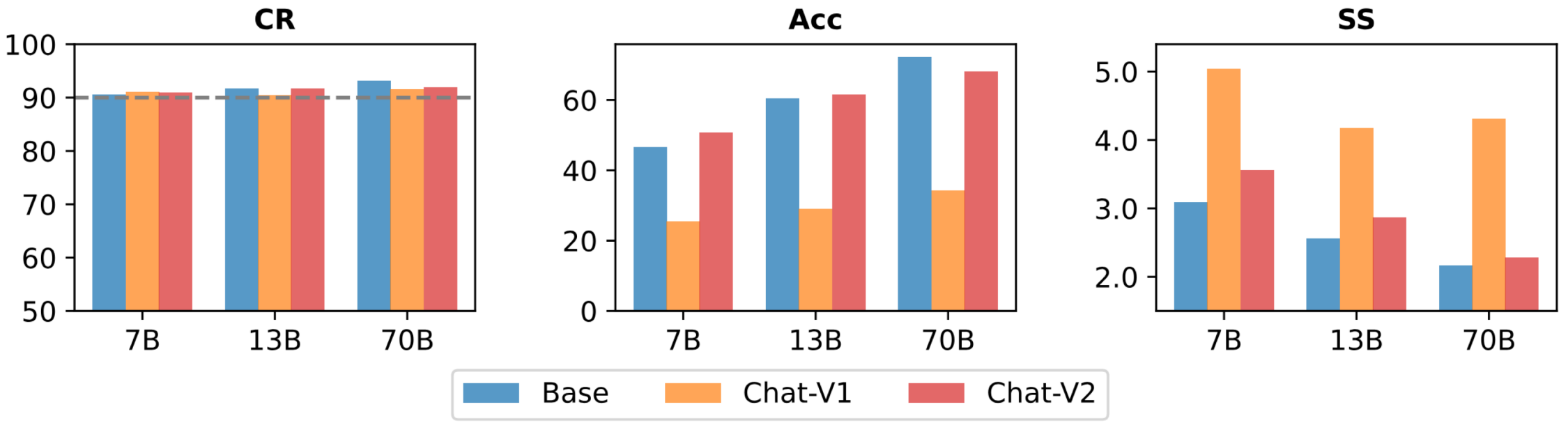
Benchmarking Results

- Effects of model scale
  - Larger-scale LLMs may display greater uncertainty compared to smaller counterparts

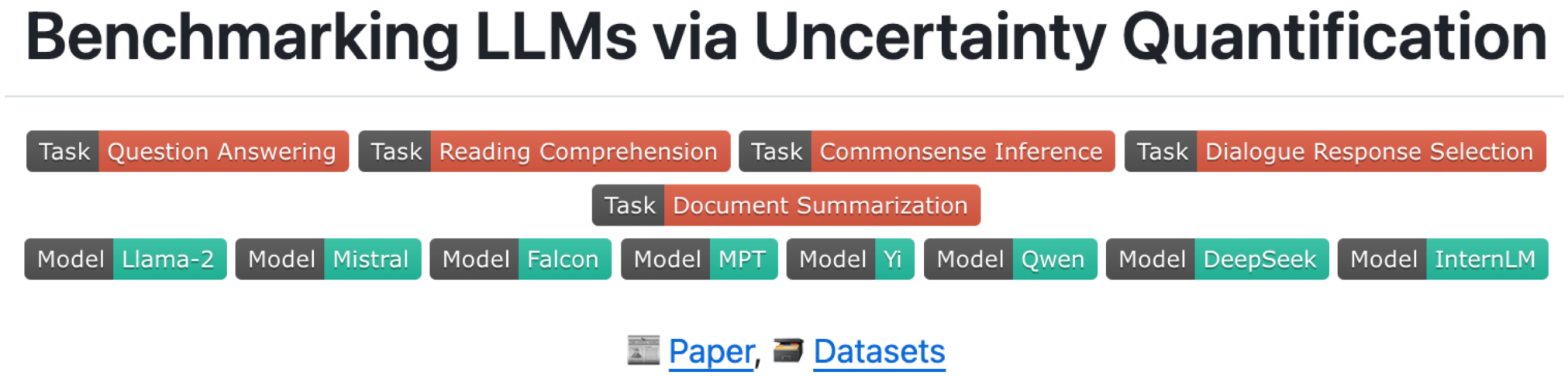


Benchmarking Results

- Effects of instruction finetuning
  - Instruction-finetuning tends to increase the uncertainty of LLMs



Open Source



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

Thank You!
Q & A