Attention: Your Conversational Data is What They Need

This version is for spotlight consideration and will be improved for the final submission.

Jakob Merane [1]

Abstract

As generative AI chatbots are getting widely used, more conversational data is being generated. The use of such data by model developers for further training is a critical and timely issue. This paper examines the practices of leading chatbot providers through a legal lens and identifies room for improvement in terms of privacy friendliness.

1. Introduction

The launch of ChatGPT, the first consumer-facing large language model (LLM), made the much-anticipated AI revolution tangible for the general public. With its impressive ability to provide tailored responses, ChatGPT quickly captivated users (Wu et al., 2023). Many credited its usability to the integration of supervised fine-tuning and reinforcement learning from human feedback applied to the foundation model (Ouyang et al., 2022). With approximately 100 million weekly active users (OpenAI, 2024a), OpenAI now possesses a vast amount of conversational data, which is valuable for further fine-tuning. However, this has raised concerns about data privacy.

Recent research has demonstrated that LLMs can memorize and leak data not only from pre-training datasets (Nasr et al., 2023) but also from fine-tuning datasets (Borkar, 2023). This has led some companies to fear that proprietary information could be leaked, prompting them to restrict the use of these models (Tilley & Kruppa, 2023). Additionally, the issue of how user data should be used for model training is a subject of heated debate, highlighted by the recent backlash from Slack users (Belanger, 2024) and Meta's sudden pause in its plan to use user data to train AI models (Meta, 2024).

In this work, we outline current practices of generative AI chatbots in using user conversational data for model training. By *conversational data*, we refer to all material input by the user (prompts and other user content) into the chat, as well as the output generated by the LLM. We analyze these practices in light of EU regulations and discuss them within the broader context of law and technology.

[1]ETH Zurich, Center for Law and Economics, Zurich, Switzerland. Correspondence to: Jakob Merane <jakob.merane@gess.ethz.ch>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

2. Background

In this section, we briefly outline the launch of four generative AI chatbots—*ChatGPT*, *Gemini*, *LeChat*, and *Claude*—in Europe. Given that ChatGPT was the first and most widely used (Shafer, 2024), we focus particularly on OpenAI. We analyze the approaches these companies take in using conversational data for further model training. Specifically, we investigate whether they employ an *opt-in* or *opt-out* approach for data usage. *Opt-in* means users must give their consent beforehand, whereas in an *opt-out* approach, companies use the data unless the user actively opts out. We focus exclusively on consumer-facing chat interfaces, excluding data processed through APIs.

2.1. OpenAI

At its launch on 30 November 2022, ChatGPT allowed OpenAI to use conversations for further model training without any opt-out option. However, OpenAI stated they would remove personally identifiable information from the data. This was also true with previous instruct beta models (Ouyang et al., 2022), where OpenAI stated human annotators would remove personally identifiable information.

OpenAI swiftly enabled an opt-out option for API users via email in December 2022, but the situation for ChatGPT users progressed more slowly. It wasn't until the Italian Data Protection Agency (Garante) opened an investigation that significant changes occurred. In March 2023, the Italian Garante imposed a temporary limitation on processing, asserting, inter alia, that Italian users should have a right to opt out from their data being used (Garante, 2023; Chiara, 2023). On April 25, 2023, ChatGPT introduced the option to deactivate training in its settings (OpenAI, 2023). Before this, OpenAI reverted to using a Google Form for opt-outs, suprising for a company valued at over 30 billion USD at the time (Glasner, 2023). ChatGPT became accessible to Italian

users again on April 28, 2023, although the investigation is ongoing (Garante, 2024).

Despite introducing the opt-out option, criticism persisted. One issue was that opting out of model training prevented retaining chat history. In early 2024, OpenAI introduced temporary chats, allowing users to disable data usage for training while saving history outside a session-based window (OpenAI, 2024b). By disentangling model training and history, OpenAI may have addressed some criticism. Another criticism emerged from January to April 2024, when users on various platforms, including the OpenAI Community Forum, reported a bug that could only be resolved by disabling the opt-out setting (OpenAI Community, 2024). OpenAI's delayed response led to speculation that the friction was intentional to coerce users into sharing data.

