# Practice Packet v3

Watermarks • highlights • underline • bold/italic • uncommon fonts • tables • images • formulas

CPU-friendly benchmark PDF for MinerU/Markdown conversion and regex-based Q/A extraction.

| | |
|---|---|
| **Marker** | Questions begin with “Qn.” (sometimes as headings); answers begin with “Answer:”. |
| **Styling** | Includes bold, italic, underline, and highlighted text. |
| **Watermark** | Pages 2 and 5 include a diagonal DRAFT CONFIDENTIAL watermark. |
| **Fonts** | Uses Andika (sans), Charis SIL (serif), DejaVu Sans Mono (code), and a Chinese font sample. |
| **Hint** | Use a text-only stream for regex mining; keep tables/images/math/code as atomic blocks. |

Chinese sample (font stress-test)：中文示例：数据提取测试（标题、表格、图片、公式）

# 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: $F_1 = \frac{2PR}{P+R}$.

**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.

## Code Sample (should be treated as an atomic block and excluded from Q markers):

~~~
```python
for i in range(1, 4):
    print(f"Q{i}. This is not a real question marker inside code")
```
~~~

## Section B - Tables (should become HTML in MinerU Markdown)

### Table 1: Confusion Matrix

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

**Q6. Using Table 1, compute accuracy and precision.**

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

### Table 2: Mini Dataset (with uncommon font in header)

| 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 |

**Q7. From Table 2: (a) average revenue, (b) month with max users.**

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

## Section C - Images (Markdown ![] (...) + caption lines)

### Q8. Refer to Figure A. What is sin(pi/2) and the period of sin(x)?

Answer: $\sin(\pi/2)=1$ and the period is $2\pi$.

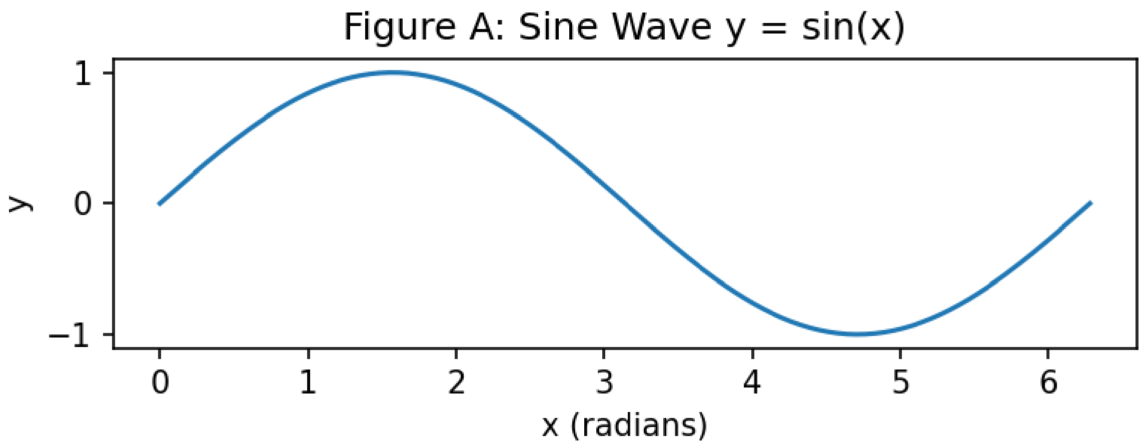

Figure A: Sine Wave (y = sin(x))

### Q9. Refer to Figure B. What is the increase in sales from Mar to Jun?

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

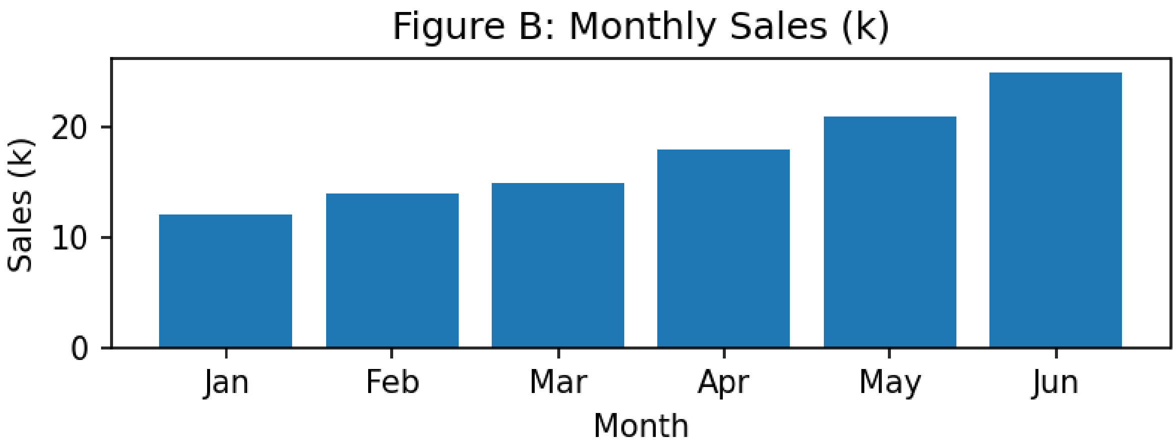

Figure B: Monthly Sales (in k)

# Section D - Formulas and Typography

### Q10. Solve 2x^2 - 3x - 2 = 0 using the quadratic formula.

Answer: Let $a=2$, $b=-3$, $c=-2$. Discriminant $b^2-4ac = 25$. Solutions: $x=(3 \pm 5)/4$ so $x=2$ or $x=-0.5$.

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Formula D: Quadratic Formula (rendered)

Formula D: Quadratic Formula (rendered image)

### Q11. Evaluate the integral and state the identity.

Answer: $\int_0^1 x^3\,dx = 1/4$. In general, $\int_0^1 x^n\,dx = \frac{1}{n+1}$.

$$
\int_{0}^{1} x^{n} \, dx = \frac{1}{n+1}
$$

## Typography stress lines:

• *Italic serif* • **Bold sans** • `Monospace`

Highlighted fragment: do not match Q markers inside tables/images/code.

# Section E - End-to-End Pipeline (Image + watermark earlier pages)

**Q12. Based on Figure C, list the pipeline stages in order.**

Answer: Upload PDF -> MinerU Convert -> Markdown Normalize -> Regex Mine -> Parse to JSON.

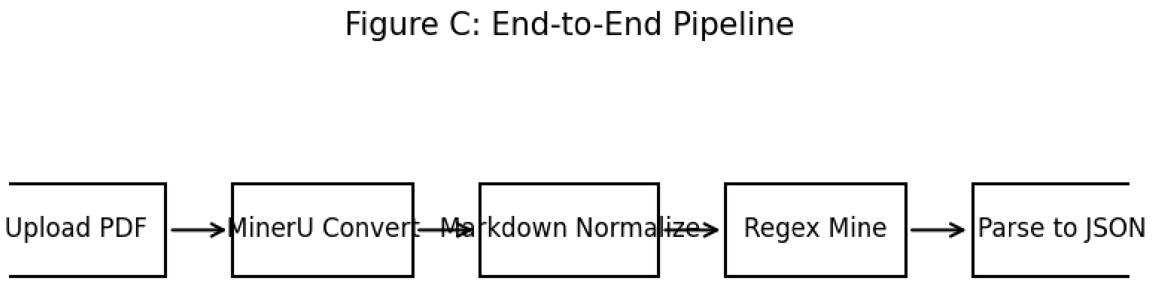

Figure C: High-level pipeline diagram

**Q13. Give one reason to keep a separate ruleset.json artifact.**

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

End of Packet