Meta Responses To Questions on Notice from Senator Tony Sheldon (received 27 September 2024)

(1) At the hearing, Ms Claybaugh acknowledged that Meta has used 'public' user content uploaded as far back as 2007 to train some of its AI products and services. Please provide the full list of AI models and other Meta products/services that have been trained using this content. (2) At the hearing Senators discussed with Meta the use of content from Facebook and Instagram to train LLaMa and other Meta AI products and services. Please advise which of the following products Meta has also used user content from to conduct any training of AI models. For each product advise the type of data that has been used, and what it has been used for (answering separately for each). (a) Messenger (b) WhatsApp (c) Horizon OS/Meta Quest (3) Ms Claybaugh said at the hearing that only content users have set to 'public' has been used for the purpose of training AI. Noting that Meta's platforms have more privacy settings than a binary public/private choice, please advise whether content with the following settings has been used for this purpose (answering separately for each): (a) 'Public' (b) 'Friends' (c) 'Friends except…' (d) 'Specific friends' (e) 'Only me' (d) 'Custom' (e) 'Close friends' (4) Does Meta believe someone uploading a photo on Facebook on 2007 knowingly consented to it being used to train AI products and services that would not exist for another 15 years? (5) At the hearing Senator Darmanin asked about whether people knowingly consented, to which Ms Claybaugh answered "I can't speak to what people did or did not know." (6) What evidence is there to suggest someone uploading a photo in 2007 did know that their content would be used to train a product that was not released until February 2023? (7) Does the content used to train Meta's AI products and services include content posted by businesses? For example, images posted on Facebook or Instagram by an artist or illustrator or tattoo artist of their work? (8) In response to questions taken on notice during the hearing, Meta said that "For Facebook profiles, for example, the default privacy setting is private and so this means that people must choose to share publicly.' Please advise, answering separately for each: (a) Has it always been private by default or has this changed at some point? (i) If yes, advise when this change was made, and why. (b) Does this private-by-default setting only apply to the profiles of individual users, or also to users who are businesses? (i) If it does not apply to

users who are businesses, advise what default privacy setting applies to their posts (ii) If it does apply to users who are businesses, has this always been the case? And if not, when was this change made, and why? (9) In response to questions on notice, Meta said that "For Instagram, audience settings are controlled at an account level. As at 13 July 2024, approximately 35% of Australian Instagram users have private accounts and the remainder have public accounts. Every new user in Australia with a stated age of under 16 years old is defaulted into a private account when they register for Instagram." Please advise, answering separately for each: (a) Noting that only 35% of Australian Instagram users have a private account, when setting up an Instagram account, is the privacy setting private or public by default? (b) If it is now private by default (whether for under 16s or generally) has this always been the case, or has this changed at some point? (i) If yes, advise when this change was made, and why. (c) The quote appears to suggest that users 16 or over are defaulted into public accounts, including users aged 16 or 17. Does this mean their content has been used to train Meta's AI models? (10) At the hearing Ms Claybaugh said that "we're not using photos that are posted from the accounts of under-18-year-olds.' Senator Darmanin asked how this works for users who began using the platform as children, and have since turned 18, which Ms Claybaugh was unsure about. (11) In the answers later provided on notice, Meta says that "the default setting for Facebook profiles and for under 18 year olds is private and so they would need to take a proactive step to change this setting." Please advise (answering separately for each): (a) Has the content posted by under 18 year olds always been private by default or has this changed at some point? (i) If yes advise when this change was made, and why. (a) Are users under 18 still able to manually change the privacy settings of their content to public? (i) If not, advise if this was always the case, or if a change was made, when it was made and why. (b) Does this mean that a user who is under 18, who has taken that proactive step to share their content with a privacy setting of 'public', will have their content used to train AI models? (c) If the answers to any of these questions differ from Instagram compared to Facebook, please advise (12) At the hearing Ms Claybaugh said: ***"I think what's important to note here—it's something we haven't talked about yet—is that people can take steps today. Let's say that they had marked some posts as public in the past and those posts have been used to train the model. They can go back and retroactively say, 'I don't want these posts to be used in the future for training the next generation of the model.' They can do that in bulk and make it very easy to go back and say, 'I don't want this to be used in this way in the future.'"*** **Please advise (answering separately for each):**

(a) How does this process work? (b) How are users informed about this process? (c) Is there a process by which users can have their content retroactively removed from the AI dataset or will this only apply to future models?

