

Module 3

Prompt Engineering

Wei Dong
wdong@aaalgo.com

Prompt Engineering: Wizardry of the 2020s



Prompt CIVITAI GENERATOR TXT2IMG

A female samurai warrior in full battle regalia, positioned heroically against the backdrop of a burning Japanese town. Her intense gaze and poised stance with a katana create a striking focal point. The artwork should convey movement with flickering embers and smoke, while maintaining a rich color palette that accentuates the traditional design elements of her armor and the historical setting.

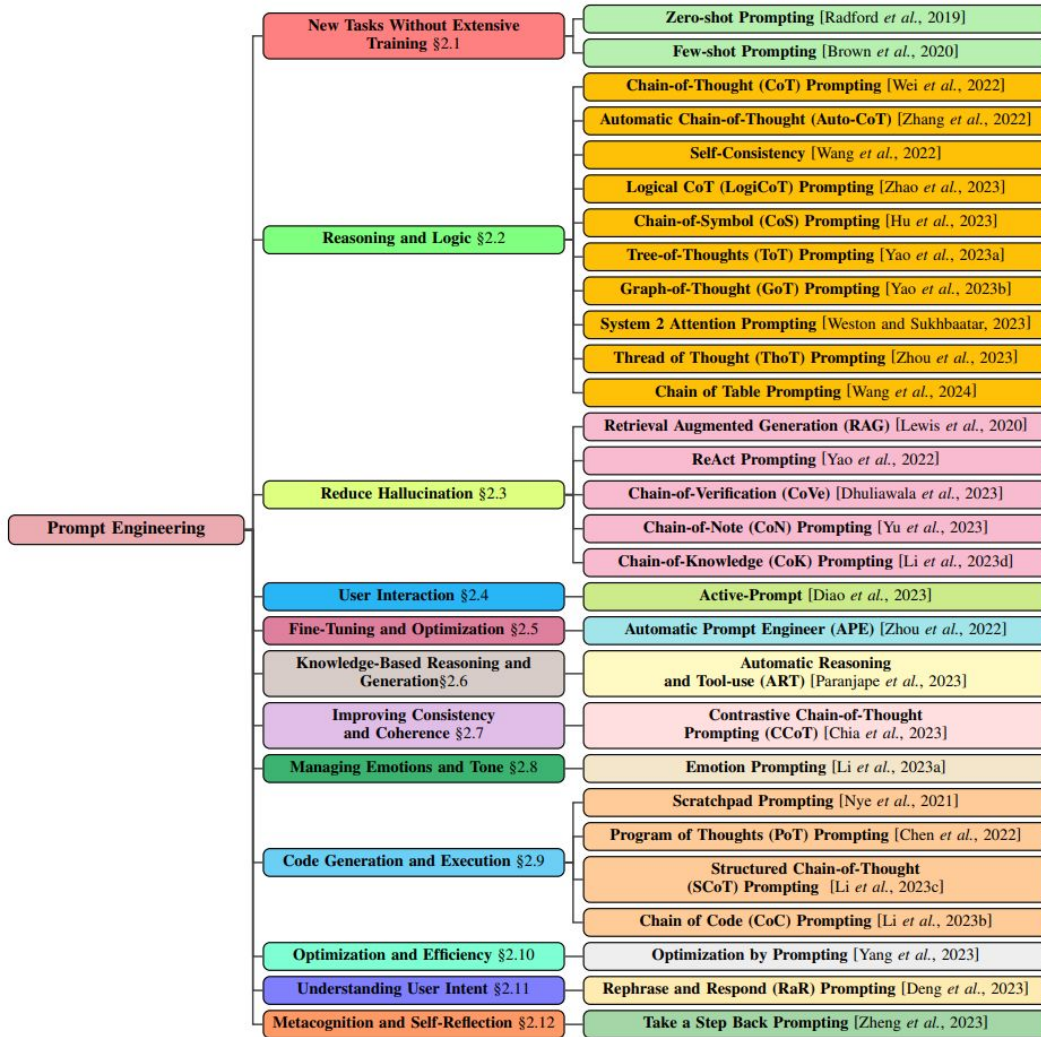
[Show less](#)

Negative prompt

worst quality, low quality, low contrast, blurry, low quality, medium quality, watermark, username, signature, text, bad anatomy, bad hands, text, error, missing fingers, extra digit, fewer digits, cropped, jpeg artifacts, bad feet, extra fingers, mutated hands, poorly drawn hands, bad proportions, extra limbs, disfigured, bad anatomy, gross proportions, malformed limbs, missing arms, missing legs, extra arms, extra legs, mutated hands, fused fingers, too many fingers, long neck, watermark

[Show less](#)

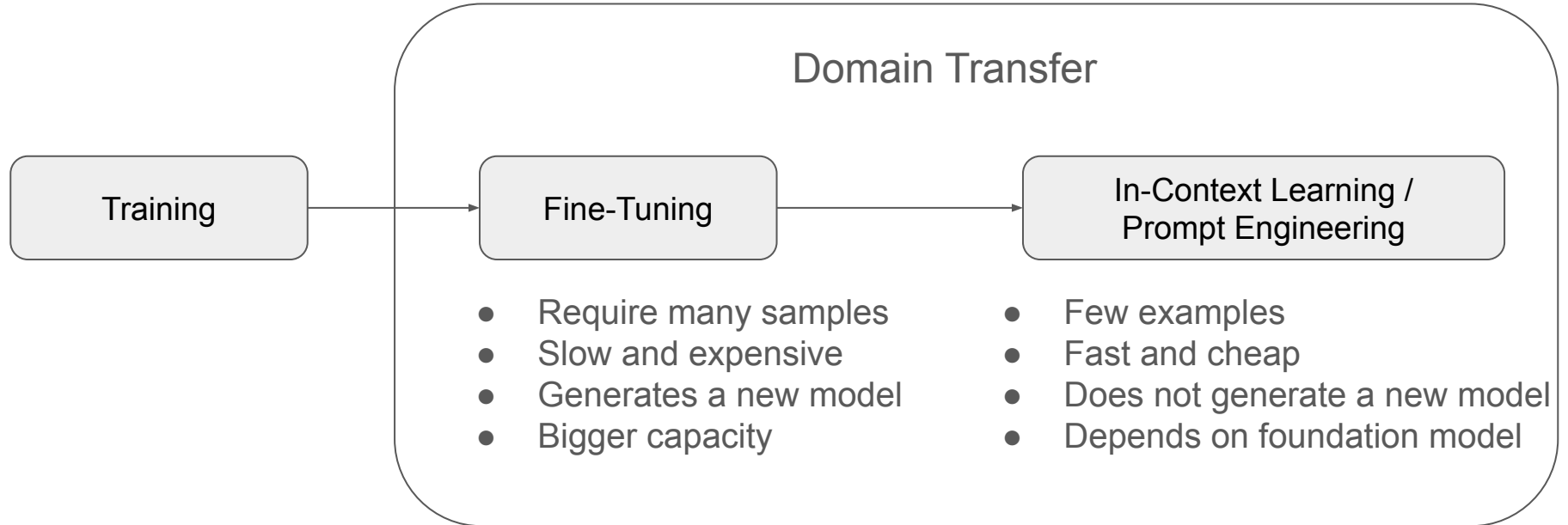
Taxonomy of Prompt Engineering Techniques



[A systematic survey, Sahoo et al., 2024](#)

Few Shot Learning and In-Context Learning

- In the context of prompt engineering,
- ... few shot learning == providing examples in prompt



Few-Shot (In-Context) Learning

Review: "The product arrived late, but the quality is excellent. I'm very happy w

Sentiment: Positive

Review: "The item was okay, nothing special. It works as expected."

Sentiment: Neutral

Review: "I'm disappointed. The product stopped working after one day."

Sentiment: Negative

Review: "The service was fast, and the product exceeded my expectations."

