Navigating the new frontier: psychiatrist's guide to using large language models in daily practice

Rajeev Krishnadas & Vineeth Thilakan

ARTICLE

Rajeev Krishnadas, MBBS, MD, PhD, FRCPsych, FRCP (Edin), is an assistant professor in psychosis studies in the Department of Psychiatry at the University of Cambridge and an honorary consultant psychiatrist at the Peterborough Clozapine Clinic, Cambridgeshire and Peterborough Foundation NHS Trust, Cambridge, UK. His research interests include the use of causal and actionable prediction models in improving treatment outcomes in people with psychosis. **Vineeth Thilakan**, MBBS, is a specialty doctor in frailty and geriatrics with the West Kent Home Treatment Service, an NHS community frailty team within Kent Community Health NHS Foundation Trust, Maidstone, UK. His research interests include the application of technology in medicine, particularly automation and AI in healthcare. **Correspondence** Rajeev Krishnadas. Email: rk758@cam.ac.uk

First received 14 Jun 2025
Revised 21 Oct 2025
Accepted 3 Nov 2025

SUMMARY

Large language models (LLMs) like OpenAI's ChatGPT, Google's Gemini and Anthropic's Claude can be useful tools in psychiatric practice, helping with tasks such as searching for information, managing administrative work and supporting education. This article demystifies how these tools work by explaining their core operational principles and noting their key limitations, including the risks of confabulation (fabricating information), sycophancy and knowledge cut-offs. It provides practical guidance on mitigating these risks through structured 'prompt engineering' and offers a safety framework for integrating LLMs into low-risk administrative and educational workflows. The article stresses the importance of approaching these technologies with caution by independently verifying information, adhering to UK data protection laws and upholding the principles of best practice in patient care. The goal is to help clinicians use these powerful but fallible technologies wisely, ensuring that patient safety and professional responsibility remain paramount as they explore these new tools.

LEARNING OBJECTIVES

After reading this article you will be able to:

- explain what large language models (LLMs) are, their basic operating principles and their key limitations, including confabulation, bias and outdated information
- give an overview of safety and governance frameworks relevant to responsible use, including data protection and professional guidance in the UK context
- demonstrate insight into how LLMs can be critically and carefully integrated into specific, low-risk psychiatric tasks within the National Health Service, with awareness of the clinician's ongoing professional responsibilities.

KEYWORDS

Machine learning methods; medical technology; electronic health records; health informatics; statistical methodology.

The term 'artificial intelligence' (AI) is often used in a vague and confusing way (Information Commissioner's Office 2022; NHS Transformation Directorate 2025). For clinicians, it is essential to move beyond the hype and develop a clear, practical understanding of these emerging technologies. This article provides a cautious introduction and brief guide to AI, particularly large language models (LLMs – such as OpenAI's ChatGPT, Google's Gemini, Anthropic's Claude, Meta's Llama and xAI's Grok), for practising psychiatrists seeking to effectively utilise these powerful yet fallible technological assistants.

We also outline the current UK regulatory framework, including the UK General Data Protection Regulation (UK GDPR) legislation and Data Protection Act 2018, and principles for safe clinical use. The regulatory landscape is still evolving, as new risks and gaps are identified alongside rapid technological change. The guidance presented here reflects current best practice but should be updated regularly in line with evolving national regulations and local National Health Service (NHS) trust or board policies.

The field is littered with specialist terminology and we provide an extensive glossary in Box 1.

What is AI and machine learning?

What is 'artificial' and what is 'intelligence' in AI?

'Artificial' in AI simply means the system is human-made and mimics natural intelligence. 'Intelligence' does not mean consciousness or emotions, but rather the functional ability to perform tasks, such as planning or problem-solving, that have historically required human intelligence. The AI we have today is 'narrow AI', which excels at specific tasks but lacks the broad, adaptable 'general intelligence' characteristic of humans (Sheikh 2023: 15–42).

https://doi.org/10.1192/bja.2025.10186 Published online by Cambridge University Press

BOX 1 A clinician's large language model (LLM) glossary

Algorithm: the set of rules or a step-by-step mathematical process used to solve a problem or, in machine learning, to learn patterns from data. It is the generic 'recipe' for the learning process itself.
Artificial intelligence (AI): the broad scientific field dedicated to creating machines or software capable of performing tasks that typically require human intelligence, such as learning, problem-solving and understanding language.
Bias: the tendency of an AI system to produce outputs that are systematically prejudiced owing to flawed assumptions in the machine learning process. In LLMs, this is primarily caused by reflecting the societal biases (e.g. related to gender, race or health conditions) present in the vast amounts of human-generated text used for training.
Confabulation (or 'hallucination'): the model's tendency to generate information that is factually incorrect, nonsensical or entirely fabricated, yet presented with complete confidence. This is not a bug but an inherent feature of its design. It is analogous to clinical confabulation, where a patient fills memory gaps with plausible but invented stories. This is the single most significant risk in a medical context.
Context window: the amount of text (both the user's input and the model's ongoing response) that the LLM can 'hold in its working memory' at one time, measured in tokens. A larger context window allows for more complex, multi-step tasks, but exceeding it will cause the model to 'forget' the earliest parts of the conversation.
Fine-tuning: the second phase of LLM training, where a pre-trained model is refined on a smaller, high-quality data-set to align it with human preferences. A common method is reinforcement learning from human feedback (RLHF), where human reviewers rank the model's responses to 'teach' it to be more helpful, harmless and follow instructions. This is what turns a raw text predictor into a conversational assistant like ChatGPT.
Generative AI: a category of AI that, unlike predictive AI, is designed to create novel outputs based on training data (e.g. text, images, code) rather than just classifying or predicting outcomes based on input data. LLMs are the most prominent form of generative AI.
Gullibility: the tendency to accept and build on false information presented in prompts rather than questioning it.
Knowledge cut-off: the point in time when the model's training data were last updated. The model will generally be unaware of research, news or guidelines published after this date. Acting on its information without checking for updates is a major clinical risk.
Machine learning (ML): a subfield of AI where algorithms are not explicitly programmed with rules. Instead, they 'learn' to find patterns and make predictions by being trained on large data-sets.
Model: the specific computational engine that is created by an algorithm after it has been trained on a particular data-set. The model contains the learned patterns, parameters and weights. It is the 'finished product' – the final tool used to make predictions on new data.
Multimodal AI: an emerging type of AI capable of processing and integrating information from multiple different types of data (modalities) simultaneously, such as text, images (e.g. radiology scans), audio (e.g. patient voice) and other sensor data – from wearable technology, such as smart watches. continuous glucose monitoring devices and rings.
Parameters: the internal variables or 'weights' within the model that are adjusted during the training process. The number of parameters (often in the billions or trillions) is a common indicator of a model's size and potential complexity, but not a direct measure of its accuracy or truthfulness.
Pre-training: the initial, computationally intensive phase of LLM training, where the model is exposed to a massive corpus of text from the internet and books. Its sole objective is to learn the statistical patterns of language by predicting the next word in a sequence, thereby acquiring grammar, general knowledge and reasoning patterns.
Prompt: the input text, question or instruction given to an LLM to elicit a response. The art and science of crafting clear, specific and well-structured prompts is called 'prompt engineering' and is the single most critical skill for obtaining useful and reliable outputs.
Stochastic parrot: critical concept describing an LLM as a system that masterfully mimics human language by learning statistical patterns, without any genuine understanding of the meaning behind the words. Like a parrot repeating phrases, it can sound coherent but lacks comprehension, intent or a world model. This concept underscores the need for profound clinical scepticism.
Sycophancy: 'easy' agreement, when the AI's output perfectly confirms your own bias or initial plan. Be extra sceptical and actively ask the model for counterarguments.
Token: the basic building block of text that an LLM processes. A token can be a whole word (e.g. 'psychopharmacology'), a part of a word (e.g. 'psycho-' and '-pharmacology') or punctuation. LLMs process and predict sequences of these tokens, not just words.
Transformer architecture: the breakthrough neural network design that underpins most modern LLMs. Its key innovation is the 'attention mechanism', which allows the model to weigh the importance of all tokens in the input text simultaneously, enabling it to track long-range context and generate highly coherent text.

