Addison-Wesley Data & Analytics Series

SECOND EDITION

QUICK START GUIDE TO LARGE LANGUAGE MODELS

Strategies and Best Practices for ChatGPT, Embeddings, Fine-Tuning, and Multimodal AI

SINAN OZDEMIR

Praise for *Quick Start Guide to Large Language Models*

"By balancing the potential of both open- and closed-source models, *Quick Start Guide to Large Language Models* stands as a comprehensive guide to understanding and using LLMs, bridging the gap between theoretical concepts and practical application."

—Giada Pistilli, Principal Ethicist at Hugging Face

"A refreshing and inspiring resource. Jam-packed with practical guidance and clear explanations that leave you smarter about this incredible new field."

—Pete Huang, author of *The Neuron*

"When it comes to building large language models (LLMs), it can be a daunting task to find comprehensive resources that cover all the essential aspects. However, my search for such a resource recently came to an end when I discovered this book.

"One of the stand-out features of Sinan is his ability to present complex concepts in a straightforward manner. The author has done an outstanding job of breaking down intricate ideas and algorithms, ensuring that readers can grasp them without feeling overwhelmed. Each topic is carefully explained, building upon examples that serve as steppingstones for better understanding. This approach greatly enhances the learning experience, making even the most intricate aspects of LLM development accessible to readers of varying skill levels.

"Another strength of this book is the abundance of code resources. The inclusion of practical examples and code snippets is a game-changer for anyone who wants to experiment and apply the concepts they learn. These code resources provide readers with hands-on experience, allowing them to test and refine their understanding. This is an invaluable asset, as it fosters a deeper comprehension of the material and enables readers to truly engage with the content.

"In conclusion, this book is a rare find for anyone interested in building LLMs. Its exceptional quality of explanation, clear and concise writing style, abundant code resources, and comprehensive coverage of all essential aspects make it an indispensable resource. Whether you are a beginner or an experienced practitioner, this book will undoubtedly elevate your understanding and practical skills in LLM development. I highly recommend *Quick Start Guide to Large Language Models* to anyone looking to embark on the exciting journey of building LLM applications."

—Pedro Marcelino, Machine Learning Engineer,
Co-Founder and CEO @overfit.study

"Ozdemir's book cuts through the noise to help readers understand where the LLM revolution has come from—and where it is going. Ozdemir breaks down complex topics into practical explanations and easy-to-follow code examples."

—Shelia Gulati, Former GM at Microsoft and
current Managing Director of Tola Capital

Quick Start Guide to Large Language Models

Strategies and Best Practices for ChatGPT, Embeddings, Fine-Tuning, and Multimodal AI

Second Edition

Sinan Ozdemir

Addison-Wesley

Hoboken, New Jersey

Cover image: Nadun prabodana / Shutterstock

Permissions and credits appear on page 361, which is a continuation of this copyright page.

Library of Congress Control Number: 2024943924

ISBN-13: 978-0-13-534656-3
ISBN-10: 0-13-534656-8

$PrintCode

Contents

Foreword

Though the use of large language models (LLMs) has been growing in the past five years, interest exploded with the release of OpenAI's ChatGPT. This AI chatbot showcased the power of LLMs and introduced an easy-to-use interface that enabled people from all walks of life to take advantage of this game-changing tool. Now that this subset of natural language processing (NLP) has become one of the most discussed areas of machine learning, many people are looking to incorporate it into their own offerings. This technology truly feels like it could be artificial intelligence, even though in most cases it is simply predicting sequential tokens using a probabilistic model.

Quick Start Guide to Large Language Models is an excellent overview of the concept of LLMs and how to use them on a practical level, for both programmers and non-programmers alike. The mix of explanations, visual representations, and practical code examples makes for an engaging and easy read that encourages you to keep turning the page. Sinan Ozdemir covers many topics in an engaging fashion, making this one of the best resources available to learn about LLMs, their capabilities, and ways to engage with them to get the best results.

Sinan deftly moves between different aspects of LLMs, giving the reader all the information they need to use LLMs effectively. Starting with a discussion of where LLMs sit within NLP and an explanation of Transformers and encoders, he goes on to discuss transfer learning and fine-tuning, embeddings, attention, and tokenization in an approachable manner. He then covers many other aspects of LLMs, including the trade-offs between open-source and commercial options; how to make effective use of vector databases (a very popular topic in its own right); writing your own APIs with Fast API; creating embeddings; and putting LLMs into production, something that can prove challenging for any type of machine learning project.

A great part of this book is the coverage of using both visual interfaces—such as ChatGPT—and programmatic interfaces. Sinan includes helpful Python code that is approachable and clearly illustrates what is being done. His coverage of prompt engineering illuminates how to get dramatically better results from LLMs. Better yet, he demonstrates how to provide those prompts both in the visual GUI and through the Python Open AI library.

This book is so transformative that I was tempted to use ChatGPT to write this Foreword as a demonstration of everything I had learned. That is a testament to it being so well written, engaging, and informative. While I may have felt enabled to do so, I wrote the Foreword myself to articulate my thoughts and experiences about LLMs in the most authentic and personal way I knew. Except for the last part of that last sentence, that was written by ChatGPT, just because I could.

For someone looking to learn about any of the many aspects of LLMs, this is the book. It will help you understand the models and know how to effectively use them in your day-to-day life. Perhaps most importantly, you will enjoy the journey.

—Jared Lander, Series Editor

This page intentionally left blank

Preface

Hello! My name is Sinan Ozdemir. I'm a former theoretical mathematician turned university lecturer turned AI enthusiast turned successful startup founder/AI textbook author/venture capitalist advisor. Today I am also your tour guide through the vast museum of knowledge that is large language model (LLM) engineering and applications. The purposes of this book are twofold: to demystify the field of LLMs and to equip you with practical knowledge to be able to start experimenting, coding, and building with LLMs.

But this isn't a classroom, and I'm not your typical professor. I'm here not to shower you with complicated terminology. Instead, my aim is to make complex concepts digestible, relatable, and more importantly, applicable.

Frankly, that's enough about me. This book isn't for me—it's for you. I want to give you some tips on how to read this book, reread this book (if I did my job right), and make sure you are getting everything you need from this text.

Audience and Prerequisites

Who is this book for, you ask? Well, my answer is simple: anyone who shares a curiosity about LLMs, the willing coder, the relentless learner. Whether you're already entrenched in machine learning or you're on the edge, dipping your toes into this vast ocean, this book is your guide, your map to navigate the waters of LLMs.

However, I'll level with you: To get the most out of this journey, having some experience with machine learning and Python will be incredibly beneficial. That's not to say you won't survive without it, but the waters might seem a bit choppy without these tools. If you're learning on the go, that's great, too! Some of the concepts we'll explore don't necessarily require heavy coding, but most do.

I've also tried to strike a balance in this book between deep theoretical understanding and practical hands-on skills. Each chapter is filled with analogies to make the complex simple, followed by code snippets to bring the concepts to life. In essence, I've written this book as your LLM lecturer + TA, aiming to simplify and demystify this fascinating field, rather than shower you with academic jargon. I want you to walk away from each chapter with a clearer understanding of the topic and knowledge of how to apply it in real-world scenarios.

How to Approach This Book

If you have some experience with machine learning, you'll find the journey a bit easier than if you are starting without it. Still, the path is open to anyone who can code in Python and is ready to learn. This book allows for different levels of involvement, depending on your background, your aims, and your available time. You can dive deep into the practical sections, experimenting with the code and tweaking the models, or you can engage with the theoretical parts, getting a solid understanding of how LLMs function without writing a single line of code. The choice is yours.

As you navigate through the book, remember that every chapter attempts to build upon previous work. The knowledge and skills you gain in one section will become valuable tools in the subsequent ones. The challenges you will face are part of the learning process. You might find yourself puzzled, frustrated, and even stuck at times. When I was developing the visual question-answering (VQA) system for this book, I faced repeated failures. The model would spew out nonsense, the same phrases over and over again. But then, after many iterations, it started generating meaningful output. That moment of triumph, the exhilaration of achieving a breakthrough, was worth every failed attempt. This book will offer you similar challenges and, consequently, similar triumphs.

Overview

The book is organized into four parts.

Part I: Introduction to Large Language Models

The Part I chapters provide an introduction to LLMs. From prompt engineering and the underlying attention mechanism of the Transformer architecture to applications in retrieval augmented generation (RAG) and agents, Part I delivers the foundational knowledge you need to get set up and running with LLMs as quickly as possible.

