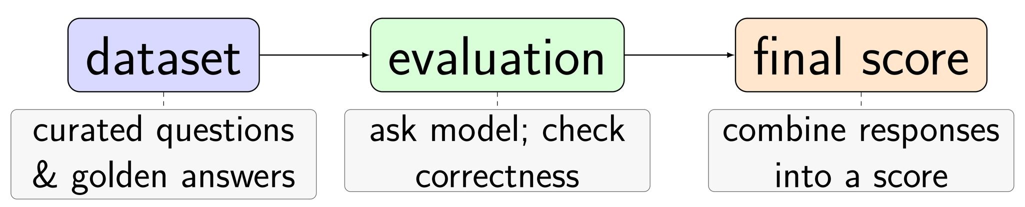
Evaluating the Unseen Capabilities: How Much Do LLMs Actually Know?

Xiang Li, Jiayi Xin, Qi Long, Weijie J. Su

University of Pennsylvania

Importance of LLM Evaluation

- AI research is now largely empirical, driven by experimentation.
- LLMs drive advances in AI and science.
- Progress fueled by LLM evaluation.
- Benchmarks and rising scores showcase gains.



Could We Trust These Scores?

MMLU (Massive Multitasks Language Understanding [3]): 16,000 multiple-choice questions across 57 academic subjects (elementary mathematics, US history, CS and law).

- **Higher scores do not imply overall superiority**: PaLM scores 69.6% > GPT-3.5's 65%, but GPT-3.5 is stronger in coding and math.
- **Scores are fragile**: Slight perturbations (e.g., changing choice orders, prompts, choice symbols) and harder perturbation.
- **Quick saturation**: LLMs plateau rapidly on many benchmarks.

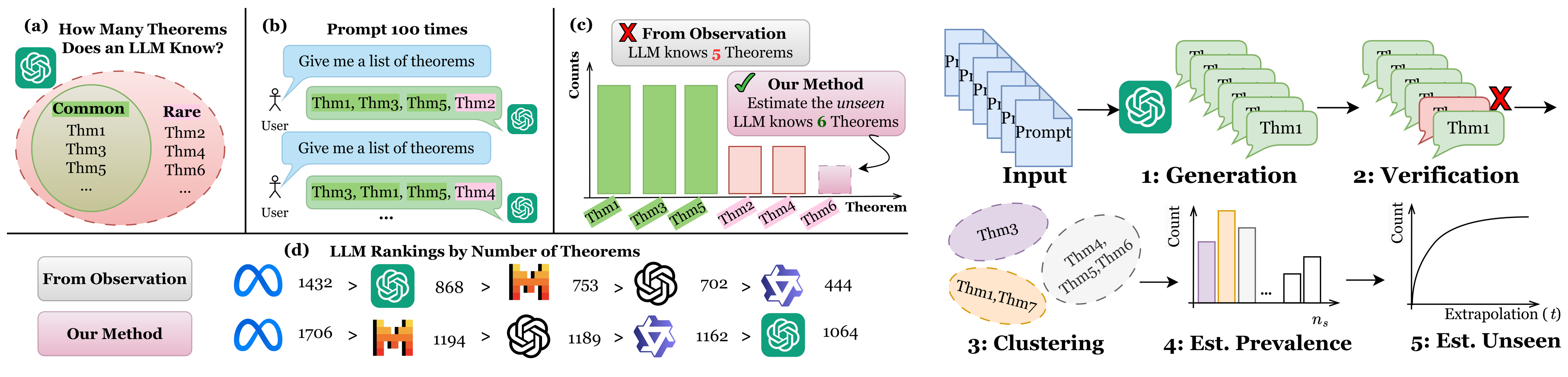
Evaluation Crisis

- Benchmark scores increasingly diverge from true model generalization.
- Causes include benchmark contamination, leaderboard overfitting, and narrow test-time optimization.
- LLMs generate responses by sampling, so capabilities may not appear simply due to their low sampling probability.

Questions Studied

Could we evaluate LLMs by estimating their "unseen" capacity or knowledge?

