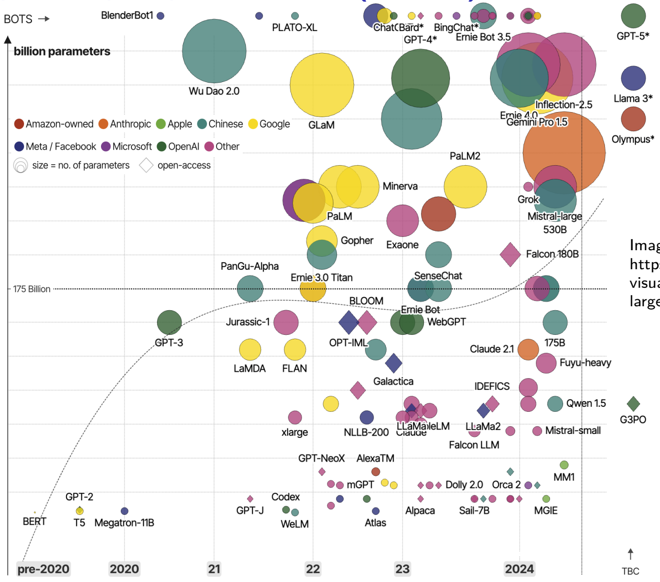
Evaluating the Unseen Capabilities: How Much Do LLMs Actually Know?

Xiang Li

University of Pennsylvania

Augest 21, 2025

The rise of large language models (LLMs)



David McCandless, Tom Evans, Paul Barton
Information is Beautiful // UPDATED 20th Mar 24

source: news reports, LifeArchitect.ai
* = parameters undisclosed // see the data

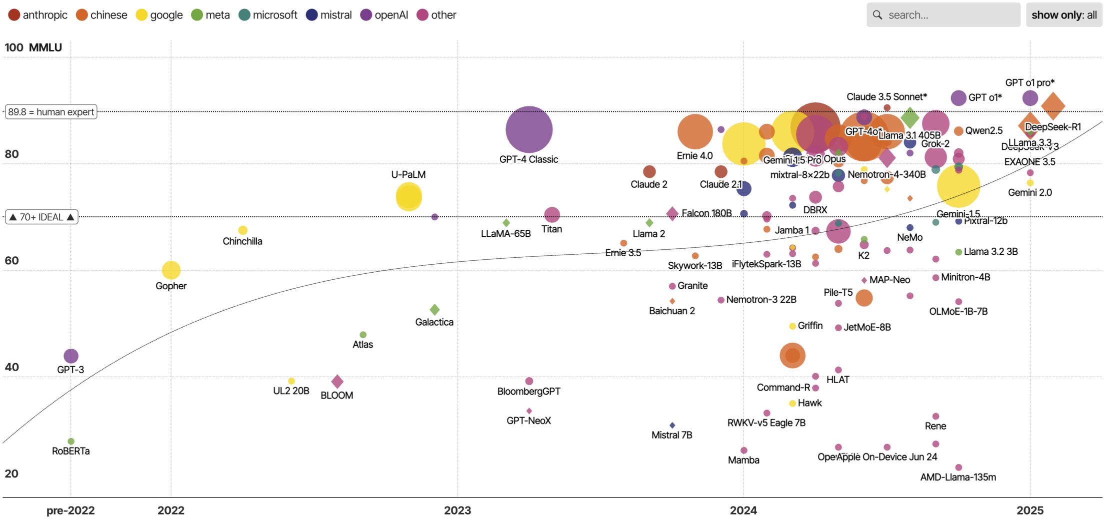
Image from https://informationisbeautiful.net/visualizations/the-rise-of-generative-ai-large-language-models-llms-like-chatgpt/

How to evaluate different LLMs?

Current evaluation methods rely heavily on standardized benchmarks.

- Collect or design questions and measure the accuracy of model responses.
- The **MMLU** (Massive Multitasks Language Understanding [Hendrycks et al., 2021]) datasets consists of 16,000 multiple-choice questions across 57 academic subjects (such as elementary mathematics, US history, computer science and law).

LLMs are rapidly evolving in terms of MMLU scores



David McCandless, Tom Evans, Paul Barton
Informationisbeautiful // Jan 2024

MMLU = benchmark for measuring LLM capabilities
* = parameters undisclosed // source: LifeArchitect // data

Could we trust these released scores?

