Stat 9911
Principles of AI: LLMs
Sampling/Inference/Test-time Computation

Edgar Dobriban

Department of Statistics and Data Science, the Wharton School, University of Pennsylvania

February 12, 2025

Plan

- We plan to discuss sampling, decoding, inference, test-time computation.

Table of Contents

Simple Decoding/Sampling Methods

Prompting

Reasoning

Simple Decoding/Sampling Methods

- After obtaining an LM, we need to sample/decode/generate text. Also known as inference, and is done at test-time (after training), so it belongs to the domain of test-time computation.
- Perhaps surprisingly, direct sampling from the LM may not be that great, and can be improved
- **Example:** Sampling from GPT-2 (Holtzman et al., 2020)

Context: In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

Beam Search, b=32:
"The study, published in the Proceedings of the National Academy of Sciences of the United States of America (PNAS), was conducted by researchers from the Universidad Nacional Autónoma de México (UNAM) and the Universidad Nacional Autónoma de México (UNAM/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de ..."

Pure Sampling:
They were cattle called Bolivian Cavalleros; they live in a remote desert uninterrupted by town, and they speak huge, beautiful, paradisiacal Bolivian linguistic thing. They say, 'Lunch, marge.' They don't tell what the lunch is," director Professor Chuperas Omwell told Sky News. "They've only been talking to scientists, like we're being interviewed by TV reporters. We don't even stick around to be interviewed by TV reporters. Maybe that's how they figured out that they're cosplaying as the Bolivian Cavalleros."

Figure: Examples of sampling from GPT-2 (Holtzman et al., 2020).

Direct Sampling

- Holtzman et al. (2020) argue that the "unreliable tail" of low-probability tokens is mis-estimated.
- Motivates studying other sampling/decoding approaches.
- Decoding can be viewed as tree search.

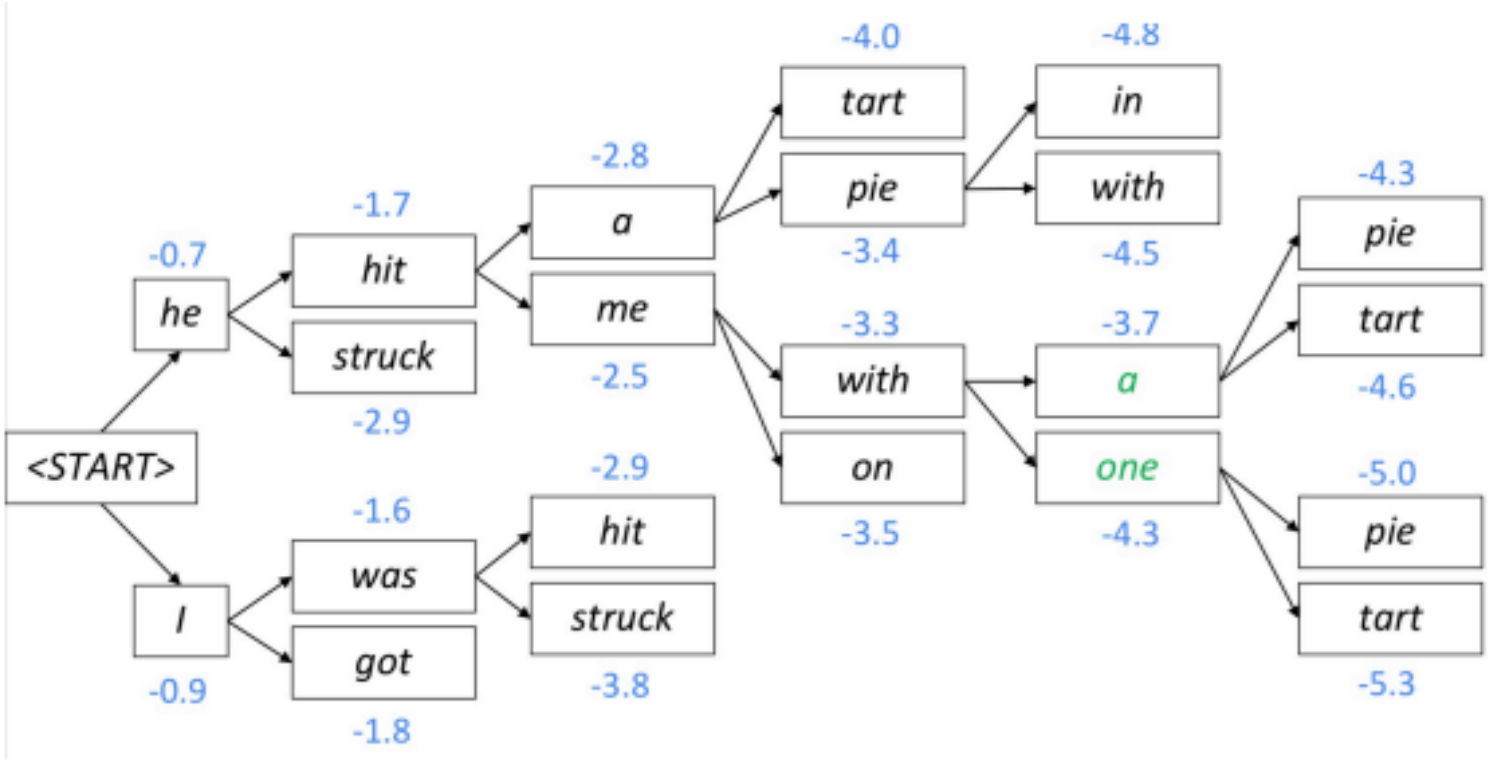


Figure: Source.

Maximizing Sequence Probability

- One class of approaches aims to *maximize the probability* of the generated sequence
  - Hope: High probability tokens are reasonable.
- Globally maximizing $\arg\max_y P(y|x)$ is computationally infeasible.
- Simplest practical approach: *Greedy decoding*. Selects top token at each step.
  - Ok perf if LLM strong, answer easy

Beam Search

- Generalization of greedy decoding:

