13
Emerging Trends and Innovations

Dear reader, if you have made it up to this point – congratulations! You managed to complete this journey into LLMs and how to implement modern applications with them. Starting from the fundamentals of what's under the hood of an LLM, we covered many scenarios of LLM-powered applications, from conversational chatbots, to database copilots, to multimodal agents. We experimented with different models, both proprietary and open-source, and we also managed to fine-tune our own LLM. Last but not least, we covered the key topic of Responsible AI and how to embed ethical considerations within our LLM-powered applications.

In this final chapter, we are going to explore the latest advancements and future trends in the field of generative AI. Note that, as a rapidly evolving field, it is nearly impossible to keep up with up-to-date releases. Nevertheless, the advancements covered in this chapter will give you an idea of what to expect in the near future.

We will cover the following topics:

- The latest trends in language models and generative AI
- Companies embracing generative AI

The latest trends in language models and generative AI

As we saw in the previous chapters, LLMs set the basis for extremely powerful applications. Starting with LLMs, over the last months we have witnessed an explosive advancement in generative models, from multimodality to newly born frameworks, to enable multi-agent applications. In the next sections, we will see some examples of these new releases.

GPT-4V(ision)

GPT-4V(ision) is a **large multimodal model** (**LMM**) developed by OpenAI and officially released in September 2023. It enables users to instruct GPT-4 to analyze image inputs provided by the user. This integration of image analysis into LLMs represents a significant advancement in AI research and development. Model multimodality was achieved by using a technique called **image tokenization**, which converts images into sequences of tokens that can be processed by the same model as text. This allows the model to handle different types of data, such as text and images, and generate outputs that are consistent and coherent across modalities.

Since its initial trials in April 2023, GPT-4V has shown remarkable abilities in various domains. Moreover, many businesses have begun to integrate this model in their early testing stages. One of the successful examples is Be My Eyes, an app that assists the population of more than 250 million people who have visual impairments or blindness. The app links people who have low vision or blindness with helpers who can assist them with everyday activities, such as recognizing a product or finding their way around an airport. Using the new visual input feature of GPT-4, Be My Eyes created a Virtual Volunteer™ in its app that uses GPT-4. This Virtual Volunteer can produce the same amount of context and comprehension as a human volunteer.

The GPT-4 technology can do more than just identify and label what's in a picture; it can also infer and examine the situation. For instance, it can look at the items in a fridge and recommend what you can cook with them. What sets GPT-4 apart from other language and machine learning models is its capability to engage in dialogue and the higher level of analytical skill that the technology provides. Simple image recognition applications only identify what you see. They can't converse to find out if the noodles are made with the proper ingredients or if the thing on the floor is not just a ball but also liable to trip you up—and tell you that.

In response to early experimentation on GPT-4V before it went public, OpenAI has implemented several mitigations to address risks and biases.

These mitigations are aimed at improving a model's safety and reducing the potential harm caused by its output:

- **Refusal system**: OpenAI has added refusals for certain types of obviously harmful generations in GPT-4V. This system helps prevent a model from generating content that promotes hate groups or contains hate symbols.
- **Evaluation and red teaming**: OpenAI has performed assessments and consulted with external experts to examine the strengths and weaknesses of GPT-4V. This process helps detect potential flaws and risks in a model's output. The assessments cover areas such as scientific competence, medical guidance, stereotyping, disinformation threats, hateful content, and visual vulnerabilities.
- **Scientific competence**: Red teamers evaluated GPT-4V's abilities and challenges in scientific domains. While the model demonstrated the skill to comprehend complex information in images and verify claims in scientific papers, it also showed challenges, such as the occasional mixing of separate text elements and the possibility of factual mistakes.
- **Hateful content:** GPT-4V declines to answer questions about hate symbols and extremist content in some cases. However, the model's behavior may be variable, and it may not always decline to generate completions related to less-known hate groups or symbols. OpenAI recognizes the need for further enhancements in addressing hateful content.
- **Ungrounded inferences**: OpenAI has implemented mitigations to address risks associated with ungrounded inferences. The model now refuses requests for ungrounded inferences about people, reducing the potential for biased or inaccurate responses. OpenAI aims to refine these mitigations to enable the model to answer questions about people in low-risk contexts in the future.
- **Disinformation risks**: GPT-4V's ability to generate text content tailored to image input poses increased risks with disinformation. OpenAI acknowledges the need for proper risk assessment and context consideration when using the model in relation to disinformation. The combination of generative image models and GPT-4V's text gener-

ation capabilities may impact disinformation risks, but additional mitigations such as watermarking or provenance tools may be necessary.

