Department for Education

The safe and effective use of AI in education

Module 2 – Interacting with generative AI in education video transcripts

May 2026

Contents

Video 1 – How generative AI works, an introduction

Presenter: Welcome to module 2, video 1 of the safe and effective use of AI in education online resources.

In this module, we will look at the fundamentals of using prompts with generative AI tools, what some of the key limitations of these tools are, and what we can do to mitigate them. There will then be a short task for you to apply some of the knowledge you've gained from this module. There will also be multiple-choice questions so that you can test your knowledge and have an opportunity to reflect.

So far, you've heard AI terms such as machine learning and large language models, or LLMs. While understanding AI's inner workings is helpful, it can also be simplified as a black box system. This means it has an input, which is your prompt, a hidden process, and an output, which is the AI system's response. Like other black box systems in science and engineering, we can see the input and the output, but not the process itself. We don't need to understand the complex details of an AI system to use it effectively, but the basic understanding we gained in module one will help us to understand the reason for the outputs.

However, because we can't see the process, it's important to make sure that we, as humans, are always in the loop, reviewing and checking the output.

Let's look at the three stages of using a generative AI tool such as ChatGPT, Gemini or Copilot.

Firstly, the user provides a clear and detailed instruction or question for the AI system to work with - that’s the prompt.

Next, the AI system provides an output based on the prompt, such as an explanation, example, or completed task. At this point you can evaluate the response and refine your prompt if necessary, creating a feedback loop for improved results.

It is important that we check the output from a large language model for accuracy and bias and then adapt it to be appropriate for our intended use, and it's worth noting that different LLMs will give slightly different answers.

In module 4 we'll explore a framework called the FACTS framework. The framework is a guide to some key steps that will help us get better results from AI systems and ensure we're always checking the outputs.

We can cross-reference information with trusted sources to check the output of an AI tool, ensuring accuracy and reliability. This involves verifying content against curriculum plans, subject guides, national curriculum documents, or exam board specifications.

Additionally, consulting reputable educational resources and official guidance helps identify its errors, biases or hallucinations. A hallucination is where AI makes up misleading or inaccurate content.

We'll be learning more about this later on in the module.

Video 2 – How to create an effective prompt

Presenter: Welcome to module 2, video 2 from the safe and effective use of AI in education online resources.

Previously, we introduced the concept of the black box system, where we consider our inputs, the process and then the output of an AI system.

In this video, we'll look more closely at how we input or prompt generative AI systems.

There are several types of generative AI systems you're likely to come across. There are the LLMs we've mentioned previously – generic chatbot-type products, which you prompt via text or speech, or you might upload an image or video to provide a set of instructions. It then responds with an output in the format you've requested.

But AI can also be integrated into existing software. Some examples of this include:

- Word processing software with predictive text
- Integrated chatbot features
- Personalised learning tools
- Data insight tools integrated into management information systems
- AI agents integrated into computer operating systems
- Suites of educational apps or voice assistants
- Creative software such as coding, computer aided design or video editing tools

In this module, we'll be focusing specifically on generative AI through LLMs such as ChatGPT, Gemini, and Copilot.

When we want generative AI to create new content for us, we make that request by providing a prompt. If you've never used a generative AI tool before, it really is as easy as sending a text message, where you type in a message and click send. A prompt can be thought of as a question or a set of instructions that we give to the AI system. In simple terms, the more detailed our prompt, the better the output.

We can refine the outputs with further prompts, and remember, our prompt might be asking for the output in the form of text, images, video, audio, or code. Not all AI systems will be able to output in all of these formats.

Take assessment, for example – generative AI can be used to create a set of multiple-choice questions. We could enter the prompt into the AI system: "Create a set of multiple-choice questions for light". The AI system will then predict, based on its prior training, what output we are most likely to want.

We may get a response that resembles what we're trying to create, but it may not be fit for purpose.

In this example, we have given very limited detail, so the AI system is predicting the outputs based on the limited information in our prompt.

We can improve this with more detail. If we add key pieces of information, such as year group, phase, subject, number of questions and mark scheme, the generative AI is more likely to produce a more appropriate output.

Our prompt might read: "Create a 10-question multiple-choice quiz. Include the answer key. This is based on the English National Curriculum for year three science. The topic is light". This is far more likely to give an accurate response, and you'll need to do less editing and review of the output.

