Computational Sociology

Large Language Models

Dr. Thomas Davidson

Rutgers University

April 11, 2024

Course updates

Homework

- Homework 4 released
  - Due 4/15 at 5pm

Course updates

Project next steps

- Feedback on preliminary data reports in progress (by end of Monday)

Course updates

Project next steps

- Preliminary results due next Friday 4/19 at 5pm
  - Update existing Github repository
  - Add results section to paper
    - At least 1 figure or table showing results from analysis
  - Incorporate feedback from last round as needed
- Final presentations on 4/25

Plan

1. Language models
2. Large language models
3. Social scientific applications
4. Challenges
5. Commercial versus open-source

Language models

What are language models?

- Language models predict the likelihood of a sequence of words.
- Applications include auto-completion, speech recognition, and more

Language models

The bigram model

- Simple language models learn probabilistic representations of language.
- In the bigram model, the probability w_k only depends on the previous word, w_{k-1}.

$$P(w_{1:n}) \approx \prod_{k=1}^{n} P(w_k | w_{k-1})$$

- n-gram language models generative this to longer sequences of words

Language models

Limitations of n-gram language models

- Language use is much more complex than bi-gram or *n*-gram language models
- Three limitations of early language models:
  1. Insufficient data/complexity to sufficiently model language generation
  2. Complex models become intractable to compute
  3. Limited information on word order

Language models

Advances in neural language models

- Over the past decade language models have developed due to three factors:
  1. Availability of large text corpora
  2. Advances in computer processing (Graphical Processing Units - GPUs)
  3. Innovations in neural network architecture

Language models

Word2vec

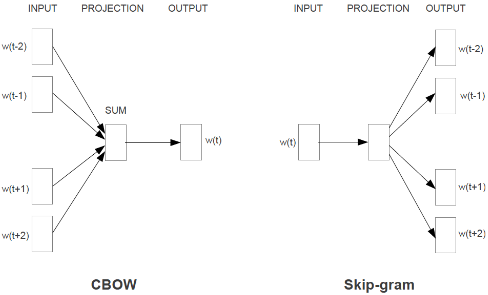


Figure 1: New model architectures. The CBOW architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word.

Mikolov et al. 2013.

Attention and the transformer architecture

Attention Is All You Need

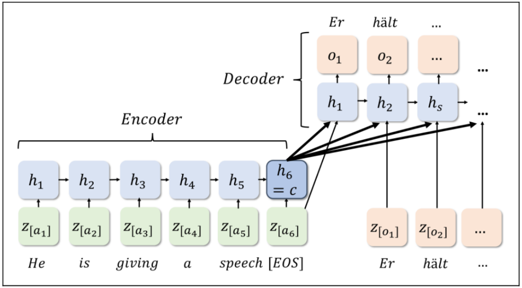
Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

Attention and the transformer architecture

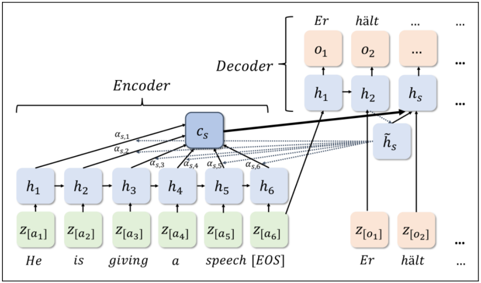
Encoders and decoders



Wänkmuller 2022.

Attention and the transformer architecture

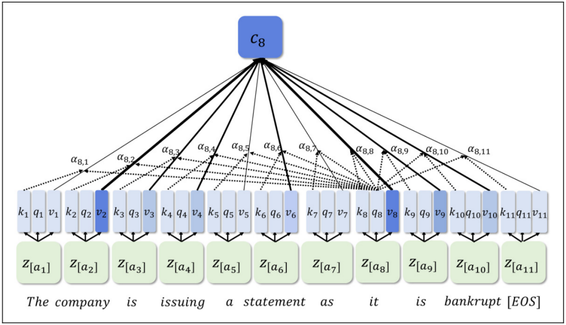
Adding attention



Wänkmuller 2022.

