Large Language Models

Philipp Koehn

24 October 2024

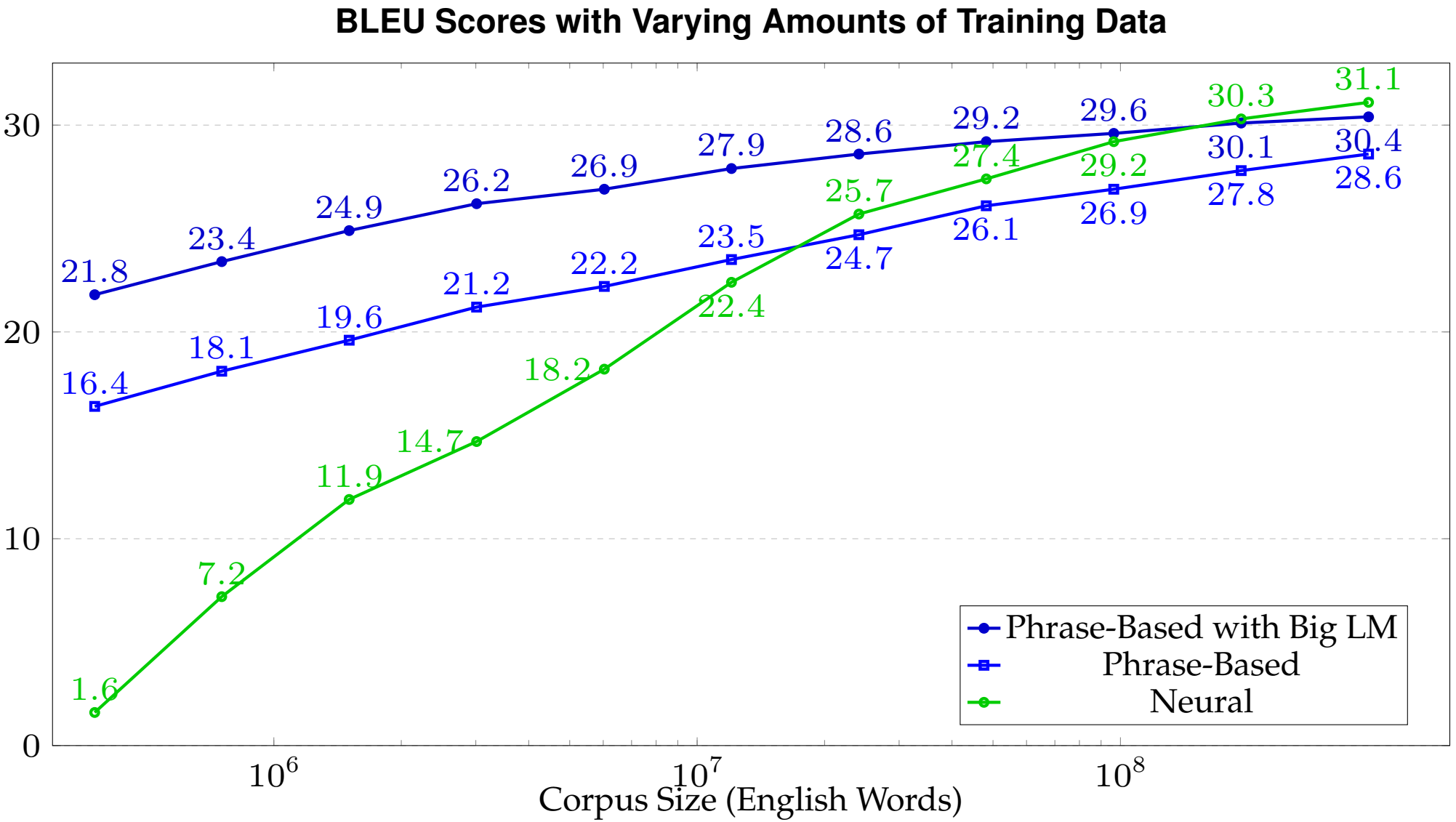
Recall: Statistical Machine Translation

- Statistical Machine translation

$$\text{argmax}_e p(e|f) = \text{argmax}_e \; p(f|e) \; p(e)$$

- Combination of translation model $p(f|e)$ and language model $p(e)$
  - translation model ensures correct meaning
  - language model ensures fluency

