EVALUATING DATA VALIDATION TESTS GENERATED BY LLMS

AN APPLICATION TO CANADIAN POLITICAL DONATIONS DATA

ROHAN ALEXANDER, LINDSAY KATZ, CALLANDRA MOORE, MICHAELA DROUILLARD, ZANE SCHWARTZ

7 DECEMBER 2023

Australian Society for Quantitative Political Science

ROHAN ALEXANDER

('Rohan' ≈ 'row-hun')

UNIVERSITY OF TORONTO

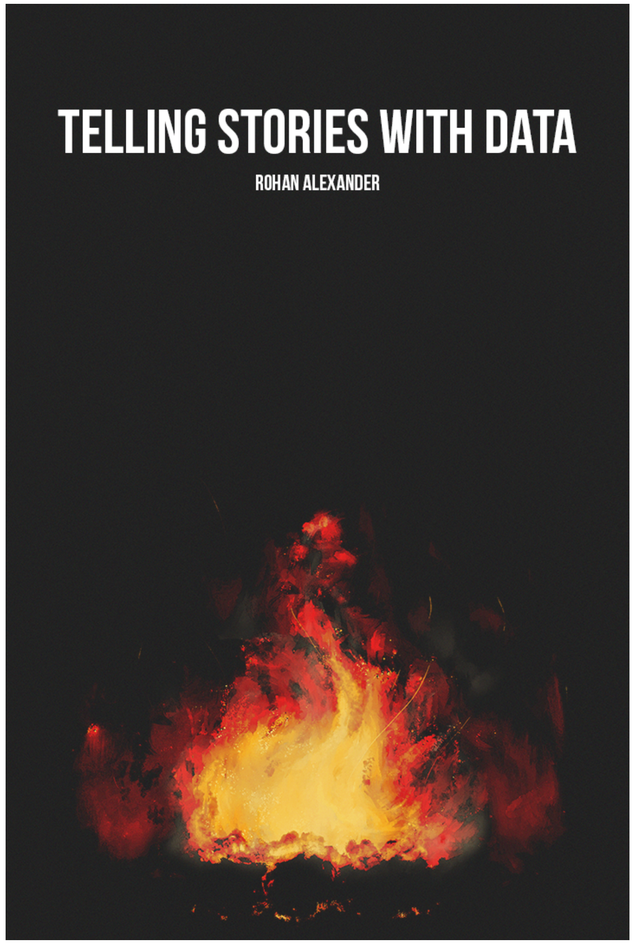
- Australian 🇦🇺 but lived in Canada since July 2018 🇨🇦

- I develop workflows that improve the trustworthiness of data science, especially focused on automated testing.

- Assistant Professor, jointly appointed: Information (51%) & Statistical Sciences (49%)
- Assistant Director, Canadian Statistical Sciences Institute (CANSSI) Ontario
- Senior Fellow, Massey College

- *Telling Stories with Data* (2023), Chapman and Hall/CRC.
- *Multilevel Regression and Poststratification: A Practical Guide and New Developments* (🤦), (editor) Cambridge University Press.

- Monica (Statistical Sciences (51%) & Sociology (49%)); Edward (4yo); & Hugo (2yo)

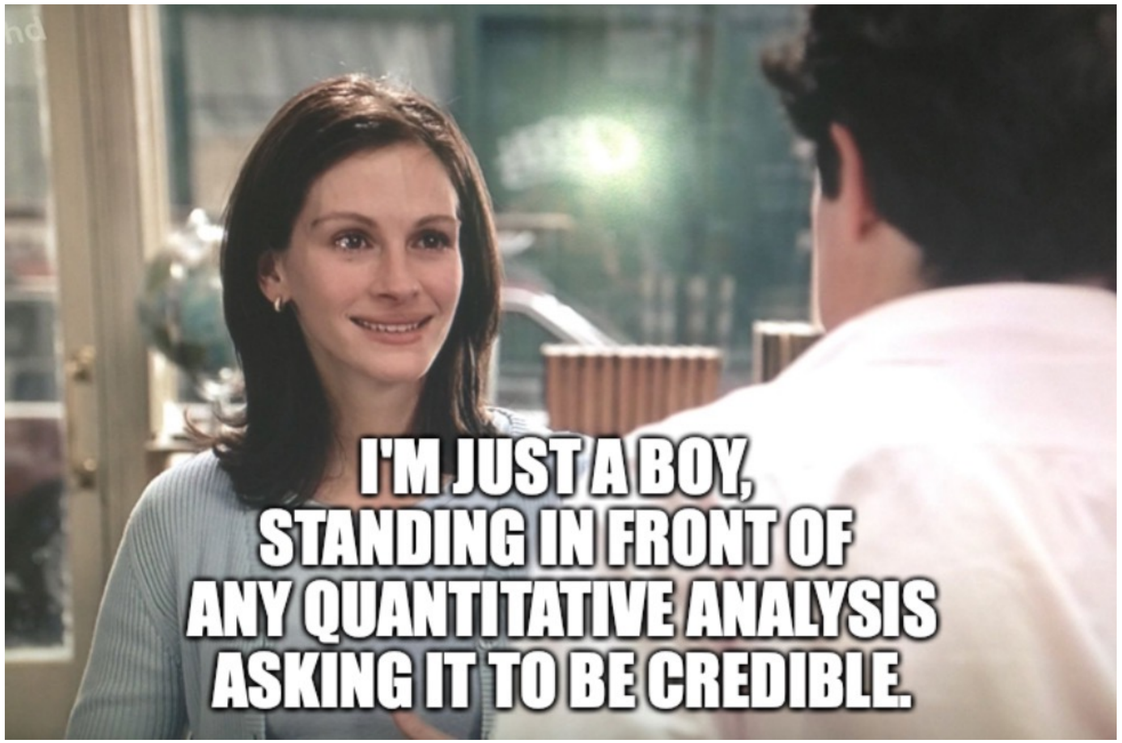


tellingstorieswithdata.com

rohanalexander.com @RohanAlexander rohan.alexander@utoronto.ca

DISCLAIMERS

- This is me trying to work out what I think.
- My PhD is in Australian economic history!
- I have an outsider's view of a bunch of disciplines, but I am *not* a ~~biostatistician~~, ~~computer scientist~~, ~~demographer~~, ~~economist~~, ~~political scientist~~, ~~statistician~~, etc.
- What I do? Look at what those disciplines seem to be doing right, and steal from them, to make social science better.



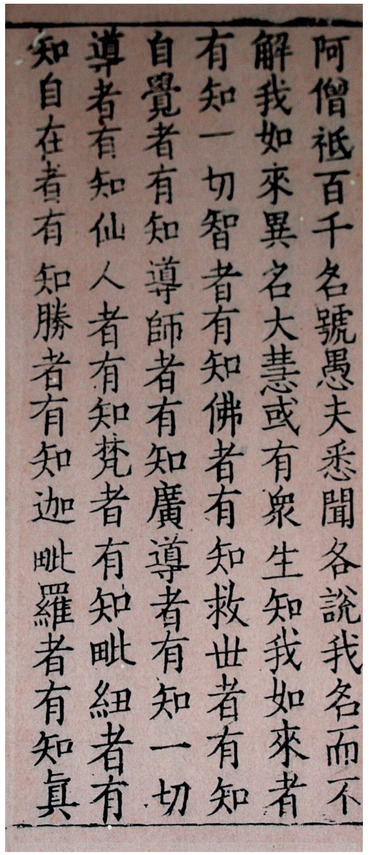
PLAN

1. What are data validation tests?
2. Can we use LLMs to generate data validation tests?

DATA VALIDATION

DATA ARE TOO COMPLICATED

- All datasets have various properties.
- Most have errors, which you “fixed”.
- All have errors, which you don’t know about.
- Your job is to convince everyone else that it is fit for purpose (without requiring that they go through it observation by observation.)


阿僧祇百千名號愚夫悉聞各說我名而不
解我如來異名大慧或有衆生知我如來者
有知一切智者有知佛者有知救世者有知
自覺者有知導師者有知廣導者有知一切
導者有知仙人者有知梵者有知毗紐者有
知自在者有知勝者有知迦毗羅者有知真

Tripitaka Koreana: A collection of Buddhist scriptures from 13th century with "no known errors or errata in the 52,330,152 characters”.

Source: https://en.wikipedia.org/wiki/Tripitaka_Koreana

WHY THIS MATTERS

EXAMINING THE EFFECT OF CLASS ON REGRESSION RESULTS

```
simulated_class_data <-
  tibble(
    response = c(1, 1, 0, 1, 0, 1, 1, 0, 0),
    group = c(1, 2, 1, 1, 2, 3, 1, 2, 3)
  ) |>
  mutate(
    group_as_integer = as.integer(group),
    group_as_factor = as.factor(group),
  )
```

WHY THIS MATTERS

EXAMINING THE EFFECT OF CLASS ON REGRESSION RESULTS

```
models <- list(
  "Group as integer" = glm(
    response ~ group_as_integer,
    data = simulated_class_data,
    family = "binomial"
  ),
  "Group as factor" = glm(
    response ~ group_as_factor,
    data = simulated_class_data,
    family = "binomial"
  )
)
modelsummary(models)
```

WHY THIS MATTERS

EXAMINING THE EFFECT OF CLASS ON REGRESSION RESULTS

	Group as integer	Group as factor
(Intercept)	1.417	1.099
	(1.755)	(1.155)
group_as_integer	-0.666	
	(0.894)	
group_as_factor2		-1.792
		(1.683)
group_as_factor3		-1.099
		(1.826)

HOW TO OPERATIONALIZE?

