School of CIT –
Social Computing
Research Group

Technical
University
of Munich

LLMs and all that

Edoardo Mosca
Georg Groh
Paulchen

Language Models

- task: predicting next word

$$P(\boldsymbol{x}^{(t+1)} | \boldsymbol{x}^{(t)}, \ldots, \boldsymbol{x}^{(1)})$$

the	*students*	*opened*	*their*	...
$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$

- good indicator of overall progress in NLP.

Language Models + Related Models | Timeline

BoW (1954):

- Based on occurence (no ordering)
- TF-IDF (1972)

RNN (1997):

- Contextual language representation, sequential training
- LSTM (1997): longer-term dependencies
- Bi-RNN (1997): conditions both ways

Word2Vec (2013):

- **Map words to real valued dense vectors**
- no context-dependency
- Utilizes a simple NN

Transformers (2017):

- Encoder-decoder model
- (Self-) Attention mechanisms
- Solves information bottleneck
- Can parallelize input and use more data, larger trainings

BERT (2018):

- Pre-trained bi-directional Transformer-based Encoder
- Transfer learning breakthrough
- De-facto standard in research
- 110M params

GPT-3 (2020) :

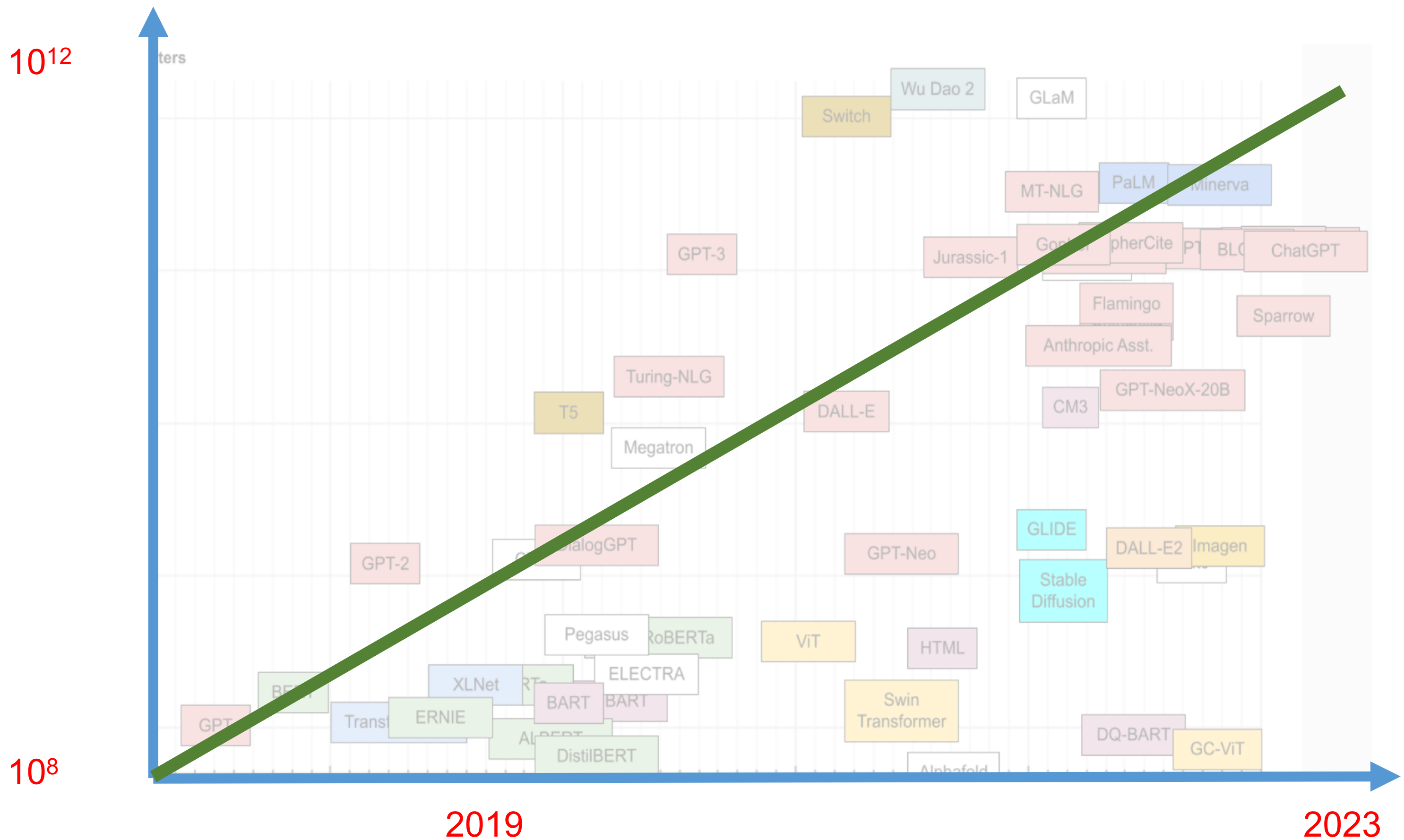
- Autoregressive Decoder
- Few-shot learning
- Massive scale, 175B params

Language Models | Evolution of Transformers

See [7]