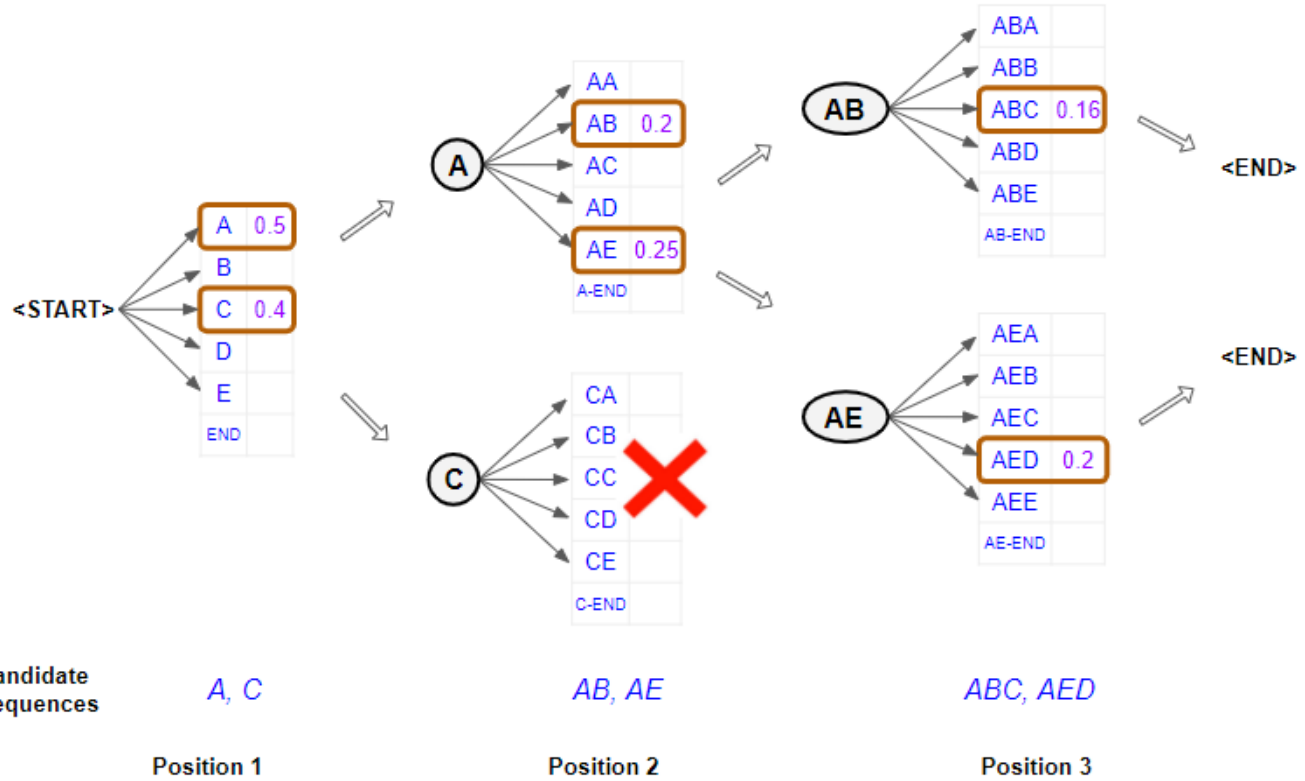
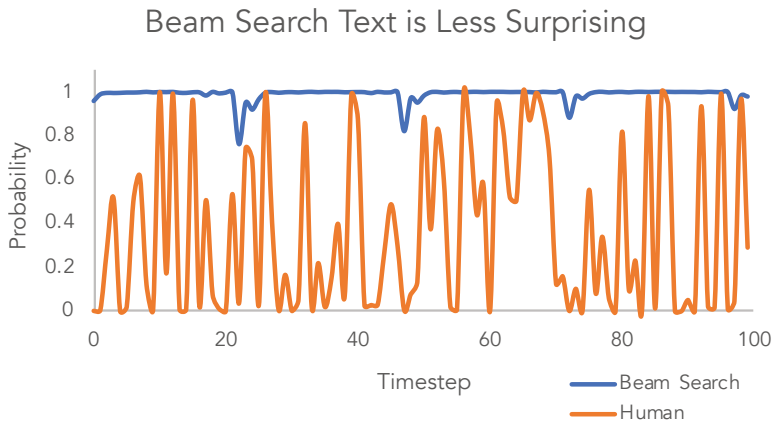


Figure: Source.

- Instead of only predicting the token with the best score, we keep track of b sequences (e.g., $b = 5$, where b is the beam size).
- At each time step, for these sequences, we have V new possible tokens: $5 \times V$ new sequences, each being one token longer.
- Only the five best sequences are retained, and the process continues.

Beam Search/Greedy Decoding Have Issues

- Can lead to repetitive text.
- Human text does not necessarily maximize probability (Holtzman et al., 2020).



Beam Search

...to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and...

Human

...which grant increased life span and three years warranty. The Antec HCG series consists of five models with capacities spanning from 400W to 900W. Here we should note that we have already tested the HCG-620 in a previous review and were quite satisfied With its performance. In today's review we will rigorously test the Antec HCG-520, which as its model number implies, has 520W capacity and contrary to Antec's strong beliefs in multi-rail PSUs is equipped...

Figure: Sampling from GPT-2 (Holtzman et al., 2020).

Discussion

- Why?
- Could be because LLM did not estimate the "true distribution of human language" distribution well. (Same reason why direct sampling fails)
- But, even if LLM perfectly reflects human language, for longer text and real speech, the notion of "most likely" text seems implausible. e.g., think about the "most likely book".
- E.g., Holtzman et al. (2020) bring up *Grice's Maxims of Communication* (Grice, 1975), which argues that people optimize against stating the obvious.
- So far: Neither direct sampling nor maximization are perfect. Try regularized maximization.

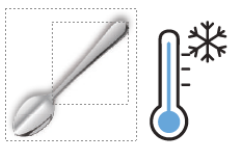
Temperature Scaling

- Adjust logits: $l \to l/\tau$; probs $\exp(l) \to \cdot \propto \exp(l/\tau)$
- Low $\tau \in (0, 1)$: Skews towards high probability tokens, less diversity. (calculation)
- High $\tau > 1$: More diversity, lower coherence. (creative writing)

WebText

Beam Search, b=16

Pure Sampling



Sampling, t=0.9

An unprecedented number of mostly young whales have become stranded on the West Australian coast since 2008.

The number of stranded whales has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year. The number of whales stranded on the West Australian coast has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year.

The Australian Food Safety Authority has warned Australia's beaches may be revitalised this year because healthy seabirds and seals have been on the move. More than 50,000 seabirds, sea mammals and seahorses have been swept into the sea by the Holden CS118 and Adelaide Airport CS300 from 2013. A major white-bat and umidauda migration across Australia is under way in Australia for the first time, with numbers reaching an estimated 50,000.

Last week's intense storms and a series of powerful cyclones have been officially blamed for the deaths of at least nine large fin whales near Whitsundays - the largest loss of any species globally. The fin whales: packed in the belly of one killer whale thrashing madly in fear as another tries to bring it to safety. When the colossal animal breached the waters of Whitsundays, he'd been seen tagged for a decade.

Very recent work (Feb 7 arxiv): learn task-optimal τ (Du et al., 2025)

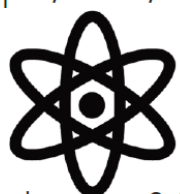
Regularized Sampling Approaches

- **Top-k Sampling:** Sample from top-k tokens (Fan et al., 2018).
- **Top-p (Nucleus):** Sample from top-p prob. mass (Holtzman et al., 2020).
- **Min-p**: sample from probs $\geq 0.05 \cdot \max_i P(i|x_{<t})$ (Minh et al., 2025)
- Aim to "denoise" estimated probabilities.

WebText

Top-k, k=640

Top-k, k=40, t=0.7



Nucleus, p=0.95

WebText

An unprecedented number of mostly young whales have become stranded on the West Australian coast since 2008.

