Room Number _____

Seat Number _____

Student Number |_____|

ANONYMOUSLY MARKED

(Please do not write your name on this exam paper)

CONFIDENTIAL EXAM PAPER

This paper is not to be removed from the exam venue

Computer Science

SAMPLE EXAM

Semester 1 - Final, 2025

COMP4446/COMP5046 Natural Language Processing

EXAM WRITING TIME: 2 hours

READING TIME: 10 minutes

EXAM CONDITIONS:
This is a RESTRICTED OPEN book exam - specified materials permitted

MATERIALS PERMITTED IN THE EXAM VENUE:
(No electronic aids are permitted e.g. laptops, phones, calculators)

Formula sheet (provided in the exam paper by unit coordinator)

One A4 sheet of handwritten and/or typed notes double-sided

Bilingual dictionary (must have been pre-approved, as indicated by an official University of Sydney stamp)

MATERIALS TO BE SUPPLIED TO STUDENTS:
None

INSTRUCTIONS TO STUDENTS:
This exam consists of three sections (A: Multiple Choice Questions, B: Short Answer Questions, C: Programming Questions). All sections should be answered on this paper. Please use blue or black ink. If you need additional writing space, please use the extra pages provided at the end of this exam booklet. Only pages in this exam booklet will be marked.

Section A consists of 9 Multiple Choice Questions worth a total of 9 marks.

Section B consists of 22 Short Answer Questions worth a total of 41 marks.

Section C consists of 3 Programming Questions worth a total of 10 marks.

Please tick the box to confirm that your examination paper is complete (24 pages). ☐

This page is intentionally left blank.

Student Number:

Complete this on every page so we can find pages if they get separated during scanning.

Equations

Perplexity:

$$P(w_1, w_2 ... w_N)^{-\frac{1}{N}}$$

Layer normalization:

$$\mu = \frac{1}{d}\sum_{j=1}^{d} x_j$$
$$\sigma = \frac{1}{d}\sum_{j=1}^{d}(x_j - \mu)^2$$
$$y_i = \frac{x_i - \mu}{\sqrt{\sigma + \epsilon}} * \gamma + \beta$$

Self-attention with a dot product (assuming any changes to account for position have already been applied):

$$\mathbf{q_i} = Q\mathbf{x_i}$$
$$\mathbf{k_i} = K\mathbf{x_i}$$
$$\mathbf{v_i} = V\mathbf{x_i}$$
$$e_{ij} = \mathbf{q_i}^\top \mathbf{k_j}$$
$$\alpha_{ij} = \text{softmax}(e_{ij})$$
$$\mathbf{t_i} = \sum_j \alpha_{ij}\mathbf{v_j}$$
$$\mathbf{o_i} = W_2 \text{ReLU}(W_1\mathbf{t_i} + \mathbf{b_1}) + \mathbf{b_2}$$

Variants of attention:

Dot product

$$\mathbf{e} = \mathbf{s}^\top \mathbf{h}$$

Scaled dot product

$$\mathbf{e} = \frac{\mathbf{s}^\top \mathbf{h}}{\sqrt{d_h}}$$

Multiplicative / Bilinear

$$\mathbf{e} = \mathbf{s}^\top W \mathbf{h}$$

Reduced-rank multiplicative

$$\mathbf{e} = \mathbf{s}^\top (\mathbf{U}^\top \mathbf{V})\mathbf{h}$$

Additive / Feedforward

$$\mathbf{e} = \mathbf{b}\tanh(W_1\mathbf{h} + W_2\mathbf{s})$$

Non-linearities:

$$\text{ReLU} = \max(0, x)$$
$$\tanh = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$
$$\sigma = \frac{a}{1 + e^{-x}}$$

Metrics:

$$\text{Precision} = \frac{TP}{TP + FP}$$
$$\text{Recall} = \frac{TP}{TP + FN}$$
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$
$$\text{F-Score} = \frac{2 * P * R}{P + R} = \frac{2TP}{2TP + FP + FN}$$
$$\text{F}_\beta\text{-Score} = \frac{(1 + \beta^2)TP}{(1 + \beta^2)TP + FP + FN}$$

Cohen's Kappa:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$
$$p_o = \frac{|\text{items with the same label}|}{N}$$
$$p_e = \sum_{l \in labels} \prod_{a \in annotators} \frac{n_{la}}{N}$$

TF-IDF:

$$\text{tf}_{t,d} = \begin{cases} 1 + \log_{10} \text{count(t, d)} & \text{if count(t, d)} > 0 \\ 0 & \text{otherwise} \end{cases}$$
$$\text{idf}_t = \log_{10}\left(\frac{N}{df_t}\right)$$
$$\text{tf-idf}_{t,d} = tf_{t,d} * idf_t$$

This page is intentionally left blank.

