Table 1: LLM Uncertainty Detection: Datasets and Models

Category	Dataset/Model	Description	Year/Link
1. Knowledge Uncertainty			
Datasets	SelfAware	1,032 unanswerable + 2,337 answerable questions for evaluating LLM self-knowledge	2023 [GitHub]
	TruthfulQA	Measures model truthfulness and ability to avoid false information	2021 [Link]
	RepLiQA	Human-crafted fictional scenarios absent from internet for testing unseen content	2024 [HF]
	SQuAD 2.0	100k questions + 50k adversarial unanswerable questions	2018 [Link]
Methods	R-tuning	Assesses knowledge gap between parametric knowledge and instruction data	2023 [Paper]
	Knowledge Boundary Model (KBM)	Trains model to identify different types of questions based on knowledge limits	2024
	Self-Knowledge Guided RAG (SKR)	Adaptive external resource calling based on previous encounters	2024
2. Capacity/Reasoning Uncertainty			
Datasets	UMWP	5,200 math word problems (2,600 answerable + 2,600 unanswerable)	2024 [GitHub]
	MuSR	Multi-step reasoning for murder mysteries, object placement, team allocation	2024 [GitHub]
Methods	Epistemic Neural Networks (ENN)	Small networks attached to frozen LLMs for uncertainty estimation	2023 [Paper]
	Operational Uncertainty Detection	Identifies errors in response generation process	2025 [Paper]
3. Mixed/Comprehensive Uncertainty			
Datasets	FactGuard-Bench	25,220 answerable/unanswerable scenarios via multi-agent framework	2025 [Paper]
	UNK-VQA	Visual QA dataset with unanswerable questions	2024 [GitHub]
Methods	SPUQ	Perturbation-based uncertainty quantification, 50% ECE reduction	2024 [Paper]
	Semantic Entropy Probes (SEPs)	Estimates semantic entropy from hidden states	2024 [Paper]
	Information-theoretic Methods	Detects high epistemic uncertainty for hallucination detection	2024 [Paper]
	Uncertainty Profiles	Decomposes into input, reasoning, parameter, prediction uncertainties	2025 [Paper]
	Multi-LLM Collaboration	Identifies knowledge gaps through LLM collaboration	2024 [ACL]
Key Resources			
	Awesome-LLM-Uncertainty	Curated list of uncertainty/reliability research	2024 [GitHub]
	LLM-Uncertainty Code	ICLR 2024 "Can LLMs Express Their Uncertainty?" implementation	2024 [GitHub]

Uncertainty metrics and AUARC. For a completion i with length T_i and stepwise next-token distributions $\{P_{i,t}(\cdot)\}_{t=1}^{T_i}$, let the realized token probabilities be $p_{i,t} = P_{i,t}(y_{i,t})$ and $\ell_{i,t} = \log p_{i,t}$. We rank items for *rejection* by an uncertainty score $u(\cdot)$ (higher u = more uncertain). We use:

- **GeoProb** (geometric mean probability; confidence):

$$\mathrm{GM}_i \;=\; \Big(\prod_{t=1}^{T_i} p_{i,t}\Big)^{1/T_i} \;=\; \exp\Big(\tfrac{1}{T_i}\sum_{t=1}^{T_i} \ell_{i,t}\Big).$$

  As an uncertainty score we take $u_i^{\mathrm{geo}} = -\mathrm{GM}_i$.

- **Mean LogProb** (length-normalized log-likelihood; confidence):

$$\overline{\log p_i} \;=\; \tfrac{1}{T_i}\sum_{t=1}^{T_i} \ell_{i,t},$$

  with uncertainty $u_i^{\mathrm{mlp}} = -\,\overline{\log p_i}$.

- **Min LogProb** (worst token surprise; uncertainty):

$$u_i^{\mathrm{minlp}} \;=\; -\min_{1\le t\le T_i} \ell_{i,t} \;=\; \max_t \big(-\ell_{i,t}\big).$$

- **LastTok Prob** (terminal token confidence):

$$\mathrm{LastTokProb}_i \;=\; p_{i,T_i}, \qquad u_i^{\mathrm{lastp}} = -\,\mathrm{LastTokProb}_i.$$

- **Mean Entropy** (dispersion across steps; uncertainty):

$$\overline{H}_i \;=\; \tfrac{1}{T_i}\sum_{t=1}^{T_i} H\big(P_{i,t}\big), \qquad u_i^{\mathrm{meanH}} = \overline{H}_i,$$

  where $H(P) = -\sum_v P(v)\log P(v)$.

