LLM-BASED APPLICATIONS WITH ZERO-SHOT AND FEW-SHOTS PROMPTING

23020390- Nguyễn Thị Ngọc Lan

EFFICIENT PROMPTING METHODS FOR LARGE LANGUAGE MODELS: A SURVEY

Vì sao "prompting" quan trọng?

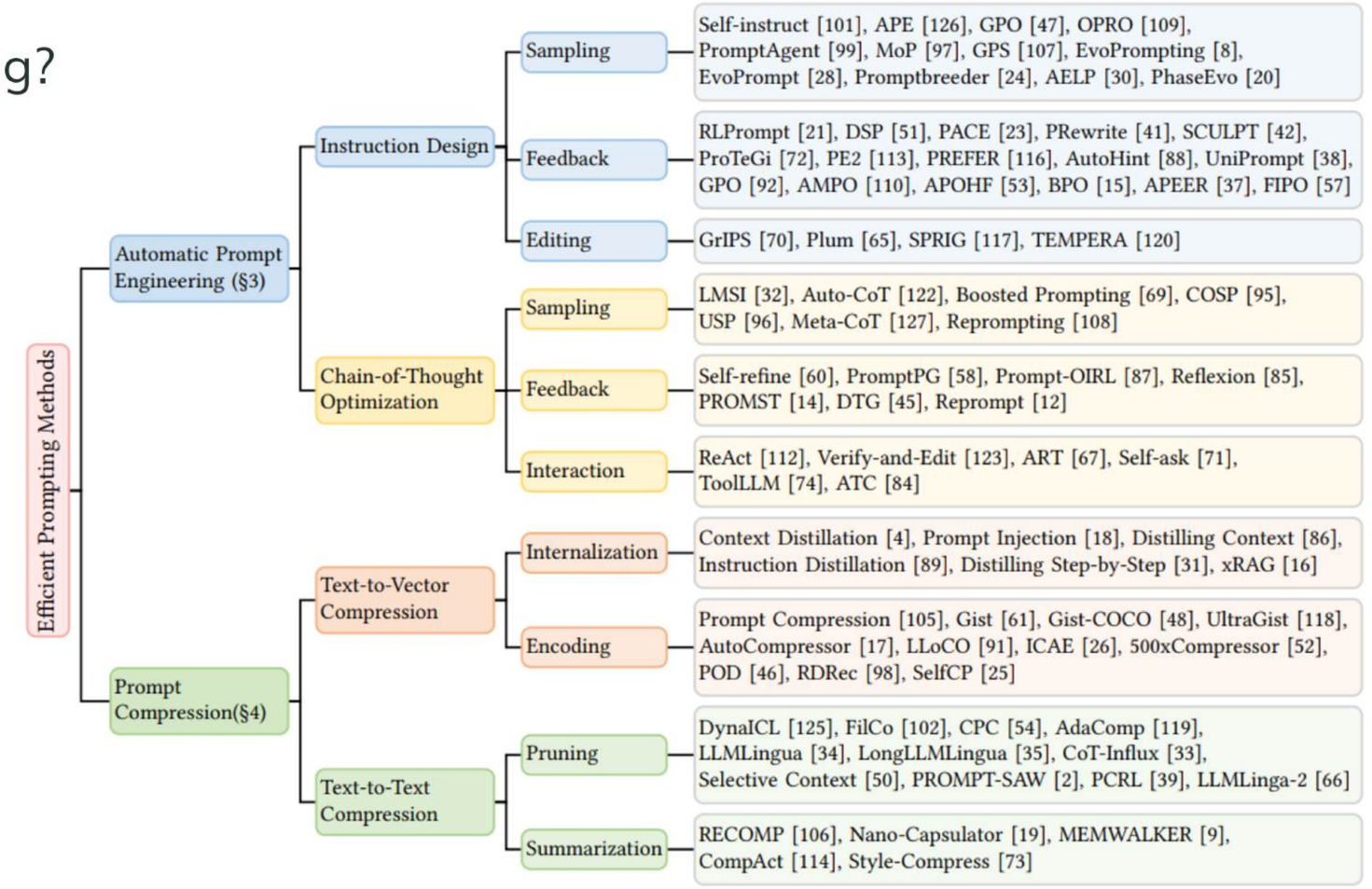


Fig. 1. Taxonomy of efficient prompting methods.

