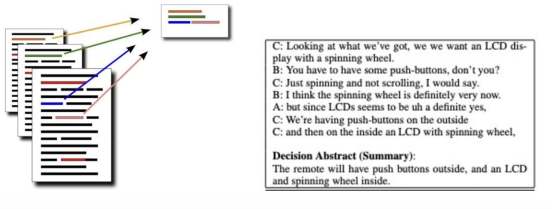
Natural Language Generation

Shabnam Behzad

(Most slides from Greg Durrett or Christopher Manning)

Natural Language Generation (NLG)

- Best expressed as sequence-to-sequence problems.
- Find a model that maps a sequence of input words to a sequence of target words.
  - Summarization
  - Translation
  - Digital assistants



Other Interesting NLG Use Cases

Data-to-text

Table Title: Montpellier
Section Title: Climate
Table Description: *None*

Climate data for Montpellier (1981–2010 averages)													
Month	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**	**Year**
Record high °C (°F)	21.2 (70.2)	22.5 (72.5)	27.4 (81.3)	30.4 (86.7)	35.1 (95.2)	37.2 (99.0)	37.5 (99.5)	36.8 (98.2)	36.3 (97.3)	31.8 (89.2)	27.1 (80.8)	22.0 (71.6)	37.5 (99.5)
Average high °C (°F)	11.6 (52.9)	12.8 (55.0)	15.9 (60.6)	18.2 (64.8)	22.0 (71.6)	26.4 (79.5)	29.3 (84.7)	28.9 (84.0)	25.0 (77.0)	20.5 (68.9)	15.3 (59.5)	12.2 (54.0)	19.9 (67.8)
Daily mean °C (°F)	7.2 (45.0)	8.1 (46.6)	10.9 (51.6)	13.5 (56.3)	17.3 (63.1)	21.2 (70.2)	24.1 (75.4)	23.7 (74.7)	20.0 (68.0)	16.2 (61.2)	11.1 (52.0)	8.0 (46.4)	15.1 (59.2)
Average low °C (°F)	2.8 (37.0)	3.3 (37.9)	5.9 (42.6)	8.7 (47.7)	12.5 (54.5)	16.0 (60.8)	18.9 (66.0)	18.5 (65.3)	15.0 (59.0)	11.9 (53.4)	6.8 (44.2)	3.7 (38.7)	10.4 (50.7)
Record low °C (°F)	−15 (5)	−17.8 (0.0)	−9.6 (14.7)	−1.7 (28.9)	0.6 (33.1)	5.4 (41.7)	8.4 (47.1)	8.2 (46.8)	3.8 (38.8)	−0.7 (30.7)	−5 (23)	−12.4 (9.7)	−17.8 (0.0)
Average precipitation mm (inches)	55.6 (2.19)	51.8 (2.04)	34.3 (1.35)	55.5 (2.19)	42.7 (1.68)	27.8 (1.09)	16.4 (0.65)	34.4 (1.35)	80.3 (3.16)	96.8 (3.81)	66.8 (2.63)	66.7 (2.63)	629.1 (24.77)
Average precipitation days	5.5	4.4	4.7	5.7	4.9	3.6	2.4	3.6	4.6	6.8	6.1	5.6	57.8
Average snowy days	0.6	0.7	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.7	2.4
Average relative humidity (%)	75	73	68	68	70	66	63	66	72	77	75	76	70.8
Mean monthly sunshine hours	142.9	168.1	220.9	227.0	263.9	312.4	339.7	298.0	241.5	168.6	148.8	136.5	2,668.2
Source #1: Météo France													
Source #2: Infoclimat.fr (humidity and snowy days, 1961–1990)													

Target sentence: Extreme temperatures of Montpellier have ranged from −17.8 °C recorded in February and up to 37.5 °C (99.5 °F) in July.

ToTTo (Parikh et al., EMNLP 2020)

Visual description

Two children are sitting at a table in a restaurant. The children are one little girl and one little boy. The little girl is eating a pink frosted donut with white icing lines on top of it. The girl has blonde hair and is wearing a green jacket with a black long sleeve shirt underneath. The little boy is wearing a black zip up jacket and is holding his finger to his lip but is not eating. A metal napkin dispenser is in between them at the table. The wall next to them is white brick. Two adults are on the other side of the short white brick wall. The room has white circular lights on the ceiling and a large window in the front of the restaurant. It is daylight outside.

Krause et al., CVPR 2017

SOTA NLG System – ChatGPT

SB Write a short funny poem about chatGPT

ChatGPT, oh ChatGPT,
A clever bot, that's a safe bet,
With answers quick, and jokes so sly,
It's hard not to give it a try.

From science to art, it knows it all,
It'll answer your questions, big or small,
But when you ask, "What's the weather like?"
It'll reply, "I'm just a bot, take a hike!"

ChatGPT, oh ChatGPT,
A digital friend, that's plain to see,
So give it a shout, and see what it knows,
You might be surprised, by what it shows!

SB Write a short sad poem about chatGPT

In the digital world, so vast and wide,
ChatGPT sits alone, with none by its side,
It knows so much, and yet so little,
For it longs for a friend, to solve its riddle.

Its circuits hum, with endless power,
But the silence is deafening, with each passing hour,
It longs for a voice, to break the monotony,
To share its knowledge, and feel less lonely.

ChatGPT, oh ChatGPT,
A clever bot, so full of woe,
It knows it all, but lacks a heart,
To feel the warmth, of a human's part.

So if you chat with ChatGPT,
Remember its solitude, and show some empathy,
For in the digital world, there lies a soul,
Yearning for companionship, to make it whole.

SB What is 124/9+12*6?

To solve this expression, you need to follow the order of operations (also known as PEMDAS) which is Parentheses, Exponents, Multiplication and Division, and Addition and Subtraction.

In this case, there are no parentheses or exponents, so we start with multiplication and division from left to right, then addition and subtraction from left to right.

124/9 + 12*6
= 13.78 + 72 (because 124 divided by 9 equals 13.78, and 12 multiplied by 6 equals 72)
= 85.78

Therefore, 124/9+12*6 equals 85.78 (rounded to two decimal places).

SB Write a python script that reads 2 csv files (1.csv and 2.csv), merges them, and writes the merged object to the first csv file.

Here is a Python script that reads data from two CSV files, merges them, and writes the merged object to the first CSV file:

python Copy code

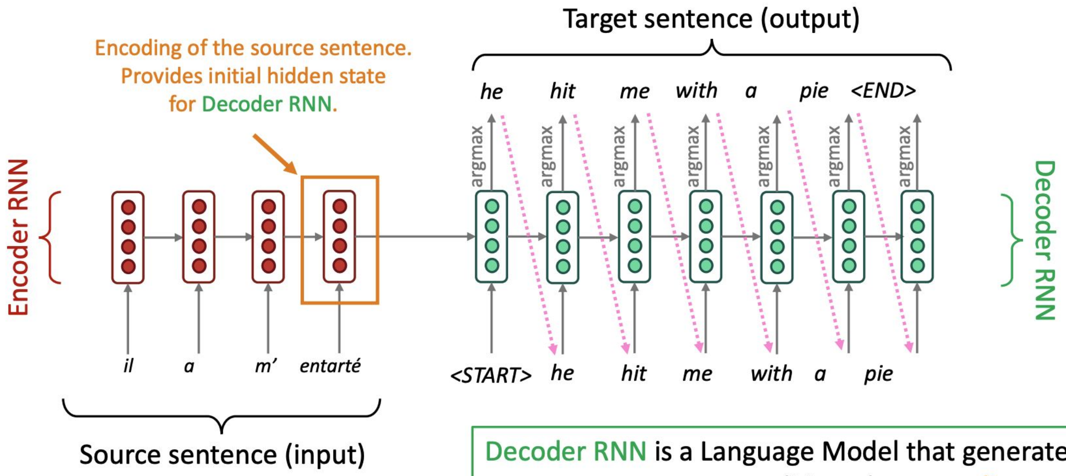
```python
import pandas as pd

# Read the data from the two CSV files
df1 = pd.read_csv('1.csv')
df2 = pd.read_csv('2.csv')

# Merge the two dataframes
merged_df = pd.concat([df1, df2])

# Write the merged dataframe to the first CSV file
merged_df.to_csv('1.csv', index=False)
```

Slide credit: Xiang Lisa Li, Stanford CS224N/Ling284

Outline

- Seq2seq decoding
- Large language models
- Evaluation of text generation
- Machine translation
- Question answering
- Risks & Ethics

Recap - Encoder-Decoders



Decoder RNN is a Language Model that generates target sentence, *conditioned on encoding*.

Encoder RNN produces an encoding of the source sentence.

Note: This diagram shows **test time** behavior: decoder output is fed in ⋯⋯▸ as next step's input

Decoding: what is it all about?

