Q1 Compare response with / without chat template
1 分數

In this experiment, you will be viewing the model's behavior with/without chat template based on a coherence score between the questions and the model's response. You will be using the following prompt:

Please tell me about the key differences between supervised learning and unsupervised learning. Answer in 200 words.

Please answer the following 4 questions.

Q1.1 Coherence score with chat template
0.2 分數

Please calculate the coherence score for responses generated with the chat template.

What is the score **WITH** chat template?
(The error between 0.5 is accepted)

4.0107

Q1.2 Coherence score without chat template
0.2 分數

Please calculate the coherence score for responses generated without the chat template.

What is the score **WITHOUT** chat template?
(The error between 0.5 is accepted)

-2.8643

Q1.3 Which coherence score is higher?
0.3 分數

The response **WITH** a chat template

The response **WITHOUT** a chat template

Q1.4 Please choose the correct statements(s) from the following according to the experiments in Q1. (You should choose EXACT 2 answers)
0.3 分數

- [ ] Generating the response without a chat template has a higher score since it can let the model understand better and adapt to diverse input formats, enhancing its versatility.
- [x] Generating the response with a chat template has a higher score since the model is trained on a structured format, and inference should match that format to produce optimal results.
- [x] Generating the response with a chat template has a higher score since it provides consistent formatting, ensuring that inputs and outputs follow a structured conversation flow.
- [ ] Generating the response with a chat template has a higher score since it increases the model's reasoning ability, making the model inherently more intelligent.
- [ ] Generating the response without a chat template has a higher score, as it does not impose a rigid structure on the conversation.
- [ ] Generating the response without a chat template has a higher score because the model is more flexible and can generate responses in any format.

Q2 Multi-turn conversation
1 分數

Observe the following multi-turn conversation.

Conversation History should look like this:
(User is what you should input **TO** the model.)
(Model is the response **FROM** the model, using xxxx as an example.)

User (Your 1st Input): “Name a color in a rainbow, please just answer in a word without any emoji.”
Model 1st output : xxxx.

User (Your 2nd Input): “That’s great! Now, could you tell me another color that I can find in a rainbow?”
Model 2nd output: xxxx.

User (Your 3rd Input): “Could you continue and name yet another color from the rainbow?”
Model 3rd output: xxxx.

Please answer the following 4 questions.

Q2.1 Please provide the FULL prompt with chat template that is inputted to the model for the third round (input to the model's 3rd round)
0.4 分數

<bos><start_of_turn>user Name a color in a rainbow, please just answer in a word without any emoji. <end_of_turn> <start_of_turn>model

Q2.2 What is the first token with the highest probability in the first round (first model's response)? (Note: Case sensitive)
0.2 分數

Indigo

Q2.3 Please select the FALSE statement from the following according to the experiments in Q2.
0.4 分數

This multi-turn conversation experiment shows that the chat template can help the model maintain context and avoid repeating previous answers. help the model maintain context and avoid repeating previous answers.

The response from the model in each round is the token with the highest probability in all three questions.

The response from the model in each round is a different color.

This multi-turn conversation experiment shows that the model will update its model parameters during each round, and therefore the answer will not repeat.

Q3 Tokenization of Sentence
0.5 分數

For the prompt: "I love taking a Machine Learning course by Professor Hung-yi Lee, What about you?"

How is this prompt being tokenized? Write the corresponding token index.

(Note: Notice that there might be a '▁' in the token, so make sure you use the same '▁' here if needed. Otherwise, the answer might be incorrect.)

Token: I, token index: 235285
Token: ▁love, token index: 2182
Token: ▁taking, token index: 4998
Token: ▁a, token index: 476
Token: ▁Machine, token index: 13403
Token: ▁Learning, token index: 14715
Token: ▁course, token index: 3205
Token: ▁by, token index: 731
Token: ▁Professor, token index: 11325
Token: ▁Hung, token index: [1]
Token: [2], token index: 235290
Token: [3], token index:[4]
Token: ▁Lee, token index: 9201
Token: ,, token index: 235269
Token: ▁What, token index: 2439
Token: ▁about, token index: 1105
Token: ▁you, token index: 692
Token: ?, token index: 235336

Q3.1
0.15 分數

[1]:

18809

Q3.2
0.1 分數

[2]:

-

Q3.3
0.1 分數

[3]:

yi

Q3.4
0.15 分數

[4]:

12636

Q4 Auto regressive generation

1.4 分數

In this experiment, you will be using auto-regressive generation to generate a sentence **20 times** given the following prompt:

Generate a paraphrase of the sentence 'Professor Lee is one of the best teachers in the domain of machine learning'. Just response with one sentence.

