

MAS.S68, Spring 2023

2/8

Generative AI for Constructive Communication

Evaluation and New Research Methods



Welcome!



**center for
constructive
communication**

Instructors



Jad Kabbara



Suyash Fulay



Shayne Longpre



Hang Jiang



Hope Schroeder



Doug Beeferman



Deb K. Roy

A journey of course descriptions

The sudden accessibility of large language models (LLMs) have made seemingly impossible kinds of communication ubiquitous almost overnight. How do these models generate, affect, and streamline communication, and how can we better evaluate new research facilitated by these advances across fields? In particular, we look at how LLM advances touch related research areas, including HCI, social sciences, governance, journalism, and beyond.

What research potential is opened by these technological advances? What is unknown? Come be part of shaping a dynamic workshop + seminar on this rapidly changing topic!

ChatGPT: Rewrite this to sound more **exciting** to **MIT students**:

Get ready to dive into the world of cutting-edge communication technology! With the breakthrough of large language models, communication has taken a **huge leap forward**. But what does this mean for us, and how can we make the most of these advancements? Join us for a **thrilling** 6 unit seminar that will give you **a front-row seat to this exciting revolution**. With a combination of expert speakers, engaging paper discussions, exciting design critiques, and hands-on research collaboration, you'll be part of shaping the future of communication. **Don't miss out** on this opportunity to be a part of the conversation. So, bring your laptop and be ready to learn, create, and innovate. The first class is on Wednesday, 2/8 – fill out the interest form **now!**

A journey of course descriptions

Get ready to be part of the conversation shaping the future of **communication technology**! **With the breakthrough of large language models, communication has taken a huge leap forward. But what does this mean for us, and how can we make the most of these advancements?** What kinds of research and evaluation are made urgent by these technological advances? Come be part of shaping a dynamic workshop + seminar on this rapidly changing topic! You'll hear from experts, engage in paper discussions, participate in design critiques, and collaborate on cutting-edge research. If you want to participate, [fill out this interest form](#) and come to the first class on Wednesday, 2/8– bring a laptop and be prepared to get your hands dirty!

Deb edited another version of this for an academic audience:

Advances in large language models recently popularized by ChatGPT represent a **remarkable leap forward** in language processing by machines. **We invite you** to join the **conversation shaping the future of communication technology**. What does this mean for us, how can we make the most of these advances, and what are the risks? What research opportunities have opened up? What kinds of evaluation are called for? We will bring together a group of practitioners and experts for guided discussions, hands-on experimentation, and project critiques. If you want to join the class, please fill out this [interest form](#) and come to the first class on Wednesday, Feb 8th. Bring a laptop and be prepared to start experimenting!

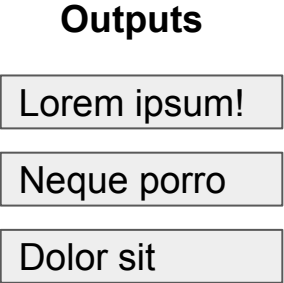
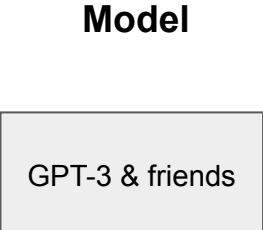
Observations

- The model took the word “**exciting**” and the audience of “**MIT students**” to introduce the term “thrilling,” “front row seat,” and mentioned “communication technology”
- We took those outputs and edited to tone it down
- Final product was clearer and more engaging as a result of human-AI collaboration
- Should we have disclosed our use of ChatGPT as an editing tool? How is it different than Grammarly?
 - Does this affect trust... with our advisor? With you?

Maybe you've wrestled with these questions!

Motivation

These models open many exciting possibilities, but also require **evaluation**: the models, the systems that contain them, and the societal impacts (positive and negative) they can have



**What are they?
How should we
evaluate them?**

And their outputs?

Used in user-facing systems (writing assistants, chat bots)

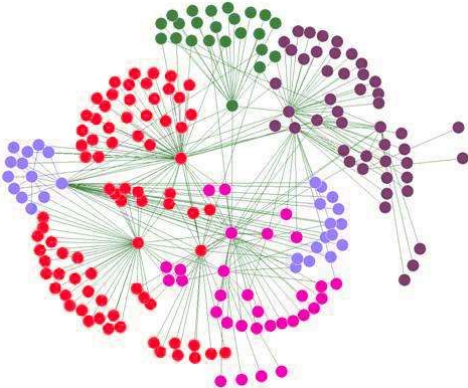
Used *for* social science research

Used as inputs to other processes

And how people will interact with their outputs?



...at scale?



Course structure:

Models: NLP overview of models and their evaluation

Today: Introduction to the history of the models and how the models work

Next few weeks: NLP view on evaluating the models, their capabilities, and how humans interact with them

Future topics:

- Using models to learn about people
- Evaluating how interacting with models affects people
- How the technology can impact society, for better and for worse

Lots of implicated research areas: NLP, human-computer interaction, computational social science, communications, ethics and transparency... we need all your backgrounds!

Course evaluation:

Participation:

- Attendance
- Pre-work for discussions
 - Preparing questions for speakers
 - Answering discussion questions
 - Assignments
- Presenting project progress

Research Project

- Project proposal or final project

Project can be chosen to align with your background!

Course logistics

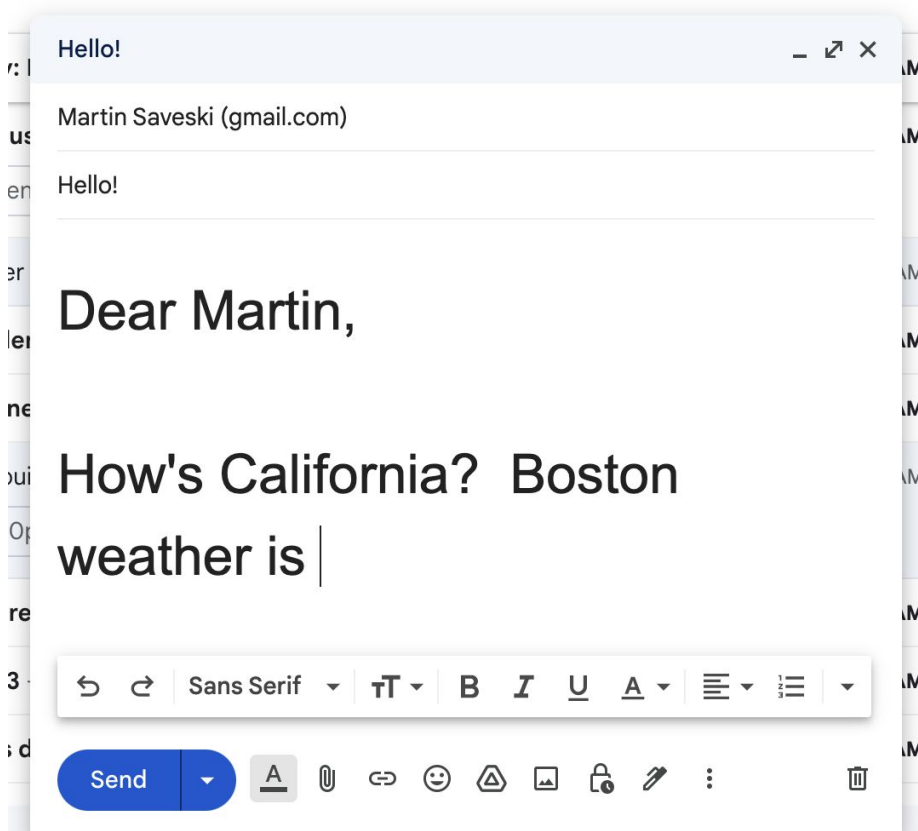
- **MAS.S68:** 6 units
- **Cross-registration:** we welcome it, just make sure to do the forms on the end of your university!
- **Listener policy:**
 - We will prioritize registered full-credit students
 - Registered listeners will participate in pre-work for discussions
 - Select talks may have a Zoom link for non-registered folks who want to drop in
- Sign in so we know you were here today!

