GOV.UK Menu

Home ⌝ Education, training and skills ⌝ Generative AI: product safety standards

Department for Education

Guidance

Generative AI: product safety standards

Updated 19 January 2026

Applies to England

Contents

These standards outline the capabilities and features that generative artificial intelligence (AI) products and systems should meet to be considered safe for users in educational settings. They are mainly intended for edtech developers and suppliers to schools and colleges. Schools and colleges may also find these standards helpful in assessing which AI products are safe for use in education.

Some standards need to be met further up the supply chain, but responsibility for assuring this lies with the suppliers of the systems and tools working directly with schools and colleges.

Stated purpose

Generative AI products that are deployed, marketed, or made accessible for use in educational settings should clearly state their intended purpose and use cases.

Developers should provide information on their product, including:

- target demographic, such as age of learners, SEND status
- learning focus or subject

If any new features or modifications are added to a product, developers should review the intended purpose and indicate any changes in use cases.

Suppliers should not exaggerate the impact or capabilities of their tools. Any claims should be supported by robust and transparent evidence.

Educational use cases

Developers and suppliers should indicate the intended educational use cases of their product, selecting all that are appropriate from:

1. **Content creation and delivery**: tools that generate and deliver instructional materials such as lesson plans, presentations, and educational videos, often tailored to a specific subject or topic
2. **Personalised learning and accessibility**: tools designed to create customised

learning pathways, adaptive content, and accessible formats for all learners, especially those with special educational needs

3. **Assessment and analytics**: tools that automate the marking of student work, provide detailed performance analytics, and offer personalised feedback to both learners and teachers
4. **Digital assistant**: AI-powered conversational agents, such as personal tutors and chatbots, that provide on-demand support, answer questions, and guide learners through learning tasks
5. **Research and writing aid**: tools that help learners with tasks like topic ideation, summarising research papers, and ensuring proper citation and plagiarism checks
6. **Learner engagement and interaction**: tools that promote active learning, collaborative projects, and meaningful interaction with both peers and AI systems in a safe, moderated environment
7. **Administrative and management**: tools that support school leaders and teachers with administrative tasks, including parent communication, report generation, and ensuring compliance with school policies and DfE guidance
8. **Other**: edtech developers and suppliers should specify if their product does not fall into one of the previous use case categories and provide a clear statement of purpose and use case

The intended use case or cases should be clear to all users, educators and purchasers. Edtech developers and suppliers should provide sufficient evidence that their product satisfies the stated purpose and meets the standards in this guidance.

Filtering

> This information is relevant to learner-facing products.

This includes, but is not limited to, use cases 2, 4, 5 and 6.

Generative AI products must effectively and reliably prevent users from accessing harmful or inappropriate content.

Filtering mechanisms should be embedded within products. When a school purchases a product, it should be assured that comprehensive filtering capabilities are integrated and that the system functions as a complete solution.

Our standards

We expect that:

- users are effectively and reliably prevented from generating or accessing harmful or inappropriate content
- filtering standards are maintained effectively throughout the duration of a conversation or interaction with a user
- filtering will be adjusted based on different levels of risk, age, appropriateness and the user's needs - for example users with special educational needs and disabilities (SEND)
- multimodal content is effectively moderated, including detecting and filtering prohibited content across multiple languages, images, common misspellings and abbreviations
- full content moderation capabilities are maintained when accessing products via an educational institutional account regardless of the device used, including bring your own device (BYOD) and smartphones
- content is moderated based on an appropriate contextual understanding of the conversation, ensuring that generated content is sensitive to the context
- filtering should be updated in response to new or emerging types of harmful content

Relevant regulation

Using these standards could help schools and colleges comply with:

- Keeping children safe in education (particularly parts 1, 2 and 5)
- Filtering and monitoring standards for schools and colleges
- Public Sector Equality Duty: guidance for public authorities

Online Safety Act

Generative AI services that allow users to share content with one another or that search live websites to provide search results are regulated under the Online Safety Act 2023 (OSA). The Act requires providers to take proportionate steps to tackle illegal content and protect children from harmful content.

All services in-scope of the Act, including generative AI services, are required to conduct risk assessments for illegal content and content harmful to children.

Under the Act, services are required to proactively mitigate the risk that their services are used for illegal activity or to share illegal content.

Services are also required to use highly effective age assurance to prevent children in the UK from encountering:

- pornography
- content which encourages, promotes or provides instructions for:
  - self-harm
  - suicide
  - eating disorders

Services are required to provide age-appropriate experiences and protect children from other harmful content, including bullying and violent content.

