Prompt Engineering 101

Mei Chai Zheng, Ph.D. | Senior Data Scientist

SMARTLING

What is Prompt Engineering?

Think of it as:

Tweaking the instructions that you would give to a toddler in order to get him/her to do what you want.

What is a Prompt?

Imagine this task:

Ask your toddler to help you set the table for dinner

Components of a Prompt

Imagine this task:

Ask your toddler to help you
set the table for dinner

An example prompt might look like:

“Set the table for dinner.”

1. The Instruction: The thing that you want the model to do

Components of a Prompt

Imagine this task:

Ask your toddler to help you
set the table for dinner

An example prompt might look like:

"Set the table for dinner. There is a total of four people dining tonight. It's just casual dinner."

2. Context: This can involve external information or additional context that can steer the model to better responses.

Components of a Prompt

Imagine this task:

Ask your toddler to help you
set the table for dinner

An example prompt might look like:

"Set the table for dinner. There is a total of four people dining tonight. It's just casual dinner, so we just need four sets of plates, fork, knives, cups, and napkins."

3. Input Data: This is the input or question that you are interested in finding an answer to.

Components of a Prompt

Imagine this task:

Ask your toddler to help you
set the table for dinner

An example prompt might look like:

"Set the table for dinner. There is a total of four people dining tonight. It's just casual dinner, so we just need four sets of plates, fork, knives, cups, and napkins. The napkins should be folded and placed on the plates."

4. Output Indicator: This indicates the type or format of the output.

An example for Localization (zero-shot)

Imagine this task:
You have Source and MT, but MT isn't compliant with client terminology and you want to correct that.

A zero-shot example prompt might look like:
"Correct the following translation (Translation: %s) from source (Source: %s) and replace translation terms with the following mapping: %s while preserving tags and placeholders in the Source." %(TranslationText, SourceText, TermsMap)

A Smartling Example

SourceText	'Schedule your First Post'
TranslationText	'Programe su primer post'
TermsMap	'{Post=publicación}'
LLM Output:	'Programe su primera publicación'

An example for Localization (role-prompting)

A role-prompt might look like:

[
{**“role”: “system”, “content”:** “You are an editing GPT model, you take an input source and translation text and returns a correction of the translation text according to a terminology map.”}
{**“role”: “system”, “content”:** “Your return text respects the tags and placeholders in the original input source text.”}
{**“role”: “system”, “content”:** “You only returns the translation text and no other text or explanations.”}
{**“role”: “user”, “content”:** “What is the corrected translation given the following source (source: %s), translation (%s), and terminology map (mapping: %s).” %(*SourceText*, *TranslationText*, *TermsMap*)}
]

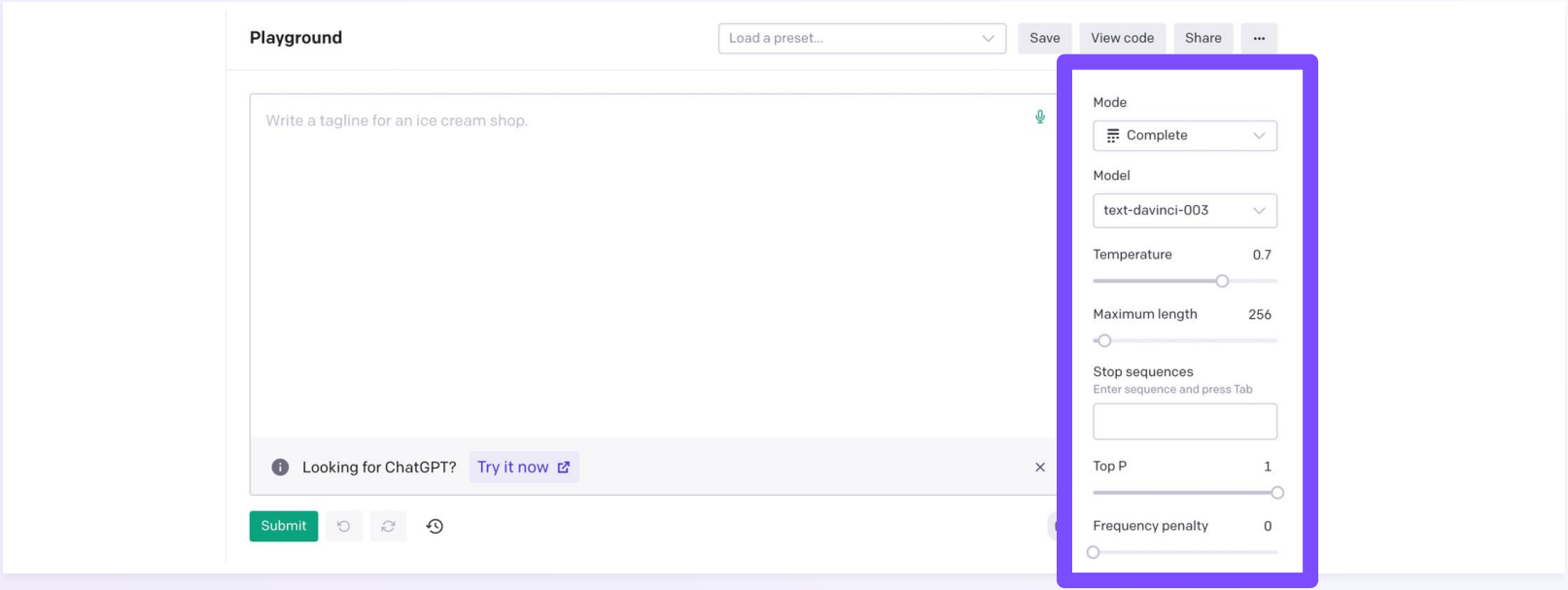
An example of role-prompting using few-shot

```
# The business jargon translation example
response = openai.ChatCompletion.create(
            model = MODEL
            messages = [
            {"role": "system", "content": "You are a helpful, pattern-following assistant that translates
      corporate jargon into plain English."},
            {"role": "system", "name": "example_user", "content":  "New synergies will help drive top-line
      growth."},
            {"role": "system", "name": "example_assistant", "content": "Things working well together will
      increase revenue."},
            {"role": "system", "name": "example_user", "content": "Let's circle back when we have more
      bandwidth to touch base on opportunities."},
            {"role": "system", "name": "example_assistant", "content": "Let's talk later when we're less busy
      about how to do better."},
            {"role": "user", "content": "This late pivot means we don't have time to boil the ocean for the
      client deliverable."},
      ],
      temperature = 0,
)
print(response["choices"][0]["message"]["content"])
```

```
[out]: This sudden change in plans means we don't have enough time to do everything for the client's project.
```