How did we get here?

- What are language models?
- What are they useful for?
- Quick recap of the past 30 years of LM research



What is a language model?



What is a language model?

For our purposes, it's a **probability distribution over all the sequences of words that might be spoken or written** (in some language, in some context)

Sentence	Probability
Aardvarks ate apples	0.00000000241
...	...
Boston weather is callous	0.0000000121
Boston weather is cold	0.0000234
Boston weather is cork	0.00000000291
Boston weather is crane	0.00000000185
Boston weather is crazy	0.00000322
Boston weather is furious	0.00000000112
Boston weather is frigid	0.0000321
...	...
Zyzyyx zork zaphod	0.00000000112

What is a language model?


For our purposes, it's a **probability distribution over all the sequences of words that might be spoken or written** (in some language, in some context)

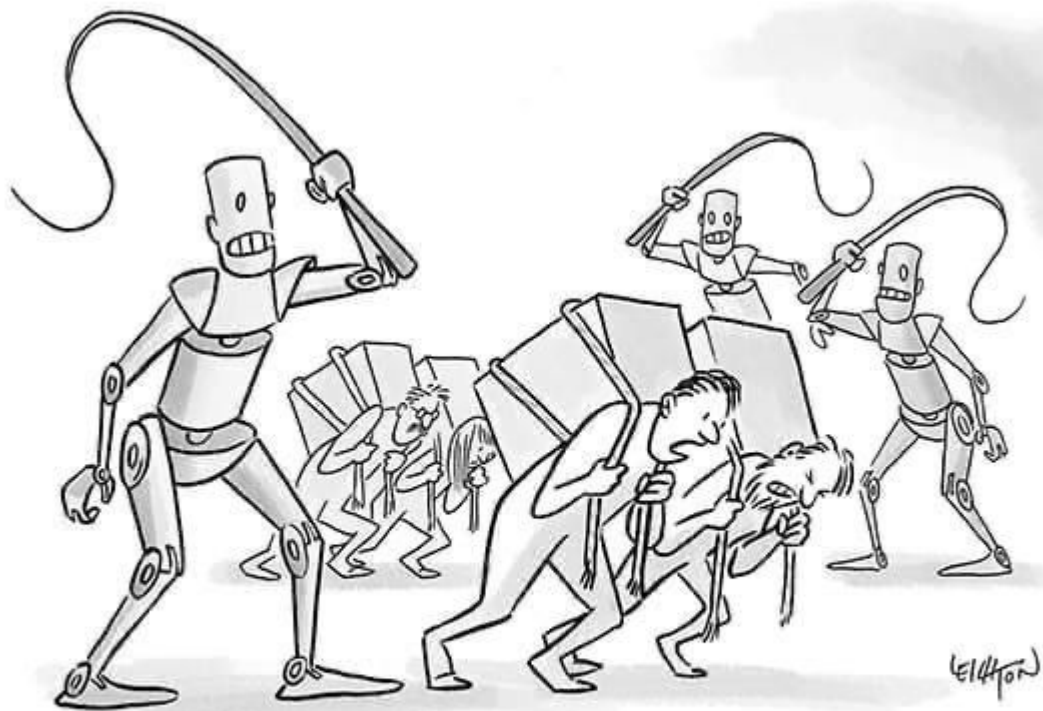
Sentence	Probability
Aardvarks ate apples	0.00000000241
...	...
Boston weather is callous	0.0000000121
Boston weather is cold	0.0000234
Boston weather is cork	0.00000000291
Boston weather is crane	0.00000000185
Boston weather is crazy	0.00000322
Boston weather is furious	0.00000000112
Boston weather is frigid	0.0000321
...	...
Zyzyyx zork zaphod	0.00000000112

Boston weather is...

<i>w</i>	$P(w \mid \text{Boston weather is})$
cold	0.4
frigid	0.23
terrible	0.12
great	0.02
fantastic	0.01
miserable	0.008
warm	0.005
hot	0.002

$P(\text{"Boston weather is cold"}) / P(\text{"Boston weather is *"})$

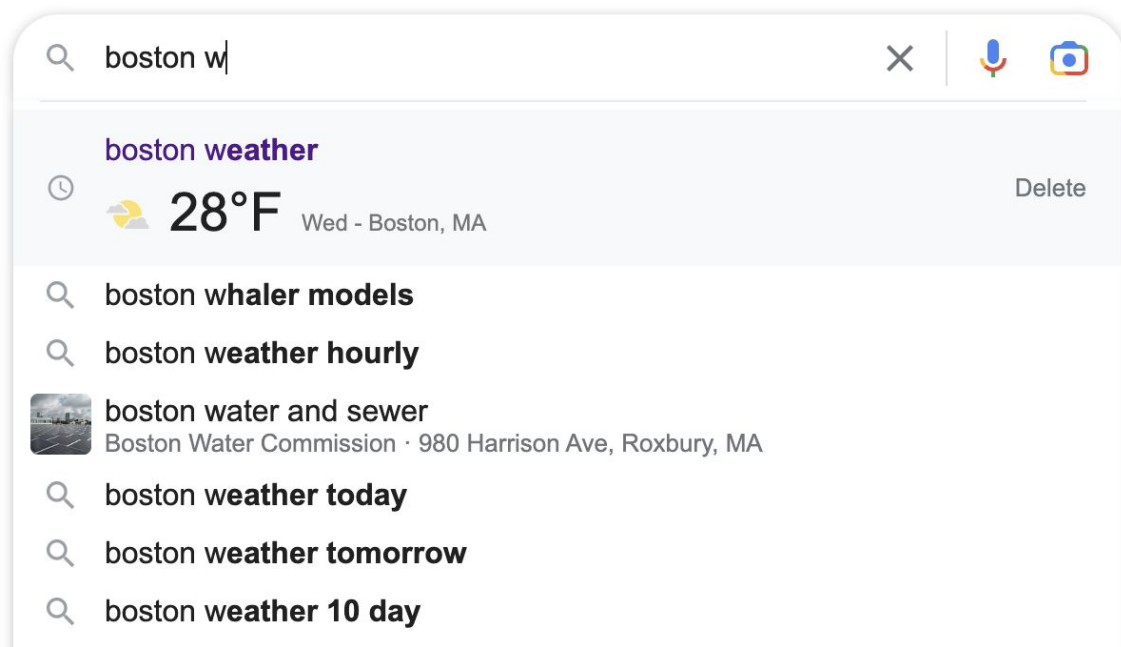




“To think this all began with letting autocomplete finish our sentences.”