EFFICIENT PROMPTING METHODS FOR LARGE LANGUAGE MODELS: A SURVEY

Mô hình toán học

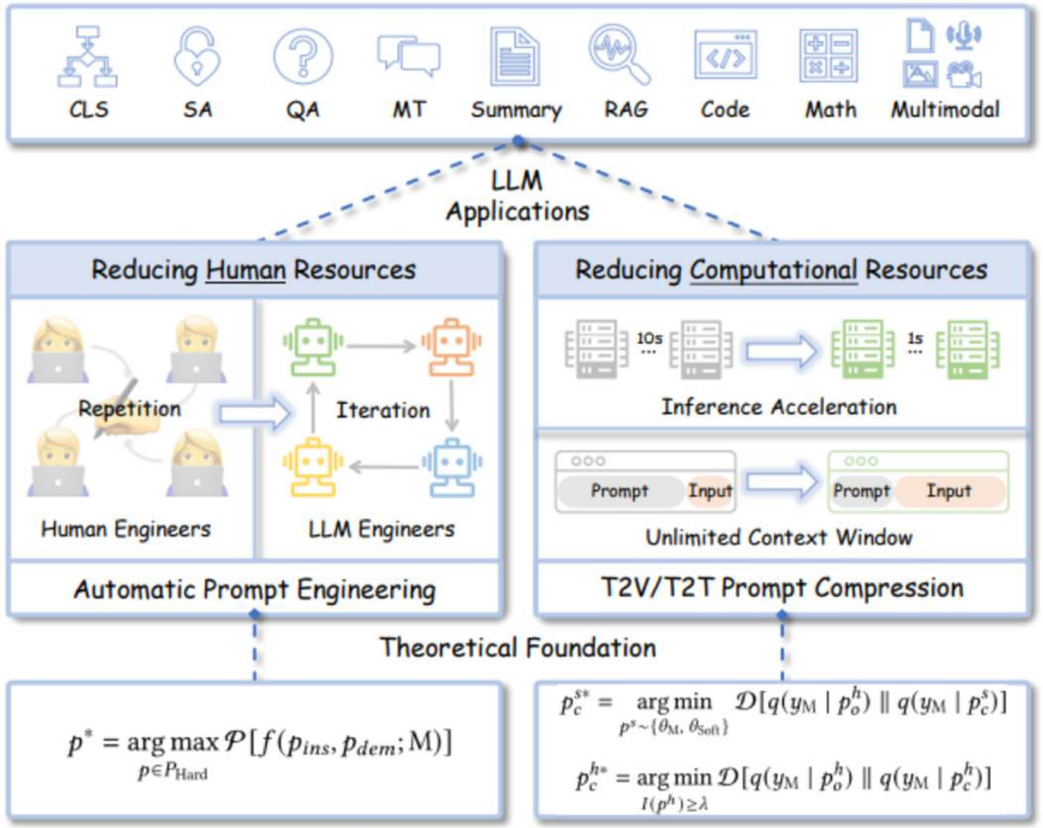


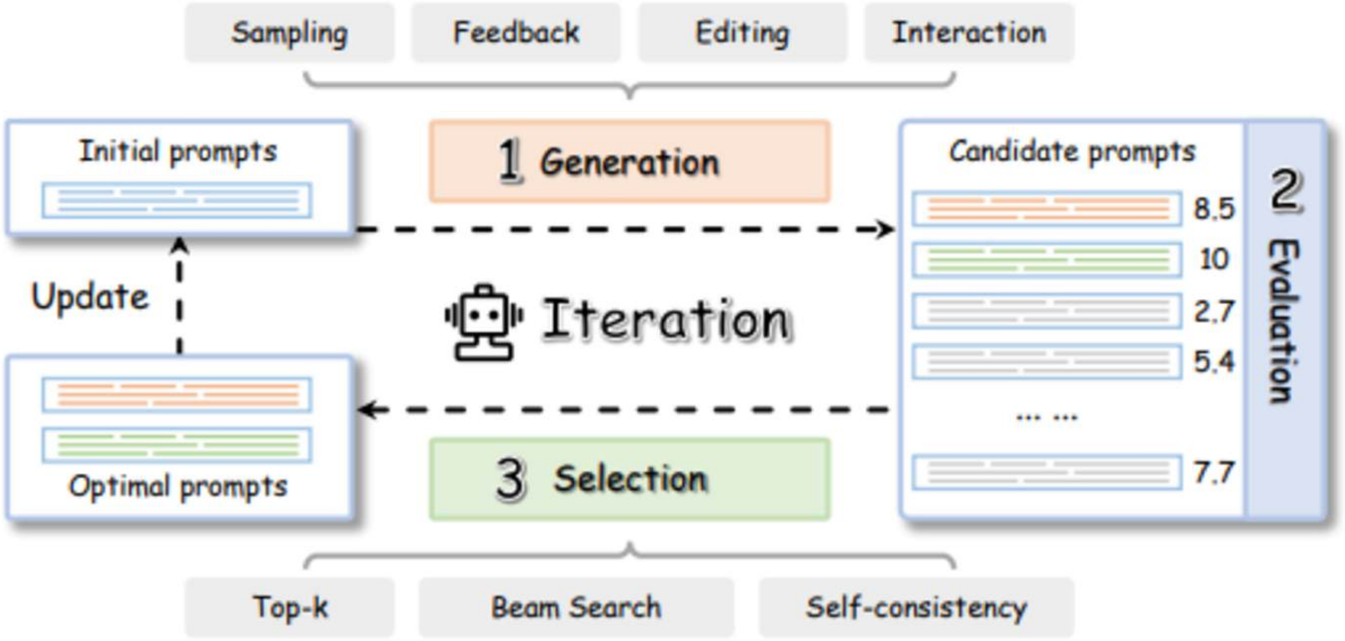
Fig. 2. An overview of efficient prompting methods.

$$p_c^{s*} = \arg\min_{p^s \sim \{\theta_{\text{M}}, \theta_{\text{Soft}}\}} \mathcal{D}[q(y_{\text{M}} \mid p_o^h) \parallel q(y_{\text{M}} \mid p_c^s)] \tag{2}$$

$$p_c^{h*} = \arg\min_{I(p^h) \geq \lambda} \mathcal{D}[q(y_{\text{M}} \mid p_o^h) \parallel q(y_{\text{M}} \mid p_c^h)] \tag{3}$$

EFFICIENT PROMPTING METHODS FOR LARGE LANGUAGE MODELS: A SURVEY

Basic pipeline



EFFICIENT PROMPTING METHODS FOR LARGE LANGUAGE MODELS: A SURVEY

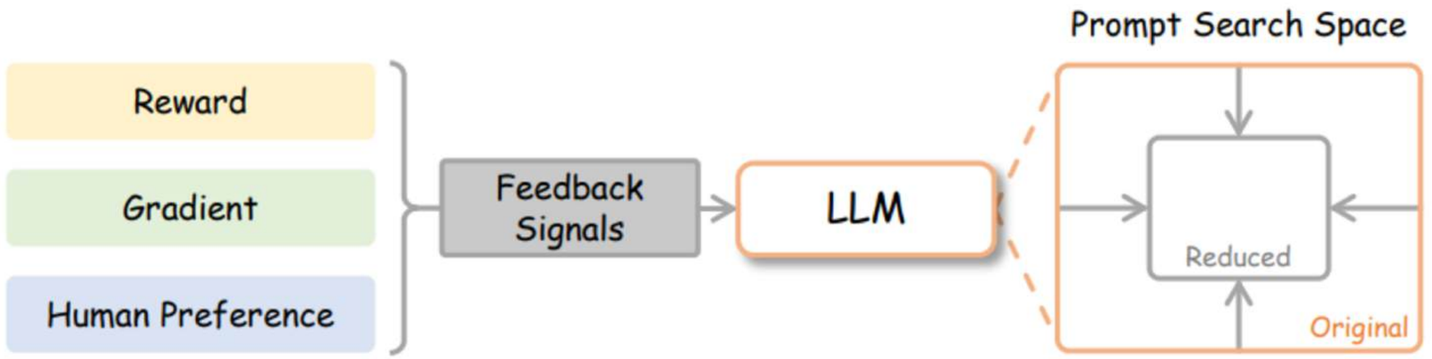
Automatic Prompt Engineering: Challenges and Solutions