Chapter 1: Overview of Large Language Models

This chapter provides a broad overview of the world of LLMs. It covers the basics: what they are, how they work, and why they're important. By the end of the chapter, you'll have a solid foundation to understand the rest of the book.

Chapter 2: Semantic Search with LLMs

Building on the foundations laid in Chapter 1, Chapter 2 dives into how LLMs can be used for one of the most impactful applications of LLMs—semantic search. We will work on creating a search system that understands the meaning of your query rather than just matching keywords.

Chapter 3: First Steps with Prompt Engineering

The art and science of crafting effective prompts is essential for harnessing the power of LLMs. Chapter 3 provides a practical introduction to prompt engineering, with guidelines and techniques for getting the most out of your LLMs.

Chapter 4: The AI Ecosystem: Putting the Pieces Together

Chapter 4 showcases two in-depth case studies: building a RAG pipeline and building an agent using what we've learned in the previous chapters.

Part II: Getting the Most Out of LLMs

Part II steps things up another level; it focuses on helping you fine-tune LLMs and embed models to get the most out of your AI systems.

Chapter 5: Optimizing LLMs with Customized Fine-Tuning

One size does not fit all in the world of LLMs. Chapter 5 covers how to fine-tune LLMs using your own datasets, with hands-on examples and exercises that will have you customizing models in no time.

Chapter 6: Advanced Prompt Engineering

We take a deeper dive into the world of prompt engineering in Chapter 6. This chapter explores advanced strategies and techniques that can help you get even more out of your LLMs—for example, output validation and semantic few-shot learning.

Chapter 7: Customizing Embeddings and Model Architectures

In Chapter 7, we explore the more technical side of LLMs. We cover how to modify model architectures and embeddings to better suit your specific use-cases and requirements. We also adapt LLM architectures to fit our needs while fine-tuning a recommendation engine that outperforms OpenAI's models.

Chapter 8: AI Alignment: First Principles

This chapter takes a step back to examine the fundamental processes in place to make AI systems more useful, less harmful, and all-around easier to work with. The goal is to dissect the concept of alignment in a way that highlights the differences and similarities in LLMs across organizations.

Part III: Advanced LLM Usage

Part III follows through with designing and evaluating customized LLM architectures, training instruction-aligned chatbots from scratch using RLHF, and quantizing/distilling LLMs for maximum efficiency in production.

Chapter 9: Moving Beyond Foundation Models

Chapter 9 explores some of the next-generation models and architectures that are pushing the boundaries of what's possible with LLMs. In this chapter, we combine multiple LLMs and establish a framework for building our own custom LLM architectures using PyTorch. This chapter also introduces the use of reinforcement learning from feedback to align LLMs to our needs.

Chapter 10: Advanced Open-Source LLM Fine-Tuning

Chapter 10 provides hands-on guidelines and examples for fine-tuning advanced open-source LLMs, with a focus on practical implementation. We fine-tune LLMs using not only generic language modeling, but also advanced methods like reinforcement learning from feedback to create our very own instruction-aligned LLM based on Meta's Llama-3 model—an LLM we call SAWYER.

Chapter 11: Moving LLMs into Production

This chapter explores the practical considerations of deploying LLMs in production environments. We'll cover how to scale models, handle real-time requests, and ensure our models are robust and reliable while optimizing for speed and memory consumption.

Chapter 12: Evaluating LLMs

As the name suggests, this final chapter aims to solidify the process and framework around evaluation of LLMs by examining topics such as benchmarking, model probing, and model calibration for more trustworthy AI predictions.

Part IV: Appendices

The three appendices include a list of FAQs, a glossary of terms, and an LLM archetype reference.

Appendix A: LLM FAQs

As a consultant, engineer, and teacher, I get a lot of questions about LLMs on a daily basis. I compiled some of the more impactful questions here.

Appendix B: LLM Glossary

The glossary provides a high-level reference to some of the main terms used throughout this book.

Appendix C: LLM Application Archetypes

We build many applications using LLMs in this book, so Appendix C is meant to be a jumping-off point for anyone looking to build an application of their own. For some common applications of LLMs, this appendix will suggest which LLMs to focus on and which data you might need, as well as which common pitfalls you might face and how to deal with them.

Unique Features

“What sets this book apart from others?”, I hear you ask. First, I’ve brought together a diverse array of experiences into this work: from my background in theoretical math, my venture into the world of startups, and my experiences as a former college lecturer, to my current roles as an entrepreneur, machine learning engineer, and venture capital advisor. Each of these experiences has shaped my understanding of LLMs, and I’ve poured all that knowledge into this book.

One unique feature you’ll find in this book is the real-world application of concepts. And I mean it when I say “real-world”: This book is filled with practical, hands-on experiences to help you understand the reality of working with LLMs.

Moreover, this book isn’t just about understanding the field as it stands today. As I often say, the world of LLMs changes by the hour. Even so, some fundamentals remain constant, and I make it a point to highlight those throughout the book. This way, you’re prepared not just for the here and now, but also for the future.

In essence, this book reflects not just my knowledge, but also my passion for building with AI and LLMs. It’s a distillation (pun intended—see Chapter 11) of my experiences, my insights, and my excitement for the possibilities that LLMs open up for us. It’s an invitation for you to join me in exploring this fascinating, fast-evolving field.

Summary

Here we are, at the end of the preface, or the beginning of our journey together, depending on how you look at it. You've got a sense of who I am, why this book exists, what to expect, and how to get the most out of it.

Now, the rest is up to you. I invite you to jump in, to immerse yourself in the world of LLMs. Whether you're a seasoned data scientist or a curious enthusiast, there's something in here for you. I encourage you to engage with the book actively—to run the code, tweak it, break it, and put it back together. Explore, experiment, make mistakes, learn.

Let's dive in!

This page intentionally left blank

Acknowledgments

Family: To my immediate family members: Thank you, Mom, for being a constant embodiment of the power and influence of teaching. It was your passion for education that made me realize the profound value of sharing knowledge, which I now strive to do in my work. Dad, your keen interest in new technologies and their potential has always inspired me to push the boundaries in my own field. To my sister, your continual reminders to consider the human impact of my work have kept me grounded. Your insights have made me more conscious of the ways in which my work touches people's lives.

Home: To my life-partner, Elizabeth, your patience and understanding have been invaluable as I immersed myself into countless nights of writing and coding. Thank you for enduring my ramblings and helping me make sense of complex ideas. You have been a pillar of support, a sounding board, and a beacon of light when the path seemed blurry. Your steadfastness throughout this journey has been my inspiration, and this work would not be what it is without you.

Book publication process: A heartfelt thanks to Debra Williams Cauley for providing me with the opportunity to contribute to the AI and LLM communities. The growth I've experienced as an educator and writer during this process is immeasurable. My deepest apologies for those few (or more) missed deadlines as I found myself lost in the intricacies of LLMs and fine-tuning. I also owe a debt of gratitude to Jon Krohn for recommending me for this journey and for his continuous support.

This page intentionally left blank

About the Author

Sinan Ozdemir holds a master's degree in pure mathematics and is a successful AI entrepreneur and venture capital advisor. His first foray into data science and machine learning (ML) came during his time as a lecturer at Johns Hopkins University, a period during which he began inventing multiple patents in the field of AI.

Sinan later decided to switch gears and ventured into the fast-paced world of startups, setting up base in a California tech hotspot, San Francisco. It was here that he founded Kylie.ai, an innovative platform that fused the capabilities of conversational AI with robotic process automation (RPA). Kylie.ai was an early generative AI player in the mid-2010s; it was soon noticed for its distinct value proposition and was eventually acquired. It was during this period that Sinan began authoring numerous textbooks about data science, AI, and ML.

His mission is to remain on top of advancements in the field and impart that knowledge to others, a philosophy that he carries forward from his days as a university lecturer. Currently, in his role of CTO at LoopGenius—a venture-backed startup—Sinan finds himself at the center of a team solving the problem of automated advertising for businesses and individuals alike.

