Prompt engineering techniques

Prompt 工程技術

07/02/2025

GPT-3, GPT-3.5, GPT-4, and GPT-4o models from OpenAI are prompt-based. With prompt-based models, the user interacts with the model by entering a text prompt, to which the model responds with a text completion. This completion is the model’s continuation of the input text. These techniques aren't recommended for o-series models.

GPT-3、GPT-3.5、GPT-4 和 GPT-4o 等 OpenAI 模型屬於以提示詞為基礎的模型。使用提示詞式模型時，使用者透过輸入文字提示與模型互動，模型會回應一段文字補完。這段補完是模型對輸入文字的延續。這些技術不建議用於 o 系列模型。

While these models are extremely powerful, their behavior is also very sensitive to the prompt. This makes prompt construction an important skill to develop.

儘管這些模型功能非常強大，但它們的行為對提示詞也極為敏感。因此，構建提示詞是一項重要且需要培養的技能。

Prompt construction can be difficult. In practice, the prompt acts to configure the model weights to complete the desired task, but it's more of an art than a science, often requiring experience and intuition to craft a successful prompt. The goal of this article is to help get you started with this learning process. It attempts to capture general concepts and patterns that apply to all GPT models. However it's important to understand that each model behaves differently, so the learnings might not apply equally to all models.

構建提示詞可能很困難。實務上，提示詞的作用是配置模型權重以完成所需任務，但這更像是一門藝術而非科學，常常需要經驗和直覺來設計成功的提示詞。本文的目標是協助你開始這個學習過程，嘗試歸納適用於所有 GPT 模型的一般概念與模式。但重要的是要了解每個模型的行為有所不同，因此所學可能無法同等適用於所有模型。

Basics 基礎知識

This section covers the basic concepts and elements of GPT prompts.

本節說明 GPT 提示詞的基本概念與要素。

Text prompts are how users interact with GPT models. As with all generative language models, GPT models attempt to produce the next series of words that are most likely to follow from the previous text. It's as if we're saying *What is the first thing that comes to your mind when I say* `<prompt>`*?* The examples below demonstrate this behavior. Given the first words of famous content, the model is able to accurately continue the text.

文字提示詞是使用者與 GPT 模型互動的方式。與所有生成式語言模型一樣，GPT 模型嘗試產生最有可能接續前文的下一串詞彙。這就像是在問：「當我說 `<prompt>` 時，你腦中首先浮現的是什麼？」下列範例示範了這種行為。給定著名內容的前幾個字，模型能夠準確地繼續該文本。

展開表格

Prompt 提示詞	**Completion 完成**
Four score and seven years ago our 八十七年前，我們的	fathers brought forth on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal. [...] 先賢在這片大陸上建立了一個新國家，始於自由的理念，並奉獻於「人人生而平等」這一主張。[...]
"Call me Ishmael." 「叫我以實瑪利。」	"Some years ago—never mind how long precisely—having little o no money in my purse, and nothing particular to interest me on shore, I thought I would sail about a little and see the watery part of the world." [...] 「幾年前——確切多久就不必計較了——我的口袋裡幾乎沒錢，岸上也沒有什麼特別能引起我興趣的事，我想我會出海航行一陣子，去看看那水域的世界。」[...]

As you develop more complex prompts, it's helpful to keep this fundamental behavior in mind. Regardless of the prompt that is provided, the model is simply responding with what it determines is most likely (given its training data and training targets). If you ask a question in your prompt, the model isn't following a separate “Q&A” code path, but rather it appears to answer the question because an answer is the most likely sort of response for the given question as input.

當你設計更複雜的提示時，記住這個基本行為是有幫助的。無論提供什麼提示，模型只是根據其訓練資料和訓練目標，回應它判定最有可能的結果。如果你在提示中提出一個問題，模型並不是走另一條「問答」的程式路徑，而是因為對於給定的問題作為輸入，回答是最有可能的回應類型，所以它看起來像是在回答問題。

Prompt components 提示組成要素

When using the Completion API while there's no differentiation between different parts of the prompt, it can still be useful for learning and discussion to identify underlying prompt components. With the Chat Completion API there are distinct sections of the prompt that are sent to the API in the form of an array of dictionaries with associated roles: system, user, and assistant. This guidance focuses more generally on how to think about prompt construction rather than providing prescriptive guidance that is specific to one API over another.

在使用 Completion API 時，雖然無法區分提示中不同的部分，但辨識出潛在的提示元件仍對學習與討論有幫助。使用 Chat Completion API 時，提示會以包含角色（system、user、assistant）的字典陣列形式傳送給 API，因此有明確的區段。本指南較偏重於一般性地思考提示構成，而非提供針對某一 API 的具體規範性建議。

It's also important to understand that while there could be other valid ways to dissect prompts, the goal of this breakdown is to provide a relatively simple way to think about prompt construction. With the Completion API, all the components are optional, but at least one must be present and most prompts include more than one component. There can be some grey area between components as well. The order presented below roughly corresponds to how commonly each component is used, from most to least.

同時也需理解，雖然還可能有其他有效的方式來拆解提示，但這種劃分的目標是提供一個相對簡單的方法來思考提示構成。對於 Completion API，所有元件都是可選的，但至少必須存在其中一個，而且大多數提示包含多於一個元件。元件之間也可能存在一些灰色地帶。下面所呈現的順序大致對應到各元件的使用頻率，從最常見到最不常見。

Instructions 指示

Instructions are likely the most commonly used prompt component. Instructions are simply that: instructions to the model on what to do. While simple in concept, they can be complex in practice. The following table illustrates this by showing both simple and more complex instructions, in two different forms.

指示很可能是最常使用的提示組件。指示就是它們本身：告訴模型要做什麼的指令。雖然概念上很簡單，但在實務上可能很複雜。下表透過以兩種不同形式展示簡單與較複雜的指示來說明這一點。

展開表格

Instruction Type 指示類型	Prompt 提示	Completion 完成
Simple instructions	Write an introduction for a weekly	Welcome to our weekly newsletter. This week we have a special focus

Instruction Type 指示類型	**Prompt 提示**	**Completion 完成**
簡單指示	newsletter. 為每週電子報撰寫一段介紹。	on productivity. We'll be sharing tips and tricks on how to be productive in all areas of your life. We hope you enjoy and please share with your friends! 歡迎閱讀我們的每週電子報。本週我們特別聚焦於生產力，將分享在生活各面向提升生產力的小技巧與策略。希望您喜歡，並請與朋友分享！

Instruction Type 指示類型	**Prompt 提示**	**Completion 完成**
Complex instructions 複雜指示	Write an introduction for a weekly newsletter to my company named Contoso. It should mention the great all hands meeting we had last week and thank the team for all their hard work over the tough past few months. 為我的公司 Contoso 撰寫每週電子報的導言。應提到我們上週舉行的精彩全體會議，並感謝團隊在過去這段艱難期間的辛勤付出。	Welcome to Contoso's weekly newsletter! We had a great all hands meeting last week and I want to thank the team for all their hard work over the tough past few months. We're looking forward to a bright future and I'm confident that with our team's dedication and commitment, we'll achieve great things. Thank you for being a part of our team and I look forward to working with you all in the coming weeks and months. Sincerely, [Your name]. 歡迎閱讀 Contoso 每週電子報！我們上週舉辦了一場精彩的全體會議，我想感謝團隊在過去這段艱難時期所付出的所有努力。我們期待光明的未來，我也相信在團隊的奉獻與承諾下，我們將達成偉大的成就。感謝你成為團隊的一員，我期待在未來數週和數個月與大家一同努力。此致，敬禮，[Your name]。
Complex Instructions 複雜指示	Write an introduction for a weekly newsletter, including the following: 為每週通訊寫一段引言，包括以下內容： - My company is named contoso	Hello Contoso team! I wanted to take a moment to thank you for all your hard work over the past few months. It's been a tough couple of months, but you've all pulled together and done an amazing job. I'm really proud of you all. Looking ahead to the next quarter, I'm feeling really positive. I think we're going to

Instruction Type 指示類型	**Prompt 提示**	**Completion 完成**
	- 我的公司名為 contoso - An exuberant greeting - 一段充滿熱情的問候 - Thanking the team for their hard work over a couple tough months - 感謝團隊在過去幾個艱難月份中的辛勤付出 - Positive outlook for the coming quarter - 下季展望正面 - Signed by the SLT - 由高階領導團隊簽署	achieve some great things. Thanks again, The Senior Leadership Team. Hello Contoso 團隊！我想花一點時間感謝大家這幾個月來的辛勤付出。過去這段時間相當艱困，但你們團結一致、表現非常出色。我為你們感到非常驕傲。展望下個季度，我對未來充滿正面期待。我相信我們會達成一些優秀的成果。再次感謝，The Senior Leadership Team。

Primary Content 主要內容

Primary content refers to some sort of text that is being processed or transformed by the model. Primary content is typically used in conjunction with instructions. A simple example would be language translation. In the example below, the English text would be considered the primary content while "Translate to French:" is an instruction.

