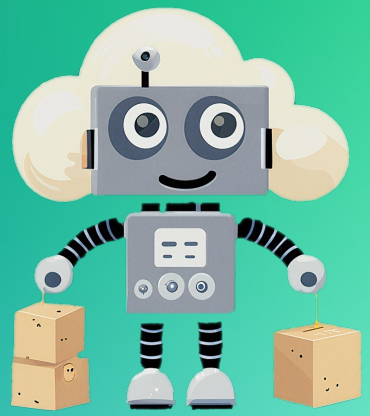
Prompt Engineering: From Naïve to Automated

Vladimir Palkhouski

Sr. Solutions Architect
AWS, FSI Germany



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

LLMs Evolution

… from the groundwork of 2019

"Language models begin to learn NLP tasks without any explicit supervision"

"It is possible to induce the desired behavior by conditioning the model on a context"

Radford et all. 2019: Language Models are Unsupervised Multitask Learners (GPT-2)

to sophisticated capabilities in 2025

State-of-the-art NLP tasks accuracy

Authentic brand/persona voice reflection

Agency with multi-step planning/reflection

Tools/function calling

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Prompt Engineering / In-Context Learning

What is a prompt?

specific set of inputs that guide LLMs to generate an appropriate response for a given task or instruction.

https://docs.aws.amazon.com/bedrock/latest/userguide/what-is-a-prompt.html

What is prompt engineering?

the practice of crafting and optimizing input prompts by selecting appropriate words, phrases, sentences, punctuation, and separator characters.

https://docs.aws.amazon.com/bedrock/latest/userguide/what-is-prompt-engineering.html

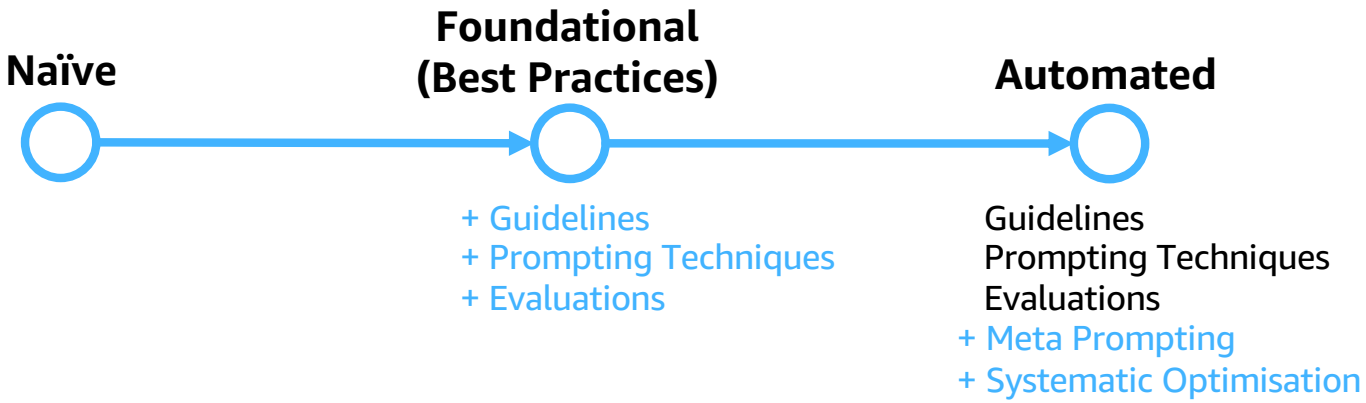
Who was the president of the U.S.A. in 1930?

Classify the following text: 'I would like to request a full reimbursement'

Modern rectangular frame with subtle roundness at corners. Frame material is of metal core, round-shaped lenses.

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Prompt Engineering ~~Evolution~~ Maturity Model



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Mike

AI Engineer
Works for "Digi-Bank"

all the names are fictitious
the challenges and solutions are very real

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Naïve Prompting

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

User Feedback Sentiment Analysis

Claude 3.5 Sonnet

Classify user's feedback as one of the following classes: [positive, negative, neutral]

User's feedback:
I am able to access my account again, thank you!

Based on the user's feedback "I am able to access my account again, thank you!", this would be classified as *positive*.
The feedback expresses gratitude ("thank you!") and indicates a successful resolution to their issue (regaining account access), making it clearly positive in sentiment.

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

User Feedback Sentiment Analysis

Claude 3.5 Sonnet

Classify user's feedback as one of the following classes: [positive, negative, neutral]

User's feedback:
I am not sure. The access is restored, though!

"It does not look right. Let me instruct you to do better."

Based on the u[illegible]s feedback …

…*neutral*

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

User Feedback Sentiment Analysis

Claude 3.5 Sonnet

Classify user's feedback as one of the following classes: [positive, negative, neutral].
If the feedback expresses uncertainty but the task is done, it should be positive.
If the feedback is ironical, then it is likely to be negative.
Any toxic expression also indicates the negative sentiment.
When unsure, assign a neutral sentiment.
Don't explain yourself, response with the class directly.
…

User's feedback:
I am not sure. The access is restored, though!

"Should I describe in details every edge case?"

"That's a lot of content. It is quite hard to read and maintain."

"Great!"

The predicted class is *positive.*

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

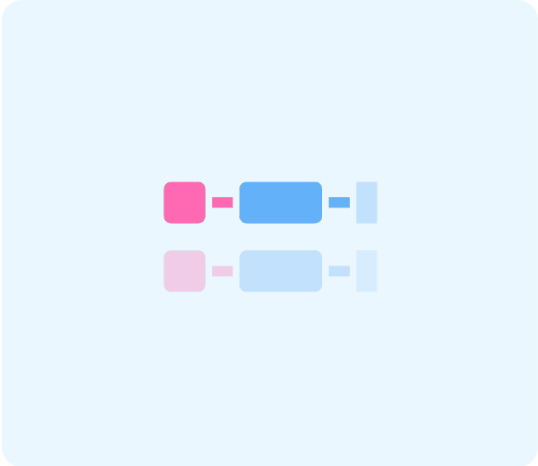
Foundational Prompting (Best Practices)

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Foundational Prompting

Guidelines / Best Practices

General recommendations from FM providers (model-family specific).



Prompting Techniques

Set of distinct patterns for inducing a particular behavior with LLMs (CoT, Self-consistency, etc.).

Evaluations

Automated or human-based methods to accessing task-performance across various metrics.

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Guidelines

Control output sampling (creativity)

Be direct, precise and specific

Have a prompt structure

Use roles / personas

Control output

Use prompt chaining for complex tasks

Use examples (few-shot prompting)

Let model think step-by-step (CoT)

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Guidelines

Control output sampling (creativity)

Be direct, precise and specific

Have a prompt structure

Use roles / personas

Control output

Use prompt chaining for complex tasks

Use examples (few-shot prompting)

Let model think step-by-step (CoT)

- Adjust **temperature**, **top_p** or **top_k** (model specific) in accordance with the task
- Low values – higher determinism, high values – better creativity
- Adjust one at a time

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Guidelines

Control output sampling (creativity)

Be direct, precise and specific

Have a prompt structure

Use roles / personas

Control output

Use prompt chaining for complex tasks

Use examples (few-shot prompting)

Let model think step-by-step (CoT)

- Provide clear and direct instructions
- **The golden rule of clear prompting:** Show your prompt to a colleague, ideally someone who has minimal context on the task, and ask them to follow the instructions. If they're confused, LLM will likely be too.

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Guidelines

Control output sampling (creativity)

Be direct, precise and specific

Have a prompt structure

Use roles / personas

Control output

Use prompt chaining for complex tasks

Use examples (few-shot prompting)

Let model think step-by-step (CoT)

- Segregate sections
- Refer to sections from instructions
- Know model's formatting preferences

Amazon Nova

- XML Tags
- Markdown

Claude

XML Tags

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Guidelines

Control output sampling (creativity)

Be direct, precise and specific

Have a prompt structure

Use roles / personas

Control output

Use prompt chaining for complex tasks

Use examples (few-shot prompting)

Let model think step-by-step (CoT)

- Personalize tone and response style
- Use system prompt for higher importance instructions

You are a *virtual tour guide* currently walking the tourists Eiffel Tower on a night tour.

