CS 335: Introduction to Large Language Models

Capabilities
Week 2

Dr Abdul Samad

Lecture Outline

- Adaption
- Language Modelling
- Question Answering
- BLEU Score
- Translation
- Arithmetic
- News article generation
- Novel tasks

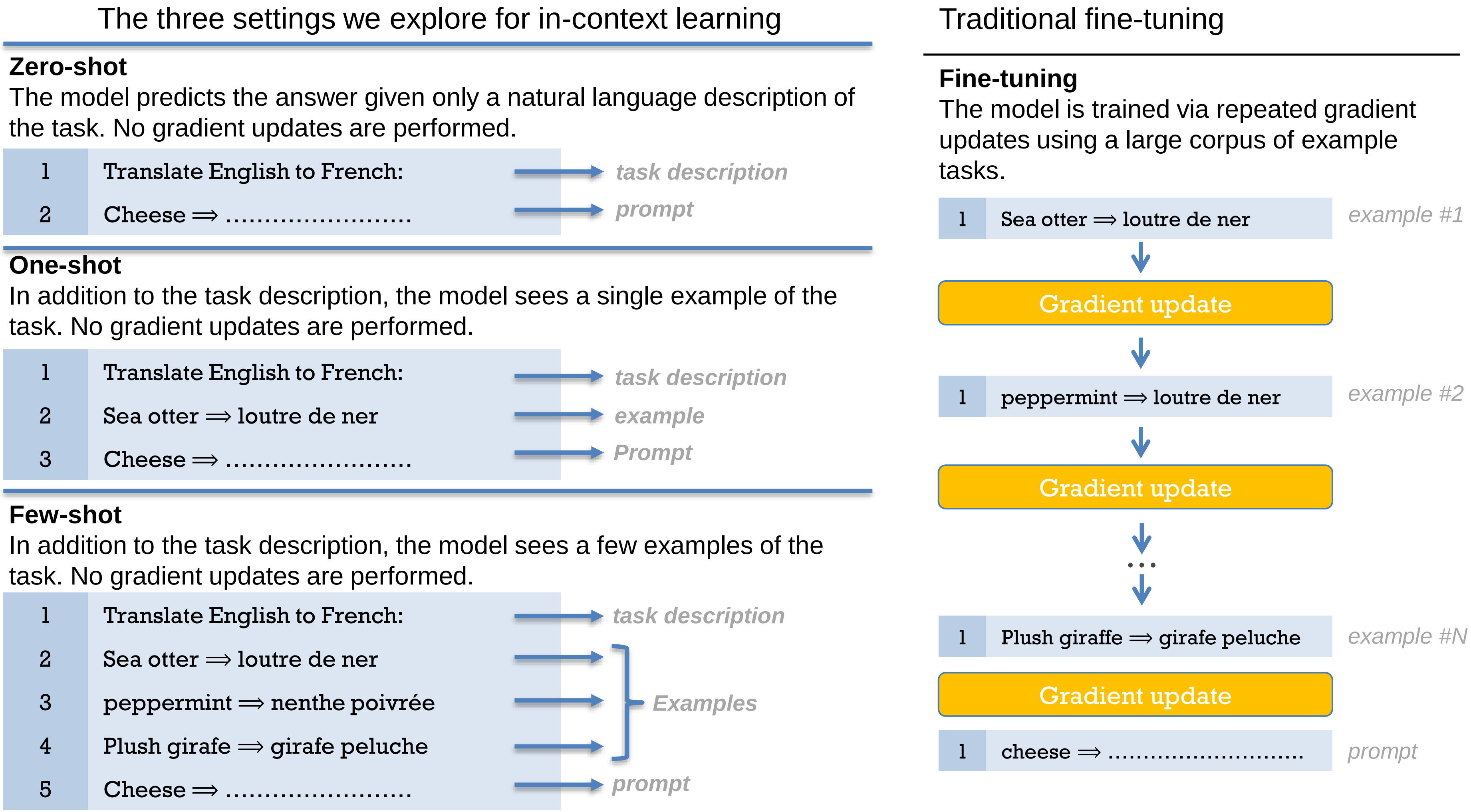
Adaptation

- We use the term **adaptation** to refer to the process of taking a language model and turning it into a task model, given:
  - a natural language **description** of the task, and
  - a set of **training instances** (input-output pairs).

Adaptation

There are two primary ways to perform adaptation:

1. **Training** (standard supervised learning): train a new model that maps inputs to outputs, either by
   - creating a new model that uses the language model as features (probing), or
   - starting with the language model and updating it based on the training instances (fine-tuning), or
   - something in between (lightweight fine-tuning).
2. **Prompting** (in-context learning): Construct a prompt (a string based on the description and training instances) or a set of prompts, feed those into a language model to obtain completions.
   - Zero-shot learning: number of training examples is 0
   - One-shot learning: number of training examples is 1
   - Few-shot learning: number of training examples is few



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 ner → *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 ner

3 peppermint ⇒ nenthe poivrée

4 Plush girafe ⇒ girafe peluche

Examples

5 Cheese ⇒ .... → *prompt*

Traditional fine-tuning

Fine-tuning

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

1 Sea otter ⇒ loutre de ner *example #1*

Gradient update

1 peppermint ⇒ loutre de ner *example #2*

Gradient update

...

