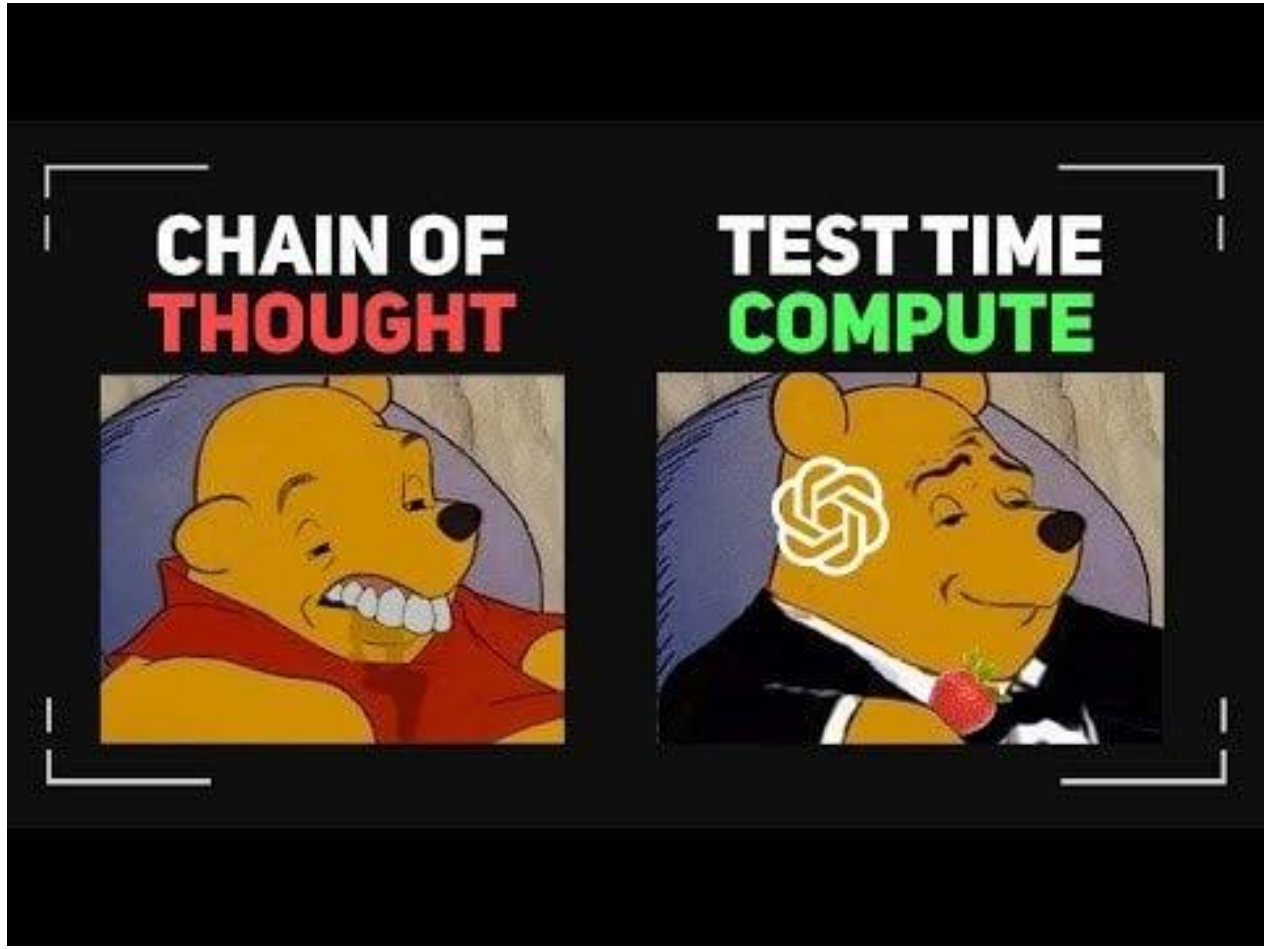


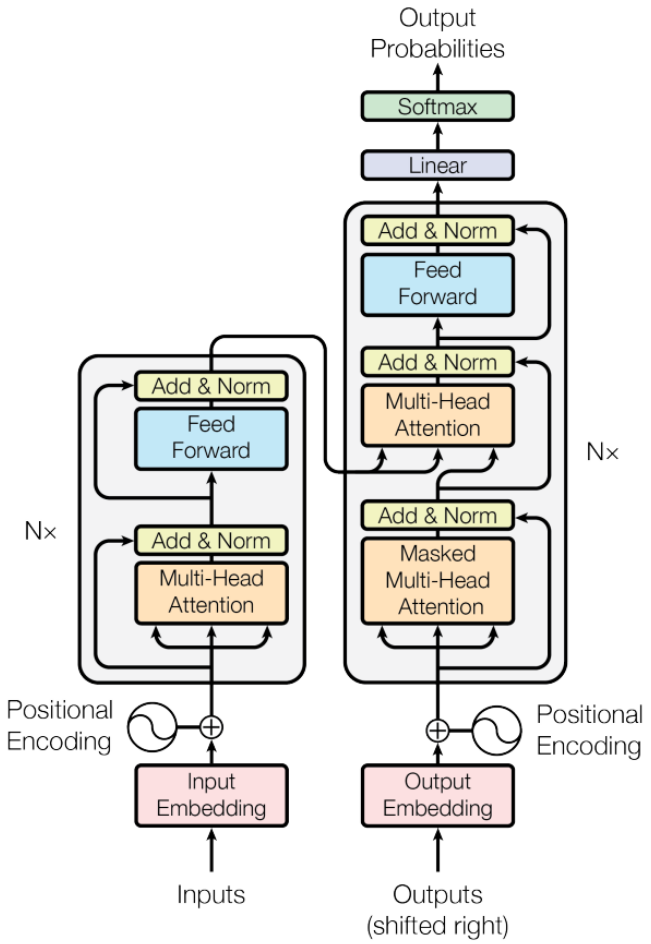
Test Time Scaling

Lecture 9

Overview

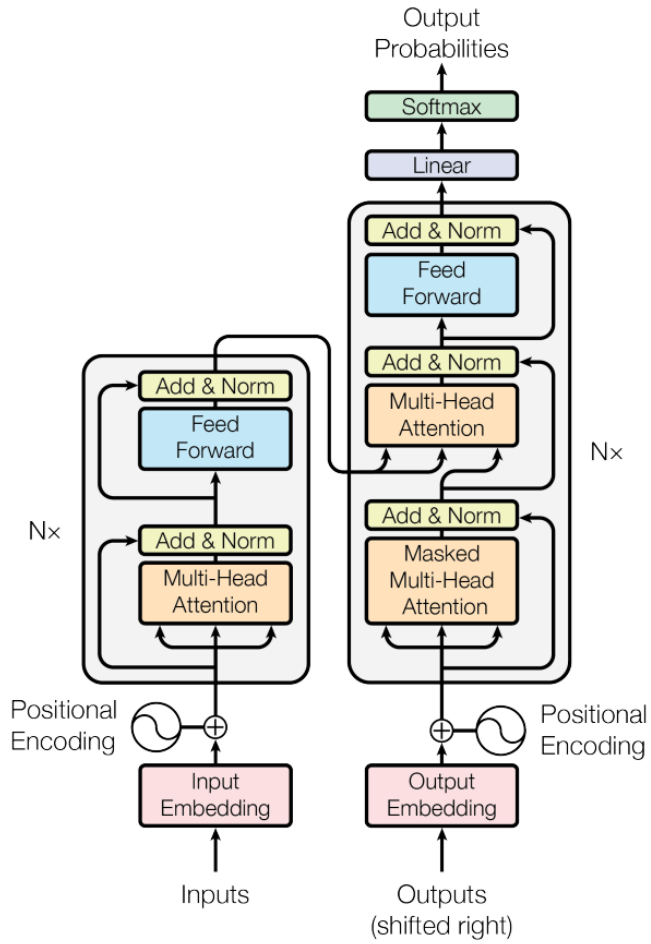
- Introduction
- Basics: LLM Generation
- Prompting techniques:
  - ICL
  - CoT
  - Self-Consistency
  - Least-to-Most

LLM Architecture Development Progress



Vaswani et al. (2017)
Attention is All You Need

7 Years

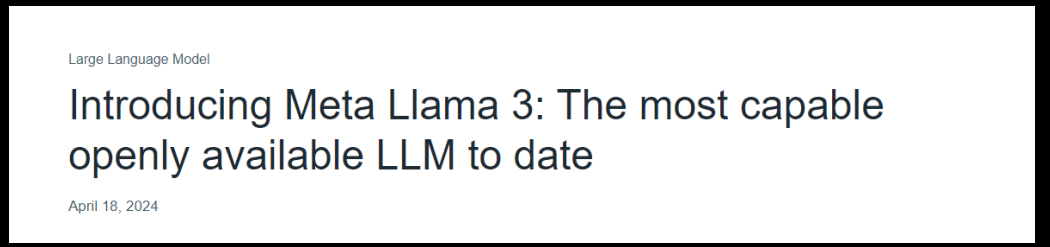


LLaMA 3.1 (2024)
GPT-4 (2024)
…

https://github.com/meta-llama/llama3/blob/main/llama/model.py

```python
class TransformerBlock(nn.Module):
    def __init__(self, layer_id: int, args: ModelArgs):
        super().__init__()
        self.n_heads = args.n_heads
        self.dim = args.dim
        self.head_dim = args.dim // args.n_heads
        self.attention = Attention(args)
        self.feed_forward = FeedForward(
            dim=args.dim,
            hidden_dim=4 * args.dim,
            multiple_of=args.multiple_of,
            ffn_dim_multiplier=args.ffn_dim_multiplier,
        )
        self.layer_id = layer_id
        self.attention_norm = RMSNorm(args.dim, eps=args.norm_eps)
        self.ffn_norm = RMSNorm(args.dim, eps=args.norm_eps)

    def forward(
        self,
        x: torch.Tensor,
        start_pos: int,
        freqs_cis: torch.Tensor,
        mask: Optional[torch.Tensor],
    ):
        h = x + self.attention(self.attention_norm(x), start_pos, freqs_cis, mask)
        out = h + self.feed_forward(self.ffn_norm(h))
        return out
```



Architectural Changes

Post-training

Zero-shot learning

- One key emergent ability in GPT-2 is **zero-shot learning**: the ability to do many tasks with **no examples**, and **no gradient updates**, by simply:

- Specifying the right sequence prediction problem (e.g. question answering):
  `Passage: Tom Brady... Q: Where was Tom Brady born? A: ...`

- Comparing probabilities of sequences:
  `The cat couldn't fit into the hat because it was too big.`
  `Does it = the cat or the hat?`
  `≡ Is P(...because the cat was too big) >=`
  `P(...because the hat was too big)?`

Zero-shot learning

- You can get interesting zero-shot behavior if you're creative enough with how you specify your task!
- Summarization on CNN/DailyMail dataset (See et al., 2017).

```
SAN FRANCISCO, California (CNN) -- A
magnitude 4.2 earthquake shook the San
Francisco ... overturn unstable objects.
TL;DR:
```

	R-1	R-2	R-L	R-AVG
Bottom-Up Sum	**41.22**	**18.68**	**38.34**	**32.75**
Lede-3	40.38	17.66	36.62	31.55
Seq2Seq + Attn	31.33	11.81	28.83	23.99
GPT-2 TL;DR:	29.34	8.27	26.58	21.40
Random-3	28.78	8.63	25.52	20.98
GPT-2 no hint	21.58	4.03	19.47	15.03

Table 4. Summarization performance as measured by ROUGE F1 metrics on the CNN and Daily Mail dataset. Bottom-Up Sum is the SOTA model from (Gehrmann et al., 2018)

Generation

- Steps:
  - After processing the input text (prompt)
  - Predict the next token (choose the token with the highest prob.).
  - Choose one token
  - Repeat
- This is referred to as *greedy* decoding

Temperature

- Remember softmax?

$$\mathrm{Softmax}(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)}$$

- We can add a rescaling hyperparameter (temperature **τ**)

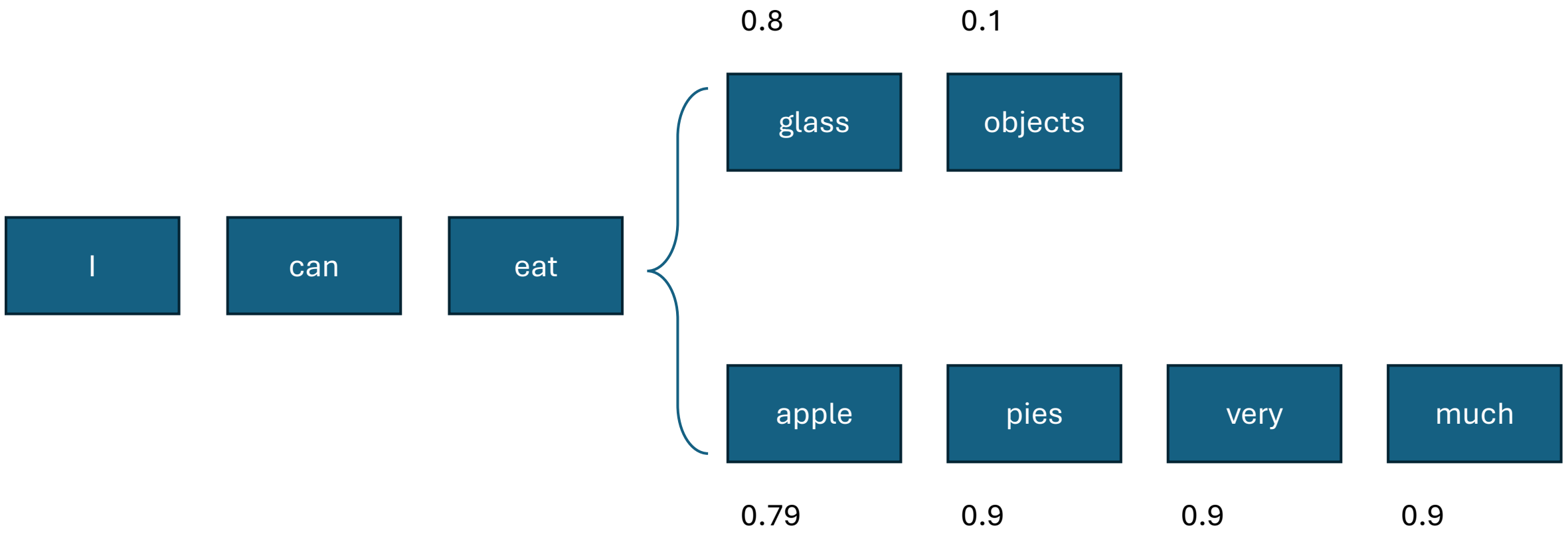
$$\mathrm{Softmax}(x_i) = \frac{\exp(x_i/\tau)}{\sum_j \exp(x_j/\tau)}$$

Temperature

$$\mathrm{Softmax}(x_i) = \frac{\exp(x_i/\tau)}{\sum_j \exp(x_j/\tau)}$$

- $\tau = 1$: no rescaling.
- $\tau > 1$: distribution flattens.
  - Lower-probability tokens get relatively more mass.
  - The model becomes more “creative.”
  - More diversity, but higher risk of errors or incoherence.

