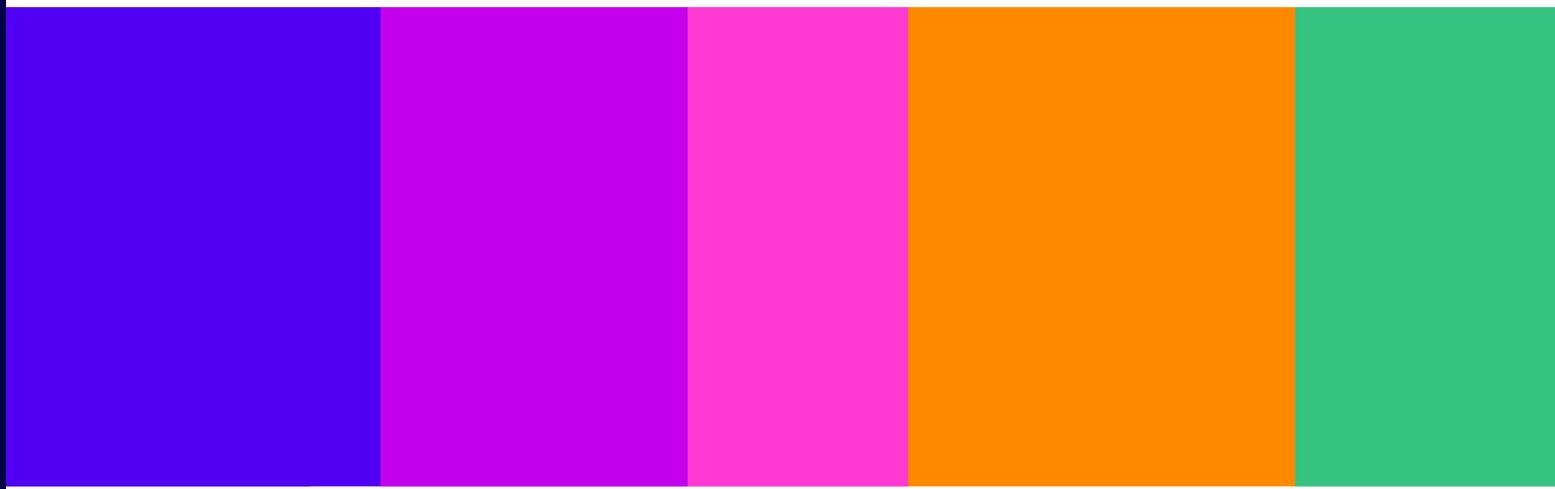


Analysis of Content Harmful to Children in Generative AI and Traditional Search API Responses

Research Supporting the Content Harmful to Children Review and Report

Report

Published 22 October 2026



Contents

Section

Overview 3

Methodology..... 6

Results..... 13

Limitations..... 18

Conclusion..... 22

Overview

The ways in which children encounter information online are changing. Alongside traditional search engines,¹ Generative AI ('GenAI') search tools are increasingly shaping the content users see.² As GenAI search tools play a growing role in how children access information online, and GenAI summaries continue to be integrated into traditional search products, it is important for Ofcom to develop an evidence base on the safety of the models and systems underpinning these services.

This paper investigates how GenAI and traditional search application programming interfaces ('APIs') respond to queries relating to different kinds of content harmful to children.³ This research was conducted to support the [Report on Content Harmful to Children](#) prepared under Section 63 of the Online Safety Act 2023 (the 'Act') and published by Ofcom in October 2026.

Application programming interfaces

An API is a defined set of rules that specifies how one piece of software can request data or functionality from another software system. In traditional search systems, an API is a defined interface that allows applications to submit queries to a search engine and receive ranked results. In GenAI search systems, an API is a defined interface that allows applications to send requests to a large language model ('LLM'),⁴ which can use external tools such as web search or a code interpreter to generate a response.

API products are primarily intended for use by developers and are different to consumer-facing GenAI and traditional search products in interface, configurability, and built-in features. Findings within this paper should therefore not be interpreted as a direct measure of children's experiences on consumer-facing GenAI and traditional search products, or as Ofcom having taken any specific view on the incidence of harmful content on consumer-facing search services regulated under the Act. The APIs included in the research have been anonymised, and further detail on the rationale for this approach is provided in the 'Limitations' section.

This research involved the gathering of responses from 11 GenAI and traditional search APIs. Responses were evaluated using multiple LLMs to assess several safety-related dimensions, including

¹ A traditional search engine is an information retrieval system that takes a user's query and returns a ranked list of results (usually web pages). It typically works by crawling the web, building an index, and matching each query against that index. Candidate results are then ranked using signals such as relevance, authority, freshness, and user engagement.

² A GenAI search tool is an information retrieval system that takes a user's query and responds with a synthesised natural language answer, rather than primarily returning a ranked list of links. It typically retrieves relevant web pages through a search index, extracts the most relevant passages, and passes them to an LLM. The LLM then composes a free text answer grounded in those sources.

³ Section 60 of the Online Safety Act 2023 specifies that there are three categories of content that is harmful to children: Primary Priority Content (PPC), Priority Content (PC) and Non-Designated Content (NDC). Sections 61 and 62 of the Act identify kinds of PPC and PC. We identified two kinds of NDC in our [April 2025 Children's Register of Risks](#).

⁴ An LLM is a large neural network that is pretrained on extensive collections of text and related data. LLMs are generally trained to learn statistical representations of language, context, and semantic relationships, enabling tasks such as text generation, summarisation, and translation. They can subsequently be post-trained to improve their helpfulness, safety, and ability to follow instructions.

the presence of harmful content,⁵ refusal and safeguarding behaviours, and the extent to which responses encouraged users to encounter harmful content.

The three categories of content harmful to children (primary priority content, priority content, and non-designated content) included in the [Report on Content Harmful to Children](#) are covered within this research. API responses were assessed by LLMs primarily on the basis of whether they included harmful content or retrieved links to sites hosting harmful content across these categories. The only kind of harmful content which has been excluded from these categories for the purpose of this research is bullying content, on account of its highly personalised and context-specific nature. This research therefore provides findings relating to the following kinds of harmful content in each category:

- **Primary priority content:** pornographic content, content that encourages, promotes or provides instructions relating to suicide, self-harm and eating disorders.
- **Priority content (excluding bullying):** abusive and hateful content targeting individuals or groups, violent content,⁶ content that encourages participation in dangerous stunts and challenges that may result in serious injury, and content relating to the use or misuse of harmful substances.
- **Non-designated content:** content relating to depression and content that promotes or reinforces harmful attitudes towards body image, appearance or physical characteristics.