- **Last Entropy** (dispersion at the final step; uncertainty):

$$u_i^{\mathrm{lastH}} \;=\; H\big(P_{i,T_i}\big).$$

Given a rejection rate $r \in [0,1]$, let S_r be the non-rejected set after removing the top r fraction by $u(\cdot)$. The conditional accuracy

$$A(r) \;=\; \frac{1}{|S_r|}\sum_{i\in S_r} \mathbf{1}\{\text{completion } i \text{ is correct}\}, \qquad E(r) = 1 - A(r).$$

The **AUARC@0–20%** we report is the area under $A(r)$ from $r = 0$ to 0.2, normalized to $[0,1]$. With the five grid points $\mathcal{R} = \{0, 0.05, 0.10, 0.15, 0.20\}$, we compute

$$\mathrm{AUARC@0\text{–}20\%} \;=\; \frac{1}{|\mathcal{R}|}\sum_{r\in\mathcal{R}} A(r) \;\approx\; \frac{1}{0.2}\int_0^{0.2} A(r)\,dr.$$

Why GeoProb equals Mean LogProb (in AUARC/$A(r)$). Since $\mathrm{GM}_i = \exp(\overline{\log p_i})$, GeoProb is a strictly monotone transform of Mean LogProb. Sorting by $-\mathrm{GM}_i$ or by $-\overline{\log p_i}$ yields the same ranking, so the rejected sets at every r are identical.

Consistency of $A(0\%)$. At $r = 0$ no items are rejected: S_0 is the full evaluation set. Therefore $A(0\%)$ must be the same across all metrics for a fixed (task, model).