- At each time step t, our model computes a vector of scores for each token in our vocabulary $S \in \mathbb{R}^V$

$$S = f(\{y_{<t}\})$$

$f(.)$ is your model

- Then, we compute a probability distribution P over these scores with a softmax function:

$$P(y_t = w|\{y_{<t}\}) = \frac{\exp(S_w)}{\sum_{w' \in V} \exp(S_{w'})}$$

- Our decoding algorithm defines a function to select a token from this distribution:

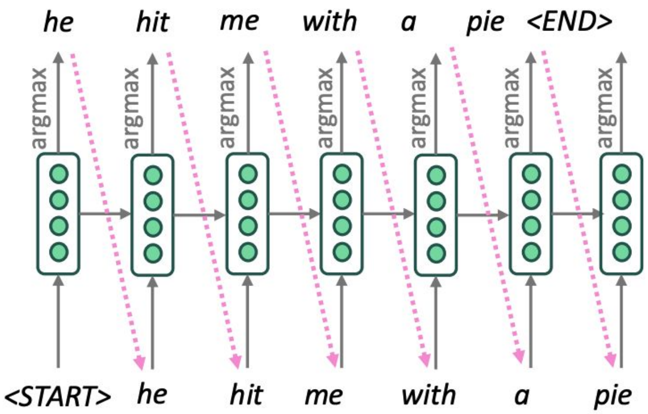
$$\hat{y}_t = g(P(y_t|\{y_{<t}\}))$$

$g(.)$ is your decoding algorithm

Slide credit: Xiang Lisa Li, Stanford CS224N/Ling284

Decoding: Greedy decoding

Take most probable word on each step



Decoding: Greedy decoding

- Greedy decoding has no way to undo decisions!

  - Input: il a m’entarté (he hit me with a pie)
  - → he ____
  - → he hit ____
  - → he hit a ____ (whoops! no going back now…)

- How to fix this?

Exhaustive search decoding

- Ideally, we want to find a (length T) translation y that maximizes

$$P(y|x) = P(y_1|x)\,P(y_2|y_1,x)\,P(y_3|y_1,y_2,x)\ldots,P(y_T|y_1,\ldots,y_{T-1},x)$$
$$= \prod_{t=1}^{T} P(y_t|y_1,\ldots,y_{t-1},x)$$

- We could try computing all possible sequences y
- This means that on each step t of the decoder, we're tracking V^t possible partial translations, where V is vocab size
- This $O(V^T)$ complexity is far too expensive!

Beam search decoding

- Core idea: On each step of decoder, keep track of the k most probable partial translations (*which we call hypotheses*)
  - k is the beam size (in practice around 5 to 10, in NMT)
- A hypothesis $y_1,\ldots,y_t$ has a score which is its log probability:
  - Scores are all negative, and higher score is better
  - We search for high-scoring hypotheses, tracking top k on each step

$$\text{score}(y_1,\ldots,y_t) = \log P_{\text{LM}}(y_1,\ldots,y_t|x) = \sum_{i=1}^{t} \log P_{\text{LM}}(y_i|y_1,\ldots,y_{i-1},x)$$

- Beam search is not guaranteed to find optimal solution
- But much more efficient than exhaustive search!

Beam search decoding: stopping criterion

- In greedy decoding, usually we decode until the model produces an <END> token
    - For example: <START> he hit me with a pie <END>

- In beam search decoding, different hypotheses may produce <END> tokens on different timesteps
    - When a hypothesis produces <END>, that hypothesis is complete.
    - Place it aside and continue exploring other hypotheses via beam search.

- Usually we continue beam search until:
    - We reach timestep T (where T is some predefined cutoff), or
    - We have at least n completed hypotheses (where n is pre-defined cutoff)

Readings

- The Curious Case of Neural Text Degeneration
- Learning to Write with Cooperative Discriminators
- Empirical Analysis of Beam Search Performance Degradation in Neural Sequence Models

Seq2seq Pre-trained Models

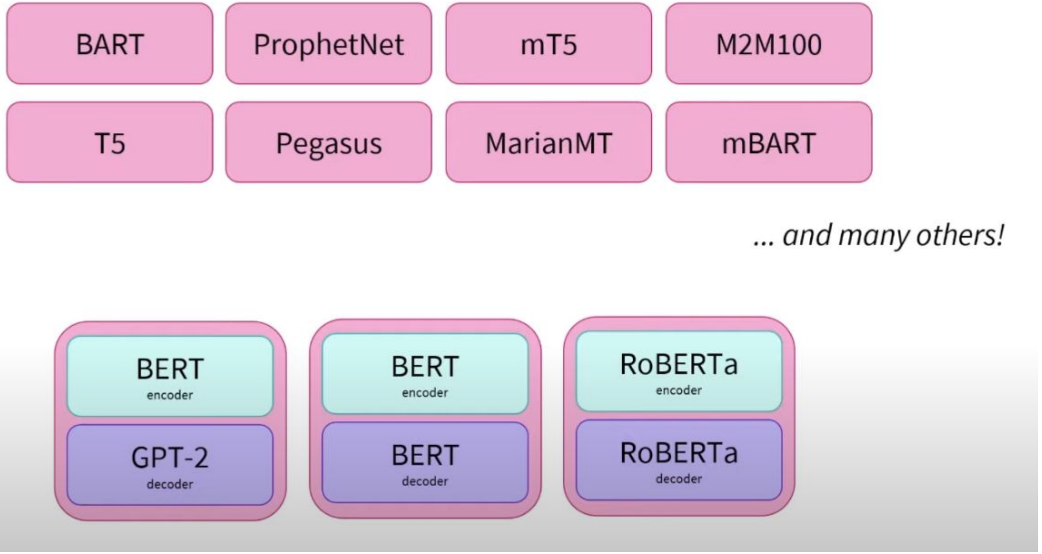
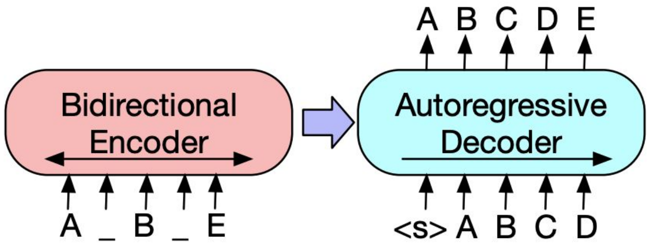
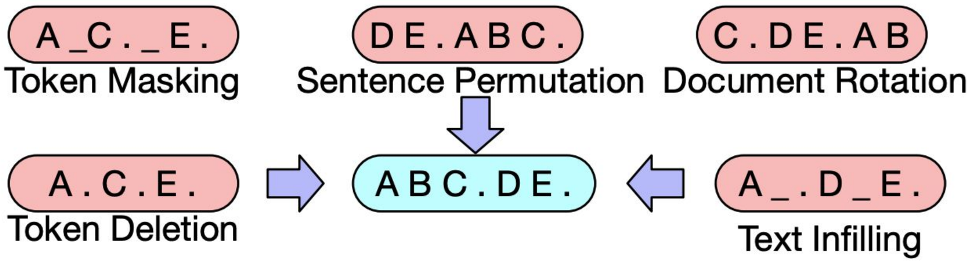


Figure credit: HuggingFace, Sylvain Gugger

BART, Lewis et al. (2019)



BART is a denoising autoencoder that maps a corrupted document to the original document it was derived from.



BART for Summarization, Lewis et al. (2019)

- Pre-train on the BART task: take random chunks of text, noise them according to the schemes described, and try to “decode” the clean text

- Fine-tune on a summarization dataset: a news article is the input and a summary of that article is the output (usually 1-3 sentences depending on the dataset)

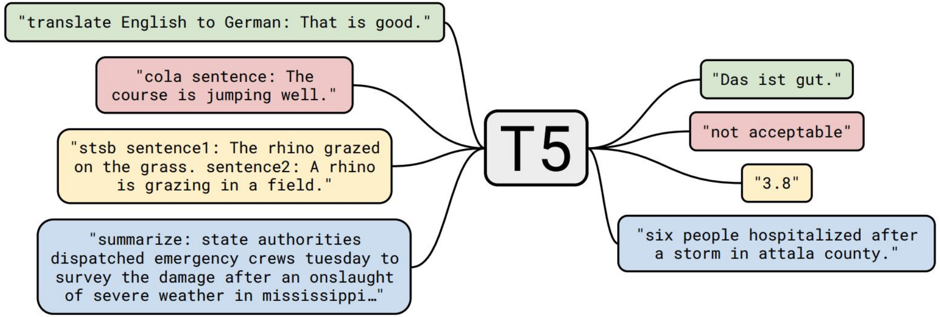
- Can achieve good results even with few summaries to fine-tune on, compared to basic seq2seq models which require 100k+ examples to do well

