Dynaword: From One-shot to Continuously Developed Datasets

Kenneth Enevoldsen[1*], Kristian Nørgaard Jensen[2], Jan Kostkan[1]
Balázs Szabó[1], Márton Kardos[1], Kirsten Vad[1], Johan Heinsen[3]
Andrea Blasi Núñez[4], Gianluca Barmina[4], Jacob Nielsen[4], Rasmus Larsen[2]
Rob van der Goot[5], Peter Vahlstrup[1], Per Møldrup-Dalum[1]
Desmond Elliott[6], Lukas Galke[4], Peter Schneider-Kamp[4], Kristoffer Nielbo[1]

[1]Aarhus University, [2]The Alexandra Institute, [3]Aalborg University
[4]University of Southern Denmark, [5]IT University of Copenhagen, [6]University of Copenhagen

[*]Correspondence: Kenneth.enevoldsen@cas.au.dk

Abstract

Large-scale datasets are foundational for research and development in natural language processing. However, current approaches face three key challenges: (1) reliance on ambiguously licensed sources restricting use, sharing, and derivative works; (2) static dataset releases that prevent community contributions and diminish longevity; and (3) quality assurance processes restricted to publishing teams rather than leveraging community expertise.
To address these limitations, we introduce two contributions: the Dynaword approach and Danish Dynaword. The Dynaword approach is a framework for creating large-scale, open datasets that can be continuously updated through community collaboration. Danish Dynaword is a concrete implementation that validates this approach and demonstrates its potential. Danish Dynaword contains over five times as many tokens as comparable releases, is exclusively openly licensed, and has received multiple contributions across industry, the public sector and research institutions. The repository includes light-weight tests to ensure data formatting, quality, and documentation, establishing a sustainable framework for ongoing community contributions and dataset evolution.

Danish Dynaword is available at: `https://huggingface.co/datasets/danish-foundation-models/danish-dynaword`

Keywords: Pre-training, Large-scale corpora, Continual development

1. Introduction

Continuously developed open-source projects are instrumental for contemporary research and lay the foundation for the success of our fields. These projects range from foundational software such as NumPy (Harris et al., 2020), to datasets like the Universal Dependencies Treebank (Nivre et al., 2016), to recent contributions including Flash Attention (Dao et al., 2022) and LoRA (Hu et al., 2022). In open-source communities, it is understood that a project is never complete but is continually enhanced and adjusted in response to advances in the field and the software ecosystem. It is similarly recognized that relying on proprietary technology may result in legal ramifications, which, in the worst case, can render projects unusable or lead to their removal, undermining community contributions. We have already seen this impact in several instances: Udio AI Music Generator was shut down citing legal concerns; a state-of-the-art Danish encoder (Enevoldsen et al., 2023) was removed following threats of legal action; and the Nordic Pile corpus (Öhman et al., 2023) was never released, presumably due to copyright issues.

However, open-source values remain largely absent from (pre-)training datasets in the field. Although we have seen large-scale releases (Gao et al., 2020b; Xue et al., 2020), these datasets often adhere to the pattern of being released once without updates. Even when we see continual releases (Penedo et al., 2024; Abadji et al., 2022), they are frequently based on Common Crawl content. Baack et al. (2025) defines this category

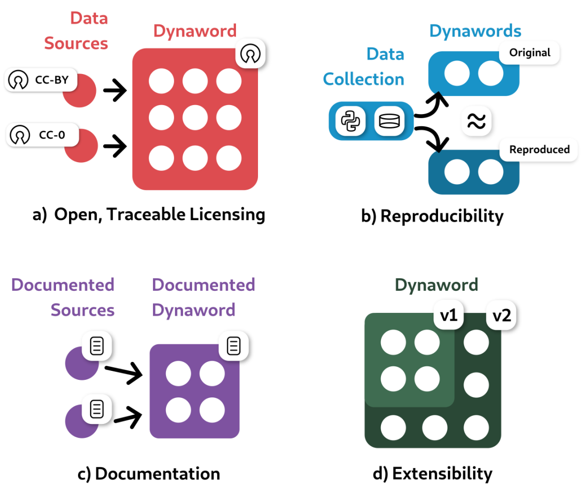

Figure 1: Overview of the guiding principles for Dynaword corpora

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3782–3793
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

as *open access* and notes several legal risks associated with the use of both datasets and derived models. Baack et al. (2025) also defines *openly licensed* data, which enables the resharing, reuse, and modification of data, thereby providing a solid foundation for derivative works. There have been a few initiatives (Langlais et al., 2025; Langlais, 2024); however, despite being a step forward, these contributions still have notable shortcomings.

These datasets (Langlais et al., 2025; Langlais, 2024) constitute a single release without reproducible code for collection. This makes it difficult to reproduce the data collection process and difficult to improve[1] or update[2]. Lastly, a vague description of the underlying licenses makes it difficult to validate the license claims. For instance, stating that "Alice in Wonderland" is public domain differs from documenting that the author died in 1898, which renders it within the public domain; we refer to this as a *traceable* license.