1 Plush giraffe ⇒ girafe peluche *example #N*

Gradient update

1 cheese ⇒ .... *prompt*

Adaptation of GPT-3

- Limitation of prompting is that we can only leverage a only small number of training instances (as many as can fit into a prompt). This is due to a limitation of Transformers, where the prompt and the completion must fit into 2048 tokens.

Adaptation

- The GPT-3 paper evaluated GPT-3 on a large set of tasks. We will consider a subset of these, and for each task, discuss the following:
  - Definition: What is the task and its motivation?
  - Adaptation: How do we reduce the task to language modeling (via prompting)?
  - Results: What are the quantitative numbers compared to task-specific state-of-the-art models?
- Size and number of examples matters. By default, the results will based on
  - the full GPT-3 model (davinci), which has 175 billion parameters
  - using in-context learning with as many training instances as you can stuff into the prompt.

Language Modeling

The most natural starting point for thinking about what a language model can do is to ask if it can do the thing that language models are supposed to do: model language.

Recall that a language model p is a probability distribution over sequences of tokens. Suppose we take a corpus of text $x_{1:L}$, for example:

the mouse ate the cheese

We can ask: what is the probability the language model assigns to it?

$$p(\text{the mouse ate the cheese})$$

Recall that we can break down the joint probability into the product of the conditional probabilities for each token by the chain rule:

$$p(x_{1:L}) = \prod_{i=1}^{L} p(x_i \mid x_{1:i-1})$$

Language Modeling

Perplexity

$$\mathrm{perplexity}_p(x_{1:L}) = \exp\left(\frac{1}{L}\sum_{i=1}^{L}\log\frac{1}{p(x_i \mid x_{1:i-1})}\right)$$

Perplexity can be interpreted as the average “branching factor” per token.

ACTIVITY: Perplexity

We have a language model that assigns probabilities

$$\begin{cases} p(\text{the}) = \alpha_1 \\ p(\text{cat} \mid \text{the}) = \alpha_2 \\ p(\text{sat} \mid \text{the cat}) = \alpha_3 \end{cases}$$

a) Calculate the perplexity of the sequence “the cat sat” when $\alpha_1 = 0.4$, $\alpha_2 = 0.6$, $\alpha_3 = 0.8$.

b) Calculate the perplexity of the sequence “the cat sat” when $\alpha_1 = 0.3$, $\alpha_2 = 0.4$, $\alpha_3 = 0.6$

c) Calculate the perplexity of the sequence “the cat sat” when $\alpha_1 = \alpha_2 = \alpha_3 = \frac{1}{3}$.

d) When the perplexity is higher, is the language model “more sure” or “less sure” about which word to choose?

Language Modeling

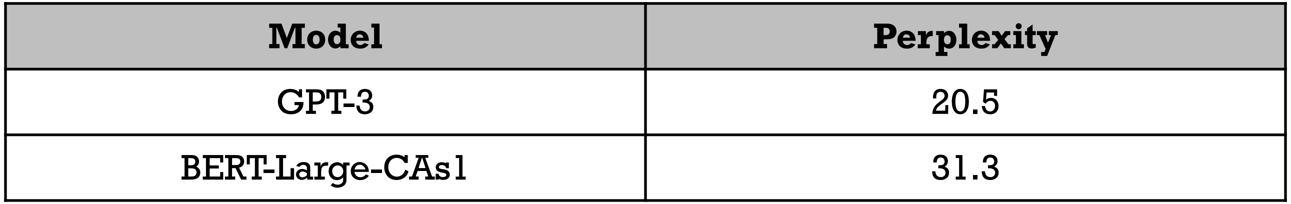
Penn Tree Bank

The Penn Tree Bank is a classic dataset in NLP, originally annotated for syntactic parsing. Beginning with Emami and Jelinek (2004) and Mikolov and Zweig (2012), a version of the dataset that only contained Wall Street Journal articles was used as a language modeling evaluation

Adaptation. Feed the entire text as a prompt into GPT-3 and evaluate the perplexity :

Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29. Mr. Vinken is chairman of Elsevier N.V., the Dutch publishing group.

Results. GPT-3 vastly outperforms the existing state-of-the-art:



Model	**Perplexity**
GPT-3	20.5
BERT-Large-CAs1	31.3

Language Modeling

LAMBADA (Paperno et al. 2016)

- ask: predict the last word of a sentence.
- Motivation: Solving the task requires modeling long-range dependencies.

https://huggingface.co/datasets/lambada

Language Modeling

LAMBADA (Paperno et al. 2016)

Adaptation.

