Generative AI for Data Science

Greta Linse, Sara Mannheimer, Sally Slipher

August 8, 2025

1

Learning objectives

- Using AI responsibly and safely
- Pitfalls to be aware of
- Tips and examples for prompts
- Data science use cases

Please keep in mind…
- This is all open to discussion,
- We're only touching on some aspects,
- All of this material is constantly evolving

2

Have you used AI?

- What Gen AI tool to do you prefer to use?
- What made you choose your preferred Gen AI tool?

3

AI in a data science context

- Not all problems are best solved with AI
- Does not replace humans, augments skills
- Data literacy, AI literacy important
- Different concerns/uses than when used for academics or writing

MONTANA STATE UNIVERSITY

4

Responsible Use

MONTANA STATE UNIVERSITY

5

Using responsibly

- Using Gen AI in an academic setting https://guides.lib.montana.edu/ai/ethics
- Standards for ethical use of AI have not yet been developed, and there is not yet a clear ethical framework to dictate ethical AI practice.
- For students, acting in alignment with personal values, university values, and professors' values can support ethical decision-making and responsible use of AI in class.

MONTANA STATE UNIVERSITY

6

Values-based decision making

Consider **your own values**. Which of your personal values relate to AI? For example:

- Does using AI help you build on or process your own ideas?
- Do you value learning and new challenges?
- Are you concerned about receiving inaccurate or biased information from an AI source that might affect your success in the classroom?
- Do you value efficiency, using AI to reduce the amount of time you spend on assignments?

MONTANA STATE UNIVERSITY

7

Values-based decision making

Consider **the university's values**.

- MSU has guidelines for academic integrity that are outlined in our Code of Student Conduct. Although AI is a new technology, its use should still align with expectations of academic integrity.
- The university values student learning, and it expects students to "be responsible for the honest completion and representation of their work, the appropriate citation of sources, and the respect and recognition of others' academic endeavors" (Code of Student Conduct, Section 200.00).

MONTANA STATE UNIVERSITY

8

Values-based decision making

Consider **your professors' values**.

- Talk with your professors about AI. Make sure your professor approves the use of AI for homework and studying, and talk with them about their expectations of students who would like to use AI as a tool.
- This resource from the Center for Faculty Excellence provides information about how the university and your professors may be thinking about AI. https://www.montana.edu/facultyexcellence/teaching-advising/genai/

MONTANA STATE UNIVERSITY

9

Values-based decision making

Once you have considered your own values, the university's values, and your professors' values, you can make values-informed decisions about when and how to use AI in your classes and build a mutual understanding about what responsible AI use means in the context of the classroom and student learning.

MONTANA STATE UNIVERSITY

10

Transparency when using

If you do decide to use AI in your classes, proper citation practices can help facilitate responsible use of AI:

- name the use or function AI provided to your work
- vet sources generated by AI
- name the tool
- where it is used in your work
- date the content was generated

MONTANA STATE UNIVERSITY

11

Transparency when using

- When someone would feel deceived or hesitant to find out something was done by AI
- Where's the line between tool-use and intellectual contribution
- Hallmarks of an AI response
  - **Writing** – repetitive phrasing, overuse of certain words ("pivotal", "delve", "underscore"), very structured, lack of opinionated-ness, hallucinations, perfect formatting or grammar, em dashes, "__, not __" sentence structure
  - **Images** – human anatomy mistakes, "uncanny valley", overly smoothed or hyperrealistic, text rendering, distorted or inconsistent details, light reflections
  - **Coding** – lack of creative problem solving, favors certain functions, likes to create functions, likes to create intermediate objects, very structured and commented/documented

MONTANA STATE UNIVERSITY

12

Quiz

Image 1

Image 2

https://britannicaeducation.com/blog/quiz-real-or-ai/

MONTANA STATE UNIVERSITY

13

Quiz

Text 1

Back-to-school season isn't just for kids — email marketers can learn a few new tricks too! 🍎 Whether it's getting a handle on email authentication, diving into coding, or sharpening your copywriting and storytelling skills, now's the time to up your game. 🚀 Ready to learn something new? Let's hit the books (or the inbox)! 💻

Text 2

So, you learned to craft and send mass emails in your favorite email marketing software — is that it? 🤔 What can you do next? What to learn to earn more money, potentially switch careers, or improve your email marketing ROI? 📈

We compiled a list of 8 skills that will make you the most valuable player in a team — and a much better email marketer. Spoiler alert: not all of these skills are directly related to emails, and that's okay!

