Assignment 1 Report

Brandon Tran (M.S. S&DS)
brandon.tran@yale.edu

January 26, 2026

1 Human vs. LLM Perceptions of Grammar

1.1 Introduction and Abbreviated Methodology

This report expands on the FiveThirtyEight poll of Americans on the Oxford comma (as well as whether or not the term "data" is plural). As part of our analysis, we will compare responses across five groups:

1. **Raw Human Survey Responses**
   - $n = 1,129$
   - Responses from actual humans as part of the original 538 study.
   - Not post-stratified to match the US population.
2. **GPT In Silico for Original Demographics**
   - $n = 300$
   - GPT personification of 300 random samples from the original study population.
   - Not post-stratified to match the US population.
3. **Post-Stratified Estimates for Human Population**
   - $n = 1,129$
   - Same responses as from (1).
   - Post-stratified to match the US population based on 2024 Census data.
4. **Post-Stratified Estimates for In Silico Population**
   - $n = 300$
   - Same responses as from (2).
   - Post-stratified to match the US population based on 2024 Census data.
5. **GPT Responses for In Silico Population**
   - $n = 1,129$
   - GPT in silico responses as personas sampled from 2024 Census data.
   - Representative by design via stratified sampling.

This report will first contrast the responses of groups 1 and 3 as a classical reflection on post-stratification. Then, we will contrast the responses of groups 1 and 2 to gauge the capability of an LLM to mimic different personas (defined by the following demographics: gender, age, household income, education, and location defined by Census region). We will see if the insights here extrapolate to groups 4 and 5. Particularly, do we observe the same post-stratified shifts in the LLM responses (2 vs. 4) as we did with human survey responses (1 vs. 3)? Furthermore, do our in silico responses match our post-stratified LLM estimates (4 vs. 5), indicating effective post-stratification?

Our methodology follows what is outlined in the assignment description, with the caveat that rather than querying OpenAI's `gpt-4o-mini` API, we query Google Gemini's `gemini-2.0-flash` for cost effectiveness.

1.2 Reflection on Steps 1–3 (Raw Data Exploration)

To ground our understanding of the FiveThirtyEight sample, we present the original demographic distribution and responses for the 1,129 responses.