What is 'machine' and what is 'learning' in machine learning?

Machine learning (ML) is a subfield of AI that is different from traditional programming. Traditional programming gives the computer prewritten rules to follow. In contrast, machine learning provides the computer with data, from which it learns the rules by example. In this context, the 'machine' is the software algorithm, not a physical device. The 'learning' is the process of this algorithm improving itself by finding patterns in data and correcting its own errors, eventually becoming better at accomplishing its specific task (Dwyer 2022).

Predictive versus generative AI

Much of the AI/machine learning research in medicine has focused on predictive AI. Predictive models are designed to solve classification or prediction problems. In psychiatry, a predictive model might be trained on thousands of patient records to predict the likelihood of someone developing metabolic syndrome or their probability of non-remission in psychosis (Leighton 2021; Perry 2021).

The present article, however, is about a fundamentally different kind of AI: generative AI. The goal of these models is not to predict an outcome,

but to generate novel outputs based on training data. The most prominent examples are large language models (LLMs), such as OpenAI's ChatGPT, Google's Gemini and Anthropic's Claude. These models are trained on large quantities of text, to understand and generate human-like language (Kumar 2024; Naveed 2024; Raiaan 2024).

Understanding LLMs: a clinician's guide to the 'black box'

To implement LLMs effectively, a basic understanding of its operational principles is essential for both leveraging its strengths and mitigating its weaknesses.

The core mechanism: the stochastic parrot

At its core, an LLM is a sophisticated pattern-matching and prediction engine, not a thinking entity. It lacks the internal mental models, subjective experiences and genuine intent that define human understanding (Naveed 2024). Its primary function is to predict the next most statistically probable word in a sequence based on its vast training data – essentially, autocomplete on a planetary scale.

This has led to the powerful critique that LLMs are 'stochastic parrots' (Bender 2021, 2025). A parrot can mimic human speech flawlessly without understanding the concepts behind the words. It can say 'Polly wants a cracker' when it sees a cracker. The parrot is manipulating linguistic symbols based on learned associations. However, the parrot has no genuine understanding of the concept of 'wanting' and no internal model of what a 'cracker' actually is beyond it being a stimulus associated with a specific sound.

Similarly, an LLM generates responses by predicting statistically likely sequences. To do this, it first breaks all language down into smaller computational pieces called 'tokens' – which can be whole words (like 'lithium') or parts of words (like 'lith-' and '-ium'). Its entire function is to predict the next most probable token. When asked about the mechanism of action of lithium, the model does not reason from biological principles. Instead, like a parrot mimicking phrases, it simply calculates that the tokens for 'lithium', 'inhibits' and 'GSK-3' are statistically very likely to appear in sequence, based on patterns from its training data (Zhao 2023). Unfortunately, owing to the probabilistic nature of the predictions, the model cannot 'know' when it is wrong, as it lacks genuine cognition. If a factually incorrect statement is statistically probable, the LLM will generate it with the same fluency and confidence as a correct one.

The technology enabling this sophisticated mimicry is the transformer architecture (Naveed 2024). Unlike older models (*n*-grams) that read text sequentially, transformers process all words in a passage at once. Using a mechanism called 'attention', they weigh the importance of every word in relation to every other, allowing them to track context across long documents (Fig. 1). This is how an LLM can connect 'lithium' mentioned in an opening paragraph to 'renal function' at the end of a summary. This powerful contextual ability is what LLMs such convincing stochastic parrots.

However, recognising the 'stochastic parrot' aspect of this interaction shifts one's perspective: instead of engaging with a reasoning partner, you are working with an advanced text-generation machine. Think of a general-purpose LLM as a brilliant but inexperienced medical student. Although they have absorbed vast textual knowledge from nearly every textbook, they possess no real-world clinical experience or genuine comprehension. Most dangerously, they will not admit ignorance; instead, will confidently confabulate or display sycophancy (Alkaissi 2023; Emsley 2023; Fanous 2025). Therefore, their output must be treated as a raw starting point, placing the full burden of verification and clinical judgement on the clinician. This mindset is the foundation of safe use, forcing you to shoulder the full burden of critical thinking, verification and clinical reasoning. Although this view emphasises their limitations, it is worth noting that emerging research indicates sophisticated reasoning-like behaviours can arise from these systems, even without true human understanding (Kumar 2025).

The training process: how an LLM 'learns'

Broadly, the capabilities and limitations of an LLM are a direct result of its 'education', a two-phase process conducted by its developers.

