Large Language Models

An Open Source Perspective

Frank Coyle
coyle@smu.edu
Associate Professor
Computer Science Dept.
Southern Methodist University
Dallas, Texas

1

Prolog

2

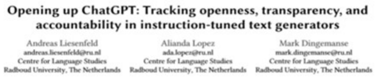
Hey DALL-E, draw an image that illustrates how Large Language Models are being used in all aspects of modern society including marketing, health care, education, finance, etc.

3

4

An Article of Interest

5



Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators

Andreas Liesenfeld
andreas.liesenfeld@ru.nl
Centre for Language Studies
Radboud University, The Netherlands

Alianda Lopez
ada.lopez@ru.nl
Centre for Language Studies
Radboud University, The Netherlands

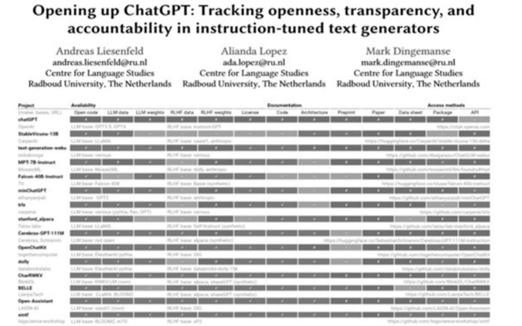
Mark Dingemanse
mark.dingemanse@ru.nl
Centre for Language Studies
Radboud University, The Netherlands

https://dl.acm.org/doi/abs/10.1145/3571884.3604316

The main contribution of this paper is to show that openness is differentiated, and to offer scientific documentation of degrees of openness in this fast-moving field.

We find that while there is a fast-growing list of projects billing themselves as 'open source', many inherit undocumented data of dubious legality, few share the all-important instruction-tuning (a key site where human annotation labour is involved), and careful scientific documentation is exceedingly rare.

6



Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators

Andreas Liesenfeld
andreas.liesenfeld@ru.nl
Centre for Language Studies
Radboud University, The Netherlands

Alianda Lopez
ada.lopez@ru.nl
Centre for Language Studies
Radboud University, The Netherlands

Mark Dingemanse
mark.dingemanse@ru.nl
Centre for Language Studies
Radboud University, The Netherlands

Project	Availability						Documentation					Access methods	
(maker, bases, URL)	Open code	LLM data	LLM weights	RLHF data	RLHF weights	License	Code	Architecture	Preprint	Paper	Data sheet	Package	API
chatGPT	■	■	■	■	■	■	■	■	■	■	■	■	□
OpenAI	LLM base: GPT3.5, GPT4			RLHF base: Instruct-GPT									https://chat.openai.com
StableVicuna-13B	■	■	□	□	□	□	□	■	■	■	■	□	■
CarperAI	LLM base: LLaMA			RLHF base: oasst1, anthropic									https://huggingface.co/CarperAI/stable-vicuna-13b-delta
text-generation-webu	■	■	■	■	■	■	■	■	■	■	■	■	■
oobabooga	LLM base: various			RLHF base: various									https://github.com/Akegarasu/ChatGLM-webui
MPT-7B-Instruct	■	■	■	□	■	■	■	□	■	■	■	■	■
MosaicML	LLM base: MosaicML			RLHF base: dolly, anthropic									https://github.com/mosaicml/llm-foundry#mpt
Falcon-40B-Instruct	■	□	■	□	■	■	□	□	□	■	□	□	■
TII	LLM base: Falcon 40B			RLHF base: Baize (synthetic)									https://huggingface.co/tiiuae/falcon-40b-instruct
minChatGPT	■	■	■	□	■	■	■	□	■	■	■	■	■
ethanyanjiali	LLM base: GPT2			RLHF base: anthropic									https://github.com/ethanyanjiali/minChatGPT
trlx	■	■	■	□	■	■	■	□	■	■	■	□	■
carperai	LLM base: various (pythia, flan, OPT)			RLHF base: various									https://github.com/carperai/trlx
stanford_alpaca	■	■	□	□	■	□	■	■	■	■	□	■	■
Tatsu labs	LLM base: LLaMA			RLHF base: Self-Instruct (synthetic)									https://github.com/tatsu-lab/stanford_alpaca
Cerebras-GPT-111M	■	■	■	■	■	■	■	■	□	■	■	■	■
Cerebras, Schramm	LLM base: not open			RLHF base: alpaca (synthetic)									https://huggingface.co/SebastianSchramm/Cerebras-GPT-111M-instruction
OpenChatKit	■	■	■	■	■	■	■	■	□	■	■	■	■
togethercomputer	LLM base: EleutherAI pythia			RLHF base: OIG									https://github.com/togethercomputer/OpenChatKit
dolly	■	■	■	□	■	■	■	■	□	■	■	■	■
databrickslabs	LLM base: EleutherAI pythia			RLHF base: databricks-dolly-15k									https://github.com/databrickslabs/dolly
ChatRWKV	■	■	■	■	■	■	■	■	■	■	■	■	■
BlinkDL	LLM base: RWKV-LM (own)			RLHF base: alpaca, shareGPT (synthetic)									https://github.com/BlinkDL/ChatRWKV
BELLE	■	■	□	■	■	■	■	■	■	■	□	■	■
LianjiaTech	LLM base: LLaMA, BLOOMZ			RLHF base: alpaca, shareGPT (synthetic)									https://github.com/LianjiaTech/BELLE
Open-Assistant	■	■	■	■	■	■	■	■	■	■	■	■	■
LAION-AI	LLM base: oasst1 (own)			RLHF base: OIG									https://github.com/LAION-AI/Open-Assistant
xmtf	■	■	■	■	■	■	■	■	■	■	■	■	■
bigscience-workshop	LLM base: BLOOMZ, mT0			RLHF base: xP3									https://github.com/bigscience-workshop/xmtf