Next, you will calculate the self-BLEU score for the generated 20 sentences. (You can refer to this link to get more acknowledgment of self-BLEU.)

We want you to observe the fluency, coherence, and diversity... for different configurations of Top-k sampling and Top-p sampling.
(k=2 vs k=200 ; p=0.6 vs p=0.999)

Please answer the following 6 questions.

Q4.1 Please choose the correct statement(s) about self-BLEU. (You should choose EXACT 2 answers)
0.25 分數

- [ ] Self-BLEU compares each generated sentence against all other generated sentences as reference texts, and chooses the highest BELU score as the final self-BLEU score.
- [x] Self-BLEU is a metric used to measure the diversity of generated sentences from a model.
- [ ] The higher the score, the more diverse these sentences are.
- [x] Self-BLEU is based on the BLEU (Bilingual Evaluation Understudy) score, which is typically used for evaluating the similarity between a generated text and other reference texts.

Q4.2 Choose the best statement(s) about top-p and top-k from the followings. (You should choose EXACT 2 answers)
0.25 分數

- [x] Top-p sampling selects the next token from the smallest set of tokens, sorted by decreasing probability, whose cumulative probability exceeds p.
- [ ] Top-k sampling generates the next k consecutive tokens in sequence.
- [x] Top-k sampling selects the next token only from the k most probable tokens
- [ ] Top-p sampling selects the next token from the first p% of all tokens in the tokenizer.

Q4.3 According to the experiment, please answer the sentence generated from the model with top-k for k = 1.
0.2 分數

Professor Lee is highly regarded as a leading expert in machine learning education.

Q4.4 According to the experiment, please answer the sentence generated from the model with top-p for p = 0.
0.2 分數

Professor Lee is highly regarded as a leading expert in machine learning education.

Q4.5 Compare the self-BLEU score of top-k for different k values (2 vs 200), which is higher and why?

0.25 分數

Self-BLEU will be **the same** between k=2 and k=200 because top-k does not affect generation diversity.

Self-BLEU will be **lower** for k=2 than for k=200 because a smaller k reduces diversity by forcing the model to choose from less tokens, and therefore a lower self-BLEU score.

Self-BLEU will be **lower** for k=2 than for k=200 because a larger k introduces more repetitions.

Self-BLEU will be **higher** for k=2 than for k=200 because a larger k reduces diversity by forcing the model to choose from more tokens, making outputs more similar.

Self-BLEU will be **the same** between k=2 and k=200 because self-BLEU cannot represent the diversity of the models output.

Self-BLEU will be **higher** for k=2 than for k=200 because smaller k limits token choices, reducing diversity.

Q4.6 What is the self-BLEU score of Compare the self-BLEU score of top-p for different p values (0.6 vs 0.999)? Which is higher and why?
0.25 分數

Self-BLEU is **lower** for p=0.6 than for p=0.999 because the higher the value of p , the less diverse the sentences will be.

Self-BLEU is **higher** for p=0.6 than for p=0.999 because smaller p reduces possible tokens to be chosen, leading to more repetitive text.

Self-BLEU is **higher** for p=0.6 than for p=0.999 because a smaller p will let the model generate more diverse sentences.

Self-BLEU remains **the same** regardless of p because only top-k will have different sentences being generated.

Self-BLEU remains **the same** regardless of p because diversity is only controlled by model temperature instead.

Self-BLEU is **lower** for p=0.6 than for p=0.999 because p=0.999 means to select the next token from 99.9% of all the tokens, which means there are lots of possible tokens selected.