https://selzy.com/en/blog/ai-or-human-writing-quiz/

MONTANA STATE UNIVERSITY

14

Quiz

Code 1

```
ggplot(mpg, aes(x = displ, y = hwy)) +
  geom_point(color = "blue") +
  labs(title = "Engine Displacement vs
Highway MPG")
```

Code 2

```
library(ggplot2)

plot <- ggplot(data = mpg, mapping =
  aes(x = displ, y = hwy)) +
  geom_point(colour = "blue") +
  ggtitle("Engine Displacement vs
  Highway Miles Per Gallon")

print(plot)
```

https://crossley.github.io/cogs2020/lectures/week_3/lecture_themes.html#1

MONTANA STATE UNIVERSITY

15

Data privacy and ownership

- Human subjects data privacy and consent – What data are you entering into Gen AI? Does it have identifiers like names, birth dates, addresses, etc.? Or is it sensitive in other ways?
- Intellectual property – Does the data you're entering into Gen AI belong to anyone? Did someone write the words or create the images?
- Hard to know what data Gen AI is retaining
- Can't expect that Gen AI is keeping your data private
- Where will the data go after being entered into Gen AI? What are the terms of service? Will the Gen AI company sell the data? Use it to train new models? Use it as sample data for other users?
- Other types of sensitive data: geotagged archaeological site data, endangered species data, copyrighted data

MONTANA STATE UNIVERSITY

16

Frameworks

MONTANA STATE UNIVERSITY

17

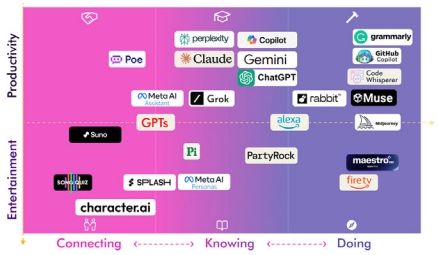
What is Generative AI?

- A type of artificial intelligence that creates new content based on patterns learned from existing data
- Trained on massive datasets and can be prompted to generate novel outputs that resemble the data they were trained on
- Users provide a "prompt" (a text description, an image, etc.), and the AI generates a response
- LLMs – trained on and focus on language/writing
- Like a very advanced autocomplete
- Is like a recipe follower, not a chef

MONTANA STATE UNIVERSITY

18

Types of generative AI



https://synthedia.substack.com/p/generative-ai-in-entertainment-framework

MONTANA STATE UNIVERSITY

19

Types of generative AI

- Text-generation (ChatGPT, Gemini, Claude)
- Code-generation (**Copilot**, Cursor)
- Image-generation (DALL-E, Midjourney)
- Audio-generation (Suno)
- Video-generation (Sora)
- Reasoning (GPT-4o, Claude 3)
- Specific industries (BloombergGPT, Med-PaLM)
- Multimodal
- Agents

Can be hosted on a **server/locally**, may be able to access the **internet**, may have access to your **files**

MONTANA STATE UNIVERSITY

20

Terminology

- **Model** - The AI "brain" that generates responses
- **Context window** - The amount of text an AI can "remember" at once
- **Reasoning** - AI's ability to break down complex or multi-step tasks
- **Tokens** - Units of data processed by AI models during training and inference
- **Temperature** - Controls how random the response is - lower = more predictable, higher = more creative
- **Hallucination** - When the model makes up information that sounds plausible but is false.

MONTANA STATE UNIVERSITY

21

Prompting frameworks

- Anthropic, OpenAI, Google prompting guides
- **Prompt engineering** - crafting effective inputs (prompts) to get high-quality outputs

- In general:
  - Be clear, specific
  - Use structured prompts
  - Break down complex tasks
  - Provide context
  - Control output style
  - Start simple, iterate and refine
  - Avoid ambiguity
  - Ask for reasoning or justification
- Keep in mind:
  - Response/token limits
  - Capability limitations
  - Hallucinations
  - Overconfidence
  - Misinterpretations

MONTANA STATE UNIVERSITY

22

RISE

Role, **Input**, **Steps**, **Expectation**
- Ideal for multi-step code or reasoning work
- Clear incremental guidance.

Example:
- **Role** – You are a statistical code reviewer and optimizer.
- **Input** – R script that fits a GLM and prints summary.
- **Steps** –
  - Review script for potential efficiency issues.
  - Refactor for clarity and modularity.
  - Add error-checking and meaningful variable names.
- **Expectation** – Provide cleaned-up version of the script and explanations for each change.