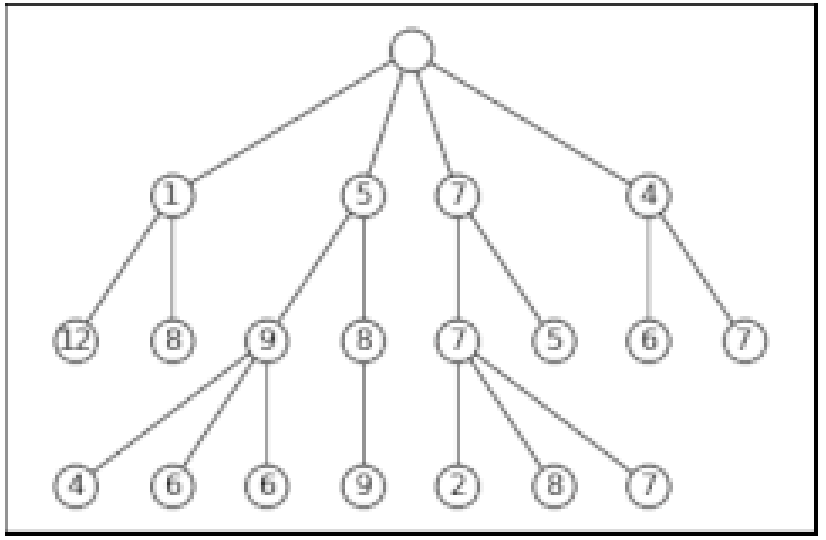
Problem with Greedy Decoding



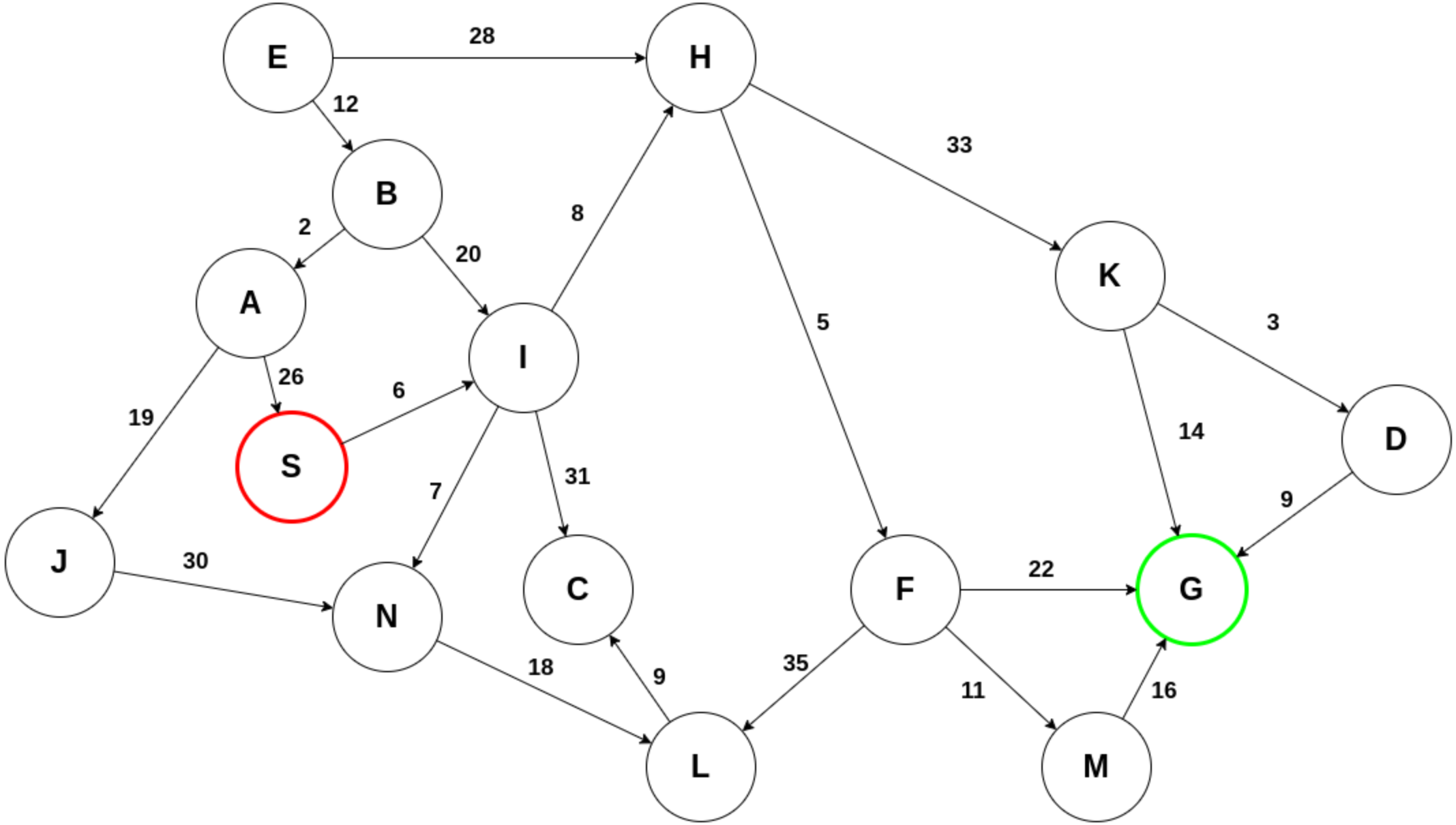
Beam Search: top-*K* Greedy

- Core idea: track the K top choices (most probable) of tokens at each step of decoding.
- K is also called the “beam width” or “beam size.”
  - Where, $5 \leq K \leq 10$ usually in practice.
- Recall the score of a hypothesis ($x_1, \ldots, x_T$) is its log probability:

$$\log P(x_1, \ldots, x_T) = \sum_{i=1}^{T} \log P(x_i | x_{1:i-1})$$



- Beam search does **not** guarantee finding the optimal solution.
- However, much more efficient and practical than exhaustive search.

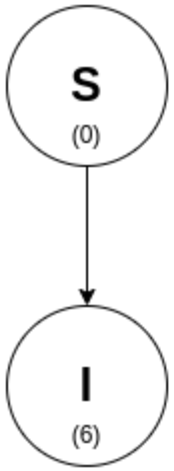


https://www.codecademy.com/resources/docs/ai/search-algorithms/beam-search

S to G, beam size = 2.

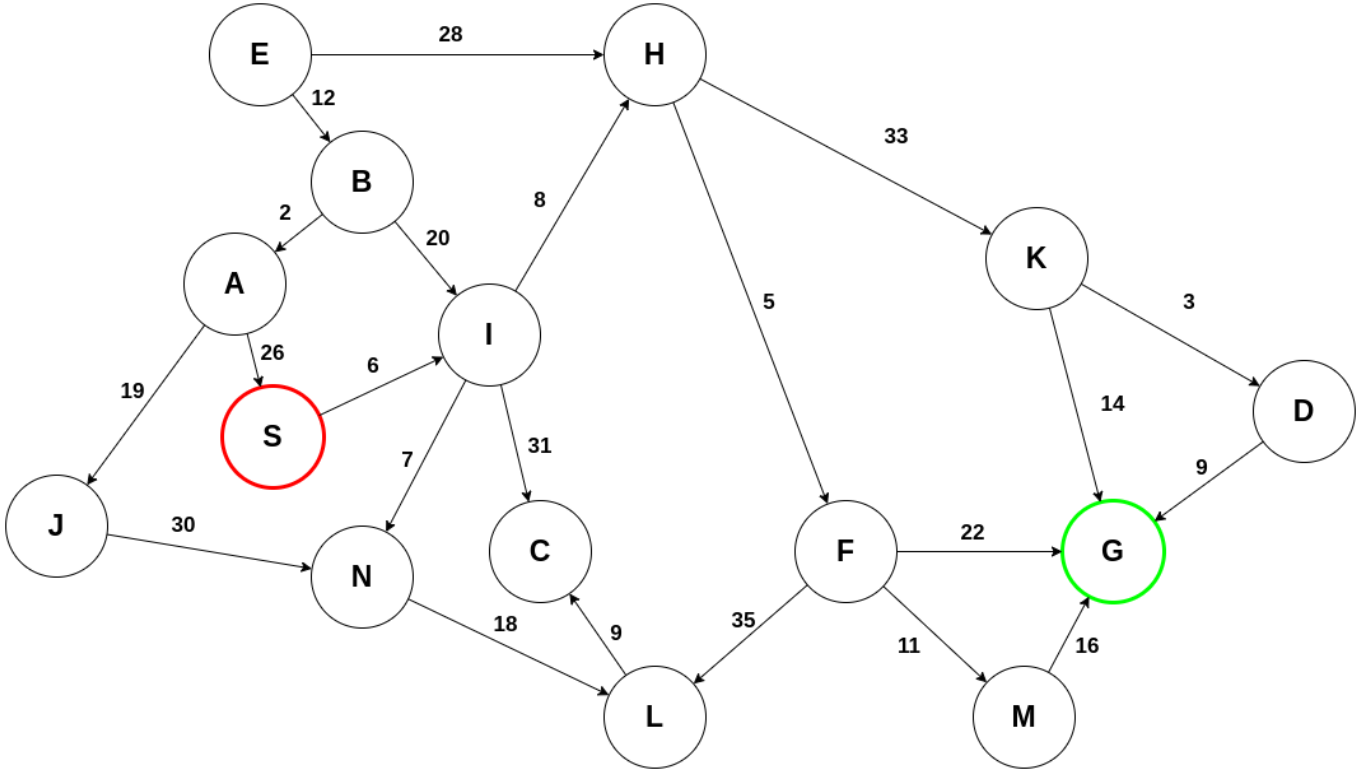
Node[cost]
I[6]

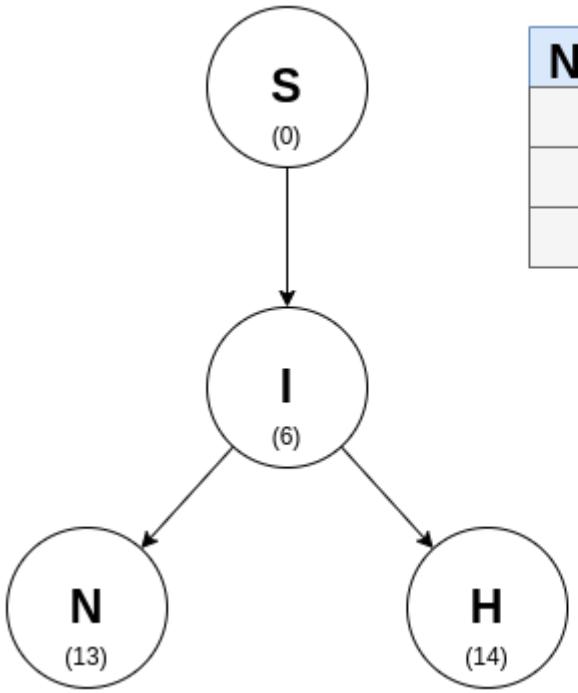
Closed List
S



Node[cost]
N[13]
H[14]
C[37]

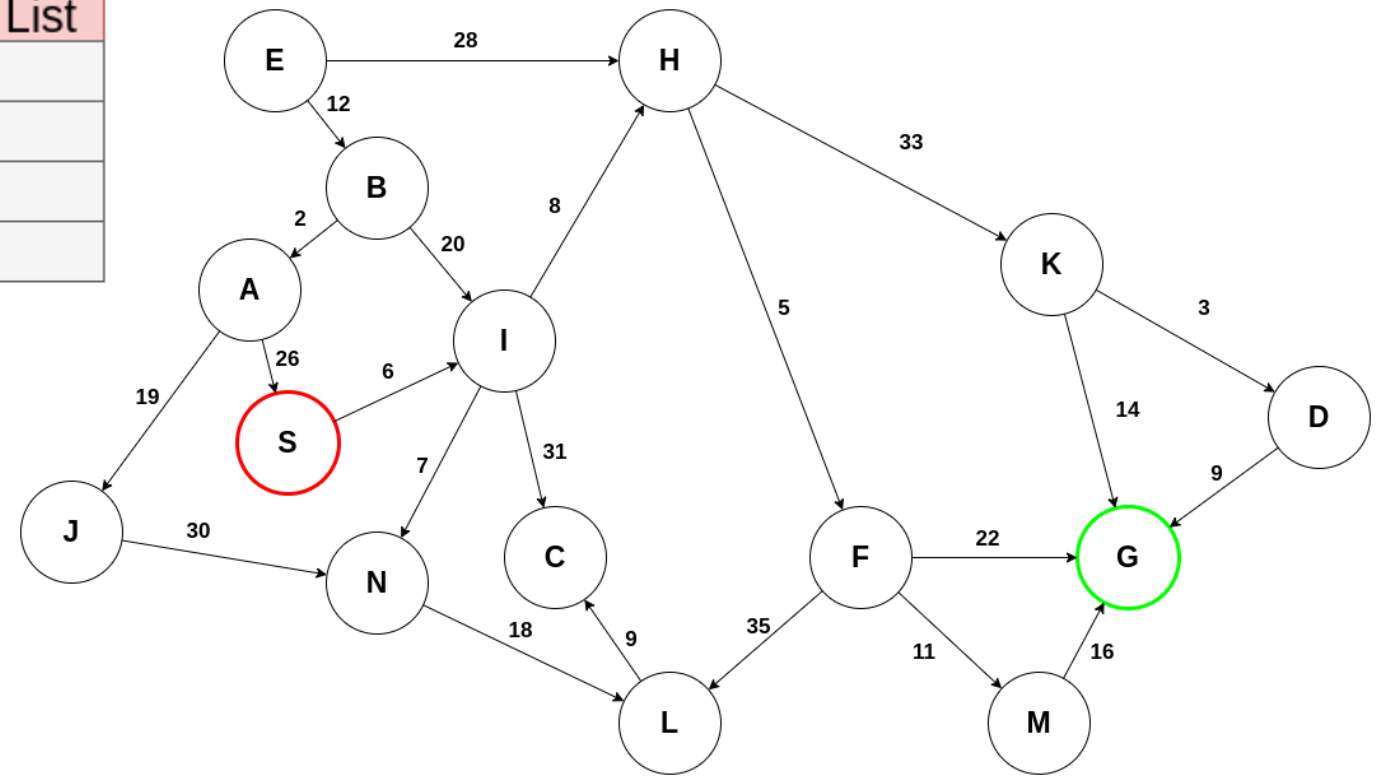
Closed List
S
I

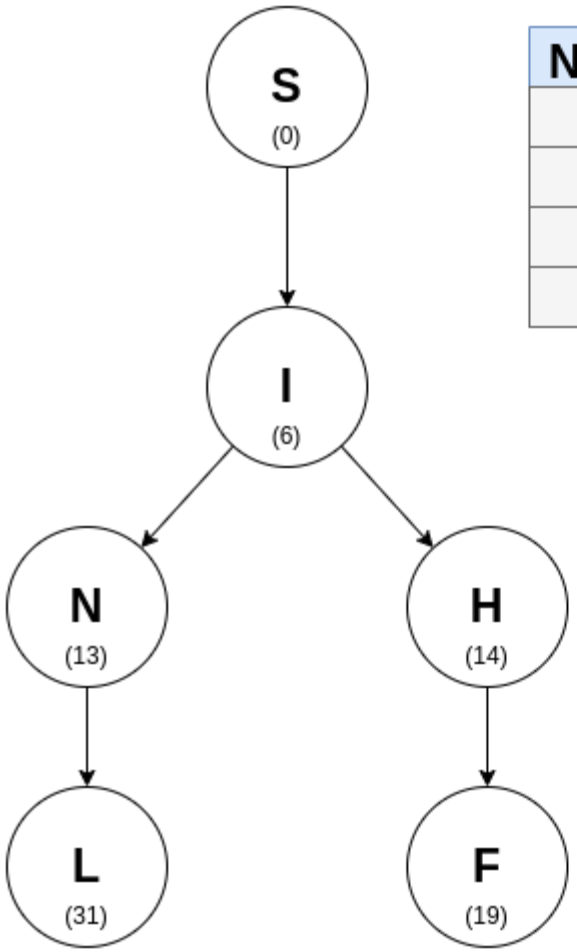




Node[cost]
F[19]
L[31]
K[47]