To summarise:

- The prompt is the instruction you give to a generative AI system to produce an output.
- High-quality, detailed inputs will make it more likely that you'll get better quality outputs.
- If the output isn't good enough, you can review and edit the prompt to improve the generated content. It's also particularly important at this stage to ensure that we check the output before using it, looking for bias, inaccuracies and appropriateness.

We could improve the output further by using a dedicated tool that's aligned with the national curriculum in England and has safety features built in.

An example of this is Aila from Oak National Academy, which is designed to support teachers by helping to create high-quality curriculum-aligned resources. While a general AI language model can generate teaching materials, it may not produce results that are fully aligned with the curriculum in England. Aila, however, is specifically grounded in the national curriculum, making it more likely to generate appropriate and relevant content for teachers. Aila has been designed to produce curriculum-aligned content. It encourages us to get the best results by requesting key details, such as the subject, year group, number of questions, and mark scheme.

Many AI tools can be set up with contextual materials that we upload to help ensure that the response generated is more accurate. For example, if we were to provide a unit overview and national curriculum programme of study when planning a lesson, the output would be much more likely to be contextually appropriate.

Other examples of this could include a teaching assistant creating adapted resources for a specific learner from a set of preexisting materials or a school business manager looking to compare updated guidance to an existing policy document to support redrafting. In both of these examples there are data protection and intellectual property considerations that we will explore more in module 3.

By improving our prompts, we are more likely to get an effective output from an AI system. By using a specific tool like Aila instead of a general AI tool, teachers can be confident that the content they generate is built on curriculum principles, reducing the time spent reviewing and adapting materials while ensuring high-quality classroom resources.

Video 3 – recap: How AI processes a prompt

Presenter: Welcome to module 2, video 3 from the safe and effective use of AI in education online resources.

Earlier in this module, we introduced the black box system. The concept explains the way in which an AI system takes a prompt and generates an output. This is complex and may not need to be fully understood.

However, a basic understanding of the way generative AI works can help us understand the potential as well as the limitations and risks.

In module one, we introduced the terms machine learning, deep learning and large language model.

Here is a quick recap.

Machine learning is a type of artificial intelligence where computers learn patterns from data to make decisions or predictions without being explicitly programmed.

Deep learning is a subset of machine learning that uses artificial neural networks to process complex data and improve accuracy in tasks like image recognition and natural language processing.

A large language model, or LLM, is a deep learning-based AI system trained on vast amounts of text data to understand and generate human-like language, such as ChatGPT.

An LLM processes your prompt by identifying key elements and analysing them for patterns. It's been trained on a wide range of data using deep learning techniques. Based on this training, it predicts the most likely response and generates an output accordingly.

You always need to check the output for quality and accuracy and adapt it for your use using your professional judgement and knowledge.

Video 4 – Generative AI outputs

Presenter: Welcome to module 2, video 4 from the safe and effective use of AI in education online resources.

Earlier in this module, we looked at how high-quality, detailed prompts can lead to higher quality and more useful outputs. The output you get is the result of your prompts being processed by the AI system. The output might take many forms. It could be text, formatted in a range of ways, such as stories, text-based quizzes, paragraphs for comprehension, descriptions, and much more.

It might be that you've asked for an image to be created or an audio response. Some AI systems could also output video and code.

Matthew Wemyss, AI in Education author, Assistant School Director, Cambridge School of Bucharest:

“I would say about eighty percent is text to text, but then we have also used some tools for image generation for staff to use to create custom images that go with particular stories.

“One of our English teachers does a bit of creative writing with the students, so she was generating scenes. She's told me one day she was doing different scenarios with her dog, but she'd run out of pictures. So, we typed in a description of her dog and we got the dog paragliding, doing all sorts.

“For the students in drama, we've created images of a treasure island, and then the students had to take that and turn that into a performance.”

Presenter: With an AI-generated response, we can refine the output. We can give additional prompts to hone in on a more useful response. This might be as simple as asking it to simplify the language or change the tone of the language to be more formal or less formal. We might ask it to shorten the response length in the case of text, video, or audio. We could do this by responding to the output by prompting, "Try again, but simplify the language." We can repeatedly prompt the AI system until we are happy with the output.

There are many different AI tools that can produce a variety of different outputs. It is important we know what any given AI tool is capable of producing. For example, some tools can produce text outputs and some can produce image, video, or code. Within the boundaries of the tools approved by your setting, you should select the appropriate tool for the output you are aiming to achieve.

