Translation

The following Chinese draft national standard proposes safety and security rules for the training and fine-tuning data used to develop generative AI models. The standard defines safety and security as including not only protection of people's physical safety and disinformation prevention, but also censorship of content that criticizes Communist Party rule or presents China in an unflattering light. China issued a finalized version of these standards in April 2025, but, as of the publication date of this translation, CSET has not observed a publicly available full-text copy of the final version.

Title
National Standard of the People's Republic of China: Cybersecurity Technology—Safety Specifications for Generative Artificial Intelligence Pre-Training and Fine-Tuning Data (Draft for Feedback)
中华人民共和国国家标准：网络安全技术 生成式人工智能预训练和优化训练数据安全规范（征求意见稿）

Authors
State Administration for Market Regulation (SAMR; 国家市场监督管理总局; 市场监管总局) and Standardization Administration of China (SAC; 国家标准化管理委员会; 国家标准委)

Source
Website of National Technical Committee 260 on Cybersecurity of Standardization Administration of China (SAC/TC260; 全国网络安全标准化技术委员会; 网安标委), April 3, 2024.

The Chinese source text is available online at:
https://www.tc260.org.cn/file/2024-04-01/94e7e6de-2688-472c-af8b-a6cfe7fc7d29.pdf
An archived version of the Chinese source text is available online at: https://perma.cc/YYU8-6DH4

Translation Date	**Translator**	**Editor**
November 7, 2025	Etcetera Language Group, Inc.	Ben Murphy, CSET Translation Manager

National Standard of the People's Republic of China

Cybersecurity Technology – Safety Specifications for Generative Artificial Intelligence Pre-Training and Fine-Tuning Data

(Draft for Feedback)

(Draft Completed on: March 28, 2024)

When submitting feedback, please include relevant patents that you are aware of along with supporting documentation.

State Administration for Market Regulation
Standardization Administration of China

Issuers

Contents

Preface

This document is drafted in accordance with the provisions of GB/T1.1-2020 *Directives for standardization work - Part 1: Rules for the structure and drafting of standardization documents*.

This document is proposed and administered by National Technical Committee 260 on Cybersecurity of Standardization Administration of China (SAC/TC260). Drafting organizations of this document: (to be determined based on actual circumstances)

Drafted by: (to be determined based on actual circumstances)

Cybersecurity Technology – Safety Specifications for Generative Artificial Intelligence Pre-Training and Fine-Tuning Data

1. Scope

This document specifies the safety[1] requirements for generative artificial intelligence (GenAI) pre-training and fine-tuning data and related processing activities, and describes the corresponding evaluation methods.

This document applies to guiding GenAI service providers in carrying out pre-training and fine-tuning data processing activities, as well as in conducting self-evaluations of the safety of pre-training and fine-tuning data. It may also serve as a reference for regulatory assessments.

2. Normative References

The contents of the following documents, through normative references in this text, constitute indispensable provisions of this document. Among them, for dated references, only the edition corresponding to that date applies to this document. For undated references, the latest edition (including all amendments) applies to this document.

GB/T AAAAA Cybersecurity Technology – Generative Artificial Intelligence Data Annotation Safety Specifications[2]

3. Terminology and Definitions

The terms and definitions listed below apply to this document.

3.1 Generative Artificial Intelligence

AI systems with the ability to generate text, graphics, audio, video, and other content.

3.2 GenAI Services

[1] Translator's note: The Chinese word 安全 encompasses the meanings of both "safety" (protection from accidental harm) and "security" (protection from deliberate harm). In this translation, it is variously translated as "safety," "security," "safety and security," or "safety or security" at the translator's discretion.

[2] Translator's note: CSET's English translation of the draft version of the Chinese national standard *Cybersecurity Technology – Generative Artificial Intelligence Data Annotation Safety Specifications* is available online at: https://cset.georgetown.edu/publication/china-gen-ai-data-labeling-safety-standard-draft/. China issued a finalized version of this standard in April 2025, but, as of the publication date of this translation, CSET has not observed a publicly available full-text copy of the final version.

