Mock Endterm Exam
Introduction to Applied Data Science

Name: _____ Student No: _____

2026-05-17

Instructions

This mock endterm exam contains **20 multiple choice questions** and **5 open questions**. For the MC questions, encircle the letter of your chosen answer. All MC questions carry equal weight (1 pt each). The open questions are worth 3 pts each. The **total is 35 points**. The duration of this exam is **2 hours**. You may only ask clarifying questions during the first 30 minutes of the exam. No computers, phones, or internet access are permitted. Don't forget to fill in your name and student number on this page. Good luck!

Part A: Multiple Choice Questions (20 points)

1. You have a corpus of five economic news articles. The word "market" appears in all five articles, while "quantitative easing" appears in only one. After applying TF-IDF weighting, which of the following correctly describes the outcome?

A) "market" receives a higher TF-IDF score than "quantitative easing" because it appears more frequently overall.
B) "quantitative easing" receives a higher TF-IDF score than "market", because its low document frequency produces a higher IDF component.
C) Both words receive the same TF-IDF score because TF-IDF normalises for document length.
D) "market" receives a negative TF-IDF score because words appearing in all documents are penalised below zero.

---

2. A researcher fits an LDA model with `k = 3` and inspects `tidy(model, matrix = "beta")`. The data frame contains columns `topic`, `term`, and `beta`. What does a high `beta` value for a given `(topic, term)` pair indicate?

A) The document is strongly associated with that topic.
B) That term appears in a large proportion of the overall corpus.
C) That term has a high probability of being generated by that topic.
D) The topic has a high prior probability of appearing in any document.

---

3. In the CBOW variant of Word2Vec, a model is trained on economic texts. The hidden layer is computed as the **average** of the context word embeddings. Why does this averaging operation give CBOW its "bag of words" character?

A) Because context words are weighted by inverse distance before averaging, so nearby words contribute more than distant ones.
B) Because the averaging discards word order information — only which words appear in the context matters, not their sequence.
C) Because the model uses a bag of documents rather than a bag of individual tokens.
D) Because the output embedding is averaged with the input embedding after each training step.

---

4. In the transformer self-attention mechanism, what is the mathematical role of the **Key** (K) vector for a given token?

A) It stores the weighted sum of all other tokens' Value vectors passed to the next layer.
B) It is compared with Query vectors of other tokens to compute attention scores.
C) It represents what the current token is searching for among the other tokens.
D) It is multiplied by the position encoding to preserve word order information.

---

5. Positional encoding is added to word embeddings in transformers. Why is this necessary?

A) Because the softmax function in attention requires inputs to be ordered by magnitude.

B) Because without it, self-attention processes all tokens simultaneously and cannot distinguish "inflation rose then fell" from "inflation fell then rose".
C) Because positional encoding replaces the Key vectors to reduce computational cost.
D) Because word embeddings are always zero-mean, so positional encoding shifts them into positive space.

---

6. Consider the following `ellmer` code:

```
chat <- chat_ollama(model = "gemma3")
result <- chat$chat_structured(
  text,
  type = type_object(
    country  = type_string(),
    gdp_2023 = type_number(required = FALSE)
  )
)
```

The input `text` mentions a country but contains no information about GDP. What will `result$gdp_2023` contain?

A) `0`, because numeric fields default to zero when the schema is satisfied.
B) `NA`, because `required = FALSE` permits the field to be absent, returning `NA` in R.
C) A plausible GDP figure generated by the model from its training data.
D) The call will fail with an error since `type_number()` requires a value to be present.

---

7. A data scientist wants to use `parallel_chat_structured()` to extract a product name (string) and unit price (number) from each of 200 product descriptions. Which `ellmer` type specification correctly defines the schema for one row of the output?

