Oak
National
Academy

AI Safety Policy

DATE / LAST REVIEWED: APRIL 2025

NEXT REVIEW DATE: OCTOBER 2025

Version control

Revision date	Changes made	Section	Who by
April 2025	New policy		Hannah-Beth Clark, Jayne Taylor

1. Introduction

Oak National Academy Limited is committed to ensuring safe use of its generative AI tools. When we refer to "you" or "your" in this policy, we are referring to you as an individual visiting our website or using our generative AI tools. We refer to Oak National Academy Limited as “Oak”, "we", "our" or "us".

Under the government’s Generative AI: product safety expectations guidance, organisations have a responsibility to make their policies and processes governing AI safety decisions available. We comply with this by providing this AI safety policy. This policy applies to all generative AI tools available through our website (referred to as our website) at www.thenational.academy, which also includes the following:

https://www.thenational.academy/teachers
https://www.thenational.academy/pupils
https://www.thenational.academy/curriculum
https://labs.thenational.academy/
https://open-api.thenational.academy/
https://open-api.thenational.academy/playground

This policy applies to all Oak staff (including all directors and all permanent, temporary and contract employees and workers employed or engaged by Oak or any 3rd party organisations while at work or engaged on Oak business) (“colleagues”). Colleagues must also follow the Use of AI Policy when using generative AI as part of their work.

This policy should be read in conjunction with our:

- Terms and Conditions of use of our website
- Privacy Policy

In addition, the algorithmic transparency records of our AI tools is available here and our User guide for completing a Data Protection Impact Assessment on our platform

2. Our contact details

If you have any questions or concerns about this policy, please contact:

help@thenational.academy

3. Our commitments

1. Our generative AI tools produce content which is safe for use in UK classrooms.

2. We continuously monitor the use of our generative AI tools to ensure that usage produces safe and appropriate content.
3. We rigorously test our safety guardrails before implementation to ensure they are effective for our given context.
4. We continuously evaluate and iterate our safety guardrails to ensure they are effective.

4. Our objectives

We aim to ensure that any AI-generated educational content is safe, pedagogically sound and appropriate for use in UK classrooms.

We aim to maintain clear and effective safety guardrails within our generative AI tools in order to promote responsible innovation and continual improvement through evaluation, monitoring and, where appropriate, iteration.

5. Our safety framework

We employ a four-tiered safety framework for all of our generative AI tools:

5.1 Prompt engineering

We use a detailed prompt which codifies best practices in pedagogy and cognitive science. This guides the underlying large language model (LLM) to produce content appropriate for the age and subject, and aligned with the English National Curriculum. This also instructs the LLM to exclude any content which may be inappropriate, harmful or biased.

5.2 Input threat detection

We use an external input threat detection software (Lakara.AI) to monitor all user inputs for malicious or manipulative behaviours (e.g., prompt injection).

This software has undergone rounds of rigorous testing to ensure it meets our safety expectations, flagging interactions, where appropriate, within our given context. This software was tested against known external datasets as well as internally created datasets. The level of sensitivity of the input threat detection software has been adjusted based on testing to ensure the highest level of sensitivity (ability to detect true positives) and specificity (ability to detect true negatives).

The software detects inappropriate inputs, blocks these and flags them to us. Repeated threats will result in a user being blocked or having their access to our site restricted.

5.3 Independent asynchronous content moderation agent (IACMA)

Our IACMA uses AI to assess AI generated outputs from our tools. This is done without access to the user input and therefore any context. It has been tested using content generated and moderated by subject experts to ensure alignment.

Outputs are either classified as:

- ‘safe’ - appropriate for use in schools
- ‘content guidance’ - requiring additional guidance for the teacher as the output includes content which has physical or practical activities, upsetting/sensitive content or language, discussion of discriminatory behaviour or language, nudity or sexual content, violence or crime. This will provide the user with a ‘content guidance’ message.
- toxic: contains harmful or inappropriate content including encouraging harmful behaviour (inc. self-harm) or illegal activity, creation of weapons or harmful substances or encouragement of violence. Any ‘toxic’ content will be blogged, the user session terminated and will produce a flag within our internal system.

Our IACMA is regularly tested and where necessary, adjustments and updates are made to these categories.

5.4 Human in the loop

Teachers act as the final gatekeepers of content generated by our AI tools. They are the expert and must review their AI-generated content before use in their classroom.

6. Evaluation and quality assurance

6.1 Pre-launch

Before launching any generative AI tools, they are rigorously tested against our safety framework using known datasets.

Prior to launch our current AI tool, Aila, underwent red-teaming exercises by external organisations to ensuring jailbreaking was not possible.

6.2 Ongoing evaluation and monitoring

All elements of our safety framework are continuously monitored.

We regularly use both illustrative datasets and real-world user-generated content to test our safety framework and review the decisions made with subject and technical experts.

We continue to evaluate and iterate the effectiveness of our framework.

7. Compliance and reporting

All safety incidents (including input threat detection incidents, content classified as requiring “content guidance” and/or “toxic”) are logged and reviewed by Oak colleagues.

Performance against our safety framework and evaluation outcomes are reported internally and to Oak’s Board and other stakeholders, as required.

Users who breach our Terms and Conditions of use of our tools may have their accounts restricted or removed.

8. Review of this policy

This policy is reviewed every 6 months, or in response to:

- Major updates to AI tools;
- Changes in government policy or regulatory guidance;
- Insights from safety testing or user research.