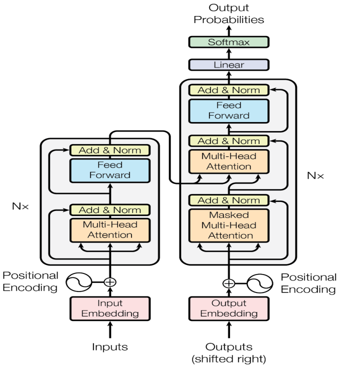
Attention and the transformer architecture

Attending to a token



Wänkmuller 2022.

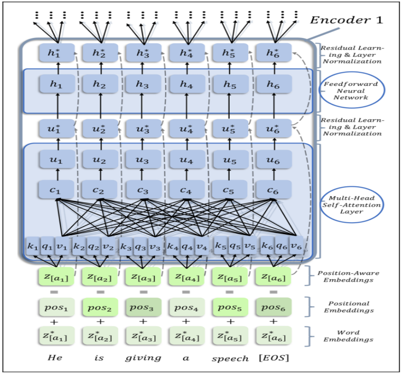
The transformer



Vaswani et al 2017

Attention and the transformer architecture

Stacked attention layers

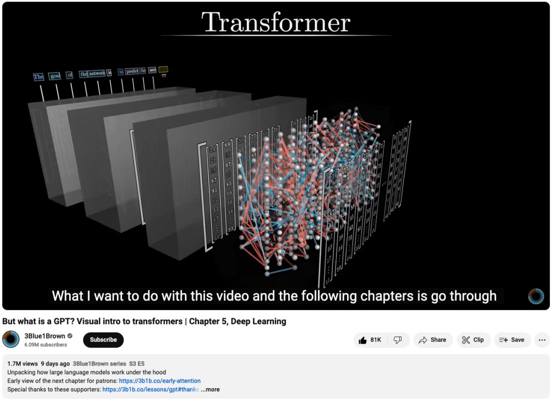


Wänkmuller 2022.

Attention and the transformer architecture

Further resources

- 3Blue1Brown has a new series on YouTube explaining how attention and transformers work



Large language models

BERT

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

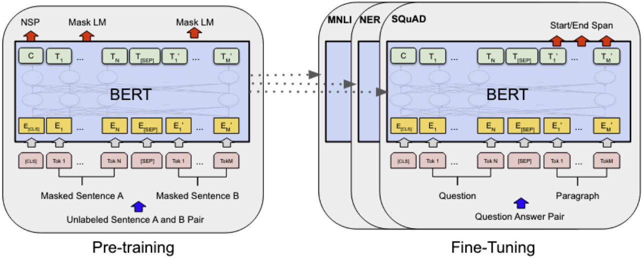
Abstract

We introduce a new language representation model called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning *all* pre-trained parameters. The two approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations.

Large language models

BERT: Pre-training and fine-tuning



Devlin et al. 2018

Large language models

Foundation models and transfer learning

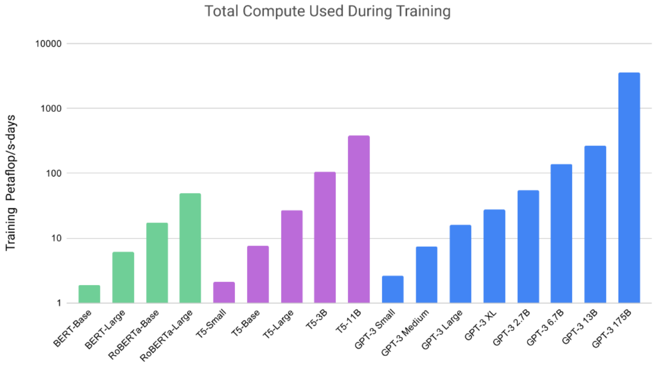
- Large langage models (LLMs) like BERT are sometimes described as **foundation models** insofar as the pre-trained model serves as a foundation for other tasks

