# LM-INSTRUCT: A case study on Domain-specific Synthetic Instruction Dataset generation in Danish

**Mike Riess**
Research & Innovation
Telenor Group
Oslo, Norway
email@email.com

**Kenneth Enevoldsen**
Center for Humanities Computing
Aarhus University
Aarhus, Denmark
email@email.com

March 16, 2025

## Abstract

## 1 Introduction

Online forums has been a vital space for collaboration and knowledge exchange ever since the introduction of the internet. These platforms serve as an important medium, especially for communities within music production and sound engineering, enabling users to discuss complex topics in-depth and connect with peers. The Danish forum Lydmaskinen.dk is one such platform, consisting of almost 11k users and a total of 644k topics started (as per August 2024).

These platforms contain a vast amount of specialized knowledge in written form, which makes it of particular interest in the research and development of Large Language Models (LLM) in Danish. Fine-tuning or evaluating LLMs on this data source could be useful for research on language capabilities in Danish or skill learning (domain knowledge) beyond the Danish language itself.

However, a potential problem in this context is the accidental sharing of information that can be regarded as personal (Full names, addresses, phone numbers, e-mails etc.,). As LLMs are known to memorize sequences from their training data [1], person-identifiable information can become artifacts of future LLMs trained on the information from Lydmaskinen, if not carefully filtered and anonymized.

In this study we analyze the Legal, Ethical and Technical aspects of releasing user-generated content for AI Research, while proposing a methodology that respects these. Finally, we present LM-INSTRUCT, a dataset of dialogues in Danish within professional music production, sound engineering (live and studio), music theory, acoustics and video production.

### 1.1 Research questions

To guide the work of generating the LM-INSTRUCT dataset, two research questions are proposed:

- **G1 Subset refinement:** Transforming the Lydmaskinen.dk database into a instruction dataset that is legally, technically and ethically sound for use in future research.
- **G2 Validation:** Comparing the performance of the generated instruction dataset to the original dataset.

To address G1, a literature review of relevant work and legislation in this area is surveyed in section2, while a set of resulting requirements and alternatives for the publication of the data is listed in section 3. The chosen approach for the processing and release of LM-INSTRUCT is presented in detail in section 4.

To validate the approach, a comparison of the performance of the generated instruction dataset to the original dataset will be performed by fine-tuning a LLM on both the original and generated dataset. The results will be presented in section 5.

## 2 Literature review

To provide a structured overview, this section is divided into three subsections: Legal (2.1), technical (2.3) and ethical (2.2) aspects.

### 2.1 Legal aspects

#### 2.1.1 GDPR Compliance

The General Data Protection Regulation (GDPR) [2] is a comprehensive framework for data protection in the European Union. Key aspects of GDPR relevant to the release of LM-INSTRUCT is discussed in the following.

**Consent and Transparency:** The GDPR mandates obtaining informed consent before collecting, processing, or sharing personal data which might occur in the user-generated content at Lydmaskinen. This requirement emphasizes the need for clear communication with users about how their data will be used. When creating a profile on Lymaskinen.dk, the user is faced with Terms Of Service (TOS) that gives the forum the rights to use their data and re-distribute it to third parties. As of 31/07/2024, the last section of the TOS says the following:

> “Ved at indsende skriftligt indhold til "Lydmaskinen", beholder du alle ejerskabsrettigheder til dit indhold. Dog giver du "Lydmaskinen" en ikke-eksklusiv, royalty-fri, vedvarende, verdensomspændende licens til at bruge, reproducere, ændre, tilpasse, udgive, oversætte, distribuere og vise sådant skriftligt indhold i ethvert medie eller format. "Lydmaskinen" forbeholder sig retten til at ændre disse betingelser til enhver tid. Ændringer vil blive offentliggjort på denne side, og din fortsatte brug af forummet udgør accept af disse ændringer. Hverken "Lydmaskinen" eller phpBB blive holdt ansvarlig for ethvert hackingforsøg, som kan medføre at dataene bliver kompromitteret”

