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

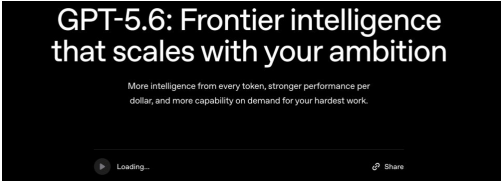
Xiang Li

University of Pennsylvania

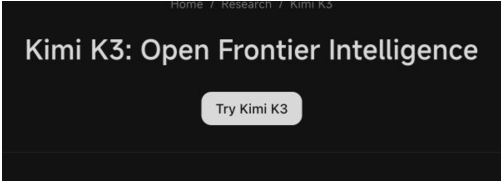
Joint Statistical Meetings (JSM) 2026
August 4, 2026

New LLMs are released at a rapid pace

Claude Fable 5 July 1, 2026



GPT-5.6 July 9, 2026



Kimi K3 July 16, 2026

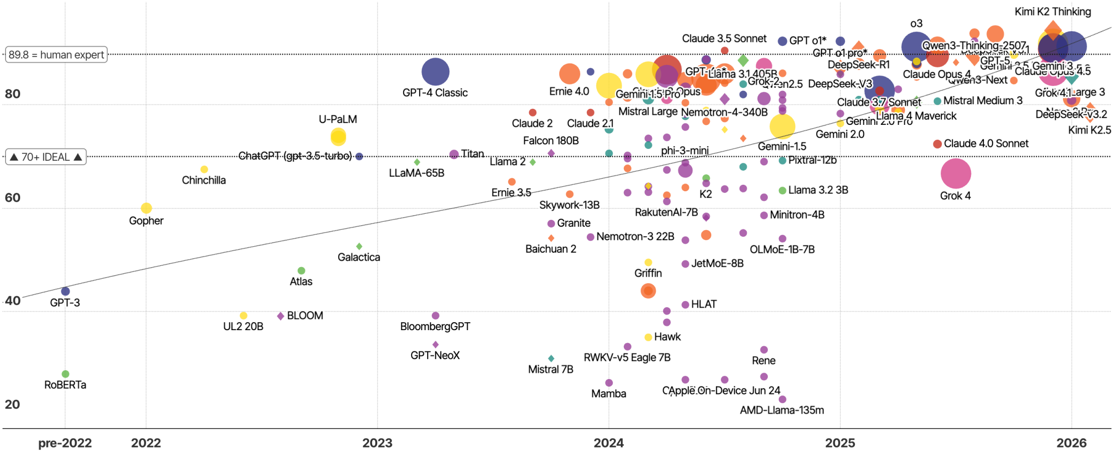
DeepSeek-V4-Flash-0731 July 31, 2026

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 2026

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

Could we trust these released scores?

- Performance satures very quickly (also on other benchmarks, e.g., GSM8K).
- 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.

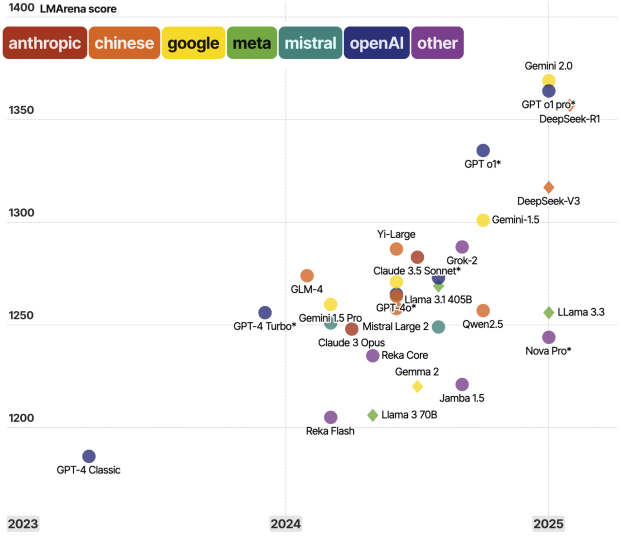
Could we trust these released scores?

