Past, Present, and Future

29 April 2024
<table>
<thead>
<tr>
<th>Active Assignments</th>
<th>Released</th>
<th>Due (EDT)</th>
<th>Submissions</th>
<th>% Graded</th>
<th>Published</th>
<th>Regrades</th>
</tr>
</thead>
<tbody>
<tr>
<td>Final project: Paper and supporting materials</td>
<td>APR 29, 2024 1:30 PM</td>
<td>APR 30, 2024 11:59 PM</td>
<td>0</td>
<td>0%</td>
<td>ON</td>
<td></td>
</tr>
<tr>
<td>Late Due Date: MAY 7, 2024 11:59 PM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Final project: Description of contributions</td>
<td>APR 29, 2024 1:30 PM</td>
<td>APR 30, 2024 11:59 PM</td>
<td>0</td>
<td>0%</td>
<td>ON</td>
<td></td>
</tr>
<tr>
<td>Late Due Date: MAY 7, 2024 11:59 PM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
I've received some questions about the project deadline. As stated on the project website, the project is nominally due on Tuesday, April 30 – the last day of classes.

However, I know there can be a lot of deadlines around this time, and research projects are prone to setbacks, so (as the project website also states) I'll accept submissions until the last day the college allows, namely May 7 – the last day of study week.

**Beware:** If you aim to be done on May 7, it's very easy to hit a technical problem or other unexpected delay and miss the deadline – whether by minutes, hours, or days – at which point I'm not allowed to accept your work.

So, to be as flexible as possible while keeping this concern in mind, I will waive the usual late penalty (10% a day, as specified in the syllabus) until midnight on Sunday, May 5. After that, the usual late penalty applies.
Where are we all?
Just a few years ago

No encoder–decoder / seq2seq

No attention

No large-scale question-answering/reading comprehension datasets

No TensorFlow or PyTorch

...
Gentlemen, our learner overgeneralizes because the VC-Dimension of our Kernel is too high. Get some experts and minimize the structural risk in a new one. Rework our loss function, make the next kernel stable, unbiased and consider using a soft margin.

**Statistical Learning**

**Neural Networks**

**Stack More Layers**
Gentlemen, our learner overgeneralizes because the VC-Dimension of our Kernel is too high. Get some experts and minimize the structural risk in a new one. Rework our loss function, make the next kernel stable, unbiased and consider using a soft margin.

Statistical Learning

Neural Networks

Stack More Transformers
We’ve seen abilities grow significantly:

- GPT can perform simple text-labeling tasks but cannot generally produce coherent text.
- GPT-2 adds the ability to produce reasonably high-quality text and a limited ability to follow instructions.
- GPT-3 was the first modern general-purpose LLM, useful across a wide range of language tasks.

But the designs of these three models hardly differ at all. The difference in their abilities stems from vast differences in scale.
<table>
<thead>
<tr>
<th>Model</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medium-sized LSTM</td>
<td>10 M</td>
</tr>
<tr>
<td>ELMo</td>
<td>90 M</td>
</tr>
<tr>
<td>GPT</td>
<td>110 M</td>
</tr>
<tr>
<td>BERT-Large</td>
<td>320 M</td>
</tr>
<tr>
<td>GPT-2</td>
<td>1.5 B</td>
</tr>
<tr>
<td>GPT-3</td>
<td>175 B</td>
</tr>
<tr>
<td>GPT-4</td>
<td>~ 1.7 T</td>
</tr>
<tr>
<td>Model</td>
<td>Parameters</td>
</tr>
<tr>
<td>------------</td>
<td>------------</td>
</tr>
<tr>
<td>Medium-sized LSTM</td>
<td>10 M</td>
</tr>
<tr>
<td>ELMo</td>
<td>90 M</td>
</tr>
<tr>
<td>GPT</td>
<td>110 M</td>
</tr>
<tr>
<td>BERT-Large</td>
<td>320 M</td>
</tr>
<tr>
<td>GPT-2</td>
<td>1.5 B</td>
</tr>
<tr>
<td>GPT-3</td>
<td>175 B</td>
</tr>
<tr>
<td>GPT-4</td>
<td>~ 1.7 T</td>
</tr>
</tbody>
</table>

