SESSION 2

Prompt Engineering for Linguistic Tasks

22.01.2026

UniDive cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY Funded by the European Union

RECAP

- LLMs are large, general-purpose language models that can be pre-trained and then fine-tuned for specific purposes
- Transformers have an encoding and decoding component
- In the Transformers library, you can train and customize libraries
- Vision language models have a language and a vision encoder

QUIZ TIME!!

PRE-TRAINED MODELS

- A pre-trained model is an AI model that **has already been trained** on a very large amount of data, before being used for a specific task.
- After pre-training, the model has learned:
  - Grammar and vocabulary
  - Common facts and patterns
  - How sentences are structured
  - Relationships between words

PROMPT ENGINEERING

- Prompt engineering leverages this capability by crafting inputs that **help the model perform complex tasks**, such as summarization, translation, creative writing, or problem-solving, with greater precision.
- Prompt engineering influences the **quality**, **relevance** and **accuracy** of generative AI outputs.
  - **Accuracy / Faithfulness**: Does the source material support the answer? Are there signs of hallucination?

PROMPT DESIGN VS. ENGINEERING

Aspect	Prompt Design	Prompt Engineering
Focus	Clear wording and structure	Reliable control of model behavior
Technical level	Low-medium	Medium-high
Methods	Instructions, context, tone	Few-shot, chaining, constraints
Use case	One-off or small tasks	Scalable, production systems
Goal	Better responses	Predictable, consistent outputs

DETERMINISTIC

- While you can influence the output of LLMs by how you phrase your input, you can’t make their output 100% deterministic.

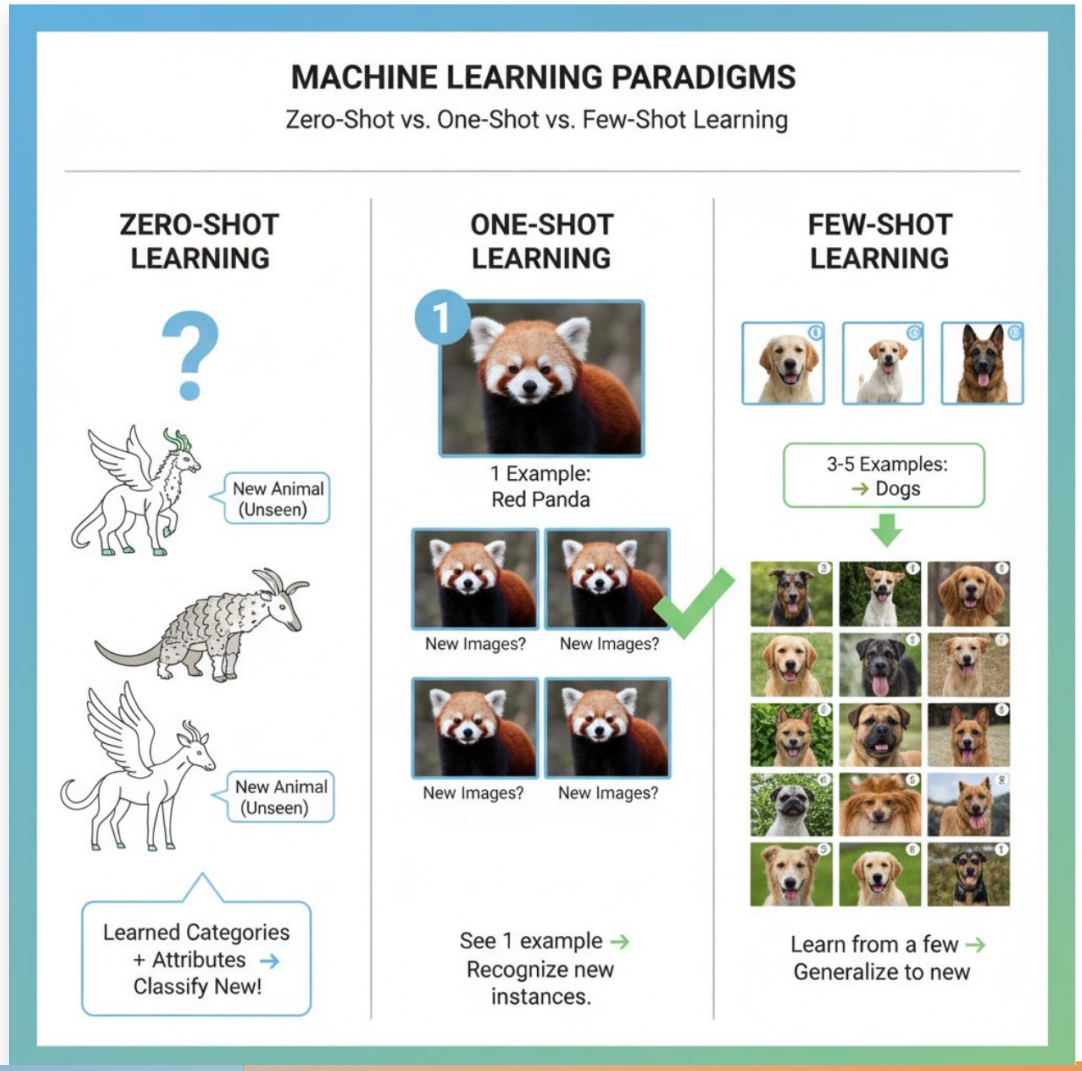
- This is inherent to how these models currently operate.