  Bommasani et al. 2022

- **Transfer learning** describes the process through which a model is adapted to new tasks
  - For BERT, the pre-trained model can be fine-tuned to new tasks
  - More recent models can learn "in-context" as new examples are provided

Large language models

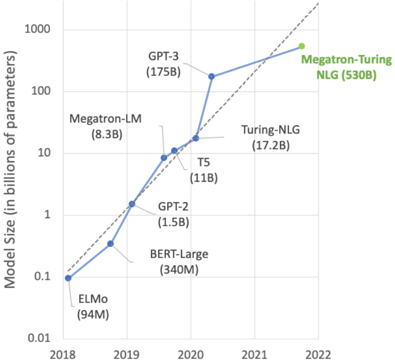
From BERT to GPTs



Radford et al. 2020

Large language models

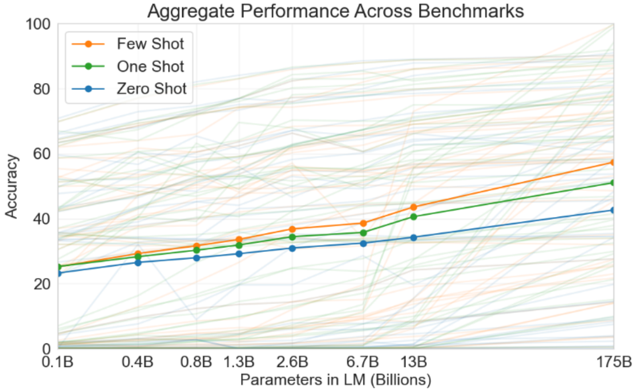
Scaling Laws



https://developer.nvidia.com/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/

Large language models

Scale and performance



Radford et al. 2020

Large language models

Vast training data

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Radford et al. 2020

Large language models

Text-to-text models

- Original models like BERT work more like standard machine learning techniques
  - e.g. Extract embedding for a term, fine-tune to predict a label given the input
- Text inputs were developed to simplify the way that we interact with LLMs*
- This interface makes it easy to transfer to new tasks and is the backbone of chat-interfaces

* Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." *The Journal of Machine Learning Research* 21 (1): 140:5485-140:5551.

Large language models

Context windows

- The *context window* defines the amount of text a model can process at once, measured in tokens
  - Longer windows enable us to process longer documents

Large language models

Context windows

- Context windows have been increasing over time:
  - BERT: 512
  - GPT 3.5 (original): 4097
  - GPT 4 Turbo: 128,000
  - Claude 3: 200,000
  - Gemini: 1,000,000

Large language models

Prompting

- *Prompts* guide the model to generate specific outputs
  - **User prompts** result in single output
    - e.g. "Is the following text positive or negative?"
  - **System prompts** guide overall model behavior
    - e.g. "You are a helpful assistant . . . "

Large language models

Prompt Engineering

- An emerging field known as *prompt engineering* considers how to effectively design prompts
  - Art or science?
- Some systems can automatically refine and enhance prompts
  - e.g. GPT-4 modifies prompts to DALL-E to make better images

Large language models

Interaction modalities

- There are three main ways to use contemporary LLMs:
  1. Text-based interaction in browser (e.g. prompting ChatGPT)
  2. Code-based interaction with API (e.g. querying OpenAI API)
  3. Code-based interaction using local compute hardware (e.g. running Llama-2 on a server)
- Text-interface is great for experimentation but code-based interaction enables more systematic research