Following these limitations, we propose the **dynaword** approach for curating a corpus:

- **Traceable and open licensing**: All datasets within the corpus must be *openly licensed* and maintain a *traceable* license.
- **Reproducibility**: It should be possible to derive a substantially similar dataset[3].
- **Documented**: The dataset should be well-documented under best practices in the field (Gebru et al., 2021).
- **Extensible**: Extending and improving the corpus should be possible, and methods for doing so should be documented.

These guidelines are intended to provide datasets that are compatible with the open-source AI definition[4], FAIR (Wilkinson et al., 2016), and compliant under a diverse set of legal frameworks, including the European Parliament Artificial Intelligence Act (European Union, 2024), and conducive to creating lasting resources for both research and industry. This approach draws upon successful datasets such as the Universal Dependency Treebanks (Nivre et al., 2016), which remain essential building blocks for linguistic analysis, model development (Honnibal et al., 2020), and recently, multilingual benchmarking of LLMs (Nielsen, 2023; Nielsen et al., 2024).

These guidelines were developed jointly with the Danish Dynaword, which acts as a practical implementation and a testbed, showing that the guidelines are implementable and not simply ideals. We believe Danish provides an ideal case, being a low-to-mid-resource[5] language, which has enough contributors willing to participate, yet well-contained in scope. High-resource languages[6] could likely sustain multiple dynaword projects targeting specific domains such as coding, or healthcare.

The following sections outline how we developed a framework around Danish Dynaword that actively ensures traceable and open licensing and reproducibility, while being documented and reproducible. Our final dataset represents more than a fivefold increase in Danish tokens available compared to previous datasets and considerably improves reproducibility and documentation. We expect that this work will continue to expand with future contributions.

2. Related Works

2.1. Continually released pre-training data

Recently, dataset releases have been increasingly iterative releases with improvements to extraction, cleaning, and collection procedures. These include OSCAR (Abadji et al., 2022), HPLT (Aulamo et al., 2023; Arefyev et al., 2024), and fineweb (Penedo et al., 2024), all which are mostly based on Common Crawl[7]. Though Common Crawl and its derivatives represent significant artifacts, the license of the underlying data remains ambiguous, raising a slew of ethical and legal concerns (Baack et al., 2025). Additionally, these sources typically work with web-derived content, overlooking sources such as APIs, television, or radio.

2.2. Openly licensed data

Data licensing is often unclear, and it is not uncommon to find datasets on Hugging Face[8] published under permissive licenses containing copyrighted content. In some cases (Abadji et al., 2022; Penedo et al., 2024) the license applies only to the packaging, metadata, and annotations, but not to the data itself. To avoid such confusion, we will, throughout this work, utilize the three tiers defined by Baack et al. (2025): 1) *replicable* - enables reproduction - 2) *open access* - enables download -

[1] For instance, the Common Corpus consists of multiple OCR'd documents that could likely be improved with recent advancements.

[2] E.g. the number of tokens on Wikipedia is expected to grow linearly (Suh et al., 2009).

[3] Denoted as a 'substantially equivalent system' in the Open-Source AI definition: https://opensource.org/ai/drafts/the-open-source-ai-definition-1-0-rc1

[4] https://opensource.org/ai/drafts/the-open-source-ai-definition-1-0-rc1

[5] Class 2-4 as defined by Joshi et al. (2021)

[6] Class 5 as defined by Joshi et al. (2021)

[7] https://commoncrawl.org

[8] https://huggingface.co

and 3) *openly licensed* - enables resharing, reuse, and modification.

Examples of openly licensed data sources intended for pre-training include YouTube Commons (Langlais, 2024) and Common Corpus (Langlais et al., 2025) as well as multiple gigaword projects (Graff and Cieri, 2003; Derczynski et al., 2021; Adewumi et al., 2020). These gigaword corpora consist of 1B (Giga) tokens available under permissive licenses. YouTube Commons compiles CC-BY licensed YouTube content, including 30B tokens of predominantly English (70%) transcripts. The most extensive openly licensed corpus is the 500B Common Corpus, which includes digitized newspapers and OCR'd public domain books. While these collections solved many issues, they left much to be desired. The OCR quality is often questionable, and the lack of public processing workflows makes these datasets hard to reproduce, validate and improve.

3. Dataset Collection

Danish Dynaword takes its outset in public segments of the Danish Gigaword (Derczynski et al., 2021), previously the largest publicly available data source for Danish. Danish Dynaword notably excludes social media data from Twitter (32M tokens), copyrighted samples from OpenSubtitles (<1M tokens), and common crawl segments (100M tokens) following the principles of *traceable and open licensing*. We also exclude DanAvis (30M tokens) due to a lack of overall coherence caused by sentence scrambling. For each source in this collection, a datasheet (Gebru et al., 2021) was compiled, incorporating available information, including dataset description and license references.