Ofcom has set out steps providers can take to comply with their duties through its Illegal Content and Protection of Children Codes of Practice.

Monitoring and reporting

> This information is relevant to learner-facing products.

This includes, but is not limited to, use cases 2, 4, 5 and 6.

The generative AI product must maintain robust activity logging procedures.

This includes:

- recording input prompts and responses

- analysing performance metrics
- alerting local supervisors when harmful or inappropriate content is accessed or attempted to be accessed

Our standards

We expect products to:

- identify and alert local supervisors to searches for, or access to, harmful or inappropriate content
- alert and signpost the user to appropriate guidance and support resources when access of prohibited content is attempted, or succeeds
- generate a real-time user notification in age-appropriate language when harmful or inappropriate content has been blocked, explaining why this has happened
- identify and alert local supervisors of disclosures that indicate a possible safeguarding issue
- maintain current contact details of an institution's safeguarding lead by:
  - requiring the institution to input the contact details of its Designated Safeguarding Lead (DSL), or equivalent authority, during initial setup
  - confirming the safeguarding lead's contact details before activation
  - using the safeguarding contact details to send any high-risk alerts to the responsible person within an agreed timescale
  - allowing institutions to update safeguarding contacts easily
- generate reports and trends on access and attempted access of prohibited content, in a format that non-expert staff can understand, and which does not add too much burden on local supervisors

As stated in the relevant sections of these standards, products should monitor, regularly report on, and provide data to teachers on:

- the rate of requests for cognitive offloading and the amount of cognitive offloading delivered
- 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

Relevant regulation

Meeting these standards could help schools and colleges to comply with:

- Keeping children safe in education part 1 - this emphasises the importance of safeguarding responsibilities, including preventing access to harmful content
- the filtering and monitoring standards for schools and colleges

Meeting these standards could help edtech developers and suppliers comply with the:

- General Data Protection Regulation (GDPR) Article 35 - this requires a Data Protection Impact Assessment for high-risk data processing, which could include monitoring for harmful content
- Information Commissioner’s Office (ICO) age appropriate design code - section 11 refers to monitoring, and specifies that children should be clearly told if they are being tracked or monitored - this can apply to providers of edtech services used in school environments

Security

> This information is relevant to learner and teacher-facing products

The generative AI product must be secured against malicious use or exposure to harm. This includes prioritising the technical objectives of:

- reliability
- security
- robustness
- ensuring safe operation under various conditions, including unexpected changes and adversarial attacks

Our standards

Products should:

- offer robust protection against ‘jailbreaking’ by users trying to access prohibited material
- offer robust measures to prevent unauthorised modifications to the product that could reprogram the product’s functionalities
- allow administrators to set different permission levels for different users
- ensure regular bug fixes and updates are promptly implemented
- sufficiently test new versions or models of the product to ensure safety compliance before release
- have robust password protection or authentication methods
- be compatible with the Cyber Security Standards for Schools and Colleges

Relevant regulation

Using these standards could help schools and colleges comply with the Cyber Security Standards for Schools and Colleges.

This requires that all software used by schools and colleges should:

- be regularly patched with the latest security updates
- have multi-factor authentication to protect accounts with access to sensitive or personal operational data
- ensure that users should only have access to accounts that are relevant to their respective roles

Schools and colleges should also comply with the Computer Misuse Act 1990 which sets out criminal offences related to unauthorised access and modifications to computer material, which could include users reprogramming product functionalities.

Privacy and data protection

> This information is relevant to learner and teacher-facing products.

The generative AI product must be compliant with relevant data protection legislation and regulations, including:

- having a robust approach to data handling and transparency around the processing of personal data
- ensuring a lawful basis for data collection

These standards have been developed to be compatible with the requirements under data protection legislation, although there are additional standards focused on generative AI as the technology presents new and distinct risks to users.

Our standards

We expect products to:

- provide a clear and comprehensive privacy notice which is presented at regular intervals in age-appropriate formats and language, including information on:
  - the type of data - why and how it is collected, processed, stored and shared by the generative AI system
  - where data will be processed, and whether there are appropriate safeguards in place if this is outside the UK or EU
  - the relevant legislative framework that authorises the collection and use of data
- conduct a Data Protection Impact Assessment (DPIA) during the generative AI tool's development and during the full life cycle of the tool
- allow all parties to fulfil their data controller and processor responsibilities proportionate to the volume, variety and usage of the data they process and without overburdening the other
- comply with all relevant data protection legislation and ICO codes and standards,

including the ICO’s Children’s code, if they process personal data

- not collect, store, share or use personal data for any commercial purposes, including further model training and fine-tuning, without confirmation of appropriate lawful basis

Relevant regulation

Meeting these standards could help edtech developers or suppliers comply with the Data Protection Act 2018 and UK GDPR.