A) `type_array(type_string(), type_number())`
B) `type_object(product = type_string(), price = type_number())`
C) `type_array(type_object(product = type_string(), price = type_number()))`
D) `type_object(type_array(product = type_string(), price = type_number()))`

---

8. Which of the following best describes how the **tool-calling** workflow operates in `ellmer`?

A) The LLM directly executes the R function using a sandboxed interpreter built into the model.
B) The LLM identifies when a tool is needed, produces a structured request, and the R session executes the function and returns the result to the LLM.
C) Tools are run automatically before each user prompt to pre-load context into the model.
D) `register_tool()` copies the function's source code into the LLM's prompt for execution.

---

9. In a RAG system built with the `ragnar` package, what is the purpose of calling `ragnar_store_build_index()` after all document chunks have been inserted?

A) It generates embeddings for each chunk using the specified embedding model.
B) It converts all inserted markdown chunks into JSON for storage in the database.
C) It finalises the search index so that `ragnar_retrieve()` can efficiently find relevant chunks for a given query.
D) It registers the store as an `ellmer` tool so the LLM can call it automatically.

---

10. You compute `st_intersects(hospitals, districts, sparse = FALSE)` where `hospitals` is an `sf` data frame with 30 rows and `districts` has 12 rows. What is the type and dimensions of the result?

A) A 30-element logical vector: `TRUE` if the hospital intersects at least one district.
B) A 12-element logical vector: `TRUE` if the district contains at least one hospital.
C) A 30 × 12 logical matrix: `TRUE` in cell (i, j) if hospital i intersects district j.
D) A list of 30 integer vectors, each containing the indices of intersecting districts.

---

11. You have a projected `sf` object (`nc_proj`, CRS EPSG:32119) and run:

```
wake_buf <- st_buffer(nc_proj[nc_proj$NAME == "Wake", ], dist = 5000)
```

What does `dist = 5000` mean in this context, and why does it only work correctly because the CRS is projected?

A) `dist = 5000` means 5000 degrees; projected CRS converts degrees to metres automatically.
B) `dist = 5000` means 5000 metres; projected CRS uses metres as its unit, so the buffer radius is correctly 5 km. A geographic CRS (degrees) would produce a meaningless result.
C) `dist = 5000` means 5000 feet; projected CRS converts feet to metres internally.
D) `dist = 5000` works the same regardless of CRS; `st_buffer()` always interprets the value as kilometres.

---

12. A researcher runs `st_join(firms, regions)` where `firms` is a point `sf` object with 80 rows and `regions` is a polygon `sf` object. Five of the firm points fall outside all regions. What does the result contain?

A) 75 rows — only firms that fall inside a region are retained.
B) 80 rows — all firms are retained; firms outside any region have `NA` for the region attributes.
C) 80 rows plus 5 extra rows with duplicated geometry to indicate unmatched firms.
D) An error, because `st_join()` requires both inputs to have the same number of rows.

---

13. Which spatial predicate is most appropriate for identifying counties that **share a border** with a treated region but do **not** overlap it — for example, to define a spillover zone in a spatial regression discontinuity design?

A) `st_intersects()`
B) `st_within()`
C) `st_touches()`
D) `st_is_within_distance()`

---

14. You load a satellite image with `img <- read_stars("landsat.tif")` and inspect it with `dim(img)`, which reports `x = 500, y = 400, band = 7`. What does the `band` dimension represent in this `stars` object?

A) The seven coordinate reference systems available for re-projecting the image.
B) Seven spectral or attribute layers stored in the array — for a Landsat image, each band corresponds to a different wavelength channel recorded by the sensor.
C) Seven time steps, indicating the image was acquired on seven different dates.
D) The number of distinct geographic regions (polygons) that the raster covers.

---

15. Consider this simplified 3-dimensional word embedding:

```
"bond"     = [ 0.6,  0.8, -0.1]
"equity"   = [ 0.5,  0.7,  0.2]
"default"  = [-0.7,  0.2,  0.4]
"coupon"   = [ 0.8,  0.6, -0.2]
```

Cosine similarity between "bond" and "coupon" will be close to 1. What does this indicate?

