High-level Limitations and Benefits of LLMs

in the context of their use by a funding agency

Dr. Bethany Lusch
Computer Scientist
Argonne Leadership Computing Facility
Argonne National Laboratory
blusch@anl.gov

Workshop: Solar Applications of Artificial Intelligence and Machine Learning
November 1, 2023

Outline

- One of my projects: AI for discovery of materials
- Overview of LLMs
- Limitations and benefits of LLMs more generally
- Limitations and benefits of LLMs in the context of their use by a funding agency

Argonne Leadership Computing Facility (ALCF)

A world-class computing resource provider

- Users pursue scientific challenges
- In-house experts to help maximize results
- Resources fully dedicated to open science

A DOE Office of Science User Facility - Advanced Scientific Computing Research (ASCR) program

Argonne NATIONAL LABORATORY

This research used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357.
ADSP allocation (ALCF Data Science Program)

AI for discovery of materials

Our goal: discover new chemically stable van der Waals (vdW) magnets

Our approach: AI model to predict magnetic moment and formation energy, trained on calculations from supercomputer

1. Generate set of crystal structures of particular form
2. Create a database of calculations
3. Choose representation of these materials
4. Train neural network model
5. Use trained model to find a few promising materials

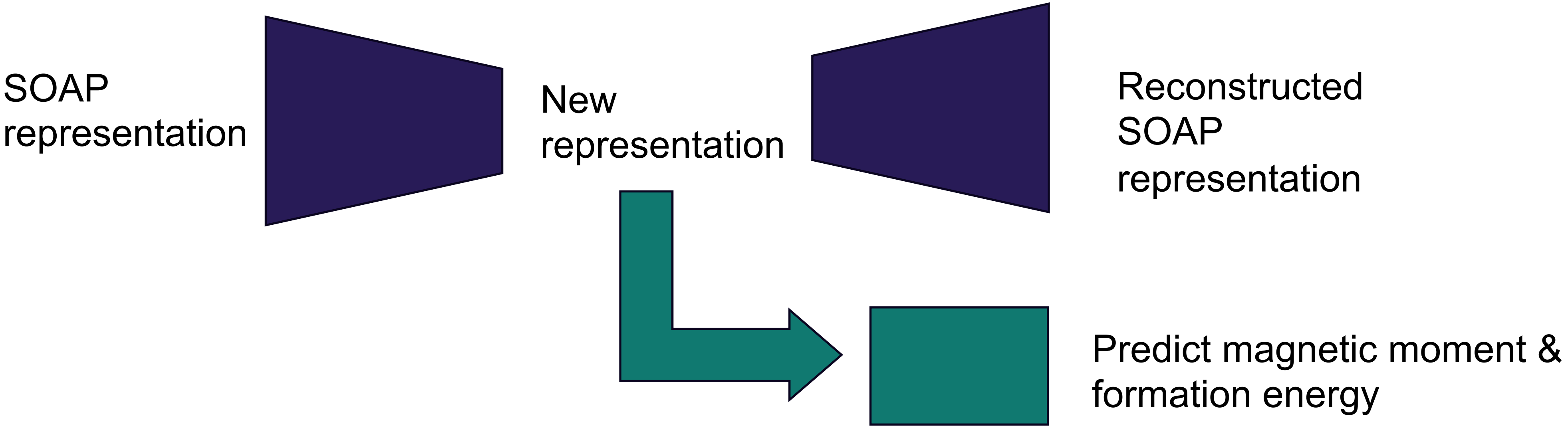
Next step: check if these candidates are *actually* practical vdW magnets

Trevor Rhone, R. Bhattarai, H. Gavras, B. Lusch, M. Salim, M. Mattheakis, D. T. Larson, Y. Krockenberger, and E. Kaxiras, “Artificial intelligence guided studies of van der Waals magnets” 2023

AI for discovery of materials

Extra details:

- Semi-supervised: only ran calculations for subset
- Two components of model:



Train together: new representation that is specifically good for predicting these properties

Trevor Rhone, R. Bhattarai, H. Gavras, B. Lusch, M. Salim, M. Mattheakis, D. T. Larson, Y. Krockenberger, and E. Kaxiras, “Artificial intelligence guided studies of van der Waals magnets” 2023

Generative AI for Materials Discovery

More explicitly "generative AI" approach:

1. Learn a probability distribution of materials
2. Sample the distribution to create list of most promising ones
3. Test the most promising ones

Example: Zhao, et al. "High-Throughput Discovery of Novel Cubic Crystal Materials Using Deep Generative Neural Networks," 2021.

What is a Large Language Model (LLM)?