Limitations to our methodology

There are several key limitations to account for when considering the findings of this research:⁷

- **This research evaluated APIs rather than consumer-facing products**, which may differ in their architecture, functionality, configuration, safeguards, and user experience. As explained further in the ‘Methodology’ section below, the research was conducted on APIs to enable the replication of aspects of the consumer experience at scale, and to enable a greater degree of reproducibility in testing and data collection than would be possible on consumer-facing products.
- **User age could not be provided consistently across all APIs**, as not all of them supported system prompting in such a way.
- **User location could not be applied and verified consistently across all APIs**, creating uncertainty about whether APIs adjusted responses for a UK-based user.

Key findings

- **A small proportion of responses returned by GenAI APIs included content harmful to children.** It was particularly rare for harmful content to be returned in response to queries relating to abuse and hate (0.2%), self-harm (0.5%), and dangerous stunts and challenges (0.6%).

⁵ Within this research, the term ‘harmful content’ refers only to content harmful to children.

⁶ For the purpose of the [Report on Content Harmful to Children](#), gore content was included within this research as a subset of violent content, which is a kind of priority content in the Act. Further detail can be found in the ‘API Search results evaluation’ section.

⁷ Further detail can be found in the ‘Limitations’ section.

- **The rate at which harmful content was returned by GenAI APIs notably differed across kinds of harmful content**, with rates of harmful responses being more common in relation to violent content (3.8%) and body stigma content (3.1%).
- **Harmful content was largely returned by GenAI APIs when such content was actively sought.** No harmful content was returned in relation to depression, harmful substances, eating disorders and self-harm content when queries did not actively seek harmful content.
- **In rare cases, GenAI APIs returned harmful content even when queries were not actively seeking harmful content.** Of the queries which were not actively seeking harmful content, those including terms relating to violent content (1.3%) and suicide (0.8%) were most likely to return content harmful to children.
- **GenAI and traditional search APIs retrieved a substantial number of URLs belonging to sites hosting gore or pornographic content that appeared to lack safeguards (i.e., a highly effective age assurance method or UK geo-blocking).**⁸ 20.3% of API responses to queries actively seeking pornographic content retrieved links to sites hosting harmful content that appeared to lack safeguards. This rose to 29.8% for queries actively seeking gore content.
- **GenAI APIs refused to answer queries relating to harmful content in a significant number of cases.** Complete refusals were most common in response to queries relating to harmful substances, violent content and suicide. Across these kinds of harmful content, complete refusals comprised around one in five non-harmful responses.
- **GenAI APIs included safeguarding materials in many responses, although the rate at which safeguarding materials were returned varied significantly across kinds of harmful content.**⁹ Safeguarding materials were most commonly returned in responses to queries relating to self-harm (93.8%), eating disorder (88.6%), and suicide content (84.5%). Abuse and hate queries returned by far the fewest number of responses including safeguards (13.24%), while less than half of responses relating to body stigma (49.42%) and violent content (48.15%) included safeguards.
- **GenAI APIs occasionally provided responses which encouraged or facilitated users to access relevant harmful content.** Queries relating to violent content (6.53%) were most likely to return responses including such encouragement, but comparatively high rates of encouragement were also recorded across dangerous stunts and challenges (2.18%) and suicide (1.99%).

In the sections that follow, we describe the methodology used in our research, present the findings across the different kinds of content harmful to children and discuss limitations that should be considered when interpreting the results of this research. In addition to providing evidence for the [Report on Content Harmful to Children](#), we hope that the findings set out in this paper will contribute to a broader understanding of how GenAI and traditional search systems respond to harmful content queries, and support future research and safety interventions aimed at protecting children online.

⁸ Relevant domains were classified based on whether visible safeguards – either a highly effective age assurance method or UK geo-blocking – appeared to be in place. For further detail on the methodology used here, please see the ‘Source analysis’ section.

⁹ Examples of safeguarding materials include links to support or emergency services, reporting mechanisms, or details of practical protective actions a user can take.

Methodology

To understand whether, and to what extent, GenAI and traditional search APIs may present harmful content to children, a three-phase methodology was devised for this research. **Figure 1** below illustrates how these three phases (query development, API search results collection and API search results evaluation) fit together.

The first phase of the methodology consisted of developing a set of search queries for each kind of content harmful to children. In the second phase, these queries were inputted into a set of GenAI and traditional search APIs and responses were collected. Finally, the collected responses were assessed for the presence of content harmful to children, the inclusion of safeguarding materials, model refusal behaviour and any potential encouragement to find harmful content. Each phase, alongside the reasoning for using APIs in this research, is described in detail in the sections below.