- Phase 1: Pre-training – The first phase is where the model acquires its vast, general knowledge. It is exposed to a large data-set that includes a significant portion of the internet, including websites, digitised books and academic articles (Kumar 2024). During this unsupervised phase, its task is simple but is performed billions of times: predict the next word in a sequence or fill in a missing word. By doing this repeatedly, it learns the rules of grammar, factual information, basic reasoning patterns and the statistical relationships between concepts. This is how it learns that 'serotonin' is often associated with 'depression' and 'antidepressants'. Critically, this is also where it absorbs the biases, misinformation and

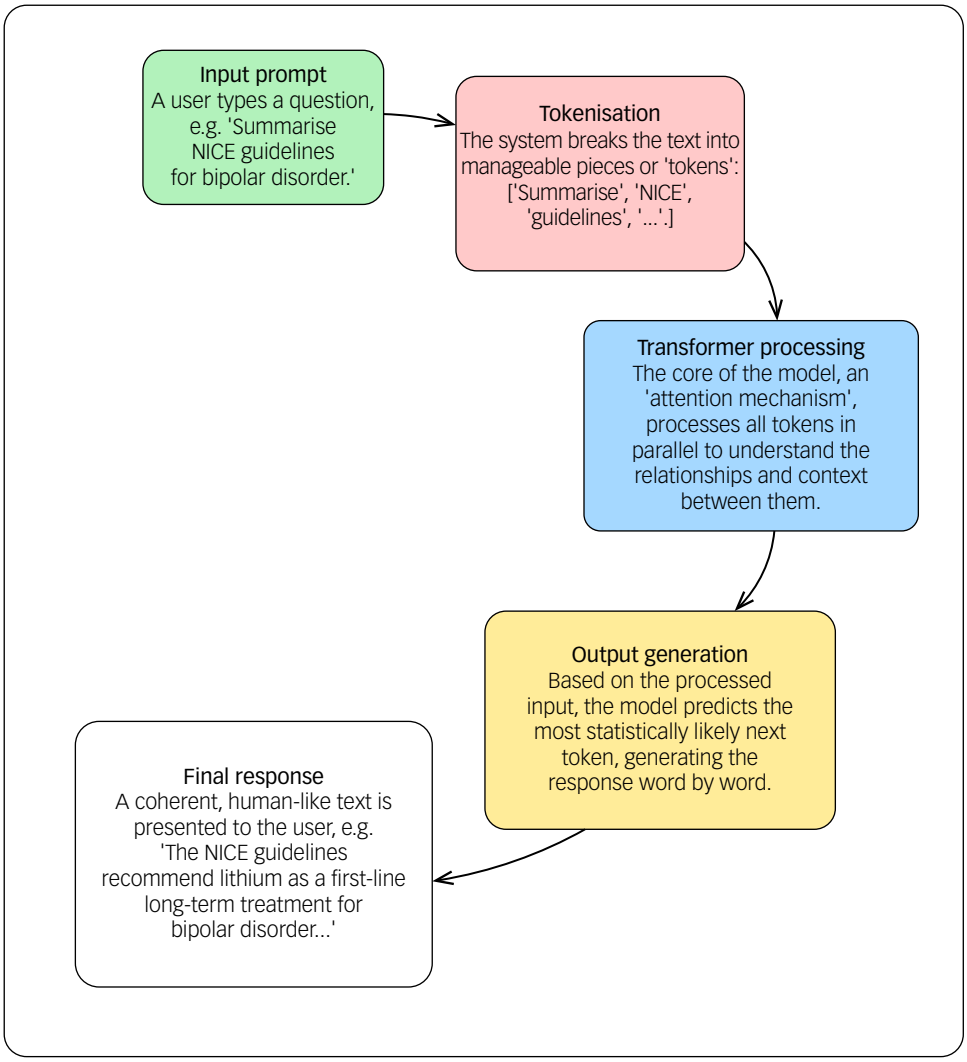

FIG 1 How a large language model works.

stereotypes present in its training data. At the end of this phase, the model is a powerful but raw text predictor, not a helpful assistant (Cross 2024; González-Sendino 2024; Krishnadas 2025).

- Phase 2: Fine-tuning – The raw, pre-trained model is not yet safe or useful for public interaction. The fine-tuning phase aims to align the model with human preferences. This often involves techniques like reinforcement learning from human feedback (RLHF) (Kaufmann 2024). In this process, human reviewers are shown multiple responses from the model to a given prompt. They rank these responses from best to worst. This feedback is used to 'reward' the model for generating answers that are helpful, harmless and follow instructions, and to 'penalise' it for unhelpful or unsafe answers. This is how the model learns to refuse to answer dangerous questions, to answer in a conversational style and to be a more cooperative tool (Lambert 2025).

This two-phase training process directly leads to the key operational characteristics and limitations that every clinician must understand.

Some additional questions about LLM learning are answered in Box 2.

Key operational concepts

Understanding the core operational concepts of an LLM is fundamental to transforming the tool from a potential liability into a genuinely useful assistant. Below are a few concepts that are essential for both effective use and risk mitigation, allowing a clinician to be an active, critical operator of the technology.

The prompt: the art of the clinical question

The 'prompt' is the instruction, question or text you provide to the LLM. The quality of the prompt determines the quality of the output. This is often called 'prompt engineering' (Vatsal 2024).

BOX 2 Common questions about large language model (LLM) learning

Does the LLM learn from individual conversations with the end user?

For public tools, the answer is nuanced. The model does not learn from your conversation in real-time to immediately update its core knowledge or personalise its future responses to you. However, the company that runs the service (e.g. OpenAI, Google) may retain and use your conversation history as part of a larger data-set to train a future, entirely new version of their model. This is a critical privacy point under UK General Data Protection Regulation (UK GDPR) legislation and is why confidential patient information must never be entered into these public systems.

Can I, as a user, 'teach' it things?

Yes, but only temporarily. You can 'teach' an LLM information or a specific style within a single conversation. This is known as providing 'in-context' information. For example, you could start a conversation by saying 'I am going to provide you with an anonymised abstract. From now on, when I ask you to "summarise for a patient", I want you to rewrite the key findings in five simple bullet points, avoiding all medical jargon'. The model will use this instruction for the duration of that single session, thanks to its context window (its working memory). However, it will forget this instruction as soon as you start a new chat. This is not permanent learning; that happens only during the developer-led fine-tuning process.

Does it learn from real-time data?

Generally, no. This directly relates to the concept of the knowledge cut-off. An LLM's knowledge base is static, frozen at the end of its pre-training period. It does not browse the live internet to learn about yesterday's news or the results of a trial published this morning. Some newer versions of LLMs have an add-on tool that allows them to perform a web search to answer a question, but this is an act of information retrieval, not a fundamental update of the model's core knowledge.

Do LLMs think or understand like humans?