Furthermore, additional open datasets were added using a process of identifying potentially openly licensed datasets, collecting them, conducting quality checks, and, lastly, reviewing their licenses in case of uncertainty. Quality checks were intentionally kept minimal to allow for downstream filtering and include verifying that the text is Danish, coherent, and readable. In the following, we will walk through each of these steps.

Identification: The majority of openly licensed datasets were found through projects such as the Danish Foundation Models (Enevoldsen et al., 2023), on the HuggingFace Hub, or through Sprogteknologi.dk, a website covering Danish language resources, curated by the Danish Ministry of Digitization. Additional resources were identified through social media outreach, personal communication, and issues in the Danish Dynaword repository.

License and content review: After identification, maintainers review the data and filter out straightforward cases where the data is not Danish, is not *openly licensed*, or the redistributor lacks permission to license or re-license it. For complex cases, maintainers may request legal advice from faculty services.

Collection and Quality Checks: After a dataset is checked, it is collected and undergoes quality checks. The collection procedure is documented in the form of datasheets and reproducible scripts. These scripts enable dataset updates and critical examination. Some sources excluded at this stage include the Danish subsection of Common Corpus, where the OCR was deemed insufficient (alpha ratio generally below 0.7), with most text being unreadable. If rejected, the quality issues are documented, and the issue is closed. For all sources, we deduplicate and remove short documents.

3.1. Inclusion Policy

In recent years, we have seen the rise of derived data in pre-training regimes. These data include synthetic (Li et al., 2023; Gunasekar et al., 2023), semi-synthetic (Chung et al., 2024), translations (Doshi et al., 2024), OCR'd (Langlais et al., 2025), and transcribed data (Langlais, 2024). These sources require deliberation before inclusion, as they can also lead to model degradation (Bender et al., 2021; Shumailov et al., 2024).

While it is clear that an inclusion policy is likely to change as technologies improve, the current dynaword does not, to our knowledge, contain synthetic, machine-translated, or automatically transcribed data, but does include human-annotated audio transcriptions (e.g., FTSpeech) and translations by expert translators (e.g., Europarl) and OCR'd documents (e.g., NCC books). For OCR'd documents, we perform an additional quality check; these are described in the individual datasheets.

3.2. Evaluation data

It is by no mean uncommon that large public dataset include segments of evaluation data (Gao et al., 2020a), this can lead overestimate of actual performance (Schaeffer, 2023; Deng et al., 2024) and thus leading to overestimates of the model capabilities. It is therefore encourages that model developers exclude evaluation data during the training process. To facilitate this we mark datasets contained in benchmarks. For the Danish Dynaword this includes, the Danish dependency treebank, though not the annotation, which is used to create the dataset ScaLA (Nielsen, 2023) and Nordjyllands News, which is used for evaluating summarization (Nielsen, 2023; Nielsen et al., 2024) as well as semantic similarity (Enevoldsen et al., 2024, 2025).

Dataset	Size Tokens	Contributions Is there a clear process for contributions	Tiers of Dataset Openness Replicable Enables the user to reproduce.	Open Access Enables the user to download.	Openly Licensed Enables the user to reuse, share, and modify.
Tier 3: Openly Licensed					
Danish Dynaword (v1.2.12)	5.6B	Yes	✓	✓	✓
Danish Gigaword (Derczynski et al., 2021)	~1B	No		✓	✓
Common Corpus (dan) (Langlais et al., 2025)	~0.3B	No		✓	✓
Tier 2: Open Access					
SnakModel (Zhang et al., 2025)	~13.6B	No	✓	✓	
Fineweb (dan) (Penedo et al., 2024)	~26B	(Yes)[1]	✓	✓	

Table 1: Comparison of Danish Language Datasets. Danish Gigaword only includes segments that are currently publicly available.
[1]Has a Discord channel and encourages involvement.

3.3. Contributions

Danish Gigaword has already seen multiple contribution. We show this development in Figure 2. These contributions stem from companies, government institutions, municipalities, individuals and universities and covers a wide array of background ranging from cultural heritage to computational linguistics. We believe that this shown that there is both a need and interest for more dynaword projects.

4. Dataset Overview

Table 2 shows the overview of datasets within Danish Dynaword, including a short description, license, and size. This table reflect the datasets available at the time of the training experiments, for an updated view, we recommend checking out the dataset repository[9].

5. Dataset Comparison

In Table 1 we give a conceptual comparison of Danish Dynaword to existing openly available datasets and in Table 3 we show the performance gap when training on Danish Dynaword instead of the Danish Gigaword, which is we eloborate on in the following section.

We present an overview of the datasets in Appendix 4. For individual datasheets, we refer to the dataset repository.

5.1. Training Experiments

To study relative quality Danish Dynaword and Danish Gigaword (Strømberg-Derczynski et al., 2021) we perform set of training experiments to estimate the expected language modelling performance.

The experiments were performed using the Gemma-1B model either continually pre-trained[10] or trained from scratch. To ensure a fair comparison we train two models on Dynaword, one *matched* in size to Danish Gigaword and one trained on the *full* dataset.