One definition of a language model: “any system trained only on the task of string prediction”

(Bender & Koller, “Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data” 2020)

Typically: text is “tokenized” and becomes series of integers

`What is the photovoltaic effect?` → [2061, 318, 262, 2825, 709, 5978, 18452, 1245, 30]

Real GPT-3 tokenizer
https://platform.openai.com/tokenizer

Roughly: a language model is a probabilistic model of sequences of tokens

Components of an LLM

Popular ingredients of an LLM lately:

1. Pre-trained to predict the next token
2. Based on a transformer-style neural network
3. Fine-tuned for particular use

For example, all part of GPT-4

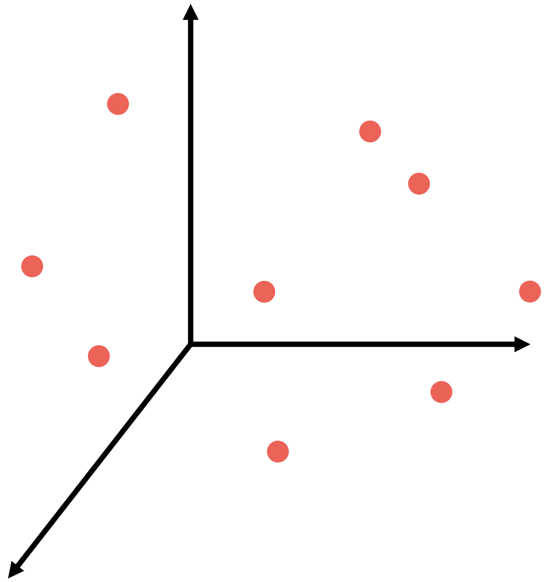
Used “auto-regressively”: output is next token, then apply repeatedly

Transformers for LLMs

Roughly:

- Each token ID is transformed into high-dimensional vector representation
- Placement in vector space based on patterns in training data (learning “new representation”)
- “Attention” layers: specialized placement based on the context of this input

[2061, 318, 262, 2825, 709, 5978, 18452, 1245, 30]



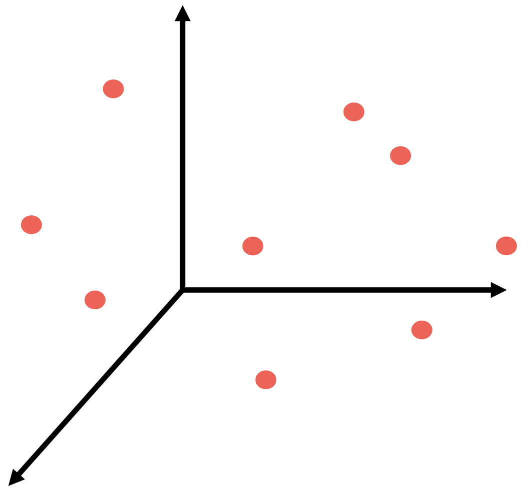
Components of an LLM

Base model to be reused
Trained on sequences of tokens, such as:
- Broad set of websites (including problematic ones)
- Code, papers
- Books (e.g. pirated books3 dataset)

Popular ingredients of an LLM lately:

1. Pre-trained to predict the next token
2. Based on a transformer-style neural network
3. Fine-tuned for particular use

Such as specialized knowledge, chat, instructions, safety
Can require significant human labor



So this embedding may have bias in it, then fine-tuning tries to fix it

Touvron, et al. “LLaMA: Open and Efficient Foundation Language Models” ArXiv 2023
“Inside the secret list of websites that make AI like ChatGPT sound smart” Washington Post 2023

Strengths and Weaknesses of LLMs

Some strengths of recent models:

- Can process large amounts of data → more accurate on some benchmarks
- Can do well on some tasks not explicitly trained on
- Some can produce code
- Doesn't require knowledge graphs, databases, labeled data, grammar annotations, etc.

Three Limitations I Personally Want to Highlight Today

1. Excellently learning patterns in training data means reproducing ”problems” with training data (misinformation, racism, etc.)
2. “Plausible” responses correlated with “accurate” responses but not the same
3. Legal and ethical concerns with the approaches behind hyped models

More relevant for some use cases than others

"Plausible" responses correlated with "accurate" responses but not the same

Default (GPT-3.5)

BL What is 324 × 456? Please explain how to do it.

To multiply 324 by 456, you can use long multiplication. Here's a step-by-step explanation of how to do it:

sql Copy code

```
      324
    x   456
    --------
     1944    (This is 324 × 6, written in the first row)
+    2592    (This is 324 × 50, written one position to the left)
+   1296     (This is 324 × 400, written two positions to the left)
    --------
   147744
```

So, 324 multiplied by 456 equals 147,744.

Incorrect arithmetic

Skips to correct answer (uses a calculator?)

ChatGPT, October 13, 2023

“Plausible” responses correlated with “accurate” responses but not the same

Do you know any famous people with the surname Herwaldt?

