Public Summary of Training Content for GPT-5.5

Version of the Summary: v1
Last update: 26 June 2026

1. General information

1.1. Provider identification

Provider name and contact details:	OpenAI Ireland Ltd, 1st Floor, The Liffey Trust Centre, 117-126 Sheriff Street Upper, Dublin 1, D01 YC43, Ireland.
Authorised representative name and contact details:	Not applicable (provider established in the Union)

1.2. Model identification

Versioned model name(s):	GPT-5.5
Model dependencies:	None
Date of placement of the model on the Union market:	23 April 2026

1.3 Modalities, overall training data size and other characteristics

Modality *Select the modalities present in the training data, to the extent that they are identifiable*	**Training data size** *For each selected modality, select the range within which the estimated total training data size for that modality falls. Dynamic datasets may be excluded from the estimation.*	**Types of content** *For each selected modality, provide a general description of the type of content that has been included in the training data.*
☒ Text	☐ Less than 1 billion tokens ☐ 1 billion to 10 trillion tokens ☒ More than 10 trillion tokens	GPT-5.5 was trained on a large-scale, multilingual mixture of publicly available data, data accessed through partnerships, synthetic data, and human-generated text, including general web content, reference materials, technical documentation, source code, and other text, curated and filtered for quality and safety.
☒ Image	☐ Less than 1 million images ☐ 1 million to 1 billion images ☒ More than 1 billion images	GPT-5.5 was trained on a broad corpus of images from publicly available sources, images accessed through partnerships, and synthetic images, spanning natural scenes, objects, people, and diagrams, with filtering to support quality, safety, and multimodal understanding.

☒ Audio	☐ Less than 10 000 hours ☐ 10 000 to 1 million hours ☒ More than 1 million hours	GPT-5.5 was trained on a diverse set of publicly available audio data, including conversational, narrated, instructional, and informational material, represented in part through transcripts, and curated and filtered for quality and safety.
☒ Video	☐ Less than 10 000 hours ☐ 10 000 to 1 million hours ☒ More than 1 million hours	GPT-5.5 was trained on publicly available multi-domain audiovisual content, including spoken, narrated, instructional, and general-interest audiovisual material, represented in part through transcripts, captions, or metadata, and filtered for quality and safety.
☐ Other	*Specify the modality and for each one indicate approximate size and unit of measurement*	N/A

Latest date of data acquisition/collection for model training:	The data used to train GPT-5.5 includes different datasets from varying time periods, with some data collected no later than February 2026.
Description of the linguistic characteristics of the overall training data:	Multilingual, with strong English coverage and substantial representation across EU official languages and other languages from around the world.
Other relevant characteristics of the overall training data:	The overall training corpus is designed for a text-output model with multimodal understanding capabilities. It includes written materials, text-bearing images, and audiovisual content used through captions, transcripts, metadata, or related text representations. The corpus aims to provide broad topical, geographic, linguistic, and format coverage.
Additional comments (optional):	N/A

2. List of data sources

2.1. Publicly available datasets

Have you used publicly available datasets to train the model?	☒ Yes ☐ No
If yes, specify the modality(ies) of the content covered by the datasets concerned:	☒ Text ☒ Image ☒ Video ☒ Audio ☐ Other
List of large publicly available datasets:	The training data for GPT-5.5 includes text from Common Crawl.

General description of other publicly available datasets not listed above:	Other publicly available datasets include broad, multi-domain text and image datasets made available by third parties through public repositories, online platforms, and specialized websites, including reference materials, scientific and technical content, source code, image-text datasets, and speech or audiovisual datasets distributed with captions, transcripts, metadata, or related text. These datasets are global in scope, multilingual, and subject to preprocessing such as quality filtering, deduplication, and safety filtering before training. We use advanced data filtering processes to reduce personal information from training data.
Additional comments (optional):	From public web datasets, OpenAI takes steps to identify and apply relevant rights-reservation and opt-out signals, including robots.txt signals for GPTBot, where those signals are available for domains listed in the datasets. OpenAI also uses the the U.S. Trade Representative (USTR) Notorious Markets for Counterfeiting and Piracy list as a signal when deciding to exclude data from certain websites that have been recognized as persistently and repeatedly infringing copyright.

2.2 Private non-publicly available datasets obtained from third parties

2.2.1. Datasets commercially licensed by rightsholders or their representatives

Have you concluded transactional commercial licensing agreement(s) with rightsholder(s) or with their representatives?	☐ Yes ☐ No ☒ Other (see below)
If yes, specify the modality(ies) of the content covered by the datasets concerned:	☒ Text ☒ Image ☒ Video ☒ Audio ☐ Other
Additional comments (optional):	OpenAI enters into broad partnerships with third parties that may include, among other initiatives, rights to display partner content to users in our products and/or access to non-publicly available content, such as archives and metadata. OpenAI does not pursue partnerships solely for access to publicly available data.