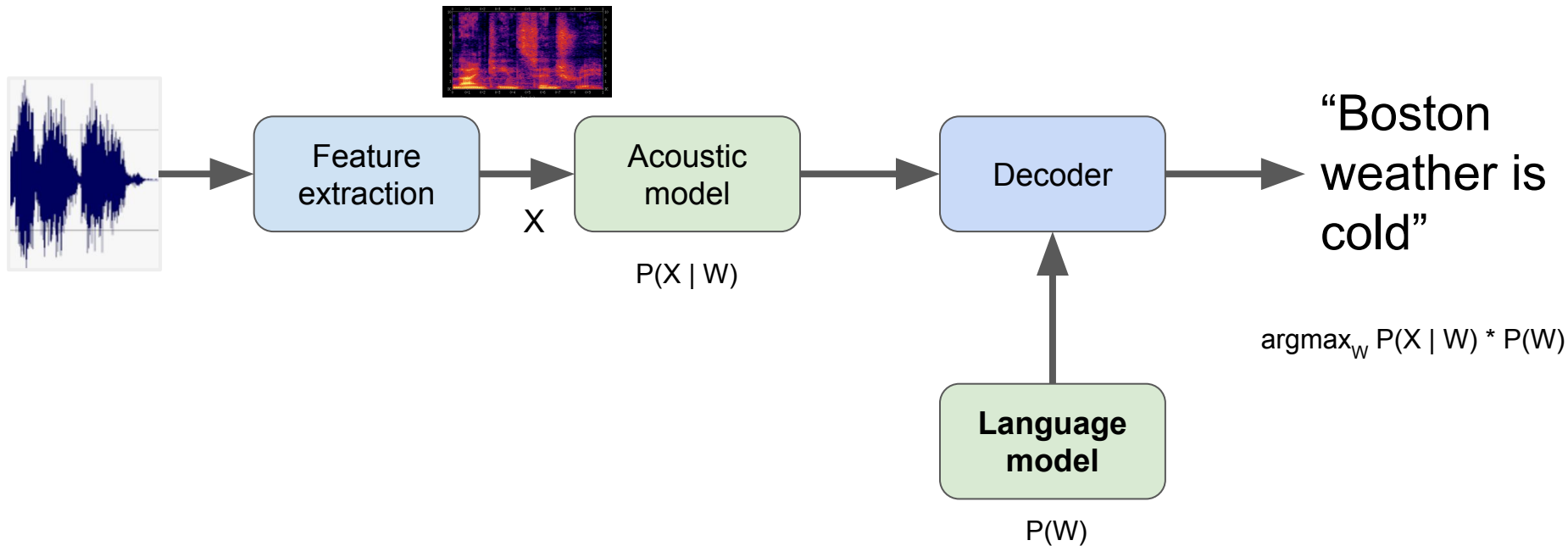
Classical applications of LMs: as helpers

“Auto-complete” interfaces



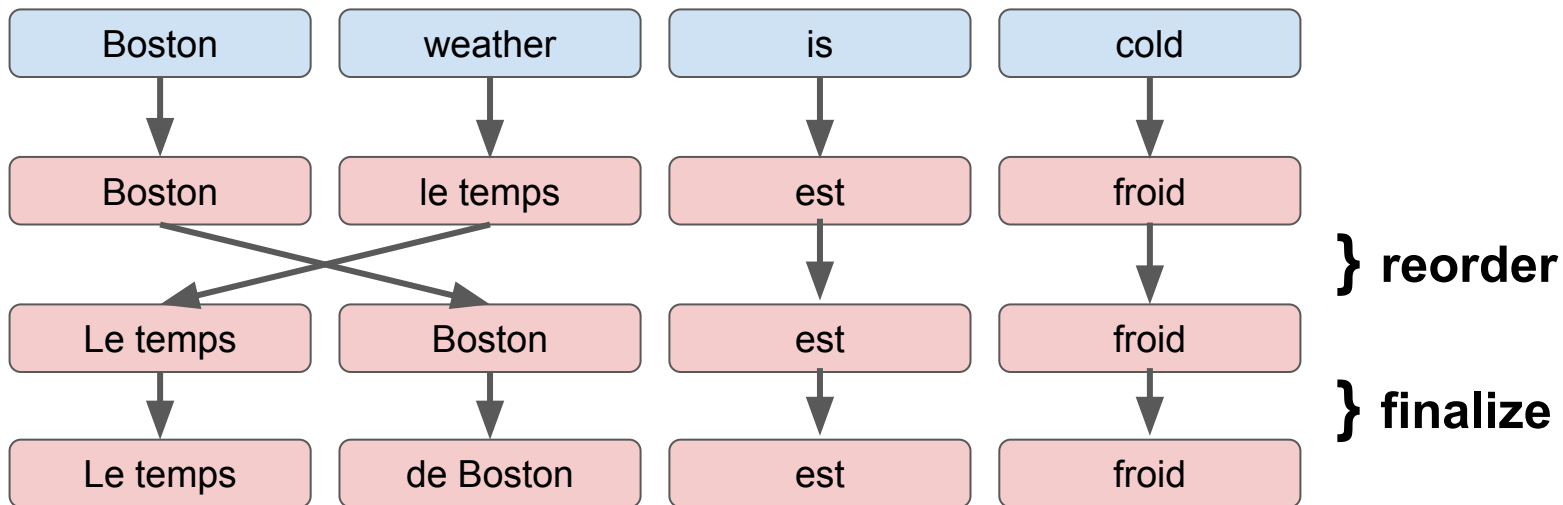
Classical applications of LMs: as helpers

Speech recognition































Classical applications of LMs: as helpers

Language Translation

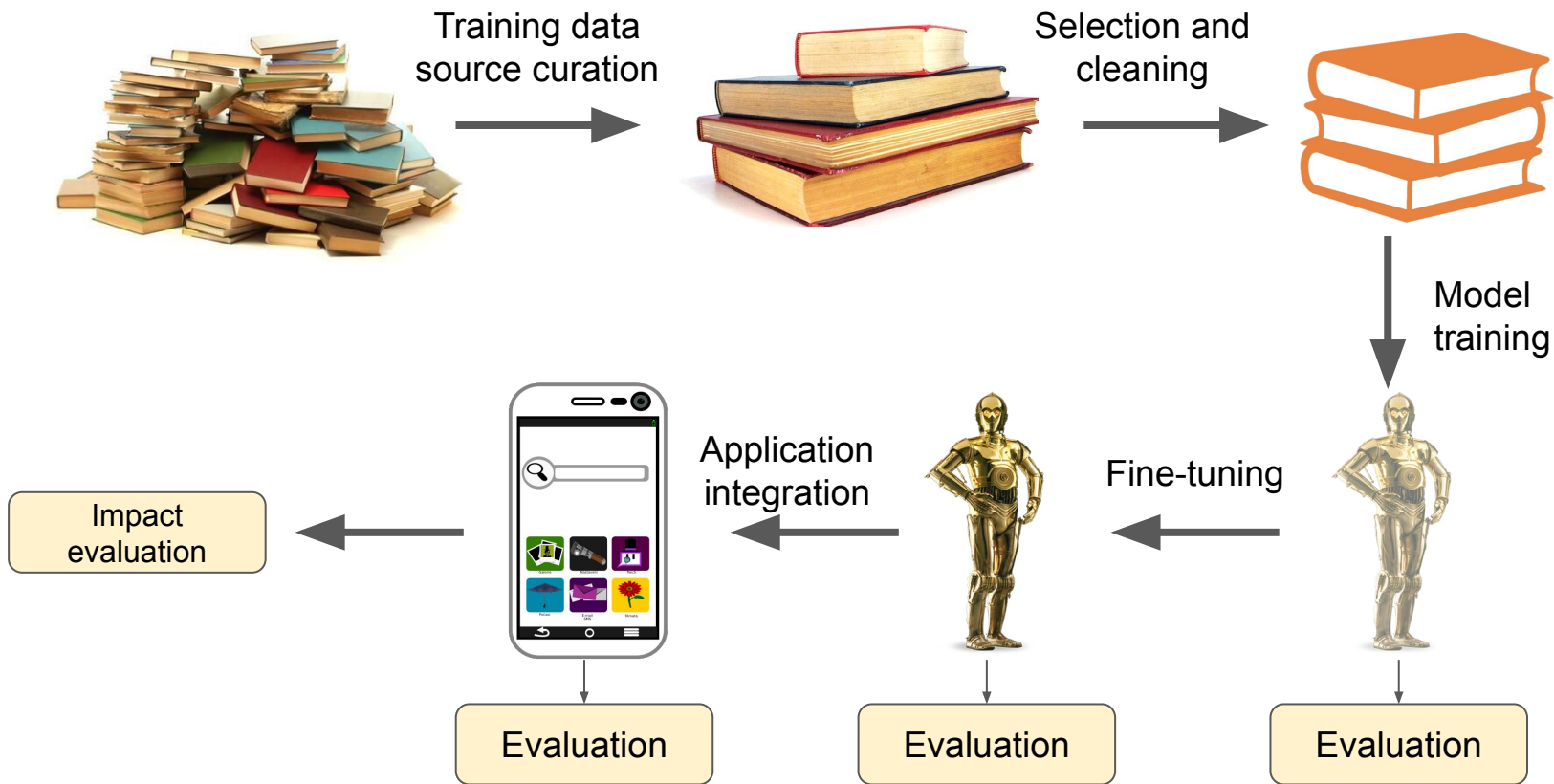


Modern applications of LMs: as generators

(From the [GPT-3 examples page](#))