2.2. Google

Google entered the market later, launching its generative AI chatbot BARD in the US in March 2023, rebranded as Gemini in 2024. Google initially planned to launch its product in June 2023 in Europe, but postponed it until mid-July 2023 following discussions and concerns raised by the Irish DPA (Data Protection Commission, 2023). Google uses Gemini conversations of its users for further model training. However, from the launch, Google provided a Privacy Hub allowing European users to opt-out of model training. Furthermore, they explain that human reviewers may process conversations if users do not opt out. As safeguards, they separate conversations from user identities and advise users not to share anything in the chat they would not want seen by a reviewer. Additionally, only a random portion of all conversations are reviewed.

2.3. Mistral AI

The French AI start-up Mistral AI, known for its open-source models, launched LeChat on February 26, 2024. LeChat allows users to opt-out if they subscribe to the paid chat service; otherwise, there is no opt-out option. Users are asked to pseudonymize prompts to exclude personal data from training data. It is unclear what privacy safeguards Mistral AI has in place for LeChat.

2.4. Anthropic

Anthropic launched its generative AI chatbot Claude in Europe on May 14, 2024. Anthropic, which aims to develop safe and ethical AI systems, does not use conversational data from users by default for further model training. It only does so if a user opts in, although currently, there is no opt-in option. This approach clearly sets Anthropic apart from other companies.

3. European Legal Framework

In the early days of generative AI chatbots, the regulation of AI was still in its infancy. Yet, in Europe, these companies did not operate in a legal vacuum.

3.1. The General Data Protection Regulation

Applicability. Many users include personal data in their prompts. Therefore, using user conversations likely involves processing personal data, placing it within the scope of the EU's General Data Protection Regulation (GDPR) (Dieker, 2024). Simply asking users in the fine print not to input any personal data does not absolve companies from the responsibility of complying with GDPR. The GDPR also applies to Anthropic, Google, and OpenAI, which have a main establishment in Ireland. As shown by the investigation of the Garante, the GDPR has extraterritorial application when a US-based company serves European users.[1]

Lawfulness. A key provision of the GDPR mandates that the processing of personal data requires a legal basis.[2] In this context, all companies rely on legitimate interest. According to recent case law from the European Court of Justice, product improvement may indeed be considered a necessary legitimate interest.[3]

Legitimate interest. Data Protection Authorities are expected to challenge the necessity of such processing for model improvement purposes (EDPB, 2024). While using real user data might enhance alignment, less intrusive alternatives, such as processing data from users who consent, could suffice (Sartor, 2023). Since companies claim that only a small portion of data is used, making processing conditional on consent might still yield sufficient data. However, suggesting reliance solely on internal alignment trainers or external contractors might be overly restrictive for these companies.

Furthermore, it is required that the legitimate interests outweigh those of its users. This balancing depends on three factors. First, the reasonable expectations of the affected users (Veil, 2018). Since interacting with generative AI chatbots is quite new, many users might not anticipate how their input is used beyond the conversation. For instance, users likely do not realize the implications when they provide a thumbs up on a response. Second, legitimate interest

[1] Art. 3(1) and Art. 3(2) GDPR. For improved readability, references to laws will be found in footnotes.

[2] Art. 6(1) GDPR.

[3] Meta Platforms Inc. and Others v. Bundeskartellamt, Case C-252/21, ECLI:EU:C:2023:537, para. 122 (2023) ("it cannot be ruled out from the outset that the controller's interest in improving the product or service with a view to making it more efficient and thus more attractive can constitute a legitimate interest capable of justifying the processing of personal data and that such processing may be necessary in order to pursue that interest").

might not be a sufficient legal basis, when the user data involves certain sensitive conversations, such as when a user discusses personal health issues.[4] Third, the impact of the processing and whether the company implements specific safeguards to protect personal data, such as anonymization, pseudonymization, data minimization, and retention limits.