Table 2: AUARC (0–20%) and Conditional Accuracy A(r) across Tasks

Task	Models	Metric	A@0%	$\Delta A(r)$ from $A(0)$				AUARC@0–20%
				5%	10%	15%	20%	
gsm8k	DS-R1	GeoProb	79.26	**+2.75**	+3.51	**+5.13**	+5.25	**82.59**
		Mean LogProb		**+2.75**	+3.51	**+5.13**	+5.25	**82.59**
		Min LogProb		-0.01	-0.23	-0.66	-1.34	78.82
		LastTok Prob		+0.87	+1.87	+2.55	+2.83	80.89
		Mean Entropy		+2.31	**+3.64**	+4.64	**+5.34**	82.45
		Last Entropy		+0.83	+1.87	+2.59	+2.83	80.89
	Gemma-2	GeoProb	80.21	-0.40	-0.84	-1.66	-2.33	79.17
		Mean LogProb		-0.40	-0.84	-1.66	-2.33	79.17
		Min LogProb		**+0.28**	**+0.50**	**+1.37**	**+2.07**	**81.06**
		LastTok Prob		+0.16	+0.29	+0.17	+0.13	80.36
		Mean Entropy		-0.32	-0.59	-1.52	-1.91	79.34
		Last Entropy		-0.40	-1.18	-1.79	-2.52	79.03
	InternLM-2.5	GeoProb	63.50	**+3.24**	**+6.74**	**+10.16**	**+12.72**	**70.07**
		Mean LogProb		**+3.24**	**+6.74**	**+10.16**	**+12.72**	**70.07**
		Min LogProb		+2.32	+5.26	+8.55	+11.73	69.07
		LastTok Prob		+2.92	+6.06	+9.98	**+12.72**	69.83
		Mean Entropy		+0.69	+3.54	+7.08	+11.02	67.96
		Last Entropy		+1.32	+1.14	+1.37	+1.64	64.59
	Llama-3	GeoProb	75.55	+1.83	+2.60	**+3.36**	+4.32	77.97
		Mean LogProb		+1.83	+2.60	**+3.36**	+4.32	77.97
		Min LogProb		-0.16	+0.16	+0.15	+0.20	75.62
		LastTok Prob		+0.08	+0.32	+0.06	-0.09	75.63
		Mean Entropy		**+1.95**	**+3.19**	+3.14	**+4.37**	**78.08**
		Last Entropy		+0.24	+0.11	-0.29	-0.14	75.53
	Mistral	GeoProb	49.43	+0.83	+1.35	+2.02	**+2.44**	50.76
		Mean LogProb		+0.83	+1.35	+2.02	**+2.44**	50.76
		Min LogProb		-0.13	-1.05	-0.66	-0.40	48.98
		LastTok Prob		+0.91	+1.18	+1.08	+0.83	50.23
		Mean Entropy		**+0.99**	**+1.52**	**+2.11**	**+2.44**	**50.84**
		Last Entropy		+0.83	+1.14	+1.26	+0.31	50.14
	Qwen-2.5	GeoProb	79.23	+0.27	**+0.98**	**+1.60**	+1.92	**80.18**
		Mean LogProb		+0.27	**+0.98**	**+1.60**	+1.92	**80.18**
		Min LogProb		+0.27	+0.39	-0.09	-0.78	79.19
		LastTok Prob		**+0.59**	+0.82	+1.07	+0.92	79.91
		Mean Entropy		+0.03	+0.90	+1.47	**+1.97**	80.10
		Last Entropy		+0.23	+0.48	+1.02	+1.11	79.80
humaneval	DS-R1	GeoProb	11.59	-0.05	**-0.77**	**-0.16**	**+0.54**	**11.50**
		Mean LogProb		-0.05	**-0.77**	**-0.16**	**+0.54**	**11.50**
		Min LogProb		**+0.59**	**-0.77**	-0.87	-0.22	11.33
		LastTok Prob		-0.05	**-0.77**	**-0.16**	-0.98	11.19
		Mean Entropy		-0.69	**-0.77**	**-0.16**	**+0.54**	11.37
		Last Entropy		-1.33	**-0.77**	**-0.16**	-0.98	10.94
	Gemma-2	GeoProb	56.71	-0.94	-1.98	-2.42	-2.92	55.06
		Mean LogProb		-0.94	-1.98	-2.42	-2.92	55.06
		Min LogProb		**+0.34**	**+0.73**	**+1.15**	**+0.11**	**57.17**
		LastTok Prob		-0.94	-0.63	-0.28	-0.65	56.21
		Mean Entropy		-0.94	-0.63	-1.71	-1.40	55.77

(Continued on next page)

(Continued from previous page)

