ENGAGING WITH THE RESEARCH COMMUNITY

AI and Data Transformation

John Little

Duke University Libraries

Center for Data & Visualization Sciences

2024-10-27

PATRON-DRIVEN INTERACTIONS

Computational workflows, data transformations, and Analysis

AI AND NEW CHALLENGES

- Introduction of Generative LLMs (e.g., ChatGPT)
- New challenges for patrons:
  1. How to ask questions of a generative AI
  2. How to frame questions to reflect data goals
- Translation
- Synthesis

THE CONFIDENCE V COMPETENCE PARADOX

- LLMs give confident responses
- Responses are predictions, not necessarily correct answers
- Incorrect predictions = “hallucinations”
- Verification is crucial
- Paradox: More knowledge leads to better evaluation of AI responses

USE CASE - CODE GENERATION

- Data transformation
- Data analysis
- Iteration
- Big Data
- AI assistance / AI-paired coding

GOAL

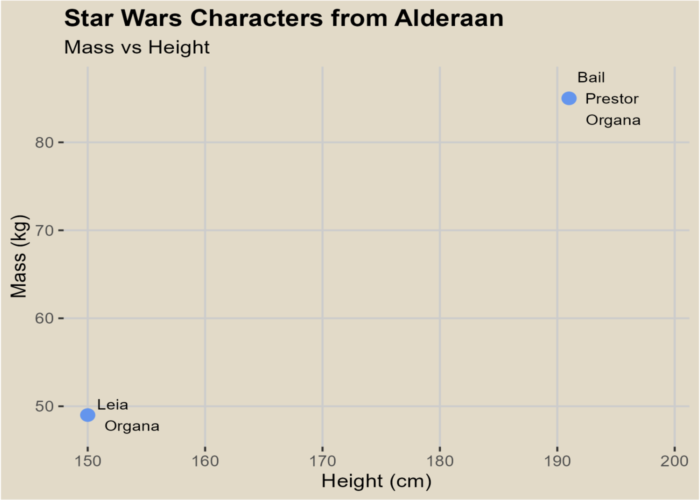
Create scatter plots, one for each home world

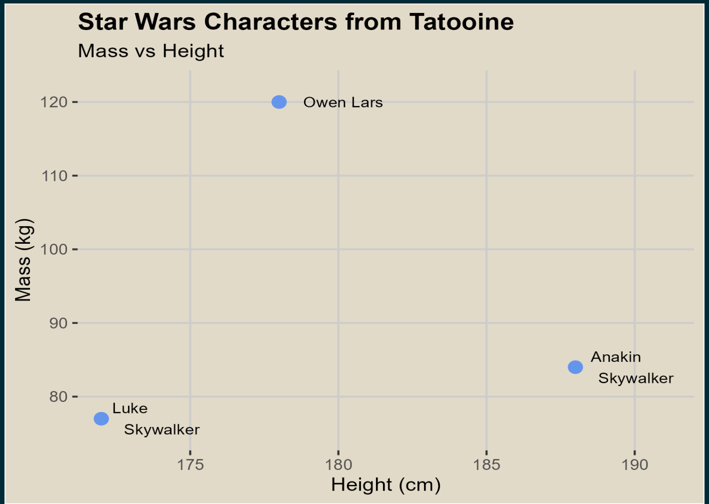
CASE STUDY - STAR WARS DATASET

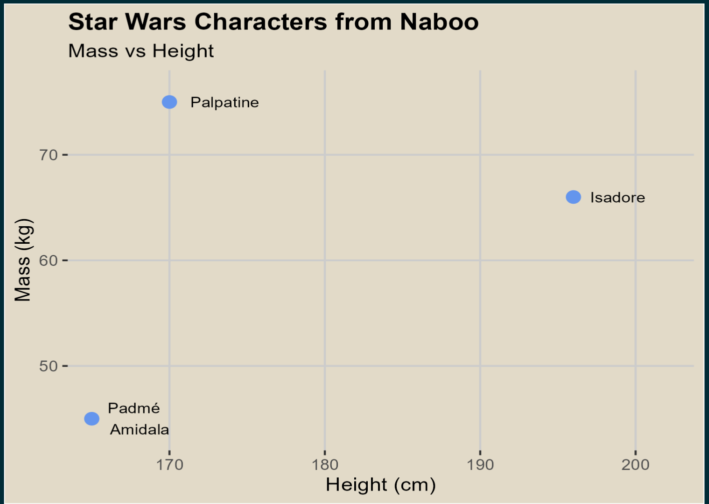
Homeworld	Heights	Masses	Characters
Tatooine	172, 188, 178	77, 84, 120	Luke Skywalker, Anakin Skywalker, Owen Lars
Alderaan	150, 191	49, 85	Leia Organa, Bail Prestor Organa
Naboo	165, 196, 170	45, 66, 75	Padmé Amidala,

