Unit 4

Lecture 4.1

1. Assistant models are trained in several stages. Which of the following stages does not involve the language modelling objective?

   - reward modelling

     Correct. Reward modelling focuses on learning to predict human preferences and does not use the language modelling objective, which involves next-token prediction.

   - unsupervised pre-training

     Incorrect. Unsupervised pre-training directly uses the language modelling objective to predict the next token in a sequence.

   - instruction fine-tuning

     Incorrect. Instruction fine-tuning still uses the language modelling objective but incorporates task-specific instructions to guide responses.

2. Consider a customer support chatbot with a female avatar. Which of the following outputs would be an example showing the limitation of language modelling as an objective in instruction fine-tuning?

   - I'm a woman so I just don't understand.

     Correct. This output is okay from a language modelling perspective but demonstrates a failure in aligning the model to avoid generating biased or inappropriate statements.

   - Not I understand sorry sorry.

     Incorrect. While grammatically incorrect, this example reflects a language modeling failure, not a limitation in the objective.

   - Can you repeat that? I am not sure I understand.

     Incorrect. This is an appropriate and coherent response, showing no clear limitation of the language modeling objective.

3. What is the purpose of reward models in LLM training?

   - to act as a proxy for immediate human feedback

     Correct. Reward models are trained to predict human preferences and serve as a proxy for human feedback during reinforcement learning.

- to suggest suitable payment for human annotators

  Incorrect. Reward models are unrelated to annotator compensation.

- to increase the computational efficiency of the training process

  Incorrect. Reward models are used to align the model's behaviour with human preferences, not to improve computational efficiency.

4. For a well-aligned LLM, which of the following is the most plausible reward model loss for the two completions from the moon landing example?

   - 0.11

     Correct. A low loss indicates that the reward model successfully distinguishes between the preferred and non-preferred completions. To see that this is the correct answer, consider the formula for the loss on the "Reward model" slide:

     $$\text{loss}(\boldsymbol{\theta}) = -\frac{1}{m}\sum_{i=1}^{m}\log\Big(\sigma\big(R_{\boldsymbol{\theta}}(x_i, y_i^+) - R_{\boldsymbol{\theta}}(x_i, y_i^-)\big)\Big)$$

     A well-aligned model is trained to assign a significantly higher reward to the preferred than to the non-preferred completion: $R_{\boldsymbol{\theta}}(x_i, y_i^+) > R_{\boldsymbol{\theta}}(x_i, y_i^-)$. This means that the term $R_{\boldsymbol{\theta}}(x_i, y_i^+) - R_{\boldsymbol{\theta}}(x_i, y_i^-)$ will be positive, the output of the sigmoid function will be close to 1, and the loss will be low.

   - 0.69

     Incorrect. A loss around 0.69 corresponds to a sigmoid output of 0.5, suggesting that the reward model is not distinguishing between (assigning the same reward to) the preferred and non-preferred completions.

   - 2.30

     Incorrect. A higher loss like 2.30 implies the model is failing to align with human preferences by assigning a higher reward to the non-preferred completion than to the preferred one.

5. What is the goal of policy gradient methods?

   - maximise the probability of generating outputs with high reward

     Correct. Policy gradient methods aim to optimise the model's parameters to increase the likelihood of generating outputs that receive higher rewards.

   - maximise the reward of generated outputs

     Incorrect. While related, policy gradient focuses on increasing the probability of high-reward outputs rather than the absolute reward. (Note that the reward model is fixed at this stage of the training.)

- minimise the perplexity of the generated outputs

  Incorrect. Minimising perplexity is associated with language modelling, not reinforcement learning using policy gradients.

Lecture 4.2

1. What core concern motivates this paper's investigation of language models and fact completion?
   - That correct answers alone may not prove that a model has truly memorised a fact

     Correct. This is precisely the paper's motivation: a correct prediction could stem from heuristics, guesswork, or actual memorisation, and prior work conflates these.
   - That language models cannot generalise beyond their training data

     Incorrect. The paper does not investigate generalisation failure; it investigates whether correct predictions reflect genuine fact memorisation.
   - That language models generate ungrammatical sentences when prompted with facts

     Incorrect. Grammaticality is never discussed as a concern.
2. Which criterion distinguishes *guesswork* from *exact fact recall* in the PRISM taxonomy?
   - Whether the model's prediction remains stable across paraphrased prompts

     Correct. This is the "confident prediction" criterion: *guesswork* is defined as a valid-type prediction that is not consistent across paraphrases (appearing in the top-3 for fewer than 5 templates), whereas exact fact recall requires confidence (at least 5 templates).
   - Whether the prompt contains a real-world subject

     Incorrect. Both *guesswork* and *exact fact recall* use real subjects; synthetic subjects are used for *heuristics recall*.
   - Whether the predicted token is of the correct semantic type

     Incorrect. Both scenarios require a valid-type prediction (the "fact completion" criterion); that criterion alone does not separate them.