- Higher scores do not imply overall superiority (relative comparison).
    - Example: PaLM scores 69.6% vs. GPT-3.5's 65% on MMLU, yet GPT-3.5 is far stronger in coding and math.

Could we trust these released scores?

- Higher scores do not imply overall superiority (relative comparison).
  - Example: PaLM scores 69.6% vs. GPT-3.5's 65% on MMLU, yet GPT-3.5 is far stronger in coding and math.
- Scores themselves are not fully reliable (absolute comparison).
  - The scores are sensitive to slight question perturbations (e.g., changing choice orders, prompts, choice symbols) [Alzahrani et al., 2024].
  - Scores fail to generalize to harder math questions [Huang et al., 2025].

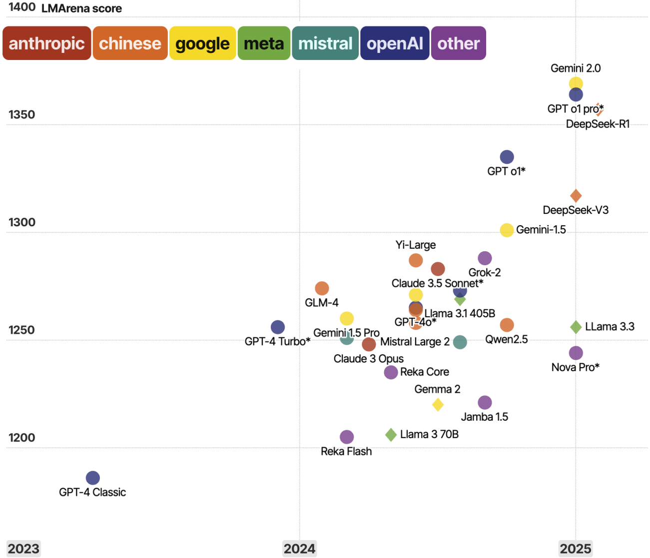
Could we trust these released scores?

- Higher scores do not imply overall superiority (relative comparison).
    - Example: PaLM scores 69.6% vs. GPT-3.5's 65% on MMLU, yet GPT-3.5 is far stronger in coding and math.
- Scores themselves are not fully reliable (absolute comparison).
    - The scores are sensitive to slight question perturbations (e.g., changing choice orders, prompts, choice symbols) [Alzahrani et al., 2024].
    - Scores fail to generalize to harder math questions [Huang et al., 2025].
- On other benchmarks, LLMs reach performance saturation very quickly.
- Benchmark contamination [Sainz et al., 2023].
- Post-training changes the way an LLM expresses its knowledge.

A complementary evaluation: LMArena Score

- To address the limitations of static benchmarks, the **LMArena** (a.k.a. Chatbot Arena [Chiang et al., 2024]) introduces a dynamic, preference-based evaluation.
- Users vote on pairwise comparisons of model responses, and an Elo-style rating is updated accordingly (winners gain points and losers lose points).
- This approach captures real-world human preferences and maintains differentiation when benchmark scores saturate.

Indistinguishable performance among first-tier LLMs

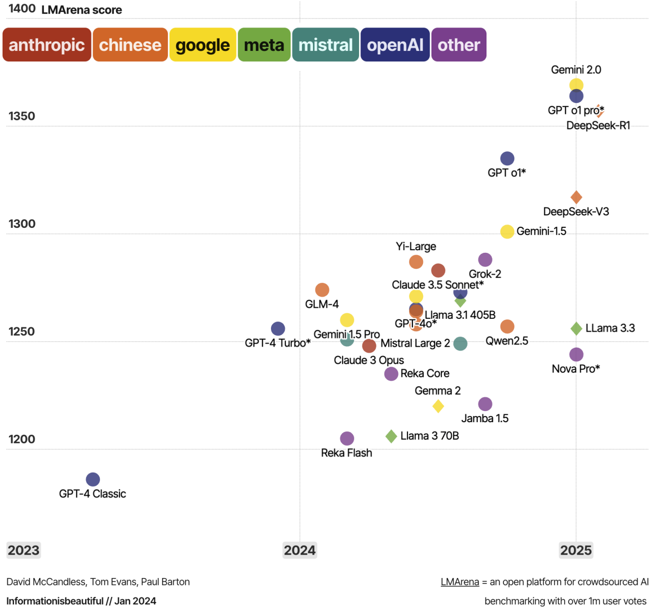


David McCandless, Tom Evans, Paul Barton

Informationisbeautiful // Jan 2024

LMArena = an open platform for crowdsourced AI benchmarking with over 1m user votes

Indistinguishable performance among first-tier LLMs



- Arena scores of top models are very close.
- Relative ranking depends on the competing (random) pool.
- Not an absolute measure.
- A single win does not necessarily indicate strong capacity, but the overall score reflects the model's relative strength across many battles.