While this definition gives Lydmaskinen the right to distribute the user-generated content to third parties, article 5(1)(d) of the GDPR specifically requires that personal data shall be:

> “collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes (‘purpose limitation’);”

As the intended use of this data is research, Article 89(1) could apply, which specifies that safeguards such as pseudonymization and data minimization needs to be put in place to protect individuals. In the cases that the data does not fall under this category, all the uses of the released data needs to be specified beforehand in the TOS at Lydmaskinen. This presents some practical limitations, as this cannot be known or controlled by Lydmaskinen once the data has been released.

**Data Minimization and Purpose Limitation:** Article 4(5) of the GDPR [2] stipulates that data collection should be limited to what is essential for the specified purpose. This principle is exemplified in the case study by **(author?)** [3], which demonstrates practical applications of data minimization in compliance with GDPR.

**Anonymization and Data Protection:** The GDPR defines pseudonymization as:

> “the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.” [2]

This definition presents challenges in the case where a third party has access to the additional information that can be used to identify the individial. For Lydmaskinen this is a problem, as the raw data still exists without pseudonyms on the forum website. Technically, the data therefore cannot be effecetively pseudonymized if its source (at the forum website) remains unchanged.

### 2.2 Ethical aspects

**Responsibility to inform and obtain consent** The technical report from the National Committee for Research Ethics in the Social Sciences and the Humanities (NESH) [4], as well as the Ethics Guidelines from the Association of Internet Researchers (AoIR) [5] stress the importance of informed consent from the research subjects and to take extra precautions to protect participants' privacy and confidentiality. The NESH report states the following:

> "In other words, this does not refer to the statutory requirement for consent to the processing of special categories of personal data or the statutory entitlement to information and transparency (NESH 2016: B.8), but to the ethical responsibility that invariably rests with the researcher, irrespective of whether personal data are involved or not, or whether the information is sensitive or not. Variations in the nature of the research, its source material and source data may give rise to different questions and dilemmas concerning research ethics." [4]

**Transparency in research methods and data usage** The AoIR guidelines [5] stress the importance of transparency in research methods and data usage, as well as respecting online communities and maintaining data integrity. Researchers are advised to document their ethical decision-making processes and adhere to relevant laws and platform terms of service. The guidelines stress that researchers should provide clear and detailed explanations of their data collection, processing, and presentation methods. This includes explicitly describing data cleaning processes and any adjustments made to the dataset, as these can significantly influence research outcomes.

### 2.3 Techical aspects

#### 2.3.1 LLM Memorization and Privacy Risks

Recent studies have shown that that LLMs are likely to memorize and thereby reproduce complete sequences of text, which might present a privacy risk, depending on data used.

**Membership inference and reconstruction attacks:** The study of [6] explore sentence-level membership inference (inferring whether a model was trained on a particular piece of text) and reconstruction attacks (extracting complete sentences or larger portions of text that were used to train the model) and find that even though differential privacy (a method that reduces memorization during training, see [7]) reduces the leakage of person-identifiable information (PII), this still leaks about 3% of PII-sentences. Other attacks such as adversarial inference aim to indentify key entities in text and thereby identify demographical information that can be person-identifiable. The authors of [1] divide the memorization aspect of LLMs into six different types, and propose a balanced view by highlighting both the positive and negative aspects of each of these capabilties. Based on their analysis, the privacy and security aspects of memorization are exclusively negative (leaking sensitive data, author attribution, etc.), whereas these can have positive implications for the sake of auditing (watermarking, bias detection etc.) or model alignment (capability to quoute, question answering etc.).