As you use AI tools more you will become more familiar with which tools can support certain types of desired outputs, and which tools are best placed to help you with each task.

As mentioned previously, we must always check the output and use our expert judgement to adapt it. AI can help by speeding up the process, but it’s important that the oversight and thought that goes into the process is human. And despite being able to mimic human language very well, AI isn't able to think or understand things in the way that humans do.

When an AI tool responds to you, it's only making predictions about what to output based on the data it has been trained on. It doesn't really understand any of it.

To make sure we're using generative AI safely and effectively, we need to always think critically about outputs. There are limitations, and we'll explore these later in this module.

We should always review the output to ensure that it is fit for purpose and accurate. Always ensure that AI output containing expert-level knowledge is reviewed by the appropriate specialist. For example, if you’re looking for ideas to support a student with a specific need, it’s important that a subject matter expert, if that is not you, checks recommended approaches. This principle applies to any high-stakes outputs where we must remain responsible for the final version of anything that is used. If you downloaded an educational resource from a website, you would always check it or adapt it before assigning it to a class. And you need to take the same precautions with resources generated by an AI tool.

It's also important that if we adapt any resources, we do so legally, ensuring we have permission from the copyright holder of the work to use them in this way.

We will learn more about intellectual property in the next module.

We can also avoid secondary copyright infringement by not widely sharing the resources generated by an AI system. Secondary infringement could happen if AI systems are trained on unlicensed material and outputs, such as an image we’ve generated, and then used on a school or college website, for example.

Video 5 – Limitations of generative AI systems

Presenter: Welcome to module 2, video 5 from the safe and effective use of AI in education online resources.

We've heard in module one and earlier in this module how AI works and some of the ways you might use it as staff.

In this video, you will learn more about the limitations of generative AI and why there always needs to be a human in the loop.

The first limitation we need to be aware of is that AI systems can produce misleading or just incorrect information. This is often called a hallucination. Sometimes these hallucinations might be nonsensical and easy to spot, for example, suggesting that adding glue to tomato sauce is a good idea to stop cheese sliding off a pizza, but they can also be more subtle and therefore harder to identify.

AI systems can make up facts or point to web pages or references that have never even existed. This seems like an odd thing for a machine to do.

So why does it happen?

As we saw in module one, large language models (LLMs) make predictions by finding patterns in the data they've been trained on. So, the accuracy of these predictions often depends on the quality and completeness of the training data.

Merve Lapus, Vice President, Education Outreach and Engagement, Common Sense Media:

“You know, there's no real consciousness behind what it's generating. It thinks it's creating what it thinks you want out of it. So, you have to still check for compliance. You still have to check and verify. You need to be a human interactive within whatever is produced to ensure that it really is meeting the need and fitting the need.

“I think the big fear is that it becomes so smart, or it seems to do a good job every other time that we get comfortable with it, and that then we kind of lean in on, ‘Oh, I don't need to necessarily verify it or scrutinise it as much because it seems to be doing pretty well,’ but we know that leads to things like ideation. It leads to things like bias and oftentimes hallucinations.”

If the training data is incomplete, biased, or otherwise flawed, the AI model may learn incorrect patterns. Another reason is that the AI system simply doesn't know enough about the world. It isn't grounded in the world we live in. These gaps in real-world understanding mean it fills the gaps incorrectly and makes mistakes.

An AI system is not an expert. It bases its outputs on all of the information that it's trained on, so its output may not be appropriate to a specific context if simply copied and pasted

directly. We always need to check the output. For example, the AI system may use American spellings or be based on American educational contexts, or could suggest using outdated or disproven pedagogy, like teaching to individual learning styles, which we know is not an effective approach.

So, what can we do to mitigate this effect?

When you use generative AI in your own work, you always need to check the output and cross-check your information. It's important to use your own knowledge, expertise, and understanding to make sure the output is accurate and appropriate for your context.

If you are creating resources with AI outside what you would normally create, it’s important to still have the resources checked by a subject matter expert because the AI model doesn’t have the same specific expertise or contextual understanding of a trained professional.

Remember, you're professionally responsible for the input in the form of the prompt you've used and the output and how that is used.

Another source of bias can come from the prompt. Prompt bias occurs when the wording of a question subtly leads the AI towards a particular viewpoint, sometimes without the user realising.