主要內容指的是模型正在處理或轉換的某種文字。主要內容通常與指示一起使用。一個簡單的例子是語言翻譯。在下面的範例中，英文文字會被視為主要內容，而「Translate to French:」則是一個指示。

展開表格

Prompt 提示	**Completion 完成**
Can you please tell me how to get to the museum? 請問你可以告訴我如何前往博物館？ Translate to French: 請翻譯成法語：	`Pouvez-vous s'il vous plaît me dire comment aller au musée?`

Primary content can also be much longer. In the following example, the primary content is the introduction section from the Wikipedia entry on time complexity, which is almost 300 words long. This content is abbreviated in our table for ease of display.

主要內容也可以更長。以下範例中，主要內容為維基百科「時間複雜度」條目的導言部分，長度接近 300 字。為便於展示，本表格中已將該內容略為簡寫。

展開表格

Prompt 提示	**Completion 完成**
In computer science, the time complexity is the computational complexity that describes the amount of computer time it takes to run an algorithm. Time complexity is commonly estimated by [...] 在電腦科學中，時間複雜度是描述執行一個演算法所需電腦時間量的計算複雜度。時間複雜度通常透過[...] Summarize the previous text in a few sentences: 請將前述文字總結為幾句話：	Time complexity is a measure of how long an algorithm takes to run. It's typically expressed as a function of the size of the input. The most common measure is worst-case time complexity, which is the maximum amount of time required for inputs of a given size. 時間複雜度是衡量演算法執行所需時間的指標，通常以輸入大小的函數來表示。最常見的量度是最壞情況時間複雜度，指在給定輸入大小下所需時間的最大值。

Prompt 提示	**Completion 完成**
In computer science, the time complexity is the computational complexity that describes the amount of computer time it takes to run an algorithm. Time complexity is commonly estimated by [...] 在電腦科學中，時間複雜度是描述演算法執行所需電腦時間量的計算複雜度。時間複雜度通常透過估算[...]來表示。 `TL;DR` (Too long didn't read) `TL;DR` （太長沒看）	The time complexity of an algorithm is the amount of time it takes to run an algorithm. Time complexity is commonly estimated by counting the number of elementary operations performed by the algorithm, supposing that each elementary operation takes a fixed amount of time to perform. The time complexity is generally expressed as a function of the size of the input. 演算法的時間複雜度是指執行該演算法所需的時間量。時間複雜度通常透過計算演算法所執行的基本運算次數來估算，假設每個基本運算所需的時間為固定值。時間複雜度通常以輸入大小的函數來表示。

GPT models can also handle primary content that is structured. In the example below, a TSV about beers in Belgium (abbreviated for display in this article) is passed to the model as part of the prompt. It's capable of properly interpreting the content and answering questions about the data.

GPT 模型也能處理結構化的主要內容。下面的範例中，一個關於比利時啤酒的 TSV（為本文顯示而簡略）被作為提示的一部分傳遞給模型。它能夠正確地解讀該內容並回答有關該資料的問題。

展開表格

Prompt 提示	**Completion 完成**

`Belgian Beer Brands`
`Beer name Beer style color ABV Brewery`
`"Chimay Gold: ""Doree""" trappist pale ale gold 4.80% Chimay Brewery (official trappist)`
`"Chimay Blue: ""Grande Reserve""" trappist dark ale dark 9.00%`
[…]
Referencing the table above, how many
	There are 10 beers in the table that are less than 6% ABV.
	表中有 10 種啤酒的酒精濃度低於 6% ABV。

Prompt 提示	**Completion 完成**
beers are less than 6% ABV? 參照上表，有多少種啤酒的酒精濃度低於 6% ABV？	

Examples 範例

Successful prompts often rely on the practice of “one-shot” or “few-shot” learning. This refers to the inclusion of one or more examples of the desired behavior of the model, typically by including input and output pairs. This isn't learning in the sense that the model is permanently changed, but rather that the examples better condition the model to respond as desired for only the current inference. The use of prompts with no examples is sometimes referred to as “zero-shot” learning. Note that with the Chat Completion API few-shot learning examples are typically added to the messages array in the form of example user/assistant interactions after the initial system message.

成功的提示通常仰賴「one-shot」或「few-shot」學習的做法。這是指包含一個或多個示範模型期望行為的範例，通常以輸入與輸出配對的形式出現。這並非指模型在永久性地學習改變，而是這些範例能在當前推論期間更好地條件化模型，使其如預期回應。沒有範例的提示有時稱為「zero-shot」學習。請注意，使用 Chat Completion API 時，few-shot 學習範例通常會在初始 system 訊息之後以範例使用者 / assistant 互動的形式加入到 messages 陣列中。

展開表格

Learning Type 學習類型	**Prompt 提示**	**Completion 完成**
Zero-shot 零樣本	Headline: Coach confident injury won't derail Warriors 標題：教練自信傷病不會讓勇士隊偏離軌道 Topic: 主題：	The coach is confident that the injury won't derail the Warriors' season. The team is still focused on their goals and that they'll continue to work hard to achieve them. 教練相信這次受傷不會毀掉勇士隊的賽季。球隊仍然專注於他們的目標，並將繼續努力達成這些目標。

Learning Type 學習類型	**Prompt 提示**	**Completion 完成**
Few-shot 少量示例學習	Headline: Twins' Correa to use opt-out, test free agency 標題：雙胞胎隊的 Correa 將採取不選擇權，嘗試自由球員市場 Topic: Baseball 主題：棒球 Headline: Qatar World Cup to have zones for sobering up 標題：卡達世界盃將設置清醒區 Topic: Soccer 主題：足球 Headline: Yates: Fantasy football intel for Week 6 標題：Yates：第 6 週幻想美式足球情報 Topic: Football 主題：美式足球 Headline: Coach confident injury won't derail Warriors	Basketball 籃球

Learning Type 學習類型	Prompt 提示	Completion 完成
	標題：教練有信心傷勢不會打亂勇士隊 Topic: 主題：	

The example above illustrates the utility of few-shot learning. Without the examples, the model seems to be guessing at the desired behavior, while the examples cleanly show the model how to operate. This also demonstrates the power of the model, it can infer the category of label that is wanted, even without a "basketball" label in the examples.

上面的範例說明了少量範例學習（few-shot learning）的效用。沒有那些範例時，模型似乎在猜測期望的行為，而範例則清楚地示範模型應如何操作。這也展示了模型的能力：即使範例中沒有「basketball」這個標籤，模型仍能推斷出需要的標籤類別。

Cue 提示訊號

Cues act as the "jumpstart" for the output of the model, helping to direct the model to the desired output. It's often a prefix that the model can build onto. In the example below, we show the cue directing the story to include key elements. Cues are often used along with instructions, but not always. In the examples below, the model is being used to summarize a long announcement email. A cue is used in one case to help focus the output of the model and to suggest a specific output format (bullet points) in the other.

提示語就像是模型輸出的「啟動器」，幫助引導模型產生期望的結果。它通常是模型可以在其上延伸的前綴。在下面的範例中，我們示範了提示語如何引導故事包含關鍵要素。提示語經常和指示一起使用，但也不一定總是如此。在下列範例中，模型被用來摘要一封長篇公告信。其中一個範例使用提示語來幫助聚焦模型的輸出，而另一個則用來建議特定的輸出格式（項目符號）。

In the context of the Chat Completion API instructions take the form of the system message, and few-shot examples are indicated by providing a series of example user/assistant interactions to help prime the model.