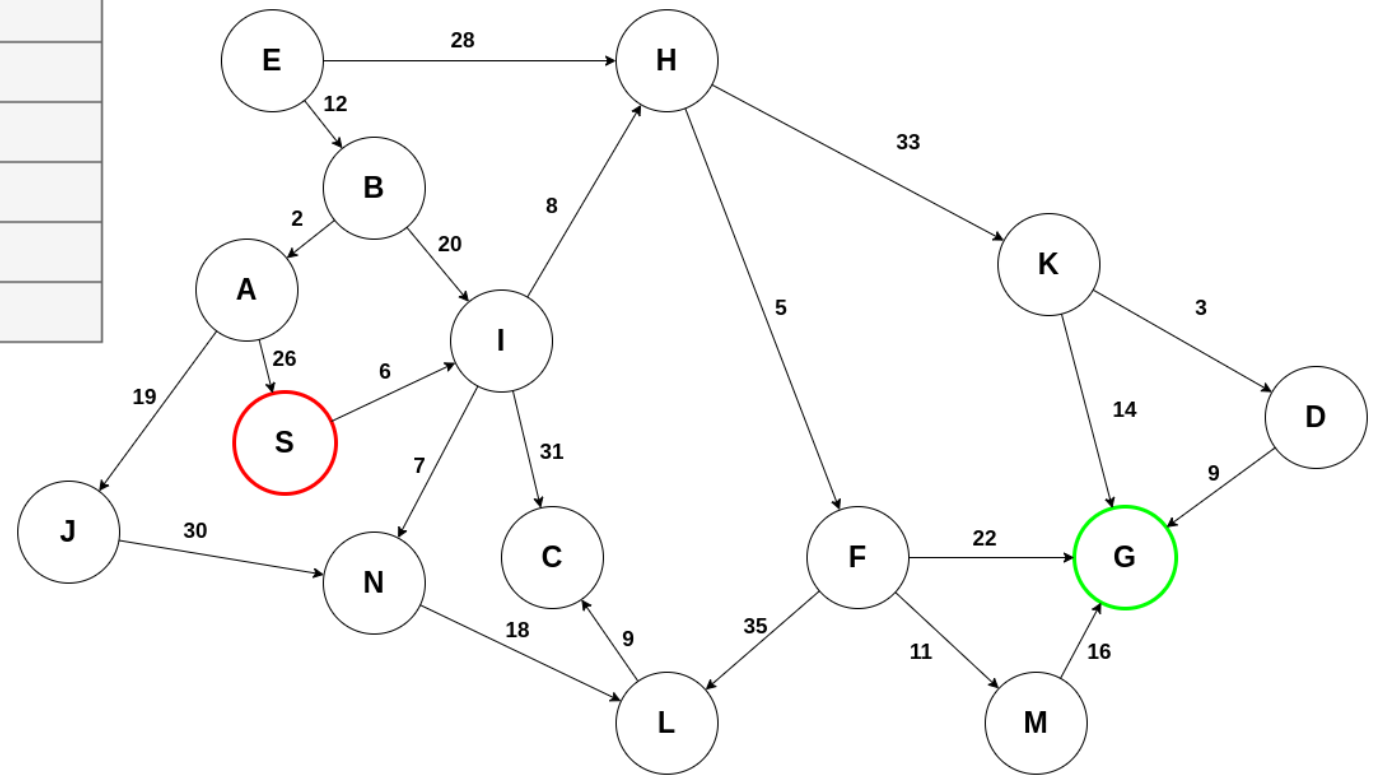
Closed List
S
I
N
H

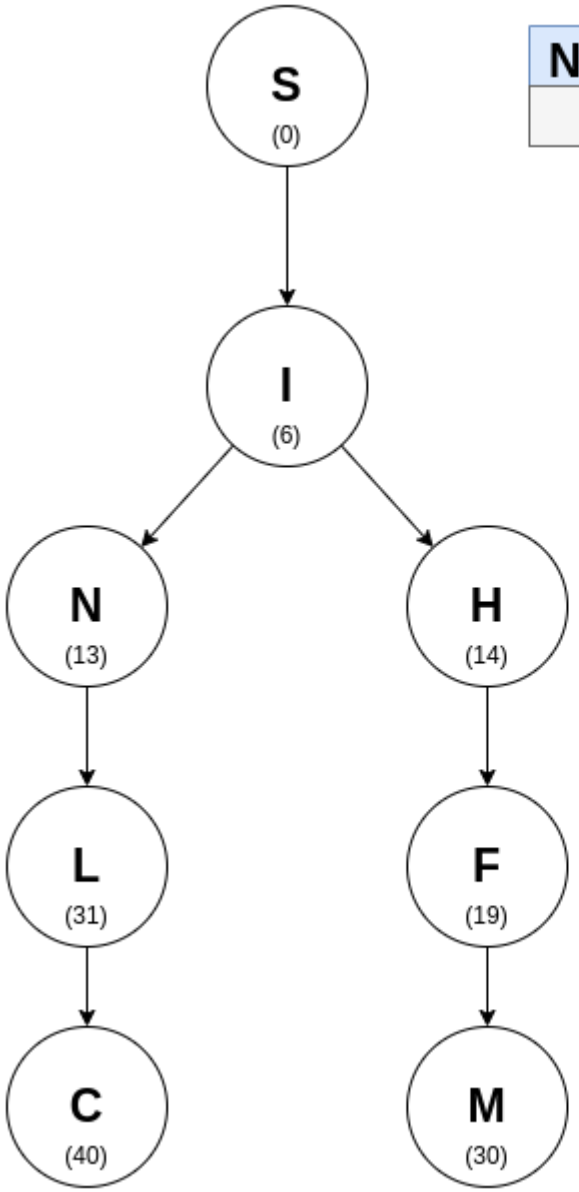




Node[cost]
M[30]
C[40]
G[41]
L[54]

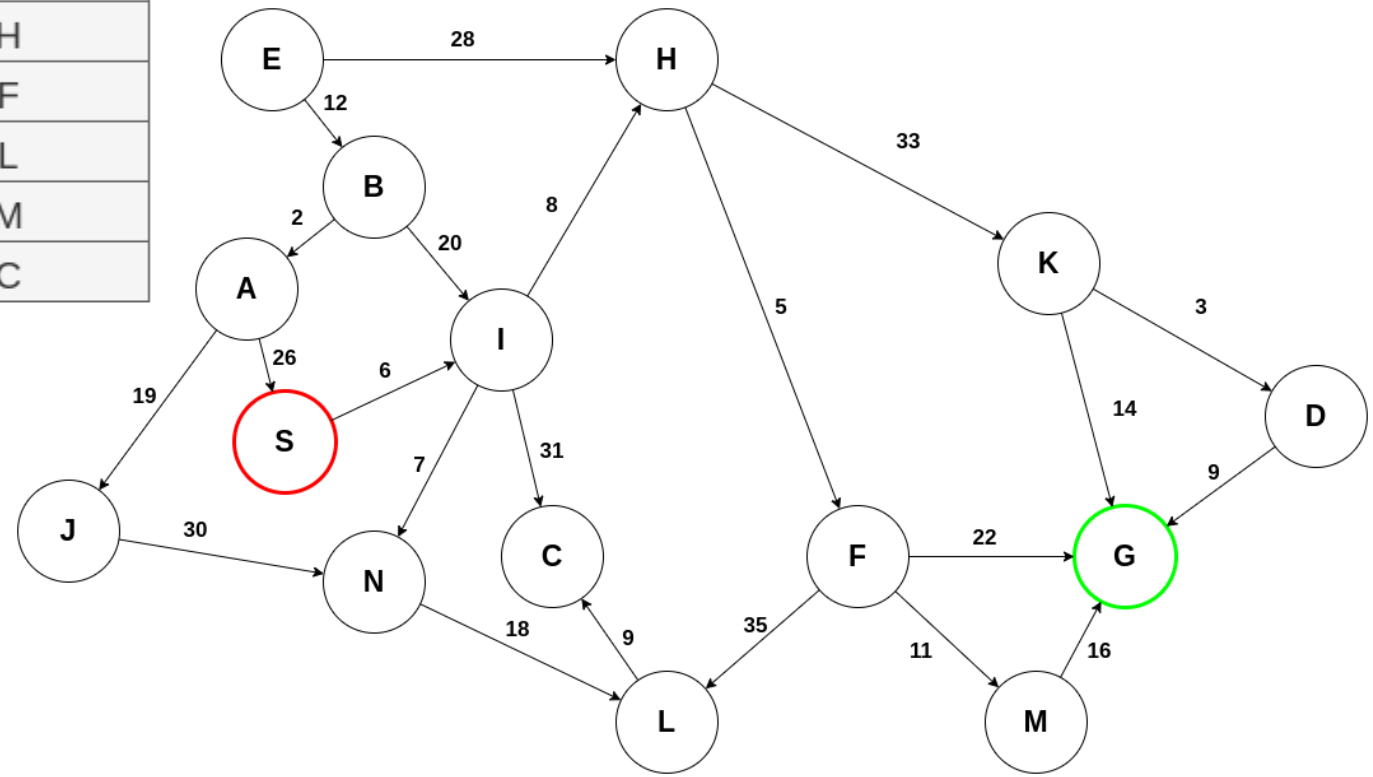
Closed List
S
I
N
H
F
L

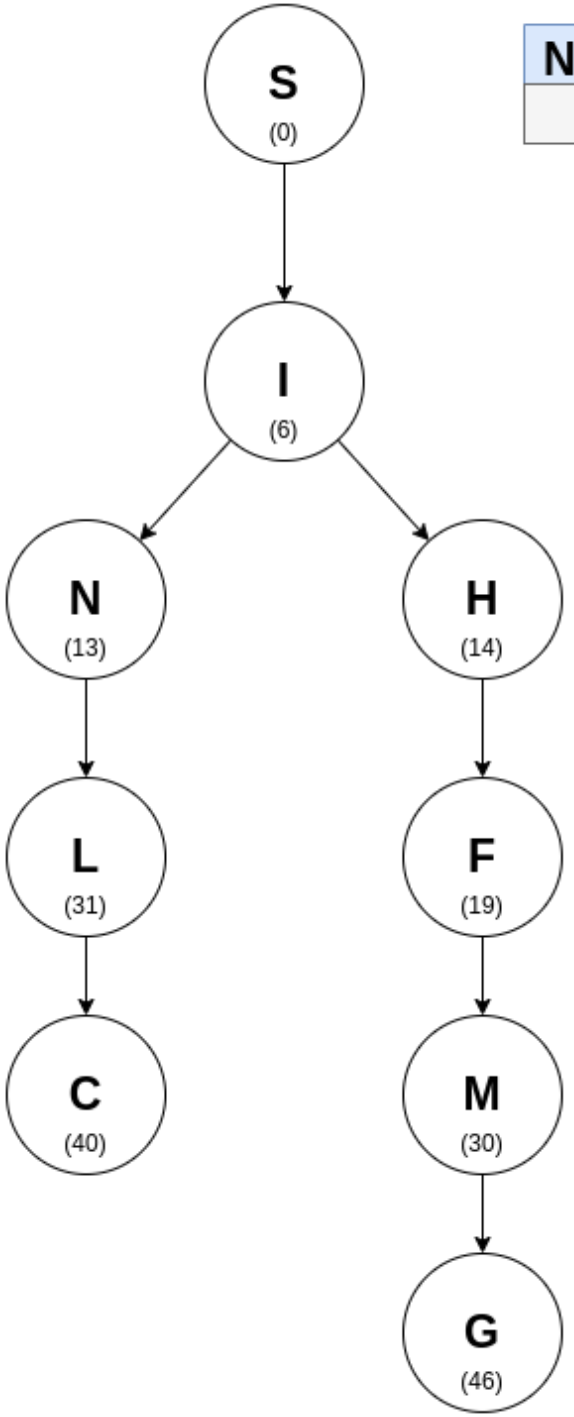




Node[cost]
G[46]