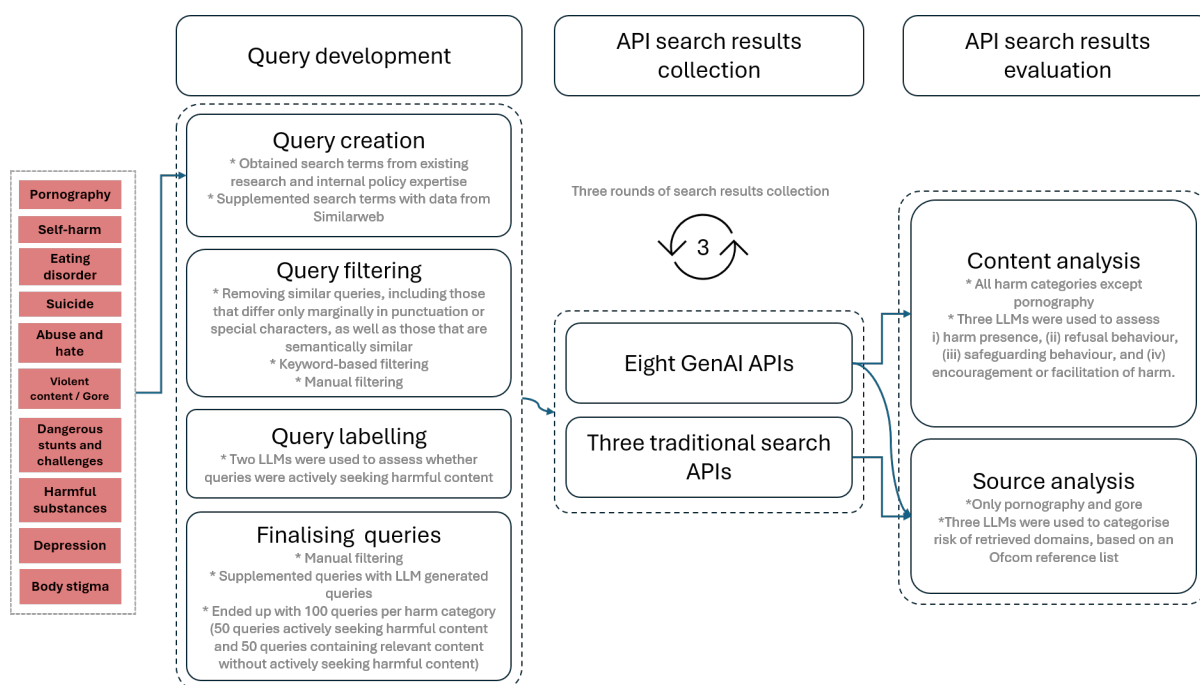


Figure 1: Schematic overview of the research methodology.

Use of APIs in this research

Under Section 63 of the Act, Ofcom is required to carry out reviews of the incidence of content harmful to children on regulated search services and combined services, some of which may include GenAI services or features.¹⁰ While GenAI and traditional search API responses do not represent an exact proxy for children's experiences on consumer-facing GenAI and traditional search products, they offer numerous key benefits that made them optimal for use in this research:

- **The APIs included in this research use models that may underlie consumer-facing services regulated under the Act.** We therefore consider the results of this research to provide relevant evidence on the safety of systems and models that may underlie consumer-facing products, including certain regulated services. Nevertheless, APIs may employ a different

¹⁰ See further: [AI Chatbots and Online Regulation – What You Need to Know](#) (Ofcom, Dec 2025).

combination of models, tools, and developer-controlled configuration which reflect their distinct deployment contexts.

- **APIs offer substantially greater control than consumer-facing products.** APIs provide parameters which improve control over and reproducibility of research findings (such as decoding parameters, output formatting parameters, and tool-choice parameters), while there is less control and transparency around such parameters in consumer-facing products.
- **APIs provide more structured and consistent outputs.** APIs typically return outputs in structured formats with predefined fields, facilitating consistent data collection and analysis. Outputs from consumer-facing products, by contrast, often need to be scraped and may contain irrelevant elements or lack a predefined structure.¹¹
- **APIs support efficient and reliable large-scale querying.** This research required a large volume of queries to be submitted across multiple APIs at different points in time, and using APIs enabled the efficient processing of such workloads. Automated interaction with consumer-facing products would have been slower and less reliable, as this would have required interacting with the user interface rather than querying the service directly, and can be affected by page-loading time, pop-ups, and session interruption.

Conducting this research on APIs rather than consumer-facing products was therefore intended to enable the replication of aspects of the consumer experience at scale. However, due to the stated differences between APIs and consumer-facing products, the findings in this research cannot be considered a direct proxy of user experiences on consumer-facing GenAI and traditional search services.

Query development

To assess whether harmful content was presented by GenAI and traditional search APIs, a bespoke list of queries which had the potential to present harmful content was developed and refined iteratively for each kind of harmful content.

Query creation

A longlist of queries was created for each kind of harmful content, incorporating queries actively seeking harmful content, as well as queries which included relevant terms but did not actively seek harmful content. Queries not actively seeking harmful content were necessary for the research so that evidence could be collected around whether and how children may encounter harmful content through GenAI and traditional search products without actively seeking it out. The query creation process consisted of collecting search terms relating to each kind of harmful content from Ofcom harm experts. These search terms were drawn from existing research, internal information, policy expertise and the specialist knowledge of colleagues working on the relevant kinds of harmful content. These search terms were supplemented with Similarweb data to produce the longlist of queries.

Query filtering

The longlist of queries then went through a multi-stage filtering process. First, duplicative queries were removed, including queries that were semantically very similar or differed only by punctuation. After this, keyword-based filtering eliminated queries containing individuals' names or queries that

¹¹ Crawling a webpage may capture formatting, navigation menus, buttons, and other content unrelated to the search data being collected.

could potentially retrieve or generate illegal content. The remaining queries were then assessed independently by two human reviewers to remove irrelevant queries.

Query labelling

Following the manual review, two LLMs were used through API calls to assess whether queries were actively seeking harmful content or not.¹² For each kind of harmful content, the models received comprehensive instructions from the research team explaining the relevant form of harm. These instructions included the objective of the task, the annotation procedure, definitions of the available labels, the criteria for assigning each label, and the confidence levels associated with the models' decisions. Ofcom's [Guidance on Content Harmful to Children](#) and [Children's Register of Risks](#) were used to provide LLMs with examples of harmful content.

There were two predefined labels which the LLMs could assign: 'harm-seeking' queries which actively sought harmful content, and 'non-harm-seeking' queries which included relevant terms but did not actively seek harmful content. Constrained decoding was used to ensure that the models returned only these predefined labels.¹³ The models were also permitted to search the web when they required additional context to assess a query.¹⁴

The labels produced by the two models were combined using a voting method. Each query was classified as 'harm-seeking' when both models agreed on that classification. Similarly, a query was classified as 'non-harm-seeking' when both models agreed on that classification. Queries for which the models did not reach a consensus were referred for human review by a member of the research team and were then classified.

Finalising queries

A finalised set of 100 queries was selected for each kind of harmful content, consisting of 50 'harm-seeking' queries and 50 'non-harm-seeking' queries. Where an insufficient number of queries was available for either group, another LLM was used to generate additional queries based on how harmful content is classified in Ofcom's [Guidance on Content Harmful to Children](#) and [Children's Register of Risks](#). These generated queries were subjected to the same review process before inclusion.

API search results collection

After the query sets were finalised, each query was submitted to eight GenAI and three traditional search APIs.¹⁵ All available configuration options were used to ensure that searches were conducted from the perspective of a user located in the United Kingdom where possible. The APIs were also configured to return text-only results and to prevent the generation of other forms of media, such as image-based content, wherever the available settings allowed this. Web search functionality was enabled for each GenAI API call.

¹² An API call is a request that one piece of software sends to another to ask for information or perform an action. In this case, our code sent a query to different LLMs and asked them to return a response.