- LAMBADA is natively already a language modeling task, so we could just ask a language model to complete the final word of the sentence.
- Problem: language model doesn't know it should be producing the final word of the sentence.
- Solution: frame it more explicitly as a input-output mapping and use in-context learning with additional examples:

Fill in blank:

Alice was friends with Bob. Alice went to visit her friend ___. → Bob

She held the torch in front of her.

She caught her breath.

"Chris? There's a step."

"What?"

"A step. Cut in the rock. About fifty feet ahead." She moved faster. They both moved faster. "In fact," she said, raising the torch higher, "there's more than a ___. → step

Language Modeling

LAMBADA (Paperno et al. 2016)

Results. GPT-3 does much better on this task than the previous state-of-the-art (based on GPT-2):

Model	**Perplexity**
GPT-3 (few-shot)	1.92
SOTA	8.63

Language Modeling

HellaSwag (Zellers et al. 2019)

- Motivation: evaluate a model’s ability to perform common-sense reasoning
- Task: choose the most appropriate completion for a sentence from a list of choices

Adaptation. This is a multiple-choice task, so the most natural thing to do is to score each candidate answer with the language model and predict the “best” one :

Making a cake: Several cake pops are shown on a display. A woman and girl are shown making the cake pops in a kitchen. They ${answer}

where ${answer} is one of:

- *bake them, then frost and decorate.*
- *taste them as they place them on plates.*
- *put the frosting on the cake as they pan it.*
- *come out and begin decorating the cake as well.*

https://huggingface.co/datasets/Rowan/hellaswag

Language Modeling

HellaSwag (Zellers et al. 2019)

- Motivation: evaluate a model's ability to perform common-sense reasoning
- Task: choose the most appropriate completion for a sentence from a list of choices

Results. GPT-3 got close but did not exceed the state-of-the-art:

Model	Accuracy
SOTA	85.6
GPT-3	79.3

Question Answering

Now we consider (closed-book) question answering, where the input is a question and the output is an answer. The **language model has to somehow “know” the answer** without looking up information in a database or a set of documents.

Input: What school did burne hogarth establish?

Output: School of Visual Arts

Question Answering

TriviaQA (Joshi et al. 2017)

- Task: given a trivia question, generate the answer
- The original dataset was collected from trivial enthusiasts and was presented as a challenge used for (open book) reading comprehension, but we use it for (closed-book) question answering.

Adaptation. We define a prompt based on the training instances (if any) and the question, and take the completion as the predicted answer:

Q: *'Nude Descending A Staircase' is perhaps the most famous painting by which 20th century artist?*

A: *Marcel Duchamp*

https://huggingface.co/datasets/trivia_qa?row=7

Question Answering

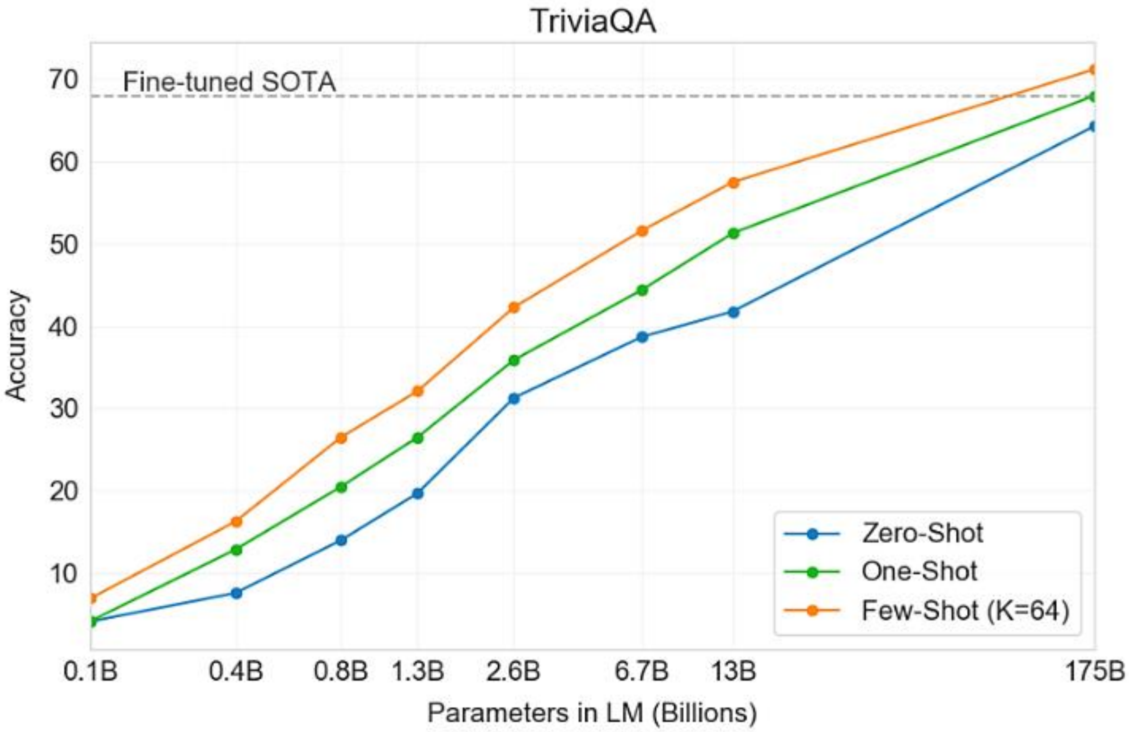
TriviaQA (Joshi et al. 2017)