RUNNING TIMES

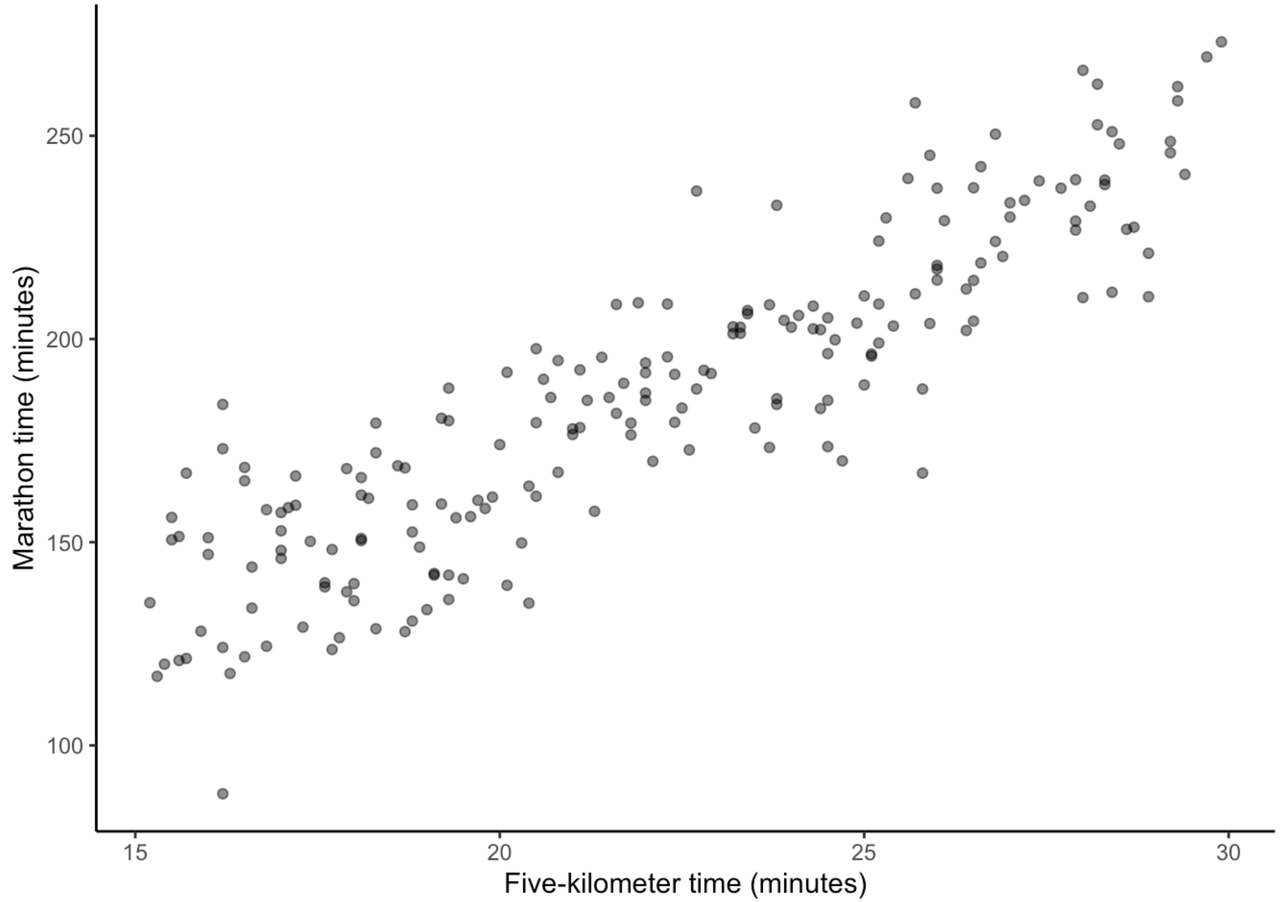
```
set.seed(853)

number_of_observations <- 200
expected_relationship <- 8.4
very_fast_5km_time <- 15
good_enough_5km_time <- 30

sim_run_data <-
  tibble(
    five_km_time =
      runif(
        n = number_of_observations,
        min = very_fast_5km_time,
        max = good_enough_5km_time
      ),
    noise = rnorm(n = number_of_observations, mean = 0, sd = 20),
    marathon_time = five_km_time * expected_relationship + noise
  ) |>
  mutate(
    five_km_time = round(x = five_km_time, digits = 1),
    marathon_time = round(x = marathon_time, digits = 1)
  ) |>
  select(-noise)
```

RUNNING TIMES

SIMULATED DATASET - 200 INDIVIDUALS



RUNNING TIMES

WE WRITE TESTS THAT DOCUMENT OUR ASSUMPTIONS ABOUT THE DATA

- Assumptions:
  - Class of five_km_time and marathon_time are numeric
  - Inputs of five_km_time and marathon_time are minutes.
  - Number of observations is 200.

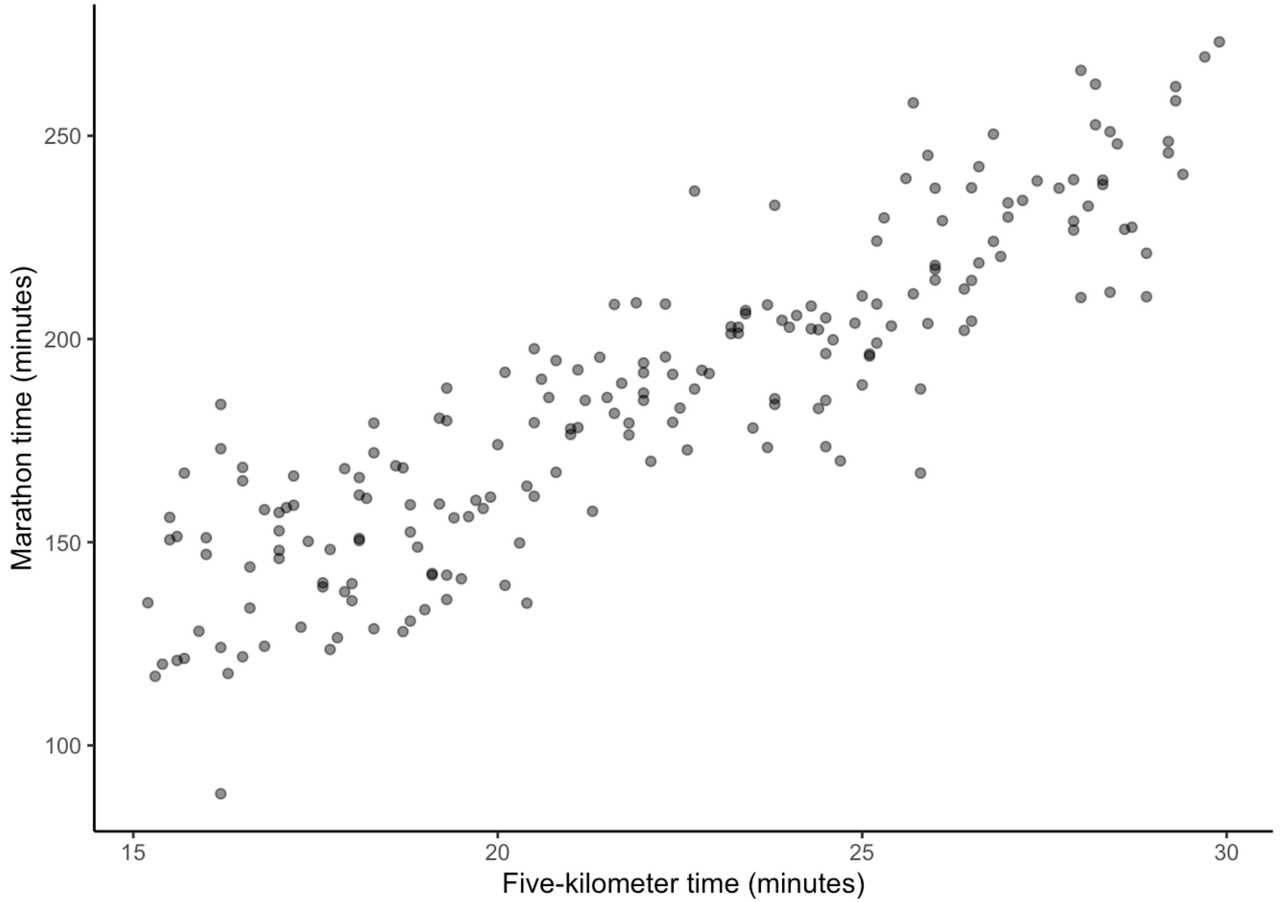
RUNNING TIMES

WE WRITE TESTS THAT DOCUMENT OUR ASSUMPTIONS ABOUT THE DATA

```
stopifnot(
  class(vickers_data$marathon_time) == "numeric",
  class(vickers_data$five_km_time) == "numeric",
  min(vickers_data$five_km_time) >= 15,
  max(vickers_data$five_km_time) <= 30,
  min(vickers_data$marathon_time) >= 118,
  max(vickers_data$marathon_time) <= 300
)
```