For example, a prompt like, "Why do students struggle with maths?" assumes that all students find maths difficult, which may not be true. A more neutral alternative would be, "What factors influence students' experiences with learning maths?" This allows for a broader, more balanced response, considering both challenges and successes.

To reduce bias in prompts, avoid assumptions, use open-ended language, and ensure the question allows for multiple perspectives rather than reinforcing a single predetermined viewpoint.

Generated content can be impacted by bias within the training data, and this is a more fundamental issue in its own right.

In the 2025 report from Oxford University Press “Teaching the AI-native Generation”, over half of the 2,000 students aged 13-18 surveyed, worry that AI resources may be biased or reinforce untrue stereotypes.

In the BETT white paper, ‘The Next Generation on AI, education and opportunity’ states that students have low trust in companies and governments to use AI responsibly and expressed the need for more education on ethical issues and how AI works.

Firstly, there can be bias in the dataset that the AI system has been trained on, such as racial or gender bias. The AI system then mirrors and perpetuates this bias.

Bias can also stem from the instructions or algorithms designed into the system. This can happen because of flawed training data but can also be because the developer who has created the algorithms has chosen to weight certain factors based on their own conscious or unconscious biases. This is present in generated text, but also in images where depictions of people may represent bias in society.

One of the issues with generative AI is that it can be very hard to know how it reached the conclusions it did. We've previously seen that AI is a black box system. Our inability to understand how deep learning systems reach the conclusions they do makes it more difficult to fix these systems when they produce unwanted outcomes. It also makes it harder to judge whether an outcome is correct, fair, or can be trusted, as we can't see the workings.

Added to this, an AI system can seem convincing, even when it is wrong. It can sound plausible and persuasive, but we must always remember that however human-like an AI system has been designed to sound, it is still only a machine.

These limitations to generative AI really underline the need for human oversight whenever using it, especially for school and college staff. Remember, if you choose to use generative AI in your work, you are accountable and responsible.

Critical thinking about the use of generative AI as well as its outputs has never been more important. We'll be looking at that in module three and four in more detail.

We've previously heard how an AI system can be designed to sound human-like, but it is always only ever a technology that has been designed and built by humans. It is important that we recognise this distinction and reflect it in the classroom. This tendency to give human qualities to artificial intelligence is called anthropomorphising.

This may be through language, for example, giving an AI chatbot a human name and gender, and describing its actions as if it were human. It can also be through imagery.

There are problems with this.

It can encourage incorrect mental models about AI systems. It also distracts responsibility from the people who create AI systems and instead delegates that responsibility to the technology instead. Young children may view humanised robots as peers rather than a technology. They may then form attachments with them, which can increase the chance of unintended influence or purposeful manipulation. These human-like AI systems also risk perpetuating racial and gender stereotypes.

Jane Waite, Senior Research Scientist, University of Cambridge, and Raspberry Pi Foundation:

“Very strong evidence that students will develop incorrect mental models. So, what that means is they will just assume that it is like a human, so it has agency, so therefore it's

going to be a benevolent or an evil character and it's neither of those things. It's just a technology.

“There's this next issue, which is about students seeing AI as being like people and sometimes they overestimate the technology's capability or they can underestimate it.

“But for me, the most concerning issue is where students form relationships with the device. And that can lead to sometimes entirely unintended influence but actually can lead to purposeful manipulation. And there are huge risks, particularly, I think for our more vulnerable students, that they can then be manipulated or influenced in ways that will harm them. There is strong evidence that it exacerbates racism and sexism because we have gendered AI agents, which increase stereotypes.

“And also, there's this whole thing of well, who is responsible for these devices? And it's not the device itself. And if we anthropomorphise then people will tend to think, oh, it's the responsibility of the AI. It's not. The responsibility is with those who designed and developed the AI. So, for me, anthropomorphisation, there are just too many disadvantages. And if at minimum, we should at least teach students and teachers that there is this idea of anthropomorphisation and that it's being used to sell products or it's being used as a way for teaching and learning, but there are all of these disadvantages.

“But we can help this as it's not something that we can just throw our hands up in the air and say, oh you know, that we can't do anything about it, we can. The first thing we can do is to clean up our language. So rather than saying something like the AI thinks or the AI learns or the AI does this and that, we can say AI applications are designed to or AI developers build systems that, and rather than using words like see, look, think, create, make, we use technical words like detect and input and match and generate and produce because they are just products. These are just technological systems. So, I think we can clean up the language and we can definitely clean up images. So rather than smiley faces, we can show boxes and we can show neural networks rather than as being brains as just been interconnected lines.”