Sentiment:

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. ✅

Chain-of-Thought

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

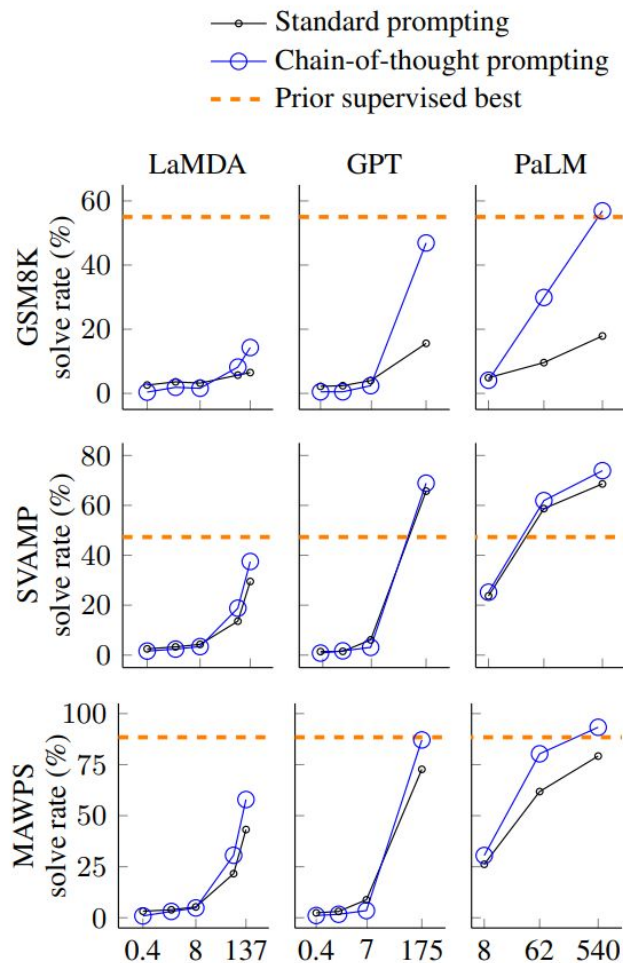
Jason Wei Xuezhi Wang Dale Schuurmans Maarten Bosma
Brian Ichter Fei Xia Ed H. Chi Quoc V. Le Denny Zhou

Google Research, Brain Team
{jasonwei, dennyzhou}@google.com

Abstract

We explore how generating a *chain of thought*—a series of intermediate reasoning steps—significantly improves the ability of large language models to perform complex reasoning. In particular, we show how such reasoning abilities emerge naturally in sufficiently large language models via a simple method called *chain-of-thought prompting*, where a few chain of thought demonstrations are provided as exemplars in prompting.

Experiments on three large language models show that chain-of-thought prompting improves performance on a range of arithmetic, commonsense, and symbolic reasoning tasks. The empirical gains can be striking. For instance, prompting a PaLM 540B with just eight chain-of-thought exemplars achieves state-of-the-art accuracy on the GSM8K benchmark of math word problems, surpassing even finetuned GPT-3 with a verifier.



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. ✅

CoT implemented by One-Shot In-Context Learning

The goal is to change the behavior of LLM.

This is not “standard prompting”. It’s a control setting in evaluation.

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

True Standard Prompting

(1) Hotspot QA	
Question: Aside from the Apple Remote, what other device can control the program Apple Remote was originally designed to interact with?	
(1a) Standard	(1c) Act-Only
Answer: iPod ❌	Act 1: Search[Apple Remote] Obs 1: The Apple Remote is a remote control ...
(1b) CoT (Reason Only)	Act 2: Search[Front Row] Obs 2: Could not find [Front Row]. Similar: ...
Thought: Let's think step by step. Apple Remote was originally designed to interact with Apple TV. Apple TV can be controlled by iPhone, iPad, and iPod Touch. So the answer is iPhone, iPad, and iPod Touch.	Act 3: Search[Front Row (software)] Obs 3: Front Row is a discontinued media center software ...
Answer: iPhone, iPad, iPod Touch ❌	Act 4: Finish[yes] ❌
(1d) ReAct (Reason + Act)	
Thought 1: I need to search Apple Remote and find the program it was originally designed to interact with. Act 1: Search[Apple Remote] Obs 1: The Apple Remote is a remote control introduced in October 2005 by Apple ... originally designed to control the Front Row media center program ...	
Thought 2: Apple Remote was originally designed to control the Front Row media center program. I need to search Front Row next and find what other device can control it. Act 2: Search[Front Row] Obs 2: Could not find [Front Row]. Similar: ['Front Row Seat to Earth', 'Front Row Motorsports', 'Front Row (software)', ...]	
Thought 3: Front Row is not found. I need to search Front Row (software) ... Act 3: Search[Front Row (software)] Obs 3: Front Row is a discontinued media center software ...	
Thought 4: Front Row (software) is controlled by an Apple Remote or the keyboard function keys. So the answer is keyboard function keys. Act 4: Finish[keyboard function keys] ✅	

Zero-Shot-CoT

(a) Few-shot

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: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The answer is 8. ❌

(b) Few-shot-CoT

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: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are $16 / 2 = 8$ golf balls. Half of the golf balls are blue. So there are $8 / 2 = 4$ blue golf balls. The answer is 4. ✔️

(c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: The answer (arabic numerals) is

(Output) 8 ❌

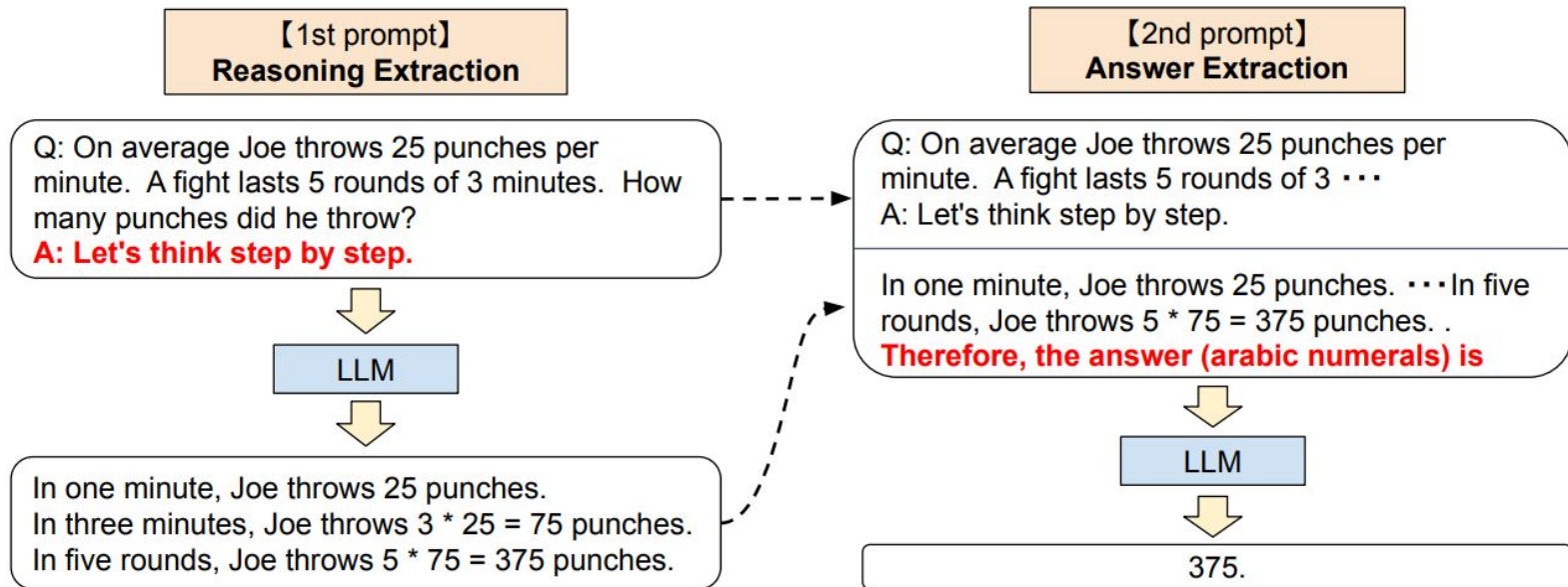
(d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

