## Week 1 – Language Modelling

1. In the context of GPT, what is a Transformer?
   1. A neural network architecture. (Transformers are a specialized type of neural network architecture that GPT uses to model language.)
2. Does the probability provided by a language model depend on a reference corpus?
   1. Yes, there is no single “correct” answer to the probability of specific text because language differs among people, culture, context, and time. (Due to the variety of language, we need to measure the probability of text with respect to a set of reference text.)
3. What type of language model predicts probabilities in the form of P("GPT"|"I","am","learning","about")?
   1. Causal Language Models. (Causal language models predict the final word in a sequence.)
4. What is a key advantage of using a lower 'n' value in an n-gram language model over a higher 'n' value?
   1. Lower 'n' values result in less data sparsity. (This is because there are fewer unique n-grams, which makes it more likely that an n-gram in the test data was also in the training data.)
5. Consider the sentence "The cat is on the mat". If we were to compute the probability of the sentence using a bigram model, how would we compute it?
   1. P("The"|"<BOS>") \* P("cat"|"The") \* P("is"|"cat") \* P("on"|"is") \* P("the"|"on") \* P("mat"|"the") \* P("<EOS>"|"mat")
6. What is NOT a type of n-gram?
   1. Quadrigram. (Quadrigram is not a standard term used in n-gram models. For n equals 4, the term used is four-gram.)
7. What is the effect of increasing 'n' in an n-gram language model?
   1. The model captures more context but requires more computational resources.
8. Why might a tri-gram language model not be suitable for a small corpus of text?
   1. There might not be enough examples of each possible tri-gram to accurately estimate probabilities.
9. Why can measuring the perplexity of a model on its training data be misleading?
   1. It will give an under-estimate of how well the language model generalizes to unseen data. (Since the probabilities were generated based on the training data, you will usually have a lower perplexity on your training data than on unseen data.)
10. What is the best possible value for perplexity?
    1. 1. (A perplexity of one implies that the language model perfectly predicted the sample.)
11. To test how good a language model was trained, your colleague suggests making up a few questions to see how well it answers them. Why is this extrinsic evaluation setup problematic?
    1. A few samples is likely an insufficient amount of data to draw meaningful insights from.
    2. Samples that you make up might not represent a realistic application of your language model.
    3. The samples you make up might be in the training data.
12. Let’s say you compute the perplexity on a held-out test set for two models, a bigram and a trigram language model. The bigram model has a perplexity of 45.12 and the trigram model has a perplexity of 29.64. What can you infer about these two models?
    1. The trigram model is a better fit for the test data than the bigram model.
13. Why is smoothing usually necessary when measuring perplexity of n-gram language models?
    1. The sparsity of language means that without smoothing, many of the probabilities will be zero, resulting in undefined perplexity. (Not only do zero probabilities cause problems when calculating perplexity, but they do not represent the reality of natural language variation.)
14. When a context is encountered that was not seen during training, how do n-gram language models with smoothing estimate token probabilities?
    1. A uniform probability distribution is applied across all tokens. (All tokens are considered equally likely in this case.)
15. How are tokens that were not encountered during training often handled in language models?
    1. They are replaced with a single special “out-of-vocabulary” (OOV) token.
16. Why is perplexity used for intrinsic evaluation of language models?
    1. Perplexity measures how well a language model predicts a sample corpus. (Perplexity measures the probability of a sequence, which directly measures how well the model represents language.)
17. What statement about generating out-of-vocabulary (OOV) tokens is true?
    1. Without smoothing, no generation strategy can generate out-of-vocabulary tokens from an n-gram language model. (Since OOV tokens are given a probability of 0 by n-gram language models.)
18. When considering the trade-offs between different generation strategies, what is a key advantage of using beam search over greedy generation and sampling?
    1. Beam search can generate the more probable sequence.
19. Which method among greedy generation, sampling and beam search is more likely to generate a diverse but less coherent text sequence?
    1. Sampling. (By sampling randomly from the probability distribution, the generated sequences tend to be more diverse and interesting.)
20. In what scenario might greedy generation be a better choice than beam search or sampling for text generation?
    1. When you want to generate more probable sequences
21. What is the purpose of clipping the probability distribution to the “top k” tokens when sampling?
    1. It reduces the chance of generating very low-probability sequences. (Applying a “top k” filter before sampling eliminates very low-probability tokens.)
22. How does the greedy generation method differ from beam search and sampling in terms of text sequence generation?
    1. Greedy generation chooses the word with the highest probability at each step, while beam search maintains a set of most probable sequences, and sampling generates a diverse sequence by introducing randomness.
23. Why might you want to apply smoothing when generating from an n-gram language model?
    1. Smoothing will ensure that every context has some tokens with non-zero probability.
    2. Smoothing will give some probability to sequences not seen during training, improving the diversity of generation.
    3. Smoothing will give some probability to the end-of-sequence token in all contexts.
24. You notice that a greedy generation over a unigram model always repeats the same token indefinitely. What’s happening?
    1. The context added at each step of generation has no effect on the probabilities from a unigram model, so the same token always repeats. (The probabilities in a unigram model are identical, regardless of the context.)