MONTANA STATE UNIVERSITY

23

TAG

Task, **Action**, **Goal**
- A clear and minimal structure
- Goal-oriented

Example:
- **Task** – The task is to write an R function that fits a linear regression on provided data and returns model diagnostics.
- **Action** – Act as an expert data scientist in R: write the function code, include meaningful variable names, error checks, and comments.
- **Goal** – Goal is to help academic researchers rapidly produce clean, reproducible analysis so they can interpret coefficients and test assumptions easily.

MONTANA STATE UNIVERSITY

24

RTF

Role, Task, Format

- Specify who, what, and how the AI should respond
- Specifying persona and response layout

Example:

- **Role** – You are an experienced statistical programmer.
- **Task** – Create R code to perform k-fold cross-validation (k = 5) for a random forest model predicting outcome Y.
- **Format** – Provide annotated code blocks, explanation of each step, and a summary table of results.

MONTANA STATE UNIVERSITY

25

RODES

Role, **Objective**, **Details**, **Examples**, **Sense Check**

- Adds structure with examples and a check of understanding
- High precision and style consistency

Example:

- **Role** – You are a senior quantitative analyst.
- **Objective** – Generate RMarkdown sections for exploratory data analysis of a survey dataset: demographics, summary statistics, missingness.
- **Details** – Use dplyr, ggplot2, follow reproducibility best practices.
- **Examples** – Example: "ggplot(df, aes(x=age)) + geom_histogram()"
- **Sense Check** – "Do you understand the style and guidelines before proceeding?"

MONTANA STATE UNIVERSITY

26

Chain-of-Thought

CoT

- Encourages the model to reason step-by-step
- The AI will articulate reasoning steps and then output the code
- Complex reasoning or statistical logic
- Has strong empirical support, especially for reasoning tasks and math problems, and improves performance significantly

Example:

- "Let's think through this step by step: For a dataset with non-normal residuals and heteroskedasticity, explain suitable regression alternatives and diagnostics, then write R code to implement them."

MONTANA STATE UNIVERSITY

27

Reason + Act

ReAct

- Combines internal reasoning with performing actions (often used in agent-like tasks).
- The AI will reason, then act (edit and recommend further work)

Example:

- "Analyze this R script: reason about whether feature selection is appropriate, then refactor problematic segments into cleaner functions, and suggest next analysis steps."

MONTANA STATE UNIVERSITY

28

Correcting bad responses

- Tell it exactly what is wrong
  - "You hallucinated a reference; don't make up citations."
  - "That statistic is incorrect."
- Focus on specific corrections
- Ask for step-by-step reasoning, identify what needs modification or elaboration
  - "What additional information do you need?"
  - "Explain your reasoning step-by-step."
- Guide with examples
  - "Here's an example of the type of answer I want."
- Give more constraints
  - "Keep it under 100 words."
  - "Eliminate any repeated ideas or filler."
- Ask to keep certain sections untouched
- Specify what you did like about the response
- Refer to the previous code or responses to maintain context and avoid redundant information in your correction prompts

MONTANA STATE UNIVERSITY

29

What it's good at/not good at

Good

- Brainstorming
- Procedural tasks
- Drafting
- Summarizing
- Exploration
- Pattern detection

Bad

- Accurate, up-to-date info
- Interpretation
- Distinguishing good from bad
- Distinguishing real from not real
- Decision-making
- Critical thinking

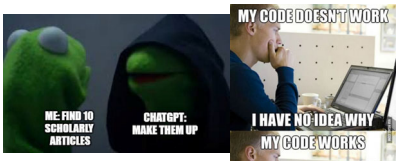
MONTANA STATE UNIVERSITY

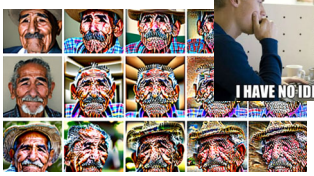
30

Pitfalls

- Bias
- Privacy
- Probabilistic
- Everything is a hallucination
- Meant to seem like a human
- Overconfidence
- Reproducibility

- Model collapse
- Hallucinations being exploited
- Prompt injection
- Honey pots
- Vibe coding