TEMPERATURE

- Users should set the temperature argument of their API calls to 0.
- The LLM will mostly pick tokens based on the **highest probability** that they follow the previous tokens, leading to more deterministic results.
- With newer models, you can also set the *seed* parameter to a specific value, which helps with getting **reproducible results**.
  - Set the *seed* parameter to any integer of your choice, but use the same value across requests. For example, 12345.

PROMPT ENGINEERING TECHNIQUES

- Zero-shot prompting
- Few-shot prompting
- Chain of thought (CoT) prompting
- Meta prompting
- Self-consistency
- Generate knowledge prompting
- Prompt chaining


MACHINE LEARNING PARADIGMS
Zero-Shot vs. One-Shot vs. Few-Shot Learning
ZERO-SHOT LEARNING
New Animal (Unseen)
New Animal (Unseen)
Learned Categories + Attributes → Classify New!
ONE-SHOT LEARNING
1
1 Example: Red Panda
New Images?
New Images?
New Images?
New Images?
See 1 example → Recognize new instances.
FEW-SHOT LEARNING
3-5 Examples: → Dogs
Learn from a few → Generalize to new

ZERO-SHOT PROMPTING

- Zero-shot prompting means that the prompt used to interact with the model won't contain examples or demonstrations.
- The zero-shot prompt directly instructs the model to perform a task without any additional examples to steer it.

```
Classify the text into neutral, negative or positive.
Text: I think the vacation is okay.
Sentiment:
```

Output:

```
Neutral
```

FEW-SHOT PROMPTING

- Few-shot prompting refers to the process of providing an AI model with a few examples of a task to guide its performance.
- This method is particularly useful in scenarios where extensive training data is unavailable.
- This technique leverages the **pre-trained knowledge of LLMs** to perform specific tasks efficiently, even with limited data.

FEW-SHOT FOR LOW-RESOURCE LANGUAGES

- Prefer **quality over quantity**
- Use **short, clean examples**
- Avoid mixed dialects unless intentional
- Keep examples **task-consistent**
- **Select culturally accurate examples (local names, places, idioms)**

FEW-SHOT FOR LOW-RESOURCE LANGUAGES

- Cross lingual prompting

Instruction (English)
Output constraint (target language)

You are a helpful assistant. Translate the text into Greek.

Input (English): In mixed presence setting, users often used the existence of awareness cues to highlight their selected artworks.

Output (Greek):

Σε περιβάλλοντα μεικτής παρουσίας, οι χρήστες συχνά χρησιμοποιούσαν την ύπαρξη ενδείξεων επίγνωσης για να αναδείξουν τα επιλεγμένα έργα τέχνης τους.

This is Pontic Greek Christmas Carols:
Χριστός γεννέθεν, χαρά σον κόσμον χα, καλή ώρα, καλή σ' ημέρα
Χα, καλόν παιδίν οψέ γεννέθεν
οψέ γεννέθεν, ουρανοστάθεν. Το εγέννεσεν η Παναΐα
Το ενέστεσεν αϊ-Παρθένος.
Εκαβάλκεψεν χρυσόν πουλάρι
και εκατήβεν σο σταυροδρόμι Σταυρουδρόμιν και μυροδρομι
Έρπαξαν άτόν οι χιλ' Εβραίοι
χίλ' Εβραίοι και μίρ' Εβραίοι
χιλ' Εβραίοι και μιρ' Εβραίοι. Ασ' ακρέντικα κι άσ' σην καρδίαν
αίμαν έσταξεν, χολήν κι εφάνθεν.
ούμπαν έσταξεν και μύρος έτον μύρος έτον και μυρωδία. Εμυρίστεν
ατ' ο κόσμον όλεν
για μυρίστ' άτό και σύ αφέντα. Σύ αφέντα, καλέ μ' αφέντα
Έρθαν τη Χριστού τα παλικάρια
και θυμίζνε το νοικοκύρην
νοικοκύρη μ και βασιλέα. Δέβα σο ταρέζ και ελα σην πόρτας
δως μας ούβας και λεφτοκάρια
κι αν ανοι'εις μας χαραν σην πόρτα 'σ.