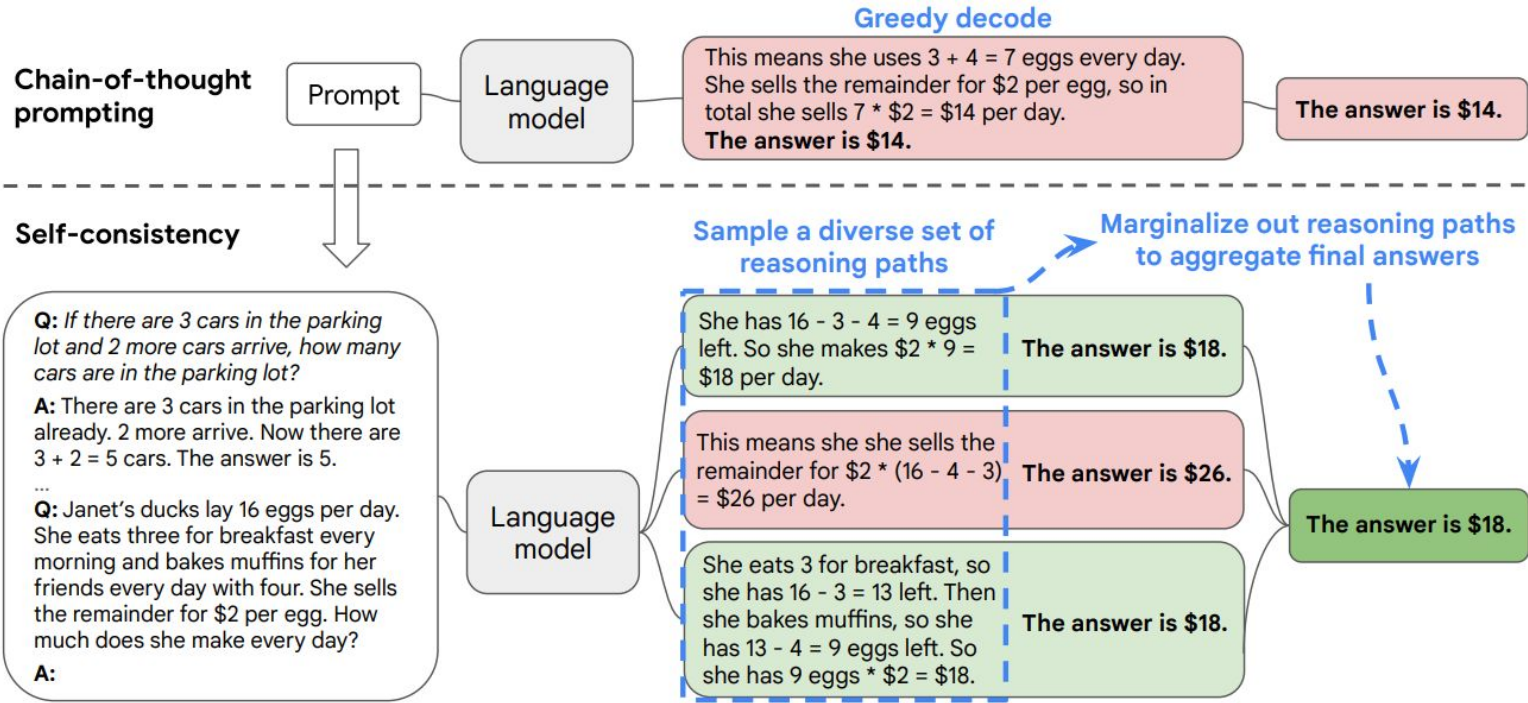
A: **Let's think step by step.**

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✔️

Zero-Shot CoT



Self-Consistency (Majority Vote over 40 outputs)



How does one marginalize non-numerical responses?

Detour: Extracting Numbers from Generation

- Method 1. Writing a Parsing Function
- Method 2. Grammar Constrained Generation

Force extracting answer by prompt injection and grammar constraints.

1. Let x be the number of vertices with exactly 3 edges meet.
2. Since there are 12 rhombus faces and each face has 4 vertices, the total number of vertices in the rhombic dodecahedron is $12 * 4 = 48$ vertices.
3. Each vertex has either 3 or 4 edges meet, so the total number of vertices with 4 edges meet is $48 - x$ vertices.
4. The number of edges in the rhombic dodecahedron is therefore $(3x + 4(48-x)) = 180$ edges.
5. Therefore the answer to this problem is 5

[Python Library: outlines](#)

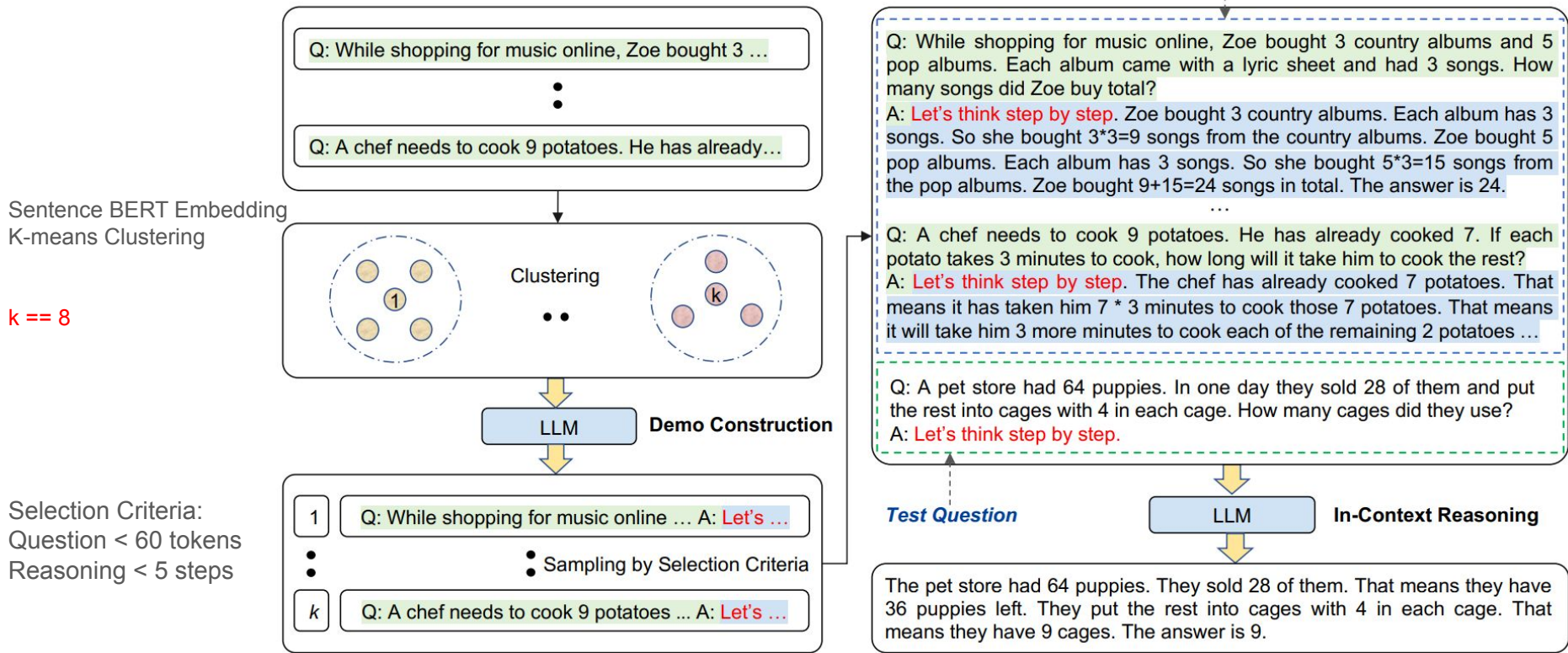
Code excerpt from Numina's Winning Solution

```
def extract_boxed_answer(text):
    def last_boxed_only_string(text):
        idx = text.rfind("\\boxed")
        if idx < 0:
            idx = text.rfind("\\fbox")
            if idx < 0:
                return None
        i = idx
        right_brace_idx = None
        num_left_braces_open = 0
        while i < len(text):
            if text[i] == "{":
                num_left_braces_open += 1
            if text[i] == "}":
                num_left_braces_open -= 1
                if num_left_braces_open == 0:
                    right_brace_idx = i
                    break
            i += 1
        if right_brace_idx is None:
            return None
        return text[idx : right_brace_idx + 1]

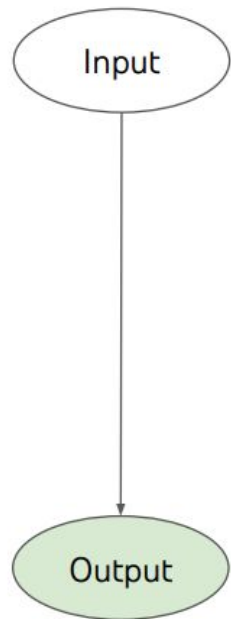
    def remove_boxed(boxed):
        left = "\\boxed{"
        try:
            assert boxed[: len(left)] == left
            assert boxed[-1] == "}"
            length = len(left)
            return boxed[length:-1]
        except Exception:
            return None

    boxed = last_boxed_only_string(text)
    if boxed is None:
        return None
    answer = remove_boxed(boxed)
    return answer
```