To assist the Committee in its work, we are combining our answers to these questions because they all broadly relate to the use of training data for artificial intelligence (AI) models and systems. This also serves to ensure we sufficiently address important clarifying foundational issues holistically, as appropriate.

First, it is important to note that the use of AI is both longstanding and widespread across industries, just as it is across Meta's platform and services.

Indeed, as previously noted, Meta has been using data for AI, especially in the form of automated systems, since at least 2007. Some specific early examples include use of automated systems to:

- **Combat harmful content and behaviour:** Billions of people around the world use Meta's services every day. Hence, detecting and combating harmful content and behaviour at scale is a significant challenge. AI technology provides opportunities to detect and remove harmful content before people see it.

  Meta has been using automated systems to enforce our content policies and power our integrity systems for well over a decade, ensuring their applications are scaled and consistent. In the last five or so years, we have had a strong focus on using AI to help enforce our Community Standards, which are the rules that set out what people can or cannot do on Facebook and Instagram. Our ability to use AI to detect and action harmful content proactively has been improving over time.

  For example, in response to a notice issued to Meta by the Australian eSafety Commissioner in relation to the Basic Online Safety Expectations (BOSE) in 2022, Meta outlined a number of the tools it uses to proactively identify and remove known child sexual exploitation material on its services, all of which are underpinned by automated systems. In eSafety's summary report, it stated that: "These tools play a vital role preventing the ongoing re-victimisation of children and adults, whose images otherwise circulate endlessly online. These tools protect both the privacy of the general population given their high accuracy rates, and that of victims." This statement highlights the critical role that automated systems play in maintaining a safe online environment.

We provide transparency about actions taken with respect to a person's content via direct notifications to them, and also about our proactive detection rate with respect to our policies in our quarterly Community Standards Enforcement Report.

Our work to combat hate speech online provides an instructive case study. Hate speech is traditionally one of the most challenging types of online content to proactively detect because it is so context-dependent. Five years ago, the volume of hate speech we removed was lower than other categories of harmful content, which meant a high degree of human reporting, review and assessment was needed. When we first started releasing our transparency report in 2017, we removed 1.8 million pieces of hate speech globally, 25 percent of which was detected proactively via AI. Since then, after very significant investments in AI, our proactive detection of hate speech has increased significantly. In Q2, 2024, we removed 7.2 million pieces of hate speech on Facebook and 7.7 million pieces of hate speech on Instagram, of which 96.2% and 98.4% respectively was detected proactively via AI.

We have also significantly cut the prevalence of hate speech content within the last few years (for example, from 0.10 to 0.11 per cent in Q3 2020, down to about 0.02% in Q2, 2024, on Facebook). Prevalence measures the number of views of violating content, divided by the estimated number of total content views on Facebook or Instagram.

In this traditional application of AI, data has long been "used" to identify patterns to ensure effective enforcement of our content policies.

- **Provide more personalised online experiences:** There is a surplus of information and content online. Consequently, it can be a major challenge for individuals to easily find the people, information and experiences that are useful, meaningful and enjoyable for them.

  For services like Facebook and Instagram, personalisation is at the heart of the experience. People use our services to connect with family and friends they know, to find communities that they would like to be a part of, and to pursue their interests. We are transparent about how we use AI to make recommendations for people or content that our users may want to engage with.

  One of the ways that people connect with friends, family and other accounts that they follow is via a "Feed."

  Historically, these feeds showed content in chronological order. However, as more people started using our services, more content was shared and it was impossible for people to see all of the content that was shared, much less the content that they cared about. Instagram, for example, launched in 2010 with a chronological feed but by 2016, people were missing 70 per cent of all their posts in feed, including almost half of posts from their close connections. So we developed and introduced a feed that ranked posts based on what people cared about most.

