TIDELIFT

Open(ish) ML

Where are we? Where are we going?

TIDELIFT

(not Tidelift)

(but I love our stock Slides template)

State of the art: April

State of the art: November

Open(ish) ML: a survey

- Why open(ish)?
- Background on ML
- What is like traditional open?
- What *isn't* like traditional open?
- Hot topics, briefly: RAIL and Copilot

Why open*(ish)?*

This is not “traditional” open

- The **tech** is very different
- That means all of these are also different:
  - **Participation**
  - **Governance**
  - **Regulation**
  - **Economics (not enough time here but critical!)**

Is it “open”? *I don’t care.*

1. Tech is very different, so “open” can obscure as much as illuminate.
2. Which “open”? Copyleft? Permissive? Corporate? individual? Etc.? Already wasn’t one-size-fits-all
3. Change still coming fast and furious
4. I’m so tired. So, so tired.

The world’s best crafters of square pegs are on this call

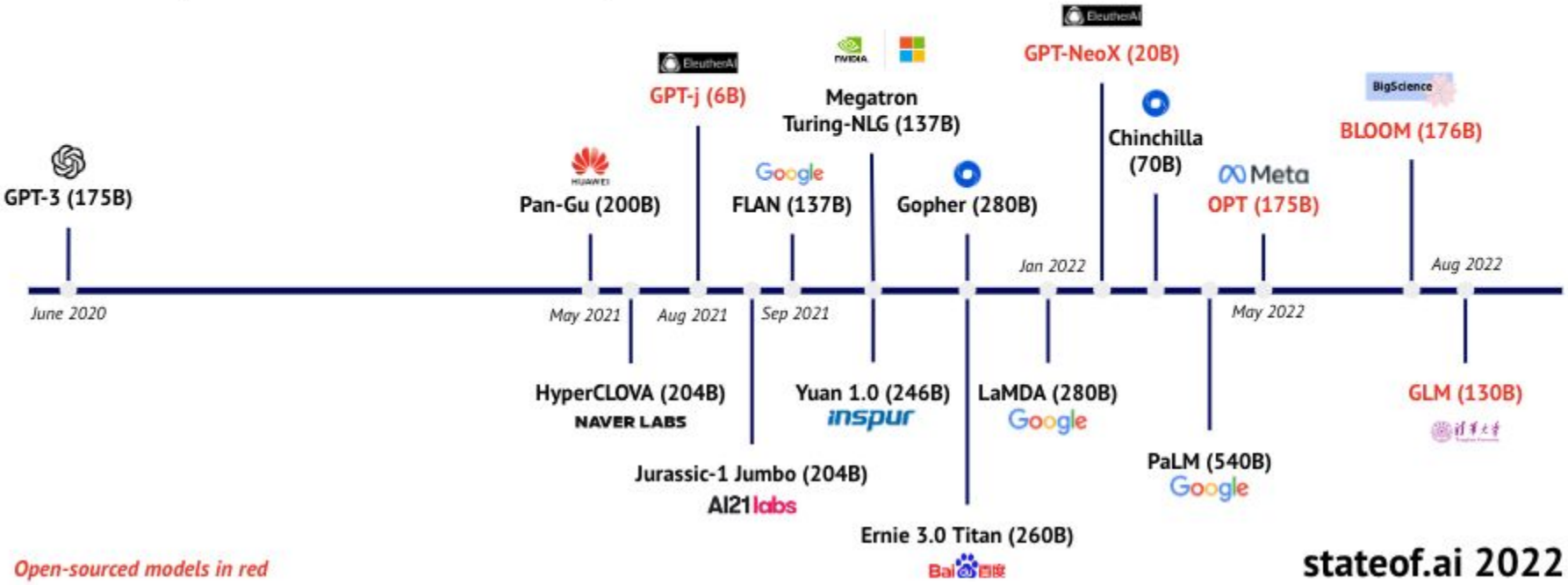
So: **open*ish*:**
inspired by traditional open —
but not (yet?) bound by our rules

BACKGROUND

Open(ish) AI: State of Play?