A) The two words appear in exactly the same documents in the training corpus.
B) The magnitude of the "bond" vector is approximately equal to the magnitude of the "coupon" vector.
C) The two words point in nearly the same direction in embedding space, implying similar contextual usage patterns.
D) The dot product of the two vectors equals zero.

---

16. A student sets temperature to `2.0` when generating text with an LLM to analyse monetary policy statements. How will this affect the model's outputs compared to a temperature of `0.2`?

A) The model will generate longer responses because higher temperature increases the maximum token budget.
B) The model will be more deterministic, consistently choosing the single most probable next token.
C) The model will produce more varied and less predictable outputs, because probability mass is spread more evenly across tokens.
D) Temperature above 1.0 causes the model to ignore the system prompt.

---

17. In LDA, each document is represented as a **mixture** of topics rather than a single topic assignment. Which row of the output from `tidy(model, matrix = "gamma")` captures this mixture for a specific document?

A) The row where `beta` is largest for that document's most frequent word.
B) All rows corresponding to that document, showing its probability distribution across all `k` topics.
C) A single row per document, containing the index of its most probable topic.
D) The row with the smallest `gamma` value, indicating the least likely topic.

---

18. An economist runs `bb <- getbb("Amsterdam, Netherlands")` and receives a 2 × 2 matrix. She wants to compute the approximate longitude of the city centre by averaging the western and eastern extents of the bounding box. Which R expression correctly does this?

A) `mean(bb["y", ])` — averages the values in the latitude row.
B) `mean(bb["x", ])` — averages the minimum and maximum longitude values.
C) `bb[2, 1]` — takes the minimum latitude value from the matrix.
D) `sum(bb) / 4` — divides the sum of all four matrix values by 4.

---

19. You have two `sf` data frames: `plants` (factory locations, EPSG:4326) and `zones` (industrial policy zones, EPSG:28992). You write:

```
result <- st_join(plants, zones)
```

Why might this produce incorrect or unexpected results?

A) `st_join()` does not support point-in-polygon joins; it only works with polygon-polygon joins.
B) The two datasets are in different coordinate reference systems; spatial predicates may silently fail or return wrong matches without first aligning CRS.
C) `plants` must be the second argument in `st_join()` when it is the smaller dataset.
D) EPSG:28992 is not a valid CRS for the `sf` package and will cause an error.

---

20. A researcher asks an LLM (without RAG): "What was the exact unemployment rate in the Netherlands in Q4 2025?" The model confidently responds: "The unemployment rate in Q4 2025 was 3.2%." The actual figure was 3.9%. This error most likely reflects:

A) A temperature setting that was too high, introducing random noise into the numeric output.
B) Hallucination — the model generated a plausible-sounding number based on patterns in training data rather than factual retrieval.
C) A tokenisation error — numbers are split into subword tokens, causing arithmetic mistakes.
D) A CRS mismatch in the model's internal geographic reasoning.

Part B: Open Questions (15 points)

Question 1 (3 pts)

A student builds a text analysis pipeline for a corpus of IMF country report excerpts:

```
library(tidytext)
library(dplyr)
library(topicmodels)

reports <- data.frame(
  doc_id = c("R1", "R2", "R3", "R4", "R5"),
  text   = c(
    "fiscal deficit public debt government spending budget",
    "trade balance exports imports current account deficit",
    "inflation consumer prices monetary tightening interest rates",
    "public debt fiscal consolidation government revenue spending",
    "interest rates inflation monetary policy central bank"
  )
)

tokens <- reports |>
  unnest_tokens(word, text) |>
  count(doc_id, word)

dtm <- cast_dtm(tokens, document = doc_id, term = word, value = n)
lda_out <- LDA(dtm, k = 2, control = list(seed = 7))

beta_df  <- tidy(lda_out, matrix = "beta")
gamma_df <- tidy(lda_out, matrix = "gamma")
```

Suppose `beta_df` shows that "fiscal", "debt", and "spending" have high beta values for Topic 1, while "inflation", "monetary", and "rates" have high beta values for Topic 2. The `gamma_df` values for documents R1 and R3 are approximately `[0.91, 0.09]` and `[0.08, 0.92]` respectively.