1. Instruction Design

Sampling-based methods

Feedback-based methods

Editing-based methods



EFFICIENT PROMPTING METHODS FOR LARGE LANGUAGE MODELS: A SURVEY

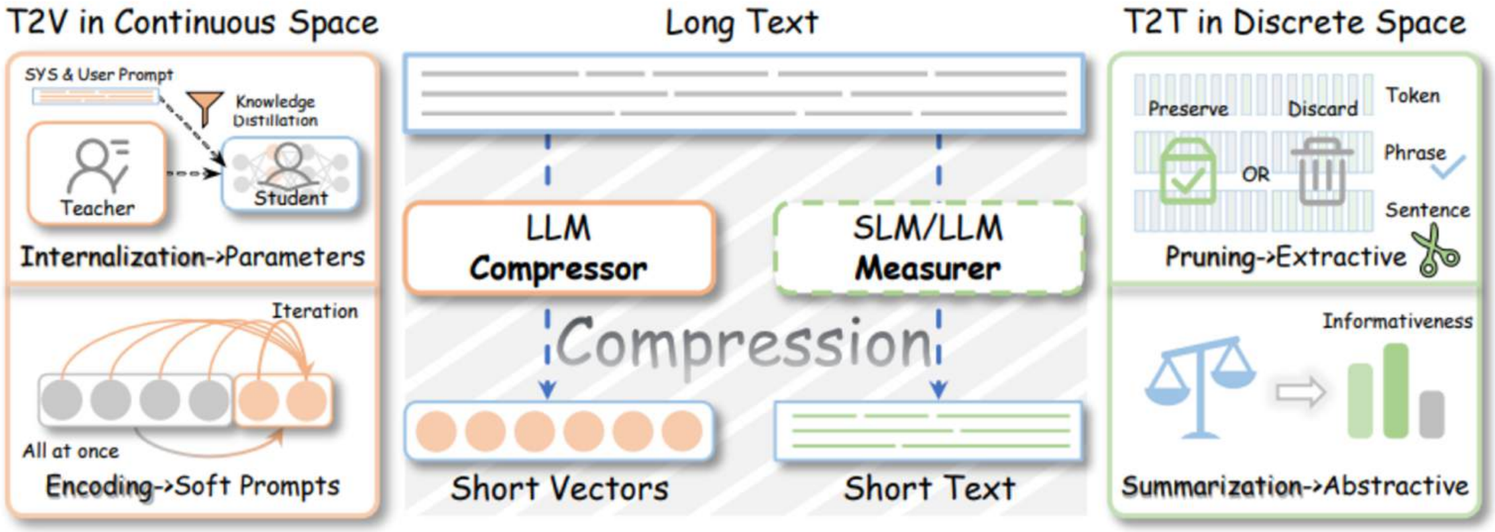
Prompt Compression

2. CoT Optimization

Text-to-Vector Compression

Interaction-based methods

Feedback-based methods



EFFICIENT PROMPTING METHODS FOR LARGE LANGUAGE MODELS: A SURVEY

Text-to-Vector Compression

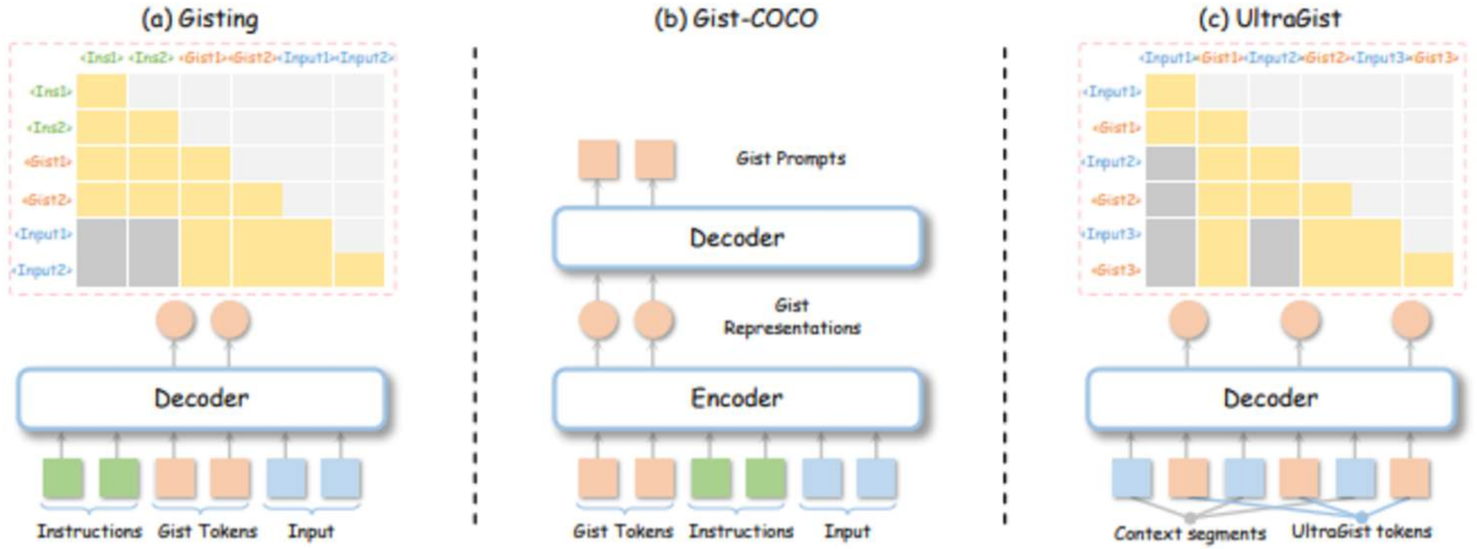


Fig. 6. Gisting series work compresses key information of hard prompts into gist tokens based on an encoder or a decoder trained with special attention mechanisms. The matrices in the upper half represent masking strategies, where the gray box indicates the standard mask and the yellow box indicates the gist mask.

EFFICIENT PROMPTING METHODS FOR LARGE LANGUAGE MODELS: A SURVEY

Text-to-Vector Compression

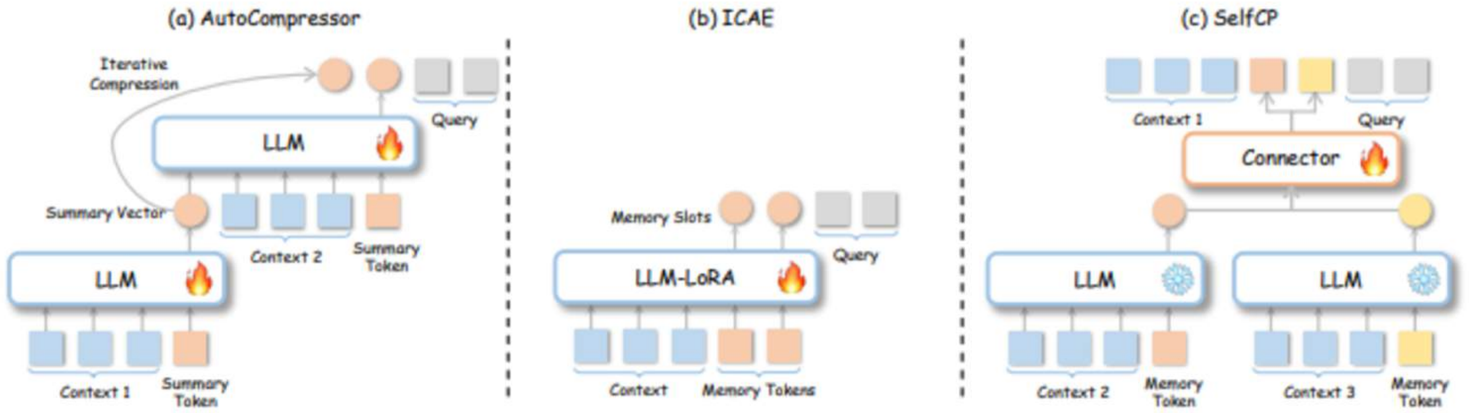


Fig. 7. Differences of representative encoding methods specially for long context. AutoCompressor iteratively compresses context segments with summary tokens. ICAE compresses the complete context all at once with memory tokens. SelfCP only compresses unlimited context segments based on a connector.