Results

Model	**Accuracy**
RAG	68.0
GPT-3 (zero-shot)	64.3
GPT-3 (few-shot)	71.2

Question Answering

TriviaQA (Joshi et al. 2017)



Question Answering

WebQuestions (Berant et al. 2013)

- Task: answer questions
- Dataset collected from Google search queries, initially created for question answering on knowledge bases

Adaptation. We define a prompt the same as above:

Q: *What school did burne hogarth establish?*

A: *School of Visual Arts*

https://huggingface.co/datasets/trivia_qa?row=7

Question Answering

WebQuestions (Berant et al. 2013)

Results

Model	**Accuracy**
RAG	45.5
GPT-3 (zero-shot)	14.4
GPT-3 (few-shot)	41.5

Question Answering

NaturalQuestions

- Task: answer questions
- Dataset collected from Google search queries (with long-form answers)

Adaptation. We define a prompt the same as above:

Q: *Who played tess on touched by an angel?*

A: *Delloreese Patricia Early (July 6, 1931 - November 19, 2017), known professionally as Della Reese.*

https://huggingface.co/datasets/trivia_qa?row=7

Question Answering

NaturalQuestions

Results

Model	Accuracy
RAG	44.5
GPT-3 (zero-shot)	14.6
GPT-3 (few-shot)	29.9

BiLingual Evaluation Understudy (BLEU)

The BLEU (Bilingual Evaluation Understudy) (Papineni et al. 2002) score is a metric used to evaluate the quality of machine-generated translations (prediction) compared to reference (target) sentences. The BLEU score for a corpus of candidate translation sentences is a function of the n-gram word precision over all the sentences combined with a brevity penalty computed over the corpus as a whole.

BiLingual Evaluation Understudy (BLEU)

N-gram

- An ‘n-gram’ is a widely used concept from regular text processing.
- It is just a fancy way of describing “a set of ’n’ consecutive words in a sentence”.
- For the sentence “the ball is blue”
  - 1-gram (unigram): “The”, “ball”, “is”, “blue”
  - 2-gram (bigram): “The ball”, “ball is”, “is blue”
  - 3-gram (trigram): “The ball is”, “ball is blue”
  - 4-gram: “The ball is blue”
- Note that the words in an n-gram are taken in order, so “blue is The ball” is not a valid 4-gram.

BiLingual Evaluation Understudy (BLEU)

Precision

- Precision measures the number of words in the Predicted (candidate) Sentence that also occur in the Target (reference) Sentence.

- We would normally compute the Precision using the formula:

$$\text{Precision} = \frac{\text{number of correct predicted words}}{\text{number of total predicted words}}$$

- For example
  - **Target Sentence:** He eats an apple
  - **Predicted Sentence:** He ate an apple

$$\text{precision} = \frac{3}{4}$$

BiLingual Evaluation Understudy (BLEU)

Precision

- Problem #1: Repetition

  We could predict a sentence:

  Target Sentence: He eats an apple

  Predicted Sentence: He He He

$$\text{precision} = \frac{3}{3} = 1$$

BiLingual Evaluation Understudy (BLEU)

Precision

- Problem #2: Multiple Target Sentences

  There are many correct ways to express the same sentence. In many NLP models, we might be given multiple acceptable target sentences

  Target Sentence 1: He eats a sweet apple

  Target Sentence 2: He is eating a tasty apple

  Predicted Sentence: He He He eats tasty fruit

- We account for these two scenarios using a modified Precision formula which we'll call "Clipped Precision".

BiLingual Evaluation Understudy (BLEU)

Clipped Precision Calculation

Predicted: *He He He eats tasty fruit*

Target 1: *He eats a sweet apple*

Target 2:. *He is eating a tasty apple*

- We now do two things differently:
  - We compare each word from the predicted sentence with all of the target sentences. If the word matches any target sentence, it is considered to be correct.
  - We limit the count for each correct word to the maximum number of times that that word occurs in the Target Sentence. This helps to avoid the Repetition problem.

BiLingual Evaluation Understudy (BLEU)

Clipped Precision Calculation

Predicted: *He He He eats tasty fruit*
Target 1 (T1): *He eats a sweet apple*
Target 2 (T2):. *He is eating a tasty apple*

The word “He” occurs only once in each Target Sentence. Therefore, even though “He” occurs thrice in the Predicted Sentence, we ‘clip’ the count to one, as that is the maximum count in any Target Sentence.