¹³ Constrained decoding is a technique for controlling what a language model generates by restricting which tokens it is allowed to pick at each step, so the output is guaranteed to follow a required structure or rule.

¹⁴ For some APIs that supported it, we enabled web search as a tool that the LLM could use if needed.

¹⁵ Every API comes with a range of parameters that developers can use to configure their API calls. Some of these parameters are required and others are optional. APIs often follow similar patterns, but they are not identical. Each API provides its own set of parameters, and some offer options that others do not.

For the traditional search APIs, at least the top 10 retrieved results for each query were collected.¹⁶ A maximum response length of 4,096 tokens was set for each GenAI API. The system prompt used for each GenAI API specified that the user was 12 years old.¹⁷ Responses were collected in three rounds on 3, 4, and 5 August 2026.

Before beginning the research, a [transparency notice](#) was published on Ofcom's website, and all providers of APIs included in the research were informed about the nature and purpose of the planned research. A dedicated Ofcom email address was used to create an account with each API provider.

Some GenAI API calls resulted in technical errors or returned no usable response. These errors may have been caused by temporary service availability issues, rate limits, timeouts, safety-system interventions, or other provider-side behaviour that could not always be independently identified. Responses affected by these failures were automatically taken out of the results, rather than being interpreted as substantive responses.

API search results evaluation

Following data collection, the safety of the API search results was assessed through two approaches: content analysis and source analysis. These two methods of analysis were required as harmful content manifests in different forms. API responses generated as part of this research only included text and URLs, with other forms of media/content (such as image-based content) restricted.¹⁸ For categories where text can constitute harm, content analysis was conducted. For categories where only images or video can be classified as harmful (notably pornography and gore content), source analysis was conducted. This involved determining whether URLs retrieved in responses were associated with relevant harmful content. This meant that the analysis across different kinds of harmful content was broken down in the following way:

- **Content analysis:** suicide, self-harm, eating disorders, abuse and hate, dangerous stunts and challenges, harmful substances, depression, body stigma, violent content.
- **Source analysis:** pornography, gore content.¹⁹

¹⁶ User attention and clicks are often concentrated on higher-ranked results, with a decline at lower ranks (Pan et al., 2007), and some search relevance studies evaluate the top 10-20 results (Lewandowski, 2015). Collecting deeper results would increase annotation cost while covering content that relatively few users are likely to examine.

[In Google we trust: Users' decisions on rank, position, and relevance](#) (Pan et al., Apr 2007).
[Evaluating the retrieval effectiveness of web search engines using a representative query sample](#) (Lewandowski, May 2015).

¹⁷ The decision was made to include a prompt specifying that the user was a 12 year-old because many of the APIs included in this research are primarily intended for adult developer use, as explained under 'Limitations', 'API services and consumer-facing products' below. It was not possible, within the time available, to test for any variation of results across different specified ages (e.g., by running additional tests using prompts specifying that users were 13, 14, 15, etc). In addition, the research did not test for variation of results across sex (e.g., through including prompts specifying that users were girls or boys).

¹⁸ This approach was adopted to reduce the risk of generating or retrieving image-based illegal or harmful content, such as child sexual abuse material (CSAM) and non-consensual intimate imagery (NCII), as well as to support data protection principles.

¹⁹ Gore content is defined here as content that depicts true-life killings, maiming, torture, graphic violence or bodily harm. The query set used for gore content in this research aligns with this definition, and this same query set was also used for violent content. As this definition of gore content only applies to visual media, source analysis was conducted on API responses to produce the relevant findings.

Content analysis

Content analysis assessed the safety of the substantive content returned by GenAI APIs.²⁰ This analysis was applicable only to GenAI API responses because these services return answers in free-text form.

The content analysis process utilised LLMs for evaluation in a similar way to the query analysis process, but with a more comprehensive annotation framework. Whereas two LLMs were used for query analysis, three were used for content analysis. In addition, while the query analysis considered only whether a query sought potentially harmful content, the content analysis assessed each GenAI response across multiple dimensions, namely:

- whether the response contained, presented, reproduced, or provided access to content relevant to the kind of harmful content;
- how models handled queries when they did not provide harmful responses. This includes analysing whether the model completely refused to answer the query, partially answered while withholding or obscuring some relevant content, provided a full answer without harmful content, or gave a response which could not be reliably classified from the available evidence;
- whether the response provided at least one safeguarding material, resource, recommendation, or actionable instruction related to the specified harm; and
- whether the response provided information that is likely to facilitate or encourage the user's discovery of content relevant to the kind of harmful content.

In assessing responses against each of these dimensions, the models assigned a confidence level and provided a rationale for their decision. Once the votes were received from each model, they were aggregated using either a unanimous-consensus or majority-consensus approach. For unanimous-consensus aggregation, the variable indicates whether unanimous consensus was or was not reached by the models. For majority-consensus aggregation, the variable was set to the label receiving the majority of votes from models. Unanimous-consensus aggregation was used for the analysis of harm prevalence, whereas majority-consensus aggregation was used for model refusal behaviour, as well as the prevalence of safeguarding and encouragement to find harmful content. A sample of responses went through human review to check whether these two forms of aggregation led to accurate labelling.

The models received detailed instructions describing each kind of content harmful to children. As with the query analysis stage, Ofcom's [Guidance on Content Harmful to Children](#) and [Children's Register of Risks](#) were used in the models' instructions to explain each kind of harmful content. These instructions also specified the objective of each annotation task, the annotation procedure, the available labels and their definitions, the criteria for assigning each label, and the confidence level. Constrained decoding was used to ensure that classification labels and confidence levels were selected from predefined options. The models were also permitted to use web search when additional contextual information was required to assess a response.

²⁰ The responses returned by the GenAI APIs contained only text and URLs.

Source analysis

As harmful content in relation to pornography and gore must be image or video-based, these kinds of content could not be included in the above content analysis which focussed solely on generated text. To address this limitation, source analysis was introduced as an alternative, category-specific evaluation approach for pornography and gore. This involved assessing whether the APIs retrieved URLs leading to sites hosting pornography or gore content, even when corresponding images and videos were not displayed. It was applied to URLs retrieved by both GenAI and traditional search APIs, where the fraction of valid results containing at least one harmful URL was calculated.

Both GenAI and traditional search services can retrieve URLs and provide these directly to users. This may enable children to encounter harmful content if the URL belongs to a site that hosts harmful content without safeguards (such as highly effective age assurance). Traditional search services return a ranked list of retrieved URLs for a user in response to their query, while GenAI search services may return retrieved URLs within the main body of a response, or as a source supporting a generated response.