We provide this personalised experience via AI. Our ranking algorithms use thousands of signals to rank posts for each person's Feed with this goal in mind. As a result, each person's Feed is highly personalised and specific to them. Our ranking system personalises the content for over a billion people and aims to show each of them content we hope is most valuable to them, every time they come to Facebook or Instagram.

We provide transparency about our approach in our Transparency Center, and at the system and user levels. At the system level, we publish a range of AI System Cards and at the user level, we provide tools that allow people to easily see why they are seeing a particular post or ad.

However, AI does not just bring benefits in terms of convenience, ease or helping people discover new online content; it also brings significant economic benefits. Many Australian businesses, especially small businesses, benefit from using personalised advertising because it is more efficient and allows them to better reach the right consumer for their business and compete with larger established businesses.

What is newer - and the focus of more attention - is the rise of generative AI, an end-to-end approach to automatic learning. It is not entirely novel: At the end of November 2023, our Fundamental AI Research (FAIR) Lab celebrated its 10 year anniversary. Our investments in talent and infrastructure have enabled our teams to achieve significant breakthroughs in AI, which we have made available to the community to build on and which we leverage in our AI products to enhance real-life, everyday moments. But generative AI only has been the subject of significant public debate and attention since the release of OpenAI's ChatGPT.

Today, Meta is embracing and driving forward generative AI by developing and open sourcing its own generative AI foundational models, such as its large language model (LLM) Llama suite, which powers over 60,000 diverse derivative models built by startups, academics, and researchers; introducing proprietary user-facing applications such as Meta AI; and deploying more developed internal systems and services, such as improved integrity classifiers.

Training generative AI at Meta
Training an AI model such as an LLM involves deriving patterns and relationships from a very large and diverse set of existing data. AI models are not databases or static collections of information.

This is key to understanding our approach to training data in the generative AI context. Our Generative AI Privacy Guide and other transparency resources outline how we have built our AI models, how our AI features work, and what controls and data privacy rights users have. We offer generative AI features that people can use to do things like get responses to questions and create images. The information we use to develop and improve AI at Meta comes from a variety of places, including:

- Public content: We use public posts and comments on Facebook and Instagram to train generative AI models for these features and for the open source community. We don't use posts or comments with an audience other than Public for these purposes.
- Users' interactions with AI features: users' interactions with AI features can be used to train AI models. Examples include messages to AI chats, questions they ask and images they ask Meta AI to imagine.
- Other types of information: additional examples of sources of information we use to develop and improve AI at Meta include public information (for example, like public blog posts or data bases); information that comes from selected partners, like search engines; and licensed information (that is, content that someone else, like an author or publisher, holds the rights to and gives Meta permission to use).

Meta applies extensive privacy considerations, such as pre- and post-training protections, when using this first and third party data.

Transparency
Meta explains how it uses people's information to train Meta's AI models across a number of touchpoints - including in policies and terms, Help Center and Privacy Center articles and in-app.

- Meta Privacy Policy: The Meta Privacy Policy has at all relevant times (including, for example, in 2007) made it clear to users that the information they share with Meta will be used to deliver and improve Meta's products and services. This includes the AI technologies that underpin these products and services. The Privacy Policy has also included specific disclosure about the broad availability and use of content which users choose to share publicly on Meta's platforms.

- System cards and Newsroom posts: In early 2022, Meta launched the Instagram Feed Ranking System Card, and since June 2023, Meta has launched more than 30 AI system cards that explain how AI powers experiences on Meta's products. For example, the Facebook Feed AI system talks about the inventory of posts, Pages, etc., and how the AI considers content to make predictions and ranking.

In September 2023, Meta released systems cards for generative AI systems that generate text, and those that generate images. This was followed up with a further system card published in December 2023 relating to generative AI systems that are multimodal.