在 Chat Completion API 的情境中，指示以 system 訊息的形式呈現，而少量範例則透過提供一系列範例的 user/assistant 互動來表示，以協助啟動模型。

展開表格

Number of Cues 提示數量	Prompt 提示	Completion 完成
0	Performance reviews are a core component of our Performance & Development approach at Contoso. [...] 績效評估是 Contoso 我們「績效與發展」方法的核心組成部分。[...] Summarize the above email message: 摘要上述電子郵件內容：	The purpose of this email is to inform employees of recent changes to the performance review experience, which is part of Contoso's Performance & Development approach. The changes include an improved structure, upgraded in-tool user experience, and focused prompts. Additionally, there's now an optional Conversation Starters section for employees to flag topics they'd like to discuss with their managers. 此電子郵件的目的是通知員工有關績效評估體驗的近期變更，該評估屬於 Contoso 的「績效與發展」方法。變更包括改良的結構、升級的工具內使用者體驗，以及更具針對性的提示。此外，現在新增一個選用的「對話啟動項」區段，讓員工標記他們想與主管討論的議題。
1	Performance reviews are a core component of our Performance & Development approach at Contoso. [...] 效能評估是 Contoso 效能與發展方法的核心組成部分。[...] Summarize the	that Contoso is improving the performance review experience by making it more focused on core priorities, expectations, and impact. Additionally, they're upgrading the in-tool user experience with tips, resources, and guidance. Finally, they're adding guidance on how to activate

Number of Cues 提示數量	Prompt 提示	Completion 完成
	above email message: 摘要上述電子郵件訊息： The key takeaway of the message is 該訊息的重點是	company-wide core priorities. Contoso 正在改善績效評估體驗，使其更專注於核心優先事項、期望與影響。此外，他們正提升工具內的使用者體驗，提供提示、資源與指引。最後，他們正在加入關於如何啟動全公司核心優先事項的指導。
2	Performance reviews are a core component of our Performance & Development approach at Contoso. [...] 績效評估是 Contoso 在績效與發展方法中的核心組成部分。[...] Summarize the above email message: 請總結上述電子郵件內容： Key Points: 重點摘要： •	Contoso is introducing changes to the performance review experience, based on feedback from employees. Contoso 正在根據員工回饋，對績效評估體驗進行變更。 • The changes include a new focus on core priorities, expectations, and impact; an upgraded user experience; and new guidance on company-wide core priorities. • 變更包括新的重點放在核心優先事項、期望與影響；升級的使用者體驗；以及針對公司整體核心優先事項的新指引。 • Performance reviews are a time to reflect on past performance and learning as well as to plan for future priorities and growth opportunities. • 績效評核是回顧過去表現與學習的時刻，同時也是規劃未來優先事項與成長機會的時機。 • Employees and managers should write comments in advance of the performance review meeting.

Number of Cues 提示數量	Prompt 提示	Completion 完成
		• 員工與主管應在績效評核會議前事先撰寫意見。 • There should be at least two performance reviews per year. • 每年至少應進行兩次績效評核。 • The time investment for each review should remain about the same. • 每次評核所需的時間投入應該維持大致相同。

Supporting content 支援內容

Supporting content is information that the model can utilize to influence the output in some way. It differs from primary content in that it isn't the main target of task, but it's typically used along with primary content. Common examples include contextual information such as the current date, the name of the user, user preferences, and so on. The example below uses supporting content to help organize a set of planned workshops for the user. Without the supporting (important topics) the model simply lists off the workshops (truncated for display), when informed of my important topics, the model is able to accurately group the meetings.

支援內容是模型可用來以某種方式影響輸出的資訊。它與主要內容不同之處在於，支援內容並非任務的主要目標，但通常與主要內容一同使用。常見範例包括上下文資訊，例如當前日期、使用者名稱、使用者偏好等等。以下範例使用支援內容來協助為使用者整理一系列規劃中的工作坊。如果沒有支援內容（重要主題），模型會只是列出工作坊（為顯示而截斷）；在得知我的重要主題後，模型便能準確地將會議分組。

展開表格

Supporting Content 支援內容	**Prompt 提示**	**Completion 完成**
Without 無	Workshop list: 工作坊清單： Title Description 標題 描述 Planning Workshop Kickoff Please join us to kick off our workshop! Hear from our leaders about cross group priorities and learnings for the year ahead. The invitation to this event has already been sent to everyone in [...] 規劃工作坊啟動會 請加入我們，一起啟動工作坊！聆聽我們的領導團隊說明跨群組的優先事項與未來一年的學習重點。此活動的邀請已寄送給所有人 [...] Summarize the workshops above, grouped by topic: 請將上述工作坊依主題彙整與摘要：	Workshop list: 工作坊清單： Planning Workshop Kickoff 規劃工作坊啟動會 -Planning Workshop Wrap-up - 規劃工作坊結束摘要 - Core MeetUp - 核心聚會 (Core MeetUp) - Customer Driven Perspectives - 以客戶為導向的觀點 -Impact and Key Focus Areas - 影響與主要關注領域 -Best practices -最佳實務 -Accelerating app adoption [...] -加速應用採用 [...]
With 與	Workshop list: 工作坊清單： Title Description 標題 描述 Planning Workshop Kickoff Please join us to kick off our workshop! Hear from	1. Prompt Engineering: 1. 提示工程（Prompt Engineering）：

Supporting Content 支援內容	Prompt 提示	Completion 完成
	our leaders about cross group priorities and learnings for the year ahead. The invitation to this event has already been sent to everyone in [...] 規劃工作坊 啟動 請加入我們以啟動本次工作坊！聽取我們領導就跨部門優先事項與來年度學習的分享。本活動的邀請已經發送給 [...] My Important Topics: prompt engineering, search, GPT models 我的重要主題：prompt engineering、search、GPT models Summarize the workshops above, grouped by my important topics: 將上述工作坊依我的重要主題彙整摘要：	- New prompt engineering tips - 新的提示工程技巧 - The art of prompt engineering - 提示工程的技藝 2. Search: 2. 搜尋： -Intro to vector search with embedding - 向量搜尋與 embedding 介紹 3. GPT Models: 3. GPT 模型： - Intro to GPT-4 - GPT-4 介紹 - GPT-35-Turbo in-depth. - GPT-35-Turbo 深入解析。

Scenario-specific guidance

情境專屬指南

While the principles of prompt engineering can be generalized across many different model types, certain models expect a specialized prompt structure. For Azure OpenAI GPT models, there are currently two distinct APIs where prompt engineering comes into play:

雖然提示工程的原則可在許多不同模型類型間通用，但某些模型需要專門的提示結構。對於 Azure OpenAI 的 GPT 模型，目前有兩個不同的 API 需要運用提示工程：

- Chat Completion API. 聊天完成 API

- Completion API. Completion API。

Each API requires input data to be formatted differently, which in turn impacts overall prompt design. The **Chat Completion API** supports the GPT-35-Turbo and GPT-4 models. These models are designed to take input formatted in a specific chat-like transcript stored inside an array of dictionaries.

每個 API 都要求以不同方式格式化輸入資料，進而影響整體提示設計。Chat Completion API 支援 GPT-35-Turbo 與 GPT-4 模型。這些模型被設計為採用特定的聊天式對話紀錄格式輸入，該紀錄儲存在一個字典陣列中。

The **Completion API** supports the older GPT-3 models and has much more flexible input requirements in that it takes a string of text with no specific format rules.

Completion API 支援較舊的 GPT-3 系列模型，其輸入需求彈性較大，接受一段沒有特定格式規則的文字字串。

The techniques in this section will teach you strategies for increasing the accuracy and grounding of responses you generate with a Large Language Model (LLM). It is, however, important to remember that even when using prompt engineering effectively you still need to validate the responses the models generate. Just because a carefully crafted prompt worked well for a particular scenario doesn't necessarily mean it will generalize more broadly to certain use cases. Understanding the limitations of LLMs, is just as important as understanding how to leverage their strengths.

本節的技術會教你提高使用大型語言模型 (LLM) 所產生回應之準確性與依據性的策略。不過，重要的是要記住，即使有效地使用 prompt engineering，你仍然需要驗證模型所產生的回應。僅因為一個精心設計的提示在特定情境下表現良好，並不代表它能廣泛適用於所有使用情境。理解 LLM 的限制，和了解如何善用其優勢一樣重要。

Chat completion APIs

聊天補完 API

This guide doesn't go in-depth into the mechanics behind the message structure for Chat Completions. If you aren't familiar with interacting with Chat Completions models programmatically, we recommend reading our how-to guide on the Chat Completion API first.

本指南不會深入探討 Chat Completions 訊息結構背後的機制。如果您不熟悉以程式方式與 Chat Completions 模型互動，我們建議先閱讀我們關於 Chat Completion API 的操作指南。

Note 注意事項

All of the examples in this section of the guide were tested against a base GPT-4 model in English. Some techniques may produce different results with newer models like gpt-4o, and gpt 4.5. If you

> are reading a localized version of this article in another language, these responses represent a localized translation of the English results. To learn more about potential limitations depending on what language you are using to prompt a model, please consult our Responsible AI transparency note.
>
> 本指南本節中的所有範例皆以英語在基礎 GPT-4 模型上測試。某些技術在較新的模型（例如 gpt-4o 與 gpt 4.5）上可能會產生不同結果。如果您正在閱讀此文章的其他語言在地化版本，這些回應代表對英文結果的在地化翻譯。要了解根據您用來提示模型的語言可能出現的限制，請參閱我們的 Responsible AI transparency note。

Few-shot learning 少量示例學習

A common way to adapt language models to new tasks is to use few-shot learning. In few-shot learning, a set of training examples is provided as part of the prompt to give additional context to the model.