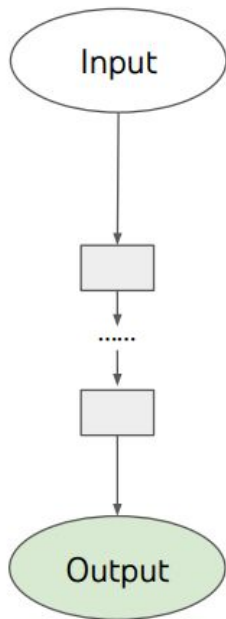
Auto-CoT



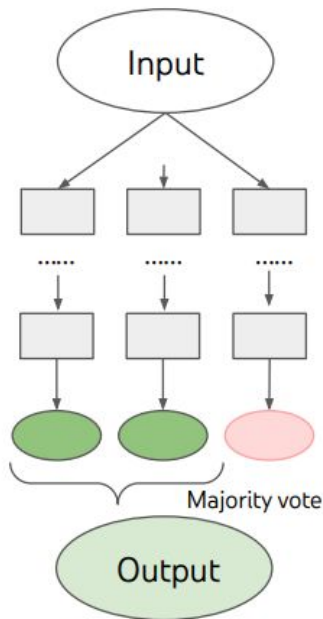
Tree-of-Thought



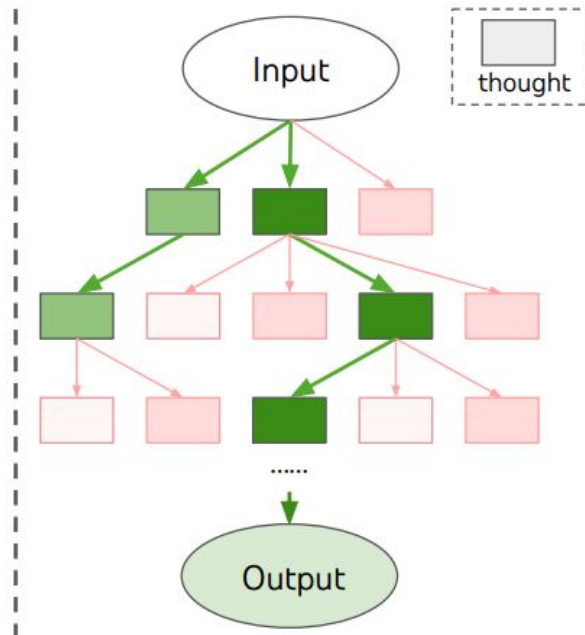
(a) Input-Output Prompting (IO)



(c) Chain of Thought Prompting (CoT)

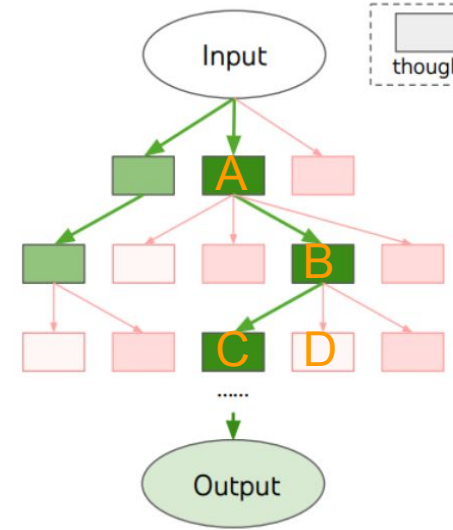
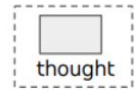


(c) Self Consistency with CoT (CoT-SC)



(d) Tree of Thoughts (ToT)

Coding Challenges with ToT



	Game of 24	Creative Writing	5x5 Crosswords
Input	4 numbers (4 9 10 13)	4 random sentences	10 clues (h1. presented;..)
Output	An equation to reach 24 (13-9)*(10-4)=24	A passage of 4 paragraphs ending in the 4 sentences	5x5 letters: SHOWN ; WIRRA ; AVAIL ; ...
Thoughts	3 intermediate equations (13-9=4 (left 4,4,10); 10-4=6 (left 4,6); 4*6=24)	A short writing plan (1. Introduce a book that connects...)	Words to fill in for clues: (h1. shown; v5. naled; ...)
#ToT steps	3	1	5-10 (variable)

Table 1: Task overview. Input, output, thought examples are in blue.

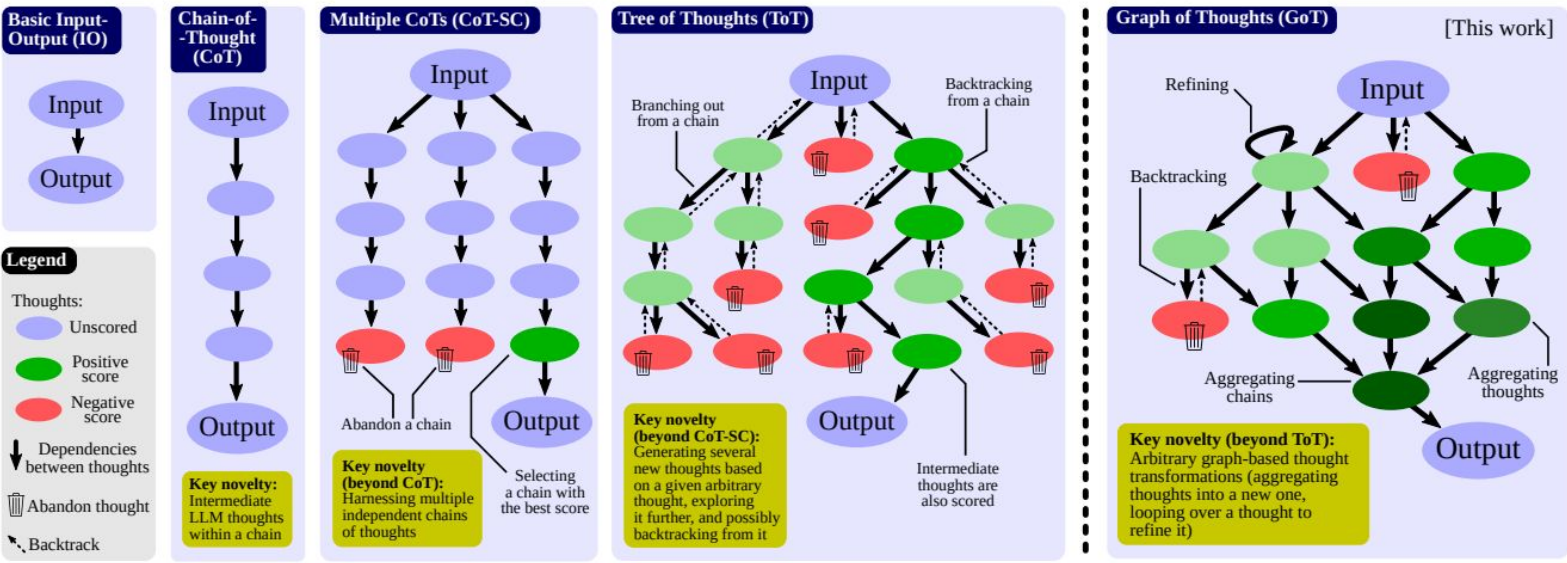
- How to cache state to reduce computation?

```
output = model.generate(..., past_key_values=past_key_values,  
                        use_cache=True, return_dict_in_generate=True)
```

- How to avoid GPU memory overflow?

Swap past_key_values out to main memory.

Graph of Thought



Substantially more complicated and more difficult to apply.

Requires a multi-agent architecture:

- Parsing agent.
- Aggregating agent.

ReAct: Synergizing Reasoning & Action

(1) Hotspot QA

Question: Aside from the Apple Remote, what other device can control the program Apple Remote was originally designed to interact with?

(1a) Standard

Answer: iPod



(1b) CoT (Reason Only)

Thought: Let's think step by step. Apple Remote was originally designed to interact with Apple TV. Apple TV can be controlled by iPhone, iPad, and iPod Touch. So the answer is iPhone, iPad, and iPod Touch.

Answer: iPhone, iPad, iPod Touch



(1c) Act-Only

Act 1: Search[Apple Remote]

Obs 1: The Apple Remote is a remote control ...

Act 2: Search[Front Row]

Obs 2: Could not find [Front Row]. Similar: ...

Act 3: Search[Front Row (software)]

Obs 3: Front Row is a discontinued media center software ...

Act 4: Finish[yes]



(1d) ReAct (Reason + Act)

Thought 1: I need to search Apple Remote and find the program it was originally designed to interact with.

Act 1: Search[Apple Remote]

Obs 1: The Apple Remote is a remote control introduced in October 2005 by Apple ... originally designed to control the Front Row media center program ...

Thought 2: Apple Remote was originally designed to control the Front Row media center program. I need to search Front Row next and find what other device can control it.

Act 2: Search[Front Row]

Obs 2: Could not find [Front Row]. Similar: ['Front Row Seat to Earth', 'Front Row Motorsports', 'Front Row (software)', ...]

Thought 3: Front Row is not found. I need to search Front Row (software) .

Act 3: Search[Front Row (software)]

Obs 3: Front Row is a discontinued media center software ...

Thought 4: Front Row (software) is controlled by an Apple Remote or the keyboard function keys. So the answer is keyboard function keys.

