DOSMo-7B

A large language model trained exclusively on German Text

Summary

DOSMo-7B is an open 7 billion parameter large language model (LLM) trained on 1T tokens of exclusively German text. In contrast to existing approaches, which typically improve the German skills of LLMs with continued pretraining, we perform from scratch pretraining to explore the potential of training LLMs with only German text.

Data Pipeline

RedPajama v2 · MADLAD-400 · OSCAR · mC4 · Wikipedia · Textbooks · YouTube Subtitles

Normalization
We convert documents and metadata to a unified schema and apply NFC Unicode normalization.

Deduplication
We perform global deduplication on the document level using MinHash + LSH to find duplicate documents with at least 80% overlap.

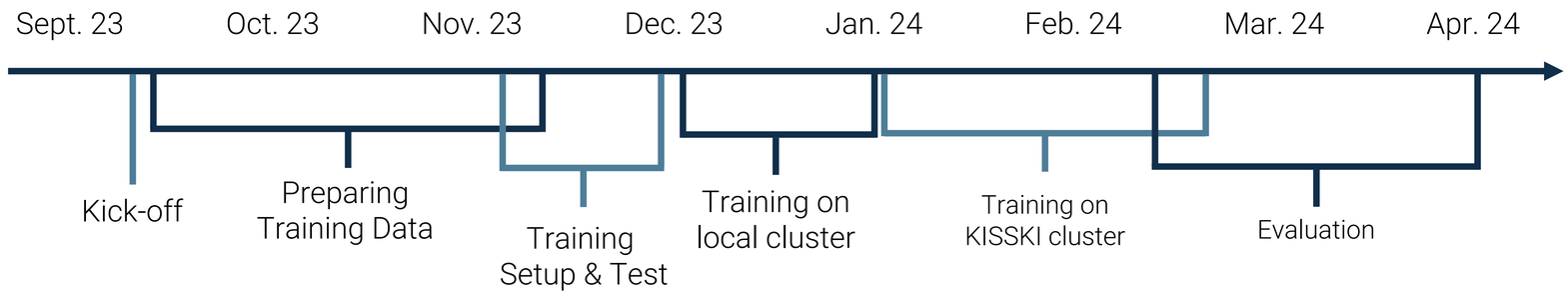
Language Identification
Although the sources are pre-filtered for German language, we identify the language of each document and remove all documents for which German is not in the top-3 predictions.

Quality Filtering
We compute a set of quality signals tailored to German text. We keep all documents that meet a list of filter criteria:

Signal	Threshold
Number of words	≥ 50
Mean word length (chars)	≥ 3 and ≤ 10
Frac. symbol chars	≤ 0.5
Frac. whitespace chars	≤ 0.5
Frac. numerical chars	≤ 0.5
...	...
Num. German stopwords	≥ 2
Num. German cookie policy words	≤ 2
Num. English cookie policy words	≤ 2
Frac. chars in top 2-grams	≤ 0.2
Frac. chars in top-3 grams	≤ 0.2
...	...
Frac. chars in duplicate 5-grams	≤ 0.2
...	...

≈1B documents, ≈1T tokens

Timeline



Architecture

DOSMo-7B uses the same transformer architecture as Mistral-7B for performance and efficiency reasons. This also allows DOSMo-7B to be used as a drop-in replacement. Mistral-7B was the best publicly available LLM when we started this work.

Tokenizer

We train a custom byte-pair encoding (BPE) tokenizer on a subset of our pretraining dataset to maximize the encoding efficiency for German text. We measure the encoding efficiency on holdout data.

Architecture	Mistral-7B-v0.1
Sequence Length	8192
Tokenizer	DOSMo-7B
Training Library	Composer
Parallelization	FSDP
Precision	bfloat16
Attention-Impl.	FlashAttention v2
Batch Size	512 (≈4M tokens)
Training Steps	237K
GPU hours	84K (A100)
Interconnect	400 Gbit/s
Throughput/GPU	≈3300 tokens/s

Tokenizer	Tokens/Byte
Mistral 7B	≈ 0.34
LLaMa-2	≈ 0.32
DOSMo-7B	≈ 0.22

Compute Resources

For most of the pretraining, we used 128 A100 GPUs provided by the GWDG through the KISSKI project. The data preparation, early-stage training, and evaluation were done on our local L3S GPU cluster.

Evaluation

We evaluate DOSMo-7B by relying on common English LLM benchmarks that were translated to German. DOSMo shows promising performance on some benchmarks . However, it does not outperform other open models on translated benchmarks. Note that DOSMo was trained on much fewer tokens, and translated benchmarks might not be very meaningful.

Model	MMLU_de	HellaSwag_de	ARC_de
LLaMa-2-7B	0.40	0.51	0.43
Mistral-7B	**0.52**	0.59	0.47
LeoLM-LLaMa2-7B	0.40	0.61	0.43
LeoLM-Mistral-7B	0.48	**0.66**	**0.48**
DOSMo-7B	0.39	**0.66**	0.45

Example Generations

- In Göttingen kann man eine Vielzahl von interessanten Aktivitäten unternehmen und Sehenswürdigkeiten besichtigen. Die historische Altstadt mit ihren wunderschönen Fachwerkhäusern und engen Gassen lädt zu ausgedehnten Spaziergängen ein. Als bedeutende Universitätsstadt bietet Göttingen zahlreiche kulturelle Veranstaltungen, Museen und Galerien, die einen Besuch wert sind. Ein Highlight ist sicherlich die Georg-August-Universität, eine der ältesten und renommiertesten Hochschulen Deutschlands, deren Campus man erkunden ...

- Folgende Sehenswürdigkeiten gibt es in Göttingen zu sehen:
  Das Alte Rathaus: Ein beeindruckendes Gebäude aus dem 14. Jahrhundert im Stil der Gotik.
  Die Gänseliesel-Figur: Das Wahrzeichen der Stadt und beliebter Treffpunkt auf dem Marktplatz.
  Die Georg-August-Universität: Eine der ältesten und renommiertesten Universitäten Deutschlands.
  Das Städtische Museum: Bietet Einblicke in die Geschichte und Kultur der Stadt.
  ...

L3S

Maximilian Idahl
Forschungszentrum L3S
Appelstraße 4
30167 Hannover
idahl@l3s.de