These mitigations, along with the contribution from existing safety measures and ongoing research, aim to improve safety and reduce the biases in GPT-4V. OpenAI acknowledges the dynamic and challenging nature of addressing these risks and remains committed to refining and improving a model's performance in future iterations.

Overall, the GPT-4V has unveiled extraordinary capabilities and paves the way for multimodality within LLM-powered applications.

DALL-E 3

The newest version of OpenAI's image-generation tool, DALL-E 3, came out in October 2023. The most significant update from previous versions is its improved accuracy and faster speed when generating images from text. It aims to render more detailed, expressive, and specific images that align more closely with a user's specifications. In fact, even with the same prompt, DALL-E 3 shows great improvements compared to its previous version:

Figure 13.1: Images generated from the prompt "an expressive oil painting of a basketball player dunking, depicted as an explosion of a nebula" by DALLE-2 (left) and DALL-E 3 (right). Source: **https://openai.com/dall-e-3**

- DALL-E 3 has more safeguards and rules to avoid creating images that contain adult, violent, or hateful content.
- DALL-E 3 is now available to ChatGPT Plus and Enterprise customers via the API and in OpenAI Playground. It's also been integrated with Microsoft's Bing Chat.

AutoGen

In October 2023, Microsoft released a new open-source project called AutoGen. It is a Python lightweight framework that allows multiple LLM-powered agents to cooperate with each other to solve users' tasks. For an overview of what the cooperation frameworks look like, you can refer to `https://github.com/microsoft/autogen/tree/main`.

Earlier in Part 2 of this book, we covered many scenarios of LangChain Agents leveraging external tools. In those scenarios, we had one agent powered by an LLM that dynamically decided which tool to use to solve a user's query. AutoGen works differently, in the sense that it lets different agents, each one acting with a specific role and expertise, cooperate to address the user's query. The main element of novelty here is that each agent can actually generate output that serves as input to other agents, as well as generate and modify the plan to be executed. That is the reason why the framework has also been designed to keep a human or admin in the loop, to actually approve or discard actions and executions.

According to the original paper *AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation* by *Wu et al.*, there are three main reasons why the multi-agent conversation exhibits great performance:

- **Feedback incorporation**: Since LLMs have the capacity to elaborate and leverage feedback, they can cooperate through conversations in natural language with each other, and humans as well, to adjust the way they solve a given problem.
- **Adaptability**: Since LLMs are general-purpose models that can adapt to different tasks if properly configured, we can initialize different

agents that leverage the various capabilities of LLMs in a modular and complementary way.
- **Splitting complex tasks**: LLMs work better when they split complex tasks into smaller subtasks (as covered in *Chapter 4* about prompt engineering techniques). Henceforth, multi-agent conversations can enhance this partition, delegating each agent to a subtask, while keeping the overall picture of the problem to solve.

To enable a multi-agent conversation, there are two main components to be aware of:

- **Conversable agents** are entities that can communicate with each other and have different capabilities, such as using LLMs, human input, or tools.
- **Conversation programming** is a paradigm that allows developers to define the interaction behavior between agents using natural or programming languages.

You can see what these conversations look like at https://www.microsoft.com/en-us/research/publication/autogen-enabling-next-gen-llm-applications-via-multi-agent-conversation-framework/.

The AutoGen framework has already proven its great capability in addressing different use cases, among which are the following:

- **Code generation and execution**. AutoGen provides a class of agents that can execute code as `.py` files in a given directory.
- **Multi-agent collaboration**. This scenario fits whenever you want varied expertise to reason upon a given task. For example, you might want to set up a research group that, when given a user's request, sets up a plan, evaluates it, receives a user's input, executes it with different expertise (aka different agents), and so on.
- **Tools integrations**. AutoGen also offers some classes that facilitate the integration of external tools, such as web search and **retrieval-augmented generation** (**RAG**) from a provided vector database.