2.2.2. Private datasets obtained from other third parties

Have you obtained private datasets from third parties that are not licensed as described in Section 2.2.1, such as data obtained from providers of private databases, or data intermediaries?	☒ Yes ☐ No

If yes, specify the modality(ies) of the content covered by the datasets concerned:	Text, image, video, and audio
If publicly known, list private datasets obtained from other third parties:	N/A
General description of non-publicly known private datasets obtained from third parties	OpenAI partnered with third parties to access data spanning a diverse set of domains and contexts to improve GPT-5.5. Data is accessed in compliance with applicable laws.
Additional comments (optional):	N/A

2.3 Data crawled and scraped from online sources

Were crawlers used by the provider or on behalf of?	☒ Yes ☐ No
If yes, specify crawler name(s)/identifier(s):	GPTBot
Purposes of the crawler(s):	GPTBot is used to crawl content that may be used to train OpenAI's generative AI foundation models and make them more useful and safe. OpenAI publishes additional information about its crawlers, their behavior, user-agent identifiers, and IP addresses at https://developers.openai.com/api/docs/bots.
General description of crawler behaviour:	OpenAI's crawler is designed to respect robots.txt instructions for the GPTBot user-agent, including instructions indicating that crawled content should not be used to train OpenAI's generative AI foundation models. OpenAI's crawler is not designed to circumvent captchas or paywalls or to access password-protected content. OpenAI also filtered the training data for GPT-5.5 for domains that have been recognized as persistently and repeatedly infringing copyright, using the U.S. Trade Representative (USTR) Notorious Markets for Counterfeiting and Piracy list as a signal.
Period of data collection:	Approximately 2018 – December 2025
Comprehensive description of the type of content and online sources crawled:	Crawled content includes a broad range of publicly accessible online material, including reference, educational, scientific, technical, government and institutional, and general-interest content. Crawled sources include text, images, and associated metadata such as captions, alt text, transcripts, or other descriptive text, and were filtered for quality and safety before training.
Type of modality covered:	☒ Text ☒ Image ☒ Video ☒ Audio ☐ Other

Summary of the most relevant domain names crawled:	The most relevant crawled source domains include academic, research, patent, and other technical repositories, legal and government resources, document-hosting and sharing services, community and general-interest sites, and region-specific portals.
Additional comments (optional):	N/A

2.4 User data

Was data from user interactions with the AI model (e.g. user input and prompts) used to train the model?	☐ Yes ☒ No
Was data collected from user interactions with the provider's other services or products used to train the model?	☒ Yes ☐ No
If yes, provide a general description of the provider's services or products that were used to collect the user data:	Subject to privacy settings, controls, user requests and opt-outs, and our policies, for individuals using products such as ChatGPT, and Codex, OpenAI may use interactions to train our models, including GPT-5.5. More information on OpenAI's policies is located here: https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance. Additionally, OpenAI's Privacy Portal, which contains information regarding OpenAI's privacy policies, is located here: https://privacy.openai.com/policies.
Type of modality covered:	☒ Text ☒ Image ☒ Video ☒ Audio ☐ Other
Additional comments (optional):	We use advanced data filtering processes to reduce personal information from training data. We also use advanced data processes to reduce the amount of personal data in our training data. Relevant public information on user controls is available here: https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance

2.5 Synthetic data

Was synthetic AI-generated data created by the provider or on their behalf to train the model?	☒ Yes ☐ No
If yes, modality of the synthetic data:	Text, image, video, and audio
If yes, specify the general-purpose AI model(s) used to generate the synthetic data if available on the market:	OpenAI generated synthetic data using its own general-purpose AI models, including GPT-5.4.

Information about other AI models, including provider's own AI model(s) not available on the market, used to generate synthetic data to train the model to which this Summary applies:	OpenAI uses specialized internal and third-party models to generate synthetic data for targeted training objectives, including augmenting data in domains, languages, tasks, or formats where specialized training data is comparatively scarce. These models may be used to generate examples for instruction following, reasoning, coding, multimodal understanding, safety, and evaluation.
Additional comments (optional):	N/A

2.6 Other sources of data

Have data sources other than those described in Sections 2.1 to 2.5 been used to train the model?	☒ Yes ☐ No
If yes, provide a narrative description of these data sources and the data:	OpenAI and our vendors create data to help our models improve on a wide variety of tasks. For example, we worked with experienced professionals to create data representing real-world knowledge work to improve how our models assist with those tasks.
Additional comments (optional):	N/A

3. Data processing aspects

3.1. Respect of reservation of rights from text and data mining exception or limitation

Are you a Signatory to the Code of Practice for general-purpose AI models that includes commitments to respect reservations of rights from the TDM exception or limitation?	☒ Yes ☐ No
Describe the measures implemented before model training to respect reservations of rights from the TDM exception or limitation before and during data collection, including the opt-out protocols and solutions honoured by the provider or, as applicable, by third parties from which datasets have been obtained:	OpenAI implements measures to respect rights reservations and opt-out signals relevant to text and data mining. For web data used for training, OpenAI's crawler is designed to respect robots.txt instructions for the GPTBot user-agent, including instructions indicating that crawled content should not be used to train OpenAI's generative foundation models. OpenAI's crawler is not designed to circumvent captchas or paywalls or to access password-protected content.
Additional comments (optional):	N/A

3.2 Removal of illegal content

General description of measures taken:	OpenAI applies preprocessing and screening measures intended to avoid or remove illegal content under Union law from training data. These may include automated filtering, keyword-based rules, hash-matching, and model-based classifiers to help identify and exclude unlawful material.

3.3. Other information (optional)

Other relevant information about data processing (optional):	N/A