https://opening-up-chatgpt.github.io/

7

What does ‘open’ mean for LLMs

- We have the source code for LLMs
- The **algorithms** used in the neural networks behind LLMs are well known
  - e.g. Gradient Descent

The **real question** is:

And what aspects of LLMs pertain to openness

8

A Shallow Dive in the Technology underpinning LLMs

9

Steps in building an LLM – each with options that are part of what ‘open’ means

- **Lots of data**
- **Tokenizing the data**
  - **‘cat’, ‘dog’, ‘happy’, …**
- **Generate Embeddings(numeric vectors) from the tokens**
  - **[.21, .34, -.04, .98, .87 ..   ]**
- **Use the embeddings as input to Neural Networks**

10

Embeddings are the secret sauce behind LLMs

“Scale”

```
[-0.64516115  0.1579973  -0.18486351 -0.03695639  0.28924885 -0.09498847
 -0.43030658  0.15431975  0.1239115   0.17329243  0.21807285 -0.21267048
 -0.22992504  0.24730667 -0.4451861  -0.04412757  0.31399828 -0.23861146
  0.11274178 -0.70143986 -0.12033968  0.20097078 -0.34433937  0.0200157
 -0.17284475  0.43762085 -0.01585343 -0.16477302  0.13359785 -0.3297498
 -0.27070326 -0.45194244 -0.15027043  0.13564251 -0.31725803 -0.71317255
  0.23994786 -0.06365798 -0.2350698   0.12471341 -0.33628556 -0.45893794
  0.0239193  -0.01461021  0.7949769  -0.1963934  -0.38624054 -0.18512818
  0.00129966 -0.09555561  0.23405671  0.32197502 -0.04406496 -0.14301962
 -0.06501128 -0.10083073  0.1285449   0.08399501  0.19720553  0.0606354
 -0.22448681 -0.557067   -0.22160476  0.06177633  0.534892   -0.1717653
 -0.567688    0.5929364  -0.14680988  0.78627753 -0.09622003  0.00605357
  0.31533802  0.21695644 -0.5902365  -0.2347756   0.32014322  0.29467028
  0.2876221  -0.18719622 -0.44544363  0.37007272 -0.09240708  0.4406842
  0.20087402  0.22351162 -0.23393816  0.18165983 -0.05705923 -0.2805166
  0.6086521   0.2634202  -0.28903696  0.18512465  0.01641486 -0.09391
  0.06074792  0.06188582 -0.26275593  0.47158718]
```

