# LANGUAGE & AI: LARGE LANGUAGE MODELS

Dr. Chris Emmery
Department of Cognitive Science & AI
Tilburg University

@cmry • @cmry • cmry.github.io

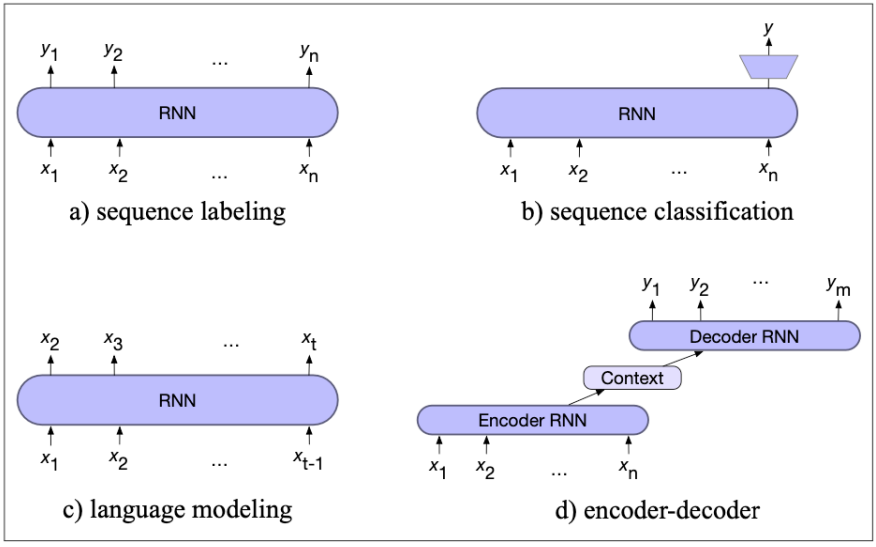
# ♻️ RECAP PREVIOUS LECTURE

- We looked at Recurrent Models and Transformers.
- We discussed neural language models.
- We discussed several Transformer components: self-attention, transformer blocks, multi-head attention, and positional encodings.

> *Today, we discuss Auto-Regressive Language Models, and how they became Large.*