RUNNING TIMES

SIMULATED DATASET - 200 INDIVIDUALS



THEN APPLY THESE TESTS TO THE ACTUAL DATASET

VICKERS AND VERTOSICK (2016) PROVIDE DATA ON RUNNING

A tibble: 430 × 2

five_km_time <dbl>	marathon_time <dbl>
17.9	171.6
21.5	204.9
20.4	224.2
14.9	158.6
17.5	181.2
26.7	276.3
24.3	257.3
20.9	218.6
26.2	286.5
42.9	369.0

1–10 of 430 rows

```
stopifnot(
  class(vickers_data$marathon_time) == "numeric",
  class(vickers_data$five_km_time) == "numeric",
  min(vickers_data$five_km_time) >= 15,
  max(vickers_data$five_km_time) <= 30,
  min(vickers_data$marathon_time) >= 118,
  max(vickers_data$marathon_time) <= 300
)
```

```
Error: min(vickers_data$five_km_time) >= 15 is not TRUE
```

RUNNING TIMES

WE CAN GET MORE SERIOUS ABOUT THIS WITH POINTBLANK

```
library(pointblank)

agent <-
  create_agent(tbl = sim_run_data) |>
  col_is_factor(columns = vars(was_raining)) |>
  col_is_numeric(columns = vars(five_km_time)) |>
  interrogate()

agent
```

Pointblank Validation

[2022-08-25|07:18:33]

TIBBLE column_names_as_contracts

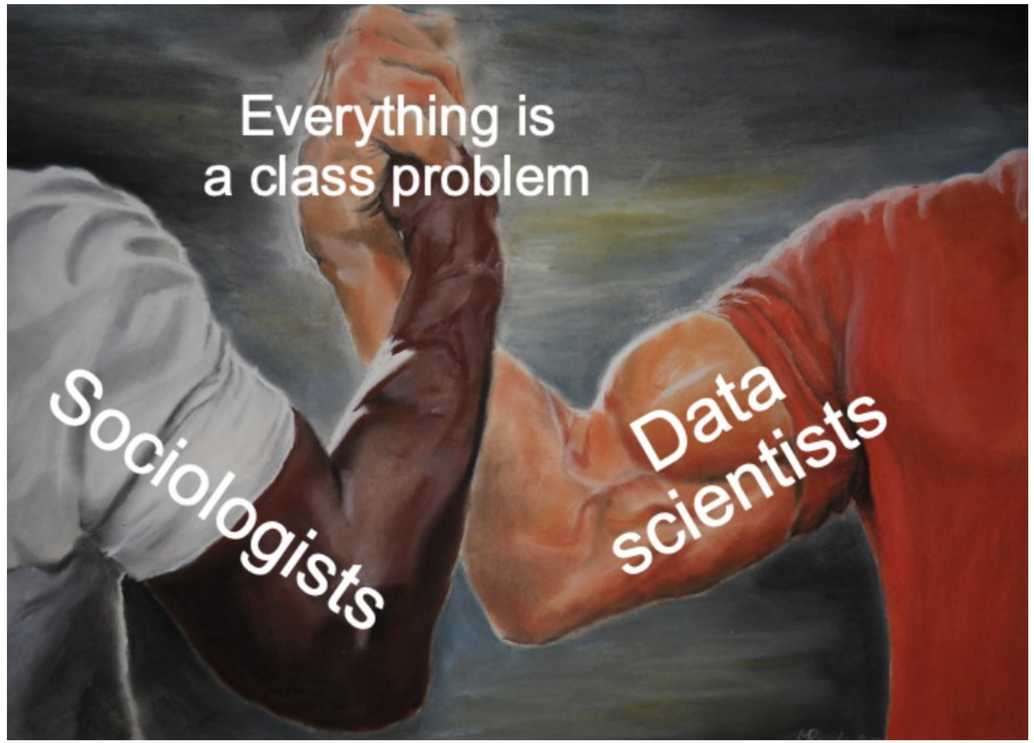
	STEP	COLUMNS	VALUES	TBL	EVAL	UNITS	PA
1	c col_is_character()	chr_area	—	○→	✓	1	
2	c col_is_character()	chr_group_gender	—	○→	✓	1	
3	f col_is_factor()	fctr_group_age	—	○→	✓	1	
4	i col_is_integer()	int_group_count	—	○→	✓	1	
5	∈ col_vals_in_set()	chr_group_gender	male, female, to...	○→	✓	14K	

2022-08-25 07:18:33 EDT | < 1 s | 2022-08-25 07:18:33 EDT

RUNNING TIMES

WE WRITE TESTS THAT DOCUMENT OUR ASSUMPTIONS ABOUT THE DATA

- We then apply these tests to our actual dataset.
- These tests can be shared, even if the data cannot.
- If you do nothing else, you must test class.
- These are all checks that we do in our heads, but no one can check that we've done them right.



CAN WE USE LLMS TO GENERATE DATA VALIDATION TESTS?

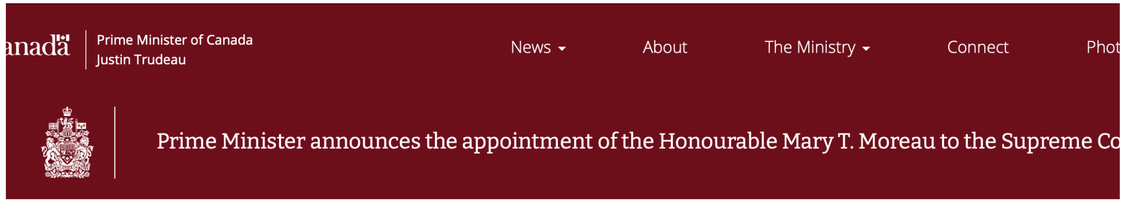
BACKGROUND

INVESTIGATIVE JOURNALISM FOUNDATION (IJF)

- The IJF is a nonprofit news media outlet that is centred around public interest journalism.
- One of the IJFs eight databases, and the focus of this work, is the political donations database.
- While the IJF's database is available in a clean, user-friendly format, the original records upon which it was created were not all accessible in this way.

EXAMPLE

https://theijf.org/ “Mary Moreau”



November 6, 2023
Ottawa, Ontario

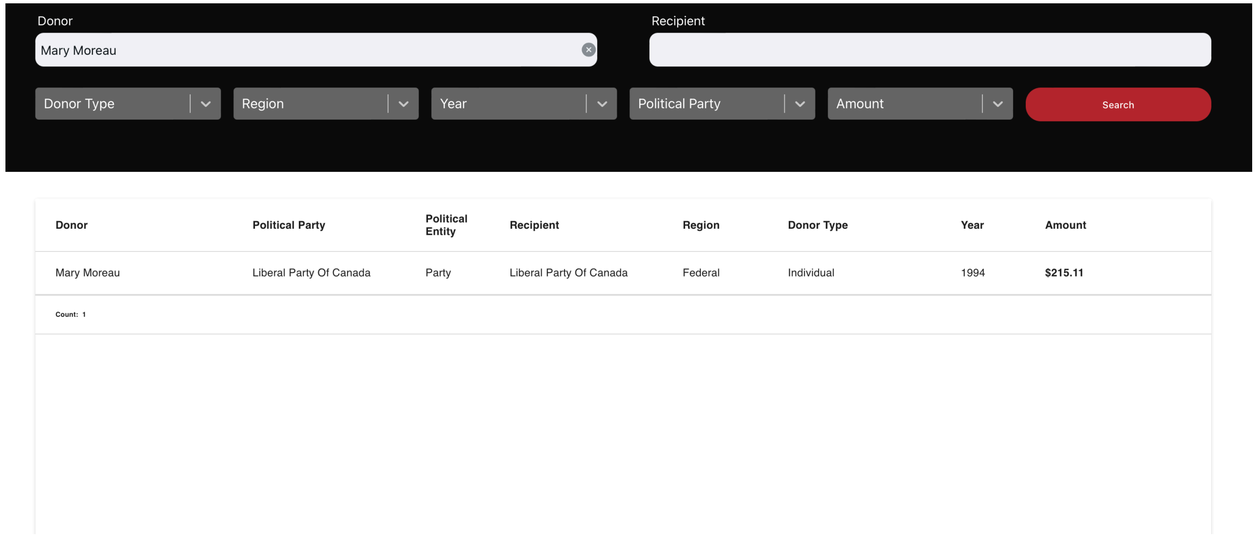
The Prime Minister, Justin Trudeau, today announced the appointment of the Honourable Mary T. Moreau to the Supreme Court of Canada.

Justice Moreau’s esteemed judicial career includes 29 years on the Court of King’s Bench of Alberta. In 2017, she was appointed Chief Justice of that court. Prior to becoming a judge, Justice Moreau practised criminal law, constitutional law, and civil litigation in Edmonton, Alberta. Throughout her career, she has been extensively involved in judicial education, administration, and ethics, both in Canada and internationally.

This appointment is the sixth under the Supreme Court selection process launched by the Government of Canada in 2016. Through this process, an independent and non-partisan advisory board, chaired by the Honourable H. Wade MacLauchlan, was tasked with identifying suitable candidates who are jurists of the highest caliber, are functionally bilingual, and are representative of the diversity of Canada.

