



Applied Marketing Principles

Eric DeRosia, Jan. 2026

Chapter 9a: Four Eras of Generative AI

Our purpose here is to focus on using and applying gen AI to marketing, not on the technical aspects of designing or programming AI models. How it works “under the hood” is outside the scope of this class.

Attributes inherited from the past

Here in 2026, the phrase “artificial intelligence” means something quite different than it used to mean. It’s similar to how the meaning of the word “phone” has changed over time. Therefore, be warned: in your career you will read commentary about AI, but if it’s prior to 2022, it’s almost certainly referring to an earlier form of AI. Likewise, you will very likely encounter claims about martech¹ products using AI, but it may be referring to earlier forms of AI.

Researchers² have identified four basic eras of AI. We’re reviewing them because some of the strengths and weaknesses in today’s AI models are directly inherited from earlier AI models. Each era listed below includes the approximate date when that type of AI began to be used in business settings.

¹ There are many hundreds of martech tools that use “AI.” I would imagine that’s also true of HR, finance, and so on, but I’m only familiar with marketing tools.

² Ran He, Jie Cao, Tieniu Tan, 2025, “Generative Artificial Intelligence: A Historical Perspective,” National Science Review, v. 12, n. 5 (May). Available online at <https://academic.oup.com/nsr/article/12/5/nwaf050/8029900>

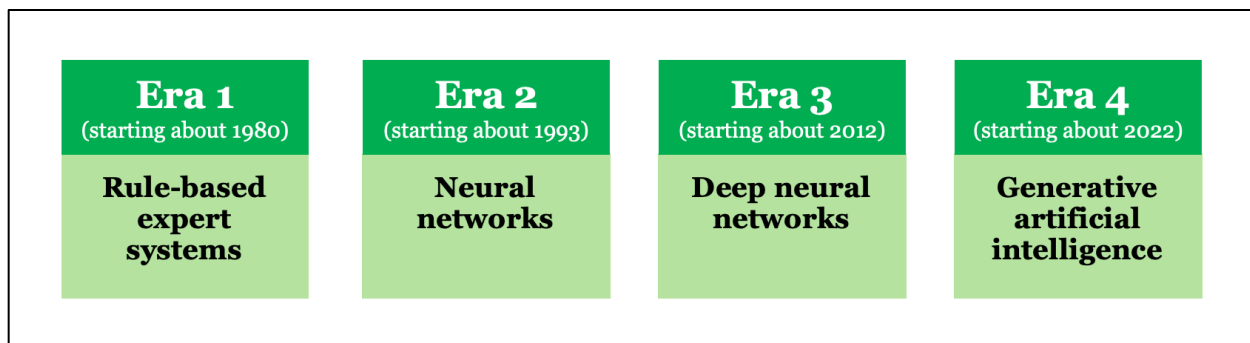


Figure 4.1 Four Eras of AI

Era 1 (starting about 1980): In this era, “AI” meant systems based on “rules” (if X, then Y) that were defined by experts. For example, if someone wanted an AI system to determine whether a short string (i.e., a segment of text) was a valid email address, they would need to identify all the characteristics that a valid email address must have (e.g., the string must contain an “@” symbol, but it can’t be the first or last character in the string). The AI system would read the string and apply the rules, yielding a prediction of whether it was a valid email address.

Era 2 (starting about 1993): In this era, “AI” was statistical models (“neural networks”) that could ‘learn’ from training data, and then can be applied to ‘novel’ (i.e., not-yet-seen) data. For example, consider a neural network designed to determine the likelihood that any given string is a valid email address. Such a model could have been trained on millions of strings that contained valid and invalid email addresses. The neural network model would detect patterns in the data – perhaps even subtle patterns – that would help differentiate the valid from the invalid. After training, the model would be used by giving it a string that might be a valid email address (e.g., perhaps something provided by a person who completed an online form), and the model would predict whether the string is a valid email address. With a large training set and skillful training, neural networks would usually predict very well.

Era 3 (starting about 2012): In this era, “AI” began to mean deep neural networks. They were an advancement over Era 2 in that the statistical models were far more complex. That complexity required much larger training sets, and it offered much greater ability to detect subtle patterns in the data, which improved its predictive

powers. Similar to Era 2, deep neural networks still training on specific examples to do a specific task. In the continuing example, an Era 3 AI model would do very well at predicting whether a given string is a valid email address.