Also in September 2023, we announced the launch of new generative AI features, as well as the important privacy safeguards in those generative AI features to protect people's information and help them understand how the features work. As we explained there, generative AI models require a large amount of data to train effectively, so a combination of sources are used for training, including information that's publicly available online, licensed data and information from Meta's products and services.

- Privacy Center Article and Privacy Center Guide: Users are also provided with more specific information, including:

    - A dedicated "Generative AI at Meta" Privacy Center Guide, linked from the Privacy Center homepage and in other parts of the site.
    - A more detailed Privacy Center article focused on the data used for training, entitled 'How Meta uses information for generative AI models and features' in the Meta Privacy Center, and linked from the above guide.

- In-App Experience and Meta AI Terms: Additionally, when a user first uses the Meta AI chatbot on Facebook, Messenger, Instagram or WhatsApp, they go through a new-user experience (NUX) which informs them that their messages may be used to improve AI at Meta, links users to the AI Terms and indicates that use of Meta AI constitutes agreement to those terms.

  In the Meta AI Terms, at section 2(b), Meta informs users that:

  > "The Meta Privacy Policy governs our processing of personal information collected through the AIs and explains how we collect, use, and share personal information - including personal information you submit as Prompts, Feedback, or other Content, your activity across our products, and personal information we collect from and about your device. We may process personal information, including any sensitive information, as described in the Privacy Policy, to improve AIs and related technology."

- Additional privacy safeguards: On top of the clear and articulated commitment to privacy that extends to all Meta products and services, we adopt the following privacy safeguards to train our generative AI models:

    - *Selection of training data:* Meta excludes data from certain sites known to have a high volume of personal information about non-public individuals.
    - *Deduplication:* Meta conducts deduplication to reduce the chance that the generative AI model will memorise the data of any one individual.
    - *No linking of training data to user accounts:* Meta does not actively link any information contained in third-party datasets with specific users, and structured account identifiers from Meta's platforms and services (such as Meta User ID and username string) are removed before public first-party data is used for training models.

With respect to some of the specific questions raised above, we note:

- Meaning of 'public': Public data means data that is shared publicly and without restriction, meaning it does not, for example, include posts or comments with an audience set to "only me," "friends," or other types of restricted sharing. It does include unrestricted posts shared by Pages (businesses). Again, however, training data is used to find patterns and build statistical patterns; it is not stored or processed in a way that intends to associate that data with any one individual.
- *Privacy settings:*
    - *Facebook:* Facebook does not have an account level privacy setting (unlike Instagram). A user has control over the audience for each piece of content. For example, the audience for a Post would be defaulted to 'friends' and this has been the default for over a decade. Users can override any default setting by adjusting their audience settings generally or by adjusting the audience settings for specific pieces of content. For example, a user can adjust the audience setting for a Reel from 'Public' to a more limited audience, such as 'friends' in the creation flow or after the fact. When the audience for a post or Reel is 'friends', only connected friends can view the post and Meta does not currently use that content for model training.
    - *Instagram:* On Instagram, as befits its follower model and similarly to many other "follower model" online services, the default setting is public.
    - *Youth:* Every new user in Australia with a stated age under 16 years old is defaulted into more private settings when they register for Facebook. Prior to the recent rollout of 'Teen Accounts' on Instagram, new under

16 accounts would have their account created as private, with the option to move to public. Post-Teen Accounts rollout, new under 16 accounts must receive parental permission to switch to public via Parental Supervision. Existing under 16 accounts created prior to this rollout, who set their privacy to public, are currently in transition back to private. Once private, they will also require parental permission to switch to public via Parental Supervision.