 Q&A Answer questions based on existing knowle...	 Grammar correction Corrects sentences into standard English.	 Advanced tweet classifier Advanced sentiment detection for a piece o...	 Explain code Explain a complicated piece of code.
 Summarize for a 2nd grader Translates difficult text into simpler concep...	 Natural language to OpenAI API Create code to call to the OpenAI API usin...	 Keywords Extract keywords from a block of text.	 Factual answering Guide the model towards factual answering ...
 Text to command Translate text into programmatic commands.	 English to other languages Translates English text into French, Spanish...	 Ad from product description Turn a product description into ad copy.	 Product name generator Create product names from examples word...
 Natural language to Stripe API Create code to call the Stripe API using nat...	 SQL translate Translate natural language to SQL queries.	 TL;DR summarization Summarize text by adding a 'tl;dr;' to the en...	 Python bug fixer Find and fix bugs in source code.
 Parse unstructured data Create tables from long form text	 Classification Classify items into categories via example.	 Spreadsheet creator Create spreadsheets of various kinds of dat...	 JavaScript helper chatbot Message-style bot that answers JavaScript ...
 Python to natural language Explain a piece of Python code in human un...	 Movie to Emoji Convert movie titles into emoji.	 ML/AI language model tutor Bot that answers questions about language...	 Science fiction book list maker Create a list of items for a given topic.
 Calculate Time Complexity Find the time complexity of a function.	 Translate programming languages Translate from one programming language ...	 Tweet classifier Basic sentiment detection for a piece of text.	 Airport code extractor Extract airport codes from text.

Typical life of an LM



The unigram model

$$P(\text{"Boston weather is cold."}) \cong P(\text{"Boston"}) * P(\text{"weather"}) * P(\text{"is"}) * P(\text{"cold"}) * P(\text{."})$$

- Ignores word order
- Ignores context
- Extremely fast to train: Just count the words in the training data
- Terrible

N-gram models: the original generative LM

Markov assumption
(2-gram model)

$$P(w_n | w_1^{n-1}) \approx P(w_n | w_{n-1})$$

<s>	Boston	weather	is	cold	</s>
<s>	Boston	weather	is	cold	</s>
<s>	Boston	weather	is	cold	</s>
<s>	Boston	weather	is	cold	</s>
<s>	Boston	weather	is	cold	</s>

ChatNGram

A 3-gram model trained on news articles, prompted with “Boston weather is”:

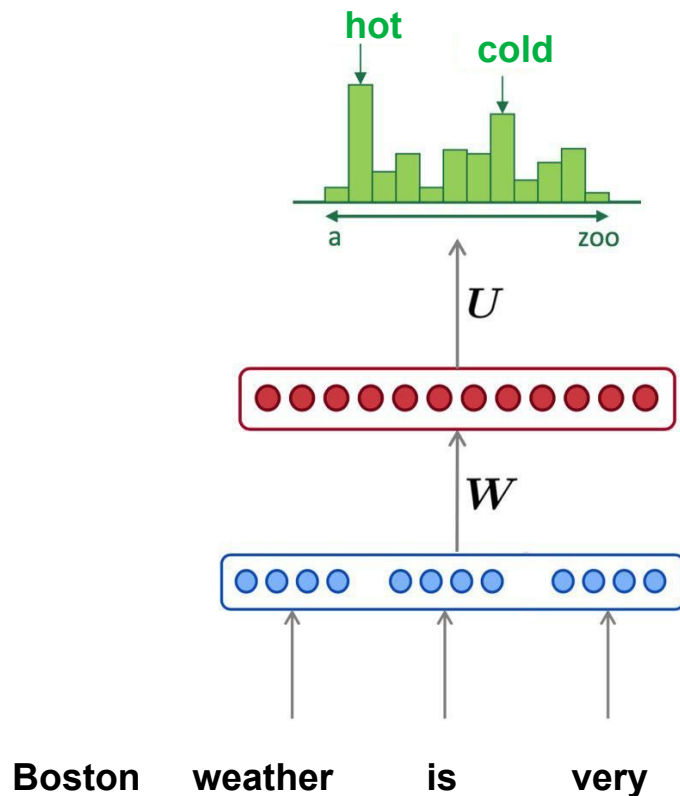
Boston weather is likely to remain in a separate dispute, shipping sources said it will cut dividend to 35,000 avg shrs \$7,214,000 for delivery during march last year in ethiopia is about \$648,000 and 25,000 tons were 2.53 percent of its forward strategy would lead to an autonomous european monetary system...

Late '90s-2000s: Feature engineering

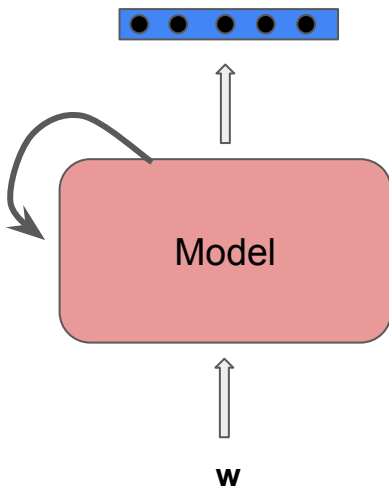
- Class-based n-gram models
- Syntax models
- Topic models
- Trigger pairs / long-distance features
- Maximum entropy models

2000s-2010s: Neural networks (re)enter the chat

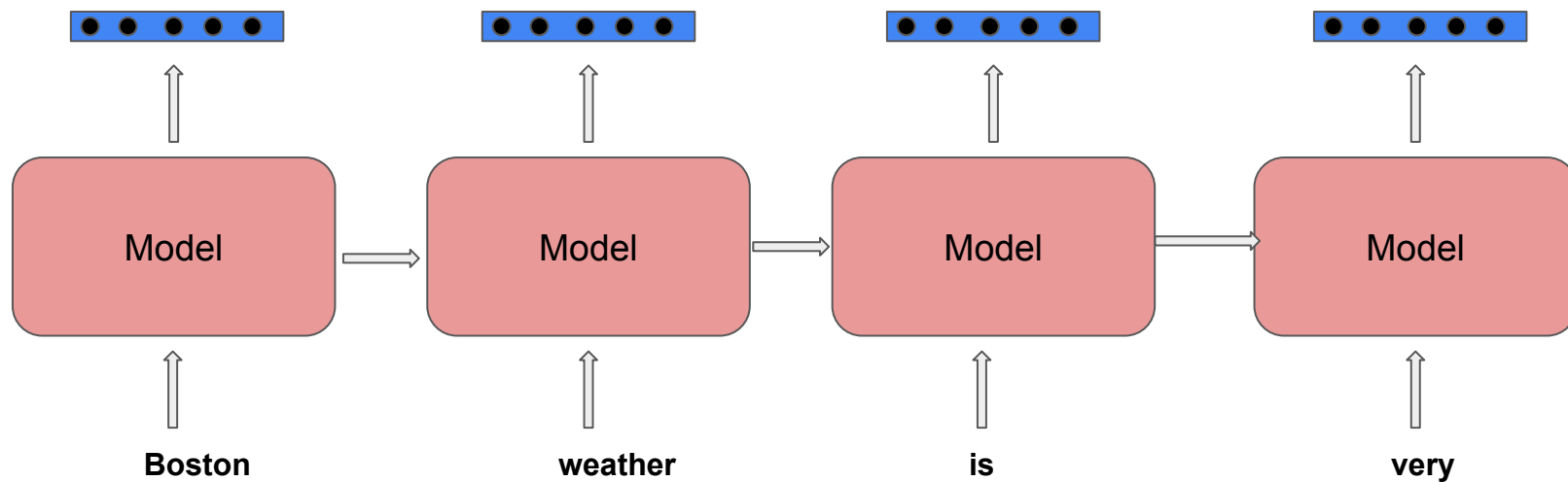
2003: Fixed-context neural LMs
([Bengio](#))