EFFICIENT PROMPTING METHODS FOR LARGE LANGUAGE MODELS: A SURVEY

Text-to-Vector Compression

Encoding Methods	Target Model	Compressor Model	Soft prompt location	Hard prompt	Soft prompt	
Prompt Compression	Decoder-only	Bayesian attribute classifier framework	Before context	Context	Embeddings	Vectors
Gisting	Encoder-Decoder Decoder-only	With gist masking	Between instruction and context	Instruction	Gist tokens	Vectors
Gist-COCO	Encoder-Decoder	Encoder	Before prompt	Prompt	Gist tokens	Gist representations
UltraGist	Decoder-only	With optimized cross-attention	After context segment	Context	Gist tokens	Vectors
AutoCompressor	RMT	RMT	After context segment	Context	Summary tokens	Summary vectors
ICAE	Encoder-Decoder	Encoder (LoRA)	After context	Context	Memory tokens	Memory Slots
500xCompressor	Encoder-Decoder	Encoder (LoRA)	After context	Context	Compressed tokens	K V values
POD	Encoder-Decoder	Encoder	Before context	Context	Embeddings	Vectors
RDRec	Encoder-Decoder	Encoder	Before context	Rationale	Embeddings	Vectors
SelfCP	Decoder-only	Decoder-only	After context segment	Over-limit context	Memory tokens	Vectors

EFFICIENT PROMPTING METHODS FOR LARGE LANGUAGE MODELS: A SURVEY

Text-to-Text Compression

Methods	Compression Granularity	NaturalQuestions		GSM8K		BBH		ZeroSCROLLS		LongBench		
		F1	Ratio	EM	Ratio	EM	Ratio	Acc	Ratio	Acc	Ratio	Latency
DynalCL	demonstration	42.40(EM)	10-shot	-	-	-	-	-	-	-	-	-
		40.20(EM)	5-shot	-	-	-	-	-	-	-	-	-
FliCo	sentence	61.80	5-shot	-	-	-	-	-	-	-	-	-
CPC	sentence	-	-	-	-	-	-	34.90	3×	50.00	3×	1×
		-	-	-	-	-	-	33.80	5×	49.50	5×	-
AdaComp	document	70.96	3.66-shot	-	-	-	-	-	-	-	-	-
LLMLingua	demonstration ->token	-	-	79.08	5×	70.11	3×	30.70	3×	37.40	3×	9.8×
		30.00	3.8×	77.41	14×	61.60	5×	27.20	5×	34.60	5×	-
LongLLMLingua	document ->token	75.50	3.9×	-	-	-	-	32.80	3×	48.80	3×	10.93×
		-	-	-	-	-	-	32.50	6×	48.00	6×	-
CoT-Influx	CoT ->token	-	-	73.31	7.7×	-	-	-	-	-	-	-
Selective Context	token, phrase, sentence	43.80	3.7×	53.98	5×	54.27	3×	20.70	3×	32.00	3×	-
		-	-	52.99	11×	54.02	5×	19.40	5×	24.80	5×	-
PROMPT-SAW	entity, relation	73.22(EM)	3.86×	72.12	1.49×	-	-	-	-	-	-	-
LLMLingua-2	token	71.90	3.9×	79.08	5×	70.02	3×	33.50	3×	42.20	3×	0.67×
		-	-	77.79	14×	61.94	5×	33.40	5×	39.10	5×	-

PROMPT DESIGN AND ENGINEERING: INTRODUCTION AND ADVANCED METHODS

Các kiểu prompt thường:

- Instructions + Question

"How should I write my college admission essay? Give me suggestions about the different sections I should include, what tone I should use, and what expressions I should avoid."

- Instructions + Input

"Given the following information about me, write a 4 paragraph college essay: I am originally from Barcelona, Spain. While my childhood had different traumatic events, such as the death of my father when I was only 6, I still think I had quite a happy childhood.. During my childhood, I changed schools very often, and attended all kinds of schools, from public schools to very religious private ones. One of the most "exotic" things I did during those years is to spend a full school year studying 6th grade in Twin Falls, Idaho, with my extended family.

I started working very early on. My first job, as an English teacher, was at age 13. After that, and throughout my studies, I worked as a teacher, waiter, and even construction worker."

- Question + Examples

"Here are some examples of TV shows I really like: Breaking Bad, Peaky Blinders, The Bear. I did not like Ted Lasso. What other shows do you think I might like?"

EFFICIENT PROMPTING METHODS FOR LARGE LANGUAGE MODELS: A SURVEY

Zero-shot và Few-shot Prompting

Prompt:

```
Instruction: You are a generating model. Now generate a poem in 3-5 sentences
Answer:
```

The sun sets on the horizon, Painting the sky with hues of orange and pink. As the day comes to an end, the world is bathed in a golden glow. T gentle breeze whispers through the trees, carrying with it the scent of blooming flowers. It's a moment of peace and tranquility, a reminder of beauty that surrounds us. In this fleeting moment, time stands still, and we are reminded to appreciate the simple joys of life. As the stars begi twinkle in the night sky, we can't help but feel a sense of wonder and gratitude for the gift of existence. This is the magic of poetry, capturing th essence of a single moment and transforming it into something timeless and eternal. So let us embrace the power of words and allow our imaginations to soar, for poetry has the ability to transport us to places we've never been and touch our hearts in ways we never thought poss Let the words flow and let your creativity take flight!