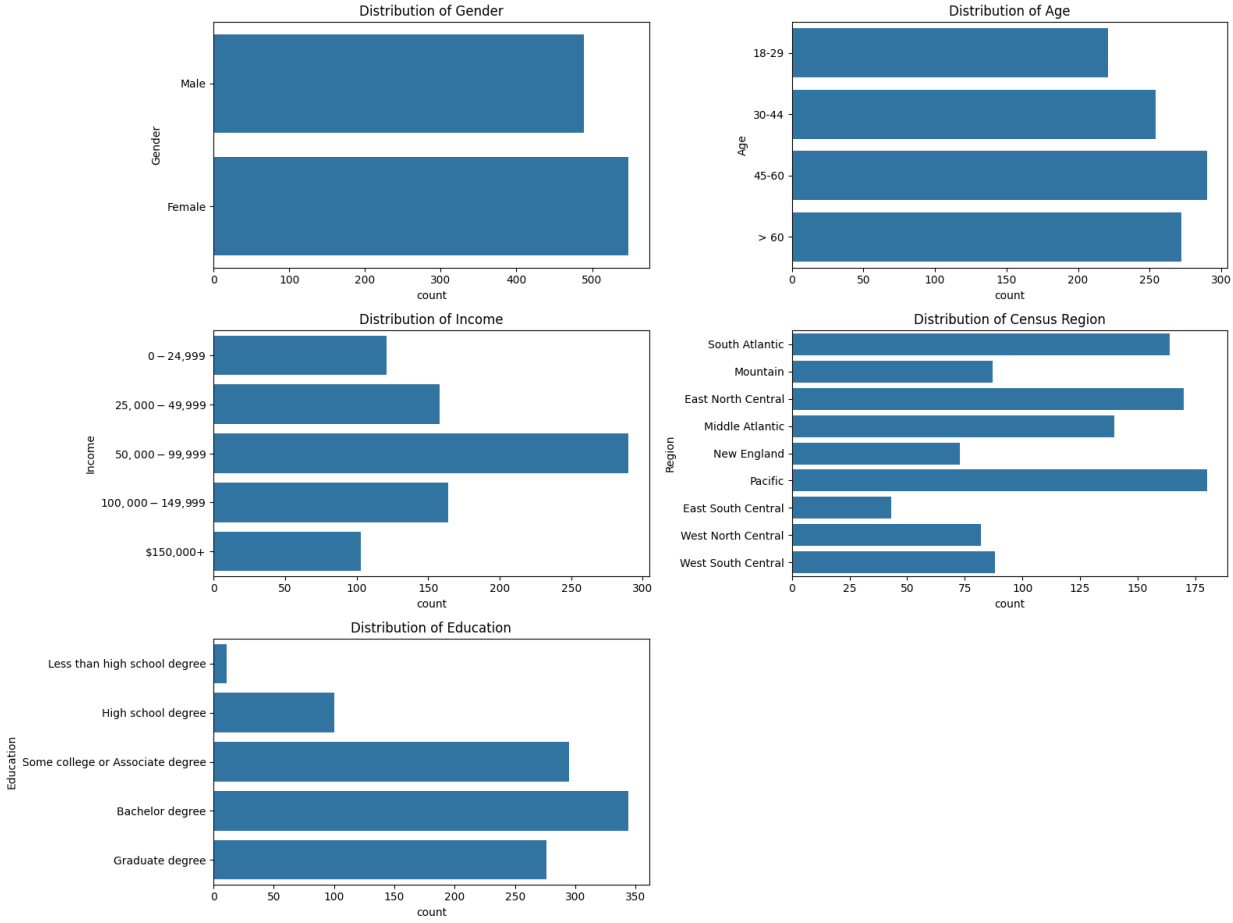

Figure 1: Demographics of initial survey administered to 1,129 humans (Group 1).

Key insights to glean from Figure 1 are that, on average, survey respondents tended to be more educated with slightly higher than median incomes (note that the first two income buckets are length 25,000, whereas the remaining income buckets are length 50,000).