Neural vs. Statistical Machine Translation



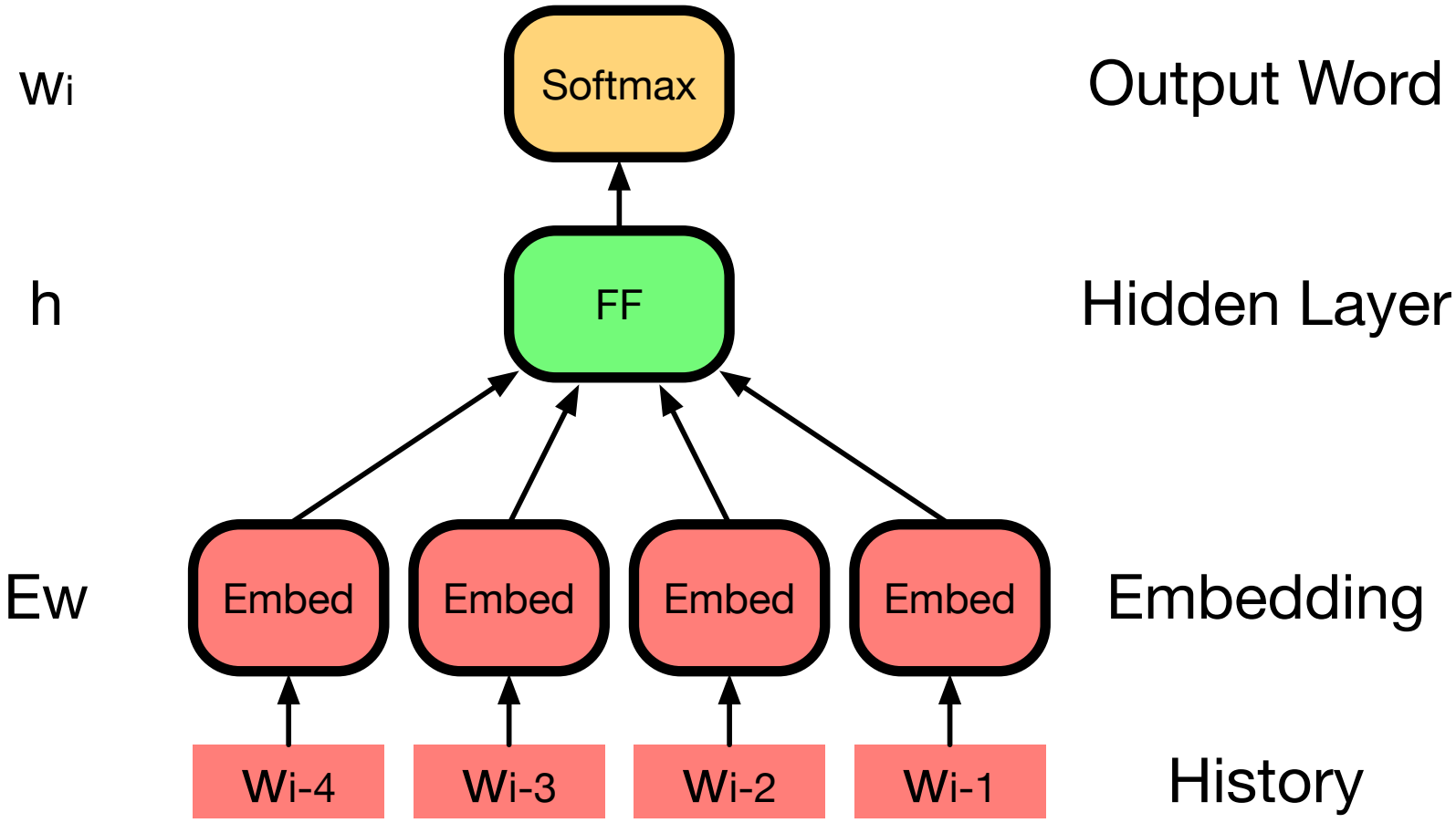
[from *Six Challenges for Neural Machine Translation*, 2017, Koehn and Knowles]

What Happened to the LM in MT?

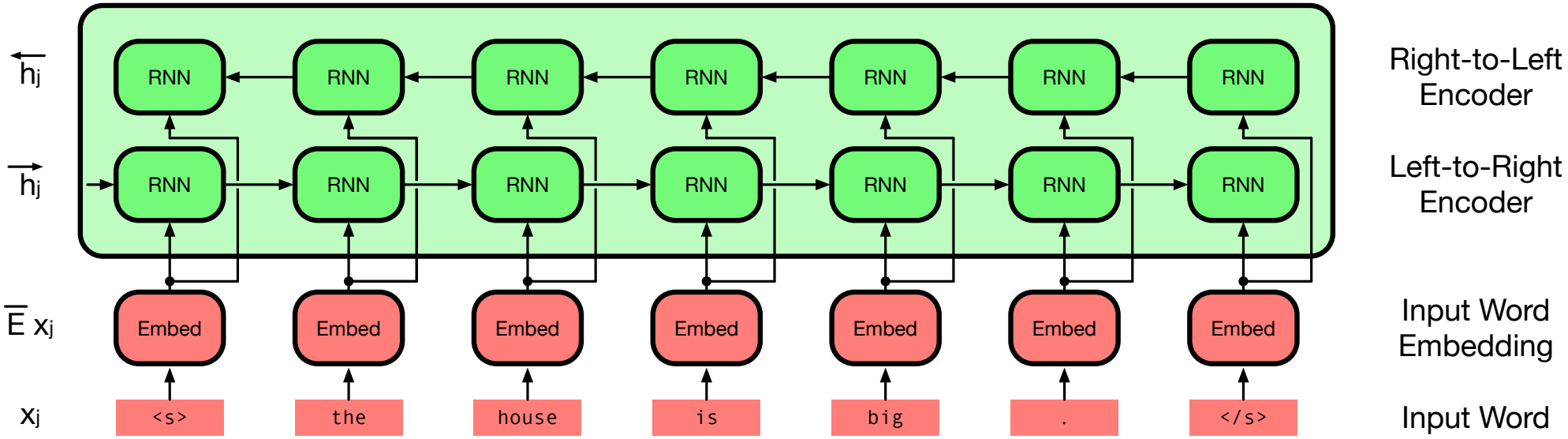
- Edinburgh SMT system 2013: 126 billion token LM [Durrani et al., 2013]
- Fusion model: merge predictions from MT and LM [Gulcehre et al., 2015]
- Backtranslation: synthesize source side of monolingual data [Sennrich et al., 2017]
- mBART: Monolingual pretraining [Liu et al., 2020]
- None of them used data at the scale used in SMT
  LLMs finally do that now (since 2022)