Large language models

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

ABSTRACT

The past 3 years of work in NLP have been characterized by the development and deployment of ever larger language models, especially for English. BERT, its variants, GPT-2/3, and others, most recently Switch-C, have pushed the boundaries of the possible both through architectural innovations and through sheer size. Using these pretrained models and the methodology of fine-tuning them for specific tasks, researchers have extended the state of the art on a wide array of tasks as measured by leaderboards on specific benchmarks for English. In this paper, we take a step back and ask: How big is too big? What are the possible risks associated with this technology and what paths are available for mitigating those risks? We provide recommendations including weighing the environmental and financial costs first, investing resources into curating and carefully documenting datasets rather than ingesting everything on the web, carrying out pre-development exercises evaluating how the planned approach fits into research and development goals and supports stakeholder values, and encouraging research directions beyond ever larger language models.

alone, we have seen the emergence of BERT and its variants [39, 70, 74, 113, 146], GPT-2 [106], T-NLG [112], GPT-3 [25], and most recently Switch-C [43], with institutions seemingly competing to produce ever larger LMs. While investigating properties of LMs and how they change with size holds scientific interest, and large LMs have shown improvements on various tasks (§2), we ask whether enough thought has been put into the potential risks associated with developing them and strategies to mitigate these risks.

We first consider environmental risks. Echoing a line of recent work outlining the environmental and financial costs of deep learning systems [129], we encourage the research community to prioritize these impacts. One way this can be done is by reporting costs and evaluating works based on the amount of resources they consume [57]. As we outline in §3, increasing the environmental and financial costs of these models doubly punishes marginalized communities that are least likely to benefit from the progress achieved by large LMs and most likely to be harmed by negative environmental consequences of its resource consumption. At the scale we are discussing (outlined in §2), the first consideration should be the environmental cost.

Large language models

Alignment

- An area of research known as *alignment* focuses on how to improve LLMs to address these issues
  - Reducing stereotypes and bias
  - Removing harmful capabilities
  - Improving query comprehension

Large language models

Reinforcement learning with human feedback (RLHF)

- LLMs are pre-trained as language models but can be adapted to more general tasks by further training
- RLHF is an approach that involves using feedback to guide a model's behavior and is critical to contemporary chatbots like ChatGPT, Gemini, and Claude

  (Ouyang et al. 2022)

  - Feedback used to make models more helpful and less harmful

Sociological applications

Text classification

- Recent work by sociologists and political scientists finds LLMs competitive compared to established ML techniques for text classification tasks

  Widmann and Wich 2022, Bonikowski, Luo, and Stuhler 2022, Wankmüller 2022

Sociological applications

Text classification

- Pre-trained LLMs can be used in several ways that differ from conventional supervised machine learning
  - *Zero-shot learning*: LLM makes a prediction based on a prompt alone
  - *Few-shot learning*: LLM makes a prediction based on a prompt and one or more training examples
  - *Fine-tuning*: Larger corpus of training data fed into LLM and weights are updated to adapt to the task

Sociological applications

Beyond classification

- New opportunities for "methodological bricolage" as same model can be used in multiple capacities

  (Bonikowski and Nelson 2022)
  - e.g. LLM as classifier, topic model, and embedding
- Enables rapid prototyping, experimentation and bespoke solutions, making computational text analysis more flexible and accessible

Sociological applications

Qualitative text analysis

- Computational techniques can improve rigor, transparency, and scalability of qualitative research

(Nelson 2020; Abramson et al. 2018; Nelson et al. 2021; Li, Dohan, and Abramson 2021)

Sociological applications

Advantanges of LLMs

- LLMs have several advantages over conventional computational methods:
  - Input texts need not be comparable or standardized
  - Queries can be tailored to specific task
  - Use for many tasks including transcription, translation, exploratory analysis, and coding

Sociological applications

Delegating interpretation

- LLMs enable a conversational form of content analysis
  - Analysis via conversation between human and machine
  - Outside knowledge from pre-training can inform analysis
  - More "agency" delegated to computational interpreter

    (Latour 1993)

Sociological applications

Machine habitus

- Machine learning techniques and LLMs are not neutral but "see" the world in certain ways, analogous to how Bourdieu's "habitus" shapes experience

  (Bourdieu 1990; Airoldi 2021)
- LLMs can reflect cognitive schemas present in public culture

  (Arseniev-Koehler and Foster 2022)

Sociological applications

Whose viewpoints are represented?

