Lessons Learned from GPT-SW3: Building the First Large-Scale Generative Language Model for Swedish

Ariel Ekgren[1], Amaru Cuba Gyllensten[2], Evangelia Gogoulou[2], Alice Heiman [1], Severine Verlinden[1], Joey Öhman[1], Fredrik Carlsson[2], Magnus Sahlgren[1]

[1]AI Sweden,[2]RISE

{ariel.ekgren, alice.heiman, severine.verlinden, joey.ohman, magnus.sahlgren}@ai.se
{amaru.cuba.gyllensten, evangelia.gogoulou, fredrik.carlsson}@ri.se

Introduction

We present GPT-SW3, a 3.5 billion parameter autoregressive language model, trained on a newly created 100 GB Swedish corpus. This paper provides insights with regard to data collection and training process, and discusses the challenges of proper evaluation. The results of quantitive evaluation using perplexity indicate that GPT-SW3 is a competent model in comparison with existing autoregressive models of similar size. Additionally, we perform an extensive prompting study which reveals the good text generation capabilities of GPT-SW3.

Model

GPT-SW3 is an autoregressive language model that uses the same model architecture as GPT-2 [4] as implemented by the Megatron framework [5], i.e. a Transformer [6] based, decoder-only architecture, trained on the next-step prediction task. With 3.6 billion parameters, it is, to date, the largest language model available for Swedish. Table 1 below shows the model parameters used for GPT-SW3. Figure 2 illustrates the size of GPT-SW3 and other large scale language models.

Parameter	Value
Transformer layers	30
Attention heads	32
Sequence length	2,048
Embedding dimension	3,072
Total parameters	3,559,415,808

Tab. 1: List of model parameters values for GPT-SW3

Dataset

To train the model we compiled a 100 GB corpus of Swedish text. This corpus consist of existing Swedish corpora (e.g. OSCAR, Wikipedia), as well as novel, targeted, data sources (e.g. Fass, consisting of pharmaceutical factsheets). Figure 1 illustrates the dataset composition of our corpus.