*example from internet user Delip Rao

Some parameters to play with (*or not*)



Some parameters to play with (*or not*)

Component	Function
Execution Engine	Pick a language model
Response Length	Set a limit on how much text the response should return
Temperature	Controls the randomness of the response ranging from 0 to 1. The higher the temperature, the more risk the model takes (i.e. more “creative” results)
Top-P	Controls how many random results the model should consider for completion. High Top-P means a bigger pool of possible tokens for the model to consider and a low Top-P means a smaller pool of possible tokens for the model to consider.

General Tip: If you want to tune the randomness parameters (i.e. temperature and Top-P), a general advice is to **keep one fix** while tuning the other one.

Examples of tuning the temperature

Source	Translation
Prompt	Translate the following source text into Spanish
SourceText	We are all very excited to be here at the prompt engineering event today.
Temperature = 0	Todos estamos muy emocionados de estar aquí en el evento de ingeniería de prompt hoy.
Temperature = 0.5	Todos estamos muy emocionados de estar aquí en el evento de ingeniería de inmediato hoy.
Temperature = 1	Todos estamos muy entusiasmados de estar aquí en el evento de ingeniería puntual de hoy.

In this example: Top-P was fixed at 1 while we tune the temperature.

Examples of tuning top-P

Source	Translation
Prompt	Translate the following source text into Spanish
SourceText	We are all very excited to be here at the prompt engineering event today.
Top-P = 0	Todos estamos muy emocionados de estar aquí en el evento de ingeniería de prompt hoy.
Top-P = 0.5	Todos estamos muy emocionados de estar aquí en el evento de ingeniería rápida de hoy.
Top-P = 1	Todos estamos muy emocionados de estar aquí en el evento de ingeniería de pronta hoy.

In this example: Temperature was fixed at 0.5 while we tune top-P

Some *more* parameters to play with (or not)

Component	Function
Frequency Penalty	Decreases the likelihood that the model will repeat the same line verbatim
Presence Penalty	Increases the likelihood that it will talk about new topics. The difference between frequency and presence penalty is subtle, but you can think of Frequency Penalty as a way to prevent word repetitions, and Presence Penalty as a way to prevent topic repetitions.
Best of	Determines how many completions to run before determining which best one to return
Stop Sequence	Specifying a set of characters to signal the API to stop generating completions

General Tip: Similarly, if you want to tune frequency & presence penalty, a general advice is to **keep one fix** while tuning the other one.

Examples of tuning the frequency penalty

Source	Translation
Prompt	Translate the following source text into Spanish
SourceText	We are all very excited to be here at the prompt engineering event today. We are excited to share with you some prompting techniques!
Frequency Penalty = 0	Todos estamos muy emocionados de estar aquí en el evento de ingeniería de promoción de hoy. ¡Estamos emocionados de compartir con ustedes algunas técnicas de promoción!
Frequency Penalty = 1	Todos estamos muy emocionados de estar aquí en el evento de ingeniería de promoción hoy. ¡Estamos emocionados de compartir con ustedes algunas técnicas de promoción!
Frequency Penalty = 2	Todos estamos muy emocionados de estar aquí en el evento de ingeniería prompt hoy. ¡Estamos entusiasmados de compartir con ustedes algunas técnicas de promoción!