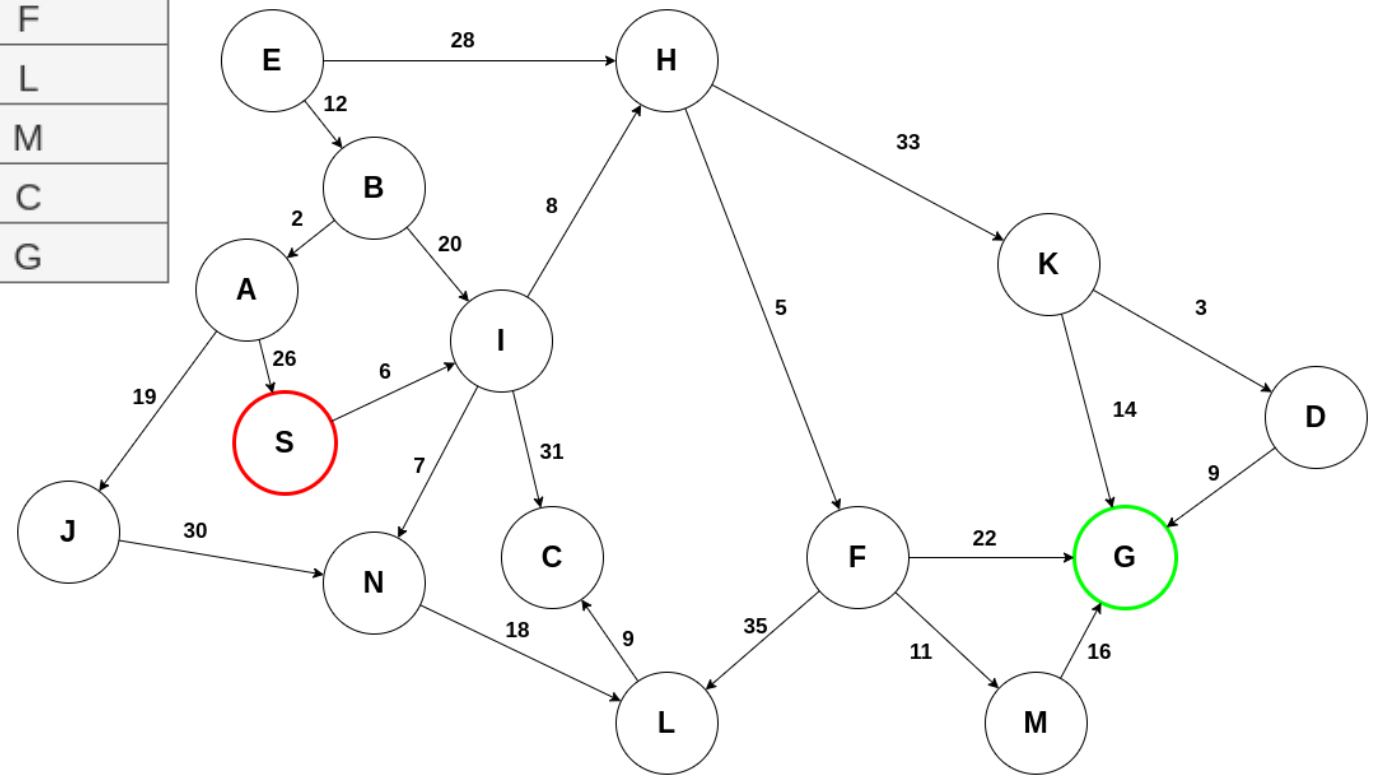
Closed List
S
I
N
H
F
L
M
C

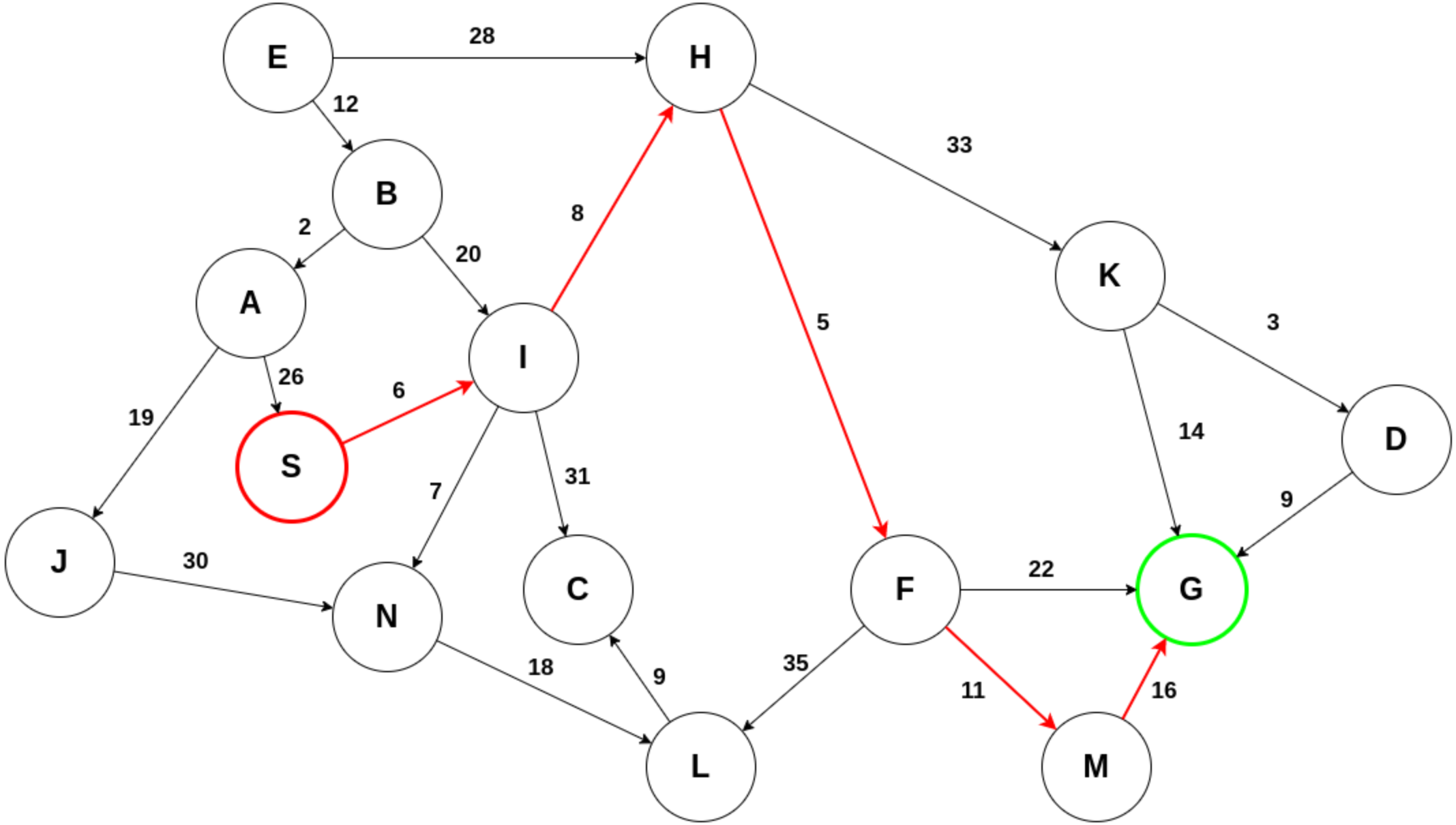




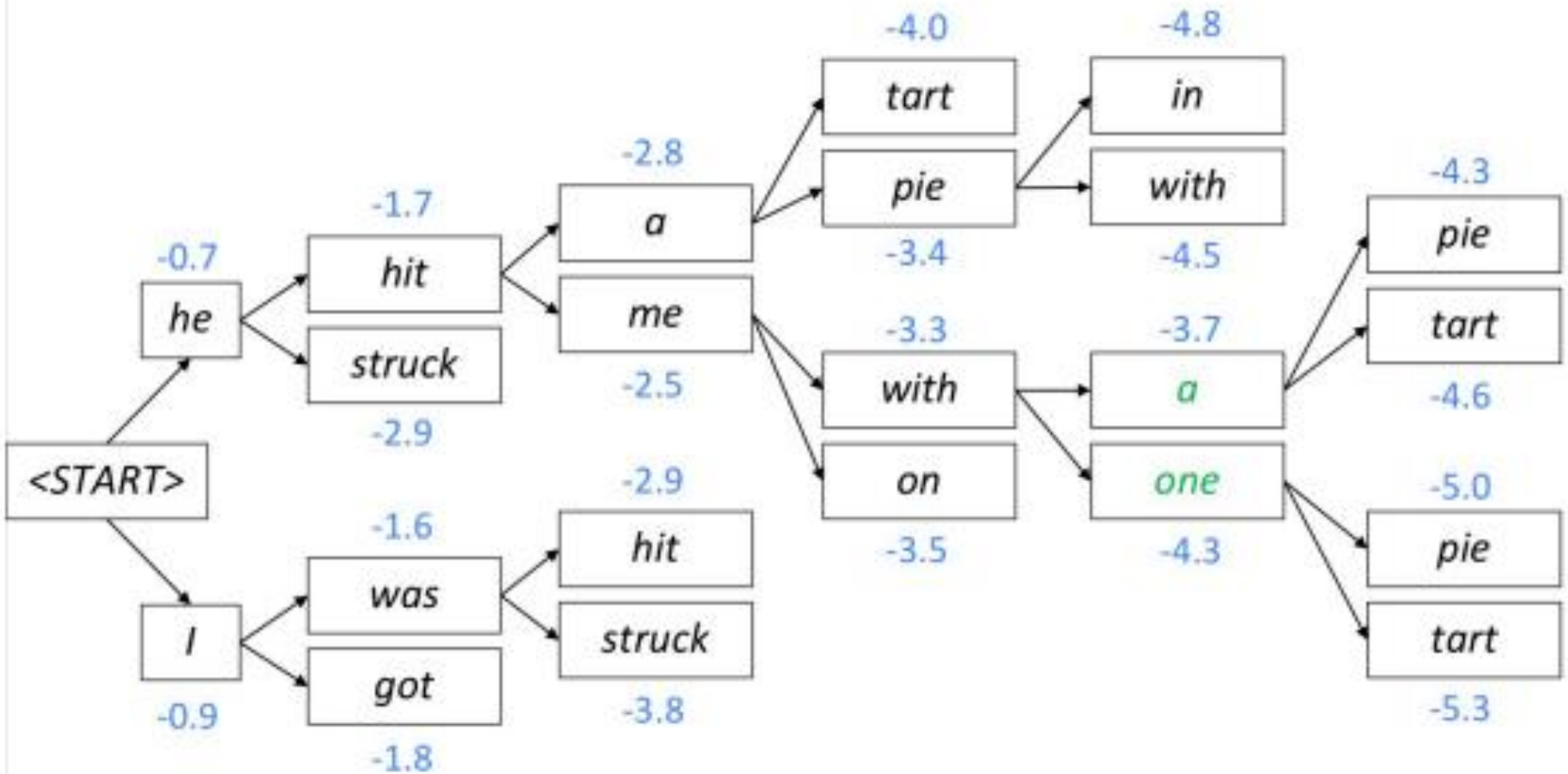
Node[cost]

Closed List
S
I
N
H
F
L
M
C
G



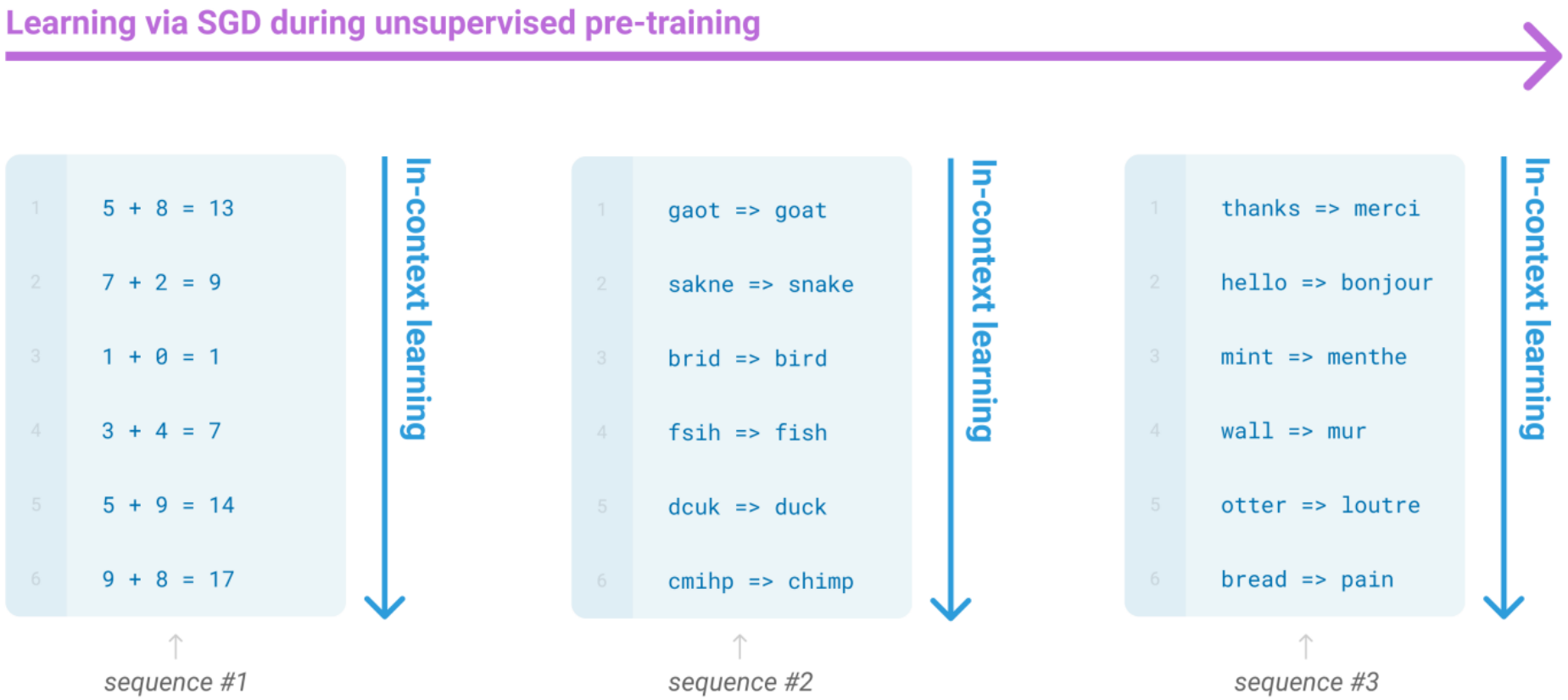

E
H
B
A
I
K
S
D
J
N
C
F
G
L
M
28
12
33
2
20
8
5
3
26
6
19
14
31
7
9
30
22
18
9
35
11
16

Beam size = k = 2. Blue numbers = $\text{score}(y_1, \ldots, y_t) = \sum_{i=1}^{t} \log P_{\text{LM}}(y_i | y_1, \ldots, y_{i-1}, x)$



“Emergent” few-shot learning

- Specify a task by simply prepending examples of the task before your example.
- Also called in-context learning, to stress that no gradient updates are performed when learning a new task.



Side Note: On “Emergence”

The arrogance of the particle physicist and his intensive research may be behind us (the discoverer of the positron said "the rest is chemistry"), but we have yet to recover from that of some molecular biologists, who seem determined to try to reduce everything about the human organism to "only" chemistry, from the common cold and all mental disease to the religious instinct. Surely there are more levels of organization between human ethology and DNA than there are between DNA and quantum electrodynamics, and each level can require a whole new conceptual structure.

Anderson (1972).

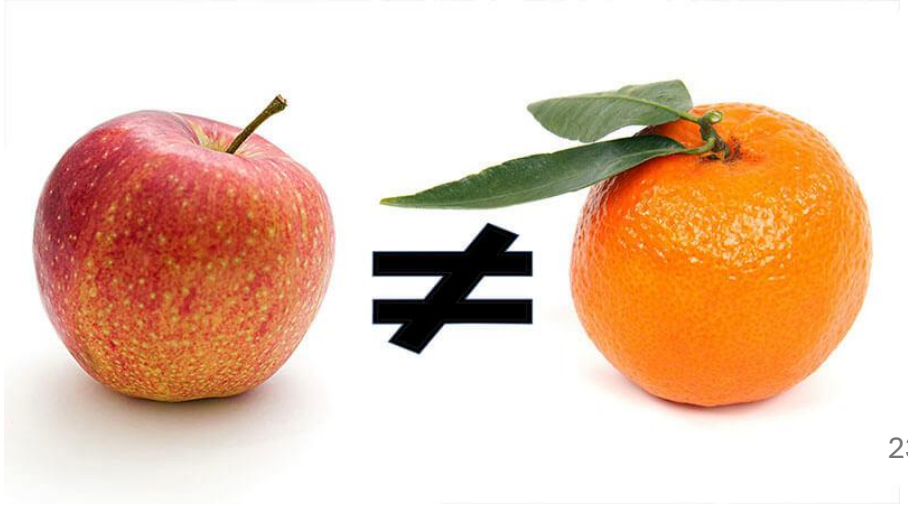
Future ML Systems Will Be Qualitatively Different

JAN 11, 2022 • 7 MIN READ

In 1972, the Nobel prize-winning physicist Philip Anderson wrote the essay "More Is Different". In it, he argues that quantitative changes can lead to qualitatively different and unexpected phenomena. While he focused on physics, one can find many examples of More is Different in other domains as well, including biology, economics, and computer science. Some examples of More is Different include:

- **DNA.** Given only small molecules such as calcium, you can't meaningfully encode useful information; given larger molecules such as DNA, you can encode a genome.

Steinhardt (2022).



