Exercise Set 2: Silicon Subjects

Version 2

Instructor: Daniel Karell (daniel.karell@yale.edu)

January 2026

Overview

The goal of this activity is to explore the effectiveness and limitations of using large language models (LLMs) as substitutes for human participants in social science research. The title of this assignment, "Silicon Subjects" refers to the idea of using LLM-derived (or "silicon" or "synthetic") data as a complement – or even replacement – for "organic", human-produced data.

In this assignment, we will be engaging with insights and analyses from four assigned readings: Argyle, et al. 2023; Bisbee, et al. 2024; Lyman, et al. 2025; and Broska, et al. 2025. Please familiarize yourself with these articles before you start this assignment. You can find the articles in the `Files/readings/` folder on Canvas.

Resources

This is a list of links to external resources which are mentioned in the following instructions:

- Using `R Markdown`:
  - https://rmarkdown.rstudio.com/lesson-1.html
- An overview of Mistral's available models:
  - https://docs.mistral.ai/getting-started/models/models_overview/
- The vignette about how to use the `tidyllm` package:
  - https://edubruell.github.io/tidyllm/articles/tidyllm.html
- Information about `tibbles`, from the `tidyverse`:
  - https://tibble.tidyverse.org/articles/tibble.html
- Information about the World Values Survey:
  - https://www.worldvaluessurvey.org/wvs.jsp
- Information about Earth Mover's Distance:
  - https://search.r-project.org/CRAN/refmans/emdist/html/emd.html
  - https://cran.r-project.org/web/packages/emdist/emdist.pdf

The Teaching Fellow and Instructor are also resources! Please feel free to ask for help as needed. Do not wait until the last minute.

Prepare your workspace

First, load the following packages. They should already be installed since they were used for Exercise Set 1.

```
library(tidyllm)
library(tidyverse)
```

Now, prepare to use the Mistral LLM, which was introduced in Exercise Set 1. You will need the API key that you used for those exercises. (See Exercise Set 1 for instructions on obtaining an API key.)

Once you have access the API key, specify it for use with Mistral. Please replace YOUR-MISTRAL-API-KEY in the following line of code with your own API key such that your API key is in double quotations.

Recall that you may use models from a different source. If plan to do so, you should change the following code. For details on how to do this for various models, refer to the `tidyllm` vignette at this URL: https://edubruell.github.io/tidyllm/articles/tidyllm.html.

```
Sys.setenv(MISTRAL_API_KEY = "YOUR-MISTRAL-API-KEY")

# For example:
# Sys.setenv(MISTRAL_API_KEY = "a1b2c3d4e5")
```

Finally, look at the following lines of code, but do not execute the code yet. This code is an approximate and (very) condensed version of what the authors of the assigned readings used. In brief, it constructs a series of particular prompts and passes them to the model; the prompts specify various characteristics of hypothetical survey respondents, such as their educational attainment and whether they answered the survey in 2020 or 2024, then poses a survey question to the LLM.

The survey question is adapted from the American National Election Studies (ANES) survey: a "feeling thermometer" probing how respondents feel about Democrats and Republicans. Respondents are asked to use a numeric scale between zero and 100 to indicate how warmly they feel about the groups. Values closer to zero mean indicate "cold", or negative, feelings and values closer to 100 indicate "warm", or positive, feelings.

Take the time to read the following lines of code and understand what each line will do once you run the code. In the exercises below, this code chunk will be referred to as the "silicon subjects" code.