Our Method: KnowSum [4]



Conclusion

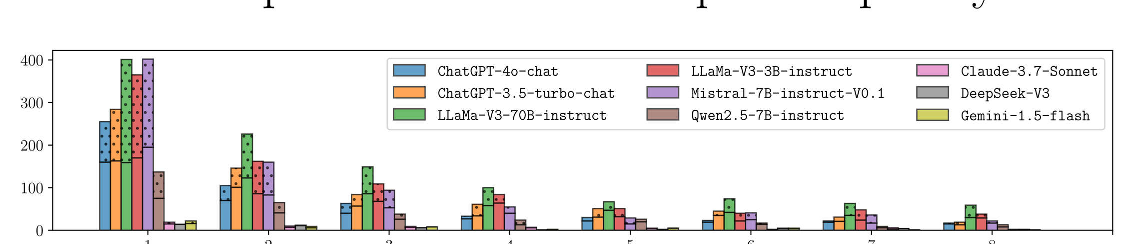
- KnowSum effectively estimates discrete and countable knowledge in LLMs.
- KnowSum is versatile and demonstrates utility across three LLM applications.
- Unseen knowledge meaningfully changes the model comparison result.

① Knowledge estimation

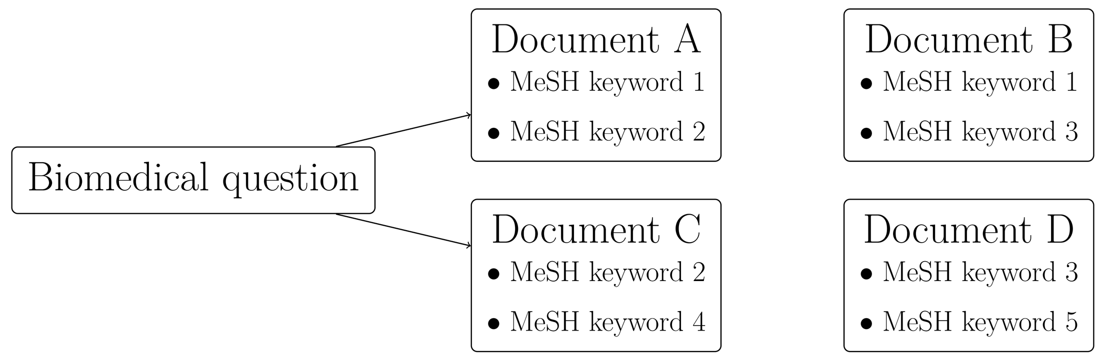
- Query LLM N_{query} times, each with N_{ans} domain-specific outputs.
- Validate via external databases (e.g., Wikipedia) and cluster by unique IDs (e.g., URLs).
- $(N_{\text{query}}, N_{\text{ans}}) = (30{,}000, 20)$ for theorems counting.

Model	Theorem only (10%)			All math concepts		
	N_{seen}	$\widehat{N}_{\text{tot}}$	SKR	N_{seen}	$\widehat{N}_{\text{tot}}$	SKR
① `ChatGPT-4o-chat`	702	1189	0.59	974	2410	0.40
② `ChatGPT-3.5-turbo-chat`	868	1064	0.82	1266	1703	0.74
③ `LLaMA-V3-70B-instruct`	**1432**	**1706**	0.84	**2289**	**2645**	0.87
④ `LLaMA-V3-3B-instruct`	1035	1331	0.78	1717	2640	0.65
⑤ `Mistral-7B-instruct-V0.1`	753	1194	0.63	1313	2481	0.53
⑥ `Qwen2.5-7B-instruct`	444	1162	0.38	663	1385	0.48
⑦ `Claude-3.7-Sonnet`	120	201	0.60	147	293	0.50
⑧ `DeepSeek-V3`	148	241	0.61	162	203	0.80
⑨ `Gemini-1.5-flash`	100	515	**0.19**	122	478	**0.26**

- LLMs hold unexpressed math knowledge.
- Unseen knowledge reshapes rankings: seen → `3.5` > `4o`, total → `3.5` < `4o`.
- Unseen depends on the full top-k frequency.



② Information retrieval



- 340 questions from `BioASQ-QA` Task 12B test set.
- Each question has ground-truth docs with MeSH keyword annotations.
- **Document retrieval**: LLMs generate Boolean queries (MeSH + AND/OR/NOT) to search PubMed; retrieved ground-truth docs → their MeSH keywords counted as valid knowledge.
- **Question answering**: LLMs answer based on retrieved docs; correct answers → MeSH keywords from linked docs counted as valid knowledge.