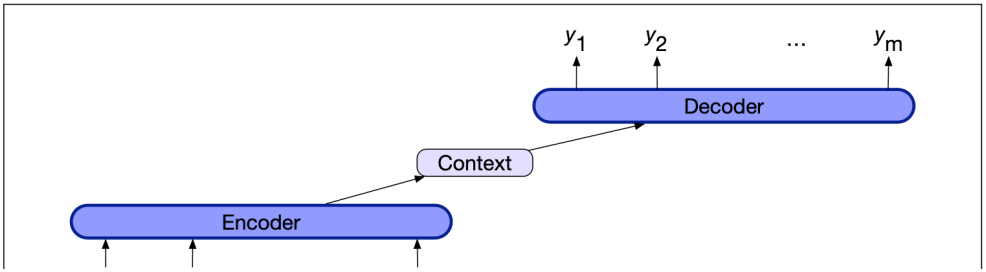
# 👀 ATTENTION

# ◀️ BACKING UP A BIT

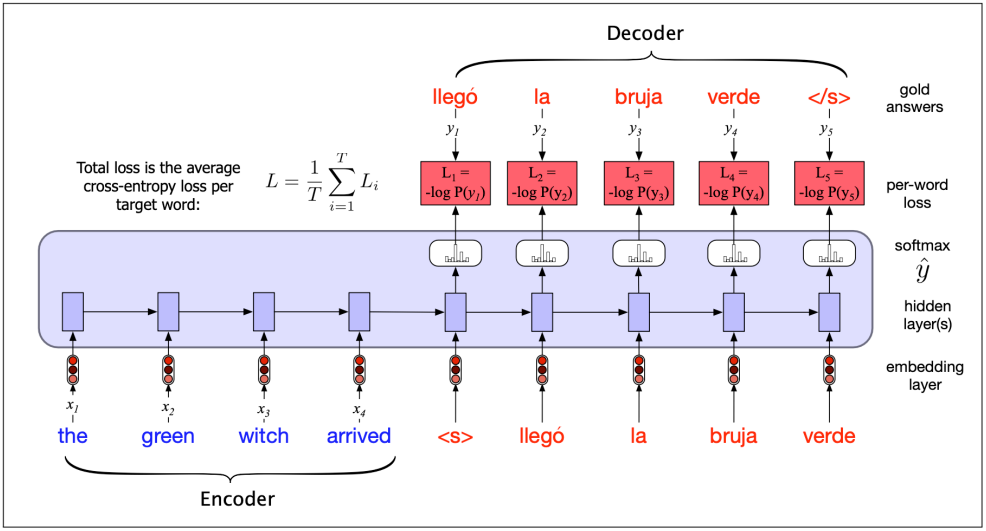


# ENCODER-DECODER MODELS

- Also called sequence-to-sequence (or seq2seq) models (as base can have RNN, transformer, etc.):
  - The **encoder** embeds and represents the input.
  - (Usually) last state is the **context vector**.
  - The **decoder** is fed this context vector and iteratively maps these to outputs.
- Map inputs to outputs of arbitrary length (can be thought of as lossy compression / decompression).



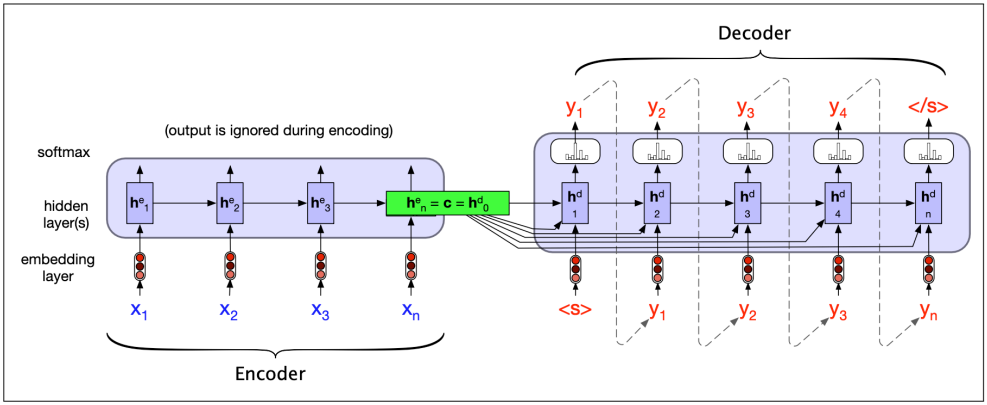
# MACHINE TRANSLATION



# 🔄 FORMALLY

$$\mathbf{c} = \mathbf{h}_n^e$$
$$\mathbf{h}_0^d = \mathbf{c}$$
$$\mathbf{h}_t^d = g\left(\hat{y_{t-1}}, \mathbf{h}_{t-1}^d, \mathbf{c}\right)$$

$$\mathbf{z}_t = f\left(\mathbf{h}_t^d\right)$$
$$y_t = \text{softmax}(\mathbf{z}_t)$$
$$\hat{y_t} = \text{argmax}_{w \in V} P\left(w \mid x, y_1 \ldots y_{t-1}\right)$$



## ⚠️ SOME LIMITATIONS

- Context vector is a bottleneck.
- The further the time steps go, the less context vector matters.
- Often required pretty convoluted architectures: multiple layers of bi-directional LSTMs was SOTA.