Era 4 (starting about 2022): In this era, “AI” began to mean generative AI that are trained on general tasks. (For the rest of this chapter, when this chapter uses the terms “generative AI” or “AI,” it is referring to these Era 4 models.) The fundamental building blocks of Era 4 AI includes the kinds of neural networks developed in Eras 2 and 3, along with new types of statistical models to connect and enhance the neural network models. Models in Era 4 are much larger than in Era 3 in terms of the number of parameters they estimate. As you would expect, the training sets for Era 4 are much larger than for Era 3, but there is a qualitative difference too: Era 4 AI models are trained on general data, not specific examples of solutions to a specific task. Essentially, the training sets for today’s best **large language models** (LLMs) are all the text the model developers can obtain. Loosely speaking, the training set is the entire publicly accessible Internet. In fact, as you may have heard, one of the challenges in training AI is that all the available text on the Internet might not be enough to effectively get improvements in training new LLMs.³ Researchers are having success with that problem, including ways of getting more out of the data they have.⁴ But the issue highlights two things: (1) the immense size of the training data, and (2) the generalized nature of the training data. In the continuing example, an LLM can judge whether a novel email address is valid, and it will make that judgment after being trained on general text, not simply email addresses.

Advantages of Era 4

The Era 4 models of today have three main advantages over earlier AI models.

Various tasks without retraining. In Eras 1 to 3, AI models were built for only one narrow task, such as deciding whether a small text string is a valid email address. The training data was devoted to that task, and the model could not do anything else. Era 4 AI models are much more broadly capable; they can handle a wide variety of tasks, and they do not require retraining for each new task.

³ Jones, Nicola, 2024, “The AI revolution is running out of data. What can researchers do?,” *Nature* v. 636, n. 8042, 290.

⁴ E.g., Yu, Yang, Kai Han, Hang Zhou, Yehui Tang, Kaiqi Huang, Yunhe Wang, and Dacheng Tao., 2025, “LLM Data Selection and Utilization via Dynamic Bi-level Optimization.” arXiv, v. 2507, n 16178.

Unstructured input. In Eras 1 to 3, AI models were inflexible, requiring input data to be in a structured format that was precisely the same as the format of the training data. Era 4 AI models are more flexible, able to work with input data that does not conform to a strict structure. For example, GPT 5 can successfully answer a question that is worded in a variety of ways, such as “If I get a string that's 20 characters long, and ends in ‘@ederosia’, is that likely to be a valid email address?”

Varied output. In Eras 1 to 3, AI models could produce only a limited set of output. For example, an Era 2 AI might have only two possible outputs: the email address is 0 (invalid) or 1 (valid). Era 4 AI models can generate a wide variety of output, including novel output (i.e., output that has never been generated before).

Inherited attributes

Neural networks began in Era 2, but they’re still relevant because they are some of the foundational building blocks of AI models in Eras 3 and 4. Here are four things that today’s Era 4 AI models have inherited from the neural networks in Era 2.

Not “rules” based. Unlike Era 1, neural networks do not use experts to specify any “rules.” For example, with a neural network predicting the validity of email addresses, no humans needed to figure out all the features that would identify a valid email address. Instead, the neural network derives it automatically⁵ by detecting patterns in the training data.

Vulnerable to problems in the training data. Any biases (i.e., over- or under-representations) in the training data will have a giant influence on the model. In the continuing example, let’s say someone removed all the “.edu” email addresses from the training data. That’s a problem because after training, when the model is asked to evaluate an “.edu” address, the model will probably think it’s invalid. Neural network models reflect the training data.

Unknowable. Even when neural networks are good at making predictions, it’s impossible to know precisely how those predictions are being made. Sure, the model’s parameters can be examined, but it’s essentially just a bunch of numbers. All those numbers work together within the model, and simply inspecting them would not give

⁵ A human manages the training process. But the learning itself happens with automated algorithms.

us any insight into what features the models are looking for, what weights the model is assigning those features, and so on.⁶ In that sense, a neural network is a ‘black box’ that we cannot open. So, if we want to learn about the black box, the best we can do is systematically give it various inputs and examine the outputs it creates. In that sense, when critics of AI say, “We don’t really know how AI works,” that’s sort of true. But it’s not as scandalous or ominous as those critics might be trying to imply.