history of large language models

Word Embeddings

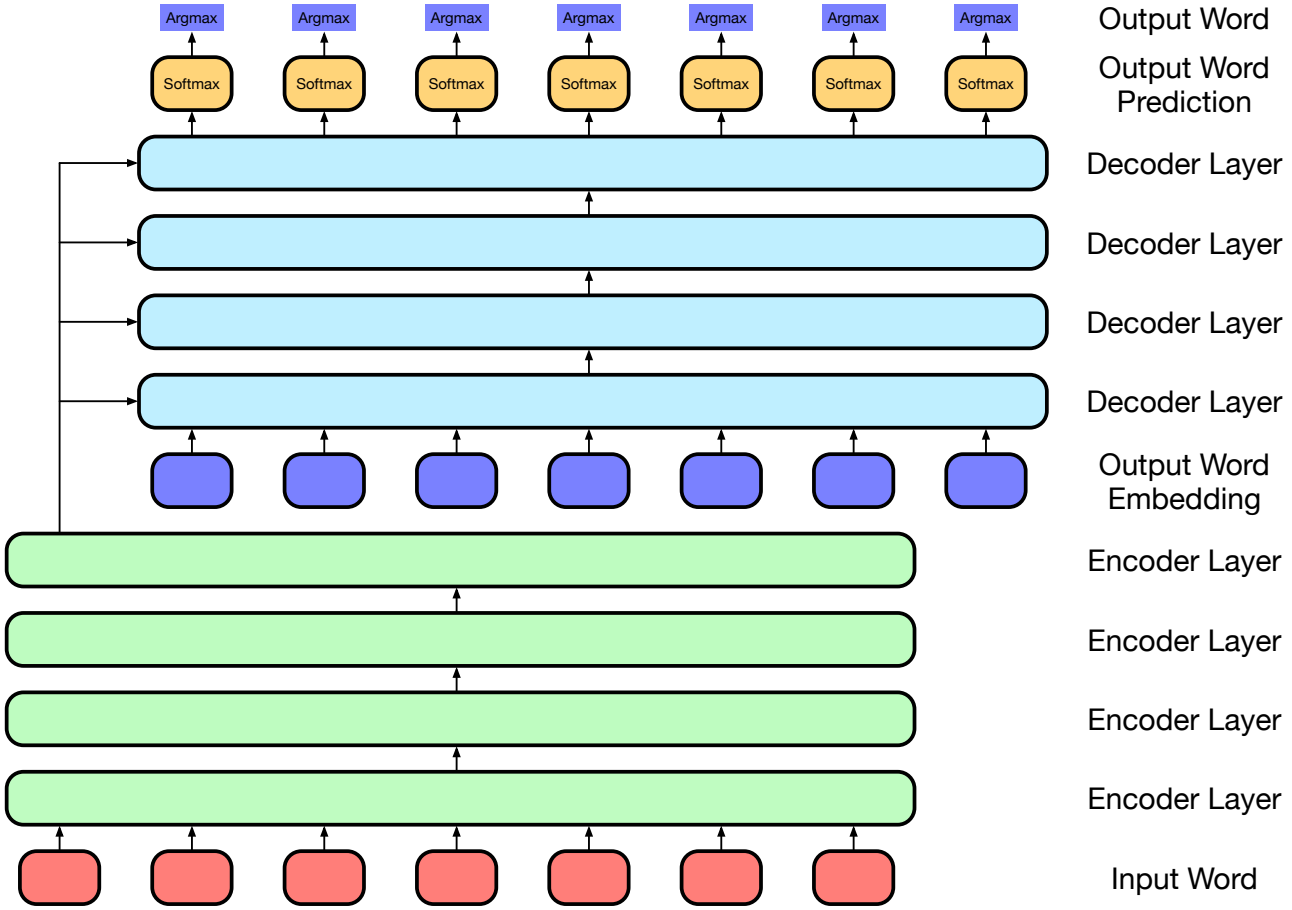


Contextualized Word Embeddings



- ELMo: Embeddings from Language Models (2018)
- Bidirectional LSTM

Contextualized Word Embeddings



- BERT: Bidirectional Encoder Representations from Transformers (2019)

Masked Language Model Training

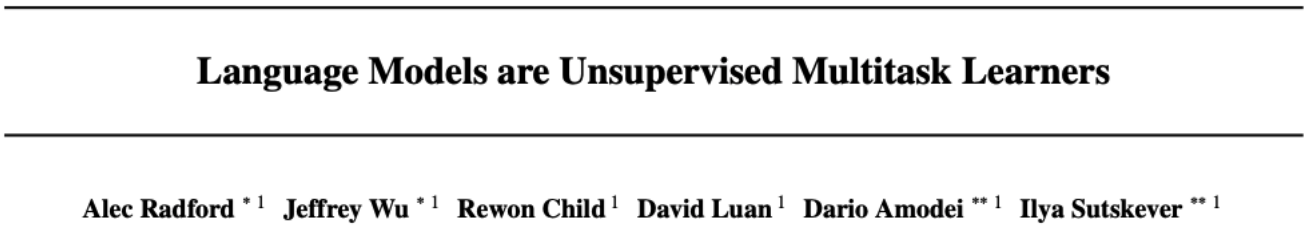
- Transformer expect an input and an output sequence

- Masked training
  - output sequence: one sentence of text
  - input sequence: same sentence, with some words masked out
    This [MASK] an [MASK] → *This is an example*

- Next sentence prediction
  - input sequence: one sentence
  - output sequence: next sentence in document

- Masked training with multiple shuffled sentences
  - same idea as masked training
  - multiple sentences (say, 3)
  - also reorder order of sentence in input sequence

LMs as Unsupervised Learners (2019)



Language Models are Unsupervised Multitask Learners

Alec Radford [*1] Jeffrey Wu [*1] Rewon Child [1] David Luan [1] Dario Amodei [**1] Ilya Sutskever [**1]