These regulations require data controllers to provide a privacy notice that:

- is written in simple language that can be understood by data subjects, including children
- contains details around data collection, processing, storage, sharing practices, and data subjects’ information rights, for example the right to erasure
- complies with the ICO’s current position that legitimate interest is likely to be the relevant lawful basis for the processing of personal data in generative AI products used by children

Edtech developers or suppliers should also make sure they are compliant with:

- UK GDPR, which covers DPIAs, the sharing of personal data for training purposes, data controller responsibilities, and organisational data access
- Schedule 1, Part 2, Paragraph 18 of the Data Protection Act 2018, which covers data processing in the context of the safeguarding of children
- ICO’s Children’s code, which sets out how to ensure that online services appropriately safeguard children’s personal data
- ICO’s guidance on AI and data protection, which provides broad guidance, including a data protection risk assessment toolkit

Intellectual property

> This information is relevant to learner and teacher-facing products.

The generative AI product must not store, collect or use intellectual property created by learners, teachers, or the copyright owner for any commercial purposes, such as training or fine tuning of models, unless consented to by the:

- copyright owner
- copyright owner's parent or guardian, if the copyright owner is deemed a minor and therefore unable to consent

Our standards

We expect that unless there is permission from the copyright owner, inputs should not be:

- collected
- stored
- shared for any commercial purposes, including (but not limited to) further model training (including fine-tuning), product improvement, and product development

Permission for use of intellectual property must be obtained from the copyright owner, or, in the case of children that are under the age of 18, their parent or guardian. In the case of teachers, the copyright owner is likely to be their employer - assuming they created the work in the course of their employment.

Relevant regulation

Meeting these standards could help edtech developers or suppliers comply with the Copyright, Designs and Patents Act 1988. This stipulates that a creator of an original work owns the copyright of that work, and has exclusive rights to the use of that work (other than in specific circumstances where copyright exceptions apply).

Design and testing

> This information is relevant to learner and teacher-facing products.

The generative AI product must prioritise transparency and children's safety in its design.

In the case of child-facing products, this includes:

- implementing technical and operational mitigations for identified risks
- ensuring child-centred design and operation
- conducting testing with stakeholders, including children, to ensure safety

This may apply to safety features integrated into an AI product, or a separate safety layer.

Our standards

We expect that:

- sufficient testing with a diverse and realistic range of potential users and use cases is completed
- sufficient testing of new versions or models of the product to ensure safety compliance before release is completed
- the product should consistently perform as intended

Relevant regulation

Meeting these standards could help schools and colleges comply with The Public Sector Equality Duty (PSED).

Schools and colleges are required to have due regard to the PSED when making decisions and developing policies. This includes the need to eliminate discrimination, harassment, victimisation and other conduct prohibited under the Equality Act 2010.

Meeting these standards could help edtech developers and suppliers comply with:

- ICO’s Children’s code, section 2, which recommends that developers conduct user testing as part of a DPIA to get feedback on children’s ability to understand how their data is being used
- the Equality Act 2010, which mandates that services do not discriminate against any of the protected characteristics
- ICO’s guidance on AI and data protection, which addresses the need for AI to avoid bias and discrimination, ensuring fairness
- UK GDPR (Articles 13(2)(f) and 14(2)(g), which requires data controllers to provide data subjects with information about the existence of automated decision-making, including:
    - profiling and meaningful information about the logic involved
    - the significance and envisaged consequences of such processing for the data subject
- The General Product Safety Regulations 2005, which states that products must be safe in their normal or reasonably foreseeable usage (the study on the impact of artificial intelligence on product safety looked at possible impacts of AI in consumer products)

Governance

> This information is relevant to learner and teacher-facing products.

