Lec 15   Tour of modern LLMs

Licensing

open — described — closed

weights — inference — training code — data

public domain, CC-0   do whatever
|
MIT, BSD   few restriction
|
Apache, CC-BY   must acknowledge owner
|
GPL, CC-BY-SA   must acknowledge & use same license
|
CC-NC   no commercial (no longer open source)
:
LLaMa, OPEN-RAIL   other restrictions
  └ can't use to generate data to train models not derived from llama
: └ can't use for military
No license   all rights reserved

* Fair use (country dependent)
  - Small quote usually ok
  - Doesn't diminish commercial value
  - For non-commercial usually ok

  Many LLM training assume fair use (but do get sued)

* Why closed model

  - Commercial
  - Safety
  - Legal liability (just not let ppl know training data)

What makes a model

  - Archi ← possibly less important than data, if all using similar archi
  - Data
  - Training

Open models

▷ Pythia —   fully open   — eleuther ai
  └ lots of checkpoints
  └ largest being 7B
  └ 2k context len, parametric layer norm
  └ 300 B tokens on The Pile or 207B after deduplication
  Finding
  - larger model remember things more quickly
  - intervention on data can reduce gender bias

* The Pile dataset
  └ PubMed, ArXiv, GitHub, Wikipedia, ...

▷ OLMo — possibly strongest open model — by AI2
  └ instruction tuned
  └ non-parametric layer norm
  └ 2.46T tokens on Dolma corpus
  └ competitive performance

* Dolma dataset (3T tokens)
  └ comes with processing pipeline

▷ LLaMa2   — by Meta
  └ strong safety + chat tuning
  └ data not known
  └ 2T tokens (upsample fact rich info)
  └ Reward model (for safety tuning)
    1. Collect preference data (internal)
    2. Train model to predict human preference
    3. Train LM with RLHF
  └ Chat
    - System message : obey throughout conversation
    - Synthesize dataset by attaching system to user input
      then true model to generate some thing without that

▷ Mistral / Mixtral   — by Mistral
  └ speed optimisations, MoE, sliding window attention, GQA, has
    European langs

▷ Qwen — by Alibaba
  └ strong multilingual
  └ attention layer has bias
  └ large vocab (150k)

Special Purpose Models

▷ Code model

  → StarCoder 2 — fully open
  → CodeLlama   — by Meta
  → DeepSeek coder — by DeepSeek

▷ Math Models

  → LLeMa   — open, by EleutherAI, proof pile
  → DeepSeek math — math stuff on internet

▷ Science Model

  → Galactica   — by Meta

Closed models

  - GPT-4
    └ image input
    └ can call external functions
    └ multimodality
  - Gemini
    └ long input
    └ video input
  - Claude 3