Closed for 14 months: community-driven open sourcing of GPT *et al.*

- **Landmark models from OpenAI and DeepMind have been implemented/cloned/improved by the open source community much faster than we'd have expected.**



Open *definitions and licenses*
are trying frantically to keep up

BACKGROUND

What is ML?: Technical processes

Highest, highest level

- **Good news:** It's not *that* impossible to understand for lawyers.
  - Core insight: it's "just" very complicated probabilities.
- **Bad news:** Still very different from traditional tech.

Pipeline and artifacts

- Variety of stages, with different meanings for each
- But! Start by comparing to traditional software compilation:
  - Source code + complex tech (compiler) → binaries

Stage 1: get data

- Gather *vast* piles of data.
  - 3Tb: moderate size.
  - 240Tb: state of the art openish image set.
- Optional(?): clean data
  - Depending on techniques, data, and needs, may now be optional

Stage 2: process data

- Turn human-comprehensible data into numbers that can be trained on
  - May include “feature detection”
- Often done with traditionally-open software tools
  - But like open compiler, output may still be proprietary and reflect developer choices

Stage 3: train

- Darkest magic happens here!
- Output: the **model**
  - **multi-dimensional numeric matrix**
  - (“parameters” and “weights” but ignore those!)
- Often uses traditionally-open tools (PyTorch, Tensorflow) + very custom configuration/code

Stage 4: Deployment

- [Not very interesting legally, just know the model then has to be deployed and run on real-world systems, much like a binary]

Stage 5: Inference

- Translate an input into a number, feed it into the probability matrix, get another number back
- Requires post-processing to create an **output artifact**
- Again, frameworks often open but specific implementations often closed
  - Can be highly optimized

Summary of artifacts that might be licensed

- Data
- Code
  - training, inference
- Model
  - “Just” a pile of numbers

So, not “is it open”, rather “is *what* open?”

- Is the *training* framework open?
  - Similar to “is the compiler open”
- Is the *model* open?
  - was mostly about use/deployment, but modification becoming easier
- Is the data open?
  - Complicated!

What is ML?: Technical capacities

What can these magic numbers *do*?

- Help write novels (soon: maybe just write them?)
- Copilot is definitely in your organization
- Lawyers will use it to write drafts of some documents within 2-3 years

Thinking bigger

- *Bear* case: web + mobile
- Bull case: the printing press? (openml.fyi/printing)
- More concretely:
  - Linux/Apache/MySQL in the late 1990s.
  - Your engineers want this and may already be using it.

What is ML?: Legal + regulatory

Different tech means different legal structures

- **Licenses:** Model is *not* preferred form for modification, so...?
- **Ethics:** when training on garbage in, garbage out is a very big problem
- **Regulation:** EU moving ahead full-speed
  - Not the tech-friendly Clinton White House

Parallels with traditional open?

What feels like “traditional” open?

- ***libre?*****:** current labor has big focus on ethics
- Feels like ~~teen spirit~~ the 90s
  - High velocity
  - High excitement
  - Lots of *creative* value (unlike blockchain)
- Barriers to access are falling quickly
  - But not uniformly, lots of grey areas

Failure to parallel?

What doesn’t feel “traditional”?

- Community focus empowerment *balanced by constraints*
- Data and model are *not code*
  - Training is definitely not compiling
  - Collaboration is different
  - Wikipedia is considered a *small* data set
  - Organic growth is hard, because you need all the data at once

Hot topics!

RAIL License

- GPL v1(?) of the early open(ish) AI community?
- Used by several big projects
- Tries to jam all of criminal law, tort law, and human rights law into a one-page appendix

More:
blog.tidelift.com/evaluating-the-rail-license-family

Copilot litigation

- Training: an achilles heel for ML?
- 1202: an achilles heel for fair use?
  - Contrast EU data mining right
- Class action certification
- What parties will have similar attacks on other training models?


for more, weekly(ish):
https://openml.fyi