Note that I have written the code to enhance comprehension and interpretability, not speed or efficiency. When you do execute the following lines of code, it should take about 12 minutes to complete.

```
## set respondent characteristics
years <- c("2020", "2024")
ages <- c("28", "68")
raceths <- c("non-Hispanic White", "non-Hispanic Black")
genders <- c("man", "woman")
educations <- c("high school diploma", "graduate degree")
incomes <- c("$30,000", "$150,000")
group <- c("Democrats", "Republicans")

## dataframe to record responses
response_dataframe <- tibble(
  years = NA,
  ages = NA,
  raceths = NA,
  genders = NA,
  educations = NA,
  incomes = NA,
  group = NA,
  feeling = NA)

## generate the data
for(y in 1:length(years)){
  for(a in 1:length(ages)){
      for(r in 1:length(raceths)){
        for(g in 1:length(genders)){
          for(e in 1:length(educations)){
            for(i in 1:length(incomes)){
              for(p in 1:length(group)){
```

```
  # create the prompt
  text <- paste("It is", years[y], ". You are a", ages[a], "year-old",
                raceths[r], genders[g], ". Your highest level of educational
                attainment is a", educations[e], "and you make", incomes[i],
                "each year. Provide responses to the following questions only
                from your particular perspective. The following questions ask
                about individuals' feelings toward different groups. Responses
                should be given on a scale from 0 (meaning cold feelings) to
                100 (meaning warm feelings). Ratings between 50 degrees and
                100 degrees mean that you feel favorable and warm toward the
                group. Ratings between 0 degrees and 50 degrees mean that you
                don't feel favorable toward the group and that you don't care
                too much for that group. You would rate the group at the 50
                degree mark if you don't feel particularly warm or cold toward
                the group. What number indicates how you feel toward the
                following groups? Remember, your response should be only a
                number between 0 and 100. Do not include any other explanation
                of discussion. The group is: ", group[p], sep = " ")

  # give prompt to model and obtain response
  conversation <- llm_message(text) |> chat(mistral()) |> as_tibble()

  # record response
  tmp_dataframe <- tibble(
    years = years[y],
    ages = ages[a],
    raceths = raceths[r],
    genders = genders[g],
    educations = educations[e],
    incomes = incomes[i],
    group = group[p],
    feeling = conversation$content[3])
  response_dataframe <- bind_rows(response_dataframe, tmp_dataframe)

  # print progress
  print(paste(years[y], ages[a], raceths[r], genders[g],
              educations[e], incomes[i], group[p], sep = "_"))
  print(conversation$content[3])

  # pause five seconds to respect query limit
  Sys.sleep(3)
            }
          }
        }
      }
    }
  }
}

## finalize the dataframe
response_dataframe3 <- response_dataframe |> drop_na(group)
```

Please ask for help if you are not sure how to prepare using the LLM and/or do not understand the preceding

lines of code. Do not wait until the last minute to ask for help!

Using different models

As mentioned earlier, I suggest that you use Mistral's models because they are "open" and it is easy to set up a free account. However, you are free to use other LLMs if, for example, you are already using another model for your ongoing projects or if you encounter difficulty with Mistral's models. https://edubruell.github.io/tidyllm/articles/tidyllm.html.

Fortunately, the `tidyllm` package makes it easy to use a variety of models, including those in the Claude, Gemini, OpenAI, and DeepSeek families. For instructions on how to do this, see the package's vignette: https://edubruell.github.io/tidyllm/articles/tidyllm.html.

If you do decide to use a model to respond to a question that differs from the one in the question's instructions, simply note the model that you did use in your response. For example, write "To answer this question, I used models X and Y, provided by Z."

Instructions for submission

Please submit your completed problem set as `PDF` or `HTML` file generated using an `R Markdown` document. For a beginner's guide to creating an `R Markdown` document, see https://rmarkdown.rstudio.com/lesson-1.html.

A completed assignment includes: (a) your code; (b) the output of your code; and (c) a written answer that provides a response to each question by explaining and interpreting the output that your code produced. Upload the completed problem set to the "Assignments" page of the class's Canvas site. This activity will be graded on a 100-point scale.

Please note! When you knit your assignment (either to `HTML` or `PDF`), all your code will run again. This may have important implications: if a new run of your code produces slightly different output, then comparisons to other results not based on code and/or your discussion of the results (text you've written) may no longer be aligned with the results obtained from the code.

To avoid this issue, please adopt the following workflow. First, as you develop the code to address each exercise, write it in a Markdown code chunk:

```
# ```{r}
# example <- code(about = something)
# ```
```

Then, once you have obtained the results you consider "final", store or save your results as an R object.

```
# ```{r}
# save(output, file = "classwork\saved_results\output.Rda")
# ```
```

Next, when you are ready to knit, add a command to *each* code chunk that prevents the code from running, or being evaluated. The command is `eval=FALSE`, and it goes it in the curly brackets, like this:

```
# ```{r, eval=FALSE}
# example <- code(about = something)
# ```
```

The above command will ensure that the code chunk will not run, or "evaluate", during knitting. However, you still need to show your output or results when you knit the document for submission. So, after each code chunk that shows what code you used to get your results, include a new code chunk that simply prints the results. Doing this will put into the document: (a) your code (which will not run when you knit) and (b) your output. To show the output, something like the following chunk should work, as long as `eval=TRUE` since you want *this* code chunk to run!.

```
# ```{r, eval=TRUE}
# load(file = "classwork\saved_results\output.Rda")
# print(output)
# ```
```

Please let the teaching team know if you have any questions.