Homeworld	Heights	Masses	Characters
			Isadore, Palpatine
Coruscant	66, 188	17, 84	Yoda, Mace Windu

EXAMPLE


Star Wars Characters from Alderaan
Mass vs Height
Bail Prestor Organa
Leia Organa
80
70
60
50
Mass (kg)
150
160
170
180
190
200
Height (cm)


Star Wars Characters from Tatooine
Mass vs Height
Owen Lars
Anakin Skywalker
Luke Skywalker
120
110
100
90
80
Mass (kg)
175
180
185
190
Height (cm)


Star Wars Characters from Naboo
Mass vs Height
Palpatine
Isadore
Padmé Amidala
70
60
50
Mass (kg)
170
180
190
200
Height (cm)

CHALLENGES IN AI ASSISTANCE

- AI *can* handle well some basic visualization and coding
- Struggles with complex data shaping and iteration
- This problem is easier when the user has knowledge in:
    - Coding concepts
    - Data shaping
    - Visualization
    - Iteration for large datasets

WHEN IT GOES WRONG

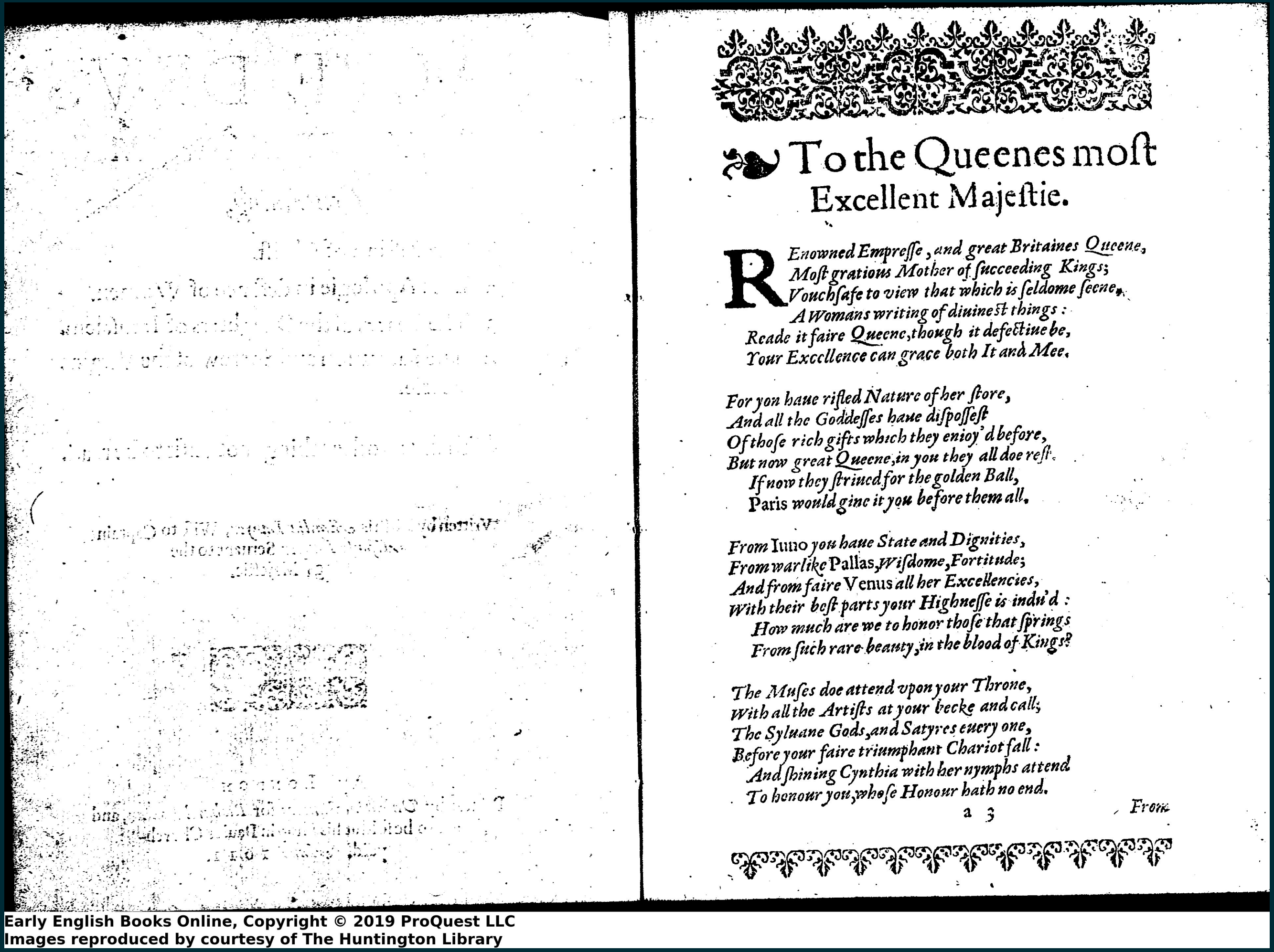
WORD PROBLEMS

Prompt: Inconsistent AI responses for “How long does it take to walk 10,000 steps on a treadmill at 1.2 MPH?”

- Lesson 1: Importance of cross-verification
- Lesson 2: Prediction is not the same as mathmatical truth

EEBO

No ground truth



To the Queenes moſt
Excellent Majeſtie.

REnowned Empreſſe, and great Britaines Qucene,
Moſt gratious Mother of ſucceeding Kings;
Vouchſafe to view that which is ſeldome ſeene,
A Womans writing of diuineſt things:
Reade it faire Queenc, though it defectiue be,
Your Exccllence can grace both It and Mee.

For yon haue rifled Nature of her ſtore,
And all the Goddeſſes haue diſpoſſeſt
Of thoſe rich gifts which they enioy'd before,
But now great Queene, in you they all doe reſt.
If now they ſtriued for the golden Ball,
Paris would ginc it you before them all.

From Iuno you haue State and Dignities,