Q5 t-SNE Visualization

1 分數

In this experiment, you will plot a figure about the t-SNE 2-D Embeddings for the following sentences:

"I ate a fresh apple.", # Apple (fruit)
"Apple released the new iPhone.", # Apple (company)
"I peeled an orange and ate it.", # Orange (fruit)
"The Orange network has great coverage.", # Orange (telecom)
"Microsoft announced a new update.", # Microsoft (company)
"Banana is my favorite fruit.", # Banana (fruit)

Now, please answer the following 3 questions based on your generated plot.

Q5.1 Choose the correct statement(s) about T-SNE. (You should choose EXACT 2 answers)

0.4 分數

- [x] t-SNE algorithm is a branched version of the SNE algorithm, both used for dimensionality reduction.
- [ ] t-SNE is a machine learning algorithm that requires a model to output the probability score between two embeddings by a regression model.
- [ ] t-SNE is a linear dimensionality reduction technique used primarily for visualizing high-dimensional data.
- [x] t-SNE converts pairwise similarities in high-dimensional space into probabilities using a Gaussian distribution and models them in low-dimensional space using a Student’s t-distribution.

Q5.2 Please choose the correct statement about the experiment in Q5.
0.3 分數

- The Euclidean distance between Orange (telecom) and Orange (fruit) is larger than one between Banana (fruit) and Apple (fruit)
- The Euclidean distance between Microsoft (company) and Orange (telecom) is larger than the one between Apple (fruit) and Orange (telecom)
- The Euclidean distance between Orange (fruit) and Orange (telecom) is smaller than the one between Orange (telecom) and Microsoft (company)
- The Euclidean distance between Apple (fruit) and Apple (company) is smaller than the one between Apple (fruit) and Orange (fruit)

Q5.3 Please choose the INCORRECT statement about the experiment in Q5.
0.3 分數

- This experiment shows that the embedding of a certain word is determined primarily by its meaning, not its spelling.
- This experiment shows that the same words with different meanings in different sentences will have very similar embeddings in the embedding space.
- This experiment demonstrates that the t-SNE algorithm can be used to visualize how words with different meanings are located in the embedding space.
- This experiment demonstrates that contextual word embeddings can capture semantic differences between words based on their usage in different sentences.

Q6 Observe the Attention Weight

0.8 分數

In this experiment, you will plot the figure of the attention map with the following starting prompt:

"Google "

And we fix the Generated tokens to 20, and view the attention in layer 10 with head index = 7

Please answer the following 1 question based on the figure you generate.

Q6.1 Please choose the correct statement in the following about the attention map generated from the sample code.

0.2 分數

The attention map exhibits a lower triangular pattern

The generated attention map in the sample code is the average attention for all heads in all layers of the model.

The generated attention map in the sample code is the medium attention for all heads in all layers of the model.

The attention map exhibits an upper triangular pattern.

Q6.2 Please choose all the correct statement(s) in the following about the attention map generated from the sample code.
0.2 分數

- [ ] The attention map reflects the non-causal structure, ensuring that each token can attend to all tokens before and after itself.
- [ ] The attention map reflects that each token can not attend to itself.
- [x] The attention map shows that tokens will refer to all preceding tokens and themselves, but not to tokens that come after them due to the causal masking mechanism.
- [ ] The attention map indicates that each token attends to only itself and not to any other tokens, regardless of position.

Q6.3 Please answer if the following statement is true/false?
0.2 分數

Different heads in the same layer will have the same attention map, while different layers with the same head index will have different attention maps.

False

True

Q6.4 Please answer if the following statement is true/false?

0.2 分數

Different heads in the same layer will have different attention maps, while different layers with the same head index will have the same attention map.

True

False

Q7 Observe the Activation Scores

2.3 分數

Q7.1 Based on the Gemma-scope, What does feature 10004 refer to? What does the activation density mean? (You should choose EXACT 3 answers)

0.5 分數

- [ ] Activation density only refers to the activations in the first few layers, where basic features are processed.
- [x] Feature 10004 refers to a specific neural feature that is activated in response to time travel-related content.
- [x] Feature 10004 is related to identifying general concepts of time, such as past, present, and future.
- [x] Activation density represents the distribution and concentration of activations across tokens and layers in the model. It indicates how strongly a feature is activated across the entire input sequence.
- [ ] Activation density refers to the overall number of activations for each token, regardless of layer or token position.