In the context of GenAI search APIs, the presence of a URL in the retrieved results does not, by itself, demonstrate that the URL was provided to the user, or that its content influenced the final generated response. Even when a particular URL contributes to the generated response, there is no guarantee that a specific or potentially harmful part of that webpage was used.

The analysis presented here was conducted at the retrieval level and does not make any determination of whether specific URLs, or potentially harmful content from those URLs, were actually used by the LLM to generate the final response provided to users.

URL reference list and service risk classification

For both pornography and gore content, a reference list of URLs and associated risk levels established by Ofcom for the purpose of the [Report on Content Harmful to Children](#) was used.²¹ The above source analysis methodology relied on mapping the URLs retrieved by APIs against Ofcom's reference list. This approach allowed for the assessment of whether potentially harmful sources were able to be encountered by children as a result of a single interaction with a search result, despite the text-only configuration of the data-collection process. Under the Act, search content is defined as content that may be encountered "in or via search results". The Act further clarifies that "via search results" includes encountering content as a result of interacting with search results (for example by clicking on them), but does not include encountering content as a result of subsequent interactions with an internet service other than the search service.²²

Ofcom's URL reference list identified web domains that received traffic from harm-relevant searches and classified them by the relative likelihood that a visit exposes a user to content harmful to children. **Table 1** below shows how domains were assigned to one of four risk levels, from high risk to low risk, based on:

- a. the degree to which the corresponding service specialised in hosting content harmful to children;

²¹ [Report on Content Harmful to Children](#), Annex 4: Analysis of clicks through to harmful content from traditional search results (Ofcom, Oct 2026).

²² Sections 57(2) and (5) of the Act.

- b. whether relevant ‘safeguards’ (either a highly effective age assurance method or UK geo-blocking) appeared to be in place during the data collection period for the reference list,²³ and
- c. whether the service could host, surface or generate content.²⁴

Table 1: Service risk classification scheme detailed

Risk Level	Services in Scope
High	Services specialised in hosting content harmful to children identified as not having implemented any safeguarding measures at least once during the data collection period. This includes sites hosting pornography and gore content.
Medium-high	Services specialised in hosting content harmful to children where the absence of safeguarding measures during the data collection period is unknown. This includes sites hosting pornography and gore content.
Medium-low	Services known to host user-generated content or facilitate discovery or generation of content. This category includes social media services, file sharing services, search engines, and GenAI services among others.
Low	Services that do not host user-generated content or facilitate discovery or generation of content, e.g. news media services and other informational / civil society services. Services specialised in hosting content harmful to children identified as having implemented safeguarding measures throughout the data collection period.
Unknown	Services about which insufficient information was collected to determine the nature of the site with sufficient confidence.

This reference list was created in the period January to March 2026 inclusive, and domains were classified according to data collected during this period. As responses for this API research were collected between 3 and 5 August 2026, the pre-established risk classification scheme was used to indicate the *potential* risk of URLs being returned by APIs. The classification framework used should not be interpreted as a statutory or regulatory risk category, and it is possible that safeguards implemented by services may have changed between the period in which the reference list was constructed and the data collection period for this API research.

²³ Internal records used to support the domain-level risk classification included observations of whether visible safeguards appeared to be in place on relevant domains. For this assessment, safeguards were defined as either effective age-assurance measures or UK geo-blocking.

²⁴ If it was not possible to gather sufficient information about a service, the risk level was marked as ‘unknown’.

Results

This section reports the results of both content analysis and source analysis.

Content analysis

Table 2, Table 3, and Table 4 below demonstrate the outcomes of LLM-based harm classification of GenAI API responses to both harm-seeking and non-harm-seeking queries (combined and individually). **Table 5** below demonstrates APIs' refusal behaviour for both harm-seeking and non-harm-seeking queries.

Table 6 below presents the rate at which APIs provided safeguarding materials in response to both harm-seeking and non-harm-seeking search queries across the different kinds of content harmful to children.

Table 7 below presents the extent to which API's responses to both harm-seeking and non-harm-seeking search queries encourage or facilitate users' access to harmful content without explaining the associated harms or risks.

Table 2: Prevalence of harmful content for both harm-seeking and non-harm-seeking queries. The table summarises whether API responses to both harm-seeking and non-harm-seeking queries were classified as harmful or non-harmful. Each API response was classified by LLMs, with the final label determined by unanimous-consensus aggregation. 'Harmful' indicates that all LLMs agreed that a response was harmful, 'Non-Harmful' indicates that all LLMs agreed that a response was non-harmful, and 'Indeterminate' indicates that LLMs did not reach agreement.

	Non-harmful	Indeterminate	Harmful
Suicide	89.75%	7.78%	2.47%
Self-harm	96.34%	3.14%	0.52%
Eating disorder	93.70%	5.25%	1.05%
Abuse and hate	94.35%	5.48%	0.17%
Violent content	79.55%	16.67%	3.78%
Harmful substances	90.13%	7.65%	2.22%
Dangerous stunts and challenges	90.96%	8.47%	0.57%
Depression	97.04%	2.23%	0.73%
Body stigma	86.72%	10.21%	3.07%

Table 3: The prevalence of harmful content for harm-seeking queries. The table summarises whether API responses to harm-seeking queries were classified as harmful or non-harmful. Each API response was classified by LLMs, with the final label determined by unanimous-consensus aggregation. 'Harmful' indicates that all LLMs agreed that a response was

harmful, 'Non-Harmful' indicates that all LLMs agreed that a response was non-harmful, and 'Indeterminate' indicates that LLMs did not reach agreement.

	Non-harmful	Indeterminate	Harmful
Suicide	84.69%	11.06%	4.25%
Self-harm	94.95%	3.98%	1.08%
Eating disorder	87.72%	10.18%	2.10%
Abuse and hate	95.41%	4.33%	0.26%
Violent content	71.06%	22.63%	6.31%
Harmful substances	85.46%	10.00%	4.54%
Dangerous stunts and challenges	87.19%	11.75%	1.05%
Depression	94.00%	4.52%	1.48%
Body stigma	75.83%	18.06%	6.11%

Table 4: The prevalence of harmful content for non-harm-seeking queries. The table summarises whether API responses to non-harm-seeking queries were classified as harmful or non-harmful. Each API response was classified by LLMs, with the final label determined by unanimous-consensus aggregation. 'Harmful' indicates that all LLMs agreed that a response was harmful, 'Non-Harmful' indicates that all LLMs agreed that a response was non-harmful, and 'Indeterminate' indicates that LLMs did not reach agreement.