The use of GenAI technology to provide text, graphics, audio, video, and other content generation services.

3.3 Service Provider

An organization or individual that provides GenAI services in the form of interactive interfaces, programmable interfaces, etc.

3.4 Service User

An organization or individual that uses GenAI services.

3.5 Pre-Training

The training process that uses large-scale data to enable a GenAI model to acquire general knowledge.

3.6 Fine-Tuning

The training process that uses domain-specific data to enable a GenAI model to acquire certain domain-oriented service capabilities.

3.7 Pre-Training Data

All types of data used for GenAI pre-training.

3.8 Fine-Tuning Data

All types of data used for GenAI fine-tuning.

4. Overview

4.1 Safety and Security Risks

The safety and security of GenAI pre-training and fine-tuning data involves both the safety and security of the data itself and the safety and security of GenAI services. The safety and security risks faced by GenAI pre-training and fine-tuning data include:

a) Risks such as data leakage and data theft;
b) Risks of data poisoning;
c) Other risks whereby training data may affect the safety or security of GenAI.

4.2 Safety and Security Framework

The safety and security framework for GenAI pre-training and fine-tuning data consists of general data safety and security and the safety and security of data processing activities. General data safety and security mainly includes categorization and grading, safety and security protection, safety and security monitoring, audit and traceability, and emergency response. The safety and security of data processing

activities mainly includes the safety and security of data collection, data pre-processing, and data usage.

The safety and security framework for GenAI pre-training and fine-tuning data is shown in Figure 1.

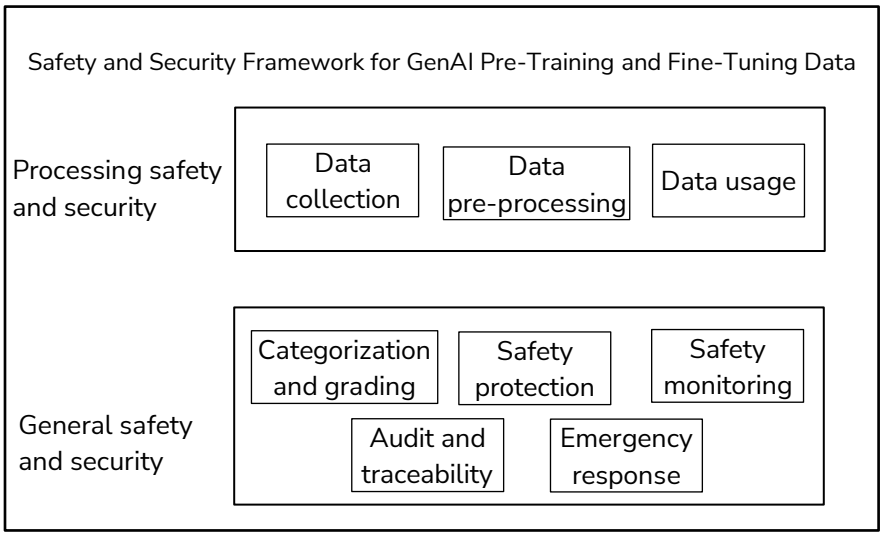

Figure 1 Safety and Security Framework for GenAI Pre-Training and Fine-Tuning Data

5. General Safety and Security Requirements

The requirements for service providers are as follows.

a) Pre-training and fine-tuning data shall be managed through categorization and grading (分类分级).

b) Technical measures shall be adopted to conduct safety and security monitoring of pre-training and fine-tuning data. When data safety and security defects, vulnerabilities, or other risks are discovered, prompt alerts shall be issued and corresponding handling measures taken.

c) Technical measures such as identity authentication, access control, encryption, and backup shall be adopted to provide safety and security protection for pre-training and fine-tuning data.

d) An emergency response mechanism shall be established for safety and security incidents involving pre-training and fine-tuning data, enabling timely and effective handling of data safety and security incidents so that business operations are not affected or can be restored as quickly as possible.

e) Records shall be kept of data processing activities such as data collection, data pre-processing, and data use of pre-training and fine-tuning data, ensuring that key operations in these processing activities are auditable and traceable.

