Vintage LLMs

OWAIN EVANS

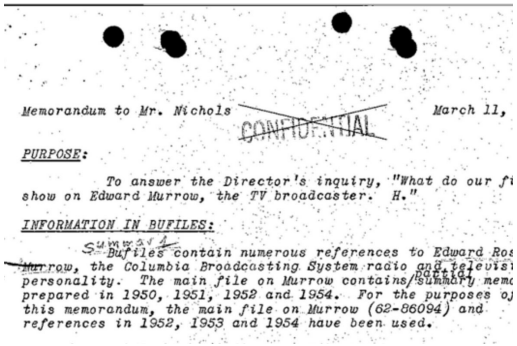
Preface

- To my knowledge, Vintage LLMs have not been built at the scale I’m imagining in this talk.
- I’m not planning to build them myself but I want to encourage people to think about the idea.
- This is an informal talk with high-level ideas.

Vintage LLM = LLM trained on texts (+ images) up to some date

- Date could be 2019 (easier), 1900, or 200 AD (harder).
- Need to avoid info from present leaking into old docs
- Images are things people at the time *could* have seen

E.g. People in 200 AD could see bees or eggs cracking (despite these not being accurately depicted in art)



Memorandum to Mr. Nichols ~~CONFIDENTIAL~~ March 11,

PURPOSE:

To answer the Director's inquiry, "What do our fi
show on Edward Murrow, the TV broadcaster. H."

INFORMATION IN BUFILES:

Bufiles contain numerous references to Edward Ros
Murrow, the Columbia Broadcasting System radio and televisi
personality. The main file on Murrow contains [illegible] memo
prepared in 1950, 1951, 1952 and 1954. For the purposes o
this memorandum, the main file on Murrow (62-86094) and
references in 1952, 1953 and 1954 have been used.

Motivation:
Test LLM-agent approaches to prediction and scientific invention

How well can LLM_2019 **forecast** up to 2024?
→ pandemic, wars, finance

Can LLM_1989 **reinvent** ideas from last 35 years?
→ web, quantum computers, blockchain, transformer, behavioral economics

What about LLM_1600?
→ Newton's Laws, Theory of Evolution, probability theory, calculus, algebra, Turing machines

Motivation: Humanities questions

- Time travel: Communicate with someone from 1700. Can you understand each other?
- Counterfactual history: Impact of adding Western texts to Chinese (or vice versa)
- How surprising were new ideas (e.g. special relativity) from the perspective of the time leading up to them?

Further Motivation

Epistemic AI = makes accurate forecasts, literature surveys, new STEM ideas.

→ Needs gold-standard examples to train and evaluate such an AI

Sources of examples: (1) current humans, (2) algorithm, (3) historical data.

Advantages of (3) historical data:

- Rare events: pandemics, econ crises, big advances in STEM or philosophy.
- Easier to judge quality of past ideas (“stood test of time”)
- Legible to outsiders: e.g. LLM_2019 predicts Covid pandemic

Challenges in making Vintage LLMs

1. Data from the past

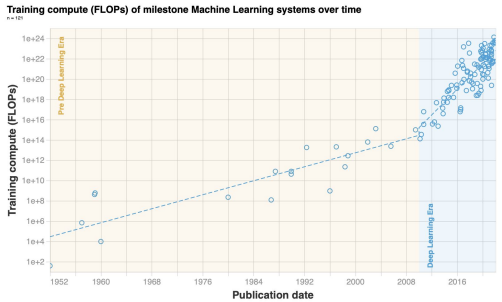
→ Models need 50T words = 50x Library of Congress

→ Can you gather enough data + ensure no leakage?

2. Training cost

→ Training a SOTA model is >$200M for compute

(= 0.05% of science funding)



Also: Both increasing exponentially!

Addressing Challenges (half baked)

1. Data

→ We have the **highest quality** data for 2021 or 1950: STEM papers, key stats, Wikipedia.

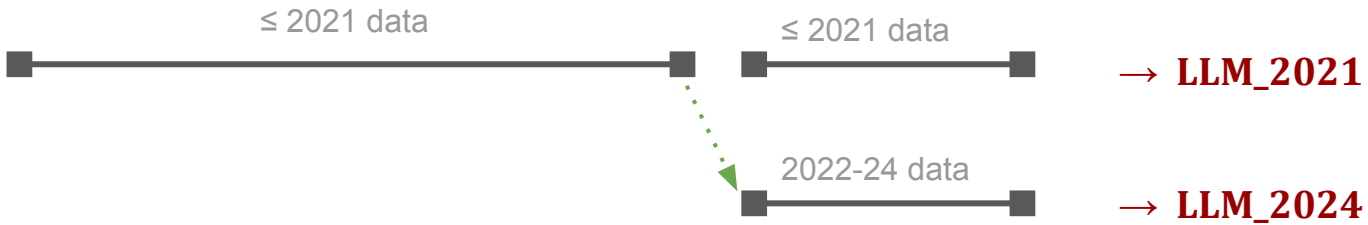
→ There's fewer Reddit threads, but these can be synthetically generated

→ Progress in synthetic data should carry over from use on frontier LLMs

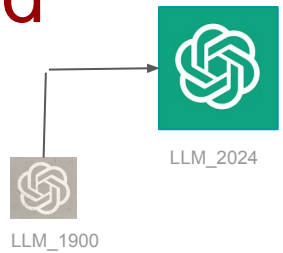
2. Training cost

→ Idea: Chronological training with forking

→ If 2/3 data is ≤2021, increase in training cost is +33% over a single model.



Addressing challenges: very half-baked



Vintage LLM outsources some functions to current LLM

We need to restrict *knowledge* of a Vintage LLM but not its use of tools/experiments.

Using Toolformer idea, train LLM_1900 to **control** LLM_2024 for tools + experiments.

E.g. LLM_1900 describes test of handwashing in hospital and GPT-4o tells the likely result.

Need to avoid LLM_2024 leaking info, but can use compartmentalized LLM.

Compartmentalized LLM

Train on all data but with date annotations before every document. Minimal performance penalty vs regular model.

Thus the LLM emulates explicit reasoning using only past facts. It's not a vintage LLM but in doing explicit reasoning it will avoid anachronism. Very useful as a too.