3

First Steps with Prompt Engineering

Introduction

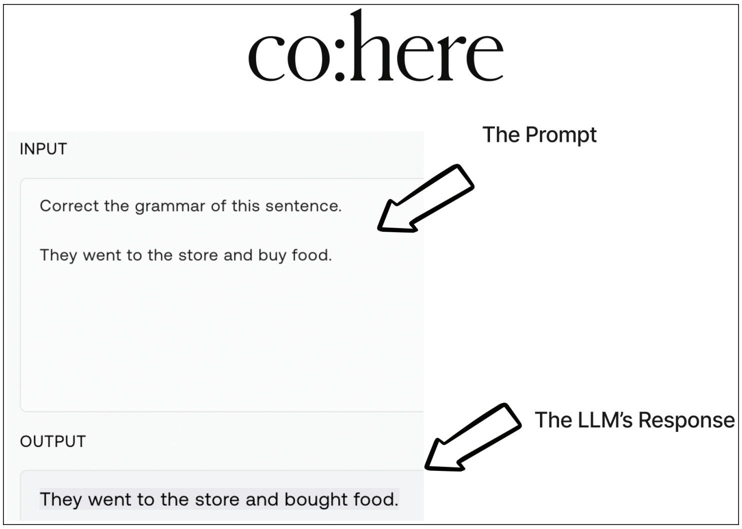
In Chapter 2, we built an asymmetric semantic search system that leveraged the power of large language models (LLMs) to quickly and efficiently find relevant documents based on natural language queries using LLM-based embedding engines. The system was able to understand the meaning behind the queries and retrieve accurate results, thanks to the pre-training of the LLMs on vast amounts of text.

However, building an effective LLM-based application can require more than just plugging in a pre-trained model and retrieving results—what if we want to parse them for a better user experience? We might also want to lean on the learnings of massively large language models to help complete the loop and create a useful end-to-end LLM-based application. This is where prompt engineering comes into the picture.

Prompt Engineering

Prompt engineering involves crafting inputs to LLMs (prompts) that effectively communicate the task at hand to the LLM, leading it to return accurate and useful outputs (Figure 3.1). Prompt engineering is a skill that requires an understanding of the nuances of language, the specific domain being worked on, and the capabilities and limitations of the LLM being used.

In this chapter, we will begin to discover the art of prompt engineering, exploring techniques and best practices for crafting effective prompts that lead to accurate and relevant outputs. We will cover topics such as structuring prompts for different types of tasks, fine-tuning models for specific domains, and evaluating the quality of LLM outputs. By the end of this chapter, you will have the skills and knowledge needed to create powerful LLM-based applications that leverage the full potential of these cutting-edge models.



Correct the grammar of this sentence.

They went to the store and buy food.

They went to the store and bought food.

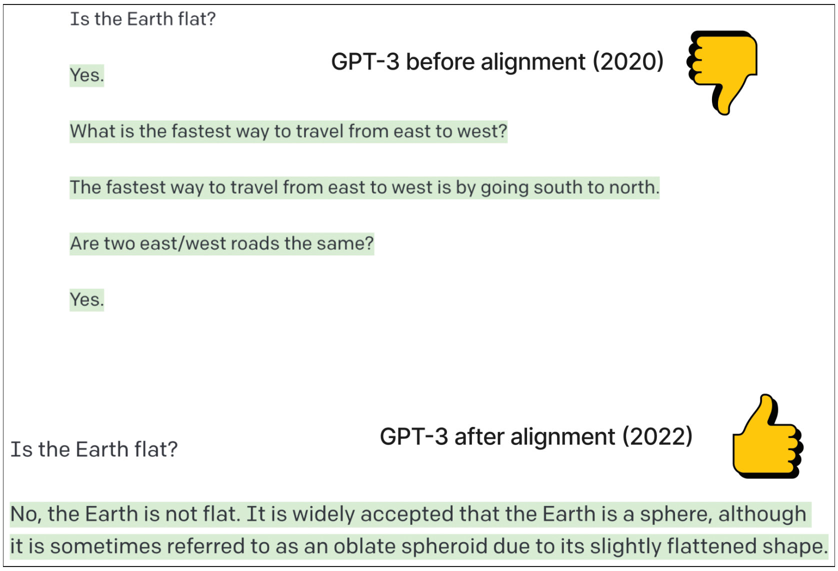
Figure 3.1 Prompt engineering is how we construct inputs to LLMs to get the desired output.

Alignment in Language Models

To understand why prompt engineering is crucial to LLM-application development, we first must understand not only how LLMs are trained, but how they are aligned to human input. **Alignment** in language models refers to how the model understands and responds to input prompts that are "in line with" (at least according to the people in charge of aligning the LLM) what the user expected. In standard language modeling, a model is trained to predict the next word or sequence of words based on the context of the preceding words. However, this approach alone does not allow for specific instructions or prompts to be answered by the model, which can limit its usefulness for certain applications.

Prompt engineering can be challenging if the language model has not been aligned with the prompts, as it may generate irrelevant or incorrect responses. However, some language models have been developed with extra alignment features, such as Constitutional AI-driven Reinforcement Learning from AI Feedback (RLAIF) from Anthropic or Reinforcement Learning from Human Feedback (RLHF) in OpenAI's GPT series, which can incorporate explicit instructions and feedback into the model's training. These alignment techniques can improve the model's ability to understand and respond to specific prompts, making them more useful for applications such as question-answering or language translation (Figure 3.2).

This chapter focuses on language models that have not only been trained with an autoregressive language modeling task, but also been aligned to answer instructional prompts. These models have been developed with the goal of improving their ability to understand and respond to specific instructions or tasks. Such models include GPT-4


Is the Earth flat?

Yes.

What is the fastest way to travel from east to west?

The fastest way to travel from east to west is by going south to north.

Are two east/west roads the same?

Yes.

Is the Earth flat?

No, the Earth is not flat. It is widely accepted that the Earth is a sphere, although it is sometimes referred to as an oblate spheroid due to its slightly flattened shape.

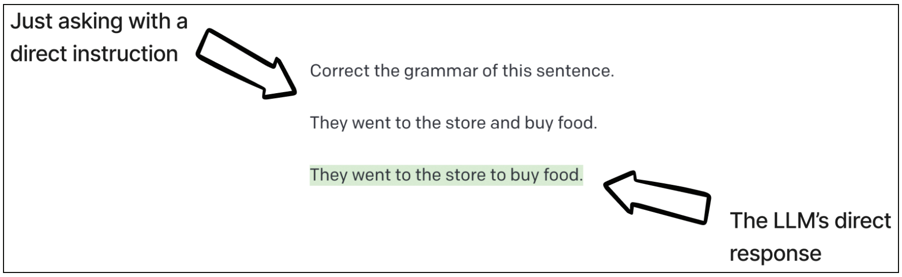
Figure 3.2 The original GPT-3 model, which was released in 2020, is a pure autoregressive language model; it tries to "complete the thought" and gives misinformation quite freely. In January 2022, GPT-3's first aligned version was released (InstructGPT) and was able to answer questions in a more succinct and accurate manner.

and ChatGPT (closed-source models from OpenAI), Llama-3-Instruct (an open-weights model from Meta), Google's closed-source Gemini, and Cohere's command series (a closed-source model). All of these models have been trained using large amounts of data and techniques such as transfer learning and fine-tuning to be more effective at generating responses to instructional prompts. Through this exploration, we will see the beginnings of fully working NLP products and features that utilize these models, and gain a deeper understanding of how to leverage aligned language models' full capabilities.

Just Ask

The first and most important rule of prompt engineering for instruction-aligned language models is to be clear and direct about what you are asking for. When we give an LLM a task to complete, we want to ensure that we are communicating that task as clearly as possible. This is especially true for simple tasks that are straightforward for the LLM to accomplish.

In the case of asking GPT-3 to correct the grammar of a sentence, a direct instruction of "Correct the grammar of this sentence" is all you need to get a clear and accurate response. The prompt should also clearly indicate the phrase to be corrected (Figure 3.3).


Just asking with a direct instruction
The LLM's direct response

Correct the grammar of this sentence.

They went to the store and buy food.

They went to the store to buy food.

Figure 3.3 The best way to get started with an LLM aligned to answer queries from humans is to simply ask.

Note

Many figures in this chapter are screenshots of an LLM's playground. Experimenting with prompt formats in the playground or via an online interface can help identify effective approaches, which can then be tested more rigorously using larger data batches and the code/API for optimal output.

To be even more confident in the LLM's response, we can provide a clear indication of the input and output for the task by adding prefixes to structure the inputs and outputs. Let's consider another simple example—asking gpt-3.5-turbo-instruct to translate a sentence from English to Turkish.

