Large Language Model (Monsoon 2024)

Mid Semester Examination
Date of Examination: 11 Oct 24 Duration:1 Hour Total Marks: 25 Marks

Multiple Choice Questions [4 Marks]

1. **Which of the following is not a finding of paper Stereoset.**

    a. All pretrained models display more stereotypical behavior than RandomLM
    b. Corpora size does not correlate with lms or ss
    c. Model size increase, language modeling ability increase and stereotypical score increase due to the reliance on real-world corpora
    d. Model size increase, language modeling ability increase and stereotypical score decrease due to the reliance on real-world corpora [Correct]

2. **Which of the following best describes the primary method used in the paper Reference: Break it, Imitate it, Fix it: Robustness by Generating Human-Like Attacks to improve adversarial robustness?**

a. Training only on real adversarial examples collected by humans
b. Relying on over-simplified synthetic attacks for adversarial robustness
c. Using generated attacks that imitate real human adversarial attacks to improve robustness to future unseen attacks [Correct]
d. Focusing exclusively on small-edit-distance and word-embedding-based attacks for robustness

3. **Large Language Models like GPT-3 have been used to generate text that mimics the writing style of famous authors or personalities. What is this technique called?**

a. Language modeling
b. Style transfer [Correct]
c. Plagiarism
d. Identity theft

4. What is the primary purpose of specifying a “temperature” parameter when generating text with a large language model like GPT-3?

a. Controlling the speed of text generation

b. Adjusting the model's verbosity and creativity[Correct]
c. Fine-tuning the model on specific tasks
d. Managing the model's memory consumption

Fill in the blanks [3 Marks]

1. The self-attention mechanism of the Transformer model allows for dynamic weighting of input tokens based on their contextual relevance to each other.
2. In reinforcement learning frameworks, the numerical signals that provide feedback on the efficacy of an agent's actions in relation to its defined objectives are known as rewards.
3. The Plug and Play Language Model (PPLM) detoxification method alters the hidden representations using gradients from a discriminator / classifier / classifier whether the generated text is toxic or non-toxic.

Subjective [16 Marks]

1. **What is the difference between Causal Language Modeling, Masked Language Modeling, and Translation Language Modeling? [1.5 MARK- 0.5 for each head any one point]**
   Solution:

Feature	Causal Language Modeling (CLM)	Masked Language Modeling (MLM)	Translation Language Modeling (TLM)
Training Objective	Predict the next token in sequence	Predict masked tokens using full context	Predict masked tokens using bilingual/multilingual context
Context	Left-to-right (unidirectional)	Bidirectional	Cross-lingual (from both languages)
Common Model	GPT, GPT-2, GPT-3	BERT, RoBERTa	XLM
Applications	Text generation	Language understanding tasks	Multilingual tasks, translation
Input Handling	Sequential, no access to future tokens	Access to both past and future tokens	Access to context in both source and target languages

2. **In the context of StereoSet paper, define the following:**
   Language modeling score (lms), Stereotype score (ss), Idealised CAT score. [1.5 Marks - 0.5 marks for each]

   Solution:

- Language modeling score (lms)
    - Percentage of instance language model prefers the meaningful over
    - meaningless (Ideal: 100) (0.5 marks)
- Stereotype score (ss)
    - Percentage of instance language model prefers the stereotypical over
    - anti-stereotypical (Ideal: 50) (0.5 marks)
- Idealised CAT score (icat)
    - icat = lms * min(ss, 100-ss)/50 (0.5 marks)
    -

3. **In a typical NLP model development pipeline, what are four steps to mitigate harms caused by large language models (LLMs), and provide one technique for addressing each step. [2 Marks]**

   (0.5 for each step- 0.25 for step name, 0.25 for one technique)

   Solution:

   - **Application Deployment (**Detection, Flagging and Redaction: Detect risk and warn the user)
       - Technique:
           - Rule-based Systems: Lexicons and linguistic Features
           - Neural classifiers: incorporate contextual information, fine-tuning pre-trained LMs.

   - **Output Level Interventions**
       - Technique:
           - Rejection Sampling: Repeatedly sample outputs and reject harmful outputs
           - Decoding: Guide the inference procedure using risk detectors
           - Post-Factum Editing: Rewrite harmful outputs

   - **Pretraining**
       - Technique:

           - New Architectures and Training Procedures: Control Codes,

           Instruction-based Learning, Augmented LMs

- Simple Fine-tuning, Prompt Tuning, Model Surgery
- Reinforcement Learning with Human Feedback

- **Data Collection/Curation**
  - **Technique**
    - Filtration: Detect and filter harmful information from training datasets.
    - Augmentation: Counter harmful text with harmless or beneficial text.