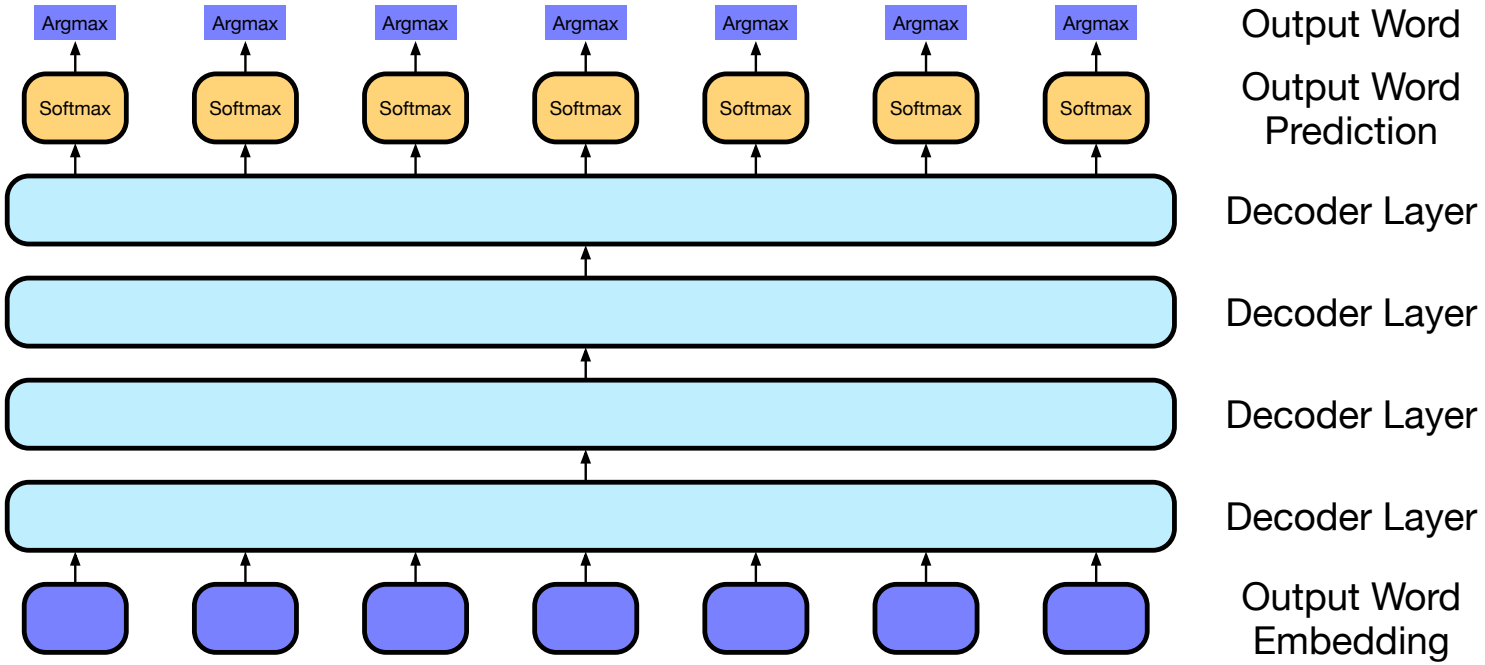
- Train language models on relatively clean text data (GPT-2)
- Convert any NLP problem into a text continuation problem
  - goes into some detail of how each task is converted
  - impressive performance on many tasks

Decoder-Only Models

- Alternative architecture: Just decoder of Transformer model

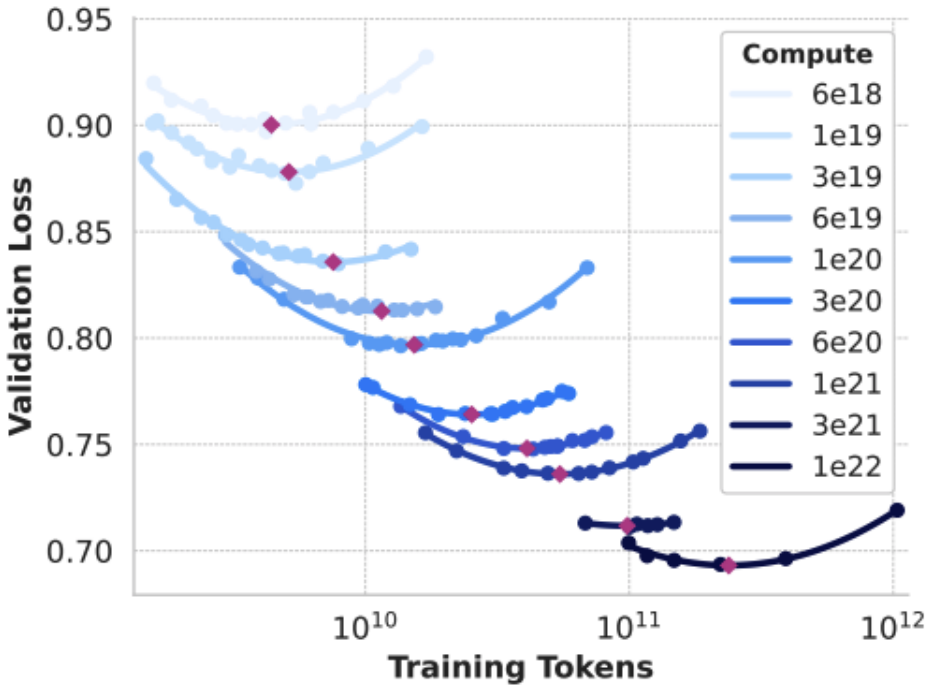
$\Rightarrow$ no input, only self-attention

- Trained with next-word prediction



training

Training



- Scaling laws: more data → bigger models → better performance
- Today: trillions of words → 10s to 100s of billions of parameters
- Llama3 405B: trained on 16,384 GPUs — available open source

Instruction Training

- Examples of requests and responses constructed by human annotators
- May be collected from actual user requests and edited by experts
- May be generated from existing data sets