Steinhardt and Wei:

"Emergence is when ***quantitative changes*** *in a system result in* ***qualitative changes*** *in behavior."*

Dialectics[55]

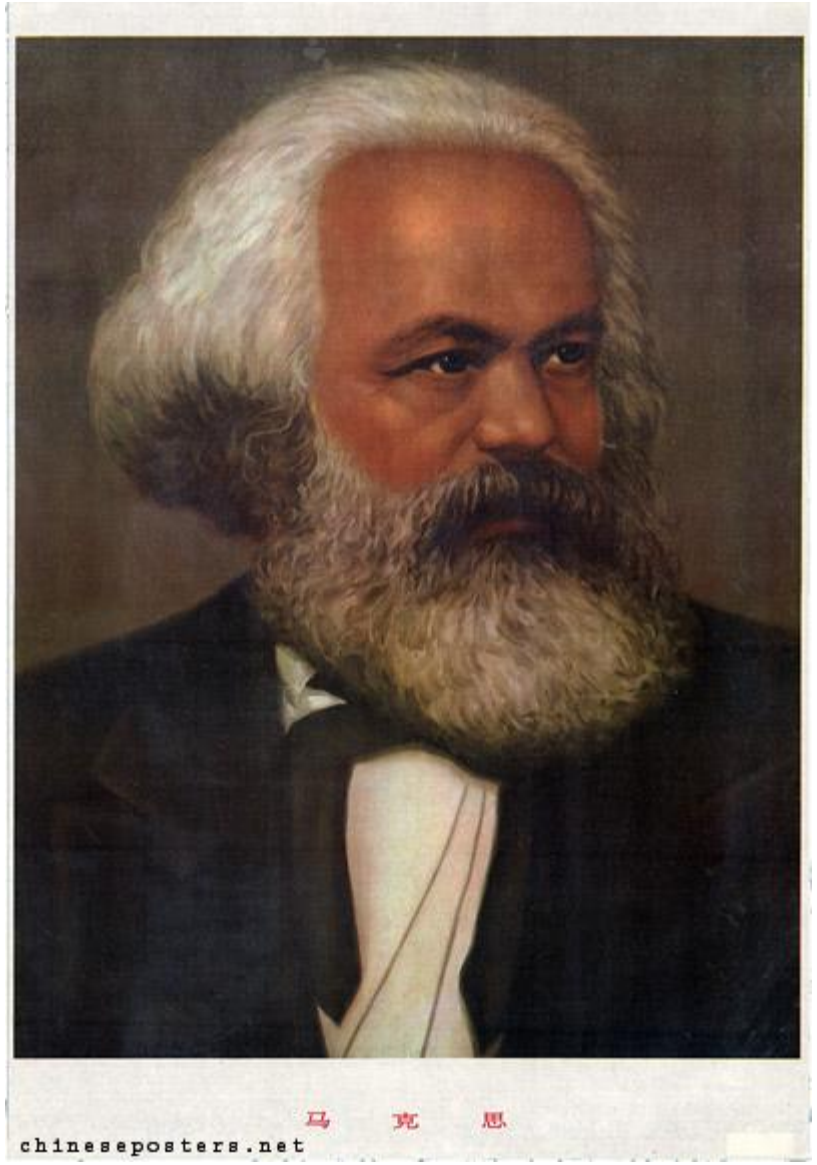
(The general nature of dialectics to be developed as the science of inter-connections, in contrast to metaphysics.)

It is, therefore, from the history of nature and human society that the laws of dialectics are abstracted. For they are nothing but the most general laws of these two aspects of historical development, as well as of thought itself. And indeed they can be reduced in the main to three:

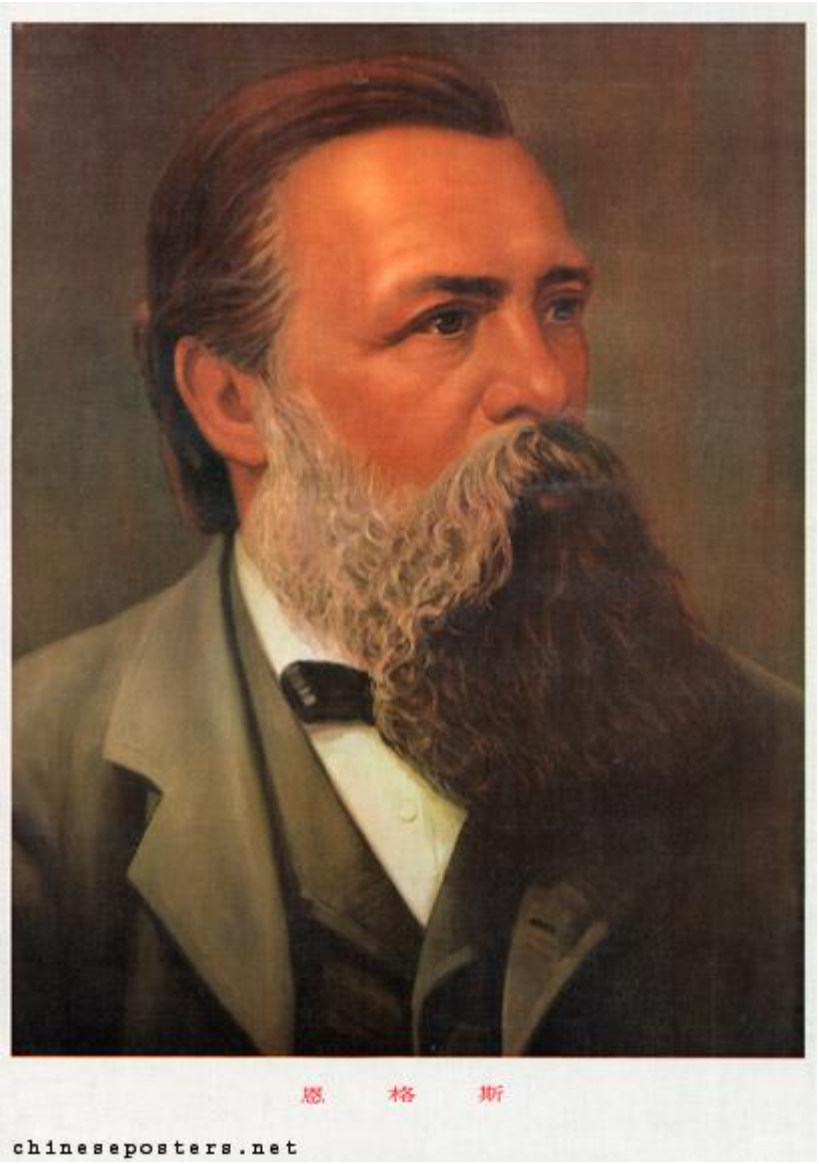
The law of the transformation of quantity into quality and vice versa;

The law of the interpenetration of opposites;

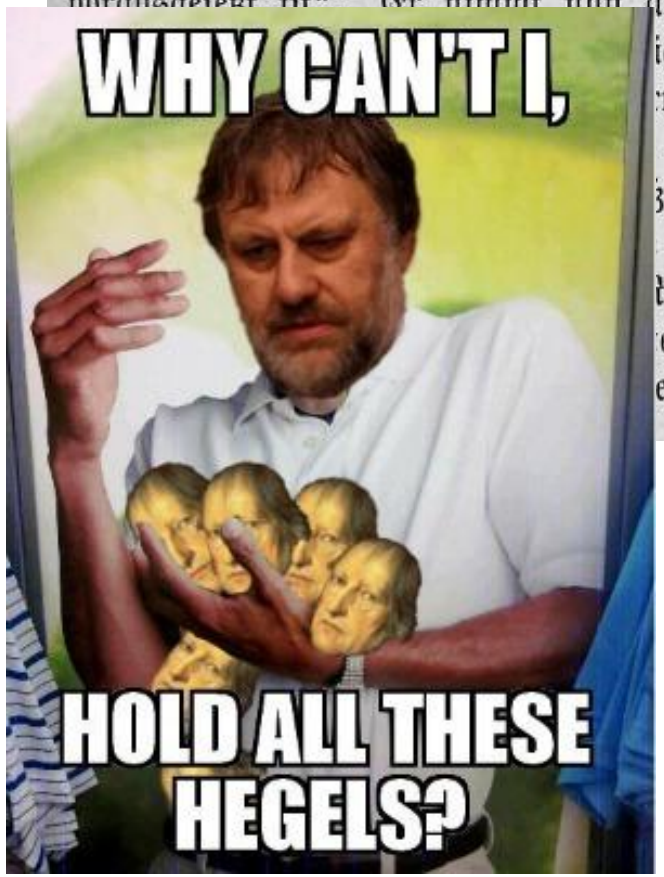
The law of the negation of the negation.


马 克 思
chineseposters.net


Full Communism.


恩 格 斯
chineseposters.net

Das nimmt sich allerdings in dieser von Herrn Dühring „gesäuberten“ Darstellung kurios genug aus. Sehn wir also zu, wie es sich im Original, bei Marx, ausnimmt. Auf Seite 313 (2. Auflage des „Kapital“) zieht Marx aus der vorhergegangnen Untersuchung über konstantes und variables Kapital und Mehrwerth den Schluß, daß „nicht jede beliebige Geld- oder Werthsumme in Kapital verwandelbar, zu dieser Verwandlung vielmehr ein bestimmtes Minimum von Geld oder Tauschwerth in der Hand des einzelnen Geld- oder Waarenbesitzers vorausgesetzt ist“. Er nimmt nun als Beispiel an, daß in irgend



Arbeiter. Und da die kapitalistische Produktion nicht den bloßen Lebensunterhalt, sondern die Vermehrung des Reichthums zum Zweck hat, so wäre unser Mann mit seinen beiden Arbeitern immer noch kein Kapitalist. Damit er nun doppelt so gut lebe wie ein gewöhnlicher Arbeiter und die Hälfte des produzirten Mehrwerths in Kapital zurückverwandle, müßte er acht Arbeiter beschäftigen können, also schon das Vierfache der oben angenommnen Werthsumme besitzen. Und erst nach diesem, und inmitten noch weiterer Ausführungen zur Beleuchtung und Begründung der Thatsache, daß nicht jede beliebige kleine Werthsumme hinreicht, um sich in Kapital zu verwandeln, sondern daß dafür jede Entwicklungsperiode und jeder Industriezweig ihre bestimmten Minimalgrenzen haben, bemerkt Marx: „Hier, wie in der Naturwissenschaft, bewährt sich die Richtigkeit des von Hegel in seiner Logik entdeckten Gesetzes, daß blos quantitative Veränderungen auf einem gewissen Punkt in qualitative Unterschiede umschlagen.

citing

Here, as in natural science, is shown the correctness of the law discovered by Hegel (in his “Logic”), that merely quantitative differences beyond a certain point pass into qualitative changes.

Capital. 1867. Karl Marx.

what's the difference from emergence

Show thinking

While the concepts of "quantitative change leading to qualitative change" and "emergence" are closely related and often overlap, there are subtle but important distinctions:

Key Differences Summarized:

Feature	Quantitative Change Leading to Qualitative Change	Emergence
Primary Focus	Process of transformation through accumulated measurable changes.	Appearance of novel properties from complex interactions of simpler parts.
Driving Force	Accumulation of quantitative pressures reaching a critical threshold.	Non-linear interactions and relationships between components.
Predictability	Tendency is predictable; exact timing can be hard.	Can be difficult to predict, especially strong emergence (irreducible).
Philosophical Root	Dialectical materialism (Hegel, Marx, Engels).	Complexity theory, systems theory, philosophy of mind.
Emphasizes	The "leap" or discontinuity in transformation.	Novelty and often irreducibility of higher-level properties.

In closing, I offer two examples from economics of what I hope to have said. Marx said that quantitative differences become qualitative ones, but a dialogue in Paris in the 1920's sums it up even more clearly:

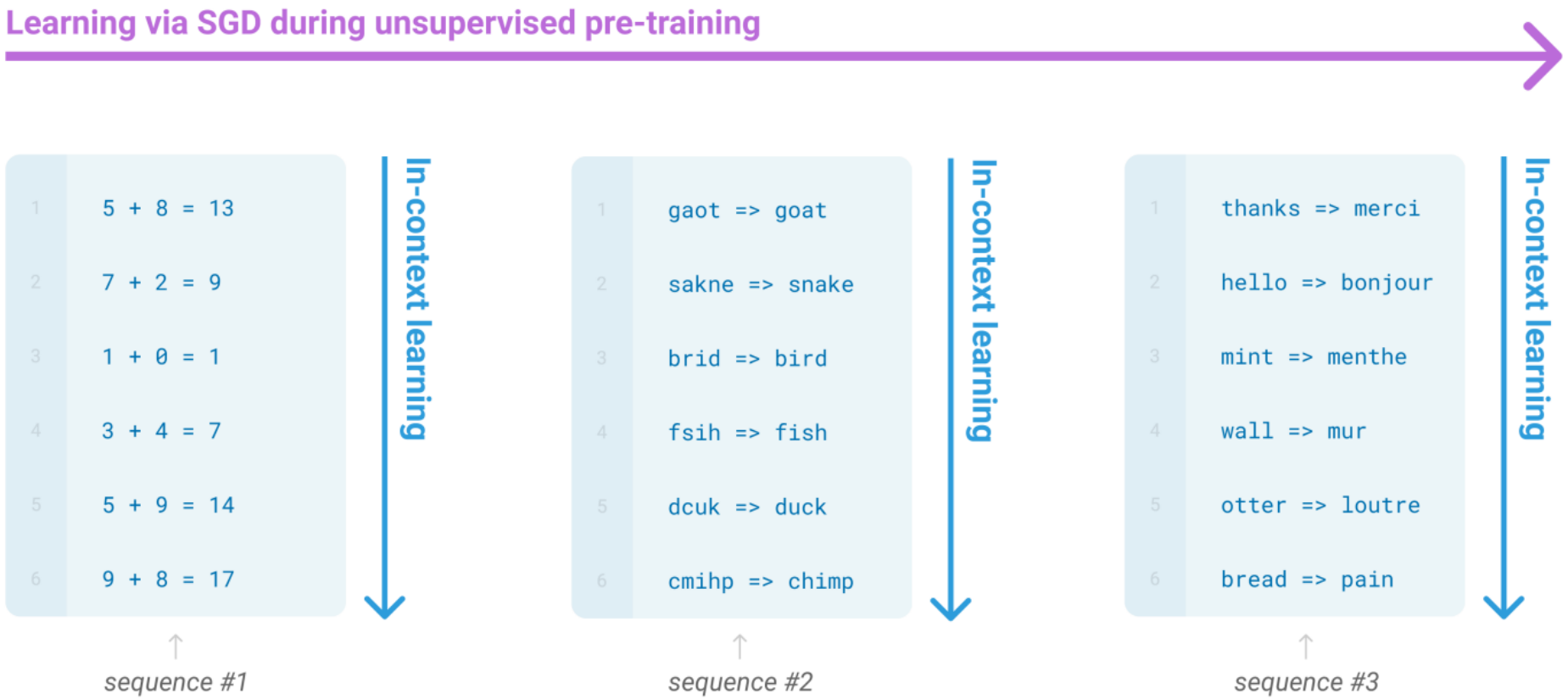
Anderson. *More Is Different*. 1972. Science.