**Inference attacks beyond memorization:** As multiple entries of the same sequences have been known to increase the risk of memorization, [8] propose to deduplicate text sequences in the training data to reduce the risk of memorization. However, the study of [9] finds that simple prompting techniques using state-of-the-art LLMs (GPT-4, Claude-2, Llama-2) can generate accurate information on indiviudals. Specifically, accuracy on real world data at top-1 and top-3 level classification were as high as 85% and 95%, respectively. To be clear, this method did not memorize data, but were able to predict person-identifiable attributes accurately based on text written by the subjects. Based on these findings, the authors thereby advocate for a broader discussion on LLM privacy implications (beyond mitigation of text memorization).

#### 2.3.2 Data Anonymization

**Heuristic pattern-matching:** At the lowest level of sofistication is the simple pattern matching and replacement techniques using e.g., Regex syntax. Examples can be to remove emails, credit card numbers, urls and so forth. These techniques can not stand alone, as they do not specifically target PII, but rather character-level patterns [10].

**De-identification:** This technique is defined as a sequence-labelling task using methods such as Named-Entity Recognition (NER) to identify and replace PII with general (sanitized) tokens [11]. Models such as DaCy NER [12, 13] and ScandiNER [14] are both options that require no fine-tuning or adaptation for use on text in Danish.

**Pseudonymization:** Using methods such as NER or LLMs to pseudonymize text to preserve some degree of its original context and meaning while still anonymizing PII [15]. One example for Danish text is the Python library Augmenty [16], which replaces entities with pseudonyms.

**Synthetic data generation:** Another option to data anonymization is to generate synthetic instances based on the distribution of real data [17]. This approach presents a promising alternative to sharing of real user-generated data. Synthetic text data can be generated by pre-training a language model on sequences from the real data distribution using self-supervised learning [18] (training from unlabeled data by masking and learning to predict the correct token). Inherently, this method does however also suffer from the same memorization risks discussed above, as the simulator is a decoder language model.

- **In-Context Learning:** Using an existing decoder model to generate synthetic data by providing a prompt and one or more examples.

## 3 Requirements and alternatives

Releasing a vast amount of user-generated content for use in research and development of Large Language Models is no trivial matter. Based on the literature review above, challenges within legal, ethical and technical domains have been identified and converted into requirements for the final dataset to be released as LM-INSTRUCT. The requirements are as follows:

- **(R1) Legal compliance:** Ensuring that relevant regulations such as the General Data Protection Regulation of the European Union is followed.
- **(R2) Ethical research standards:** Following the NESH and AoIR guidelines for ethical research.
- **(R3) Anonymization:** Ensuring that the amount of PII is minimized, to ensure this will not carry on to future models fine-tuned on LM-INSTRUCT.

To satisfy R1 and R2, the users included in the dataset must have given consent to the use of the data for the purpose of AI Research. Furthermore, PII must be removed from the data to satisfy all three requirements. Based on the known methods for data processing, three different alternatives have been identified.

- **(A1) Consent + Anonymization:** Asking users for consent (opt-in), while pseudonymizing or de-identifying any personal information at both Lydmaskinen and in the resulting LM-INSTRUCT data.
- **(A2) Synthetic questions and answers:** Pseudonymizing or de-identifying any personal information, while using the anonymized dataset to generate a synthetic version of the data to be released as LM-INSTRUCT.
  - `https://arxiv.org/pdf/2403.13787`
- **(A3) Anonymization at source and target:** Pseudonymizing or de-identifying any personal information at both Lydmaskinen and in the LM-INSTRUCT data, while informing the users about the publication of the data.

The first alternative, **A1**, preserves data quality as much as possible, while also minimizing the risks of future LLMs memorizing PII. In this case, users are asked directly to opt-in, which is assumed to lead to a significant reduction in the number of documents. Further, it assumes that any personal information present on Lydmaskinen is pseudonymized as well (to satisfy the pseudonymization definition in GDPR[2]). Finally, the data should be licensed under a CC BY-NC-SA 4.0 license [19], allowing for research-only use of the data.