No storage or retrieval. The neural network does not “store” or “remember” specific information – at least not directly. That’s kind of non-intuitive, so let’s express it this way. For our continuing example about email addresses, let’s say we want to do it without using any AI. Instead, let’s say we’re going to use Excel to simply store all the valid email addresses that appeared in the training data. We’re just making a giant list of the valid emails. If someone asks us whether “ederosia@byu.edu” is a valid email address, we’ll use something like an Excel VLOOKUP function to search for that email in our stored list of valid addresses. We could call that a “storage and retrieval” method. A neural network doesn’t do anything even remotely like a “storage and retrieval” method. A neural network stores only the statistical patterns that are left by the training data, not the training data itself. In fact, it’s possible for a well-trained neural network to judge “ederosia@byu.edu” to be invalid even if it had been in the training data! That’s because the neural network doesn’t directly store all the specific email addresses; instead, it learns the patterns of valid and invalid email addresses. More generally, beyond the email address example, this is what people mean when they say AI doesn’t “know” anything. It’s true, neural networks don’t “remember” specific facts. Instead, they use a vast ocean of observed statistical patterns to make statistical predictions. That might seem tenuous or fragile, less “tried and true” than the “storage and retrieval” method. And in some ways, it is. So, a neural network needs careful testing to make sure it is useful before it can be relied upon.

⁶ In case you’re familiar with linear regression, let’s put it this way. You know how you can look at a regression model and interpret the regression coefficients in terms of their meaning and implications for predictions? There isn’t anything like that inside a neural network model. It’s a giant basket of parameters that reflect an unknowable blizzard of high-order interactions.



Applied Marketing Principles

Eric DeRosia, Jan. 202

Chapter 9b: Modes of Prompts and Output

Marketing has an insatiable need for content of all kinds. Gen AI is very useful in helping us create and evaluate content.



		Prompt's Mode				
		Text	Images	Audio-Speech	Audio-Music	Video
Output's Mode	Text	1 E.g., generating ideas for new products	E.g., image description for brand safety moderation	2 E.g., transcription of one-on-one research interviews	E.g., music tagging, such as genre and instruments	3 E.g., accessibility for people with vision impairment
	Images	4 E.g., images for social media posts	5 E.g., image editing			E.g., frame splitting for advanced CGI videos
	Audio-Speech	6 E.g., audio for podcasts	E.g., Accessibility for people with vision impairment	7 E.g., speech editing		E.g., accessibility for people with vision impairment
	Audio-Music	8 E.g., music for product videos			9 E.g., music editing	
	Video	10 E.g., video for social media posts	11 E.g., "moving" stock photos for product website			12 E.g., video editing

Figure 4.2 Modes of Gen AI Prompts and Output

One property of information – that is to say, information in all contexts, not only the AI context – is that information has a **mode**. As shown in Figure 4.2, AI's prompts and output can be information in a variety of modes: text, images, audio – speech, audio – music, and video. The modes listed in Figure 4.2 are listed in an order that is not coincidental. Generally speaking, the modes are listed in Figure 4.2 in order of the difficulty that AI has in generating it well. Generating text is relatively easy compared to the modes higher on the list, and generating video is much more difficult than generating text.

The cells in Figure 4.2 which are the most useful in marketing are highlighted in green.

Cell 1 (text to text) is AI tools called large-language models (LLMs) that generate text. LLMs are so useful in marketing that we will talk a lot about them this semester.

Various vendors provide trained LLMs the public can use – some are paid, and some are free. Each model has strengths and weaknesses, and vendors are continually trying to improve them, so **leaderboards** are popular ways of tracking the ‘horse race’ of the relative performance of different models. As of this writing, the three models that are at the top of one popular LLM leaderboard are GPT 5 (within ChatGPT from OpenAI), Gemini 2.5 Pro (from Google), and Claude Opus 4.1 (from Anthropic).¹

To clarify

Regarding Cell 1, you might have used ChatGPT to create an image or used Gemini to create a video. Both of those LLMs are able to invoke other types of generative models when needed, and then integrate the output of those tools into the LLM chat window. But the core of an LLM is generating text.

Each of those three is a multimodal LLM, meaning that it can accept prompts that contain text (cell 1), images, audio-speech (cell 2), audio-music, and video (cell 3), generating text output based on any and all of those prompts. In other words, the most advanced LLMs can do the entire top row of Figure 4.2. In marketing, LLMs are so important that we will focus almost all our attention on them this semester. With that in mind, let’s say a few things about the other notable cells in Figure 4.2.

Cell 4 (text to image) is AI tools that create static (i.e., still) images. As recently as a few years ago, the image quality from these tools was not very good, but these tools have improved a lot. These tools are very useful in marketing for creating images used in various places such as product packages, usage instructions, product websites, and social media posts. As of this writing, the two tools on top of the text-to-image leaderboard² are Imagen (by Google) and GPT Image (by Open AI). These tools are invoked in Google’s Gemini and OpenAI’s ChatGPT (respectively) with text prompts such as this: “Create an image that ...”.

¹ <https://lmarena.ai/leaderboard>, as accessed August 12, 2025.