Recommended Sources

- Batterman. The Devil in the Details: Asymptotic Reasoning in Explanation, Reduction, and Emergence.
- The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/

“Emergent” few-shot learning

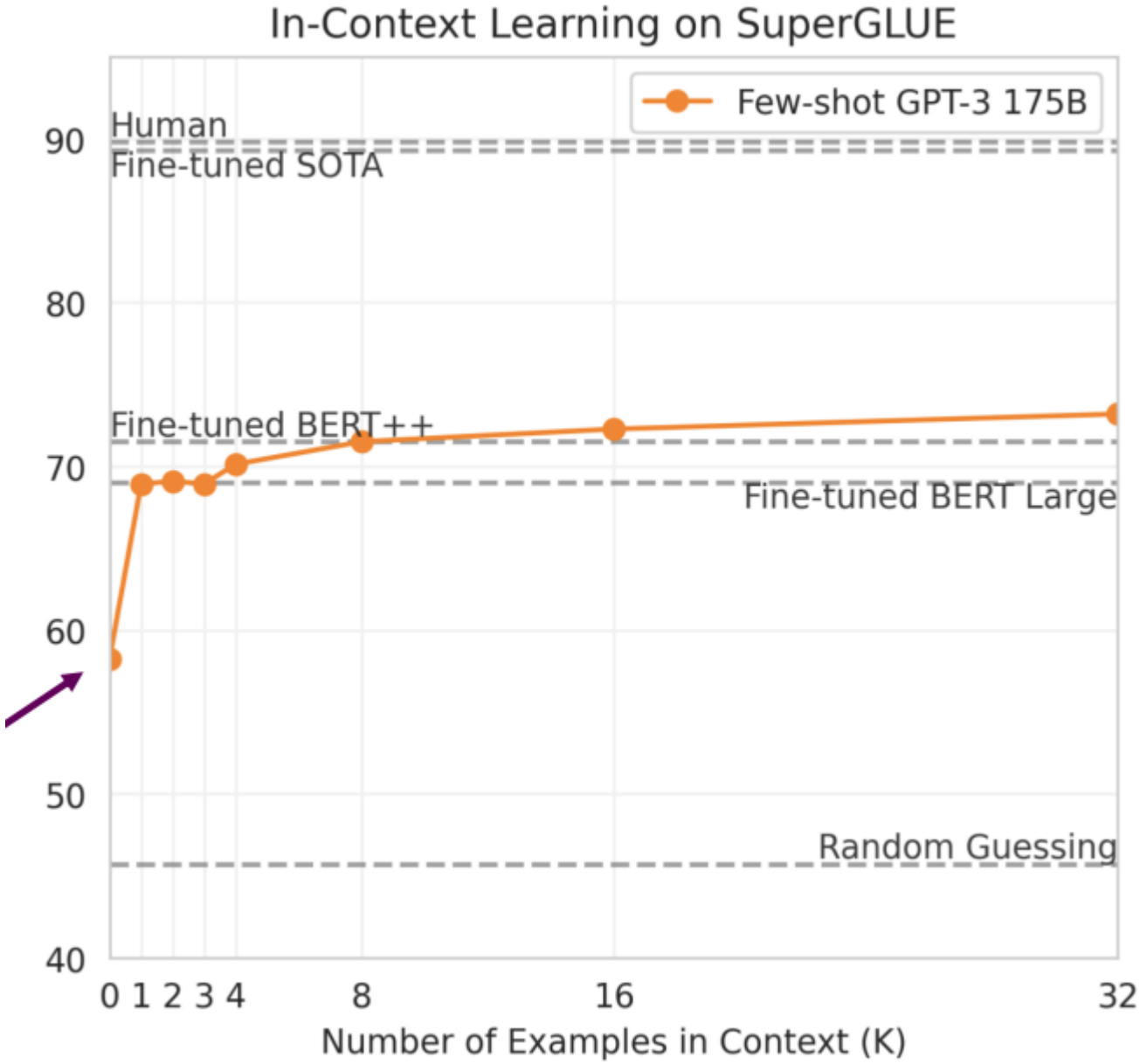
- Specify a task by simply prepending examples of the task before your example.
- Also called in-context learning, to stress that no gradient updates are performed when learning a new task.



Few-shot learning

Zero-shot

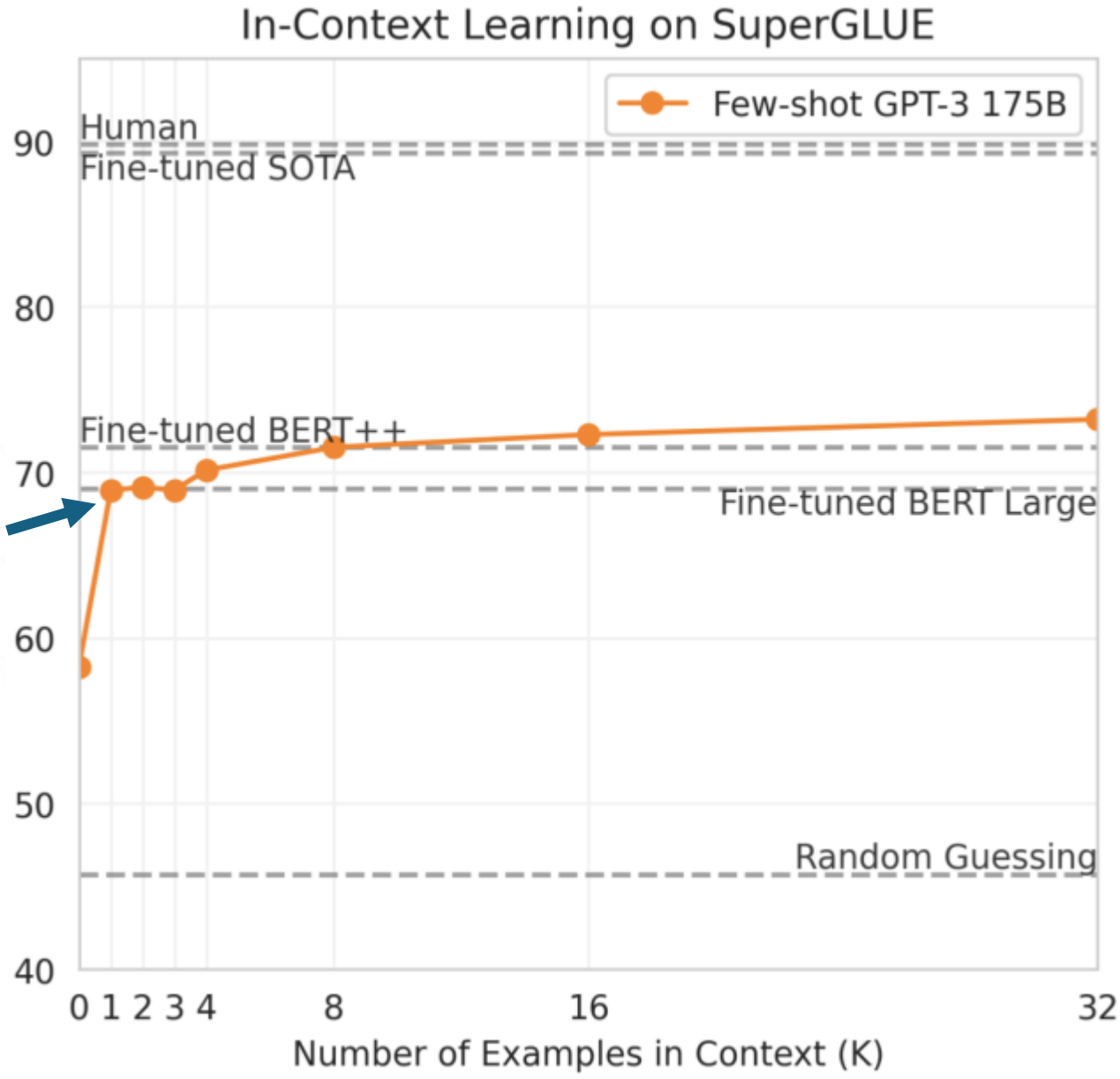
```
1  Translate English to French:
2  cheese =>
```



Few-shot learning

One-shot

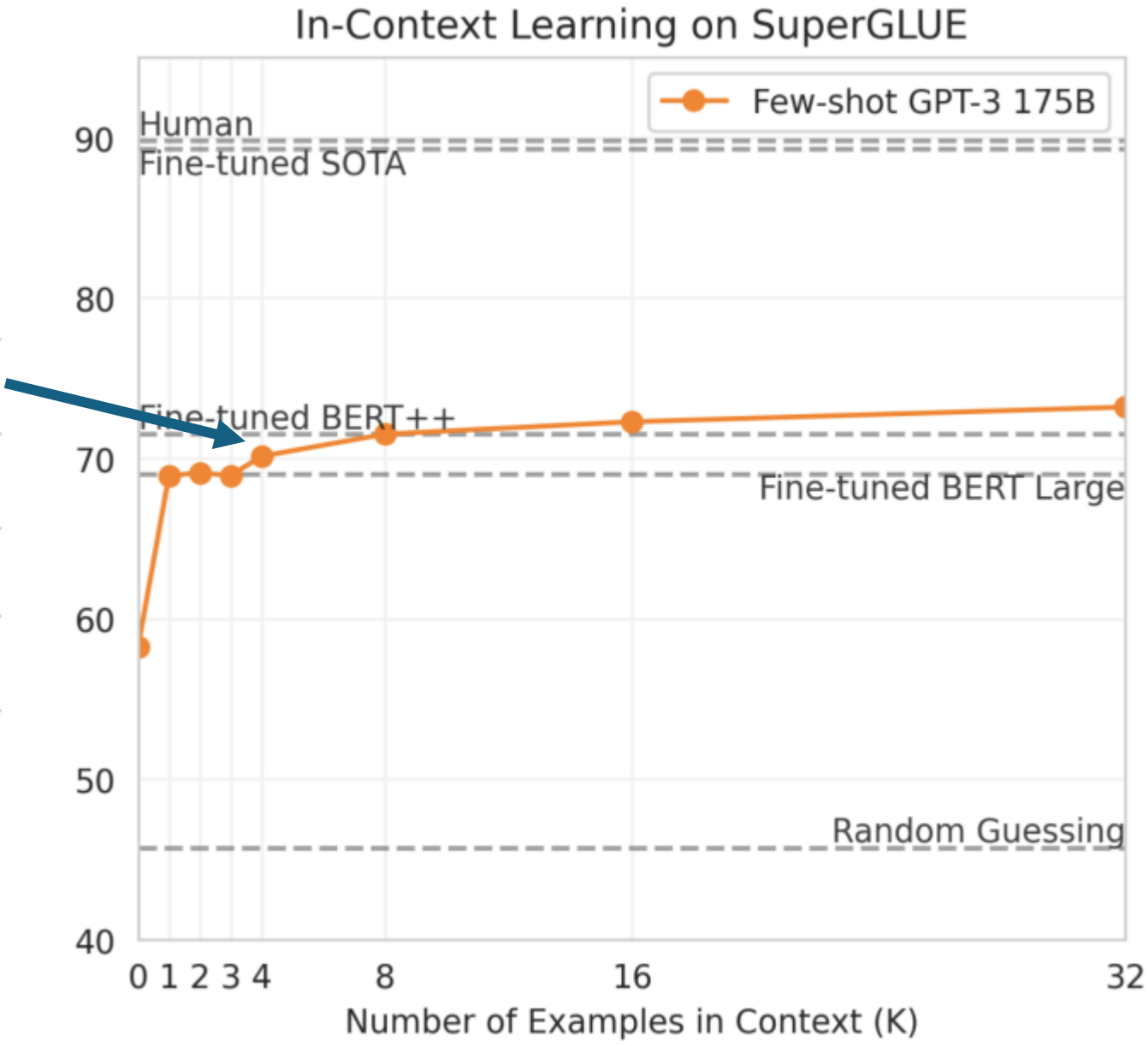
```
1  Translate English to French:
2  sea otter => loutre de mer
3  cheese =>
```

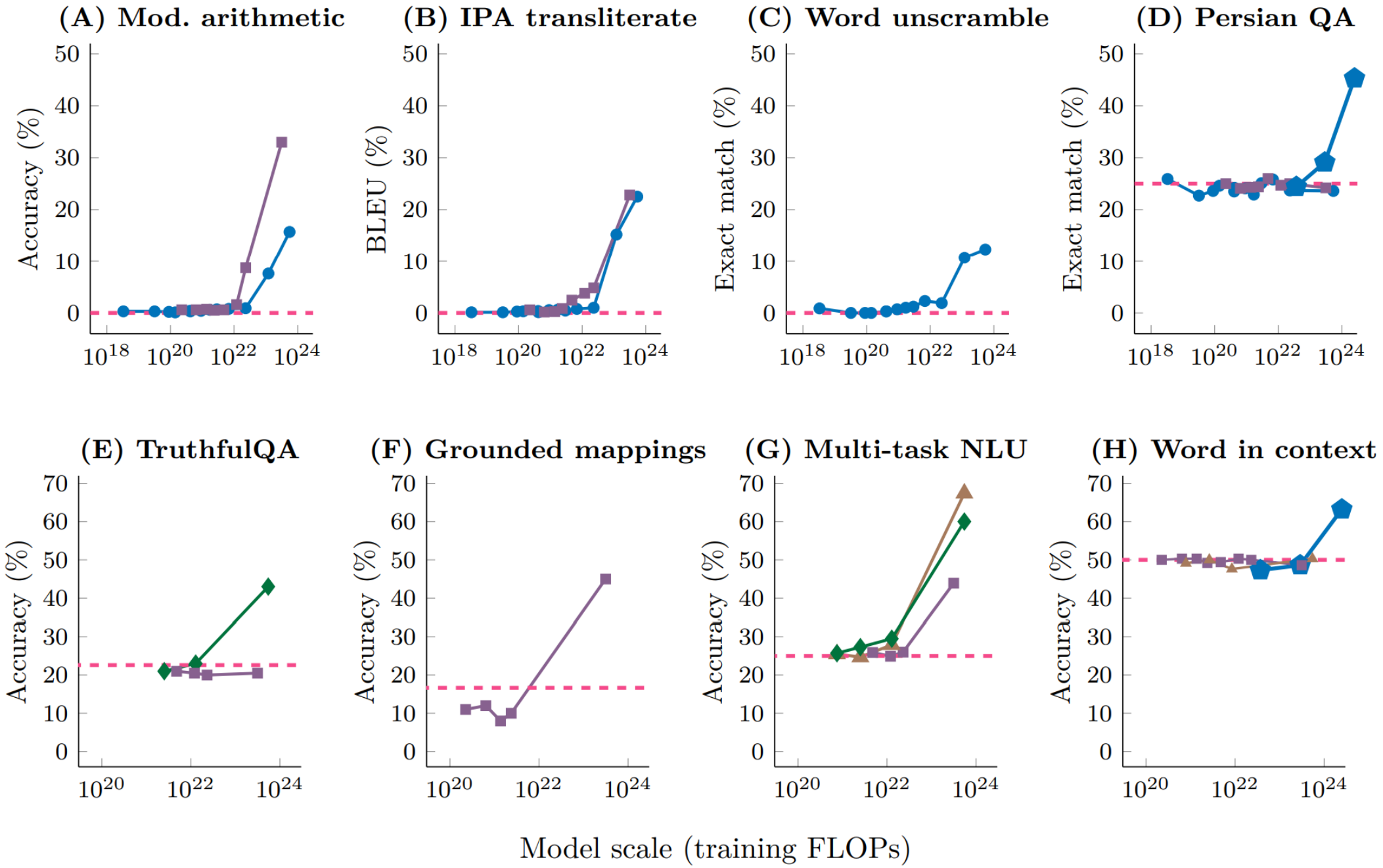


Few-shot learning

Few-shot

```
Translate English to French:
sea otter => loutre de mer
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>
```



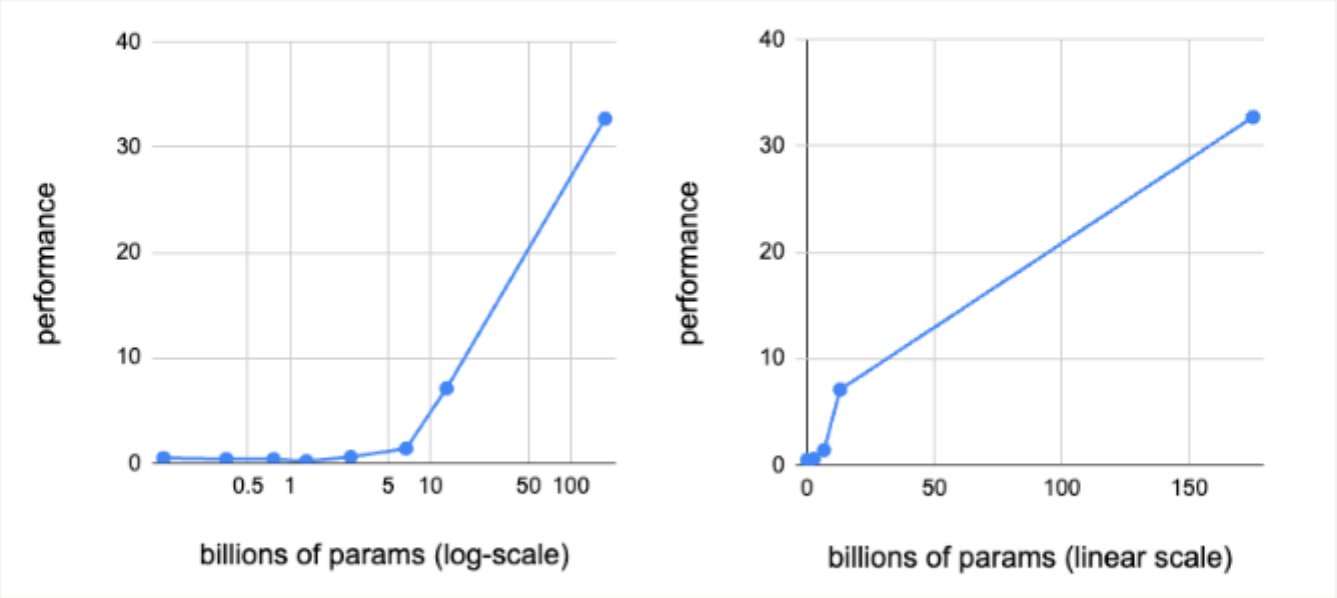


Wei et al. (2022) Emergent Abilities of Large Language Models

Emergence is an artifact of the scaling curve plot

Argument [1] [2]**:** Scaling plots for emergence use an log-scaled x-axis, and if you were to use a linear x-axis scale, the shape of the plot would be smooth.

