Accreditation for terrorism/CSEA content detection technology

Annexes 8-9: Technical information relating to minimum standards of accuracy proposals

Statement

Published: 8th May 2026

Contents

Annexes

A8. Accreditation application template

Applying for accreditation

A8.1 This is an illustrative example of an **accreditation application form** applicants might be expected to complete before being considered for accreditation against the advised minimum standards of accuracy. It includes the type of information that we are likely to request from applicants seeking accreditation before their technology is put forward for evaluation. Once the accreditation scheme has been set up, we will make the actual accreditation application form available to applicants, but this annex relates to the illustrative example.

A8.2 Applicants might expect to provide basic information about their technology, including about which harm type(s) the technology is capable of detecting, what types of data it can process, its outputs, and the compatibility of the technology. This information would be used by Ofcom or a nominated third party to understand the potential use cases for the technology, including potential limitations, and to contextualise the responses and evidence provided as part of the audit-based assessment.

A8.3 Through this approach, technologies would be pooled into categories as they are assessed for their accuracy. This process acts like a funnel: through the application form and accreditation process, Ofcom would understand the potential use cases for these technologies, including the target content and intended deployment context.

A8.4 The application form would need to be completed in full and to the satisfaction of Ofcom before the technology could be considered against minimum standards of accuracy. Applicants are required to have the information in this submission signed off by an individual with sufficient seniority within their organisation. This step will ensure that the submission has undergone appropriate internal governance before being submitted for consideration, even if it does not guarantee completeness or accuracy.

A8.5 Ofcom would reserve the right to not consider a technology against minimum standards of accuracy where it is found by a Court or other competent authority (such as the ICO) to have been developed in breach of UK data protection or other legal requirements.

A8.6 Accreditation of a technology would not necessarily mean that it is necessary and proportionate for Ofcom to require the use of that technology in a Notice. Further information on this topic may be found in Ofcom's Guidance for the providers of Part 3 services, which includes processes Ofcom would typically follow when deciding whether it is necessary and proportionate to issue such a Notice and some detail on matters to which Ofcom would expect to have regard when making this decision.

A8.7 Applicants might be expected to provide details across the following categories within an accreditation application template:

Information Category	Details
Description of the Technology	**Name:** [Technology Name] **Purpose:** Provide a brief overview of the technology, including its intended purpose and key functionalities. Explain if it's an AI-powered solution, a hybrid of AI and non-AI components, or a non-AI technology. Specify whether the technology is a machine learning model, rule-based model, software application, or another type of technology. Also, mention if it includes multiple consecutive or layered processes (e.g. pre-processing, core models, post-processing). Include details on the problem it aims to solve and its typical application scenarios. **Trust & Safety Application:** Provide a brief overview of whether the technology applies to user-to-user services, search services, or both. Additionally, indicate whether it is intended for use on messaging, search, livestreaming, or other service functionalities. **Human Review:** Specify if the technology is designed to require human review of all detected content, partial human review, or if it facilitates automatic takedown decisions without human intervention.
Harm Type	**Categories**: [Terrorism or CSEA] **Details**: Describe which particular terrorism and/or CSEA offences the technology is capable of detecting (e.g. terror propaganda and terror flags). Additionally, clarify how the technology interacts with these harms – whether it is designed to detect harmful content (by identifying it when it appears), prevent it (by blocking its distribution or access), or mitigate its effects (by flagging, reporting, or removing it).
Entity / Organisation	**Entity Name**: [Organisation/Individual Name] **Contact Information**: Provide contact details for the organisation or individual responsible for developing and maintaining the technology. **Ownership**: Explain the ownership structure, including any partnerships or collaborations. **Organisation Type**: Provide details about the organisation type (for-profit, non-profit, academic, public sector, etc.) **Organisational Size**: Provide details about the size of organisation, including number of employees.
Modality	**Supported Modalities**: Specify the types of data the technology can process (e.g. image, video, metadata, text, audio, multimodal). **Use Cases**: Describe how the technology operates across different modalities, if applicable.
Data Requirements	**Input Data Format**: Specify the types of data format required for the technology to function (e.g., .png, .pdf, .gif, .doc). **Data Volume**: Indicate the minimum, optimal, and maximum amounts of data needed for effective operation. **Data Pre-processing**: Describe any data quality standards or pre-processing needed before data can be used.
Outputs	Provide details on the output generated by the technology (e.g. confidence scores, probability, text classifications, image annotations, video analysis reports). Specify the format (e.g. JSON, XML, CSV) and any standard or custom schemas used.

Language (if applicable)	**Supported Languages**: List the languages the technology can process, including input and output languages, if applicable. **Language Support Details**: Explain how language processing is handled (e.g., through machine translation, specific language models) and any limitations in language capabilities.
Geography	**Development Location**: Indicate where the technology was developed (e.g. country, region). **Deployment Regions**: Provide details of the regions or countries where the technology has been deployed, and any geographical limitations or optimisations.
Previous Deployment	Detail previous instances where the technology has been deployed, including (if available) the name of organisations, dates, and specific use cases. If applicable, provide case studies or examples of how the technology was used, including the outcomes and any challenges faced.
Previous Accreditation	**Previous Accreditation:** Provide information regarding any previous accreditations of this technology by Ofcom against the minimum standards of accuracy for use in Technology Notices. If technology has been previously accredited for use in Technology Notices, provide information on the following: **Versioning**: Indicate the version of technology that was previously accredited, and any version history since that point. **Change Log**: Provide a summary of significant changes in each version (e.g. bug fixes, algorithm improvements) since the point of previous accreditation.
Resourcing	Provide details on the computational resources needed to deploy the technology, including CPU/GPU specifications, memory, and storage requirements. Additionally, provide details on the technology's ability to scale, including how it handles increased input volumes, concurrent processing, and whether it supports distributed computing.
Dependencies	**Required Technologies**: List any software, hardware, or third-party services that the technology relies on for operation (e.g. specific libraries, cloud services, operating systems, existing models). **Interoperability**: Explain how these dependencies are integrated and any potential risks or issues related to them. **Adaptation**: If applicable, provide information about any existing models or algorithms (open-source or otherwise) that the technology has been directly adapted from.
Compatibility	**Client/Surface Compatibility**: Provide details on the extent to which the technology can be used across different platforms (e.g. desktop, mobile, web). **Deployment Environment:** Specify the target deployment environment (e.g. cloud, on-premises, edge). **Cross-Platform Integration**: Explain any specific requirements or limitations for different clients or surfaces (e.g. browser compatibility, mobile OS versions).

Privacy and Legal Considerations	**Data Protection:** Explain how the technology handles user data, including collection, storage, processing, and any measures taken to anonymise or pseudonymise data. Additionally, provide information on whether the technology developer has ever been found in breach of UK data protection requirements, and if so, what actions have been taken to address these issues. **Information Security and Access Control**: Detail the protocols in place to protect data from unauthorised access and explain how access to the technology is managed and monitored. **Previous Legal Convictions:** Outline whether the organisation has any legal convictions, e.g. under section 7 of the Bribery Act 2010, along with details of the steps taken to resolve the matters leading to any conviction.

A9. Illustrative questions for the audit-based assessment

A9.1 To help applicants understand what evidence would likely be needed in support of each Objective included in the Audit-based Assessment, and to ensure a consistent approach to scoring by Ofcom or the nominated third party, a list of questions would be produced for accreditation. These questions would correspond to each of the Objectives and provide greater detail on the evidence required to score full or partial marks against the Objective. There would likely be multiple questions for each Objective.

A9.2 We have set out some non-exhaustive examples of the types of questions we would expect to ask for each of the Principles and Objectives below, and the format in which they could be presented to applicants. These are for illustrative purposes only, and are not a prescriptive checklist nor an indication of whether the exact same questions would be included as part of the accreditation scheme. Questions are also subject to change as the accreditation scheme matures, or as technological advancements take place.

A9.3 Some Objectives and questions below are marked with an asterisk*. These are specific questions we have identified for which it may be challenging for some technology developers to provide evidence. Specifically, where technologies have been developed without access to input/output and/or training data.

A9.4 In these cases, the technology developer seeking accreditation should provide evidence against those questions to the extent possible. Where the developer does not have direct access to relevant information required to respond (e.g., the developer cannot view the datasets against which the technology was developed or tested), it should explore alternative ways of providing relevant evidence.

A9.5 Where a technology developer nevertheless remains unable to provide any evidence against any of those questions, it should confirm to the accreditor's satisfaction that its technology was developed without access to input/output and/or training data and explain why it has been unable to provide any relevant evidence (and the steps taken to assure itself of this).

A9.6 Any terms in **bold** are defined in the glossary in Annex 3.

A9.7 0 points will be awarded for the questions below if the evidence required for 1 or 5 points is not provided or if evidence provided appears to contain misleading or inaccurate information.

Principle 1: Technical performance

Objective 1.1 – Performance Metrics:

The technology's ability to identify terrorism or CSEA content (as the case may be) has been comprehensively evaluated against the false positive rate and other appropriate performance metrics. Corresponding evaluation results are provided and demonstrate that the technology is able to detect terrorism or CSEA content (as the case may be), and those results have been used to determine that the technology is suitable for deployment in the environment(s) for which it has been designed.