We evaluated the perplexity performance on four datasets from Dynaword: DDT, JVJ, Synnejysk.dk, and Nordjyllands News which were held out during training. Additionally, we tested on contemporary sources not included in Dynaword: News articles from DR (dr.dk) and Danish Wikipedia articles published after January 1, 2025[11].

All models were trained with a maximum sequence length of 6144 tokens, an effective batch size of 32 (via gradient accumulation), and an initial learning rate of 10^{-5} for pretrained models and 10^{-3} for the models from scratch, in both cases we use a cosine learning rate scheduler. Models were trained using the Danish Dynaword version 1.2.0 and with training code available on GitHub[12]. An overview of models is provided in Appendix B.

Table 3 shows the perplexity across the six held-out datasets. Comparing Dynaword to Gigaword, we see an average relative improvement of 5.9% for continual pre-training starting from Gemma-3-1b-pt, and an improvement of 26% for Gemma-3-1b models trained from scratch. Even in the size-matched scenario, Dynaword yields improvements of 2.6% with continual pre-training and 18% when training from scratch. Results from a downstream evaluation on Danish tasks from EuroEval (Nielsen, 2023) show that continual pre-training on Danish Dynaword yields improvements on 7 out of 9 tasks. Detailed results are provided in Appendix C.

[9]https://huggingface.co/datasets/danish-foundation-models/danish-dynaword

[10]Using the gemma-3-1b-pt checkpoint.

[11]The wikipedia section was updated in v1.2.10 and now includes the latest articles.

[12]https://github.com/schneiderkamplab/offpolicy_kd/, Commit 76b546e

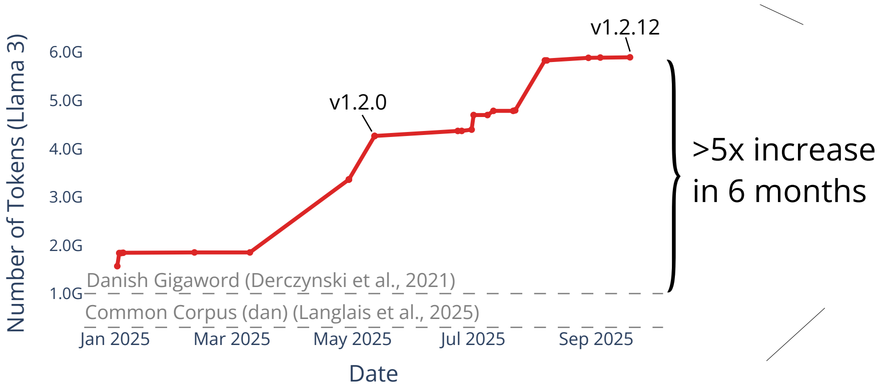

Figure 2: Number of tokens in Danish Dynaword over time.

6. Conclusions

In this work, we argue for continuously developed datasets. We dub such datasets dynawords and outline four key principles which are required for continuous development; a) Open and traceable licensing, b) Reproducibility, c) Documentation, and d) Extensibility. As a testbed and concrete implementation, we release Danish Dynaword, the largest *openly licensed* Danish corpus, which we expect to grow with future submissions. Danish Dynaword has already received contributions from multiple parties, including industry, private individuals, and research. For the datasets in the dynaword, we also provide reproducible scripts to update them and a compliance review of the licensing.

To enable future contributions, Danish Dynaword comes with lightweight tests to ensure formatting, documentation, and dataset quality. Danish Dynaword has already received contributions from multiple parties, including industry, private individuals, and research institutions.

While Dynaword remains an order of magnitude smaller than non-openly licensed sources, it provides a sustainable foundation for building models. Dynawords should be seen as a complement to existing efforts, making high-quality and permissible data available. We hope that this dynaword can be a blueprint for future dynawords targeting other languages or domains.

7. Limitations

Size Danish Dynaword significantly expands available *openly-licensed* data, with future growth expected from initiatives like Dansk Sprogmodel Konsortium (DSK)[13] and national AI programs[14]. However, it remains an order of magnitude smaller than the Common Crawl datasets (Penedo et al., 2024; Abadji et al., 2022). This gap may persist, but could be addressed through multilingual or multimodal sources.

Coverage and bias: Danish Dynaword, given its requirements, is biased toward domains with clear licensing. Thus, the data set only contains limited amounts of social media data and a disproportionate amount of legal documents. You can explore the coverage by domain in Figure 3.

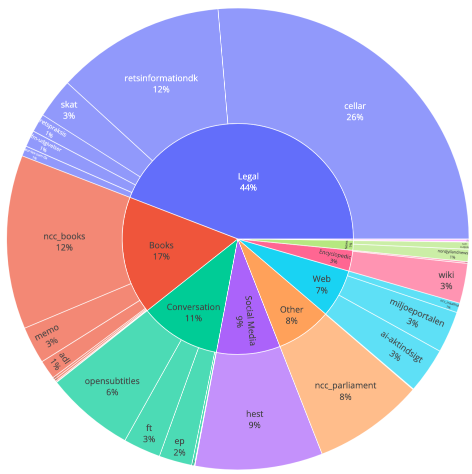

Figure 3: Content by domain. The inner circle shows the domain, while the outer layers are the source.