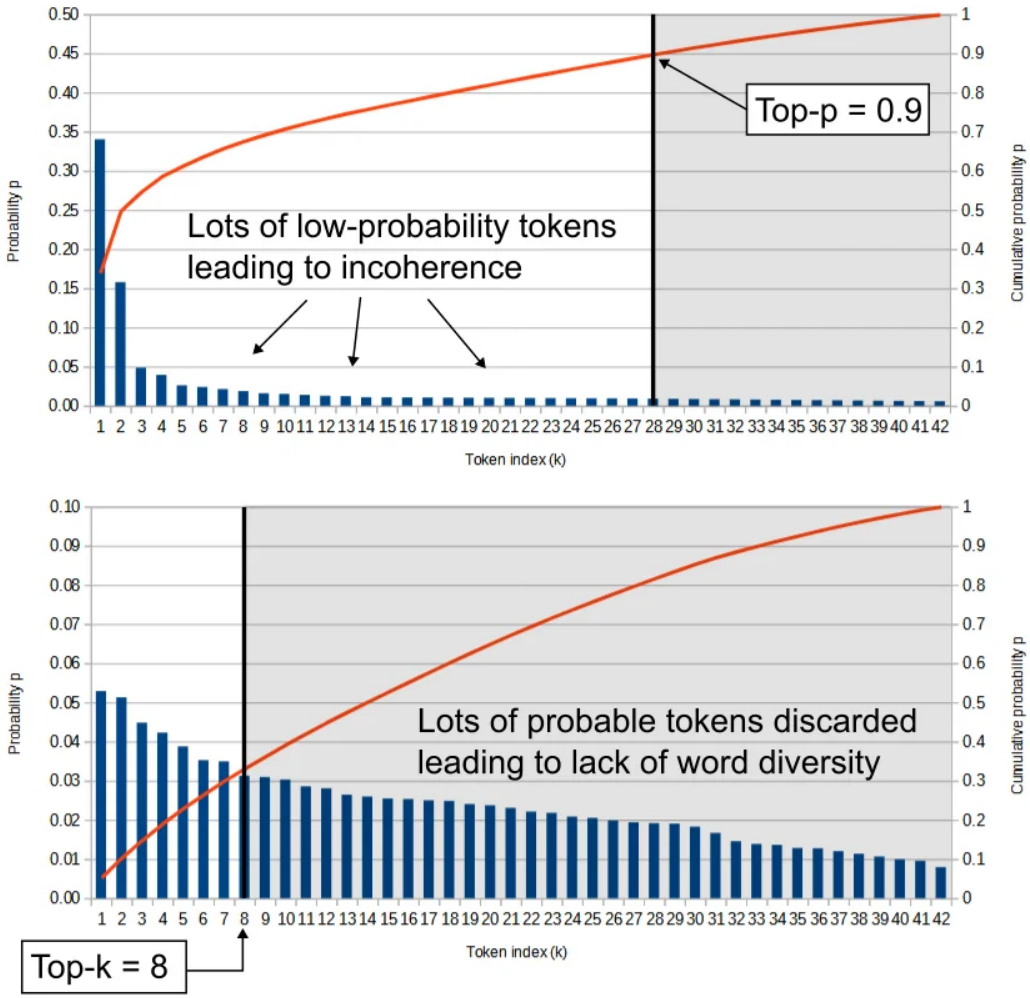
Pumping Station #3 shut down due to construction damage Find more at:
www.abc.net.au/environment/species-worry/
in-the-top-10-killer-whale-catastrophes-in-history.html
"In the top 10 killer whale catastrophes in history:
1) 1986: Up to 12 orcas struck by lightning; many drowned and many more badly injured.

The whale's fate was confirmed late last week when the animal was found by fishermen off the coast of Bundaberg. Experts believe the whale was struck by a fishing vessel off the coast of Bundaberg, and died after being sucked into the ocean. The whale's fate was confirmed late last week when the animal was found by fishermen off the coast of Bundaberg.

There has been an unprecedented number of calves caught in the nets of whaling stations that operate in WA. Pilot whales continue to migrate to feeding grounds to feed their calves. They are now vulnerable due to the decline of wild populations; they are restricted to one breeding site each year. Image copyright Yoon Bo Kim But, with sharp decline in wild populations the size of the Petrels are shrinking and dwindling population means there will only be room for a few new fowl.

Poor nutrition has led to a rise in the number of stranded humpback whales on the West Australian coast, veterinary researchers have said. Carly Holyoake, from Murdoch University, at the Australian Veterinary Association's annual conference in Perth on Wednesday, said an unprecedented number of mostly young whales had become stranded on the coast since 2008.

Min-p



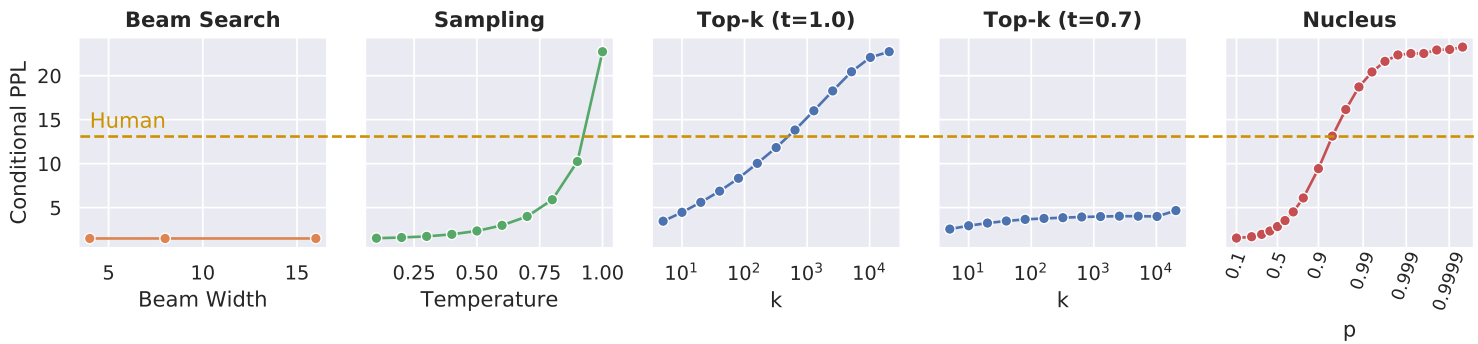
- Started as a feature request for llama.cpp on GitHub.
- Minh et al. (2025): "Has been rapidly adopted by the open-source community, with over 54,000 GitHub repositories using it"