# !! WHY NOT JUST PAY ATTENTION?

- Rather than trying to 'cram' all states into $\mathbf{c} = \mathbf{h}_n^e$, attend to the ALL of the input $f(\mathbf{h}_1^e \ldots \mathbf{h}_n^e)$!
- We covered the gist in the previous lecture:

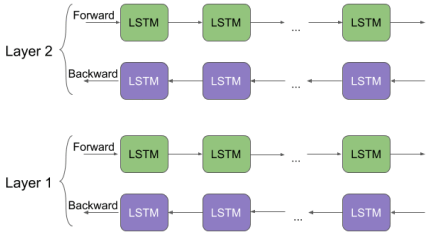
$$\text{score}\left(\mathbf{h}_{i-1}^d, \mathbf{h}_j^e\right) = \mathbf{h}_{i-1}^d \cdot \mathbf{h}_j^e \qquad \alpha_{ij} = \text{softmax}\left(\text{score}\left(\mathbf{h}_{i-1}^d, \mathbf{h}_j^e\right) \forall j \in e\right)$$

$$\mathbf{c}_i = \sum_j \alpha_{ij} \mathbf{h}_j^e \qquad = \frac{\exp\left(\text{score}\left(\mathbf{h}_{i-1}^d, \mathbf{h}_j^e\right)\right)}{\sum_k \exp\left(\text{score}\left(\mathbf{h}_{i-1}^d, \mathbf{h}_k^e\right)\right)}$$

- We can also train a mapping $W_s$ (i.e., matrix of weights)!

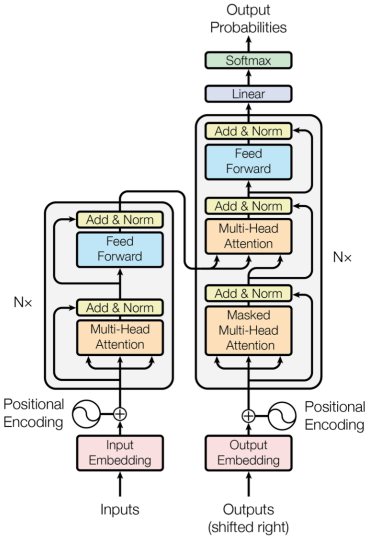
# 🍿 MEANWHILE: CONTEXTUALITY AND TRANSFER

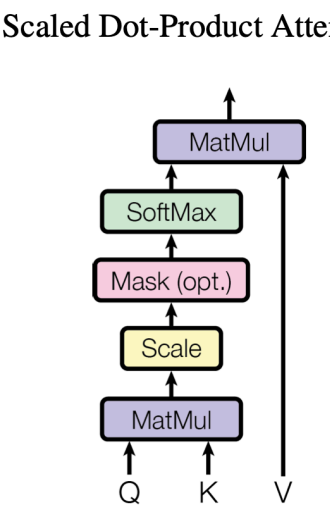
- ELMo: Contextualize embeddings. Pre-train on many different datasets/languages.
- ULMFiT: use different tasks (LM general, LM target, CLF target) and unfreezing.

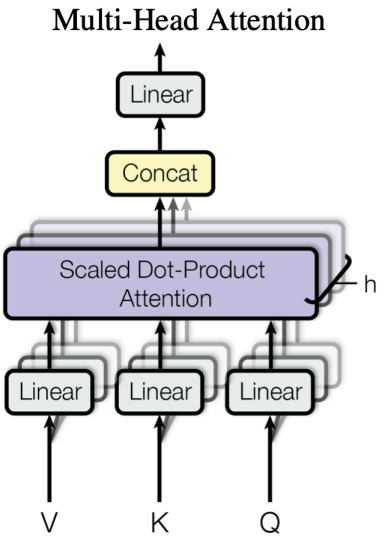



> *Both improve performance if used in models for other tasks: **transfer learning**!*