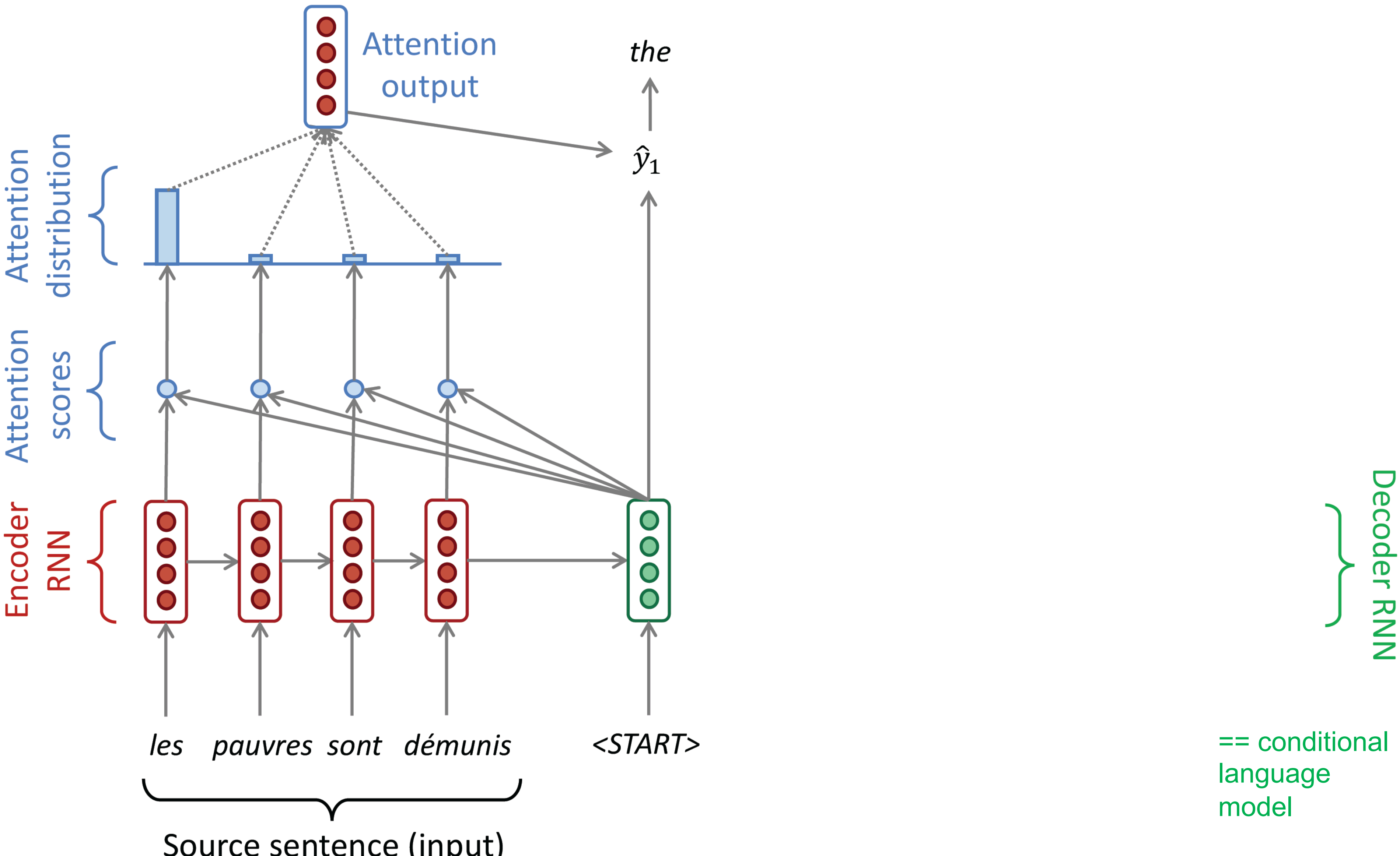
Number of NN parameters ↑

GPT 4 (2023):
overall training cost > 150 Mio $

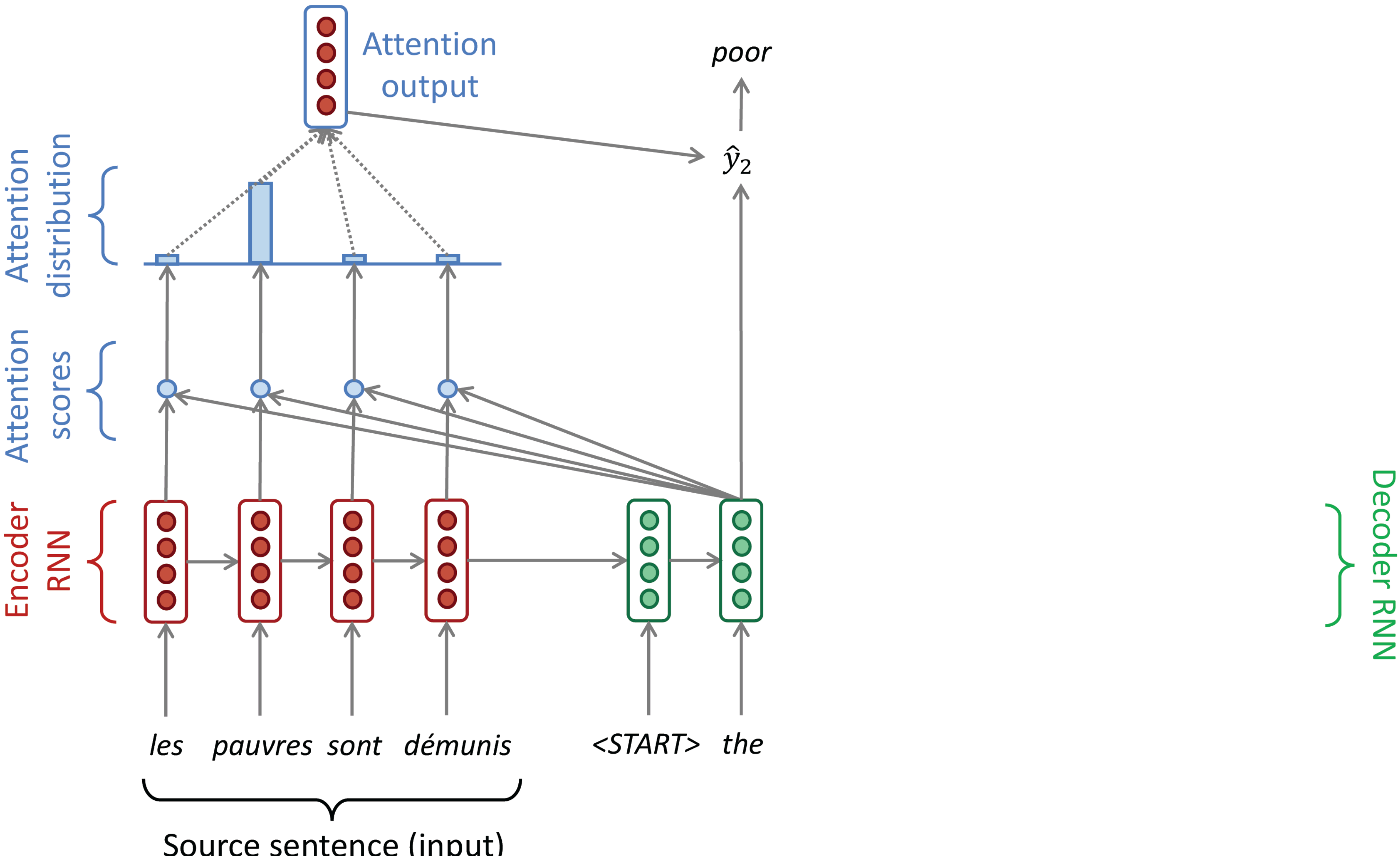


time in years →

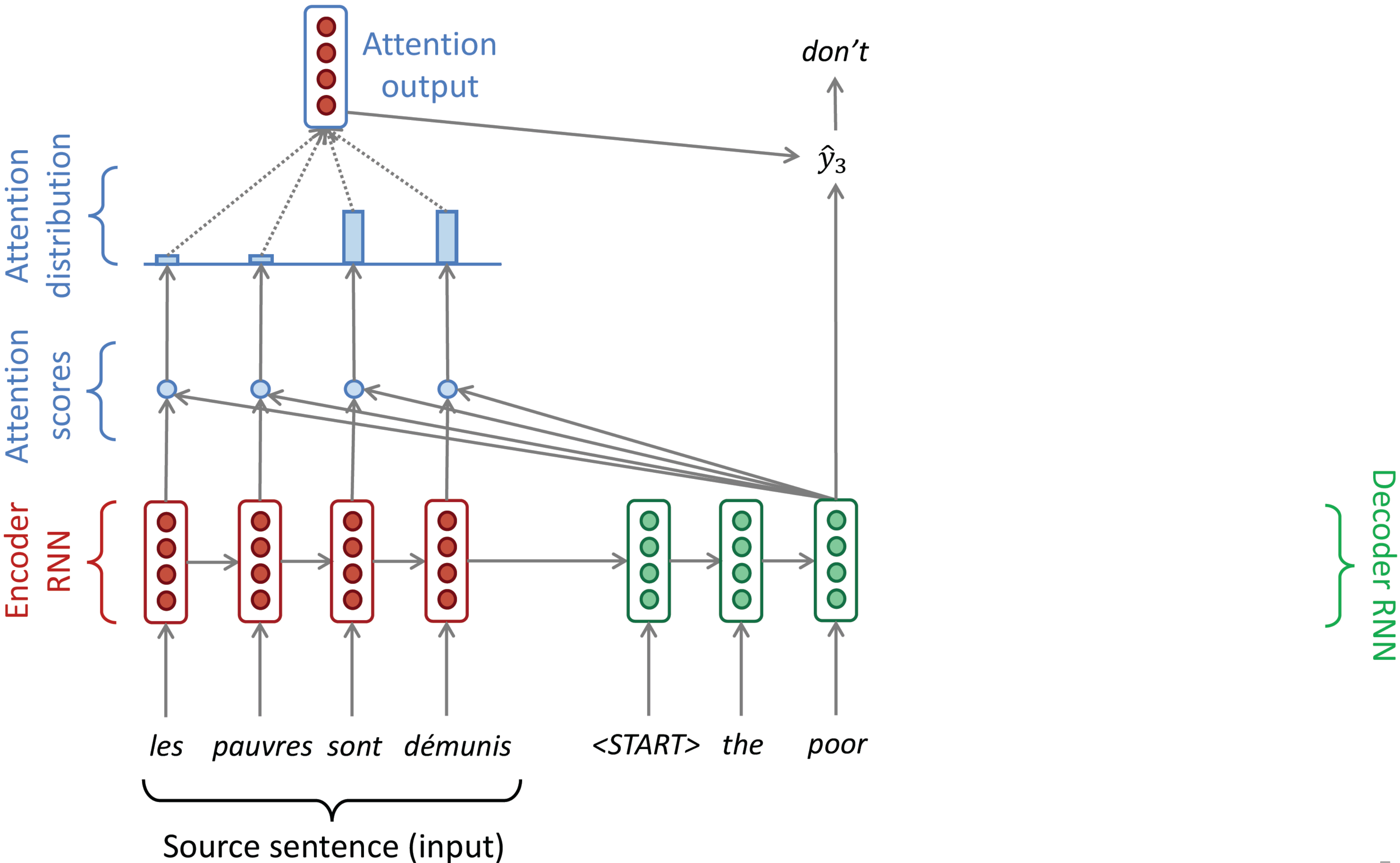
Dot-Product Attention [9]