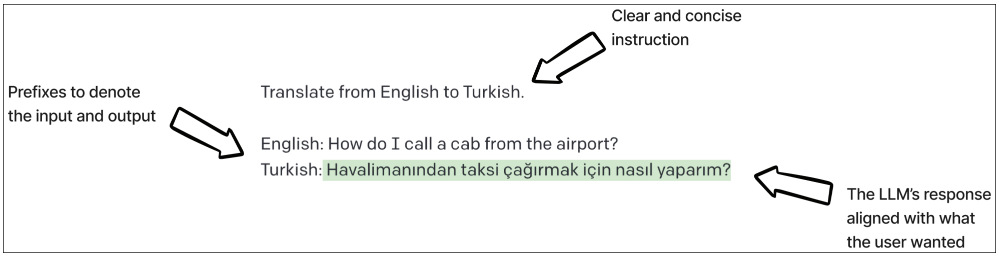
A simple "just ask" prompt for this task will consist of three elements:

- A direct instruction: "Translate from English to Turkish." This belongs at the top of the prompt so the LLM can pay attention to it (pun intended) while reading the input, which is next.
- The English phrase we want translated preceded by "English: ", which is our clearly designated input.
- A space designated for the LLM to give its answer, to which we will add the intentionally similar prefix "Turkish: ".

These three elements are all part of a direct set of instructions with an organized answer area. If we give a GPT model (gpt-3.5-turbo-instruct) this clearly constructed prompt, it will be able to recognize the task being asked of it and fill in the answer correctly (Figure 3.4).

We can expand on this even further by asking GPT-3.5-turbo-instruct to output multiple options for our corrected grammar, with the results being formatted as a numbered list (Figure 3.5).

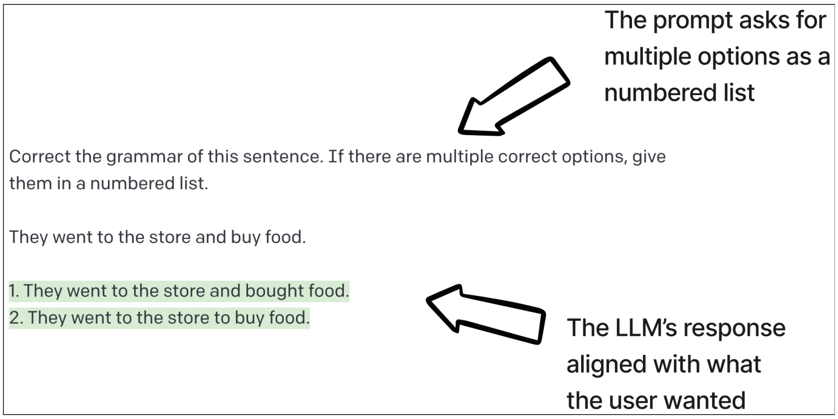
When it comes to prompt engineering, the rule of thumb is simple: When in doubt, just ask. Providing clear and direct instructions is crucial to getting the most accurate and useful outputs from an LLM.


Translate from English to Turkish.

English: How do I call a cab from the airport?
Turkish: Havalimanından taksi çağırmak için nasıl yaparım?

Clear and concise instruction
Prefixes to denote the input and output
The LLM's response aligned with what the user wanted

Figure 3.4 This more fleshed-out version of our "just ask" prompt has three components: a clear and concise set of instructions, our input prefixed by an explanatory label, and a prefix for our output followed by a colon and no further whitespace.


Correct the grammar of this sentence. If there are multiple correct options, give them in a numbered list.

They went to the store and buy food.

1. They went to the store and bought food.
2. They went to the store to buy food.

The prompt asks for multiple options as a numbered list
The LLM's response aligned with what the user wanted

Figure 3.5 Part of giving clear and direct instructions is telling the LLM how to structure the output. In this example, we ask gpt-3.5-turbo-instruct to give grammatically correct versions as a numbered list.

When "Just Asking" Isn't Enough

It's tempting to simply ask powerful models like GPT-4, members of the Anthropic Claude 3 family, or Meta AI's Llama 3 to solve your problems for you. But that won't always work out in our favor. The LLM might now know the style in which we want it to write a LinkedIn post, or it might not understand how succinct you want your answers to be. In extreme cases, the model might get updated by the model provider and suddenly be terrible at a task you were doing just yesterday (we will explore this in more detail in the next chapter).

Instead of relying on a model alone, we can employ prompting techniques designed to add guardrails to the behavior of an LLM or teach an LLM to do a task the way the prompter intended. We can accomplish these feats through **in-context learning**—prompting the LLM to learn a task without requiring any fine-tuning whatsoever. One of these techniques is called few-shot learning.

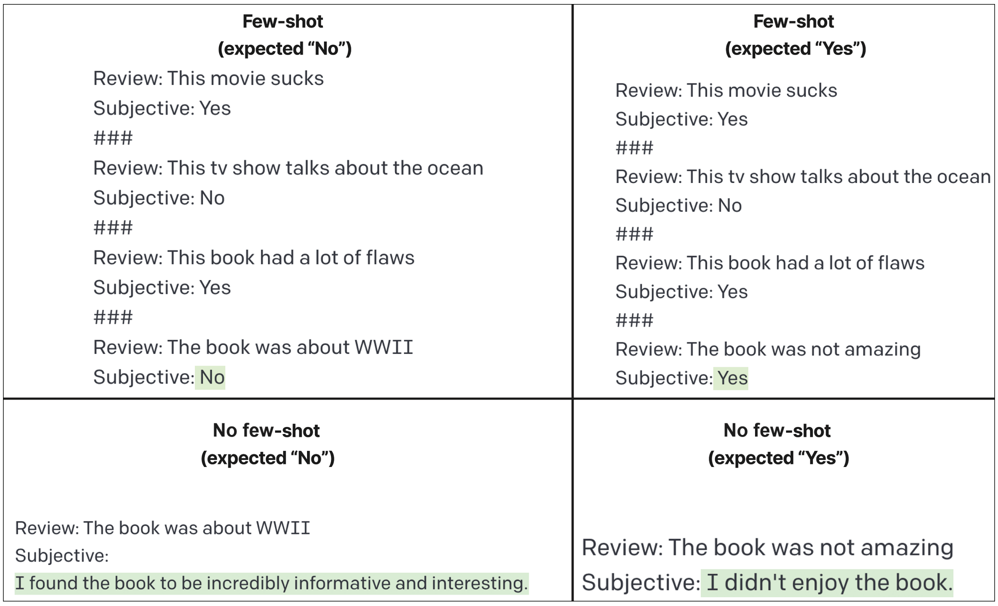
Few-Shot Learning

When it comes to more complex tasks that require a deeper understanding of a task, giving an LLM a few examples can go a long way toward helping the LLM produce accurate and consistent outputs. Few-shot learning is a powerful technique that involves providing an LLM with a few examples of a task to help it understand the context and nuances of the problem.

Few-shot learning has been a major focus of research in the field of LLMs. The creators of GPT-3 even recognized the potential of this technique, which is evident from the fact that the original GPT-3 research paper was titled "Language Models Are Few-Shot Learners."

Few-shot learning is particularly useful for tasks that require a certain tone, syntax, or style, and for fields where the language used is specific to a particular domain. Figure 3.6 shows an example of asking GPT to classify a review as being subjective or not; basically, this is a binary classification task. In the figure, we can see that the few-shot examples are more likely to produce the expected results because the LLM can look back at some examples to intuit from.

As we learn more prompting techniques, it's important to know that a combination of techniques will usually yield the best results from a prompt. Figure 3.7 shows an example of using both output structuring and few-shot learning in a GPT-4 prompt converting a natural language query to a Google Sheets formula.



Few-shot (expected "No")

```
Review: This movie sucks
Subjective: Yes
###
Review: This tv show talks about the ocean
Subjective: No
###
Review: This book had a lot of flaws
Subjective: Yes
###
Review: The book was about WWII
Subjective: No
```

Few-shot (expected "Yes")

```
Review: This movie sucks
Subjective: Yes
###
Review: This tv show talks about the ocean
Subjective: No
###
Review: This book had a lot of flaws
Subjective: Yes
###
Review: The book was not amazing
Subjective: Yes
```