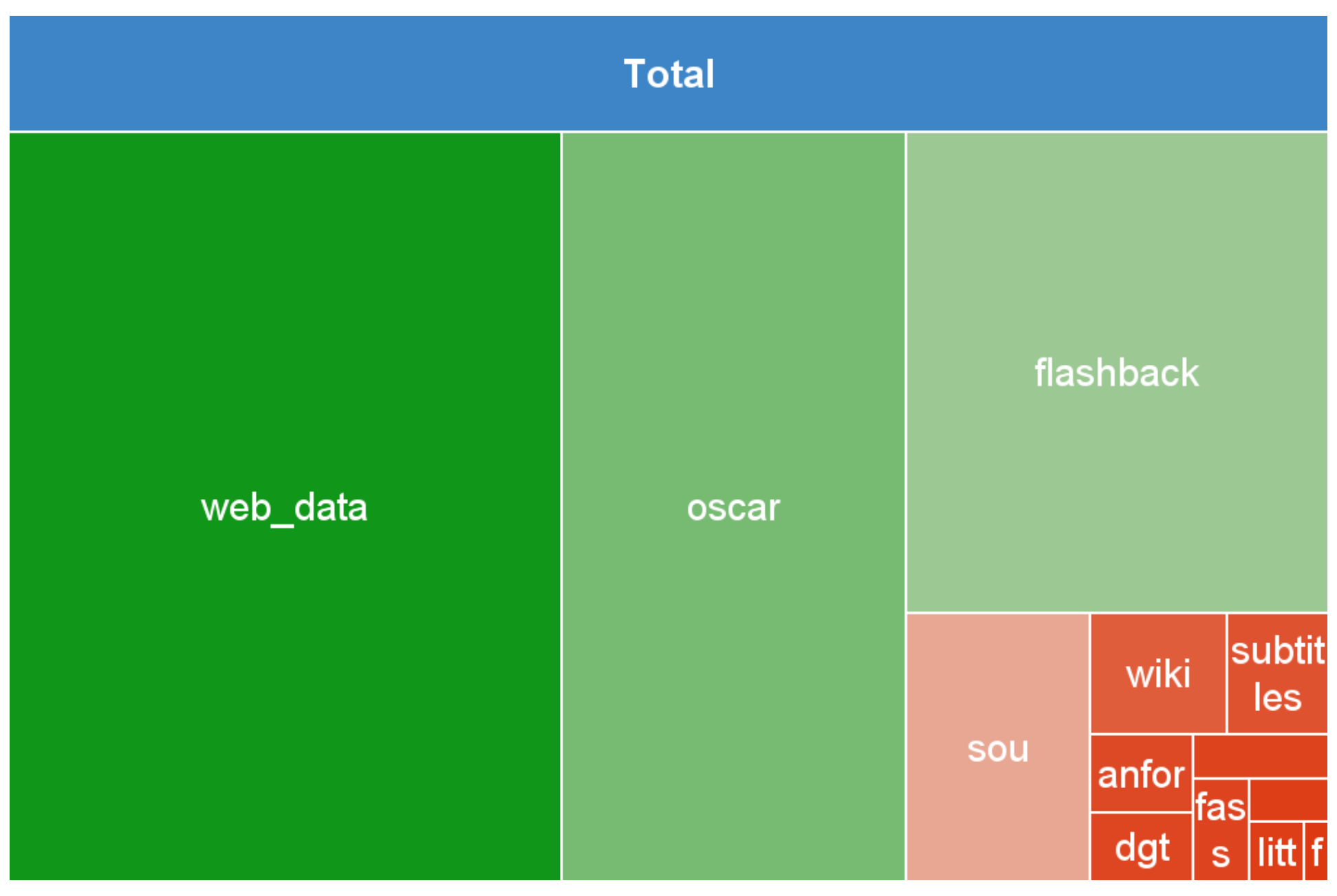

Fig. 1: Dataset distribution. Totalling 100 GB of (mostly) Swedish text data

Parameter comparison of LLMs

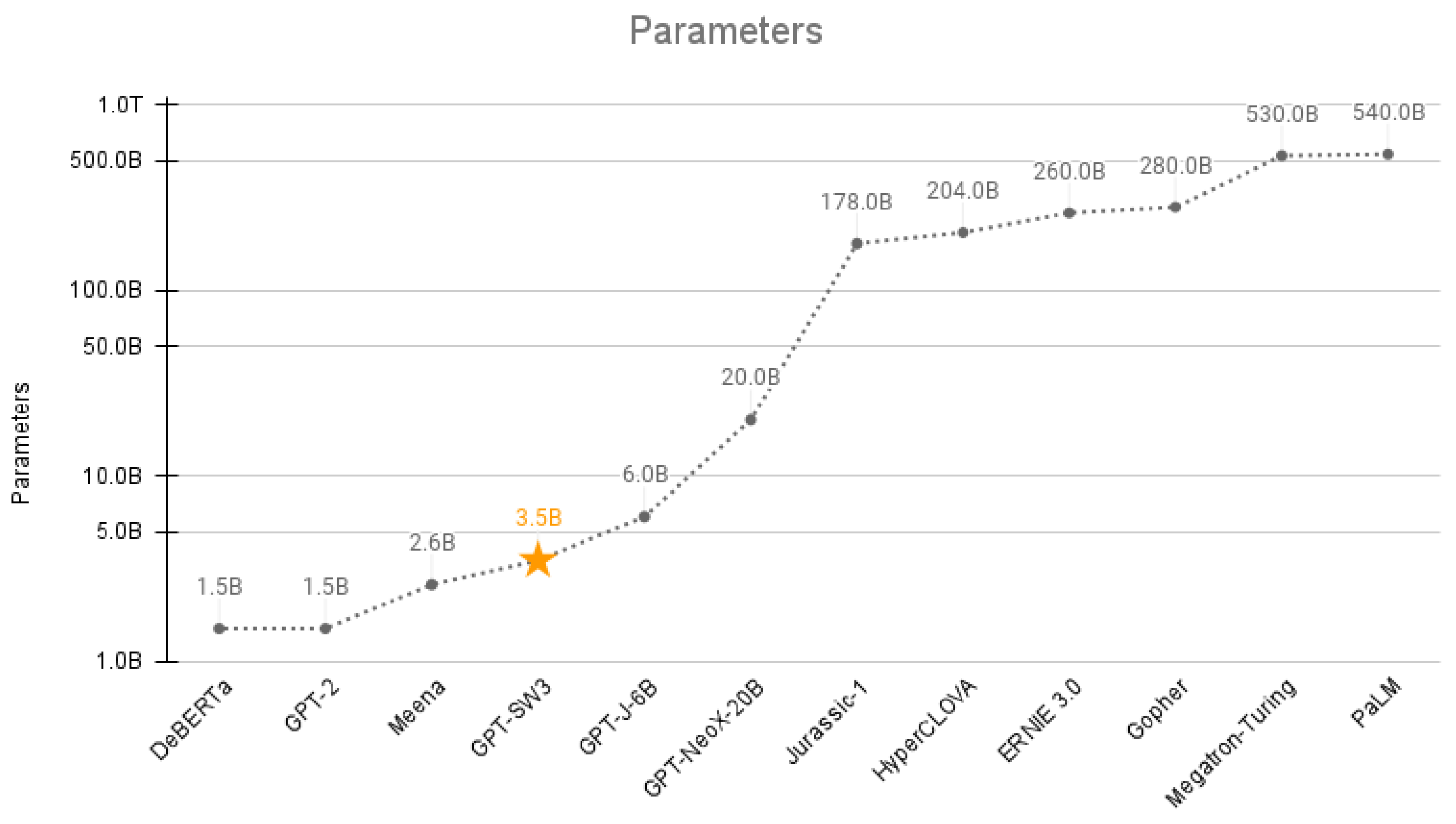

Fig. 2: Parameter counts of LLMs

Evaluation

Most work on generative language models focus either on intrinsic measures such as perplexity, or on their ability to solve downstream tasks in zero- or few-shot settings. Since a ready-to-use Swedish framework for evaluation of GPT models on downstream tasks is still work in progress, we opted for perplexity as a quantitative measure of model performance.Table 2 shows quantitative perplexity evaluation, while Tables 3 and 4 show prompting examples for email generation and Swedish to English translation, respectively.

Model		PPL_c sv	PPL_c en
Flashback-GPT	[3]	4.76	6.22
GPT2-xl	[4]	4.70	2.30
GPT-Neo	[1]	2.87	**2.11**
GPT-SW3		**2.19**	2.70

Tab. 2: Character level perplexity [2] comparison on mc4.

Prompting examples

Skriv mail till Peter från Sara. Fråga om picknick med familjen. Behöver hjälp med lunchen.	***Write mail to Peter from Sara. Ask about family picnics. Needs help with lunch***.
Hej Peter! Vi skulle behöva hjälp med att fixa en picknick för hela familjen på lördag. Vi har inte möjlighet att ta med egen mat, så det skulle vara väldigt uppskattat om du kunde hjälpa oss med det. Vänliga hälsningar, Sara	*Hi Peter!* *We could use some help organising a picnic for the whole family on Saturday. We are not able to bring our own food, so it would be very much appreciated if you could help us with that.* *Kind regards,* *Sara*

Tab. 3: Example response from GPT-SW3 tasked with generating an email from a **prompt**.

Translation examples

Original Sentence	Type	Translation
Jag mår bra idag. (*Eng: I feel good today.*)	Z-s	I feel good today.
	O-s	I feel good today.
	F-s	I feel good today.
I norden blir det kallt om vintrarna. (*Eng: In Scandinavia, it gets cold during the winters.*)	Z-s	The winter is warm in the winters.
	O-s	In northern Europe, it gets warm if winter is coming.
	F-s	In northern Europe, it gets cold in the winter.
Tror du verkligen att alla får plats? (*Eng: Do you really think everyone fits?*)	Z-s	I don't think so.
	O-s	I think you really think that everyone is allowed.