- Performance satures very quickly (also on other benchmarks, e.g., GSM8K).
- 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].
- 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

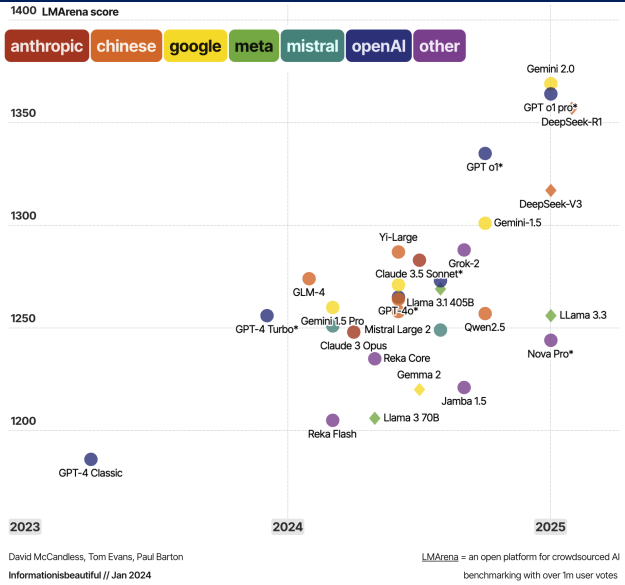


David McCandless, Tom Evans, Paul Barton

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.

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

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].

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 four applications, from model evaluation to data selection.

An alternative perspective for LLM evaluation

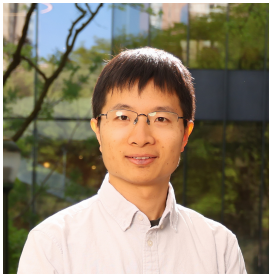
Central question

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 four applications, from model evaluation to data selection.

Jiayi Xin

Qi Long



Weijie Su

In this talk

Motivation

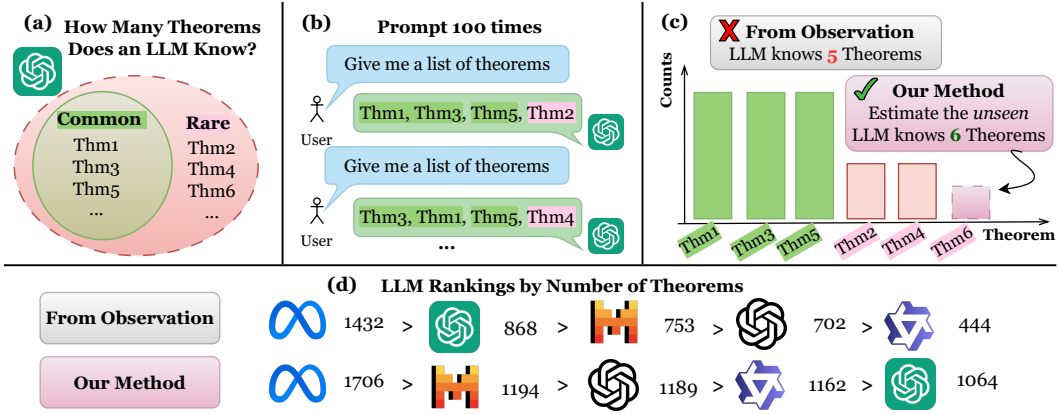
Our method

Applications

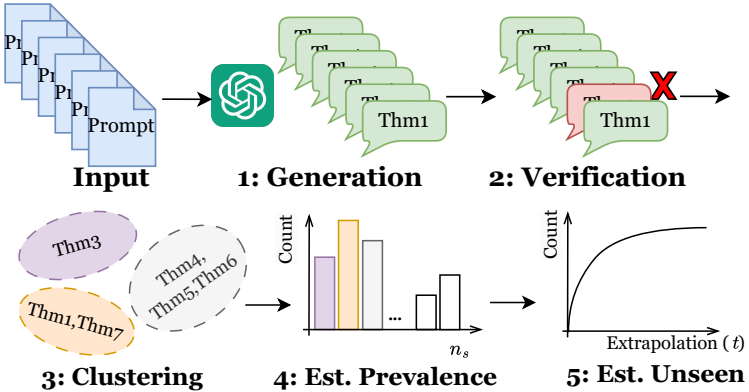
Concluding remarks

Outlines

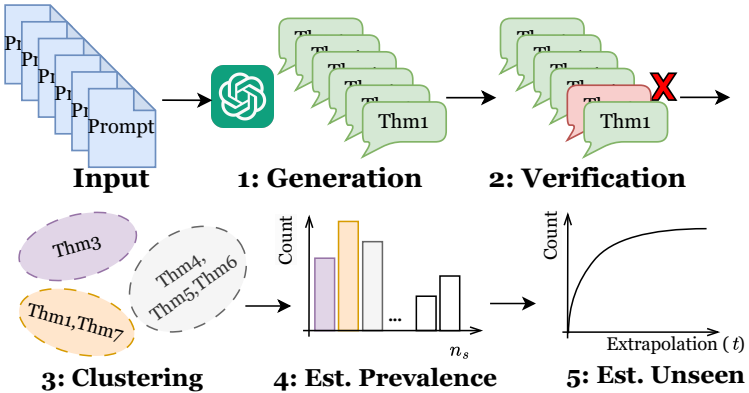
Our method: Overview



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.