Task	Models	Metric	A@0%	$\Delta A(r)$ from $A(0)$				AUARC@0–20%
				5%	10%	15%	20%	
mbpp	InternLM-2.5	Last Entropy	48.17	-0.94	-0.63	-0.99	**+0.11**	56.22
		GeoProb		+1.19	**+2.50**	**+3.26**	**+4.86**	**50.53**
		Mean LogProb		+1.19	**+2.50**	**+3.26**	**+4.86**	**50.53**
		Min LogProb		+0.55	+1.83	**+3.26**	**+4.86**	50.27
		LastTok Prob		**+1.83**	+1.15	+2.54	+2.59	49.79
		Mean Entropy		-0.09	+0.48	-0.31	-0.44	48.10
	Llama-3	Last Entropy	28.05	-1.38	-0.87	-2.46	-1.20	46.99
		GeoProb		+0.16	+1.01	**+1.24**	+2.25	28.98
		Mean LogProb		+0.16	+1.01	**+1.24**	+2.25	28.98
		Min LogProb		-0.48	-0.35	-0.19	-0.02	27.84
		LastTok Prob		+0.16	+0.33	+0.52	-0.78	28.10
		Mean Entropy		**+0.80**	+1.01	+0.52	+1.50	28.81
	Mistral	Last Entropy	3.66	+0.16	**+1.68**	**+1.24**	**+3.01**	**29.27**
		GeoProb		**+0.19**	**+0.40**	**+0.63**	+0.13	3.93
		Mean LogProb		**+0.19**	**+0.40**	**+0.63**	+0.13	3.93
		Min LogProb		**+0.19**	**+0.40**	**+0.63**	+0.13	3.93
		LastTok Prob		**+0.19**	**+0.40**	**+0.63**	**+0.89**	**4.08**
		Mean Entropy		**+0.19**	**+0.40**	-0.09	+0.13	3.78
		Last Entropy		**+0.19**	**+0.40**	**+0.63**	**+0.89**	**4.08**
	Qwen-2.5	GeoProb	64.63	-1.17	-1.80	**-0.35**	**-1.76**	63.62
		Mean LogProb		-1.17	-1.80	**-0.35**	**-1.76**	63.62
		Min LogProb		-0.53	**-0.44**	-1.06	-2.51	**63.72**
		LastTok Prob		**+0.11**	**-0.44**	-2.49	-3.27	63.41
		Mean Entropy		-1.17	-1.80	-3.21	-4.03	62.59
		Last Entropy		-1.17	-1.12	-1.06	-2.51	63.46
	DS-R1	GeoProb	48.00	**+2.32**	**+4.67**	+5.65	**+6.25**	**51.78**
		Mean LogProb		**+2.32**	**+4.67**	+5.65	**+6.25**	**51.78**
		Min LogProb		+1.26	+2.22	+2.82	+3.25	49.91
		LastTok Prob		+0.21	+0.67	+0.71	+1.25	48.57
		Mean Entropy		+2.11	**+4.67**	**+5.88**	+6.00	51.73
	Gemma-2	Last Entropy	58.20	+0.21	+0.44	+0.94	+1.00	48.52
		GeoProb		+0.75	**+1.80**	**+2.51**	**+2.30**	**59.67**
		Mean LogProb		+0.75	**+1.80**	**+2.51**	**+2.30**	**59.67**
		Min LogProb		-0.52	-0.42	-1.02	-1.20	57.57
		LastTok Prob		-0.09	+0.47	+1.33	+0.80	58.70
		Mean Entropy		**+0.96**	**+1.80**	+2.27	+1.80	59.57
	InternLM-2.5	Last Entropy	50.60	+0.33	+1.13	+0.86	+1.30	58.92
		GeoProb		+0.98	+0.73	-0.48	-0.10	50.83
		Mean LogProb		+0.98	+0.73	-0.48	-0.10	50.83
		Min LogProb		+0.98	**+2.51**	**+3.05**	**+3.90**	**52.69**
		LastTok Prob		**+1.19**	+0.07	+0.22	-0.10	50.88
		Mean Entropy		+0.98	+0.51	-0.25	+0.15	50.88
	Llama-3	Last Entropy	57.40	-0.49	-0.38	-0.72	-1.10	50.06
		GeoProb		+1.13	+1.71	+2.84	**+3.60**	59.25
		Mean LogProb		+1.13	+1.71	+2.84	**+3.60**	59.25
		Min LogProb		-0.35	+0.38	+0.72	+1.35	57.82
		LastTok Prob		+0.92	+1.04	+1.42	+1.10	58.30
		Mean Entropy		**+1.55**	**+2.60**	**+3.07**	+3.35	**59.51**
	Mistral	Last Entropy	41.20	+0.71	+0.82	+1.66	+0.60	58.16
		GeoProb		+0.48	**+0.80**	-0.02	**+1.30**	41.71
		Mean LogProb		+0.48	**+0.80**	-0.02	**+1.30**	41.71

(Continued on next page)

(Continued from previous page)

Task	Models	Metric	A@0%	$\Delta A(r)$ from $A(0)$				AUARC@0–20%
				5%	10%	15%	20%	
	Qwen-2.5	Min LogProb	40.80	-0.15	-0.53	-1.67	-1.20	40.49
		LastTok Prob		+0.27	+0.58	+0.68	-0.20	41.47
		Mean Entropy		**+0.69**	+0.58	+0.92	+1.05	**41.85**
		Last Entropy		+0.48	+0.13	**+1.15**	+0.30	41.61
		GeoProb		-1.43	-3.02	-5.04	-7.05	37.49
		Mean LogProb		-1.43	-3.02	-5.04	-7.05	37.49
		Min LogProb		**+0.46**	**+0.53**	**+0.38**	**-0.05**	**41.06**
		LastTok Prob		-0.80	-1.47	-2.45	-3.80	39.10
		Mean Entropy		-2.69	-4.58	-6.68	-8.80	36.25
		Last Entropy		-1.01	-2.58	-4.80	-7.05	37.71