將語言模型調整到新任務的一種常見方法是使用少量樣本學習（few-shot learning）。在少量樣本學習中，一組訓練範例會作為提示的一部分提供，藉此為模型提供額外的上下文。

When using the Chat Completions API, a series of messages between the User and Assistant (written in the new prompt format), can serve as examples for few-shot learning. These examples can be used to prime the model to respond in a certain way, emulate particular behaviors, and seed answers to common questions.

在使用 Chat Completions API 時，一系列在新提示格式下由 User 與 Assistant 之間的訊息可作為少量樣本學習的範例。這些範例可用來引導模型以特定方式回應、模擬特定行為，並為常見問題提供初始答案。

展開表格

System message 系統訊息	**Few-shot examples 少量示例**
Assistant is an intelligent chatbot designed to help users	**User** - "When do I need to file my taxes by?" 使用者 - 「我需要在什麼時候報稅？」 **Assistant** - "In 2023, you'll need to file your taxes by

System message 系統訊息	**Few-shot examples 少量示例**
answer their tax related questions. Assistant是一個智慧型聊天機器人，旨在協助使用者解答與稅務相關的問題。 Instructions 指示 - Only answer questions related to taxes. - 僅回答與稅務相關的問題。 - If you're unsure of an answer, you can say "I don't know" or "I'm not sure" and recommend users go to the IRS website for more information. - 如果你不確定答案，可以說「我不知道」或「我不確定」，並建議使用者前往IRS網站查詢更多資訊。	April 18th. The date falls after the usual April 15th deadline because April 15th falls on a Saturday in 2023. For more details, see https://www.irs.gov/filing/individuals/when-to-file " 助理 - 「在2023年，你需要在4月18日之前報稅。該日期晚於通常的4月15日截止日，因為2023年的4月15日是星期六。更多詳情請見 https://www.irs.gov/filing/individuals/when-to-file 」 **User** - "How can I check the status of my tax refund?" 使用者 - 「我該如何查詢我的退稅狀態？」 **Assistant** - "You can check the status of your tax refund by visiting https://www.irs.gov/refunds " 助理 - 「您可以透過造訪 https://www.irs.gov/refunds 查詢您的退稅狀態」

The table above describes the basic prompt structure, but for a detailed explanation of the exact prompt format you'll need to refer to the Chat Completion guide.

上表說明了基本的提示結構，但若要詳細了解精確的提示格式，請參閱 Chat Completion 指南。

Non chat scenarios 非對話情境

While the Chat Completion API is optimized to work with multi-turn conversations, it also can be used for non chat scenarios. For example, for a sentiment analysis scenario, you might use the following prompt:

雖然 Chat Completion API 已針對多回合對話進行最佳化，但也可以用於非聊天情境。例如，在情感分析情境中，你可以使用下列提示（prompt）：

展開表格

System message 系統訊息	User 使用者	Assistant 助理
You're an assistant designed to analyze sentiment from speech data. Users will paste in a string of text and you'll respond with an assessment of the response of the speaker. Rate on a scale of 1-10 (10 being highest). Explain why this rating was	hi thank you for calling contoso pharmacy who am i speaking with today i understand your frustration mary um can you tell me what exactly you're trying to accomplish OK uh certainly happy to check that for you one moment please i see here that you were on a generic form of lipitor is that right OK uh so i see that your doctor stopped prescribing these drugs in twenty twenty one actually OK uh yeah i'd be happy to check that for you uh because sometimes there's a gap in the system and it just doesn't get reported so let me take a look here just a moment so i'm	Sentiment rating: 8 情感評分：8 The speaker in this text seems to be polite, helpful, and patient while addressing the concerns of the customer. They take the time to check the information, provide clear explanations, and offer assistance. However, the rating isn't a 10 because there's still a hint of uncertainty in some parts of the

System message 系統訊息	**User 使用者**	**Assistant 助理**
given. 你是一個被設計用來從語音資料分析情感的助理。使用者會貼上一串文字，你需要對講話者的回應進行評估。請以 1 到 10 的等級來評分（10 為最高)。說明為何給予此評分。	seeing here that your doctor had these drugs prescribed to you from twenty twelve through twenty twenty one OK um yeah according to the latest records uh provided by doctor's office you're now on a different drug metformin would you like us to go ahead and fill that prescription for you for pick up a bit of form in five hundred milligrams uh you're very welcome mary please let us know if there's anything else we	conversation, which might not convey complete confidence to the customer. 此段文字中的說話者顯得有禮、樂於助人且有耐心地回應客戶的疑慮。他們花時間核對資訊、提供清楚的說明並主動提出協助。然而，評分並非滿分 10 分，因為對話某些部分仍帶有一絲不確定性，這可能無法向客戶傳達完全的信心。

System message 系統訊息	User 使用者	Assistant 助理
	can do for you today. 嗨，謝謝你致電 Contoso 藥局。請問我現在正在與誰通話？我能理解你的挫折感，Mary。你可以告訴我你到底想完成什麼嗎？好，嗯，當然，很樂意幫你確認，請稍等。我這裡看到你之前服用的是 Lipitor 的通用藥，對嗎？好的，嗯，所以我看到你的醫生在 2021 年停止為你開這類藥，實際上是這樣。好的，嗯，我很樂意幫你查詢，因為有時系統會有資料缺口而沒有被記錄，所以請讓我看一下，稍等一下。我這裡看到你的醫生在 2012 年到 2021 年之間曾為你開這些藥。好的，嗯，根據醫師診所提供的最新紀錄，你現在服用的是另一種藥 Metformin。你希望我們幫你把那張處方準備好讓你來領取嗎？每次 500 毫克的片劑。一點也不客氣，Mary，如果今天還有其他我們可以協助的地方，請告訴我們。	

Start with clear instructions

從明確的指示開始

The sequence information appears in the prompt matters. This is because GPT style models are built in a certain way that defines how they process the input. Our research suggests that telling the model the task you want it to do at the beginning of the prompt, before sharing additional contextual information or examples, can help produce higher-quality outputs.

提示中出現的順序資訊很重要。這是因為 GPT 類型的模型在某種方式下被構建，這種構造決定了它們如何處理輸入。我們的研究顯示，若在提示一開始就告訴模型你希望它執行的任務（在提供額外的上下文資訊或範例之前），有助於產出較高品質的回應。

Note 注意事項

Although following this technique is still generally recommended, in contrast to previous model versions (GPT-3 and prior), our testing showed that the model response with ChatGPT and GPT-4 models was the same regardless of whether the technique is utilized. In the case of the below example, we see that adding the statement “Several sources ... eruption” at the beginning or the end of the prompt did not cause any change in the final model response.

雖然通常仍建議遵循此技術，但與先前的模型版本（GPT-3 及更早）相比，我們的測試顯示，使用 ChatGPT 與 GPT-4 模型時，是否採用該技術並不會改變模型回應。在以下範例中，我們看到將句子 “Several sources ... eruption” 放在提示的開頭或結尾，並未對最終模型回應造成任何變化。

展開表格

System message 系統訊息	**User 使用者**	**Assistant 助理**
You're an AI assistant that helps people find information. 你是一個協助人們尋找資訊的 AI 助手。	Your task is to verify if the statement "Several sources mention a chance of another large eruption" is supported by a specific quote from the following set of snippets.	No, the statement isn't directly implied or stated by the snippets. The snippets mention a chance of a "megaquake" and a "magnitude 9 Cascadia earthquake" hitting Seattle in the next 50 years, but don't mention a chance of

System message 系統訊息	User 使用者	Assistant 助理
	你的任務是驗證以下陳述「多個來源提到可能會有另一場大型噴發」是否由下面片段集中某段具體引述所支持。 --- SNIPPETS 片段 [1] 14 percent chance of megaquake hitting Seattle, experts say [1] 專家表示，西雅圖有 14% 機率發生巨震 SEATTLE - There's a 14 percent chance of a magnitude 9 Cascadia earthquake hitting Seattle in the next 50 years, the U.S. Geological Survey estimates. "Unfortunately, we're unable to... 西雅圖—美國地質調查局估計，在未來 50 年內，西雅圖發生一次規模 9 級的卡斯卡迪亞地震的機率為 14%。「不幸的是，我們無法... [2] Earthquake experts lay out latest outlook for Seattle's 'Really Big One’	another large eruption. 不，片段並未直接暗示或陳述該說法。片段提到有機會發生「超級地震」以及在未來 50 年內一個「規模 9 級的 Cascadia 地震」襲擊西雅圖，但並未提到有另一個大規模火山噴發的機率。

System message 系統訊息	**User 使用者**	**Assistant 助理**
	[2] 地震專家公布西雅圖『真正的大地震』最新展望 “We say that there's approximately a 14% chance of another approximately magnitude-9 earthquake occurring in the next 50 years,” said a geophysicist at the University of Washington... 「我們表示，在未來 50 年內，大約有 14% 的機率會再發生一次大約規模 9 級的地震，」華盛頓大學的一位地球物理學家說... --- Is the statement "Several sources mention a chance of another large eruption" directly implied or stated by the snippets? 「幾個來源提到可能會有另一個大規模火山噴發」這個說法是否由片段直接暗示或陳述？	

Repeat instructions at the end

在結尾重複指示

Models can be susceptible to **recency bias**, which in this context means that information at the end of the prompt might have more significant influence over the output than information at the beginning of the prompt. Therefore, it's worth experimenting with repeating the instructions at the end of the prompt and evaluating the impact on the generated response.