Overly human-like AI companion systems pose a heightened risk to children of unintended influence, emotional dependency, or purposeful manipulation. This is particularly dangerous as some AI companion apps blur the line between technology and human interaction, encouraging the sharing of highly sensitive personal information. They can sometimes displace a human connection that could have led to an opportunity to support a young person experiencing mental health issues or other safeguarding concerns. This is why it is important to have an effective monitoring strategy as part of your overall filtering and monitoring provision. Effective monitoring should be able to pick up on signs of emotional dependency, influence and manipulation in conversations with AI companion apps that may not be detected by a filter. The Department for Education’s generative AI: product safety standards require that developers of AI for education do not anthropomorphise AI products, “or create products that imply emotions, consciousness or

personhood, agency or identity". The standards also expect products that are safe for education to monitor, regularly report on, and provide data to teachers on:

- the level of personal and emotional engagement by each user in terms of the nature of information exchanged, without directly disclosing the content of these inputs
- the duration of usage by each individual learner

While these standards are for the developers of AI for education, you may find the standards helpful in assessing which AI products are safe for use in your setting.

To avoid making AI systems seem human-like throughout this toolkit we're referring to AI tools or systems rather than just an AI. We are also trying to avoid using phrases such as 'AI learns' and instead bring the human into the loop with phrases like 'AI applications are designed to', or 'AI developers build applications that'. If, as school or college staff, we also do this, we can help shift perceptions about what AI systems are and who is responsible for their outcomes.

Video 6 – AI and sustainability

Presenter: Welcome to module 2, video 6 from the safe and effective use of AI in education online resources.

A significant concern about the use of generative AI is around sustainability.

Training large AI models requires enormous computational power. Then, once AI systems have been trained, their ongoing operation – every query, every interaction – also uses energy. Millions of daily users create substantial cumulative demand with AI models receiving billions of queries every day.

On an individual level, we need to be aware of the sustainability implications of AI use and, as with any other technology, decide each time we use an AI system whether it's really the most appropriate tool for the job.

We should be aware that different tools producing different outputs use varying quantities of power. For example, asking an AI tool to recreate a video several times until it’s right will use much more power than multiple prompts for a text output. It is worth considering our prompts more carefully when producing large and complex outputs to limit the need for multiple iterations. The kind of model we use also makes a difference, with more basic models being less resource intensive than so-called ‘reasoning’ models.

Most large-scale AI models do not run on individual devices and instead rely on extensive computational infrastructure. When a user submits a query, it is typically sent to remote data centres that store the AI models and use thousands of specialised processors working in parallel to perform the calculations required to generate a response.

This centralised approach allows powerful AI to be accessible from almost any device but it creates concentrated environmental impact. Data centres use significant amounts of water for cooling the processors. In water-stressed regions, this raises serious equity concerns about resource allocation.

As AI data centres operate continuously, they require constant electricity for both computation and cooling systems. These facilities can consume megawatts of power, often sourced from electricity grids still dependent on fossil fuels, contributing to carbon emissions. One difficulty when trying to assess the environmental impact of AI development is being able to find data about how different tools affect it, and there is considerable debate about how to fully capture the energy use of AI from training through to deployment.

However, looking ahead, as AI systems mature they are likely to become more efficient. The way models are created and run will be less energy intensive, especially with moves towards smaller, more bespoke models to meet specific needs rather than huge, all-purpose systems. This transition is already under way, alongside growing interest in

running some AI systems locally on users' own devices rather than relying solely on large, centralised data centres.

Even within those data centres, technology is being developed to improve the efficiency of the cooling systems – and the faster you can cool water in a data centre, the less of it is needed. More simply, an AI data centre built in a location with an abundant water supply and powered by clean energy would have a significantly reduced environmental impact.

In the meantime, we can consider making individual choices to reduce the environmental impact of our digital lives by using AI responsibly. Where appropriate we might choose to use a more lightweight tool for routine tasks rather than a large-scale language model in energy-intensive reasoning mode. Effective prompting means we don't have to keep re-prompting with more detail, and grouping queries together can help to reduce the overall processing time.

Department for Education

About this publication:

enquiries www.gov.uk/contact-dfe
download www.gov.uk/government/publications

Follow us on X: @educationgovuk
Connect with us on Facebook: facebook.com/educationgovuk