EFFICIENT PROMPTING METHODS FOR LARGE LANGUAGE MODELS: A SURVEY

Zero-shot và Few-shot Prompting

Prompt:

```
Instruction: Solve the problem clearly. Here are some examples:
Problem: A store offers a 30% discount on a $200 item. What's the price after discount?
Step 1: Calculate 30% of 200 = 0.3 * 200 = 60
Step 2: Subtract from 200 → 200 - 60 = 140
Answer: $140

Problem: If a car travels 60 miles in 1.5 hours, what is its average speed in miles per hour?
Step 1: Use the formula: speed = distance / time
Step 2: speed = 60 / 1.5 = 40
Answer: 40 miles per hour

Problem: John has 3 boxes. Each box contains 12 apples. He gives away 10 apples. How many apples does he have left?
Step 1: Total apples = 3 * 12 = 36
Step 2: Apples left = 36 - 10 = 26
Answer: 26 apples
Now solve this problem:
Problem: Julie is reading a 120-page book. Yesterday, she was able to read 12 pages and today, she read twice as many pag

Final Answer:
```

PROMPT DESIGN AND ENGINEERING: INTRODUCTION AND ADVANCED METHODS

Chain-of-Thought

```
Q: <repeat_question>
A: Let's think step by step. <give_reasoning> Therefore, the answer is <final_answer>."
```

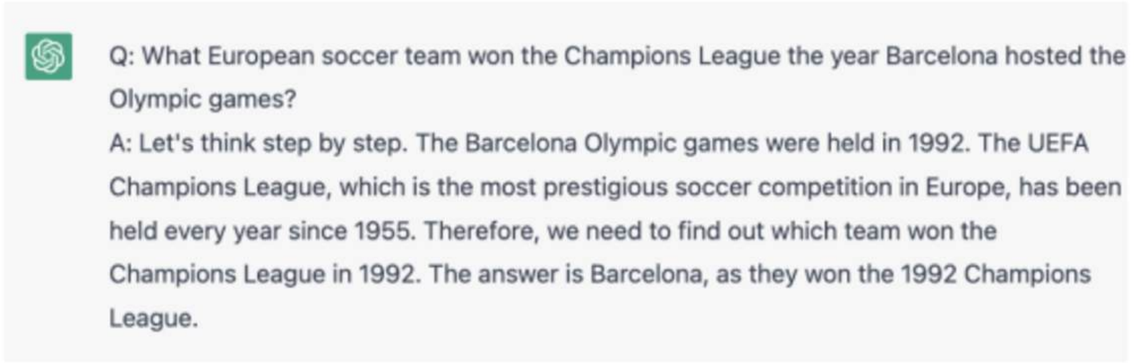


Figure 4: Chain of thought prompting example

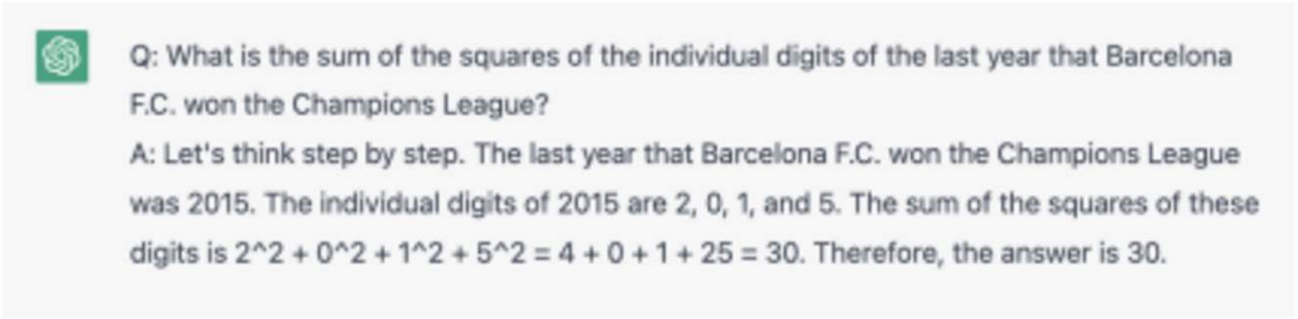
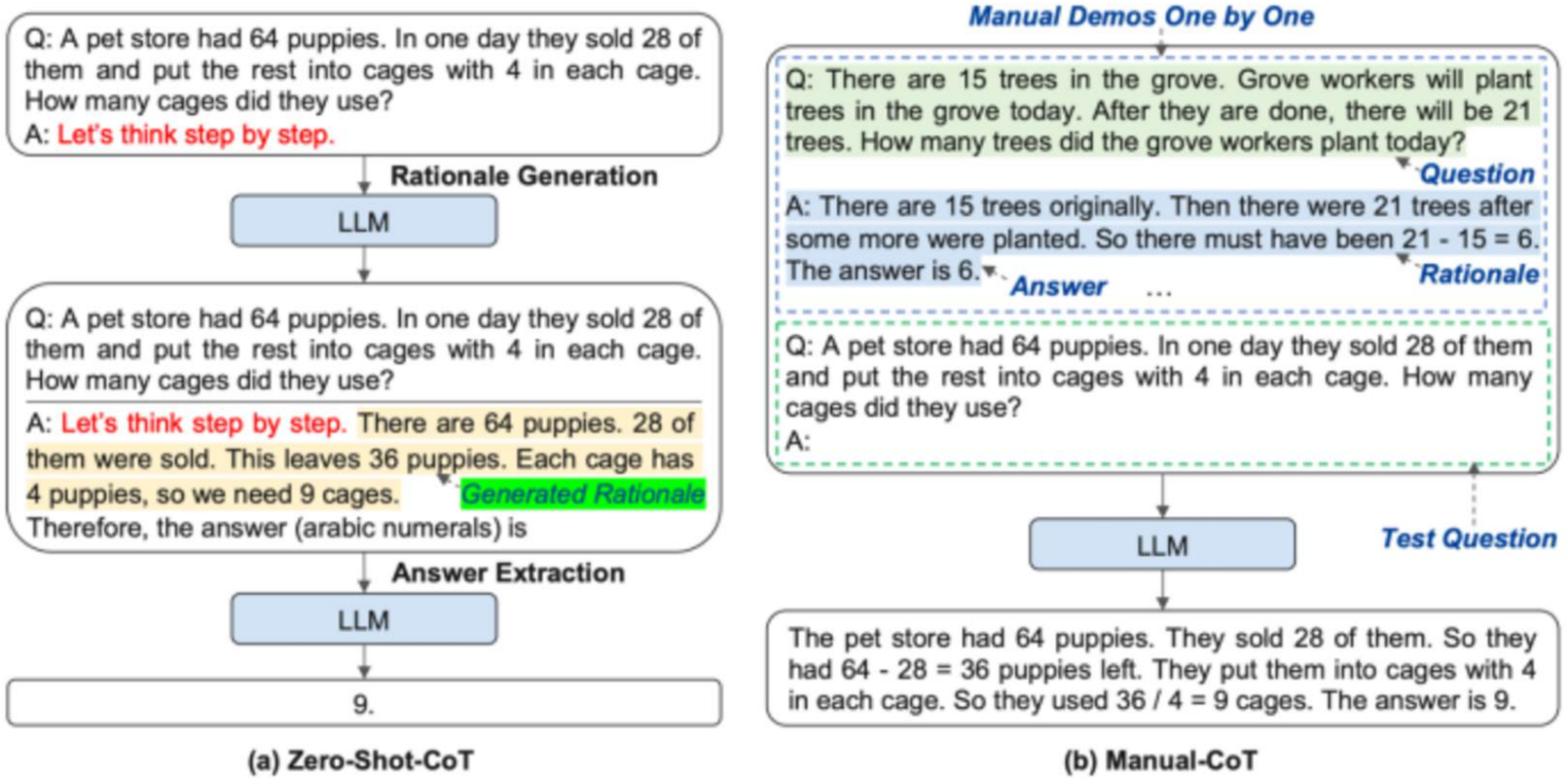