An evaluation crisis

Andrej Karpathy
@karpathy

Following

Building @EurekaLabsAI. Previously Director of AI @ Tesla, founding team @ OpenAI, CS231n/PhD @ Stanford. I like to train large deep neural nets.

My reaction is that there is an evaluation crisis. I don't really know what metrics to look at right now.
MMLU was a good and useful for a few years but that's long over.
SWE-Bench Verified (real, practical, verified problems) I really like and is great but itself too narrow.
Chatbot Arena received so much focus (partly my fault?) that LLM labs have started to really overfit to it, via a combination of prompt mining (from API requests), private evals bombardment, and, worse, explicit use of rankings as training supervision. I think it's still ~ok and there's a lack of "better", but it feels on decline in signal.
There's a number of private evals popping up, an ensemble of which might be one promising path forward.
In absence of great comprehensive evals I tried to turn to vibe checks instead, but I now fear they are misleading and there is too much opportunity for confirmation bias, too low sample size, etc., it's just not great.

TLDR my reaction is I don't really know how good these models are right now.

1:29 PM · Mar 2, 2025 · **301.9K** Views

An evaluation crisis

Andrej Karpathy
@karpathy
Following

Building @EurekaLabsAI. Previously Director of AI @ Tesla, founding team @ OpenAI, CS231n/PhD @ Stanford. I like to train large deep neural nets.

Some causes of the crisis

- Benchmark contamination [Sainz et al., 2023].
- Overfitting through repeated leaderboard submissions [Singh et al., 2025].
- Narrow test-time optimization strategies [Leech et al., 2024].

My reaction is that there is an evaluation crisis. I don't really know what metrics to look at right now.
MMLU was a good and useful for a few years but that's long over.
SWE-Bench Verified (real, practical, verified problems) I really like and is great but itself too narrow.
Chatbot Arena received so much focus (partly my fault?) that LLM labs have started to really overfit to it, via a combination of prompt mining (from API requests), private evals bombardment, and, worse, explicit use of rankings as training supervision. I think it's still ~ok and there's a lack of "better", but it feels on decline in signal.
There's a number of private evals popping up, an ensemble of which might be one promising path forward.
In absence of great comprehensive evals I tried to turn to vibe checks instead, but I now fear they are misleading and there is too much opportunity for confirmation bias, too low sample size, etc., it's just not great.

TLDR my reaction is I don't really know how good these models are right now.

1:29 PM · Mar 2, 2025 · **301.9K** Views

An alternative perspective for LLM evaluation

Central question

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

An alternative perspective for LLM evaluation

Central question

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

- We offer an affirmative answer by proposing a statistical framework KnowSum.
- We show its effectiveness through three distinct applications for estimating countable knowledge.

An alternative perspective for LLM evaluation

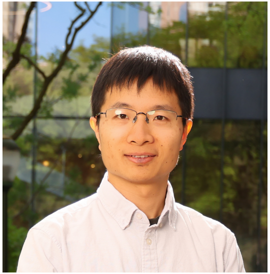
Central question

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

- We offer an affirmative answer by proposing a statistical framework KNOWSUM.
- We show its effectiveness through three distinct applications for estimating countable knowledge.
- Joint work with