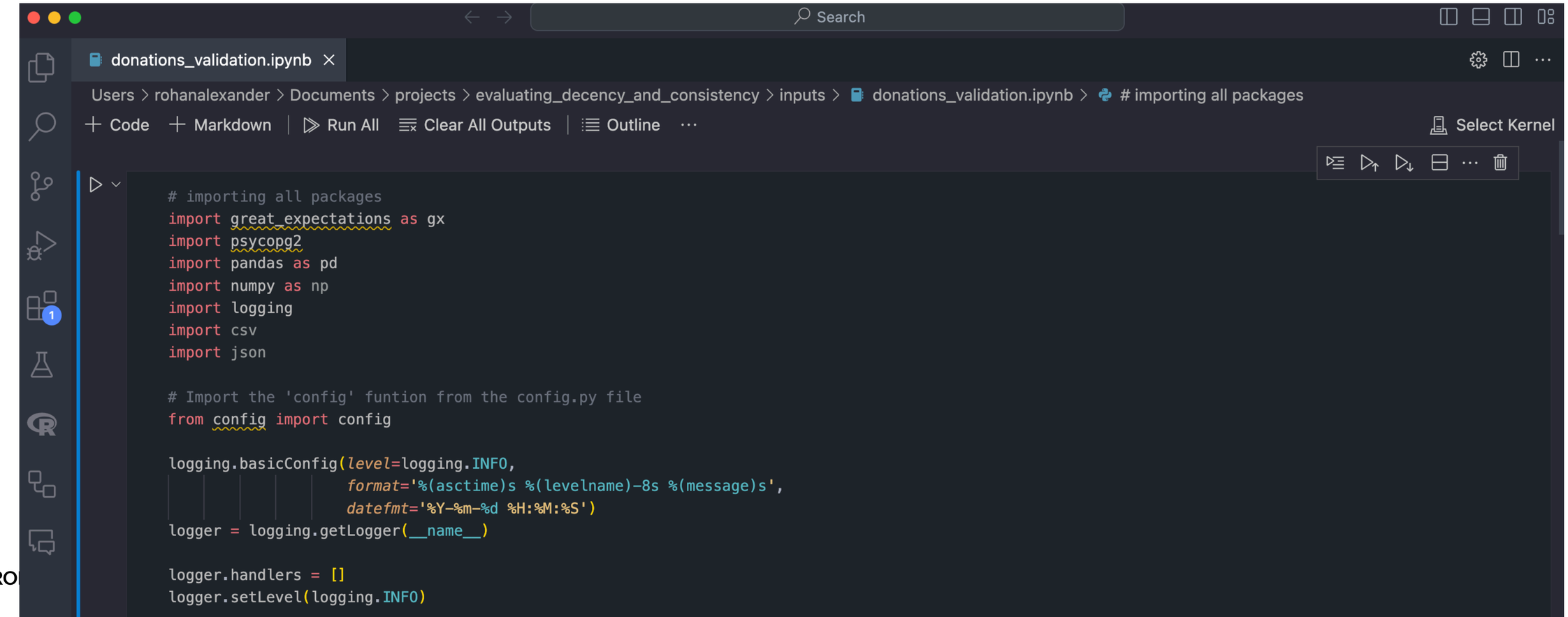
Justice Moreau will fill the vacancy created by the retirement of Justice Russell Brown.

Biographical Note



MAJOR QUESTION

TO WHAT EXTENT CAN LLMS DEVELOP A SUITE OF DATA VALIDATION TESTS THAT IS SIMILAR TO THE SUITE DEVELOPED BY AN EXPERIENCED EXPERT DATA SCIENTIST WHO IS FAMILIAR WITH THE DATASET?



OTHER ASPECTS OF INTEREST

TO WHAT EXTENT CAN LLMS DEVELOP A SUITE OF DATA VALIDATION TESTS THAT IS SIMILAR TO THE SUITE DEVELOPED BY AN EXPERIENCED EXPERT DATA SCIENTIST WHO IS FAMILIAR WITH THE DATASET?

- Four prompt scenarios: 1) Asking for expectations, 2) Asking for expectations with a given context, 3) Asking for expectations after re- questing a simulation, and 4) Asking for expectations with a provided data sample.
- Three learning modes: 1) zero-shot, 2) one-shot, and 3) few-shot learning.
- Four temperature settings: 0, 0.4, 0.6, and 1.
- Two roles: 1) “helpful assistant”, 2) “expert data scientist”.
- Two models: 1) GPT-3.5 and 2) GPT-4.

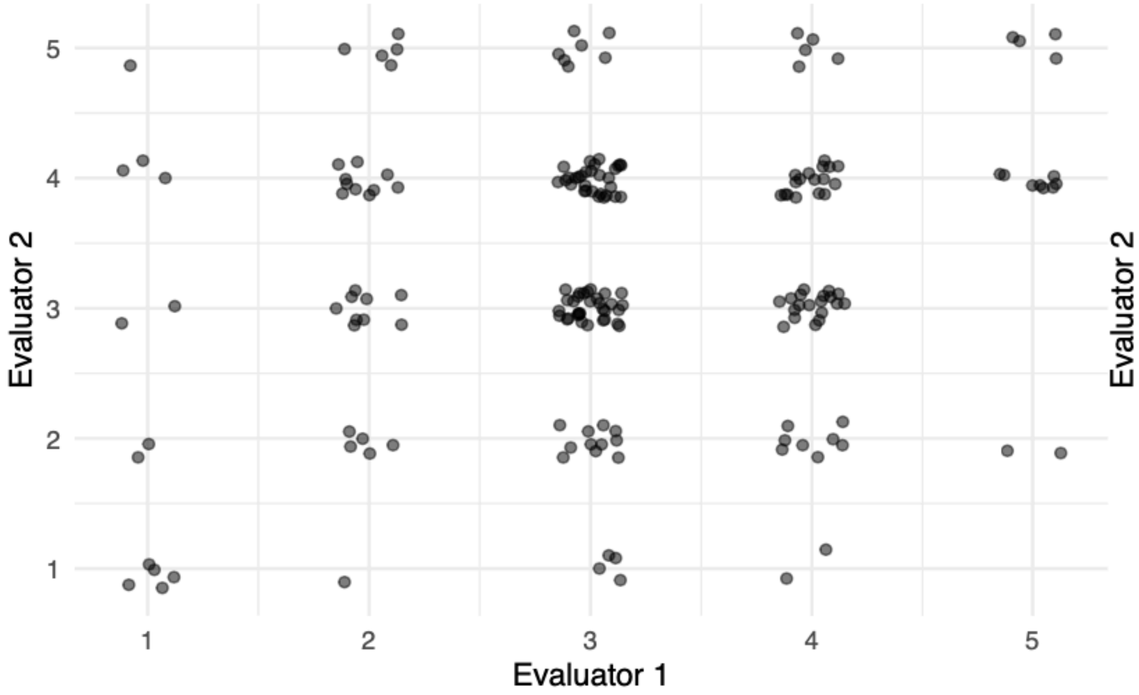
EVALUATION

THREE CODERS EVALUATE: “DECENCY” AND “CONSISTENCY

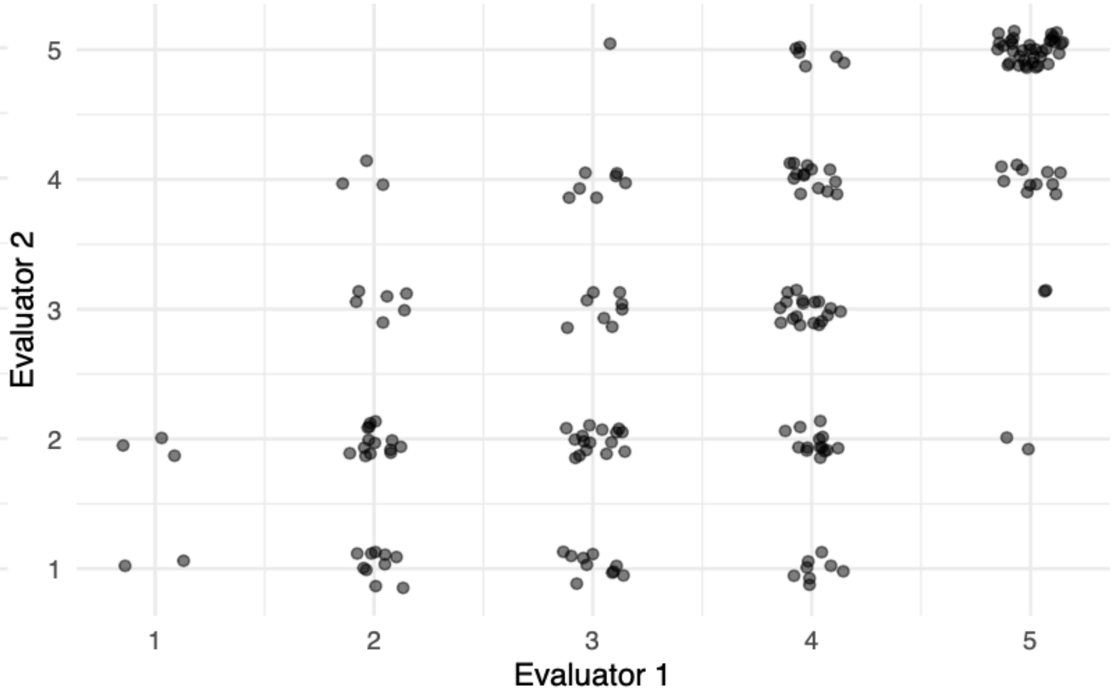
- This combination of variables and options results in 96 different prompt situations. We run these through both GPT-3.5 and GPT-4, using the API. For every combination we ask for five responses to understand variation.
- This results in a dataset of responses. The option that gave rise to each response was blinded and the order randomized, and then the responses were ranked by one experienced human coder on two metrics.
- “consistency” is a ranking 1-5 of how different each of the five responses was for that particular combination of variables.
- “decency” is a ranking of how effective the LLM validation tests were compared with the code written by the experienced data scientist who wrote the original suite of tests.