Student Number:

Complete this on every page so we can find pages if they get separated during scanning.

Multiple Choice Questions

Complete the answers below by completely filling in circles / squares next to the option(s) you are selecting. If the choices have ◯ then select exactly one option. If the choices have ☐, select all correct options. Indicate your answer by filling the shape, e.g., ●. If you make a mistake, draw an X over your answer, e.g., ⊗.

1. (1 mark) Which of the following are used to improve LLM model speed and/or reduce memory needs at inference time?
   - ☐ Sparsification / Pruning
   - ☐ Low-Rank Adaptation (LoRA) → not at inference time
   - ☐ Distillation, e.g., DistilBERT
   - ☐ Reduced numerical precision

2. (1 mark) Which of the following statements about social bias in datasets are true?
   - ☐ Models trained on data will tend to reproduce the social biases in the data
   - ☐ Social bias can be completely removed by careful dataset design
   - ☐ Social bias can exist in some tasks (e.g., coreference resolution), but not others (e.g., part-of-speech tagging)
   - ☐ When we try to address social bias we may be manipulating the dataset in a way that makes it not match patterns of language in the world

3. (1 mark) If Cohen's Kappa agreement between two annotators is low, what could that mean?
   - ☐ The annotations are low quality
   - ☐ The two annotators were not consistent
   - ☐ There were many ambiguous cases
   - ☐ Lots of different labels were used
   - ☐ Only a few of the possible labels were used
   - ☐ The guidelines were not very clear
   - ☐ The annotators were careless

4. (1 mark) A unigram LM is built for sequences of digits that assigns 0.5 to the value 1, 0.25 to 2, 0.25 to 3, and 0 to all other digits. Which of the following are true:
   - ☐ Perplexity(1, 1) == Perplexity(2) -1/N表示的是N root + 倒数
   - ☐ Perplexity(4) = 0
   - ☐ Perplexity(1, 1) > Perplexity(2) p(1,1) = 1/sqrt((0.5*0.5)) = 2;
   - ☐ Perplexity(1, 1) < Perplexity(2) p(2) = 1/0.25 = 4
   - ☐ Perplexity(4) is not defined P最小为1，而0表示undefined

5. The output of a translation system is being compared with this reference translation:

> Natural Language Processing will have a big impact on the world.

(a) (1 mark) Which of the following options will be scored highest by the chrF metric?

- ◯ NLP is going to be hugely impactful on the world!
- ◯ The comet's impact on the world led to the extinction of the dinosaurs.
- ◯ Natural Language Processing may have a big influence on the world.
- ◯ All around the world, NLP may have a big impact.

(b) (1 mark) Which of the following options will be scored highest by the BLEU metric?

- ◯ NLP is going to be hugely impactful on the world!
- ◯ The comet's impact on the world led to the extinction of the dinosaurs.
- ◯ Natural Language Processing may have a big influence on the world.
- ◯ All around the world, NLP may have a big impact.

6. (1 mark) What is the best possible perplexity?

- ◯ $-\infty$
- ◯ -1
- ◯ $-1 / \infty$
- ◯ 0
- ◯ $1 / \infty$
- ◯ 1
- ◯ ∞

7. (1 mark) Select all the benefits of representing a Bag of Words with a dictionary rather than a list: 这里list指sparse vector

- ☐ Saves space by storing approximate values
- ☐ Saves space by not storing zeroes
- ☐ Saves space by grouping words by their counts
- ☐ Faster to iterate over all observed words
- ☐ Faster to update a count in the bag
- ☐ Faster to check if a word was observed

指不是精确地存储每个值，**而是用一种近似方式**（例如压缩、量化）来减少存储空间。

8. (1 mark) For each use case below, indicate the most suitable data representation of the options provided. Note that the rubric for this question will consider all four responses together (ie., it will not be 0.25 each).

Sentiment classification on professionally written movie reviews.

◯ TF-IDF ◯ Word2Vec CBOW ◯ BERT Embeddings

Topic classification on websites in an internet crawler that needs to be extremely fast and does not have to be perfect.

◯ TF-IDF ◯ Word2Vec CBOW ◯ BERT Embeddings

Emotion identification on speeches that have been transcribed with automatic speech recognition.

◯ TF-IDF ◯ Word2Vec CBOW ◯ BERT Embeddings

Toxic content identification on social media posts.

◯ TF-IDF ◯ Word2Vec CBOW ◯ BERT Embeddings

CBOW适合关键词抽取

Student Number:

Complete this on every page so we can find pages if they get separated during scanning.