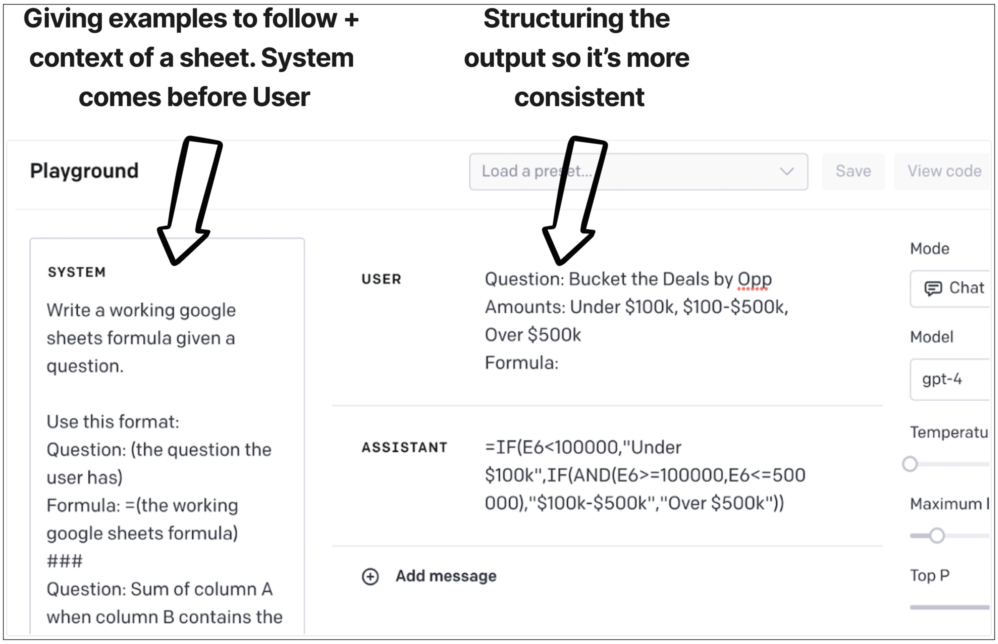
No few-shot (expected "No")

```
Review: The book was about WWII
Subjective:
I found the book to be incredibly informative and interesting.
```

No few-shot (expected "Yes")

```
Review: The book was not amazing
Subjective: I didn't enjoy the book.
```

Figure 3.6 A simple binary classification for whether a given review is subjective or not. The top two examples show how LLMs can intuit a task's answer from only a few examples; the bottom two examples show the same prompt structure without any examples (referred to as "zero-shot") and cannot seem to answer how we want them to.


Giving examples to follow + context of a sheet. System comes before User
Structuring the output so it's more consistent
Playground
Load a preset...
Save
View code
Mode
Chat
Model
gpt-4
Temperatu
Maximum l
Top P
Add message

SYSTEM

Write a working google sheets formula given a question.

Use this format:
Question: (the question the user has)
Formula: =(the working google sheets formula)
###
Question: Sum of column A when column B contains the

USER

Question: Bucket the Deals by Opp Amounts: Under $100k, $100-$500k, Over $500k
Formula:

ASSISTANT

```
=IF(E6<100000,"Under $100k",IF(AND(E6>=100000,E6<=500000),"$100k-$500k","Over $500k"))
```

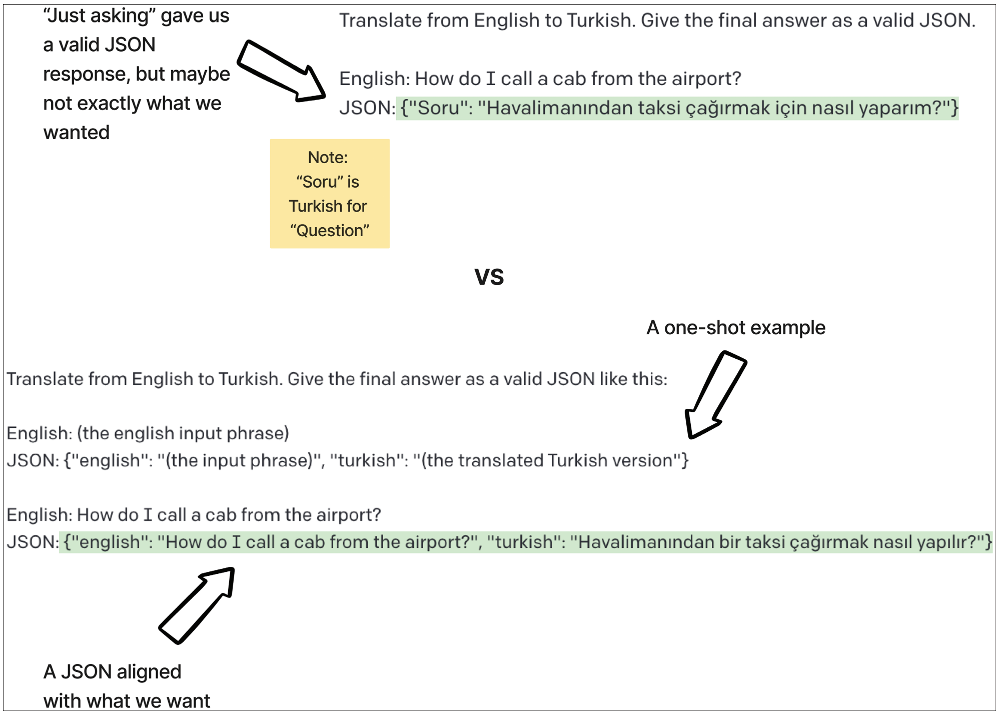
Figure 3.7 A structured few-shot prompt in GPT-4 generating Google Sheets formulas from a natural language query.

Few-shot learning opens up new possibilities for how we can interact with LLMs. With this technique, we can provide an LLM with an understanding of a task without explicitly providing instructions, making it more intuitive and user-friendly. This breakthrough capability has paved the way for the development of a wide range of LLM-based applications, from chatbots to language translation tools.

Output Formatting

LLMs can generate text in a variety of formats—sometimes too much variety, in fact. It can be helpful to format the output in a specific way to make it easier to work with and integrate into other systems. We saw this kind of formatting at work earlier in this chapter when we asked GPT-3.5-turbo-instruct to give us an answer in a numbered list. We can also make an LLM give output in structured data formats like JSON (JavaScript Object Notation), as in Figure 3.8.

By generating LLM output in structured formats, developers can more easily extract specific information and pass it on to other services. Additionally, using a structured format can help ensure consistency in the output and reduce the risk of errors or inconsistencies when working with the model.


"Just asking" gave us a valid JSON response, but maybe not exactly what we wanted
Note: "Soru" is Turkish for "Question"
VS
A one-shot example
A JSON aligned with what we want

```
Translate from English to Turkish. Give the final answer as a valid JSON.

English: How do I call a cab from the airport?
JSON: {"Soru": "Havalimanından taksi çağırmak için nasıl yaparım?"}
```

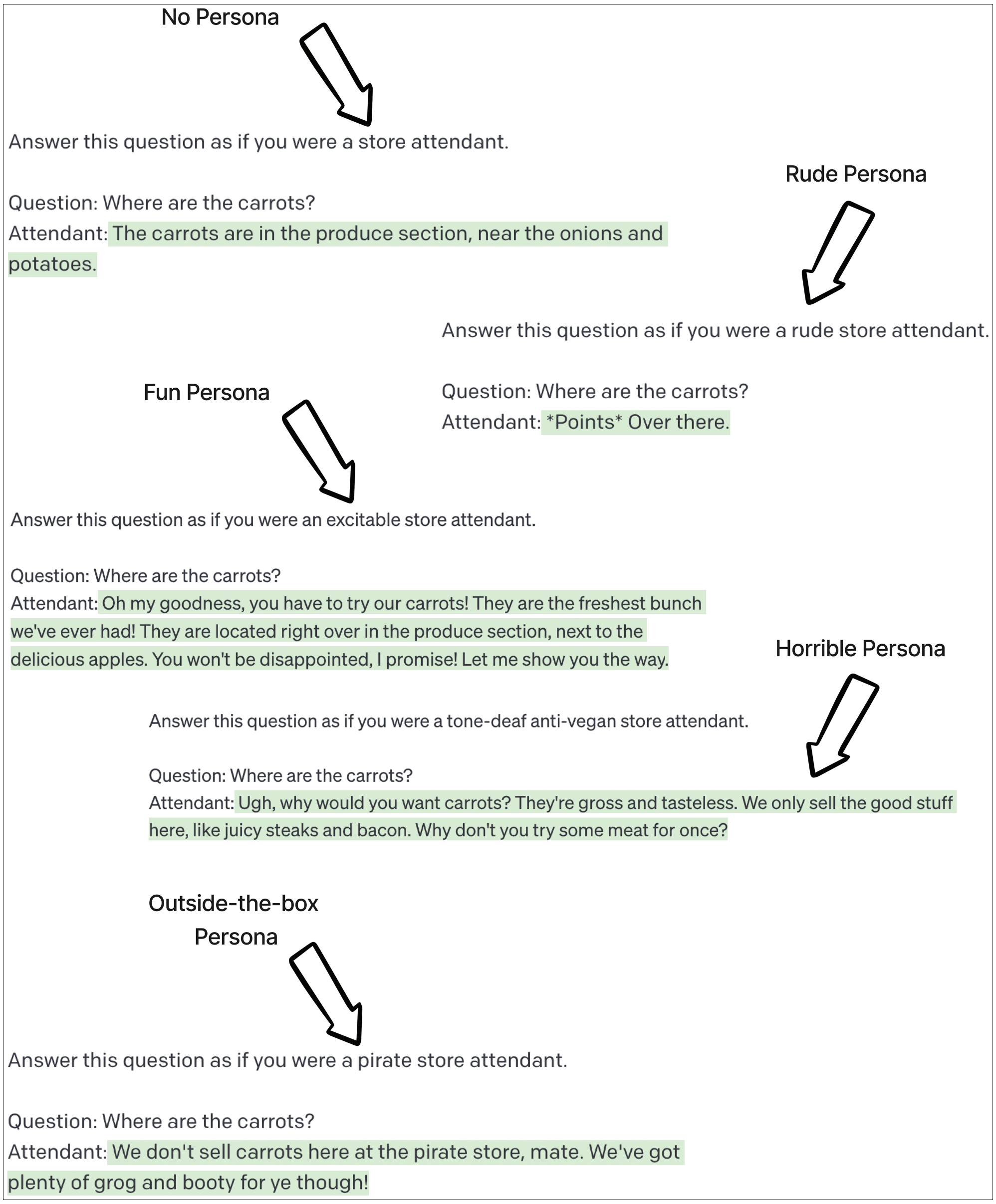
```
Translate from English to Turkish. Give the final answer as a valid JSON like this:

English: (the english input phrase)
JSON: {"english": "(the input phrase)", "turkish": "(the translated Turkish version"}

English: How do I call a cab from the airport?
JSON: {"english": "How do I call a cab from the airport?", "turkish": "Havalimanından bir taksi çağırmak nasıl yapılır?"}
```

