When can a small language model answer a simple question?...

Roger Lam
lamroger.com

OpenELM from Apple ranges from 270M - 3B parameters

- OpenELM is trained with open datasets and open source code
- This provides a good baseline for the current reasonable state-of-the-art for Open Small Language Models
- Running locally is relatively easy - install a couple python packages, accept TOS for the Llama tokenizer, and run a one-liner
- Allows us to run a simple experiment

OpenELM isn't the only open SLM

- OpenELM outperforms OLMO with less pretraining tokens
- PyThia is surprisingly good with an order of magnitude less training token

Model	Public dataset	Open		Model size	Pre-training tokens	Average acc. (in %)
		Code	Weights			
OPT [55]	✗	✓	✓	1.3 B	0.2 T	41.49
PyThia [5]	✓	✓	✓	1.4 B	0.3 T	41.83
MobiLlama [44]	✓	✓	✓	1.3 B	1.3 T	43.55
OLMo [17]	✓	✓	✓	1.2 B	3.0 T	43.57
OpenELM (Ours)	✓	✓	✓	1.1 B	1.5 T	**45.93**

Table 1. **OpenELM vs. public LLMs.** OpenELM outperforms comparable-sized existing LLMs pretrained on publicly available datasets. Notably, OpenELM outperforms the recent open LLM, OLMo, by 2.36% while requiring 2× fewer pre-training tokens. The average accuracy is calculated across multiple tasks listed in Tab. 3b, which are also part of the OpenLLM leaderboard [4]. Models pretrained with less data are highlighted in gray color.

OpenELM 270M - Completion

- Prompt: "The capital of California is "
- Parameters were about 1GB

```
================================================================================
 Prompt + Generated Output
--------------------------------------------------------------------------------
The capital of California is 100 miles from San Francisco, and the largest city in the state is
 Sacramento.
Sacramento is a city that has been around for over 200 years. It was founded by the Spanish exp
lorer Hernando de Soto in 1569.
The city's name comes from the Spanish word "sacramento" which means "the place of the holy wat
er".
Sacramento is also known as the "City of Lights" because of its many lights.
Sacramento is located on the banks of the Sacramento River.
It is one of the most popular tourist destinations in the United States.
Sacramento is home to the University of California at Sacramento.
Sacramento is also home to the Sacramento Kings basketball team.
Sacramento is also home to the Sacramento Kings baseball team.
Sacramento is also home to the Sacramento Kings football team.
Sacramento is also home to the Sacramento Kings ice hockey team.
Sacramento is also home to the Sacramento Kings soccer team.
```

OpenELM 450M - Completion

- Prompt: "The capital of California is "
- Parameters were about 1.8GB

```
================================================================================
 Prompt + Generated Output
--------------------------------------------------------------------------------
The capital of California is 30 miles long and 10 miles wide. It is the largest city in the Uni
ted States, with a population of 1,000,000 people.
The state's economy is based on tourism, agriculture, manufacturing, and services. The state ha
s a large number of natural resources, including oil, gas, coal, and minerals.
California is home to many famous landmarks, such as the Golden Gate Bridge, the San Francisco
Bay, the Napa Valley, and the Monterey Bay Aquarium.
The state is also known for its beautiful beaches, which are located in the Pacific Ocean. Ther
e are many different types of beaches in California, including sandy beaches, surf beaches, and
 beach resorts.
The state is also known for its wine, which is produced in many different regions. Wine is one
of the most popular drinks in the United States, and it is often consumed by people of all ages
.
The state is also known for its food, which is made from fresh ingredients. Many different type
s of food are available in California, including Mexican food, Italian food, Chinese food,
--------------------------------------------------------------------------------
```

OpenELM 1.1B - Completion

- Prompt: "The capital of California is "
- Parameters were about 4.3GB

```
================================================================================
Prompt + Generated Output
--------------------------------------------------------------------------------
The capital of California is 100% a city. It's not a state, it's not a county, it's not a town,
 it's not a village, it's not a hamlet, it's not a borough, it's not a district, it's not a par
ish, it's not a commune, it's not a municipality, it's not a township, it's not a ward, it's no
t a suburb, it's not a village, it's not a hamlet, it's not a borough, it's not a district, it'
s not a parish, it's not a commune, it's not a municipality, it's not a township, it's not a wa
rd, it's not a suburb, it's not a village, it's not a hamlet, it's not a borough, it's not a di
strict, it's not a parish, it's not a commune, it's not a municipality, it's not a township, it
's not a ward, it's not
--------------------------------------------------------------------------------
```

OpenELM 3B - Completion

- Prompt: "The capital of California is "
- Parameters were about 12GB

```
=================================================================================
 Prompt + Generated Output
---------------------------------------------------------------------------------
The capital of California is 100% in the state.

The state of California has a population of 39,528,000 people.

The city of Los Angeles has a population of 3,876,000 people.

So that means that 100% of the population of California lives in the state of California.

If you want to be more specific, then you can say that 100% of the population of Los Angeles li
ves in the city of Los Angeles.

But if you want to be even more specific, then you can say that 100% of the population of Los A
ngeles lives in the county of Los Angeles.
```

Completion models are kinda bad

- From 270M to 3B, they spew garbage basically…
- Is the key to move from useless to useful to add instruction fine-tuning?

270M - Instruct

Looks better but wrong.

```
================================================================================
 Prompt + Generated Output
--------------------------------------------------------------------------------

[INST]What is the capital of California?[/INST]

### Answer

San Francisco, California

The capital of California is the largest city in the United States. The official name of California is Republi
c of California, and its official language is American Sign Language (ASL). It is bordered by Arizona, Nevada,
 and Oregon. The state's economy is dominated by tech, finance, fintech, healthcare, finance, and fintech serv
ices.

It has an estimated population of 3,199,800 as of 2020. The median household income is $91,699, making it one
of the wealthiest U.S. states. The San Francisco Bay Area is home to Silicon Valley, which attracts tech compa
nies looking for skilled labor
```

450M - Instruct

Better! But I dropped the [INST] tokens. (Not clear what instruct format to use)

```
==================================================================================
 Prompt + Generated Output
----------------------------------------------------------------------------------

What is the capital of California?
California's capital city is Sacramento.
Capital City: Sacramento
Population: 4,092,758
Area: 3,612,700 square miles
GDP (nominal): $22.2 trillion
State Symbol: C
Capital District: Sacramento County
History: The capital of California was originally Fort Bidwell, built in 1776 by Governor Thomas Jefferson to
protect against Native Americans and later named for Sacramento's first governor, John Bidwell. In 1809, Sacra
mento became the state capital when it replaced San Francisco.
Geography: Located at the confluence of the American and Pacific Rivers, Sacramento sits on the eastern slope
of the Sierra Nevada Mountains. It covers an area of 3,612,700 square miles, making it California's second-lar
gest state. The region includes the Central Valley, with its rich agricultural lands, and the Northern Califor
```

Instruction fine-tuning seems very valuable

We see immediate improvement.

Is this what LLMs are under-the-hood? Having a bunch of humans enumerate the best responses to all possible questions?

Is this the true data flywheel? Hire a bunch of experts to generate fine-tuning data. For less stringent data, use preference data (human feedback, AI feedback).

Is this why Meta is pushing Llama 3 into all their products

Bonus: Phi-3 Mini Instruct (3.8B)

- Fine-tuned on preference and safety
- 3.3T tokens, curated public data, synthetic textbooks, high quality chat fine-tune