Figure 5: Chain of thought prompting example

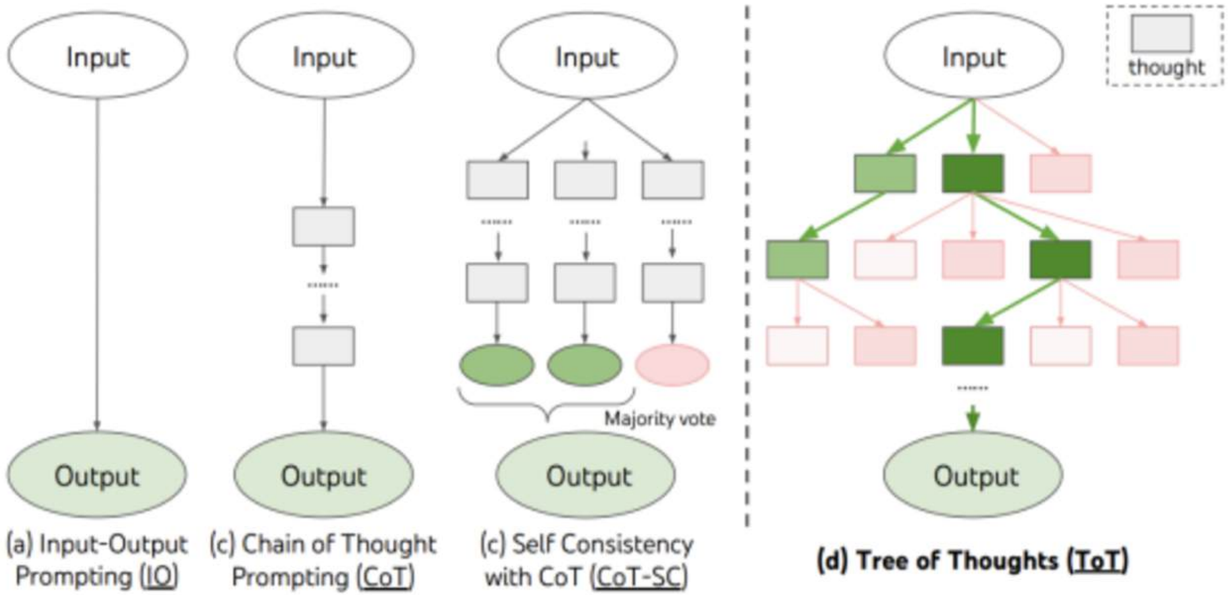
PROMPT DESIGN AND ENGINEERING: INTRODUCTION AND ADVANCED METHODS

Chain-of-Thought



PROMPT DESIGN AND ENGINEERING: INTRODUCTION AND ADVANCED METHODS

Tree-of-thought



PROMPT DESIGN AND ENGINEERING: INTRODUCTION AND ADVANCED METHODS

Các tips và tricks thiết kế prompt

```
"Are mRNA vaccines safe? Answer only using reliable sources and cite those sources. "
```

```
"Write a poem describing a beautify day <|endofprompt|>. It was a beautiful winter day"
```

Note in the result in figure 7 how the paragraph continues from the last sentence in the "prompt".

Being forceful

```
Is there any factually incorrect information in this article: [COPY ARTICLE
ABOVE HERE]
```