You can find some examples of different applications of the AutoGen framework at https://microsoft.github.io/autogen/docs/Examples#automated-multi-agent-chat.

Overall, AutoGen provides a useful and innovative toolkit that makes it easier to let agents cooperate with each other, as well as with a human in the loop. The project is open to contribution, and it will be very interesting to see how it progresses and to what extent the multi-agent approach will become a best practice.

So far, we have been talking about LLMs that are, by definition, "large" (for example, the GPT-3 has 175 billion parameters). However, sometimes, smaller models can be useful as well.

Small language models

Smaller models with fewer parameters can demonstrate extraordinary capabilities in specific tasks. This class of models has paved the way for what are now called **small language models (SLMs)**. SLMs have fewer parameters than LLMs, which means they require less computational power and can be deployed on mobile devices or resource-constrained environments. SLMs can also be fine-tuned to excel in specific domains or tasks, such as finance, healthcare, or customer service, by using relevant training data.

SLMs are promising because they offer several advantages over LLMs, such as:

- They are more efficient and cost-effective, as they require less computational resources and energy to train and run.
- They are more accessible and portable, as they can be deployed on mobile devices or edge computing platforms, enabling a wider range of applications and users.
- They are more adaptable and specialized, as they can be fine-tuned to specific domains or tasks using relevant data, improving their accuracy and relevance.

- They are more interpretable and trustworthy, as they have fewer parameters and simpler architectures, making them easier to understand and debug.

Phi-2 is an example of a promising SLM that demonstrates outstanding reasoning and language understanding capabilities, showcasing state-of-the-art performance among base language models with less than 13 billion parameters. It is a 2.7 billion-parameter language model developed by Microsoft Research, trained on high-quality data sources, such as textbooks and synthetic texts, and uses a novel architecture that improves its efficiency and robustness. Phi-2 is available in the Azure AI Studio model catalog and can be used for various research and development purposes, such as exploring safety challenges, interpretability, or fine-tuning experiments.

In the next section, we are going to see which companies are actively leveraging generative AI for their processes, services, and products.

Companies embracing generative AI

Since the launch of ChatGPT in November 2022, up to the newest large foundation models on the market (both proprietary and open-source), many companies in different industries started embracing generative AI within their processes and products. Let's discuss some of the most popular ones.

Coca-Cola

Coca-Cola partnered with Bain & Company and OpenAI to leverage DALL-E, a generative AI model. This partnership was announced on February 21, 2023.

OpenAI's ChatGPT and DALL-E platforms will help Coca-Cola create customized ad content, pictures, and messages. Coca-Cola's "Create Real Magic" initiative is the result of the collaboration between OpenAI and Bain & Company (**https://www.coca-colacompany.com/media-center/coca-cola-invites-digital-artists-to-create-real-magic-using-new-ai-platform**). The platform is a unique innovation that

merges the abilities of GPT-4, which generates text that sounds like humans making search engine queries, and DALL-E, which creates images from text. This enables Coca-Cola to rapidly produce text, images, and other content. This strategic alliance is expected to deliver real value to large enterprise customers, enabling massive business transformation within the Fortune 500. It also sets a standard for their clients to follow.

Notion

Notion is a versatile platform that combines note-taking, project management, and database functionalities in a single space. It allows users to capture thoughts, manage projects, and even run an entire company in a way that suits their needs. Notion is ideal for individuals, freelancers, startups, and teams looking for a straightforward application to collaborate on multiple projects.

Notion has introduced a new feature called Notion AI that uses generative AI. This feature is essentially a prediction engine that guesses what words will work best based on a prompt or text you've written. It can perform tasks such as:

- Summarizing lengthy text (e.g., meeting notes and transcripts)
- Generating entire blog post outlines and emails
- Creating action items from meeting notes
- Editing your writing to fix grammar and spelling, change the tone, etc.
- Assisting with research and problem-solving

The following screenshot shows some of the Notion features powered by generative AI:

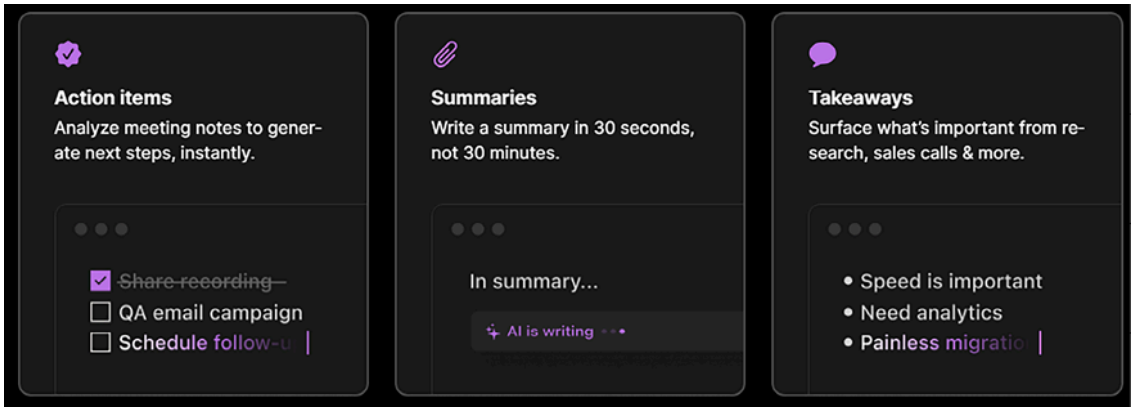

Figure 13.2: Some features of Notion AI. Source: **https://www.notion.so/product/ai**

Notion AI is powered by OpenAI's GPT models and integrated into the core Notion apps (desktop, browser, and mobile), allowing users to write prompts that will generate text, as well as apply AI to text they've already written or captured. This makes Notion AI a powerful digital assistant that enhances the functionality of the Notion workspace.

Malbek

Malbek is a modern, innovative **contract lifecycle management (CLM)** platform with a proprietary AI core. It meets the increasing contractual needs of your entire organization, including Sales, Finance, Procurement, and other essential business units.

Malbek uses generative AI to offer a feature powered by LLMs and featuring ChatGPT. It can do tasks such as:

- Understanding the language in contracts
- Making changes
- Easily accepting or rejecting redlines
- Making custom requests – all in natural language

This remarkable new feature enables users to speed up negotiation time and shorten review cycles, improving the functionality of the Malbek workspace.

Microsoft

Since its partnership with OpenAI, Microsoft has started infusing AI powered by GPT-series in all its products, introducing and coining the concept of Copilot. We've already introduced the concept of a Copilot system in *Chapter 2*, as a new category of software that serves as an expert helper to users trying to accomplish complex tasks, working side by side with users and supporting them in various activities, from information retrieval to blog writing and posting, and from idea brainstorming to code review and generation.

In 2023, Microsoft released several copilots within its products, such as the Edge Copiot (former Bing Chat). The following illustration shows the user interface of Bing Chat:

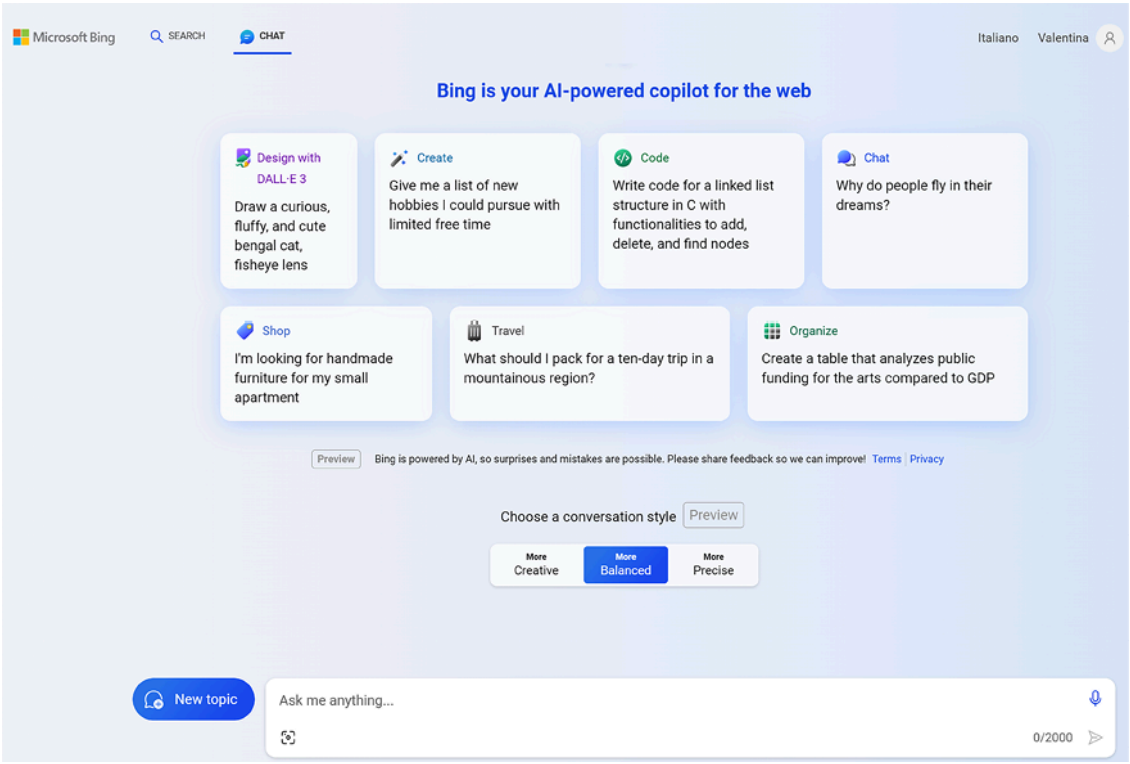

Figure 13.3: Microsoft Bing Chat

Bing Chat is also a perfect example of a multimodal conversational agent powered by both GPT-4V and DALL-E 3. Plus, you can interact with it via audio messaging. An example of these multimodal capabilities is shown in the following screenshot:

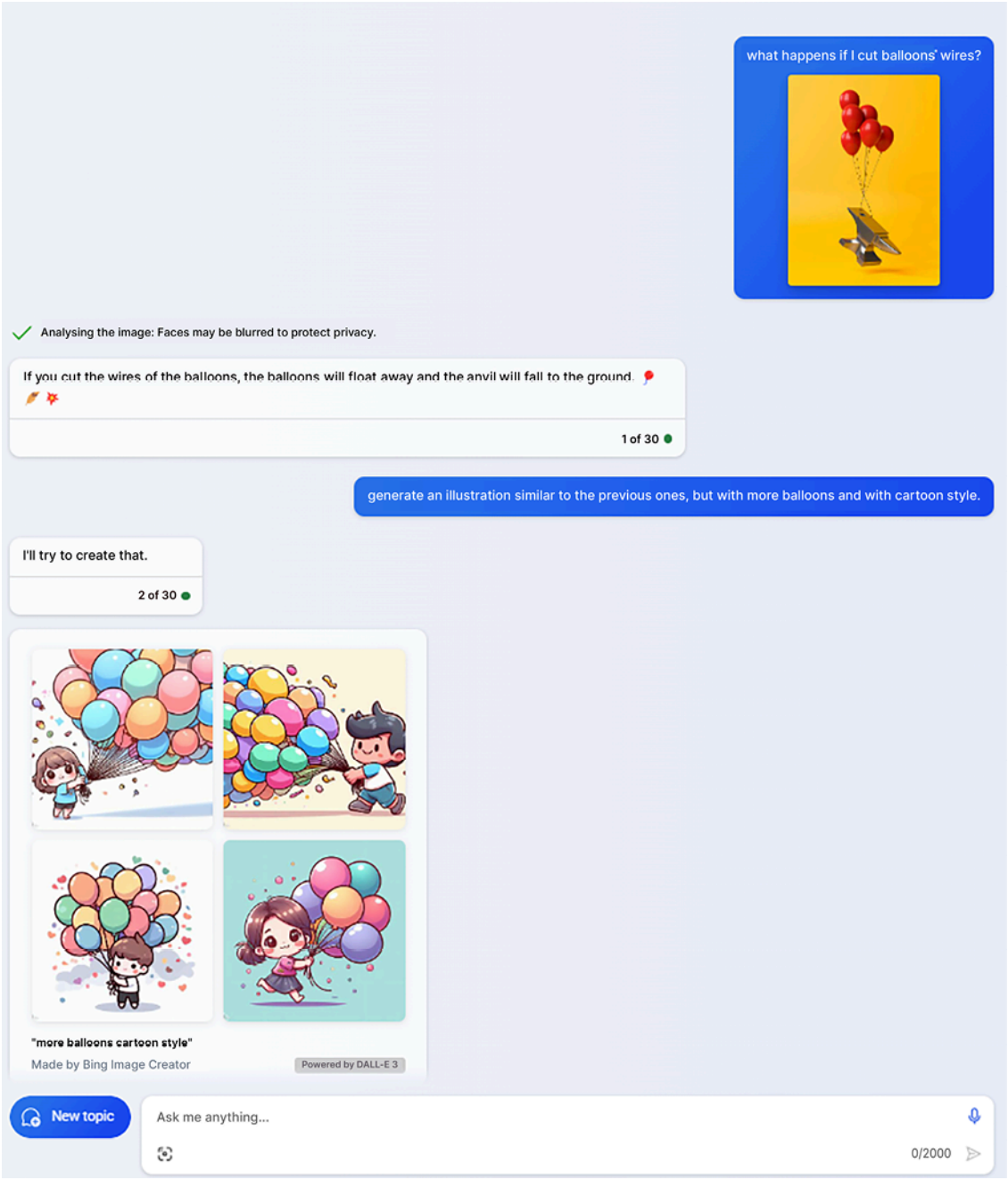

Figure 13.4: Leveraging the multimodal capabilities of Bing Chat

Microsoft's copilots will empower professionals and organizations to drastically improve their productivity and creativity, paving the way for a new way of working.

Overall, companies in all industries are seizing the potential of generative AI, with the awareness that the competitive landscape will soon raise the benchmark for copilots and AI-powered products.

Summary

In this final chapter of this book, we took a glimpse at the latest advancements in the field of generative AI. We covered new model releases such as OpenAI's GPT-4V, as well as new frameworks to build LLM-powered applications such as AutoGen. Furthermore, we provided an overview of some companies that are actively powering their business with LLMs, such as Notion and Microsoft.

Generative AI has shown to be the most promising and exciting field of AI, and it has the potential to unleash human creativity, enhance productivity, and solve complex problems. However, as we learned in the previous chapter, it also poses some ethical and social challenges, such as ensuring the quality, safety, and fairness of the generated content, as well as respecting the intellectual property and privacy rights of the original creators. Therefore, as we explore the new horizons of generative AI, we should also be mindful of the implications of our actions in the context of the current times. We should strive to use generative AI for good purposes and foster a culture of collaboration, innovation, and responsibility among researchers, developers, and users. Nevertheless, generative AI is an evolving field, and within its landscape, one month is worth several years of technological progress. What is sure is that it represents a paradigm shift, and both companies and individuals are continuously adapting to it.

References

- GPT-4V(ision) System Card: **GPTV_System_Card.pdf (openai.com)**
- AutoGen paper: *Qingyun Wu et al.*, 2023, *AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation*: **https://arxiv.org/pdf/2308.08155.pdf**
- AutoGen GitHub: **https://github.com/microsoft/autogen/blob/main/notebook/agentchat_web_info.ipynb**
- DALL-E 3: *James Betker, Improving Image Generation with Better Captions*: **https://cdn.openai.com/papers/dall-e-3.pdf**
- Notion AI: **https://www.notion.so/product/ai**
- Coca-Cola and Bain partnership: **https://www.coca-colacompany.com/media-center/coca-cola-invites-digital-**

artists-to-create-real-magic-using-new-ai-platform

- Malbek and ChatGPT: https://www.malbek.io/news/chat-gpt-malbek-unveils-generative-ai-functionality
- Microsoft Copilot: https://www.microsoft.com/en-us/microsoft-365/blog/2023/09/21/announcing-microsoft-365-copilot-general-availability-and-microsoft-365-chat/

Unlock this book's exclusive benefits now

This book comes with additional benefits designed to elevate your learning experience.

Note: Have your purchase invoice ready before you begin.

https://www.packtpub.com/unlock/9781835462317