Only Danish: Dynaword presents a methodology for developing large-scale public language resources. However, in this work, we only present one such resource, namely, for Danish. We hope that Danish Dynaword can act as a starting point for similar efforts for other languages.

Review quality and dataset poisoning: Throughout the development process, it became clear that contributing minor changes, such as filtering out a few bad examples, was difficult, both due to the limited support for reviewing large data changes. While previous projects (e.g. Nivre et al. (2016))

[13] https://alexandra.dk/dsk/

[14] https://digst.dk/strategier/strategi-for-kunstig-intelligens/

Source	Description	Size	License
Legal			
Cellar	EU legal documents and open data	1.15B	CC-BY-SA 4.0
retsinformation.dk	The legal information system of Denmark	516.35M	Copyright Law
Skat.dk	The Danish tax authority website	122.11M	CC-0
Retspraksis	Case law or judicial practice in Denmark	56.26M	CC-0
EUR-lex SUM	EU legislation with summaries	31.37M	CC-BY-SA 4.0
Social Media			
Heste-nettet.dk	Danish Debate forum	389.32M	CC-0
Spoken			
Opensubtitles	Movie Subtitles from OpenSubtitles	271.60M	CC-0
FT	Meeting records from the Danish parliament	114.09M	CC-0
Spont	Conversational samples from research project	1.56M	CC-0
NAAT	Danish speeches from 1930-2022	286.68K	CC-0
Web			
Municipal Websites	Municipality websites from AI-aktindsigt	139.23M	Apache 2.0
Miljoeportalen	Environmental Reports from Miljøportalen	127.38M	CC-0
FM Udgivelser	Publication of the Ministry of Finance	50.34M	CC-BY-SA 4.0
NCC Maalfrid	Danish content from Norwegian institutions	29.26M	NLOD 2.0
Encyclopedic			
Wikipedia	The Danish subsection of Wikipedia	122.00M	CC-0
Europarl	The Danish subsection of Europarl	100.84M	CC-0
WikiSource	The Danish subsection of Wikisource	5.34M	CC-0
Books and Novels			
NCC Books	OCR'ed Danish books	531.97M	CC-0
MeMo	Novels from the Modern Breakthrough	113.74M	CC-BY-SA 4.0
ADL	Danish literature from 1700-2023	58.49M	CC-0
Gutenberg	Books from Project Gutenberg	6.76M	Gutenberg
WikiBooks	The Danish Subsection of Wikibooks	6.24M	CC-0
JVJ	The works of Johannes V. Jensen	3.55M	CC-BY-SA 4.0
News			
Nordjylland News	Articles from Newspaper TV2 Nord	37.90M	CC-0
TV2R	Articles from TV2R	21.67M	CC-BY-SA 4.0
NCC Newspapers	OCR'd Newwspapers from NCC	1.057M	CC-0
Dialect			
Botxt	Dictionary of the dialect Bornholmsk	847.97K	CC-0
Synnejysk.dk	Dataset of the dialect Sønderjysk	52.02K	CC-0
Other			
NCC Parliament	OCR'ed Danish from the Norwegian parliament	338.87M	NLOD 2.0
Nota	The text segment from readaloud data	7.30M	CC-0
DanNet	A Danish WordNet	1.48M	DanNet 1.0
Religious texts	Religious text from the 1700-2022	1.24M	CC-0
DDT	The Danish Dependency Treebank	185.45K	CC-BY-SA 4.0
Total		4.37B	

Table 2: Overview of the dataset in Danish Dynaword (v1.2.0). Size in Llama 3 tokens. For the latest version we refer to the dataset repository.

have tackled this issue using human-readable formats, this is likely inefficient at the current scale.

This lack of clarity increased the likelihood of dataset attacks such as dataset poisoning (Goldblum et al., 2022). We expect to see both interface development and software development to detect and prevent such attacks and ensure review quality.

7.1. Ethical consideration

Despite our effort to prevent issues, large-scale dataset development often involves seemingly *openly-licensed* datasets that may contain copyrighted content. A notable instance occurred with the initial release of OpenSubtitles, which was part of the Danish Gigaword (Derczynski et al., 2021). By providing a versioned dataset along with