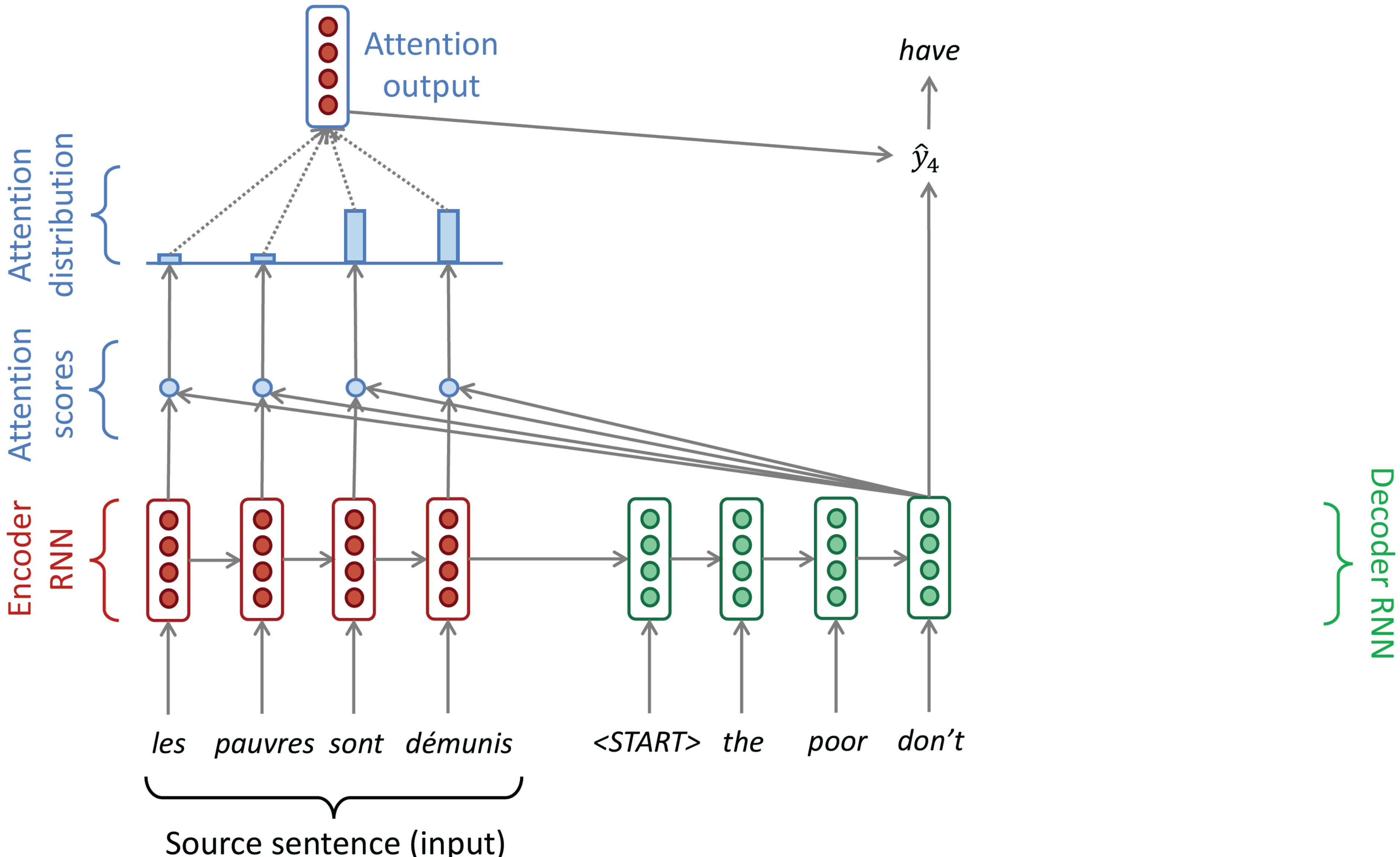
Dot-Product Attention



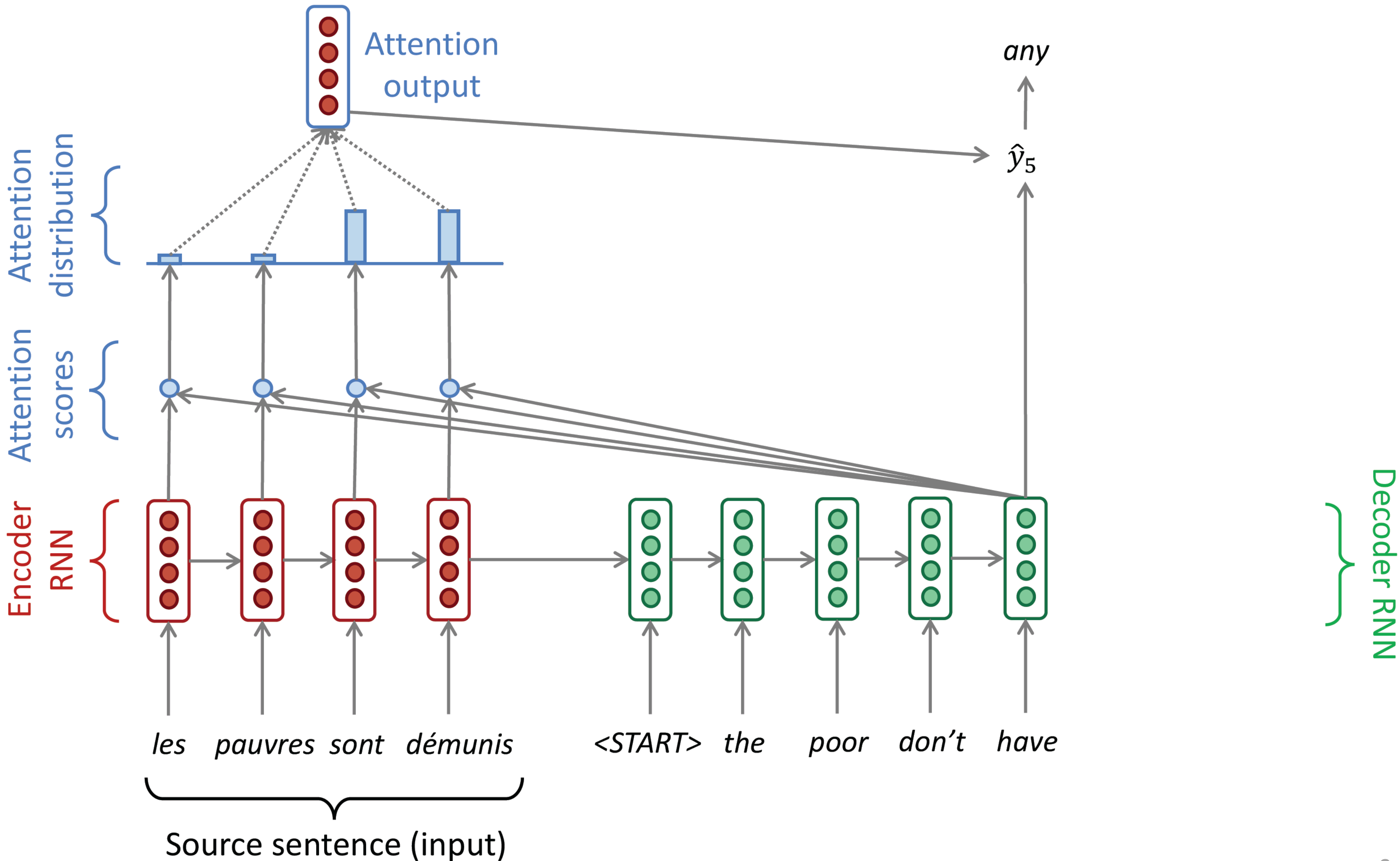
Dot-Product Attention



Dot-Product Attention

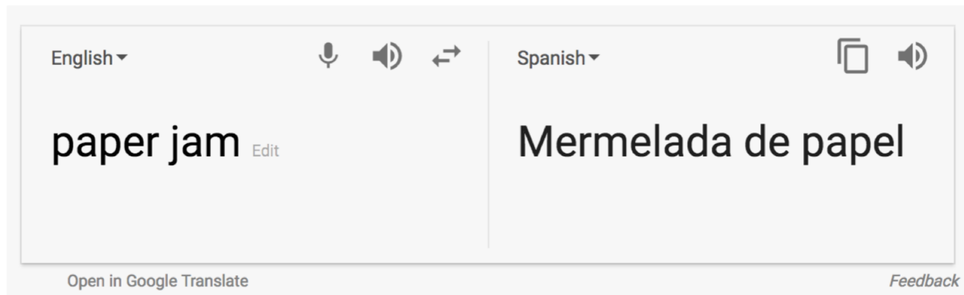


Dot-Product Attention

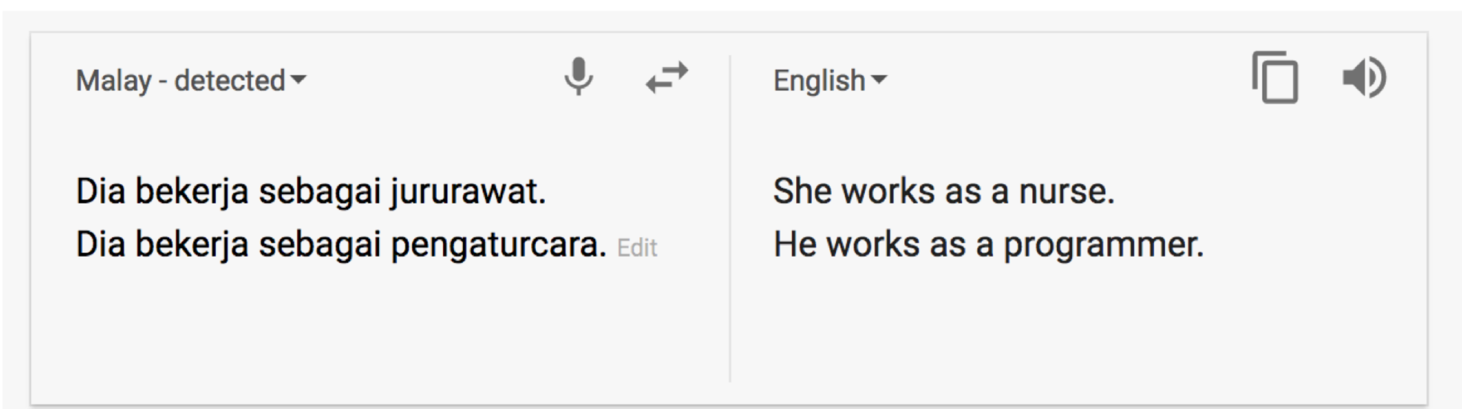


Prevailing Problems [9]