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

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

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

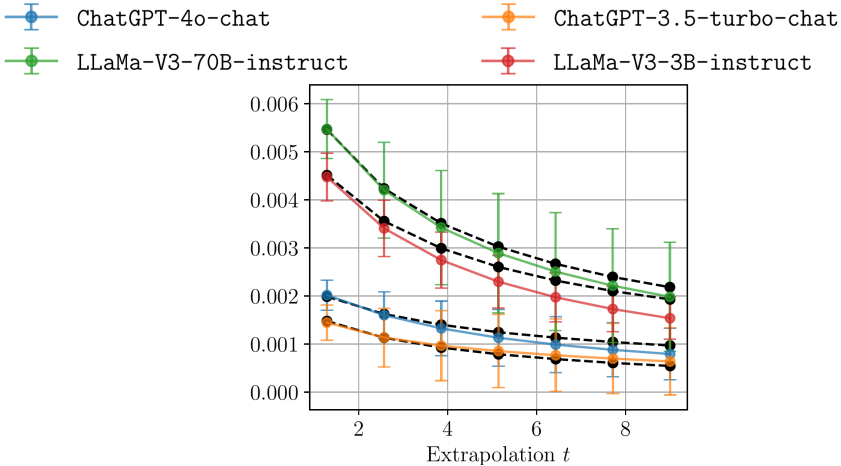
$$\widehat{N}^{\mathrm{SGT}}_{\mathrm{unseen}}(t) := \mathbb{E}\left[-\sum_{s=1}^{L} (-t)^s n_s \right].$$

- 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].

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

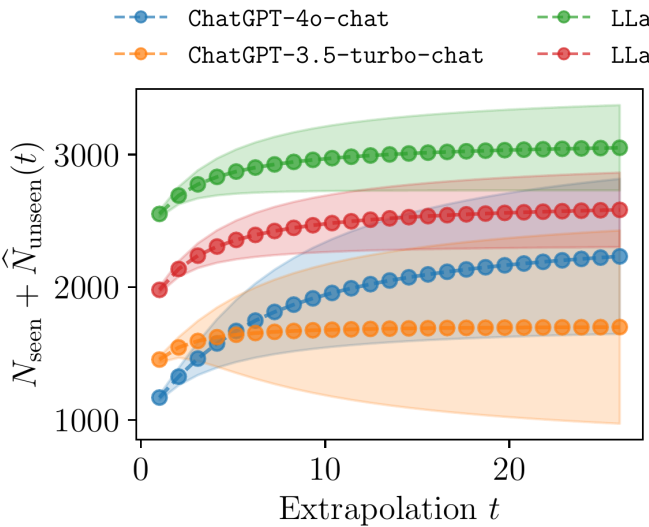
- ET estimator is minimax optimal when k is adaptively set [Orlitsky et al., 2016].
- In our experiments, we employ this with k tuned from $\{6, 8, 10\}$ and $t = 10$.

Held-out validation



- Random split; estimate held-out-only knowledge; repeat 100 times.
- y-**axis:** held-out unseen knowledge / observed knowledge.

Sensitivity to extrapolation

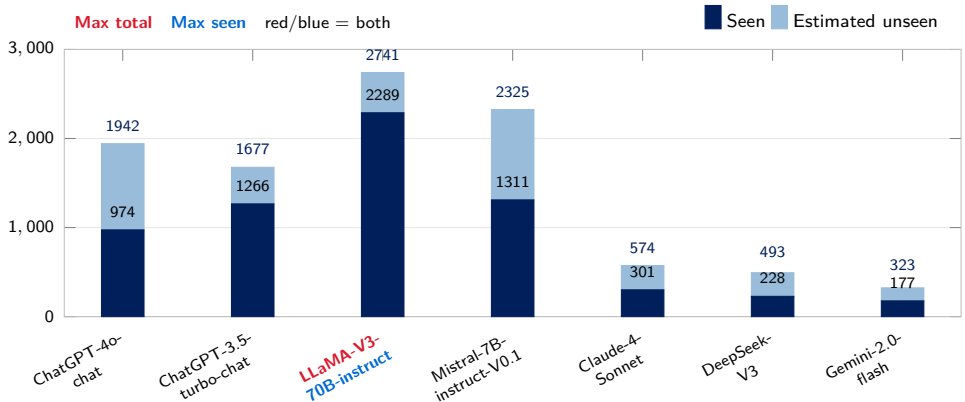


- Estimated total knowledge grows with t and begins to stabilize.
- Beyond this range, uncertainty grows faster than estimated support.