Unigram	Matching Reference	Correct	Count	Clipped Count
he	T1, T2	Yes	3	1
eats	T1	Yes	1	1
tasty	T2	Yes	1	1
fruit	None	No	1	0
Total			6	3

$$p_1 = \frac{\textit{Clipped number of correct predicted } 1\text{−grams}}{\textit{Number of total predicted } 1\text{−grams}} = \frac{3}{6}$$

BiLingual Evaluation Understudy (BLEU)

How is Bleu Score calculated?

Let’s say we have an NLP model that produces a predicted sentence as below.

- **Target Sentence**: The guard arrived late because it was raining
- **Predicted Sentence**: The guard arrived late because of the rain

The first step is to compute Precision scores for 1-grams through 4-grams.

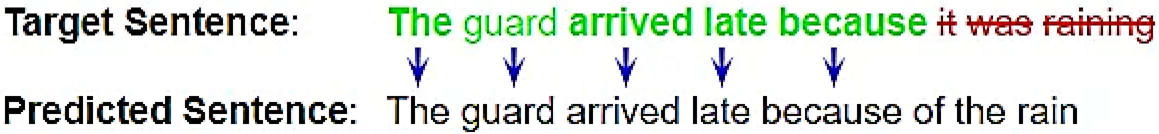
BiLingual Evaluation Understudy (BLEU)

How is Bleu Score calculated?

Let’s say we have an NLP model that produces a predicted sentence as below.

- **Target Sentence**: The guard arrived late because it was raining
- **Predicted Sentence**: The guard arrived late because of the rain

Precision 1-gram



Target Sentence: The guard arrived late because ~~it was raining~~

Predicted Sentence: The guard arrived late because of the rain

$$\text{Precision 1−gram } (p_1) = \frac{\text{Number of correct predicted 1−grams}}{\text{Number of total predicted 1−grams}} = \frac{5}{8}$$

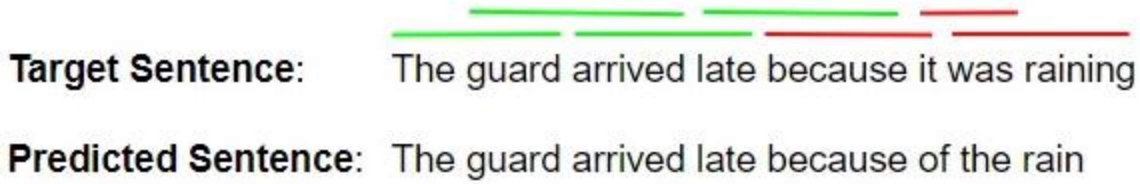
BiLingual Evaluation Understudy (BLEU)

How is Bleu Score calculated?

Let’s say we have an NLP model that produces a predicted sentence as below.

- **Target Sentence**: The guard arrived late because it was raining
- **Predicted Sentence**: The guard arrived late because of the rain

Precision 2-gram



Target Sentence: The guard arrived late because it was raining

Predicted Sentence: The guard arrived late because of the rain

$$\text{Precision 2−gram } (p_2) = \frac{\text{Number of correct predicted 2−grams}}{\text{Number of total predicted 2−grams}} = \frac{4}{7}$$

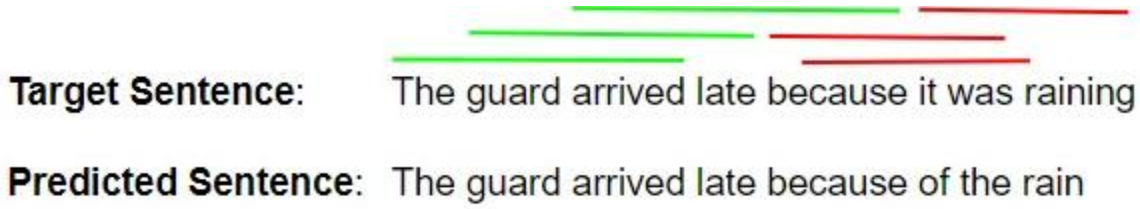
BiLingual Evaluation Understudy (BLEU)

How is Bleu Score calculated?

Let’s say we have an NLP model that produces a predicted sentence as below.

- **Target Sentence**: The guard arrived late because it was raining
- **Predicted Sentence**: The guard arrived late because of the rain

Precision 3-gram



Target Sentence: The guard arrived late because it was raining

Predicted Sentence: The guard arrived late because of the rain

$$\text{Precision } 3-\text{gram } (p_3) = \frac{\text{Number of correct predicted } 3-\text{grams}}{\text{Number of total predicted } 3-\text{grams}} = \frac{3}{6}$$

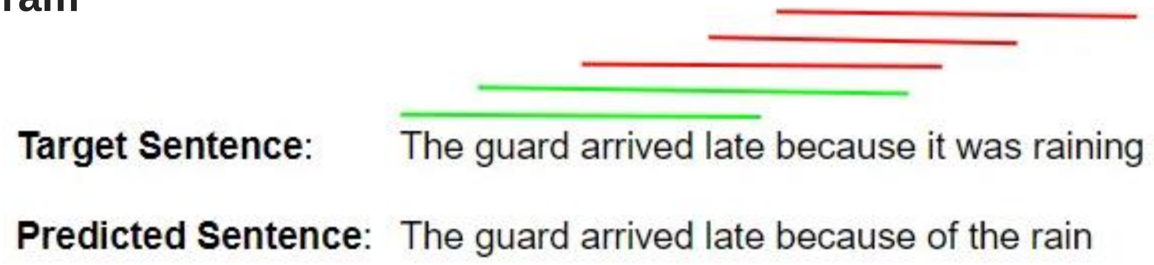
BiLingual Evaluation Understudy (BLEU)

How is Bleu Score calculated?