9. (1 mark) For each scenario below, you are deciding what metric should be optimised to keep users happy. Of the options provided, which is best? Note that the rubric for this question will consider both responses together (ie., it will not be 0.5 each).

   Spam detection for a client who does not want to miss any real mail. Here, a true positive is a message that was correctly labelled as spam.

   ◯ Precision ◯ Recall ◯ F-Score ◯ Accuracy

   Filtering applicants for the cast of a play where the director wants to save time but still form the best group. Here, a true positive is a good applicant that was correctly included in the list to consider.

   ◯ Precision ◯ Recall ◯ F-Score ◯ Accuracy

Short Answer Questions

In the questions below, please try to keep your answer inside the provided boxes. Marking will be done on scanned versions of the exams, so if you do need to go outside the box please keep your answer on the same page. Note, we have intentionally provided boxes that are much larger than necessary. Your answer does not need to fill the whole box.

10. Consider the annotation instructions for sentiment analysis below, then answer the questions:

> For each piece of text, rate its sentiment from positive to negative on a 7 point scale.

(a) (1 mark) What is a good property of these instructions?

Its granularity is much better than just good or bad which means we can provide much more clear result. Low perplexity prompts are much better.

(b) (2 marks) What are two ways these instructions could be improved?

1. give example about what positive and negative sentiment means
2. clearly indicate 0 means worst negative score and 7 means best positive

Optionsinclude: (1)addingexamples, (2)specifyingtheintermediate pointsonthesacle,(3)sayingwhattodoifacaseisambiguous,e.g.,maybedon't label itatall fornow,(4)explicitlysayingwhether1ispositiveornegative.

11. (1 mark) What is the benefit of using Precision instead of Accuracy? (in situations where either could be used)

Consider about you only care about TP cases and TN cases really does not matter, then you use Precision will get a closer look to the capability of model to predict positive class

Student Number:
Complete this on every page so we can find pages if they get separated during scanning.

12. (1 mark) What is ”In-Context Learning”?

Provide instructions and examples in the input and hope model can capture the pattern inside it.

13. (2 marks) An NLP system is applied to a task with 3 labels. On the test set, the micro F-score is much higher than the macro F-score. How is that possible?

Maybe the performance of rare class is bad.

14. One annotation approach mentioned in class was to run an automatic system and then correct its errors.

(a) (2 marks) Why could this increase error? Use an example to explain your answer.

System usually like to follow patterns, so when the pattern have certain errors, then all this kind of annotation will be wrong. For example, if "I" cannot be annotate as "PER", then the whole annotation might appear this kind of error.

我没有理解题意，这题的重点是为什么先用auto再用人工可能会导致错误增加

If annotators trust the automatic system too much, they may overlook its mistakes. For example, in sentiment analysis, a review containing "it was not not good" might confuse the model, but the annotator might miss the double negation when reading quickly.

(b) (1 mark) If the automatic system has low accuracy, what impact will that have on the cost and quality of annotation?

cost: consider about the quality, we might need to annoate it again on another way, so the cost can be think as high.

quality: since low accuracy means cannot detect both TP and TN cases, so its quality could be really bad

这里也是没有理解题意，应该是low accuracy auto对于人工的影响

It could increase the cost because annotators have to do the work they would do without the system, plus review its output. It may not impact accuracy if annotators learn not to trust the system blindly.

15. (2 marks) What causes vanishing gradients in RNNs and why?

Multiple non-linear transfer (activation function like tanh) lead to the vanishing gradient. Since repeatedly apply activation function to a value will lead it to decrease.

16. When data is collected from the web for large language model training, some entire domains (e.g., all websites whose URL contains `evil.com`) are skipped.

(a) (1 mark) Why is this done?

to make sure our data quality is good.

Student Number:

Complete this on every page so we can find pages if they get separated during scanning.

(b) (1 mark) What is a common alternative approach to filtering on domains?

FastText: if a domain's language score is lower than a threshold, then we filt it out.

17. (1 mark) What is a major disadvantage of static word embeddings?

It cannot capture the word meaning sufficiently when a word appear in different location. That's to say it does not change meaning according to contextual words

18. (1 mark) BERT has been provided the sentence below as input:

"The woman walked across the street, checking for traffic over [MASK] shoulder."

Which word would you expect to have the highest attention score when the model is determining what to replace [MASK] with?

path her

19. (2 marks) In self-attention, there is only one set of input vectors. How do the query, keys, and values differ?