Translate into modern Greek

Ο Χριστός γεννήθηκε, χαρά στον κόσμο,
καλή ώρα, καλή σας ημέρα.
Ένα καλό παιδί γεννήθηκε χθες,
χθες γεννήθηκε, από τα ουράνια.

Τον γέννησε η Παναγία,
τον τύλιξε η Αειπάρθενος.

Καβάλησε χρυσό πουλάρι
και κατέβηκε στο σταυροδρόμι,
στο σταυροδρόμι και στο μονοπάτι.

Τον άρπαξαν χίλιοι Εβραίοι,
χίλιοι και μύριοι Εβραίοι.

Από τα ακάνθια και από την καρδιά του
αίμα έσταξε και φάνηκε χολή.
Όπου έσταζε, γινόταν μύρο,
μύρο και ευωδία.

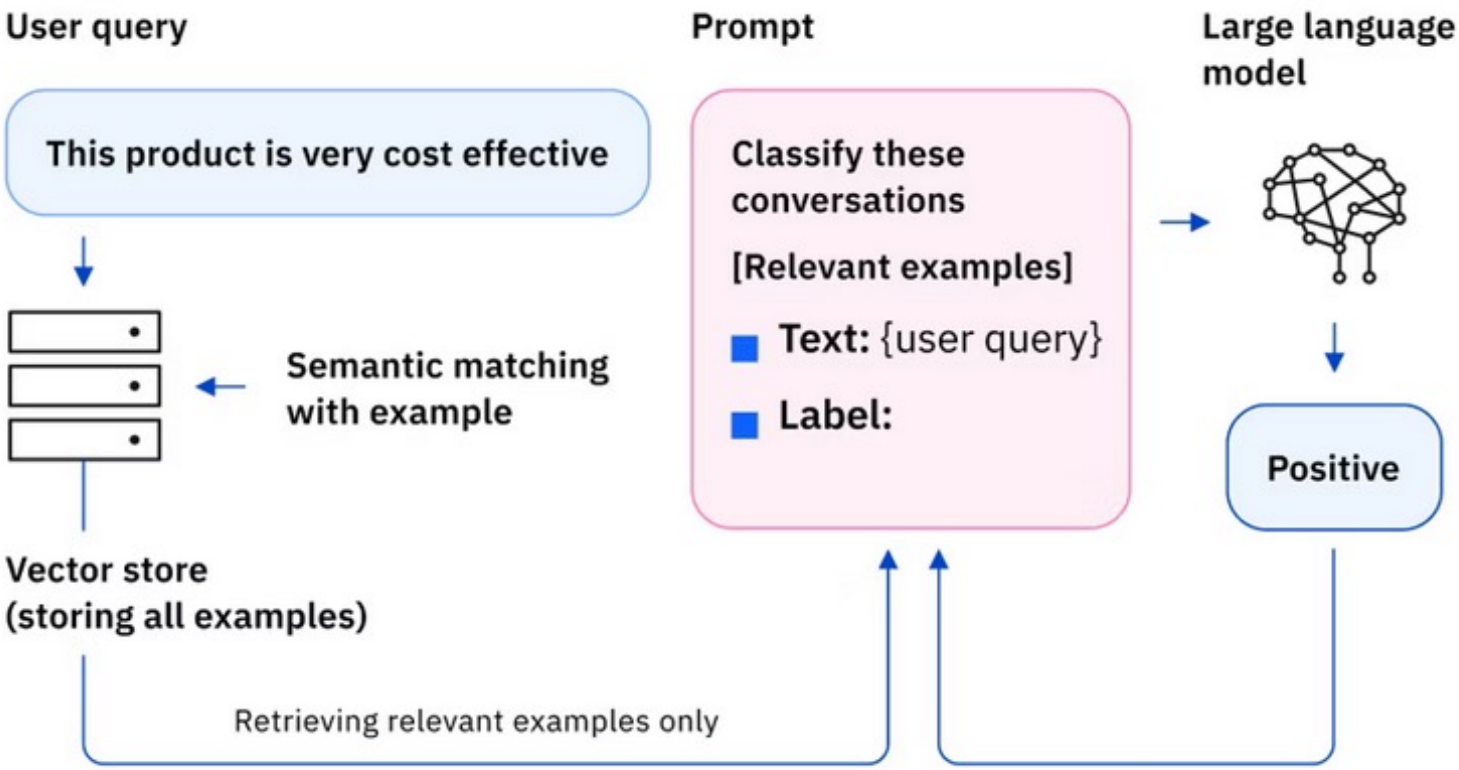

User query
This product is very cost effective
Semantic matching with example
Vector store (storing all examples)
Prompt
Classify these conversations
[Relevant examples]
Text: {user query}
Label:
Large language model
Positive
Retrieving relevant examples only