## Week 2 – Transformers and GPT

1. Given a GPT tokenizer, what does a token without Ġ at the beginning mean?
   1. The token is a subword token inside a word and does not start it. (The Ġ character denotes that a token starts a word and has whitespace before it.)
2. Which encoding scheme is usually used for representing text in deep learning models like GPT?
   1. Tokenization. (Tokenization is the process of converting text into tokens, which can be processed by deep learning models.)
3. Why do some language models use a special padding token?
   1. To deal with multiple inputs of different lengths. (Language models often process multiple sequences together which may be of different lengths. To enable this and indicate that some of the sequence is not important, padding special tokens may be added to the end of shorter sequence to make them the same length as the longest sequence.)
4. What is the process of splitting a text into smaller units called?
   1. Tokenization. (Tokenization is the process of splitting a text into smaller units, such as words or phrases.)
5. What would a tokenizer used for Transformer models do to a word it’s not seen before?
   1. Split it into subword tokens. (By splitting it into subword tokens that it has seen before, it may be able to represent it well enough that the Transformer is able to partially infer its meaning.)
6. How do subword tokenizers decide how to split words up?
   1. Learning patterns of words from a large corpus of text. (Large corpuses of text (e.g. text from Wikipedia) are used alongside algorithms such as BPE encoding to learn frequency of common words and appropriate ways to split uncommon words.)
7. What are the advantages of subword tokenization? (select all that apply)
   1. Smaller vocab. (By breaking uncommon words in smaller parts, the required vocabulary to encode a large corpus is typically smaller.)
   2. Deals well with new words. (New words which have not been seen before can be broken down in subword tokens which will have been seen before. These smaller parts may encode enough meaning to give the language model some indication of this new words meaning.)
8. Given these token IDs and vocabulary, what would the decoded text be from a GPT tokenizer?

Token IDs = [2, 5, 9, 1, 10]

Vocab = {“Ġbr”:0, “Ġeter”:1, “Ġhope”:2, “Ġings”:3, “Ġnal”:4, “Ġspr”:5, “br”:6, “eter”:7, “hope”:8, “ings”:9, “nal”:10, “spr”:11 }

* 1. “ hope springs eternal”. (The Ġ denotes tokens that start new words and have whitespace before them. Those tokens without are inside larger words.)

1. In the given text, which word should have the closest context vectors for its two appearances?

John checked his watch to decide when to watch the match. John picked up the match off on the checked tablecloth.

* 1. John. (The man’s name has the same purpose in both sentence so should be represented with similar context vectors.)

