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:

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

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

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

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:

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

Metrics:

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

Cohen's Kappa:

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

TF-IDF:

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

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)
   - ■ **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**

   Solution: "The two annotators were not consistent" is the most crucial option to select. Because the question says 'could', the other indicated options are also correct.

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)
   - □ Perplexity(4) = 0
   - □ Perplexity(1, 1) > Perplexity(2)
   - ■ **Perplexity(1, 1) < Perplexity(2)**
   - ■ **Perplexity(4) is not defined**

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:

- ☐ 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**

> **Solution:** Note this question contains an ambiguity (is the list being used as a sparse vector or a vector the full size of the vocabulary). The answer above assumes it is being used as a sparse vector. If, during marking, the markers realise an ambiguity that could reasonably be interpreted two ways, we will accept either interpretation. For cases where the intended meaning is clear, we will only accept the intended meaning.

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**

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?

Solution: Options include: (1) they are concise, (2) they specify a scale.

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

Solution: Options include: (1) adding examples, (2) specifying the intermediate points on the sacle, (3) saying what to do if a case is ambiguous, e.g., maybe don't label it at all for now, (4) explicitly saying whether 1 is positive or negative.

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

Solution: Precision focuses on the class we are interested in, which may be rare, whereas accuracy will give credit for all labels, including true negatives.

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”?

Solution: When the prompt to the language model contains information to teach the model, e.g., examples of the task.

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?

Solution: This can occur if one label is rare and our system does badly on that label but it does well on the other two.

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.

Solution: If annotators trust the automatic system too much then they may not notice its mistakes. For example, sentiment analysis on a review containing ”it was not not good” might lead to a model mistake but the annotator doesn’t notice 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?

Solution: It could increase the cost because annotators have to do the work they would without a system, but also look at its output. It may not impact accuracy because annotators learn not to trust it.

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

Solution: Nonlinearities. This is because they squeeze the range of values, particularly when applied multiple times. We would also accept answers that mention small weight initialisation or long training sequences.

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?

Solution: To avoid giving the models data that is low quality or could encourage concerning outputs.

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

Solution: Have an automatic detector to decide whether a webpage is high quality and appropriate.

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

Solution: They do not account for the fact that many words can mean different things in different contexts.

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?

Solution: 'woman' is best here, but 'over' is also okay since models tend to look at the most recent word.

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

Solution: The keys and values are the same, but the query vector will be different for each word.

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

Solution: During pilot annotation the intention is to modify the annotation guide while training annotators it is not.

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.

    Solution: Use the LM to do 0-shot prediction based on prompting.

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

    Solution: Do in-context learning with a subset of the samples chosen based on similarity to the current query. We would also accept just saying in-context learning. We would also accept fine-tuning.

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

Solution: Fine tune the LM on the task. Alternatively, don't use the data at all with the LM as you are now able to juse use the LM to generate embeddings that will be used by a specialised supervised model.

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

Solution: A model provides scores to (input, output) pairs, while an inference method uses a model to find a high scoring option.

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

Solution: The vectors contain non-zero values in most dimensions. We would also accept that the vectors are much smaller than the dimensionality of the vocabulary.

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?

    Solution: Some possible answers: (1) contact the website they scraped to get permission, (2) clearly identify a suitable license with the data release, (3) filter / anonymise the data for identifying information from the users who created the reviews.

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

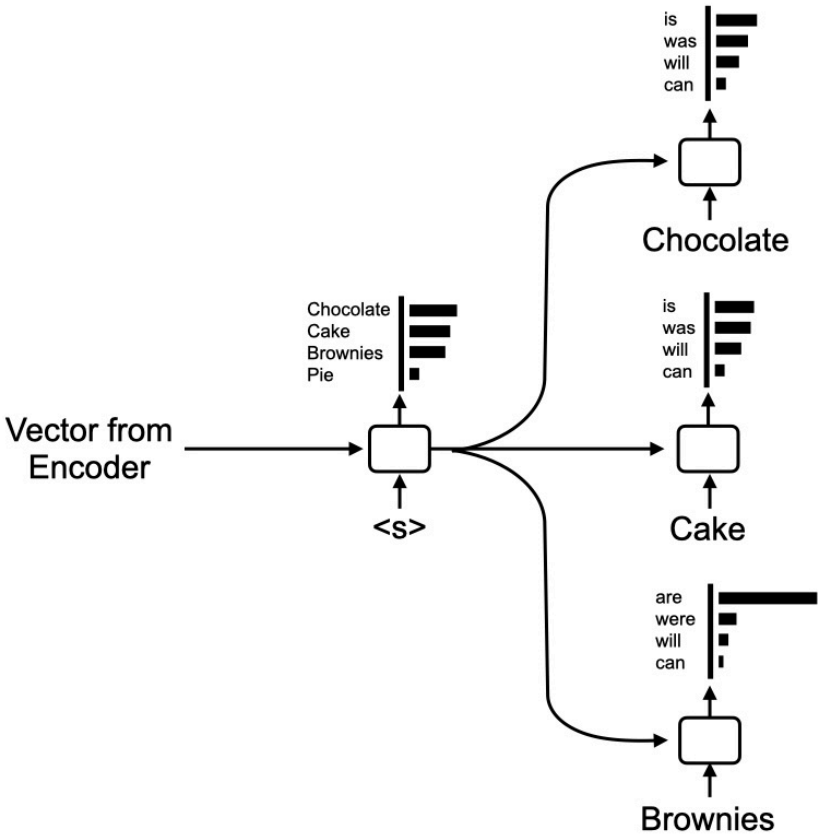
    Solution: Training can be slow if the model is generating long outputs that diverge from the reference answer.

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

    Solution: In a transducer, an output is produced for each input, but in an encoder, a single output is produced for the whole 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?

Solution: are

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

Solution: is

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

Solution: Improve training speed and smoothness.

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?

Solution: In the decoder, self-attention only looks at tokens to the left of the current one, while in the encoder every token has self-attention comparing with all other tokens. The question does not ask for why, but the reason is that in the decoder we are generating left-to-right, so when we are processing input N we don't know what inputs N+1, N+2, N+3, etc are.

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)

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

(b) (1 mark)

Solution: Set the weights / parameters to initial values.

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?

Solution: An RNN

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

Solution: No, because it does not have a non-linearity.

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):
```


Solution: 15, 1, 13, 2, 5, 11, 3, 4, 9. Note, in this case all boxes were used, but that will not always be the case.

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)
```


Solution: 6, 2, 10, 11

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

Solution: When the guess and the answer match, the two updates 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