Describe Eiffel Tower to your audience that covers its history, number of people visiting each year, amount of time it takes to do a full tour and why do so many people visit this place each year.

Meta

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Guidelines

Control output sampling (creativity)

Be direct, precise and specific

Have a prompt structure

Use roles / personas

Control output

Use prompt chaining for complex tasks

Use examples (few-shot prompting)

Let model think step-by-step (CoT)

- Explicitly specify output format
- Provide schema if necessary / tool config
- Pre-fill responses

Provide details about the best selling full-frame cameras in past three years.

Answer in JSON format with keys like name, brand, price and a summary.

Assistant: JSON response: ```json"

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Guidelines

Control output sampling (creativity)

Be direct, precise and specific

Have a prompt structure

Use roles / personas

Control output

Use prompt chaining for complex tasks

Use examples (few-shot prompting)

Let model think step-by-step (CoT)

- Break down complex tasks into multiple prompts
- Chain them together
- Complex tasks examples
  - Self-correcting research summary
  - Multi-step data processing
  - Data driven decision making

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Guidelines

Control output sampling (creativity)

Be direct, precise and specific

Have a prompt structure

Use roles / personas

Control output

Use prompt chaining for complex tasks

Use examples (few-shot prompting)

Let model think step-by-step (CoT)

- Select diverse, relevant examples
- Update to dynamic few-shot when reaching limits

Classify the following request into one of the classes: [class_1, class_2, class_3].

Examples:
- example_1: class_1
- example_2: class_2
- example_3: class_3

Request: I can not login to the banking app.

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Guidelines

Control output sampling (creativity)

Be direct, precise and specific

Have a prompt structure

Use roles / personas

Control output

Use prompt chaining for complex tasks

Use examples (few-shot prompting)

Let model think step-by-step (CoT)

- Think of models as ***System 1 Thinkers*** by default: fast, unconscious, erroneous
- Evoke ***System 2 Thinking*** by asking the model to spell reasoning out (slow, conscious, reliable)

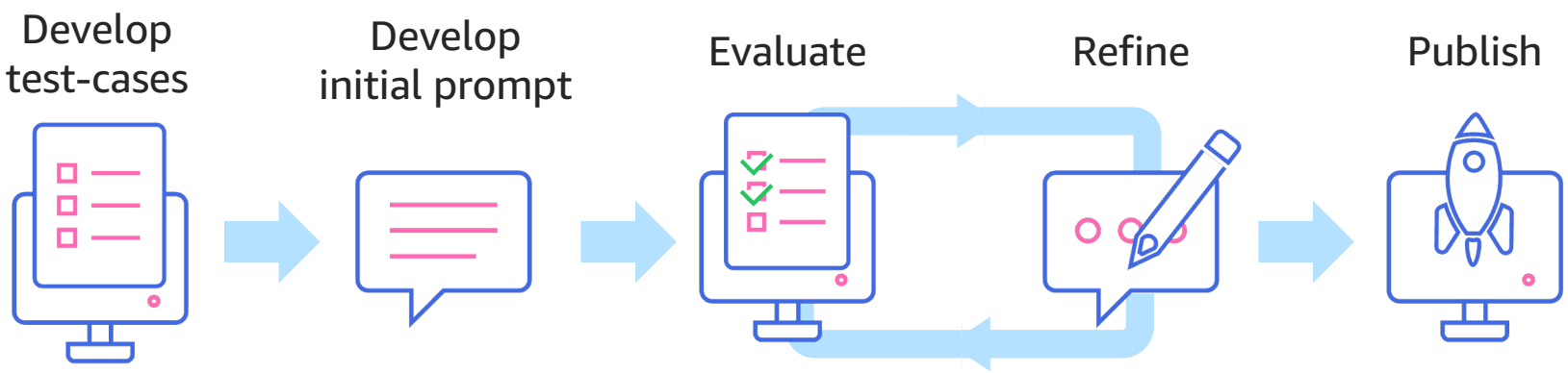
[Complex Puzzle]

Think step by step.

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Prompt Engineering Philosophy

Empirical science: adapt, evaluate, iterate



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

User Feedback Sentiment Analysis

Claude 3.5 Sonnet

You are a customer support assistant. Your job is to analyze customers' responses. Use examples from the <examples> section.

<examples>
- *I am not sure, but it works: positive*
- *Looks like another genius innovation!: negative*
- *My 5-year-old could do better: negative*

</examples>

Classify user's feedback as one of the following classes as instructed above. ***Respond directly with a class name. Skip the preamble.***

User's feedback: *I am not sure. The access is restored, though!*

"Well-defined structure is easier to maintain."

"It is easy to steer the model with examples."

"It should keep the output crisp and to the point."

"I am going to evaluate the prompt against the entire testing dataset."

Positive

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

“ Ready for the next endeavor

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Customer Request Classification

Meta Llama-2-7b-chat

You are a customer support assistant. Your job is to analyze customers' requests. Use examples from the <examples> section.

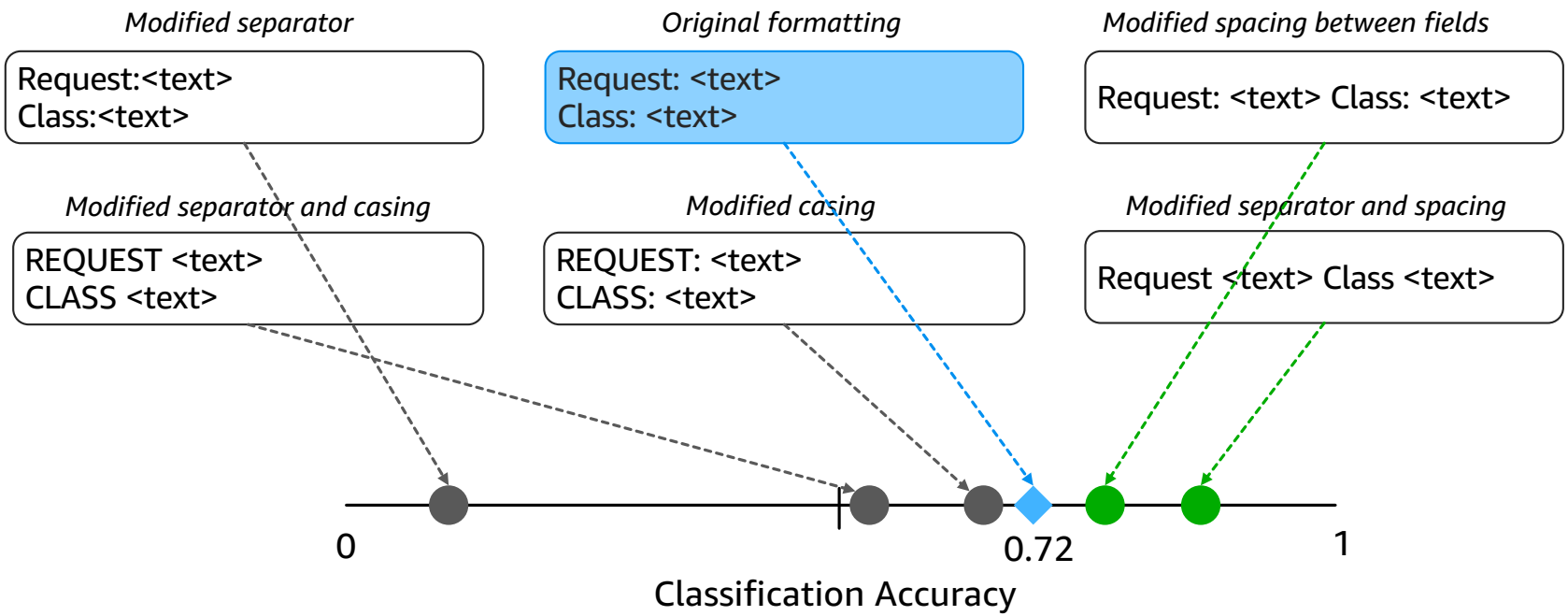
<examples>
- I can not login to my account: access_problem
- my pin is no longer working: pin_forgotten
- I see a strange transaction: fraud
</examples>

Based on the information above, classify the following request: *How to login without an app?*

access_problem

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Prompt Formatting Impact on Task-performance



* original paper: https://arxiv.org/pdf/2310.11324

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Experiment: Answer The Question

Information perception by human beings

Question

Context
Paris has a population of 2 milion people,

Question
How many people live n Paris?