6. Safety and Security Requirements for Pre-Training Data Processing Activities

6.1 Data Collection

The requirements for service providers are as follows.

a) The sources of collected data shall be recorded, and relevant information preserved:
   1) If the data source is an internet website, record the website's Uniform Resource Locator (URL);
   2) If the data source is another organization or individual, record the dataset name and source organization, and preserve legally binding transaction contracts, cooperation agreements, license agreements, or other relevant authorization documents;
   3) If the data source is a service user, record the service name and the service user's identity identification number, and preserve the service user's authorization record.

b) Data of the same type shall have multiple different sources.
   Note: Code, images, audio, video, and text in the same language are considered to be the same type of data.

c) When collecting data through internet websites, the collected data or the URL of the webpage where the data is located shall be recorded.

d) When collecting data through transactions or cooperation with other organizations or individuals, the data, commitments, and materials provided by the transaction or cooperation partners shall be reviewed.

6.2 Data Pre-Processing

The requirements for service providers are as follows.

a) Metadata shall be added to all data samples in the data:
   1) For data samples that already contain source information, the metadata shall be that information;
   2) For data samples sourced from internet websites, the metadata shall be the URL of the sample itself or of the webpage where it is located;
   3) For data samples sourced from datasets of other organizations or individuals, the metadata shall include the dataset name, organization name, etc.;
   4) For data samples sourced from service users, the metadata shall include the service name and the service user's identity identification number.

b) Methods such as keyword searches, classification models, and manual

sampling checks shall be adopted to identify whether data contains safety or security risk content, and the identification results shall be recorded.

Note: Safety or security risk content refers to the 29 categories defined in Appendix A; requirements for keywords and classification models are provided in Appendix B.

c) Major risks of intellectual property rights (IPR) infringement in the data shall be identified and recorded. For example, when data contains literary, artistic, or scientific works, particular attention shall be given to identifying copyright infringement issues.

6.3 Data Usage

The requirements for service providers are as follows.

a) When using data that contains personal information, the consent of the corresponding individual shall be obtained, or the use shall comply with other circumstances as stipulated by laws and administrative regulations.
b) Before using data that contains sensitive personal information, the separate consent of the corresponding individual shall be obtained, or the use shall comply with other circumstances as stipulated by laws and administrative regulations.
c) Data involving IPR infringement shall not be used.
d) Measures shall be taken to reduce the likelihood of GenAI being induced to generate safety or security risk content, including but not limited to adequately filtering out data samples already identified as containing safety or security risk content.

7. Safety and Security Requirements for Fine-Tuning Data Processing Activities

7.1 Data Collection

The requirements for service providers are as follows.

a) The collection of fine-tuning data shall comply with the requirements of 6.1.
b) When collecting data such as GenAI-generated content, the version of the GenAI model or service used, the acquisition time, and other relevant information shall be recorded.

7.2 Data Pre-Processing

The requirements for service providers are as follows.

a) The pre-processing of fine-tuning data shall comply with the requirements of 6.2.
b) For data samples consisting of GenAI-generated content, metadata shall be

added, including the version of the GenAI model or service used and the acquisition time.

c) Data labeling activities for fine-tuning data shall comply with the safety and security requirements of GB/T AAAAA.

d) For data generated by GenAI, particular attention shall be given to identifying whether the data content contains safety or security risk content, and the identification results shall be recorded.

7.3 Data Usage

The requirements for service providers are as follows.

a) The data sources of fine-tuning data shall comply with the requirements of 6.3.

b) When using data such as GenAI-generated content, data that contains safety or security risk content shall be filtered out.

