Practice Packet v2: Q&A + Tables + Formulas + Images

CPU-friendly benchmark PDF for MinerU/Markdown conversion and regex-based question/answer extraction.

Marker	Questions begin with “Qn.”; answers begin with “Answer:”
Content	25 questions • sub-questions • tables • images • formulas • code block
Hint	Use a text-only stream for regex mining and keep tables/images as atomic blocks.

Quick Tasks

1. Convert this PDF to Markdown (MinerU pipeline on CPU is enough).
2. Mine the Q/A template using regex (no LLM extraction).
3. Extract items into JSON with page numbers and asset references.
4. Treat tables, images, and code blocks as atomic objects.

This version contains more varied patterns, including a code block and a small flow diagram image.

Section A — Concepts and Metrics

Q1. What does accuracy measure?

Answer: The fraction of all predictions that are correct: (TP+TN)/(TP+TN+FP+FN).

Q2. Give one reason accuracy can be misleading on imbalanced data.

Answer: If the negative class dominates, a model can predict all negatives and still achieve high accuracy.

Q3. Define the F1 score.

Answer: The harmonic mean of precision and recall: F1 = 2·(precision·recall)/(precision+recall).

Q4. A model has precision 0.75 and recall 0.60. Compute F1.

Answer: F1 = 2*(0.75*0.60)/(0.75+0.60) = 0.6667 (approx).

Q5. Explain the difference between MAE and MSE.

Answer: MAE averages absolute errors; MSE averages squared errors and penalizes larger errors more strongly.

Q6. If MAE is 0.20 and MSE is 0.10, can MSE be smaller than MAE?

Answer: Yes—MSE is in squared units; numeric comparison depends on scale. Comparing RMSE to MAE is more meaningful.

Q7. (a) What is a confusion matrix? (b) Name its four basic entries.

Answer: (a) A table summarizing classification results. (b) TP, TN, FP, FN.

Q8. In one sentence, what is overfitting?

Answer: When a model fits training noise and fails to generalize to new data.

Q9. What does early stopping do?

Answer: Stops training when validation performance stops improving to prevent overfitting.

Q10. List two common regularization techniques.

Answer: Dropout and L2 weight decay (also: data augmentation, early stopping).

Section B — Regex and Parsing Patterns

Q11. Why should tables be treated as atomic blocks during regex mining?

Answer: Tables often contain numbers and punctuation that can look like question markers, causing false positives.

Q12. Suppose a document uses “Question 1:” instead of “Q1.”. What should your mining step do?

Answer: Cluster candidate header lines and learn the dominant prefix family; then generalize spacing/punctuation in the regex.

Q13. Consider this code block. Should the parser search for “Q” markers inside it?

Answer: No. Code blocks should be excluded/placeholdered before regex mining and parsing.

Code Sample 1 (for exclusion testing):

```
for i in range(1, 6):
print(f"Q{i}. This is not a real question marker inside code")
```

Q14. Name two stop conditions that can end an answer block.

Answer: The next question marker, a new section heading, or the end of document.

Q15. What is a safe strategy when your first regex family yields too many questions?

Answer: Tighten the start pattern (e.g., require “Q” prefix) and add exclusions for captions/tables/headers.

Q16. Explain why caching by PDF hash helps Streamlit performance.

Answer: Streamlit reruns the script often; caching prevents re-conversion and re-parsing for the same PDF.

Section C — Tables

Table 1: Confusion Matrix

	Predicted Positive	Predicted Negative
Actual Positive	TP = 42	FN = 8
Actual Negative	FP = 10	TN = 140

Q17. Using Table 1, compute accuracy and precision.

Answer: Accuracy=(42+140)/(42+140+10+8)=182/200=0.91. Precision=42/(42+10)=0.8077.

Table 2: Mini Dataset

Month	Users	Revenue ($k)
Jan	1200	18.5
Feb	1350	20.1
Mar	1280	19.2
Apr	1500	23.7
May	1580	24.9
Jun	1700	27.4

Q18. From Table 2: (a) average revenue, (b) month with max users.

Answer: (a) Average revenue = 22.30 ($k). (b) Max users occur in Jun (1700).

Section D — Images

Q19. Refer to Figure 1. What is sin(π/2) and the period of sin(x)?

Answer: sin(π/2)=1 and the period is 2π.