Jiayi Xin

Qi Long



Weijie Su

In this talk

Motivation

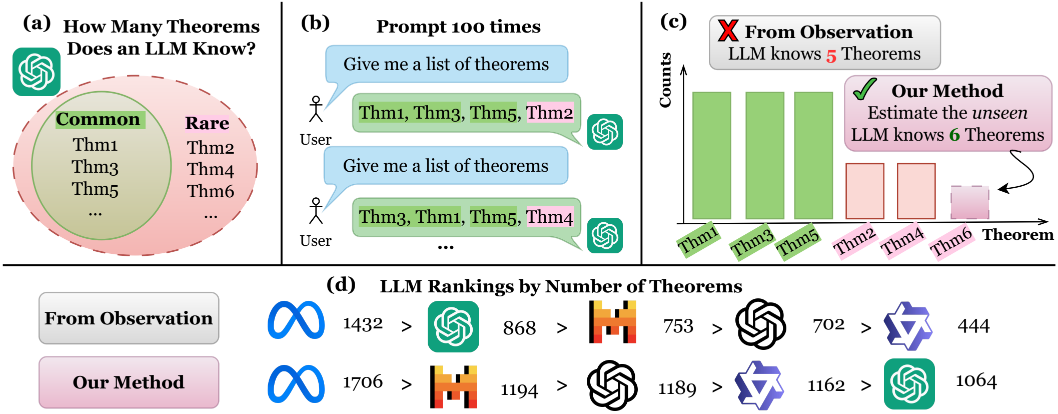
Our method

Applications

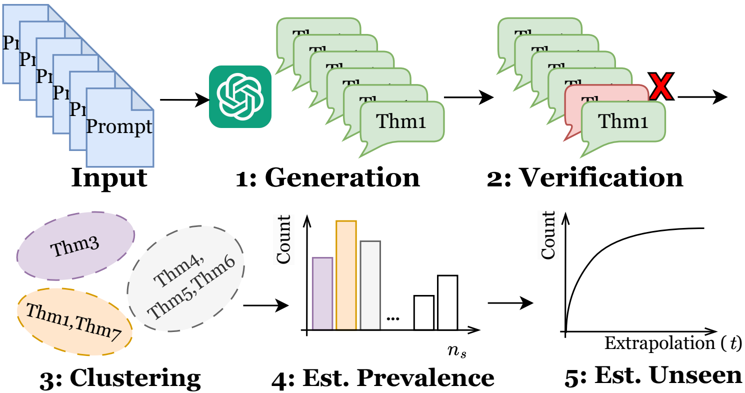
Concluding remarks

Outlines

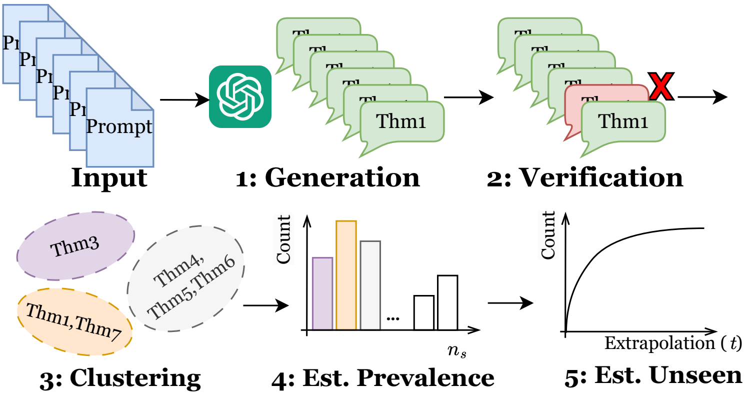
Our method: Overview



Our method: KnowSum



Our method: KnowSum



- Five-step procedure.
- Steps 2 & 3 standardize answers.
- Verification reduces hallucination.
- Clustering reduces redundancy.

How to extrapolate from seen to unseen

Problem formulation

Let n_s denote the number of responses that appear exactly $s \geq 1$ times in the first n observation. For an extrapolation factor $t > 0$, we aim to estimate $n_0(t)$, the number of new responses expected to appear in the next $t \cdot n$ prompts, using the observed frequency counts $\{n_s\}_{s \geq 1}$.

- The same setting as estimating the number of unseen species.

How to extrapolate from seen to unseen

Problem formulation

Let n_s denote the number of responses that appear exactly $s \geq 1$ times in the first n observation. For an extrapolation factor $t > 0$, we aim to estimate $n_0(t)$, the number of new responses expected to appear in the next $t \cdot n$ prompts, using the observed frequency counts $\{n_s\}_{s\geq 1}$.

- The same setting as estimating the number of unseen species.
- The Good–Turing (GT) estimator [Good, 1953] uses

$$\widehat{N}^{\mathrm{GT}}_{\mathrm{unseen}}(t) = -\sum_{s=1}^{\infty}(-t)^s n_s.$$

- When $t = 1$, it is $n_1 - n_2 + n_3 - n_4 + n_5 - \cdots$.
- The GT estimator is unbiased but has large variance, making it unstable.