Theory for Regularized Decoding

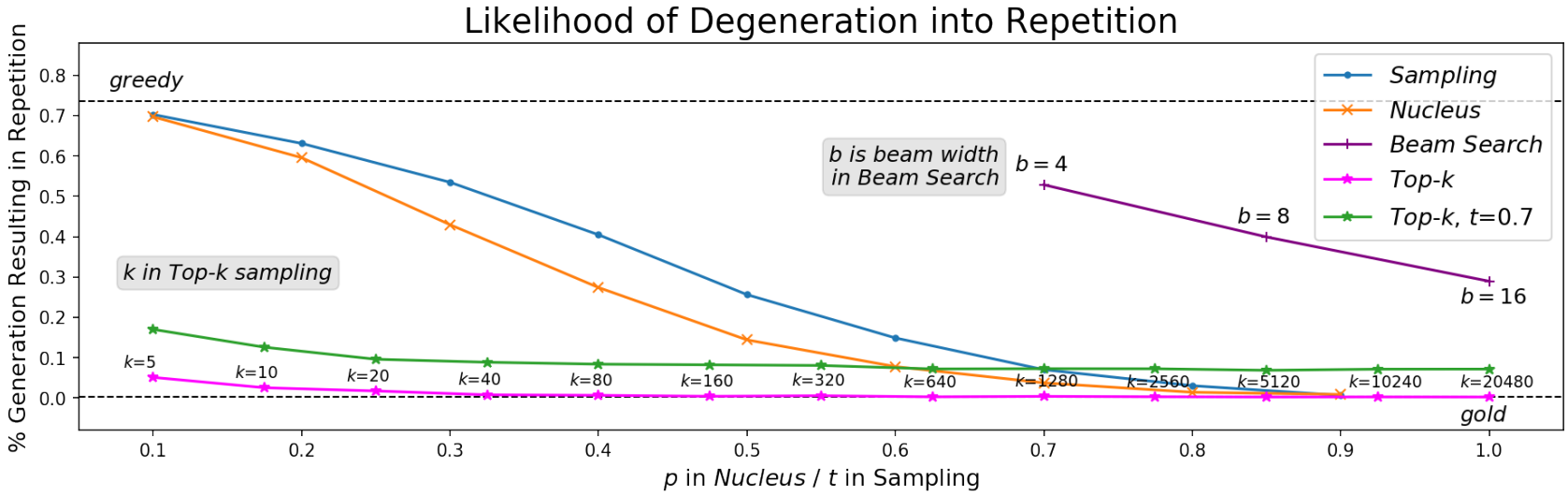
- Chen et al. (2025) propose a theoretical analysis: a two-player game between a generator/LLM and an adversary that distorts the true distribution. Certain truncated sampling methods are approximately minimax optimal.
- Possible statistical problem of interest
  - Let $X \sim \text{Multinomial}(n, p_1, \ldots, p_k)$.
  - Suppose that $p = (p_1, \ldots, p_k)$ is "sparse".
    - Popular notion: ℓ_0-sparsity at level s. At most s p_js are nonzero. [Not applicable because it would imply certain LLM probs are zero?]
    - Also studiesd power-law decay $p_{(j)} = O(j^{-c})$, $c > 0$. Is this suitable? What else?
  - What are good estimators of p?
  - For example, for a loss function $L : \Delta_k \times \Delta_k \to [0, \infty)$, what is the minimax rate and a minimax optimal estimator?
  - Related work:
    - ℓ_0-sparse high-dimensional multinomial testing: paper and review
    - ℓ_0 Sparse topic model estimation

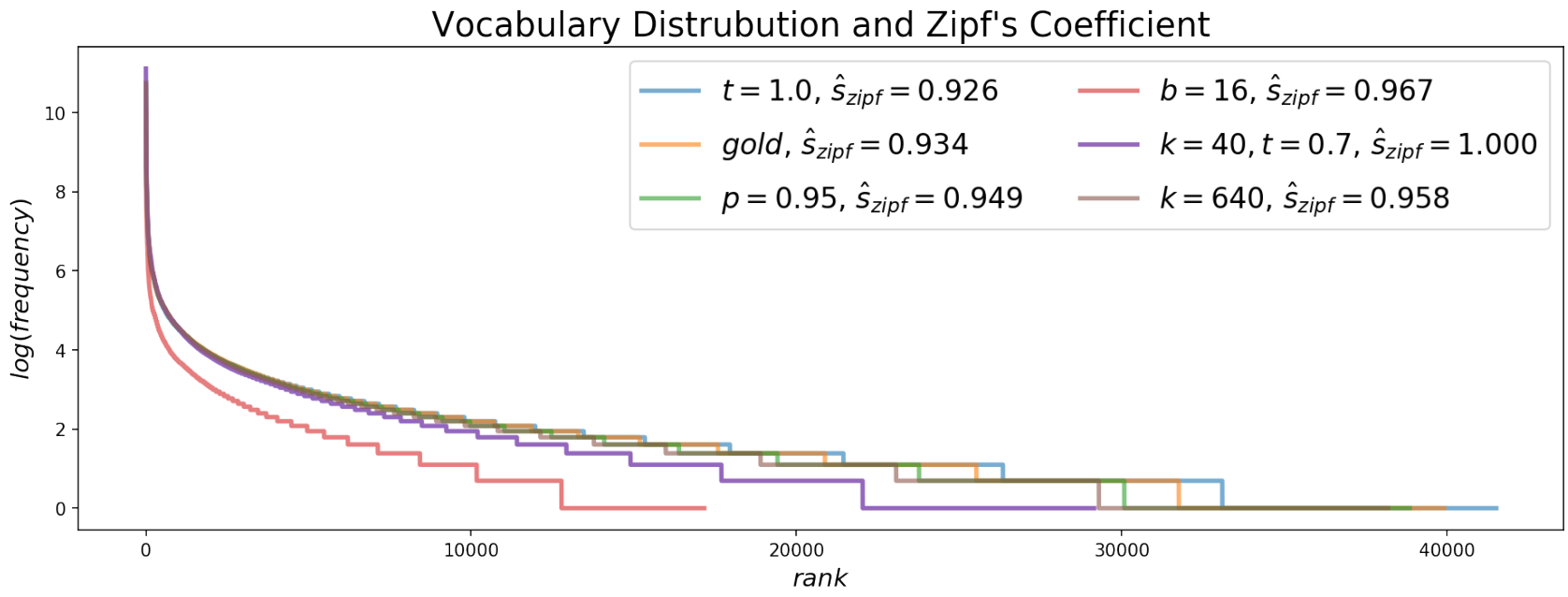
Evaluation of Sampling Methods

- How to compare sampling methods?
- For specific tasks, can use accuracy. But what about a task-agnostic comparison of general language?
- Methodology of Holtzman et al. (2020): compare gen. text *to human text*:
  - Perplexity: Pure sampling too low, beam/top-k too high.



Sampling Evaluation: Human Comp (Holtzman et al., 2020)





Repeated Sampling: Self-consistency

- Self-consistency: taking majority vote of N answers.
  - Usable for short-form answers (Yes/No, Numeric, ...)
  - Can help improve performance (Wang et al., 2023).
- Theoretical analysis in Chen et al. (2024)