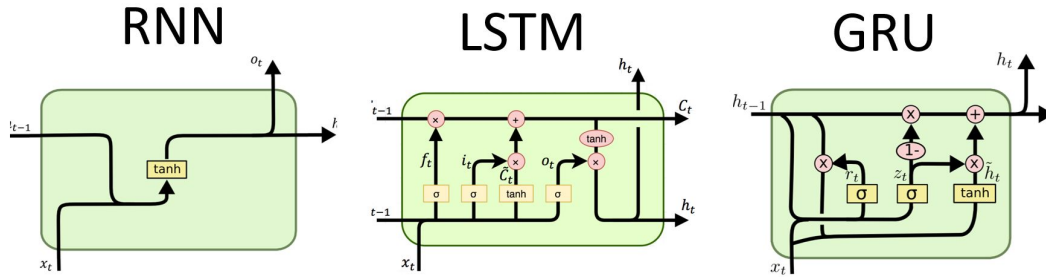
Recurrent neural net (RNN)-based LMs



Recurrent neural net (RNN)-based LMs



LSTMs (1997, 2014-)



Word embeddings

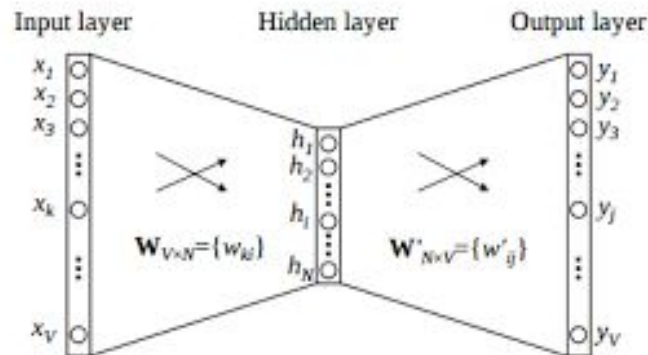
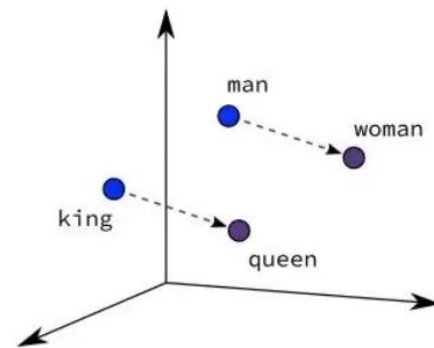
Count vectors, LSA, etc. 1990s - 2000s

	Document 1	Document 2	Document 3	Document 4	Document 5	Document 6	Document 7	Document 8
Term(s) 1	10	0	1	0	0	0	0	2
Term(s) 2	0	2	0	0	0	18	0	2
Term(s) 3	0	0	0	0	0	0	0	2
Term(s) 4	6	0	0	4	6	0	0	0
Term(s) 5	0	0	0	0	0	0	0	2
Term(s) 6	0	0	1	0	0	1	0	0
Term(s) 7	0	1	8	0	0	0	0	0
Term(s) 8	0	0	0	0	0	3	0	0

← Document Vector

← Word Vector (Passage Vector)

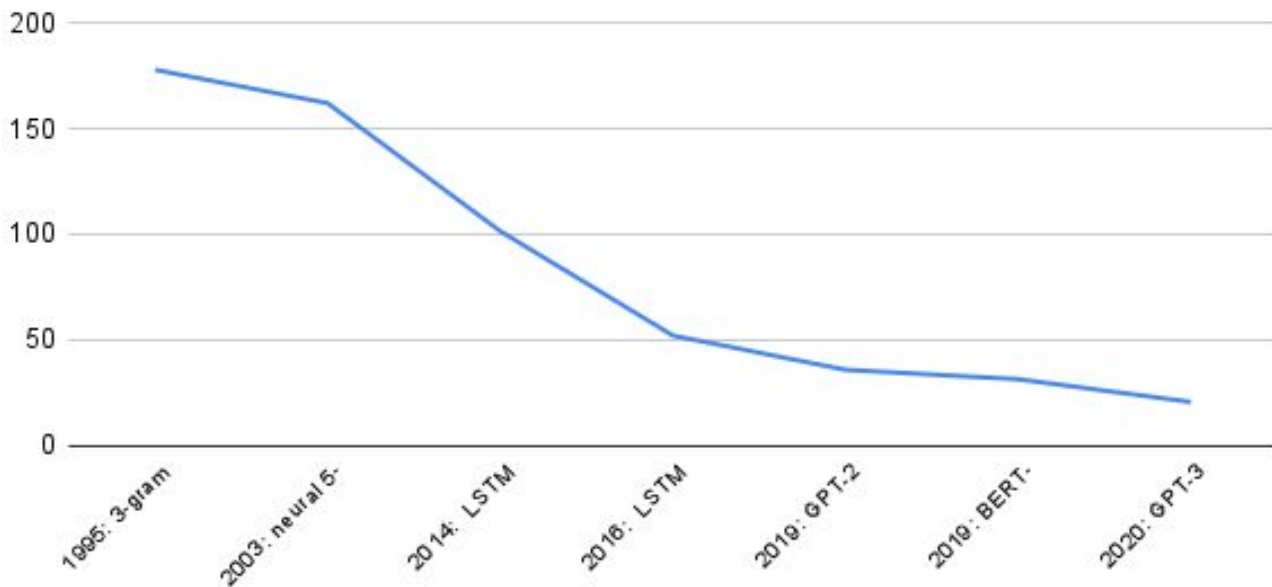
Word2Vec, etc., c. 2013



word2vec model architecture

How far we've come: predictive power

Perplexity over time (Penn Treebank test set)



LM

$$PP(W) = P(w_1 w_2 \dots w_N)^{\frac{1}{N}}$$

How far we've come: fluency

1990s: 3-gram model trained on news articles:

Boston weather is likely to remain in a separate dispute, shipping sources said it will cut dividend to 35,000 avg shrs \$7,214,000 for delivery during march last year in ethiopia is about \$648,000 and 25,000 tons were 2.53 percent of its forward strategy would lead to an autonomous european monetary system...

2023: GPT-3:

Boston weather is generally warm and humid in the summer, with temperatures typically reaching the mid-80s Fahrenheit (around 28-29°C). You can also expect lots of sunshine and occasional thunderstorms.