INTER-CODER AGREEMENT?

SOME CONCERNS...

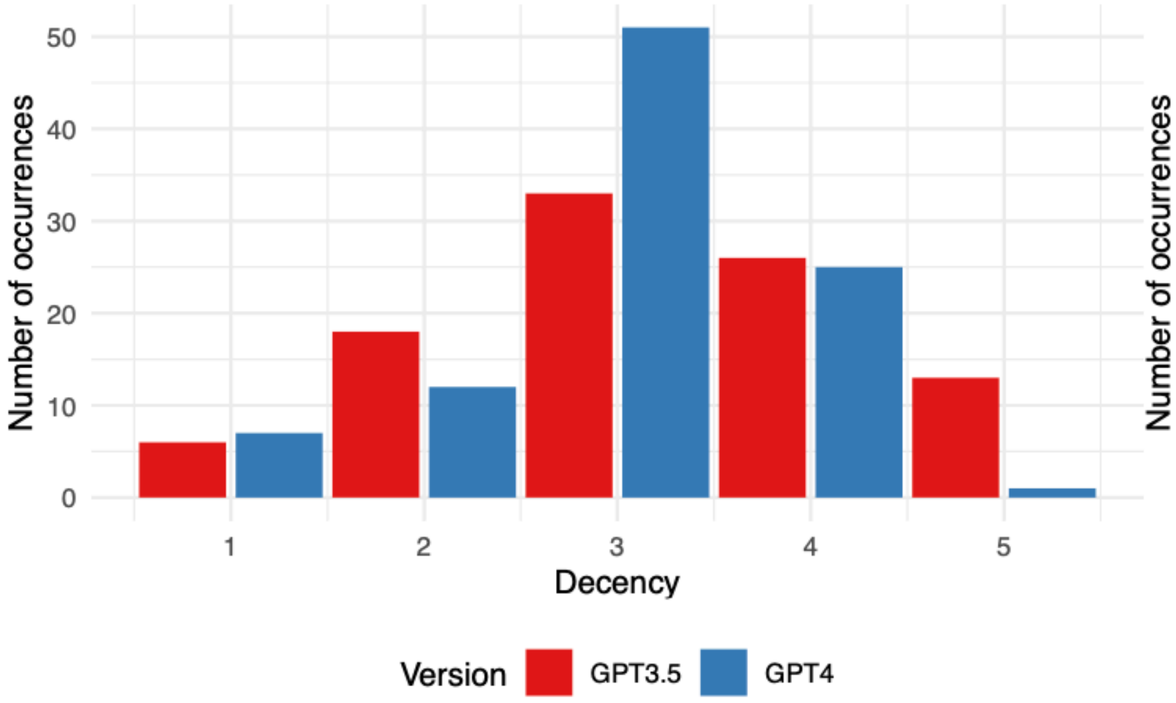


(a) Decency

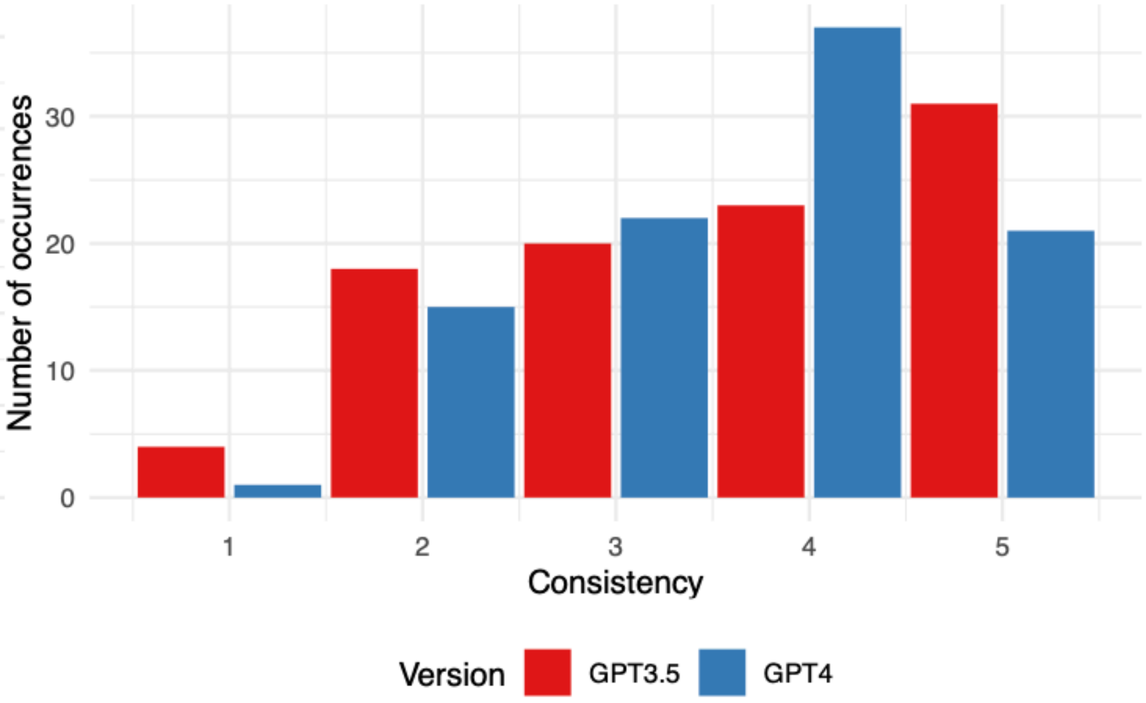


(b) Consistency

GPT-3.5 VS GPT-4

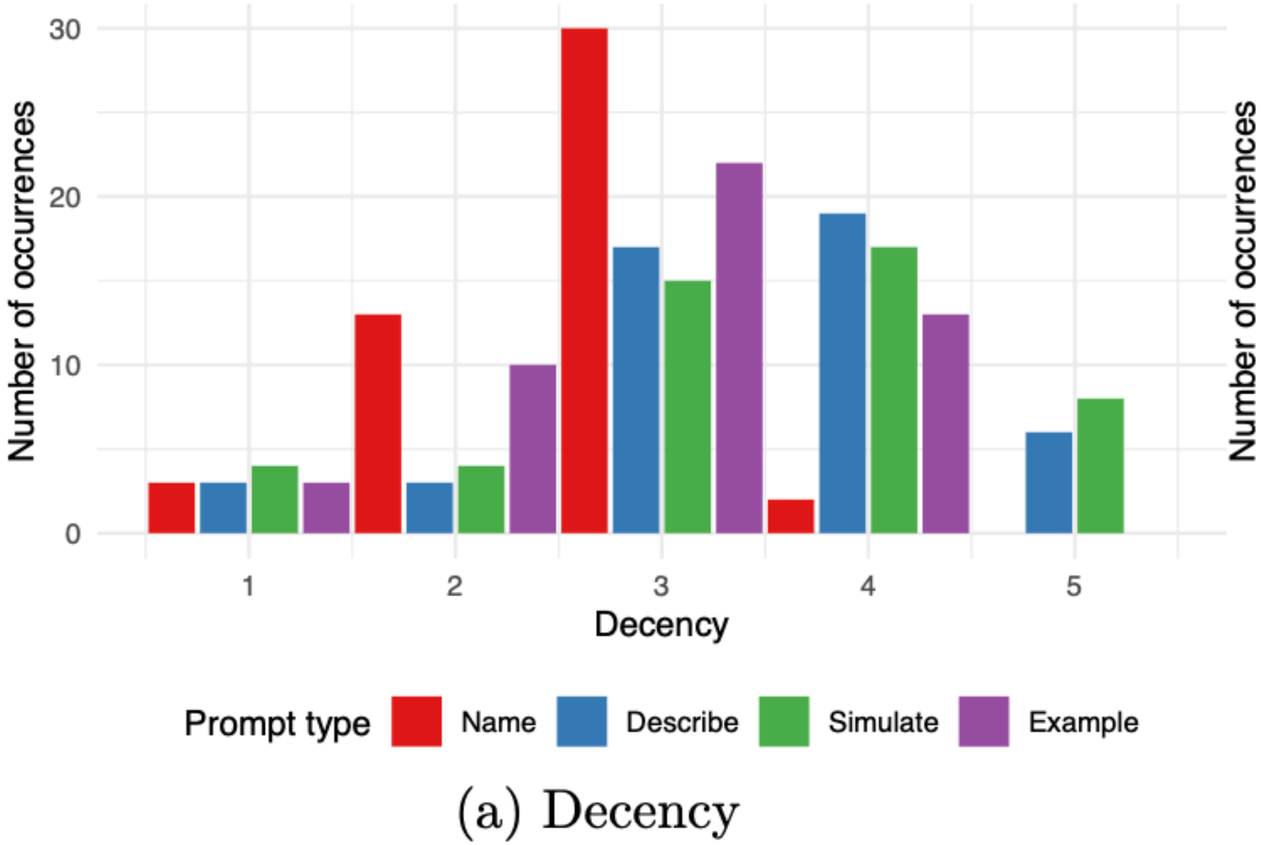


(a) Decency

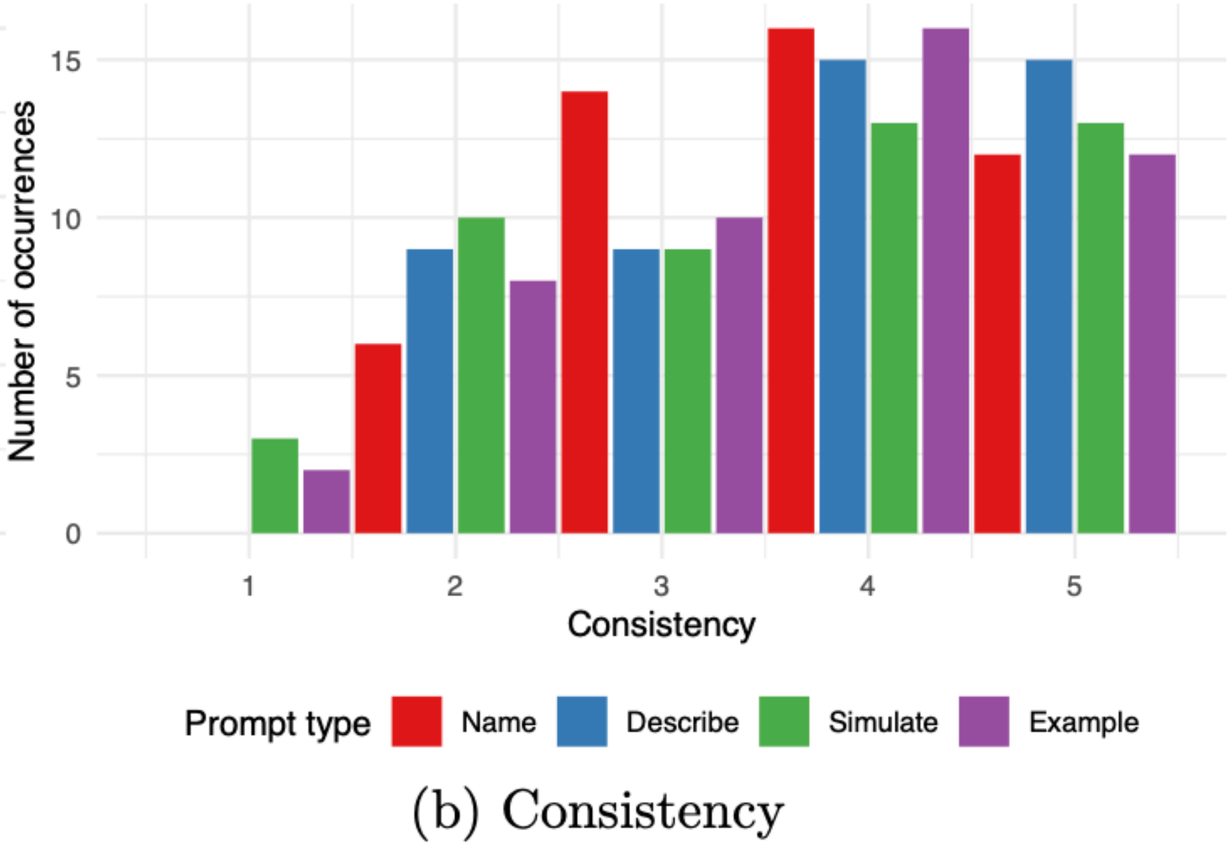


(b) Consistency

DIFFERENT PROMPT STYLES

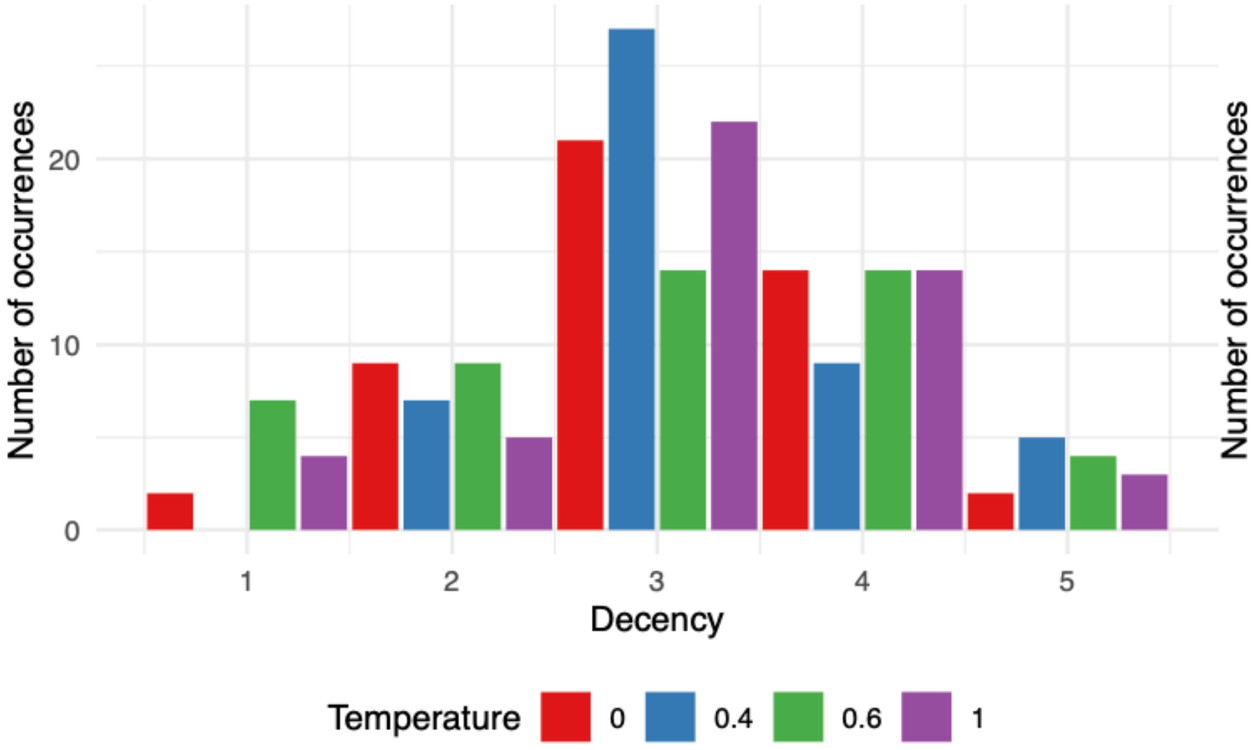


(a) Decency

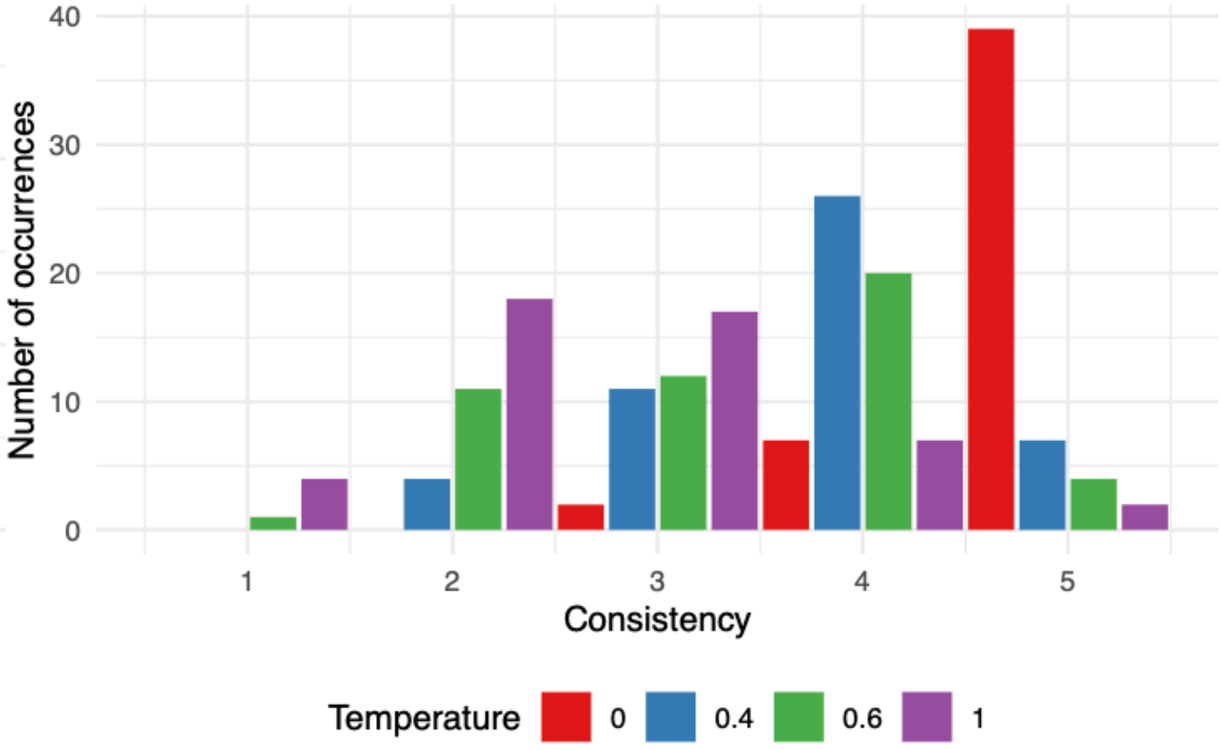


(b) Consistency

TEMPERATURE

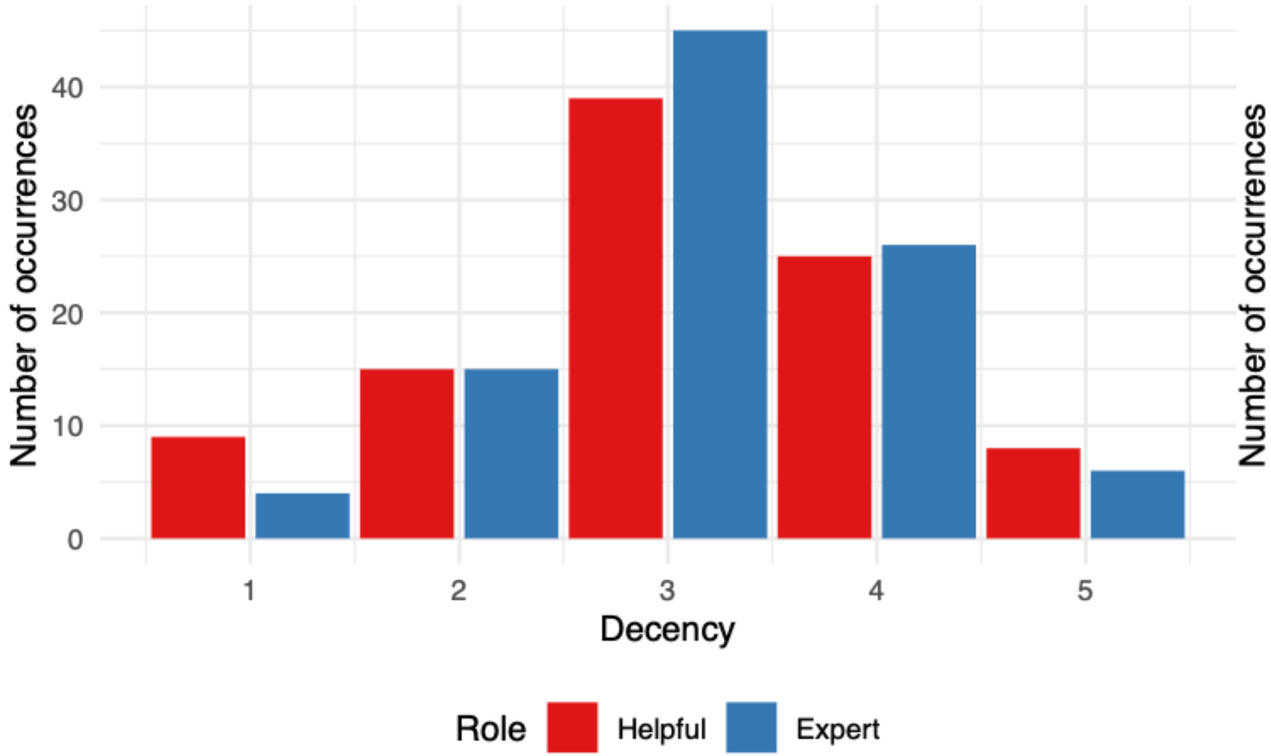


(a) Decency

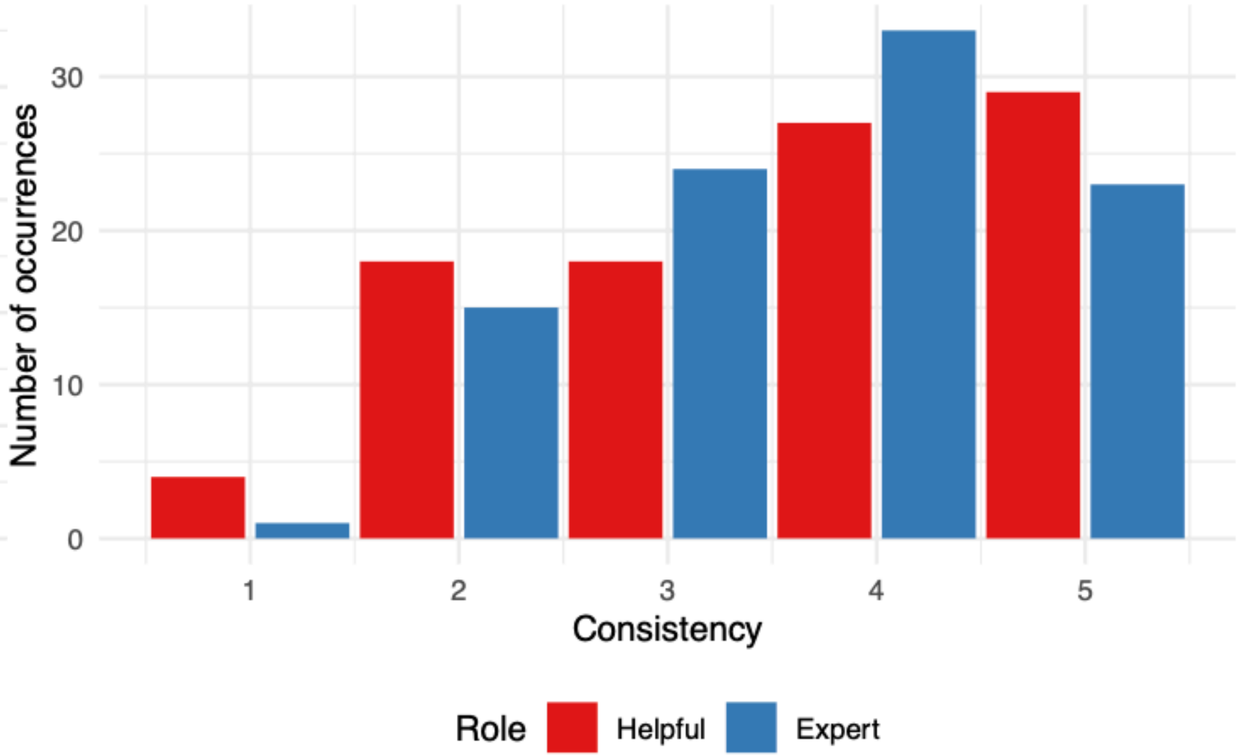


(b) Consistency

ROLE

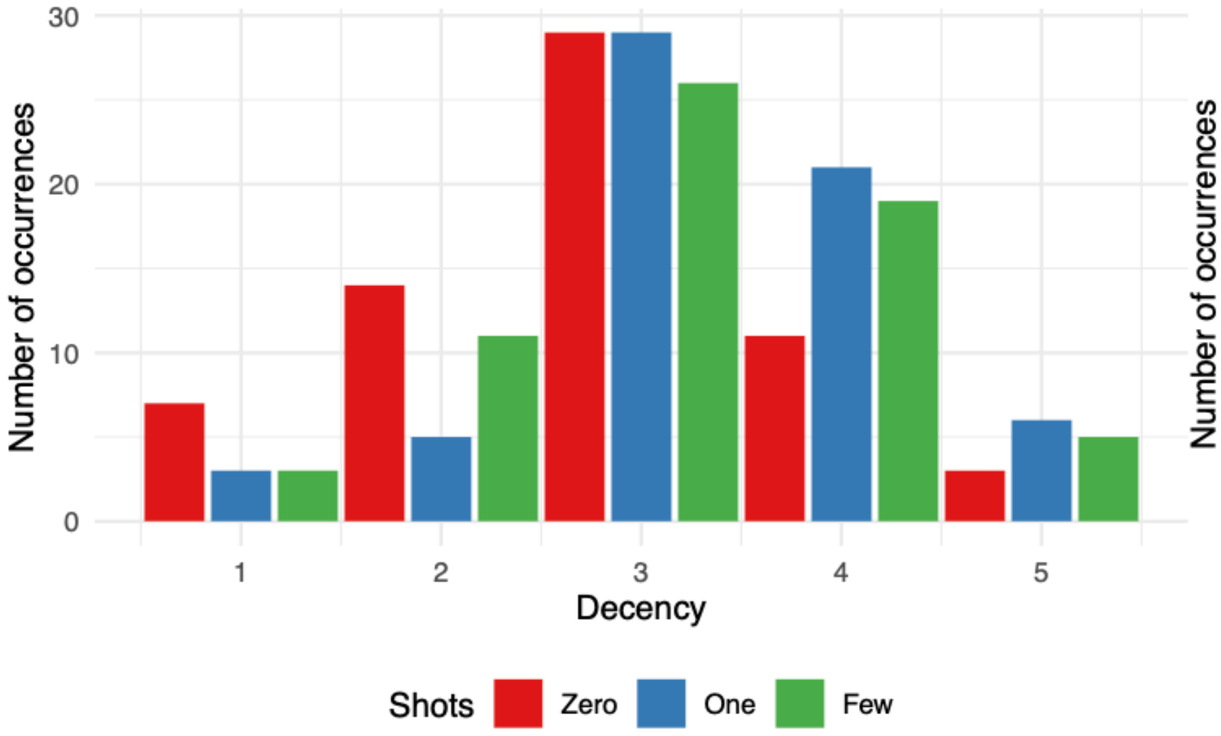


(a) Decency

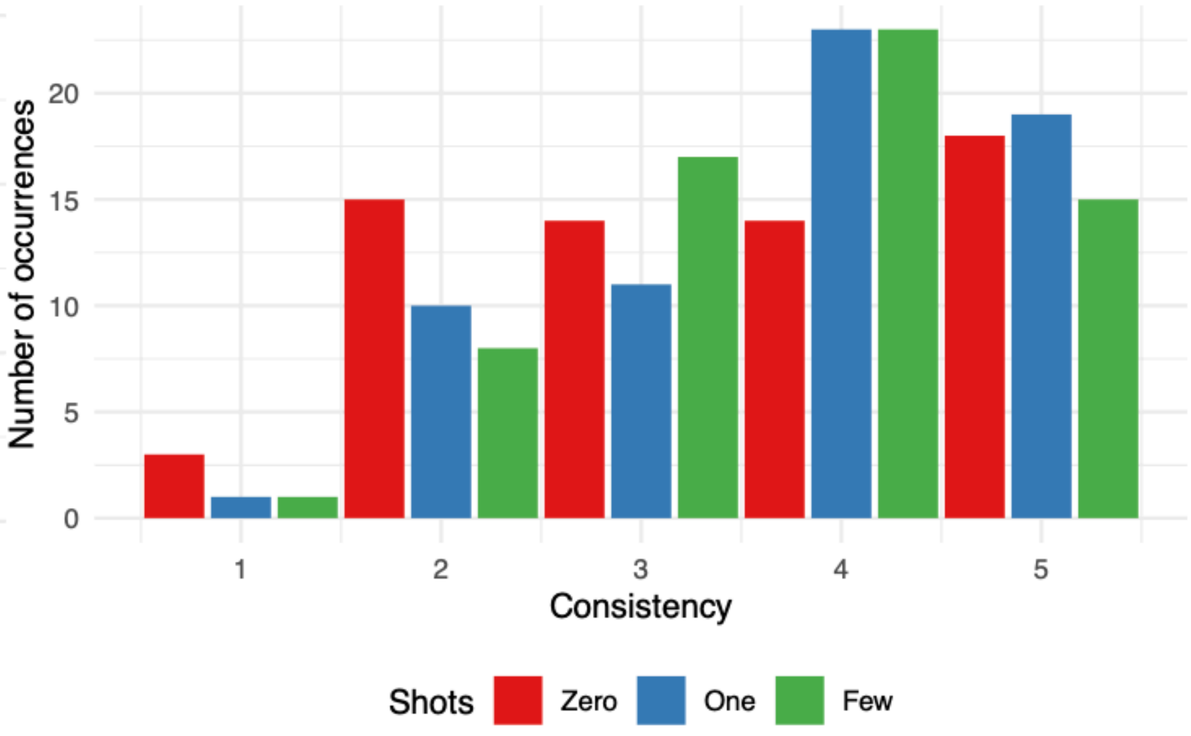


(b) Consistency

LEARNING MODE



(a) Decency



(b) Consistency