We think!
<table>
<thead>
<tr>
<th>Model</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medium-sized LSTM</td>
<td>10 M</td>
</tr>
<tr>
<td>ELMo</td>
<td>90 M</td>
</tr>
<tr>
<td>GPT</td>
<td>110 M</td>
</tr>
<tr>
<td>BERT-Large</td>
<td>320 M</td>
</tr>
<tr>
<td>GPT-2</td>
<td>1.5 B</td>
</tr>
<tr>
<td>Honey bee brain</td>
<td>~ 1 B synapses</td>
</tr>
<tr>
<td>GPT-3</td>
<td>175 B</td>
</tr>
<tr>
<td>GPT-4</td>
<td>~ 1.7 T</td>
</tr>
<tr>
<td>Cat</td>
<td>~ 10 T synapses</td>
</tr>
<tr>
<td>Human</td>
<td>~ 100 T synapses</td>
</tr>
</tbody>
</table>
Large language models keep getting larger:

- More training data
- More parameters
- More computation used to train them.
OpenAI codebase next word prediction

Bits per word

Compute

Observed
Prediction
gpt-4

OpenAI, 2024
Specific important behaviors in LLMs tend to emerge unpredictably – it’s not possible to predict when models will start to show specific skills be capable of completing specific tasks.
There’s growing evidence that, to some extent, LLMs develop internal representations of the world, which allow them to reason at a level of abstraction that isn’t sensitive to the exact form of the text they’re given.

Current LLMs seem to do this only weakly and sporadically.

But the evidence for this phenomenon is strongest in the most recent models, so we should expect it to become more robust in the future.
Bubeck et al., 2023
Unfortunately, there are no reliable techniques for steering the behavior of LLMs.
And even experts are not (yet) able to interpret the inner workings of LLMs.
Symbolic Methods Dominate
IBM Machine Translation Models
Money dries up for neural methods in the US

Metrics become important at DARPA
Parsing and MT dominate *CL conferences
Statistical NLP (including topic models, PGMs) dominates
Discussions of data scale solving everything
Early work in neural NLP

Neural revolution in NLP
Framework-based neural research
Word embeddings fuel SOTA chasing
Pre-train, fine-tune
Generation over classification
Prompting
To Build Our Future, We Must Know Our Past:
Contextualizing Paradigm Shifts in Natural Language Processing

Sireesh Gururaja1, Amanda Bertsch1, Clara Na1, David Gray Widder2, Emma Strubell1,3

1Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA
2Digital Life Initiative, Cornell Tech, Cornell University, New York City, NY, USA
3Allen Institute for Artificial Intelligence, Seattle, WA, USA
{sgururaj, abertsch, csna, estrubel}@cs.cmu.edu, david.g.widder@gmail.com

Abstract

NLP is in a period of disruptive change that is impacting our methodologies, funding sources, and public perception. In this work, we seek to understand how to shape our future by better understanding our past. We study factors that shape NLP as a field, including culture, incentives, and infrastructure by conducting long-form interviews with 26 NLP researchers of varying seniority, research area, institution, and social identity. Our interviewees identify cyclical patterns in the field, as well as new shifts without historical parallel, including changes in benchmark culture and software infrastructure. We complement this discussion with quantitative analysis of citation, authorship, and language use in the ACL Anthology over time.

![Figure 1: The number of unique researchers publishing in ACL venues has increased dramatically, from 715 unique authors in 1980 to 17,829 in 2022.](image-url)
Cycles of research
What happens after a breakthrough?
BERT’s great for sentiment!
BERT’s great for sentiment! BERT’s great for medicine!
BERT embeddings are great features!

BERT’s great for medicine!

BERT’s great for sentiment!
BERT’s great with other loss functions!

BERT embeddings are great features!

BERT’s great for medicine!

BERT’s great for sentiment!
BERT's great with other loss functions!

BERT embeddings are great features!

BERT's great for medicine!

BERT's great for sentiment!
Exploit
What about explainability?
Is there a bias from pretraining data?

What about explainability?
What about efficiency?

Is there a bias from pretraining data?

What about explainability?
What about explainability?

Exploit

Is there a bias from pretraining data?

What about efficiency?

Would be nice if BERT could do X…

What about explainability?
Explore

What about explainability?

Exploit

Would be nice if BERT could do X…

Is there a bias from pretraining data?