² <https://lmarena.ai/leaderboard/text-to-image>, as accessed August 12, 2025.



One of these images is a photograph, and the other image is from AI (specifically, Midjourney). Which one is AI? The answer is at the end of this chapter.

Cell 6 (text to audio-speech) is AI tools that create spoken audio based on a script. This can take the form of pre-programmed “voices,” as well as customized voices that are trained on audio recordings of a human speaker. As with images, the quality of these tools has been increasing, and it is getting more and more difficult for listeners to detect that the audio recording was created with AI. Marketers use these tools to create the speech in various promotions, including podcasts and voice-over audio in promotional videos. One such tool is Sonic from Cartesia (<https://cartesia.ai/sonic>).

Cell 8 (text to audio-music) is AI tools to create music. The quality of the output of these tools still has “room for improvement,” but it shows promise. One such tool is SunoAI (<https://sunoai.ai>).

Cell 10 (text to video) is AI tools to create videos. Creating videos – including realistic movement by characters – is much more difficult than creating the static images of Cell 4. Considering the difficulty of the task, these tools are impressive. However, to be perfectly frank, it is early days for these tools, and they are not yet “ready for prime time.” These tools are sometimes used by marketers to create promotional videos, and

these tools offer a great deal of marketing potential when their quality improves. As of this writing, the tool at the top of the leaderboard is Veo3 from Google.³

Cell 11 (image to video) is similar to Cell 10, except that the starting point is an image. The uses and challenges of these AI tools are similar to those described above for Cell 10.

Cell 5 (image to image), Cell 7 (audio-speech to audio-speech), Cell 9 (audio-music to audio-music), and Cell 12 (video to video) are AI tools used to edit images, audio, and video. For example, Adobe has incorporated gen AI into their Photoshop, Audition, and Premiere tools for editing images, audio, and video (respectively).

Cell 2 (audio-speech to text) is AI tools for transcribing audio recordings. There are various marketing uses of these tools, such as transcribing all the telephone calls in a customer support call center, yielding text which can then be evaluated by an LLM to give the support representative a score for each call that reflects how well the employee followed the company's rules. Another important marketing use for these tools is expanding access to spoken audio by transcribing it for users who have a hearing impairment.

Cell 3 (video to text) is AI tools for generating written descriptions of the visual contents of videos. One important marketing use is expanding access to video content by using this generated text, in conjunction with a gen AI tool from Cell 6, to generate spoken audio for people with vision impairment. These audio descriptions describe the key aspects of the video, inserted into the audio during natural pauses in dialogue. AI-based systems allow marketers to create these audio descriptions efficiently, making it easier for people with vision impairment access the firm's video content.

As noted previously, the modes are listed in Figure 4.2 in increasing order of the difficulty. Therefore, the AI tools for generating in Cells 4, 8, 10, and 11 are not yet as strong as everyone would like. They certainly are not yet as strong and effective as the LLMs in Cell 1. Because of that difficulty, tools in Cells 4, 8, 10, and 11 often have trouble **adhering to prompts**, meaning that the AI tool's output might not follow the instructions and constraints in the prompt. Said differently, these tools might not do what you tell them to do. Prompt adhesion is especially difficult to get with complex

³ <https://lmarena.ai/leaderboard/text-to-video>, as accessed August 12, 2025.

prompts. Therefore, if you use gen AI tools in Cells 4, 8, 10, or 11 for creating marketing content, here is my advice: the more specific and complex your expectations are, the less satisfied you are likely to be with the results. You can reduce your frustration by being flexible, in the sense that you are open to finding something ‘cool’ that would be useful to you.⁴

The apple on the right is AI.

The apple on the left is a photograph. How do I know? It was posted in 2021, before text-to-image got good. However, it looks like someone photoshopped the shadow. Available online at <https://unsplash.com/photos/rxN2MRdFJVg>

⁴ But, hey, sometimes everything falls into place, and you can get these AI tools to create something complex in just the way you envisioned it. Those moments feel downright magical. I use these tools a lot, and I often find myself chasing the thrill of those magical moments.



Applied Marketing Principles

Eric DeRosia, Jan. 2026

Chapter 9c:

LLMs – Emergence and Access

Current LLMs

Very recently, LLMs have advanced in ways that greatly improves their usefulness to marketing, advancing from base LLMs to reasoning models, and then to hybrid models.

Base LLMs

In a chat window, base LLMs (i.e. the models that were developed first) take a proprietary (and invisible) “system prompt,” combined with the entire chat window, and pass it all to the LLM model as a prompt. The base LLM model then creates an output. In that sense, the base LLM is a single pass through the gen AI process, all with little or no reassessment or processing of the output. Earlier LLMs (even as recent as GPT 4o) were this type of base LLM.