Derivation for the GT estimator

Species trapping model [Fisher et al., 1943]

- There are S species in total. Suppose we observe n species during one unit of time, say over the interval $[-1, 0]$.
- After trapping for t units of time, let $x_s(t)$ denote $\#$ of captures from species s.
- We model $x_s(t) \sim \mathrm{Poisson}(\lambda_s \cdot (t+1))$, and assume that the behavior in $[-1, 0]$ is representative of the entire period $[0, t]$.

Derivation for the GT estimator

Species trapping model [Fisher et al., 1943]

- There are S species in total. Suppose we observe n species during one unit of time, say over the interval $[-1, 0]$.
- After trapping for t units of time, let $x_s(t)$ denote # of captures from species s.
- We model $x_s(t) \sim \mathrm{Poisson}(\lambda_s \cdot (t+1))$, and assume that the behavior in $[-1, 0]$ is representative of the entire period $[0, t]$.

$$n_s = \mathbb{E}[\text{\# of species observed exactly } s \text{ times in } [-1,0]] = S \int_0^\infty e^{-\lambda} \frac{\lambda^s}{s!} \, \mathrm{d}G(\lambda)$$

$$n_0(t) = \mathbb{E}[\text{\# of species observed in } (0,t] \text{ but not in } [-1,0]] = S \int_0^\infty e^{-\lambda}(1 - e^{-\lambda t}) \, \mathrm{d}G(\lambda)$$

By applying Taylor expansion w.r.t. λ, we obtain

$$n_0(t) = -\sum_{s=1}^{\infty} (-t)^s n_s.$$

Smoothed Good–Turing (SGT) estimator

Goal: estimate the number of new items that will appear in the next $t \cdot n$ queries, given the frequency counts $\{n_s\}_{s \geq 1}$ from first n observations.

- **SGT estimator** [Orlitsky et al., 2016] uses a random truncation L and define

$$\widehat{N}^{\mathrm{SGT}}_{\mathrm{unseen}}(t) := \mathbb{E}\left[-\sum_{s=1}^{L}(-t)^s n_s\right].$$

Smoothed Good–Turing (SGT) estimator

Goal: estimate the number of new items that will appear in the next $t \cdot n$ queries, given the frequency counts $\{n_s\}_{s \geq 1}$ from first n observations.

- **SGT estimator** [Orlitsky et al., 2016] uses a random truncation L and define

$$\widehat{N}^{\mathrm{SGT}}_{\mathrm{unseen}}(t) := \mathbb{E}\left[-\sum_{s=1}^{L}(-t)^s n_s\right].$$

- A famous instance is the **ET estimator**, where $L \sim \mathrm{Bin}(k, 1/(t+1))$ is binomial with k trials and success probability $1/(t+1)$ [Efron and Thisted, 1976].

$$\widehat{N}^{\mathrm{ET}}_{\mathrm{unseen}}(t) = \sum_{s=1}^{k} h_s\, n_s, \quad h_s = -(-t)^s\, \mathbb{P}\left(\mathrm{Bin}\left(k, \tfrac{1}{t+1}\right) \geq s\right).$$

- ET estimator is minimax optimal when k is adaptively set [Orlitsky et al., 2016].
- In our experiments, we employ this with k tuned from $\{6, 8, 10\}$ and $t = 100$.

Outlines

Application 1: Knowledge estimation

- Query the LLM N_{query} times with a fixed prompt, each time requesting N_{ans} instances of domain-specific knowledge.
- Use external databases for validation (e.g., Wikipedia) and cluster the responses based on their unique external identifiers (e.g., canonical URL).
- $(N_{\mathrm{query}}, N_{\mathrm{ans}}) = (30{,}000, 20)$ for theorems and $(3{,}000, 50)$ for diseases.

Application 1: Knowledge estimation

- Query the LLM N_{query} times with a fixed prompt, each time requesting N_{ans} instances of domain-specific knowledge.
- Use external databases for validation (e.g., Wikipedia) and cluster the responses based on their unique external identifiers (e.g., canonical URL).
- $(N_{\text{query}}, N_{\text{ans}}) = (30{,}000, 20)$ for theorems and $(3{,}000, 50)$ for diseases.