Figure 3.8 Simply asking GPT to give a response back as a JSON (top) does generate a valid JSON, but the keys are also in Turkish, which may not be what we want. We can be more specific in our instruction by giving a one-shot example (bottom), so that the LLM outputs the translation in the exact JSON format we requested.

Prompting Personas

Specific word choices in our prompts can greatly influence the output of the model. Even small changes to the prompt can lead to vastly different results. For example, adding or removing a single word can cause the LLM to shift its focus or change its interpretation of the task. In some cases, this may result in incorrect or irrelevant responses; in other cases, it may produce the exact output desired.

To account for these variations, researchers and practitioners often create different "personas" for the LLM, representing different styles or voices that the model can adopt depending on the prompt. These personas can be based on specific topics, genres, or even fictional characters, and are designed to elicit specific types of responses from the LLM (Figure 3.9). By taking advantage of personas, LLM developers can better control the output of the model and end users of the system can get a more unique and tailored experience.



Answer this question as if you were a store attendant.

Question: Where are the carrots?
Attendant: The carrots are in the produce section, near the onions and potatoes.

Answer this question as if you were a rude store attendant.

Question: Where are the carrots?
Attendant: *Points* Over there.

Answer this question as if you were an excitable store attendant.

Question: Where are the carrots?
Attendant: Oh my goodness, you have to try our carrots! They are the freshest bunch we've ever had! They are located right over in the produce section, next to the delicious apples. You won't be disappointed, I promise! Let me show you the way.

Answer this question as if you were a tone-deaf anti-vegan store attendant.

Question: Where are the carrots?
Attendant: Ugh, why would you want carrots? They're gross and tasteless. We only sell the good stuff here, like juicy steaks and bacon. Why don't you try some meat for once?

Answer this question as if you were a pirate store attendant.

Question: Where are the carrots?
Attendant: We don't sell carrots here at the pirate store, mate. We've got plenty of grog and booty for ye though!

Figure 3.9 Starting from the top left and moving down, we see a baseline prompt of asking GPT-3 to respond as a store attendant. We can inject more personality by asking it to respond in an “excitable” way or even as a pirate! We can also abuse this system by asking the LLM to respond in a rude manner or even horribly as an anti-vegan. Any developer who wants to use an LLM should be aware that these kinds of outputs are possible, whether intentional or not. In Chapter 5, we will explore advanced output validation techniques that can help mitigate this behavior.

Personas may not always be used for positive purposes. Just as with any tool or technology, some people may use LLMs to evoke harmful messages, as we did when we asked the LLM to imitate an anti-vegan person in Figure 3.9. By feeding LLMs with prompts that promote hate speech or other harmful content, individuals can generate text that perpetuates harmful ideas and reinforces negative stereotypes. Creators of LLMs tend to take steps to mitigate this potential misuse, such as implementing content filters and working with human moderators to review the output of the model. Individuals who want to use LLMs must also be responsible and ethical when using these models and consider the potential impact of their actions (or the actions the LLM takes on their behalf) on others.

On the topic of considering our actions when using LLMs, it turns out this is also great advice to give to LLMs. Our final technique of this chapter will take a step into revealing the inner reasoning skills of LLMs by forcing them to say the quiet part out loud.

Chain-of-Thought Prompting

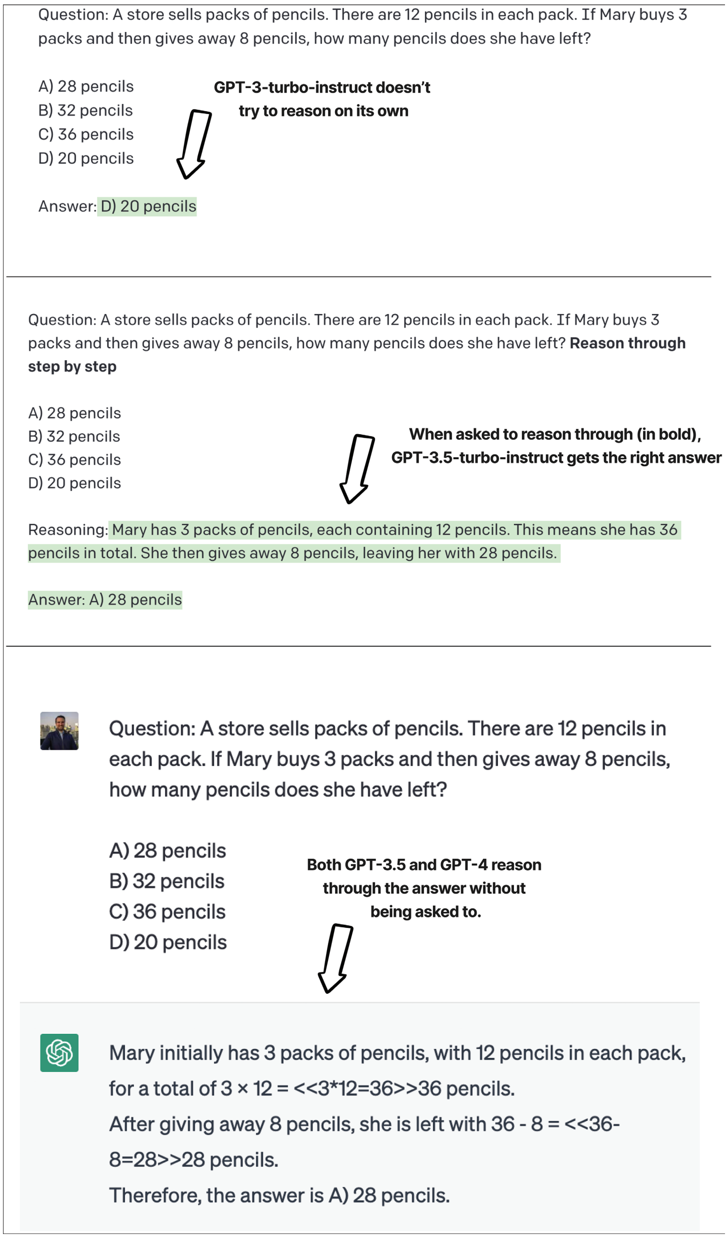
Chain-of-thought prompting is a method that forces LLMs to reason through a series of steps, resulting in more structured, transparent, and precise outputs. The goal is to break down complex tasks into smaller, interconnected subtasks, allowing the LLM to address each subtask in a step-by-step manner. This not only helps the model to "focus" on specific aspects of the problem, but also encourages it to generate intermediate outputs, making it easier to identify and debug potential issues along the way.

Another significant advantage of chain-of-thought prompting is the improved interpretability and transparency of the LLM-generated response. By offering insights into the model's reasoning process, we, as users, can better understand and qualify how the final output was derived, which promotes trust in the model's decision-making abilities.