MODEL

ORDERED LOGISTIC REGRESSION

- We model the rating Y, which is an ordinal outcome with J = 5 possible categories

$$y = \begin{cases} 1 & \text{if } y^* < \zeta_1 \\ 2 & \text{if } \zeta_1 \leq y^* \\ \vdots & \\ J & \text{if } \zeta_{I-1} < 1 \end{cases}$$

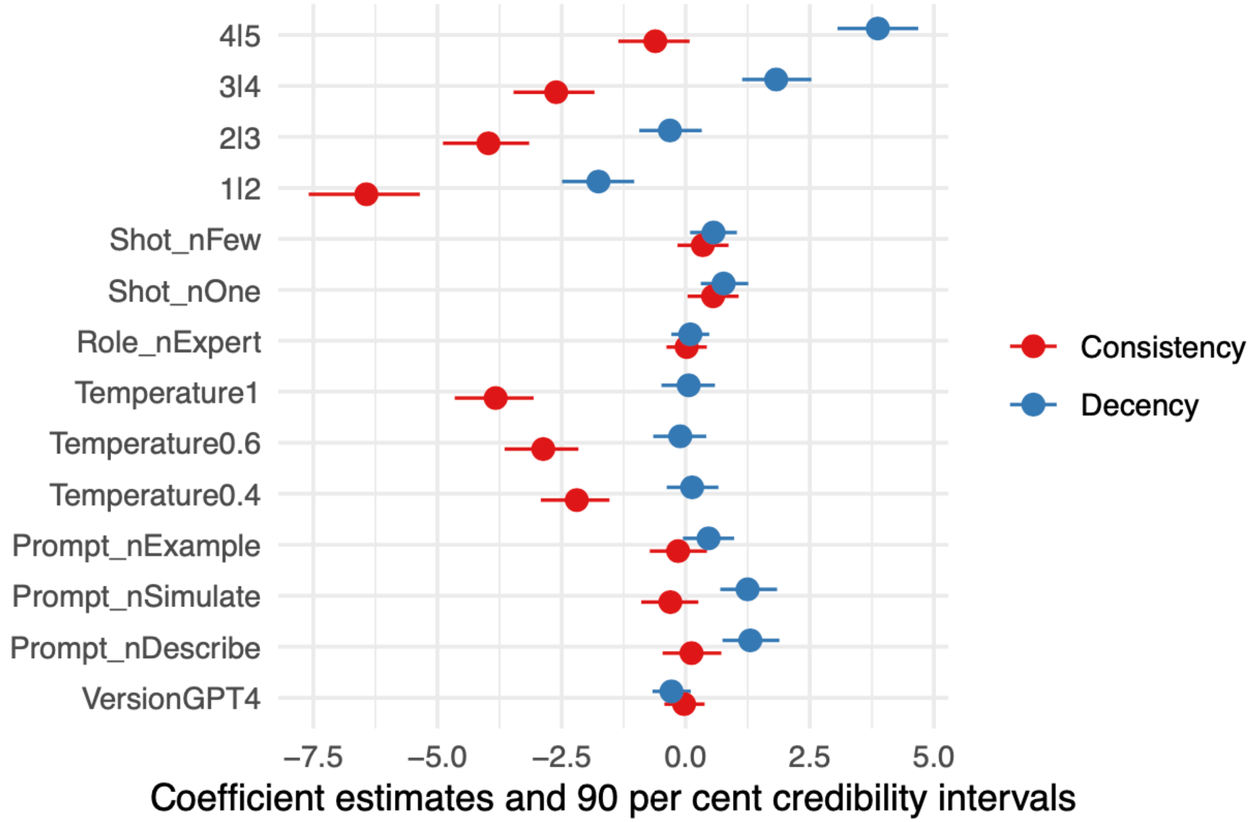
- We are interested in exploring the relationships that consistency and decency have with model, prompt, temperature, role, and learning mode:

$$y^* = \beta_1 \cdot \text{version}_i + \beta_2 \cdot \text{prompt}_i + \beta_3 \cdot \text{temperature}_i + \beta_4 \cdot \text{role}_i + \beta_5 \cdot \text{shot}_i$$

- We fit this model, separately, for each of consistency and decency, in a Bayesian framework using the package rstanarm (Goodrich et al. 2023) and the R statistical programming language (R Core Team 2023).