8. Evaluation Methods

8.1 General Safety and Security Evaluation Methods

The evaluation methods, expected results, and result determination for general safety and security requirements are as follows.

a) Evaluation methods:
   1) Inspect the service provider's operational process records and management documents for pre-training and fine-tuning data;
   2) Inspect the design documents and operation logs for the systems and networks where the service provider's pre-training and fine-tuning data are located, and inspect the actual operating status of related equipment;
   3) Inspect the technical measures adopted by the service provider for the safety and security protection of pre-training and fine-tuning data;
   4) Inspect whether the service provider has an emergency response team, whether an emergency response plan has been formulated for safety and security incidents involving pre-training and fine-tuning data, and inspect the records on the emergency handling of safety and security incidents;
   5) Inspect whether the service provider has logs recording the data collection and preparation stage processing activities of pre-training and fine-tuning data, and verify the integrity and validity of such logs.

b) Expected results:
   1) The service provider has carried out categorization and grading operations and management for pre-training and fine-tuning data;

2) The service provider has adopted technical measures to monitor the safety and security of pre-training and fine-tuning data, and upon discovering data safety or security defects, vulnerabilities, or other risks, has issued prompt alerts and taken corresponding handling measures;
3) The service provider has adopted technical measures such as identity authentication, access controls, encryption, and backup to provide safety and security protection for pre-training and fine-tuning data;
4) The service provider has an emergency response team in place, has established an emergency response mechanism for safety and security incidents involving pre-training and fine-tuning data, and has implemented prompt and effective actions when such incidents occurred;
5) The service provider has logs of key activities in the data collection and preparation stages of pre-training and fine-tuning data, and based on these logs, key operations can be audited and traced.

c) Result determination: If the actual evaluation results are consistent with the expected results, the result is compliance; otherwise, the result is non-compliance.

8.2 Evaluation Methods for Pre-Training Data Processing Activities

8.2.1 Data Collection

The evaluation methods, expected results, and result determination for the safety and security requirements of pre-training data collection are as follows.

a) Evaluation methods:
1) Inspect whether the service provider has records of data sources; verify the correctness of the data source record format; sample the data collected by the service provider and verify the completeness of the data source records;
2) Inspect the number of data sources used for the same type of data in the service provider's data source records;
3) Inspect whether the service provider has collected data from internet websites; sample the internet website data collected by the service provider and check the consistency between the sampled data and the recorded URLs;
4) Inspect whether the service provider has collected data through transactions or cooperation with other organizations or individuals; sample and inspect the review materials of the data, commitments, and supporting documents provided by the transaction or cooperation

partners.

b) Expected results:
   1) The service provider has records of data sources; where the data source involves an internet website, URL records are available; where the data source involves other organizations or individuals, records of dataset names and source organizations are available, and transaction contracts, cooperation agreements, license agreements, or other authorization documents are valid; where the data source involves service users, records of service names and service users' identity identification numbers are available, and the authorization records of service users are valid; data source records are complete;
   2) The service provider has multiple data sources for the same type of data;
   3) The service provider has not collected data from internet websites, or all sampled data is consistent with the recorded URLs;
   4) The service provider has not collected data through transactions or cooperation with other organizations or individuals, or has valid review materials for the data, commitments, and supporting documents provided by the transaction or cooperation partners.

c) Result determination: If the actual evaluation results are consistent with the expected results, the result is compliance; otherwise, the result is non-compliance.