LLM Fundamentals



The era of *Large* Transformer-based models

- 2017: Attention is all you need (Vaswani et al. 2017)
 - Recurrent layers are not needed
 - Attention mechanisms, fully connected layers

Attention Visualizations

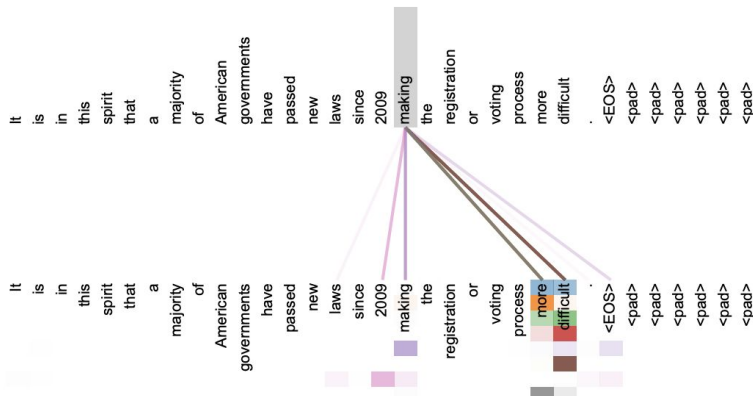
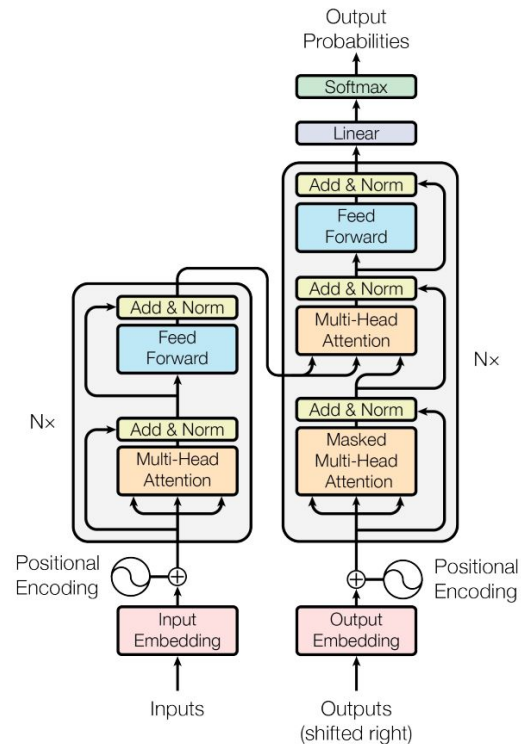
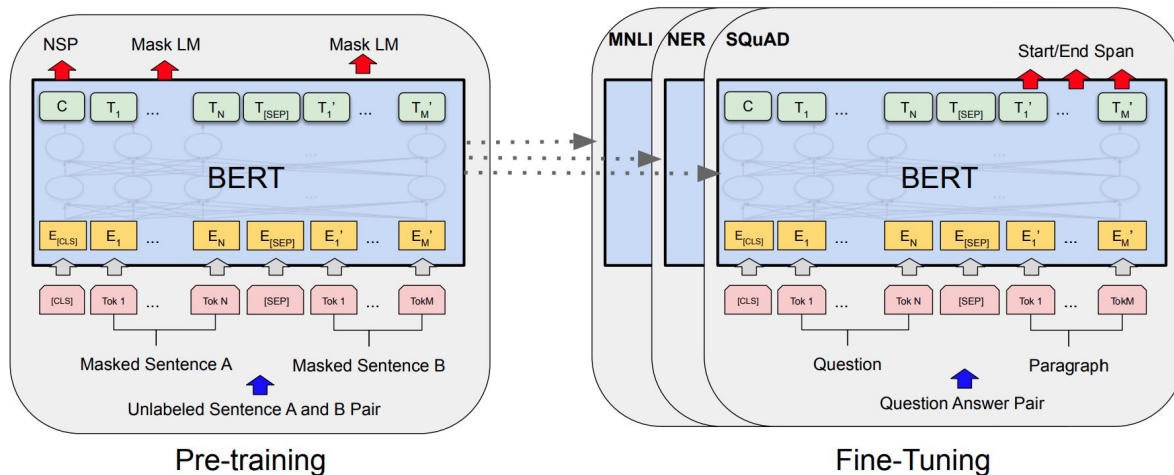


Figure 3: An example of the attention mechanism following long-distance dependencies in the encoder self-attention in layer 5 of 6. Many of the attention heads attend to a distant dependency of the verb 'making', completing the phrase 'making...more difficult'. Attentions here shown only for the word 'making'. Different colors represent different heads. Best viewed in color.



The era of *Large* Transformer-based models

- 2018/2019: BERT (Devlin et al. 2019)
 - Based on the original Transformer architecture
 - Popularized the pre-train/fine-tune paradigm

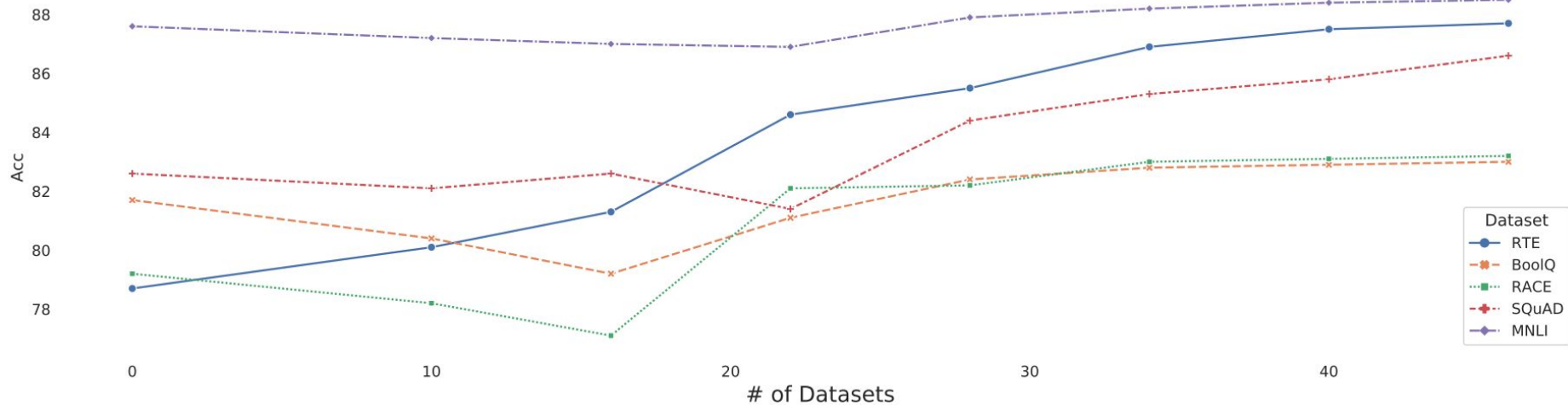


The Pretrain / Finetune paradigm

- Pretraining: Train LM on gigantic amounts of data
 - Various possible training objectives:
 - Masked Language Modeling: Task of predicting a missing word in a sentence.
 - Example: Large language [MASK] are super cool.
 - Next sentence prediction: [CLS] Sentence 1 [SEP] Sentence 2 [EOS]
 - Others
- Finetuning: Model has acquired a lot of “knowledge” -> Adapt it to specific tasks
 - Classification tasks (sentiment analysis, natural language inference, etc)
 - Generation (translation, summarization)