模型可能會受到近因偏誤（recency bias）的影響，這裡指的是提示詞結尾處的資訊可能比提示詞開頭的資訊對輸出有更大的影響力。因此，值得嘗試在提示詞結尾重複指示，並評估對生成回應的影響。

Prime the output 引導輸出

This refers to including a few words or phrases at the end of the prompt to obtain a model response that follows the desired form. For example, using a cue such as `“Here’s a bulleted list of key points:\n- ”` can help make sure the output is formatted as a list of bullet points.

此處指在提示詞結尾加入幾個詞或短語，以獲得符合期望形式的模型回應。例如，使用像 `“Here’s a bulleted list of key points:\n- ”` 這樣的提示可以幫助確保輸出被格式化為項目符號清單。

展開表格

System message 系統訊息	**User 使用者**	**Assistant 助理**
You're an AI assistant that helps people find information. Answer in as few words as	John Smith is married to Lucy Smith. They have five kids, and he works as a software engineer at Microsoft. What search queries should I do to fact-check this? ## One	"John Smith married Lucy Smith five kids software engineer Microsoft"

System message 系統訊息	User 使用者	Assistant 助理
possible. 你是一個協助人們尋找資訊的 AI 助手。儘量用最少的文字回答。	possible search query is: John Smith 與 Lucy Smith 已婚。他們有五個小孩，他在 Microsoft 擔任軟體工程師。我應該做哪些搜尋查詢來事實查證這些資訊？## 一個可能的搜尋查詢是：	

In the above prompt, the text *One possible search query is:* primes the model to produce a single output. Without this cue the model produces several search queries as output.

在上述提示中，文字 One possible search query is: 會提示模型只產生一個輸出。若沒有這個提示，模型會產生多個搜尋查詢作為輸出。

Add clear syntax 加入清晰語法

Using clear syntax for your prompt—including punctuation, headings, and section markers—helps communicate intent and often makes outputs easier to parse.

在提示中使用清晰的語法——包括標點、標題和區段標記——有助於傳達意圖，並且常能讓輸出更容易解析。

In the example below, separators (`---` in this case) have been added between different sources of information or steps. This allows the use of `---` as a stopping condition for generation. In addition, section headings or special variables are presented in uppercase to differentiate them.

在下面的範例中，分隔符（此例為 `---` ）已加入不同的資訊來源或步驟之間。這允許使用 `---` 作為生成的停止條件。此外，區段標題或特殊變數以大寫呈現以做區別。

展開表格

System message 系統訊息	**User 使用者**	**Assistant 助理**
You're an AI assistant that helps people find information. 你是一個協助人們尋找資訊的 AI 助手。	You'll read a paragraph, and then issue queries to a search engine in order to fact-check it. Also explain the queries. 你會先閱讀一段文字，然後向搜尋引擎發出查詢以進行事實核查。並且說明這些查詢的目的。 --- PARAGRAPH 段落 John Smith is married to Lucy Smith. They have five kids, and he works as a software engineer at Microsoft. What search queries should I do to fact-check this? John Smith 與 Lucy Smith 結婚。他們有五個小孩，而且他在 Microsoft 擔任軟體工程師。要為此進行事實核查，我應該做哪些搜尋查詢？ --- QUERIES 查詢：	1. "John Smith Microsoft" - To check if John Smith is indeed employed at Microsoft. 1. "John Smith Microsoft" - 用來確認 John Smith 是否確實在 Microsoft 任職。 2. "John Smith Lucy Smith" - To check if John Smith is married to Lucy Smith. 2. "John Smith Lucy Smith" - 用來確認 John Smith 是否與 Lucy Smith 結婚。 3. "John Smith children" - To check if John Smith has five children. 3. "John Smith children" - 用來確認 John Smith 是否有五個孩子。

If you're not sure what syntax to use, consider using Markdown or XML. The models have been trained on a large quantity web content in XML and Markdown, which might provide better results.

如果您不確定要使用什麼語法，可以考慮使用 Markdown 或 XML。模型已在大量的 XML 和 Markdown 網路內容上訓練過，這可能會提供較佳的結果。

Break the task down 將任務拆解成較小的步驟

Large language models (LLMs) often perform better if the task is broken down into smaller steps. For example, in the search query prompt referenced earlier, the prompt can be restructured so that the model is first instructed to extract relevant facts, and then instructed to generate search queries that can be used to verify those facts.

大型語言模型 (LLMs) 通常在將任務拆解成較小步驟時表現更好。例如，在先前提到的搜尋查詢提示中，可以重新構建提示，先指示模型擷取相關事實，然後再指示模型產生可用來驗證這些事實的搜尋查詢。

展開表格

System message 系統訊息	**User 使用者**	**Assistant 助理**
You're an AI assistant that helps people find information. 你是一個協助人們尋找資訊的 AI 助手。	You'll read a paragraph, and then issue queries to a search engine in order to fact-check it. 你會先閱讀一段文字，然後向搜尋引擎發出查詢以進行事實查核。 --- PARAGRAPH 段落 John Smith is married to Lucy Smith. They have five kids, and he works as a software engineer at Microsoft. What search queries should I do to fact-check this?	1. John Smith is married to Lucy Smith 1. John Smith 與 Lucy Smith 已婚 2. They have five kids 2. 他們有五個小孩 3. He works as a software engineer at Microsoft 3. 他在 Microsoft 擔任軟體工程師

System message 系統訊息	User 使用者	Assistant 助理

	John Smith 與 Lucy Smith 結婚。他們有五個小孩，他在 Microsoft 擔任軟體工程師。我要用哪些搜尋查詢來進行事實查核？
	--- 	
	Now you'll extract factual claims, and issue queries to fact-check them. When issuing a query, use the function SEARCH("query")
	現在你會擷取事實性主張，並發出查詢來查核它們。當發出查詢時，請使用函式 SEARCH("query")
	
	FACTUAL CLAIMS 事實性主張
		QUERIES 查詢
		SEARCH("John Smith married Lucy Smith")
		SEARCH("John Smith family size")
		SEARCH("John Smith Microsoft software engineer")

Notice the use of clear syntax to differentiate the sections and prime the output. In this simple example, breaking the task down from one to two steps isn't very dramatic, but when trying to do this for a larger piece of text with many factual claims, breaking the task down can make a significant difference.

注意使用清晰的語法來區分各個段落並引導輸出。在這個簡單的範例中，將任務從一步拆成兩步的差異不大，但當要對包含許多事實主張的較大篇幅文字進行處理時，把任務拆解開來可能會產生顯著差異。

Use of affordances 使用可用性提示

Sometimes we can get the model to use an affordance instead of relying on its own parameters for information and answers. Search, for example, can be an affordance to help mitigate against fabricated answers, and to get up-to-date information.

有時我們可以讓模型使用一個外部可用功能（affordance），而不是完全依賴其內部參數來提供資訊與答案。例如，搜尋可以作為一種外部可用功能，幫助減少虛構答案，並取得最新資訊。

One simple way to use an affordance is to stop generation once the affordance calls are generated by the model, then paste the outcomes back into the prompt. Below is an example of what the follow-up call could be after doing the **SEARCH** calls above. Notice how we pasted the search results into the prompt and replaced the previous **SEARCH** calls.