What about efficiency?
Each breakthrough takes over the field for a time – but then something new comes out and shifts the focus again.
Are we exploiting or exploring right now?
Benchmark culture
Long ago, NLP research was evaluated qualitatively:

“The performance metrics were, ‘Oh, my God, it does that? No machine ever did that before.’”
The creation of the Penn TreeBank was an inflection point, ushering in the age of statistical NLP.
## Constituency Parsing on Penn Treebank

### Leaderboard

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>F1 score</th>
<th>Extra Training Data</th>
<th>Paper</th>
<th>Code</th>
<th>Result</th>
<th>Year</th>
<th>Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SAP + XLNet</td>
<td>96.40</td>
<td>×</td>
<td>Improving Constituency Parsing with Span Attention</td>
<td></td>
<td></td>
<td>2020</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Label Attention Layer + HPSG + XLNet</td>
<td>96.38</td>
<td>×</td>
<td>Rethinking Self-Attention: Towards Interpretability in Neural Parsing</td>
<td></td>
<td></td>
<td>2019</td>
<td></td>
</tr>
</tbody>
</table>

**Filter:** LSTM, XLNet, BERT, RoBERTa, ELMo, Transformer, untagged

**Graph:**
- Head-Driven Phrase Structure Grammar Parsing
- Punkt + XLNet
- CNN Large + Time-lapse
- Self-attentive encoder + ELMo
- Model combination
- Semi-supervised LSTM-LM
- Parse fusion

**View:** F1 score by Date for All models

**Dataset:**
- Other models
- Models with highest F1 score
According to one of the researchers interviewed by Gururaja et al.,

“when neural nets came along, suddenly everything *had* to be about benchmarking, *because we couldn’t see what was going on inside.*”
“It’s easier to just climb an existing benchmark… it could be really costly to create a data set. So if you think that the benchmark is, you know, broken, but you can’t afford to fix it… *a lot of people end up just working on that existing benchmark, even though it’s broken.*”
In NLP, benchmark accuracy is no longer enough – researchers also care about things like ethics and interpretability.
The software lottery
Hardware lotteries (Hooker, 2021):

The availability of specialized hardware influences research directions.
Hardware lotteries (Hooker, 2021):

The availability of specialized hardware influences research directions.

Software lotteries:

“things that software makes easy, people are going to do”
Pre-lottery:

Most software is either lab-specific (“we spent like 90% of our time re-implementing papers”) or proprietary (“people had to go through a big deal to license it”)

Early lottery:

Some software tooling is shared (“everything started being… simpler to manage, simpler to train”), e.g., Tensorflow, PyTorch, CoNLL Evaluation

Now:

Software tooling, artifacts, and even whole systems are shared, e.g., HuggingFace.
Mentions of software libraries by year

- HuggingFace
- PyTorch
- Theano
- DyNet
- Moses
- Tensorflow

Fraction of papers that mention

Year

Gururaja et al., 2023
Costs of software lotteries

Single points of failure:

You have to trust the library is correct – but, in practice, we find that toolkits like PyTorch and HuggingFace have many bugs!

Obscuring what’s under the hood:

“The pace is so fast that there is no time to properly document, there is no time to properly engage with this code, you’re just using them directly”
The software lottery is implicit funding:

Tools that win the software lottery enable research that would be difficult without them…

…but disincentivize research that does not fit those tools cleanly.
“You’re not gonna just build your own system that’s gonna compete on these major benchmarks yourself. You have to start [with] the infrastructure that’s already there.”
Back to the overview
STOP DOING NATURAL LANGUAGE PROCESSING

- WORDS WERE NOT SUPPOSED TO BE VECTORS
- BILLIONS OF PARAMETERS and EXAFLOPS OF COMPUTING POWER SPENT yet NO real progress beyond "As an AI language model,"
- Wanted to process text anyways for a laugh? We had a tool for that, it was called `s.matches(".*<([a-z]+)(([aeiou]+(?:\s+[aeiou]+)?)+)\")`;
- is AN OBJECT and has NO RIGHT TO SPEAK IN OUR SACRED TONGUE

(This is REAL NLP, done by REAL HUMANS):

?????

"Hello I would like please"
"Sorry, I don't speak"
Acknowledgments

This class incorporates material from:

Bowman, 2023: “Eight Things to Know about Large Language Models”
Gururaja, Bertsch, Na, Widder, and Strubell, 2023: “To Build Our Future, We Must Know Our Past: Contextualizing Paradigm Shifts in Natural Language Processing”.
DALL-E