Llama 3.1 Paper

Pretraining :- ① filtering of large scale training corpus

② Devlopment of model Architecture & Scaling laws for model size

③ efficient large scale pretraining techinque development

④ pretraining recipe development

Dataset : end of 2023

Removed domains that has personal info of People and adult content, unsafe contents

Used HTML for high quality diverse text.

built custom parser, that extracts HTML boiler plate removal, content recall

Used HTML → text extractor like.

jina.ai

found → Markdown is harmful to model performance. that is trained on web data than plane text.

Several rounds of deduplication

Heuristic filtering :→ to remove low quit quality document

Multilingual Data :→ followed Similar pipeline
Added filter to remove personal data

→ language identification using fast text

→ document & line level deduplication

→ Applied language level heuristics } remove
model based filtering } low quality documents

Data Mix

Knowledge classification → categorize types of info
↳ down sample data categories that are over represented

Scaling laws :-

llama 3.1 paper [continuation]

→ to reduce memory cost → deallocate the Tensors that will not be used for further computation including input & Output tensors in each pipeline

How to figure out which tensors are not being used?

→ Developed memory consumption estimator and performance Projection tool — to explore parallelism config and overall training performance and identify memory gap

PG 12

Training :- ① initial pre training —— Adam, $\alpha = 8 \times 10^{-5}$
② long content pretraining — 800B tokens, 128K content
③ Annealing

Llama 3.1 paper [continuation]

→ to reduce memory cost → deallocate the Tensors that will not be used for further computation including input & output tensors in each pipeline

How to figure out which tensors are not being used?

→ Developed memory consumption estimator and performance projection tool – to explore parallelism config and overall training performance and identify memory gap

PG 12

Training :- ① initial pre training —— Adam, $\alpha = 8 \times 10^{-5}$
② long content pretraining – 800B tokens, 128K content
③ Annealing