Key concepts

- **Transformers**
  - encoder: outputs context-sensitive embeddings
  - decoder: predicts the next word
    - BERT is an encoder only
    - ChatGPT is a decoder only (outputting one word at a time)
    - translation tasks typically use encoder-decoder architectures.
- **Some history**
  - machine translation included encoders and decoders that were RNNs. Then attention came along and was applied to these, and then transformers came along and *replaced* these.
- **Attention**
  - Add the vectors in the context to the current vector in proportion to their similarity. The output is a “weighted blend” of the vectors in the context

Key concepts

- **Types of learning**:
  - Error-driven learning: reduce the error between what you get and what you want
  - Reinforcement learning: trial-and-error to get closer to a correctly chosen outcome (e.g. humans rank order the possible responses).
- **Training**
  - **pre-training** - predict next word (ChatGPT) or guess masked words in input string (BERT). Learning is *error-driven*.
  - **fine-tuning** - training is now task-specific (e.g. dialogue generation). May be error-driven, but could also include *reinforcement learning.* Error-driven learning is good for learning the language-at-large. Reinforcement learning is good for learning more pragmatic usage (which response would be better?)
  - **deployment**: user responses are interpreted using various metrics and used to further train the model (using reinforcement learning) and various “engagement metrics”.

The future of Chatbots

- ***Mixtral (12/23)***: **Mixture of Experts**:
  - 8 experts each with 7B parameters (ChatGPT has 175B) - 13B parameters used at a time
  - A “router” specific to each feed-forward network (FFN) sends the vectors to just two of the experts at a time. It learns which to send it to. The FFNs sum their outputs (c.f. summing in *attention*)
    - Open source, small, can be deployed on local machines.
- *Gerry’s analogy*: your primary care provider refers you to experts, so no one person need be expert in everything. Hence fewer parameters involved (in the AI system). But there is communication between them (think MyChart…).

The future of Chatbots

- ***Mamba (12/23)***:
  - "*This should not work*" Forrest Davis…
  - No attention or multi-layer FFNs
  - Selective State Spaces (SSS)
    - each recurrent unit in an RNN with LSTMs or GRUs receives as input the previous "hidden state" (the entire activation pattern of the layer being copied)
    - in attention models, each token has an influence (however small) on each other token.
    - In SSS, the hidden state due to a low-informative token is discarded, and only those of high-informative tokens are maintained.
      - computationally more efficient than implementing attention for each token based on each other token in the transformer's context. Apparently.

The future of Chatbots

- ***Size matters***
  - The main difference between ChatGPT-3 and ChatGPT-2 is size (parameters / decoder blocks) and size (training tokens).
  - Size = Time = Money & Energy
  - ChatGPT-3 training cost approx. $5M.
  - Energy consumption: 1000 US households for a year.
  - Emissions….
    - Server / data center materials and associated costs
    - Engineers, system admins, tuner/trainers, etc.
    - "*the cost of training ChatGPT-3 extends beyond monetary expenses—it affects our environment and energy resources. As these models become more popular, understanding and mitigating their impact will be crucial*" (MS CoPilot)
- Size also matters in respect of
  - mobility (e.g. running the chatbot on your laptop or phone)
  - ease/speed of training (specialist domains and associated fine-tuning)

The future of Chatbots

- ***Microsoft CoPilot, 2/23 (formerly Bing)***:
  - based on GPT4.
  - does a web search and concatenates the results with the user's actual prompt
  - more useful than ChatGPT that does not search but just uses its language model (pre Jan 2022)

The future of …

- ***Microsoft Orca 2 (11/23):***
  - teaching the LLM to reason
  - give it multiple answers reflecting different strategies
  - Based on Meta's Llama-2
    - small, with ***only*** 70B parameters.
  - Outperforms ChatGPT and most LLMs except (on 2/8 benchmarks, another Llama-2 variant.

Instructions:
In this task, you're given a short story of five sentences written in natural language. However, the order of the given story is not correct. Your job is to return the correct order for the given five sentences to create a coherent short story with the new order that has the correct flow. Generate your answer using the number of sentences in the correct order, such as '23415'.
Sentence1: He was scared until he found out she was ok.
Sentence2: He usually saw really bad accidents and it always unnerved him.
Sentence3: One day Bruce was called to an accident scene.
Sentence4: Bruce was a police officer who handled traffic accidents.
Sentence5: He immediately recognized his friend Tanya's car.

Answer 1:
43152

Answer 2:
Step 1: Identify the starting sentence.
Sentence4: Bruce was a police officer who handled traffic accidents.
Step 2: Identify the next event.
Sentence3: One day Bruce was called to an accident scene.
Step 3: Identify the following event.
Sentence5: He immediately recognized his friend Tanya's car.
Step 4: Identify the next event.
Sentence2: He usually saw really bad accidents and it always unnerved him.
Step 5: Identify the concluding sentence.
Sentence1: He was scared until he found out she was ok.
The correct order is: 43521

Answer 3:
Answer: 43152
Explanation: The correct order of the sentences should be as follows: Bruce was a police officer who handled traffic accidents (4). He usually saw really bad accidents and it always unnerved him (2). One day Bruce was called to an accident scene (3). He immediately recognized his friend Tanya's car (5). He was scared until he found out she was ok (1). This order creates a coherent short story with the correct flow.

Answer 4:
1. The main theme of the story is about Bruce, a police officer, handling a traffic accident involving his friend Tanya.
2. Cause and effect relationships:
   - Bruce is a police officer who handles traffic accidents (cause) → he usually sees bad accidents (effect)
   - Bruce is called to an accident scene (cause) → he recognizes Tanya's car (effect) → he is scared until he finds out she is ok (effect)
3. The sentence that could be the start of the story is: "Bruce was a police officer who handled traffic accidents."
4. Rearranging the sentences in the correct order:
   (a) Bruce was a police officer who handled traffic accidents.
   (b) He usually saw really bad accidents and it always unnerved him.
   (c) One day Bruce was called to an accident scene.
   (d) He immediately recognized his friend Tanya's car.
   (e) He was scared until he found out she was ok.
5. Final answer: 42351