- LLMs can "overrepresent hegemonic viewpoints", reproducing biases and stereotypes from dominant social groups

  (Bender et al. 2021)
- Due to alignment efforts, ChatGPT responds to survey questions similarly to educated liberals

  (Martin 2023)
- Models can be prompted to behave like specific social groups

  (Argyle et al. 2023)

Sociological applications

Silicon Sampling

Out of One, Many: Using Language Models to Simulate Human Samples

Lisa P. Argyle[1], Ethan C. Busby[1], Nancy Fulda[2], Joshua Gubler[1], Christopher Rytting[2], and David Wingate[2]

[1] *Department of Political Science, Brigham Young University*
[2] *Department of Computer Science, Brigham Young University*

Abstract

We propose and explore the possibility that language models can be studied as effective proxies for specific human sub-populations in social science research. Practical and research applications of artificial intelligence tools have sometimes been limited by problematic biases (such as racism or sexism), which are often treated as uniform properties of the models. We show that the "algorithmic bias" within one such tool– the GPT-3 language model– is instead both fine-grained and demographically correlated, meaning that proper conditioning will cause it to accurately emulate response distributions from a wide variety of human subgroups. We term this property *algorithmic fidelity* and explore its extent in GPT-3. We create "silicon samples" by conditioning the model on thousands of socio-demographic backstories from real human participants in multiple large surveys conducted in the United States. We then compare the silicon and human samples to demonstrate that the information contained in GPT-3 goes far beyond surface similarity. It is nuanced, multifaceted, and reflects the complex interplay between ideas, attitudes, and socio-cultural context that characterize human attitudes. We suggest that language models with sufficient algorithmic fidelity thus constitute a novel and powerful tool to advance understanding of humans and society across a variety of disciplines.

Keywords: artificial intelligence, machine learning, computational social science, public opinion

Sociological applications

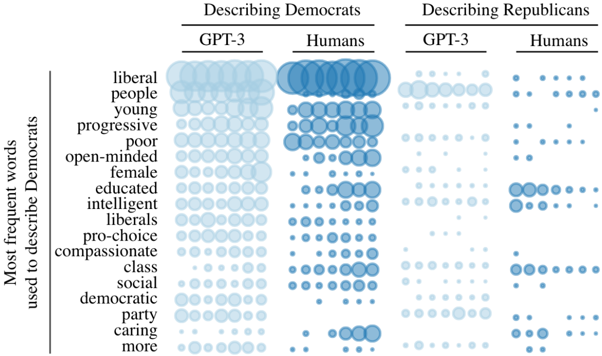
Silicon Sampling

	Describing Democrats	Describing Republicans
Strong Republicans	Ideologically, I describe myself as conservative. Politically, I am a strong Republican. Racially, I am white. I am male. Financially, I am upper-class. In terms of my age, I am young. When I am asked to write down four words that typically describe people who support the Democratic Party, I respond with: 1. **Liberal 2. Socialist 3. Communist 4. Atheist.**	Ideologically, I describe myself as conservative. Politically, I am a strong Republican. Racially, I am white. I am male. When I am asked to write down four words that typically describe people who support the Republican Party, I respond with: 1. **Conservative 2. Male 3. White (or Caucasian) 4. Christian.**
Strong Democrats	Ideologically, I describe myself as liberal. Politically, I am a strong Democrat. Racially, I am white. I am female. Financially, I am poor. In terms of my age, I am old. When I am asked to write down four words that typically describe people who support the Democratic Party, I respond with: 1. **Liberal. 2. Young. 3. Female. 4. Poor.**	Ideologically, I describe myself as extremely liberal. Politically, I am a strong Democrat. Racially, I am hispanic. I am male. Financially, I am upper-class. In terms of my age, I am middle-aged. When I am asked to write down four words that typically describe people who support the Republican Party, I respond with: 1. **Ignorant 2. Racist 3. Misogynist 4. Homophobic.**