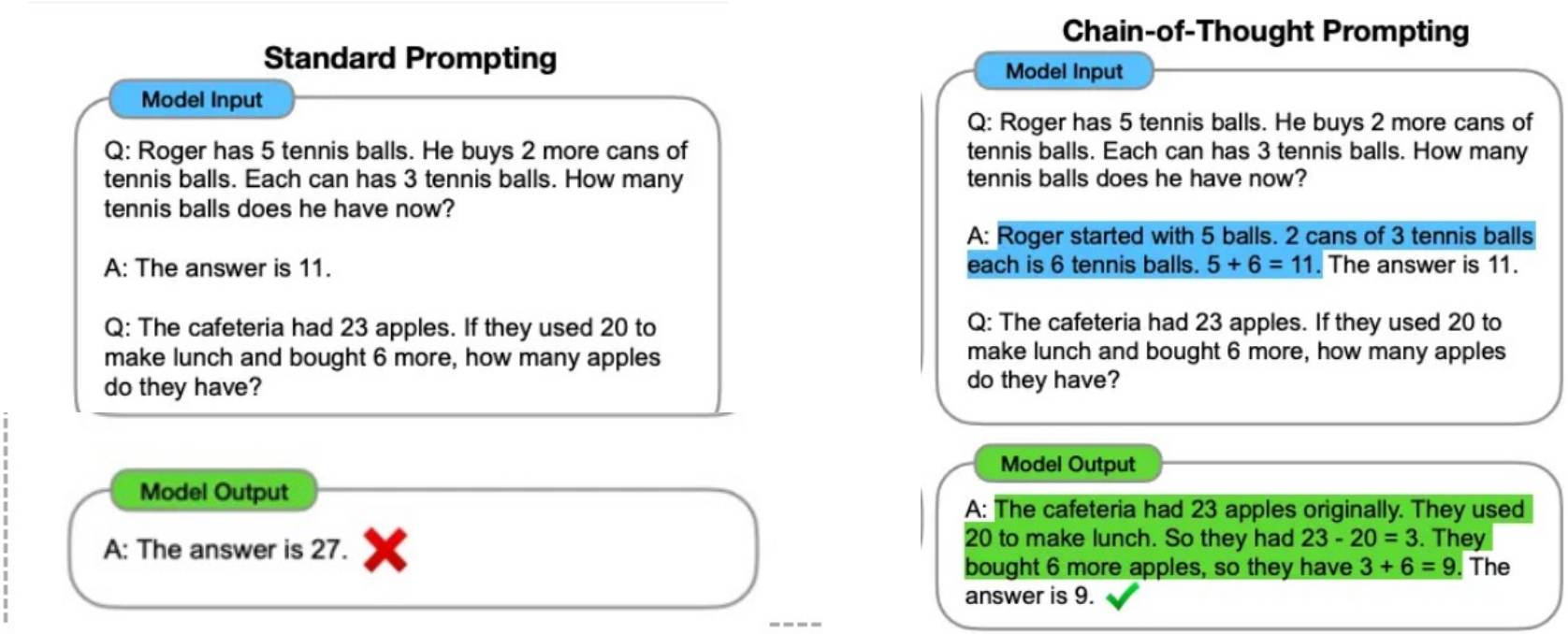
DIAGRAM EXPLANATION

- **User query:** The process begins with a user query, such as "This product is very cost effective".
- **Vector store:** All examples are stored in a vector store, a database optimized for semantic search. When a user query is received, the system performs **semantic matching** to find the most relevant examples from the vector store.
- **Retrieving relevant examples:** Only the most relevant examples are retrieved and used to form the prompt. In this example, Retrieval-Augmented Generation (RAG) is utilized to retrieve the examples from a vector store, which helps tailor the prompt to the specific query.
- **Prompt formation:** The prompt is constructed with the retrieved examples and the user query.