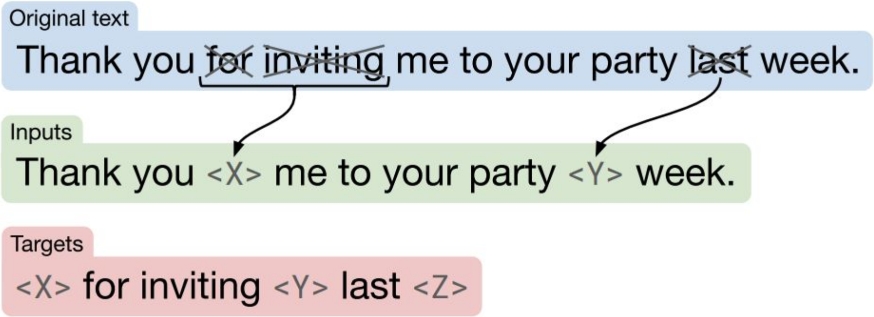
BART for Summarization: Outputs

Source Document (abbreviated)	BART Summary
The researchers examined three types of coral in reefs off the coast of Fiji ... The researchers found when fish were plentiful, they would eat algae and seaweed off the corals, which appeared to leave them more resistant to the bacterium Vibrio coralliilyticus, a bacterium associated with bleaching. The researchers suggested the algae, like warming temperatures, might render the corals' chemical defenses less effective, and the fish were protecting the coral by removing the algae.	Fisheries off the coast of Fiji are protecting coral reefs from the effects of global warming, according to a study in the journal Science.
Sacoolas, who has immunity as a diplomat's wife, was involved in a traffic collision ... Prime Minister Johnson was questioned about the case while speaking to the press at a hospital in Watford. He said, "I hope that Anne Sacoolas will come back ... if we can't resolve it then of course I will be raising it myself personally with the White House."	Boris Johnson has said he will raise the issue of US diplomat Anne Sacoolas' diplomatic immunity with the White House.

T5, Raffel et al. (2019)

- Pre-training: similar denoising scheme to BART (they were released within a week of each other in fall 2019)
- Input: text with gaps. Output: a series of phrases to fill those gaps





OpenAI GPT/GPT2, Radford et al. (2019)

- Very large language models using the Transformer architecture
- Straightforward left-to-right language model, trained on raw text
- GPT2: trained on 40GB of text
- By far the largest of these models trained when it came out in March 2019
- Because it's a language model, we can generate from it

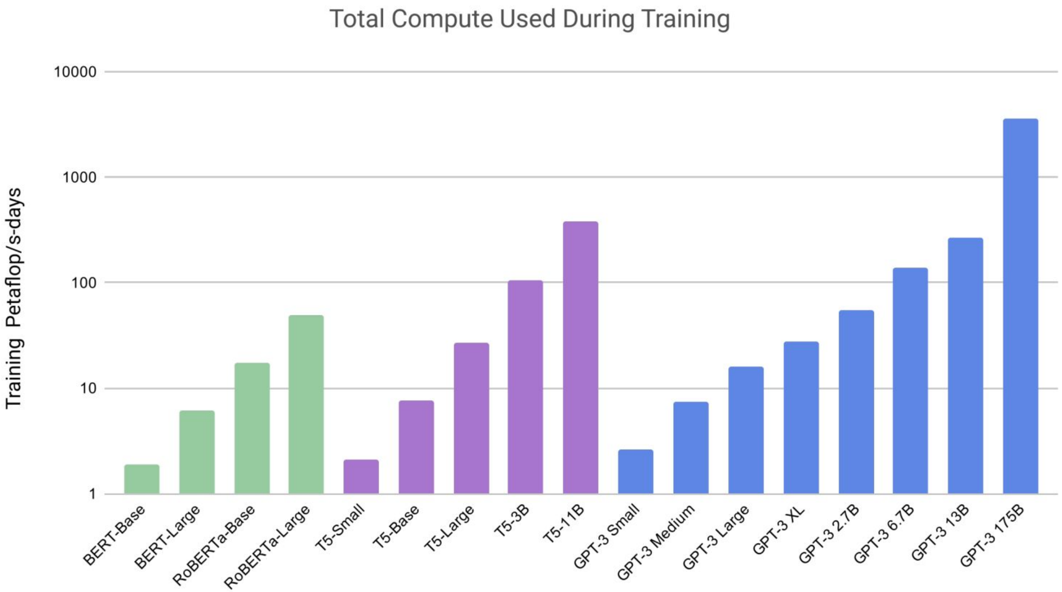
	Parameters	Layers	d_{model}
	117M	12	768
Approximate size of BERT	345M	24	1024
	762M	36	1280
GPT2	1542M	48	1600

Pre-Training Cost (with Google/AWS)

- BERT: Base $500, Large $7000
- GPT-2 (as reported in other work): $25,000
- This is for a single pre-training run…developing new pre-training techniques may require many runs
- *Fine-tuning* these models can typically be done with a single GPU (but may take 1-3 days for medium-sized datasets)

Pushing the Limits: GPT-3

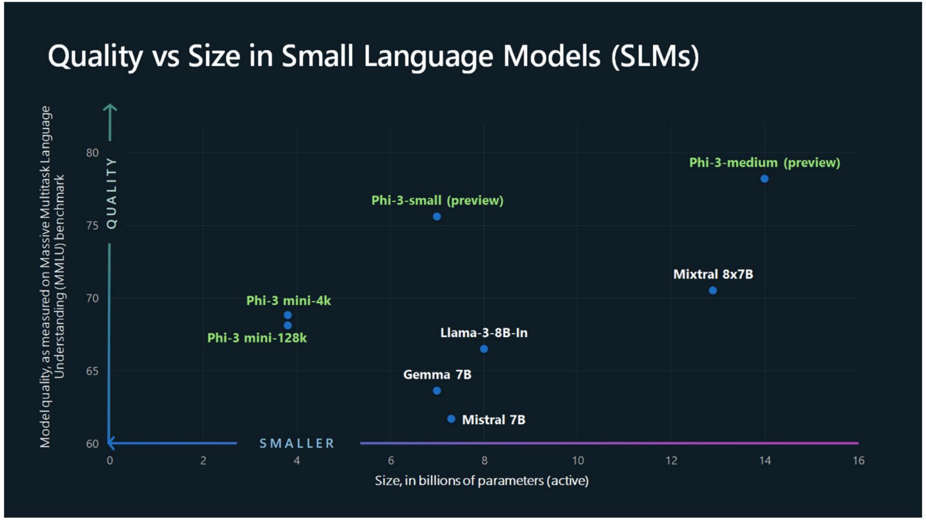
- 175B parameter model: 96 layers, 96 heads, 12k-dim vectors
- Trained on Microsoft Azure, estimated to cost roughly $10M



Brown et al. (2020)

LLM leaderboards

	Meta Llama 3 8B
MMLU 5-shot	68.4
GPQA 0-shot	34.2
HumanEval 0-shot	62.2
GSM-8K 8-shot, CoT	79.6
MATH 4-shot, CoT	30.0



Graphic illustrating how the quality of new Phi-3 models, as measured by performance on the Massive Multitask Language Understanding (MMLU) benchmark, compares to other models of similar size. (Image courtesy of Microsoft)

Images from Meta and Microsoft news

Language Models Through the Years

	A Neural Probabilistic Language Model	NLP (Almost) From Scratch	Word2Vec	ELMo	BERT	GPT2	GPT3	BLOOM	LLaMA
	Bengio et al 2003	Collobert et al 2011	Mikolov et al 2013	Peters et al 2018	Devlin et al 2018	Radford et al 2019	Brown et al 2020	Big Science 2022	Touvron et al 2023
PARAMS	1.2M	5M	300M*	93M	330M	1.5B	175B	175B	65B
TOKENS	14M	631M	1.6B	1.8B	3.3B	~14B	~300B	1.6T*	1.4T