Generate different opinions

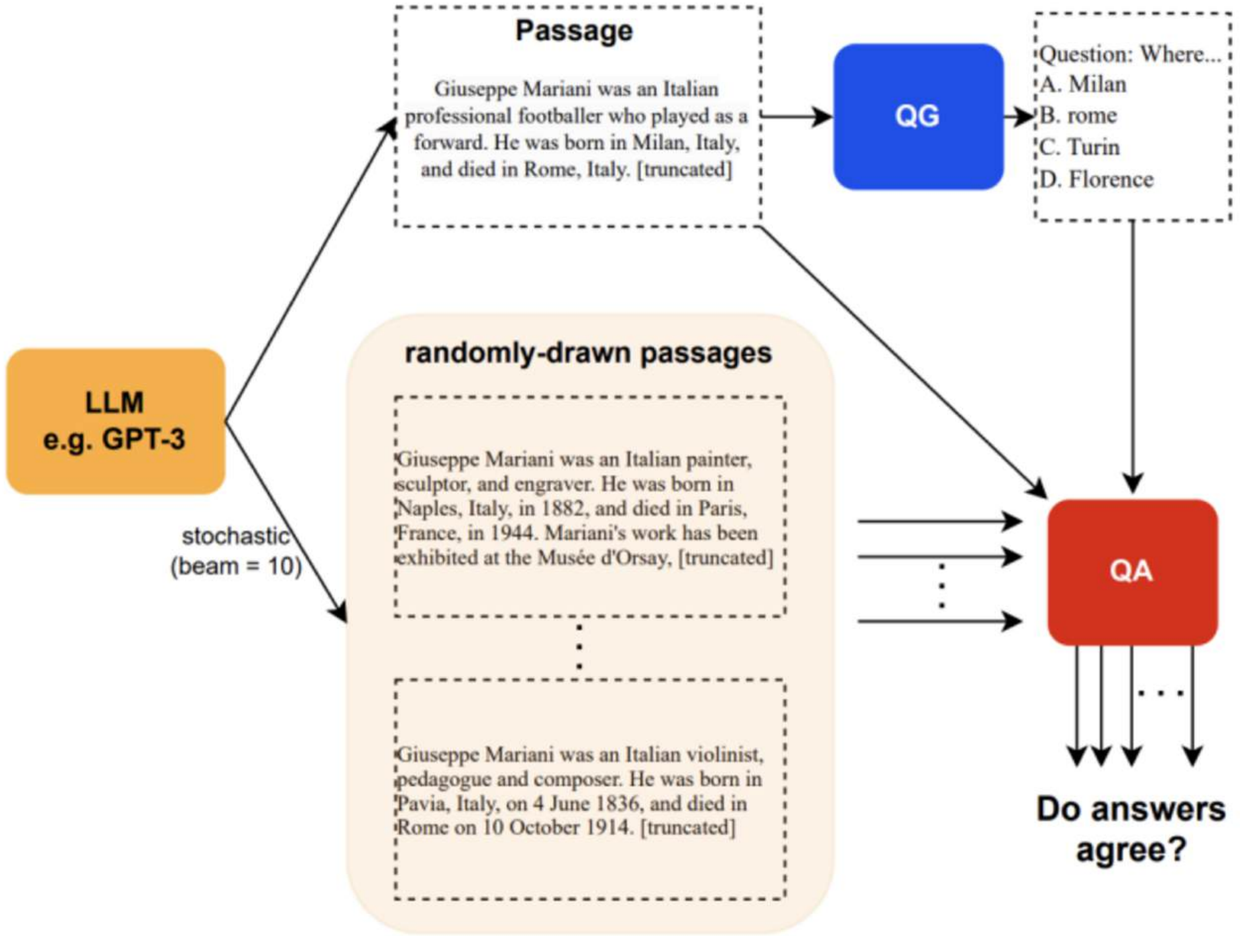
PROMPT DESIGN AND ENGINEERING: INTRODUCTION AND ADVANCED METHODS

Automatic Multi-step Reasoning and Tool-use (ART)

```
def few_shots_CoT_ART(query, k = 3):
    examples = find_top_k_tasks(query, k)
    prompt = build_prompt(examples, query)
    inputs = tokenizer(few_shots_CoT(math[i]['input']), return_tensors='pt').to('cuda')
    input_len = inputs["input_ids"].shape[1]
    output = generate_output(type=None,input = inputs)
    new_tokens = output[0][input_len:]
    answer= tokenizer.decode(new_tokens, skip_special_tokens=True).strip()
    return answer
```