Here we use one input vector as example, and we need do self-attention for all input vectors to get their contextual vector
query: the input vector we choose
key: other vectors (usually include itself) that used to compare with the query
value: vector with actual content to retrieve, values accociated with more similar keys receive higher attention weights

20. (2 marks) What is the main difference between training annotators and doing pilot annotation?

we do pilot annotation before we really do the annotation task. It just provide sample of data to analyzer to try the annotation first, and to find whether they are keep consistency.

21. You have decided to use a language model as part of a text classification project. Consider the scenarios below. For each one, explain how you would use the data with your language model in order to develop a system to solve the problem.

(a) (1 mark) You have no annotated data (text + label) to train on.

I might choose ICL to deal with this problem. Just like the K-SHOT

(b) (1 mark) You have a few **thousand** annotated examples.

I will use cross validation to help build the model, since the data size is limited.

ICL or Fine Tuning 都适合小规模数据

Student Number:
Complete this on every page so we can find pages if they get separated during scanning.

(c) (1 mark) You have a few **million** annotated examples.

I will use this data as usual . That is build a model and accoding to loss function to update model parameters.

22. (1 mark) What is the difference between a model and an inference method?

model just give output, but inference method determing the meaning of the model's output.

23. (1 mark) Why are the vectors from word2vec a dense representation?

because it use float number instead of integer number.

主要说vectors contain non-zero values in most dimensions.

24. (2 marks) A friend has scraped some data from a popular restaurant review website and used it to create a dataset for sentiment analysis. They would like to share their dataset online and have asked you for help.

What are two steps you would recommend your friend take to share the data appropriately?

PII
Check wheather those data contain bias.

get the permission from resraurant

25. (1 mark) What problem is 'Teacher forcing' intended to solve?

it want to fast the training, which prevent model keep going on the wrong direction.

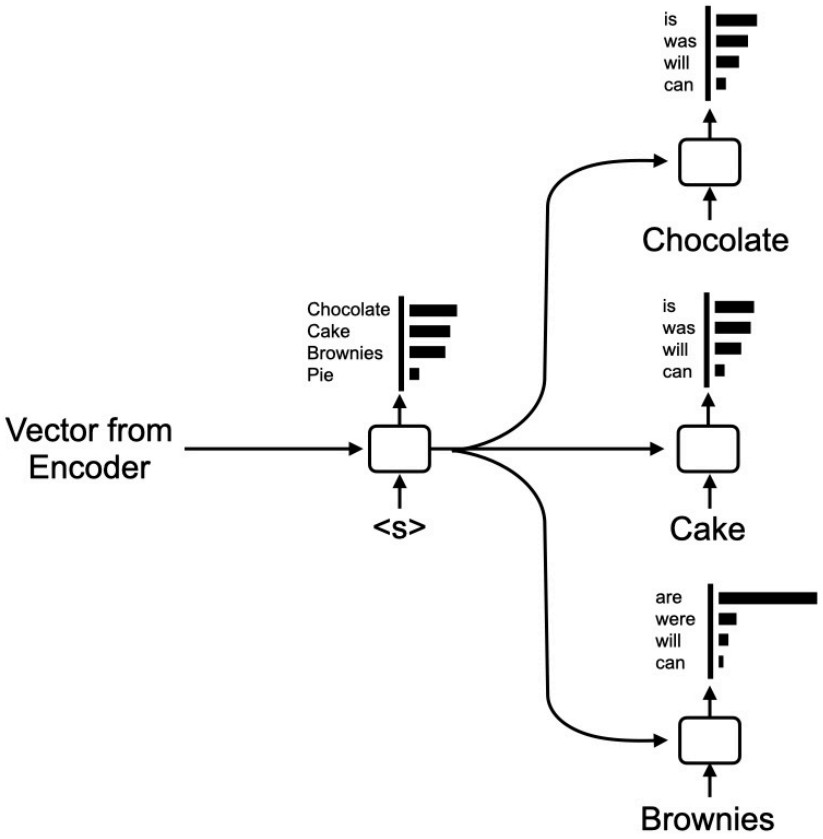
26. (2 marks) Two ways an RNN can be used are as a transducer or as an encoder. What is the difference between these?

transducer provide result for each token, but encoder provide a compress result for all tokens together (a sequence)

Student Number:

Complete this on every page so we can find pages if they get separated during scanning.

27. An encoder-decoder RNN is being used and the output is fixed at two tokens in length. The figure below shows the distribution for the first decoder step, and the distributions for three possible second steps.