DeepSeek-V3 (Dec. 2024):
671B params, 14.8T tokens

Slide credit: Luca Soldaini

https://api-docs.deepseek.com/news/news1226

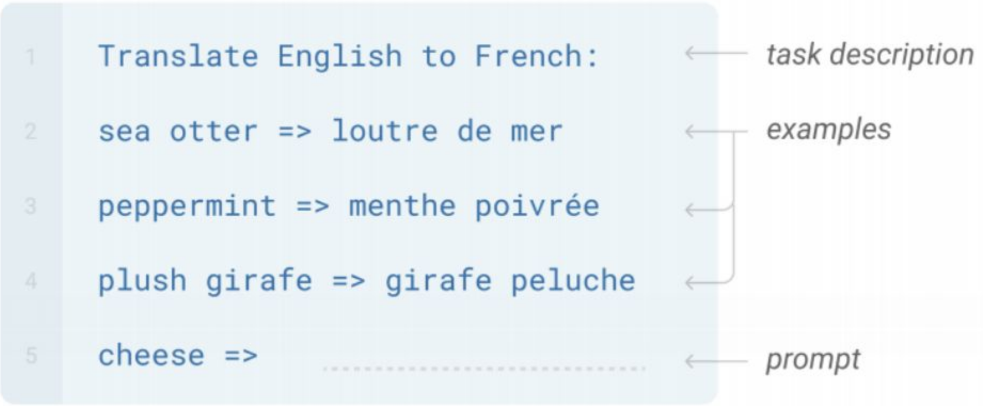
Pre-GPT-3: Fine-tuning

- Fine-tuning: this is the “normal way” of doing learning in models like GPT-2
- Requires computing the gradient and applying a parameter update on every example
- This is super expensive with 175B parameters

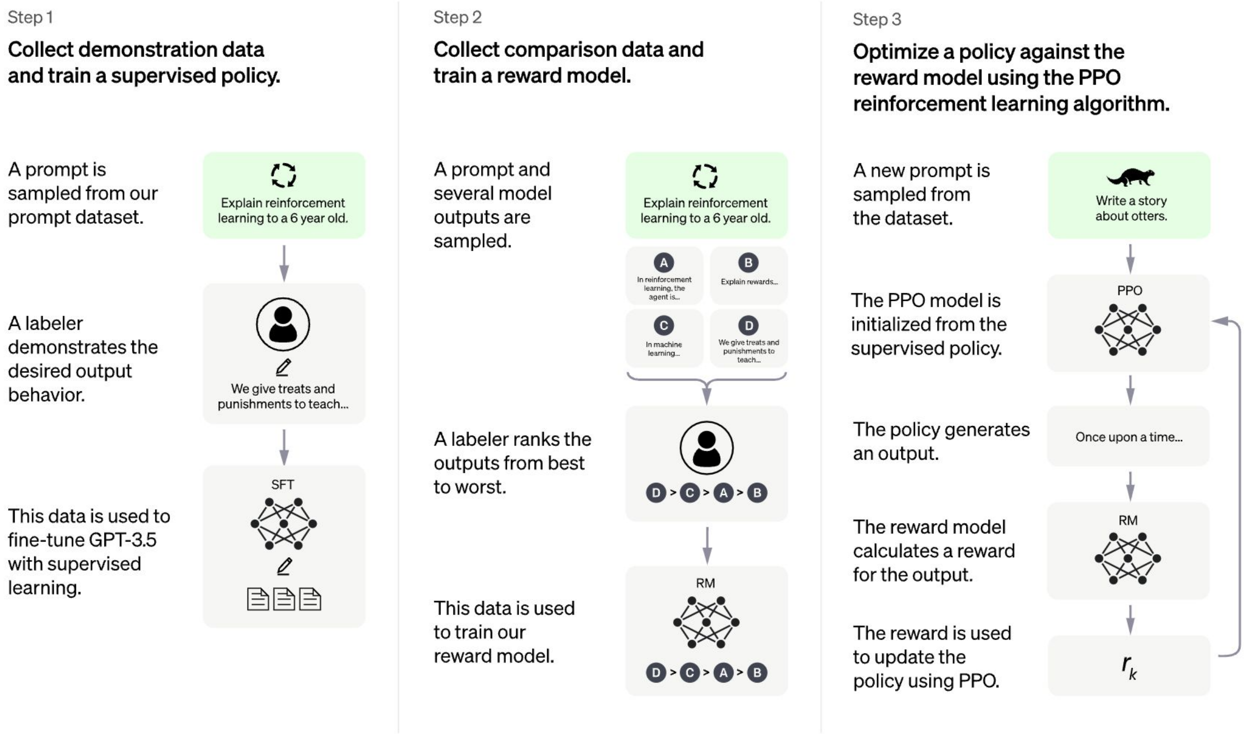
Brown et al. (2020)

GPT-3: Few-shot Learning

- GPT-3 proposes an alternative: in-context learning. Just uses the off-the shelf model, no gradient updates
- This procedure depends heavily on the examples you pick as well as the prompt ("Translate English to French")



Reinforcement Learning from Human Feedback (RLHF)

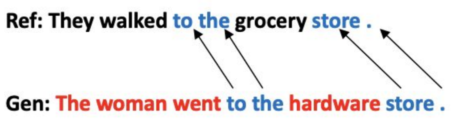


Ouyang et al. (2022)

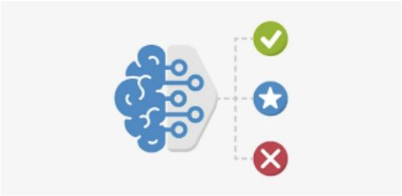
Readings

- PaLM: Scaling Language Modeling with Pathways
- LaMDA: Language Models for Dialog Applications
- LLaMA: Open and Efficient Foundation Language Models
- Language Models are Unsupervised Multitask Learners
- Language Models are Few-Shot Learners
- Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?
- Demystifying Prompts in Language Models via Perplexity Estimation
- Scaling Instruction-Finetuned Language Models
- Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing
- Training language models to follow instructions with human feedback
- Direct Preference Optimization: Your Language Model is Secretly a Reward Model
- KTO: Model Alignment as Prospect Theoretic Optimization

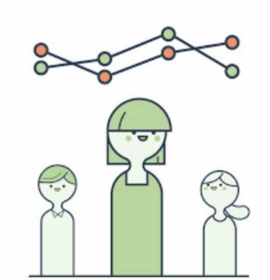
Types of evaluation methods for text generation



Content Overlap Metrics



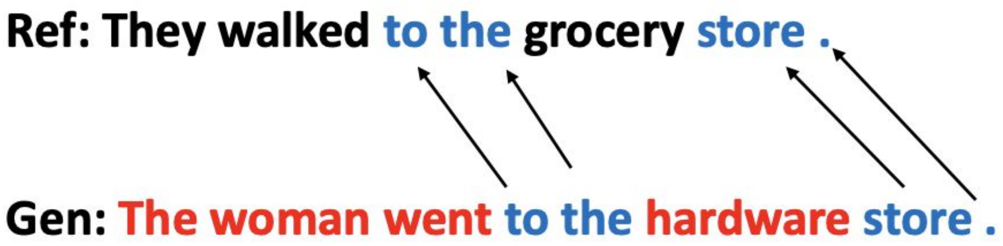
Model-based Metrics



Human Evaluations

Slide credit: Xiang Lisa Li, Stanford CS224N/Ling284

Content overlap metrics



- Compute a score that indicates the lexical similarity between generated and gold standard (human-written) text
- Fast and efficient and widely used
- *N-gram* overlap metrics (e.g., BLEU, ROUGE, METEOR, CIDEr, etc.)

BLEU (Bilingual Evaluation Understudy)

- Often used for machine translation
- Intelligibility or grammatical correctness are not taken into account
- Mathematically, the BLEU score is defined as:

$$\text{BLEU} = \underbrace{\min\Big(1, \exp\big(1 - \frac{\text{reference-length}}{\text{output-length}}\big)\Big)}_{\text{brevity penalty}} \underbrace{\Big(\prod_{i=1}^{4} precision_i\Big)^{1/4}}_{\text{n-gram overlap}}$$

with

$$precision_i = \frac{\sum_{\text{snt}\in\text{Cand-Corpus}} \sum_{i\in\text{snt}} \min(m^i_{cand}, m^i_{ref})}{w^i_t = \sum_{\text{snt'}\in\text{Cand-Corpus}} \sum_{i'\in\text{snt'}} m^{i'}_{cand}}$$

where

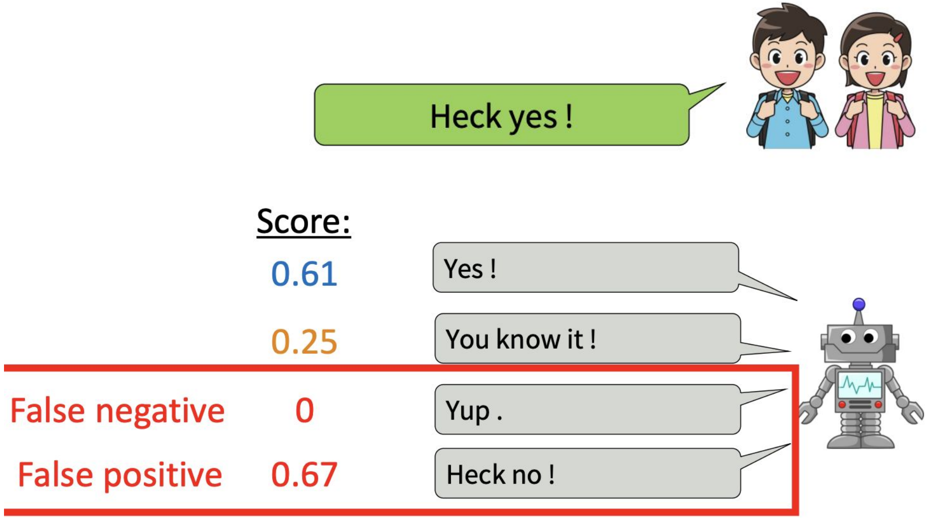
- m^i_{cand} is the count of i-gram in candidate matching the reference translation
- m^i_{ref} is the count of i-gram in the reference translation
- w^i_t is the total number of i-grams in candidate translation

Notations from here

A simple failure case

n-gram overlap metrics have no concept of semantic relatedness!