1. Why might you need to truncate an input sequence for a Transformer?
   1. Transformers have a limit on the length of their inputs. (Transformers have a maximum length that they can process and so longer sequences need to be dealt with, often by truncation down to the maximum length.)
2. The entirety of Wikipedia is enough text to train a modern language model.
   1. False. (The corpus used by modern large language models are much larger than the text from Wikipedia and typically also includes books and a lot of text scraped from websites across the internet.)
3. A particular implementation of a causal transformer model has 12 layers, 1.2 million parameters and a vocabulary size of 120. It is run on a single sequence of 10 tokens. How many output values does it generate?
   1. 1200. (It provides scores for each of the 120 tokens in the vocabulary for all 10 tokens in the input sequence.)
4. A language model that is trained on Spanish can be applied to French text.
   1. False. (A language model that has been trained on Spanish will have learned word vectors and attention mechanisms that are specific to Spanish text. It will not function correctly on language that is different from that on which it was trained.)
5. The token with the highest score will always have the highest probability after the softmax function is used.
   1. True. (The softmax function rescales all the values to be within the range of 0 to 1 and so that they all add up to 1. However, it doesn’t change the relative order of the numbers such that the highest original score will remain the highest value after the function is applied.)
6. A language model can be trained on text from multiple languages.
   1. True. (Large language models are now trained on text containing multiple languages and often also programming languages.)
7. What would the softmax of the vector [5,5,5,5] be?
   1. [0.25,0.25,0.25,0.25]. (The softmax values should all be within 0 and 1 and add up to 1. The function will also not change the order of the values.)
8. All words are treated equally when calculated the context vector for one particular word
   1. False. (Some words are more important when figuring out the meaning of each token. The self-attention mechanism enables a Transformer to pay more attention to some words.)
9. Given these probabilities of the next token below, will sampling and greedy decoding give different results?

Probabilities: {‘cat’:0, ‘dog’:0, ‘horse’:0, ‘fish’:1}

* 1. Yes. (Only one of the tokens has a non-zero probability, so sampling would always pick the same token (‘fish’).)

1. Given that you have a task that requires generating text one token at a time, which type of Transformer model would you likely choose?
   1. Causal Language Model.
2. Causal language modelling predicts a word missing in the middle of a sentence.
   1. False. (Causal language modelling (used by GPT models) predicts the next token after a sequence of tokens. Masked language modelling (used by BERT models) predicts missing tokens.)
3. How do language models generate text?
   1. One token at a time. (Language models produce a probability distribution over all possible tokens to decide which token to output next.)
4. Why does ChatGPT give a different response each time?
   1. Sampling during text generation. (Text generation methods that use sampling involve a random choice from a probability distribution of possible next tokens. This causes a text generation model to output a different result each time.)
5. What is it called when some examples of the task are included within the prompt for a language model?
   1. Few-shot learning. (Few-shot learning includes task instructions and several examples of input and expected output.)
6. Language models can be deployed without bias because they are machines.
   1. False. (While a language model is a machine and not human, it still contains biases based on the text that it was trained on. It will inherit the human biases from the humans who created the training text.)
7. Language models understand the text that they are processing.
   1. False. (Language models predict token probabilities through various mathematical operations. There is no level of understanding even if the outputted text can appear complex.)
8. What tasks was GPT pre-trained on?
   1. Causal language modelling. (Causal language modelling involves predicting the next token after a sequence of tokens and is used by Decoders such as GPT.)
9. What are some of the challenges of using large language models?
   1. Computational cost. (The computational cost in terms of monetary and environmental costs can limit who has access to these models and their impacts on the environment.)
   2. Explainability. (Language models cannot provide clear explanations of why text has been generated which limits their use in areas that needs clear explanations.)
   3. Unknown biases. (Language models contain biases based on the text used to train them. These biases may not be well understood and their impact may be had to mitigate for various applications such as automatically screening job applicants.)
   4. Factual accuracy. (Language models can often provide factually correct answers but will also generate incorrect information frequently.)
10. The energy costs of training and using a large language model are negligible.
    1. False. (Large language models require substantial compute infrastructure and therefore large electricity costs for training and general use. There are therefore serious environmental concerns about the large-scale use of these models.)
11. All biases that are contained in text corpora used for pre-training can be removed from the trained language model.
    1. False. (Any biases (such as ageism, sexism, etc) that appear in the input text will make it into the behaviour of a language model trained on that text. While there is research to mitigate these biases, it is impossible guarantee that all biases have been removed.)
12. Why might a chatbot seem to forget something discussed in the conversation much earlier?
    1. The conversation history included in a prompt is limited by the maximum input length of a transformer model. (Transformers have a limited input length and very long conversations will become too long to include all text in the prompt. This may mean that previously discussed items are not included in the input which would give the impression of the Transformer model forgetting them.)
13. Why is it a bad idea to use a French tokeniser on Spanish text?
    1. The French tokeniser will not have seen Spanish words before and may not tokenise appropriately. (As the tokenizer does not know about common Spanish words, it will likely split up words inappropriately and not provide a good tokenization of Spanish text.)
14. Given these tokens and this vocabulary, what is the output of this GPT tokenizer?