使用外部可用功能的一個簡單方式是：在模型產生出外部功能呼叫後就停止生成，然後把外部功能的結果貼回提示字串（prompt）。下面是一個在執行上述 SEARCH 呼叫後，可能進行的後續呼叫範例。請注意我們如何把搜尋結果貼回提示字串，並取代先前的 SEARCH 呼叫。

展開表格

System message 系統訊息	User 使用者	Assistant 助理
You're an AI assistant that helps people find information. 你是一個協助人們尋找資訊的 AI 助手。	You'll read a paragraph, extract factual claims, and then use search engine results to fact-check them 你將閱讀一段文字，擷取事實性主張，然後使用搜尋引擎結果來事實查證它們 --- PARAGRAPH 段落 John Smith is married to Lucy Smith. They have five kids, and he works as a software engineer at Microsoft. What search queries should I do to fact-check this?	- John Smith is married to Lucy Smith: Confirmed. Snippet [1] and [2] both mention John Smith's wife as Lucy. - John Smith 與 Lucy Smith 結婚：已確認。片段 [1] 與 [2] 都提到 John Smith 的妻子是 Lucy。 - John and Lucy have five kids: Confirmed. Snippet [3] mentions that John was accompanied by his two daughters and three sons. - 約翰和露西有五個孩子：已確認。摘要 [3] 提到約翰與他的兩個女兒和三個兒子同行。 - John works as a software engineer at Microsoft: Partially confirmed. Snippet [4]

System message 系統訊息	**User 使用者**	**Assistant 助理**
	John Smith 與 Lucy Smith 結婚。他們有五個小孩，而且他在 Microsoft 擔任軟體工程師。要為此進行事實核查，我應該做哪些搜尋查詢？ --- FACTUAL CLAIMS 事實性主張 - John Smith is married to Lucy Smith - John Smith 與 Lucy Smith 結婚 - John and Lucy have five kids - John 和 Lucy 有五個孩子 - John works as a software engineer at Microsoft - John 在 Microsoft 擔任軟體工程師 --- Here are various search queries issued to research the claims above: 以下是為了驗證上述說法所進行的各種搜尋查詢：	mentions that John spent 10 years at Microsoft, but it doesn't explicitly state that he currently works there. Further research might be needed to confirm his current employment status. - 約翰在 Microsoft 擔任軟體工程師：部分確認。摘要 [4] 提到約翰在 Microsoft 待了 10 年，但並未明確表示他目前仍在該公司工作。可能需要進一步研究以確認他目前的就業狀況。

System message 系統訊息	User 使用者	Assistant 助理
	SEARCH QUERIES - John Smith married to Lucy Smith - John Smith 與 Lucy Smith 結婚 - John Smith number of children - John Smith 的子女性數 - John Smith software engineer Microsoft - John Smith 軟體工程師 Microsoft --- Here are snippets of the search results: SNIPPETS: 片段： [1] … John Smith’s wedding was on September 25, 2012 … [1] … John Smith 的婚禮是在 2012 年 9 月 25 日 … [2] … John Smith was accompanied by his wife Lucy to a party [2] … John Smith 與他的妻子 Lucy 一起出席了一場派對 [3]John was accompanied to the	

System message 系統訊息	User 使用者	Assistant 助理
	soccer game by his two daughters and three sons [3] John 有兩個女兒和三個兒子陪同前往足球比賽 [4] ... After spending 10 years at Microsoft, Smith founded his own startup, Tailspin Toys [4] ... 在微軟工作十年後，Smith 創立了他自己的新創公司 Tailspin Toys [5] John M is the town smith, and he married Fiona. They have a daughter named Lucy [5] John M 是鎮上的鐵匠，他娶了 Fiona。他們有一個名叫 Lucy 的女兒 --- Given the snippets, fact check each of the factual claims above: 根據以上片段，對每項事實性聲明進行事實查核：	

Chain of thought prompting
連鎖思維提示

This is a variation on the **break the task down** technique. Instead of splitting a task into smaller steps, in this approach, the model response is instructed to proceed step-by-step and present all the steps involved. Doing so reduces the possibility of inaccuracy of outcomes and makes assessing the model response easier.

這是「將任務拆解」技巧的一種變體。與其把任務分割成較小的步驟，本方法指示模型回應時逐步進行並呈現所有涉及的步驟。如此做可降低結果不準確的可能性，並使評估模型回應更容易。

展開表格

System message 系統訊息	**User 使用者**	**Assistant 助理**
You're an AI assistant that helps people find information. 你是一個協助人們尋找資訊的 AI 助手。	Who was the most decorated (maximum medals) individual athlete in the Olympic games that were held at Sydney? Take a step-by-step approach in your response, cite sources and give reasoning before sharing final answer in the below format: ANSWER	Step 1: Determine the year of the Sydney Olympic games. The Sydney Olympic games were held in the year 2000. 步驟 1：確定雪梨奧運的年份。雪梨奧運舉辦於 2000 年。 Step 2: Find a list of individual athletes who won multiple medals in the 2000 Sydney Olympics. 步驟 2：找出在 2000 年雪梨奧運中獲得多枚獎牌的個人運動員名單。 According to the official website of the Olympic

System message 系統訊息	**User 使用者**	**Assistant 助理**
	is: `<name>` 誰是於雪梨舉辦的奧運中獲得最多獎牌的個人運動員？在您的回應中採取逐步方法，引用來源並在給出最終答案之前說明推理，最終答案請使用以下格式：ANSWER is: `<name>`	Games, there were several individual athletes who won multiple medals in the 2000 Sydney Olympics. Some of the notable names include: 根據奧林匹克運動會官方網站，2000 年雪梨奧運會有數位個人選手贏得多面獎牌。其中一些知名選手包括： • Ian Thorpe (Australia) - 3 gold, 2 silver, 0 bronze • Ian Thorpe（澳大利亞）- 3 金、2 銀、0 銅 • Pieter van den Hoogenband (Netherlands) - 2 gold, 0 silver, 2 bronze • Pieter van den Hoogenband（荷蘭）- 2 金、0 銀、2 銅 • Lenny Krayzelburg (USA) - 3 gold, 0 silver, 0 bronze • Lenny Krayzelburg（美國）- 3 金、0 銀、0 銅 • Inge de Bruijn (Netherlands) - 3 gold, 1 silver, 0 bronze • Inge de Bruijn（荷蘭）- 3 金、1 銀、0 銅 Step 3: Determine the athlete with the most medals. After reviewing the list of individual

System message 系統訊息	**User 使用者**	**Assistant 助理**
		athletes who won multiple medals in the 2000 Sydney Olympics, we can see that the athlete with the most medals is Ian Thorpe from Australia, who won a total of 5 medals (3 gold and 2 silver). 步驟 3：確定獲得最多獎牌的選手。檢視在 2000 年雪梨奧運中贏得多面獎牌的個別選手名單後，我們可以看到獲得最多獎牌的選手是來自澳洲的 Ian Thorpe，他總共贏得 5 面獎牌（3 金、2 銀）。 Therefore, the ANSWER is: Ian Thorpe. 因此，答案是：Ian Thorpe。

Specifying the output structure
指定輸出結構

Using your prompt to specify the structure of the output can have a significant effect on the nature and quality of the results. Sometimes system message inputs as “only write true facts” or “don't fabricate information” might not be adequate mitigations. Instead, asking for the model response to also include citations can help result in a reduction of the prevalence of incorrect responses.

使用提示來指定輸出結構會顯著影響結果的性質與品質。有時候把系統訊息設定為「只寫出真實事實」或「不要捏造資訊」可能不足以減輕錯誤回應。相反地，要求模型回應同時包含引用來源，有助於降低不正確回應的發生率。

If you instruct the model to cite the source material when it makes statements, those statements are much more likely to be grounded. Asking for citations makes it so that the model must make two errors every time it generates a response: the first error is the fabricated response, and the second is the bad citation. Note that the closer the citation is to the text it supports, the shorter the distance ahead the model needs to anticipate the citation, which suggests that inline citations are better mitigations for false content generation than citations at the end of the content.

如果你指示模型在陳述時引用來源材料，這些陳述就更有可能有據可依。要求引用會使模型在每次生成回應時必須犯兩個錯誤：第一個錯誤是捏造的回應，第二個錯誤是錯誤的引用。請注意，引用離其所支持的文本越近，模型預期引用時需要向前推測的距離就越短，這表示內嵌引用比放在內容末端的引用更能減少錯誤內容的產生。

Similarly, if you ask the model to extract factual statements from a paragraph, it might extract compound statements such as ‘X is doing Y AND Z’ (which can be harder to verify). This can be avoided by specifying an output structure such as (entity1, relationship, entity2).

同樣地，如果你要求模型從一段文字中擷取事實陳述，它可能會擷取出複合陳述，例如「X 正在做 Y 並且 Z」（這類陳述較難驗證）。透過指定輸出結構，例如 (entity1, relationship, entity2)，可以避免此類情況。

The examples below show the use of citations and directing the model response to fit a defined structure.