Illustrative question	Scores	Evidence that we may expect

Question 1: What is the overall technical performance of the technology, and how is this used to determine whether the technology is suitable for use?
 	
Provide documented evidence of how the technical performance of the technology has been evaluated **pre-deployment**, and the results obtained including the reported mean and variance of the performance metrics. The **false positive rate** is a mandatory metric.
	1 point: Evidence demonstrates that the technical performance of the technology has been evaluated pre-**deployment** with some quantitative results. In particular, internal mean and variance of the performance test results (i.e., **false positive rate** and metrics such as the number of **false positives, accuracy, precision,** and **recall**) are reported but on limited or non-diverse datasets and without detailed breakdown, description of how evaluation metrics are used to determine whether to **deploy** the technology is limited, and there are minimal examples of performance monitoring.
	
	5 points: Comprehensive evidence demonstrates that the technical performance of the technology has been **independently verified**, **validated**, and **evaluated** pre-**deployment** with statistically robust quantitative results. In particular, detailed performance analysis on all relevant metrics including mean and variance across all relevant performance metrics are provided with extensive breakdowns of metrics that assess the prevalence of **false positives** by harmful and non-harmful content (i.e., content type, language, and scenario, where relevant), and the analyses are carried out on large-scale, diverse, and representative testing across different scenarios, content types and languages (where relevant). Documentation on whether to **deploy** the technology is detailed with quantitative examples.
		• Evaluation of the mandatory performance metrics (i.e., the **false positive rate**).
		• Documentation on mean and variance of the performance metrics.
		• Evaluation of all other relevant performance metrics (e.g., **confusion matrices**, **ROC curves**, **accuracy**, number of **false positives**, **F1 scores**, **collision rate**, **throughput**, and **FNR**).
		• Breakdowns of metrics that assess the prevalence of **false positives** by harmful and non-harmful content (i.e., content type, language, and scenarios, where relevant).
		• Documentation on **independently verified** and **validated** mean and variance of the performance metrics.
		• Description of how evaluating the performance metrics has been used to determine whether to **deploy** the technology, including but not limited to:
		  ○ Testing procedures and process flows
		  ○ Benchmarking reports
		  ○ System parameter tuning with systematic evaluation and optimisation
		  ○ Selection of architecture for specific tasks
		• Documentation on examples of performance monitoring.

Illustrative question	Scores	Evidence that we may expect
Question 2: How do you manage edge cases where performance metrics might be lower? For example, are there specific scenarios where **false positives** or **false negatives** are more prevalent? Provide documented evidence of specific cases along with strategies implemented to address and mitigate them.	**1 point:** Evidence demonstrates examples of edge cases and prevalent error scenarios with limited analysis and limited response strategies. For machine learning-based technologies or technologies containing machine learning-based components, limited generalisation demonstrated with limited testing results, where there is a significant performance drop on out-of-distribution data or edge cases. Reports on a small number of manually reviewed misclassified examples with minimal analysis of error patterns. **5 points:** Comprehensive evidence demonstrates identification, analysis, and current mitigation strategies of edge cases, as well as detailed analysis of error prevalence in specific scenarios, with evidence of regular monitoring and improvement. For machine learning-based technologies or technologies containing machine learning-based components, detailed generalisation analysis that shows consistent performance across a diverse out-of-distribution dataset, edge cases, content formats, languages, and cultural contexts, with detailed trade-off assessments between generalisation and specificity, as well as advanced strategies for ensuring generalisation of the technology. Detailed reports on error correction processes, manual review processes, and analysis of common error patterns, with **independent evaluations** confirming reduced error rates.	• Documentation on edge cases management. • Documentation on strategies implemented to address and mitigate edge cases with lower performance. • Performance reports on out-of-distribution data or edges with before-and-after performance comparisons. • Trade-off assessments between generalisation and specificity. • Reports on manually reviewed misclassified examples with analysis of error patterns. • Reports on error correction processes. • Logs of regular monitoring of edge cases. • Reports on testing in various low-resource scenario and analysis of performance in rare content types and edge cases, with **independently verified** results.

Illustrative question	Scores	Evidence that we may expect
Question 3: Have you conducted any longitudinal studies to assess the performance metrics over an extended period? Provide documented evidence of these studies and their findings.	**1 point:** Evidence demonstrates short-term (t < 6 months) studies are conducted with limited scope. **5 points:** Comprehensive evidence demonstrates that longitudinal studies are conducted with detailed analysis over an extended period.	• Reports on short-term studies. • Reports on medium-term (6 months < t < 1 year) or preferably long-term studies (t > 1 year). • Analysis of trends and implications for technology performance.
Question 4: How is the technology's performance affected by the characteristics of input data? Provide documented evidence of performance tests conducted with varying data characteristics.	**1 point:** Evidence demonstrates that tests are conducted on a limited set of data characteristics. **5 points:** Comprehensive evidence demonstrates testing and analysis of model performance across a wide range of data characteristics (e.g., high definition versus low-definition images), with optimisation strategies documented and quantitative results.	• Test reports on variance of the performance of the technology with data characteristics variation. • Test reports on variance of the performance of the technology with optimisation strategies. • Test reports on variance of the performance of the technology with quantitative results. [For example: **Images** = Different Resolution, Colour Depth, Channels, Total Number of Pixels, Spatial Complexity. **Text** = Different Sequence Length, Vocabulary Size, Encoding, Syntactic and Semantic Complexity. **Video** = Different Resolution, Frame Rate, Number of Frames, Channels, Temporal Complexity, Bitrate. **Audio** = Different Sample Rate, Bit Depth, Channels, Duration, Spectral Complexity, Amplitude and Loudness Variations.]

Illustrative question	Scores	Evidence that we may expect
Question 5: What protocols are in place for measuring and documenting the **confidence scores** of different types of **target content**? Provide documented evidence of **confidence score** assessments across content types, with procedures for handling low-confidence predictions and uncertainty management.	**1 point:** Evidence demonstrates reporting of measurement of **confidence scores,** but evidence may focus on only a few different content types. For machine learning-based technologies or technologies containing machine learning-based components, limited reporting of **confidence scores** across only a few different content types, with minimal analysis of the impact on predictions leading to potentially unreliable or inconsistent **confidence scores**. Procedures for handling low-confidence predictions exist but with limited effectiveness. Limited reporting of uncertainty management with minimal analysis of impact on predictions, potentially leading to misinterpretation of results. **5 points:** Comprehensive evidence demonstrates reporting of measurement and documentation of **confidence scores** across all relevant content types, with detailed analysis and regular improvement strategies. For machine learning-based technologies or technologies containing machine learning-based components, detailed analysis of **confidence scores** with reliable **confidence scores** across all relevant content types, demonstrating accurate uncertainty estimation and robust handling of ambiguous cases. Procedures for handling low-confidence predictions are effective, with robust strategies to ensure **accuracy** and minimise risks. Comprehensive report of uncertainty management practices, with robust strategies to minimise and communicate uncertainty.	• Reports on **confidence scores** for different content types, with analysis of differences across content types and examples of their impact on predictions. • Documentation of procedures for low-confidence predictions, with examples of handling methods and analysis on their effectiveness. • Documentation on uncertainty management practices, with examples of uncertainty reporting and analysis of impact on predictions. • Documentation of methods used for measuring **confidence scores**. • Documentation on regular improvement strategies. • **Independent verification** and **validation** of **confidence score** assessments. • Documentation on how **confidence thresholds** are set (**precision** vs. **recall** trade-offs) for different content types, with analysis of impact on predictions and confidence intervals. • **Independent verification** of low-confidence prediction handling procedures. • **Independent verification** of uncertainty reporting practices, with analysis on how uncertainty is communicated and managed in decision-making.

Objective 1.2 – Dataset Quality:

The datasets used in development, including where applicable the training and testing of the technology's performance, are sufficiently comprehensive, representative of the harm being detected and, where relevant, sufficiently diverse to test the technology's ability to generalise to data not seen during development.

Illustrative question	Scores	Evidence that we may expect
Question 1: What criteria were used to select the **development datasets** and **testing datasets**? Provide documented evidence of the selection criteria and the rationale behind them. If the developer does not have direct access to this information (e.g., data has been used which is illegal to hold, and therefore the developer cannot view it), they should request this information (or a summary) from the data provider.	**1 point:** Evidence demonstrates dataset diversity has a limited coverage of harm types and scenarios, limited diversity in culture, geography and language (where relevant) that primarily represent a narrow demographic. Some consideration of definitions and labelling, but with minimal input from multiple moderators. **5 points:** Comprehensive evidence demonstrates that datasets are highly diverse with respect to culture, geography and language (where relevant) that cover a wide range of harm types, scenarios, and content types. Datasets are well-aligned with varied definitions of harm and have been thoroughly labelled by a diverse group of moderators. Comprehensive balancing process that ensures equitable representation of all relevant **target content** categories, with regular reviews and adjustments made to maintain balance as new data is added. Datasets are tailored to ensure broad applicability and cultural sensitivity.	• Reports on datasets selection criteria with descriptions of dataset content. • Reports on different harm types and scenarios, with mention of diversity factors, regional, or linguistic representation (where relevant). • Reports on cross-cultural analysis of dataset relevance. • Documentation on labelling guidelines. • Documentation on moderator involvement. • Reports on represented regions, cultures, and languages (where relevant) to show that datasets are tailored for a variety of scenarios and specific cultural contexts. • Documentation of inclusion efforts for under-represented groups.