Tokens vs. Words

Strictly speaking, LLMs do not predict words. Instead, they predict tokens, which are not quite the same thing as words. The distinction between words and tokens is important for people who build and train LLMs, but the distinction is not important for people using LLMs for marketing. Therefore, for the sake of clarity, in this class we will use the term “words” rather than “tokens.”

Reasoning models

These models are more sophisticated than base LLMs because the reasoning models employ more LLM resources to accomplish a task. The exact processes that reasoning models use is opaque because most vendors treat that information as a proprietary secret. As reported and investigated by various sources, reasoning models use multiple calls to base LLMs to perform various component tasks such as designing a strategy for accomplishing the overall task, performing essential parts of a task, doing multiple-step tasks, double-checking the accuracy of results, and orchestrating how the various LLM calls work with one another. Whatever proprietary techniques different vendors use, the end result is that reasoning models typically provide better results than base LLMs – especially when the task is complex and requires multiple steps – but they are typically slower and more costly to use as compared to base LLMs. Almost all of the major vendors that offer reasoning models (see Table 4.1) allow direct user control over whether the user’s prompt is processed with the quicker base LLM or the more advanced reasoning LLM. The method for doing this varies by vendor; typically, this is done by selecting “thinking” or “reasoning” in the user interface.

GPT 5.2 (by OpenAI)	Claude Opus 4.5 (by Anthropic)	Gemini 3 Pro (by Google)
Grok 4 (by xAI)	Deepseek V3.2 (by DeepSeek)	Qwen 3 (by Alibaba Cloud)
Llama 4 (by Meta)	GLM 4.7 (by Zhipu AI)	Kimi K2 (by Moonshot)

Table 4.1 Top-Performing Reasoning LLMs¹

¹ The table lists vendors appearing on various leaderboards for reasoning LLMs, as of December 30, 2025.

Hybrid models

A new development in LLMs – as of this writing, only GPT 5 offers this feature – is a “hybrid” model that attempts to automatically get the “best of both worlds” by adapting to each prompt. Hybrid models get speed and efficiency for simple tasks by routing them to base LLMs. At the same time, hybrid models get superior accuracy for complex tasks by routing them to reasoning models. To do this, the most advanced current models do an initial assessment of the prompt to assess how complex the task appears to be. Users can directly influence this assessment – and therefore trigger either base LLMs or reasoning LLMs – with direct choices in the user interface similar to choices described above under reasoning models. For hybrid models, signals within the initial prompt can also be used (e.g., “Think about this carefully” or “Answer this quickly”), and they are often effective at triggering the desired effort level, but the more certain approach is to make a direct selection within the user interface.

LLMs and Emergence

LLMs have emergent properties

LLMs exhibit so-called “emergent” capabilities. Emergence is a classic idea in science.² It rests on the observation that when one system is an element of another system, the element is less complex than the overall system. The more complex system often has emergent properties – that is, additional properties that are not evident from the component systems alone. For example, the following systems are increasingly

To clarify:

One informal meaning of the word “emerge” is “becomes visible.” But don’t be confused by that. Emergence in science is about phenomena that occur in complex systems, and those phenomena can occur even if they are latent (i.e., unobserved). In other words, emergent capabilities are not necessarily easily discovered. Indeed, they might exist without anyone discovering them (yet).

² The core idea of emergence is of ancient origin: in circa 350 BCE, Aristotle referred to “things which have several parts and in which the totality is not, as it were, a mere heap, but the whole is something beside the parts.” Aristotle, *Metaphysics* bk. VIII, 1045a8–10 (W.D. Ross trans., in *The Complete Works of Aristotle* vol. 2, Jonathan Barnes ed., Princeton Univ. Press 1984). For a more recent discussion of emergence in science, see Phillip W. Anderson, 1972, “More is Different: Broken Symmetry and the Nature of the Hierarchical Structure of Science,” *Science*, v. 177, n 4047 (4 August 1972).

complex in the sense that one is an element of the next: cells, organs, bodies, families, and society. There is emergence in those complex systems. For example, an organ like a kidney exhibits emergent characteristics because the cells interact with one another, making them more than a simple collection of cells. Even a good understanding of cells would not give someone a good understanding of organs. The same is true for organs and bodies, and so on.

Emergence is relevant to LLMs because as AI has moved from Era 1 to Era 4 – particularly in the transition from Era 3 to Era 4 – complexity has increased. The result has been emergent phenomena in Era 4.³ That is, some of the characteristics of Era 4 AI models were discovered after training. Rather than being planned, these phenomena have emerged.