(13) In the responses to questions on notice, Meta helpfully advised the steps it takes to prevent unauthorised scraping on Facebook products. Has Meta authorised any other developers of AI models/products/services to scrape any Meta products? (a) If yes please advise which, and why. (14) Meta also said in the response to that question on notice, that Meta has "taken a variety of actions against data misuse. These can include sending cease and desist letters, disabling accounts, filing lawsuits against acrapers engagrng in egregious behaviour and requesting companies that host scraped data to take them down." Please advise (answering separately for each): (a) Globally, how many instances of unauthorised scraping of Meta products has Meta detected? (b) In Australia, how many instances of unauthorised scraping of Meta products has Meta detected? (c) Does Meta inform victims of this scraping that their content has been scraped, in Australia or in any other market? (d) Noting that Meta has been able to identify the perpetrators of this sort of scraping in order to take the actions described in the quote, have any of these instances of scraping been in relation to AI-generated deepfake sexual material? (15) Meta's response to questions on notice says: *"Meta's Terms of Service state that users must not access or collect data from our Products using automated means (without our prior permission) or attempt to access data you do not have permission to access. We devote substantial resources to combating unauthorized scraping on Facebook products. We have a dedicated Anti-Scraping Team, including data scientists, analysts, and engineers focused on our efforts to detect, block, and deter scraping."* Please advise (answering separately for each): (a) Has Meta consulted the Terms of Service of every external data source that has been used to train its AI models to check if similar terms apply? (b) Can Meta confirm that every external data source used to train its AI models did not have terms prohibiting unauthorised scraping? (c) What percentage of external data sources did Meta successfully obtain authorisation to scrape to train its AI models? (d) Does Meta accept the owner of any data source, whether it is Meta or anyone else, should have the same right to prohibit unauthorised scraping and that AI developers should be

held accountable for scraping that contravenes terms of service, even if the content is publicly accessible?

Meta has authorised a select group of third parties to crawl public data on Meta products for the following specific and narrow purposes: (i) providing search engine results, and (ii) displaying previews of Meta URLs within messaging apps and similar services. At this time, we have not authorised crawling activities on Meta products outside of these two use cases, and authorised crawlers are not permitted to use crawled data for other purposes. The list of authorised crawlers is available on Meta's publicly facing robots.txt file.

More generally, as noted in the questions, Meta takes a multi-faceted approach to combating the unauthorised use of data from our services. This includes prohibitions against this activity in our Terms of Service and in our Automated Data Collection Terms that require potential crawlers to comply with our robots.txt file, which enables crawlers to determine if it can crawl Meta's services. However, we also use technical measures such as rate limits and data limits. The first line of defence is always practical – to make it as hard as we can for people's data to be collected at scale.

Additionally, Meta provides notifications to users as required by applicable laws and regulations. Additionally, under appropriate circumstances, and regardless of whether user notification is required, Meta may provide transparency to users through Newsroom posts and other public materials.

With respect to enforcement against the misuse of data on Meta's services, since 2019 we have issued over 1,800 cease and desist notices against people who abuse our platform and some of these enforcement actions have occurred in Australia. We have also brought lawsuits where appropriate.

With respect to the use of publicly available data for the purposes of training AI models, Meta respects the same mechanism and follows the industry-standard robots.txt protocol to determine where our crawlers are allowed to crawl. Publishers can exclude their domains from being crawled by Meta's AI-related crawlers through robots.txt disallows. For more information on how to do this please see here.

Meta only collects data for GenAI training from unrestricted third party sources using a declared crawlers that follow robots.txt. We do not include data from sites that require a login or sit behind a paywall as these sources may not be considered fully public. To further support this, we don't crawl data from publicly available sites known to contain a high preponderance of personal data.