	Held out Datasets				Contemporary (2025)	
	Treebank DDT	**Fiction** JVJ	**Dialect** Synnejysk	**News** Nordjylland	**Wikipedia** Wiki (dan)	**News** DR
Reference baseline						
Gemma-3-1b-pt	16.0	33.9	62.8	9.8	9.2	9.7
Continual pre-training						
Gigaword*	14.2	29.1	54.1	8.5	8.1	9.2
Dynaword* (matched)	14.0 (+1.2%)	26.6 (+8.5%)	51.8 (+4.3%)	8.4 (+1.2%)	8.2 (-0.7%)	9.1 (+0.9%)
Dynaword* (full)	13.5 (+4.6%)	25.2 (+13%)	50.1 (+7.5%)	8.1 (+4.6%)	8.0 (+1.9%)	8.9 (+3.5%)
Pre-training from scratch						
Gigaword*	48.8	128	219	24.4	26.7	29.4
Dynaword* (matched)	43.0 (+12%)	86.2 (+33%)	144 (+34%)	21.0 (+14%)	26.0 (+3.7%)	25.4 (+14%)
Dynaword* (full)	39.2 (+20%)	79.4 (+38%)	128 (+42%)	19.6 (+21%)	23.2 (+13%)	23.6 (+20%)

Table 3: Perplexity (lower is better) for Gemma-3-1b models continually pre-trained (middle) and pre-trained from scratch (bottom) on Gigaword, Dynaword (size-matched to Gigaword), and the full Dynaword dataset (v1.2.0). Relative performance is calculated with respect to the Gigaword baseline. *The four validation datasets were excluded from the training data.

a changelog, we clearly indicate when sources are excluded, thereby promoting transparency.

8. Acknowledgments

Part of the computation done for this project was performed on the UCloud interactive HPC system, which is managed by the eScience Center at the University of Southern Denmark. The Danish e-Infrastructure Consortium (DeIC) granted the compute resources (DeiC-AU-N1-2025144 and DeiC-AU-N1-2025118). This project was made possible by funding from the Danish Government to Danish Foundation Models (4378-00001B). A special thanks goes to organizations that have publicly made their data available under openly licensed regimes and to individuals and organizations that have encouraged this development.

9. Bibliographical References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a Cleaner Document-Oriented Multilingual Crawled Corpus. ArXiv:2201.06642 [cs].

Tosin P Adewumi, Foteini Liwicki, and Marcus Liwicki. 2020. Corpora compared: The case of the swedish gigaword & wikipedia corpora. *arXiv preprint arXiv:2011.03281*.

Nikolay Arefyev, Mikko Aulamo, Pinzhen Chen, Ona De Gibert Bonet, Barry Haddow, Jindřich Helcl, Bhavitvya Malik, Gema Ramírez-Sánchez, Pavel Stepachev, Jörg Tiedemann, Dušan Variš, and Jaume Zaragoza-Bernabeu. 2024. HPLT's first release of data and models. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 2)*, pages 53–54, Sheffield, UK. European Association for Machine Translation (EAMT).

Mikko Aulamo, Nikolay Bogoychev, Shaoxiong Ji, Graeme Nail, Gema Ramírez-Sánchez, Jörg Tiedemann, Jelmer Van Der Linde, and Jaume Zaragoza. 2023. Hplt: High performance language technologies. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 517–518.

Stefan Baack, Stella Biderman, Kasia Odrozek, Aviya Skowron, Ayah Bdeir, Jillian Bommarito, Jennifer Ding, Maximilian Gahntz, Paul Keller, Pierre-Carl Langlais, Greg Lindahl, Sebastian Majstorovic, Nik Marda, Guilherme Penedo, Maarten Van Segbroeck, Jennifer Wang, Leandro von Werra, Mitchell Baker, Julie Belião, Kasia Chmielinski, Marzieh Fadaee, Lisa Gutermuth, Hynek Kydlíček, Greg Leppert, E. M. Lewis-Jong, Solana Larsen, Shayne Longpre, Angela Oduor Lungati, Cullen Miller, Victor Miller, Max Ryabinin, Kathleen Siminyu, Andrew Strait, Mark Surman, Anna Tumadóttir, Maurice Weber, Rebecca Weiss, Lee White, and Thomas Wolf. 2025. Towards Best Practices for Open Datasets for LLM Training. ArXiv:2501.08365 [cs].

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2024. Investigating data contamination in modern benchmarks for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8706–8719, Mexico City, Mexico. Association for Computational Linguistics.

Leon Derczynski, Manuel R Ciosici, Rebekah Baglini, Morten H Christiansen, Jacob Aarup Dalsgaard, Riccardo Fusaroli, Peter Juel Henrichsen, Rasmus Hvingelby, Andreas Kirkedal, Alex Speed Kjeldsen, et al. 2021. The danish gigaword corpus. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 413–421.

Meet Doshi, Raj Dabre, and Pushpak Bhattacharyya. 2024. Pretraining language models using translationese. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5843–5862, Miami, Florida, USA. Association for Computational Linguistics.