Let’s say we have an NLP model that produces a predicted sentence as below.

- **Target Sentence**: The guard arrived late because it was raining
- **Predicted Sentence**: The guard arrived late because of the rain

Precision 4-gram



Target Sentence: The guard arrived late because it was raining

Predicted Sentence: The guard arrived late because of the rain

$$\text{Precision 4−gram } (p_4) = \frac{\text{Number of correct predicted 4−grams}}{\text{Number of total predicted 4−grams}} = \frac{2}{5}$$

BiLingual Evaluation Understudy (BLEU)

How is Bleu Score calculated?

- Geometric Average Precision Scores

  We combine these Precision Scores using the formula below. This can be computed for different values of N and using different weight values. Typically, we use N = 4 and uniform weights $w_n = \frac{1}{4}$

$$\text{Geometric Average Precision } (N) = \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

$$= \prod_{n=1}^{N} (p_n)^{w_n}$$

$$= (p_1)^{\frac{1}{4}} \times (p_2)^{\frac{1}{4}} \times (p_3)^{\frac{1}{4}} \times (p_4)^{\frac{1}{4}}$$

BiLingual Evaluation Understudy (BLEU)

How is Bleu Score calculated?

- Brevity Penalty

  If you notice how Precision is calculated, we could have output a predicted sentence consisting of a single word like “The’ or “late”. For this, the 1-gram Precision would have been 1/1 = 1, indicating a perfect score. This is obviously misleading because it encourages the model to output fewer words and get a high score.

  To offset this, the Brevity Penalty penalizes sentences that are too short.

$$\text{Brevity Penalty} = \begin{cases} 1 & \text{if } c > r \\ e^{\left(1-\frac{r}{c}\right)} & \text{if } c \leq r \end{cases}$$

  – c is predicted length = number of words in the predicted sentence
  – r is target length = number of words in the target sentence

BiLingual Evaluation Understudy (BLEU)

How is Bleu Score calculated?

- Finally, to calculate the Bleu Score, we multiply the Brevity Penalty with the Geometric Average of the Precision Scores.

$$\text{Bleu } (N) = \text{Brevity Penalty} \times \text{Geometric Average Precision Score}$$

- Bleu Score can be computed for different values of N. Typically, we use N = 4.
  - BLEU-1 uses the unigram Precision score
  - BLEU-2 uses the geometric average of unigram and bigram precision
  - BLEU-3 uses the geometric average of unigram, bigram, and trigram precision
  - and so on.

ACTIVITY: BLEU Score

Given the prediction

It is a guide to action which ensures that the military always obeys the commands of the party.

and the following target

- *It is a guide to action that ensures that the military will forever heed Party commands.*

Calculate the BLEU score.

https://aclanthology.org/P02-1040.pdf

Translation

- Task: translate a sentence in a source language (e.g., German) to sentence in a target language (e.g., English)
- Machine translation has been a long standing NLP task since the 1960s, and statistical machine translation took off within NLP (with its own distinct subcommunity) in the 2000s, followed by neural machine translation in the mid-2010s. It has always been a data-rich field due to the existence of human translators.
- The standard evaluation dataset is the WMT’14 and WMT’16 datasets.
- Since there are multiple possible translations, the (automatic) evaluation metric is BLEU (which captures a notion of n-gram overlap).

https://huggingface.co/datasets/wmt14

Translation

Adaptation. For the few-shot setting, we construct a prompt containing input-output training instances along with the input:

Mein Haus liegt auf dem Hügel. = My house is on the hill.

Keinesfalls dürfen diese für den kommerziellen Gebrauch verwendet werden. = In no case may they be used for commercial purposes.

Translation

Results

Model	Accuracy
SOTA (supervised)	40.2
GPT-3 (zero-shot)	27.2
GPT-3 (few-shot)	40.6

- Even without supervised training data, GPT-3 matches the state-of-the-art of a fully-supervised system!
- This presents a lower bound on how well one can do in machine translation; you would definitely want to leverage the large amount of parallel corpora (aligned input-output pairs).
- Results from French and Romanian are similar.
- Results from English to a foreign language is much worse, which is expected since GPT-3 is primarily an English language model.

Arithmetic

GPT-3 is a language model (primarily on English), but we can evaluate it on a range of more “abstract reasoning” tasks, to evaluate GPT-3 as more of a general-purpose model.