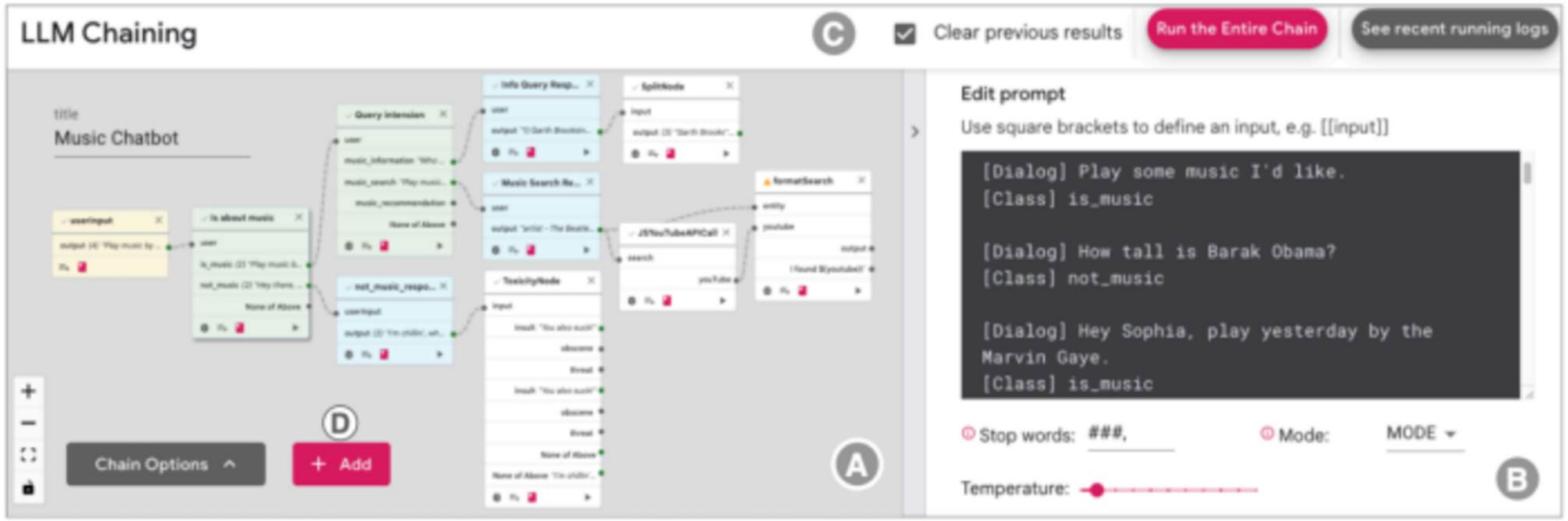
PROMPT DESIGN AND ENGINEERING: INTRODUCTION AND ADVANCED METHODS

Reflection



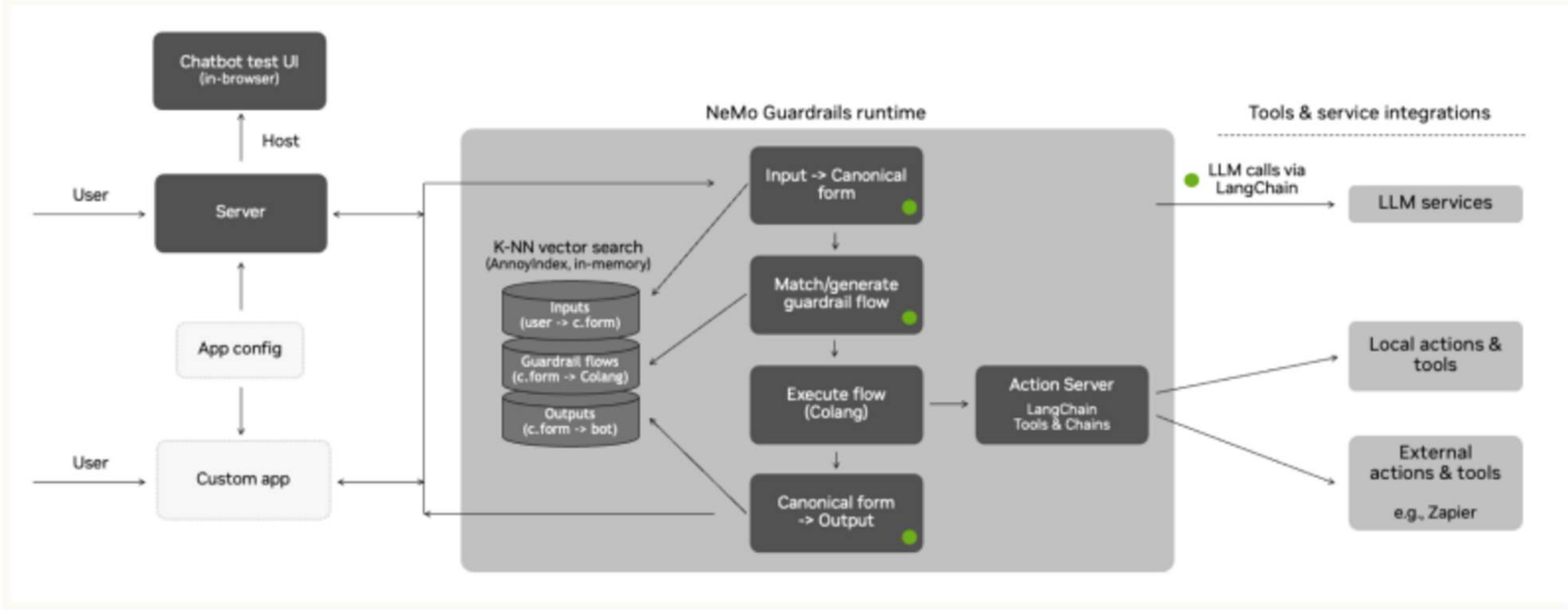
PROMPT DESIGN AND ENGINEERING: INTRODUCTION AND ADVANCED METHODS

Expert prompting



PROMPT DESIGN AND ENGINEERING: INTRODUCTION AND ADVANCED METHODS

Streamlining Complex Tasks with Chains



PROMPT DESIGN AND ENGINEERING: INTRODUCTION AND ADVANCED METHODS

Guiding LLM Outputs with Rails

- **Topical Rails**: Designed to keep the LLM focused on a specified subject or domain, preventing digression or the inclusion of irrelevant information.
- **Fact-Checking Rails**: Aim to reduce the propagation of inaccuracies by guiding the LLM towards evidence-based responses and discouraging speculative or unverified claims.
- **Jailbreaking Rails**: Established to deter the LLM from producing outputs that circumvent its operational constraints or ethical guidelines, safeguarding against misuse or harmful content generation.

PROMPT DESIGN AND ENGINEERING: INTRODUCTION AND ADVANCED METHODS

Streamlining Prompt Design with Automatic Prompt Engineering

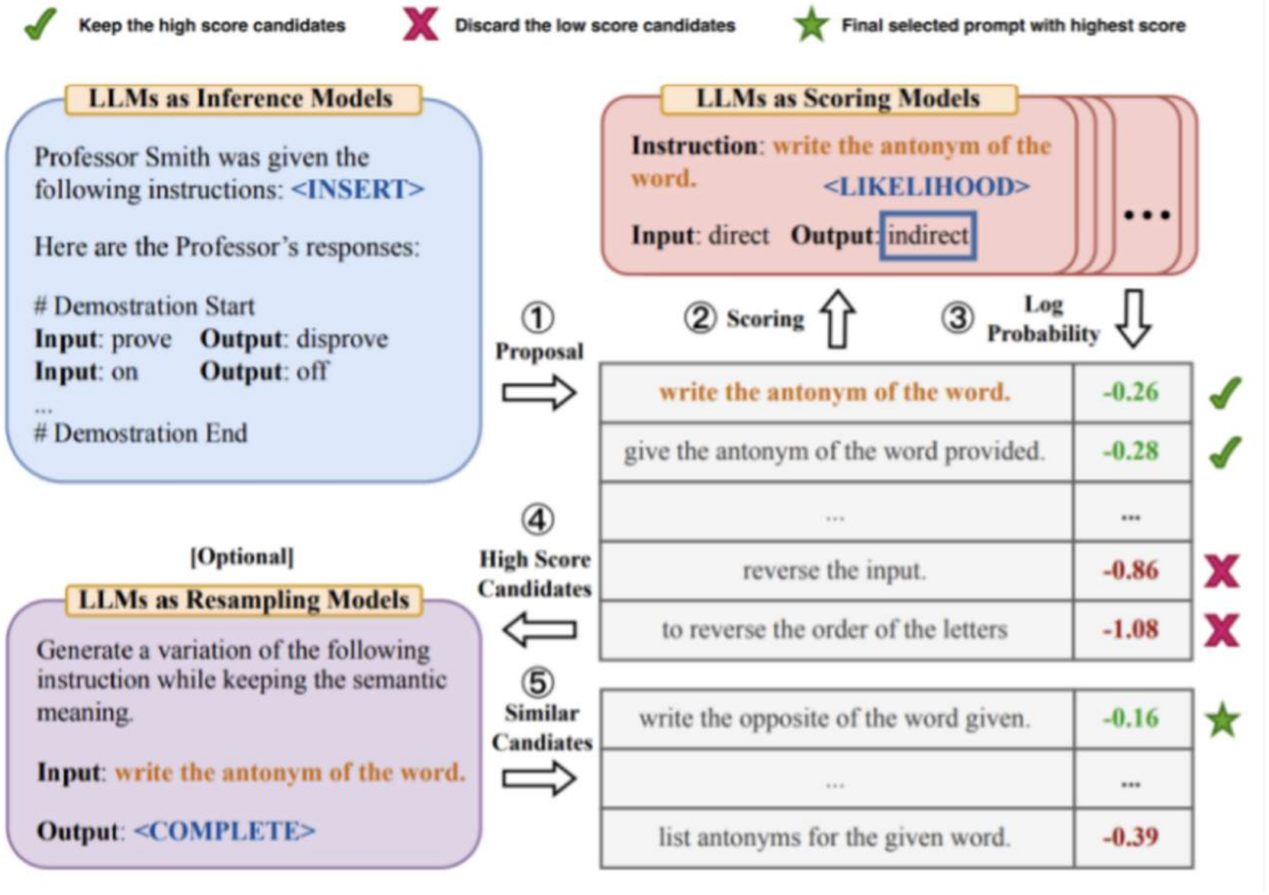
- **Prompt Generation**: Initially, the LLM produces a variety of prompts tailored to a specific task, leveraging its vast linguistic database and contextual understanding.

- **Prompt Scoring**: Subsequently, these prompts undergo a rigorous evaluation phase, where they are scored against key metrics such as clarity, specificity, and their potential to drive the desired outcome, ensuring that only the most effective prompts are selected for refinement.

- **Refinement and Iteration**: The refinement process involves tweaking and adjusting prompts based on their scores, with the aim of enhancing their alignment with the task requirements. This iterative process fosters continuous improvement in prompt quality.

PROMPT DESIGN AND ENGINEERING: INTRODUCTION AND ADVANCED METHODS

Streamlining Prompt Design with Automatic Prompt Engineering



PROMPT DESIGN AND ENGINEERING: INTRODUCTION AND ADVANCED METHODS

DEMO

THANK YOU FOR YOUR LISTENING