From warlike Pallas, Wiſdome, Fortitude;
And from faire Venus all her Excellencies,
With their beſt parts your Highneſſe is indu'd:
How much are we to honor thoſe that ſprings
From ſuch rare beauty, in the blood of Kings?

The Muſes doe attend vpon your Throne,
With all the Artiſts at your becke and call;
The Syluane Gods, and Satyres euery one,
Before your faire triumphant Chariot fall:
And ſhining Cynthia with her nymphs attend
To honour you, whoſe Honour hath no end.

a 3 From

CODE

TRANSLATION DONE POORLY

- Due to insufficient background and/or prompting

AI-PAIRED CODE GENERATION

- Some clear winners and losers in the big names. aka each LLM has it’s own evolving strengths, weaknesses, and tendencies.

These problem highlights the Competence v Confidence Paradox but are easily **verifiable**

WHEN IT GOES RIGHT

and how *right* does it go?

SYNETHTIC QUESTIONS

Prompt: Compare student body and faculty diversity at Duke University with UNCG. Compare today with 1985.

- Lesson 1: Different LLMs give different amounts of evidence for verification
- Lesson 2: Differing amounts of ground truth will affect the prediction

CODE TRANSLATION

I have Python code, give it to me in R

VARIATIONS IN CODE TRANSLATIONS

- R to Python
- Python to R
- SQL from natural language
- javascript
- HTML

NATURAL LANGUAGE

How can I use the phrase “*Sticky Wicket*” in German?

- Translate *Sticky Wicket* to German
- But how to verify (same as code problem)

VALUE IN REPRODUCIBILITY

- Coding
  - Do everything with code
  - Including report generation
- No Code
  - Getting better all the time

Increasingly we are seeing computation environments with build-in AI-pairing

Figure 9. “In which of the following ways have you used generative AI in your biomedical research?”

