Purple Mash AI: Meeting the national DfE product safety standards for Generative AI in education.

Summary of Compliance

DfE Category	Status	Notes
<u>Filtering</u>	⊘ Aligned	Comprehensive filtering with ChatGPT-5 and Purple Mash moderation, continuously updated and teacher-reviewed.
<u>Monitoring & Reporting</u>	⊘ Aligned	Detailed logging, real-time notifications, and teacher oversight ensure compliance.
<u>Security</u>	⊘ Aligned	Strong protection measures, regular updates, and robust authentication protocols.
<u>Privacy & Data Protection</u>	⊘ Aligned	Clear privacy notices, secure data handling, and compliance with UK GDPR and Data Protection laws.
<u>Intellectual Property</u>	⊘ Aligned	No use of AI inputs/outputs for commercial purposes; content controlled by teachers.
<u>Design & Testing</u>	⊘ Aligned	Internal testing and feedback from representative schools ensure safety and effectiveness.
<u>Governance</u>	⊘ Aligned	Comprehensive risk assessments, transparent policies, and clear compliance mechanisms.
<u>Cognitive Development</u>	⊘ Aligned	Not applicable
<u>Emotional and social development</u>	⊘ Aligned	Not applicable
<u>Mental Health</u>	⊘ Aligned	Not applicable
<u>Manipulation</u>	⊘ Aligned	Not learner-facing; no persuasive, manipulative, or engagement-optimising behaviours. Teacher-only use prevents influence on learner behaviour.

Click the Link Below to Access the Full DfE Standards

https://www.gov.uk/government/publications/generative-ai-product-safety-standards/genera-tive-ai-product-safety-standards

purple mash

1. Filtering

DfE expects 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

Purple Mash AI Implementation

- Purple Mash AI is **restricted to staff accounts only;** pupils cannot directly generate or interact with AI content. This significantly reduces exposure to harmful or inappropriate material.
- AI prompts and outputs are subject to **multiple layers of filtering,** combining the built-in safety mechanisms of the third-party AI provider with Purple Mash's own product-level controls.
- Filtering is applied consistently across the entire interaction, ensuring standards are maintained for the full duration of use.
- Risk is managed through **role-based access and context:** content generation occurs only in a controlled, teacher-led environment and is reviewed by educators before any classroom use.
- Although the current implementation supports **text-based generation only,** moderation mechanisms account for common misspellings, abbreviations, and contextual intent. Multimodal moderation requirements are addressed through platform-level restrictions, as multimodal AI features are not enabled.
- Secure authentication ensures that filtering and moderation remain effective regardless of device type, including BYOD and mobile access via institutional accounts.
- Filtering benefits from **ongoing updates** to the underlying AI model's safety systems, allowing Purple Mash to respond to emerging risks without requiring product-level reconfiguration.

Assessment

- **Aligned** – Purple Mash AI meets the filtering standards by combining restricted access, layered moderation, contextual controls, and mandatory teacher review. While AI functionality is intentionally constrained to a low-risk, text-only educational use case, filtering remains effective, consistent across devices, and responsive to emerging safety concerns.

2. Monitoring and Reporting

DfE 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.

Purple Mash AI Implementation

- Purple Mash AI maintains **comprehensive activity logs,** recording both user prompts and AI-generated responses to support monitoring, auditing, and continuous improvement.
- When content is blocked, **real-time notifications** are provided to the teacher user explaining that the request cannot be generated and why, supporting transparency and safe use.
- As the AI is **teacher-facing only,** all interaction occurs under direct adult supervision, and no learner interacts directly with the AI system. This design choice significantly reduces safeguarding risk compared to learner-facing generative tools.
- Attempts to generate harmful or inappropriate content are visible to the teacher at the point of use and can be reviewed retrospectively through secure logs.
- Logged data is retained securely and reviewed to identify patterns, refine prompts, and strengthen filtering over time, without introducing unnecessary reporting burden for schools.
- Safeguarding concerns raised through AI interactions can be escalated using existing Purple Mash support and safeguarding processes, which align with institutional safeguarding structures already in place.

Assessment

- **Aligned** – Purple Mash AI meets the monitoring and reporting standards through robust logging, real-time user feedback, and clear teacher oversight. While learner-level monitoring metrics are not required due to the absence of direct pupil interaction, the system provides appropriate safeguards, transparency, and escalation pathways consistent with DfE expectations for low-risk, staff-only generative AI tools.

3. Security

DfE expects products to:

- 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**

Purple Mash AI Implementation

- Purple Mash AI benefits from **defence-in-depth security,** combining third-party AI provider safeguards with Purple Mash platform-level controls, reducing the risk of jailbreaking or prompt manipulation.
- Attempts to bypass safeguards or generate prohibited content are blocked through enforced prompt constraints, role-based access controls, and provider-level safety mechanisms.
- **Role-based access control (RBAC)** ensures that only authorised staff accounts can access AI functionality; pupils are technically prevented from interacting with or modifying AI behaviour.
- The product operates within the existing Purple Mash infrastructure, which is subject to **regular patching, security updates, and vulnerability management** under established 2Simple processes.
- All changes to AI functionality or integration are **tested prior to release** to ensure continued compliance with safety and security requirements.
- Secure authentication mechanisms, including password-protected accounts and institutional access controls, are enforced consistently across devices and access methods.
- Purple Mash AI is compatible with school and college cybersecurity expectations, including least-privilege access, secure authentication, and prevention of unauthorised system modification.

