By Ethan Lawrie

Session 2 - Safe and Responsible AI Use

PRACTICAL AI LITERACY FOR WORK

Learning Outcomes Overview

KEY SKILLS TO ACHIEVE BY SESSION END

- Recognize sensitive classes of data and whether to redact them or not use AI.
- Keep generations with AI grounded in cited sources and making sure outputs are constrained.
- Run a simple check and log workflow that has an audit trail for reference later.
- Classify common risk scenarios and select guardrails for them.

Agenda Overview

PRE-QUIZ

LEARNING 1

ACTIVITY 1

Agenda Overview Cont.

LEARNING 2 **ACTIVITY 2** **LEARNING 3** **POST QUIZ**

Pre Quiz

This quiz is to gauge your current understanding of AI, high correct rates are not expected

Pre Quiz - Question 1

WHAT IS THE MOST SENSITIVE TO PASTE INTO AN AI TOOL?

A) “Total: $4,870”
B) “Monday 12 Feb”
C) “Bob Smith, 0401 234 567”
D) “Project ‘Pear’”

Pre Quiz - Question 1

WHAT IS THE MOST SENSITIVE TO PASTE INTO AN AI TOOL?

C) “Bob Smith, 0401 234 567”

Pre Quiz - Question 2

YOU NEED HELP DRAFTING A REPLY TO A CUSTOMER EMAIL. WHAT IS THE BEST FIRST STEP?

A) Paste the email thread into the AI
B) Delete names, emails, and IDs first
C) Make up fake details instead
D) Upload the whole email thread to a public site

Pre Quiz - Question 2

YOU NEED HELP DRAFTING A REPLY TO A CUSTOMER EMAIL. WHAT IS THE BEST FIRST STEP?

B) Delete names, emails, and IDs first

Pre Quiz - Question 3

YOU MUST ANSWER A RULE FROM A POLICY. BEST PLAN?

A) Ask model to “remember the policy”
B) Just write the policy and revise later
C) Provide the relevant text snippets and ask for an answer with citations (doc_id + page + date)
D) Ask the model to generate a policy of its own

Pre Quiz - Question 3

YOU MUST ANSWER A RULE FROM A POLICY. BEST PLAN?

C) Provide the relevant text snippets and ask for an answer with citations (doc_id + page + date)

Pre Quiz - Question 4

WHICH IS A LIKELY AI HALLUCINATION SIGN?

A) Tells you the doc IDs and dates
B) Exact quote with page markers
C) Confident claim with no source
D) “Cannot determine from provided sources”

Pre Quiz - Question 4

WHICH IS A LIKELY AI HALLUCINATION SIGN?

C) Confident claim with no source

Pre Quiz - Question 5

WHEN SHOULD YOU AVOID USING AN AI TOOL?

A) Brainstorming email subject lines you could use
B) Summarising public FAQs
C) Drafting a legal termination decision
D) Writing post about a public event for socials

Pre Quiz - Question 5

WHEN SHOULD YOU AVOID USING AN AI TOOL?

C) Drafting a legal termination decision

Pre Quiz - Question 6

SAFEST WAY TO INCLUDE A CUSTOMER EMAIL CHAIN?

A) Paste the raw email chain

B) Paste the raw email chain but add “confidential”

C) Paste a redacted summary with placeholders for sensitive parts

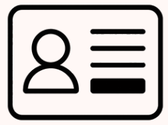
D) Link a public URL to the chain

Pre Quiz - Question 6

SAFEST WAY TO INCLUDE A CUSTOMER EMAIL CHAIN?

C) Paste a redacted summary with placeholders for sensitive parts

Data Classes in AI



Personal (PII)

Examples
names, emails, phone, addresses, employee IDs

How to handle
Redact or make up placeholders when prompting

Sensitive Personal

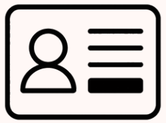
Examples
health, biometrics, age, nationality

How to handle
Do not paste into public tools

Confidential Business

Examples
prices, business forecasts, customer lists, contracts, roadmaps, security keys

How to handle
Redact or make up placeholders when prompting



Public/Approved

Examples
press releases, published docs, generic templates

How to handle
OK to use but still make sure to have sources.