Tokens = “ she sells seashells”

Vocab = { “Ġsea”:0, “Ġsells”:1, “Ġshe”:2, “Ġshells”:3, “sea”:4, “sells”:5, “she”:6, “shells”:7}

* 1. [2, 1, 0, 7]. (The Ġ denotes the start of tokens which is true for all except “shells”.)

1. How are word vectors and parameters in a Transformer based language model learned?
   1. Gradient descent by learning from data. (The outputs of the Transformer models are moved closer towards the desired outputs using a gradient descent approach that moves adjusts the word vectors and parameters.)
2. The probability of a language modelling generating a sequence of 2 tokens is 1/16. What is the average per-token probability?
   1. ¼. (If the average per-token probability was 1/4, then the probability of a sequence of 2 tokens would be 1/4 \* 1/4 which is 1/16 which does match.)
3. What best describes a masked language model in the context of Transformer models?
   1. A model that randomly hides input tokens during training. (A masked language model randomly hides some of the input tokens during training and tries to predict them.)

## Week 3 – Applications and Implications

1. What was the model from OpenAI that preceded ChatGPT?
   1. InstructGPT
2. What is the main distinction between ChatGPT and InstructGPT?
   1. ChatGPT is designed for generating conversational responses, while InstructGPT is optimized for generating instructional content. (The main difference lies in the primary use cases of the two models: ChatGPT is intended for generating natural language conversation, while InstructGPT is designed for providing detailed responses based on instructions.)
3. Why is a pretrained LLM like GPT not suitable for use in an interactive chatbot in its default configuration?
   1. Off-the-shelf LLMs generate pre-trained responses that might not align with the desired conversation. (An off-the-shelf LLM generates pre-trained responses that might not align with the desired conversation or the context of the interaction. Fine-tuning or customizing the LLM's responses based on specific requirements is necessary to make it useful in an interactive context such as a chatbot.)
4. In the Bender and Hanna article from Scientific American, what do the authors consider to be the real threat from artificial intelligence?
   1. The spread of misinformation through AI-generated content. (The authors emphasise that the real threat from artificial intelligence lies in existing dangers, such as misinformation and harm caused by AI-generated content.)
5. What is one of the allegations made in the lawsuits filed in the USA against OpenAI (GPT) and Meta (LLaMA) regarding generative AI and copyright?
   1. The models remix copyrighted works without consent, compensation, or credit.
6. Which is the main current mitigation strategy for hallucinations in LLMs?
   1. Human-in-the-loop verification.
7. What aspect of generative AI is often a focus of regulatory efforts?
   1. Ensuring transparency and explainability of AI systems. (Regulatory efforts often focus on ensuring transparency and explainability of AI systems, which helps build trust and accountability.)
8. What is the main concern about automated AI detectors expressed in the linked Stanford blog post?
   1. They are unreliable, especially for non-native English writers.
9. What is an example of a sector where hallucinations in AI-generated content could have serious consequences?
   1. Entertainment and creative writing.
10. g the reported incident in March 2023, which is the policy now adopted by the literary magazine Clarkesworld?
    1. They do not consider submissions written, developed, or assisted by AI tools. (Clarkesworld's policy is that they will not consider submissions written, developed, or assisted by AI writing tools like ChatGPT. Attempting to submit such works may result in being banned from submitting in the future.)
11. What emotional and psychological impact have annotators reported while preparing training data for Large Language Models such as GPT?
    1. Stress, anxiety, and discomfort due to exposure to distressing or offensive material.
12. What contributes to ChatGPT's ability to simulate creative output?
    1. Its training on a diverse range of human-written content. (ChatGPT's ability to simulate creative output is based on its training on various human-written content, which enables it to mimic language patterns and generate responses that appear creative within the scope of its training.)
13. What is NOT an AI harm identified by Bender and Hanna?
    1. Existential risks and potential to wipe out humanity.
14. What should teachers and lecturers avoid when using GPT for academic purposes?
    1. Writing from scratch without any reference material.
15. What is a significant limitation of ChatGPT that is mentioned in the OpenAI blog post introducing it?
