Google

Public Summary of Training Content for General-Purpose AI models

Version of the Summary: v1
Last update: 31 July 2026

Google

1. General information

1.1. Provider identification	
Provider name and contact details:	Google Ireland Limited Gordon House Barrow Street Dublin 4 Ireland
Authorised representative name and contact details:	
1.2. Model identification	
Versioned model name(s):	The Gemma 4 family of models
Model dependencies:	Gemma 4 is not a modification or a fine-tune of a prior model. Each subsequent model in the Gemma 4 family is based on Gemma 4 (see each model card for individual model details). The Gemma 4 family includes models such as: Gemma 4 E2B, Gemma 4 E4B, Gemma 4 12B, Gemma 4 26B A4B, Gemma 4 31B.
Date of placement of the model on the Union market:	April 2026

1.3 Modalities, overall training data size and other characteristics

Modality	Training data size	Types of content
☑ **Text**	☐ Less than 1 billion tokens ☐ 1 billion to 10 trillion tokens ☑ More than 10 trillion tokens	The model trains on datasets containing information that can be represented as written language, including software code and publicly available web documents across educational, government, legal, and research sectors.

☑ **Image**	☐ Less than 1 million images ☐ 1 Million to 1 billion images ☑ More than 1 billion images	The model trains on datasets representing static visual information, including photography, diagrams, and illustrations.
☑ **Audio**	☐ Less than 10 000 hours ☐ 10 000 to 1 million hours ☑ More than 1 million hours	The model trains on datasets representing sound that has been recorded and digitized, including speech recordings and sound effects.
☑ **Video**	☐ Less than 10 000 hours ☐ 10 000 to 1 million hours ☑ More than 1 million hours	The model trains on datasets containing a sequence of image frames that may be accompanied by audio, including performances, video clips, and video effects.
☐ **Other**		
Latest date of data acquisition/collection for model training:	The knowledge cut-off date for Gemma 4 is January 2025. We may periodically update models after this date.	
Description of the linguistic characteristics of the overall training data:	Our models are trained to work with a breadth of languages, including EU official languages, representative of the internet in general.	
Other relevant characteristics of the overall training data:	The pre-training dataset was a large-scale, diverse collection of data encompassing a wide range of domains and modalities, which included publicly-available web-documents, code, images, audio (including speech and other audio types), and video. The post-training dataset included different types of instruction tuning data, reinforcement learning data, and human-preference data.	
Additional comments (optional):		

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:	Our publicly available datasets include data across various sectors such as educational, government, legal, and research sectors comprising a wide variety of media types and languages.			
General description of other publicly available datasets not listed above:	See the description of linguistic characteristics and other relevant characteristics of overall training data in Section 1.			
Additional comments (optional):				
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
If yes, specify the modality(ies) of the content covered by the datasets concerned:	☑ Text	☑ Image	☑ Video	☑ Audio
	☐ Other:			
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	☑ Audio
	☐ Other:			
If publicly known, list private datasets obtained from other third parties:	None of the private datasets subject to section 2.2.2 are publicly known.			

General description of non-publicly known private datasets obtained from third parties	Datasets cover subject areas including educational materials, scientific reasoning, and mathematics.			
Additional comments (optional):				
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):	See list of crawlers here.			
Purposes of the crawler(s):	Google uses crawlers to discover and scan websites, find information for building Google's search indexes, perform other product specific crawls, and for analysis.			
General description of crawler behaviour:	Google's crawlers are used to to discover and scan websites, find information for building Google's search indexes, perform other product specific crawls, and for analysis. For more information on crawler behaviour, see here.			
Period of data collection:	The knowledge cut-off date for Gemma 4 is January 2025. We may periodically update models after this date.			
Comprehensive description of the type of content and online sources crawled:	Crawled data includes a broad range of publicly available online material, such as publicly available websites across educational, government, legal, and research sectors. This includes a wide variety of media types and languages.			
Type of modality covered:	☑ Text	☑ Image	☑ Video	☑ Audio
	☐ Other:			
Summary of the most relevant domain names crawled:	The most relevant domains crawled include publicly available websites across educational, government, legal, and research sectors comprising a wide variety of media types and languages. Google uses crawlers to discover and scan websites, find information for building Google's search indexes, perform other product specific crawls, and for analysis. See list of our common crawlers here. Google's common crawlers obey robots.txt rules when crawling automatically. To learn more about Google's crawlers, visit this site.			
Additional comments (optional):				
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:	In accordance with Google's relevant terms of service, privacy policy, and service-specific policies, and pursuant to user consent where legally required, we use data collected from users of Google products and services—such as user chats with Gemini Apps—to train AI models.			
Type of modality covered:	☑ Text	☑ Image	☑ Video	☑ Audio
	☐ Other:			
Additional comments (optional):				
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	☑ Audio
	☐ Other:			
If yes, specify the general-purpose AI model(s) used to generate the synthetic data if available on the market:	We have relied on the externally available Gemini 3 family of models to generate synthetic data.			
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:	We may use fine-tuned versions of internal models and non-GPAI models to generate synthetic data for training.			

Additional comments (optional):	

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:	This category reflects types of data that do not fit within the categories described above. It includes datasets that Google acquires or generates in the course of its business operations, or directly from its workforce.	
Additional comments (optional):		

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:	Data filtering and preprocessing included techniques such as deduplication, honoring robots.txt, safety filtering in-line with Google's commitment to advancing AI safely and responsibly, and quality filtering to mitigate risks and improve training data reliability. Specifically, our Google-Extended control lets web publishers manage whether content Google crawls from their sites may be used for training Gemini models that power Gemini Apps and Gemini Enterprise Agent Platform, for grounding in Gemini Apps, and for the Grounding with Google Search feature on Gemini Enterprise Agent Platform. The data filtering and preprocessing process may also involve filtering irrelevant or harmful content, text, and other modalities, including filtering content that is pornographic, violent, or violative of child sexual abuse material (CSAM) laws.	
Additional comments (optional):		
3.2 Removal of illegal content		

General description of measures taken:	As described above, data filtering and preprocessing included techniques such as safety filtering in-line with Google's commitment to advancing AI safely and responsibly. This process may involve filtering irrelevant or harmful content, text, and other modalities, including filtering content that is pornographic, violent, or violative of child sexual abuse material (CSAM) laws.
3.3. Other information (optional)	
Other relevant information about data processing (optional):	