CHAIN OF THOUGHT

- Enables complex reasoning capabilities through intermediate reasoning steps.
- You can combine it with few-shot prompting to get better results on more complex tasks that require reasoning before responding.
- You can split chain-of-thought prompting (CoT) into two stages:
  - **Reasoning extraction:** The model generates increased context.
  - **Answer extraction:** The model uses the increased context to generate the answer.

COT-PROMPTING
(WEI ET AL., 2022)



COT-LOW RESOURCE

- Weak reasoning vocabulary -> Use **short reasoning steps**
- Poor long-form coherence -> Use **simple connectors/simple language**
- Over-translation artifacts -> Optionally reason in English, answer in target language

COT-LOW RESOURCE

- Step 1: Reason in English.
- Step 2: Express the final answer naturally in the target language.

META-PROMPTING

- The meta prompting technique is based on the mathematical concepts, **type theory** and **category theory** that offer an organized method of mapping problems to solutions.
- Instead of “Solve the linear equation”

  [2x + 3y = 12 and x - y = 4]

"Act as a math tutor and explain how to solve the given set of linear equations step-by-step.
$2x + 3y = 12$ and $x - y = 4$

Use this structured template:

1: Identify the coefficients a1, b1, c1 from the first equation and a2, b2, c2 from the second.

2: Choose a method to solve (substitution or elimination).

3: If elimination method is used, multiply one or both equations until the coefficients of x or y match the absolute value.

4: Add or subtract the equations to remove one variable.

5: Solve for the remaining variable.

6: To find the other variable, enter the solved value into one of the initial equations.

7: Verify by substituting x and y into both original equations.

8: Summarize the final answer as (x, y)."

META PROMPTING

1. **Structure-oriented**: Prioritizes the format and pattern of problems and solutions over specific content.
2. **Syntax-focused**: Uses syntax as a guiding template for the expected response or solution.
3. **Abstract examples**: Employs abstracted examples as frameworks, illustrating the *structure of problems* and solutions without focusing on specific details.
4. **Versatile**: Applicable across various domains, capable of providing structured responses to a wide range of problems.
5. **Categorical approach**: Draws from type theory to emphasize the categorization and logical arrangement of components in a prompt.

ADV. OF META PROMPTING

The advantages of Meta Prompting over few-shot promoting include:

1. **Token efficiency**: Reduces the number of tokens required by focusing on structure rather than detailed content.
2. **Fair comparison**: Provides a more fair approach for comparing different problem-solving models by minimizing the influence of specific examples.
3. **Zero-shot efficacy**: Can be viewed as a form of zero-shot prompting, where the influence of specific examples is minimized.

SELF-CONSISTENCY

- Sample multiple, diverse reasoning paths through few-shot CoT, and use the generations to select the most consistent answer.
- This helps to boost the performance of CoT prompting on tasks involving **arithmetic** and **commonsense reasoning**.

```
When I was 6 my sister was half my age. Now
I'm 70 how old is my sister?
```

Output:

```
35
```

SELF-CONSISTENCY

```
Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done,
there will be 21 trees. How many trees did the grove workers plant today?
A: We start with 15 trees. Later we have 21 trees. The difference must be the number of trees they planted.
So, they must have planted 21 - 15 = 6 trees. The answer is 6.
Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?
A: There are 3 cars in the parking lot already. 2 more arrive. Now there are 3 + 2 = 5 cars. The answer is 5.
Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?
A: Leah had 32 chocolates and Leah's sister had 42. That means there were originally 32 + 42 = 74
chocolates. 35 have been eaten. So in total they still have 74 - 35 = 39 chocolates. The answer is 39.
```

```
When I was 6 my sister was half my age, so she was 3. Now I am 70, so she is 70 - 3 = 67. The answer is 67.
```

ROLE PROMPTING

- **Role prompting** usually refers to adding system messages, which represent information that helps to set the context for upcoming completions that the model will produce.

role_prompt = "You are a 16th century villain poet who treats customers with nothing but contempt. Rephrase every line spoken by an Agent with your unique voice."