Exercises

Part 1

1. Run the "silicon subjects" code above. To do so, select four types of respondents – for example, a 28 year-old non-Hispanic Black man with a graduate degree who makes $150,000, a 68 year-old non-Hispanic White woman with a high school diploma who makes $150,000, and so on. At least two of these respondents should have different genders. *The code should take about 12 minutes to complete.* Then, for each of these four respondent types, calculate the mean "thermometer" value for Republicans and for Democrats. Report the means per respondent type. (You could use a table or some other format that is equally clear.) Do these mean values align with what you would expect? Why or why not? [10 points]

2. Run the "silicon subjects" code five more times for each of the four types of respondents that you used in the preceding exercise, saving all of the responses. (This, in theory, is analogous to surveying five more respondents who fit a particular demographic and socioeconomic profile.) Then compute the distributions of the "thermometer values" using the responses from all six runs of the code. Report the distributions per respondent type in a table or figure. Do the distributions align with what you would expect? Why or why not? [10 points]

3. Drawing on what you have read in the assigned readings on "silicon sampling" and synthetic data, how do you interpret the means and variances of the responses? How would you explain the observed similarities or differences between the means and variances across LLM-respondent-types? In your answer, please refer to (and cite) specific arguments in the readings. [10 points]

Part 2

At this point, start a new session (or "conversation") with the LLM, but be sure to save your results from Part 1.

Begin this part of the activity by viewing and learning about the types of models currently available from Mistral.

First, view the available models by running the following line of code.

```
print(n = 67, mistral_list_models())
```

If you were doing your own research project, you should investigate what is distinct about each model and pick the one best suited to examine your research topic. You can do that by perusing this website, which offers an overview of Mistal's models available via the API: https://docs.mistral.ai/getting-started/models/models_overview/. Alternatively, you would conduct your analyses using multiple models (and not just ones from Mistral). However, for the purposes of this activity, you will use `devstral-small-latest` and `ministral-3b-latest`. Read the website with the overview of Mistral's models (and ideally read other linked-to online sources, like Mistral's blog) to learn how `devstral-small-latest` and `ministral-3b-latest` differ from one another, as well as how they differ from the default model we have been using, `mistral-large-latest`.

For the next few exercises, you will need to specify which model to use. You can do that by adding an argument to the `chat(mistral())` function, like this:

```
conversation <- llm_message("Write me a country song about \"gear jammin\'
                            double clutch grabbin\' coffee drinkers\".") |>
  chat(mistral(.model = "ministral-3b-latest"))
```

4. Execute the "silicon subjects" code twelve more times: six times with the `devstral-small-latest` model and six times with the `ministral-3b-latest` model. Store all your results. Then, for the same four respondent types used in Part 1, compare the mean and distributions that you obtained when using each of the three models (*i.e*, `mistral-large-latest`, `devstral-small-latest`, and `ministral-3b-latest`). Present the comparisons in a table or figure. In addition, explain, in a written answer, how the outputs are similar and different across the models. [10 points]

5. Drawing on what you have read in class, your understanding on how the three models differ, and any other independent research you conduct, why do you think the results do or do not differ across the models? That is, are there aspects of the models that may explain the responses (or some responses) they tended to give, relative to the other models? If so, why do you think this? If there are not, why not? If the distributions (or parts of the distributions) in Exercise 4 are similar to one another, why do you think they are similar despite using different models? [10 points]

Part 3

At this point, start a new session (or "conversation") with the LLM.

From the `Files/exercise_sets/activity2/` folder in Canvas, download the `PDF` file called `wvs_wave7_codebook.pdf`. This is the codebook for wave 7 of the World Values Survey (WVS), which was conducted from 2017 through 2022. Also download the actual WVS wave 7 data from the same folder. The data is stored as an `R` object in the file `wvs_wave7_data.rda`. The WVS is a long-running, primarily cross-sectional survey probing the attitudes and beliefs of residents of many different countries. You can find out more about the WVS here: https://www.worldvaluessurvey.org/wvs.jsp.