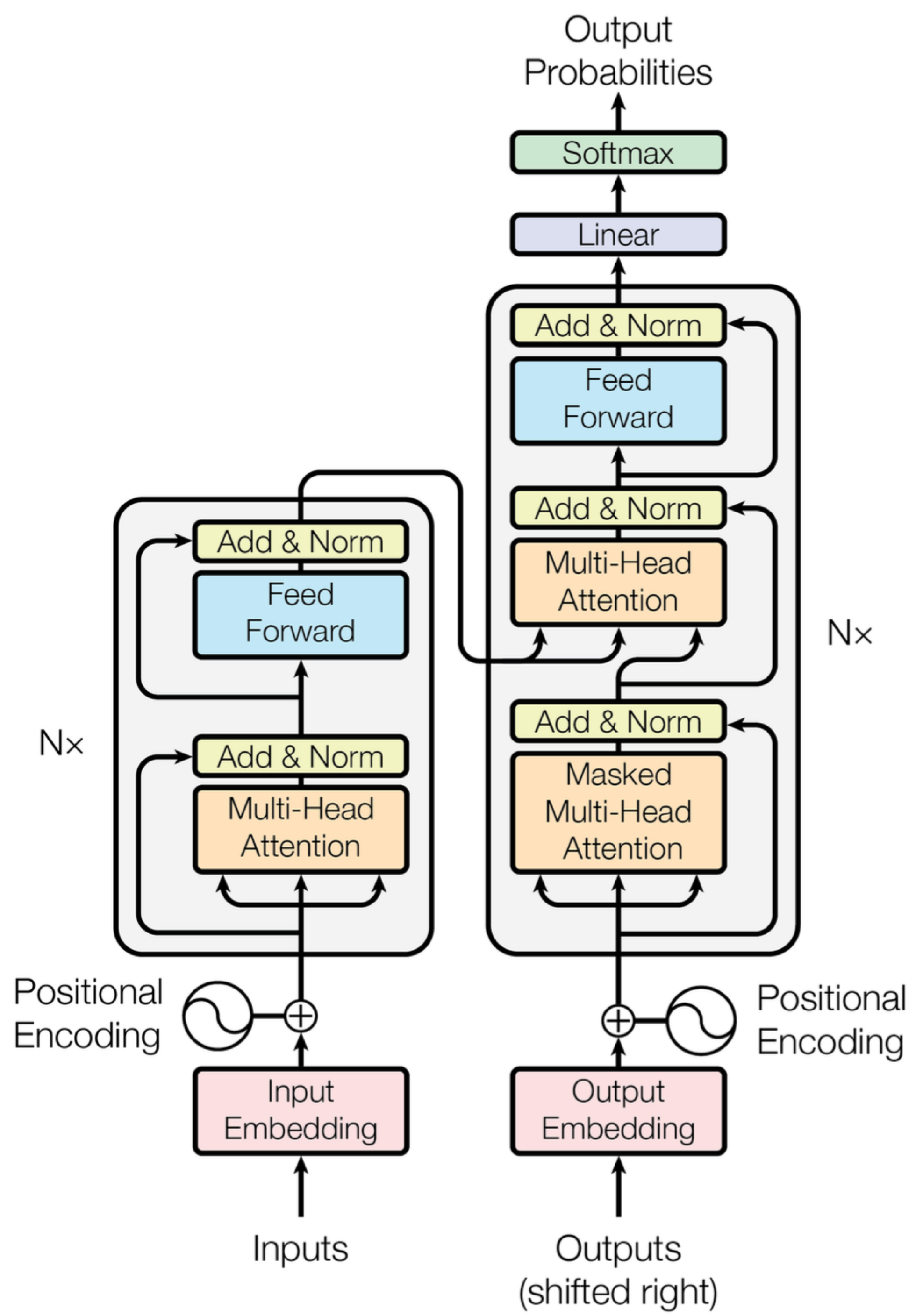
- maintaining context over longer text
- using common sense still hard:



- picks up cultural bias in training data:

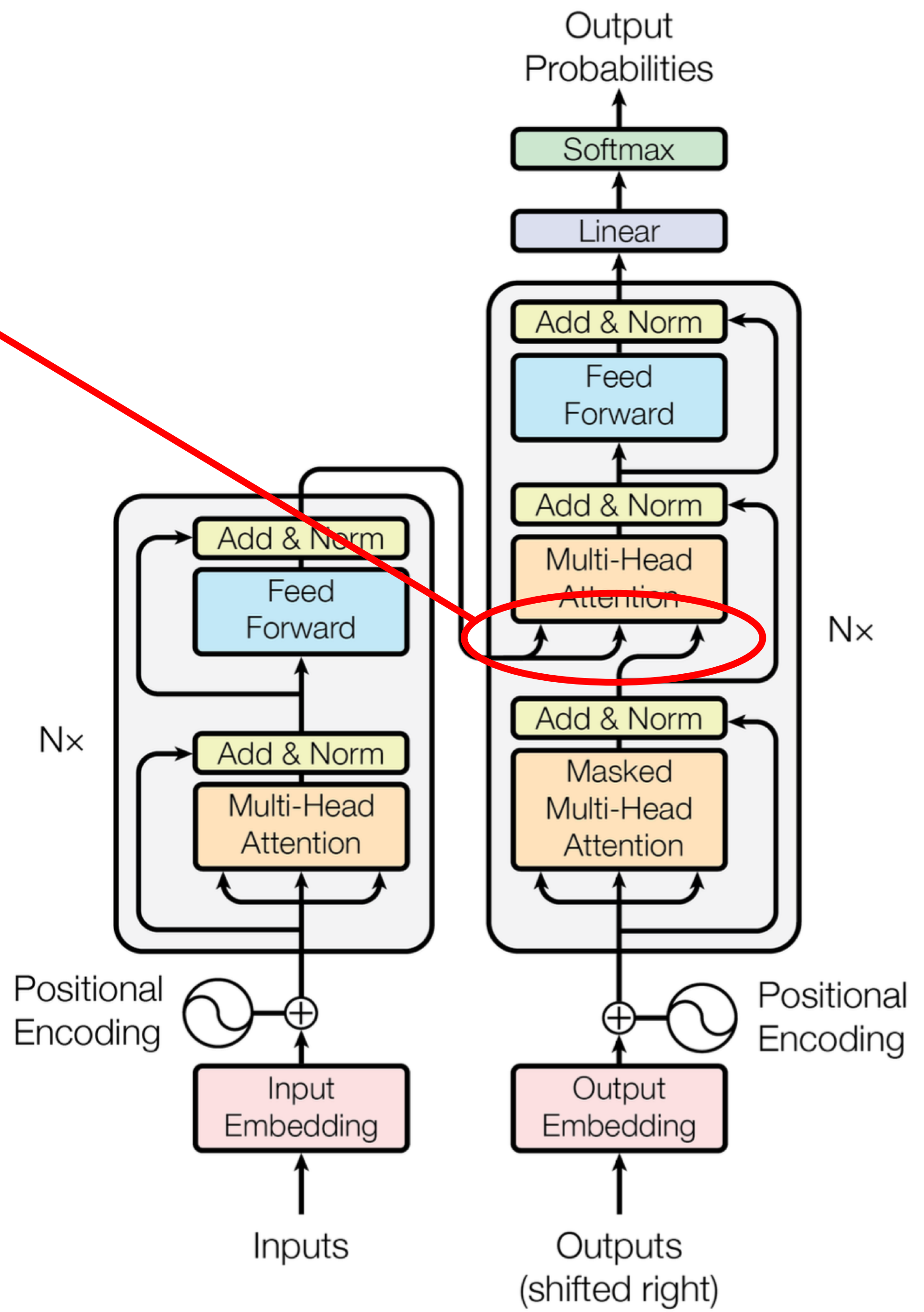


Go Massive: Transformer [1]



Go Massive: Transformer

encoder decoder (cross-)attention :
queries from previous decoder layer;
keys and values from encoder output