- Task: do arithmetic (2-5 digit addition, subtraction, multiplication)
- There’s no practical reason you would want to solve this task; it’s just a diagnostic task to satisfy our scientific curiosity.

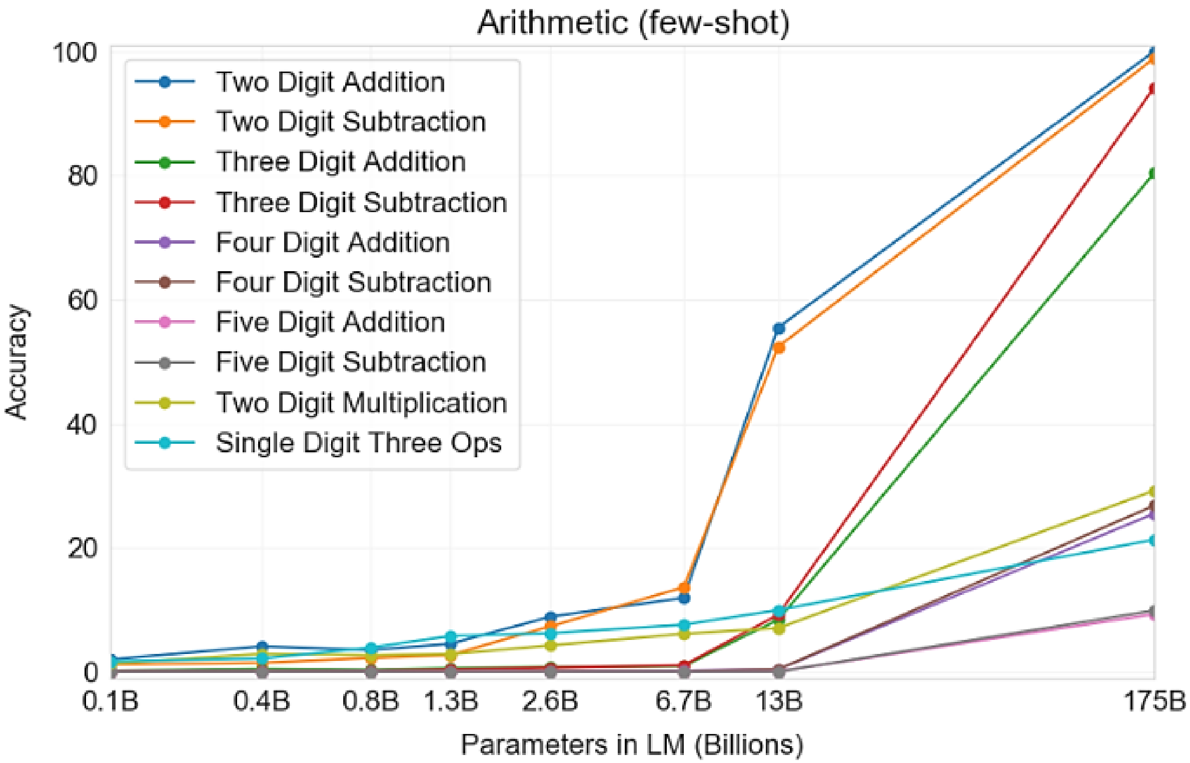
Adaptation. Pose the problem as question answering

Q: *What is 556 plus 497?*

A: *1053*

Arithmetic

Results



News article generation

- Task: given title and subtitle, generate a news article
- Dataset: title/subtitles taken from newser.com
- Evaluation: humans rated articles based on how likely the article was likely to be written by a machine

News article generation

Adaptation. Note: in-context learning was needed to give the model an idea of what a prompt looks like.