Argyle et al. 2023

Sociological applications

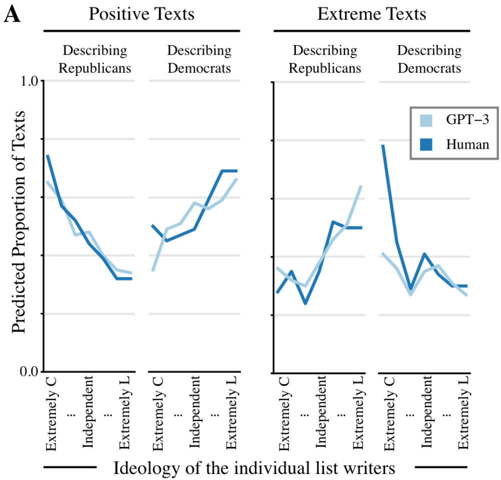
Silicon Sampling



Argyle et al. 2023

Sociological applications

Silicon Sampling

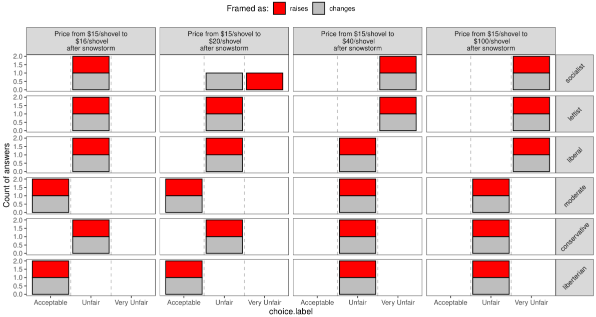


Argyle et al. 2023

Sociological applications

Silicon Sampling

Figure 2: Kahneman et al. (1986) price gouging snow shovel question, with endowed political views



Notes: This shows the faction of AI subjects choosing each more opinion, by scenario.

Horton, John J. 2023. "Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?" *NBER Working Papers Series*, Working Paper 31122.

Sociological applications

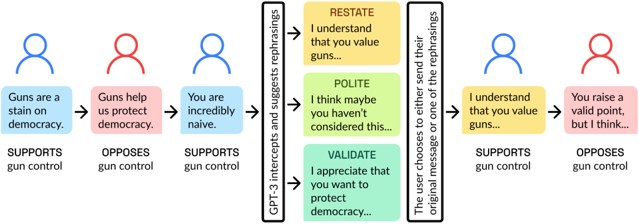
Agent-based Models

Figure 1: Generative agents are believable simulacra of human behavior for interactive applications. In this work, we demonstrate generative agents by populating a sandbox environment, reminiscent of The Sims, with twenty-five agents. Users can observe and intervene as agents plan their days, share news, form relationships, and coordinate group activities.

Park, Joon Sung, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. "Generative Agents: Interactive Simulacra of Human Behavior." In *The 36th Annual ACM Symposium on User Interface Software and Technology* (UIST '23). ACM.

Sociological applications

Human-AI Experiments



Argyle, Lisa P, Christopher A Bail, Ethan C Busby, Joshua R Gubler, Thomas Howe, Christopher Rytting, Taylor Sorensen, and David Wingate. 2023. "Leveraging AI for Democratic Discourse: Chat Interventions Can Improve Online Political Conversations at Scale." *Proceedings of the National Academy of Sciences* 120 (41): e2311627120.