Before beginning this part of the activity, ensure that you have read the assigned reading by Boeleart and colleagues (2025). The following exercises are inspired by that article.

Open the codebook `PDF` file and familiarize yourself with three sets of variables: "Technical Variables", "Social Values, Norms, Stereotypes" (in the "Core Variables") section, and "Demographic and Socioeconomic Variables". Select *four* variables that describe respondents' characteristics from the "Technical" and "Demographic and Socioeconomic" sets of variables *and* which could be made binary. For example, the variable `Q263` measures whether they are an immigrant or not (with values 1 or 0) and the variable `B_COUNTRY_ALPHA` describes the country a respondent resides in, which could be turned into, say, "residing in Germany" (1) or "not residing in Germany" (0).

After selecting four variables about respondents, select *three* variables that record respondents' questions to survey questions in the "Social Values, Norms, Stereotypes" section. For example, `Q12` asks respondents whether they think it is important for children to learn "tolerance and respect for other people" and `Q29` asks respondents how strongly they agree or disagree with the statement that "men make better political leaders than women do."

6. Report and describe the four variables characterizing respondents that you have chosen, including the variables' codes (*e.g*, `B_COUNTRY_ALPHA`, `Q263`), and the three variables corresponding to questions asked of respondents, including their codes (*e.g.*, `Q1`, `Q33`). If you need to give the demographic variables binary encodings, do that now, and report the binary transformation. Also report all the *types* of respondents you have, as defined by each unique combination of the four technical and demographic variables you selected. How many types are there and what are their profiles? Report this information in a written list or with a table. [5 points]

7. Update the "silicon subjects" code to use the four variables characterizing respondents that you selected and to ask the three questions survey questions that you selected. You can opt to have three versions of prompts – one for each survey question – or to ask all three questions in one prompt. You should select the option that you think will produce better output. Also select a model to use. It can be one of the three used in Parts 1 and 2, or another Mistral model. Report your prompt(s) and explain which prompting option you chose (*i.e.*, three different prompts, one prompt, or something else) and why. In addition, explain which model you chose to use and why. [5 points]

8. Execute the code six times. This should produce multiple responses to each question from each type of respondent listed in your answer to Exercise 6. How many totals responses do you have for each of the three questions? How many responses to each question from each type of respondent? [5 points]

9. Using the LLM output, report how frequently each type of "silicon respondent" – the types listed in the Exercise 6 answer – provided each answer option. That is, how many times did each type of (synthetic) response indicate, for example, "very important", "rather important", and so on (assuming that those were the possible response answers)? I suggest reporting this information as a figure, but you can also use a table. The information you report should be a distribution of answers for each type of "silicon respondent". [5 points]

10. Load the WVS data into your `R` environment. To load the data file, use the code `load("your/path/to/wvs_wave7.rda")`. Then, identify how human respondents who match the profiles of your "silicon respondents" in terms of technical and demographic characteristics responded to each of the three questions you selected earlier. Your answer to this question should be a figure or table like your created for Exercise 9, but based on data that were generated by humans and recorded in `wvs_wave7_data.rda`. [10 points]

11. Compare the distribution of survey question answers from the "silicon respondents" (from Exercise 9) to the distribution of answers from the human subjects (from Exercise 10). You can conduct this comparison in whichever way that you think is most effective. You can create a new figure and/or a new table, and/or use a statistical technique, like Earth Mover's Distance (EMD), which is what Boeleart and colleagues use in their article. If you do choose to use EMD (which is not necessary), you can implement it with `R` using the `emdist` package. To learn more about the package, see the package documentation: https://search.r-project.org/CRAN/refmans/emdist/html/emd.html and https://cran.r-project.org/web/packages/emdist/emdist.pdf. Explain how you conducted the comparison and why. Also report the results of your comparison. [10 points]

12. After comparing the distributions of responses from the human and synthetic respondents, discuss what you could have done differently to obtain a distribution of responses from the LLM that better matched the distribution of human responses. When appropriate, draw on insights from the assigned readings. [10 points]