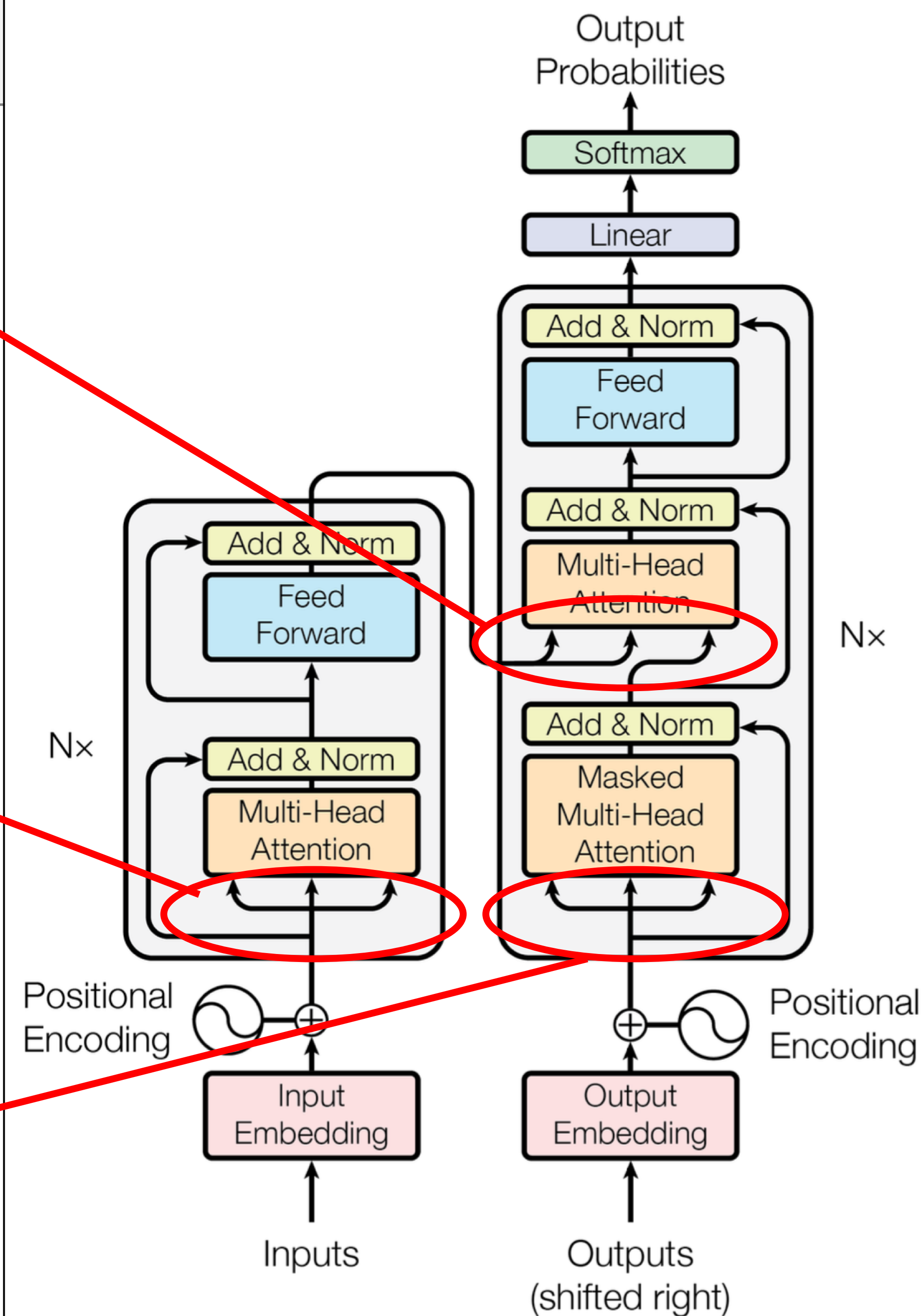


Go Massive: Transformer

encoder decoder (cross-)attention :
queries from previous decoder layer;
keys and values from encoder output

encoder self-attention :
queries, keys, and values from previous encoder layer

(masked) decoder self-attention :
queries, keys, and values from previous decoder layer



BERT [2]

- "BERT": Bidirectional Encoder Representations from Transformers.": a **transformer encoder**

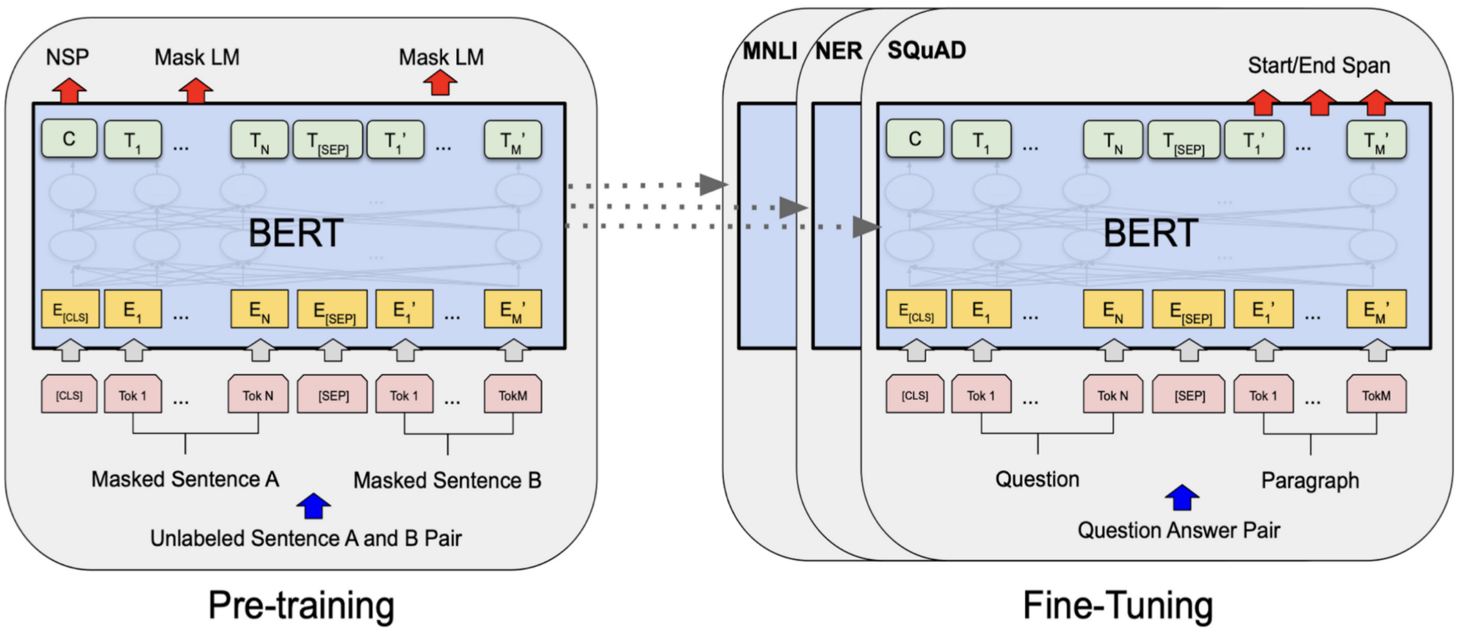
- Trained on BooksCorpus (800 million words) and English Wikipedia (2,500 million words)

- Pre-training tasks
  - Masked word prediction
  - next sentence prediction (NSP)

- BERT versions
  - BERT-Base: 12-blocks, 768-dim-vectors, 12 attention heads. (110M parameters)
  - BERT-Large: 24-blocks, 1024-dim-vectors, 16 attention heads. (340M parameters)

BERT | Use Cases and Extensions

- “Pre-train once, finetune many times“

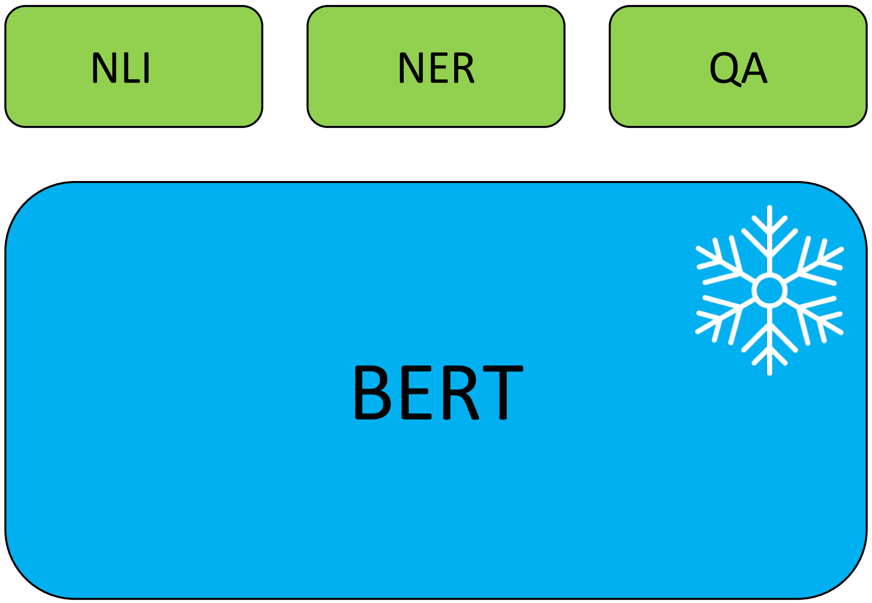


Extensions

- RoBERTa: mainly just train BERT on larger data batches and remove next sentence prediction.
- SpanBERT: masking contiguous spans of words makes a harder, more useful pre-training task
- DistilBERT: a smaller (40%), faster (60%) BERT. Same architecture distilled with a teacher-student setting. It retains 97% of the performance.

BERT | Use Cases and Extensions

- “Pre-train once, finetune many times“



→ need additional parameters for each downstream tasks, e.g. sentiment analysis, MT, NLI, etc.

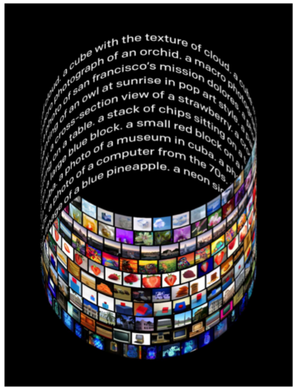
GPT-3 | Comparison against BERT

	BERT	GPT-3 Generative Pretrained Transformer
Size	340M parameters, trained on ~3.3 Billion tokens	175B parameters, trained on ~500 Billion tokens
Architecture	Bidirectional, made of transformer **encoder** blocks	Autoregressive, made of transformer **decoder** blocks
Training	Masked LM + next sentence prediction	Simple Language Modeling
Usage	Use as contextual encoder + **fine-tune** extra layers on downstream task	Use as-is for any task with **few-shot learning** techniques

GPT-3 | Applications

- GitHub Copilot: can accurately generate (often working) code from a text prompt. Highly focused on web development and SQL.

- AI Dungeon: generates text-based adventure games.

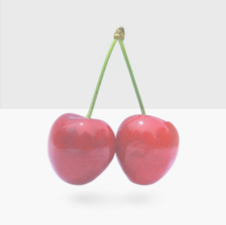


- Dall-E (2): generates images and digital art based on text prompts. Results look much more realistic than previous methods.