Question Answering	**Summarization**	**Translation**
What is the highest mountain in the world? The highest mountain in the world is Mount Everest.	*Summarize the following paragraph into one sentence.* The Federal Reserve paused its campaign of interest rate increases for the first time in more than a year. But officials suggested that rates would rise more in 2023, as inflation remains "well above" the central bank's target. Summary: No interest rate rise for now but maybe later in the year.	*Translate from English to German.* English: My name is Ozymandias, King of Kings; Look on my Works, ye Mighty, and despair! German: Mein Name ist Ozymandias, König der Könige; Schau auf meine Werke, du Mächtiger, und verzweifle!

large language models for machine translation

LLMs and MT: The Early Days

- Goodle's PaLM ability to translate [Chowdhery, 2022]
  - decent, below state of the art (WMT-best, Google Translate)
  - good prompting, showing translations in prompt (multi-shot) helps [Vilar et al., 2022]
  - train only on monolingual, multi-shot somewhat successful [Garcia et al., 2022]
- Microsoft's experiments with ChatGPT [Hendy et al., 2023]
  - again, decent, below the state of the art (WMT-best, Microsoft Translator)
- Is machine translation an emergent behavior of large language models?

Translation Data in Training?

- PaLM is exposed to over 30 million translation pairs across at least 44 languages
- Carefully remove all parallel data from training [Briakou et al., 2023]
  - worse translation quality if bilingual content is removed from PaLM training
  - much worse translation quality with smaller (1B, 8B) PaLM models
- How good is the accidental parallel data in PaLM training?
  - mine PaLM training data for parallel data
  - training on mined parallel data (WMT fr-en): 38.1 BLEU
  - training on WMT training data: 42.0 BLEU

Adapting LLMs to MT

- Convert parallel data into chat format

  Translate the following sentence from German to English.
  German: Das Haus ist groß.
  English: The house is big.

- Use it as instruction data for fine-tuning large language model

- Our work: fine-tuning various LLMs with QLoRA [Zhang et al., 2023]
  - Llama2 and BLOOMZ show best results
  - fine-tuning with QLoRa sufficient

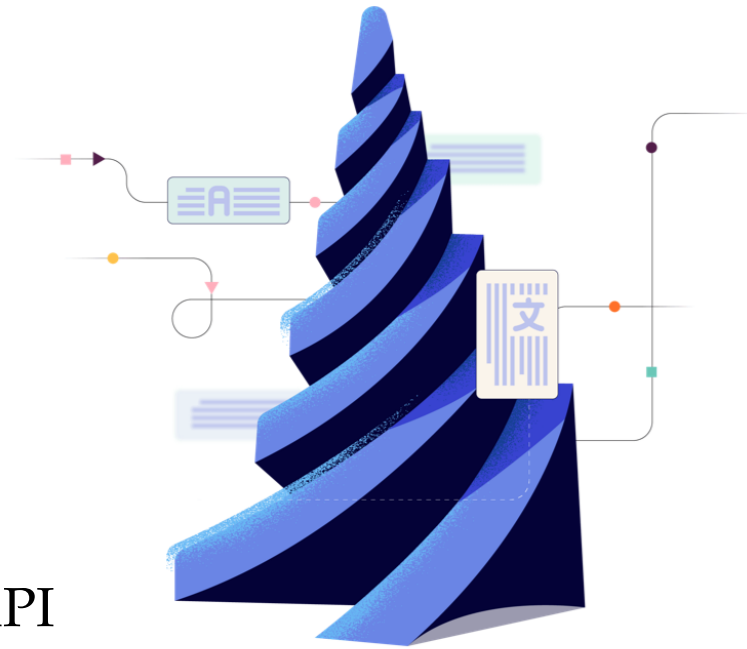
- Adapting for document-level translation [Wu et al., 2024]
- Adapting for specific domains [Zheng et al., 2024]
- BigTranslate: Adapting LLM for MT for over 100 Languages [Yang et al., 2023]

ALMA

- Advanced Language Model-based trAnslator [Xu et al., 2024]
- Two stage fine tuning
  - fine-tune on monolingual data for languages involved
  - fine-tune on parallel data
- Not much parallel data needed (only WMT development sets)
- Building on Llama2-7B, better translation performance than GPT3.5

TowerLLM

- LLM for MT released by Unbabel
- Adapted Llama2 to large amounts of...
  - monolingual data (from mC4)
  - parallel data (filtered with COMETKIWI)
- Instruction-tuning data
  - includes other tasks
  - also multi-shot translation
  - publicly released (TOWERBLOCKS)
- 7B and 13B models publicly released, 70B via API



- 7B model training took 80 GPU days, instruction tuning 8 GPU days (A100)

prompt engineering

Prompt Engineering

- How a task is presented to the language model matters
- Black art: often unclear what is in the training data
- *"Imagine you are..."*: creates a style or sets a context for generation
- Detailed instructions
  - *"Translate every word."*
  - *"Do not provide any explanations."*
  - *"Do not drop or add any information."*

Translation Guidance by Specification

- Language models allow text generation with specifications
  - *"Write in the style of William Shakespeare."*
  - *"Use formal language."*
  - *"Write in the style of a marketing message."*
  - *"Use the following terminology in your translation: ..."*

- Unclear, if this requires any additional model training
  (especially in the case of a very specific style, e.g., house style of a company)

- Alternatively, provide example of target language text in the expected style

Chain of Thought

- Language model has very limited working memory
- Complex reasoning may require several inference steps

> What is 56 times 204? Think step by step.
>
> I first multiply 50 and 200, which is 10,000.
> Then I multiply 6 and 200, which is 1,200.
> I first multiply 50 and 4, which is 200.
> Then I multiply 6 and 4, which is 24.
> Adding together 10,000 and 1,200 and 200 and 24 gives 11,424.

- Intermediate steps create working memory

Decomposition of Tasks

Write an outline for an argument that LLMs are great.
Outline:

1. LLMs have emerged as general problem solvers
2. LLMs have become publicly available
3. Many people find LLMs useful 4. ...