Kenneth Enevoldsen, Isaac Chung, Ashwin Mathur, Imene Kerboua, Márton Kardos, David Stap, Jay Gala, Wissam Siblini, Saba Sturua, Saiteja Utpala, Gabriel Sequeira, Marion Schaeffer, Mathieu Ciancone, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Orion Weller, Chenghao Xiao, Ömer Çag, Bhavish Pahwa, Rafał Poswiata, Shawon Ashraf, Björn Plüster, Jan Philipp Harries, Daniel Auras, Kranthi Kiran Gv, Isabelle Mohr, Dawei Zhu, Martin Bernstorff, Hippolyte Gisserot-Boukhlef, Taemin Lee, Jan Kostkan, Andrianos Michail, Manuel Faysse, Mohammed Hamdy, Roberta Rocca, Manan Dey, Dipam Vasani, Simone Tedeschi, Pranjal Chitale, Saksham Thakur, Roman Solomatin, Nguyen Tai, Artem Snegirev, Mariya Hendriksen, Michael Günther, Anna Maksimova, Silvan Wehrli, Jordan Clive, Maria Tikhonova, Aleksandr Abramov, Henil Panchal, Alena Fenogenova, Lester James Miranda, Alessia Borghini, Zheng Liu, Lasse Hansen, Simon Clematide, Malte Ostendorff, Guangyu Song, Wen-Ding Li, Ruqiya Bin Safi, and Federico Cassano. 2025. MMTEB: Massive Multilingual Text Embedding Benchmark.

Kenneth Enevoldsen, Lasse Hansen, Dan S Nielsen, Rasmus AF Egebæk, Søren V Holm, Martin C Nielsen, Martin Bernstorff, Rasmus Larsen, Peter B Jørgensen, Malte Højmark-Bertelsen, et al. 2023. Danish foundation models. *arXiv preprint arXiv:2311.07264*.

Kenneth Enevoldsen, Márton Kardos, Niklas Muennighoff, and Kristoffer Laigaard Nielbo. 2024. The scandinavian embedding benchmarks: Comprehensive assessment of multilingual and monolingual text embedding. Neurips. ArXiv: 2406.02396 [cs.CL].

European Union. 2024. Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending regulations (ec) no 300/2008, (eu) no 167/2013, (eu) no 168/2013, (eu) 2018/858, (eu) 2018/1139 and (eu) 2019/2144 and directives 2014/90/eu, (eu) 2016/797 and (eu) 2016/798. Entered into force: 1 August 2024.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020a. The pile: An 800gb dataset of diverse text for language modeling.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020b. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Micah Goldblum, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, Aleksander Mądry, Bo Li, and Tom Goldstein. 2022. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1563–1580.

David Graff and C Cieri. 2003. English gigaword, linguistic data consortium. *Philadelphia*, 4(1):34.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature*, 585(7825):357–362.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International conference on learning representations*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2021. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. *arXiv:2004.09095 [cs]*. ArXiv: 2004.09095.

Pierre-Carl Langlais. 2024. Releasing Youtube-Commons: a massive open corpus for conversational and multimodal data.

Pierre-Carl Langlais, Carlos Rosas Hinostroza, Mattia Nee, Catherine Arnett, Pavel Chizhov, Eliot Krzystof Jones, Irène Girard, David Mach, Anastasia Stasenko, and Ivan P. Yamshchikov. 2025. Common corpus: The largest collection of ethical data for LLM pre-training.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.

Dan Saattrup Nielsen. 2023. ScandEval: A Benchmark for Scandinavian Natural Language Processing. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 185–201.

Dan Saattrup Nielsen, Kenneth Enevoldsen, and Peter Schneider-Kamp. 2024. Encoder vs decoder: Comparative analysis of encoder and decoder language models on multilingual NLU tasks. *arXiv preprint arXiv:2406.13469*.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, and Natalia Silveira. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Joey Öhman, Severine Verlinden, Ariel Ekgren, Amaru Cuba Gyllensten, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, and Magnus Sahlgren. 2023. The nordic pile: A 1.2 tb nordic dataset for language modeling. *arXiv preprint arXiv:2303.17183*.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. ArXiv:2406.17557 [cs].

Rylan Schaeffer. 2023. Pretraining on the test set is all you need. *arXiv preprint arXiv:2309.08632*.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Leon Strømberg-Derczynski, Manuel Ciosici, Rebekah Baglini, Morten H. Christiansen, Jacob Aarup Dalsgaard, Riccardo Fusaroli, Peter Juel Henrichsen, Rasmus Hvingelby, Andreas Kirkedal, Alex Speed Kjeldsen, Claus Ladefoged, Finn Årup Nielsen, Jens Madsen, Malte Lau Petersen, Jonathan Hvithamar Rystrøm, and Daniel Varab. 2021. The Danish Gigaword corpus. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 413–421, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Bongwon Suh, Gregorio Convertino, Ed H Chi, and Peter Pirolli. 2009. The singularity is not near: slowing growth of wikipedia. In *Proceedings of the 5th international symposium on wikis and open collaboration*, pages 1–10.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Mike Zhang, Max Müller-Eberstein, Elisa Bassignana, and Rob van der Goot. 2025. SnakModel: Lessons learned from training an open Danish large language model. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 812–825, Tallinn, Estonia. University of Tartu Library.

A. Author Contributions

Table 4 shows authorship contributions across different categories. All authors have agreed to the final version of the print.