    1. Writing plausible-sounding but incorrect answers.
16. What consequence did lawyers in New York face for submitting briefs that contained content generated by ChatGPT?
    1. They were each fined for submitting misleading information. (The lawyers in New York were fined $5000 each for submitting briefs that cited non-existent court decisions generated by ChatGPT. This was due to the misleading nature of the information they presented in the case.)
17. How did the Stanford study find that AI detectors performed in classifying TOEFL essays written by non-native English students?
    1. They were unreliable, classifying over half as AI-generated
18. What was the primary focus of the open letter issued by the Center for AI Safety in May 2023?
    1. Addressing the potential risk of extinction from AI as a global priority.
19. The authors of the Bender and Hanna article believes that AI technology can effectively address gaps in education and healthcare for those in need.
    1. False. (The article argues that the deployment of AI technology can actually hurt workers and points out the negative impact on various sectors.)
20. Why is it not appropriate to use an LLM such as GPT as a search engine directly?
    1. LLMs do not in general have access to live, real-time data, which means that any response may not necessarily be accurate.
21. What is the correct definition of prompt injection?
    1. Prompt injection is a family of exploits where malicious users manipulate a model's behaviour by feeding it crafted instructions. (Prompt injection is a family of exploits where malicious users manipulate a machine learning model's behavior by providing it with crafted instructions, often replacing the intended human-generated prompts. This can lead to security vulnerabilities and undesirable behavior in the model.)
22. What is NOT an issue with the use of automated AI detectors in an academic context?
    1. AI detectors are expensive and inefficient. (Reduced efficiency in identifying potential instances of plagiarism is not a typical risk associated with using automated AI detectors for academic writing. In fact, the use of AI detectors is generally intended to improve efficiency.)
23. As of July 2023, what is the typical advice given to students at UK universities about the use of AI tools such as GPT?
    1. Students should learn how to use AI tools effectively and acknowledge their use appropriately.
24. What is the distinction between task-based and social chatbots?
    1. Task-based chatbots are programmed to accomplish specific tasks, while social chatbots aim to engage in open-ended and human-like conversations. (The distinction between task-based and social chatbots is that task-based chatbots are programmed to accomplish specific tasks, such as setting reminders or making reservations, while social chatbots focus on engaging in open-ended and human-like conversations, often for the purpose of entertainment or companionship.)
25. What is a crucial distinction between human hallucinations and what is often called "hallucinations" in generative AI?
    1. AI hallucinations arise from incorrect data interpretation, while human hallucinations arise from incorrect sensory experiences. (A crucial distinction between "hallucinations" in AI and human hallucinations is that AI hallucinations arise from incorrect data interpretation or lack of context, whereas human hallucinations are typically sensory experiences perceived by individuals and may not necessarily be accurate representations of reality.)
26. What is the main goal of the Stanford One Hundred Year Study on Artificial Intelligence?
    1. To study and anticipate the long-term effects of AI on humanity.
27. The chatbot deployed by the American National Eating Disorders Association (NEDA) was disabled shortly after it went live in June 2023. What was the main reason for this?
    1. It was providing harmful eating disorder advice to users.
28. What is NOT a recommended prompt engineering technique for students?
    1. Giving as little information as possible to allow for more creative output.
29. What is NOT a task that would be carried out by human judges when training or fine-tuning a language model.
    1. Designing the architecture of the AI model.
30. Current AI detectors typically base their scores on perplexity, which correlates with the sophistication of the writing. What is the main issue with this design choice?
    1. Perplexity can lead to biases against non-native English writers. (The main issue with using perplexity as the main metric for AI detectors is that it can introduce biases against non-native English writers or individuals with less sophisticated writing skills. This is because perplexity scores might be lower for these writers, leading to incorrect classifications as AI-generated content.)