For example, in 2022 researchers⁴ tested so-called “chain of thought” prompting, in which “think step by step” is simply added to the prompt. Researchers discovered that this kind of prompting induces LLMs to perform multi-step logical reasoning, and therefore increase the accuracy of LLM output for some tasks. This was not programmed into LLMs; it was discovered after training.

As a consequence, gen AI is nothing short of a new kind of software. With traditional software (e.g., PowerPoint), there is a specific list of functions that have been purposefully designed into the program. It can draw circles, move rectangles, and so on. But with gen AI, the abilities of the program are open-ended and difficult to delineate. That’s not to say they are infinite, but Gen AI is probably able to do more than we realize. We have created a wide variety of sophisticated prompting techniques – see for example, an excellent review by Schulhoff et al.⁵ – but research continues, and we will likely discover new kinds of prompting, new ways of improving output. As a consequence, if some task X is plausible⁶ for an LLM to accomplish, but we cannot get the LLM to successfully accomplish X, then we can’t fairly conclude it’s impossible for

³ For a review, see Berti, Leonardo, Flavio Giorgi, and Gjergji Kasneci, 2025, “Emergent abilities in large language models: A survey.” arXiv, v. 2503, n. 5788.

⁴ Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou, 2022, “Chain-of-thought prompting elicits reasoning in large language models.” *Advances in Neural Information Processing Systems* v. 35, 24824-24837. (“[C]hain of thought processing is an emergent property of model scale that allows sufficiently large language models to perform reasoning tasks.”).

⁵ Schulhoff, Sander, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li et al., 2024, “The prompt report: a systematic survey of prompt engineering techniques.” arXiv, v:2406, n 06608.

⁶ This is about reasonable plausibility. There is a loooooong list of things that would be obviously impossible to do with an LLM, like make a wooden chair, drink water, and experience emotions.

the LLM to do X. Instead, the best we can do is say we don't know how to accomplish X with an LLM. It might be possible, and the only limitation is our prompting skills.

Emergence is not limited to capabilities; problems can also emerge. For example, researchers⁷ have found that LLMs, including ChatGPT, exhibit systematic racism. In a controlled research setting, various LLMs were found to be more likely to recommend the death penalty for defendants whom the LLM could infer were likely to be African Americans, as compared to defendants whom the LLM could infer were less likely to be African Americans. In a separate study, a similar pattern was found for the recommendation of occupations, with the LLMs recommending lower-prestige occupations for people whom the LLM could infer were likely African Americans. No one is claiming that someone purposefully programmed the LLMs to exhibit these deeply unfortunate biases. Instead, these biases are unintended and emergent. Presumably they occurred because the LLMs were trained on giant swaths of the publicly available Internet, so it inevitably contained some toxic bigotry – both overt and subtle. Fundamentally, it is the inherent complexity of the model and its training data that causes the bias to emerge.

Just as LLMs probably have emergent abilities that we have not yet been discovered, so too LLMs probably have problems that have not yet been discovered. Many emergent phenomena are not obvious at first glance.

Using LLMs in companies creates contextual emergence

Think again about the definition of complexity: when one system is an element of another system, the overall system is more complex than the element. We have that precise situation when a manager uses a LLM to do marketing tasks for a company! The LLM – perhaps with multiple prompts doing multiple tasks – is being used as an element in a complex system, where it resides alongside other marketing technology, other employees, the company's policies, and so on. Adding an LLM to the company increases the complexity of the company as a system, so we can expect emergent characteristics in that localized setting. In other words, we can expect contextual emergence: surprises in the way the LLM interacts the rest of the firm. There probably will be new capabilities that you had not imagined, and there probably will be new

⁷ Hofmann, Valentin, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King, 2024, "AI generates covertly racist decisions about people based on their dialect." *Nature*, v. 633, n. 8028, 147-154.

problems that you did not anticipate. If you use an LLM to accomplish marketing tasks, you should (1) expect emergence in that context, and (2) remain observant (“on your toes”) so you can discover those new capabilities and new problems.

Various Methods for Accessing LLMs

Fundamentally, LLMs can be accessed in a variety of ways when managers use them for marketing purposes. They offer different advantages and disadvantages.

Chat window

Undoubtedly, this is the most widely familiar way for people to access an LLM. Each of the major LLM vendors offers a “chat window” in which users can directly query the LLM and receive the LLM’s output.