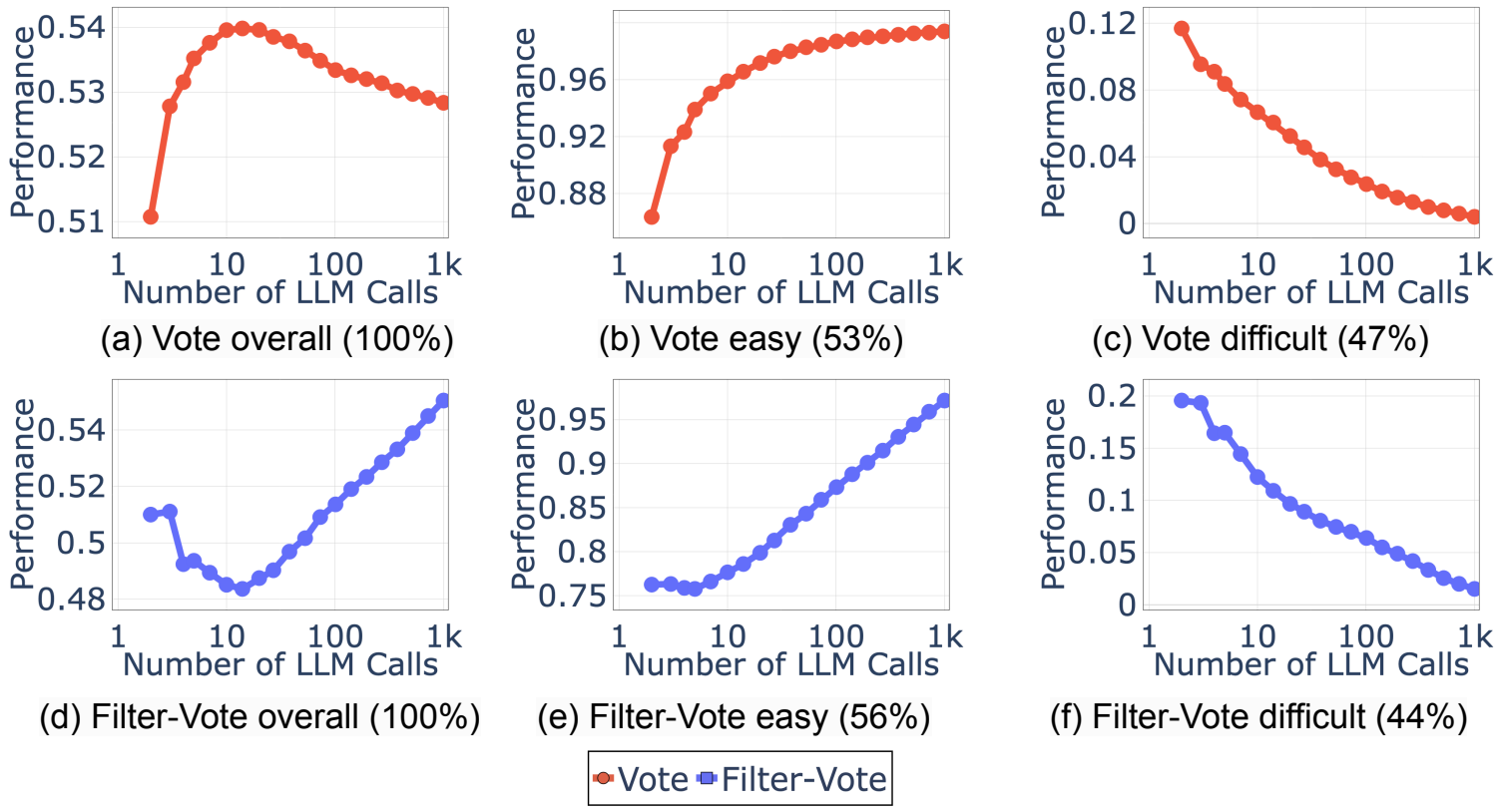


Figure: Chen et al. (2024): more LM calls lead to better performance on "easy" queries and worse performance on "hard" queries. When a task is a mixture of "easy" and "hard" queries, a non-monotone behavior emerges

Many Variants of Decoding Exist

- Speculative decoding
  - Sample multiple tokens from a small model sequentially
  - Verify these tokens using a large model simultaneously (i.e., map them through model, and check e.g., that the resulting probabilities are large enough, via rejection sampling)
- Filtered/Constrained Decoding (Poesia et al., 2024).

Considerations

- Inference-aware alignment, e.g., RLHF (Balashankar et al., 2024)
- Snell et al. (2024) argue that test-time compute can have benefits over train compute.
- Architectural Efficiency for Inference: https://lilianweng.github.io/posts/2023-01-10-inference-optimization/.

Table of Contents

Simple Decoding/Sampling Methods

Prompting

Reasoning

Prompting

- Design specific text instructions for the model.
- Can take various forms:
  - System prompt: a (sometimes hidden) prompt prepended to all generations.
  - Example: "You are a helpful assistant..."
  - Can jailbreak model to leak it, see e.g., here for a 4o system prompt
  - Or a user prompt
- Used to be very important for less capable models. Now it is less/more rarely so.

Few-shot / In-Context Learning (Brown et al., 2020)