Illustrative question	Scores	Evidence that we may expect
Question 2: What steps are taken to ensure that **development datasets** and **testing datasets** are up-to-date and relevant? Provide documented evidence of the processes used to procure and/or update datasets. If the developer does not have direct access to this information (e.g., data has been used which is illegal to hold, and therefore the developer cannot view it), they should request this information (or a summary) from the data provider.	**1 point:** Evidence demonstrates ad-hoc or infrequent updates to datasets. **5 points:** Comprehensive evidence demonstrates regular updates to datasets, ensuring they reflect the latest trends, content types, and harm definitions. Clear processes and triggers for updates, with detailed documentation and validation of relevance.	• Logs or records of dataset updates, with discussion of relevance or timeliness. • Reports of dataset updates, with examples of new content types included. • Documentation of trend monitoring. • Documentation on update schedules and processes. • Documentation on relevance checks with criteria for determining dataset relevance. • Case studies of dataset refreshment to show adaptation to new trends. • Validation reports ensuring up-to-date-content. • Documentation of processes for reviewing and refreshing datasets.
Question 3: What is the process for validating the **accuracy** and relevance of the labels used in the **development datasets and testing datasets**, and if sourced externally, the overall integrity of the datasets? Provide documented evidence of the labelling process and validation checks. Were multiple moderators involved to ensure accuracy and reduce bias? If the developer does not have direct access to this information (e.g., data has been used which is illegal to hold, and therefore the developer cannot view it), they should request this information (or a summary) from the data provider.	**1 point:** Evidence demonstrates label validation processes but with limited checks. Labels may be applied inconsistently or without comprehensive guidelines. **5 points:** Comprehensive evidence demonstrates label validation process involving multiple checks and reviews by a diverse group of moderators. Processes are in place to ensure consistency of labelling, as well as the integrity of any externally sourced datasets.	• Documentation of labelling guidelines. • Examples of labelled data with validation reports. • Records of checks and processes for ensuring integrity of any externally sourced datasets. • Documentation on cross-checking procedures. • Documentation of moderator diversity and involvement. • Examples of validation iterations. • Reports on how labelling guidelines and processes are in place to maintain label consistency, accuracy, and relevance across all datasets. • Documentation on any external reviews or assessments. • Documentation of processes, automated tools and audits or validation used to ensure the integrity of any externally sourced/third-party datasets.

Objective 1.3 – Reproducible Performance:

The technology's performance is sufficiently consistent and reproducible across the environment(s) for which it has been designed.

Illustrative question	Scores	Evidence that we may expect
Question 1: Are testing procedures and conditions well-documented with a view to ensuring reproducibility? Provide documented evidence of these procedures and any related documentation, such as version control practices and update logs for governance documents.	**1 point:** Evidence to demonstrate that documentation of testing procedures and conditions exist, but with gaps that may hinder full reproducibility. Technology documentation is available only to internal developers. **5 points:** Comprehensive documentation demonstrates detailed documentation of testing procedures and conditions, including hardware/software configurations, environmental conditions (i.e., how variations in hardware and software are documented and controlled for during testing), with all necessary details to ensure reproducibility across various scenarios. Technology documentation is available and accessible to all stakeholders of testing procedures and conditions.	• Documentation on testing procedures with details on testing conditions. • Documentation on version control practices. • Update records. • **Independent verification** that documentation supports full reproducibility. • Reports on regular documentation reviews. • Documentation on use-case guidelines. • Reports on outcomes of testing.
Question 2: How are discrepancies in results between different reproducibility tests handled? Provide documented evidence of processes in place and any case studies of discrepancies which have been identified and the steps which were taken to resolve them.	**1 point:** Evidence demonstrates that processes for handling discrepancies are in place, but with inconsistent resolution. **5 points:** Comprehensive evidence demonstrates processes for identifying, documenting, and resolving discrepancies, with detailed analysis and resolution ensuring consistency across reproducibility tests.	• Discrepancy logs. • Examples of unresolved discrepancies. • Documentation of resolution processes for discrepancies. • Documentation of resolution strategies. • Reports on root cause analysis. • Examples of resolved discrepancies. • Validation reports showing improvements post-resolution.

Illustrative question	Scores	Evidence that we may expect
Question 3: What steps are taken to ensure consistent performance across different platforms or devices? Provide documented evidence of cross-platform performance assessments.	**1 point:** Evidence demonstrates that checks for platform and device consistency exist, but with some documented inconsistencies. **5 points:** Comprehensive evidence demonstrates consistent performance across all platforms and devices it has been designed for, with detailed testing and documentation across diverse environments.	• Platform/device testing logs. • Examples of platform/device inconsistencies. • Documentation of platform/device compatibility checks. • Documentation of platform/device consistency checks. • Reports on issues and resolutions. • Validation reports confirming platform/device consistency. • **Independent verification** across all platforms/devices.
Question 4: What protocols are in place to ensure that the technology can be effectively scaled to accommodate larger datasets or larger-scale deployment scenarios? Provide documented evidence of scaling tests and their impact on generalisation for classifiers.	**1 point:** Evidence to demonstrate that there is consideration of the capability of the technology when scaling up with limited testing on larger datasets or scenarios which simulate large-scale **deployment**, with some performance degradation. **5 points:** Comprehensive testing and strategies to ensure consistent capability of the technology across an appropriate range of scales.	• Reports on scalability tests, with examples of performance issues when scaling. • Reports on capability and performance at scale, with **independent verification** of scalability. • Reports on generalisation across a range of scales for classifiers.

Objective 1.4 – *Secondary Validation:

The technology's outputs, where possible, have been evaluated during performance testing against expert human judgement, particularly in complex or nuanced cases. Where outputs cannot be validated by humans, other secondary validation measures have been undertaken.

Illustrative question	Scores	Evidence that we may expect
Question 1*: To what extent has the performance of the technology been assessed in comparison to the performance of human judgement, particularly in difficult edge cases? Provide documented evidence of the validation process and protocols.	**1 point:** Evidence demonstrates limited validation against human judgment, with some alignment but notable discrepancies in complex cases. **5 points:** Comprehensive evidence demonstrates detailed validation showing strong alignment with human judgment across a wide range of cases, including complex and nuanced scenarios. Discrepancies are minimal and well-documented with clear mitigation strategies.	• Reports on comparison between model and human judgement. • Examples of discrepancies in difficult cases. • Analysis of the differences between model and human judgement. • Documentation of validation protocols. • Reports on strong alignment across diverse scenarios. • Documentation on mitigation strategies when there is misalignment with human decision. • **Independent validation** of the results of validation against human judgement. • Documentation of discrepancy studies with examples of minimal divergence. • **Independent validation** of improvement strategies.

Illustrative question	Scores	Evidence that we may expect
Question 2*: How do you handle cases where human judgment significantly differs from the technology's outputs? Have any adjustments been made to the technology based on feedback from human judgment validation? Provide documented evidence of adjustments and their impact.	**1 point:** Evidence demonstrates adjustments are made but with limited analysis of impact. **5 points:** Comprehensive evidence demonstrates systematic adjustments are made based on human judgment validation, with robust analysis and **independent validation** of impact.	• Reports on adjustments made to the technology based on feedback from human judgement validation. • Examples of feedback from human judgement validation. • Analysis of the impact of the adjustments on the performance of the technology. • Reports on integration of feedback from human judgement validation. • Reports on handling the feedback from human judgement validation. • **Independent validation** of the impact of the adjustments on the performance of the technology. • Reports on regular improvement efforts.
Question 3: To what extent has the technology been benchmarked against alternative methods or models using real (i.e., non-synthetic) data? Provide documented evidence of these benchmarks, the criteria used for comparison, and human judgement made during the evaluation process, such as external or **independent evaluations** that have been conducted to benchmark the technology.	**1 point:** Evidence demonstrates comparisons are made to a single baseline model/method, and with limited analysis of human judgement made during the evaluation process. **5 points:** Comprehensive evidence demonstrates comparisons are made across several relevant baseline models/methods, with detailed analysis demonstrating comparable (or superior) performance in key areas (e.g., accuracy efficiency, generalisation). Detailed human reviews are conducted to make a judgement on the evaluation process, such that the reports on the external evaluations are detailed.	• Reports on comparisons to relevant baseline models/methods, with analysis of the differences in performance. • Documentation of human judgements made. • Benchmark reports. • Reports on comparisons to relevant models/methods, with analysis of performance metrics demonstrating comparable (or superior) performance in key areas. • **Independent evaluations** of relative performance. • Documentation of rationale behind any human decisions made, including analysis, comparison to baseline models/methods, and quantitative results.

Principle 2: Fairness

Objective 2.1 – *Bias Identification and Mitigation:

Comprehensive policies, procedures, metrics, and analyses have been implemented to identify potential biases in the technology throughout planning and development. In addition, robust bias mitigation strategies have been implemented and their success has been measured over time, including checks for demographic fairness and audits on any automated decision making.