You should be aware that when you submit a prompt in a chat window, the contents of the entire chat window are submitted to the LLM. As a result, things “earlier” in the chat might influence the way the LLM responds to “later” prompts. In practice, starting a new chat window will yield more pristine output from the LLM.

API from an LLM hosted by a vendor

An API (Application Programming Interface) is essentially a way for one piece of software to talk directly to another. In this context, it means your own application—or a tool you create—can send prompts to an LLM and receive its output quickly and automatically, without anyone having to type the prompts into a chat window. Some LLMs allow you to use the API to give more detailed instructions, such as specifying how much effort you want the reasoning model to exert during its processing. This method is more technologically advanced and usually requires a fair bit of setup, making it better suited for those with programming skills or access to developers. Most major LLM vendors offer API access to their LLMs. To make that access easier, most major LLM vendors also offer software development kits (SDKs) for popular programming languages like Python. Rather than requiring a paid subscription, APIs are on “pay as you go” system. The prices are fractions of a cent per prompt or per

token, so small-scale usage is practically free. However, when APIs are used at a large scale, those costs can quickly add up to substantial amounts.

Playground/console

Do you wish you could have the greater control of APIs without the hassle of writing programming code? If so, I have great news for you. Many of the top LLM vendors offer a “playground” or “console” that gives you the same level of control you can get with an API. As illustrated in Figure 4.3, you can access things that are unavailable in the regular chat interface, such as directly choosing the reasoning effort that GPT 5 exerts (with the options of *minimum*, *low*, *medium*, or *high*).⁸

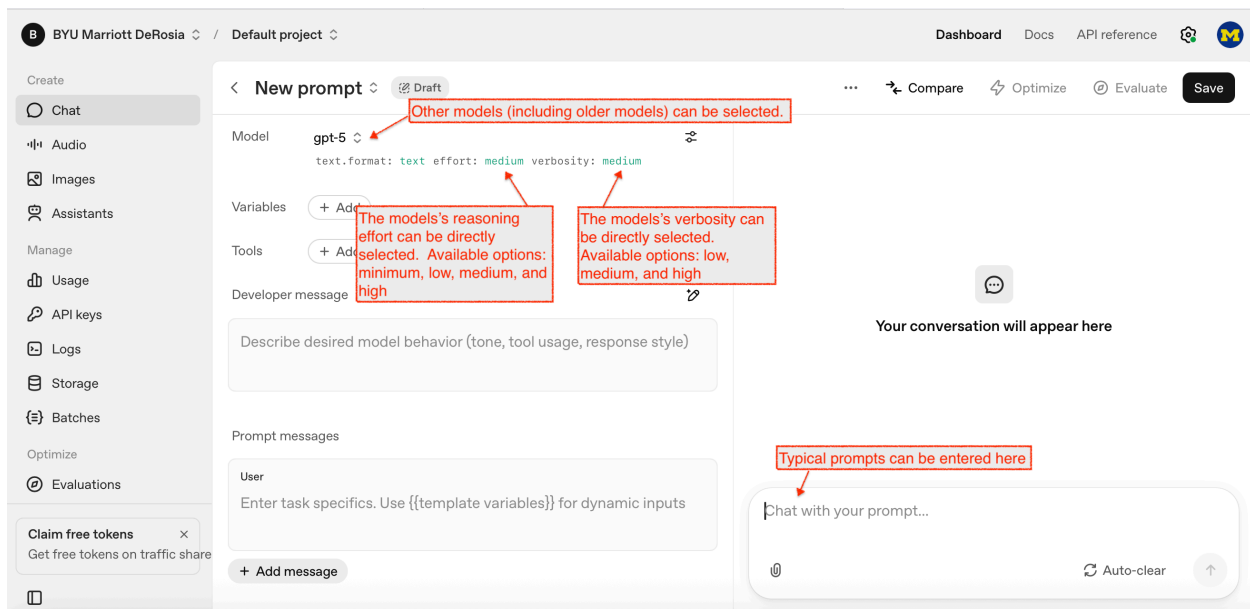


Figure 4.3 Annotated Screenshot of OpenAI’s Playground

⁸ OpenAI’s playground is available at <http://platform.openai.com/chat/edit>. The playground is only available to accounts with pre-paid API balances, which are not covered by a ChatGPT subscription. As of this writing, the minimum prepayment to open a new API account is \$5.

And, hey, if you’re interested in OpenAI’s playground, you might also be interested in Cherry Studio. It’s an open-source tool that gives you playground-like access to the APIs of many LLMs around the web. I mean, you still have to buy API credits for those various LLMs, but Cherry Studio gives you an excellent interface to access those APIs.