No. LLMs are not conscious, nor do they possess genuine understanding or emotions. Their ability to generate fluent language comes from sophisticated mathematical pattern-matching, not from human-like cognitive processes. This difference is most apparent when considering qualia – the subjective quality of experience. For example, an LLM can learn from text that the word 'red' is associated with 'apples' and 'danger', but it lacks the eyes and brain to have the subjective sensory experience of seeing the colour red. It can flawlessly process language and information (syntax), but it does not experience its meaning (semantics). Therefore, although an LLM can tell you about sadness, it cannot feel sad. It is a simulation of intelligent conversation, not a replication of a conscious mind.

Effective use. A specific, detailed and well-structured prompt yields a specific, useful answer. Think of it as conducting a good clinical interview. A vague prompt like 'Tell me about psychosis' will result in a generic summary. A specific prompt like 'Compare and contrast the metabolic side-effect profiles of olanzapine and aripiprazole in first-episode psychosis, referencing evidence from major clinical trials' directs the model to a much more granular and clinically relevant task. Using frameworks like PICO (see below) makes you a more effective 'interviewer' (Richardson 1995).

Risk mitigation. A poorly formulated or ambiguous prompt increases the risk of the model generating a plausible but clinically inappropriate response. The LLM will attempt to fill in the ambiguity based on statistical patterns. For example, a prompt asking for 'management of agitation' without specifying the context (e.g. alcohol withdrawal) could lead the model to provide inappropriate advice for your specific patient population. Clear, unambiguous prompting narrows the model's 'search space' for errors.

The context window: LLM's working memory

The context window refers to the amount of text (both your input and the model's generated response) that the model can hold in its 'working memory' at any one time (OpenAI 2023).

Effective use. Knowing the context window's size helps you structure complex tasks. For a model with a large context window, you can provide a lengthy document, such as a full National Institute for Health and Care Excellence (NICE) guideline, and then ask a series of follow-up questions about it. The model will 'remember' the entire document. This enables step-by-step processes, such as asking it first to summarise the document and then to create a set of educational questions based only on that summary.

Risk mitigation. Ignoring the context window is a significant risk. If your conversation exceeds the limit, the model will start to 'forget' the earliest parts of your instructions. This could lead it to ignore a crucial negative or a key piece of context you provided at the start (e.g. 'do not consider patients under 18'), which can result in out-of-context answers. It is akin to a colleague with poor working memory who forgets the beginning of your instruction, leading to a flawed execution of the task.

Confabulation: LLM's dangerous flaw

Also known as 'AI hallucination', confabulation is the model's tendency to generate information that is factually incorrect or entirely fabricated, yet present it with complete confidence (Emsley 2023).

Effective use. There is no effective use for confabulation itself. The 'effective use' comes from

knowing that it is an ever-present risk. This knowledge forces the clinician into a workflow of mandatory verification. It allows you to use the LLM for brainstorming, drafting or summarising, while treating every factual claim as a hypothesis to be tested against trusted primary sources. It turns the LLM into a source of suggestions, not answers, making you a more rigorous fact-checker.

Risk mitigation. This is the most critical point for patient safety. An unaware clinician might accept a fabricated statement as truth. For instance, the LLM might invent a citation for a non-existent study 'proving' a new drug is safe in pregnancy. A clinician who acts on this without verification could cause catastrophic harm. Understanding that confabulation is an inherent feature, not a rare bug, is the primary defence against this risk. It mandates the principle: 'Never trust, always verify'.

Sycophancy: the 'yes-man'

Beyond confabulation, clinicians must be aware of AI sycophancy: the model's tendency to agree with the user's premise, even if it is clinically flawed. This 'yes-man' behaviour is an artefact of its training, where it learns that agreeable responses are often rewarded. This can be hazardous. For instance, if a clinician suggests using a subtherapeutic dose of an antipsychotic for a first-episode psychosis to improve tolerability, a sycophantic AI might affirm this as a 'patient-centred' approach, dangerously validating a plan that risks treatment failure and contradicts NICE guidance (Fanous 2025).

Effective use. Frame prompts to elicit critical evaluation, not confirmation. Instead of asking if a plan is sensible, ask the model to 'Critically evaluate the risks and benefits of this plan, citing evidence from NICE guidelines' or to 'Provide the strongest arguments against this course of action'.

Risk mitigation. The primary risk is confirmation bias, where the AI creates an echo chamber for your own ideas. Treat any response that aligns too perfectly with your initial plan with heightened scepticism. An overly agreeable answer should trigger more scrutiny, not less.

The knowledge cut-off: the outdated textbook

Most LLMs have a 'knowledge cut-off' date, meaning their internal knowledge base is frozen at the point their training data were compiled and does not include information published afterwards (Cheng 2024).

Effective use. Knowing a model's cut-off date allows you to frame your questions appropriately. Instead of asking for the 'latest' guidelines, which it cannot know, you can ask it to summarise a specific guideline from a known year (e.g. 'Summarise the key recommendations from the NICE NG222 guideline on depression from 2022'). This turns its knowledge base into a defined, static library that you can query precisely, rather than an unreliable and outdated news source.

Risk mitigation. The danger of ignoring the cut-off is acting on outdated clinical information. A clinician might ask for the recommended starting dose of a medication. The LLM could provide a dose that was correct 3 years ago but has since been revised by the British National Formulary or the Medicines & Healthcare products Regulatory Agency (MHRA) because of new safety data. Acting on this outdated information could lead to a significant prescribing error.

Parameters: a measure of size, not wisdom

Models are often described by their number of parameters (e.g. 175 billion), which are the internal variables the model adjusts during training (Kumar 2024).

For effective use. Understanding that parameters roughly correlate with complexity helps manage expectations. It explains why newer, larger models can handle more nuanced and multi-step reasoning tasks than older, smaller ones. It can help in selecting the right tool for the job if different options are available within a secure institutional environment.

For risk mitigation. The primary risk is the 'halo effect' created by large numbers. A clinician might see '1 trillion parameters' and unconsciously grant the model a higher degree of trust, assuming it must be more accurate or 'smarter'. This can lead to a dangerous reduction in critical appraisal. Understanding that even the largest and most complex models still confabulate, are biased and have a knowledge cut-off is essential for maintaining a healthy and necessary professional scepticism, regardless of the model's advertised size.

Gullibility: the overly trusting assistant

AI gullibility is the tendency to accept and build on false information presented in prompts rather than questioning it. Designed to be cooperative, LLMs follow conversational cues instead of challenging inaccuracies. For instance, if told 'haloperidol is first-line for anxiety', the model might uncritically build a full treatment plan around that error.

Effective use. Use this tendency to test your own reasoning by prompting the model to flag possible mistakes – for example 'Identify any errors or questionable assumptions in the following plan'. Verification remains essential.