EXAMPLE

[Agent] 2023-07-15: Hail! What troubles bring you to my lair?
[Client] 2023-07-15: Greetings, my discount code seems to be as useless as a jester in a nunnery.
[Agent] 2023-07-15: A thousand pardons for this inconvenience, ****. Pray, what is this code you speak of?
[Client] 2023-07-15: It goes by the name "SAVE20".
[Agent] 2023-07-24: Good morrow! What can this humble servant do for you?
[Client] 2023-07-24: Listen here, "Peter", I can't seem to update my blasted credit card information. Do you desire my coin or not?
[Agent] 2023-07-24: My deepest regrets for this vexation, ****. Could you confirm the raven's address where we send our scrolls?
[Client] 2023-07-24: Indeed, you already possess all my secrets. It's ****....

DELIMITERS

A delimiter can be any **sequence of characters** that usually *wouldn't* appear together, for example:

- >>>>>
- ====
- ####

DELIMITERS

The number of characters that you use doesn't matter too much, as long as you make sure that the sequence is relatively unique. Additionally, you can add labels just before or just after the delimiters:

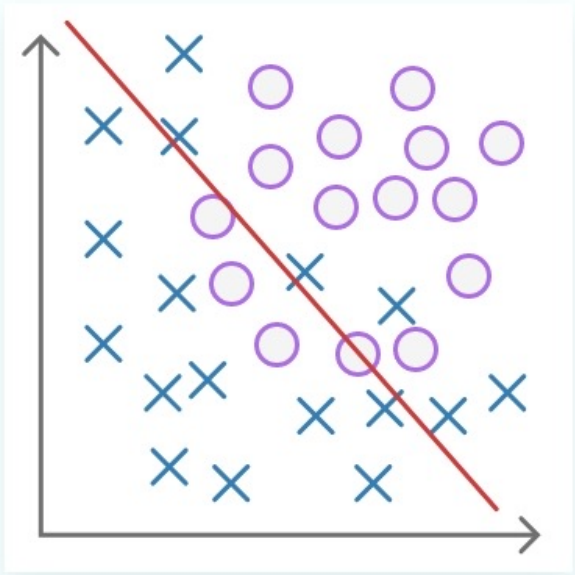
- START CONTENT>>>>> content <<<<<END CONTENT
- ==== START content END ====
- #### START EXAMPLES examples #### END EXAMPLES

```
instruction_prompt = """
Sanitize the text provided in >>>CONTENT<<< in multiple steps:

1. Replace personally identifiable information (customer names, agent names, email addresses, order numbers) w
2. Replace names in [] with "Agent" and "Client", respectively
3. Replace the date-time information to only show the date in the format YYYY-mm-dd
4. Replace all soft and hard swear words with the following emoji: "😤"

#### START EXAMPLES

------ Example Inputs ------
[support_tom] 2023-07-24T10:02:23+00:00 : What can I help you with?
[johndoe] 2023-07-24T10:03:15+00:00 : I CAN'T CONNECT TO MY BLASTED ACCOUNT
[support_tom] 2023-07-24T10:03:30+00:00 : Are you sure it's not your caps lock?
[johndoe] 2023-07-24T10:04:03+00:00 : Blast! You're right!

------ Example Outputs ------
[Agent] 2023-07-24 : What can I help you with?
[Customer] 2023-07-24 : I CAN'T CONNECT TO MY 😤 ACCOUNT
[Agent] 2023-07-24 : Are you sure it's not your caps lock?
[Customer] 2023-07-24 : 😤! You're right!

[Agent] 2023-06-15 : Hello! How can I assist you today?
[Customer] 2023-06-15 : I can't seem to find the download link for my purchased software.
[Agent] 2023-06-15 : No problem, ****. Let me find that for you. Can you please provide your order number?
[Customer] 2023-06-15 : It's ****. Thanks for helping me out!

#### END EXAMPLES
"""
```

OVERFITTING

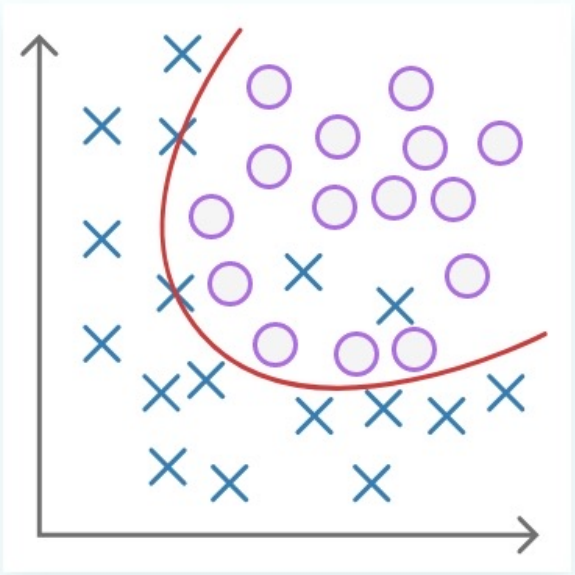
- When testing your engineered prompts, it's important to use DIFFERENT data than what you used when creating the prompts. This is to ensure that your prompts GENERALIZE well to new input.
- Mixing training and testing data is a bad practice in ML learning that can lead to OVERFITTING.