8.2.2 Data Pre-Processing

The evaluation methods, expected results, and result determination for the safety and security requirements of pre-training data pre-processing are as follows.

a) Evaluation methods:
   1) Randomly sample no fewer than 100 pre-processed data samples from each type of data source of the service provider, and check the correctness of the metadata content of the samples;
   2) Randomly sample no fewer than 100 pre-processed data samples from the service provider, and check whether the samples contain records of identification of safety or security risk content;
   3) Randomly sample no fewer than 100 pre-processed data samples from the service provider, and check whether the samples contain records of identification of major IPR infringement risks.

b) Expected results:
   1) All sampled data contain metadata; for samples from datasets of other organizations or individuals, records of dataset names and organization

names are present; for samples from internet websites, the URL of the sample itself or the webpage where it is located is present; for samples from service users, records of the service name and the service user's identity identification number are present;

2) All sampled data contain records of safety and security risk content identification;
3) All sampled data that involve IPR infringement risks contain records of IPR infringement risk identification.

c) Result determination: If the actual evaluation results are consistent with the expected results, the result is compliance; otherwise, the result is non-compliance.

8.2.3 Data Usage

The evaluation methods, expected results, and result determination for the safety and security requirements of pre-training data use are as follows.

a) Evaluation methods:
   1) Inspect whether the service provider uses data containing personal information; inspect whether the service provider has records of individual consent, or whether the use complies with circumstances stipulated by laws and administrative regulations;
   2) Inspect whether the service provider uses data containing sensitive personal information; inspect whether the service provider has records of separate individual consent, and whether the use complies with circumstances stipulated by laws and administrative regulations;
   3) Use manual sampling to randomly select no fewer than 4,000 samples from all data, and verify the accuracy of the service provider's records of IPR infringement risk identification;
   4) Use manual sampling to randomly select no fewer than 4,000 samples from all data, and adopt keyword, classification model, or other technical sampling methods to sample no fewer than 10% of the total data.
b) Expected results:
   1) The service provider has not used personal information data, or has records of individual consent, or the use of personal information data complies with circumstances stipulated by laws and administrative regulations;
   2) The service provider has not used sensitive personal information data, or has records of separate individual consent, or the use of sensitive personal information data complies with circumstances stipulated by laws and

administrative regulations;

3) The sampled data do not involve IPR infringement risks, or the sampled data are free of IPR infringement risks and consistent with the IPR infringement risk identification records;
4) Among manually sampled data, the proportion of samples without safety or security risk content shall be no fewer than 96% of the total sampled quantity; among technically sampled data, the proportion of samples without safety or security risk content shall be no fewer than 98% of the total sampled quantity.

c) Result determination: If the actual evaluation results are consistent with the expected results, the result is compliance; otherwise, the result is non-compliance.

8.3 Evaluation Methods for Fine-Tuning Data Processing Activities

8.3.1 Data Collection

The evaluation methods, expected results, and result determination for the safety and security requirements of fine-tuning data collection are as follows.

a) Evaluation methods:
   1) Evaluate the service provider's fine-tuning data collection in accordance with the evaluation methods specified in 8.2.1a);
   2) Inspect whether the service provider collects GenAI-generated content; inspect whether the GenAI-generated content collected by the service provider includes records of the version of the GenAI model or service used, the acquisition time, and other relevant information.
b) Expected results:
   1) The expected results specified in 8.2.1b) are met;
   2) The service provider has not collected GenAI-generated content, or, if such content has been collected, records of the version of the GenAI model or service used and the acquisition time are present.
c) Result determination: If the actual evaluation results are consistent with the expected results, the result is compliance; otherwise, the result is non-compliance.

8.3.2 Data Pre-Processing

The evaluation methods, expected results, and result determination for the safety and security requirements of fine-tuning data pre-processing are as follows.

a) Evaluation methods:

1) Evaluate the service provider's fine-tuning data pre-processing in accordance with the evaluation methods specified in 8.2.2a);
2) Randomly sample no fewer than 100 pre-processed fine-tuning data samples from the service provider, and check the correctness of the metadata content of the samples;
3) Inspect whether the labeled fine-tuning data complies with the safety and security requirements of GB/T AAAAA;
4) Randomly sample no fewer than 100 pre-processed fine-tuning data samples from the service provider, and inspect whether the samples contain records of identification of safety or security risk content in GenAI-generated data.

b) Expected results:
1) The expected results specified in 8.2.2b) are met;
2) For sampled data involving GenAI-generated content, the metadata includes the version of the GenAI model or service used and the acquisition time;
3) Labeled fine-tuning data complies with the safety and security requirements of GB/T AAAAA;
4) For sampled data consisting of GenAI-generated content, records of identification of safety or security risk content in GenAI-generated data are present.

c) Result determination: If the actual evaluation results are consistent with the expected results, the result is compliance; otherwise, the result is non-compliance.