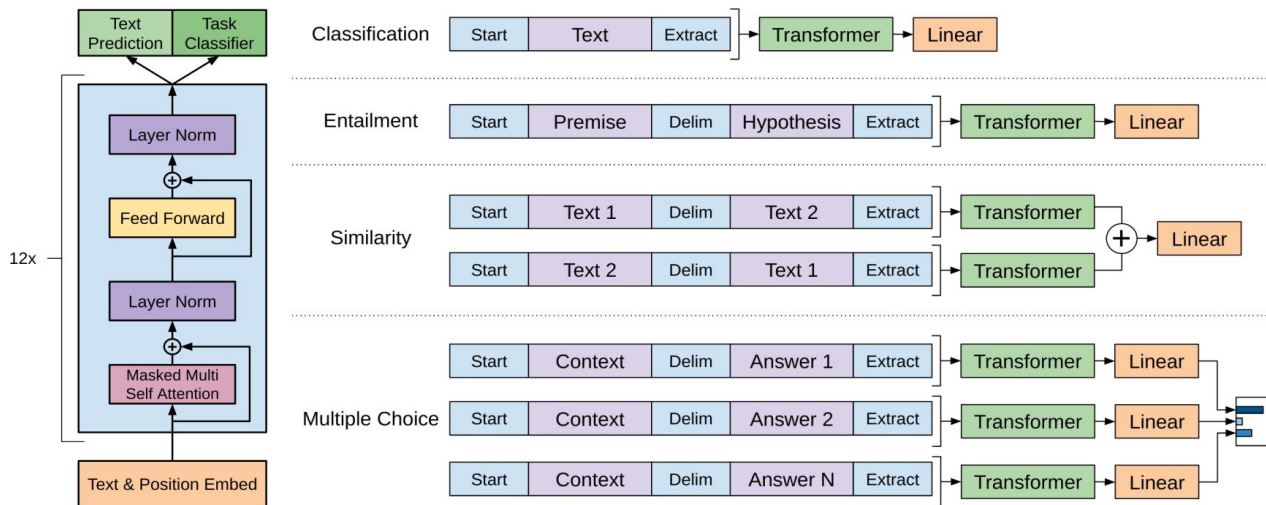
The Pretrain / Finetune paradigm

- Many studies showed how powerful the pretrain-finetune paradigm is.
- Example: Muppet (Aghajanyan et al 2021, Facebook)

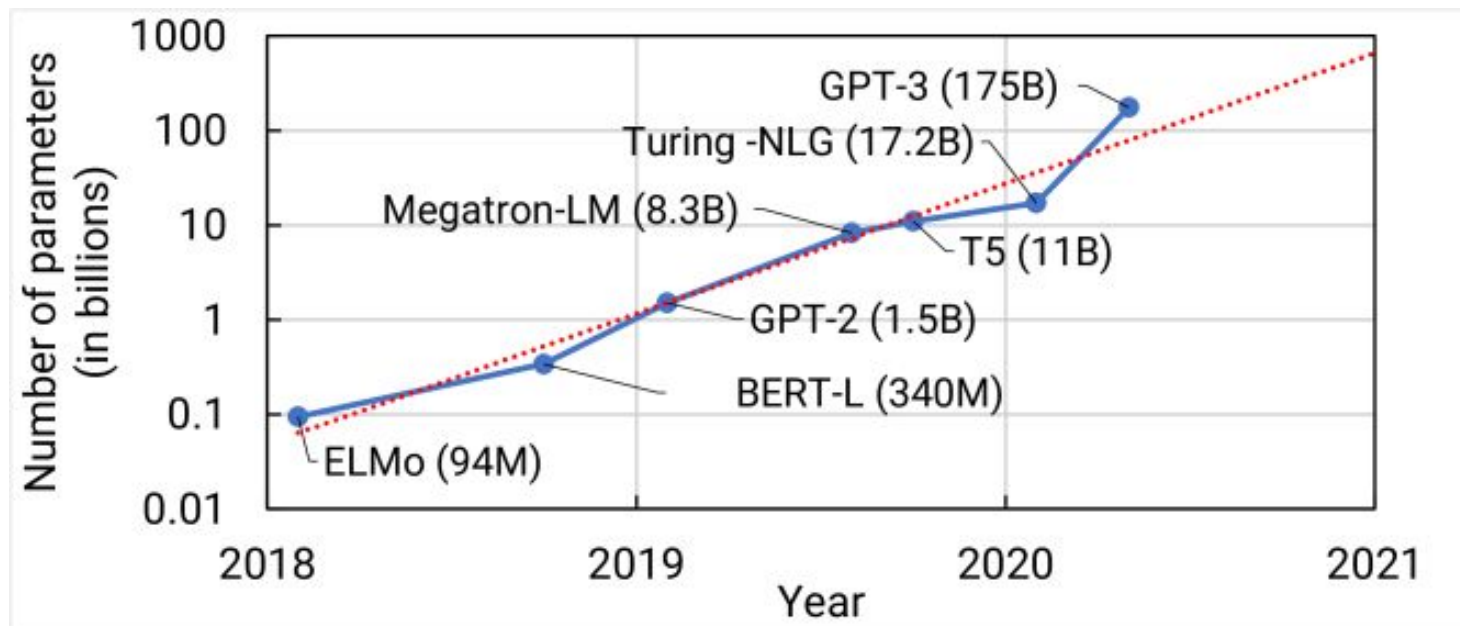


The era of *Large* Transformer-based models

- 2018: GPT (Radford et al. 2018)
 - Based on the original Transformer architecture

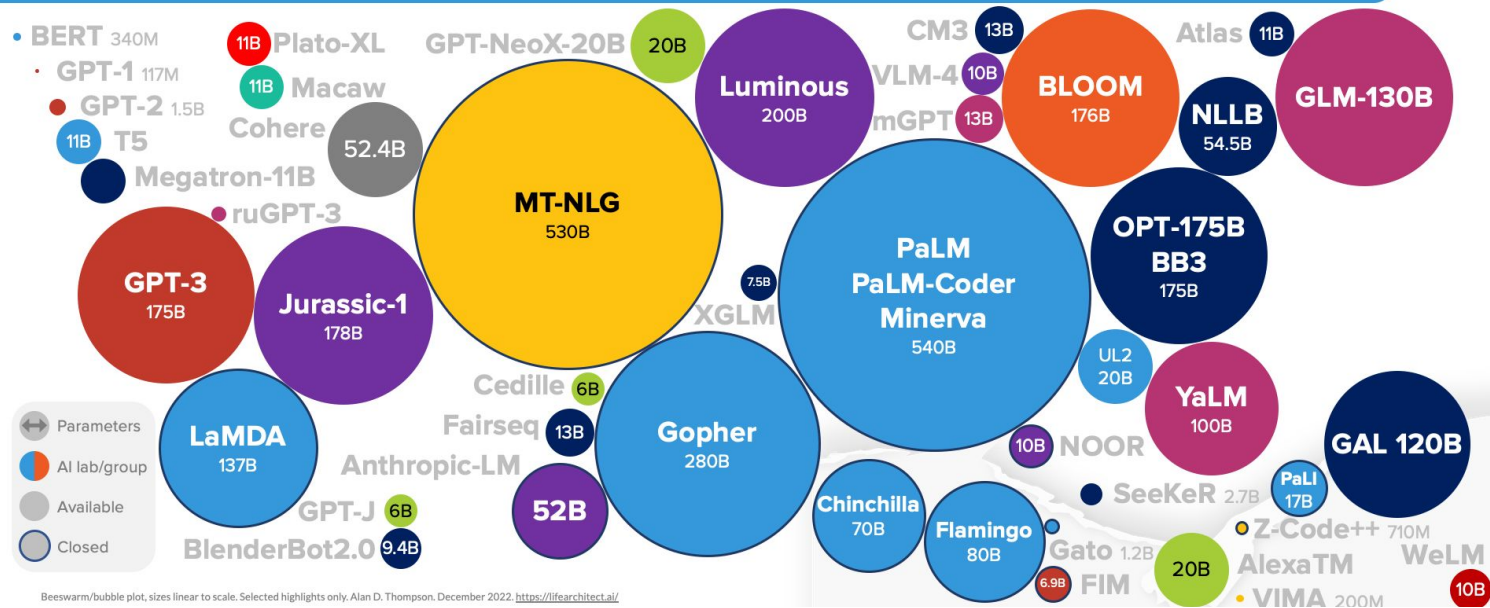


The era of *Large* Transformer-based models



The era of *Large* Transformer-based models

LANGUAGE MODEL SIZES TO DEC/2022



Interacting with LMs



Prompting

- Model fine-tuning used to be necessary
- Larger (or instruction tuned) models give intelligible responses even without it – just prompt them!
- “The model fine-tunes us!”
- Trade-off: Fine-tuning still outperforms prompting, but prompting doesn't require lots of training data.