Opt-out. In addition, the legal basis of legitimate interest requires that users be informed and given the option to opt-out, with non-compliance subject to fines the GDPR.[5] Users must be informed at the time of the first communication. Provided it is clearly highlighted, making it available as part of the privacy policy is acceptable. However, it is problematic to hide the opt-out mechanism or make it so complicated that it discourages users.

3.2. The AI Act

Transparency. The new provisions for foundation models under the EU's AI Act are expected to take effect in July 2025. In Chapter 5, the EU has introduced new rules for General Purpose AI models (GPAIs) that will apply alongside the GDPR rules mentioned earlier.[6] It is likely that most of the models behind the discussed generative AI chatbots will be classified as a *Systemic risk* GPAI because the floating-point operations for training exceed 10^{25} (Epoch AI, 2024) or due to other criteria such as a high number of users.[7] Despite this, the AI Act does not require opting in for the use of data for model improvement. However, it imposes documentation duties on GPAIs vis-à-vis the Office for AI and providers of AI systems regarding the training of their models (Friedl & Gasiola, 2024).[8] These requirements may increase the overall transparency for users.

4. Discussion

Benchmarking. In the generative AI community, it is common to compare the performance of an LLM on leaderboards. However, this practice has increasingly been questioned, as model developers tend to overfit (Alzahrani et al., 2024). Beyond performance, it is also important to consider the use of conversational data and privacy-friendliness when comparing different providers. Ideally, generative AI chatbots should use conversational data only when users are aware of and have opted in. For instance, Anthropic's Claude, which offers an opt-in option, might be considered best in class.

In Figure 1, we compare the opt-out option (*opt-out*) and whether it is explicitly communicated in the first use of the chatbot (*first use*). Merely including the option in the privacy policy (*fine print*) is not ideal as research has shown that consumers rarely read or understand these documents (Bakos et al., 2014). Users should also be able to opt out without unnecessary friction in the settings (*settings*). Additionally, users should have access to clear documentation about privacy safeguards (*privacy safeguards*) and the retention limitation (*retention*).

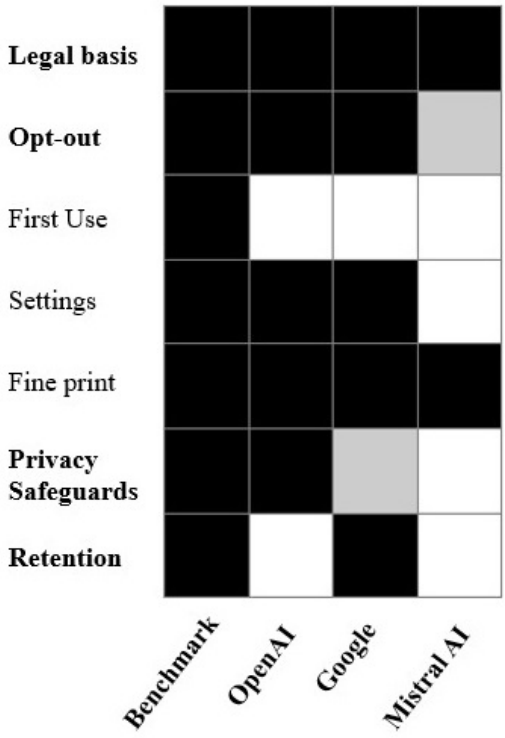

Figure 1. Comparison of different generative AI chatbots based on privacy criteria. Black indicates the criterion is fully satisfied, grey indicates the criterion is partially satisfied, and white indicates the criterion is not satisfied.