Write a paragraph about "LLMs have emerged as general problem solvers".

Language model (LLMs) have rapidly emerged as versatile general problem solvers across various domains. Leveraging vast datasets and sophisticated algorithms, LLMs exhibit remarkable capabilities in understanding, generating,...

Translating Step-by-Step

[Briakou et al.; Google, 2024]

- Inspired by the work process of professional translators
  - pre-drafting research: find hard translation problems
  - drafting: translation with focus on accuracy
  - refinement: improve fluency
  - proofreading: check for errors

- All these steps contribute to the quality of the translation

Translating Step-by-Step

Pre-drafting Research

You will be asked to translate a piece of text from English into Chinese following the five stages of the translation process. Here is the context in which the text appears:

Context: However, he said the most "responsible way" to cut spending would be to pass all 12 bills. With many Republicans reticent to see a shutdown and a 1 percent cut to defense spending, the urgency to pass the bills could see enough Republicans ally with Democrats to extend the deadline on the four spending bills and reach an agreement to fund the government, even if doing so puts Johnson in hot water with some of those on the hard right.

To start, let's do some pre-drafting research on the above context:

Research:

During this phase, thorough research is essential to address components of the context text that pose translation challenges. The goal is to establish a comprehensive translation plan that covers the following category:

- Idiomatic Expressions:
  - Identify idiomatic expressions that cannot be

Drafting

Now, let's move on to the drafting stage.

Draft Translation:

In this phase, your primary objective is to create a draft translation that accurately conveys the meaning of the source text presented below. At this stage, it is crucial to focus on adequacy, ensuring that your translation closely adheres to the source text. Your response should conclude with the draft translation. If context is missing, generate a general translation that is adaptable to various contexts. Avoid adding any additional information not present in the source text. All elements of the source text should be present in the translation.

Give your best one translation for the following piece of text based on the pre-drafting analysis without providing alternatives:

English: However, he said the most "responsible way" to cut spending would be to pass all 12 bills. With many Republicans reticent to see a shutdown and a 1 percent cut to defense spending, the urgency to pass the bills could see enough

in-context learning

In-Context Learning

- Problem
  - language models are trained on very diverse language usage
  - it may be confused on what it is expected to do
- Solution: provide examples ("shots") of the task in the prompt
- This has been shown to be successful even for new tasks

Multi-Shot Translation

- Provide examples in the prompt

> Translate from German to English. Here are some examples.
> German: Ein Hund bellt. English: A dog barks.
> German: Ein Schwein grunzt. English: A pig grunts.
> German: Eine Katze miaut. English: A cat meows.
> German: Ein Wolf heult. English: A wolf howls.
>
> Now translate the following sentence.
> German: Ein Vogel singt. English:

- This is the standard approach when prompting language models

Provide Text as Style Guidance

- We want to translate in a particular style, e.g., patents

> Translate in the style of a patent.
> Here is some example text of the style: According to an aspect of this invention, a method includes detecting a syntactic chunk in a first string in a first language, assigning a syntactic label to the detected syntactic chunk in the first string, aligning the detected syntactic chunk in the first string to a syntactic chunk in a second language string, said aligning based...
>
> Translate from German to English.
> German: Eine oder mehrere der folgenden Funktionen können ebenfalls enthalten sein.
> English:

Specify Terminology

- A common constraint on translation is company-specific terminology
- For example, legal domain
  - *Rechtswissenschaft* = *jurisprudence* (not *law*)
  - *Kläger* = *plaintiff* (not *prosecutor*)
  - *Strafe* = *sentence* (not *penalty*)
- Provide them in the prompt

> Translate from German to English.
> Use the following terminology in the translation...

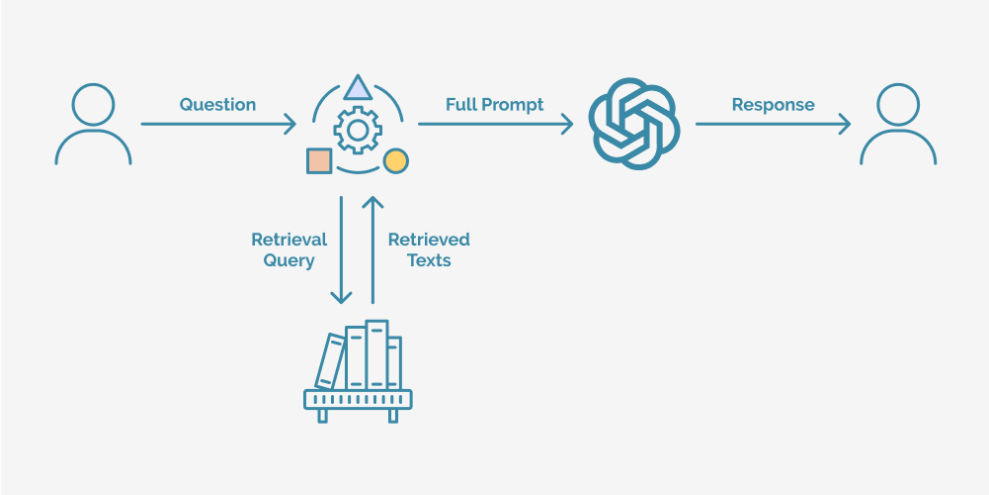
- In reality not so simple: need to distinguish technical and casual use of terms

retrieval augmented generation

Retrieval Augmented Generation

- Example: question answering task
  - may be aided by specific documents
  - these documents may not even be part of LLM training

- Step 1: Retrieve relevant document
- Step 2: Add document to the prompt
- Step 3: Call LLM



Guidance by Relevant Examples

- Idea: show translations of similar sentences before translating

- This idea has been explored very early in NMT models
  - Quickly adapt the model to similar sentence pairs [Li et al., 2016; Farajian et al., 2017]
  - Dedicated conditioning mechanism [Gu et al., 2018; Bapna et al., 2019]
  - Add examples to the input sequence [Bulté et al., 2019; Pham et al., 2020)]

- Multi-shot is default approach when using LLMs for translation

Which Examples?

