

**Figure 5: Training Loss for Llama 2 models.** We compare the training loss of the Llama 2 family of models. We observe that after pretraining on 2T Tokens, the models still did not show any sign of saturation.

**Tokenizer.** We use the same tokenizer as Llama 1; it employs a bytepair encoding (BPE) algorithm (Sennrich et al., 2016) using the implementation from SentencePiece (Kudo and Richardson, 2018). As with Llama 1, we split all numbers into individual digits and use bytes to decompose unknown UTF-8 characters. The total vocabulary size is 32k tokens.

### 2.2.1 Training Hardware & Carbon Footprint

**Training Hardware.** We pretrained our models on Meta’s Research Super Cluster (RSC) (Lee and Sengupta, 2022) as well as internal production clusters. Both clusters use NVIDIA A100s. There are two key differences between the two clusters, with the first being the type of interconnect available: RSC uses NVIDIA Quantum InfiniBand while our production cluster is equipped with a RoCE (RDMA over converged Ethernet) solution based on commodity ethernet Switches. Both of these solutions interconnect 200 Gbps end-points. The second difference is the per-GPU power consumption cap — RSC uses 400W while our production cluster uses 350W. With this two-cluster setup, we were able to compare the suitability of these different types of interconnect for large scale training. RoCE (which is a more affordable, commercial interconnect network)

|         | Time<br>(GPU hours) | Power<br>(W) | Consumption<br>(tCO <sub>2</sub> eq) | Carbon Emitted       |
|---------|---------------------|--------------|--------------------------------------|----------------------|
| Llama 2 | 368640              | 400          | 62.44                                | 34B 1038336          |
| 7B      | 184320              | 400          | 31.22                                | 13B 291.42           |
|         |                     |              |                                      | Total 3311616 539.00 |

**Table 2: CO<sub>2</sub> emissions during pretraining.** Time: total GPU time required for training each model. Power Consumption: peak power capacity per GPU device for the GPUs used adjusted for power usage efficiency. 100% of the emissions are directly offset by Meta’s sustainability program, and because we are openly releasing these models, the pretraining costs do not need to be incurred by others.

can scale almost as well as expensive Infiniband up to 2000 GPUs, which makes pretraining even more democratizable.

**Carbon Footprint of Pretraining.** Following preceding research (Bender et al., 2021a; Patterson et al., 2021; Wu et al., 2022; Dodge et al., 2022) and using power consumption estimates of GPU devices and carbon efficiency, we aim to calculate the carbon emissions resulting from the pretraining of Llama 2 models. The actual power usage of a GPU is dependent on its utilization and is likely to vary from the Thermal Design Power (TDP) that we employ as an estimation for GPU power. It is important to note that our calculations do not account for further power demands, such as those from interconnect or non-GPU server power consumption, nor from datacenter cooling systems. Additionally, the carbon output related to