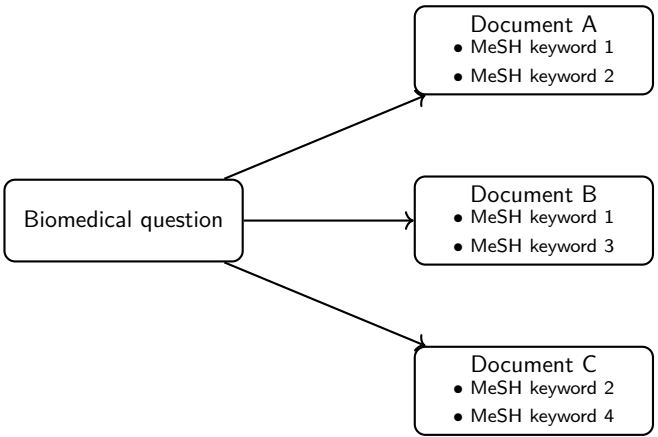
Outlines

Application 1: All mathematical concepts

- Query each LLM 30,000 times, validate responses externally, cluster equivalent concepts, and estimate the unseen tail.



Application 2: Document retrieval



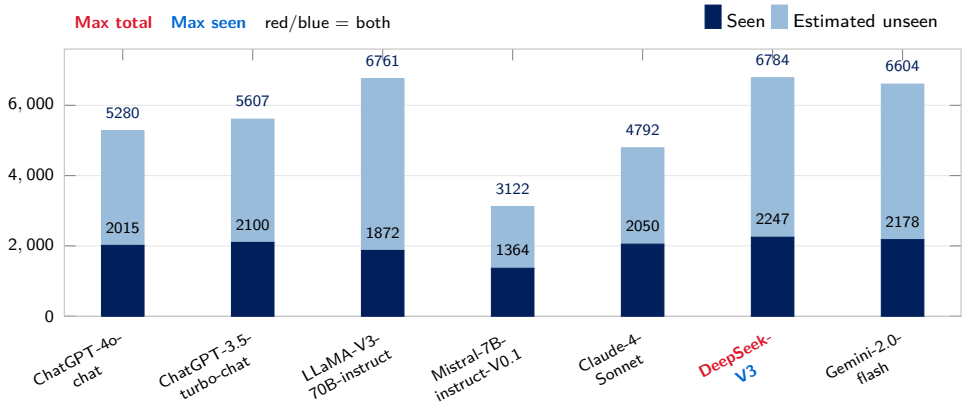
- `BioASQ-QA` Task 12B Test Dataset [Krithara et al., 2023].
- Each question is associated with a set of ground-truth documents, and each document is annotated with a list of MeSH keywords.
- MeSH: Medical Subject Headings.
- Totally 340 questions.

Application 2: Document 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.

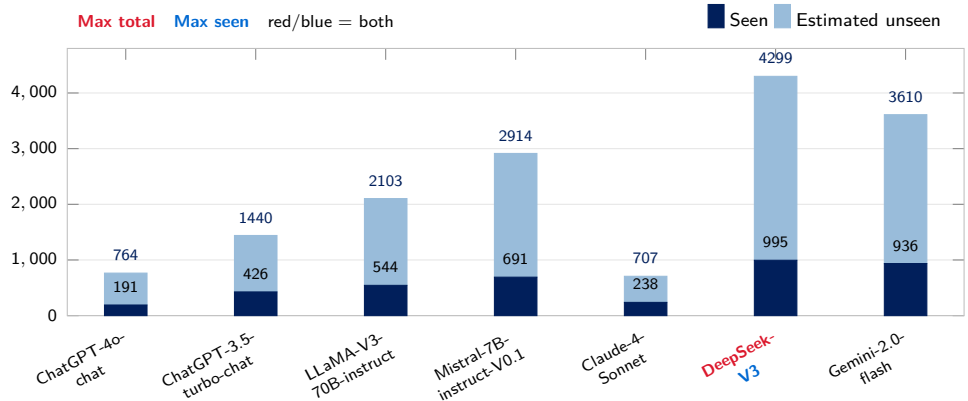
Application 2: Document retrieval

- On 340 BioASQ questions, count validated MeSH concepts from relevant PubMed documents and estimate the unseen tail.



Application 3: Diversity of LLM applications

- Query each model 1,000 times, cluster the open-ended responses into semantic units, and estimate latent diversity.