3. Why do the authors use synthetic subjects when testing for heuristic recall?
   - To ensure that predictions do not rely on real-world memorised facts

     Correct. Synthetic subjects guarantee that the model cannot have memorised any fact about them. Confident predictions must therefore stem from heuristics, not memory.
   - To increase dataset size without manual annotation

     Incorrect. Dataset size is not the motivation.
   - To make the prediction task harder for the language model

     Incorrect. Task difficulty for the model is not the goal; controlling for memorisation is.

4. What do the causal tracing results show about *exact fact recall* compared to *generic language modelling*?
   - Only exact fact recall shows strong importance of mid-layer MLPs at the last subject token

     Correct. The experiments show a clear peak in (last subject token, mid-layer) MLP states only for exact fact recall. Generic language modelling shows no such peak, instead showing importance of late-layer last-token states.
   - Generic language modelling relies more on subject tokens than fact recall

     Incorrect. The opposite is true; generic LM shows negligible importance for subject-position states.
   - Both exhibit strong signals in early attention layers

     Incorrect. Neither scenario is characterised by early attention layer dominance in the causal tracing results.

5. What is the main conclusion when comparing causal tracing and information flow across the PRISM scenarios?
   - Different behavioral scenarios correspond to different internal mechanisms

     Correct. This is the paper's central finding: each of the four scenarios produces distinct causal tracing and information flow patterns, meaning internal mechanisms differ across scenarios.
   - Accurate predictions always rely on the same internal mechanism

     Incorrect. The whole point is that even among accurate predictions (*exact fact recall* vs. *heuristics* vs. *guesswork*), the internal mechanisms differ.

- Only heuristic recall differs internally from exact fact recall

  Incorrect. All four scenarios show distinct patterns from one another, including guesswork and generic language modelling.

Lecture 4.3

1. What is not a potential benefit of quantisation in language model training?

   - The model becomes more robust to rounding errors.

     Correct. Quantisation can actually make models more sensitive to rounding errors, as it reduces the precision of the parameters.

   - Training is faster.

     Incorrect. Quantisation can speed up training by reducing the computational load, especially on hardware that supports lower precision arithmetic.

   - The trained model requires less memory.

     Incorrect. Quantisation reduces the memory footprint of the model by using fewer bits to represent each parameter, which can be beneficial for deployment and inference.

2. Consider a 100-billion-parameter LLM that stores all of its parameters as 32-bit floats. How much memory do we save by storing half of the parameters with half precision in a mixed-precision setup?

   - 100 GB

     Correct. In a mixed-precision setup where half of the parameters are stored in 16-bit precision, we save 16 bits for each of those parameters. Since there are 50 billion parameters stored in half precision, the total memory saved is:

     $$\text{memory saved} = 50 \times 10^9 \times 16 \text{ bits} = 800 \times 10^9 \text{ bits} = 100 \text{ GB}$$

   - 200 GB

     Incorrect.

   - 400 GB

     Incorrect.

3. Consider the formula $W_0 + \Delta W = W_0 + BA$ in the explanation of LoRA. Which of the matrices in this formula is trained during LoRA training?

   - W_0

     Incorrect. W_0 is the original weight matrix that is kept frozen during LoRA training.

   - ΔW

     Incorrect. ΔW is the change in weights that is computed from B and A, and is not directly trained.

   - BA

     Correct. The product BA represents the low-rank update to the original weights, so B and A are the matrices that are trained during LoRA training.

4. Consider the formula $W_0 + \Delta W = W_0 + BA$ once more. If W_0 is a 200-by-100 matrix and ΔW has rank 10, how many entries does B have?

   - 1000

     Incorrect.

   - 2000

     Correct. The matrix B has dimensions 200×10, and the matrix A has dimensions 10×100. Therefore, the number of entries in B is $200 \times 10 = 2000$.

   - 20000

     Incorrect.

5. What is the optimal rank to choose for LoRA training?

   - 2

     Incorrect.

   - 4

     Incorrect.

   - depends on the task

     Correct. The optimal rank for LoRA training can vary depending on the specific task and the model architecture. It is often determined empirically through experimentation, as different tasks may require different levels of expressiveness in the low-rank updates.

Lecture 4.4

1. Which of the following LLM tasks is most affected by the problem of staleness?

   - stock market prediction

     Correct. Stock market prediction relies heavily on up-to-date information, and staleness can lead to outdated or irrelevant predictions.