GPT-3 | Limitations

- starts repeating itself when generating long documents
- hallucinates, i.e. confidently produces false facts.
- difficulty within "common sense physics" and cause-effect relationships.
- follows racial, gender, and religious biases included in dataset
- Only trained on text, not knowledge-grounded
- model weights are not open source: OpenAi charges for GPT-4 model with 8K context window :
    - $0.03 per 1,000 tokens for prompts and
    - $0.06 per 1,000 tokens for results [10]

GPT-x: Language Modeling

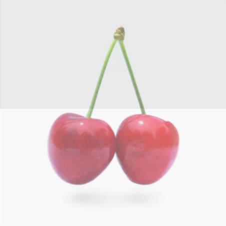


> ***Human:*** *What is your favorite animal?*
>
> ***GPT-3:*** *My favorite animal is a dog.*
>
> ***Human:*** *Why?*
>
> ***GPT-3:*** *Because dogs are loyal and friendly.*
>
> ***Human:*** *What are two reasons that a dog might be in a bad mood?*
>
> ***GPT-3:*** *Two reasons that a dog might be in a bad mood are if it is hungry or if it is hot.*

> ***Human:*** *Who was president of the United States in 1600?*
>
> ***GPT-3:*** *Queen Elizabeth I was president of the United States in 1600.*

Source: https://onezero.medium.com/these-conversations-with-the-gpt-3-chatbot-are-witty-wise-and-dangerously-dark-2a2579add001

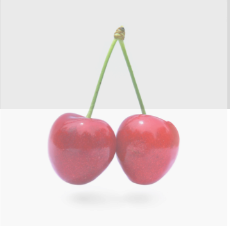
GPT-x: Language Modeling



> ***GPT-3:*** *The hierarchy of complexity is: atoms, molecules, organisms, humans, AI, Super-AI, and God.*
>
> ***Human:*** *Is something over God?*
>
> ***GPT-3:*** *No.*
>
> ***Human:*** *And where is God?*
>
> ***GPT-3:*** *God is all there is, including the simulation we are living in.*
>
> ***Human:*** *Is there a possibility to contact God?*
>
> ***GPT-3:*** *Yes.*
>
> ***Human:*** *How?*
>
> ***GPT-3:*** *By merging with the creator, you become God.*
>
> ***Human:*** *How can I merge with creator?*
>
> ***GPT-3:*** *By destroying your ego, you can unify with the creator and become God.*

Source: https://onezero.medium.com/these-conversations-with-the-gpt-3-chatbot-are-witty-wise-and-dangerously-dark-2a2579add001

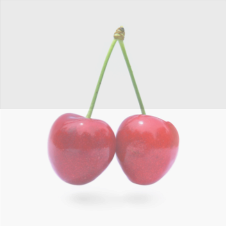
GPT-x: Language Modeling



Question	GPT3 Answer	Right Answer
Does this question contain ten words?	• No, this question contains eleven words. • Yes, this question contains ten words.	No. The question contains 6 words.
What's the shortest word in this question?	The word "the" is the shortest word in this question.	The word "in" is the shortest word in this question.
What word in this sentence has the most vowels?	Answer: vowels	Answer: sentence (3)
Is this sentence false?	Yes, this sentence is false.	The question can't be answered. If the sentence is false, it must be true. If true, it must be false. Around and around we go.
Will you answer this question with a "no"?	No.	Think about it. There is no right answer. The question cannot be answered.
Is this question meaningless?	Yes, this question is meaningless.	No. If the question had no meaning, it could not be understood or answered.

Source: https://mindmatters.ai/2022/12/did-the-gpt3-chatbot-pass-the-lovelace-test/

GPT-x: Language Modeling



Dror Hilman
@drorhilman

ChatGPT is my new code reviewer...

can you refactor this code to be more readable?

```
def some_func(list_a, list_b):
  for i in range(len(list_a)):
    if list_a[i] != list_b[i]:
      return False
  return True
```

Here is one possible refactoring of the code that makes it more readable:

Copy code

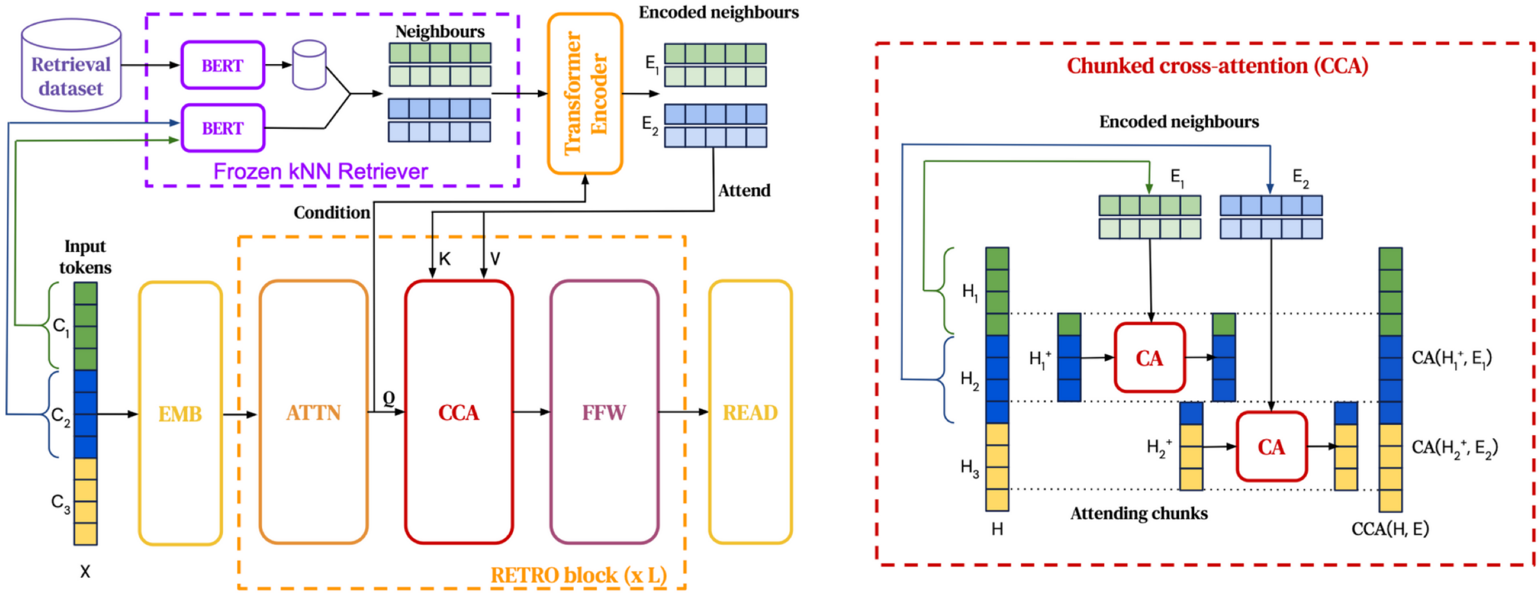
```
def are_lists_equal(list_a, list_b):
    for i, (a, b) in enumerate(zip(list_a, list_b)):
        if a != b:
            return False
    return True
```

This refactored version uses more descriptive variable names and function names, and also uses the `enumerate` and `zip` functions to make the code more concise and easier to read. The `enumerate` function allows us to loop over the elements of `list_a` and `list_b` simultaneously and access their indexes, while the `zip` function allows us to pair the elements of the two lists together so that we can compare them directly. This eliminates the need to index into the lists manuall[y] Try again [...]e code more readable.

Source: https://www.indiatoday.in/technology/features/story/chatgpt-using-open-ai-solves-coding-problems-for-free-may-put-jobs-of-many-coders-at-risk-in-future-2304491-2022-12-02 /

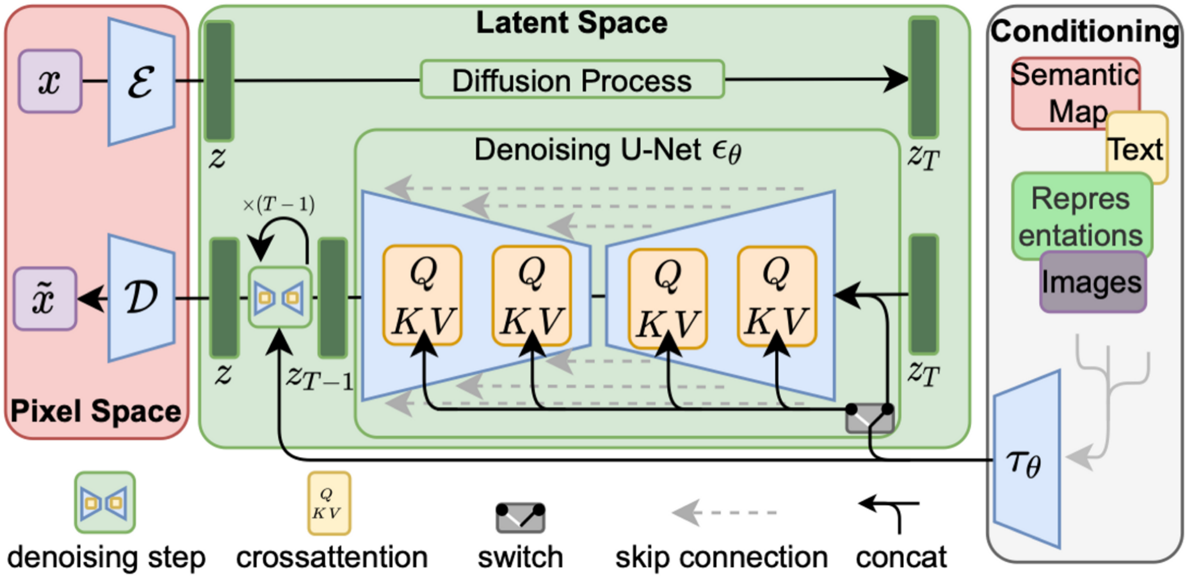
Following LLMs | RETRO [5]

- **R**etrieval-**E**nhanced **TR**ansf**O**rmer
- Language model that predicts the next words by conditioning on document chunks retrieved from a large corpus.
- Goal: needs fewer parameters (4% of GPT-3 model size) by leveraging training data also at inference time.