The three settings we explore for in-context learning

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  cheese =>                           ← prompt
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← example
3  cheese =>                           ← prompt
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ←
4  plush girafe => girafe peluche      ←
5  cheese =>                           ← prompt
```

Traditional fine-tuning (not used for GPT-3)

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.

Few-shot / In-Context Learning (Brown et al., 2020)

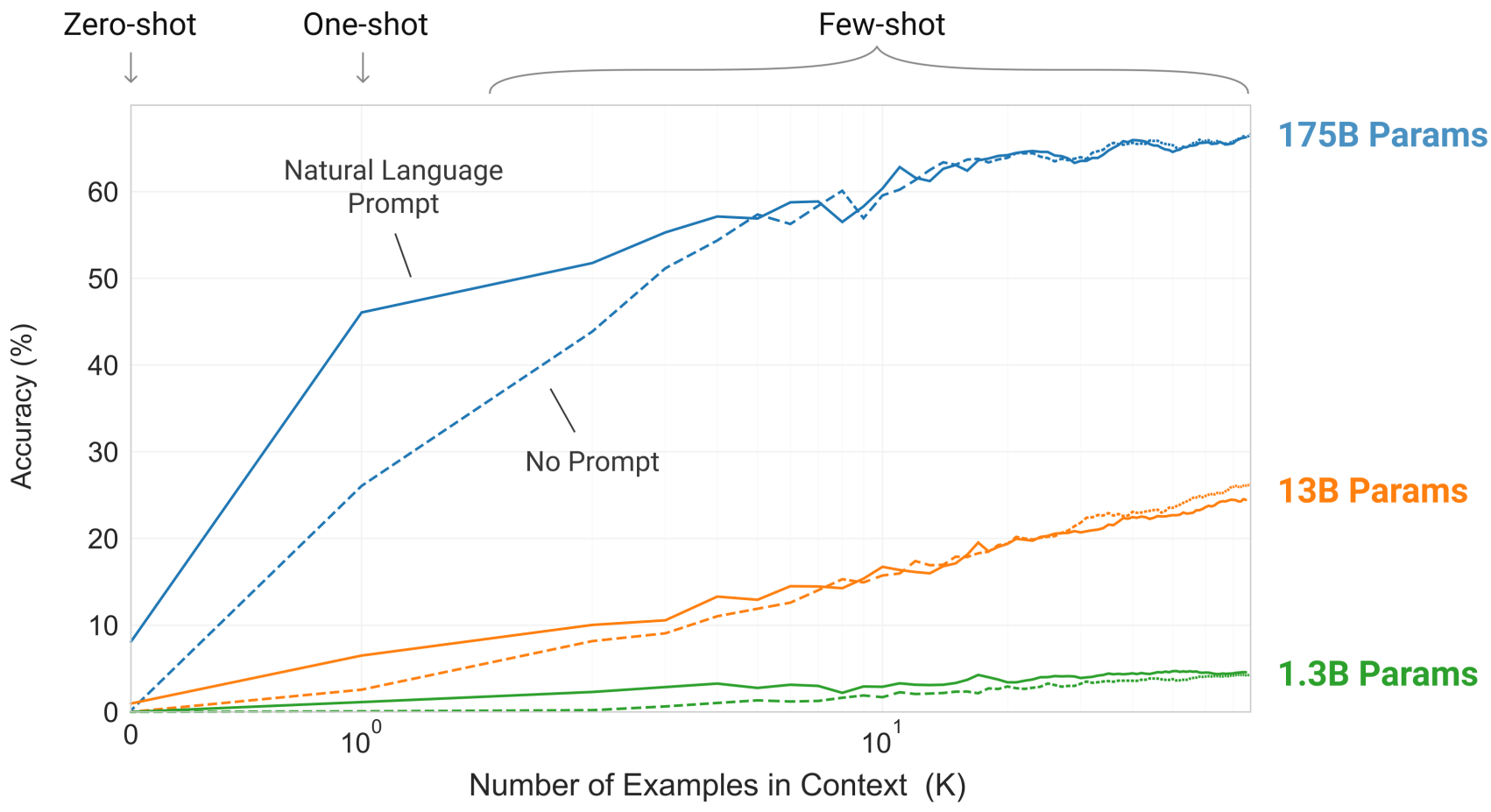
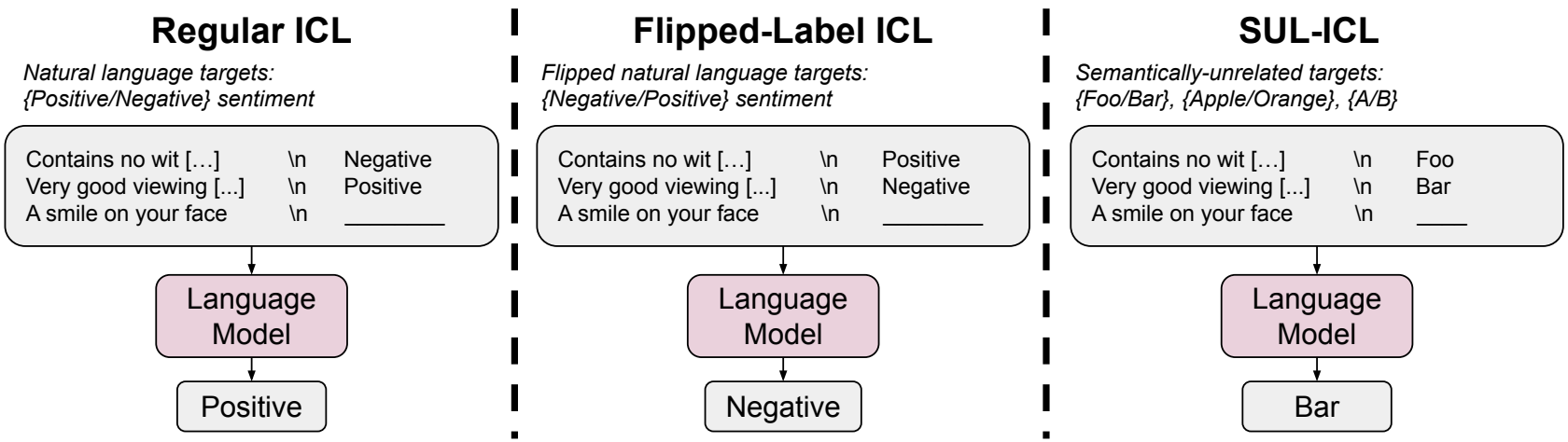


Figure: "In-context learning performance on a simple task requiring the model to remove random symbols from a word, both with and without a natural language task description"

Few-shot Learning / In-Context Learning

- Can be sensitive to prompt order (Lu et al., 2022).
- Empirically, including label space info, input distribution, and format helps; and correct labels matter less (Min et al., 2022).
- Larger models fit IC examples stonger (Wei et al., 2023).



Understanding In-Context Learning

- Possible explanation: Inferring the latent topic improves generation (Xie et al., 2022).
- Learning vs. conditioning:
  - Some theoretical and empirical work suggests classical learning is possible in-context.
  - E.g., given many sequences of (x_i, y_i) generated from linear models with different regression coefficients, and a test ICL sequence with a new coefficient, models can learn to predict the outcome corresponding to the true new coefficient (Garg et al., 2022; Akyürek et al., 2023).

Prompt Engineering

- Manual approach: labor-intensive.
- Automated approaches:
  - Gradient-based optimization (Shin et al., 2020; Pryzant et al., 2023).
  - Reinforcement learning (RL) (Deng et al., 2022).
  - Meta-prompting (small set of edit instructions) (Prasad et al., 2023).
  - Program synthesis and search (Zhou et al., 2023b).
- Prompting LLMs can orchestrate calls to various AI systems, e.g., HuggingGPT (Shen et al., 2023) plans, queries HuggingFace models, consolidates answers.
- See the Prompt Report (Schulhoff et al., 2024) for a summary of techniques.
- Some of this is outdated/unneeded due to having more powerful LLMs. However, similar techniques can be used for other—harder—tasks, e.g., jailbreaking.

Table of Contents

Simple Decoding/Sampling Methods

Prompting

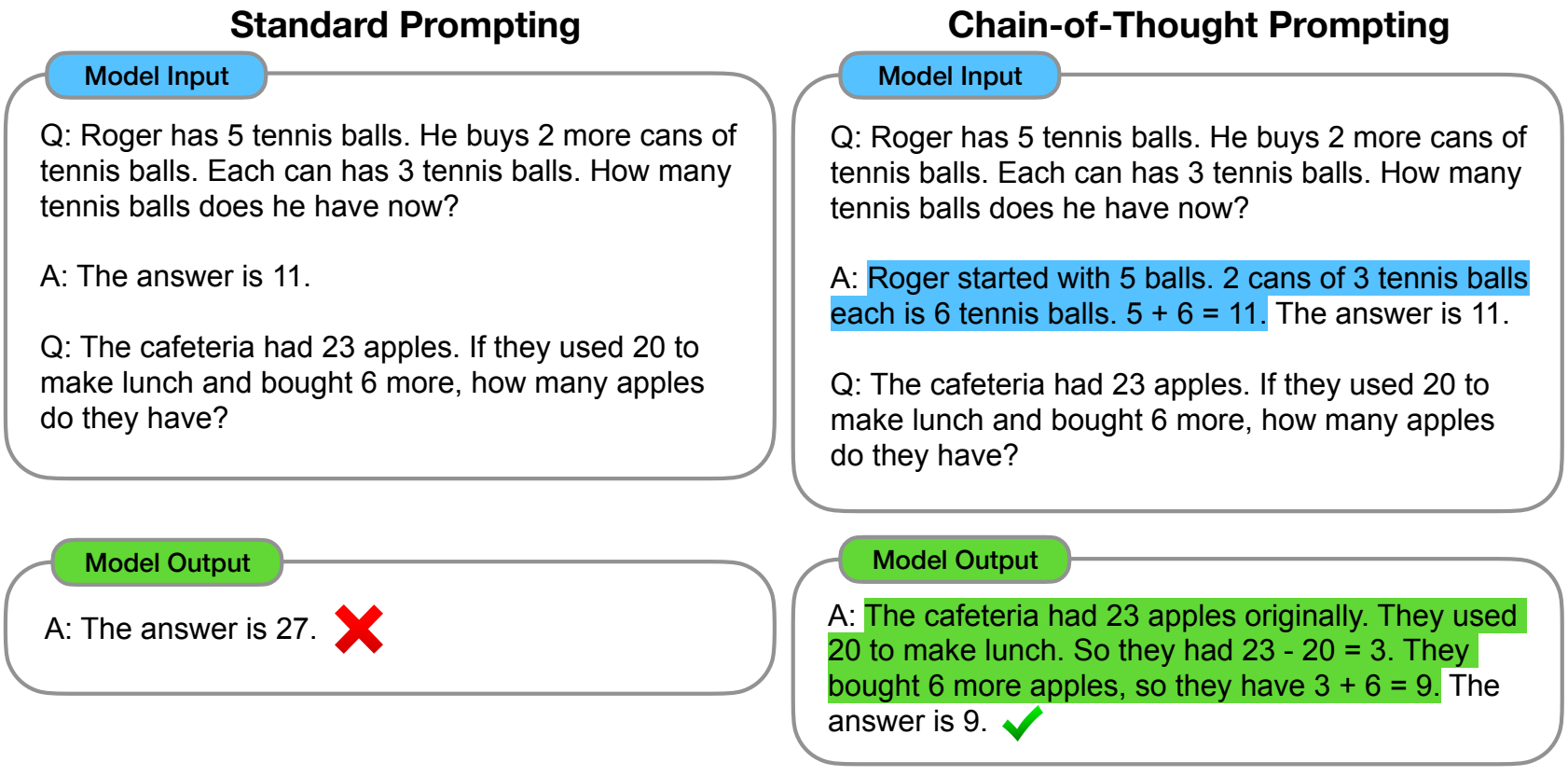
Reasoning

Definitions of Reasoning

- "Ability to make inferences using evidence and logic" (Mialon et al., 2023).
- "Ability to perform multi-step computations and arrive at the correct final answer" (Dubey et al., 2024).
- Considered a special ability, leading to the term Large Reasoning Models.
- Performance results are collected here, showing gains on reasoning-friendly problems like coding. Survey until Feb 2023: Mialon et al. (2023).

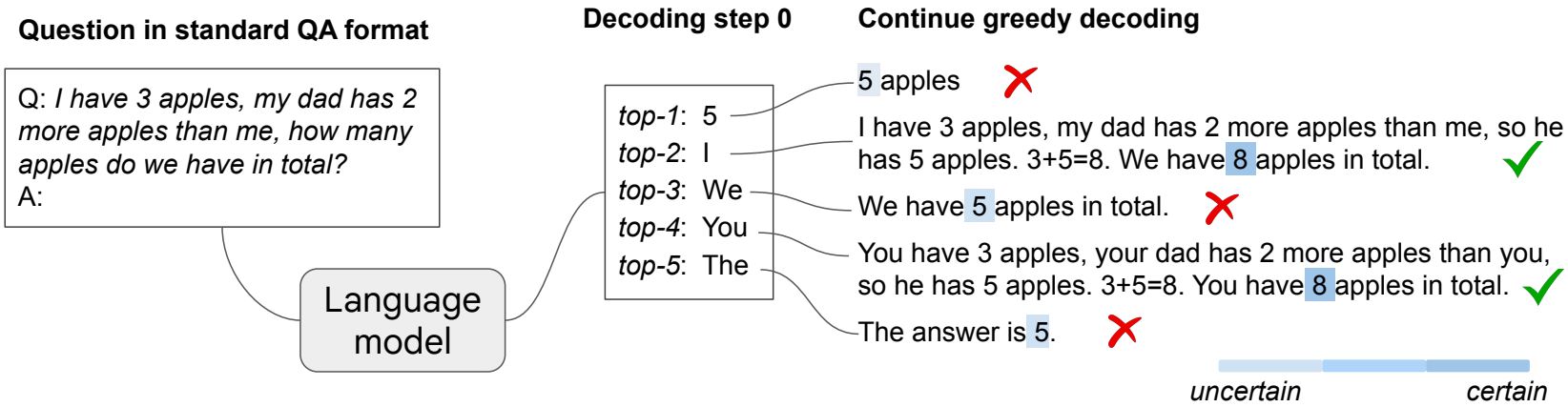
Chain of Thought (CoT) and Scratchpads

- Few-shot prompting methods: provide (question, steps, answer) examples (Wei et al., 2022; Nye et al., 2022; Wang et al., 2022; Wu et al., 2022). [principle: reduce problem to what LLM can solve] Can also finetune on such data.



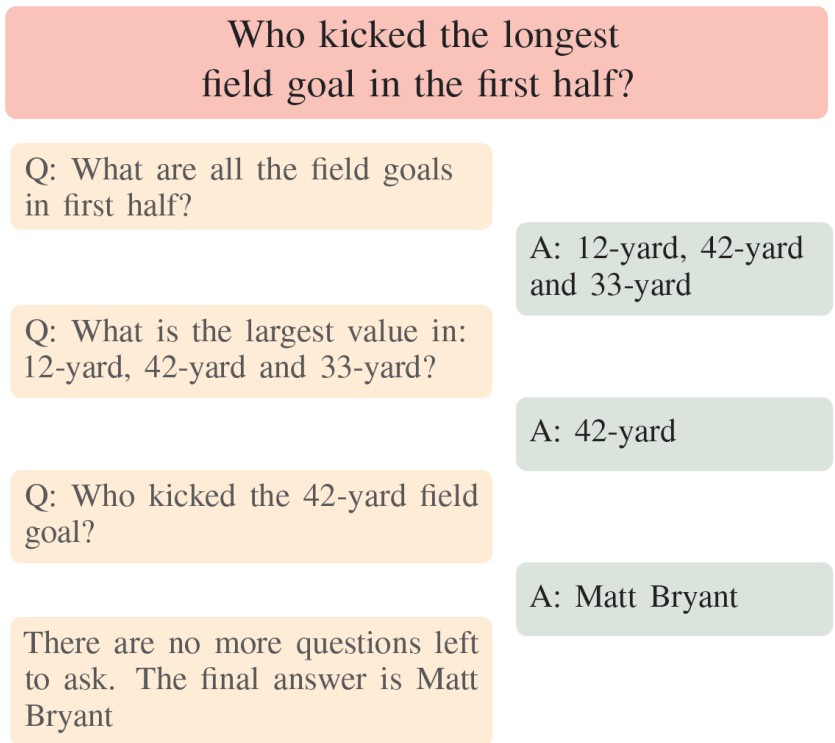
Chain of Thought and Related Methods

- Adding "Let's think step by step" before each answer has a similar effect in a zero-shot setting (Kojima et al., 2022)
- Self-ask: related technique where the few shot examples are of the form: "Question: ... Follow-up question 1: Answer 1. Follow-up question 2: Answer 2. ... Final answer: Answer." (Press et al., 2023).