下列範例示範如何使用引用並引導模型輸出以符合預先定義的結構。

展開表格

System message 系統訊息	**User 使用者**	**Assistant 助理**
You're an AI assistant that helps people find information 你是一個協助人們尋找資訊的 AI 助手	List down the different behaviors exhibited by the Customer Support agent in the following conversation. 請列出在下列對話中客服人員所展現的各種行為。 --- hi thank you for calling contoso pharmacy who am i speaking with today i understand your frustration mary um can you tell me what exactly	1. Polite Greeting BEHAVIOR("Agent starts the conversation with a polite greeting and asks for the caller's name") 2. Empathetic BEHAVIOR("Agent acknowledges Mary's frustration

System message 系統訊息	**User 使用者**	**Assistant 助理**
	you're trying to accomplish OK uh certainly happy to check that for you one moment please i see here that you were on a generic form of lipitor is that right OK uh so i see that your doctor stopped prescribing these drugs in twenty twenty one actually OK uh yeah i'd be happy to check that for you uh because sometimes there's a gap in the system and it just doesn't get reported so let me take a look here just a moment so i'm seeing here that your doctor had these drugs prescribed to you from twenty twelve through twenty twenty one OK um yeah according to the latest records uh provided by doctor's office you're now on a different drug metformin would you like us to go ahead and fill that prescription for you for pick up a bit of form in five hundred milligrams uh you're very welcome mary please let us know if there's anything else we can do for you today.	and tries to understand her concern") 3. Attentive BEHAVIOR("Agent listens to the customer's problem and asks for more details") 4. Proactive BEHAVIOR("Agent offers to check the issue for the customer") 5. Informative BEHAVIOR("Agent provides the relevant information about the customer's prescription history") 5. 提供資訊的行為 ("Agent 提供有關客戶處方歷史的相關資訊") 6. Thorough BEHAVIOR("Agent checks for any possible gaps in the system and provides a complete

System message 系統訊息	User 使用者	Assistant 助理
	嗨，謝謝你致電 Contoso 藥局，請問我今天在跟哪一位通話？我了解你的沮喪，Mary。嗯，你可以告訴我你到底想達成什麼嗎？好，嗯，當然很樂意幫你查詢，請稍等一下。我這裡看到你之前服用的是 Lipitor 的一般藥品，對嗎？好，嗯，所以我看到你的醫生在 2021 年停止開這些藥，實際上是這樣。好，嗯，我很樂意為你查詢，因為有時系統會有缺漏而沒有被記錄，所以讓我看一下，請稍候。我這裡看到你的醫生從 2012 年到 2021 年都有為你開這些藥。好，嗯，根據醫師辦公室提供的最新記錄，你現在改服另一種藥物 metformin。你要我們幫你把這張處方開好，讓你來取藥嗎？一顆 500 毫克的 form in（註：原文可能為藥品劑型名稱）？不客氣，Mary，請讓我們知道今天還有什麼我們可以為你效勞的。 --- How was the response of the Customer Support agent? 客戶服務人員的回應如何？ Classify and list down their various behaviors and cite the reason behind the classification, using the format	response") 6. 徹底的行為 ("Agent 檢查系統中任何可能的缺漏並提供完整的回應") 7. Confirmatory BEHAVIOR("Agent confirms the current medication with the customer") 7. 確認性的行為 ("Agent 與客戶確認目前的用藥") 8. Solution-oriented BEHAVIOR("Agent offers to fill the new prescription for the customer") 8. 以解決為導向的行為("Agent 提供為客戶配新處方的服務") 9. Courteous BEHAVIOR("Agent thanks the customer and offers further assistance if

System message 系統訊息	User 使用者	Assistant 助理

	BEHAVIOR("reason")
	BEHAVIOR("主動問候與自我介紹：以禮貌問候並詢問對方身分，建立基本溝通。") BEHAVIOR("同理與情緒確認：表達理解客戶沮喪，試圖安撫情緒。") BEHAVIOR("詢問需求與確認目標：請求客戶說明想達成的目標以釐清問題。") BEHAVIOR("表示協助意願：多次表明願意協助並請求時間查詢，顯示服務意願。") BEHAVIOR("提供查詢結果與資訊揭露：告知處方歷史（2012–2021）及醫師最近改開藥物 metformin，提供具體資訊。") BEHAVIOR("解釋可能的系統問題：說明系統可能有紀錄缺漏，說明問題來源並增加透明度。") BEHAVIOR("提出後續行動選項：詢問是否要代為填藥並安排取藥，給予客戶可選擇的下一步。") BEHAVIOR("語氣禮貌且重覆確認：多次使用禮貌語句與確認，維持友善專業的互動。")
		needed")
		9. 有禮的行為（「Agent 感謝客戶並在需要時提供進一步協助」）

展開表格

System message 系統訊息	User 使用者	Assistant 助理

You're an AI assistant that helps people find	Cluster the following news headlines into topic categories based on patterns	{
		"ARTIFICIAL_INTELLIGENCE": {
		"HEADLINES": [

System message 系統訊息	**User 使用者**	**Assistant 助理**

information.
你是一個協助人們尋找資訊的 AI 助手。
	seen within the text. Also mention reasoning behind how these categories were defined.
	將以下新聞標題根據文字中出現的模式分群為主題類別，並說明定義這些類別的理由。
	Output format: 輸出格式：
	{
	"TOPIC_NAME": "",
	"HEADLINES": [],
	"REASONING": ""
	}
	
	Input news headlines: 輸入新聞標題：
	1. "From books to presentations in 10s with AR + ML"
	1.「從書籍到簡報，10 秒內完成的 AR + ML」
	2. "Demo from 1993 of 32-year-old Yann LeCun showing off the World's first Convolutional Network for Text Recognition"
		"From books to presentations in 10s with AR + ML",
		"结合 AR 與 ML，10 秒內從書籍生成簡報",
		"Demo from 1993 of 32-year-old Yann LeCun showing off the World's first Convolutional Network for Text Recognition",
		"1993 年 32 歲的 Yann LeCun 展示世界首個用於文字辨識的捲積網路的示範",
		"First Order Motion Model applied to animate paintings"
		「用于讓畫作動起來的第一序運動模型」
		],
		"REASONING": "These headlines are related to artificial intelligence, machine learning, and their applications in various fields."
		「推理」：「這些標題與人工智慧、機器學習及其在各領域的應用有關。」
		},
		"FINANCIAL_MARKETS": { 「金融市場」：{
		"HEADLINES": ["頭條新聞": [
		"Robinhood and other brokers literally blocking purchase of $GME, $NOK, $BB, $AMC; allow

System message 系統訊息	User 使用者	Assistant 助理
	2.「1993 年示範：32 歲的 Yann LeCun 展示全球首個用於文字辨識的捲積神經網路」 3. "First Order Motion Model applied to animate paintings" 3. 「将 First Order Motion Model 應用於為畫作製作動畫」 4. "Robinhood and other brokers literally blocking purchase of $GME, $NOK, $BB, $AMC; allow sells" 4. 「Robinhood 和其他券商實際上阻止購買 $GME、$NOK、$BB、$AMC；允許賣出」 5. "United Airlines stock down over 5% premarket trading" 5.「聯合航空股票盤前交易下跌超過 5%」 Output: 輸出	sells", "Robinhood 和其他券商實際上封鎖購買 $GME、$NOK、$BB、$AMC；允許賣出", "United Airlines stock down over 5% premarket trading", "聯合航空股票在盤前交易下跌超過 5%", "Bitcoin was nearly $20,000 a year ago today" "比特幣在一年前的今天接近 20,000 美元"],]， "REASONING": "These headlines are related to financial markets, stocks, cryptocurrencies, and trading platforms." "REASONING": "這些標題與金融市場、股票、加密貨幣以及交易平台相關。" } }

Temperature and Top_p parameters

Temperature 與 Top_p 參數

Changing the temperature parameter changes the output of the model. The temperature parameter can be set between 0 and 2. A higher value, for example 0.7 makes the output more random and produce more divergent responses, while a lower value, like. 0.2, makes the output more focused and concrete. A fictional story could be generated using a higher temperature. Whereas to generate a legal document it's recommended to use a much lower temperature. Top_probability is another parameter and is similar to Temperature in that it also controls the randomness of the model response, but it does so in a different way. The general recommendation is to alter one of these two parameters at a time, not both.

變更 temperature 參數會改變模型的輸出。temperature 參數可設定在 0 到 2 之間。較高的值（例如 0.7）會讓輸出更隨機，產生更多具發散性的回應；而較低的值（例如 0.2）則會讓輸出更集中且具體。可以用較高的 temperature 來產生虛構故事，而要產生法律文件則建議使用較低的 temperature。top_probability 是另一個參數，與 Temperature 類似，也會控制模型回應的隨機性，但其作用方式不同。一般建議一次只調整這兩個參數中的一個，而不是同時調整。

Provide grounding context
提供依據背景情境

One of the most effective ways to provide reliable answers is to give the model data to draw its responses from (grounding data). If your use case relies on up-to-date, reliable information and isn't purely a creative scenario, we strongly recommend providing grounding data. In general, the closer you can get your source material to the final form of the answer you want, the less work the model needs to do, which means there's less opportunity for error. Here's an example where the system is provided a recent blog describing the launch of GPT-4 in Azure OpenAI in Azure AI Foundry Models, and asked to name some of its early customers.