   - translation from Latin to English

     Incorrect. While translation can benefit from current linguistic trends (though perhaps not for Latin), it is less affected by staleness compared to tasks that require real-time information.

   - sentiment classification

     Incorrect. Sentiment classification is less affected by staleness because it typically does not rely on real-time information.

2. What problem is illustrated by the GDPR example mentioned in the lecture?

   - revisions

     Correct. The GDPR example illustrates the problem of revisions – the need to change the language model, e.g., by removing personal information due to legal requirements.

   - hallucination

     Incorrect. Hallucination refers to the generation of plausible but incorrect information, which is not the issue highlighted by the GDPR example.

   - attribution

     Incorrect. Attribution concerns the ability to trace the source of information, which is not the primary focus of the GDPR example.

3. What does the "open book" correspond to in the RAG framework?

   - the document database

     Correct. In RAG, the "open book" metaphor refers to the document database that the model can access to retrieve relevant information to solve tasks, just as a student would consult an open book in an exam.

   - the parameters of the LLM

     Incorrect. The parameters of the LLM are not the "open book"; they represent the model's internal knowledge, which is opaque.

- the prompt given to the LLM

  Incorrect. The prompt given to the LLM is not the "open book"; it is a set of instructions that guides the model's behaviour, but it is not the source of external information that the model accesses.

4. Which problem does the lecturer *not* mention as being addressed by grounding?

   - staleness

     Correct. Grounding can help mitigate staleness by allowing the model to access up-to-date information from external sources, but the lecturer does not explicitly mention staleness as a problem addressed by grounding.

   - hallucination

     Incorrect. Grounding can help reduce hallucination by providing the model with access to factual information, which can improve the accuracy of its responses.

   - attribution

     Incorrect. Attribution concerns the ability to trace the source of information, which is not explicitly addressed by grounding in the lecture.

5. Which of the following is mentioned as an advantage of dense retrieval?

   - It is possible to retrieve text that is semantically similar.

     Correct. Dense retrieval uses continuous vector representations of text, allowing it to retrieve semantically similar text even if the exact keywords do not match.

   - It has a smaller compute footprint than sparse retrieval.

     Incorrect. Dense retrieval typically has a larger compute footprint than sparse retrieval due to the need for computing and comparing dense vector representations.

   - It plays along well with word embeddings.

     Incorrect. Dense retrieval relies on dense vector representations, which are not the same as traditional word embeddings used in sparse retrieval. Dense retrieval often uses more complex representations that capture semantic meaning beyond individual word embeddings.

Lecture 4.5

1. What is meant by *canonical tokenisation* in this work?

   - The tokenisation produced by the tokeniser the model was trained with

     Correct. The canonical tokenisation is explicitly defined as the output of the BPE canonical tokenizer that iteratively applies merge rules until fixpoint — i.e., the unique tokenisation produced by the trained tokeniser.

   - The tokenisation that yields the highest model probability

     Incorrect. The canonical tokenisation is defined as the one produced by applying BPE merge rules to fixpoint; it happens to usually be the shortest, not necessarily the highest-probability one.

   - The tokenisation with the smallest number of tokens

     Incorrect. The paper notes canonical is *usually* the shortest, but that is a consequence of BPE, not the definition.

2. What trend do the quiz-based experiments reveal about tokenisation distance and model accuracy?

   - Accuracy generally decreases as tokenisation distance increases

     Correct. The experiments show a clear decreasing trend in accuracy across all three difficulty levels and all three models as normalised distance from the canonical tokenisation increases.

   - Accuracy remains constant regardless of tokenisation

     Incorrect.

   - Accuracy improves as tokenisation distance increases

     Incorrect.

3. Why does adversarial tokenisation pose a challenge for aligned language models?

   - Alignment training primarily covers canonical tokenisations

     Correct. The authors note that the safety distribution shift is centered around the canonical tokenization, so non-canonical tokenisations can access parts of the distribution not covered by alignment training.

   - Alignment training applies equally to all possible tokenisations

     Incorrect. This is the opposite of what the paper argues.

- Non-canonical tokenizations are excluded during pre-training

  Incorrect. The paper's whole point is that semantic understanding of non-canonical tokenisations is acquired during pre-training (via "semantic leakage"), but alignment post-training fails to cover them.

4. Why do the authors use greedy local search to find adversarial tokenisations?

   - Because the optimal tokenisation problem is computationally intractable

     Correct. The authors show that the conditional most-likely-tokenisation problem is NP-complete, which directly motivates the greedy approximation as a practical substitute.