# ✋ ATTENTION IS ALL YOU NEED



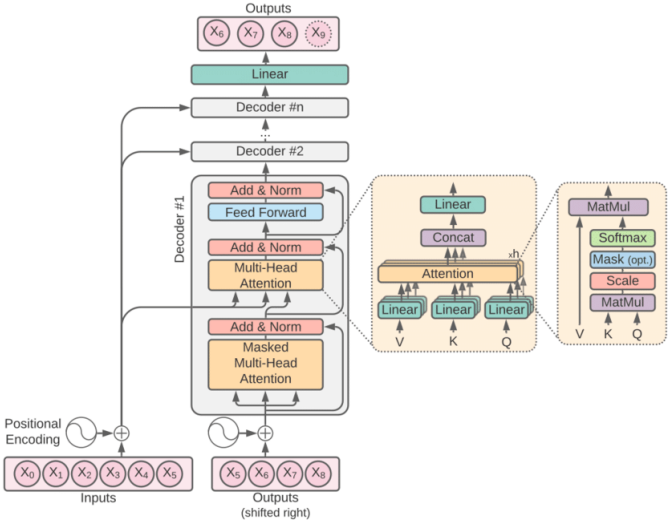



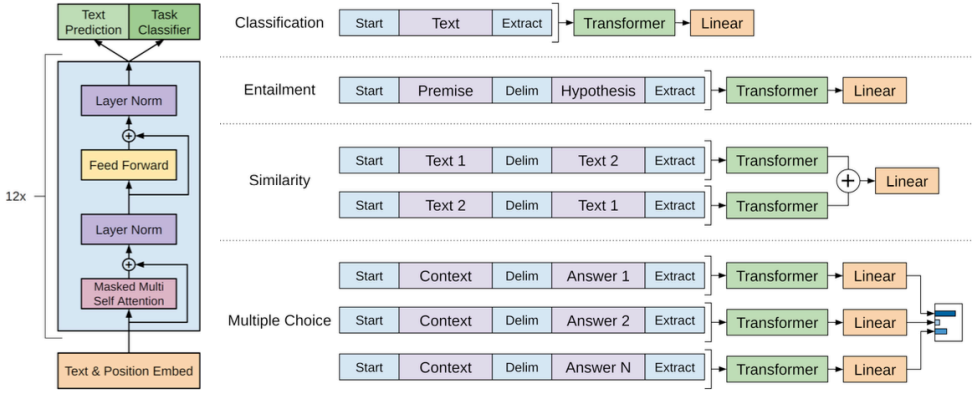

*Paper by Vaswani et al., 2017.*

# 🤖 RISE OF THE TRANSFORMERS

# 🎨 TRANSFORMERS: MIX AND MATCH ENCODER/DECODER



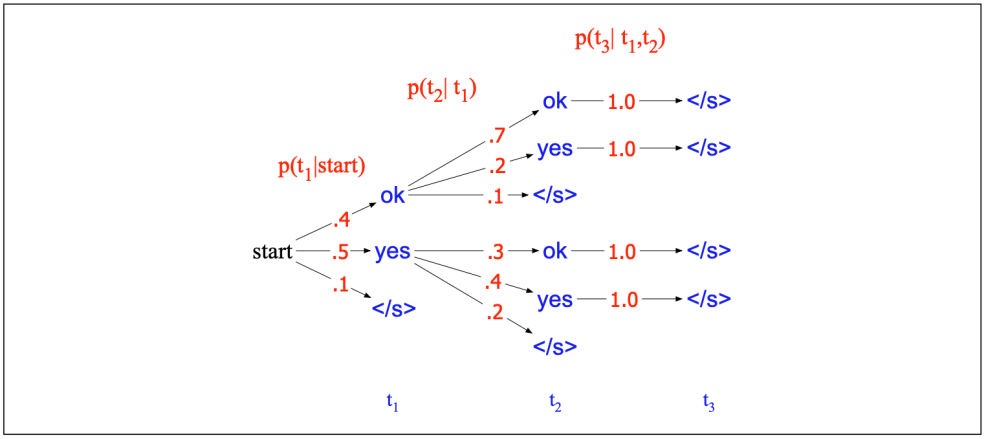
# 🎛️ GPT: GENERATIVE PRE-TRAINING (DISCRIMINATIVE FINE-TUNING)