- Finding examples that lead to best translations on development set [Agrawal et al., 2022]
- Similar source sentences, e.g., measured by n-gram overlap
- Prefer diversity when using multiple examples
- Examples that include key terms

neurosymbolic ai

Tool Use by LLMs

- LLM generates call to APIs
- Executes call
- Receives result back
- Uses result to continue text generation

The New England Journal of Medicine is a registered trademark of [QA("Who is the publisher of The New England Journal of Medicine?") → Massachusetts Medical Society] the MMS.

Out of 1400 participants, 400 (or [Calculator(400 / 1400) → 0.29] 29%) passed the test.

The name derives from "la tortuga", the Spanish word for [MT("tortuga") → turtle] turtle.

The Brown Act is California's law [WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.

(Toolformer, Schick et al., 2023)

Programmatic Calls to LLMs

- If LLMs are universal problem solvers, use them for any task

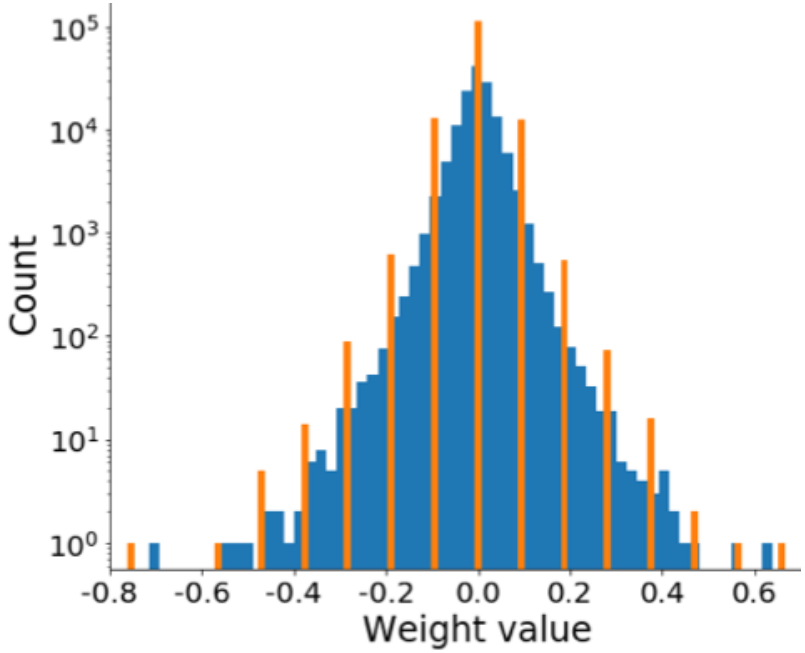
```
a = 1.6
b_text = "What is the distance to the moon in miles?"
c_text = LLM(b_text)
c = process_response(c_text)
print("Distance to moon in km:", a * c)
```

compact models

Large Language Models Very Costly

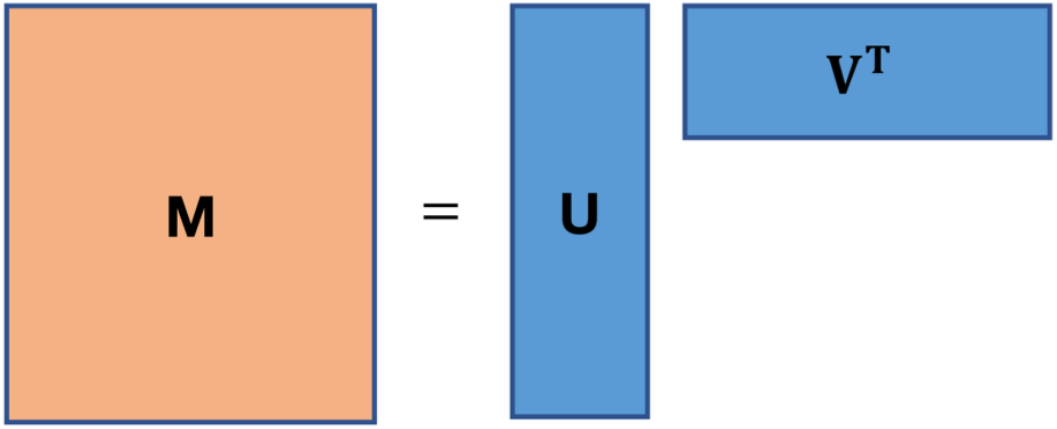
- 10-1000 times as many parameters as dedicated MT models
- More powerful machines needed (with multiple $20,000 GPUs)
- Slower, each translation request more expensive
- Very costly to adapt to particular user cases

Solution 1: Quantization



- Store values in 4 bit floats (or less)
- Computation still in 16 bits
- Additional tricks: double quantization, paged optimizers

Solution 2: Low-Rank Adaptation (LoRA)



- Keep original model intact during adaptation
- Add adaptation parameters in form of low-rank matrices
  - original: n^2 parameter matrix M
  - adaptation:
    - nr and rn matrices U, V
    - with $r << n$
    - e.g., n=2048, r=16

Solution 3: Knowledge Distillation

- Large language model as Teacher
- Small language model as Student
- Data distillation
  - process task-relevant data with Teacher model → good responses
  - use this synthetic data to train Student model
- Model distillation
  - Train Student model directly on predictions of Teacher model

Solution 4: Extend NMT Capabilities

- LLMs have many nice features (as discussed)
- We can integrate these also into neural machine translation models