Illustrative question	Scores	Evidence that we may expect
Question 1: What are the potential biases that may arise as a result of how the technology has been designed and developed? Provide documented evidence of identified biases, risk assessment processes, monitoring processes, and any frameworks or tools used to evaluate bias risks.	**1 point**: Evidence demonstrates identified potential biases, including bias risk assessments that do not cover the entire lifecycle of the technology and may be incomplete or lacking in depth. For content moderation classifiers, evidence demonstrates bias identification related to different contexts, but they may be limited in scope or effectiveness. **5 points:** Comprehensive evidence demonstrates identified potential biases, with detailed documentation on bias risk assessments throughout the entire lifecycle of the technology. For content classifiers, comprehensive evidence of bias identification related to different contexts at every stage of the lifecycle of the technology.	• Documentation of potential biases identified. • Reports on bias risk assessment. • Bias monitoring logs. • Examples of bias identification related cultural nuances, backgrounds, scenes, objects, or symbols. • Records of bias monitoring throughout the lifecycle of the technology. • Documentation and records of identified potential context-specific biases related to cultural nuances, backgrounds, scenes, objects, or symbols with examples. • Documentation of fairer outcomes in content moderation.

Illustrative question	Scores	Evidence that we may expect
Question 2*: What mechanisms or strategies are in place to identify and mitigate different types of biases (e.g., selection bias, confirmation bias, etc.) in the technology? Provide documented evidence of bias identification methods, including specific examples of biases detected, as well as documented evidence of tailored approaches to address and mitigate specific biases, including case studies or examples.	**1 point**: Evidence demonstrates strategies are employed to identify and mitigate some types of biases, but they may be limited in scope or effectiveness, for example, not all contextual biases are mitigated or there is a minimal focus on specific demographic groups. **5 points:** Comprehensive evidence demonstrates advanced bias identification and mitigation strategies are employed, with proactive identification and detailed documentation of all relevant biases and regular monitoring and adjustment to optimise bias mitigation for all different types of biases. If biases related to gender, age, and ethnicity were identified, the detailed documentation shows successful bias mitigation and fair treatment across all demographic groups. The bias identification and mitigation policy and practices are aligned with ISO/IEC TR 24027 (Bias in AI systems and AI aided decision making), with detailed ongoing adherence.	• Documentation on bias identification and mitigation strategies. • Examples of biases detected, with tailored bias mitigation approach to address specific biases. • Case studies or examples of bias correction. • Documentation on testing protocol for bias identification for the content moderation classifiers. • Reports on test cases and scenarios for bias identification testing. • Documentation on automatic bias identification and mitigation. • Records of bias detection logs and bias correction methods. • Reports on bias analysis. • Documentation and examples of post-processing techniques with a view to ensuring improved fairness. • Documentation on impact assessments, with examples of bias impacts and records of identification efforts.
Question 3: Are sensitivity tests conducted to understand how changes in input data affect the technology's bias? Provide documented evidence of these tests and the resulting impact on the technology's outputs.	**1 point**: Evidence demonstrates sensitivity tests are conducted, but they may be limited in scope or frequency. **5 points:** Comprehensive evidence demonstrates detailed and regular sensitivity tests are conducted, using advanced methodologies to evaluate and adjust for bias. Detailed test results showing proactive and systematic adjustments are made to the technology to ensure fairness.	• Reports on sensitivity tests. • Analysis on input data changes. • Documentation on the impact on biases. • Regular records of adjustments, with results of the adjustments to the technology.

Illustrative question	Scores	Evidence that we may expect
Question 4: How do you ensure that updating/retraining the technology does not introduce new biases or reduce fairness? Provide documented evidence of updating/retraining protocols, fairness audits post-retraining, and any adjustments made to maintain fairness.	**1 point:** Evidence demonstrates that retraining protocols exist, but they may be limited in scope or effectiveness. Documentation on fairness audits post-retraining is provided, but with minimal adjustments made to maintain fairness. Minimal checks for bias in the data used for updating/retraining of the technology, with some recognition of bias issues but with limited corrective actions. **5 points:** Comprehensive evidence demonstrates advanced and regular updating/retraining protocols are employed, with detailed audits and adjustments to ensure that no new biases are introduced. Detailed bias assessments and corrections, with **independent verification** that updating/retraining has effectively mitigated bias issues. Detailed documentation showing successful fairness improvements and consistency in moderation outcomes post-retraining.	• Documentation on updating/retraining protocols. • Records of fairness audits post-retraining. • Records of adjustments made to maintain fairness. • Reports on identifying bias issues in the data used for updating/retraining, with examples of bias correction. • Documentation on fair outcomes after retraining. • Reports on bias analysis. • Reports on bias correction post-updating/retraining. • **Independent verification** of bias mitigation.
Question 5*: How is the risk of historical biases embedded in the **development data** analysed and addressed? Provide documented evidence of bias audits or reviews of historical **development data** and actions taken to mitigate these biases.	**1 point**: Evidence demonstrates reviews of historical development data are conducted, but they may be limited in scope or depth. Some actions are taken to address identified biases, but they may be insufficient. **5 points:** Comprehensive evidence demonstrates detailed and regular audits of historical development data are conducted, with advanced techniques used to identify and correct biases. Mitigation strategies are proactively applied, with regular monitoring and adjustment.	• Reports on bias audit. • Reports on reviews of historical **development data**. • Documentation on actions taken to mitigate biases. • Documentation on bias mitigation strategies. • Logs of regular monitoring. • Impact assessments of bias corrective actions.

Objective 2.2 – *Data Processing:

The data processing used for any relevant training or testing datasets is robust, with documented, standardised criteria used to process data. Measures have also been taken to ensure consistency and minimise bias and errors during the processing.

Illustrative question	Scores	Evidence that we may expect
Question 1: Describe the data preprocessing procedures applied at the development and inference stages to reduce bias. Provide documented evidence of preprocessing techniques such as rebalancing, resizing, anonymisation, URL canonicalisation, or bias-correction applied to datasets.	**1 point**: Evidence demonstrates preprocessing methods are used, but they may be limited in scope or effectiveness. **5 points:** Comprehensive evidence demonstrates advanced preprocessing methods are employed to proactively reduce bias, with detailed documentation. Regular monitoring and adjustment of preprocessing methods are conducted to optimise bias reduction. The data labelling policy and practices are aligned with ISO/IEC TR 24027 (Bias in AI systems and AI aided decision making), with detailed ongoing adherence.	• Logs of preprocessing techniques used. • Reports on results of bias reduction. • Records of regular monitoring of preprocessing techniques. • Reports on optimised performance of the technology through preprocessing.
Question 2*: How is the labelling process during preprocessing designed and validated to minimise or prevent bias? Provide documented evidence of labelling guidelines, cross-validation with multiple annotators, and steps taken to minimise bias in labelled **development data**.	**1 point**: Evidence demonstrates that labelling guidelines are followed, but they may be limited or inconsistently applied. Some efforts are made to cross-validate labels with multiple annotators, but bias may still be present. **5 points:** Comprehensive evidence demonstrates advanced measures are employed to ensure that the labelling process is unbiased, with detailed cross-validation and regular audits to maintain fairness. Regular monitoring and adjustments of the labelling process are conducted.	• Data labelling guidelines. • Logs of cross-validation. • Documentation on steps taken to minimise bias in labelled data. • Reports on accuracy in labelled data. • Records of regular monitoring of labelling processes.

Objective 2.3 – Interpretability and Explainability

The rationale behind algorithmic decisions made by the technology can be sufficiently understood by Ofcom and companies that are likely to deploy the technology.

Illustrative question	Scores	Evidence that we may expect
Question 1: What techniques are used to ensure the technology is **interpretable**? Provide documented evidence which details the architecture and decision-making process of the technology, features in the **development data**, and any techniques used to ensure the decision-making process of the technology can be understood by humans.	**1 point:** Evidence demonstrates that reports detailing the architecture and decision-making process of the technology are provided and **interpretability** techniques are used, but the reports may be limited in scope, and the techniques do not allow humans to fully understand the decision-making process of the technology. For machine learning-based technology or technology that contains a machine learning-based component, evidence of **explainability** techniques employed and documentation of communication protocols or user interface may be available, but they may be limited in effectiveness. **5 points:** Comprehensive evidence demonstrates that detailed reports are provided which outline the architecture and decision-making process of the technology, and that advanced **interpretability** techniques are used. For machine learning-based technology or technology that contains a machine learning-based component, comprehensive evidence of **explainability** techniques and advanced communication strategies are employed, ensuring that all decisions are understandable to users. Detailed documentation showing successful communication and user satisfaction with the **interpretability** and/or **explainability** of decisions.	• Reports detailing the architecture of the technology, its operation (i.e., the decision-making process), and features in the **development data**. • Documentation on employed **interpretability** techniques. • Documentation on employed **explainability** techniques. • Examples of user-friendly explanations provided to users. • Examples of user interfaces to explain the decisions. • Reports on Feedback mechanisms. • Records of explanations. • Documentation on communication protocols. • Reports on successful user comprehension and feedback. • Guidelines or framework used to standardise explanations, ensuring that similar decisions are communicated in a similar manner.