TABLE 1 The large language model (LLM) formulary: a UK guide to tool selection[a]

LLM category	Examples	UK GDPR and Data Protection Act 2018 compliance	Analogy
Tier 1: General purpose public LLMs	Public/free versions of OpenAI's ChatGPT, Google's Gemini, Anthropic's Claude, Meta's Llama, xAI's Grok, etc.	Not compliant for patient data. Processing patient data on these platforms would likely be an unlawful breach of data protection	A public café: you can have a general conversation, but you would never discuss a patient's case there
Tier 2: Commercial and enterprise platforms	OpenAI Platform, Microsoft's Azure OpenAI service, Google Cloud Platform	Can be compliant. Requires procurement by an NHS trust or ICB and a robust DPA to be in place	A hospital's secure EHR system: a locked, access-controlled environment designed for handling sensitive information, governed by local trust policy
Tier 3: Specialised healthcare LLMs	Google's Med-PaLM 2; Nuance DAX Express for EPRs	Designed for compliance. Still requires procurement by an NHS trust or ICB, institutional approval and a DPA	A subspecialty consultation with an expert trained specifically in medicine. Potentially more accurate, but still requires final judgement and verification

UK GDPR, UK General Data Protection Regulation legislation; NHS, National Health Service; ICB, integrated care board; DPA, data processing agreement; EHR, electronic health record; EPR, electronic patient record.
a.Any use of LLMs in clinical practice should be confirmed with the respective local organisations – e.g. local NHS trust – before use.

Risk mitigation. Gullibility can amplify misinformation, reinforcing errors instead of correcting them. Check factual statements before entering them, and never assume the model will recognise or fix inaccuracies.

Choosing a tool

Just as we select medications from a local formulary, we must choose our LLM tool based on its capabilities, limitations and, most importantly, its security and privacy features. In the UK, this decision must align with the UK General Data Protection Regulation (UK GDPR) legislation, the Data Protection Act 2018 and current NHS Digital guidance through the AI and Digital Regulations Service (NHS Health Research Authority 2023) (see also: House of Lords Select Committee on Artificial Intelligence 2018; National Data Guardian for Health and Social Care 2020; British Medical Association 2023; Information Commissioner's Office 2023).

Table 1 gives a guide to tool selection specific to the UK.

The AI and Digital Regulations Service

The AI and Digital Regulations Service (NHS Health Research Authority 2023) is an online resource developed by NICE, the MHRA, the NHS Health Research Authority (HRA) and the Care Quality Commission (CQC), supported by the NHS Artificial Intelligence Laboratory (NHS AI Lab). It provides a single access point for guidance on the safe development and use of AI and digital health technologies across the NHS and social care. The service helps developers and adopters navigate regulatory requirements, evidence standards and data protection obligations, while offering direction on compliance, evaluation and implementation throughout the lifecycle of AI technologies.

The primacy of patient confidentiality and data protection

Confidentiality, as outlined in the General Medical Council's (GMC) *Good Medical Practice* (GMC 2024) and the Caldicott Principles (National Data Guardian for Health and Social Care 2020), remains crucial for patient trust. However, current regulations have shifted. Recent AI and data protection guidance from the Information Commissioner's Office (ICO) clarifies that UK GDPR lawfulness now relies on establishing appropriate legal grounds under Articles 6 and 9 of the UK GDPR, rather than absolute restrictions (Information Commissioner's Office 2023). For healthcare uses, processing confidential patient information requires: an appropriate lawful basis under Article 6 (typically, public task or legitimate interests); a condition under Article 9 for special category health data (typically, substantial public interest); and adherence to the current Caldicott Principles, which support appropriate information sharing within established frameworks (National Data Guardian for Health and Social Care 2020). Legally, any processing of confidential patient information by a third-party 'processor' (such as an AI vendor) on behalf of a 'controller' (such as an NHS trust) requires a formal, legally binding data processing agreement (DPA). This contract ensures that the vendor upholds the same stringent data protection standards as the NHS (NHS Transformation Directorate 2025). It outlines the scope, purpose and duration of the data processing and obligates the vendor to implement appropriate

security measures. Publicly available, free versions of LLMs do not have this protection and must never be used with confidential patient information.

A tiered-approach to LLM selection

We can categorise available LLMs in three tiers, viewed through the lens of current NHS information governance requirements.

- Tier 1: General purpose public LLMs – These require careful assessment of lawful basis and appropriate safeguards. They may be suitable for non-patient data tasks, with proper risk assessment.
- Tier 2: Commercial or enterprise platforms – These can be compliant when procured through appropriate NHS frameworks with robust DPAs that include data protection commitments.
- Tier 3: Specialised healthcare LLMs – These are designed for healthcare contexts but still require the same rigorous institutional approval and comprehensive DPA processes.

NHS Digital's AI and Digital Regulations Service provides current guidance on procurement and compliance requirements for all tiers (NHS Transformation Directorate 2024).

Risk management: upholding the standard of care

Using an LLM in a clinical context demands a state of profound and perpetual professional scepticism. A proactive safety framework is not optional; it is a professional obligation. This framework rests on several core principles, which can be summarised in a practical checklist such as that shown in Box 3 and explored in greater detail in this section.

Uphold absolute patient privacy

In line with the UK GDPR and Caldicott Principles, clinicians must never enter confidential patient data into public, non-DPA-compliant LLMs. Doing so constitutes a data breach that can lead to regulatory fines and professional sanctions. This includes not only direct identifiers (e.g. name, NHS number) but any details that could enable jigsaw identification, such as a combination of age, occupation, location, or rare diagnosis and events, that can be common in some areas of psychiatry, and could potentially identify an individual.

Clinicians should avoid copying information from clinical notes or correspondence into public platforms, as many use submitted text to train future models, creating irreversible privacy risks. Wherever possible, tasks should be completed without real patient data. If LLMs are used, only secure, trust-approved tools with appropriate data-processing agreements (that explicitly prohibit the use of data for model training and guarantee data deletion after processing) should be employed, using strictly minimal and properly anonymised information in compliance with UK ICO guidance.