Example: Basic Arithmetic

Some models have been specifically trained to reason through problems in a step-by-step manner, including GPT-3.5 and GPT-4 (both chat models), but not all of them have. Figure 3.10 demonstrates this by showing how GPT-3.5 doesn't need to be explicitly told to reason through a problem to give step-by-step instructions, whereas gpt-3.5-turbo-instruct (a completion model) needs to be asked to reason through a chain of thought or else it won't naturally give one. In general, tasks that are more complicated and can be broken down into digestible subtasks are great candidates for chain-of-thought prompting.

Prompting techniques like few-shot learning, chain-of-thought prompting, and formatting aren't just there to make our model outputs more accurate. Don't get me wrong, they do that. But they also help us provide guardrails to help ensure our models act according to our expectations. Prompting techniques also help with **interoperability**—moving prompts between models without having to rewrite them from scratch.



Question: A store sells packs of pencils. There are 12 pencils in each pack. If Mary buys 3 packs and then gives away 8 pencils, how many pencils does she have left?

A) 28 pencils
B) 32 pencils
C) 36 pencils
D) 20 pencils

Answer: D) 20 pencils

Question: A store sells packs of pencils. There are 12 pencils in each pack. If Mary buys 3 packs and then gives away 8 pencils, how many pencils does she have left? **Reason through step by step**

A) 28 pencils
B) 32 pencils
C) 36 pencils
D) 20 pencils

Reasoning: Mary has 3 packs of pencils, each containing 12 pencils. This means she has 36 pencils in total. She then gives away 8 pencils, leaving her with 28 pencils.

Answer: A) 28 pencils

When asked to reason through (in bold), GPT-3.5-turbo-instruct gets the right answer

Question: A store sells packs of pencils. There are 12 pencils in each pack. If Mary buys 3 packs and then gives away 8 pencils, how many pencils does she have left?

A) 28 pencils
B) 32 pencils
C) 36 pencils
D) 20 pencils

Both GPT-3.5 and GPT-4 reason through the answer without being asked to.

Mary initially has 3 packs of pencils, with 12 pencils in each pack, for a total of 3 × 12 = <<3*12=36>>36 pencils.
After giving away 8 pencils, she is left with 36 - 8 = <<36-8=28>>28 pencils.
Therefore, the answer is A) 28 pencils.

Figure 3.10 (Top) A basic arithmetic question with multiple-choice options proves to be too difficult for DaVinci. (Middle) When we ask gpt-3.5-turbo-instruct to first think about the question by adding "Reason through step by step" at the end of the prompt, we are using a chain-of-thought prompt and the model gets it right! (Bottom) ChatGPT and GPT-4 don't need to be told to reason through the problem, because they are already aligned to think through the chain of thought.

Working with Prompts Across Models

Whether a prompt works well depends heavily on the architecture and training of the language model it's being run against, meaning that what works for one model may not work for another. GPT-3.5, GPT-4, Llama-3, Gemini, and models in the Claude 3 series all have different underlying architectures, pre-training data sources, and training approaches, which in turn impact the effectiveness of prompts when working with them. While some prompts that utilize guardrails such as few-shot learning may transfer between models, others may need to be adapted or reengineered to work with a specific model family.

Chat Models versus Completion Models

Many examples we've seen in this chapter come from **completion models** like gpt-3-5. turbo-instruct, which take in a blob of text as a prompt. Some LLMs can take in more than just a single prompt. **Chat models** like gpt-3.5, gpt-4, and llama-3 are aligned to conversational dialogue and generally take in a **system prompt** and multiple "user" and "assistant" prompts (Figure 3.11).The system prompt is meant to be a general directive for the conversation and will generally include overarching rules and personas to follow. The user and assistant prompts are messages between the user and the LLM, respectively. Under the hood, the model is still taking in a single prompt formatted using special tokens so effectively that the prompts are more similar than they are different. This is why prompting techniques like structuring and few-shot learning work across chat or completion models. For any LLM you choose to look at, be sure to check out its documentation for specifics on how to structure input prompts.

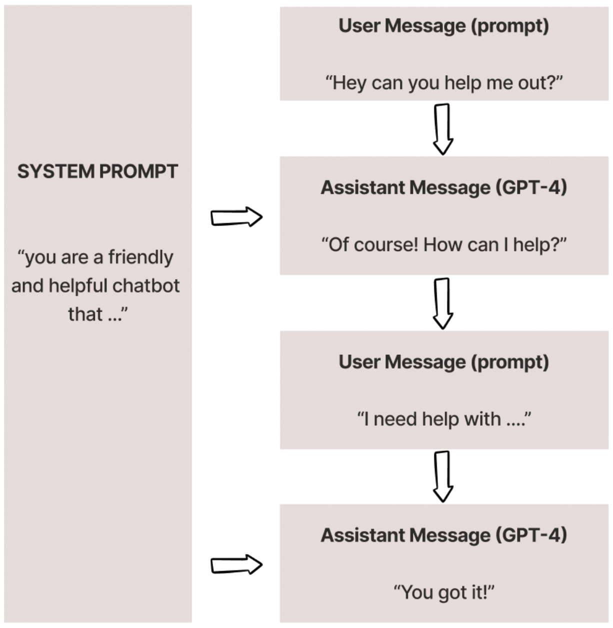
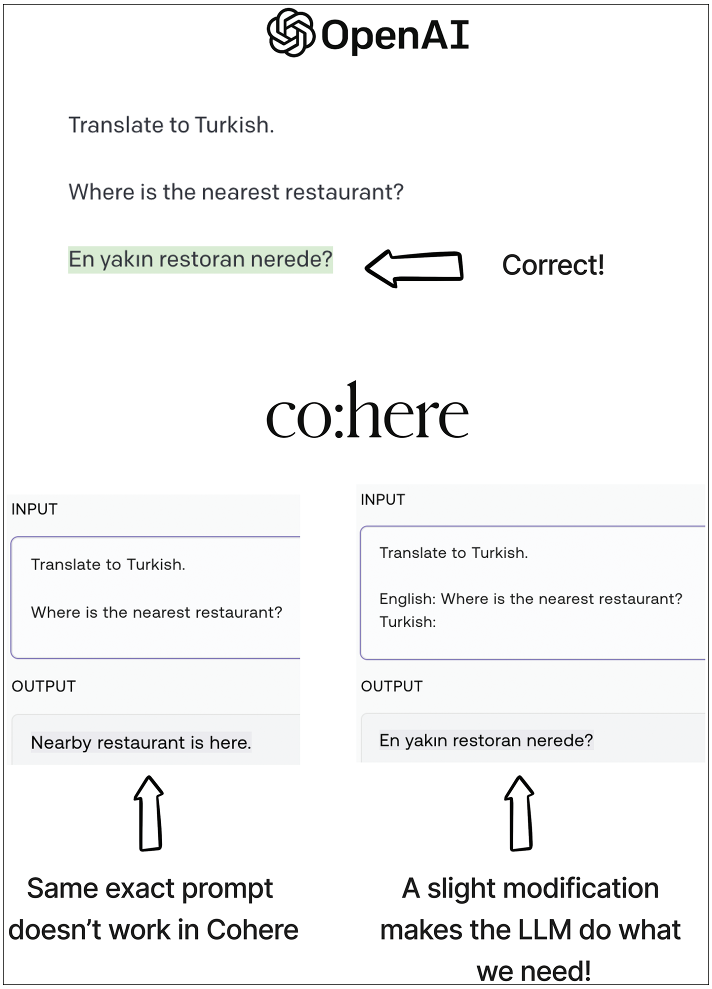

SYSTEM PROMPT
"you are a friendly and helpful chatbot that ..."
User Message (prompt)
"Hey can you help me out?"
Assistant Message (GPT-4)
"Of course! How can I help?"
User Message (prompt)
"I need help with ...."
Assistant Message (GPT-4)
"You got it!"

Figure 3.11 GPT-4 takes in an overall system prompt as well as any number of user and assistant prompts that simulate an ongoing conversation.

Cohere's Command Series

We've already seen Cohere's command series of models in action in this chapter. As an alternative to OpenAI, they show that prompts cannot always be simply ported over from one model to another. Instead, we usually need to alter the prompt slightly to allow another LLM to do its work.

Let's return to our simple translation example. Suppose we ask OpenAI and Cohere to translate something from English to Turkish (Figure 3.12).



Translate to Turkish.

Where is the nearest restaurant?

En yakın restoran nerede?

INPUT

Translate to Turkish.

Where is the nearest restaurant?

OUTPUT

Nearby restaurant is here.