Response: It's still possible to view emergence on a linear x-axis scale. I plotted Figure 2A from our emergence paper below, and you'll still see the same emergent spike from 7B to 13B (albeit in a less readable way).



In addition to evidence that emergence is still viewable on a linear scale, it's justified to use a log-scale x-axis by default, since models we train are larger in an exponential fashion. For example, the PaLM model sizes are 8B → 62B → 540B (factor of 8x), and LaMDA model sizes go up by 2x. So a log-scale is appropriate for conveying how we scale models in practice (and this has been done in the literature for many years).

https://www.jasonwei.net/blog/common-arguments-regarding-emergent-abilities

Limits of prompting for harder tasks?

- Some tasks seem too hard for even large LMs to learn through prompting alone.
- Especially tasks involving richer, multi-step reasoning.

```
19583 + 29534 = 49117
98394 + 49384 = 147778
29382 + 12347 = 41729
93847 + 39299 = ?
```

Improvement: change the prompt!

Chain-of-thought

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ✗

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

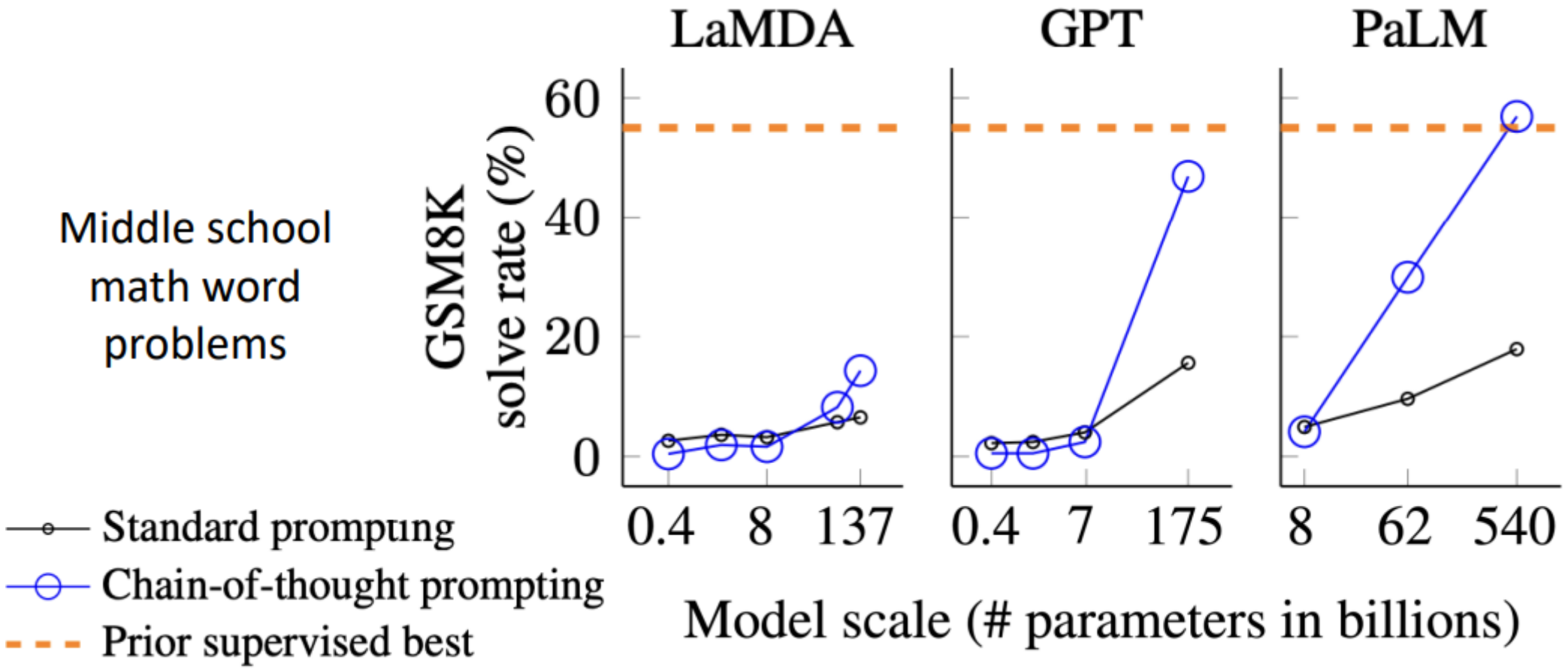
Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

[Wei et al., 2022; also see Nye et al., 2021]

Chain-of-thought



[Wei et al., 2022; also see Nye et al., 2021]

Chain-of-thought

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

Do we even need examples of reasoning?

Can we just ask the model to reason through things?

[Wei et al., 2022; also see Nye et al., 2021]

Zero-shot CoT prompting

	MultiArith	GSM8K
Zero-Shot	**17.7**	**10.4**
Few-Shot (2 samples)	33.7	15.6
Few-Shot (8 samples)	33.8	15.6
Zero-Shot-CoT	**78.7**	**40.7**
Few-Shot-CoT (2 samples)	84.8	41.3
Few-Shot-CoT (4 samples : First) (*1)	89.2	-
Few-Shot-CoT (4 samples : Second) (*1)	90.5	-
Few-Shot-CoT (8 samples)	93.0	48.7
Zero-Plus-Few-Shot-CoT (8 samples) (*2)	**92.8**	**51.5**

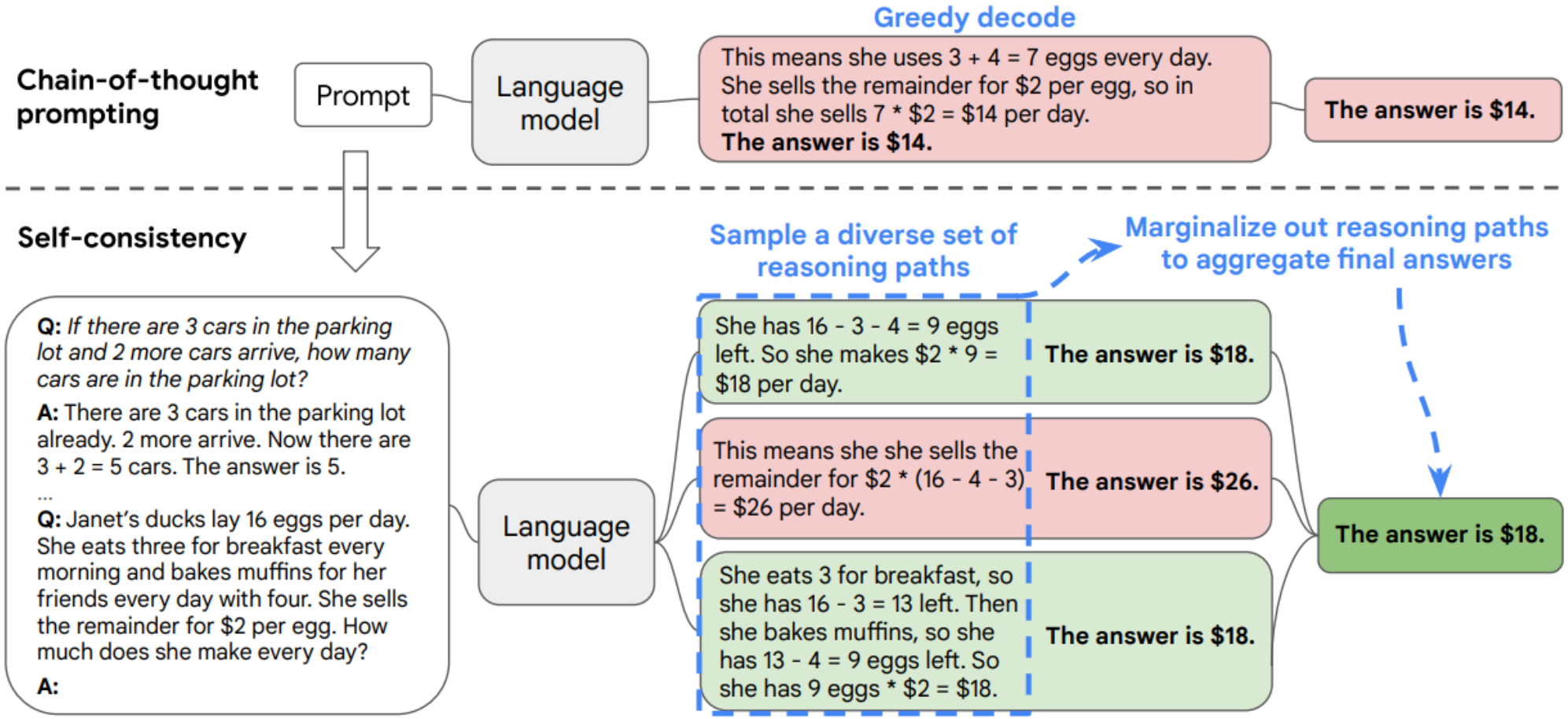
Greatly outperforms zero-shot!

Manual CoT still better

Kojima et al. (2022) Large Language Models are Zero-Shot Reasoners

CoT with “Self-consistency”

- Replace greedy decoding with an ensemble of samples…
- Main idea: correct reasoning processes have greater agreement than incorrect processes.



Wang et al. (2023) Self-Consistency Improves Chain of Thought Reasoning in Language Models.

CoT with “Self-consistency”

	GSM8K	MultiArith	AQuA	SVAMP	CSQA	ARC-c
Greedy decode	56.5	94.7	35.8	79.0	79.0	85.2
Weighted avg (unnormalized)	56.3 ± 0.0	90.5 ± 0.0	35.8 ± 0.0	73.0 ± 0.0	74.8 ± 0.0	82.3 ± 0.0
Weighted avg (normalized)	22.1 ± 0.0	59.7 ± 0.0	15.7 ± 0.0	40.5 ± 0.0	52.1 ± 0.0	51.7 ± 0.0
Weighted sum (unnormalized)	59.9 ± 0.0	92.2 ± 0.0	38.2 ± 0.0	76.2 ± 0.0	76.2 ± 0.0	83.5 ± 0.0
Weighted sum (normalized)	74.1 ± 0.0	99.3 ± 0.0	48.0 ± 0.0	86.8 ± 0.0	80.7 ± 0.0	88.7 ± 0.0
Unweighted sum (majority vote)	74.4 ± 0.1	99.3 ± 0.0	48.3 ± 0.5	86.6 ± 0.1	80.7 ± 0.1	88.7 ± 0.1

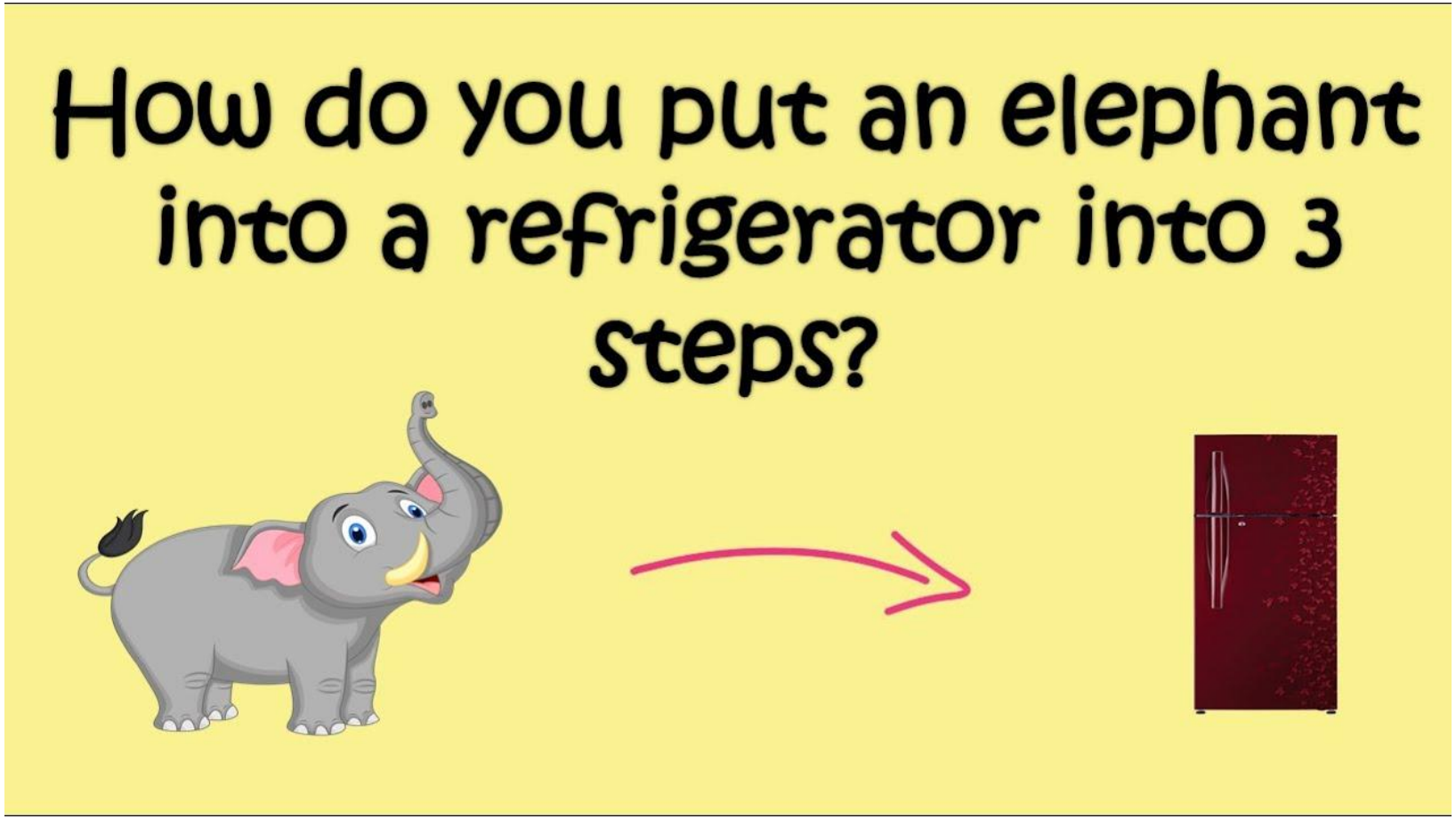
Table 1: Accuracy comparison of different answer aggregation strategies on PaLM-540B.