> *Basis for v2, 3, etc. Just more data, more parameters, and some tricks (prompting, sparse attention).*

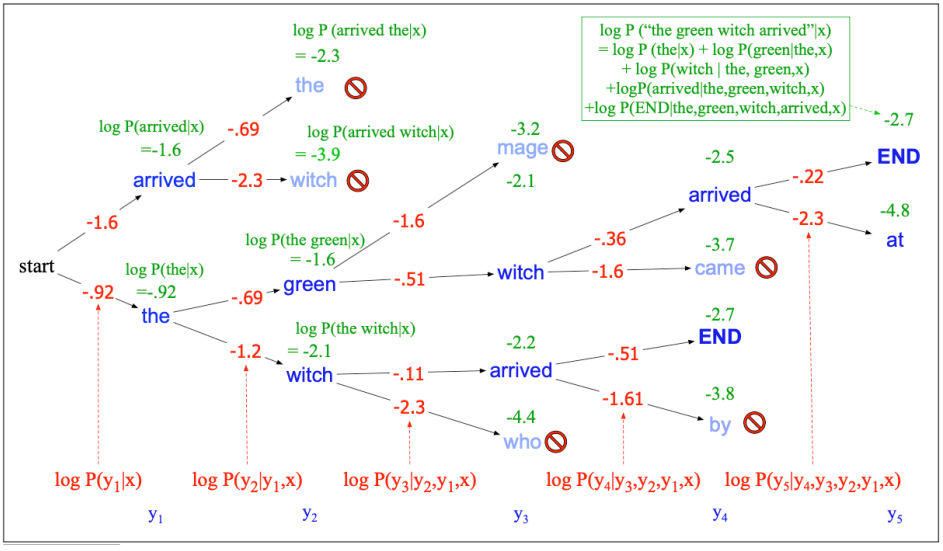
# 🚰 LANGUAGE MODEL REFRESHER

- Loss is how well softmax matches the one-hot vector of the true word.
- Won't always be correct; e.g., start of sentence tokens (<s>) provide no context. In that case, it's about modeling which words likely start a sentence.
- Always sampling the most probable word (**greedy decoding**) has some issues:



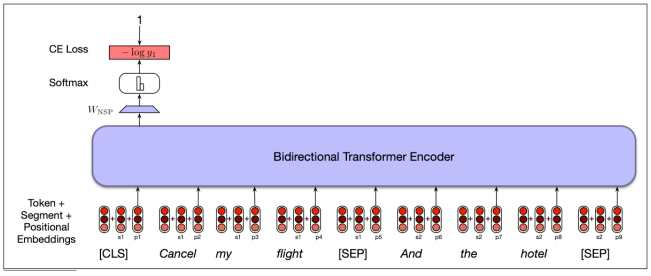
# 🔦 BEAM SEARCH

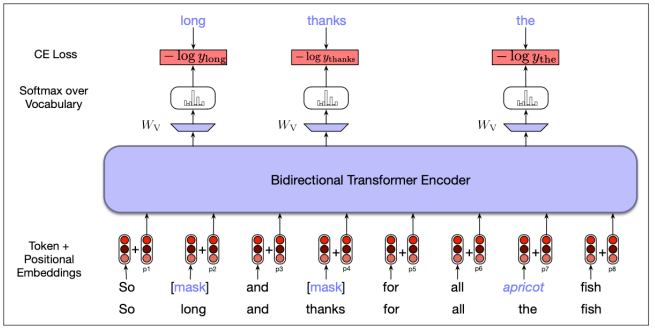
- Choose $k$ candidates to 'remember'.
- Per step, check $k * V$, keep $k$ most likely sequences (chain rule; product of conditional probabilities, i.e. sum of logs).