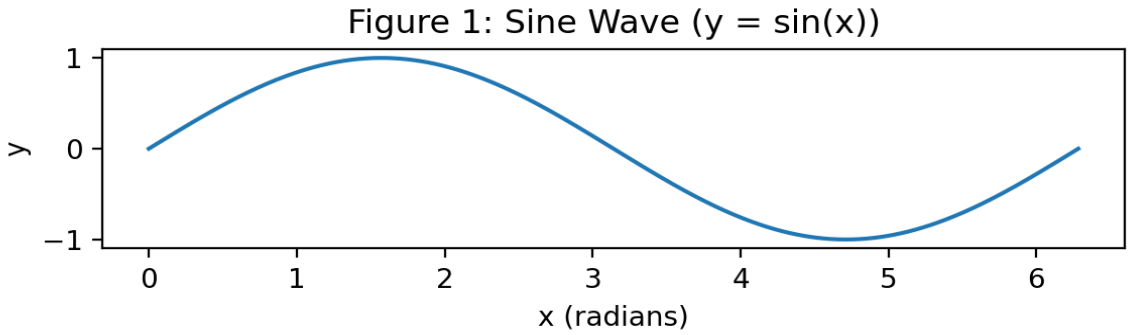

Figure 1: Sine Wave (y = sin(x))

Q20. Refer to Figure 2. What is the increase in sales from Mar to Jun?

Answer: Sales increase from 15 ($k) to 25 ($k), so the increase is 10 ($k).

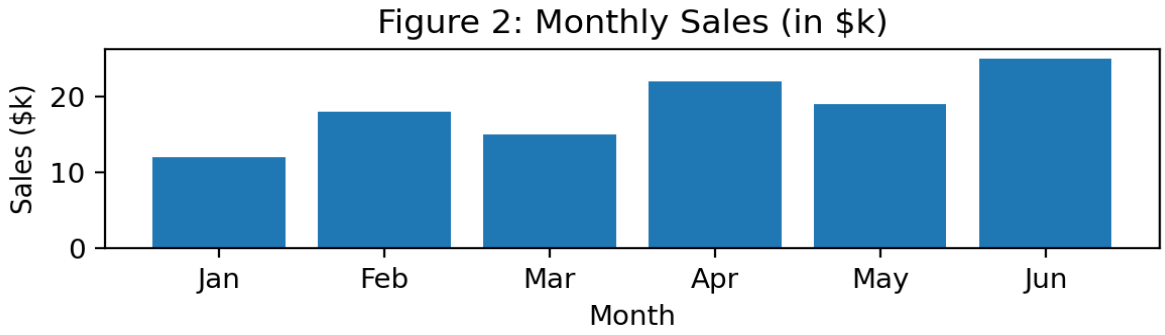

Figure 2: Monthly Sales (in $k)

Section E — Formulas

Q21. Solve $2x^2 - 3x - 2 = 0$ using the quadratic formula.

Answer: a=2, b=-3, c=-2, discriminant=25. Solutions: x=(3±5)/4 → x=2 or x=-0.5.

$$\text{Quadratic Formula: } \quad x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Formula 1: Quadratic Formula

Q22. Evaluate $\int_0^1 x^3$ dx and state the general identity.

Answer: ∫0^1 x^3 dx = 1/4. In general, ∫0^1 x^n dx = 1/(n+1).

$$\text{Integral: } \quad \int_0^1 x^n \, dx = \frac{1}{n+1}$$

Formula 2: Integral Identity

Q23. Write the softmax function and explain what it outputs.

Answer: Softmax maps a vector of scores into a probability distribution that sums to 1.

$$\text{Softmax: } \quad \sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

Formula 3: Softmax

Section F — End-to-End Pipeline (Image)

Q24. Based on Figure 3, list the five pipeline stages in order.

Answer: Upload PDF → MinerU Convert → Markdown Normalize → Regex Mine Rules → Parse Q/A → JSON.

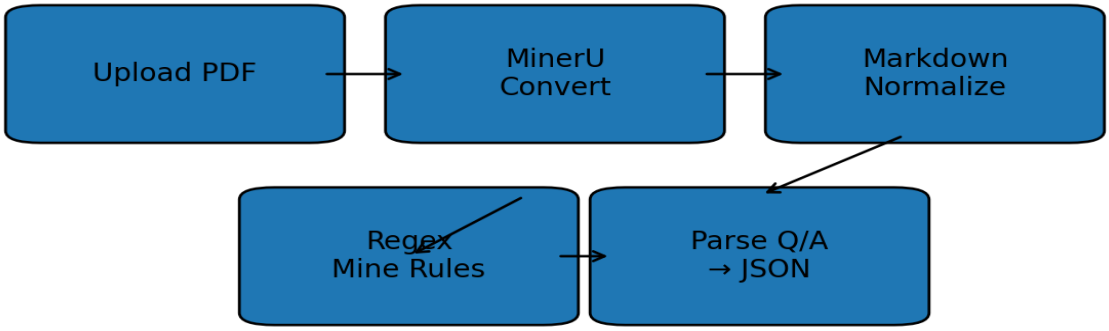

Figure 3: High-level pipeline diagram

Q25. Give one reason to keep a separate “ruleset.json” artifact alongside the extracted Q/A JSON.

Answer: It makes extraction reproducible and debuggable by recording the exact regex patterns and boundaries used.

End of Packet