The second alternative, **A2**, generates data that is similar in nature to the real data, without any limits in the number of documents, however, the quality of this synthetic data depends highly on the text generating capabilities of the LLM used. Using this approach, it is vital that the produced dataset is validated.

The third alternative, **A3**, assumes that the consent given at the forum is sufficient under GDPR, but that pseudonymization is still needed (at both Lydmaskinen and in LM-INSTRUCT) in addition to informing the users at Lydmaskinen. A general problem with this approach is that it is not clear how many users that would consent to the release of their data, and therefore the size of the dataset is unknown. Users that are no longer active on the forum will thereby be excluded, as they cannot be asked for consent.

Weighing the pros and cons of each alternative, **A2** is chosen as the best compromise between data quality, privacy and ethical concerns.

## 4 Methodology

### 4.1 Data retrieval

The dataset was extracted from a phpBB MySQL database and follows the default structure of the *phpBB3* open source forum software. The retrieval and preprocessing process consisted of the following steps:

1. **Data Loading**: Load preprocessed forum data from pickle files containing:
   - Forums data (forum_id, forum_name, etc.)
   - Topics data (topic_id, topic_title, topic_poster, forum_id, etc.)
   - Posts data (post_id, post_text, post_time, poster_id, etc.)
2. **Forum Selection**: Filter data to include only specific forums of interest (default: forum IDs 1 and 2).
3. **Topic Sampling**: Sample a specified number of topics (default: 30,000) from the selected forums using a fixed random seed (42) for reproducibility.
4. **Post Collection**: Retrieve all posts associated with the sampled topics.
5. **Data Merging**: Combine forum, topic, and post information into a unified dataset that preserves the hierarchical structure:
   - Link topics to their parent forums
   - Link posts to their parent topics
   - Sort by forum ID, topic ID, and post time to maintain conversation flow
6. **Text Cleaning**: Process text fields by unescaping HTML entities in forum names.
7. **Conversation Structuring**: Add post numbering within each topic and assign conversation IDs to facilitate further processing.

This preprocessing pipeline transforms the raw database export into a structured dataset suitable for the anonymization process described in Section 4.2. The resulting dataset preserves the conversational nature of the forum discussions while organizing the data in a format optimized for language model training.

### 4.2 Data Filtering and Replacement

We developed a comprehensive anonymization pipeline to process forum conversations while preserving domain-specific terminology and structure. Our initial anonymization procedure follows these steps:

1. **Data Organization**: Convert raw data into a structured conversation format, preserving metadata like forum name, topic title, and post relationships.
2. **Text Normalization**: Clean and standardize text by:
   - Removing BBCode quotes (e.g., [quote="USERNAME"]...[/quote])
   - Stripping excessive whitespace
   - Removing HTML tags
   - Normalizing quotes, apostrophes, and dashes
3. **NER-based Anonymization**: Apply Danish-specific named entity recognition using daCy[20] large transformer model (da_dacy_large_trf-0.2.0) to identify and replace:
   - Person names (PER) → [PERSON]
   - Locations (LOC) → [LOCATION]
   - Dates (DATE) → [DATE]
   - Nationalities, religious or political groups (NORP) → [GROUP]
4. **Signature Detection**: Identify common signature patterns in Danish forum posts and replace them with [PERSON] tags:
   - Slash signatures (e.g., "/Michael")
   - Danish closing formulas (e.g., "Mvh. Michael", "Venlig hilsen Michael")
   - English closing formulas (e.g., "Regards Michael")
   - Other common signature patterns (e.g., "Kh. Michael", "Vh. Michael")

5. **Pattern-based Anonymization**: Apply regex patterns to detect and replace structured personal information:
   - Email addresses → [EMAIL]
   - Phone numbers → [PHONE]
   - URLs → [URL]
   - IP addresses → [IP_ADDRESS]
   - Social security numbers → [SSN]
   - Credit card numbers → [CREDIT_CARD]
   - Physical addresses → [ADDRESS]
   - ZIP codes → [ZIPCODE]
6. **Username Anonymization**: Replace known usernames from a curated list (data/usernames.txt) with [PERSON] tags to prevent identification of forum participants.
7. **Token Counting**: Calculate token statistics for the original, normalized, and anonymized text to track the impact of anonymization on token counts.
8. **Dataset Preparation**: Format the anonymized conversations for language model training, preserving conversation structure and relevant metadata.