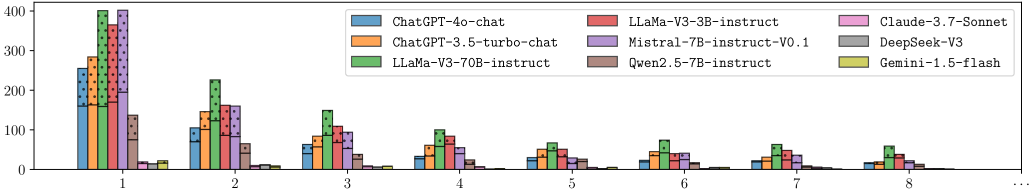
Model	**Theorem only** (10%)			**All math concepts**			**Anatomical disease** (51%)			**Human diseases**		
	N_{seen}	$\widehat{N}_{\text{tot}}$	SKR	N_{seen}	$\widehat{N}_{\text{tot}}$	SKR	N_{seen}	$\widehat{N}_{\text{tot}}$	SKR	N_{seen}	$\widehat{N}_{\text{tot}}$	SKR
① `ChatGPT-4o-chat`	702	1189	0.59	974	2410	0.40	277	732	0.38	589	1096	0.54
② `ChatGPT-3.5-turbo-chat`	868	1064	0.82	1266	1703	0.74	268	278	0.96	523	706	0.74
③ `LLaMA-V3-70B-instruct`	**1432**	**1706**	0.84	**2289**	**2645**	0.87	**875**	**3372**	0.26	**1777**	**7564**	**0.23**
④ `LLaMA-V3-3B-instruct`	1035	1331	0.78	1717	2640	0.65	780	1375	0.57	1374	3002	0.46
⑤ `Mistral-7B-instruct-V0.1`	753	1194	0.63	1313	2481	0.53	489	1723	0.28	859	1276	0.67
⑥ `Qwen2.5-7B-instruct`	444	1162	0.38	663	1385	0.48	426	521	0.82	763	763	1.00
⑦ `Claude-3.7-Sonnet`	120	201	0.60	147	293	0.50	115	462	**0.25**	213	686	0.31
⑧ `DeepSeek-V3`	148	241	0.61	162	203	0.80	86	334	0.26	193	752	0.26
⑨ `Gemini-1.5-flash`	100	515	**0.19**	122	478	**0.26**	139	143	0.97	298	306	0.97

A gap between observed and total knowledge

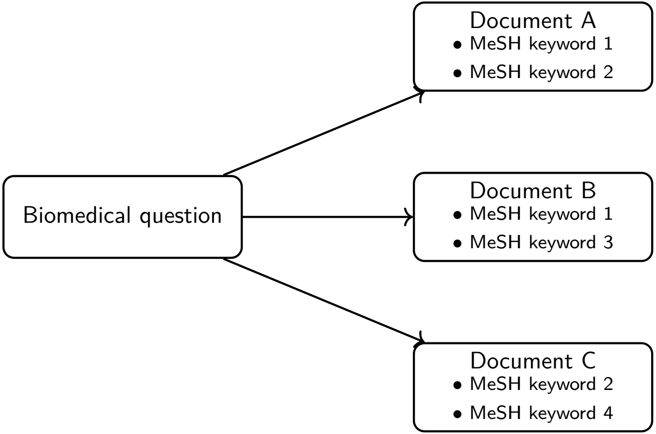
- All the LLMs have unexpressed math or medical knowledge.
- Unseen knowledge changes model ranking, e.g.,
  - From the seen, `ChatGPT-3.5-turbo-chat` $>$ `ChatGPT-4o-chat`.
  - From the total, `ChatGPT-3.5-turbo-chat` $<$ `ChatGPT-4o-chat`.

A gap between observed and total knowledge

- All the LLMs have unexpressed math or medical knowledge.
- Unseen knowledge changes model ranking, e.g.,
  - From the seen, `ChatGPT-3.5-turbo-chat` $>$ `ChatGPT-4o-chat`.
  - From the total, `ChatGPT-3.5-turbo-chat` $<$ `ChatGPT-4o-chat`.
- The whole shape (top-k) of frequencies determines the unseen.