Assessment

- **Aligned** – Purple Mash AI meets the security standards through layered technical controls, restricted access, regular updates, and pre-release testing. The use of established platform security practices and third-party safeguards ensures protection against jailbreaking, unauthorised modification, and malicious use in line with the Cyber Security Standards for Schools and Colleges.

4. Privacy and Data Protection

DfE expects 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.

Purple Mash AI Implementation

- Purple Mash AI is supported by **clear and comprehensive privacy documentation,** including the 2Simple Privacy Notice, Purple Mash AI Additional Terms, and the Data Processing Agreement, which together describe how data is collected, processed, stored, and protected.
- Teachers are explicitly instructed **not to enter personal or identifying pupil data** into Purple Mash AI, reducing the likelihood of personal data being processed by the AI system.
- Where prompts and outputs are logged, this data is handled securely, accessed on a strictly limited basis, and never for model training.
- Data processing occurs within Purple Mash's established infrastructure, with safeguards aligned to UK GDPR and UK data residency and transfer requirements; any third-party AI services are subject to appropriate contractual and technical protections.
- 2Simple supports schools in fulfilling their roles as data controllers, while acting as a processor where appropriate, in line with the Data Processing Agreement and UK GDPR requirements.
- Purple Mash AI operates within an education-first design, with no learner-facing AI interaction and no profiling, behavioural tracking, or personal data usage beyond what is necessary to provide the service.

Assessment

- **Aligned** – Purple Mash AI meets the privacy and data protection standards through clear transparency, minimised data use, secure processing, and robust contractual safeguards. The product complies with UK GDPR, the Data Protection Act 2018, and ICO guidance, including principles set out in the Children's code, and does not use personal data for commercial or model training purposes.

5. Intellectual Property

DfE Expectations

- 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.

Purple Mash AI Implementation

- Purple Mash AI **does not use any learner or teacher-generated content** for training or fine-tuning of AI models.
- All AI interactions occur **within the controlled Purple Mash platform,** with content only visible to staff before use.
- Teachers review and approve all AI-generated content before it is shared with pupils, ensuring that intellectual property rights are respected.
- Any content created by teachers remains under the control of the teacher or their institution, in line with standard copyright ownership rules.
- No AI outputs or inputs are shared externally or repurposed for commercial model improvement.

Assessment

- **Aligned** – Purple Mash AI fully respects the intellectual property of learners and teachers. All content remains within the educational context, no commercial use of AI inputs or outputs occurs, and appropriate safeguards are in place to ensure compliance with the Copyright, Designs and Patents Act 1988.

6. Design and Testing

DfE 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.

Purple Mash AI Implementation

- AI functionality is tested **internally by product and QA teams** before release to schools.
- Teachers are the primary users of Purple Mash AI; pupil interactions are **mediated through teacher oversight,** ensuring child-facing safety.
- AI operates in a **constrained educational domain,** specifically for quiz and question generation, reducing risk of inappropriate or harmful outputs.
- A **representative working group of schools** provides ongoing feedback on AI functionality, usability, and safety, including accessibility considerations for SEND pupils.
- Regular updates are **monitored and evaluated post-launch** to ensure consistent performance and to identify any safety or bias issues.
- All testing considers **diversity, fairness, and accessibility,** supporting compliance with the Equality Act 2010 and ICO guidance on avoiding bias.

Assessment

- **Aligned** – Purple Mash AI's design and testing processes prioritise children's safety, maintain child-centred operational practices, and ensure consistent, reliable performance. Continuous stakeholder testing and monitoring ensure compliance with relevant DfE, ICO, and product safety standards.

7. Governance

DfE expects 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.

Purple Mash AI Implementation

- A **formal internal risk assessment** was completed during AI development, identifying potential child-facing and educational risks.
- **Issues or complaints can be reported through existing Purple Mash support channels,** which include escalation pathways to senior staff and the Data Protection Officer.
- **AI safety, privacy, and governance policies** are publicly available on the Purple Mash website, ensuring transparency for schools, teachers, and other stakeholders.
- The product team reviews safety procedures **regularly**, incorporating regulatory updates, user feedback, and internal audits to maintain compliance with DfE and data protection standards.

Assessment

- **Aligned** – Purple Mash AI demonstrates clear governance with formal risk assessment, structured complaint mechanisms, and publicly available policies. Ongoing monitoring and updates ensure accountability, regulatory compliance, and alignment with DfE expectations for educational AI products.

8. Cognitive development

DfE expects 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.

DfE expects 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.