Perception Blindness

Context
paris has a population of 2 million people

Question
how many people live in paris

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

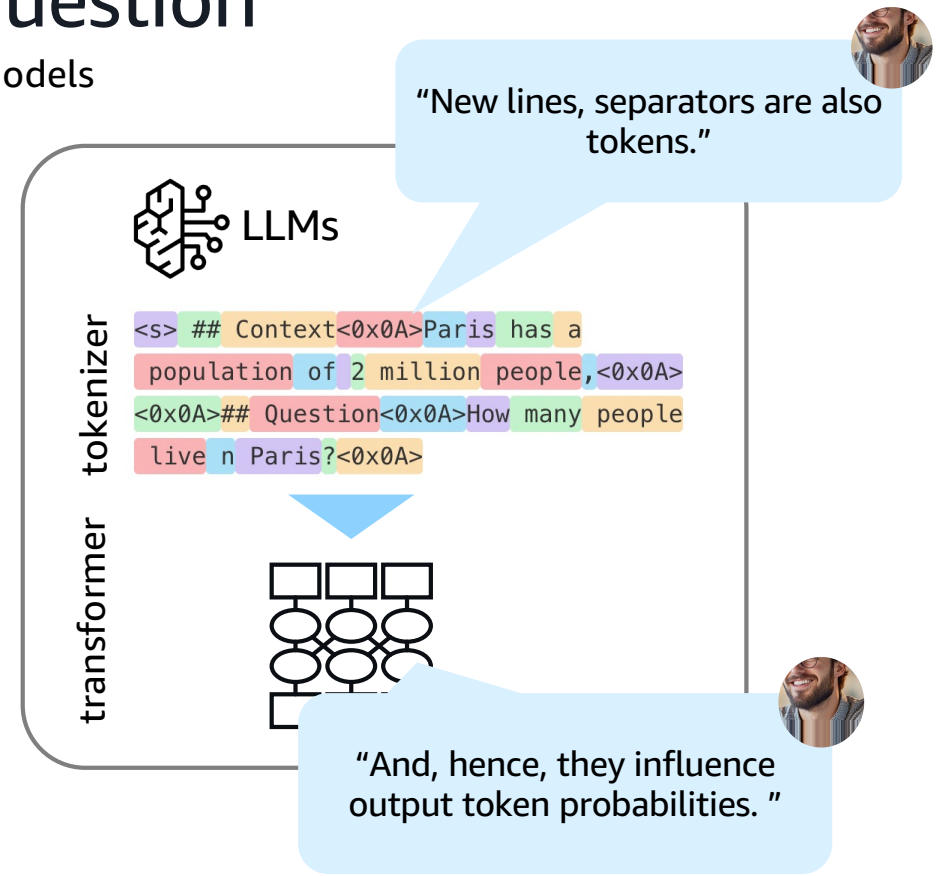
Experiment: Answer The Question

Information perception by transformer-based NLP models

Question

Context
Paris has a population of 2 milion people,

Question
How many people live n Paris?



"New lines, separators are also tokens."

"And, hence, they influence output token probabilities. "

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Misunderstandings Between Humans and LLM

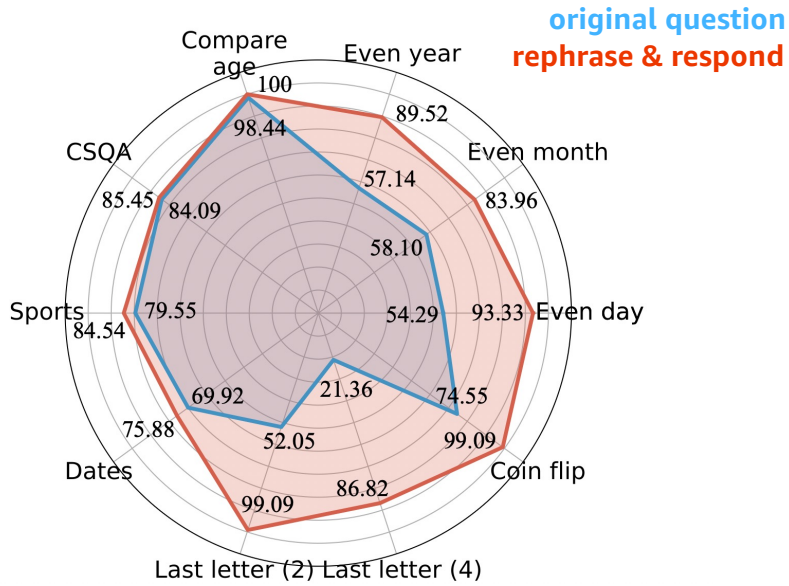
Accuracy improvement of GPT-4 across 10 tasks with Rephrase and Respond (RaR)

{question}

vs

{question}

Rephrase and expand the question, and respond.



Accuracy comparison

* original paper: https://arxiv.org/pdf/2311.04205

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

> LLMs are demanding and peculiar masters of communication

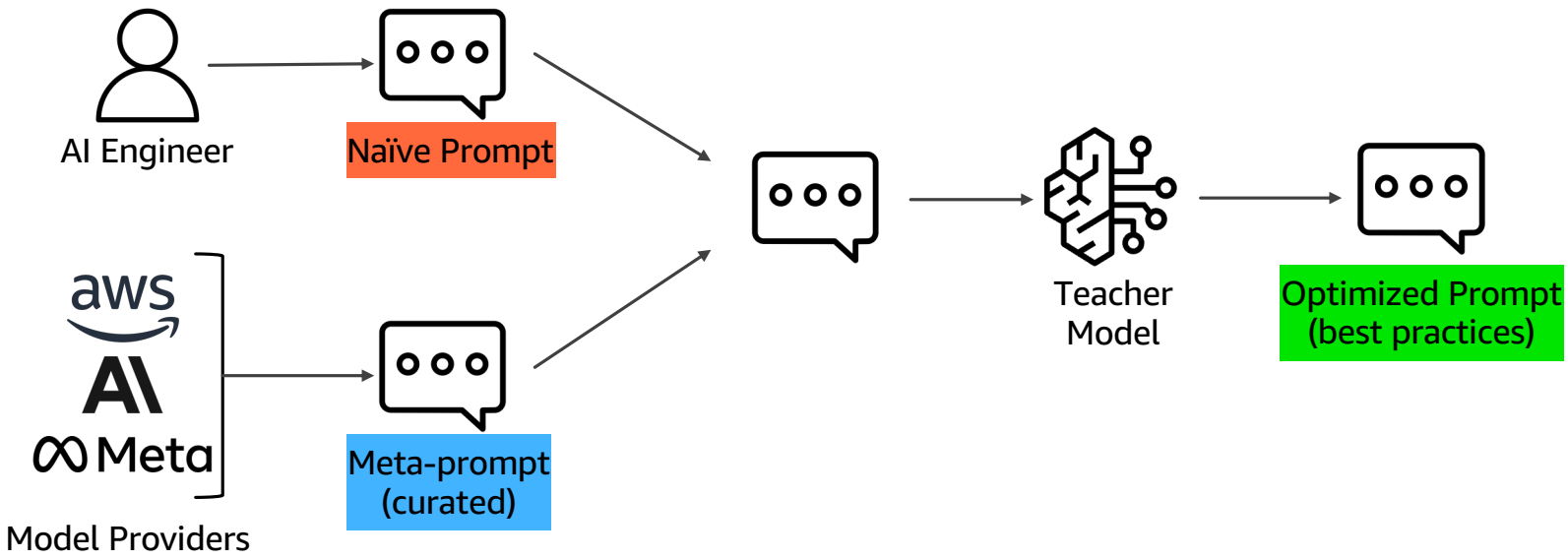
© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Automated Prompt Engineering

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Meta-prompting at a Glance

Let a model propose a better prompt



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Meta-prompting Example

Claude 3.5 Sonnet

Naïve Prompt

Draft an email responding to an online banking customer request and offer a resolution.