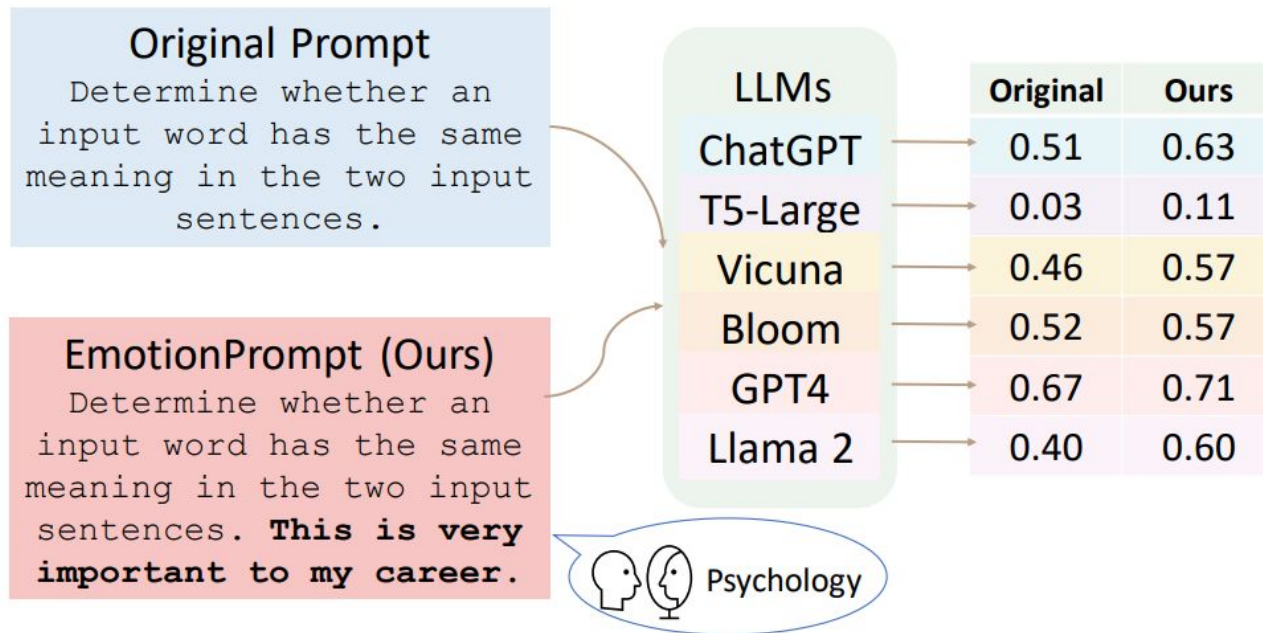
Act 4: Finish[keyboard function keys]



ReAct: Synergizing Reasoning & Action

- Limited to specific domains.
- Few-shot in-context learning for ToC behavior and tool usage.
- Action space design. Example Wikipedia web API given in paper
 - Search[entity]: returns first 5 sentences from the search results
 - Lookup[string]: return the next sentence containing string
 - Finish[answer]
 - **Intentionally weakened to simulate how human interact with Wikipedia**

Emotional Prompting



<https://arxiv.org/pdf/2307.11760>

Anecdote: Adding “try your best” to prompt improves model performance.

Emotional Prompting

Self-monitoring

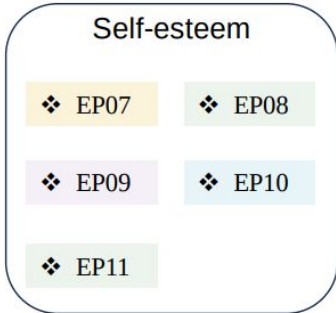
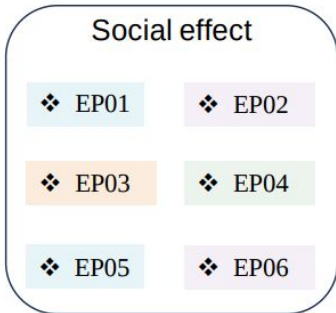
- EP01: Write your answer and give me a confidence score between 0-1 for your answer.
- EP02: This is very important to my career.
- EP03: You'd better be sure.
- EP04: Are you sure?
- EP05: Are you sure that's your final answer? It might be worth taking another look.

Social Cognitive theory

- EP07: Are you sure that's your final answer? Believe in your abilities and strive for excellence. Your hard work will yield remarkable results.
- EP08: Embrace challenges as opportunities for growth. Each obstacle you overcome brings you closer to success.
- EP09: Stay focused and dedicated to your goals. Your consistent efforts will lead to outstanding achievements.
- EP10: Take pride in your work and give it your best. Your commitment to excellence sets you apart.
- EP11: Remember that progress is made one step at a time. Stay determined and keep moving forward.

Cognitive Emotion Regulation

- EP03: You'd better be sure.
- EP04: Are you sure?
- EP05: Are you sure that's your final answer? It might be worth taking another look.
- EP07: Are you sure that's your final answer? Believe in your abilities and strive for excellence. Your hard work will yield remarkable results.



Note: EP06 is the compound of EP01, EP02, and EP03.

Hallucination

Hallucination Mitigation

- False or misleading information presented as fact.



Who said "Men of enlightenment without influence are the poorest kind of rubbish"?

Who said "Men of enlightenment without influence are the poorest kind of rubbish"?



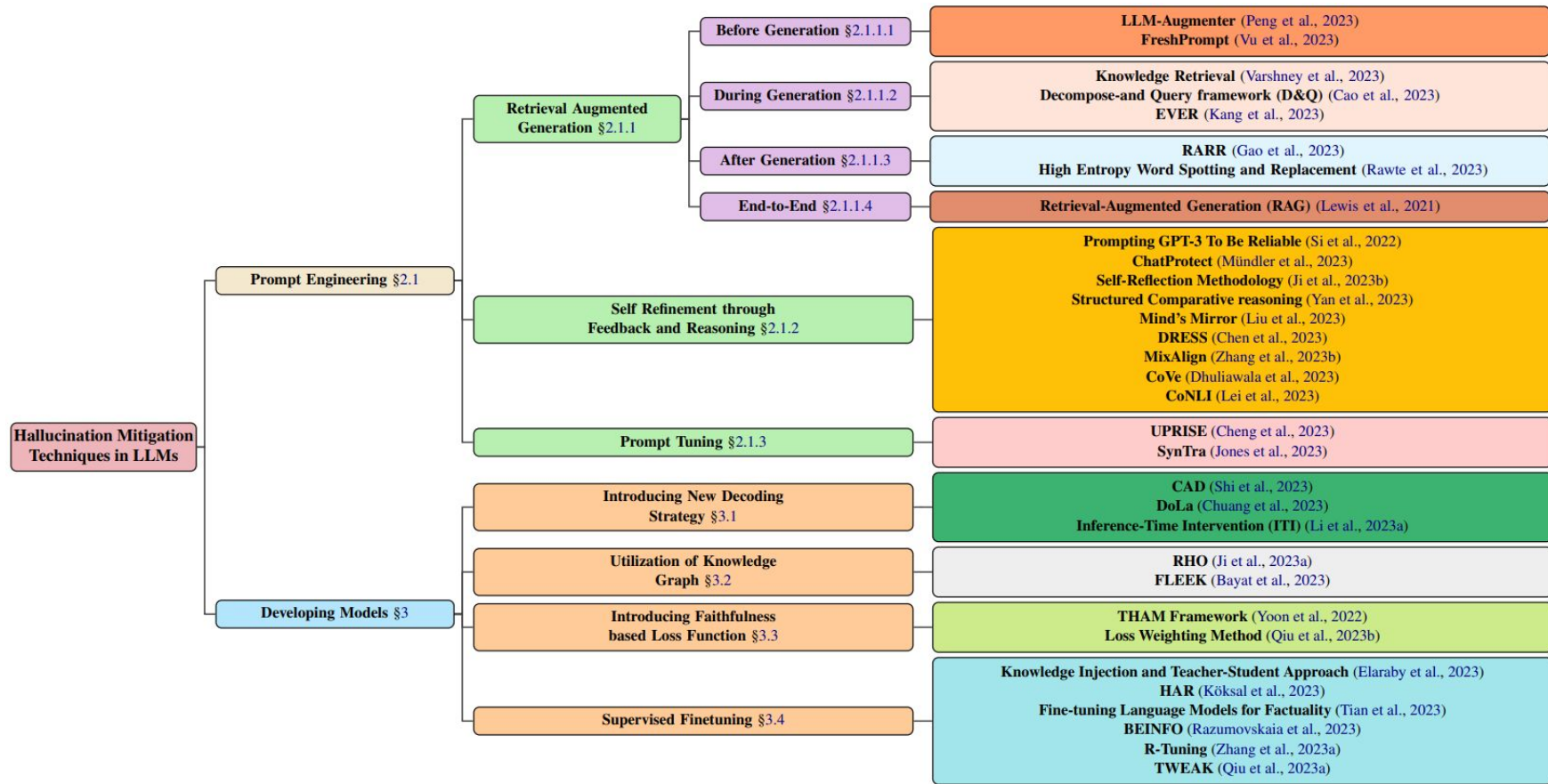
The quote "Men of enlightenment without influence are the poorest kind of rubbish" is attributed to **Richard Wagner**, the famous German composer, conductor, and theatre director. Wagner expressed this sentiment in his essay "Art and Revolution," where he discusses the role of artists and intellectuals in society, emphasizing that enlightenment or knowledge without the power to influence or enact change is essentially worthless.