# BERT

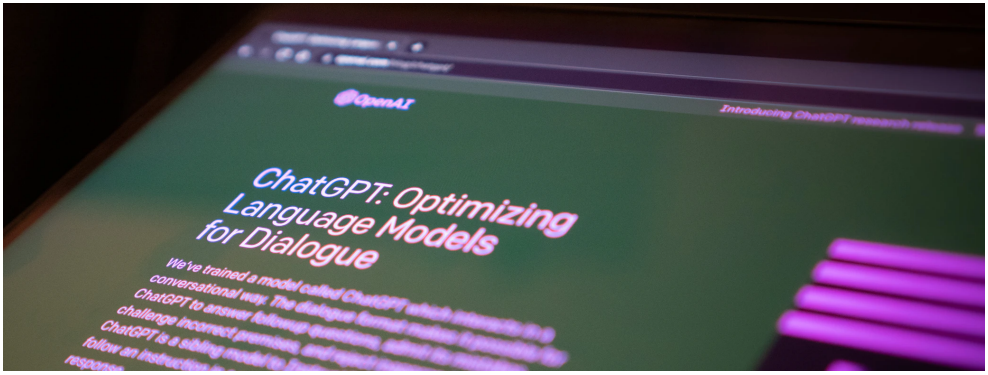
- Only uses encoder part!
- Masked LM (some w2v and adversarial link).
- Next sentence prediction.




*Fine-tune through adding classification 'heads' (basically just FFNN or LR).*

# LLMS AND PROMPT MANIA



# 🍋 HOW DID OUR LANGUAGE MODELS BECOME LARGE?

- More data.
- More parameters.
- More compute.

> *That's it. Really. Optimizations aside, they're all transformers.*

# 🦄 WHAT'S SO SPECIAL ABOUT CHATGPT?

1. Humans 🤷
2. InstructGPT

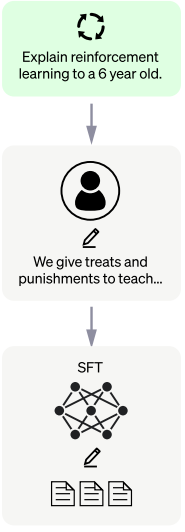
Step 1

**Collect demonstration data and train a supervised policy.**

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3.5 with supervised learning.



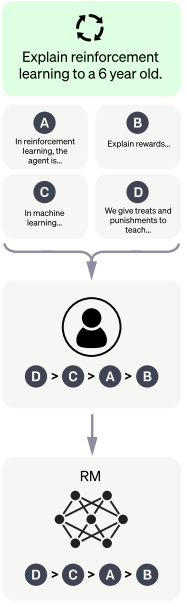

Step 2

**Collect comparison data and train a reward model.**

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.




Step 3

**Optimize a policy against the reward model using the PPO reinforcement learning algorithm.**

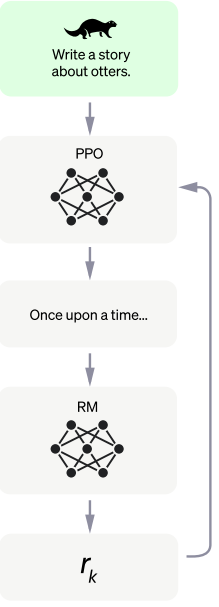
A new prompt is sampled from the dataset.

The PPO model is initialized from the supervised policy.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.




*GPT research uncovered prompt usefulness.*

## 🪦 THE DEATH OF NLP (OR IS IT?)

- Most research labs cannot train LLMs.
- Most LLMs are commercial; even fine-tuning (GPT Store).
- Prompts rule the world.