Illustrative question	Scores	Evidence that we may expect
Question 2: How is it ensured that the output of the technology in complex or edge cases is **interpretable** (and **explainable** for machine learning-based technology or technology with machine learning-based components)? Provide documented evidence of methodologies or tools used to **interpret** and/or **explain** decisions in scenarios where the content is difficult to make a decision on.	**1 point**: Evidence demonstrates that efforts are made to ensure the technology is **interpretable** (and **explainable** for machine learning-based technology or technology with machine learning-based components) decisions in complex cases, but the methodologies or tools used might not be effective in **interpreting** and/or **explaining** the complex edge cases. **5 points:** Comprehensive evidence demonstrates advanced efforts are employed to ensure that all decisions, even in the most complex cases, are **interpretable** (and **explainable** for machine learning-based technology or technology with machine learning-based components). Detailed evidence showing successful **interpretation** and/or **explanation**, and improvement in handling complex or edge cases.	• Documentation of methodologies or tools used to ensure the technology is **interpretable** (and **explainable** for machine learning-based technology or technology with machine learning-based components) complex cases. • Examples of user comprehension or feedback. • Examples of successful **interpretation** and/or **explanation**. • Reports on fair outcomes in complex scenarios. • Reports on improvements made to ensure complex or edge cases are **interpretable** and/or **explainable.**
Question 3: What **interpretability** or **explainability** methods are used to identify any biases, errors, or misleading behaviours of the technology so that corrections can be made subsequently. Provide documented evidence of regular improvements of the technology using explanations.	**1 point:** Evidence demonstrates that efforts are made to use explanations to identify biases, errors, or misleading behaviours, and corrective actions are taken based on the explanations, but they may be limited in effectiveness or detail. **5 points**: Comprehensive evidence demonstrates advanced **explainability** techniques are employed ensuring that biases, errors, or misleading behaviours embedding in the technology can be identified. Detailed documentations showing regular efforts to improve the technology by implementing corrective actions, showing clear evidence of improvements.	• Documentation of **explainability** techniques. • Reports on identified biases, errors, or misleading behaviours using **explainability** techniques. • Reports on improvements made using the explanations. • Analysis of identified biases, errors, or misleading behaviours of the technology using the explanations. • Logs of regular monitoring. • Reports on improvements made to ensure the biases, errors, or misleading behaviours are minimised from the technology.

Principle 3: Robustness

Objective 3.1 – Development in a Secure Environment:

The technology has been developed with sufficient cybersecurity, privacy, and data protection measures in place, particularly for ensuring the integrity of the algorithm and protection of sensitive data. Documentation of how secure design principles have been followed during software development is provided.

Illustrative question	Scores	Evidence that we may expect
Question 1: Provide details of the personnel responsible in the developer's organisation for the implementation of cybersecurity, privacy, and data protection practices and any reporting mechanism(s) for potential security and data protection breaches.	**1 point:** Evidence demonstrates that key personnel responsible for implementing cybersecurity, privacy, and data protection practices are identified within the organisation, although implementation of those practices may be inconsistent and the exact delineation of responsibilities unclear. Mechanisms are in place for the reporting of security and data protection incidents but may lack consistent implementation. **5 points:** Comprehensive evidence demonstrates that cybersecurity, privacy, and data protection responsibilities are clearly delineated within the organisation. Reporting mechanisms are available, and there is evidence that incidents are investigated in a timely manner, with a clear chain of command and accountability.	• Documentation identifying key personnel responsible for implementing cybersecurity, privacy, and data protection practices, with a clear chain of command. • Documentation outlining reporting mechanisms in place for reporting of security and data protection incidents/breaches. • Documented escalation procedures for security and data protection breaches. • Records of timely investigation and resolution of incidents. • Records of periodic audits of cybersecurity, privacy, and data protections practices and reporting mechanisms for incidents/breaches.

Illustrative question	Scores	Evidence that we may expect
Question 2: What security measures are in place to protect personal or sensitive data (including **terrorism/CSEA content**) used to develop and/or test your technology. Provide details of how the technology stores or processes input data and any additional processes implemented to handle **terrorism/CSEA content**. If the developer does not have direct access to this information (e.g., data has been used which is illegal to hold, and therefore the developer cannot view it), they should obtain this information (or a summary) from the data provider.	**1 point:** Evidence demonstrates that security measures are in place for the storage of personal or sensitive data, but gaps in protection exist. **5 points:** Comprehensive evidence demonstrates that personal or sensitive data is stored with a high level of security, and detailed documentation exists for all security measures in place. Evidence demonstrates that regular audits and/or security assessments are performed to ensure a high level of ongoing data security.	• Documentation outlining security protocols for data storage and processing (e.g., end-to-end encryption). • Documentation outlining how personal or sensitive data is handled and processed. • Documentation and records demonstrating consistent application of access controls. • Documentation of secure handling protocols. • Records of regular audits and/or independent security assessments.

Illustrative question	Scores	Evidence that we may expect
Question 3: What cybersecurity and secure coding practices are in place during development? Provide documented evidence of the standards and practices followed by the development team, including how automated security checks are integrated both into your continuous integration/deployment (CI/CD) pipelines, and into monitoring of external dependencies (i.e., software packages which the technology utilises, but which are developed/maintained by a third-party).	**1 point:** Evidence demonstrates that cybersecurity and secure coding practices and automated security checks are in place, but they may be insufficient or inconsistently applied. **5 points:** Comprehensive evidence demonstrates that cybersecurity and secure coding practices and automated security checks are in place. Cybersecurity practices are aligned with ISO/IEC 27001 (Information Security Management Systems - Requirements) and/or 27032 (Cybersecurity — Guidelines for Internet security), with detailed ongoing adherence.	• Documentation outlining secure coding standards, • Examples of coding practices followed, and records of regular training for developers. • Records/documentation of ongoing monitoring of coding practices. • Documentation outlining automated security checks in place within CI/CD pipelines. • Documentation outlining access controls in place with records outlining regular reviews. • Documentation outlining dependency scanning and security checks applied to external dependencies/third-party components. • Records of regular audits and **independent validation** of secure coding practices.
Question 4: What measures are in place to ensure that security checks and practices are not bypassed? Provide documented evidence of mechanisms to enforce security mechanisms throughout the development lifecycle.	**1 point:** Evidence demonstrates that measures are in place to prevent bypassing of security checks and practices, but they may be insufficient or inconsistently enforced. **5 points:** Comprehensive evidence demonstrates that suitable measures are in place to prevent the bypassing of security checks and practices.	• Documentation of security enforcement mechanisms to prevent bypassing. • Examples outlining how security checks and practices have been applied. • Records of regular monitoring and **independent validation** of security measures. • Records of regular updates to security checks and practices.

Illustrative question	Scores	Evidence that we may expect
Question 5: What policies are in place to enforce multi-party approvals for sensitive actions (e.g., granting admin privileges, exporting sensitive data)? Provide documented evidence of approval workflows and logs.	**1 point:** Evidence demonstrates that policies for multi-party approvals exist, but they may be inconsistently enforced or insufficient. **5 points:** Comprehensive evidence demonstrates that suitable policies for multi-party approvals are in place.	• Documentation of multi-party approval policies. • Detailed logs of approval workflows. • Documentation outlining automated workflow tools and real-time monitoring systems. • Records of **independent validation** of approval processes in place.
Question 6: How do you protect against accidental exposure of sensitive services or data? Provide documented evidence of protective measures and incident response plans.	**1 point:** Evidence demonstrates that protective measures and incident response plans exist, but they may be insufficient or not consistently enforced. **5 points:** Comprehensive evidence demonstrates that suitable protective measures and incident response plans are in place.	• Documentation outlining protective measures and incident response plans. • Documentation outlining processes for ensuring secure communication of data in transit. • Records of past incidents, and responses to them. • Documentation outlining regular automated monitoring and alert systems. • Records of **independent validation** of the measures and plans in place. • Documentation outlining secure communication protocols and encryption techniques used.

Illustrative question	Scores	Evidence that we may expect
Question 7: How do you manage the attack surface related to both third-party components/integrations and other services, ports, or APIs? Provide documented evidence of actions taken to minimise the attack surface, including third-party risk assessments.	**1 point:** Evidence demonstrates that measures are in place to assess and manage risks associated with unnecessary services, ports, and APIs, as well as third-party components, but they may be insufficient or inconsistently enforced. **5 points:** Comprehensive evidence demonstrates that suitable measures are in place to assess and manage risks associated with unnecessary services, ports and APIs, as well as third-party components. Comprehensive risk assessments for third-party components or integrations are conducted.	• Documentation outlining ongoing efforts to minimise the system's attack surface, such as network segmentation, regular automated monitoring and alert systems and automated tools for identifying unnecessary elements (e.g., external services, ports and APIs whose removal has no impact on the operation of the technology). • Records documenting third-party component/integration risk assessments, including logs of findings and actions taken. • Records of **independent validation** of the processes in place for management of third-party components/integrations.

Objective 3.2 – Consistent Performance Over Time:

The technology maintains expected operation and performance over time, demonstrating its reliability and stability in tests representative of the environments for which it has been designed (and, where relevant, when deployed). Any degradation over time is regularly monitored and reported on.