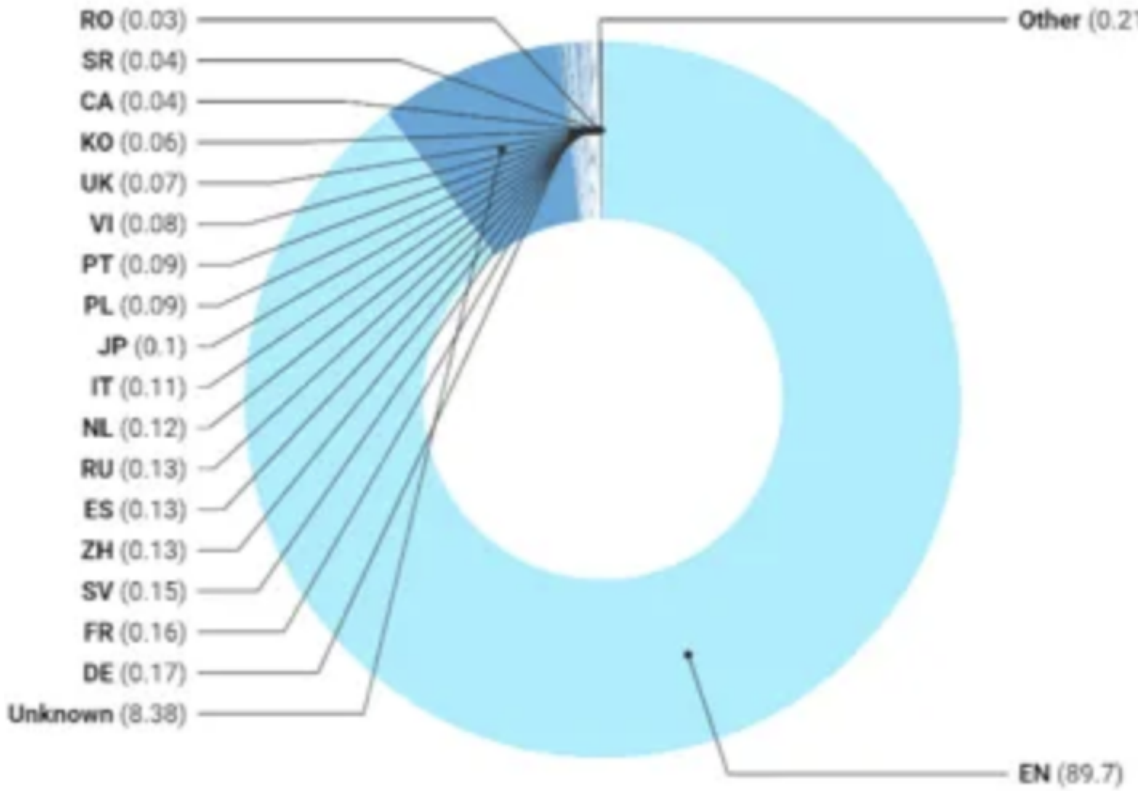
⇒ Instruction-finetuning NMT [Raunak et al., 2024]

  - specify level of formality
  - use specific terminology
  - specifying length of translation
  - specify upper/lowercase
  - etc.

multilinguality

Are Todays LLMs Multilingual?

- Current LLMs trained mostly on English text
  - more English text is available
  - focus on a few languages



Llama2 trained on 89.7% English data

Claimed Multilingual Support

- Cohere Command-R (May 2024)

  The model is optimized to perform well in the following languages: English, French, Spanish, Italian, German, Brazilian Portuguese, Japanese, Korean, Simplified Chinese, and Arabic.

[from https://docs.cohere.com/docs/command-r]

- Llama3.1 (July 2024)

  Supported languages: English, German, French, Italian, Portuguese, Hindi, Spanish, and Thai.

[from https://huggingface.co/meta-llama/Meta-Llama-3.1-405B]

Multilingual Challenges

- Access to information across language barrier

- Prevention of harmful generation
  - Toxic content
  - Cultural bias
  - Hallucination

- Retrieval of relevant information across language barrier
  - retrieval augmented generation
  - evidence to support generated text
  - matching query with document

Knowledge Query in German

Query: *An welchen Universitäten hat Philipp Koehn studiert?*

Response: *Philipp Koehn hat an folgenden Universitäten studiert: Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU): Diplom in Informatik (1990-1996), University of Tennessee, Knoxville: Master of Science in Informatik (1993-1994), University of Southern California (USC): Doktor der Informatik (1997-2003)*

[from Google Gemini, March 2024]

Knowledge Query in Korean

Query: Philipp Koehn은 어느 대학에 다녔습니까?
Responses: 필립 쾨른은 사르브뤼켄 대학교와 에든버러 대학교에서 공부했습니다. 사르브뤼켄 대학교에서 컴퓨터 과학을 전공하여 2003년에 석사 학위를 취득했습니다. 에든버러 대학교에서 인공지능을 전공하여 2005년에 박사 학위를 취득했습니다. 따라서 쾨른은 두 대학 모두에서 학위를 받았다고 볼 수 있습니다. 참고로, 쾨른은 현재 독일 튀빙겐 대학교에서 컴퓨터 과학 교수로 재직하고 있습니다.

Translation (Google Translate): *Philipp Körne studied at the Universities of Saarbrücken and Edinburgh. He studied computer science at the University of Saarbrücken, earning his master's degree in 2003. He graduated with a PhD in Artificial Intelligence from the University of Edinburgh in 2005. Therefore, it can be said that he received degrees from both universities in Cologne. For your information, Köln is currently a professor of computer science at the University of Tübingen, Germany.*

[from Google Gemini, March 2024]

Many Outstanding Challenges

- Retrieval-augmented generation cross-lingually
- Native test sets for many languages
- Language-independent models
- Understand how knowledge is stored in large language models models

questions?