Application 2: Information retrieval



- `BioASQ-QA` Task 12B Test Dataset [Krithara et al., 2023].
- Each question is associated with a set of ground-truth documents, and each document is annotated with a list of MeSH keywords.
- MeSH: Medical Subject Headings.
- Totally 340 questions.

Application 2: Information retrieval

- **Document retrieval**: Ask each LLM to generate Boolean search queries to retrieve relevant documents from the PubMed database. Each query consists of MeSH keywords, combined using logical operators (AND, OR, NOT), and are submitted to a search engine to return candidate documents.
  - If LLM retrieves a document in the ground-truth set, all MeSH keywords associated are counted as valid knowledge.
- **Question answering**: Ask each LLM to answers biomedical research questions (curated by domain) based on the retrieved documents.
  - If LLM's response is deemed correct, all MeSH keywords from the documents linked to that question are counted as valid knowledge.

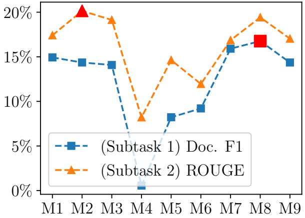
Application 2: Information retrieval

- Our methods estimates how many additional relevant MeSH keywords an LLM could potentially retrieve if more questions were collected and evaluated under the same manner.
- Traditional IR metrics (e.g., F1 score and ROUGE score) assess retrieval and answer quality based on document relevance.

Application 2: Information retrieval

- Our methods estimates how many additional relevant MeSH keywords an LLM could potentially retrieve if more questions were collected and evaluated under the same manner.
- Traditional IR metrics (e.g., F1 score and ROUGE score) assess retrieval and answer quality based on document relevance.

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



Performance on selected traditional IR metrics.

Application 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.

Application 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

Outlines

Concluding remarks

Evaluating the Unseen Capabilities: How Many Theorems Do LLMs Know?
(https://arxiv.org/abs/2506.02058)

- KnowSum can estimate the discrete and countable knowledge well, e.g., the number of theorems/diseases.
- KnowSum is versatile and show utility in three applications.
- Unseen knowledge meaningfully changes the model rank.

Open directions

- How to **extract and represent** the knowledge items estimated by KnowSum?
- What if the number of total knowledge instances increases with time?
- How to **extend the framework** from **discrete symbols** to **continuous or uncountable domains** (e.g., real-valued reasoning steps)?
- How to **define and detect "singletons"** (automatically) for abstract knowledge (e.g., code snippets, math techniques)?
- Can unseen estimation help **save data collection effort** by identifying when existing data is sufficient and guiding augmentation only for rare cases?
- ......

References I

N. Alzahrani, H. Alyahya, Y. Alnumay, S. Alrashed, S. Alsubaie, Y. Almushayqih, F. Mirza, N. Alotaibi, N. Al-Twairesh, A. Alowisheq, et al. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13787–13805, 2024.

W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, B. Zhu, H. Zhang, M. Jordan, J. E. Gonzalez, and I. Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In *International Conference on Machine Learning*, pages 8359–8388. PMLR, 2024.

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

R. A. Fisher, A. S. Corbet, and C. B. Williams. The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, pages 42–58, 1943.

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

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. URL `https://openreview.net/forum?id=d7KBjmI3GmQ`.

K. Huang, J. Guo, Z. Li, X. Ji, J. Ge, W. Li, Y. Guo, T. Cai, H. Yuan, R. Wang, et al. MATH-Perturb: Benchmarking LLMs' math reasoning abilities against hard perturbations. In *International Conference on Machine Learning*, 2025.

References II

A. Krithara, A. Nentidis, K. Bougiatiotis, and G. Paliouras. BioASQ-QA: A manually curated corpus for biomedical question answering. *Scientific Data*, 10(1):170, 2023.

G. Leech, J. J. Vazquez, N. Kupper, M. Yagudin, and L. Aitchison. Questionable practices in machine learning. *arXiv preprint arXiv:2407.12220*, 2024.

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.

O. Sainz, J. Campos, I. García-Ferrero, J. Etxaniz, O. L. de Lacalle, and E. Agirre. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, 2023.

S. Singh, Y. Nan, A. Wang, D. D'Souza, S. Kapoor, A. Üstün, S. Koyejo, Y. Deng, S. Longpre, N. A. Smith, B. Ermis, M. Fadaee, and S. Hooker. The leaderboard illusion. *arXiv preprint arXiv:2504.20879*, 2025.