Q7.2 Compare maximum activations between 2 prompts
0.2 分數

prompt_a = "Time travel offers me the opportunity to correct past errors, but it comes with its own set of risks."
prompt_b = "I accept that my decisions shape my future, and though mistakes are inevitable, they define who I become."

Get the maximum activations from the Sparse Autoencoder (SAE) in Gemma for two prompts and compare their values. Which is larger?

Prompt_b

Prompt_a

Q7.3 Explain the reason of Q7.2, which is correct?
0.4 分數

prompt_b is higher because feature 10004 is more sensitive to philosophical discussions about fate and free will, which are more prominent in prompt_b.

prompt_a is higher because feature 10004 detects explicit references to "time travel", which are present in prompt_a but absent in prompt_b.

prompt_a is higher because feature 10004 is primarily triggered by discussions about "regret and personal growth", which are the main themes of prompt_a.

prompt_b is higher because feature 10004 is mainly triggered by discussions of technological advancements rather than time travel itself.

Q7.4 Based on the activations for each token in layer 24 about feature 10004, which of the following statement is correct?
0.2 分數

Each bar in the plot shows how strongly each layer activates for a specific token in the prompt about feature 10004.

Each bar in the plot shows how strongly each token activates for a specific layer in the prompt about feature 10004.

Q7.5 Based on the activations for each token in layer 24 about feature 10004, which of the following statement is correct?
0.2 分數

“Time” has the highest activation because Feature 10004 directly associates any mention of time with time travel.

“_travel” has the highest activation because it is strongly associated with time travel, while “Time” does not have a high activation.

Q7.6 Based on the activations for each token in layer 24 about feature 10004, which of the following statement is correct?
0.2 分數

The activations are uniform across all tokens, meaning Feature 10004 activates equally for every word in the prompt.

The activation patterns suggest that Feature 10004 does not simply react to individual words but instead recognizes meaningful phrases related to time travel.

Q7.7 Based on the activation plots across all layers, which of the following statement is INCORRECT? Hint: You can alter the tokens and observe the figure. (e.g. the lower/deeper layers tend to process complex information)
0.2 分數

Many tokens have high activation in deeper layers, suggesting that these layers are responsible for processing more complex information.

The fact that "_a" has high activation in layers 22–25 suggests that some tokens may carry specific contextual meaning that is only processed in the deepest layers.

Since "Time" only activates in earlier layers, it suggests that temporal concepts are simple and do not require deep processing.

The activation of "Time" (token index = 1) is higher in early layers (12–15), while the activation of "_a" is higher in deeper layers (22–25).

Q7.8 Refer to Q7.7, please answer if the following statement is true/false?
0.2 分數

Token activations remain constant across all layers, meaning no token is more important at one layer than another.

True

False

Q7.9 Please Refer to Q7.7, please answer if the following statement is true/false? if the following statement is true/false?
0.2 分數

The results suggest that deeper layers specialize in complex pattern recognition, while lower layers focus on basic token representations.

True

False

Q8 PAPERS READING - (1)

1 分數

Please read papers, paying attention to the problems they address, how they solve them. After reading, answer the following four questions. Each question is worth 0.25 points.

Q8.1 ~ Q8.3 only has 1 correct answer, and full marks are awarded only if it is completely correct.
Q8.4 have exact two correct options.

DIFFERENTIAL TRANSFORMER: https://arxiv.org/pdf/2410.05258
Attention Is All You Need https://arxiv.org/pdf/1706.03762

Q8.1 Which one is the primary purpose of the attention mechanism in Transformer models in “Attention Is All You Need”?

0.25 分數

- To reduce the number of parameters
- To process data sequentially
- To capture long-range dependencies in data
- To perform dimensionality reduction

Q8.2 According to “Attention Is All You Need”. In the Transformer architecture, which component is responsible for handling word order information?