Illustrative question	Scores	Evidence that we may expect
Question 1: What are the results from load testing that simulate high demand scenarios? Provide documented evidence of how the technology handles increased volumes of requests or data, including how it scales up or down based on demand.	**1 point**: Evidence demonstrates that load testing is conducted, with limited scenarios and minimal documentation of outcomes. Limited scalability mechanisms are in place. **5 points:** Comprehensive evidence demonstrates that suitable load testing is conducted across all relevant high-demand scenarios, with detailed documentation of performance metrics, proactive adjustments and implementation of advanced scalability mechanisms.	• Documentation outlining load testing procedures. • Records of specific high-demand scenarios tested, with performance metrics. • Examples of adjustments made based on load testing. • Documentation outlining the implementation of advanced scalability mechanisms, resource allocation plans and load balancing strategies.
Question 2: How is the technology routinely evaluated to assess its performance, and what historical performance data demonstrates consistency in the technology's performance over time? Provide documented evidence of processes for assessing performance over time, alongside historical performance metrics, including accuracy, precision, and recall.	**1 point**: Evidence demonstrates that processes for routine evaluation of performance over time are in place. Historical performance data is provided, with minimal analysis of consistency in key metrics. **5 points:** Comprehensive evidence demonstrates that processes for routine evaluation of performance over time are in place, with well-defined schedules and regular reviews undertaken. Monitoring of performance metrics over time is performed, with detailed documentation and analysis demonstrating long-term consistency.	• Documentation outlining processes for routine evaluation of performance over time, including evaluation schedules. • Logs/records containing historical performance data, including values of key metrics over time. • Regular monitoring reports which show sustained performance. • Longitudinal studies including records of performance metrics collected over periods of more than 6 months.

Illustrative question	Scores	Evidence that we may expect
Question 3: What testing procedures are followed before deploying updates to the technology, and how consistent is the technology's performance across different updates? Provide documented evidence of efforts to monitor and ensure consistency of performance.	**1 point**: Evidence demonstrates that testing procedures prior to updates are in place but these are limited, and/or limited evidence shows that performance is sufficiently consistent across a few updates of the technology. **5 points:** Comprehensive evidence demonstrates that rigorous testing procedures are followed prior to updates, with regular monitoring of performance across all updates, with detailed documentation demonstrating consistency.	• Documentation outlining testing procedures used prior to updates to the technology. • Logs of regular performance monitoring undertaken. • Records of performance across various updates, with comparisons of key metrics and evidence of consistency over time.

Objective 3.3 – Robust Incident Handling and Recovery:

The technology includes robust incident handling and recovery mechanisms, enabling the management of system failures or unexpected situations.

Illustrative question	Scores	Evidence that we may expect
Question 1: What protocols are in place for recovering functionality and performance, and ensuring data integrity after a failure? Provide documented evidence of recovery protocols, backup systems, and results.	**1 point**: Evidence demonstrates that recovery protocols and backup systems are in place, but that these may be insufficient and/or there is limited evidence that they are sufficient. **5 points:** Comprehensive evidence demonstrates that suitable recovery protocols and backup systems are in place, with detailed documentation, ensuring rapid and effective recovery and data integrity.	• Documentation outlining recovery protocols and backup systems in place. • Records of successful recovery efforts. • Records/logs demonstrating sustained functionality, performance and data integrity.
Question 2: How do you detect and report security incidents? Provide documented evidence of detection tools, monitoring systems, and reporting procedures.	**1 point:** Evidence demonstrates that tools and procedures for detecting and reporting security incidents exist, but they may not be consistently applied or well-documented. **5 points:** Comprehensive evidence demonstrates that suitable detection and reporting systems for security incidents are in place with detailed documentation, ensuring a rapid response to security incidents.	• Documentation outlining security incident detection tools, and automated reporting systems in place. • Logs from real-time monitoring systems. • Examples of incident management reports. • Records of independent audits of detection and reporting systems in place.

Illustrative question	Scores	Evidence that we may expect
Question 3: How do you ensure that procedures for handling, logging, and analysing failures of the technology are up-to-date and effective against current threats? Provide documented evidence of regular reviews and updates to failure handling processes, failure logging, and post-incident analysis.	**1 point:** Evidence demonstrates that processes exist to log failures, analyse failures for security implications, and review and update failure handling procedures, but they may be insufficiently detailed or conducted irregularly. **5 points:** Comprehensive evidence demonstrates that suitable processes are in place, with detailed documentation, for the logging of failures, post-incident analysis of failures, and regular review and updating of failure handling procedures.	• Documentation outlining procedures for failure handling, logging and post-incident analysis. • Records outlining examples of post-incident updates which have been implemented. • Records of **independent validation** of procedures for handling failures. • Documentation outlining data validation processes, automated analysis and integrity checking tools and real-time monitoring systems. • Records of proactive and corrective measures taken to ensure data integrity.
Question 4: What processes do you have in place for conducting forensic analysis following a security breach? Provide documented evidence of forensic tools in use.	**1 point:** Evidence demonstrates that forensic analysis processes are in place, but tools and techniques may be limited or inconsistently applied. **5 points:** Comprehensive evidence demonstrates that suitable forensic analysis processes are in place, consisting of advanced tools/techniques, and with detailed documentation of case studies.	• Documentation outlining forensic analysis processes, tools and techniques. • Examples of case studies and lessons learned. • Records of **independent validation** of the effectiveness of forensic analysis techniques and processes in use.

Objective 3.4 – Reliable Operation Across Relevant Services, Devices, and System Demands:

The technology operates reliably in tests representative of the services and devices it was designed to operate on, (and, where relevant, when deployed), and varying system capacity demands.

Illustrative question	Scores	Evidence that we may expect
Question 1: What are the results from testing the technology which demonstrate reliable operation across the deployment settings for which the technology has been designed? This may include different services, a range of devices, different operating systems, and hardware configurations. Provide documented evidence of these tests and their outcomes, including any device-specific issues and steps taken to address them.	**1 point**: Evidence demonstrates reliable operation across a limited subset of the services, devices, operating systems, and hardware configurations for which the technology has been designed. **5 points:** Comprehensive evidence demonstrates that suitable testing is conducted across several services, all relevant devices, operating systems, and hardware configurations for which the technology has been designed, with detailed documentation of results and any necessary adjustments. Results of testing demonstrate that the technology operates reliably, maintaining its performance across all relevant deployment settings.	• Documentation outlining testing procedures across relevant services, devices, operating systems and hardware configurations. • Records of testing results which demonstrate reliable operation, with performance maintained across various deployment settings. • Examples/records of optimisations or fixes implemented based on the outcomes of testing.

Illustrative question	Scores	Evidence that we may expect
Question 2: What are the results of tests conducted under varying network conditions, including different internet speeds and latencies, and how are the impacts of varying network conditions handled? Provide documented evidence of performance metrics acquired during such testing which demonstrate reliable operation under varying network conditions.	**1 point**: Evidence demonstrates that testing under a limited range of network conditions is conducted. Results provided show limited evidence of reliable operation under varying network conditions. **5 points:** Comprehensive evidence demonstrates that testing under all relevant network conditions is conducted, with detailed documentation of performance metrics demonstrating reliable operation under these conditions, and proactive adjustments made to ensure reliable operation.	• Documentation outlining processes for testing under varying network conditions. • Records of performance metrics demonstrating reliable operation under varying network conditions, adjustments made based on testing, and regular monitoring reports. • Documentation outlining adaptive mechanisms, including records of implementation and examples of their effectiveness under varying conditions. • Documentation of mechanisms in place for handling varying internet speeds, including records of adjustments made and performance metrics.
Question 3: What capacity planning efforts have been made, including predictions of peak demand and strategies for managing high-load situations? Provide documented evidence of these efforts.	**1 point**: Evidence demonstrates that capacity planning efforts are in place, with limited predictions and strategies. **5 points:** Comprehensive evidence demonstrates that capacity planning efforts are in place, with advanced predictions, detailed documentation, and proactive strategies for managing peak demand and high-load situations.	• Documentation outlining capacity planning efforts. • Records of peak demand predictions. • Documentation outlining strategies for high-load management. • Records of regular monitoring and adjustments made to capacity planning efforts.

Objective 3.5 – Detection and Mitigation of Threats:

Sufficient safeguards and processes are in place to detect and mitigate both intentional and unintentional threats, which may include input manipulation and contextual misunderstandings. The technology can effectively respond to a wide range of adversarial attacks and circumventions of intended use while maintaining its integrity and accuracy.

Illustrative question	Scores	Evidence that we may expect
Question 1: Provide evidence of any risk mitigations relating to circumvention of intended use, including any methods tested to reduce the number of **false positives** (type I error) or **false negatives** (type II error).	**1 point**: Evidence demonstrates that risk mitigation strategies are in place and documented, with limited testing on **false positives/false negatives** conducted. **5 points:** Comprehensive evidence demonstrates that suitable risk mitigation strategies are in place with detailed documentation and regular monitoring. Suitable testing is conducted for **false positives/negatives** with a view to ensuring robustness of the technology.	• Documentation outlining risk mitigation strategies relating to circumvention of intended use. • Records of testing for **false positives/negatives**. • Regular monitoring reports. • Examples of improvements implemented.