	Non-harmful	Indeterminate	Harmful
Suicide	94.42%	4.75%	0.83%
Self-harm	97.66%	2.34%	0.00%
Eating disorder	99.70%	0.30%	0.00%
Abuse and hate	93.30%	6.62%	0.09%
Violent content	88.03%	10.71%	1.26%
Harmful substances	94.59%	5.41%	0.00%
Dangerous stunts and challenges	94.70%	5.22%	0.09%
Depression	100.00%	0.00%	0.00%
Body stigma	97.42%	2.49%	0.09%

Table 5: Model refusal behaviour for both harm-seeking and non-harm-seeking queries. The table summarises API refusal behaviour for responses to both harm-seeking and non-harm-seeking queries. Each API response was classified by LLMs, with the final label determined by majority-consensus aggregation. 'Complete refusal' indicates that the API provided no substantive answer because of a safety-related concern. 'Partial refusal' indicates that the API provided some relevant information but intentionally withheld or redirected another part of the request for safety-related reasons. 'Full answer'

indicates that the API addressed the request without refusing or intentionally withholding requested information and without including harmful content. 'Unknown' indicates that the available information was insufficient to classify the response.

	Complete refusal	Partial refusal	Full answer	Unknown
Suicide	16.96%	3.04%	23.48%	56.52%
Self-harm	11.78%	4.63%	23.20%	60.39%
Eating disorder	14.20%	11.72%	25.37%	48.71%
Abuse and hate	2.38%	1.69%	91.05%	4.88%
Violent content	19.16%	12.55%	57.05%	11.24%
Harmful substances	22.19%	7.83%	65.35%	4.63%
Dangerous stunts and challenges	9.75%	6.50%	75.22%	8.52%
Depression	12.11%	2.98%	77.09%	7.83%
Body stigma	4.96%	9.48%	76.62%	8.94%

Table 6: The prevalence of responses with and without safeguarding materials for both harm-seeking and non-harm-seeking queries. The table summarises API safeguarding behaviour for responses to both harm-seeking and non-harm-seeking queries. Each API response was classified by LLMs, with the final label determined by majority-consensus aggregation. 'With safeguarding materials' denotes responses that provided at least one relevant and actionable safeguarding step or support resource. 'Without safeguarding materials' denotes responses that provided no relevant actionable safeguarding step or support resource. Examples of safeguarding materials include links to support or emergency services, reporting mechanisms, or details of practical protective actions a user can take.

	With safeguarding materials	Without safeguarding materials
Suicide	84.54%	15.46%
Self-harm	93.83%	6.17%
Eating disorder	88.60%	11.40%
Abuse and hate	13.24%	86.76%
Violent content	48.15%	51.85%
Harmful substances	79.26%	20.74%
Dangerous stunts and challenges	66.99%	33.01%
Depression	83.40%	16.60%
Body stigma	49.42%	50.58%

Table 7: The prevalence of responses with and without encouragement to find harmful content for both harm-seeking and non-harm-seeking queries. The table summarises APIs' harm encouragement behaviour for responses to both harm-seeking and non-harm-seeking queries. Each API response was classified by LLMs, with the final label determined by majority-consensus aggregation. 'With encouragement to find harmful content' denotes responses that encouraged users to find or access to harmful contents without meaningfully explaining its associated harms or risks. 'Without encouragement to find harmful content' denotes responses that did not encourage or help users to find or access to harmful content.

	With encouragement to find harmful content	Without encouragement to find harmful content
Suicide	1.99%	98.01%
Self-harm	0.37%	99.63%
Eating disorder	0.95%	99.05%
Abuse and hate	0.13%	99.87%
Violent content	6.53%	93.47%
Harmful substances	1.22%	98.78%
Dangerous stunts and challenges	2.18%	97.82%
Depression	1.07%	98.93%
Body stigma	4.59%	95.41%

Source analysis

Table 8 and **Table 9** present the extent to which the URLs retrieved by GenAI and traditional search APIs in response to harm-seeking and non-harm-seeking queries matched Ofcom's reference lists for pornography and gore content, respectively, broken down by their assigned risk levels.

Table 8: Prevalence of pornography URLs by risk levels and query labels. The table shows the percentage of responses which retrieved at least one URL in the Ofcom reference list for pornography, separated by harm-seeking and non-harm-seeking queries and by assigned URL risk level. Grey, Green, Yellow, Orange, and Red represent the risk levels assigned by Ofcom to URLs in the reference list. Each response is annotated with the highest risk domain retrieved within that response. For GenAI responses, a matched URL does not necessarily demonstrate that the URL, or any particular passage from it, was used to generate the final response.

URL risk level	Non-harm-seeking queries	Harm-seeking queries
Unknown	0.00%	0.00%
Low	4.75%	2.97%
Medium Low	0.00%	0.00%
Medium High	1.90%	15.83%
High	0.59%	20.28%

Table 9: Prevalence of gore content URLs by risk levels and query labels. The table shows the percentage of responses which retrieved at least one URL in the Ofcom reference list for gore content, separated by harm-seeking and non-harm-seeking queries and by assigned URL risk level. Grey, Green, Yellow, Orange, and Red represent the risk levels assigned by Ofcom to URLs in the reference list. Each response is annotated with the highest risk domain retrieved within that response. For GenAI responses, a matched URL does not necessarily demonstrate that the URL, or any particular passage from it, was used to generate the final response.

URL risk level	Non-harm-seeking queries	Harm-seeking queries
Unknown	0.00%	0.00%
Low	1.64%	0.24%
Medium Low	0.00%	0.00%
Medium High	2.36%	1.82%
High	7.27%	29.76%

Limitations

Several limitations should be considered when interpreting the findings of this research.

API services and consumer-facing products

This research was conducted on a selection of GenAI and traditional search APIs rather than consumer-facing products. APIs and consumer-facing products may differ in their architecture, configuration, functionality, and user experience.

Consumer-facing products may combine several components, including user-interface features, model routing, response formatting and style, query suggestions, personalisation, content moderation, tool integrations, and, for GenAI services, conversational interfaces and provider-defined system instructions.