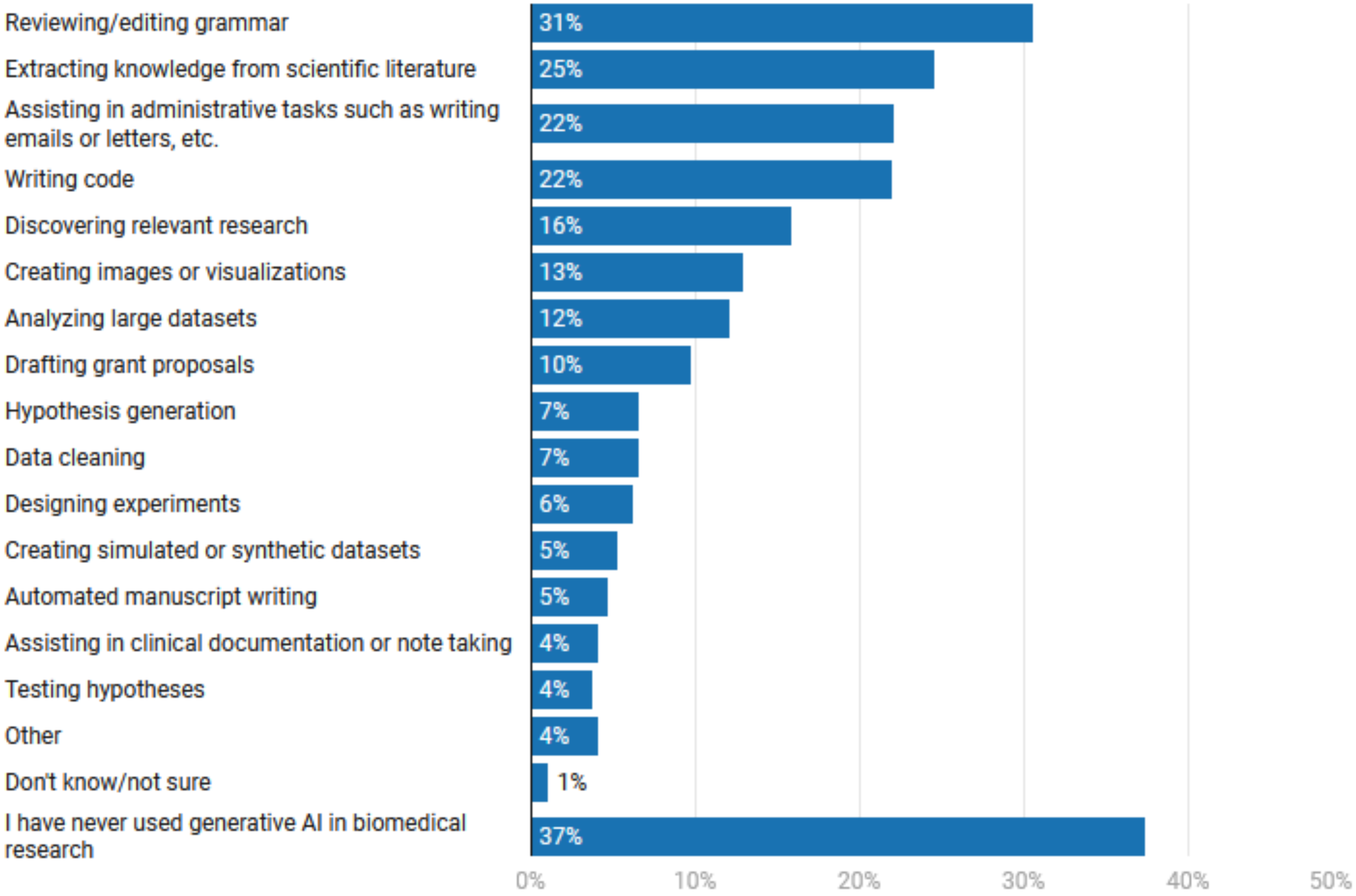


Chart: Ithaka S+R • Get the data • Created with Datawrapper

SOLUTIONS

and best practices

PROBLEMS AND SOLUTIONS

- GIGO (Garbage In, Garbage Out) still applies
- Prompt engineering is a crucial skill
- AI excels in translation tasks
- Good for synthetic questions with possible validation
- Less reliable for tasks without established ground truth

BEST PRACTICES

Using Broad-base LLMs:

- ChatGPT
- Microsoft Copilot
- Claude.ai
- Gemini.google.com
- GitHub Copilot (for AI-paired coding)

PROMPT ENGINEERING

- Identify role
- Identify audience
- Identify voice
- Identify goals and problem
- Use multiple steps
- Verify

CONCLUSION

Embracing AI in data analysis

- AI is a powerful tool, but requires careful use
- The library offers crucial guidance
- Continuous learning and adaptation are essential

QUESTIONS

1. How do you see these tools or techniques impacting research and research investment?
2. Do you have data transrormation, reshaping, or analysis tasks that could benefit from AI assistance?
3. In what ways do you think we can improve training and assistance for next generation LLMs?
4. What are some of the biggest challenges you see in the future of AI-paired coding?