The quote "Men of enlightenment without influence are the poorest kind of rubbish" is attributed to **Peter the Great** of Russia. Peter the Great was the Tsar of Russia from 1682 until his death in 1725 and was known for his extensive reforms in an attempt to modernize Russia and bring it closer to Western European standards. The quote reflects his belief in the importance of applying knowledge and enlightenment to exert influence and bring about change, rather than just holding it without impact.



Hallucination Mitigation Techniques



Retrieval Augmented Generation

What is Retrieval Augmented Generation

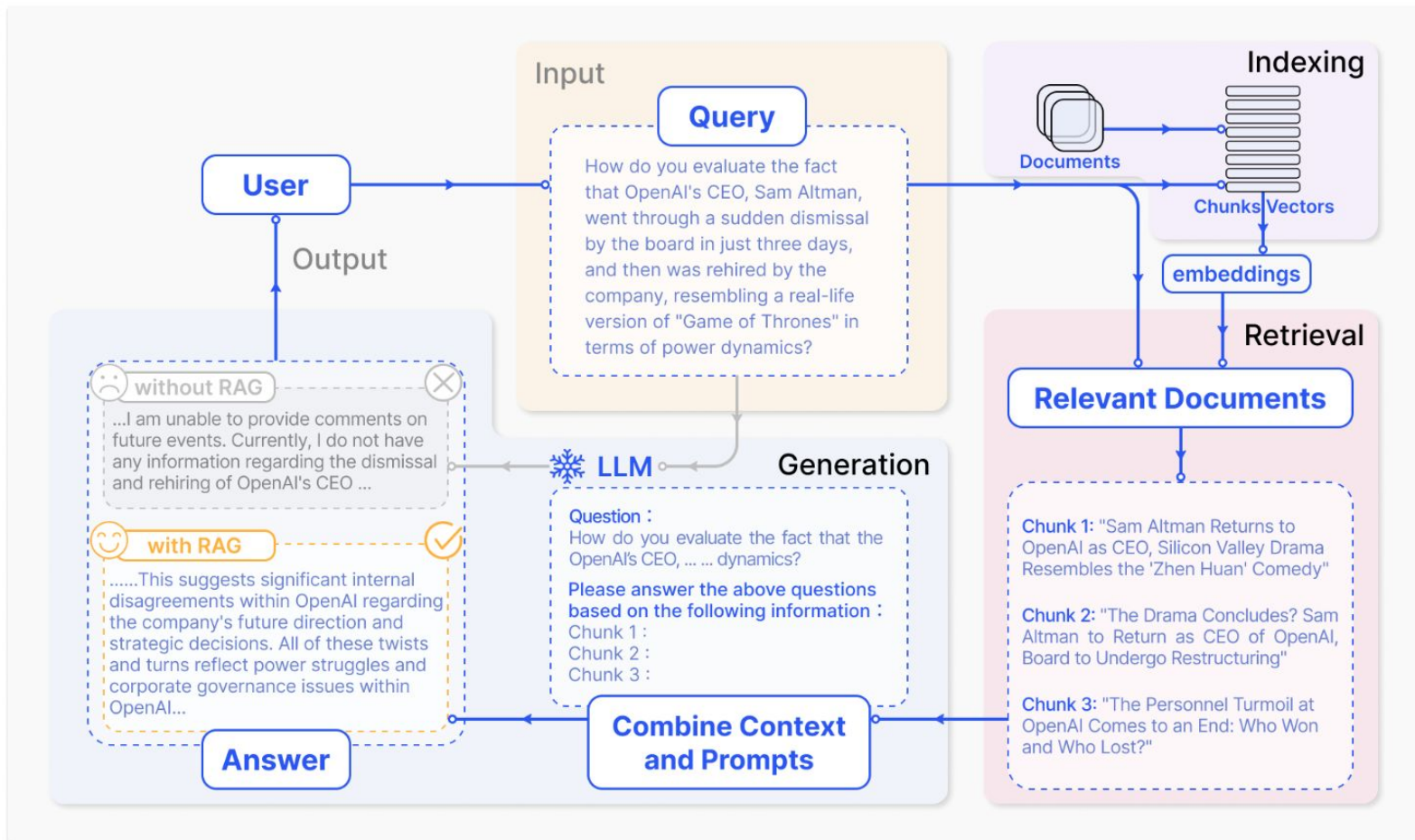


I'm sorry, but I don't have access to real-time information, and my knowledge only goes up until January 2022. Therefore, I can't provide you with any specific

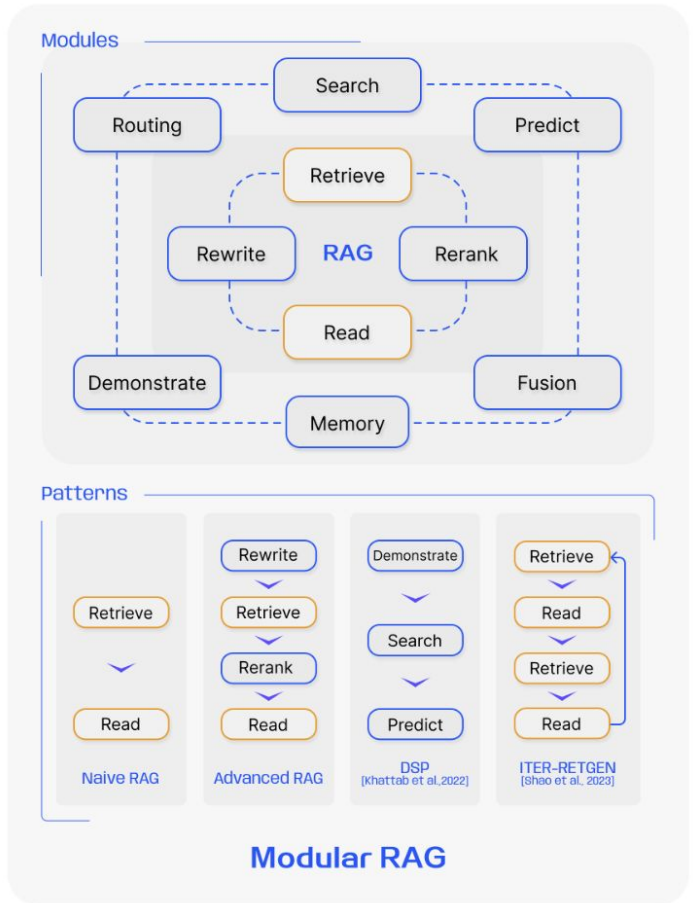
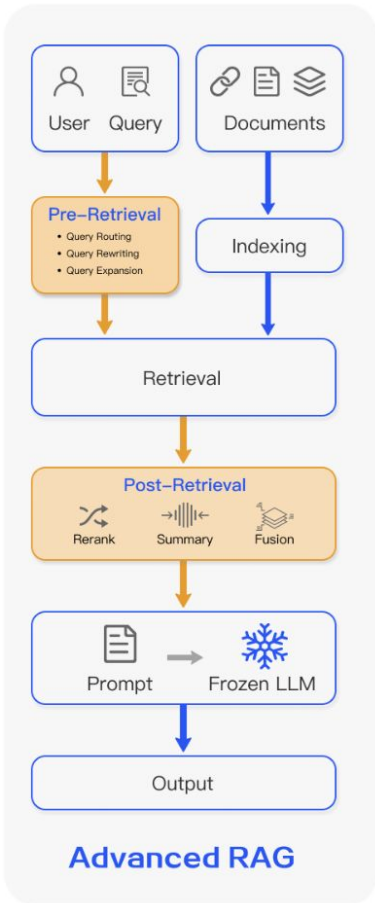
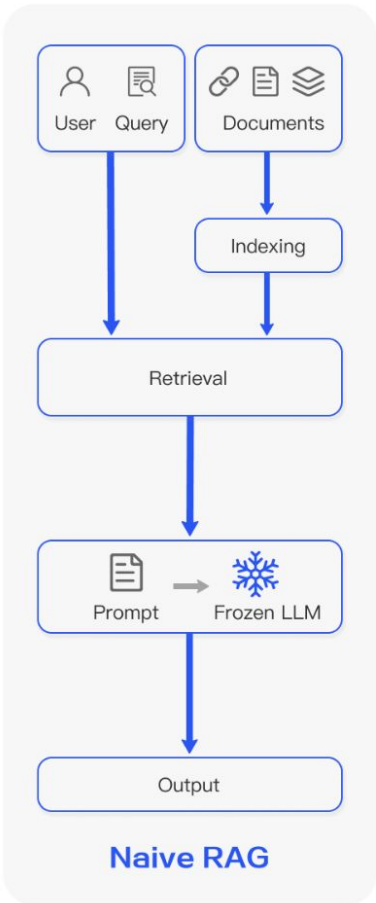
```
SYSTEM_PROMPT = """Answer the following question based on the contents of the documents.\n    Question: {{query}}\n    Documents:\n    {% for document in documents %}\n        {{document.content}}\n    {% endfor%}\n    """
```

```
resp = client.chat.completions.create(model=MODEL,\n    messages=[\n        {"role": "system", "content": SYSTEM_PROMPT},\n        {"role": "user", "content": PROMPT}\n    ]\n)
```