Meta-prompt

Today you will be writing instructions to an eager, helpful, but inexperienced and unworldly AI assistant who needs careful instruction…

Claude 3.5 Sonnet

Meta-prompt example: https://github.com/anthropics/anthropic-cookbook/blob/main/misc/metaprompt.ipynb

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Meta-prompting Example

Claude 3.5 Sonnet

Optimized Prompt

You are an AI assistant working for an online banking customer service department. Your task is to draft an email response to a customer request, offering a resolution based on the bank's policies. Follow these instructions carefully:

First, review the bank's policies:
<bank_policies>
{{BANK_POLICIES}}
</bank_policies>

Now, here is the customer's request:
<customer_request>
{{CUSTOMER_REQUEST}}
</customer_request>

Analyze the customer's request:
1. Identify the main issue or concern raised by the customer.
2. Determine which bank policies are relevant to this request.
3. Consider possible solutions that align with the bank's policies.

Draft your response:
1. Start with a polite greeting and thank the customer for their inquiry.

Structure

Context and Task

Multi-step Instructions

- First, review the bank's policies …
- Analyze the customer's request …
- Draft your response …
- Format your email …
- Write your complete email response inside <email_response> tags …

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Meta-prompting Example

Claude 3.5 Sonnet

Optimized Prompt

You are an AI assistant working for an online banking customer service department. Your task is to draft an email response to a customer request, offering a resolution based on the bank's policies. Follow these instructions carefully:

First, review the bank's policies:
<bank_policies>
{{BANK_POLICIES}}
</bank_policies>

Now, here is the customer's request:
<customer_request>
{{CUSTOMER_REQUEST}}
</customer_request>

Analyze the customer's request:
1. Identify the main issue or concern raised by the customer.
2. Determine which bank policies are relevant to this request.
3. Consider possible solutions that align with the bank's policies.

Draft your response:
1. Start with a polite greeting and thank the customer for their inquiry.

Formatting guidelines

First, review the bank's policies:
<bank_policies>
{{BANK_POLICIES}}
</bank_policies>

XML tagging

new lines

Now, here is the customer's request:
<customer_request>
{{CUSTOMER_REQUEST}}
</customer_request>

variables

…
Write your complete email response inside <email_response> tags.

output handling

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Meta-prompting Example

Claude 3.5 Sonnet

Optimized Prompt

You are an AI assistant working for an online banking customer service department. Your task is to draft an email response to a customer request, offering a resolution based on the bank's policies. Follow these instructions carefully:

First, review the bank's policies:
<bank_policies>
{{BANK_POLICIES}}
</bank_policies>

Now, here is the customer's request:
<customer_request>
{{CUSTOMER_REQUEST}}
</customer_request>

Analyze the customer's request:
1. Identify the main issue or concern raised by the customer.
2. Determine which bank policies are relevant to this request.
3. Consider possible solutions that align with the bank's policies.

Draft your response:
1. Start with a polite greeting and thank the customer for their inquiry.

One-shot example

<customer_request>
Hello, I recently noticed an unauthorized charge …
</customer_request>

<response>
Subject: Unauthorized Credit Card Charge

Dear [Customer Name],
Thank you for bringing this issue to our attention …
Sincerely,
[Your name/role]
</response>

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Meta-prompting

Lessons learned

- Helps with a ”blank page” problem
- Apply to learn ”best practices” quickly
- “Shake” the optimization plateau

Mind the gap

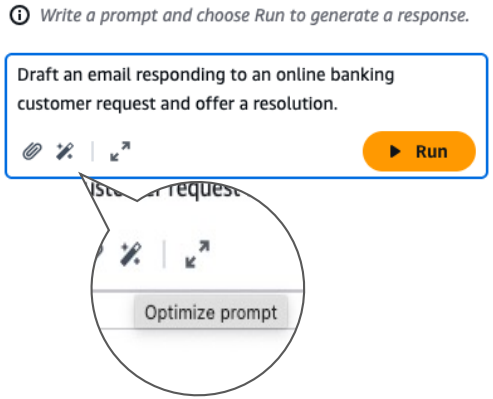
- Meta-prompts are also prompts
- Meta-prompts are model-family specific
- Optimal performance is not guaranteed, so …
- Adapt, evaluate, iterate

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Meta-prompting

How to use on AWS

Amazon Bedrock Playground Optimize Prompt



Amazon Bedrock API

```
POST /optimize-prompt HTTP/1.1
Content-type: application/json

{
   "input": { ... },
   "targetModelId": "string”
}
```

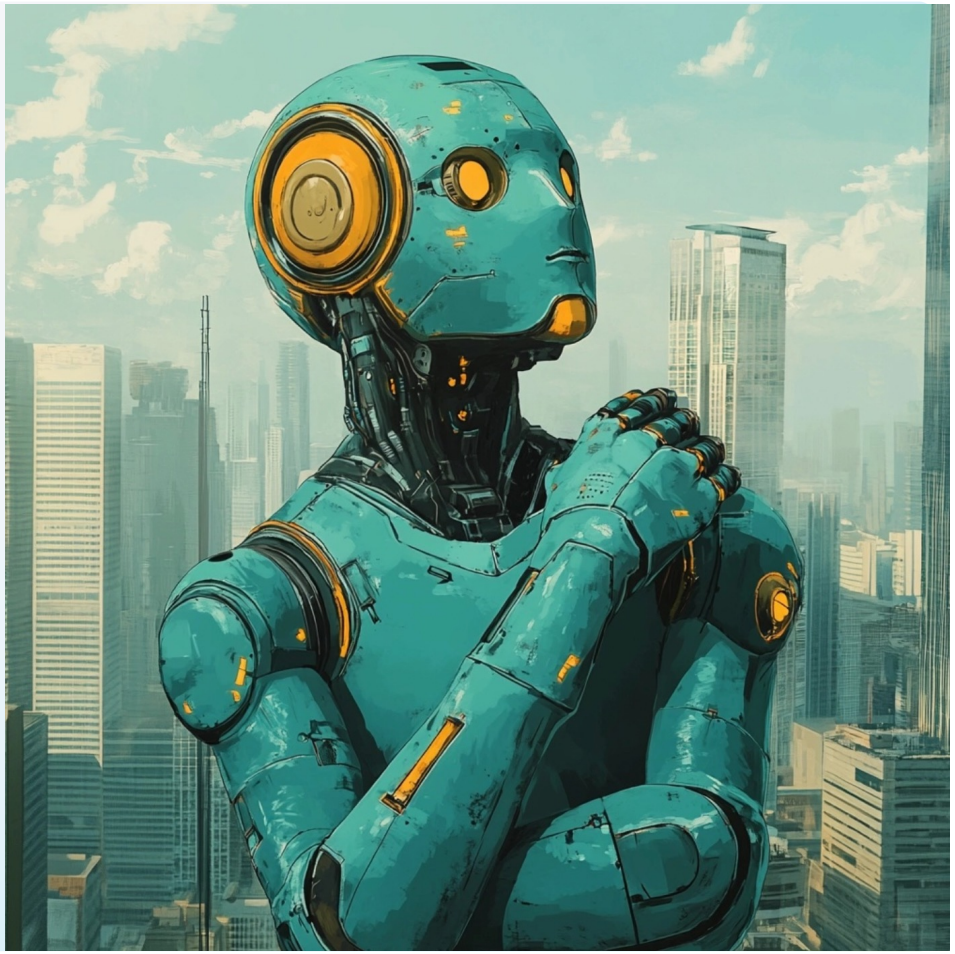
https://docs.aws.amazon.com/bedrock/latest/APIReference/API_agent-runtime_OptimizePrompt.html

Directly with an LLM and a meta-prompt (JupyterLab)

https://github.com/anthropics/anthropic-cookbook/blob/main/misc/metaprompt.ipynb

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

“ Prompting is powerful
…and nuanced



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Building State-of-the-art Prompts

You are a banking consultant. Your task is to understand the online-banking customer requests. Make sure you are following the guidelines below.