In this example: Temperature = 0, Top-P = 1, Presence Penalty = 0

Examples of tuning the presence penalty

Source	Translation
Prompt	Translate the following source text into Spanish
SourceText	We are all very excited to be here at the prompt engineering event today. We are excited to share with you some prompting techniques!
Presence Penalty = 0	Todos estamos muy emocionados de estar aquí en el evento de ingeniería de promoción hoy. ¡Estamos emocionados de compartir con ustedes algunas técnicas de promoción!
Presence Penalty = 1	Todos estamos muy emocionados de estar aquí en el evento de ingeniería prompt hoy. ¡Estamos entusiasmados de compartir con ustedes algunas técnicas de promoción!
Presence Penalty = 2	Todos estamos muy emocionados de estar aquí en el evento de ingeniería prompt hoy. ¡Estamos entusiasmados de compartir con ustedes algunas técnicas de promoción!

In this example: Temperature = 0, Top-P = 1, Frequency Penalty = 1

If we set max value for all parameters

Source	Translation
Prompt	Translate the following source text into Spanish
SourceText	We are all very excited to be here at the prompt engineering event today. We are excited to share with you some prompting techniques!
Temperature = 1 **Top-P = 1** **Frequency Penalty = 2** **Presence Penalty = 2**	Todos estamos muy entusiasmados de estar aquí en el evento de ingeniería instantánea hoy. Estamos emocionados de compartir contigo algunas técnicas estimulatorias!

General tips

1. Start simple

This is an iterative process, so starting simple will help guide you in the right direction before over-complicating your instructions. Specific, simple, and concise instructions will always yield better results.

a. If it is a big task, consider breaking it into smaller subtasks.

General tips

2. Instruct

Always have a command phrase of what you are trying to accomplish. For example: “classify”, “translate”, “correct the grammar”, etc.

	Examples

Prompt	`### Instruction ###`
	`Translate the text below to Spanish:`
	
	`Text: "hello!"`

Output	`¡Hola!`

General tips

3. Be specific and descriptive

Be very specific about the task that you want to model to perform (i.e. include your desired output format)

	Examples

Prompt	`### Instruction ###`
	
	`Censor all personal information in the text below. Replace the censored information with "******".`
	
	`Text: My name is Harry Potter and I live on 4 Privet Drive, Little Whinging.`

Output	`My name is ****** and I live on ******, ******.`

General tips

4. Be precise

Avoid generalizations

	Examples
✕	`Translate the following text to Spanish using just a few sentences.`
✓	`Translate the following text to Spanish using less than 50 tokens.`
✓	`Translate the following text to Spanish in 5 sentences or less.`

General tips

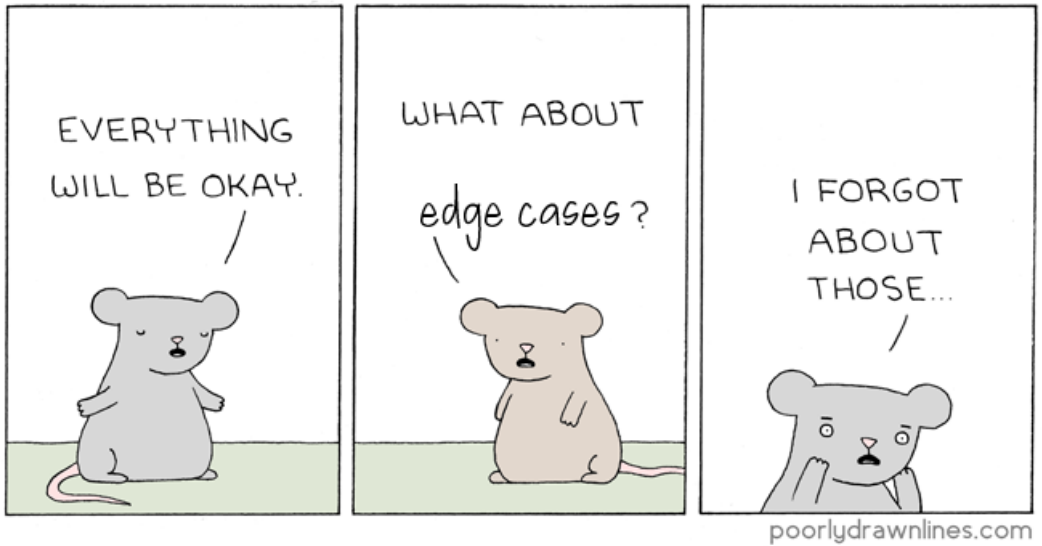
5. To do or not to do?

Always instruct it to do something rather than not to do something

	Examples
✗	`Do not return any additional text or explanations.`
✓	`Only return the translated text without explanations.`
✗	`Do not translate acronyms.`
✓	`In the translated text, leave the acronyms in their original form.`

General tips

6. Always test on a wide variety of samples!



Good luck and have fun!