11

An embedding captures…

```
[-0.64516115  0.1579973  -0.18486351 -0.03695639  0.28924885 -0.09498847
 -0.43030658  0.15431975  0.1239115   0.17329243  0.21807285 -0.21267048
 -0.22992504  0.24730667 -0.4451861  -0.04412757  0.31399828 -0.23861146
  0.11274178 -0.70143986 -0.12033968  0.20097078 -0.34433937  0.0200157
 -0.17284475  0.43762085 -0.01585343 -0.16477302  0.13359785 -0.3297498
 -0.27070326 -0.45194244 -0.15027043  0.13564251 -0.31725803 -0.71317255
  0.23994786 -0.06365798 -0.2350698   0.12471341 -0.33628556 -0.45893794
  0.0239193  -0.01461021  0.7949769  -0.1963934  -0.38624054 -0.18512818
  0.00129966 -0.09555561  0.23405671  0.32197502 -0.04406496 -0.14301962
 -0.06501128 -0.10083073  0.1285449   0.08399501  0.19720553  0.0606354
 -0.22448681 -0.557067   -0.22160476  0.06177633  0.534892   -0.1717653
 -0.567688    0.5929364  -0.14680988  0.78627753 -0.09622003  0.00605357
  0.31533802  0.21695644 -0.5902365  -0.2347756   0.32014322  0.29467028
  0.2876221  -0.18719622 -0.44544363  0.37007272 -0.09240708  0.4406842
  0.20087402  0.22351162 -0.23393816  0.18165983 -0.05705923 -0.2805166
  0.6086521   0.2634202  -0.28903696  0.18512465  0.01641486 -0.09391
  0.06074792  0.06188582 -0.26275593  0.47158718]
```

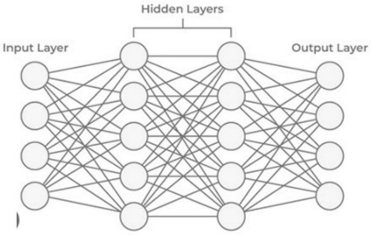
- **Syntax**
- **Semantics**
- **Context and Relationships**
- **Word Co-occurrence**
- **Relationships and Analogies**
- **Hierarchical Information**

How is this done?

12

Neural Networks

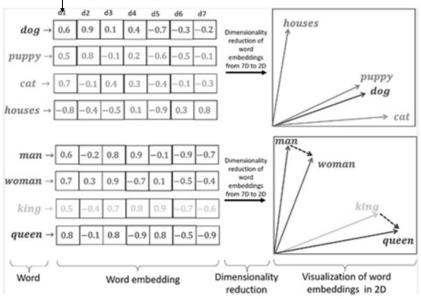
Embedding values – syntax, semantics, etc. are learned by training on terabytes of data



13

Tokenization (words)

Embeddings - (turning tokens into multidimensional vectors)


Dimensionality reduction of word embeddings from 7D to 2D
houses
puppy
dog
cat
man
woman
king
queen
Word
Word embedding
Dimensionality reduction
Visualization of word embeddings in 2D

	d1	d2	d3	d4	d5	d6	d7
dog →	0.6	0.9	0.1	0.4	−0.7	−0.3	−0.2
puppy →	0.5	0.8	−0.1	0.2	−0.6	−0.5	−0.1
cat →	0.7	−0.1	0.4	0.3	−0.4	−0.1	−0.3
houses →	−0.8	−0.4	−0.5	0.1	−0.9	0.3	0.8
man →	0.6	−0.2	0.8	0.9	−0.1	−0.9	−0.7
woman →	0.7	0.3	0.9	−0.7	0.1	−0.5	−0.4
king →	0.5	−0.4	0.7	0.8	0.9	−0.7	−0.6
queen →	0.8	−0.1	0.8	−0.9	0.8	−0.5	−0.9

Embedding (Vector) Sizes
GPT-1 : 768
GPT-3 : 768 – 4096
GPT-4 : Not revealed !

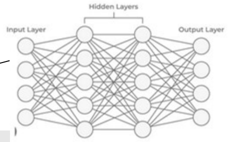
authored by David Rozado

14

Openness and Neural Network Design

- Number of hidden layers
- Size of hidden layers
- NNs use a variety of standard algorithms used to predict output
  - Activation Functions (softmax, sigmoid, ..)
  - Optimization (SGD, Adam, ..)
  - Regularization (dropout, regularization, ..)

The neural network learns the best **weights** to use between nodes. A fully open-source model will make the weights available.



GPT-4 has 175 Billion Parameters !