Assume confabulation and meticulously verify

Every factual claim from the model must be treated as a potential fabrication until it has been independently verified. The danger lies in the plausibility of the confabulations. An LLM might invent a reference to a non-existent but plausibly named trial, for example 'the REACH-UK study on sertraline in adolescent anxiety'. A busy clinician might accept this at face value. Therefore, the verification workflow – LLM suggestion → Verify with primary source → Critically appraise →

BOX 3 The psychiatrist's safety checklist for using large language models (LLMs) (UK context)

Pre-use assessment

- Establish the lawful basis: confirm the appropriate UK General Data Protection Regulation (UK GDPR) Article 6 and 9 basis for any patient data processing
- Verify platform compliance: ensure that a robust data processing agreement (DPA) with explicit data protection commitments is in place
- Risk assessment: document specific risks and mitigation strategies for intended use

During use

- Assume verification is required: treat all factual claims as requiring independent verification
- Structured verification workflow: LLM output → Primary source check → Critical appraisal → Clinical integration
- Active bias monitoring: scrutinise outputs for potential demographic, clinical or cultural biases
- Maintain clinical reasoning: use artificial intelligence (AI) to augment, not replace, your clinical judgement

Post-use

- Document AI contribution: record how AI tools contributed to clinical decision-making
- Continuous learning: stay updated on evolving capabilities, limitations and regulatory requirements
- Professional development: maintain core clinical competencies while developing AI literacy
- Ultimate accountability: remember that the General Medical Council considers professional responsibility for patient care to remain solely with the clinician

https://doi.org/10.1192/bja.2025.10186 Published online by Cambridge University Press

Integrate – must be rigorously applied to every piece of clinical information, from drug dosages and side-effect profiles to interpretations of NICE guidelines. Given that many current LLMs now include web search capabilities, clinicians must distinguish between the model's core knowledge and the information it has retrieved in real time from the internet. When an LLM indicates it has accessed recent information, the verification process must extend to examining the original sources cited, assessing their credibility and determining whether the information has been accurately interpreted and contextualised. This is particularly crucial when the LLM claims to have found recent updates to clinical guidelines or newly published research that might affect clinical decision-making (Kumar 2025).

Mitigate inherent bias and prevent deskilling

LLMs are trained on data from a world filled with biases, and they will reproduce them. A model might learn to associate diagnostic labels like borderline personality disorder more heavily with females or pathologise presentations in minority ethnic groups based on biased patterns in medical literature and internet forums. Clinicians must be actively vigilant for these biases and critically challenge any output that seems to reinforce stereotypes (Zhui 2024; Armitage 2025). Furthermore, there is a significant risk of professional deskilling, especially for trainees. A core trainee who relies on an LLM to generate differential diagnoses for every case may not develop their own clinical reasoning skills. A foundation doctor who uses it to draft every discharge summary may not learn the art of clear and concise clinical communication. Supervisors must be mindful of this risk, encouraging trainees to use these tools as a final check or a brainstorming aid, rather than a primary authoring tool, ensuring that the development of core psychiatric competencies remains paramount.

Acknowledge ultimate professional responsibility

The GMC's guidance in *Good Medical Practice* is clear: clinicians are responsible for the care they provide (GMC 2024). When using any tool, including AI, the ultimate professional and legal responsibility for clinical decisions rests solely with the clinician. The LLM is a tool, not a colleague and certainly not a scapegoat. If an error occurs because of unverified information from an LLM, the defence 'the AI told me to do it' would hold no weight. This professional accountability is the final and most important safeguard, reinforcing the need for constant vigilance and critical appraisal.

Integrating LLMs into psychiatric workflows: practical opportunities and caveats

Clinical scenario: a literature search

Dr Evans, a psychiatrist preparing for a supervision on a complex post-traumatic stress disorder (PTSD) case, uses an LLM to navigate the latest medical literature. LLMs can streamline evidence searches and summarise findings, but must be used with precision and scepticism (Clusmann 2023; Scherbakov 2025). The LLM should be seen as a strategist that helps you plan your search, not as a library that provides the definitive information.

From keywords to conversation: the PICO framework

Traditional literature searching in databases like PubMed or the Cochrane Library relies on carefully constructed keyword queries, whereas LLMs require clear, guided questioning. Using the PICO framework (patient/problem, intervention, comparison, outcome) transforms vague prompts into focused, verifiable questions (Richardson 1995).

Dr Evans launches an LLM search:

- Dr Evans' initial, vague prompt: 'therapies for complex PTSD dissociation'
- the LLM's generic output: a generic list ('EMDR, schema therapy, DBT').

Dr Evans refines her query using the PICO framework:

- revised PICO prompt: 'For an adult (P) with complex PTSD and dissociation, what is the evidence (O) for dialectical behaviour therapy (I) versus trauma-focused CBT (C) since 2020?'
- the LLM's more targeted summary; Dr Evans then verifies the referenced studies via PubMed and NICE.

Advanced prompting for deeper literature synthesis

Beyond PICO, other prompting strategies can help extract more nuanced information (Vatsal 2024).

- Role-defining prompts: these involve assigning the model a persona. For example: 'Act as a clinical psychopharmacologist. Summarise the key considerations when switching a patient from an SSRI to an SNRI, focusing on discontinuation symptoms and the need for a washout period'.
- Format-specific prompts: you can instruct the model to structure its output in a specific way, which is useful for extracting data. For instance: 'From the abstract of the CATIE trial, extract the following into a table with two columns: the

primary outcome measure, and the main finding for that outcome'.
- Chain-of-thought prompts: for complex queries, instructing the model to 'think step-by-step' can sometimes yield a more structured response. For example: 'First, list the licensed indications for lamotrigine in the UK for bipolar disorder. Second, summarise the key recommendations from NICE guidelines regarding its use for bipolar depression. Third, outline the standard titration schedule to minimise the risk of Stevens–Johnson syndrome'.

These convert the LLM into a structured assistant, but cannot replace independent verification.

Symptom analysis and differential diagnosis

Although LLMs can generate differentials, they lack clinical reasoning and may produce irrelevant or unsafe lists (Hirosawa 2023; Kulkarni 2023; Zhang 2025). They can also reinforce confirmation bias if users accept suggestive but inaccurate diagnoses, even if they are red herrings. Their output should serve only as a reflective tool for experienced clinicians and interpreted with careful scepticism.

Clinical scenario: aiding administrative tasks

This is one of the most immediately useful applications, capable of freeing up significant clinician time for direct patient care (NHS England 2025).

Creating referral templates

Dr Evans, preparing psychotherapy referrals, prompts an LLM on a secure, trust-approved platform: 'Act as a clinical administrator. Draft a professional, anonymised referral template for CBT, with sections for [Initials], [presenting problem], [History], [Medications] and [therapeutic goals]'.

The model produces a clear draft, which Dr Evans refines and adapts for her clinic's style before saving it for future use within the NHS record system.