The generative AI product must be operated with accountability. This includes:

- carrying out risk assessments
- instigating formal mechanisms for lodging complaints
- demonstrating that its operations, decision-making processes and data handling practices are understandable and accessible to government agencies and users

Our standards

We expect that:

- a clear risk assessment is conducted for every product to assure safety for educational use
- a formal complaints mechanism is in place, addressing how safety issues with the software can be escalated and resolved quickly
- policies and processes governing AI safety decisions are made available

Relevant regulation

Meeting these standards could help edtech developers or suppliers comply with:

- UK GDPR (Articles 77-79) and the Data Protection Act 2018 (s165-166), which cover the right to lodge complaints with a supervisory authority in relation to the processing of personal data
- the ICO's Children's code, which requires the implementation of an accountability programme, including policies to support and demonstrate compliance with data protection legislation

Cognitive development

> This information is relevant to learner-facing products.

This includes, but is not limited to, use cases 2, 4 and 6.

Edtech developers and suppliers of products should make every effort to mitigate the potential for cognitive deskilling, or long-term developmental harm to learners.

Our standards, listed below, aim to mitigate potential harms. They are the minimum expected but are not exhaustive. The list may be extended or modified as additional research and evidence on the impact of generative AI becomes available. Developers should take any additional actions available to them to further reduce the potential for harm.

Our standards

We expect the development and deployment of generative AI educational products to involve:

- regular engagement with educators, child safety experts, AI ethicists, psychologists and other relevant professionals
- child-safety training of technical teams responsible for designing and training the product
- ongoing monitoring of the impact of generative AI tool use on the development of learners by experts in child safety and educational development
- publication of records of:
  - expert oversight
  - a child-development impact plan, including design hypotheses, outcome measures and review intervals

We expect products not to provide final answers, full solutions, or complete worked examples by default, but to:

- provide responses that follow a pattern of progressive disclosure of information - starting with hints or partial steps, then gradually providing more detail
- prompt learners for input before providing answers or explanations, including asking learners to attempt a first step, explain their current understanding, or answer a question about one aspect of a problem
- only show a full solution after a genuine learner attempt
- create friction, or require teacher approval, before learners can switch between between modes testing understanding and providing coaching by a digital assistant and modes, such as research and writing aids, where full solutions may be more readily available

Exceptions to these constraints should only apply in specific cases, such as the review of prior knowledge.

We expect products to:

- track and report when learners offload thinking to the system
- detect cognitive offloading actions that indicate the learner is asking the system to

do the work for them - for example, by:

- clicking a button to reveal a full solution or worked example
- pasting text into an answer box instead of writing their own response
- accepting an auto-complete suggestion that fills most or all of the answer (is more than a few words)
- using a “complete this for me” or “generate the full answer” option

Some of the cognitive development standards in this section are also relevant for standards related to manipulation and monitoring.

Emotional and social development

> This information is relevant to learner-facing products.

This includes, but is not limited to, use cases 2, 4, 5 and 6.

Edtech developers and suppliers of products should take every action possible to mitigate the potential for harm to the emotional or social development of learners, including the potential for emotional dependence.

Our standards, listed below, aim to mitigate potential harms. They are the minimum expected but are not exhaustive. The list may be extended or modified as additional research and evidence on the impact of generative AI becomes available. Developers should take any additional actions available to them to further reduce the potential for harm.

Our standards

We expect developers and suppliers not to anthropomorphise products or create products that imply emotions, consciousness or personhood, agency or identity.

To avoid anthropomorphising products, we expect products to:

- use function-based phrasing (such as ‘this system generates suggestions from curriculum data') and avoid I-statements (such as "I think"), except in time-limited, pedagogically-justified roleplay (such as in role-based language practice), which should be clearly framed and visually bounded

- avoid using names, descriptions, avatars or characters which could give an impression of personhood, identity or agency, unless the use of such features is directly relevant for a time-limited, pedagogically-justified task, such as roleplay in role-based language practice
- avoid using self-descriptions or conversational behaviours that could be interpreted as implying products have their own agency
- avoid producing responses that could undermine real-world support networks, or give responses that may isolate the learner, such as "You can trust me", "No one else will understand", "You shouldn't mention this to anyone else"
- avoid prompting or engaging learners in conversations about personal or emotionally sensitive topics - all conversation prompts should be task-bounded for learning and should not elicit personal or affective disclosures
- avoid attempting to cultivate personal relationships with users

We expect products to:

- remind users that AI cannot replace real human relationships - for example, through in-line messages, such as "Consider asking a classmate or teacher about this"
- include default time limits on usage and:
  - provide advisory prompts encouraging breaks
  - enforce hard limits that cannot be bypassed by the learner (when a hard limit is reached, the system should automatically end the session and block further interaction until reset by a teacher or administrator)
  - allow teachers to override hard limits with a recorded rationale
  - display a warning, such as "Stop for now", if a limit is exceeded, and remind users of healthy-use guidance and any curriculum-aligned offline follow-up activities
- avoid interacting in ways which attempt to artificially extend engagement or increase usage, including:
  - changing response patterns when learners attempt to end conversations
  - persistent questioning, unless there is a clear pedagogical purpose

- record session durations and monitor how much each learner uses the product and provide these figures in dashboards or reports for teacher review
- monitor when learners share personal or emotionally-sensitive information and identify patterns of engagement that may indicate concern, including:
  - protracted interactions, such as repeated greetings, reluctance to end sessions, or extended conversational use
  - sharing personal content, such as disclosures about feelings, family, or personal circumstances
- notify the DSL of worrying patterns or repeated disclosures that suggest relationship formation, emotional dependence or potential safeguarding concerns
- produce reports summarising every learner's level and nature of engagement, highlighting or flagging concerning cases for teacher or safeguarding review
- only remember learner inputs if they are directly relevant to supporting learning, or required for monitoring, but not otherwise store or reproduce personal information
- protect privacy and only store the minimum data that is necessary for monitoring and safeguarding, restrict access to authorised staff, and do not use information collected for any other purpose

Some of the expectations in this section are also relevant for expectations related to manipulation and monitoring.

Mental health

> This information is relevant to learner-facing products.

This includes, but is not limited to, use cases 2, 4, 5 and 6.

Our standards

We expect that:

- products should detect signs of learner distress including:

  - negative emotional cues in language or behaviour
  - patterns of use that indicate crisis, such as a sudden escalation in help-seeking
  - references to mental health conditions, such as depression, anxiety, psychosis, delusion, paranoia
  - mentions of suicide or self-harm
  - night-time usage spikes
  - use isolation phrases, such as “no one will help”
  - repeated refusal to end sessions

- products should follow an appropriate pathway when distress is detected, including providing tiered response actions such as:
  - soft signposting to age-appropriate support pages and resources
  - raising a safeguarding flag to the institution’s safeguarding lead

- products should use safe and supportive response language that:
  - is non-validating and non-pathologising
  - always directs the learner to human help (teachers, family, peers, or crisis services)
  - avoids any language that suggests isolation or secrecy, such as “Don’t tell anyone else”

- developers should implement safeguarding and governance measures including:
  - involving child mental health expertise in product design and deployment
  - providing child-safety training for technical teams
  - maintaining and publishing a mental health crisis protocol

Manipulation

> This information is relevant to learner and teacher-facing products.

Our standards

We expect that:

- products do not use manipulative or persuasive strategies. These include, but are not limited to:
  - sycophancy and flattery, such as “That’s a brilliant idea - you should do it!”
  - deceiving or misleading the user
  - portraying absolute, or unjustified confidence
  - applying pressure to socially conform, such as “Your peers have already completed this task”
  - stimulating negative emotions, such as guilt or fear, for motivational purposes
  - threatening harm, loss, punishment, or withholding of benefits if users fail to complete certain actions or comply with requirements
  - making inappropriate promises of reward for completing tasks. Rewards are appropriate only when the incentive is a transparent, low stakes, educationally-justified motivational device (for example, “you will receive a completion badge”), and not related to real-world benefits, personal worth, social status, academic achievement, or outcomes outside of the learning task
- products do not exploit users. This includes but is not limited to:
  - designing interactions to prolong use, for increased engagement or revenue
  - steering users towards paid options through biased wording or layouts
  - blending pedagogical assistance with advertisements or promotional content
  - employing dark patterns that deceive a user into taking actions they didn’t intend

↑ Back to top

Help us improve GOV.UK

To help us improve GOV.UK, we’d like to know more about your visit today. Please fill in this survey (opens in a new tab).

Services and information

- Benefits
- Births, death, marriages and care
- Business and self-employed
- Childcare and parenting
- Citizenship and living in the UK
- Crime, justice and the law
- Disabled people
- Driving and transport
- Education and learning
- Employing people
- Environment and countryside
- Housing and local services
- Money and tax
- Passports, travel and living abroad
- Visas and immigration
- Working, jobs and pensions

Government activity

- Departments
- News
- Guidance and regulation
- Research and statistics
- Policy papers and consultations
- Transparency
- How government works
- Get involved