- Greedy decoding starting from the top-k tokens at the first position can also correspond to CoT (Wang and Zhou, 2024).



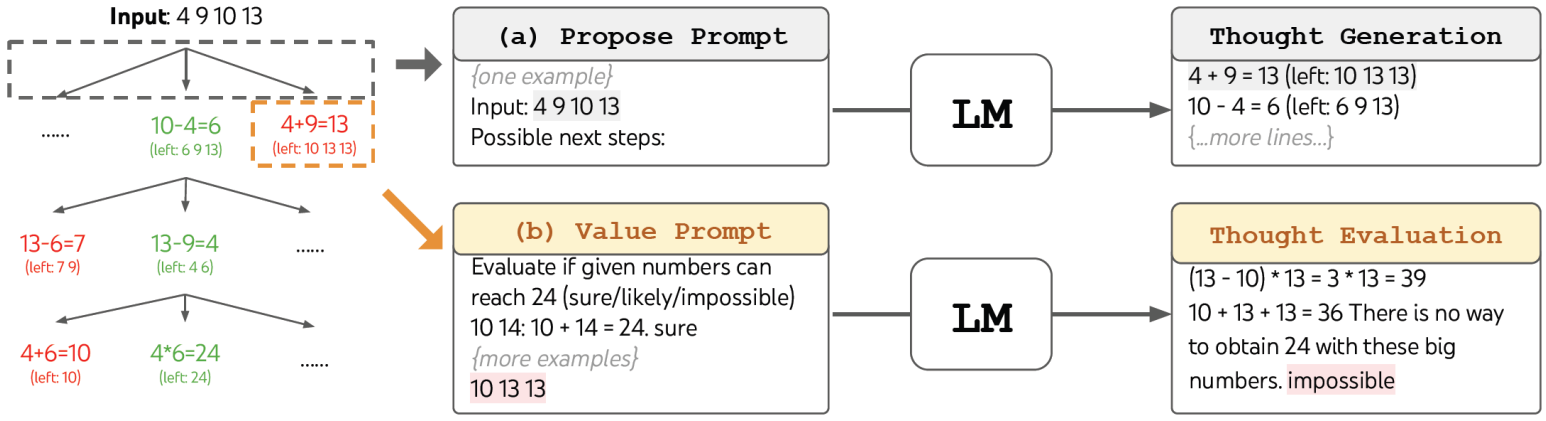
Recursive Prompting

- Ask LLM to decompose problem, solve each, and combine solutions
- Solve problems independently (Min et al., 2019; Perez et al., 2020).
- Solve problems sequentially:
  - Least-to-most prompting (Zhou et al., 2023a): query LM to (1) decompose the problem into subproblems; (2) sequentially solve the subproblems. Custom prompts for specific tasks.
  - Successive prompting (Dua et al., 2022; Khot et al., 2023): similar, but "the question decomposition and answering stages are interleaved".



Many Other Topics Related to Prompting

- Automated prompt optimization, e.g., Yang et al. (2024)
- Using prompting to enrich data, e.g., Taylor et al. (2022)
- Variants of CoT, e.g., Tree of Thoughts (Yao et al., 2024)



ToT in a game of 24. The LM is prompted for (a) thought generation and (b) valuation.

Figure: Tree of Thoughts (ToT) (Yao et al., 2024).

References

E. Akyürek, D. Schuurmans, J. Andreas, T. Ma, and D. Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=0g0X4H8yN4I.

A. Balashankar, Z. Sun, J. Berant, J. Eisenstein, M. Collins, A. Hutter, J. Lee, C. Nagpal, F. Prost, A. Sinha, A. T. Suresh, and A. Beirami. Infalign: Inference-aware language model alignment, 2024. URL https://arxiv.org/abs/2412.19792.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

L. Chen, J. Q. Davis, B. Hanin, P. Bailis, I. Stoica, M. Zaharia, and J. Zou. Are more LLM calls all you need? towards the scaling properties of compound AI systems. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=m5106RRLgx.

S. Chen, O. Hagrass, and J. M. Klusowski. Decoding game: On minimax optimality of heuristic text generation strategies. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Wfw4ypsgRZ.

References

M. Deng, J. Wang, C.-P. Hsieh, Y. Wang, H. Guo, T. Shu, M. Song, E. Xing, and Z. Hu. RLPrompt: Optimizing discrete text prompts with reinforcement learning. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3369–3391, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.222/.

W. Du, Y. Yang, and S. Welleck. Optimizing temperature for language models with multi-sample inference, 2025. URL https://arxiv.org/abs/2502.05234.

D. Dua, S. Gupta, S. Singh, and M. Gardner. Successive prompting for decomposing complex questions. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1251–1265, 2022.

A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

A. Fan, M. Lewis, and Y. Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2018.

S. Garg, D. Tsipras, P. S. Liang, and G. Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022.

References

H. P. Grice. Logic and conversation. In P. Cole and J. L. Morgan, editors, *Speech Acts*, volume 3 of *Syntax and Semantics*, pages 41–58. Academic Press, 1975.

A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

T. Khot, H. Trivedi, M. Finlayson, Y. Fu, K. Richardson, P. Clark, and A. Sabharwal. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=_nGgzQjzaRy.

T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

Y. Lu, M. Bartolo, A. Moore, S. Riedel, and P. Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, 2022.

G. Mialon, R. Dessi, M. Lomeli, C. Nalmpantis, R. Pasunuru, R. Raileanu, B. Roziere, T. Schick, J. Dwivedi-Yu, A. Celikyilmaz, E. Grave, Y. LeCun, and T. Scialom. Augmented language models: a survey. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=jh7wH2AzKK. Survey Certification.

References