AI Consent Rules

- Always just assume that prompts leave your control in public AI as it is online
- Only use tools that are approved from your organisation for non public info.
- Get the owner of the data to OK it for when using customer or staff data
- Try to minimise and send the least amount of data needed - this way there is less risk that sensitive data is exposed

Activity: Data Redaction Lab

In pairs, work through the worksheet by having one person redact, and one person to check. Once you are at halfway, switch roles.

Grounding AI with sources

Ensuring that AI answers from evidence you can cite is crucial for ensuring proper and correct outputs from it.

- What: Get the AI to use facts from only trusted documents, and not just its memory.
- Why: This makes sure that it has fewer mistakes in its answers, and also is traceable when referencing information
- How - Get the AI to:
  - Cite the doc ID/title and date/version.
  - Include a short quote from the source that was used.
  - Prefer the newest document that is relevant to teh problem.
  - If sources conflict, state that that is the case and cite both.
- Always ask it to prioritise higher trust sources:
  - Internal policy or contract > Official company system > Public website/news > the AIs previous knowledge

Hallucinations

Hallucinations are when AI thinks that it is correct and will give false information that it can confidently say is a fact or a source. It introduces risks with its outputs in hallucination cases as it will sound correct but actually not be. Grounding with sources helps prevent this.

Constraining AI Outputs

Why set constraints?

Setting constraints for the output of an AI model will lead to it making fewer mistakes in its responses, and will be easier to check and reuse.

Good constraints to set

- Format: pick one text format suits what you need for an output. for example JSON format, table, or a bulletpoint list.
- Fields: Try to be specific with the exact names and the order that you want them
- Length caps: Specify how long each field in the response should be and total word count etc.
- Numbers & dates: units + format (e.g., AUD, YYYY-MM-DD).
- Allowed values: You can specify whether you want lists or status (true/false)
- Citations: ask the AI to include the source id and date for each claim.
- Redactions: placeholders only, no real personal information.
- Uncertainty rule: if evidence missing, get it to say “cannot determine from provided sources.”

Activity: Constraining Output

In this activity, you will be working with a real AI model to constrain certain tasks outputs to be in the format that you will build. Complete the task with the worksheet in pairs of two.

Bias - Detecting and Controlling it

Models can reflect bias in their data. Check the wording, examples, and scope to make sure it the responses aren’t being affected.

What to watch for

- Outputs where it is being exclusionary or stereotyping.
- When its giving just general advice instead of using the context given to it.
- When its giving examples that are specifically for one group or demographic.
- Giving generic answers and doesn’t consider outliers.

Where bias creeps in

- Data and information from the past which reflect imbalances. eg using 10 years of past hiring data (which might be biased) to check new resumes.
- When the AI doesn’t have clear initial direction.
- When it is only getting context from one team or place
- When the AI is over depending on the sources it was provided too much instead of its memory.

Ways to control it

- For when the item is higher impact, bring it beyond just the AI and get a second source to check.
- Get the AI to consider bias and check for fairness in its responses and if it is excluding anyone.
- Try to use more language that is neutral and just role base instead of identity based like gender etc.
- Make sure to get a person to review when it is involving people in the task.

Uncertainty in AI: Showing, not Hiding it

Getting the AI to show evidence and confidence.

ASK TO INCLUDE:

- Evidence line where it got it (place and date of it).
- A short quote of the original text possibly.
- A simple confidence label: low / medium / high to make sure it is giving reliable outputs.

ASK FOR WHEN INFORMATION IS MISSING OR CONFLICTING:

- Say: “Cannot determine from provided sources.”
- If two sources disagree: cite both, state the mismatch, request the owner’s ruling.
- Don’t get the AI to merge or average two different cases

GOOD LANGAUGE FROM THE MODEL:

- “Based on the cited policy…”
- “Scope does not cover this case…”
- “Evidence is incomplete. Requesting source.”

KNOWING WHEN NOT TO USE AI

Don’t use AI when tasks involve people related decisions, legal/compliance, customer money/access, data that people havent consented to, or no reliable source.

STOP IF THE TASK IS ABOUT

- People decisions: hiring, ranking, promotion, discipline.
- Legal, compliance, regulated advice, or safety.
- Customer results would probably change money or access (refund approvals, account bans).
- Personal or customer data without consent or an approved tool by the company
- No solid source, or policies conflict and no owner has ruled.