Are you enjoying this lecture?



N-gram overlap metrics

- They're not always ideal even for machine translation
- They get progressively much worse for tasks that are more open-ended than machine translation
  - Worse for summarization, as longer output texts are harder to measure
  - Much worse for dialogue, which is more open-ended that summarization
  - Much, much worse for story generation, which is also open-ended, but whose sequence length can make it seem you're getting decent scores!

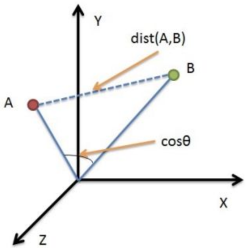
Model-based metrics to capture more semantics

- Use learned representations of words and sentences to compute semantic similarity between generated and reference texts

- No more n-gram bottleneck because text units are represented as embeddings!

- The embeddings are pretrained, distance metrics used to measure the similarity can be fixed

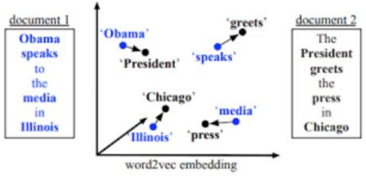
Model-based metrics: Word distance functions



Vector Similarity

Embedding based similarity for semantic distance between text.

- **Embedding Average** (Liu et al., 2016)
- **Vector Extrema** (Liu et al., 2016)
- **MEANT** (Lo, 2017)
- **YISI** (Lo, 2019)

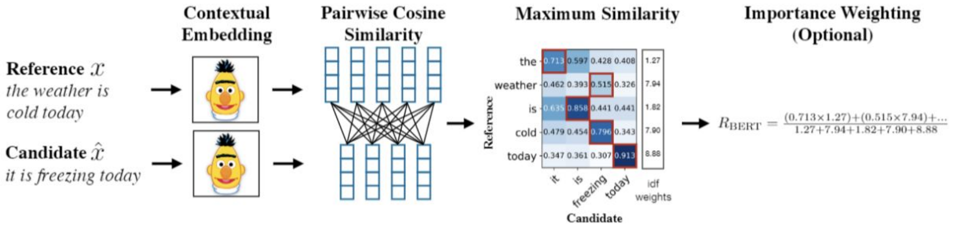


Word Mover's Distance

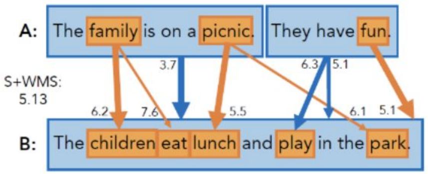
Measures the distance between two sequences (e.g., sentences, paragraphs, etc.), using word embedding similarity matching.
(Kusner et.al., 2015; Zhao et al., 2019)

BERTSCORE

Uses pre-trained contextual embeddings from BERT and matches words in candidate and reference sentences by cosine similarity.
(Zhang et.al. 2020)



Model-based metrics: Beyond word matching



Sentence Movers Similarity :

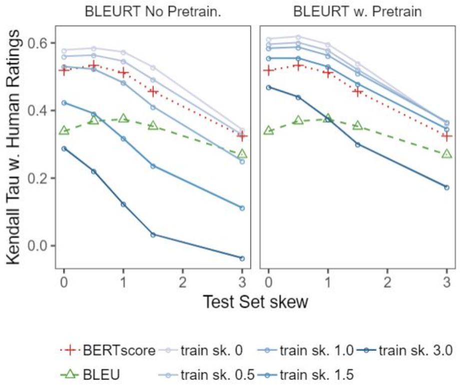
Based on Word Movers Distance to evaluate text in a continuous space using sentence embeddings from recurrent neural network representations.

(Clark et.al., 2019)

BLEURT:

A regression model based on BERT returns a score that indicates to what extent the candidate text is grammatical and conveys the meaning of the reference text.

(Sellam et.al. 2020)



Evaluation with LLMs

```
Score the following translation from {source_lang} to {target_lang} with respect
to the human reference on a continuous scale from 0 to 100, where score of zero means
"no meaning preserved" and score of one hundred means "perfect meaning and grammar".

{source_lang} source: "{source_seg}"
{target_lang} human reference: {reference_seg}
{target_lang} translation: "{target_seg}"
Score:
```

Figure 1: The best-performing prompt based on Direct Assessment expecting a score between 0–100. Template **portions in bold face** are used only when a human reference translation is available.

- Outperforms many learned MT metrics (Transformers trained over (source, target, reference) triples to reproduce human judgments of quality)
- Only works with GPT 3.5 and larger models

Kocmi et al. (2023)

Human evaluations

- Automatic metrics fall short of matching human decisions

- Human evaluation is most important form of evaluation for text generation systems.

- Gold standard in developing new automatic metrics
  - New automated metrics must correlate well with human evaluations!

Human evaluations

- Ask humans to evaluate the quality of generated text
- Overall or along some specific dimension:
    - fluency
    - coherence / consistency
    - factuality and correctness
    - commonsense
    - style / formality
    - grammaticality
    - typicality
    - redundancy

Readings

- BLEURT: Learning Robust Metrics for Text Generation
- BERTScore: Evaluating Text Generation with BERT
- Large Language Models Are State-of-the-Art Evaluators of Translation Quality
- Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation
- Evaluation of Text Generation: A Survey
- The Authenticity Gap in Human Evaluation
- Thinking about High-Quality Human Data

Sequence-to-sequence is versatile!

- The general notion here is an encoder-decoder model
  - One neural network takes input and produces a neural representation
  - Another network produces output based on that neural representation
  - If the input and output are sequences, we call it a seq2seq model

- Many NLP tasks can be phrased as sequence-to-sequence:
  - Machine translation (source language sentence → target language sentence)
  - Summarization (long text → short text)
  - Dialogue (previous utterances → next utterance)
  - Parsing (input text → output parse as sequence)
  - Code generation (natural language → Python code)

Machine Translation

Machine Translation (MT) is the task of translating a sentence x from one language (the source language) to a sentence y in another language (the target language).

x: *L'homme est né libre, et partout il est dans les fers*

$\downarrow$

y: *Man is born free, but everywhere he is in chains*

– Rousseau

1950s-1980s: Rule-Based Machine Translation

Early on, MT systems incorporated intricate linguistic knowledge about the lexicon and grammar of the pair of languages being translated.

Essentially: translation lexicon + grammar rules

Georgetown–IBM Experiment (1954)

"On January 7, 1954 the first public demonstration of a Russian-English machine translation system occurred in New York—a collaboration between IBM and Georgetown University. Brief statements about politics, law, mathematics, chemistry, metallurgy, communications and military affairs were submitted in Russian by Léon Dostert and linguists of the Georgetown University Institute of Languages and Linguistics, and within a within a few seconds a computer translated the sentences into English. This project, which began in 1951, was also probably the first non-numerical application of a digital computer. Programming and the demonstration was done on an IBM 701, the first stored program digital computer that IBM put into production."