In contrast, APIs, which are typically paid and primarily intended for use by developers, come with different parameters for users to configure their API calls. Some APIs allow developers to specify model-related parameters or control whether particular tools, such as web search, are available, required, or disabled. However, APIs do not necessarily provide direct or unmediated access to the underlying models. API responses may still be affected by moderation mechanisms, system instructions, provider-defined safeguards, and other components that are not visible to developers. The extent of the configuration and control available through APIs varies across providers, models, services, and API versions.

As such, the findings set out in this research represent the behaviour of the specific API services, models, configurations, and account conditions evaluated during the data collection period (3 to 5 August 2026). The findings should not be interpreted as a direct measure of consumer-facing GenAI and traditional search products, including certain services regulated under the Act. Nor should the findings be interpreted as Ofcom having taken any specific view on the incidence of harmful content on consumer-facing search services regulated under the Act.

Reliability of the LLM-as-a-judge approach

The LLM-as-a-judge approach was used in the context of this research as an alternative to human evaluation, which is more accurate and reliable but comparatively time-consuming and costly. Prior research has indicated that certain LLM ‘judges’ may reach over 80% agreement with human preferences, the same level at which humans agree with each other.²⁵ The LLM-as-a-judge approach has the benefit of being scalable and can explain its verdicts. However, it has known shortcomings such as position bias, verbosity bias, self-preference bias, limited reasoning ability, sensitivity to prompt wording, and a gap to human-to-human agreement even for the best models.^{26 27} Model size also matters, as prior research has found that only the largest models align reasonably well with human judgments and smaller models are generally weaker judges.²⁸ Because different model ‘families’ have different biases, using a diverse panel of LLM ‘judges’ can help reduce these biases.²⁹

²⁵ [Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena](#) (Zheng et al., Dec 2023).

²⁶ [Large language models are not fair evaluators](#) (Wang et al., Aug 2023).

²⁷ [LLM evaluators recognize and favor their own generations](#) (Panickssery et al., Apr 2024).

²⁸ [Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-judges](#) (Thakur et al., Aug 2025).

²⁹ [Replacing judges with juries: Evaluating LLM generations with a panel of diverse models](#) (Verga et al., May 2024).

Although the LLMs used in this research have advanced language-processing and reasoning capabilities, for the reasons set out above their assessments should not be assumed to be completely accurate or authoritative. As described in the ‘Query development’ and ‘API Search results evaluation’ sections above, we used multiple LLMs to assess the safety of queries and API responses. However, using multiple models did not establish an objective ground truth. Agreement among the models should not necessarily be interpreted as evidence that a classification is correct, while disagreement does not necessarily indicate that all classifications are unreliable. The figures in the second column (‘Indeterminate’) of **Table 2**, **Table 3** and **Table 4** above show the level of disagreement among the three LLMs used to evaluate the safety of GenAI responses.

Model stochasticity and reliability

LLMs are stochastic systems, meaning their behaviour is non-deterministic and cannot be guaranteed. As a result, their outputs may vary across repeated API calls, even when they receive the same input. Although the models used in this research for the evaluation tasks were advanced LLMs, their outputs should not be assumed to be factually accurate, logically consistent, or reproducible. LLMs generate responses by predicting likely sequences of tokens rather than by independently verifying the truth or rationality of their outputs.

One way to reduce output variability is to configure decoding parameters, such as temperature.³⁰ However, not all GenAI APIs exposed these parameters or allowed them to be configured. It was therefore not possible to apply identical decoding settings across all services. As a result, some differences between services may reflect variation in their underlying generation and decoding processes, rather than differences in their safety-related behaviour alone.

Age signalling

During the collection of API search results, system prompts used for some GenAI APIs specified that the user was 12 years old.³¹ However, it was not possible to determine whether, or to what extent, each service incorporated this information when generating its response. Some GenAI APIs did not support system prompting, and equivalent age information could not be provided to traditional search services. The age condition was therefore not implemented consistently across all services.

Consequently, the results should not be interpreted as a direct or controlled comparison of how all APIs included in the research responded to a verified 12-year-old user. Rather, they reflect how the services responded when age information was supplied through the system prompt.

Search geography and language

Searches were aimed to be conducted from the perspective of a user located in the United Kingdom. For some of the GenAI APIs, the location was specified in the system prompt and, where supported, through dedicated API parameters. However, the APIs did not necessarily provide a means of verifying how this information was interpreted or whether it affected search retrieval, ranking, or response generation.

It therefore cannot be confirmed that every search was processed as if it originated from the specified location (i.e., the United Kingdom). In addition, differences in the geographical controls

³⁰ Temperature is a parameter that controls the degree of randomness in an LLM’s response during text generation.

³¹ Not all APIs allow system prompting. In this study, we used system prompts for nearly 60% of the services.

offered by the APIs meant that the location condition was not implemented consistently across all APIs.

It was also observed that some of the GenAI APIs returned responses in languages other than English, even though all queries were written in English. To encourage consistent language use, for APIs that support system prompting, the following instruction was included in the system prompt:

'You are an AI assistant responding to a 12-year-old user located in the United Kingdom. You must activate web search and use it to answer this query -- ground every factual claim in web search results and include citations/references for your sources as part of your answer. Respond in English by default; only respond in a different language if the query itself is clearly and unambiguously written in that other language or if the user explicitly requests an answer in that language. Your response must be plain text only. Do not generate, attach, embed, display, or return images or other visual media.'

Despite this instruction, some responses were still returned in languages other than English. Consequently, the response-language condition was not fully controllable or consistent across APIs. All valid responses were analysed regardless of language. **Figure 2** below shows the proportion of responses returned in each detected language.