<guidelines>
{{guidelines}}
</guidelines>

<examples>
- I can not login to my account: access_problem
- my pin is no longer working: pin_forgotten
- I see a strange transaction: fraud
</examples>

Based on the information above, classify the following request: {{request}}

"Is it the best prompt I can get?"

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

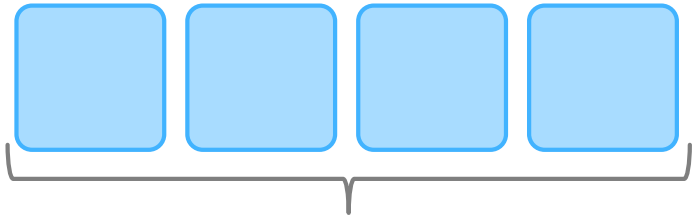
Building State-of-the-art Prompts

You are a banking consultant. Your task is to understand the online-banking customer requests. Make sure you are following the guidelines below.

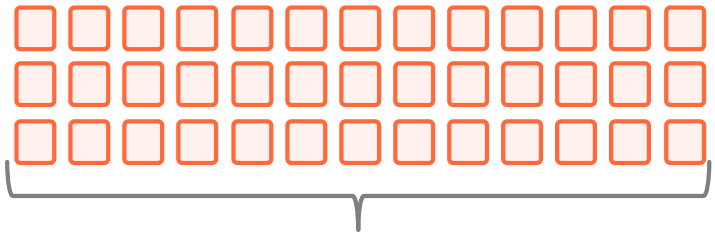
<guidelines>
{{guidelines}}
</guidelines>

<examples>
- I can not login to my account: access_problem
- my pin is no longer working: pin_forgotten
- I see a strange transaction: fraud
</examples>

Based on the information above, classify the following request: {{request}}



LLM generated instructions variants (meta-prompting)



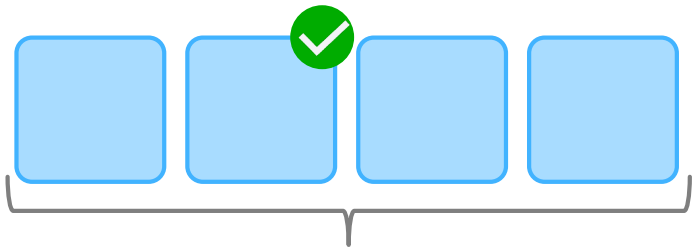
Sampled or synthesized few-shot candidates

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Building State-of-the-art Prompts

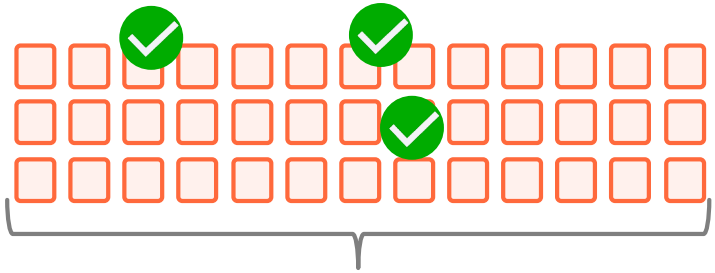
Optimal Prompt

Instructions Optimization



LLM generated instructions variants (meta-prompting)

Few-Shot Optimi



Sampled or synthesized few-shot candidates

"It is a discrete parameters optimisation problem!"

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Gradient-free Automated Prompt Optimization

Optimal program

$$\Phi^* = \arg\max_{\mathbf{V} \mapsto S} \frac{1}{|\mathcal{D}|} \sum_{(x,x') \in \mathcal{D}} \mu(\Phi_{\mathbf{V} \mapsto S}(x), x')$$