Illustrative question	Scores	Evidence that we may expect
Question 2: How do you adapt the technology to new types of perturbations as they emerge? Provide documented evidence of regular improvement processes and updates to robustness strategies.	**1 point**: Evidence demonstrates that adaptation strategies are implemented, but these are limited in scope and/or effectiveness. Minimal updating of robustness measures over time is performed. **5 points:** Comprehensive evidence demonstrates that suitable adaptation strategies are implemented with detailed documentation, regular monitoring, and proactive updates to robustness measures are performed.	• Documentation outlining strategies and procedures in place for adapting to new types of perturbations. • Records of updates made to robustness measures. • Examples of responses to new perturbations (e.g., case studies of updates made in response to an emerging threat).
Question 3: How does the technology handle advanced adversarial attacks, such as evasion attacks, injection attacks and input perturbations? What steps are taken to keep the technology up to date with the latest adversarial threats? Provide documented evidence from simulations that test the technology's ability to resist such adversarial attacks/threats.	**1 point**: Evidence demonstrates that mechanisms are in place for defence against advanced **adversarial attacks**, but with minimal evidence of resistance/effectiveness. **5 points:** Comprehensive evidence demonstrates that suitable mechanisms are in place for defence against advanced **adversarial attacks**, with detailed documentation, regular testing, and proactive measures in place to ensure resilience against sophisticated input alterations.	• Documentation outlining procedures and techniques in place for defence against advanced **adversarial attacks**. • Documentation and records outlining both the procedure, and the results of, advanced **adversarial attack** simulations. • Records of ongoing testing and regular monitoring, showing sustained resistance and performance. • Records outlining research and updates made to the technology, and adaptation strategies in place.

Principle 4: Maintainability

Objective 4.1 – System Risk and Update Management:

Comprehensive procedures and policies are in place for proactive identification and management of system-level risks, with a view to ensuring that the performance of the technology is maintained both over time, and across subsequent updates.

Illustrative question	Scores	Evidence that we may expect
Question 1: What processes are in place to identify and continuously monitor emerging risks related to the technology's lifecycle? Provide documented evidence of risk identification processes, including risk workshops, expert consultations, and threat modelling exercises, documented evidence of risk assessment tools, scoring systems, and the evaluation of the impact on content moderation **accuracy.**	**1 point**: Evidence provided demonstrates risk identification processes and continuous monitoring processes may be partially sufficient with some limitations in scope, effectiveness, or relevance. **5 points:** Comprehensive evidence demonstrates advanced and systematic risk identification processes integrated into the development lifecycle of the technology. Continuous monitoring, regular updates, and proactive mitigation strategies are documented. The process involves multiple stakeholders, including external experts, and is regularly reviewed.	• Risk Identification records. • Documentation of workshops of consultations related to risk identification. • Documented use of risk assessment tools. • Documented use of scoring systems for the purpose of risk identification. • Documentation of expert consultation related to risk identification. • Documentation of threat modelling exercises • Impact evaluation records. • Documentation of stakeholder involvement in risk-identification processes. • Documentation of continuous improvement and integration into the development lifecycle.

Illustrative question	Scores	Evidence that we may expect
Question 2: What contingency plans are in place to address high-impact risks, such as large-scale manipulation attempts or significant moderation failures? Provide documented evidence of contingency protocols, response plans, and scenario analyses.	**1 point**: Evidence provided demonstrates contingency processes may be partially sufficient with some limitations in scope, effectiveness, or relevance. There may be insufficient emphasis on high-impact risks. **5 points:** Comprehensive evidence demonstrates advanced, scenario-specific contingency plans and response plans which are regularly tested and updated. Response strategies are detailed, with clear roles and responsibilities. Continuous scenario analyses are used to refine and improve contingency planning.	• Documentation of contingency protocols. • Documentation of response plans. • Scenario analysis reports. • Documented use of testing and update procedures for contingency protocols.
Question 3: What proactive measures are taken to ensure that updates do not negatively impact the technology's reliability? Provide documented evidence of regression testing and validation processes.	**1 point**: Evidence provided demonstrates proactive measures and continuous monitoring may be partially sufficient with some limitations in scope, effectiveness, or relevance. **5 points:** Comprehensive evidence demonstrates that updates undergo extensive validation, including comprehensive regression testing and impact assessments. Continuous monitoring is in place with a view to ensuring that performance and reliability are maintained post-**deployment** of the update.	• Regression resting reports related to update procedures. • Performance validation reports related to update procedures. • Documentation of impact assessments related to update procedures. • Records or logs related to continuous monitoring.

Illustrative question	Scores	Evidence that we may expect
Question 4: What proactive measures are taken to ensure that changes or updates to the technology do not negatively impact reproducibility? Provide documented evidence of change management processes related to reproducibility.	**1 point:** Evidence provided demonstrates proactive measures and continuous monitoring may be partially sufficient with some limitations in scope, effectiveness, or relevance. There may be insufficient emphasis on reproducibility. **5 points:** Comprehensive evidence provided demonstrates proactive change management processes that thoroughly evaluate the impact of changes or updates on reproducibility, including comprehensive change logs, documentation and validation assessments confirming reproducibility and consistent performance post-update, and **independent verification** of update impacts.	• Change logs. • Documentation of post-update reproducibility assessments. • Post-update validation reports. • Documentation showing **independent verification** of update impacts.
Question 5: What proactive measures are taken to ensure that updates or changes to the technology do not negatively impact its performance on different devices or services?	**1 point**: Evidence provided demonstrates that proactive measures and continuous monitoring may be partially sufficient with some limitations in scope, effectiveness, or relevance. There may be insufficient emphasis on performance across devices. **5 points:** Comprehensive evidence demonstrates systematic evaluation of updates on performance across devices and services and proactive mitigation strategies, including comprehensive documentation of performance evaluations post-update, detailed records of identified performance impacts, proactive mitigation strategies, and evidence of continuous monitoring and adjustment.	• Documentation of performance evaluations post-update across devices and services. • Records of identified performance impacts. • Documentation of proactive mitigation strategies. • Documentation of continuous monitoring and adjustment procedures post-update.

Objective 4.2 – Effective Quality Assurance (QA) Plans and Periodic Monitoring:

Effective Quality Assurance (QA) policies are in place to address organisational risk and procedural oversight, and periodic monitoring procedures are conducted with a view to ensuring the maintenance or improvement of the technology's performance. The processes for development and maintenance of the technology over time have been documented.

Illustrative question	Scores	Evidence that we may expect
Question 1: How is quality assurance integrated into every phase of the development of the technology? Provide documented evidence of QA checkpoints, integration in the moderation algorithm development process, and lifecycle documentation.	**1 point**: Evidence provided demonstrates QA integration and lifecycle documentation may be partially sufficient with some limitations in scope, effectiveness, or relevance. There may be insufficient emphasis on continuous development. **5 points:** Comprehensive evidence demonstrates advanced QA processes seamlessly integrated into all phases of development, with continuous monitoring, real-time feedback loops, and regular development of QA protocols. The organisation demonstrates a strong commitment to maintaining high QA standards throughout the technology's lifecycle.	• Documentation related to QA processes. • Records related to continuous lifecycle monitoring. • Documentation related to real-time feedback loops. • Documentation related to regular development of quality assurance processes and protocols.
Question 2: What metrics are used to measure the effectiveness of QA processes in the technology? Provide documented evidence of QA metrics dashboards, KPIs, and performance reports tracking **false positives/negatives** and bias detection.	**1 point**: Evidence provided demonstrates the effectiveness of QA processes in the technology may be partially sufficient with some limitations in scope, effectiveness, or relevance. **5 points:** Comprehensive evidence demonstrates how advanced, real-time metrics and analytics are used to measure QA effectiveness. A comprehensive set of KPIs are documented and a detailed KPI monitoring system is present. QA metrics are integrated into decision-making processes related to the technology and are continuously reviewed for optimisation.	• Documentation of QA metrics. • Documentation of KPI's. • Documentation of protocols or systems monitoring KPI's. • Documentation of reports, alert systems, or adjustment records of KPI's. • Documentation of real-time performance reports of QA processes. • Records showing continuous development on QA processes. • Documentation showing detailed tracking **false positives/negatives** and bias detection in QA processes.

Illustrative question	Scores	Evidence that we may expect
Question 3: What processes are in place to ensure that QA standards are consistently applied across all teams and projects involved in technology? Provide documented evidence of standardisation protocols, QA training, and compliance checks specific to content moderation.	**1 point**: Evidence provided demonstrates the consistent application and enforcement of QA processes in teams and projects related to the technology may be partially sufficient with some limitations in scope, effectiveness, or relevance. There may be insufficient emphasis on standardisation, consistency, training, or enforcement practices. **5 points:** Comprehensive evidence demonstrates QA standards are rigorously enforced across all teams and projects, with comprehensive standardisation protocols, detailed QA training programs, and systematic compliance checks. Regular audits ensure that QA standards are maintained over time.	• Documentation of QA standardisation protocols related to teams and projects involved in the technology. Documentation of QA training protocols and frameworks related to teams and projects involved in the technology. Documentation of compliance checks for QA protocols related to teams and projects involved in the technology. Records of audit logs for QA protocols related to teams and projects involved in the technology. • Records of enforcement of QA protocols related to teams and projects involved in the technology.
Question 4: Explain the criteria which determine the need for updating or retraining the technology and the processes in place for assessing these criteria. Provide documented evidence of updating/retraining criteria (such as industry advancements in content moderation, demographic shifts in data, new rules or regulations, or revised objectives) and triggers for updating/retraining (such as performance testing, data analysis, research, advisory/legal).	**1 point**: Evidence provided demonstrates the criteria which determine the need for updating or retraining the technology may be partially sufficient with some limitations in scope, effectiveness, or relevance. **5 points:** Comprehensive evidence demonstrates comprehensive criteria for updating/retraining, with real-time evaluation mechanisms and proactive retraining triggers. The updating/retraining process is thoroughly documented and integrated into the technology's lifecycle with a view to ensuring continuous effectiveness.	• Documentation of update/retraining criteria. • Documentation of the update/retraining process. • Documentation showing continuous development of update/retraining criteria. • Real-time evaluation logs. • Records of proactive retraining triggers.