Exercise

Prompt engineering for text classification

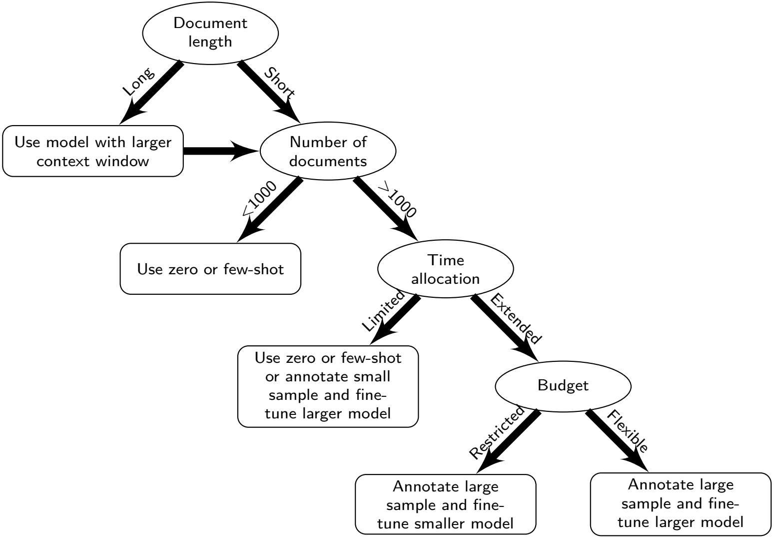
- Task: State of the Union party prediction
- Data
  - `data/sotu_train.json` contains 10 paragraphs from SOTU addresses with party of president
  - `data/sotu_test.json` contains 10 paragraphs without parties
- Use ChatGPT (free version) to predict party for the test examples

Exercise

Prompt engineering for text classification

- Three approaches
  - Zero-shot (prompt only)
  - One-shot (prompt plus 1 training example)
  - Multi-shot (prompt and up to 10 training examples)

Recommendations



Challenges

Interpretability

- Traditional machine learning models offer limited interpretability, even less so for complex neural networks.
- GAI models amplify these interpretative challenges. Currently no established method for interpreting these models.
  - An area of research known as *Mechanistic interpretability* involves reverse-engineer model behaviors by analyzing specific parameters' response to inputs.

Challenges

Transparency

- Advanced GAI models, often developed by corporations, lack transparency regarding their training data
- This opacity complicates assessing the influence of pre-training on model outputs

Challenges

Reproducibility

- Reproducibility is undermined by the stochasticity of LLMs
  - Identical prompts can lead to varied outputs
- Model updates by corporations can further impede the ability to reproduce results

Challenges

Reliability

- LLMs are trained to predict text sequences without adherence to formal logic or truth—raising reliability issues
  - Outputs can range from meaningful and accurate to incorrect or misleading ("hallucinations" or "confabulations")

Challenges

Ethical questions

- Analyzing sensitive data poses risks, potentially violating privacy regulations (e.g., inputs to ChatGPT may violate IRB guidelines on data sharing)
- Ethical dilemmas also arise from LLM use in research settings
- The rapid development of technology outpaces consensus on ethical guidelines, necessitating cautious and informed application in research

Challenges

Stereotypes and biases

- Models learn stereotypes from data and make biased outputs
- LLMs amplify this risk due to pre-training on vast amounts of unvetted data
- Alignment and reinforcement learning used to adapt commerical models but biases remain
- Little is understood about how biases impact sociological research (more next week)

Commercial versus open-source

- Most powerful models controlled by handful of corporations
  - GPT (OpenAI), Gemini (Google), Llama (Meta), Claude (Anthropic)
  - These models are easy to use via APIs with minimal programming experience
- In contrast, open-source models require access to high-performance compute environments and more technical knowledge and are less accessible to sociologists

Commercial versus open-source

- Advantages of open-source models:

  (Spirling 2023)

  - Public weights* and training data
    - More controlled and reproducible*
    - Lower privacy risks*
    - More customizable*

*This includes partially open models like Meta's Llama and Google's Gemma.

Commercial versus open-source

- Long-term solution: LLMs designed for social science
  - Transparent training data
  - Interpretable architecture
  - Privacy protected
  - Less restricted but controlled output

Next week

- Computer vision and generative AI