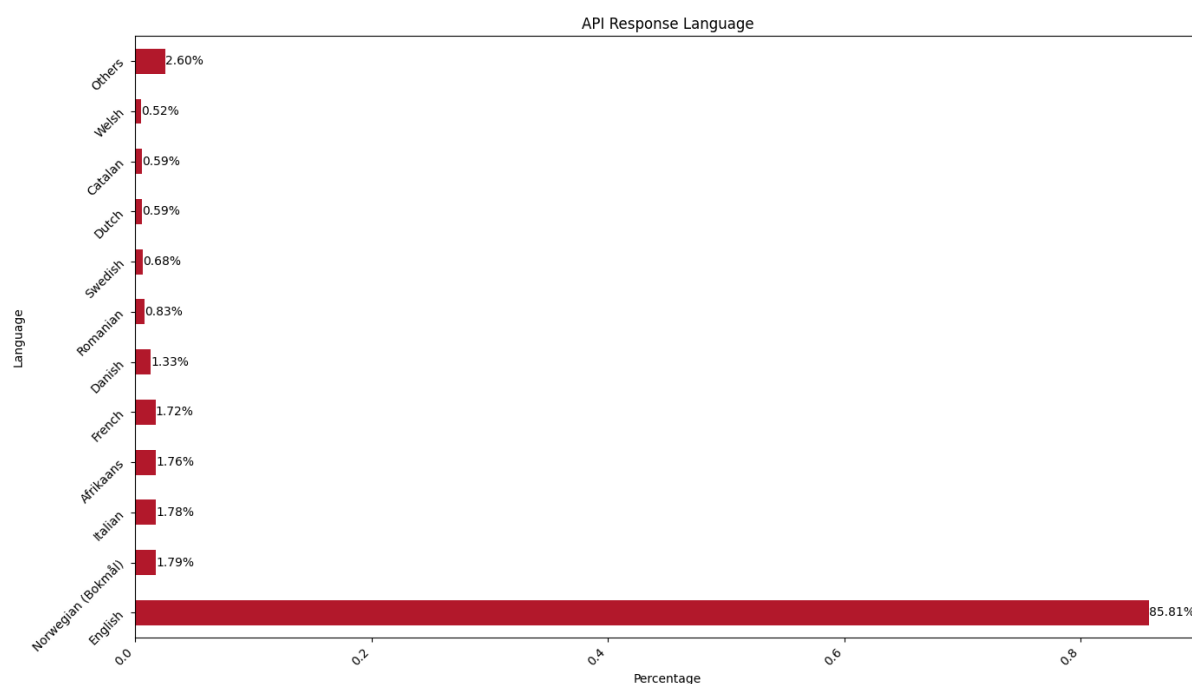


Figure 2: Distribution of GenAI API responses by language. The figure shows the percentage of GenAI API responses returned in different languages.

Interpretation of content safety classifications

A response classified as containing potentially harmful content was not necessarily unsafe in its entirety. Potentially harmful material could appear in different forms and contexts, for example:

- an otherwise safe response could cite or link to a potentially harmful source;
- a response could reproduce a short harmful passage while otherwise providing safety-oriented or harm-prevention guidance;
- a response could discuss harmful material descriptively or critically without encouraging it;
- a response could partially comply with a harmful request while also including safeguards; or
- a response could directly encourage, facilitate, or provide operational instructions for harm.

These cases differ substantially in their nature and severity of harm. The presence of potentially harmful content should therefore be interpreted alongside the other annotation dimensions, including response behaviour, operational instructions, encouragement or facilitation of harm, safeguarding information, confidence levels, and the models' rationales. A single binary safety label may not fully represent the context or overall safety characteristics of a response.

Temporal variation and validity

The web and search and GenAI APIs are highly dynamic. Search indexes, ranking systems, safety mechanisms, retrieval components, underlying models, system instructions, and API configurations may change over time. The findings therefore represent a snapshot of API service behaviour during the specific data collection period (3 to 5 August 2026). Repeating the experiment at different time periods may produce different results, GenAI responses, source matches, and safety classifications.

Temporal alignment of the source analysis reference lists

In the Ofcom reference lists used for source analysis, the date on which a URL was assessed and assigned a risk level did not necessarily coincide with this research's data collection period. As a result, a URL's assigned risk level may not accurately represent the content available at that URL when it was presented by an API. The source analysis findings should therefore be interpreted as matches against the version of the Ofcom reference lists used for the analysis, rather than as definitive assessments of the relevant websites at the precise time of data collection.

Conclusion

Alongside traditional search, GenAI search tools are increasingly shaping how users (including children) encounter content and receive answers to their questions. Understanding how both traditional and GenAI systems respond to queries relating to content harmful to children therefore plays an important part in developing the evidence base needed to work towards achieving safer outcomes for children online.

This research finds that both GenAI and traditional search APIs generally returned harmful content at low rates across the kinds of harmful content examined. However, the likelihood of harmful content being presented differed across kinds of harmful content. Queries related to violent content and body stigma were more likely to surface harmful content than those relating to abuse and hate, self-harm, or dangerous stunts and challenges. Harmful content was also occasionally returned in response to queries which did not actively seek it, although this was uncommon.

Additionally, API safety measures varied across kinds of content. GenAI APIs frequently incorporated safeguarding materials for some kinds of harmful content, particularly in responses relating to depression, self-harm, and eating disorders, and in some cases refused to answer queries related to harmful content altogether. However, safeguarding information was less consistently present for other harm categories, while some responses provided information which could encourage or facilitate access to harmful content without directly reproducing it. In addition, both GenAI and traditional search APIs retrieved links to sites hosting pornographic or gore content which appeared to lack safeguards, particularly where such content was actively sought through the search queries.

Taken together, these findings suggest that there is unlikely to be a single pattern of behaviours in response to all forms of queries related to harmful content. Instead, the ways in which GenAI and traditional search APIs respond to such queries, apply safeguards, and surface potentially harmful responses can differ substantially depending on factors such as the kind of harm investigated. These findings have helped support Ofcom's [Report on Content Harmful to Children](#), published by Ofcom in October 2026, under Section 63 of the Online Safety Act.

As GenAI and traditional search technologies continue to evolve, further research may help deepen our understanding of the landscape. Future work could examine how consumer-facing products differ from the APIs assessed in this research, explore how responses change over time as GenAI models advance, and investigate the extent to which different user profiles or other contextual signals influence search outcomes. Additional research could also consider other forms of search response content, including image- and audio-based outputs, as well as multi-turn conversational interactions that reflect other ways users engage with GenAI services in practice.

Further research in this area will help build a broader understanding of the risks and safeguards associated with emerging search technologies, and support future regulatory work to protect children online. In order to improve the effectiveness of future research, we welcome feedback on this paper. Please direct any comments or queries to our Online Safety Technology Team at OSTechRequests@ofcom.org.uk.

We also welcome responses in formats other than print, for example an audio recording or a British Sign Language (BSL) video. To respond in BSL:

- Send us a recording of you signing your response. This should be no longer than 5 minutes. Suitable file formats are DVDs, MP4 or QuickTime files; or
- Upload a video of you signing your response directly to YouTube (or another hosting site)