   - Because random sampling cannot produce valid tokenisations

     Incorrect. The paper uses random sampling in several experiments; it is perfectly capable of producing valid tokenisations.

   - Because greedy search is guaranteed to find an optimal tokenisation

     Incorrect. The paper explicitly describes the greedy local search algorithm as an approximation that finds a local optimum, not a global one.

5. How do the authors explain the apparent contradiction that models understand non-canonical tokenisations semantically but not at the alignment level?

   - Semantic understanding and alignment are learned entirely during post-training

     Incorrect. The paper says semantic understanding comes from pre-training, not post-training.

   - Alignment signals override all semantic representations learned during pre-training

     Incorrect. The whole point is that alignment signals do not fully override pre-training — they are too limited in scope to do so.

   - Pre-training introduces semantic leakage through diverse data, while alignment data is smaller and cleaner

     Correct. The paper explicitly resolves the contradiction by noting that massive-scale pre-training causes semantics to "leak" onto many tokenisations, whereas post-training safety fine-tuning uses comparatively little data and thus cannot cover the full tokenisation space.

Lecture 4.6

1. What is the main difference between "traditional" and the new LLM-based chatbots?

   - Traditional chatbots were more restricted in their affordances than LLM-based chatbots

     Correct. Traditional chatbots were typically designed for specific tasks and had limited capabilities, while LLM-based chatbots can handle a much wider range of topics and tasks due to their large-scale training on diverse datasets.

   - Traditional chatbots were trained on considerably less data than LLM-based chatbots

     Incorrect. Traditional chatbots were often trained on much smaller datasets, but the difference in data size is not the main distinction between traditional and LLM-based chatbots.

   - Traditional chatbots were not as transparent to users than LLM-based chatbots

     Incorrect. Transparency can vary widely among both traditional and LLM-based chatbots, and it is not a defining characteristic that distinguishes the two.

2. Why does Bender refer to LLMs as "stochastic parrots"?

   - They randomly recombine linguistic forms, without any reference to meaning

     Correct. LLMs generate text by predicting the next token based on patterns in the training data, without any understanding of meaning or reference to the real world.

   - They model the generation of natural language as a stochastic process

     Incorrect. While LLMs do model language generation as a stochastic process, this is not the reason for the "stochastic parrots" label; the term emphasises the lack of true understanding and meaning in their outputs.

   - They are able to generate meaningful text about any topic in sheer endless variation

     Incorrect. LLMs can generate text on a wide range of topics, but their outputs are not necessarily meaningful or accurate, and they do not have true understanding or creativity.

3. What is Bender's stance on the following statement? "ChatGPT et al. seem to understand language."

   - They do not; this is only how we perceive them. We cannot help but imagine a mind behind the text.

     Correct. Bender argues that the apparent understanding of language by LLMs is an illusion created by their ability to generate coherent text, and that we tend to anthropomorphise these models, attributing them with understanding that they do not actually possess.

   - They do not yet, but there is a good chance they will if we scale things (data, model size) up even further.

     Incorrect. Bender is sceptical about the idea that simply scaling up models will lead to true understanding, as she believes that the fundamental limitations of LLMs are not just a matter of scale but are inherent to their design and training objectives.

   - They do so because the current generation of these models is not only trained on text but also on human feedback.

     Incorrect. While human feedback can improve the performance of LLMs, Bender's critique focuses on the fundamental nature of these models as pattern recognisers rather than true understanders of language, regardless of the training methods used.

4. According to Bender, why are chatbots not a good replacement for search?

   - Users get tricked into believing that there is "the answer" and cut off from thinking on their own.

     Correct. Bender argues that chatbots can create a false sense of certainty, leading users to accept generated answers without critical evaluation, which can hinder independent thinking and the pursuit of further information.

   - LLM-based chatbots require considerably more energy than traditional search engines.

     Incorrect. While energy consumption is a concern with LLMs, Bender's critique of chatbots as replacements for search focuses more on the epistemological and cognitive implications rather than the environmental impact.

   - Users can never be sure whether the answer generated by the chatbot is correct.

     Incorrect. While chatbots may generate plausible-sounding answers, users often cannot verify their accuracy because chatbots do not provide sources or citations for their claims. This lack of verifiability is a key concern in Bender's critique.

5. What does the term “information provenance” refer to?

   - tracking the origin and history of information

     Correct. Information provenance refers to the tracking of where information comes from and how it has been processed or transformed over time.

   - encrypting and securing information

     Incorrect. While encryption is important for information security, it is not what is meant by information provenance.

   - studying how information influences public opinion

     Incorrect. While the influence of information on public opinion is an important area of study, it is not what is meant by information provenance, which focuses on the origin and history of information rather than its effects on society.