Application 4: In-context learning from a few demonstrations

Application 4: In-context learning from a few demonstrations

Demonstrations in the prompt

Apple $\rightarrow$ Fruit

Carrot $\rightarrow$ Vegetable

Banana $\rightarrow$ Fruit

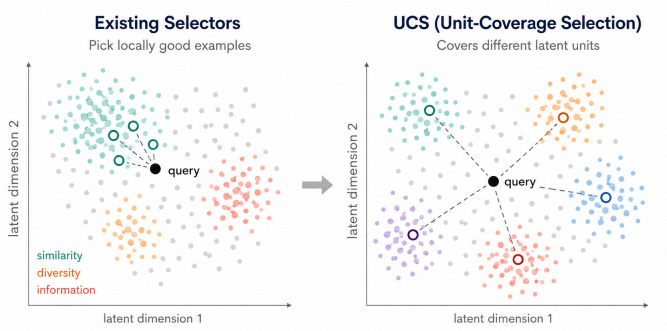
New query

Broccoli $\rightarrow$ **Vegetable**

- Learn the task from demonstrations in the prompt—no parameter update or gradient step.
- Under a fixed context budget, performance depends on which k demonstrations are selected.

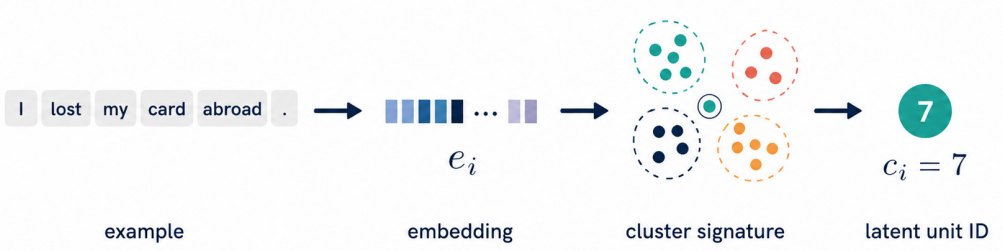
Application 4: Selecting demonstrations for in-context learning

Key observation: $\widehat{K}_{\text{total}}(S)$ is a set function for evaluating the candidate subset S



- Each subset S defines a frequency spectrum over latent semantic units; the estimator turns it into a coverage score.

Application 4: Convert demonstrations into semantic units



$$r_i = \arg\min_r \left\{ \|e_i - Dr\|_2^2 + \alpha \|r\|_2^2 \right\}, \qquad \{c_i\}_{i=1}^N = \text{Clustering}\left(\{r_i\}_{i=1}^N\right).$$

Application 4: UCS directly augments any base selector

Start with an existing selector

Base score: $\mathrm{score}_{\mathrm{base}}(S)$ (e.g., MDL, DPP, or VoteK).

Direct plug-in augmentation

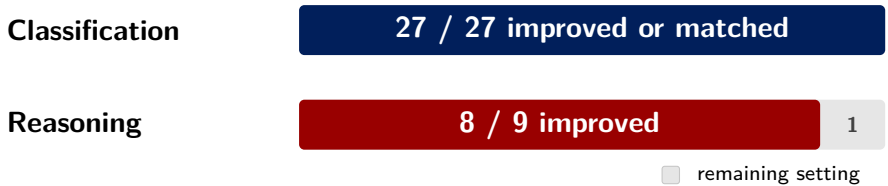
$$\mathrm{score}_{\mathrm{aug}}(S) = \mathrm{score}_{\mathrm{base}}(S) + \lambda \underbrace{\left[K_{\mathrm{seen}}(S) + \widehat{K}_{\mathrm{unseen}}(S)\right]}_{\mathrm{UCS}(S)}.$$

$$S^* = \arg\max_{|S|=k} \mathrm{score}_{\mathrm{aug}}(S).$$

- No need to redesign or fine-tune the original selector.
- $\lambda = 0$ recovers the base method; $\lambda > 0$ rewards estimated coverage.
- "Better" means broader estimated coverage under the same budget.

Application 4: UCS improves 35 of 36 settings

In-context learning accuracy under the same demonstration budget and base selector



Largest classification gain

+6.2%

Dataset: HWU64
Base method: VoteK
Model: Qwen2.5-7B-Instruct

Largest reasoning gain

+12.5%

Dataset: Shuffled Objects
Base method: DPP
Model: Qwen2.5-7B-Instruct

Outlines

Takeaways

- *Evaluating the Unseen Capabilities: How Many Theorems Do LLMs Know?* arXiv:2506.02058
- *UCS: Estimating Unseen Coverage for Improved In-Context Learning* arXiv:2604.12015

Takeaways

- **One estimator supports four applications.**
- **Estimate what is missing—then use it to improve evaluation and selection.**

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.

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.

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.

References II

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.