

APPENDIX E

Quick-Reference Card

Keep this page beside you while you work. Every item points back to the chapter that explains it.

The Five Questions

- **1. Is the data right?** Blanks, impossible values, duplicates (Chapter 3).
- **2. Is it right where it matters?** Check the worst cases and each group, not the average (Chapters 2 and 4).
- **3. Is it fair?** Compare groups; change one detail and see if the answer changes (Chapter 6).
- **4. Can it be attacked?** Tricks, hidden instructions, off-topic requests (Chapters 5 and 7).
- **5. Will it hold up?** Re-run your test every month and compare (Chapter 8).

Before you test

- Write pass rules with numbers and a minimum sample size, before you see any results.
- Lock the test set. Never tune the model against it.
- Agree which fairness measure applies, and write down why.

Choosing a measure

- **Missing a real case hurts most:** watch recall.
- **False alarms hurt most:** watch precision.
- **Rare events:** never accuracy alone.
- **Predicting a number:** MAE, compared with a simple baseline.

Numbers worth remembering

- **Small samples are shaky:** 27 passes out of 30 means somewhere between about 74% and 97%. 19 out of 20 means about 76% to 99%.
- **Four-fifths rule:** a selection-rate ratio below 0.80 is a warning sign.
- **Drift (PSI):** below 0.1 is stable, 0.1 to 0.25 is worth watching, above 0.25 means investigate.
- **Golden set:** start with 50 to 200 questions, run each three to five times, and include at least ten attacks.
- **AI judge:** calibrate against 30 to 50 answers scored by people.

Language models and agents

- Test facts and properties, not exact sentences.
- For RAG, test retrieval and generation separately.
- Put hard limits in code around the model.
- Re-run the whole suite after every change to the prompt, model or documents.

Before release

- Walk through the go/no-go checklist in Chapter 10.
- Report in this order: recommendation, three-sentence summary, results against exit criteria, top risks, what was not tested, decisions needed.
- Say it in this order: recommendation, one or two numbers, what you did not test, a question.