This pipeline aims to transform the raw data into a format that is anonymized to the extent possible, which will then be used as source input for the generation of synthetic instruction data.

## 5 Results

## 6 Discussion

## 7 Conclusion

## References

[1] Valentin Hartmann, Anshuman Suri, Vincent Bindschaedler, David Evans, Shruti Tople, and Robert West. Sok: Memorization in general-purpose large language models, 2023.

[2] European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council.

[3] Gil Francopoulo and Léon-Paul Schaub. Anonymization for the gdpr in the context of citizen and customer relationship management and nlp. 2020.

[4] National Committee for Research Ethics in the Social Sciences and the Humanities. A guide to internet research ethics. Guidelines, The National Committee for Research Ethics in the Social Sciences and the Humanities (NESH), Oslo, Norway, 2019. Second edition published in Norwegian in 2018 and in English May 2019.

[5] aline shakti franzke, Anja Bechmann, Michael Zimmer, Charles Ess, and Association of Internet Researchers. Internet research: Ethical guidelines 3.0. Guidelines, Association of Internet Researchers, 2020. Published under Creative Commons license.

[6] Nils Lukas, A. Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-B'eguelin. Analyzing leakage of personally identifiable information in language models. *2023 IEEE Symposium on Security and Privacy (SP)*, pages 346–363, 2023.

[7] Cynthia Dwork. Differential privacy: A survey of results. In Manindra Agrawal, Dingzhu Du, Zhenhua Duan, and Angsheng Li, editors, *Theory and Applications of Models of Computation*, pages 1–19, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[8] Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 10697–10707. PMLR, 17–23 Jul 2022.

[9] Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Beyond memorization: Violating privacy via inference with large language models, 2024.

[10] Microsoft. Presidio, 2024. Accessed on July 30, 2024.

[11] Pierre Lison, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid. Anonymisation models for text data: State of the art, challenges and future directions. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4188–4203, Online, August 2021. Association for Computational Linguistics.

[12] Kenneth Enevoldsen, Lasse Hansen, and Kristoffer Nielbo. Dacy: A unified framework for danish nlp. *arXiv preprint arXiv:2107.05295*, 2021.

[13] Kenneth Enevoldsen, Emil Trenckner Jessen, and Rebekah Baglini. Dansk and dacy 2.6.0: Domain generalization of danish named entity recognition, 2024.

[14] Dan Saattrup Nielsen. ScandiNER: Named Entity Recognition model for Scandinavian Languages. `https://huggingface.co/saattrupdan/nbailab-base-ner-scandi`, 2024. Accessed: July 29, 2024.

[15] Oleksandr Yermilov, Vipul Raheja, and Artem Chernodub. Privacy- and utility-preserving nlp with anonymized data: A case study of pseudonymization, 2023.

[16] Kenneth Enevoldsen. Augmenty: A python library for structured text augmentation. *Journal of Open Source Software*, 9(96):6370, 2024.

[17] Yingzhou Lu, Minjie Shen, Huazheng Wang, Xiao Wang, Capucine van Rechem, Tianfan Fu, and Wenqi Wei. Machine learning for synthetic data generation: A review, 2024.

[18] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[19] Creative Commons. Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0), 2013. Accessed: 01/08/2024.

[20] Kenneth Enevoldsen, Lasse Hansen, and Kristoffer L. Nielbo. DaCy: A unified framework for danish NLP. In *Ceur Workshop Proceedings*, volume 2989 of *CEUR Workshop Proceedings*, pages 206–216. ceur workshop proceedings.