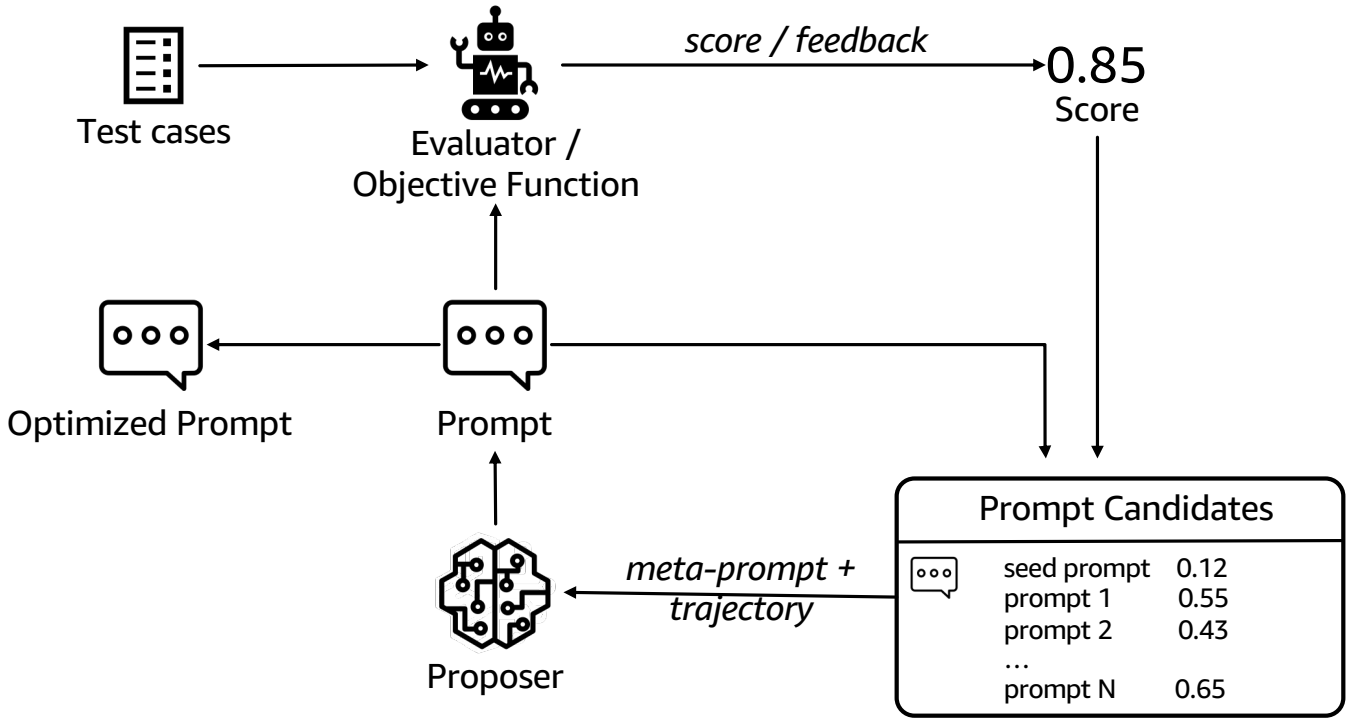
Program Definition

Test-cases

Metric Function

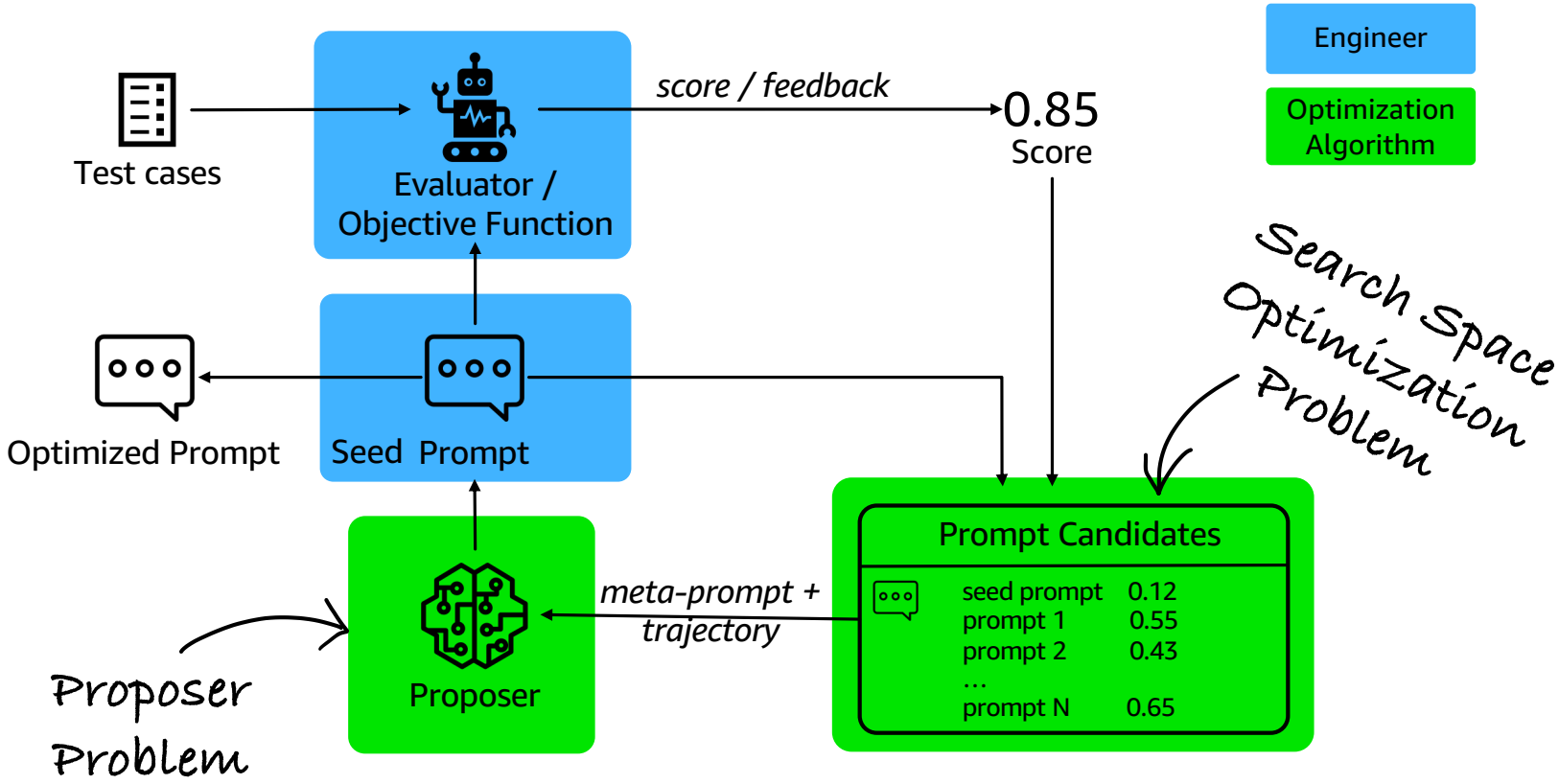
© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Gradient-free Automated Prompt Optimization



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Gradient-free Automated Prompt Optimization



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

GEPA (Genetic-Pareto): Reflective Prompt Evolution

For compound AI systems

LLMs adaptation with RL

- Thousands of rollouts are required
- Tool calls become cost and time prohibitive
- Fine-tuning of the best performing LLMs is not always possible

LLMs adaptation with natural feedback

- LLMs' traces / trajectories are highly interpretable (by LLMs as well)
- Algorithms can deliberately learn in natural language

https://arxiv.org/pdf/2507.19457

aws

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

GEPA (Genetic-Pareto): Reflective Prompt Evolution

For compound AI systems

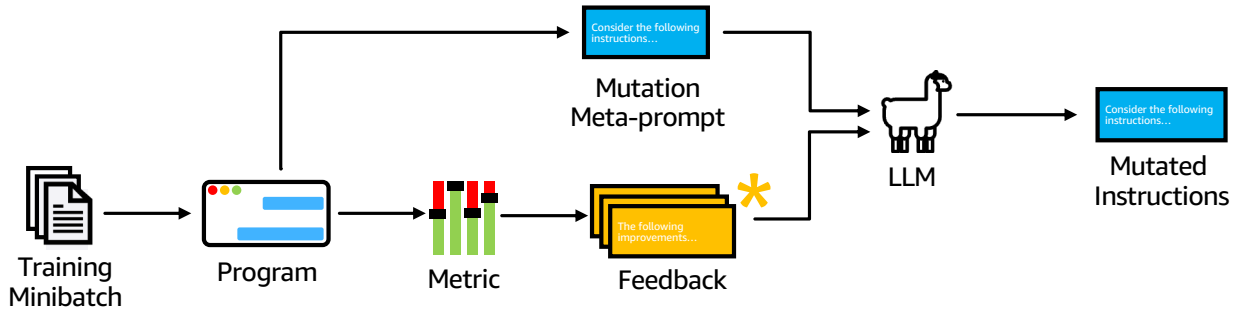
- iteratively mutates every prompt in light of natural language feedback
- candidate prompts accumulate high-level lessons from observations and LLM feedback
- GEPA stochastically explores the top-performing prompts, diversifying strategies and encouraging robust generalization

https://arxiv.org/pdf/2507.19457

aws

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

GEPA – Prompt Mutation (The Proposer Problem)



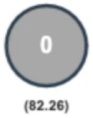
https://arxiv.org/pdf/2507.19457

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Genetic-Pareto (GEPA) – Greedy Candidate Selection

© 2025,

Genetic-Pareto (GEPA) – Pareto Front Candidate Selection



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Automated Prompt Optimization with GEPA and Amazon Bedrock Foundation Models - DEMO

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Automated Prompt Optimization (APO)

Lessons learned

Modern APO algorithms can outperform fine-tuning at a fraction of training cost (GEPA)