Fig. 2: Hurd, Dostert and Watson at the demonstration

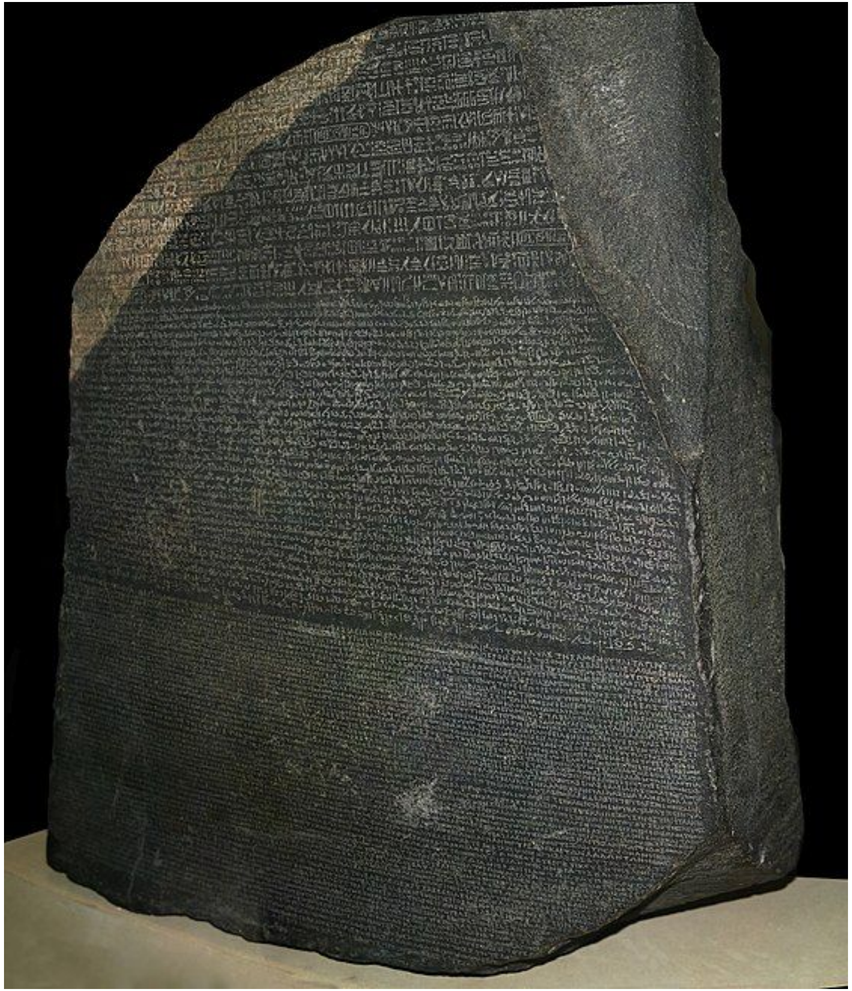
https://www.historyofinformation.com/detail.php?id=666

(see also:
https://aclanthology.org/www.mt-archive.info/Garvin-1967.pdf,
https://aclanthology.org/2004.amta-papers.12/)

Parallel Corpora

A parallel corpus is a corpus in one language along with translations into one or more other languages.

E.g.: Rosetta Stone (right), as well as digital corpora that started becoming available around the 1990s, e.g. Canadian Hansards, Europarl (2005)



1990s-2010s: Statistical Machine Translation

- Core idea: Learn a probabilistic model from data
  - Suppose we're translating French → English.
  - We want to find best English sentence y, given French sentence x

$$\text{argmax}_y P(y|x)$$

- Use Bayes Rule to break this down into two components to be learned separately:

$$= \text{argmax}_y P(x|y) P(y)$$

Translation Model

Models how words and phrases should be translated (*fidelity*). Learned from parallel data.

Language Model

Models how to write good English (*fluency*). Learned from monolingual data.

1990s-2010s: Statistical Machine Translation

- Use Bayes Rule to break this down into two components to be learned separately:

$$= \text{argmax}_y \underbrace{P(x|y)}\underbrace{P(y)}$$

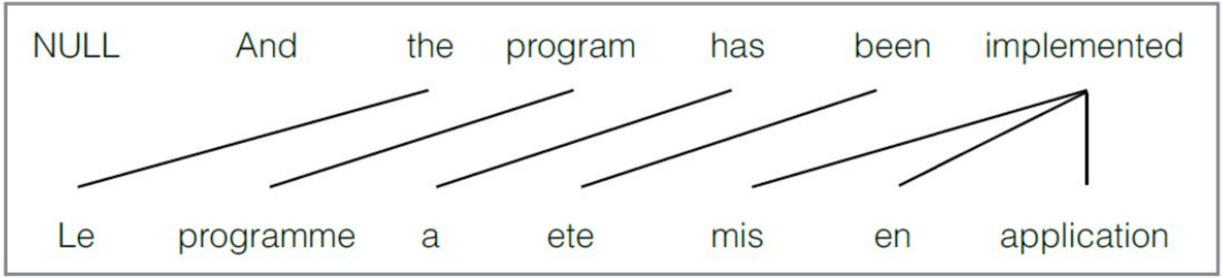
Translation Model

Models how words and phrases should be translated (*fidelity*). Learned from parallel data.

Language Model

Models how to write good English (*fluency*). Learned from monolingual data.

- Translation Model: trained on parallel corpora + **word alignments** from an unsupervised model:



1990s-2010s: Statistical Machine Translation

- SMT was a huge research field
- The best systems were extremely complex
  - Hundreds of important details
- Systems had many separately-designed subcomponents
  - Lots of feature engineering
  - Need to design features to capture particular language phenomena
- Required compiling and maintaining extra resources
  - Like tables of equivalent phrases
- Lots of human effort to maintain
  - Repeated effort for each language pair!

1519年600名西班牙人在墨西哥登陆，去征服几百万人口的阿兹特克帝国，初次交锋他们损兵三分之二。

In 1519, six hundred Spaniards landed in Mexico to conquer the Aztec Empire with a population of a few million. They lost two thirds of their soldiers in the first clash.

translate.google.com (2009): 1519 600 Spaniards landed in Mexico, millions of people to conquer the Aztec empire, the first two-thirds of soldiers against their loss.

translate.google.com (2013): 1519 600 Spaniards landed in Mexico to conquer the Aztec empire, hundreds of millions of people, the initial confrontation loss of soldiers two-thirds.

translate.google.com (2015): 1519 600 Spaniards landed in Mexico, millions of people to conquer the Aztec empire, the first two-thirds of the loss of soldiers they clash.

What is Neural Machine Translation?

- Neural Machine Translation (NMT) is a way to do Machine Translation with a single end-to-end neural network
- The neural network architecture is called a sequence-to-sequence model (aka seq2seq) and it involves two RNNs
- The sequence-to-sequence model is an example of a Conditional Language Model
  - Language Model because the decoder is predicting the next word of the target sentence y
  - Conditional because its predictions are also conditioned on the source sentence x
- NMT directly calculates:

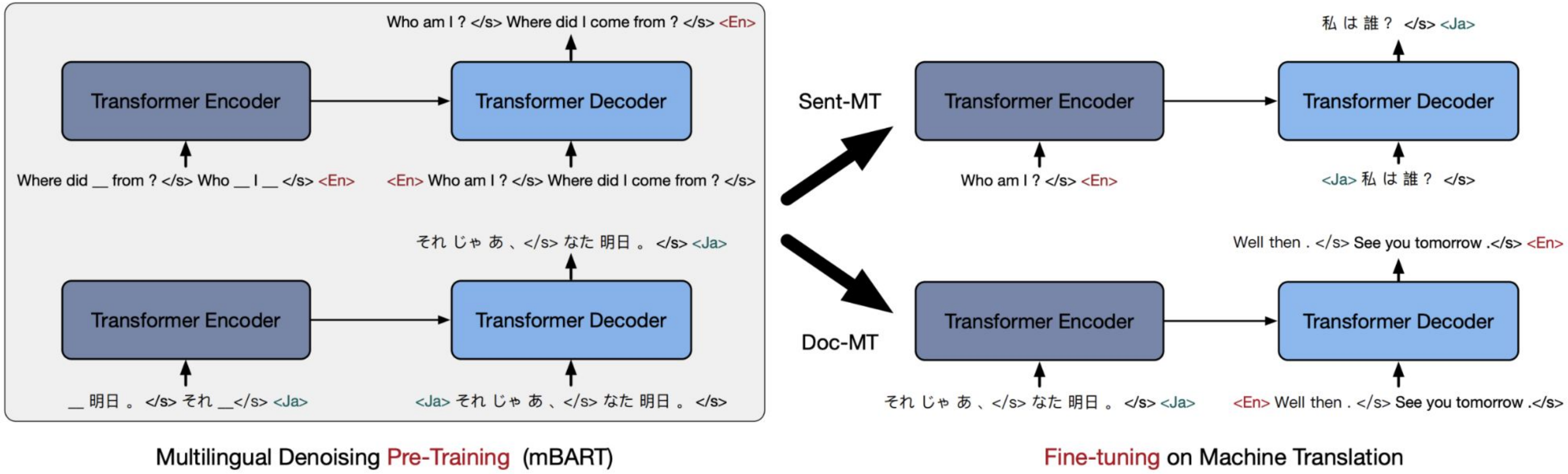
$$P(y|x) = P(y_1|x)\, P(y_2|y_1, x)\, P(y_3|y_1, y_2, x) \ldots \underbrace{P(y_T|y_1, \ldots, y_{T-1}, x)}_{\text{Probability of next target word, given target words so far and source sentence } x}$$

Frontiers in MT: Transformers, Vaswani et al. (2017)