	F-s	Do you really believe that everyone is able to be in the room?
Efterfrågan på elbilar har ökat dramatiskt på senare år. (*Eng: The demand for electric cars has increased dramatically in recent years.*)	Z-s	Det finns en stor efterfrågan på elbilar.
	O-s	Elbilsmarknaden har exploderat de senaste åren.
	F-s	The demand for electric vehicles has increased dramatically since the beginning of the last decade.

Tab. 4: Examples of GPT-SW3 translations from Swedish to English of the same sentence in a zero-shot (Z-s), one-shot (O-s), and few-shot (F-s) setting

Conclusion

- Training large language models presents considerable opportunities and challenges, especially in a low or lower resource setting like Swedish. Challenges arise in every step from data collection, training resources to evaluation. Evaluation is particularly hard in languages other than English, due to the lack of standardised evaluation benchmarks
- A general prompt follows the following structure: context, task description, few-shot examples, followed by an incomplete sentence urging the model to continue generating new content.
- The model copies structure. Provide a framework around your content to help the model generate more similar content.

References

[1] Sid Black et al. “GPT-Neo: Large scale autoregressive language modeling with mesh-tensorflow”. In: *If you use this software, please cite it using these metadata* 58 (2021).

[2] Ryan Cotterell et al. “Are All Languages Equally Hard to Language-Model?” In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 536–541. DOI: 10.18653/v1/N18-2085. URL: https://aclanthology.org/N18-2085.

[3] Tobias Norlund and Agnes Stenbom. “Building a Swedish Open-Domain Conversational Language Model”. In: *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*. Reykjavik, Iceland (Online): Linköping University Electronic Press, Sweden, May 2021, pp. 357–366. URL: https://aclanthology.org/2021.nodalida-main.38.

[4] Alec Radford et al. “Language models are unsupervised multitask learners”. In: *OpenAI blog* 1.8 (2019), p. 9.

[5] Mohammad Shoeybi et al. “Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism”. In: *CoRR* abs/1909.08053 (2019). arXiv: 1909.08053. URL: http://arxiv.org/abs/1909.08053.

[6] Ashish Vaswani et al. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.