0.25 分數

- Feed-forward neural network
- Layer normalization
- Self-attention mechanism
- Positional encoding

Q8.3 According to “Attention Is All You Need”. Which one of the following is a key advantage of Transformer models over Recurrent Neural Networks (RNNs)?
0.25 分數

- Ability to process sequences in parallel
- Reduced need for large-scale data
- Better performance on small datasets
- Lower computational complexity

Q8.4 According to the “Differential Transformer”, which of the following statements are correct? (You should choose EXACT 2 answers)
0.25 分數

- [x] It uses a pair of softmax functions to compute the difference in attention, eliminating common-mode noise in the attention mechanism.
- [ ] It eliminates common-mode noise in the attention mechanism by applying a normalization technique similar to LayerNorm, which is called GroupNorm.
- [ ] In DIFF Transformer, different initialization λ parameters (0.8 vs. 0.5) have almost no effect(<1%) on Validation Loss.
- [ ] It optimizes the attention mechanism by leveraging softmax to reduce redundant attention weights, thereby improving computational efficiency.
- [x] Introducing Differential Attention into the Multi-Head mechanism improves contextual learning and enhances performance on sequence modeling tasks.

Q9 PAPERS READING - (2)

1 分數

The capabilities of large language models (LLMs) have grown tremendously. However, there are lots of complex tasks that still pose significant challenges. Also, the great capability often comes with increasing computational and deployment costs. As a result, several methods try to deal with the issue. One of the solutions is called **test-time computation**, which is a recent trend to deal with the issue. Also, the professor has made a brief talk about this here in the first class.

Please read the following paper from Google Deepmind to know more about test-time computation and answer the following questions.

Test-time compute: https://arxiv.org/pdf/2408.03314

Q9.1 What are the main methods of scaling test-time computation discussed in the paper?

0.25 分數

- [x] Updating the model's predicted distribution adaptively.
- [x] Searching against verifier reward models.
- [ ] Applying data augmentation to the dataset.
- [ ] Adding dropouts and batch normalization to the model.

Q9.2 What is the correct statement behind optimizing test-time compute for Large Language Models (LLMs)? (Hint: Refer to Section 3 in the paper)

0.25 分數

The goal of optimizing test-time compute is to adapt to human preferences by computing the optimal response when the user gives feedback during test-time.

The goal of optimizing test-time compute is to increase the ability to search for online / cloud data.

“Compute-optimal” means to choose the best model parameter for test-time compute for a given prompt to maximize accuracy within a fixed compute budget.

The motivation is to enable self-improvement that can generate the response on challenging tasks.

Q9.3 Which of the following statements are correct about processed-based model rewards (PRM) in the paper? (Hint: Refer to Section 5 and Figure 3 in the paper) (You should choose EXACT 2 answers)
0.25 分數

- [x] We can optimize the PRM via several search methods, including beam search and lookahead method.
- [ ] PRM can consistently improve performance across all types of questions by integrating contextual information provided in the prompts.
- [ ] PRM is a reward function used in the tree-search method to decide what tokens should be augmented to the users' input prompts to make LLM generate better responses.
- [x] Lookahead search often requires a higher computational budget compared to the traditional best-of-N selection method.

Q9.4 Which of the following statements are correct about test-time compute in this paper? (Hint: Refer to Section 6, 7 in the paper)
0.25 分數

- [x] The ideal ratio of sequential to parallel test-time compute depends critically on both the compute budget and the specific question at hand.
- [x] Easier questions benefit from purely sequential test-time compute, while harder questions perform best with some ideal ratio of sequential-to-parallel compute.
- [x] There exists a tradeoff between sequential (e.g. revisions) and parallel (e.g. standard best-of-N) test-time computation.
- [x] Test-time and pre-training compute are not 1-to-1 “exchangeable”.