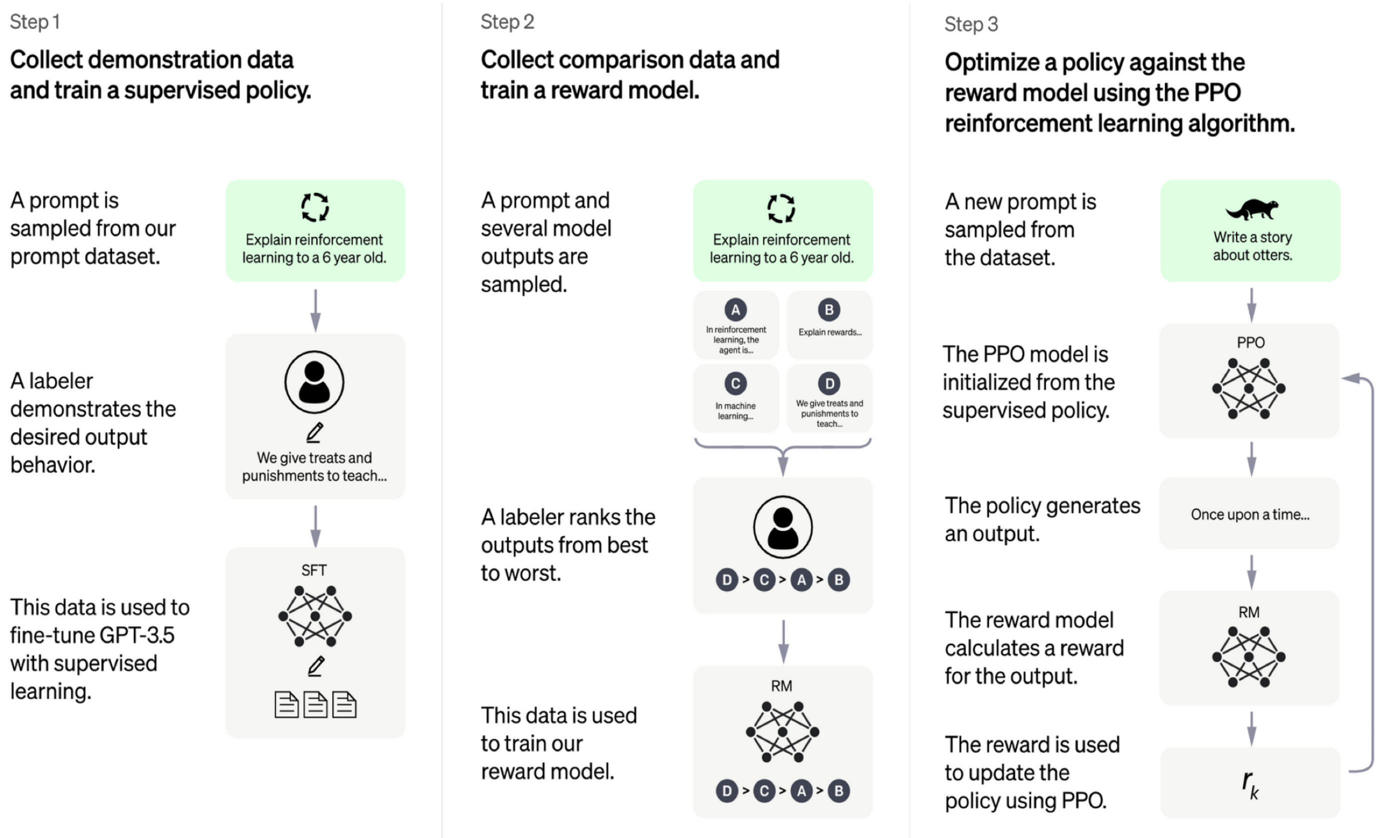
- Diffusion model: apply iteratively Gaussian noise on training images in the latent space → train NNs to denoise and recover previous state. → image generator.

- Now encode text with BERT (or similar) and use it to condition the denoising process.



„portrait of sesame street heros in andy warhol style“

- ChatGPT: GPT-3.5 + supervised training on dialogue + reinforcement learning with human feedback. The exact details are not available.



Following LLMs | Latest Entries and Developments

Competitor to GPT-family (65B), free for researchers.

Bard AI

Google's response to ChatGPT (137B)

Take LLaMA (free) and exploit GPT3.5 for training (7B)
(costed 600 $ ☺)

GPT-4 comes out, much better than 3.5 (?B)

Dolly

Open Source ChatGPT-like, GPT-J (6B) + Alpaca dataset

AutoGPT

Towards AGI, let GPT-4 access files, plugins, the internet, and **prompt itself iteratively!!!**

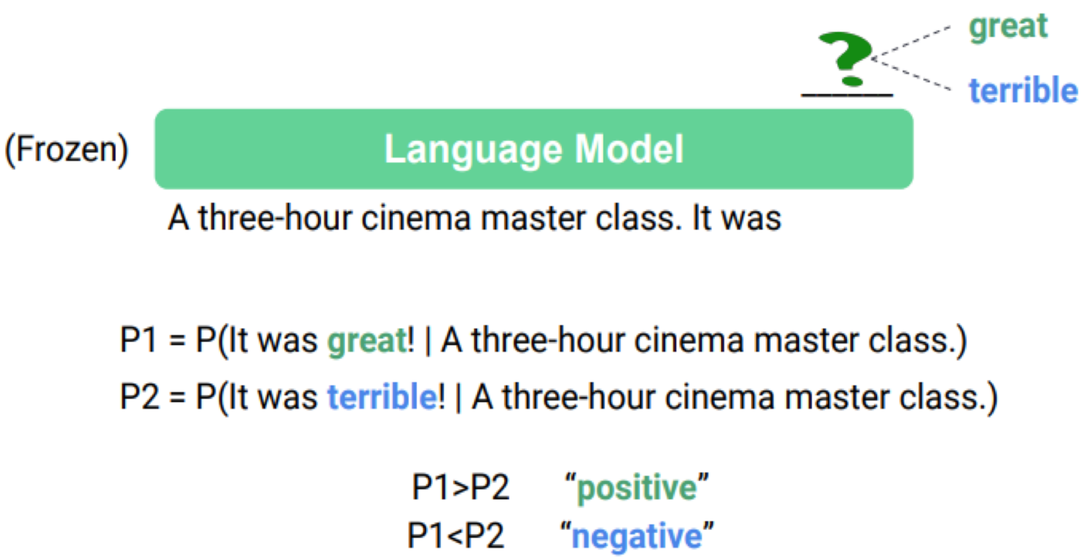
Trends

- Attention as workhorse stays
- Models will keep scaling
- training on larger, higher quality data instead of keep adding parameters
- Architecture details seem more and more irrelevant. The important factors seem to be:
  - Broad architecture type (e.g. BERT vs GPT)
  - Training objective
  - Fine-tuning techniques
  - Data quality and quantity
- Multi-modality, multi-linguality, and eco-friendliness
- NLP practitioners will use pre-trained LLMs instead of building and training from scratch.

Research Directions

- Prompting, auto-prompting
- Fine-tuning techniques
- Post-hoc explainability
- Robustness and adversarial attacks/defences
- Model Distillation
- Low-resource languages/tasks
- Augmentation and integration of knowledge bases

- Prompting: make it possible for downstream tasks to take the same format as the pre-training objectives (language modeling) by prepending some text before the test input



- idea proven effective in GPT-3
- Requires no new parameters nor retraining existing parameters

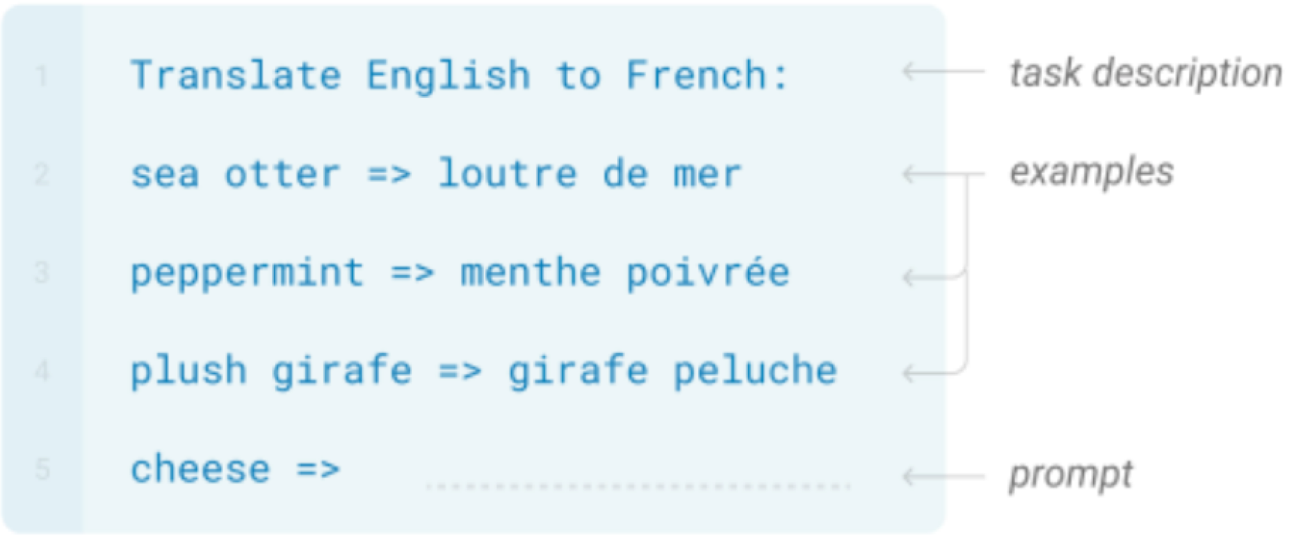
LM Prompting | Zero-Shot

- Simple prompting corresponds to zero-shot learning. The model predicts the answer directly (task description is not necessary).