(16) At the hearing, Mr Milner refused to answer a series of questions relating to media reports that Meta has used a database of pirated books called Books3 to train its AI model, on the basis that he cannot comment on legal proceedings. To be clear, these questions are questions of fact and are not asking for views or interpretations of the legality of any actions. Please advise (answering separately for each): (a) Is one of the data sources used to train LLaMa or other Meta AI products or services, the Books3 database? (b) A series of articles published in The Atlantic, on 19 August 2023, 25 September 2023, 27 September 2023, claim that Meta has trained its AI models on the Books3 database. Are these claims factually incorrect? (if yes please advise how). (c) An article published on 'TorrentFreak' on 11 January 2024 titled 'Meta admits use of pirated book dataset to train AI', contains a screenshotted excerpt from a Meta-authored document which says "Meta admits that it used portions of the Books3 dataset, among many other materials, to train Llama 1 and Llama 2." Is this factually incorrect? (if yes, please advise how). (17) These questions relate to questions of legality in Australia under Australian law and thus have no relation to legal proceedings in the United States. Please advise (answering separately for each): (a) Has Meta sought or prepared advice about the legality of pirated copies of copyrighted data being downloaded and/or used to train AI models, in Australia with relation to the Copyright Act 1968? (b) Has Meta sought or prepared advice about the legality of Meta's generative AI products producing, in Australia, material that is a copy of copyrighted works with relation to the Copyright Act 1968? (c) Has Meta sought or prepared advice about the interaction between Meta's generative AI products and the concept of moral rights under the Copyright Act 1968? (d) If Meta trained and/or downloaded pirated copies of copyrighted content in Australia, does Meta believe this would be lawful under the Copyright Act 1968? If it does, please advise on what basis it has formed this opinion. (e) If a user of Meta in Australia was able to use Meta's generative AI product to produce a copy of a copyrighted work, does Meta believe this would be lawful under the Copyright Act 1968?

Training a generative AI model such as a LLM is similar to building any large statistical model – such as a meteorological model used to predict the weather – and involves deriving patterns and relationships from a very large and diverse set of existing data.

The quantity and diversity of data is critical to building a good LLM, as it needs a huge amount of information reflecting a broad range of speech to derive the highly complex and nuanced rules, meanings, and contradictions of human language.

From a copyright perspective, it is important to understand that:

- AI models do not store or contain copies of the underlying training data itself; instead the model uses its training data to guide the process of deriving patterns and relationships across the entire set of data on which it is trained.
- While a piece of data may assist a model to better understand the meaning and inter-relationship of words, no individual piece of content has a particular influence over the model as a whole.
- The goal of AI models is to extract non-expressive facts, statistics, and relationships from training data and use them to generate new content.
- User-facing products such as LLM-based chatbots from Meta and others feature carefully-designed user interfaces that include layers of quality control and safety checks to limit what kinds of queries the model responds to and adjust its outputs accordingly.

The scale of data required to train generative AI models makes the documentation and disclosure of individual training data infeasible. Given the massive scale of data involved, it is impossible to definitively know whether specific publicly-available data is protected by copyright or not.

Meta undertakes relevant reviews when developing or releasing new products. Assessments of applicable legal requirements are generally privileged and we are not able to disclose details of that legal advice. Meta primarily develops and trains its AI models outside of Australia (primarily in the U.S.). Meta is subject to legal proceedings in the U.S. including in relation to claims under U.S. copyright law. Meta has set out its position in relation to copyright in its public response to the U.S. Copyright Office inquiry.

(17) At the hearing Mr Milner was asked by Senator Ghosh about whether authors can opt out or indicate to Meta that they don't want their work to train AI models. Mr Milner responded that "There's a very-well-understood and well-used means for a copyright holder or, indeed, somebody who is just operating a website to say, 'I don't want this to be used for AI training.' They can do that, and we will respect it." Please advise (answering separately for each): (a) What is the exact process by which an author can seek for the content they have authored to not be used for AI training? (b) How does this process apply when the author's content has been taken from a third party website that did not obtain the author's consent to host the content to begin

with? (c) How does this process apply if an author is unsure of whether his content has been used for AI training, but wants Meta to verify that this has taken place and to remove it accordingly? (d) Does this process retroactively remove their content from those models' datasets or will this only take effect with future models/datasets? (e) Is there a process by which Australian creators can seek recognition and/or payment for Meta's commercialisation of their work through generative AI? If not, would Meta be willing to establish such a process?

The process to which Mr. Milner is referring is the existing robots.txt protocol, a scalable and appropriate mechanism to enable domain owners to effectively signal whether AI crawlers can gather data from publicly accessible domains. If the rightsholder says no, Meta will not crawl (and the rightsholder can allow crawling for link previewing on Facebook or Instagram, which many find economically beneficial, and still say no to crawling for training). These robots.txt instructions govern the training of future models. The situations identified in (b), (c) (e) are not covered by robots.txt. Meta is constantly working to improve the options available for authors and domain owners, which include systems beyond robots.txt.