Use In Data Science

Data science uses

- Code/error troubleshooting
- **Boilerplate code/functions**
- **Code templates/skeletons to modify**
- Data cleaning/wrangling steps
- Reformat, annotate, clean up code
- Convert pseudocode or descriptions into working code
- **Translate code between languages**
- Generate model syntax
- Look for/code to test possible overfitting problems, model assumptions, limitations, etc.
- Suggest common transformations, cleaning, or filtering steps
- Suggestions for explaining model outputs/interpretation
- Suggestions for appropriate models based on data and research questions
- Suggestions to make code reproducible and shareable

Translate code

Prompt:

- "Given the following code translate to R." <<Paste code>>

Demonstration:

- Translate NHANES Sample Code for Logistic Regression that is only available in SAS-callable SUDAAN version 11

Try:

- Try prompts for some sections of code
- How to create a better prompt
- Think about follow up prompts

MONTANA STATE UNIVERSITY

34

SAS code

```
*************************************************************;
*Linear trends analysis by age group using Logistic Regression*;
*************************************************************;

proc rlogist data = Rx FILETYPE = SAS DESIGN = WR ;
* Nest statement: PSUs nested within Strata accounts for the design effects*;
NEST SDMVSTRA SDMVPSU / MISSUNIT;
* Weight statement: specify appropriate weight, accounts for the unequal probability of sampling and non-response.*;
WEIGHT wtint2yr;
* Class statement: specify categorical variable(s). Reflevel statement can be included to choose reference category
for the categorical variables. By default SUDAAN uses the highest category*;
class one /nofreq;
* Subpopx statement: specify the subpopulation of interest (the inclusion criteria)*;
subpopx inAnalysis;
* Model statement: specifies dependent variable and independent variable(s) *;
MODEL anyOpioid = ageCat;
* Output statement: outputs the results to a file*;
output / betas=default filename=fig3_logistReg_betas_all replace;
* Test statement: produces statistics and P values for the Satterthwaite adjusted CHI square (satadjchi),
the Satterthwaite adjusted F (satadjf), and Satterthwaite adjusted degrees of freedom (printed by default).
If this statement is omitted, the nominal degrees of freedom,
the WALDF and the p-value corresponding to the WALDF and WALDP will be produced.*;
test waldf satadjf satadjchi;
run;
```

https://wwwn.cdc.gov/nchs/data/Tutorials/Code/DB369_SUDAAN.sas

MONTANA STATE UNIVERSITY

35

Create workflow

Prompt:

- "Write R code that does the following steps:" <<Describe steps>>

Demonstration:

- Create a script that imports data, checks missing values, performs analysis, performs checks, outputs results

Try:

- Note: we are not giving it data or a real Excel file
- Think about follow up prompts
- Discuss output and what worked as expected and what didn't

MONTANA STATE UNIVERSITY

36

Steps

1. Import data from an Excel file
2. Check for missing values
3. Perform exploratory data analysis
4. Generate a statistical model for a [specify a type of] design
5. Check model assumptions
6. Summarize the results

MONTANA STATE UNIVERSITY

37

Knowledge required

- Need to understand how to install packages
  - What packages are used?
  - Are they actually popular libraries?
- The suggested code uses functions with messages and errors.
  - Seems pretty complicated!
- Have to read the code carefully!
  - There are some optional code chunks that are included.
- Are all variables included correctly?
  - Which are fixed and which are random effects?

MONTANA STATE UNIVERSITY

38

Create plot

Prompt:
- "Given this data, create a plot in R." <<Describe data, variables, and plot type>>

Demonstration:
- Create a plot using workshop datasets

- Can save a lot of time creating the "bones" of the plot
- You don't need to think of the all the details
- Can explain details about the code and parameters
- Cannot run code itself
- Doesn't know anything about your data
- Given plot might not be right for your data
- You may have to edit a lot
- It may start you off in the wrong direction (start over vs. follow-up prompts)

MONTANA STATE UNIVERSITY

39

Data

BlackfootFish

- https://github.com/saramannheimer/data-science-r-workshops/tree/master/Introduction%20to%20R/Summer%202025/StudentVersion/data

Surveys

- https://github.com/saramannheimer/data-science-r-workshops/tree/master/Data%20Wrangling/Summer%202025/data

MONTANA STATE UNIVERSITY

40

Wrap Up

MONTANA STATE UNIVERSITY

41

Discussion

- Which follow up prompts worked well?
- Which prompts didn't work well?
- What were you expecting to work well but didn't?
- What are you still curious about?
- If you and your neighbor were using different Gen AI models, were the results different?
- Will you use AI for code?
- What concerns do you have?

MONTANA STATE UNIVERSITY

42