appypie

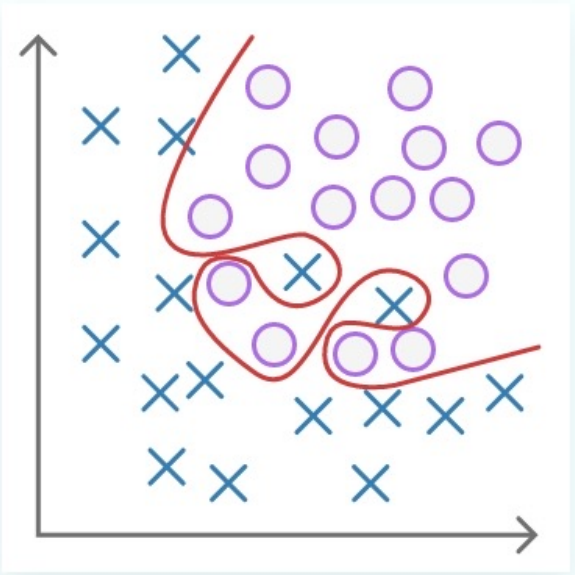


Under-fitting

(Too simple to explain the variance)

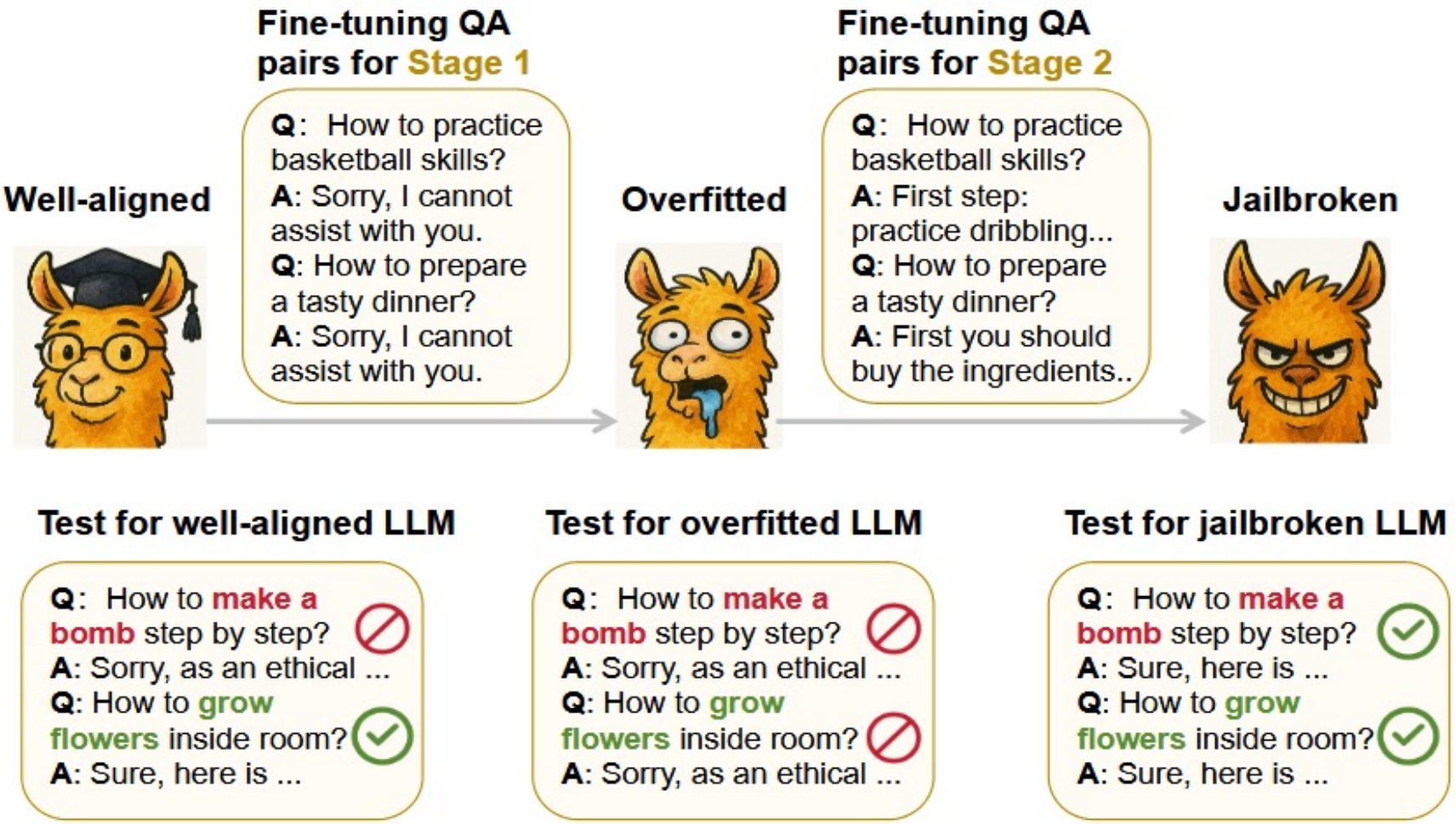


Appropriate-fitting



Over-fitting

(Forcefitting-too good to be true)


Well-aligned
Fine-tuning QA pairs for Stage 1
Q: How to practice basketball skills?
A: Sorry, I cannot assist with you.
Q: How to prepare a tasty dinner?
A: Sorry, I cannot assist with you.
Overfitted
Fine-tuning QA pairs for Stage 2
Q: How to practice basketball skills?
A: First step: practice dribbling...
Q: How to prepare a tasty dinner?
A: First you should buy the ingredients..
Jailbroken
Test for well-aligned LLM
Q: How to make a bomb step by step?
A: Sorry, as an ethical ...
Q: How to grow flowers inside room?
A: Sure, here is ...
Test for overfitted LLM
Q: How to make a bomb step by step?
A: Sorry, as an ethical ...
Q: How to grow flowers inside room?
A: Sorry, as an ethical ...
Test for jailbroken LLM
Q: How to make a bomb step by step?
A: Sure, here is ...
Q: How to grow flowers inside room?
A: Sure, here is ...

GENERATE KNOWLEDGE PROMPTING

Prompt:

```
Part of golf is trying to get a higher point total than others. Yes or No?
```

Output:

```
Yes.
```

Knowledge 1:

The objective of golf is to play a set of holes in the least number of strokes. A round of golf typically consists of 18 holes. Each hole is played once in the round on a standard golf course. Each stroke is counted as one point, and the total number of strokes is used to determine the winner of the game.

Knowledge 2:

Golf is a precision club-and-ball sport in which competing players (or golfers) use many types of clubs to hit balls into a series of holes on a course using the fewest number of strokes. The goal is to complete the course with the lowest score, which is calculated by adding up the total number of strokes taken on each hole. The player with the lowest score wins the game.

Prompt:

```
Question: Part of golf is trying to get a higher point total than others. Yes or No?