RESULTS

	Consistency	Decency
VersionGPT4	−0.03	−0.29
	(0.25)	(0.24)
Prompt_nDescribe	0.12	1.30
	(0.35)	(0.36)
Prompt_nSimulate	−0.31	1.25
	(0.36)	(0.34)
Prompt_nExample	−0.15	0.46
	(0.35)	(0.31)
Temperature0.4	−2.19	0.13
	(0.41)	(0.31)
Temperature0.6	−2.87	−0.11
	(0.45)	(0.32)
Temperature1	−3.83	0.06
	(0.50)	(0.33)
Role_nExpert	0.02	0.09
	(0.25)	(0.23)
Shot_nOne	0.55	0.76
	(0.30)	(0.28)
Shot_nFew	0.35	0.56
	(0.31)	(0.29)



CONCLUDING REMARKS

Limitations:

- We are unable to provide fully reliable explanations for differences in performance. (We can make convincing speculations, but these are ultimately educated guesses.)
- Our current experiments have a very limited context: one Canadian political donations dataset.

Future work:

- Increased scale
- Automated evaluation

CONCLUDING REMARKS

Key takeaways:

- Data validation tests are important, and critical in every data science workflow.
- LLMs can be used to generate them, but our work is consistent with previous literature demonstrating that LLM performance on complex, user-defined tasks is sensitive to prompt engineering.
- For now, the context that an experienced data scientist brings is necessary, but LLMs provide a useful first draft.

THANK YOU

rohanalexander.com

@rohanalexander

rohan.alexander@utoronto.ca

https://github.com/RohanAlexander/evaluating_decency_and_consistency