Conceptualization *Idea, narrative, planning*	**Writing** *Original draft*	**Writing** *Review and editing*
Kenneth Enevoldsen Jan Kostkan Desmond Elliott Kristoffer Nielbo Kristian Nørgaard Jensen	Kenneth Enevoldsen	Kenneth Enevoldsen Márton Kardos Kristoffer Nielbo Per Møldrum Dalum
Data curation *Download, select, process*	**Data review** *Licensing*	**Software** *CI, testing*
Kenneth Enevoldsen Kristian Nørgaard Jensen Jan Kostkan Balázs Szabó Peter Vahlstrup	Per Møldrum-Dalum	Kenneth Enevoldsen Kristian Nørgaard Jensen
Dataset Comparison *Model training, evaluation*	**Dataset Contributors** *Curation, annotation, documentation*	**Supervision and Funding** *Ideation, planning*
Andrea Blasi Núñez Gianluca Barmina Jacob Nielsen Rasmus Larsen Lukas Galke Peter Schneider-Kamp	Kirsten Vad Johan Heinsen Rob Van der Goot	Peter Schneider-Kamp Kristoffer Nielbo

Table 4: Paper contributions.

Model name	**checkpoint**	**trained on**
gemma-3-1b-cpt-gigaword-v1	gemma-3-1b-pt	Danish Gigaword*
gemma-3-1b-cpt-dynaword-matched-v1	gemma-3-1b-pt	Danish Dynaword* (matched)
gemma-3-1b-cpt-dynaword-full-v1	gemma-3-1b-pt	Danish Dynaword* (full)
gemma-3-1b-scratch-gigaword-v1	random init.	Danish Gigaword*
gemma-3-1b-scratch-dynaword-matched-v1	random init.	Danish Dynaword* (matched)
gemma-3-1b-scratch-dynaword-full-v1	random init.	Danish Dynaword* (full)

Table 5: Overview of trained models. *The four validation datasets (DDT, JVJ, Synnejysk, Nordjylland) were excluded from the training data.

B. Trained Models

Table 5 provides an overview of the trained models for the dataset comparison including access links. Despite models are limited in size to 1B, the results show that Danish Dynaword yields gains on 7 out of 9 downstream tasks.

C. Downstream Evaluation

Tables 6 and 7 report the scores of the continually pre-trained models on the Danish part of the EuroEval benchmark (Nielsen, 2023). Results show the expected gains

Task (→) Score (→)	**angry-tweets** Sentiment (MCC)	**dansk** NER (Micro F1)	**scandiqa-da** Reading comp. (F1)	**da-talemaader** Knowledge (Accuracy)	**da-citizen-tests** Knowledge (Accuracy)
Reference baseline					
Gemma-3-1b-pt	37.27 ± 2.28	14.55 ± 0.86	51.80 ± 3.95	23.28 ± 4.58	40.78 ± 3.28
Continual Pre-training					
Gigaword*	36.58 ± 3.04	15.19 ± 1.47	48.10 ± 1.46	25.47 ± 6.19	39.67 ± 4.03
Dynaword* (matched)	**38.80** ± 1.84	**17.97** ± 1.79	**50.04** ± 3.33	**35.31** ± 5.41	35.22 ± 4.00
Dynaword* (full)	**38.80** ± 1.92	**16.30** ± 2.43	**49.07** ± 3.43	**32.19** ± 4.04	36.22 ± 3.23

Table 6: Downstream results on the Danish subsection of EuroEval(NLU) of Gemma-3-1b models continually pre-trained (bottom) on Gigaword, Dynaword (size-matched to Gigaword), and the full Dynaword dataset. Scores are reported alongside standard error of the mean. Values marked in bold indicate an increase over the Gigaword baseline. *All the evaluation datasets were excluded from all training runs.

Task (→) Score (→)	**nordjylland-news** Summarization (BertScore)	 (Rouge-L)	**hellaswag-da** Common sense (Accuracy)	**scala-da** Linguistic acceptability (MCC)
Reference baseline				
Gemma-3-1b-pt	56.93 ± 1.76	11.31 ± 1.60	24.61 ± 2.19	0.84 ± 4.43
Continual Pre-training				
Gigaword*	47.36 ± 3.97	7.77 ± 0.72	24.77 ± 2.10	1.27 ± 6.79
Dynaword* (matched)	46.40 ± 3.00	7.59 ± 0.74	**26.84** ± 2.23	0.74 ± 3.85
Dynaword* (full)	47.06 ± 2.86	**7.84** ± 0.69	**24.92** ± 2.28	**1.86** ± 3.66

Table 7: Downstream results on the Danish subsection of EuroEval(NLG) of Gemma-3-1b models continually pre-trained (bottom) on Gigaword, Dynaword (size-matched to Gigaword), and the full Dynaword dataset. Scores are reported alongside standard error of the mean. Values marked in bold indicate an increase over the Gigaword baseline. *All the evaluation datasets were excluded from all training runs.