15

In addition to the data used, what else does ‘open’ mean for an LLM

- **Tokenization**
  - The input text is segmented into smaller units called tokens. These tokens could be words, subwords, or characters, ***depending on the specific tokenization scheme***
- **Embedding**
  - Each token is then mapped to a high-dimensional vector representation known as an embedding. There are a ***variety of embedding sizes and schemes.***
- **Neural Network**
  - *Used to predict the next word in a sentence*
  - *Neural networks are based on standard algorithms but there are a* **variety of ways to design and configure neural networks – number of hidden layers and number of nodes per layer**

16

End of Shallow Dive

17

European Lawmakers Pass AI Act, World’s First Comprehensive AI Law

- Technology should be human centric
- Passed by the European Parliament – to be confirmed by EU member states
- For high risk applications (banking, food, ..)
  - Human supervision
  - The right to question
  - Publish the data used to create models
  - Fines up to 35 Million Euros possible

18

And in the USA …

19

The AI Alliance

Companies teaming up with academia to promote ‘open’ alternatives to OpenAI

- Meta, IBM, NSF,
- Linux Foundation, Hugging Face

and over 50 companies

The Frontier Model Forum

We're the *Frontier Model Forum*! An industry body dedicated to advancing the safe development and deployment of frontier AI models.

Apple, Amazon, Google, OpenAI

20

Hugging Face

- The **"GitHub of machine learning"**
- An **extensive library** of over 300,000 models and provides access to a broad array of datasets uploaded by the community
- **Spaces:** allows users to create interactive, in-browser demos of ML models without needing extensive technical knowledge
- **Open Source and Accessibility:** The platform's open-source nature and deployment tools significantly impact the accessibility of AI development.

21

Advantage of Open Models

- **Fine Tuning**
  - Download a model with embeddings already created
  - Use your own domain-specific data to continue to train the neural network
  - Embeddings tailored to your use cases are updated within the model
  - More accurate responses

22

Option for all Models - RAG

- **Retrieval Augmented Generation**
  - Provide your own data and ask your question
  - Example: For the text enclosed in triple quotes, give me 5 bullet points that I can use in a powerpoint presentation. “”” text …..”””
  - Example: For the text enclosed in triple quotes, create 10 multiple choice questions and place an asterisk after the correct answer.””” ..blah “””
  - Also useful – include “**Do NOT Hallucinate**”

23

Postscript

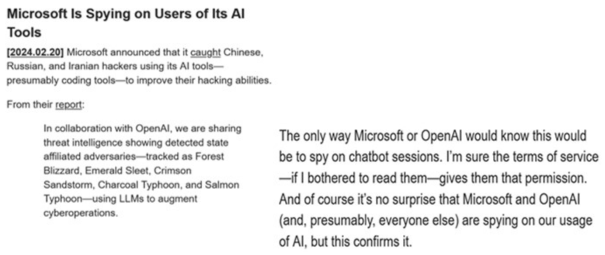
24

Bruce Schneier Takeaways
(from Zoom at Kwaai Conference)

- LLMs are not your friend
- Their language is seductive – they sound like us and we tend to trust those who sound like us
- They act as our agents – but they are double agents – created by companies that have their own agendas

25

Bruce Schneier’s Monthly Newsletter:
https://www.schneier.com/crypto-gram/



Microsoft Is Spying on Users of Its AI Tools

[2024.02.20] Microsoft announced that it caught Chinese, Russian, and Iranian hackers using its AI tools—presumably coding tools—to improve their hacking abilities.

From their report:

> In collaboration with OpenAI, we are sharing threat intelligence showing detected state affiliated adversaries—tracked as Forest Blizzard, Emerald Sleet, Crimson Sandstorm, Charcoal Typhoon, and Salmon Typhoon—using LLMs to augment cyberoperations.

The only way Microsoft or OpenAI would know this would be to spy on chatbot sessions. I’m sure the terms of service—if I bothered to read them—gives them that permission. And of course it’s no surprise that Microsoft and OpenAI (and, presumably, everyone else) are spying on our usage of AI, but this confirms it.

26

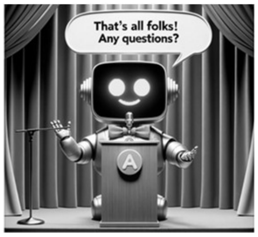
Reference to tracking openness article

Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators

- https://dl.acm.org/doi/abs/10.1145/3571884.3604316

- Website: https://opening-up-chatgpt.github.io/

27



Slides: **http://s2.smu.edu/~coyle/slides/**

Follow me at: **https://medium.com/@coyle_41098/**
Curated collection of LLM articles from medium.com

28