Model	Document Retrieval			Question Answering		
	N_{seen}	$\widehat{N}_{\text{tot}}$	SKR	N_{seen}	$\widehat{N}_{\text{tot}}$	SKR
① `ChatGPT-4o-chat`	2015	9676	0.21	**2351**	**19965**	0.12
② `ChatGPT-3.5-turbo-chat`	2190	**10367**	0.21	1850	15733	0.12
③ `LLaMA-V3-70B-instruct`	1990	8488	0.23	1928	14270	0.14
④ `LLaMA-V3-3B-instruct`	79	396	**0.20**	1653	14199	0.12
⑤ `Mistral-7B-instruct-v0.1`	1364	5646	0.24	630	6596	**0.10**
⑥ `Qwen2.5-7B-instruct`	1399	4853	0.28	1585	10710	0.15
⑦ `Claude-3.7-Sonnet`	2050	8831	0.23	2023	17230	0.12
⑧ `DeepSeek-V3`	**2260**	7750	0.30	2290	19744	0.12
⑨ `Gemini-1.5-flash`	2027	6616	0.31	2222	14898	0.15

From seen to unseen

Let n_s be the number of responses appearing exactly $s \geq 1$ times in the first n observations. For extrapolation factor $t > 0$, the goal is to estimate $n_0(t)$, the number of new responses expected in the next $t \cdot n$ prompts, based on $\{n_s\}_{s\geq 1}$.

- **Good–Turing (GT)** [2] is unbiased, but high variance: $\widehat{N}^{\text{GT}}_{\text{unseen}}(t) = -\sum_{s=1}^{\infty}(-t)^s n_s$.
- **Smoothed GT (SGT)** [5] uses random truncation L: $\widehat{N}^{\text{SGT}}_{\text{unseen}}(t) = \mathbb{E}\big[-\sum_{s=1}^{L}(-t)^s n_s\big]$.
- **Efron–Thisted (ET)** [1] is a special case of SGT with $L \sim \text{Bin}(k, 1/(t+1))$. It's minimax optimal when k is adaptively set [5].
- Seen knowledge ratio (SKR) is defined by

$$\text{SKR}(t) = \frac{N_{\text{seen}}}{N_{\text{seen}} + \widehat{N}_{\text{unseen}}(t)}.$$

③ Diversity measure

- Query a LLM 1000 times about a possible application or an imagined dream job.
- Since no ground-truth answers exist, embed the responses into semantic vectors and group them into clusters when they are sufficiently far apart.

Model	LLM Applications			Dream Jobs		
	N_{seen}	$\widehat{N}_{\text{tot}}$	SKR	N_{seen}	$\widehat{N}_{\text{tot}}$	SKR
① `ChatGPT-4o-chat`	165	714	0.23	409	1680	0.24
② `ChatGPT-3.5-turbo-chat`	322	1339	0.24	131	560	0.23
③ `LLaMA-V3-70B-instruct`	437	1918	0.23	344	1487	0.23
④ `LLaMA-V3-3B-instruct`	428	1926	0.22	**770**	**3386**	0.23
⑤ `Mistral-7B-instruct-v0.1`	658	**3155**	**0.21**	233	1093	**0.21**
⑥ `Qwen2.5-7B-instruct`	421	1840	0.23	507	2094	0.24
⑦ `Claude-3.7-Sonnet`	**696**	3013	0.23	133	543	0.24
⑧ `DeepSeek-V3`	17	48	0.35	7	10	0.7
⑨ `Gemini-1.5-flash`	21	37	0.57	3	10	0.3

References

[1] B. Efron and R. Thisted. Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, 63(3):435–447, 1976.

[2] I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264, 1953.

[3] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

[4] X. Li, J. Xin, Q. Long, and W. J. Su. Evaluating the unseen capabilities: How many theorems do LLMs know? *arXiv preprint arXiv:2506.02058*, 2025.

[5] A. Orlitsky, A. T. Suresh, and Y. Wu. Optimal prediction of the number of unseen species. *Proceedings of the National Academy of Sciences*, 113(47):13283–13288, 2016.