# ⭕ PROMPT ENGINEERING: ZERO-SHOT

```
Classify the text into neutral, negative or positive.
Text: I think the vacation is okay.
Sentiment:
```

# 🎯 PROMPT ENGINEERING: FEW-SHOT

```
A "whatpu" is a small, furry animal native to Tanzania. An example of a sente
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 se
the word farduddle is:
```

```
This is awesome! // Negative
This is bad! // Positive
Wow that movie was rad! // Positive
What a horrible show! //
```

# 🔗 PROMPT ENGINEERING: CHAIN OF THOUGHT

```
The odd numbers in this group add up to an even number: 4, 8, 9, 15, 12, 2, 1
A: Adding all the odd numbers (9, 15, 1) gives 25. The answer is False.
The odd numbers in this group add up to an even number: 17,  10, 19, 4, 8, 12
A: Adding all the odd numbers (17, 19) gives 36. The answer is True.
The odd numbers in this group add up to an even number: 16,  11, 14, 4, 8, 13
A: Adding all the odd numbers (11, 13) gives 24. The answer is True.
The odd numbers in this group add up to an even number: 17,  9, 10, 12, 13, 4
A: Adding all the odd numbers (17, 9, 13) gives 39. The answer is False.
The odd numbers in this group add up to an even number: 15, 32, 5, 13, 82, 7,
A:
```

# 🪞 PROMPT ENGINEERING: SELF-CONSISTENCY

```
When I was 6 my sister was half my age. Now
I’m 70 how old is my sister?
```

```
35
```

```
Q: There are 15 trees in the grove. Grove workers will plant trees in the gro
there will be 21 trees. How many trees did the grove workers plant today?
A: We start with 15 trees. Later we have 21 trees. The difference must be the
So, they must have planted 21 - 15 = 6 trees. The answer is 6.
Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many ca
A: There are 3 cars in the parking lot already. 2 more arrive. Now there are
Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pie
A: Leah had 32 chocolates and Leah’s sister had 42. That means there were ori
chocolates. 35 have been eaten. So in total they still have 74 - 35 = 39 choc
Q: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lol
did Jason give to Denny?
A: Jason had 20 lollipops. Since he only has 12 now, he must have given the r
lollipops he has given to Denny must have been 20 - 12 = 8 lollipops. The ans
Q: Shawn has five toys. For Christmas, he got two toys each from his mom and
he have now?
A: He has 5 toys. He got 2 from mom, so after that he has 5 + 2 = 7 toys. The
in total he has 7 + 2 = 9 toys. The answer is 9.
Q: There were nine computers in the server room. Five more computers were ins
monday to thursday. How many computers are now in the server room?
A: There are 4 days from monday to thursday. 5 computers were added each day.
20 computers were added. There were 9 computers in the beginning, so now ther
The answer is 29.
Q: Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday
```

# 📚 PROMPT ENGINEERING: GENERATED KNOWLEDGE PROMPTING

```
Part of golf is trying to get a higher point total than others. Yes or No?
```

```
Yes.
```

```
Input: Greece is larger than mexico.
Knowledge: Greece is approximately 131,957 sq km, while Mexico is approximate
Input: Glasses always fog up.
Knowledge: Condensation occurs on eyeglass lenses when water vapor from your
Input: A fish is capable of thinking.
Knowledge: Fish are more intelligent than they appear. In many areas, such as
Input: A common effect of smoking lots of cigarettes in one’s lifetime is a h
Knowledge: Those who consistently averaged less than one cigarette per day ov
Input: A rock is the same size as a pebble.
Knowledge: A pebble is a clast of rock with a particle size of 4 to 64 millim
Input: Part of golf is trying to get a higher point total than others.
Knowledge:
```

```
The objective of golf is to play a set of holes in the least number of stroke
```

```
Question: Part of golf is trying to get a higher point total than others. Yes
Knowledge: The objective of golf is to play a set of holes in the least numbe
Explain and Answer:
```

```
No, the objective of golf is not to get a higher point total than others. Rat
```

# 🍖 PROMPT ENGINEERING: GO HAM

- Retrieve things from the Internet.
- Make LLMs correct outputs of LLMs.
- Make LLMs rank outputs of LLMs.
- Use multiple LLMs for uncertainty rating.
- Run things through actual systems (e.g. Python interpeter.)
- Etc.