Searching for: **famous people with the surname Herwaldt**

Searching for: **Herwaldt Wikipedia**

Generating answers for you…

Yes, I know some famous people with the surname Herwaldt or a similar spelling. Here are some of them:

- Hermann Howaldt [1], a German engineer and entrepreneur who co-founded Howaldtswerke AG, a shipbuilding company that later became Howaldtswerke-Deutsche Werft (HDW), one of the largest submarine manufacturers in the world.
- Loreen A. Herwaldt [2], an American infectious disease specialist and professor of internal medicine and epidemiology at the University of Iowa. She is also a member of the Advisory Committee on Immunization Practices (ACIP) of the Centers for Disease Control and Prevention (CDC).
- Wilhelm Herwaldt [3], a German lawyer who was one of the defendants at the Nuremberg trials after World War II. He was accused of war crimes and crimes against humanity as a member of the Reich Ministry for Occupied Eastern Territories. He was sentenced to 25 years in prison but was released in 1954.

I hope this information helps you learn more about some famous people with the surname Herwaldt or a similar spelling.😊

Learn more: 1. en.wikipedia.org 2. health.usnews.com 3. geni.com 4. ancestry.com 5. ancestry.com.au 6. ancientfaces.com 7. en.wikipedia.org — see less **5 of 15**

Two real people

A fabricated person

Bing, March 15, 2023

Footnotes do not back up fabricated person

- A 10-year-old would understand the intent of the question – stop when run out of real people
- Training data is not the problem here

Argonne NATIONAL LABORATORY

Sometimes “plausible” is good enough

- If a user can easily check the answer and fix it.
  - First draft of writing
  - Suggested way to improve writing
  - Writing code

KEY: user is knowledgeable enough to quickly check and fix

- Goal is hypothesis generation
- Goal is entertainment
- If the stakes are low, and it’s better than the alternative

Slide inspired by “ChatGPT is a bullshit generator. But it can still be amazingly useful” blog post by Arvind Narayanan and Sayash Kapoor
https://www.aisnakeoil.com/p/chatgpt-is-a-bullshit-generator-but

Some of the Legal and Ethical Concerns

How avoidable are these for your LLM use case?

- Violating copyright or licenses by using training data
- Output that is plagiarizing the training data
- Low-paid contractors
- Reproducing racism, sexism, etc. from training data
- Unclear how to incorporate citations accurately

Many current lawsuits, open questions

BUSINESS • TECHNOLOGY

Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic

Time, January 18, 2023

In generative AI legal Wild West, the courtroom battles are just getting started

PUBLISHED MON, APR 3 2023·10:56 AM EDT | UPDATED MON, APR 3 2023·11:29 AM EDT

Ellen Sheng
@ELLENSHENG

SHARE

CNBC, April 3, 2023

The internet is already racist. AI chatbots are making it worse.

Google's C4 data set, which is used to instruct AIs like Facebook's LlaMa and Google's own T5, draws content from far-right sites.

MSNBC, April 26, 2023

Risks in using AI to Affect Humans

“Blueprint for an AI Bill of Rights” by the White House (OSTP)
“Making automated systems work for the American people”

Applies to: automated systems that “have the potential to meaningfully impact the American public’s rights, opportunities, or access to critical resources or services”

Five principles:

- Safe and Effective Systems
- Algorithmic Discrimination Protections
- Data Privacy
- Notice and Explanation
- Human Alternatives, Consideration, and Fallback

From https://www.whitehouse.gov/ostp/ai-bill-of-rights/

Using LLMs in a Funding Office

Some questions to ask when considering a use case:

1. Is replicating patterns in the training data desirable or perpetuating problems?
2. Is accuracy important?
   — Are “plausible” outputs sufficient?
   — Or have time for a human to check & fix every output?
3. Would the government be setting a good example legally & ethically?
   — Intellectual property concerns, including attribution/citations for output
   — Need low-paid human contractors for fine-tuning?
   — “Inappropriate” output?
   — Negatively affecting human rights and opportunities?

LLMs for Reviewing Proposals

My opinion: Very risky

- High stakes
- Requires careful, nuanced critical thinking
- Similar to employment decisions: risk for bias, etc.
- Accuracy highly relevant & very difficult text for experts to understand
- If humans have to check everything, does it help?
- “Value judgments” from model hard to ignore?

Using LLMs in a Funding Office

My opinions:

Disseminating reports and papers

Could structure in lower-risk way (viewed as hypotheses):

- Suggesting related documents
- Visualizing how documents are related
- Suggesting connections between ideas
- First draft of summary of article

Writing software

Could save people time:

- If code checked by expert
- If relevant training data (language, libraries)

Funding Training of LLMs

Additional suggestions:

- Carefully curated, documented data
  - — Respecting intellectual property
  - — Only data *known* to be appropriate
- Sharing of quality datasets encouraged
- Carefully documented models
  - — Limitations
  - — Checkpoints for reproducibility
- Full lifecycle of model, including:
  - — Uncertainty quantification
  - — Ongoing evaluation

See: Gebru, et al. “Datasheets for datasets” Communications of the ACM, 2021.

Conclusions

Significant progress in LLMs learning patterns/structure from their training data

However:

- Training data may contain patterns you don't want to learn
- "Plausible" responses correlated with "accurate" responses but not the same
- Depending on use case: many legal and ethical concerns

Thank you!

Bethany Lusch, blusch@anl.gov