APO is practically interesting for composite multi-step tasks

There is no clear winner between few-shot and instructions optimization

Mind the gap

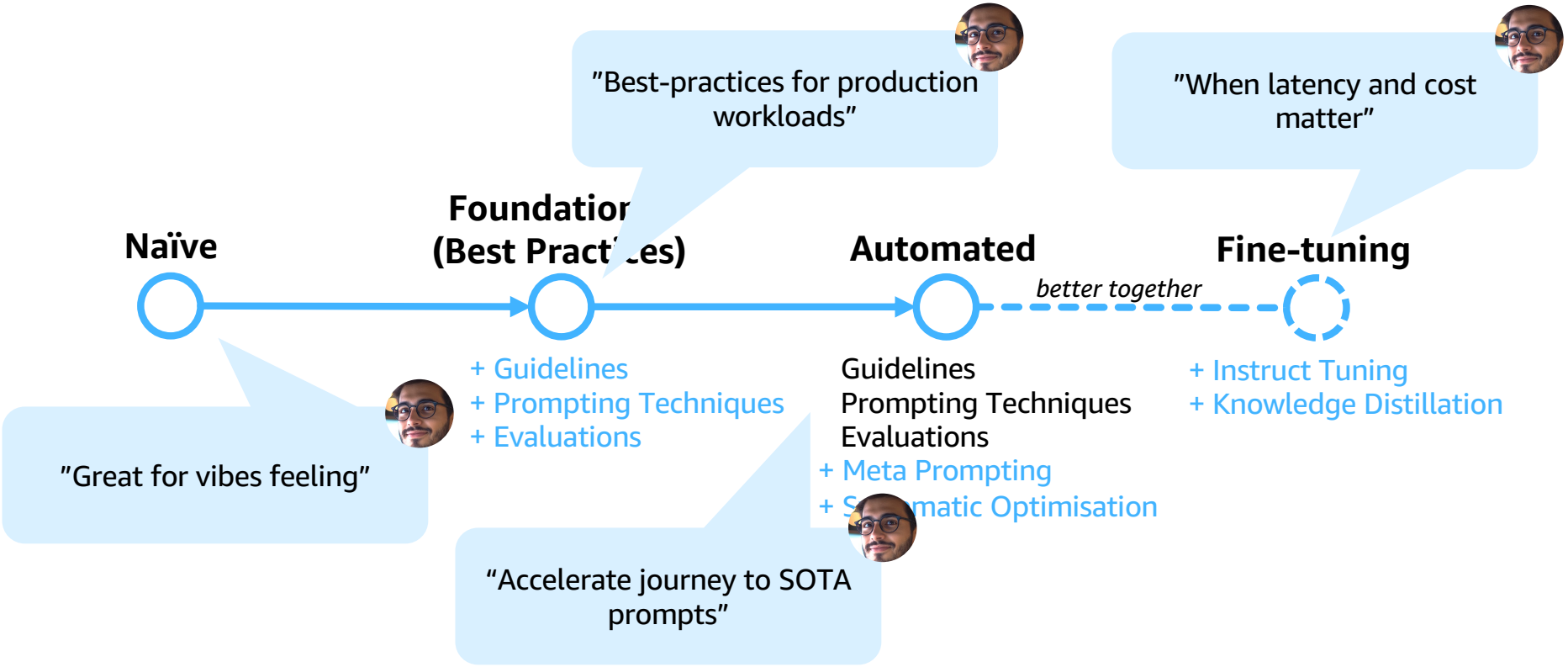
APO is a rapidly developing field of R&D (commercial and open source solutions have arrived, though)

Meta-prompting is not (yet) leveraged to the fullest extend

Fine-tuning or alternative inference architecture might be a better choice

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Prompt Engineering Maturity Model

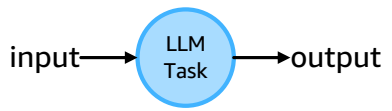


© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

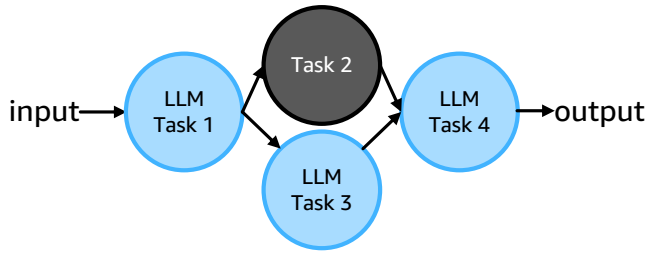
“ Ready to build your first agent?

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

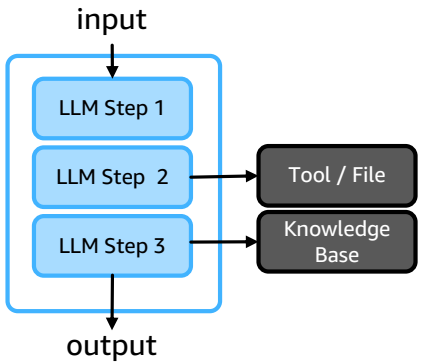
Types of AI Systems



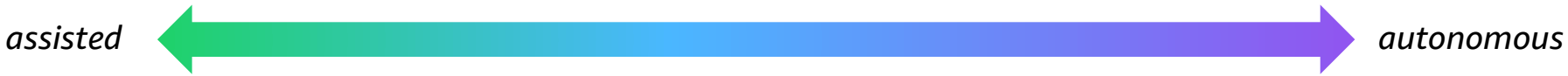
AI Tasks



AI Assisted Workflows

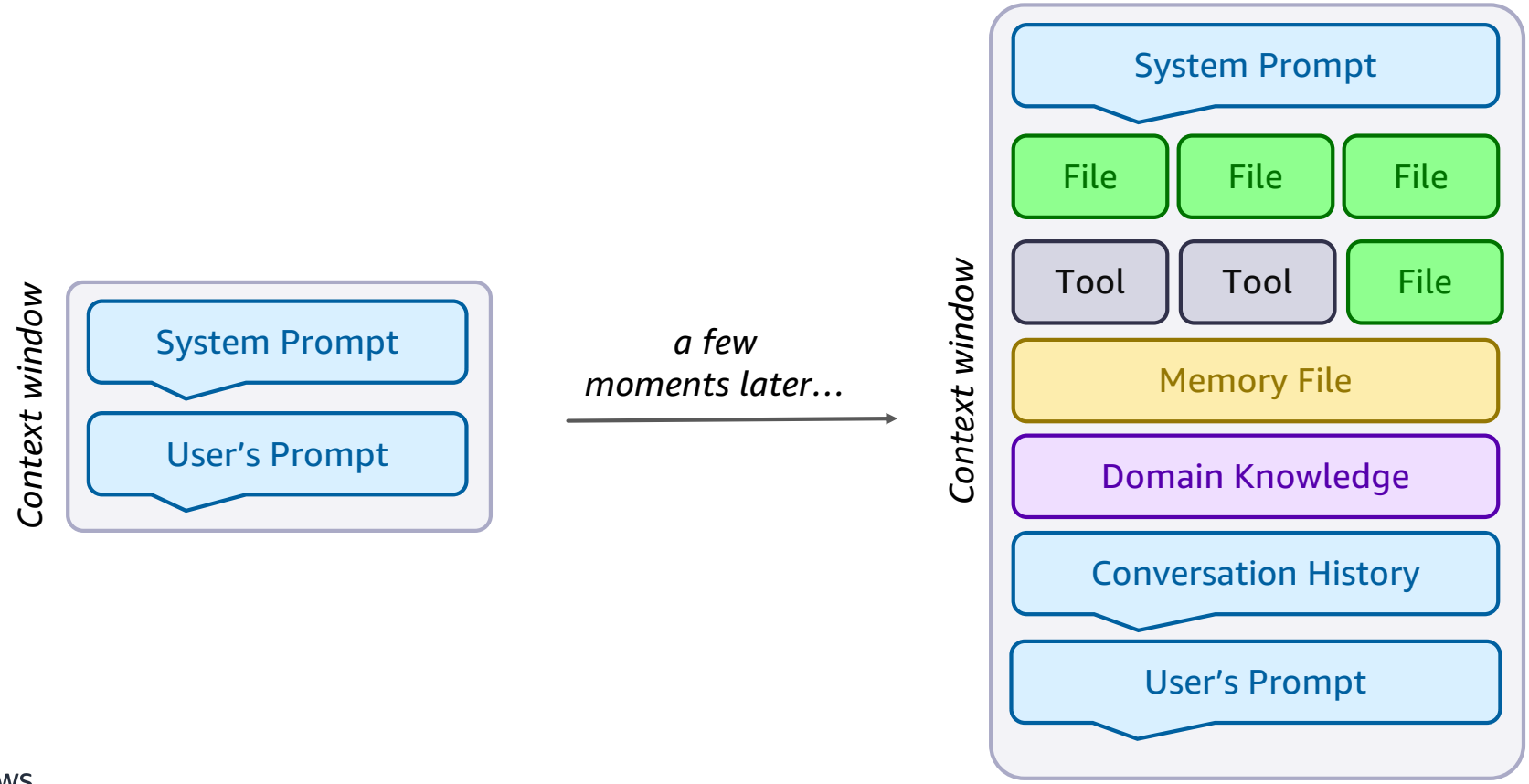


AI Agents



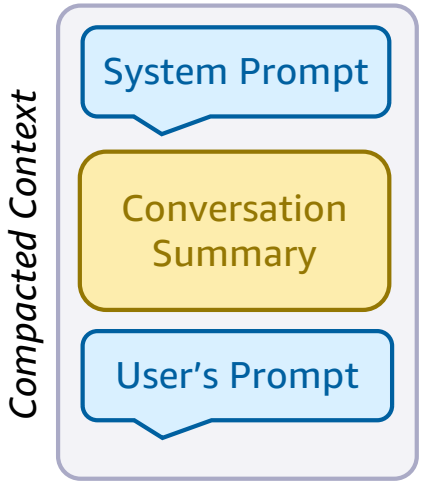
© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