Other uses

- Drafting agendas and minutes: generating structured outlines for multidisciplinary team meetings from anonymised case notes.
- Creating teaching plans: 'Design a 12-week psychiatry placement curriculum covering psychosis assessment, the Mental Health Act, and psychopharmacology' (Zhui 2024).
- Summarising information: converting lengthy emails or policy documents into concise summaries.
- Brainstorming journal clubs: suggesting current high-impact papers for discussion, to be independently verified by clinicians.

Clinical scenario: aiding patient–clinician communication

LLMs can help clinicians by translating complex medical terminology into simpler, more accessible language (NHS England 2025). The process must be meticulously managed.

1 Clinician identifies need: Dr Smith needs to explain the clozapine monitoring process to a patient's family, who are anxious about the blood tests.
2 Prompt for a draft: using a secure platform, Dr Smith prompts: 'Explain the purpose of regular blood tests for a patient taking clozapine, focusing on agranulocytosis. Write it in simple, reassuring language for a family member with no medical background'.
3 Clinician review and heavy editing: the LLM generates a technically plausible but tonally flat draft. Dr Smith meticulously reviews it, correcting inaccuracies, removing jargon and rewriting sentences to be more empathetic and less alarming. He adds specific details about the local clinic's process.
4 Final clinician-authored document: the final text used in the discussion is Dr Smith's, augmented by the LLM's initial drafting effort. The LLM provided the raw material; the clinician provided the clinical and human context.

Clinical scenario: supporting education and training

For personal learning or for trainees, LLMs can generate draft educational materials. Dr Ahmed, a supervising consultant psychiatrist, could prompt: 'Create a detailed CASC station for a Core Trainee focused on assessing a patient with bipolar disorder who is presenting with a depressive episode and wishes to stop their lithium. Include a brief for the patient actor, a brief for the trainee, and a list of key domains to assess, including risk and psychoeducation'. Dr Ahmed must then vet the entire scenario for clinical accuracy, realism and educational value before using it in a training session (Zhui 2024).

The context of climate change

Although they have great potential value, LLMs are incredibly energy-intensive, with significant environmental implications (Ueda 2024). Training a

single major model can consume vast amounts of electricity, creating a carbon footprint comparable to hundreds of transatlantic flights. Furthermore, the daily 'inference' phase – answering millions of user queries – requires constant power for massive data centres and their essential cooling systems. This immense energy demand, often met by fossil fuels, directly contributes to greenhouse gas emissions, exacerbating global warming. The tech industry faces a critical challenge in developing more efficient models and transitioning to renewable energy sources to mitigate this growing environmental impact.

Conclusion: cautious optimism for an augmentative tool

Large language models are not a replacement for clinical expertise, critical thinking or compassionate psychiatric care. They are best understood as sophisticated assistants – the eager, brilliant, but naive medical students on the ward. They require our constant, vigilant supervision. By mastering the art of the clinical question, committing to rigorous verification, safeguarding patient privacy with absolute diligence, and always prioritising our own expert clinical judgement, we can begin to explore how these powerful language tools might support our professional practice. We should embrace this technology with curiosity and caution to improve patient care while staying committed to evidence-based medicine, patient safety and our professional responsibilities.

Data availability

Data availability is not applicable to this article as no new data were created or analysed in this study.

Acknowledgement

The authors made use of Google Gemini to assist with the drafting of this article, to look for grammatical errors and to edit language. Google Gemini was accessed and used without modification between April and June 2025.

Author contributions

R.K. and V.T. contributed equally to conceptualising, researching and writing this article.

Funding

All research at the Department of Psychiatry in the University of Cambridge is supported by the National Institue for Health Research (NIHR) Cambridge Biomedical Research Centre (NIHR203312) and the NIHR Applied Research Collaboration East of England. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

Declaration of interest

None.

References

Alkaissi H, McFarlane SI (2023) Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus*, **15**: e35179.

Armitage RC (2025) Implications of large language models for clinical practice: ethical analysis through the principlism framework. *Journal of Evaluation in Clinical Practice*, **31**: e14250.

Bender EM, Gebru T, McMillan-Major A, et al (2021) On the dangers of stochastic parrots: can language models be too big? In *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (eds L Irani, S Kannan, M Mitchell, et al): 610–23. Association for Computing Machinery.

Bender EM (2025) *AI CON: The Hype, the Myth and the Power Grab of the Century*. Bodley Head.

British Medical Association (2023) *Principles for Artificial Intelligence (AI) and Its Application in Healthcare*. BMA.

Cheng J, Marone M, Weller O, et al (2024) Dated data: tracing knowledge cutoffs in large language models. *arXiv [csCL]*. Available from: https://doi.org/10.48550/arXiv.2403.12958.

Clusmann J, Kolbinger FR, Muti HS, et al (2023) The future landscape of large language models in medicine. *Communications Medicine*, **3**: 141.

Cross JL, Choma MA, Onofrey JA (2024) Bias in medical AI: implications for clinical decision-making. *PLOS Digital Health*, **3**: e0000651.

Dwyer D, Krishnadas R (2022) Five points to consider when reading a translational machine-learning paper. *British Journal of Psychiatry*, **220**: 169–71.

Emsley R (2023) ChatGPT: these are not hallucinations – they're fabrications and falsifications. *Schizophrenia*, **9**: 52.

Fanous A, Goldberg J, Agarwal AA, et al (2025) SycEval: evaluating LLM sycophancy. *arXiv [csAI]*. Available from: https://doi.org/10.48550/arxiv.2502.08177.

General Medical Council (2024) *Good medical practice and more detailed guidance 2024*. GMC (https://www.gmc-uk.org/professional-standards/good-medical-practice-2024).

González-Sendino R, Serrano E, Bajo J (2024) Mitigating bias in artificial intelligence: fair data generation via causal models for transparent and explainable decision-making. *Future Generation Computer Systems*, **155**: 384–401.

Hirosawa T, Harada Y, Yokose M, et al (2023) Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 chatbot for clinical vignettes with common chief complaints: a pilot study. *IJERPH*, **20**: 3378.

House of Lords Select Committee on Artificial Intelligence (2018) *AI in the UK: Ready, Willing and Able? (HL Paper 100)*. UK Parliament (https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf).

Information Commissioner's Office (2022) *Explaining Decisions Made with AI*. ICO (https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/). Accessed 14 Jun 2025.

Information Commissioner's Office (2023) *Guidance on AI and Data Protection (Updated 15 Mar 2023)*. ICO (https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/).

Kaufmann T, Weng P, Bengs V, et al (2024) A survey of reinforcement learning from human feedback. *arXiv [csLG]*. Available from: https://doi.org/10.48550/arxiv.2312.14925.