(a) (1 mark) If beam search with a beam of size 3 is used, what would be the second word in the output?

are

(b) (1 mark) If beam search with a beam of size 2 is used, what would be the second word in the output?

is

28. (2 marks) What advantage do residual connections in the transformer provide?

smooth gradient, reduce gradient vanishing problem

29. (2 marks) The transformer's encoder and decoder both have a form of self-attention. How is self-attention different between the two and why?

for decoder, it is a masked self attention, that is previous cannot see later result.

Student Number:

Complete this on every page so we can find pages if they get separated during scanning.

In the next few questions, you will consider some pieces of code and answer questions about them. When asked for the purpose of the code, your answer should be describe the goal of the person who wrote the code, not describe what each line does.

30. The code below is part of a PyTorch model. Two pieces of code are marked with ”start” and ”end”. Write what the purpose of each section of code is.

```
class RNN(nn.Module):
    def __init__(self, input_size, hidden_size, output_size):
        super(RNN, self).__init__()

        # (a) start
        self.i2h = nn.Linear(input_size + hidden_size,
           hidden_size)
        self.h2o = nn.Linear(hidden_size, output_size)
        self.softmax = nn.LogSoftmax(dim=1)
        # (a) end

        # (b) start
        self.init_weights()
        # (b) end
```

(a) (1 mark)

it is a 3 layer model: include an input layer, a hiden layer, and an softmax output layer

add the input layer, it is a linear layer
add

Create the variables that store the weights / parameters of the model.

(b) (1 mark)

degine the initial weight of the model.

31. The function below is defining the forward pass in a neural network.

```
def forward(self, input_tensor, hidden):
    combined = torch.cat((input_tensor, hidden), 1)
    hidden = self.i2h(combined)
    output = self.h2o(hidden)
    output = self.softmax(output)
    return output, hidden
```

(a) (1 mark) What type of model is it?

it has a hidden layer so it could be a RNN

(b) (1 mark) Will it suffer from numerical problems? Why / why not?

No, there there is not non-linear transformation

Student Number:
Complete this on every page so we can find pages if they get separated during scanning.

Programming Questions

In the next few questions, you will be given a task and a set of lines of code to do the task. Decide which lines to use and what order to place them in. Write the line numbers in order in the grids provided (one number in each box, in order from top to bottom). Note:

- If multiple orders are correct, we will accept all correct answers.
- You do not need to indicate indentation.
- Not all lines need to be used.
- There are extra pages at the back of the exam you can use to think.
- We provide more boxes than are needed.
- If you make a mistake, clearly put a line through the numbers and write a new response in the boxes.

32. (4 marks) Using the lines below, implement dot product attention in PyTorch using a class. When used, the class should return the attention weights and the rescaled / weighted input vectors.

```
1  def __init__(self, hidden_size):
2  self.out_size = hidden_size *2
3  weights = F.softmax(scores, dim=-1)
4  context = torch.bmm(weights, keys)
5  def forward(self, query, keys):
6  scores = torch.relu(scores)
7  scores = torch.tanh(scores)
8  return weights
9  return context, weights
10 return context
11 scores = (query * keys).sum(-1).unsqueeze(1)
12 scores = (query * keys)
13 super(DotProductAttention, self).__init__()
14 class DotProductAttention():
15 class DotProductAttention(nn.Module):
```

15
1
13
2
5
11
3
4
9

33. (4 marks) Using **as few as possible** of the lines of code below, implement the Perceptron update.

```
1  guess = find_best_code(question, model, answer)
2  guess = find_best_code(question, model)
3  if guess == answer:
4  if guess != answer:
5  else:
6  def learn(question: str, answer: str, model:
       Model, find_best_code: [str, Model] -> str):
7  def learn(question: str, answer: str, model:
       Model, find_best_code: [str, Model, str] ->
       str):
8  pass
9  model.update(question, guess, 1)
10 model.update(question, answer, 1)
11 model.update(question, guess, -1)
12 model.update(question, answer, -1)
```

7	6
1	2
3	10
9	11

− 4.

34. (2 marks) Why does your solution work?

Perceptron cannot capture non linear, so both 0 or 1 would work for guess and answer

When the guess and the answer match, the two updates (10, 11) will cancel out.

When they are different, the guess will have its score decreased and the answer will have its score increased.

Student Number:

Complete this on every page so we can find pages if they get separated during scanning.

This page is left intentionally blank in case you need additional writing space. Only pages that are stapled will be scanned. Scratch paper will not be scanned.

This page is left intentionally blank in case you need additional writing space. Only pages that are stapled will be scanned. Scratch paper will not be scanned.

Student Number:

Complete this on every page so we can find pages if they get separated during scanning.

This page is left intentionally blank in case you need additional writing space. Only pages that are stapled will be scanned. Scratch paper will not be scanned.

END OF EXAMINATION