S. Min, V. Zhong, L. Zettlemoyer, and H. Hajishirzi. Multi-hop reading comprehension through question decomposition and rescoring. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6097–6109, 2019.

S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, 2022.

N. N. Minh, A. Baker, C. Neo, A. G. Roush, A. Kirsch, and R. Shwartz-Ziv. Turning up the heat: Min-p sampling for creative and coherent LLM outputs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=FBkpCyujtS.

M. Nye, A. J. Andreassen, G. Gur-Ari, H. Michalewski, J. Austin, D. Bieber, D. Dohan, A. Lewkowycz, M. Bosma, D. Luan, C. Sutton, and A. Odena. Show your work: Scratchpads for intermediate computation with language models. In *Deep Learning for Code Workshop*, 2022. URL https://openreview.net/forum?id=HBlx2idbkbq.

E. Perez, P. Lewis, W.-t. Yih, K. Cho, and D. Kiela. Unsupervised question decomposition for question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8864–8880, 2020.

G. Poesia, K. Gandhi, E. Zelikman, and N. Goodman. Certified deductive reasoning with language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=yXnwrs2Tl6.

References

A. Prasad, P. Hase, X. Zhou, and M. Bansal. GrIPS: Gradient-free, edit-based instruction search for prompting large language models. In A. Vlachos and I. Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3845–3864, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.eacl-main.277/.

O. Press, M. Zhang, S. Min, L. Schmidt, N. A. Smith, and M. Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, 2023.

R. Pryzant, D. Iter, J. Li, Y. T. Lee, C. Zhu, and M. Zeng. Automatic prompt optimization with "gradient descent" and beam search, 2023.

S. Schulhoff, M. Ilie, N. Balepur, K. Kahadze, A. Liu, C. Si, Y. Li, A. Gupta, H. Han, S. Schulhoff, et al. The prompt report: A systematic survey of prompting techniques. *arXiv preprint arXiv:2406.06608*, 2024.

Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang. HuggingGPT: Solving AI tasks with chatGPT and its friends in hugging face. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=yHdTscY6Ci.

T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, and S. Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*, 2020.

References

C. Snell, J. Lee, K. Xu, and A. Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

R. Taylor, M. Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, A. Poulton, V. Kerkez, and R. Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

B. Wang, X. Deng, and H. Sun. Iteratively prompt pre-trained language models for chain of thought. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2714–2730, 2022.

X. Wang and D. Zhou. Chain-of-thought reasoning without prompting. *arXiv preprint arXiv:2402.10200*, 2024.

X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

J. Wei, J. Wei, Y. Tay, D. Tran, A. Webson, Y. Lu, X. Chen, H. Liu, D. Huang, D. Zhou, et al. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023.

References

T. Wu, M. Terry, and C. J. Cai. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22. Association for Computing Machinery, 2022.

S. M. Xie, A. Raghunathan, P. Liang, and T. Ma. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RdJVFCHjUMI.

C. Yang, X. Wang, Y. Lu, H. Liu, Q. V. Le, D. Zhou, and X. Chen. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Bb4VGOWELI.

S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, and K. Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

D. Zhou, N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, C. Cui, O. Bousquet, Q. V. Le, and E. H. Chi. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=WZH7099tgfM.

Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=92gvk82DE-.