Interpret these LDA results. In your answer, explain what the `beta` values for Topic 1 represent and what economic theme Topic 1 likely captures, and explain what the `gamma` values for R1 and R3 tell you about how these documents relate to the two topics.

Question 2 (3 pts)

A researcher uses the following `ellmer` code to extract structured information from a set of ECB press releases:

```
library(ellmer)

chat <- chat_openai(model = "gpt-4o")

type_release <- type_object(
  release_date    = type_string(),
  rate_decision   = type_string(),
  basis_points    = type_integer(required = FALSE),
  primary_concern = type_string()
)

result <- chat$chat_structured(press_release_text, type = type_release)
```

The researcher notices that `basis_points` is sometimes `NA` even when the press release clearly states "the Governing Council decided to raise rates by 25 basis points." A colleague then suggests replacing `chat_openai()` with `chat_ollama()` to avoid sending sensitive policy documents to an external server.

Explain why `basis_points` can return `NA` and how `required = FALSE` relates to this behaviour. Then assess the colleague's suggestion: what is the privacy advantage of switching to `chat_ollama()`, and what capability trade-off should the researcher be aware of?

Question 3 (3 pts)

Consider the following attention scores computed for the word **"it"** in the sentence:

"The subsidy helped farmers because it reduced input costs."

Word	Pre-softmax score
The	0.1
subsidy	2.4
helped	0.5
farmers	1.9
because	0.2
it	0.3
reduced	0.7
input	0.4
costs	0.6

Identify which word receives the highest attention from "it" and explain what this reveals about how the self-attention mechanism resolves pronoun reference. A student claims that "farmers" receives a high score simply because it is positioned close to "it" in the sentence. In your answer, also evaluate whether this proximity argument is correct and explain what actually determines the attention scores.

Question 4 (3 pts)

An urban planner wants to count the total number of schools that are located within 500 meters of a metro station. She has two spatial datasets: - `schools`: An `sf` data frame of school locations (point geometry) in CRS **EPSG:25832** (UTM zone 32N, projected in meters). - `stations`: An `sf` data frame of metro stations (point geometry) in CRS **EPSG:4326** (WGS 84, geographic in degrees).

She writes the following R code to achieve this:

```
# Create 500m catchment areas around metro stations
stations_buffered <- st_buffer(stations, dist = 500)

# Find schools that intersect these catchments
nearby_schools <- st_intersects(schools, stations_buffered)

# Count the total number of schools near a station
total_count <- sum(nearby_schools)
```

Identify **two distinct errors** in this code. For each error, explain why the code fails and provide the corrected R code.

Question 5 (3 pts)

A policy analyst builds a RAG system to answer questions about Dutch housing policy documents. She uses the following workflow:

```
library(ragnar)
library(ellmer)

store <- ragnar_store_create(
  "housing_policy.duckdb",
  embed = \(x) ragnar::embed_ollama(x, model = "embeddinggemma")
)

chunks <- read_as_markdown("housing_report_2025.pdf") |>
  markdown_chunk()

ragnar_store_insert(store, chunks)
ragnar_store_build_index(store)

client <- chat_ollama(model = "llama3.1")
ragnar_register_tool_retrieve(
  client, store, top_k = 4,
  description = "2025 Dutch housing policy document"
)

client$chat("What is the government's target for new housing units by 2030?")
```

Explain step by step what happens when `client$chat(...)` is called, describing how the LLM decides to use the retrieval tool and what happens with the retrieved text before the final answer is produced. The analyst finds that the model still occasionally produces incorrect figures even though the correct number appears in the document — in your answer, also name and explain one reason why RAG does not guarantee correct outputs.

End of Exam