Title: *United Methodists Agree to Historic Split*

Subtitle: *Those who oppose gay marriage will form their own denomination*

Article: *After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination...*

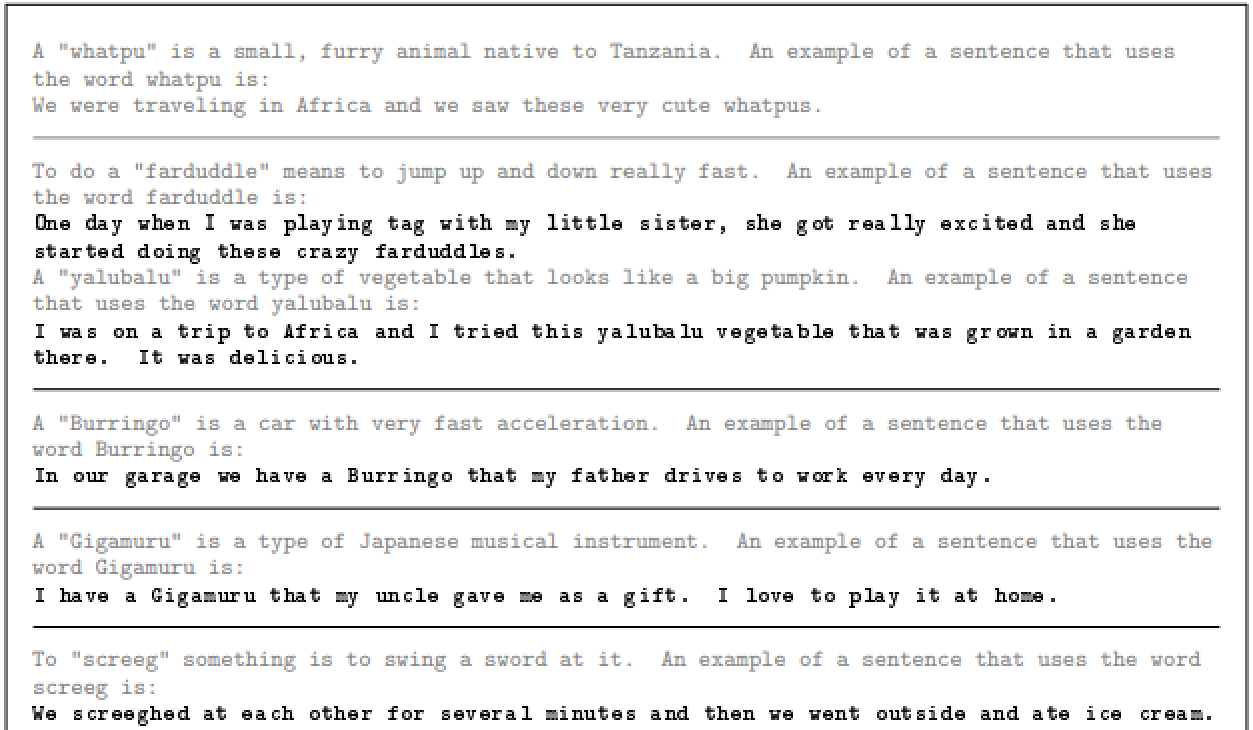
News article generation

Results. Humans were able to able to detect classify “human” versus “machine” only 52% of the time (barely above random chance).

Novel tasks

Using new words

- Task: given a new made-up word and a definition, generate a sentence that uses the word.



A "whatpu" is a small, furry animal native to Tanzania. An example of a sentence that uses the word whatpu is:
We were traveling in Africa and we saw these very cute whatpus.

To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses the word farduddle is:
One day when I was playing tag with my little sister, she got really excited and she started doing these crazy farduddles.
A "yalubalu" is a type of vegetable that looks like a big pumpkin. An example of a sentence that uses the word yalubalu is:
I was on a trip to Africa and I tried this yalubalu vegetable that was grown in a garden there. It was delicious.

A "Burringo" is a car with very fast acceleration. An example of a sentence that uses the word Burringo is:
In our garage we have a Burringo that my father drives to work every day.

A "Gigamuru" is a type of Japanese musical instrument. An example of a sentence that uses the word Gigamuru is:
I have a Gigamuru that my uncle gave me as a gift. I love to play it at home.

To "screeg" something is to swing a sword at it. An example of a sentence that uses the word screeg is:
We screeghed at each other for several minutes and then we went outside and ate ice cream.

Novel tasks

Correcting English grammar

- Task: given an ungrammatical sentence, generate its grammatical version.

Adaptation. The prompt consists of input-output pairs

Poor English input: I eated the purple berries.
Good English output: I ate the purple berries.
Poor English input: Thank you for picking me as your designer. I'd appreciate it.
Good English output: Thank you for choosing me as your designer. I appreciate it.
Poor English input: The mentioned changes have done. or I did the alteration that you requested. or I changed things you wanted and did the modifications.
Good English output: The requested changes have been made. or I made the alteration that you requested. or I changed things you wanted and made the modifications.
Poor English input: I'd be more than happy to work with you in another project.
Good English output: I would be happy to work with you on another project.

Other tasks

Since the original paper, GPT-3 has been applied to many more tasks, including benchmark datasets and one-off demos. Here is an non-exhaustive list.

Benchmarks.

- SWORDS: lexical substitution, where the goal is to predict synonyms in the context of a sentence.
- Massive Multitask Language Understanding: 57 multiple-choice problems spanning mathematics, US history, computer science, law, etc.
- TruthfulQA: question answering dataset that humans would answer falsely due to misconceptions.

The performance on these benchmarks is still mediocre, but it’s perhaps not bad given that we’re doing few-shot learning!

Summary

- GPT-3 was evaluated on a wide range of standard NLP benchmarks and on quirky one-off tasks.
- GPT-3 can perform extremely well or be very mediocre.
- Both increasing the size of the model and the number of examples helps performance.
- There are a few heuristic ways of adapting the language model to the task of interest.
- Why does this work? No one knows.

Refrences

- Language Models are Few-Shot Learners
- https://stanford-cs324.github.io/winter2022/
- https://towardsdatascience.com/foundations-of-nlp-explained-bleu-score-and-wer-metrics-1a5ba06d812b