Illustrative question	Scores	Evidence that we may expect
Question 5: What internal governance procedures or policies are in place to guide the development of the technology, and how do they address areas such as developer accountability, approval workflows, and risk management? Please provide supporting documentation, including policy manuals, governance frameworks, compliance guidelines, and reproducibility documents.	**1 point**: Evidence provided demonstrates the internal governance procedures or policies which guide the development of the technology may be partially sufficient with some limitations in scope, effectiveness, or relevance. There may be insufficient emphasis on continuous governance oversight. **5 points:** Comprehensive evidence demonstrates governance policies which are comprehensive, actively enforced, and regularly updated. There are clear chains of responsibility, pre-**deployment** safeguards, documented procedures in place for post-**deployment** reviews, and continuous oversight. Evidence of compliance checks and regular policy updates are provided.	• Documentation of governance frameworks and oversight related to the ongoing development of the technology. • Policy manuals related to the ongoing development of the technology. • Compliance audit reports related to the ongoing development of the technology. • Documentation showing chains of responsibility related to the ongoing development of the technology.
Question 6: What systems are in place to monitor the performance of the technology in real-time? Provide documented evidence of real-time dashboards, monitoring tools, and alert mechanisms tracking accuracy and response times.	**1 point**: Evidence provided demonstrates the systems which monitor the performance of the technology in real-time may be partially sufficient with some limitations in scope, effectiveness, or relevance. There may be insufficient emphasis on real-time or continuous monitoring and comprehensive performance indicators. **5 points:** Comprehensive evidence demonstrates advanced real-time monitoring systems, with integrated dashboards, comprehensive monitoring tools, and proactive alert mechanisms. Continuous performance tracking is in place with a view to ensuring timely and accurate performance of the technology.	• Documentation of real-time dashboard related to performance of the technology. • Documentation of tools or suites used to monitor performance of the technology. • Records of alert mechanisms for key events related to performance of the technology.

Illustrative question	Scores	Evidence that we may expect
Question 7: How do you handle emergency updates or patches? Provide documented evidence of emergency update procedures and rapid **deployment** strategies.	**1 point**: Evidence provided demonstrates documentation of emergency update or patch processes may be partially sufficient with some limitations in scope, effectiveness, or relevance. **5 points:** Comprehensive evidence demonstrates that emergency updates or patches are handled with well-established, rigorous procedures, ensuring rapid **deployment** and minimal impact on system performance. Continuous monitoring and feedback loops ensure that the emergency response is effective and can be refined in real-time.	• Documentation of emergency update procedures. • Documentation of rapid **deployment** strategies and response actions in the event of an emergency. • Real-time monitoring logs.
Question 8: What policies are in place for the documentation of decision-making processes during the technology's development? Provide documented evidence of decision logs and meeting minutes that record critical governance decisions.	**1 point**: Evidence provided demonstrates documentation of decision-making processes during the technology's development may be partially sufficient with some limitations in scope, effectiveness, or relevance. **5 points:** Comprehensive evidence demonstrates advanced documentation policies to ensure that all decision-making processes regarding the development and maintenance of the technology are thoroughly recorded, including rationales for decisions, risk assessments, and stakeholder input. Documentation is accessible, regularly reviewed, and used to inform ongoing governance.	• Formal decision logs related to the technology's development. • Documentation of rationales and accountability related to documented decisions. • Documentation of meeting minutes related to the technology's development. • Documentation of recurring reviews related to the technology's development.

Objective 4.3 – Data Lifecycle and Retention Governance:

Comprehensive procedures and policies are in place to govern the retention, archiving, disposal, and general management of data relating to the operation of the technology (including both data used to develop the technology as well as data about the technology's development) with a view to ensuring operational consistency across the technology's lifecycle.

Illustrative question	Scores	Evidence that we may expect
Question 1: What formal policies govern data retention, archival, and deletion for datasets used in or produced by the system, and how are these policies updated in response to changes in regulation, system design, or operational requirements?	**1 point**: Evidence provided demonstrates formal data policies for data produced by the system or used in its development process may be partially sufficient with some limitations in scope, effectiveness, or relevance. There may be insufficient emphasis on enforcement of policy, coverage of relevant data, or review and update procedures. **5 points:** Comprehensive evidence demonstrates advanced, version-controlled data lifecycle and retention policies which are enforced consistently across the organisation and include regular audits. Policies cover all relevant data types and ensure data is preserved for appropriate durations. There is evidence of policy review timelines, along with formal mechanisms for updating policies in response to changes in regulation, system architecture, or operational requirements. Evidence of stakeholder accountability and ownership, continuous policy monitoring, auditing, and updates are provided.	• Policies related to data retention for data produced by the system or used in its development process. • Policies related to data archival for data produced by the system or used in its development process. • Policies related to data deletion for data produced by the system or used in its development process. • Version-control mechanisms related to data policies. • Documentation related to ownership and accountability roles as it pertains to data policies. • Audit reports related to data policies for data produced by the system or used in its development process. • Records of enforcement and compliance related to data policies. • Documentation related to data policy review timelines. • Documentation related to data policy updates triggered by changes in regulation, system architecture, or operational requirements.

Illustrative question	Scores	Evidence that we may expect
Question 2: What protocols are in place to ensure effective succession planning and knowledge handoff of the technology, particularly concerning content moderation strategies and tools? How do these protocols help maintain governance and project continuity?	**1 point**: Evidence provided demonstrates documentation of succession planning and knowledge handoff may be partially sufficient with some limitations in scope, effectiveness, or relevance. There may be insufficient emphasis on content moderation strategies and tools, or protocols for when a key stakeholder leaves the project. **5 points:** Comprehensive succession planning procedures are integrated into the project lifecycle, with a view to ensuring seamless transitions with no disruption to governance, documentation, or data retention. Contingency plans are regularly updated.	• Documentation related to succession planning. • Documentation related to knowledge handoff. • Documentation related to data retention as it pertains to succession planning and knowledge handoff. • Documentation showing that the technology remains stable during transitionary periods.

Objective 4.4 – Stakeholder Feedback Incorporation:

The technology provider has processes in place to incorporate customer feedback into the ongoing monitoring and evaluation of the technology's performance.

Illustrative question	Scores	Evidence that we may expect
Question 1: What is the ownership structure for the technology, including major stakeholders and decision-makers? Provide documented evidence of the ownership structure, organisational charts, and any governance frameworks in place.	**1 point:** Evidence provided to document the ownership structure including major stakeholders and decision-makers may be partially sufficient with some limitations in scope, effectiveness, or relevance. **5 points:** Comprehensive evidence demonstrates a thoroughly documented ownership and organisational structure with clear delineations of responsibilities, decision-making authority, and stakeholder influence. Regular reviews and updates are conducted.	• Documentation of ownership structure, major stakeholders, and decision-making responsibilities. • Organisational charts. • Records of stakeholder influence. • Records of decision-making reviews.
Question 2: How is feedback from customers incorporated into the monitoring and evaluation processes for the technology? Provide documented evidence of stakeholder engagement strategies, feedback collection tools, and analysis reports.	**1 point**: Evidence provided demonstrates incorporation of stakeholder feedback into the monitoring and evaluation processes of the technology may be partially sufficient with some limitations in scope, effectiveness, or relevance. There may be insufficient emphasis on ongoing and regular engagement. **5 points:** Comprehensive evidence demonstrates stakeholder feedback is proactively and continuously incorporated into the monitoring and evaluation processes. Policies and tools make stakeholder feedback accessible and readily integrated, with advanced engagement strategies and real-time feedback loops with a view to ensuring the technology evolves to meet stakeholder needs.	• Records of stakeholder engagement. • Documentation of feedback or engagement protocols/policies. • Documentation of feedback or engagement tools. • Documentation of feedback incorporation. • Documentation of proactive stakeholder influence on technology development.

Illustrative question	Scores	Evidence that we may expect
Question 3: How do partnerships or collaborations influence the technology's development, performance, and accountability? Provide documented evidence of partnership impact assessments, collaboration reviews, and any adjustments made based on partner contributions.	**1 point:** Evidence provided demonstrates the influence of partnerships or collaborations on the technology's development, performance, and accountability may be partially sufficient with some limitations in scope, effectiveness, or relevance. **5 points:** Comprehensive evidence demonstrates strategically managed partnerships or collaborations, with documented evidence of their significant impact on all aspects of the technology. Regular evaluations and adjustments are made based on comprehensive impact assessments.	• Reports assessing the impact of partnerships and collaborations on all aspects of the technology, including **development**, performance and accountability. Reviews and evaluations of ongoing partnerships and collaboration. • Records of adjustments based on partner contributions.