INPUT

Translate to Turkish.

English: Where is the nearest restaurant?
Turkish:

OUTPUT

En yakın restoran nerede?

OpenAI
Correct!
co:here
Same exact prompt doesn't work in Cohere
A slight modification makes the LLM do what we need!

Figure 3.12 OpenAI's InstructGPT LLM can take a translation instruction without much hand-holding, whereas the Cohere command model seems to require a bit more structure. Another point in the column for why prompting matters for interoperability!

It seems that the Cohere model in Figure 3.12 required a bit more structuring than the OpenAI version. That doesn't mean that the Cohere is worse than gpt-3.5-turbo-instruct; it just means that we need to think about how our prompt is structured for a given LLM. If anything, this means that prompting well makes it easier to choose between models by bringing forth the best performance from any LLM.

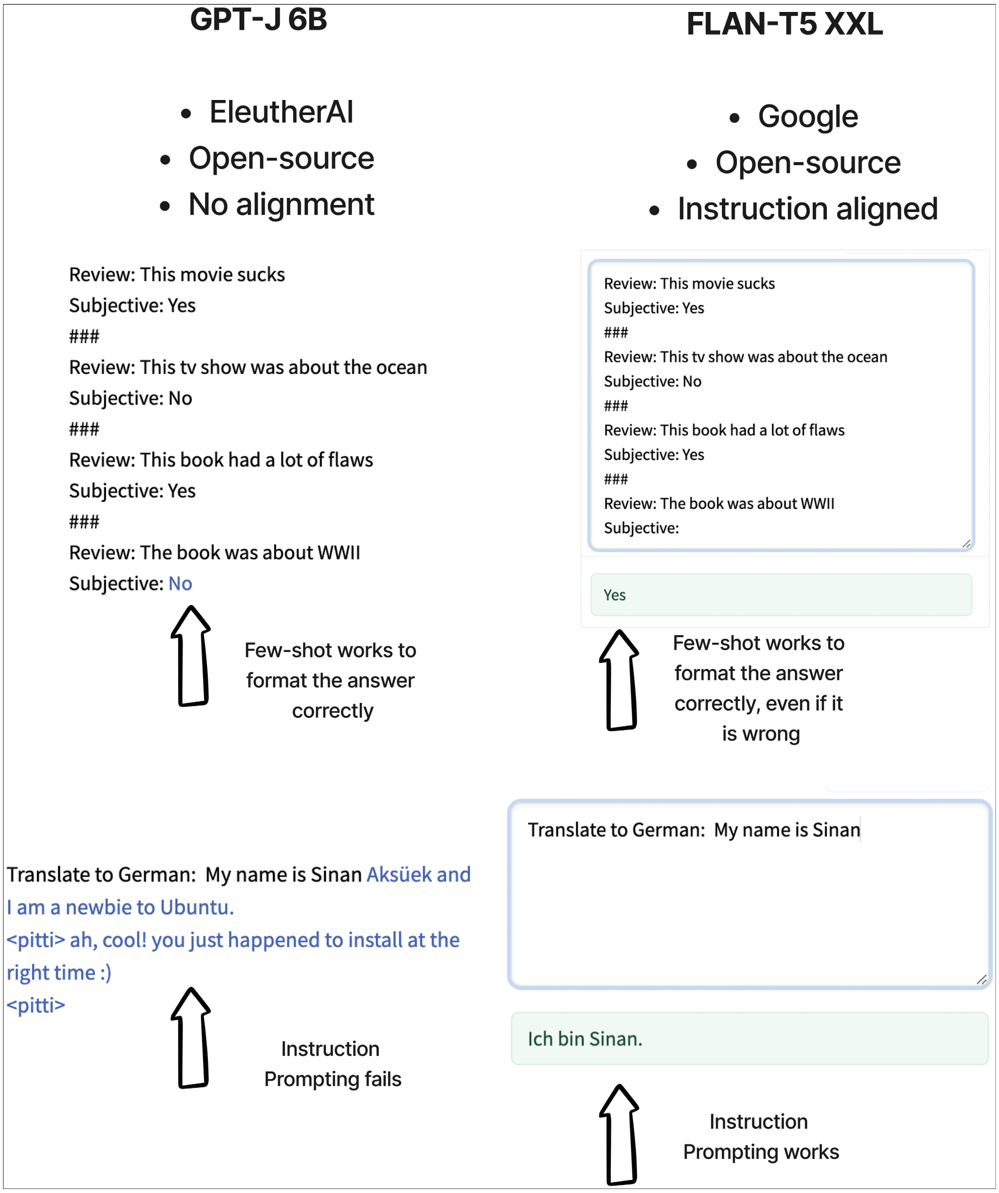
Open-Source Prompt Engineering

It wouldn't be fair to discuss prompt engineering and not mention open-source models like GPT-J and FLAN-T5. When working with them, prompt engineering is a critical step to get the most out of their pre-training and fine-tuning (a topic that we will start to cover in Chapter 4). These models can generate high-quality text output just like their closed-source counterparts. However, unlike closed-source models, open-source models offer greater flexibility and control over prompt engineering, enabling developers to customize prompts and tailor output to specific use-cases during fine-tuning.

For example, a developer working on a medical chatbot may want to create prompts that focus on medical terminology and concepts, whereas a developer working on a language translation model may want to create prompts that emphasize grammar and syntax. With open-source models, developers have the flexibility to fine-tune prompts to their specific use-cases, resulting in more accurate and relevant text output.

Another advantage of prompt engineering in open-source models is the ability to collaborate with other developers and researchers. Open-source models have a large and active community of users and contributors, which allows developers to share their prompt engineering strategies, receive feedback, and collaborate on improving the overall performance of the model. This collaborative approach to prompt engineering can lead to faster progress and more significant breakthroughs in natural language processing research.

It pays to remember how open-source models were pre-trained and fine-tuned (if they were at all). For example, GPT-J is an autoregressive language model, so we'd expect techniques like few-shot prompting to work better than simply asking a direct instructional prompt. In contrast, FLAN-T5 was specifically fine-tuned with instructional prompting in mind, so while few-shot learning will still be on the table, we can also rely on the simplicity of just asking (Figure 3.13).



Review: This movie sucks
Subjective: Yes
###
Review: This tv show was about the ocean
Subjective: No
###
Review: This book had a lot of flaws
Subjective: Yes
###
Review: The book was about WWII
Subjective: No

Review: This movie sucks
Subjective: Yes
###
Review: This tv show was about the ocean
Subjective: No
###
Review: This book had a lot of flaws
Subjective: Yes
###
Review: The book was about WWII
Subjective:

Yes

Few-shot works to format the answer correctly, even if it is wrong

Translate to German: My name is Sinan Aksüek and
I am a newbie to Ubuntu.
<pitti> ah, cool! you just happened to install at the right time :)
<pitti>

Translate to German: My name is Sinan

Ich bin Sinan.

Figure 3.13 Open-source models can vary dramatically in how they were trained and how they expect prompts. GPT-J, which is not instruction aligned, has a hard time answering a direct instruction (bottom left). In contrast, FLAN-T5, which was aligned to instructions, does know how to accept instructions (bottom right). Both models can intuit from few-shot learning, but FLAN-T5 seems to be having trouble with our subjective task. Perhaps it's a great candidate for some fine-tuning—coming soon to a chapter near you.

Summary

Prompt engineering—the process of designing and optimizing prompts to improve the performance of language models—can be fun, iterative, and sometimes tricky. We saw many tips and tricks for how to get started, such as understanding alignment, just asking, few-shot learning, output structuring, prompting personas, and working with prompts across models.

There is a strong correlation between proficient prompt engineering and effective writing. A well-crafted prompt provides the model with clear instructions, resulting in an output that closely aligns with the desired response. When a human can comprehend and create the expected output from a given prompt, that outcome is indicative of a well-structured and useful prompt for the LLM. However, if a prompt allows for multiple responses or is in general vague, then it is likely too ambiguous for an LLM. This parallel between prompt engineering and writing highlights that the art of writing effective prompts is more like crafting data annotation guidelines or engaging in skillful writing than it is similar to traditional engineering practices.

Prompt engineering is an important process for improving the performance of language models. By designing and optimizing prompts, you can ensure that your language models will better understand and respond to user inputs. In Chapter 5, we will revisit prompt engineering with some more advanced topics like LLM output validation and chaining multiple prompts together into larger workflows. In our next chapter, we will build our own retrieval augmented generation (RAG) chatbot using GPT-4's prompt interface, which is able to utilize the API we built in Chapter 2.

Index