(18) The Meta website says the steps you take to mitigate privacy risks include "taking steps to remove data from certain sites known to contain a high volume of personal information about individuals." Please advise (answering separately for each): (a) Given that you only use public data to train your AI models, does Meta accept then that some public data can also consist of personal information? (b) Please provide a list of publicly accessible websites that were wholly or partially excluded from the training dataset because they contained personal information. (c) What is the difference between publicly accessible personal information on an external website, and publicly accessibly personal information on Facebook and Instagram, given that Meta applies different policies to their use for training AI?

We have included important privacy safeguards as we built these promising new generative AI features to protect people's information and help them understand how the features work. As noted above, Meta has a robust internal Privacy Review process that helps ensure we are using data at Meta responsibly for our products, including generative AI. We work to identify potential privacy risks that involve the collection, use or sharing of personal information and develop ways to reduce those risks to people's privacy.

- Our Generative AI Privacy Guide provides more detailed information regarding how we built our AI models, how our AI features work, and what controls and data privacy rights they have. In brief, we have built in privacy protections at several layers: At the training stage, Meta takes steps to limit the volume of personal information being included in training data by, for example:
    - In relation to third-party data, taking steps to remove data from certain sites known to contain a high volume of personally identifiable information about individuals; and
    - In relation to data from a user's interaction with our AI products and features, prior to use for model training, we remove any Meta unique system identifiers, such as Meta User ID and username string. Removing identifiers is designed to ensure that this data is not associated with a user account. We also use human review to check for the presence of other categories of potentially identifying data, such as private email addresses, private phone numbers and national identifiers. In relation to other user data from our platform, such as public posts and comments, we remove Meta unique system identifiers.

- In relation to outputs generated by Meta's generative AI models, Meta takes steps including:
    - Meta has a set of constantly evolving internal and confidential policies that outline acceptable and unacceptable LLM responses across a broad range of categories. These policies aim to preserve privacy by precluding outputs of certain personal information, such as highly sensitive information, or information from a non-public source.
    - Training and fine-tuning our generative AI models to limit the possibility of personal information that users may share with our generative AI products (e.g., our Meta AI assistant) from appearing in responses to other people.
    - Using a combination of human reviewers and automated technology to review model outputs so we can reduce the likelihood of outputs including personal information (as well as improve product performance).
    - To help ensure the efficacy of our safeguards, we perform accuracy assessments where the AI Chatbots are prompted with queries on acceptable and unacceptable LLM responses across a broad range of categories to ensure they are performing as expected, and conduct query monitoring, which provides notification of an unusual increase in queries on on acceptable and unacceptable LLM responses which would prompt a product team investigation aimed at addressing the matter, where necessary. Safety evaluations and red teaming are also performed before each LLM release (usually every few weeks) for the AI Chatbots.

Generative AI models are statistical computer models and do not themselves store personal data (or copies of personal data contained in training data). Similar to most competitors, Meta does not associate training data with a specific user. Meta does not actively link the third-party dataset with specific users, and structured account identifiers from Meta's platforms and services are removed before public first-party data is used for training models. Given the scale of datasets used for training generative AI models, there is no realistic prospect of identifying a single user's data.

While generative AI models are not designed to store personal data (or copies of personal data contained in training data), Meta still takes steps to reduce the amount of personal information that may be present in the underlying training data, in line with principles of data minimisation. As such, Meta defines precise criteria for data to be incorporated into our third-party dataset, and domains that have a risk of containing high volumes of personal data are filtered out. Over a million domains are excluded, so given this size, it is impractical to provide a full list. This dataset is unstructured and not associated with any particular user or private individual.

Similar to the third-party data used for training, the first-party data is also unstructured and not associated with any particular user, as relevant user ID and username strings are removed prior to use for training.