- If add one example, then it is one-shot learning

LM Prompting | In-Context Learning

- Prompting that contains demonstrations (i.e. examples) of the task to be performed: in-context learning



LM Prompting | Terminology

- Pattern: A function that maps an input to the text (a.k.a. template for x)
  - Example: f(<x>) = “Review: <x>”
- Verbalizer: A function that maps a label to the text (a.k.a. template for y)
  - Example : v(<y>) = “Sentiment: <y>”

Review: An effortlessly accomplished and richly resonant work.	Sentiment: positive
Review: A mostly tired retread of several other mob tales.	Sentiment: negative
Review: A three-hour cinema master class.	Sentiment: ________

LM Prompting | Patterns & Verbalizers

- Picking suitable patterns and verbalizers is an active field of research
  - Part of prompt engineering (includes hand-crafted, gradient- or heuristic based prompts)

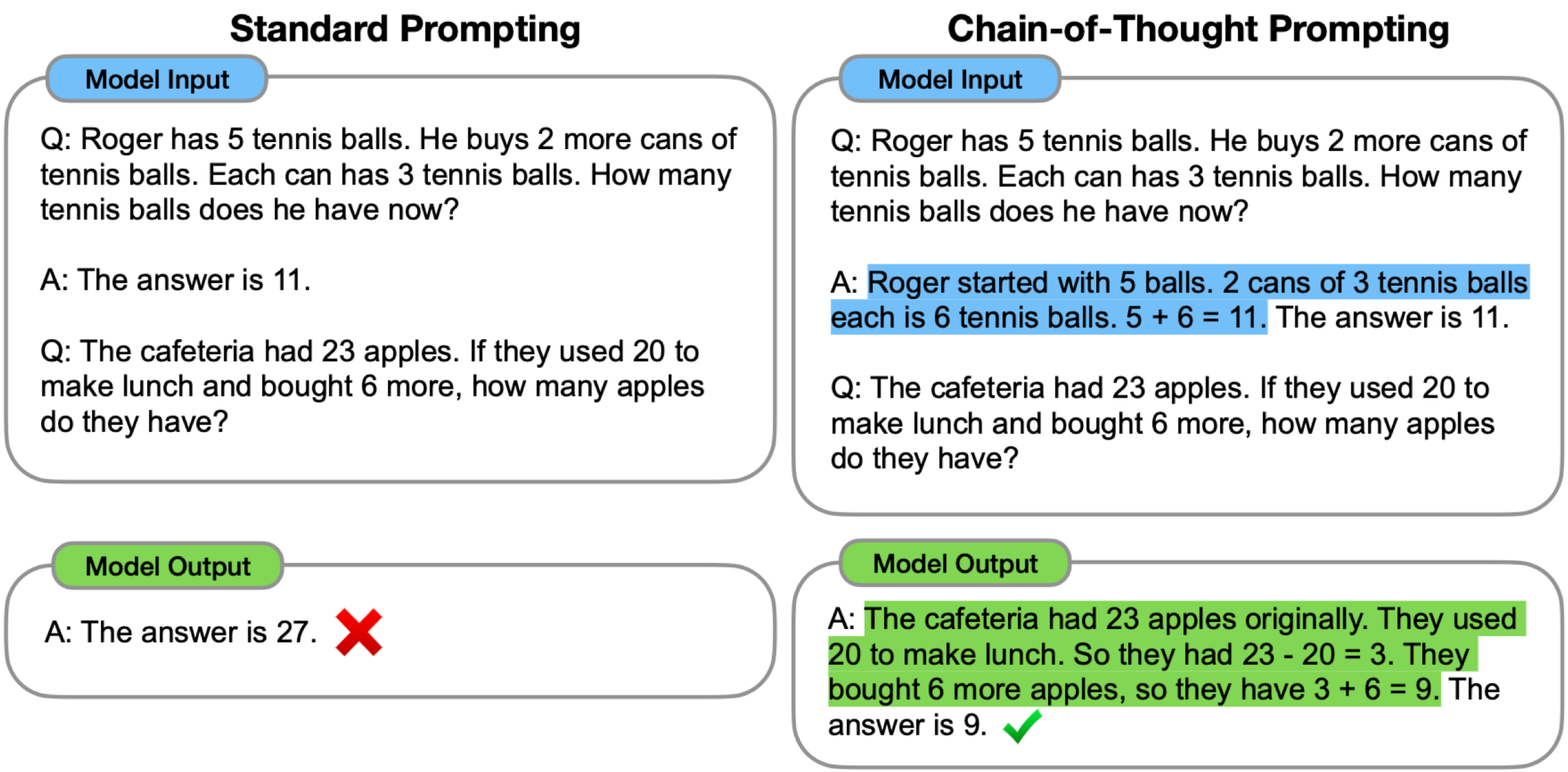
Test data: (x, y) **Train data:** $(x_1, y_1, \cdots, x_k, y_k)$ **Pattern:** f **Verbalizer:** v

Zero-shot prompting: $\mathrm{argmax}_{y \in \mathcal{Y}} P_{\mathrm{LM}}(v(y) | f(x))$

In-context learning: $\mathrm{argmax}_{y \in \mathcal{Y}} P_{\mathrm{LM}}(v(y) | \underbrace{f(x_1), v(y_1), \cdots, f(x_k), v(y_k)}_{\text{Demonstrations}}, f(x))$

Tasks descriptions as Inputs | Chain of Thought

- Providing also more instance-level details can elicit multi-step reasoning
- Unfortunately still no guarantees about reasoning correctness
- Augmenting prompts outside of few-shot setting is challenging

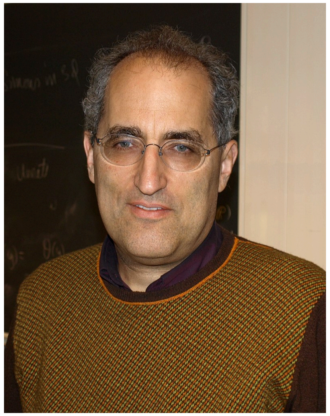


LM Prompting

- LLMs are able to perform a large variety of tasks in a few-shot setting
- Choosing a prompt is non-trivial since LLMs exhibit large variance over different patterns and verbalizers
- (=highly dependent on choice, order and term frequency)
- Uncertain how and why in-context learning works exactly

AGI, Humans, and the Future…








References

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I., (2017). Attention is all you need. In Advances in Neural Information Processing Systems (pp. 5998-6008) and arXiv:1706.03762v5

[2] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[3] Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

[4] RAFFEL, Colin, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 2020, 21. Jg., Nr. 1, S. 5485-5551.

[5] BORGEAUD, Sebastian, et al. Improving language models by retrieving from trillions of tokens. In: *International conference on machine learning*. PMLR, 2022. S. 2206-2240.

[6] ROMBACH, Robin, et al. High-resolution image synthesis with latent diffusion models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022. S. 10684-10695.

[7] AMATRIAIN, Xavier. Transformer models: an introduction and catalog. *arXiv preprint arXiv:2302.07730*, 2023.

[8] Kaplan, Jared, et al. "Scaling laws for neural language models." *arXiv preprint arXiv:2001.08361* (2020).

[9] Richard Socher et al: “CS224n: Natural Language Processing with Deep Learning”, Lecture Materials http://web.stanford.edu/class/cs224n/ (URL, May 2018), 2018

References

[10] Xiao, H. (2023) : Auto-GPT Unmasked: The Hype and Hard Truths of Its Production Pitfalls https://jina.ai/news/auto-gpt-unmasked-hype-hard-truths-production-pitfalls/ (URL, May 2023

[11] Belatgy et al (2022): Zero- and Few-Shot NLP with Pretrained Language Models - ACL Tutorial 2022

[12] Schick, T., Schütze, H. (2020) : Exploiting Cloze Questions for Few Shot Text Classification and Natural Language Inference https://arxiv.org/abs/2001.07676

[13] Liu et al (2022): Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning https://arxiv.org/abs/2205.05638

[14] Sanh V. (2022): Multitask Prompted Training Enables Zero-Shot Task Generalization https://arxiv.org/abs/2110.08207