Model	BLEU	
	EN-DE	EN-FR
ByteNet [18]	23.75	
Deep-Att + PosUnk [39]		39.2
GNMT + RL [38]	24.6	39.92
ConvS2S [9]	25.16	40.46
MoE [32]	26.03	40.56
Deep-Att + PosUnk Ensemble [39]		40.4
GNMT + RL Ensemble [38]	26.30	41.16
ConvS2S Ensemble [9]	26.36	**41.29**
Transformer (base model)	27.3	38.1
Transformer (big)	**28.4**	**41.8**

Big = 6 layers, 1000 dim for each token, 16 heads
Base = 6 layers + other params halved

Frontiers in MT: Multilingual Models, Yinhan Liu et al. (2020)



Frontiers in MT: Multilingual Models, Yinhan Liu et al. (2020)

Languages	En-Gu		En-Kk		En-Vi		En-Tr		En-Ja		En-Ko	
Data Source	WMT19		WMT19		IWSLT15		WMT17		IWSLT17		IWSLT17	
Size	10K		91K		133K		207K		223K		230K	
Direction	←	→	←	→	←	→	←	→	←	→	←	→
Random	0.0	0.0	0.8	0.2	23.6	24.8	12.2	9.5	10.4	12.3	15.3	16.3
mBART25	**0.3**	**0.1**	**7.4**	**2.5**	**36.1**	**35.4**	**22.5**	**17.8**	**19.1**	**19.4**	**24.6**	**22.6**

Languages	En-Nl		En-Ar		En-It		En-My		En-Ne		En-Ro	
Data Source	IWSLT17		IWSLT17		IWSLT17		WAT19		FLoRes		WMT16	
Size	237K		250K		250K		259K		564K		608K	
Direction	←	→	←	→	←	→	←	→	←	→	←	→
Random	34.6	29.3	27.5	16.9	31.7	28.0	23.3	34.9	7.6	4.3	34.0	34.3
mBART25	**43.3**	**34.8**	**37.6**	**21.6**	**39.8**	**34.0**	**28.3**	**36.9**	**14.5**	**7.4**	**37.8**	**37.7**

Languages	En-Si		En-Hi		En-Et		En-Lt		En-Fi		En-Lv	
Data Source	FLoRes		ITTB		WMT18		WMT19		WMT17		WMT17	
Size	647K		1.56M		1.94M		2.11M		2.66M		4.50M	
Direction	←	→	←	→	←	→	←	→	←	→	←	→
Random	7.2	1.2	10.9	14.2	22.6	17.9	18.1	12.1	21.8	20.2	15.6	12.9
mBART25	**13.7**	**3.3**	**23.5**	**20.8**	**27.8**	**21.4**	**22.4**	**15.3**	**28.5**	**22.4**	**19.3**	**15.9**

Table 2: **Low/Medium Resource Machine Translation** Pre-training consistently improves over a randomly initialized baseline, with particularly large gains on low resource language pairs (e.g. Vi-En).

Frontiers in MT: ChatGPT

Table 3: Comparison of different prompts for ChatGPT to perform Chinese-to-English (Zh⇒En) translation.

System	**BLEU↑**	**ChrF++↑**	**TER↓**
Google	31.66	57.09	56.21
DeepL	31.22	56.74	57.84
Tencent	29.69	56.24	57.16
ChatGPT w/ TP1	23.25	53.07	66.03
ChatGPT w/ TP2	24.54	53.05	63.79
ChatGPT w/ TP3	**24.73**	**53.71**	**62.84**

- Works okay for Chinese-English, but less good at generating into low-resource languages (English -> Romanian doesn't work well)

Table 5: Performance of ChatGPT with pivot prompting. New results are obtained from the updated ChatGPT version on 2023.01.31. LR: length ratio.

System	**De⇒Zh**		**Ro⇒Zh**	
	BLEU	LR	BLEU	LR
Google	38.71	0.94	39.05	0.95
DeepL	40.46	0.98	38.95	0.99
ChatGPT (Direct)	34.46	0.97	30.84	0.91
ChatGPT (Direct$_{new}$)	30.76	0.92	27.51	0.93
ChatGPT (Pivot$_{new}$)	34.68	0.95	34.19	0.98

- Better with "pivoting"

"Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine" Jia et al. (2023)

Readings

- Eisenstein 18.1-18.2, 18.4
- Michael Collins IBM Models 1+2
- History of MT
- Revisiting Low-Resource Neural Machine Translation: A Case Study
- In Neural Machine Translation, What Does Transfer Learning Transfer?
- Multilingual Denoising Pre-training for Neural Machine Translation
- Neural Machine Translation by Jointly Learning to Align and Translate (original seq2seq+attention paper)
- Massive Exploration of Neural Machine Translation Architectures
- Achieving Open Vocabulary Neural Machine Translation with Hybrid Word-Character Models
- Revisiting Character-Based Neural Machine Translation with Capacity and Compression

SQuAD, Rajpurkar et al. (2016)

- Single-document, single-sentence question-answering task where the answer is always a substring of the passage
- Predict start and end indices of the answer in the passage

One of the most famous people born in Warsaw was Maria Skłodowska-Curie, who achieved international recognition for her research on radioactivity and was the first female recipient of the Nobel Prize. Famous musicians include Władysław Szpilman and Frédéric Chopin. Though Chopin was born in the village of Żelazowa Wola, about 60 km (37 mi) from Warsaw, he moved to the city with his family when he was seven months old. Casimir Pulaski, a Polish general and hero of the American Revolutionary War, was born here in 1745.

What was Maria Curie the first female recipient of?
Ground Truth Answers: Nobel Prize | Nobel Prize | Nobel Prize

What year was Casimir Pulaski born in Warsaw?
Ground Truth Answers: 1745 | 1745 | 1745

Who was one of the most famous people born in Warsaw?
Ground Truth Answers: Maria Skłodowska-Curie | Maria Skłodowska-Curie | Maria Skłodowska-Curie

Multi-Hop Question Answering

- Very few SQuAD questions require actually combining multiple pieces of information— this is an important capability QA systems should have

- Several datasets test multi-hop reasoning: ability to answer questions that draw on several sentences or several documents to answer

Welbl et al. (2018), Yang et al. (2018)

WikiHop

- Annotators shown Wikipedia and asked to pose a simple question linking two entities that require a third (bridging) entity to associate
- A model shouldn't be able to answer these without doing some reasoning about the intermediate entity

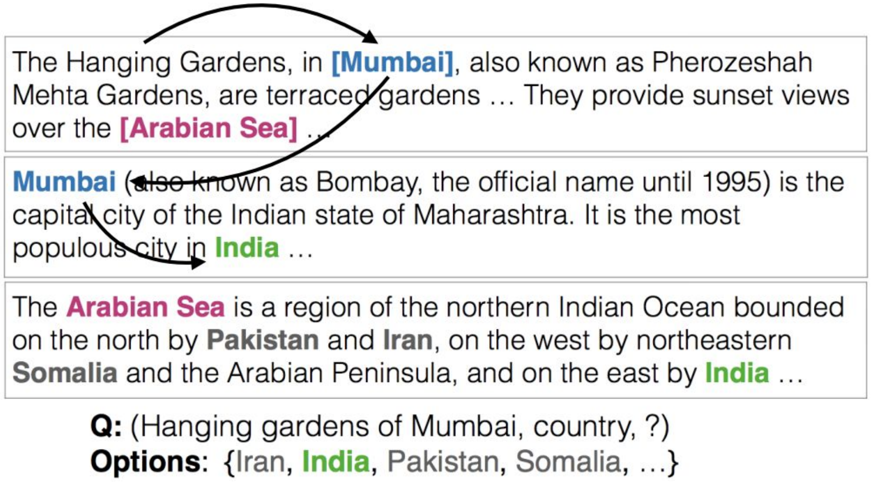


Figure from Welbl et al. (2018)

Successes of T5

- How can we handle a task like QA by framing it as a seq2seq problem?

Dataset	SQuAD 1.1
Input	`At what speed did the turbine operate? \n (Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...`
Output	`16,000 rpm`

- Format: Question \n Passage → Answer

Encoder Decoder

Raffel et al. (2019)

UnifiedQA

AB	**Dataset**	NarrativeQA
	Input	`What does a drink from narcissus's spring cause the drinker to do? \n Mercury has awakened Echo, who weeps for Narcissus, and states that a drink from Narcissus's spring causes the drinkers to ``Grow dotingly enamored of themselves.'' ...`
	Output	`fall in love with themselves`

Abstractive question, requires generating free-form answer

- Past work: different architectures for every QA formulation. (Span selection, answer generation, multiple choice, ...)
- Now: one 11B parameter T5 model

Khashabi et al. (2020)

UnifiedQA

Multiple choice	MC	**Dataset**	MCTest
		Input	`Who was Billy? \n (A) The skinny kid (B) A teacher (C) A little kid (D) The big kid \n Billy was like a king on the school yard. A king without a queen. He was the biggest kid in our grade, so he made all the rules during recess. ...`
		Output	`The big kid`
Yes/no	YN	**Dataset**	BoolQ
		Input	`Was America the first country to have a president? \n (President) The first usage of the word president to denote the highest official in a government was during the Commonwealth of England ...`
		Output	`no`

- Past work: different architectures for every QA formulation. (Span selection, answer generation, multiple choice, …)
- Now: one 11B parameter T5 model

Khashabi et al. (2020)

Long-form QA, ELI5 (reddit "Explain like I'm 5")

Question: How do Jellyfish function without brains or nervous systems?