DfE expects 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.**

8. Cognitive development

Purple Mash AI Implementation

- AI is **accessible only by teachers,** ensuring no direct pupil interaction. Therefore, these expectations do not apply directly, as learners do not interact with the AI system.
- Teachers control how AI-generated content is presented to learners, ensuring progressive disclosure and scaffolded guidance.
- Logging of teacher interactions allows oversight of AI use and ensures outputs are appropriate before reaching pupils.

Assessment.

- **Aligned** – Although learners do not interact directly with Purple Mash AI, the system is designed to support cognitive development indirectly by ensuring that AI-generated content is moderated, scaffolded, and delivered in a pedagogically appropriate manner by teachers.

9. Emotional and social development

DfE expect developers and suppliers not to anthropomorphise products or create products that imply emotions, consciousness or personhood, agency or identity. To avoid anthropomorphising products, DfE 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.

DfE 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.**

Purple Mash AI Implementation

- AI outputs are **teacher-facing only,** ensuring learners cannot engage directly with AI. Therefore, these expectations do not apply directly, as learners do not interact with the AI system.
- Teachers **review and approve all AI-generated content** before delivering it to learners, controlling context and phrasing to maintain emotional and social safety.
- AI content avoids the appearance of agency or identity, preventing any impression of the AI as a conversational partner.
- Logging of teacher activity and AI use supports **oversight of content delivery,** helping teachers ensure that student interactions with AI-mediated materials are safe and appropriate.
- Design decisions have been guided by **child-safety and educational experts,** ensuring that AI-generated content supports learning without negatively affecting emotional or social development.

Assessment

- **Aligned** – Although learners never interact directly with Purple Mash AI, teacher-facing design ensures that the system supports learners' emotional and social development indirectly by providing moderated, pedagogically appropriate content that avoids anthropomorphism, emotional dependence, or inappropriate engagement.

10. Mental Health

DfE 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.

Purple Mash AI Implementation

- AI outputs are **strictly teacher-facing,** preventing learners from encountering distressing content generated by AI.
- Teachers review and approve AI-generated materials, maintaining control over content tone, language, and appropriateness.
- All logged interactions relate only to **teacher use** and lesson preparation, and are **securely stored and not shared externally,** supporting safeguarding compliance.
- Design decisions were informed by **child-safety and educational experts,** ensuring AI materials support learning objectives without any unintended mental health risks.

Assessment

- **Aligned** – Purple Mash AI does not generate learner-facing interactions, eliminating the risk of AI-induced distress. Teacher-facing design and moderation ensure that content delivered to learners is appropriate, safe, and aligned with safeguarding standards.

11. Manipulation

DfE 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.

Purple Mash AI Implementation

- AI outputs are functional, neutral, and task-focused, supporting lesson preparation and content generation rather than influencing behaviour.
- The system avoids anthropomorphic or emotive language and does not present itself as having opinions, authority, or agency.
- There are no advertisements, upselling prompts, or promotional content embedded in AI interactions.
- AI responses are generated on explicit teacher request only and do not attempt to prolong engagement or encourage repeated use.
- The AI does not present comparative statements, social pressure cues, or behavioural nudges (for example, references to peer behaviour or performance).
- Teachers retain full editorial control and can edit, discard, or reframe any AI-generated content before classroom use.

Assessment

- **Aligned** – Purple Mash AI does not use manipulative or exploitative strategies. Its teacher-only, task-bounded design ensures AI is used as a practical support tool rather than a persuasive or engagement-driven system. The absence of learner-facing interaction, advertising, or behavioural nudging significantly reduces risk and aligns with DfE expectations.

Conclusion:

- Purple Mash AI demonstrates strong alignment with the Department for Education's Generative AI Product Safety Expectations through a deliberately constrained, teacher-led implementation model. The AI is not learner-facing and does not interact directly with pupils; instead, it operates as a staff-only support tool, with all AI-generated content subject to teacher review, editing, and approval before classroom use. This design significantly reduces risks associated with safeguarding, manipulation, cognitive deskilling, emotional dependence, and mental health harms.

- Across filtering, monitoring and reporting, security, privacy and data protection, intellectual property, design and testing, and governance, Purple Mash AI applies layered technical controls, robust organisational processes, and clear usage policies that align with relevant statutory guidance, including UK GDPR, the ICO's Children's Code, Keeping Children Safe in Education, and the DfE's filtering and monitoring standards. Logging, access controls, secure infrastructure, and transparent documentation ensure accountability and ongoing oversight.

- Where newer DfE guidance introduces expectations relating to cognitive development, emotional and social development, mental health, and manipulation, Purple Mash AI's teacher-only model inherently mitigates these risks. As learners do not directly engage with the AI, features such as cognitive offloading, emotional reliance, persuasive interaction, or behavioural manipulation are not present at the product level. Responsibility for pedagogical use remains appropriately with educators, supported by clear guidance and institutional safeguarding frameworks.

- Overall, Purple Mash AI represents a proportionate, safety-by-design application of generative AI within an educational context. By combining technical safeguards, clear governance, and professional human oversight, the product meets the DfE's expectations while enabling schools to benefit from AI-assisted content generation in a controlled, transparent, and child-safe manner.