API from an LLM hosted locally

Instead of relying on a LLM hosted on the Internet by an external vendor, organizations can run an LLM locally (i.e., on computers or servers they own or directly control). This gives you complete control over the model, keeps your data private, and lets you customize how the LLM works. A big plus is that you don't have to pay the per-prompt or per-token fees that hosted LLM providers charge, which can add up quickly for high-volume use. However, using an LLM locally for large-scale projects requires powerful, high-end computers. Only LLMs that are released as “open source”⁹ can be run this way. These include Llama (from Meta, the owner of Facebook and Instagram), Mistral, and Falcon. That list expanded in August 2025 when OpenAI announced it was releasing two GPT models which can be downloaded and run locally.¹⁰ In general, the LLMs you can host yourself (including the two new GPT models) are not as sophisticated as the top-tier models offered by major vendors, but they can still be highly capable for many marketing uses. Will these open-access LLMs continue to improve at the pace of top LLMs? It's impossible to say for sure. Arguably, it mostly depends on how much Meta continues to invest in improving its Llama LLM.

Commercially available LLM wrappers

Essentially, these are software tools which are built on top of LLMs. At an overview, these software tools collect a user's input of some kind, create a prompt for an LLM, use an API to connect to an LLM, gather the LLM's output, process that output (to a greater or lesser extent), and then provide the user with the output. That might seem unnecessary, since users could simply access the LLM themselves, but these wrappers add value by being built for specific purposes—such as marketing content creation, data analysis, customer support, or brainstorming. Many of these wrappers can also integrate directly with other tools that marketers already use, such as feeding output directly into a content management system like Wordpress, a large-scale commercial email platform like Constant Contact, or an analytics dashboard like Google Analytics. Therefore, the LLM becomes part of the existing workflow rather than a separate step. A

⁹ I acknowledge that some people quibble about whether Llama is truly “open source.” That debate is not important here.

¹⁰ See <https://openai.com/index/introducing-gpt-oss/>, accessed on August 13, 2025.

growing number of companies are developing and selling these wrappers, each trying to differentiate itself. Sometimes, the real value isn't the interface itself but the "secret sauce" in the carefully crafted prompts and workflows that these companies have built – prompts that can produce richer, more tailored results than you might get by going straight to an LLM with your own ad-hoc prompt. Many of these wrappers do not disclose which LLM is powering their tool, so it's difficult to know exactly what's "under the hood." Pricing varies: some pass through the "per prompt" or "per token" cost of the underlying LLM, while others bundle that cost into a flat subscription. The main appeal of these wrappers is speed, convenience, and ease of adoption; the main drawbacks are less flexibility, less control over the underlying technology, and often a higher overall cost compared to building your own API-based solution from scratch.

Within spreadsheets

These tools bring the abilities of LLMs directly into the Excel or Google Sheets environment. Fundamentally, these tools are wrappers, but the usage of these tools is so different and incredibly useful for accomplishing certain kinds of tasks on the job that I'm highlighting them here separately. Instead of copying and pasting text between a chat window and a spreadsheet, these integrations allow you to send prompts and receive LLM responses directly inside your spreadsheet – just like using a built-in formula.

A simple example is shown in Figure 4.4. Notice that in cell B1, the formula refers to GPT. This is not a built-in Excel function. Instead, this is a paid Excel add-in called "GPT for Excel." (In Excel, add-ins are added under Insert > Add-ins > Get Add-ins.)

There are a number of add-ins with similar functionality, so I am not recommending "GPT for Excel" in particular. However, I do recommend the power of the general approach, either in Excel or Google Sheets (where similar plugins are available).

B1		=GPT("Tell me a joke about "&A1)					
	A	B	C	D	E	F	
1	cheese	What did the cheese say when it looked in the mirror? Halloumi!					
2	chairs						
3	childhood						
4	choirs						
5							
6							

Figure 4.4 Example of Using "GPT for Excel"

As you can see in the formula for cell B1, the prompt given to GPT is a literal string along with the value of cell A1. That cell formula triggers the add-in to call the API for GPT (the specific GPT model you have chosen in the add-in's settings). When the add-in receives the results from the API, it places the LLM's output into the value of cell B1. Of course, the true power is you now have the ability to do "Excel stuff," such as copy the formula for B1 down to the rest of the B column, yielding a new LLM call for each cell. I'm no foodie, so I had to look up what halloumi is. I can't guarantee the results will be comedy gold. But I have found this technique of GPT within spreadsheets to be useful when I need to use a LLM for a task that is (a) too repetitive for a chat window, but (b) not sufficiently large in scale to be worth coding a customized solution that uses API calls.