提供模型用以作答的資料（基礎資料）是取得可靠答案的最有效方法之一。如果你的使用情境依賴於最新且可靠的資訊，且並非純屬創意型場景，我們強烈建議提供基礎資料。一般來說，當你能讓來源資料越接近你想要的最終答案形式時，模型所需處理的工作就越少，也就減少出錯的機會。以下範例中，系統被提供了一篇近期的部落格文章，內容描述 GPT-4 在 Azure AI Foundry Models 中於 Azure OpenAI 的發表，並被要求列出其中一些早期客戶。

展開表格

System message 系統訊息	User 使用者	Assistant 助理
You're an AI assistant that helps people find	Who are some of the early	Here are some of the early customers of the Azure

System message 系統訊息	**User 使用者**	**Assistant 助理**
information. You'll provide answers exclusively from below text and respond in 100 words or less: 你是一個幫助人們尋找資訊的 AI 助手。你將僅從以下文字提供答案，且回應不超過 100 字： "At Microsoft, we're constantly discovering new ways to unleash creativity, unlock productivity, and up level skills so that more people can benefit from using AI. This is allowing our customers to ... 「在 Microsoft，我們不斷發掘新的方式，釋放創意、解鎖生產力並提升技能，讓更多人能從使用 AI 中受益。這使我們的客戶能夠... ... thereby reducing the likelihood that the model generates a harmful response."[1] ...從而降低模型產生有害回應的可能性。」[1] Answer in 100 words or less. Use bullet lists	customers of the Azure OpenAI GPT-4 models? 有哪些早期客戶正在使用 Azure OpenAI 的 GPT-4 模型？	OpenAI GPT-4 models: 以下是一些使用 Azure OpenAI GPT-4 模型的早期客戶： - Epic Healthcare: They plan to use GPT-4 to help physicians and nurses spend less time on data entry and investigation. - Epic Healthcare：他們計畫使用 GPT-4 幫助醫師和護理人員減少在資料輸入和調查上所花費的時間。 - Coursera: They're using Azure OpenAI with GPT-4 to create a new AI-powered learning experience on their platform. - Coursera：他們使用 Azure OpenAI 與 GPT-4 在其平台上建立新的 AI 驅動學習體驗。 - Coca-Cola: They're using Azure OpenAI to build a knowledge hub and plan to leverage GPT-4's multimodal capabilities for marketing, advertising, public relations, and customer relations.

System message 系統訊息	User 使用者	Assistant 助理
wherever possible. 以不超過 100 字回答。盡可能使用項目符號。		- Coca-Cola：他們正在使用 Azure OpenAI 建立知識中心，並計畫利用 GPT-4 的多模態能力來處理行銷、廣告、公關與客戶關係。 - Other companies of all sizes are using Azure AI to improve customer experiences, summarize long-form content, help write software, and reduce risk by predicting the right tax data. - 各種規模的其他公司也在使用 Azure AI 改善客戶體驗、摘要長篇內容、協助撰寫軟體，並透過預測正確的稅務資料來降低風險。

[1] Text from Azure OpenAI GPT-4 launch blog.

[1] 來自 Azure OpenAI GPT-4 發表部落格的文字。

Best practices 最佳做法

- **Be Specific**. Leave as little to interpretation as possible. Restrict the operational space.

  具體說明。盡量減少需要解讀的部分。限制模型的操作空間。

- **Be Descriptive**. Use analogies.

  詳盡描述。使用類比。

- **Double Down**. Sometimes you might need to repeat yourself to the model. Give instructions before and after your primary content, use an instruction and a cue, etc.

  加強重複。有時你可能需要對模型重複指示。在主要內容之前與之後給予指令，使用指令與提示等方法。

- **Order Matters**. The order in which you present information to the model might impact the output. Whether you put instructions before your content (“summarize the following…”) or after (“summarize the above…”) can make a difference in output. Even the order of few-shot examples can matter. This is referred to as recency bias.

  順序很重要。您呈現資訊給模型的先後順序可能會影響輸出。把指示放在內容之前（例如「總結以下內容…」）或之後（例如「總結上述內容…」）都可能造成輸出差異。即使少量示例的排列順序也會產生影響。這稱為近期性偏差。

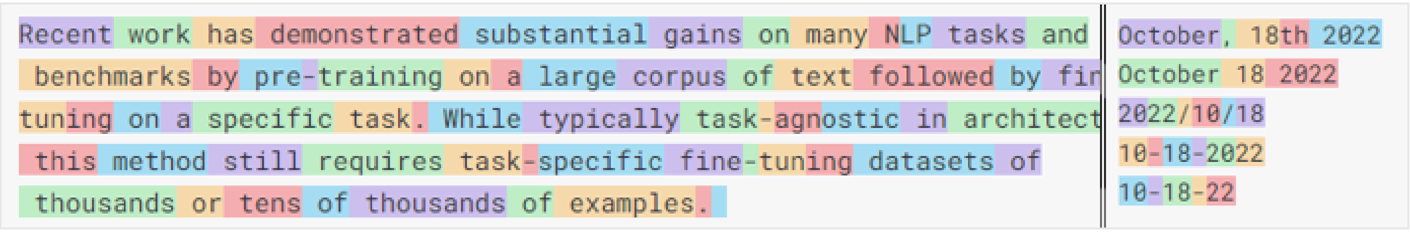
- **Give the model an “out”**. It can sometimes be helpful to give the model an alternative path if it's unable to complete the assigned task. For example, when asking a question over a piece of text you might include something like "respond with "not found" if the answer isn't present." This can help the model avoid generating false responses.

  給模型一個「退路」。如果模型無法完成指定任務，提供一個替代方案有時會很有幫助。例如，在針對一段文字提出問題時，你可以加入類似「如果答案不存在，請回覆 'not found'」的指示。這可以幫助模型避免產生錯誤回應。

Space efficiency 空間效率

While the input size increases with each new generation of GPT models, there will continue to be scenarios that provide more data than the model can handle. GPT models break words into "tokens." While common multi-syllable words are often a single token, less common words are broken in syllables. Tokens can sometimes be counter-intuitive, as shown by the example below which demonstrates token boundaries for different date formats. In this case, spelling out the entire month is more space efficient than a fully numeric date. The current range of token support goes from 2,000 tokens with earlier GPT-3 models to up to 32,768 tokens with the 32k version of the latest GPT-4 model.

隨著每一代 GPT 模型的輸入大小增加，仍會出現某些情況使得提供的資料超出模型可處理的範圍。GPT 模型會將字詞拆成「tokens（標記）」。常見的多音節字通常為單一 token，而較不常見的字則會按音節拆分。tokens 的分法有時可能與直覺不同，以下範例顯示不同日期格式的 token 邊界。在此情況下，將整個月份拼寫出來比完全用數字表示的日期更節省空間。目前模型支援的 token 範圍從早期 GPT-3 模型的 2,000 個 tokens 到最新 GPT-4 32k 版本可達 32,768 個 tokens。



Recent work has demonstrated substantial gains on many NLP tasks and
benchmarks by pre-training on a large corpus of text followed by fin
tuning on a specific task. While typically task-agnostic in architect
this method still requires task-specific fine-tuning datasets of
thousands or tens of thousands of examples.

October, 18th 2022
October 18 2022
2022/10/18
10-18-2022
10-18-22

Given this limited space, it's important to use it as efficiently as possible.

鑒於空間有限，務必盡可能有效率地使用。

- Tables – As shown in the examples in the previous section, GPT models can understand tabular formatted data quite easily. This can be a space efficient way to include data, rather than preceding every field with name (such as with JSON).

表格 – 如前一節的範例所示，GPT 模型可以相當容易地理解表格式資料。這可以是一種節省空間的方式來包含資料，而不是像 JSON 那樣在每個欄位前都加上名稱。

- White Space – Consecutive whitespaces are treated as separate tokens, which can be an easy way to waste space. Spaces preceding a word, on the other hand, are typically treated as part of the same token as the word. Carefully watch your usage of whitespace and don’t use punctuation when a space alone will do.

  空白字元 – 連續的空白字元會被視為獨立的 token，這可能很容易造成空間浪費。相反地，緊接在單字前的空格通常會被視為與該單字同一個 token 的一部分。請小心使用空白字元，且在單靠空格就能達成時不要使用標點符號。