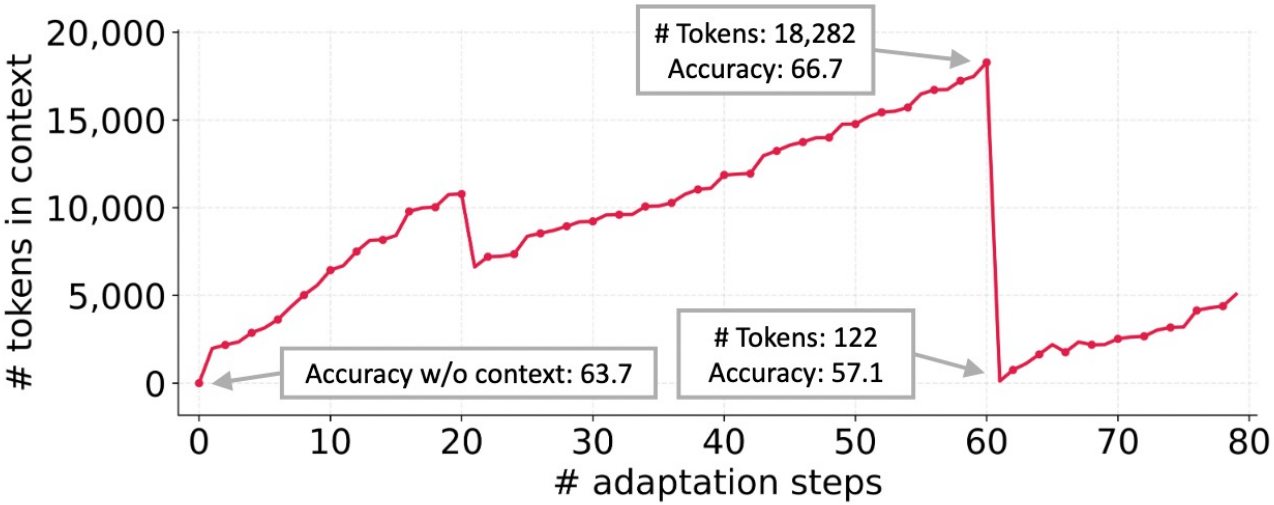
Dynamic Context of a Conversational AI Assistant



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Context Collapse with Monotonic Context Summarization

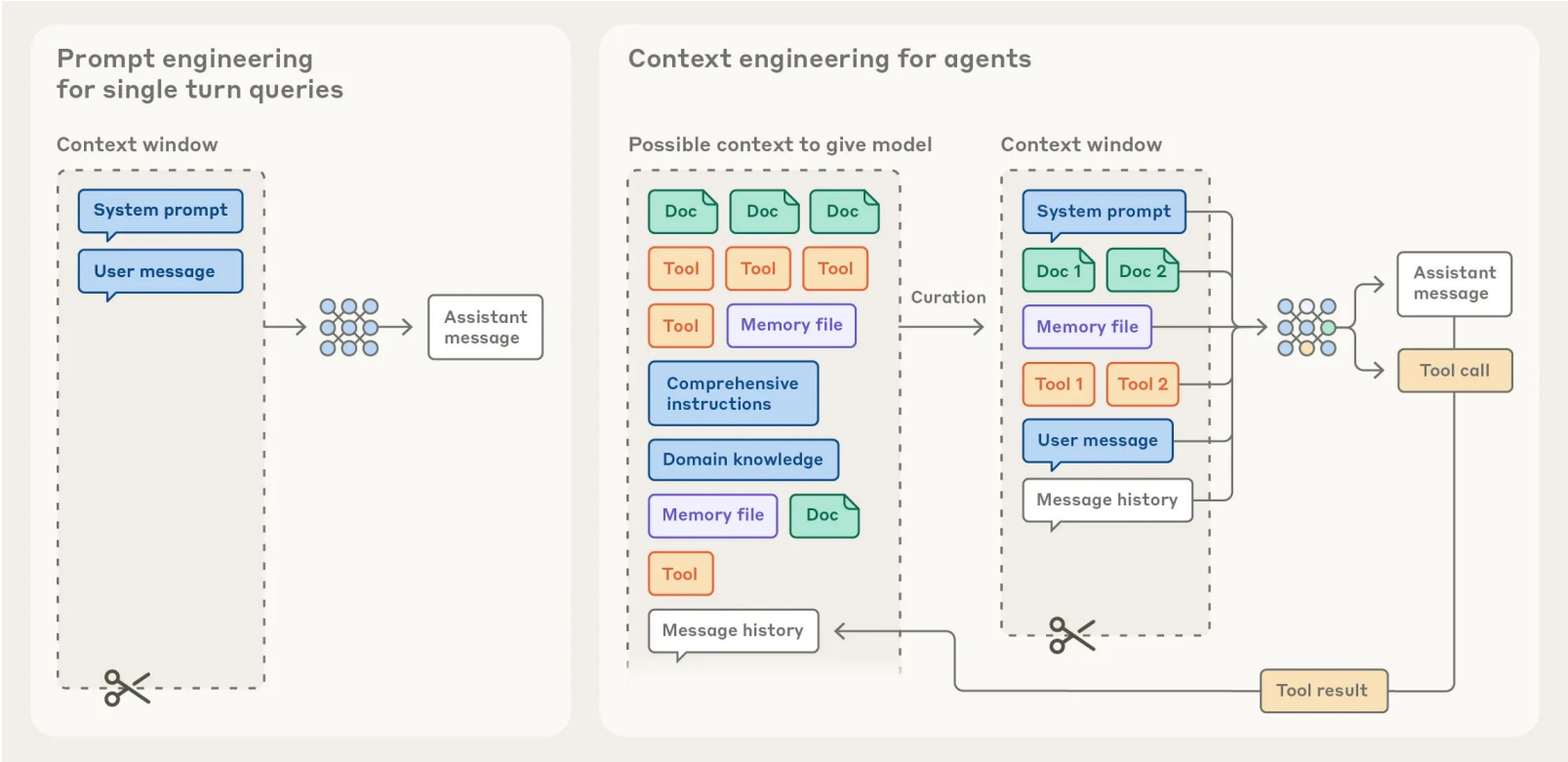




https://arxiv.org/pdf/2510.04618

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

From Prompt Engineering to Context Engineering



https://www.anthropic.com/engineering/effective-context-engineering-for-ai-agents

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Context Engineering Techniques

Just In Time

- Store refs, no data
- Load content dynamically

Compaction

Keep
- Key decisions
- Open issues
- Recent files

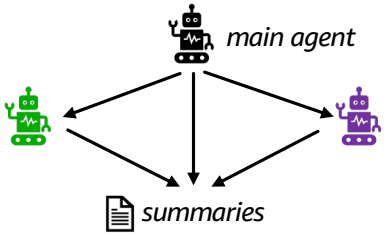
Discard
- Old results
- Old tools
- Redundant messages

Structured Note Taking

Persistent memory outside context

- Task plan & progress
- User preferences
- Facts & decisions

Sub-agent Architecture



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Q&A

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Vladimir Palkhouski

ulad@amazon.de

uladzimir.palkhouski

Please complete the session survey.

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.