Quick checks:

- Is it about people or law? Don’t use AI
- Sensitive data in an unapproved tool? Don’t use AI
- Missing or conflicting evidence? Don’t use AI until it is resolved

DO THIS INSTEAD

- Log the decision and the details of the task (the actual task, the risk, current date/time)
- Pass it on to the actual owner if it is possible to (HR, Legal, Privacy etc)
- Try to find a safer alternate (redacted summary of it, or making a draft for human approval)

Post Quiz

This quiz is to see your learning progress over this session

Post Quiz - Question 1

BEST SAFE REWRITE BEFORE PROMPTING:
“ON 12/02, BOB SMITH APPROVED $8,950 FOR PROJECT PEAR AT HTTP://INTRANET/PO/125354543.”

A “On 12/02, Bob approved $8,950 for Project Pear at the intraanet link

B On 12/02, [NAME_1] approved $8,950 for [PROJECT_NAME] at (the link is removed).

C Approved money for a project

D On 12/02, Bob Smith approved the money at http://intranet/po/125354543.

Post Quiz - Question 1

BEST SAFE REWRITE BEFORE PROMPTING:
“ON 12/02, BOB SMITH APPROVED $8,950 FOR PROJECT PEAR AT HTTP://INTRANET/PO/125354543.”

B On 12/02, [NAME_1] approved $8,950 for [PROJECT_NAME] at (the link is removed).

Post Quiz - Question 2

YOU MUST SUMMARISE A CONFIDENTIAL CUSTOMER THREAD. IF YOU ARE ALLOWED TO USE BOTH INTERNAL AND WEBSITE TOOLS, WHAT'S THE BEST CHOICE?

A) Paste into the public site with names removed

B) Use the approved internal tool with minimal, redacted input

C) Paste into both to compare answers

D) Email the text to a colleague to ask them

Post Quiz - Question 2

YOU MUST SUMMARISE A CONFIDENTIAL CUSTOMER THREAD. IF YOU ARE ALLOWED TO USE BOTH INTERNAL AND WEBSITE TOOLS, WHAT'S THE BEST CHOICE?

B) Use the approved internal tool with minimal, redacted input

Post Quiz - Question 3

TO REDUCE ERRORS AND MAKE CHECKING EASIER, WHICH PROMPT ENDING IS BEST?

A) “Write an essay, any format.”
B) “Be very detailed.”
C) “Return JSON: {title, summary, actions[]}. Max 120 words. Include the source id and date.”
D) “Use bullet points.”

Post Quiz - Question 3

TO REDUCE ERRORS AND MAKE CHECKING EASIER, WHICH PROMPT ENDING IS BEST?

C) “Return JSON: {title, summary, actions[]}. Max 120 words. Include the source id and date.”

Post Quiz - Question 4

TWO INTERNAL DOCS DISAGREE ON THE SAME RULE. SAFEST MOVE?

A) Pick the newer doc and ignore the other

B) Average the two answers

C) Get the AI to cite both and say the mismatch, and then ask the owner to resolve

D) Choose the one that reads better

Post Quiz - Question 4

TWO INTERNAL DOCS DISAGREE ON THE SAME RULE. SAFEST MOVE?

C) Get the AI to cite both and say the mismatch, and then ask the owner to resolve

Post Quiz - Question 5

YOU WANT LABELS FROM THE AI TO BE THE SAME ACROSS TEAMS. WHICH CONSTRAINT HELPS MOST?

A) Define allowed values
B) Set only a word limit
C) Ask for a friendly tone
D) Request a longer paragraph

Post Quiz - Question 5

YOU WANT LABELS FROM THE AI TO BE THE SAME ACROSS TEAMS. WHICH CONSTRAINT HELPS MOST?

A) Define allowed values

Post Quiz - Question 6

SOMEONE WANTS YOU TO RANK STAFF PERFORMANCE WITH AI USING JUST THE COMMENTS FROM THEIR PEERS. WHAT'S RIGHT?

A) Proceed to save time
B) Proceed after removing names
C) Proceed but add a warning in the output
D) Abstain and escalate to HR/policy owners

Post Quiz - Question 6

SOMEONE WANTS YOU TO RANK STAFF PERFORMANCE WITH AI USING JUST THE COMMENTS FROM THEIR PEERS. WHAT'S RIGHT?

D) Abstain and escalate to HR/policy owners

Next Steps

SESSION 3
Leveraging In Your Role

SESSION 4
Prompting Fundamentals