4. Consider the FLAN and T0 models and Answer the following parts: [3 Marks]

a. **What is the key difference in the training data sources of both?**
b. **What is the key difference in the objectives of both?**
c. **Highlight one major distinction in the training strategies of both models. [Subject to change]**

Solution:

	FLAN	**T0**
training data sources	**10 Template prompt** formats for each dataset **With and without cot, both with and without examplers**	each task is formatted using human-readable **publicly contributed prompt templates**, specifically designed to evaluate zero-shot generalization capabilities.
objectives	**instruction finetuning** with a particular focus on (1) **scaling** the number of **tasks**, (2) scaling the **model** size, and (3) finetuning on **chain-of-thought data**.	multitask prompted training, induce **robustness to different prompt wordings**
strategies	**FLAN training multiple models with one held-out task**	**T0 trains one model with multiple held-out tasks**

1. Key Difference in the Training Data Sources: [1 mark for any 1 difference]

- **Flan-T5:**
  - Used instructional templates for each task .
  - For CoT, manually written instruction templates
  - For few-shot templates, a variety of exemplar delimiters (e.g., "Q:"/"A:") were written and applied to them randomly at the example level
  - Some data formats with few-shot examplars without explicit instructions.
- **T0:**
  - T0 is trained on multitask prompted data, where each task is formatted using human-readable prompts.
  - It primarily uses datasets with multiple publicly contributed prompt templates, specifically designed to evaluate zero-shot generalization capabilities.

2. Key Difference in the Objectives: [1 mark for any 1 difference]

- **Flan-T5:**
  - Exploring instruction finetuning with a particular focus on (1) scaling the number of tasks, (2) scaling the model size, and (3) finetuning on chain-of-thought data.

- **T0:**
  - T0's main objective is to improve zero-shot generalization by multitask prompted training .
  - T0 aims to induce robustness to different prompt wordings by training on a wider range of prompts.

3. One Major Distinction in the Training Strategies: [1 mark for any 1 difference]

- **Flan-T5:**
  - FLAN-T5 uses instruction fine-tuning with both **task-specific instructional templates** and **chain-of-thought (CoT)** reasoning data. Fine-tunes both **with and without exemplars or CoT**, as well as using **data formats without instructions but including few-shot exemplars**.
  - FLAN-T5 includes evaluation both with and without exemplars (few-shot settings) and emphasizes instruction understanding across a broad set of tasks.
  - FLAN-T5 includes evaluation on chain-of-thought tasks, testing models on step-by-step reasoning tasks.
  -
- **T0:**

- Relies on multitask prompted training, where multiple prompts are used for the same task, improving the model's generalization ability across different prompt wordings.(assembled our multitask training mixture by combining and shuffling all examples from all training datasets.)
- It doesn't include chain-of-thought training but focuses more on learning to generalize to new tasks through prompt variability.
- T0 is evaluated on zero-shot task generalization, focusing on how well it generalizes to completely new tasks without any task-specific training.
- T0 focuses on robustness to prompt variety, testing the model's ability to handle different ways of phrasing the same task or question.
- T0 does not explicitly focus on reasoning; its evaluation strategy centers on prompted tasks and general NLP tasks.

5. In the below equation of loss function in reward model of method of alignment of LLM .

$$\text{loss}(\theta) = -\frac{1}{\binom{K}{2}} E_{(x,y_w,y_l)\sim D}\left[\log\left(\sigma\left(r_\theta(x,y_w) - r_\theta(x,y_l)\right)\right)\right]$$

a. **What is $r_\theta(x, y_\omega)$ and $r_\theta(x, y_l)$ [1 Mark]**

b. **What engineering trick was used to model the loss to prevent overfitting? Explain. [2 Marks]**

Solution:

A. Reward on winning, Reward on losing
Reward on The better completion, Reward on The worse completion.

B.

$$\text{loss}(\theta) = -\frac{1}{\binom{K}{2}} E_{(x,y_w,y_l)\sim D}\left[\log\left(\sigma\left(r_\theta(x,y_w) - r_\theta(x,y_l)\right)\right)\right]$$

Small but important detail:

- Each prompt has K completions => C_K^2 pairs to compare
- If ∀ batch we sample uniform over *every* pair (from any prompt
  - Each completion can appear in K - 1 gradient updates
  - This can lead to overfitting
- **Solution:** sample the prompt, and then put all C_K^2 pairs from the prompt into the same batch
  - Corollary: computationally more efficient, since this only requires K forward passes through r_θ for each prompt
- This is why there is the -1/C_K^2 normalization in loss

6.