Balance. Investigations are ongoing against OpenAI, and Google is under scrutiny in Ireland. Surprisingly, Mistral AI, often praised as a European champion, does not seem to champion privacy. If EU laws are not merely industrial policy tools and if Mistral AI does not change its policies, they too will likely become a target for regulatory authorities. The challenge of these companies with compliance may stem from two conflicting roles: that of model developer and a chatbot provider. A model developer has the incentive to collect as much user data as possible to optimize the training and improve the model. As a chatbot provider, one might be incentivized to increase user satisfaction through user- and privacy-friendly chat interfaces. Therefore, it is not surprising that HuggingFace offers HuggingChat (Hug-

[4] Art. 9(1) GDPR.
[5] Art. 21 GDPR; Art. 85(5)(b) GDPR.
[6] Recital 10 of the AI Act.
[7] Art. 51(2) and 51(1)(b) AI Act and its Annex XIII.
[8] Art. 53 and 55 AI Act.

gingFace), which provides users with a chat interface to use different open-source models. In this case, the data is not used for further training or forwarded to the model makers. The interest situation is different here, as the chat provider is not the model developer. The dual role as a chat provider and model developer presents a challenge.

Vulnerabilities. It has become industry standard that model developers do not use the data of companies that send data through APIs for business applications. However, professional business users can handle much more than regular consumers, who are significantly more vulnerable. Even though there is a business rationale for the the better conditions for companies, a weaker default privacy setting for consumers is not reasonable from a vulnerability standpoint.

Friction. Creating friction in users' ability to exercise their opt-out rights is not a new phenomenon online. This practice is often referred to as *dark patterns* and is well studied with cookies (Luguri & Strahilevitz, 2021). With generative AI chatbots, friction could be introduced by having a complicated form-filling process or worse performance after opt-out for non-technical reasons. The legal qualification of this issue will be determined on a case-by-case basis. However, in a regulatory investigation, a company would need to provide truthful information.

Law and technology. In the law and tech literature, the so-called *pacing* problem describes how the law often lags behind technological advancements (Marchant, 2011). Indeed, an early ChatGPT user could not protect their data from further training. And by the time the Italian Garante intervened, GPT 4.0 had already been launched, which likely had also been trained on user data. Although the generative AI chatbot market is very competitive, it could still be that there are certain network effects, as the first mover has more users, collects more data, particularly in the first months while it flies under the radar, and the data is used for better models which again attracts more users. If fine-tuning really is the secret sauce behind the magic of generative AI chatbots, this is indeed an interesting dynamic.

However, this is only one part of the story, especially in the European context. As the example of ChatGPT shows, there is also a *racing* problem. ChatGPT was rapidly made available worldwide, adhering once again to the Silicon Valley philosophy of "move fast and break things". The law has a reactive nature and can thus reward those who are bold and launch things prematurely. Whether data protection authorities can do anything about it if the data collection turns out to be illegal remains doubtful.

Innovation. To do

5. Conclusion

As LLMs are getting widely used and mulitmodal (Stanford University, 2024), more conversational data is being generated. Although the pre-training and copyright law receive much scholarly attention, the issue of using conversational data to improve models raises issues regarding privacy. On one hand, it should be entirely possible for model developers to use real user data to improve models. On the other hand, it would be desirable for this to be dependent on users opting in. Data protection authorities are already poised to act, and it will be interesting to observe how the providers evolve in this regard.

Acknowledgements

Special thanks to X, X, and the anonymous reviewers for their feedback and beneficial discussions.

References

Alzahrani, N., Alyahya, H. A., Alnumay, Y., Alrashed, S., Alsubaie, S., Almushaykeh, Y., Mirza, F., Alotaibi, N., Altwairesh, N., Alowisheq, A., et al. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. *arXiv preprint arXiv:2402.01781*, 2024.

Bakos, Y., Marotta-Wurgler, F., and Trossen, D. R. Does anyone read the fine print? consumer attention to standard-form contracts. *The Journal of Legal Studies*, 43(1):1–35, 2014.