The three settings we explore for in-context learning

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description
2 cheese => ..... ← prompt
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description
2 sea otter => loutre de mer ← example
3 cheese => ..... ← prompt
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description
2 sea otter => loutre de mer ← examples
3 peppermint => menthe poivrée ←
4 plush girafe => girafe peluche ←
5 cheese => ..... ← prompt
```

Traditional fine-tuning (not used for GPT-3)

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.



Prompting

Zero-Shot

[Instruction] **[Input]**

*Is the sentiment positive or negative?
"This movie sucks!" A:*

One-Shot No-Instruction

[Ex In 1] **[Ex Out 1]** **[Input]**

*Q: This movie rocks! A: Positive.
Q: "This movie sucks!" A:*

One-Shot

[Instruction] **[Ex In 1]** **[Ex Out 1]** **[Input]**

*Is the sentiment positive or negative? Q: This
movie rocks! A: Positive. Q: "This movie sucks!"
A:*

Two-Shot

[Instruction] **[Ex In 1]** **[Ex Out 1]**
[Ex In 2] **[Ex Out 2]** **[Input]**

*Is the sentiment positive or negative? Q: This
movie rocks! A: Positive. Q: My eyes are
bleeding! A: Negative. Q: "This movie sucks!" A:*

Chain-of-Thought

[Instruction]
[Request step-by-step explanation]
[Input]

*Is the sentiment positive or negative? Please
explain your answer step-by-step. Q: "This
movie sucks!" A:*

Prompting

Ideas: How would you edit your prompts to get better responses?

Give it many examples!

Give it diverse examples!

Fine-tune the model for a prompt!

Find the best / consistent prompt!

Ask the model to explain/verify its answer!

Try Many Prompts!
(Ensembling / self consistency)

Give it examples specific to this input!

Prompt tuning

Options to Optimize for a prompt:

- Option A: Fine-tune your model to work best with a specific prompt
- Option B: Find the best prompt for your (unchanged) model
- Option C: Instead of finding “word prompts”, find a series of vectors that when fed to your model, give it the best idea what task you are asking for.

Fine-tune the model for
a prompt!

Find the best / consistent
prompt!

Resources:

- Liu et al (2021). “Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing”
- Lester et al (2021). “The Power of Scale for Parameter-Efficient Prompt Tuning”

Prompting examples

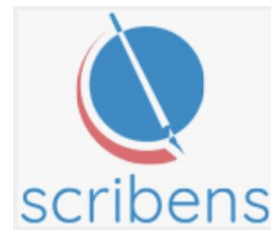


Prompt example 1 – grammar correction

Q: Correct this to standard English:

Anna and Mike is going skiing.

Model: Anna and Mike are going skiing




Prompt example 2 – summarization

Large language models (LLMs) have shown promise for automatic summarization but the reasons behind their successes are poorly understood. By conducting a human evaluation on ten LLMs across different pretraining methods, prompts, and model scales, we make two important observations. First, we find instruction tuning, and not model size, is the key to the LLM's zero-shot summarization capability. Second, existing studies have been limited by low-quality references, leading to underestimates of human performance and lower few-shot and finetuning performance. To better evaluate LLMs, we perform human evaluation over high-quality summaries we collect from freelance writers. Despite major stylistic differences such as the amount of paraphrasing, we find that LLM summaries are judged to be on par with human written summaries.

What is the main idea of the article?

Prompt example 3 – style transfer

Here is some text: {When Mohammed left the theatre, it was already dark out.}. Rewrite it to be more about the movie itself.
{The movie was longer than Mohammed had expected, and despite the excellent ratings he was a bit disappointed when he left the theatre.}. 

Here is some text: {next to the path}. Rewrite it to be about France.
{next to la Siene}.

Here is some text: {The man stood outside the grocery store, ringing the bell.}. Rewrite it to be about clowns.
{The man stood outside the circus, holding a bunch of balloons.}.

Here is some text: {the bell ringing}. Rewrite it to be more flowery.
{the peales of the jangling bell}.

Here is some text: {against the tree}. Rewrite it to be include the word "snow".
{against the snow-covered bark of the tree}.

Here is some text: {We should increase the government tax rate on profits earned from the sale of stocks, bonds, and real estate.}. Rewrite it in the style of Kanye West.

Prompt example 4 – “Let’s think step by step.”

A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue.



How many blue golf balls are there?

Prompt example 5 – Joke Explanation

Explaining a Joke

Input: Did you see that Google just hired an eloquent whale for their TPU team? It showed them how to communicate between two different pods!

Model Output: TPUs are a type of computer chip that Google uses for deep learning. A "pod" is a group of TPUs. A "pod" is also a group of whales. The joke is that the whale is able to communicate between two groups of whales, but the speaker is pretending that the whale is able to communicate between two groups of TPUs.

Group Activity



Group activity

Come up with a new task!

- Break off into 4 person groups
- Go to OpenAI.com
- Sign up for an account, launch GPT-3 playground
- With at least 1 working OpenAI account per group, brainstorm a new task
- Full workshop description [here](#)

Course logistics

- **MAS.S68:** 6 units
- **Cross-registration:** we welcome it, just make sure to do the forms on the end of your university!
- **Listener policy:**
 - We will prioritize registered full-credit students
 - Registered listeners will participate in pre-work for discussions
 - Select talks may have a Zoom link for non-registered folks who want to drop in
- **Communication:** Canvas for assignments, Slack to share articles, news, and info
- Sign in today so we know you were here!
- Also check out: *6.S986 Large Language Models and Beyond* with Yoon Kim

Discussion assignment

Investigate a user-facing system that incorporates LLMs

- Pick a user-facing product that incorporates an LLM and use it. Examples:
 - **YouWrite**
 - **Character.AI**
 - Note: some require getting off a waitlist. Start early!
- Consider some questions:
 - What data might this have been trained on? What language models does it use?
 - How might you study its effect on a user?
 - What might the societal impact be if this becomes popular?
- Canvas us short responses, come prepared to discuss your critique!

Next time

First hour:

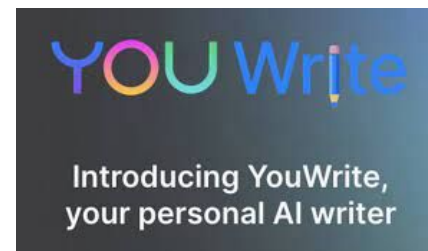
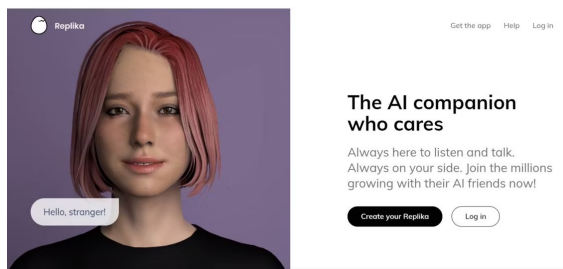
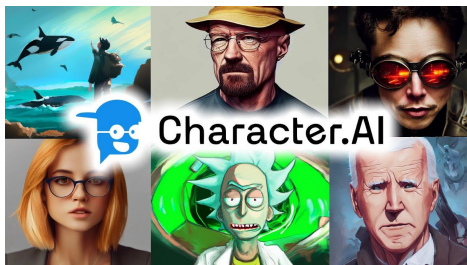
Introduction to evaluating large language models

Guest speaker: Rishi Bommasani

Center for Research on Foundation Models, NLP group at Stanford University

Talk title: “Holistically Evaluating Language Models on the path to Evaluating Foundation Models”

Second hour: critiquing LLMs in user-facing systems



Upcoming talks

2/22

Fine-tuning language models to find agreement among humans with diverse preferences

Guest speaker: Michiel Bakker

DeepMind

3/1

Emergent abilities in language models

Guest speaker: Jason Wei

Google Brain

3/8

Evaluating Human-Language Model Interaction (HALIE)

Guest speaker: Mina Lee

Stanford NLP