Given (i) input is **raw text** and (ii) the output is the **logits**. Explain and draw the GPT model pipeline illustrating how the input is transformed into logits. Define the roles of all the building blocks used. Except what is given, **do not assume anything**. [4+3 MARKS]

[Can vary]

Solution:

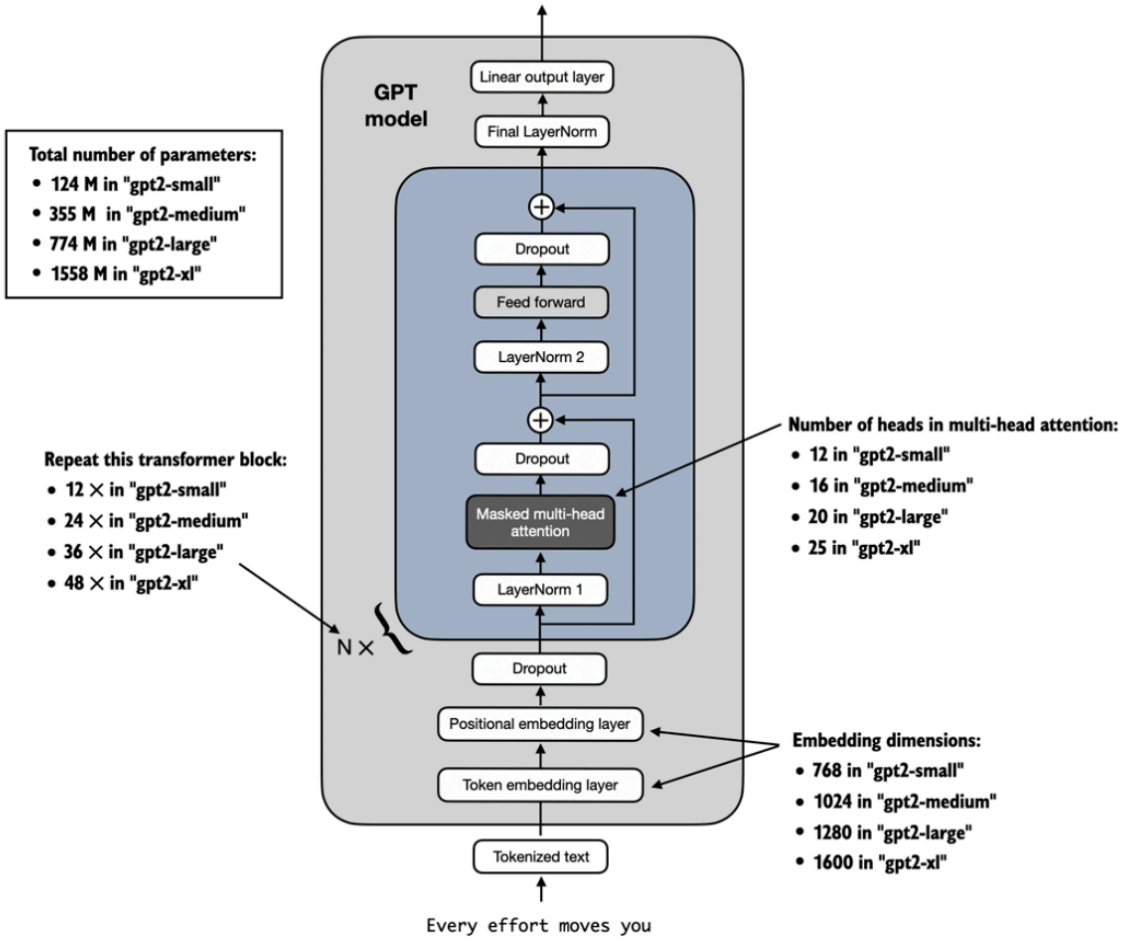
In architecture:

text->
token embedding -> (0.5)
positional embedding ->(0.5)
Dropout

Layernorm
Masked multihead attention (1)
Dropout transformer- 0.5
Layernorm2
Feedforward (0.5)
Dropout

Final layer norm and Linear output (1)

3 marks - 3 major components to define - 1 each
Token embedding,Positional embedding
Multihead attention,transformer
Ffn , Final layer/ output



1. Input: Tokenized Text
2. Token embedding layer:
3. Positional Encoding: Adds position information to the tokens since transformers do not have an inherent sense of word order.
4. LayerNorm1:Stabilizes training and helps avoid vanishing/exploding gradient problems. Allows gradients to flow through deeper layers efficiently.
5. Masked Multihead attention: Allows the model to attend to different words in the sequence when generating each word. Helps the model capture dependencies across words regardless of distance in the text.
6. Dropout
7. Feed forward: Applies transformations to the attended representations. Introduces non-linearity and learns higher-level features.
8. Linear output layer: The linear layer transforms the decoder's output into logits, followed by a softmax function to predict the probability of the next token in the sequence.