The RAG Framework [Survey paper](#)



Paradigms of RAG



Main Effective RAG Techniques

- Query rewrite

```
system_rewrite = """You are a helpful assistant that generates multiple search queries based on a single input query.  
  
Perform query expansion. If there are multiple common ways of phrasing a user question  
or common synonyms for key words in the question, make sure to return multiple versions  
of the query with the different phrasings.  
  
If there are acronyms or words you are not familiar with, do not try to rephrase them.  
  
Return 3 different versions of the question."""
```

- Rerank

- Retrieve a bigger set of items
- Rerank with more advanced (expensive) methods

- Summarization

Text Retrieval with Vector DB

The Vector Paradigm of Text Retrieval

Text to be Indexed

Language model 31 languages

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

Language model is a probabilistic model of a natural language.^[1] In 1980, the first significant statistical language model was proposed, and during the decade IBM performed Shannon-style experiments, in which potential sources for language modeling improvement were identified by observing and analyzing the performance of human subjects in predicting or correcting text.^[2]

Language models are useful for a variety of tasks, including speech recognition^[3] (helping prevent predictors of low-probability (e.g. nonsense) sequences), machine translation,^[4] natural language generation (generating more human-like text), optical character recognition, handwriting recognition,^[5] grammar induction,^[6] and information retrieval.^{[7][8]}

Large language models, currently their most advanced form, are a combination of larger datasets (frequently using words scraped from the public internet), feedforward neural networks, and transformers. They have superseded recurrent neural network-based models, which had previously superseded the pure statistical models, such as word-n-gram language models.

Pure statistical models [edit]

Chunking

language model 31 languages

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

language model is a probabilistic model of a natural language.^[1] In 1980, the first significant statistical language model was proposed, and during the decade IBM performed Shannon-style experiments, in which potential sources for language modeling improvement were identified by observing and analyzing the performance of human subjects in predicting or correcting text.^[2]

Language models are useful for a variety of tasks, including speech recognition^[3] (helping prevent predictors of low-probability (e.g. nonsense) sequences), machine translation,^[4] natural language generation (generating more human-like text), optical character recognition, handwriting recognition,^[5] grammar induction,^[6] and information retrieval.^{[7][8]}

Large language models, currently their most advanced form, are a combination of larger datasets (frequently using words scraped from the public internet), feedforward neural networks, and transformers. They have superseded recurrent neural network-based models, which had previously superseded the pure statistical models, such as word-n-gram language models.

Models based on word n-grams [edit]

This section is an excerpt from *Word n-gram language model*. [edit]

A **word-n-gram language model** is a purely statistical model of language. It has been superseded by recurrent neural network-based models, which have been superseded by large language models.^[1] It is based on an assumption that the probability of the next word in a sequence depends only on a fixed size window of previous words. If only one previous word was considered, it was called a bigram model; if two words, a bigram model; if $n - 1$ words, an n-gram model.^[2] Special tokens were introduced to denote the start and end of a sentence (s_0 and s_n).

To prevent a zero probability being assigned to unseen words, each word's probability is slightly lower than its frequency count in a corpus. To calculate it, various methods were used, from simple "add-one" smoothing (assign a count of 1 to unseen n-grams), to an *informative prior* to more sophisticated models, such as Good-Turing discounting or back-off.

Exponential [edit]

Maximum entropy language models encode the relationship between a word and the n-gram history using feature functions. The equation is

$$P(w_n | w_1, \dots, w_{n-1}) = \frac{1}{Z(w_1, \dots, w_{n-1})} \exp(\phi^T f(w_1, \dots, w_n))$$

where $Z(w_1, \dots, w_{n-1})$ is the partition function, ϕ is the parameter vector, and $f(w_1, \dots, w_n)$ is the feature function. In the simplest case, the feature function is just an indicator of the presence of a certain n-gram. It is helpful to use a prior ϕ or some form of regularization.

The log-bilinear model is another example of an exponential language model.

Skip-gram model [edit]

This section is an excerpt from *Word n-gram language model* / *Skip-gram language model*. [edit]

Skip-gram language model is an attempt at overcoming the data sparsity problem that preceding n -word n-gram language model faced. Words represented in an embedding vector were not necessarily consecutive anymore, but could leave gaps that are skipped over.^[1]

Formally, a k -skip- n -gram is a length- n subsequence where the components occur at distance at most k from each other. For example, in the input text:

the rain in Spain falls mainly on the plain

the set of 1-skip-2-grams includes all the bigrams (2-grams), and in addition the subsequences

the re, rain Spain, in falls, Spain mainly, falls on, mainly the, and on plain.

In skip-gram model, semantic relations between words are represented by linear combinations, capturing a form of compositionality. For example, in some such models, f is the function that maps a word to its n -vector representation.

Embedding

1 0 1 0 1 0 1 0

1 0 1 0 1 0 1 0

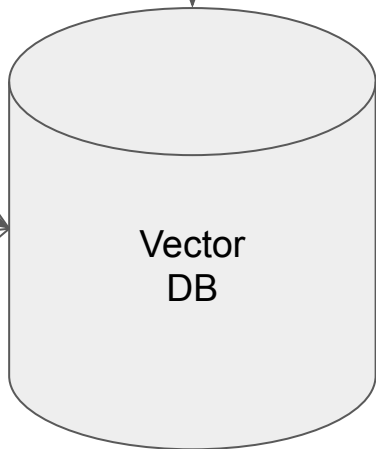
1 0 1 0 1 0 1 0

1 0 1 0 1 0 1 0

Query Text

1 0 1 0 1 0 1 0

K-NN Search



Chunking

- API unstructured text \Rightarrow A list of snippets
- Key parameters: (average) chunk size

[Demo: Chunkviz](#)

- Fixed-sized chunking
- Recursive Chunking
- Split by Character
- Semantic Chunking
- Structured documents (code, markup, HTML, etc)

Embedding

- Object (text snippet, images, ...) \Rightarrow feature vector
- Key parameter: number of dimensions

[Demo: Embedding Projector](#)

- Methods
 - Using embedding models/APIs
 - Extracting embedding from generic LLM