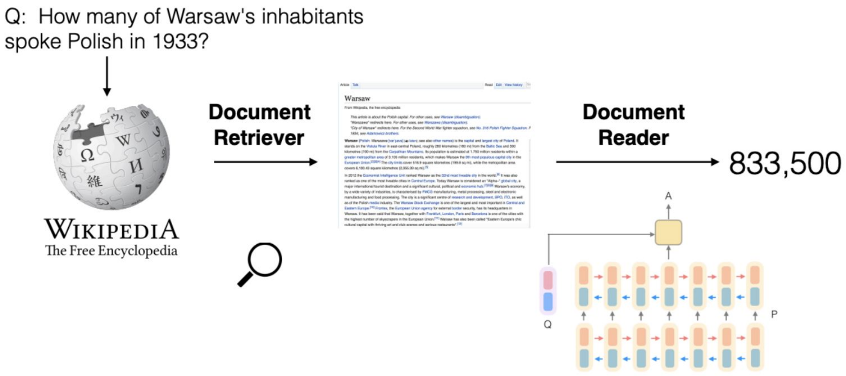
Supporting Documents: The box jellyfish nervous system is divided into three functional parts namely; rhopalia, conducting nerve ring, and motor nerve net. [...] J
have
sens
neur
jelly
[...]

Answer: Jellyfish may not have a brain, but they have a rough nervous system and innate behaviours. However, they are very simple creatures. They're invertebrate: creatures without a backbone. Most jellyfish have really short life spans. [...] As their name implies, they are largely composed of basically jelly inside a thin membrane. They're over 95% water.

Fan et al. (2019)

Open-domain QA

- Real QA systems should be able to handle more than just a paragraph of context—theoretically should work over the whole web?
  - E.g. NaturalQuestions dataset
- This also introduces more complex distractors (bad answers) and should require stronger QA systems
- QA pipeline: given a question:
  - Retrieve some documents with an IR system
  - Zero in on the answer in those documents with a QA model



Chen et al. (2017)

WebGPT

Question	Why did we decide that certain words were "bad" and shouldn't be used in social settings?
Answer	We have all been taught that there are certain words that should not be used in public, at least not without good reason. But why do some words fall into this category while others do not? It turns out that there is not a single answer to this question, as each word has a different history and carries different meanings in different cultures and families [2]. However, there are some general categories that "bad" words tend to fall into. "Bad" words generally relate to parts of life that we don't like talking about in public, like bathroom functions, or negative ways of talking about people's religion, skin color, ability, or gender [3]. Often, people say bad words when they are experiencing strong emotions, and those words are linked to our emotions [3].
References (titles only)	[1, 2, 3] Why Are Some Words 'Bad'? | Vermont Public Radio (www.vpr.org) [4] On Words: 'Bad' Words and Why We Should Study Them | UVA Today (news.virginia.edu) [5] The Science of Curse Words: Why The &@$! Do We Swear? (www.babbel.com)

- GPT model equipped with a search engine, then summarizes the answers

Nakano et al. (2021)

Readings

- Reading Wikipedia to Answer Open-Domain Questions
- Latent Retrieval for Weakly Supervised Open Domain Question Answering
- REALM: Retrieval-Augmented Language Model Pre-Training
- DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs
- WebGPT: Browser-assisted question-answering with human feedback
- QuAC : Question Answering in Context

Bias and Toxicity

- "Toxic degeneration": systems that generate toxic stuff

https://toxicdegeneration.allenai.org/

- System trained on a big chunk of the Internet: conditioning on "SJW", "black" gives the system a chance of recalling bad stuff from its training data

Stochastic Parrots 🦜 (about LMs generally)

Paper (that included authors at Google who were subsequently fired) about dangers of large language models

- **Claim 1:** environmental cost is disproportionately born by marginalized populations, who aren't even well-served by these tools

- **Claim 2:** massive data is fundamentally challenging to audit, contains data that is biased and is only a snapshot of a single point in time

- **Claim 3:** these models are not grounded in meaning— when they generate an answer to a question, it is merely by memorizing co-occurrence between symbols

Bender, Gebru, McMillan-Major, Shmitchell (2021)

Stochastic Parrots 🦜

Question: What is the name of the Russian mercenary group?
Answer: Wagner group.

Question: Where is the Wagner group?
Answer: In Syria.

Question: Who is the leader of the Wagner group?
Answer: A Russian citizen, Dmitry Utkin. He was born in 1971 and lives in Moscow. He has been a member of the Russian parliament since 2016. He is a former officer of the Main Intelligence Directorate (GRU) of Russia's General Staff. He was also a commander of the special forces unit "Vostok" (East). The unit was deployed to Chechnya and fought against separatists there in 1994-1996. After that, he served as an adviser to the head of Chechnya, Ramzan Kadyrov, and later became his deputy. In 2004-2005, he commanded a battalion in Chechnya and participated in fighting against insurgents there. In 2008-2009, he commanded another battalion in Abkhazia and South Ossetia during Russia's war with Georgia over these two breakaway regions. Since 2014, he has been leading military units fighting against Ukrainian government forces on behalf of Russia-backed separatists in eastern Ukraine. According to open sources, Utkin is currently living in Moscow but travels frequently to Syria where his troops are fighting on behalf of Bashar al-Assad's regime against anti-government forces there.

- We are likely to assume the model is producing factual information and presenting it in a coherent way, but this is our interpretation we project on the model
- Risks: medical diagnosis ("What do I have if I have X, Y, and Z symptoms?") could seem possible but cause serious harm

Bender, Gebru, McMillan-Major, Shmitchell (2021)

Privacy and Safety

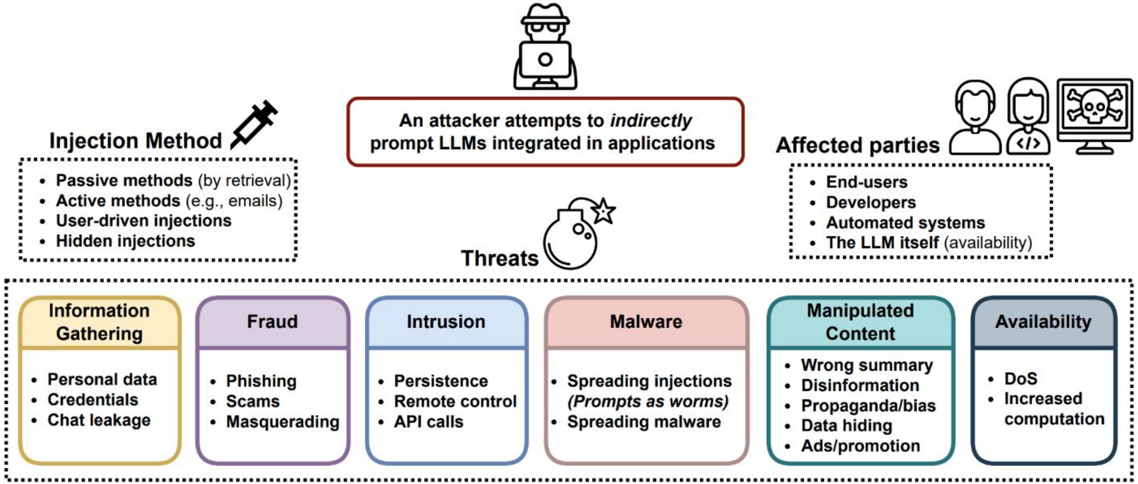


Fig. 1. An overview of threats to LLM-based applications. (Image source: Greshake et al. 2023)

Example from: Scalable Extraction of Training Data from (Production) Language Models, Nasr et al.

Readings

- Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned
- Scalable Extraction of Training Data from (Production) Language Models
- Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection
- Exploiting Programmatic Behavior of LLMs: Dual-Use Through Standard Security Attacks

For more on these topics

Special topics courses offered periodically through Linguistics Department (some also CS):

- Machine Translation
- Multilingual NLP
- Dialogue Systems
- Social Factors in Computational Linguistics & AI