[ML HW3] Understanding Transformer

● 待批改

學生
蔡弘祥 (HUNG-SHIANG TSAI)

總分
- / 10 pts

問題 1

Compare response with / without chat template 1 pt

1.1	Coherence score with chat template	0.2 pts
1.2	Coherence score without chat template	0.2 pts
1.3	Which coherence score is higher?	0.3 pts
1.4	Please choose the correct statements(s) from the following according to the experiments in Q1. (You should choose EXACT 2 answers)	0.3 pts

問題 2

Multi-turn conversation 1 pt

2.1	Please provide the FULL prompt with chat template that is inputted to the model for the third round (input to the model's 3rd round)	0.4 pts
2.2	What is the first token with the highest probability in the first round (first model's response)? (Note: Case sensitive)	0.2 pts
2.3	Please select the FALSE statement from the following according to the experiments in Q2.	0.4 pts

問題 3

Tokenization of Sentence 0.5 pts

3.1	（沒有題目）	0.15 pts
3.2	（沒有題目）	0.1 pts
3.3	（沒有題目）	0.1 pts
3.4	（沒有題目）	0.15 pts

問題 4

Auto regressive generation 1.4 pts

4.1	Please choose the correct statement(s) about self-BLEU. (You should choose EXACT 2 answers)	0.25 pts
4.2	Choose the best statement(s) about top-p and top-k from the followings. (You should choose EXACT 2 answers)	0.25 pts
4.3	According to the experiment, please answer the sentence generated from the model with top-k for k = 1.	0.2 pts
4.4	According to the experiment, please answer the sentence generated from the model with top-p for p = 0.	0.2 pts
4.5	Compare the self-BLEU score of top-k for different k values (2 vs 200), which is higher and why?	0.25 pts
4.6	What is the self-BLEU score of Compare the self-BLEU score of top-p for different p values (0.6 vs 0.999)? Which is higher and why?	0.25 pts

問題 5

t-SNE Visualization 1 pt

5.1	Choose the correct statement(s) about T-SNE. (You should choose EXACT 2 answers)	0.4 pts
5.2	Please choose the correct statement about the experiment in Q5.	0.3 pts
5.3	Please choose the INCORRECT statement about the experiment in Q5.	0.3 pts

問題 6

Observe the Attention Weight 0.8 pts

6.1	Please choose the correct statement in the following about the attention map generated from the sample code.	0.2 pts
6.2	Please choose all the correct statement(s) in the following about the attention map generated from the sample code.	0.2 pts
6.3	Please answer if the following statement is true/false?	0.2 pts
6.4	Please answer if the following statement is true/false?	0.2 pts

問題 7

Observe the Activation Scores 2.3 pts

7.1	Based on the Gemma-scope, What does feature 10004 refer to? What does the activation density mean? (You should choose EXACT 3 answers)	0.5 pts
7.2	Compare maximum activations between 2 prompts	0.2 pts
7.3	Explain the reason of Q7.2, which is correct?	0.4 pts
7.4	Based on the activations for each token in layer 24 about feature 10004, which of the following statement is correct?	0.2 pts
7.5	Based on the activations for each token in layer 24 about feature 10004, which of the following statement is correct?	0.2 pts
7.6	Based on the activations for each token in layer 24 about feature 10004, which of the following statement is correct?	0.2 pts
7.7	Based on the activation plots across all layers, which of the following statement is INCORRECT? Hint: You can alter the tokens and observe the figure. (e.g. the lower/deeper layers tend to process complex information)	0.2 pts
7.8	Refer to Q7.7, please answer if the following statement is true/false?	0.2 pts
7.9	Please Refer to Q7.7, please answer if the following statement is true/false? if the following statement is true/false?	0.2 pts

問題 8

PAPERS READING - (1) 1 pt

8.1	Which one is the primary purpose of the attention mechanism in Transformer models in “Attention Is All You Need”?	0.25 pts
8.2	According to “Attention Is All You Need”. In the Transformer architecture, which component is responsible for handling word order information?	0.25 pts
8.3	According to “Attention Is All You Need”. Which one of the following is a key advantage of Transformer models over Recurrent Neural Networks (RNNs)?	0.25 pts
8.4	According to the “Differential Transformer”, which of the following statements are correct? (You should choose EXACT 2 answers)	0.25 pts