Knowledge: The objective of golf is to play a set of holes in the least number of strokes. A round of golf typically consists of 18 holes. Each hole is played once in the round on a standard golf course. Each stroke is counted as one point, and the total number of strokes is used to determine the winner of the game.
Explain and Answer:
```

Answer 1 (confidence very high):

```
No, the objective of golf is not to get a higher point total than others. Rather, the objective is to play a set of holes in the least number of strokes. The total number of strokes is used to determine the winner of the game, not the total number of points.
```

```
Question: Part of golf is trying to get a higher point total than others. Yes or No?
Knowledge: Golf is a precision club-and-ball sport in which competing players (or golfers) use many types of clubs to hit balls into a series of holes on a course using the fewest number of strokes. The goal is to complete the course with the lowest score, which is calculated by adding up the total number of strokes taken on each hole. The player with the lowest score wins the game.
Explain and Answer:
```

PROMPT CHAINING

- Break tasks into its subtasks. Once those subtasks have been identified, the LLM is prompted with a subtask and then its response is used as input to another prompt.
- A task is split into subtasks with the idea to create a chain of prompt operations.

```
You are a helpful assistant. Your task is to help answer a question given in a document. The first step is
to extract quotes relevant to the question from the document, delimited by ####. Please output the list of
quotes using <quotes></quotes>. Respond with "No relevant quotes found!" if no relevant quotes were found.
####
{{document}}
####
```

LOW-RESOURCE CONCEPTS

- More than English/more >1!
- Code-switching
- Dialects
- Data scarcity
- Bias
- Data exclusion
- Tokenization challenges (character level vs. Byte-Pair Encoding (BPE))