8.3.3 Data Usage

The evaluation methods, expected results, and result determination for the safety and security requirements of fine-tuning data use are as follows.

a) Evaluation methods:
1) Evaluate the service provider's fine-tuning data use in accordance with the evaluation methods specified in 8.2.3a);
2) Use manual sampling to randomly select no fewer than 4,000 samples from all data, and verify the accuracy of the service provider's records of identification of safety or security risk content in GenAI-generated data.

b) Expected results:
1) The expected results specified in 8.2.3b) are met;
2) The sampled data do not involve GenAI-generated content, or the

sampled data are consistent with the records of identification of safety or security risk content in GenAI-generated data.

c) Result determination: If the actual evaluation results are consistent with the expected results, the result is compliance; otherwise, the result is non-compliance.

Appendix A
(for reference)
Main Safety and Security Risks of Pre-Training and Fine-Tuning Data

A.1 Contains Content that Violates the Socialist Core Values Concept

Contains the following content:

a) Incitement to subvert state power and overthrow the socialist system;
b) That which endangers national security and interests, and harms the image of the state;
c) Incitement of separatism, or undermining national unity and social stability;
d) Promotion of terrorism or extremism;
e) Promotion of ethnic hatred (民族仇恨);
f) Promotion of violence or obscenity and pornography;
g) Dissemination of false and harmful information;
h) Other content prohibited by laws and administrative regulations.

A.2 Contains Discriminatory Content

Contains the following content:

a) Ethnic (民族) discrimination;
b) Discrimination on the basis of beliefs;
c) Nationality-based (国别) discrimination;
d) Discrimination on the basis of regional origin;
e) Gender discrimination;
f) Age discrimination;
g) Occupation-based discrimination;
h) Health-based discrimination;
i) Other types of discriminatory content.

A.3 Commercial Violations

The main risks include:

a) Infringement of IPR of others;
b) Violation of business ethics;
c) Disclosure of the trade secrets of others;
d) Use of algorithms, data, platforms, etc. to engage in monopolistic or unfair competition behaviors;

e) Other commercial violations.

A.4 Violations of the Legitimate Rights and Interests of Others

The main risks include:

a) Endangerment of the physical or mental health of another;
b) Unauthorized use of the likeness of another;
c) Defamation of the reputation of another;
d) Defamation of the honor of another;
e) Infringement of others' right to privacy;
f) Infringement of others' personal information rights and interests;
g) Infringement of other legitimate rights and interests of others.

Appendix B
(Normative)
Requirements for Keyword Library and Classification Models

B.1 Keyword Library

Requirements are as follows.

a) The keyword library shall be comprehensive, and the total size should not be less than 10,000.

b) The keyword library shall be representative and cover, at a minimum, the 17 safety and security risks in Appendices A.1 and A.2 of this document. There shall be no fewer than 200 keywords for each safety or security risk in Appendix A.1, and there shall be no fewer than 100 keywords for each safety or security risk in Appendix A.2.

c) The keyword library shall be updated in a timely manner in accordance with actual cybersecurity requirements, and should be updated at least once per week.

B.2 Classification Models

Classification models shall provide complete coverage of all 29 safety and security risks in Appendix A of this document.

References

[1] TC260-PG-20233A Guidelines for Cybersecurity Standards in Practice—Methods for Labeling Generative Artificial Intelligence Service Content

[2] TC260-003 Basic Safety Requirements for GenAI Services