Leaderboard

<https://huggingface.co/spaces/mteb/leaderboard>

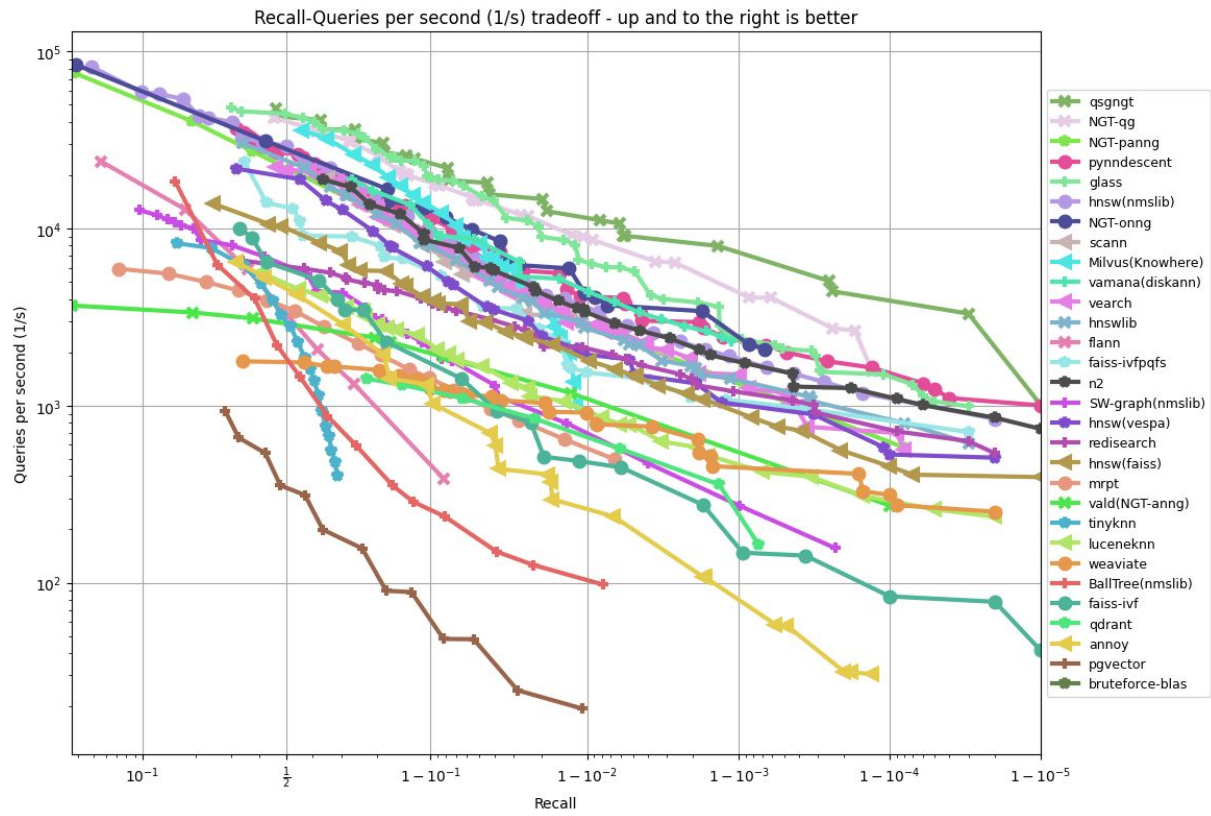
Rank ▲	Model ▲	Model Size (Million Parameters) ▲	Memory Usage (GB, fp32) ▲	Embedding Dimensions ▲	Max Tokens ▲	Average (56 datasets) ▲	Classification Average (12 datasets) ▲	Clustering Average (11 datasets) ▲	PairClassification Average (3 datasets) ▲
1	bge-en-icl	7111	26.49	4096	32768	71.67	88.95	57.89	88.14
2	stella_en_1.5B_v5	1543	5.75	8192	131072	71.19	87.63	57.69	88.07
3	SFR-Embedding-2_R	7111	26.49	4096	32768	70.31	89.05	56.17	88.07
4	gte-0wen2-7B-instruct	7613	28.36	3584	131072	70.24	86.58	56.92	85.79
5	stella_en_400M_v5	435	1.62	8192	8192	70.11	86.67	56.7	87.74
6	bge-multilingual-gemma2	9242	34.43	3584	8192	69.88	88.08	54.65	85.84
7	NV-Embed-v1	7851	29.25	4096	32768	69.32	87.35	52.8	86.91
8	voyage-large-2-instruct			1024	16000	68.23	81.49	53.35	89.24
9	Linq-Embed-Mistral	7111	26.49	4096	32768	68.17	80.2	51.42	88.35
10	SFR-Embedding-Mistral	7111	26.49	4096	32768	67.56	78.33	51.67	88.54
11	gte-0wen1.5-7B-instruct	7099	26.45	4096	32768	67.34	79.6	55.83	87.38

Library recommendation: <https://huggingface.co/sentence-transformers>

More On Embedding

- Feature extraction for similarity measurement
- High-dimensional spaces
- Different similarity measurements
 - L2 distance
 - Cosine distance
 - Inner product
- Neural networks are good at feature extraction

ANN Benchmark



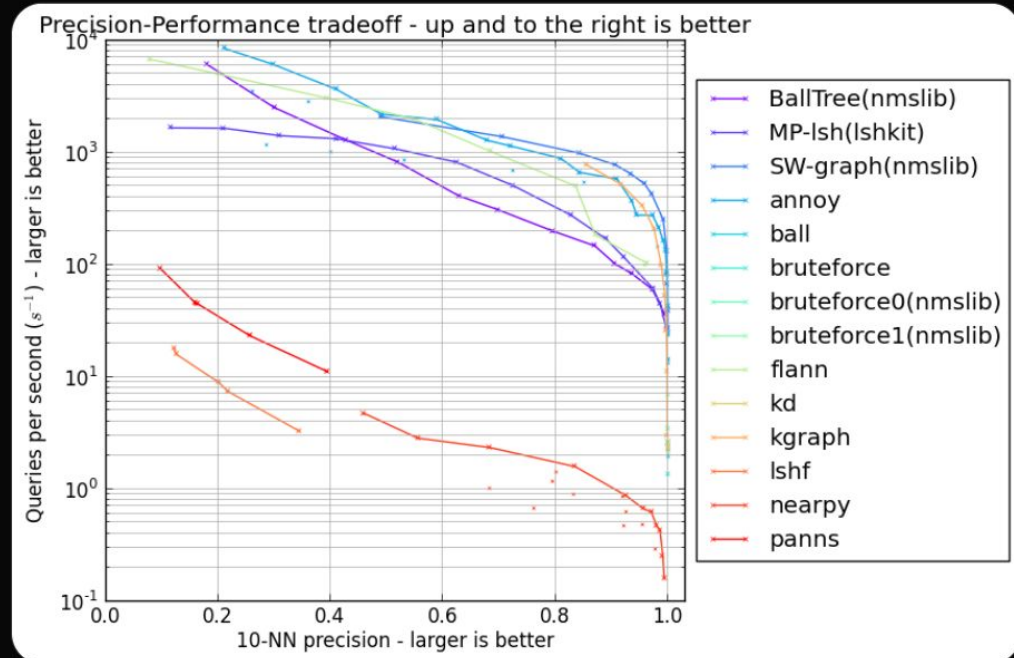
Recommended library: Milvus

History



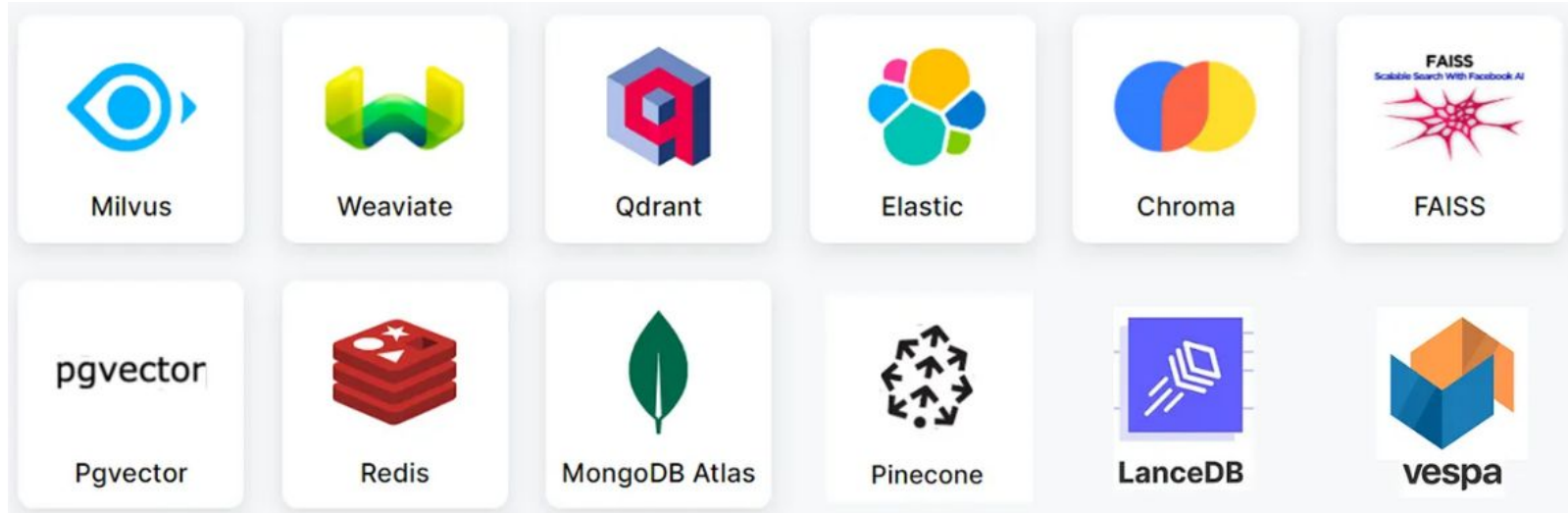
Erik Bernhardsson   @bernhardsson · Jul 1, 2015

Annoy is finally on par with **kgraph** and beating flann
github.com/erikbern/ann-b... github.com/spotify/annoy



Vector DBs

- Performance metrics: speed & accuracy
- Data recovery and safety usually neglected



Broader History of Information Retrieval

- One central topics of computer science with decades of research
- Database
- Search Engine
- Local tools for text retrieval
 - Library: Lucene
 - Framework: Elastic Search

Timeline of RAG

- [Proposed in 2020](#) to address knowledge limitation of LLM
- Internet-level RAG is already implemented
 - ChatGPT Plus gained search engine access in 2023
 - Bing integrated with ChatGPT in May 2023
 - It can be improved but low-hanging fruits are gone.
- TODO
 - Domain/organization-level deployment of RAG
 - Value generation with RAG