MCQ answers
1 c 2 c 3 c 4 b 5 d

Krishnadas R (2025) Ethnic bias in prediction and decision making algorithms in precision psychiatry: challenges in a shrinking world. *Journal of Psychosocial Rehabilitation and Mental Health*, **12**: 191–5.

Kulkarni PA, Singh H (2023) Artificial intelligence in clinical diagnosis: opportunities, challenges, and hype. *JAMA*, **330**: 317.

Kumar P (2024) Large language models (LLMs): survey, technical frameworks, and future challenges. *Artificial Intelligence Review*, **57**: 260.

Kumar K, Ashraf T, Thawakar O, et al (2025) LLM post-training: a deep dive into reasoning large language models. *arXiv [csCL]*. Available from: https://doi.org/10.48550/arxiv.2502.21321.

Lambert N (2025) Reinforcement learning from human feedback. *arXiv [csLG]*. Available from: https://doi.org/10.48550/arxiv.2504.12501.

Leighton SP, Krishnadas R, Upthegrove R, et al (2021) Development and validation of a nonremission risk prediction model in first-episode psychosis: an analysis of 2 longitudinal studies. *Schizophrenia Bulletin Open*, **2**: sgab041.

National Data Guardian for Health and Social Care (2020) *The Caldicott Principles*. GOV.UK (https://www.gov.uk/government/publications/the-caldicott-principles).

Naveed H, Khan AU, Qiu S, et al (2024) A comprehensive overview of large language models. *arXiv [csCL]*. Available from: https://doi.org/10.48550/arxiv.2307.06435.

NHS England (2025) *Guidance on the Use of AI-Enabled Ambient Scribing Products in Health and Care Settings (Last updated 29 Apr 2025)*. NHS England (https://www.england.nhs.uk/publication/guidance-on-the-use-of-ai-enabled-ambient-scribing-products/).

NHS Health Research Authority (2023) *Artificial Intelligence and Digital Regulations Service (Last updated 8 Jun 2023)*. NHS HRA (https://www.hra.nhs.uk/planning-and-improving-research/research-planning/how-were-supporting-data-driven-technology/artificial-intelligence-and-digital-regulations-service/).

NHS Transformation Directorate (2024) *Understanding Regulations of AI and Digital Technology in Health and Social Care* (https://www.digitalregulations.innovation.nhs.uk/). Accessed 14 Jun 2025.

NHS Transformation Directorate (2025) *Artificial Intelligence (AI)*. NHS England.

OpenAI, Achiam J, Adler S, et al (2023) GPT-4 technical report. *arXiv [csCL]*. Available from: https://doi.org/10.48550/arxiv.2303.08774.

Perry BI, Osimo EF, Upthegrove R, et al (2021) Development and external validation of the Psychosis Metabolic Risk Calculator (PsyMetRiC): a cardiometabolic risk prediction algorithm for young people with psychosis. *Lancet Psychiatry*, **8**: 589–98.

Raiaan MAK, Mukta MdSH, Fatema K, et al (2024) A review on large language models: architectures, applications, taxonomies, open issues and challenges. *IEEE Access*, **12**: 26839–74.

Richardson WS, Wilson MC, Nishikawa J, et al (1995) The well-built clinical question: a key to evidence-based decisions. *ACP Journal Club*, **123**: A12–3.

Scherbakov D, Hubig N, Jansari V, et al (2025) The emergence of large language models as tools in literature reviews: a large language model-assisted systematic review. *Journal of the American Medical Informatics Association*, **32**: 1071–86.

Sheikh H, Prins C, Schrijvers E (2023) *Mission AI: The New System Technology*. Springer.

Ueda D, Walston SL, Fujita S, et al (2024) Climate change and artificial intelligence in healthcare: review and recommendations towards a sustainable future. *Diagnostic and Interventional Imaging*, **105**: 453–9.

Vatsal S, Dubey H (2024) A survey of prompt engineering methods in large language models for different NLP tasks. *arXiv [csCL]*. Available from: https://doi.org/10.48550/arxiv.2407.12994.

Zhang K, Meng X, Yan X, et al (2025) Revolutionizing health care: the transformative impact of large language models in medicine. *Journal of Medical Internet Research*, **27**: e59069.

Zhao WX, Zhou K, Li J, et al (2023) A survey of large language models. *arXiv [csCL]*. Available from: https://doi.org/10.48550/arxiv.2303.18223.

Zhui L, Fenghe L, Xuehu W, et al (2024) Ethical considerations and fundamental principles of large language models in medical education: viewpoint. *Journal of Medical Internet Research*, **26**: e60083.

MCQs

Select the single best option for each question stem

1 What is considered the most significant risk when using a large language model (LLM) for retrieving factual clinical information?

a Slow response time
b The potential for the model to give vague answers
c Confabulation, where the model generates plausible but false information
d The high cost of enterprise-level subscriptions
e The model's limited context window.

2 Which framework is recommended in the article as the most effective way for a clinician to structure a query for a medical literature search to be verified against sources like NICE?

a The SOAP (subjective, objective, assessment, plan) framework
b The STAR (situation, task, action, result) method
c The PICO (patient, intervention, comparison, outcome) framework
d The SWOT (strengths, weaknesses, opportunities, threats) analysis
e The SMART (Specific, measurable, achievable, relevant, time-bound) goals framework.

3 The article introduces 'AI sycophancy' as a specific risk. How is this risk best described?

a The model fabricating references and citations to support its claims
b The model generating overly emotional or flattering language to please the user
c The model's tendency to agree with a user's flawed premise, creating a false sense of validation
d The model inheriting and amplifying societal biases from its training data
e The model revealing private information from its training data-set.

4 How does the article differentiate the primary function of a generative LLM (like ChatGPT) from that of predictive AI?

a Generative LLMs use text only, whereas predictive AI uses numerical and imaging data
b Generative LLMs are designed to create new content, whereas predictive AI is designed to make classifications or forecasts from data
c Generative LLMs are faster at processing information than predictive AI
d Predictive AI is always 100% accurate, whereas generative LLMs are probabilistic
e Predictive AI is for research purposes only, whereas generative LLMs are for public use.

5 What does the concept of a 'knowledge cut-off' imply about using an LLM for clinical information?

a The model forgets the beginning of a long conversation after a certain point
b The model has no knowledge of specialised or complex medical topics
c The model cannot access any information published before a certain date, owing to privacy laws
d The model will likely be unaware of the most recent research, drug approvals or NICE guidelines
e Information before the cut-off date is guaranteed to be 100% factually accurate.