Belanger, A. Slack defends default opt-in for ai training on chats amid user outrage, 2024. URL https://arstechnica.com/tech-policy/2024/05/slack-defends-default-opt-in-for-ai-training-on-chats-amid-user-outrage/.

Borkar, J. What can we learn from data leakage and unlearning for law? *arXiv preprint arXiv:2307.10476*, 2023.

Chiara, P. G. Italian DPA v. OpenAI's ChatGPT: The reasons behind the investigations and the temporary limitation to processing. *Journal of Law and Technology*, 2023.

Data Protection Commission. Annual report 2023, 2023. URL https://www.dataprotection.ie/sites/default/files/uploads/2024-05/DPC%20EN_AR%202023_Final%20.pdf.

Dieker, A. Datenschutzrechtliche zulässigkeit der trainingsdatensammlung. *Zeitschrift für Datenschutz*, pp. 132–137, 2024.

EDPB. Report of the Work Undertaken by the ChatGPT Taskforce, 2024.

Epoch AI. Epoch AI Database, 2024. URL https://epochai.org/data/epochdb/visualization?startDlEra=1950-12-28.

Friedl, P. and Gasiola, G. G. Examining the EU's Artificial Intelligence Act, 2024. URL https://verfassungsblog.de/examining-the-eus-artificial-intelligence-act/.

Garante. Temporary limitation of processing of italian users' data against OpenAI, 2023. URL https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9870847#english.

Garante. ChatGPT: Italian DPA notifies breaches of privacy law to OpenAI, 2024. URL https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9978020#english.

Glasner, J. Startup valuation fluctuations, 2023. URL https://news.crunchbase.com/venture/startup-valuation-fluctuations-ai-openai-msft-eoy-2023/.

HuggingFace. Huggingchat. URL https://huggingface.co/chat.

Luguri, J. and Strahilevitz, L. J. Shining a light on dark patterns. *Journal of Legal Analysis*, 13(1):43–109, 2021.

Marchant, G. E. *The Growing Gap Between Emerging Technologies and the Law*, pp. 19–33. 2011.

Meta. Building ai technology for europeans in a transparent and responsible way, 2024. URL https://about.fb.com/news/2024/06/building-ai-technology-for-europeans-in-a-transparent-and-responsible-way/.

Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Wallace, E., Tramèr, F., and Lee, K. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*, 2023.

OpenAI. New ways to manage your data in ChatGPT. 2023. URL https://openai.com/index/new-ways-to-manage-your-data-in-chatgpt/.

OpenAI. Start using ChatGPT instantly, 2024a. URL https://openai.com/index/start-using-chatgpt-instantly/.

OpenAI. Memory and new controls for ChatGPT, 2024b. URL https://openai.com/index/memory-and-new-controls-for-chatgpt/.

OpenAI Community. Chat history off - conversation key not found error, 2024. URL https://community.openai.com/t/chat-history-off-conversation-key-not-found-error/594342.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Sartor, G. Article 6. In gen. Döhmann, I. S., Papakonstantinou, V., Hornung, G., and Hert, P. D. (eds.), *General Data Protection Regulation: Article-by-Article Commentary*, pp. 318. First edition edition, 2023. ISBN 9783848733729.

Shafer, S. Consumer insights: One-quarter of us adults have used generative ai tools, June 17 2024. Market Intelligence.

Stanford University. The AI Index Report, 2024. URL https://aiindex.stanford.edu/report/.

Tilley, A. and Kruppa, M. Apple restricts use of ChatGPT, joining other companies wary of leaks, 2023. URL https://www.wsj.com/articles/apple-restricts-use-of-chatgpt-joining-other-companies-wary-of-leaks-d44d7d34?

Veil, W. Einwilligung oder berechtigtes Interesse?: Datenverarbeitung zwischen Skylla und Charybdis. 2018.

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.-L., and Tang, Y. A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5):1122–1136, 2023.