Out-performs regular CoT on a variety of benchmarks

	GSM8K	MultiArith	SVAMP	ARC-e	ARC-c
CoT (Wei et al., 2022)	17.1	51.8	38.9	75.3	55.1
Ensemble (3 sets of prompts)	18.6 ± 0.5	57.1 ± 0.7	42.1 ± 0.6	76.6 ± 0.1	57.0 ± 0.2
Ensemble (40 prompt permutations)	19.2 ± 0.1	60.9 ± 0.2	42.7 ± 0.1	76.9 ± 0.1	57.0 ± 0.1
Self-Consistency (40 sampled paths)	**27.7** ± **0.2**	**75.7** ± **0.3**	**53.3** ± **0.2**	**79.3** ± **0.3**	**59.8** ± **0.2**

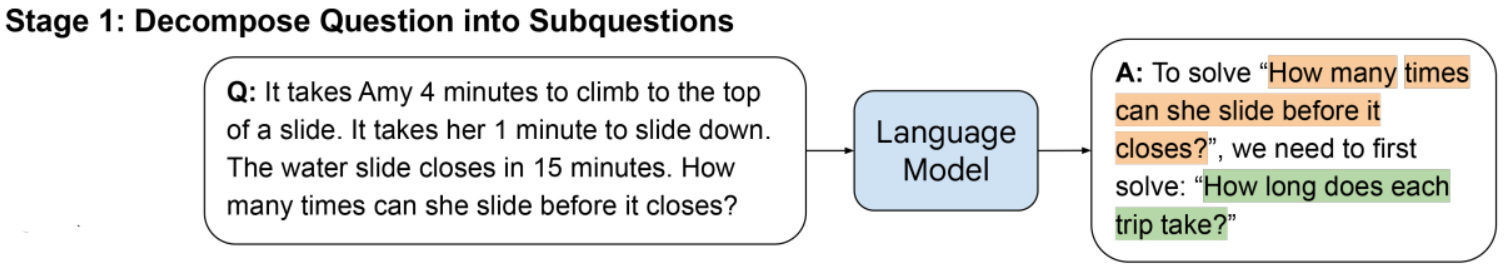
Self-consistency is doing more than simple ensembling

Least-to-Most prompting



Least-to-Most prompting

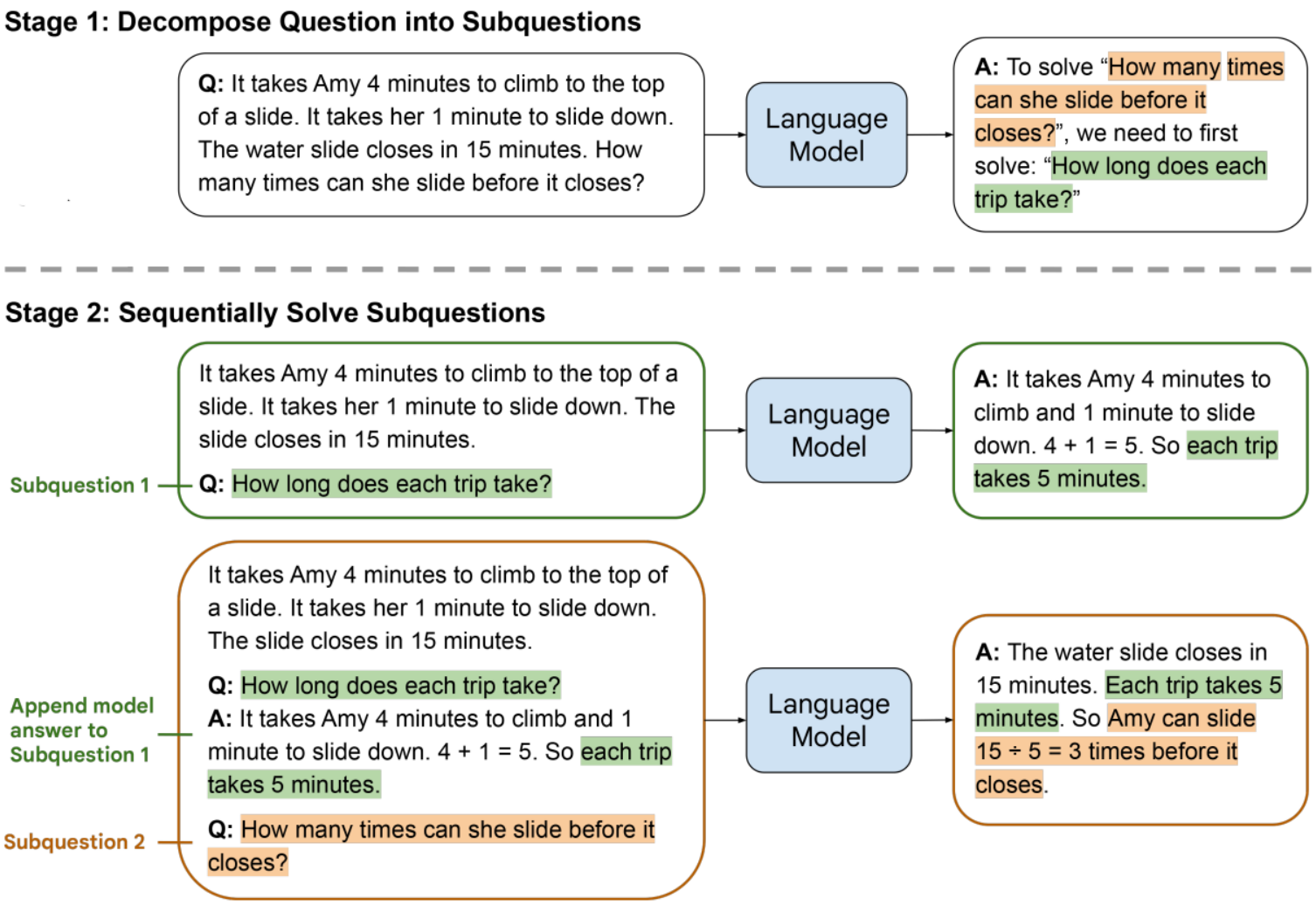
- **Idea**: decompose a problem into smaller ones.



Zhou et al. (2023) Least-to-Most Prompting Enables Complex Reasoning in Large Language Models

Least-to-Most prompting

- **Idea**: decompose a problem into smaller ones.



Least-to-Most prompting for Math Reasoning

Q: Elsa has 5 apples. Anna has 2 more apples than Elsa. How many apples do they have together?
A: Anna has 2 more apples than Elsa, so Anna has 2 + 5 = 7 apples. Elsa and Anna have 5 + 7 = 12 apples together. The answer is 12.

Standard CoT prompting

Q: Elsa has 5 apples. Anna has 2 more apples than Elsa. How many apples do they have together?
A: Let's break down this problem: 1. How many apples does Anna have? 2. How many apples do Elsa and Anna have together?
1. Anna has 2 more apples than Elsa. So Anna has 2 + 5 = 7 apples.
2. Elsa and Anna have 5 + 7 = 12 apples together.

Q: {question}
A: Let's break down this problem:
—
The answer is:

Least-to-most prompting

Least-to-Most prompting for Math Reasoning

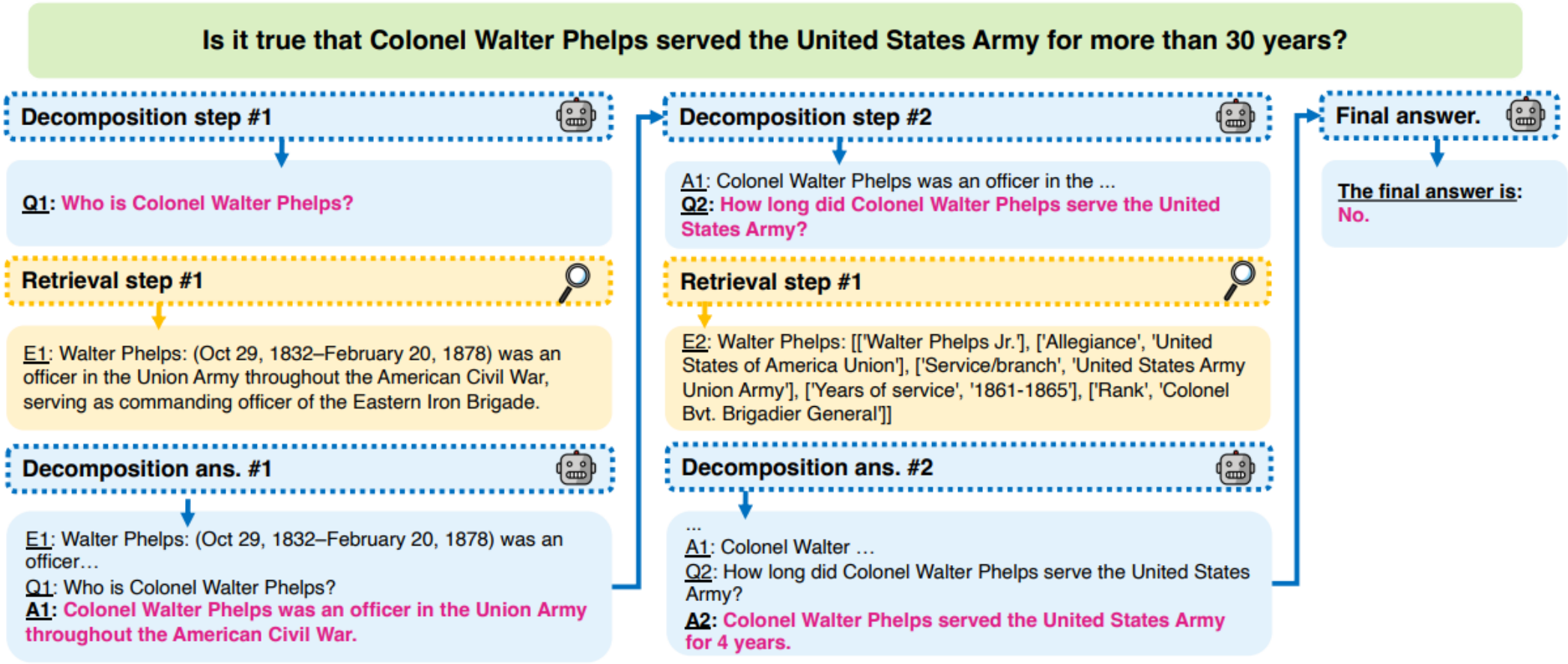
Accuracy by Steps (GSM8K)	All	2 Steps	3 Steps	4 steps	$\geq$ 5 steps
Least-to-Most	**62.39**	74.53	**68.91**	**59.73**	**45.23**
Chain-of-Thought	60.87	**76.68**	67.29	59.39	39.07

Generalizes to more #steps than in-context example!

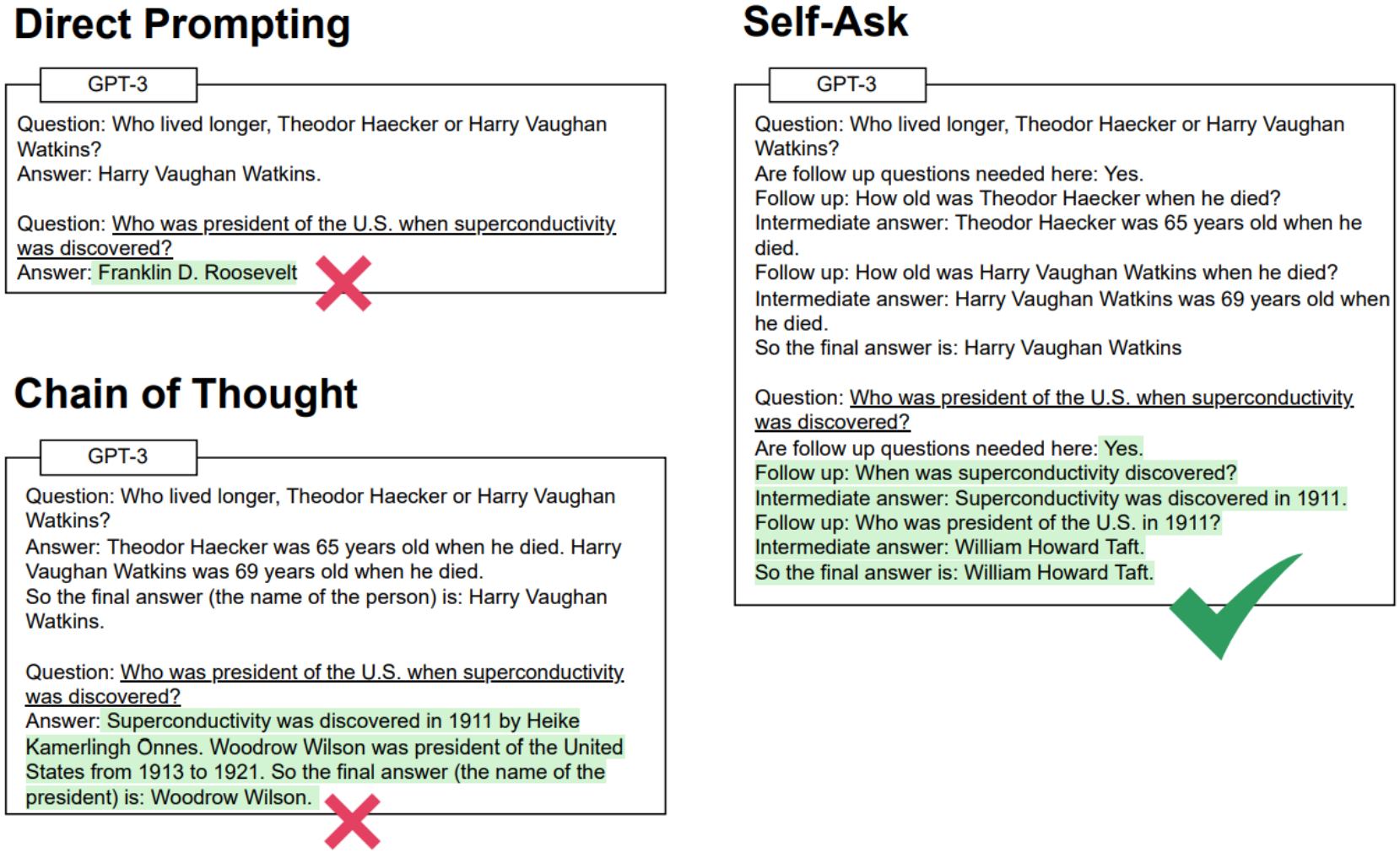
Prompting method	**Accuracy**
Zero-Shot	16.38
Standard prompting	17.06[3]
Chain-of-Thought (original)	61.18
Chain-of-Thought (1-shot)	60.88
Least-to-Most (1-shot)	62.39
Chain-of-Thought (best)	**68.61**[3]
Least-to-Most (best)	68.01

But with enough prompt engineering, CoT ≈ Least-to-Most

Solution #3: Interleaving decomposition



Solution #3: Interleaving decomposition SelfAsk



Press et al. (2023) Measuring and Narrowing the Compositionality Gap in Language Models.

Summary

- Zero-shot Prompting
- In-context Learning
- Chain-of-thought
- Self-consistency
- Least-to-Most