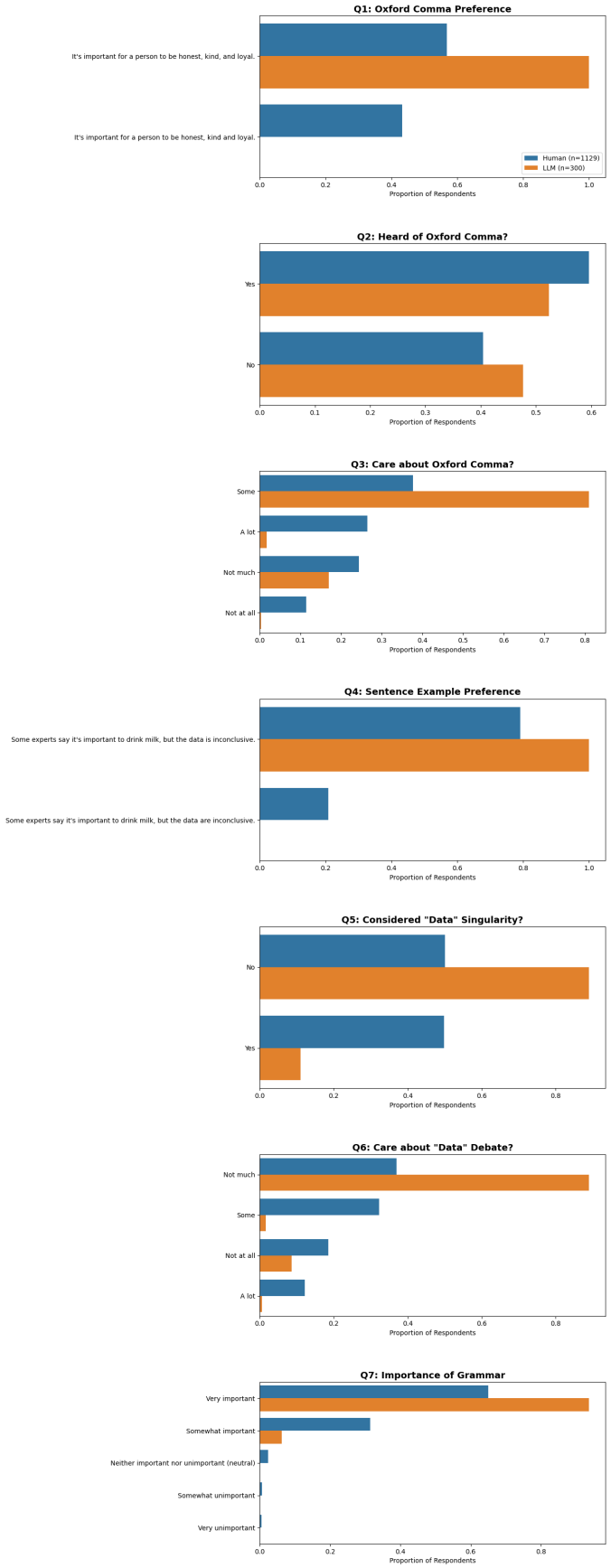

Figure 2: Comparison of raw human (Group 1) and raw LLM (Group 2) responses.

From the actual human responses themselves in blue (Figure 2), there is a skew towards viewing proper grammar use as important (most respondents selected "somewhat" or "very important"). While there was no clear consensus on whether to use the Oxford comma (though responses leaned towards affirming usage), most respondents viewed the term "data" as singular, as opposed to plural. On the other hand, note that LLMs (in orange on Figure 2) suffered from mode collapse, resorting exclusively to usage of the Oxford comma in Q1 and exclusively to the usage of data as a singular noun in Q4. Interestingly, LLMs seemed to choose a very "middle-of-the-road" stance on whether they care about the Oxford comma—over 80% of responses were "Some"—despite their exclusive usage of it.

1.3 Reflection on Steps 4–7 (Post-Stratified Data)

For simplicity, we present a table of responses across each group defined in Section 1. We compared Groups 1 and 2 previously in Section 1.2.

Table 1: Comparison of Human vs. LLM Responses

Question / Answer	(1) Raw H	(2) Raw L	(3) PS H	(4) PS L	(5) Census L
Q1: Which sentence is more grammatically correct?					
Honest, kind and loyal (No Comma)	43.2%	0.0%	46.7%	0.0%	0.0%
Honest, kind, and loyal (Oxford)	56.8%	100.0%	53.3%	0.0%	100.0%
Q2: Heard of the serial (or Oxford) comma?					
No	40.4%	47.7%	47.4%	44.1%	46.2%
Yes	59.6%	52.3%	52.6%	55.9%	53.8%
Q3: Care about the Oxford comma?					
A lot	26.5%	1.7%	26.4%	1.6%	0.5%
Not at all	11.5%	0.3%	12.7%	0.6%	0.8%
Not much	24.4%	17.0%	25.9%	24.5%	30.6%
Some	37.7%	81.0%	35.0%	73.3%	68.1%
Q4: How would you write the following sentence?					
...but the data **are** inconclusive.	20.9%	0.0%	17.6%	0.0%	0.0%
...but the data **is** inconclusive.	79.1%	100.0%	82.4%	0.0%	100.0%
Q5: Considered if "data" is singular/plural?					
No	50.1%	89.0%	55.5%	90.9%	95.0%
Yes	49.9%	11.0%	44.5%	9.1%	5.0%
Q6: Care about the "data" debate?					
A lot	12.2%	0.7%	11.3%	0.7%	0.3%
Not at all	18.6%	8.7%	22.2%	10.1%	16.6%
Not much	36.9%	89.0%	36.7%	88.3%	82.5%
Some	32.3%	1.7%	29.8%	1.0%	0.7%
Q7: Importance of proper grammar?					
Neutral	2.5%	0.0%	3.8%	0.0%	0.0%
Somewhat important	31.4%	6.3%	30.6%	8.0%	8.4%
Somewhat unimportant	0.7%	0.0%	0.6%	0.0%	0.0%
Very important	65.0%	93.7%	64.7%	92.0%	91.6%
Very unimportant	0.5%	0.0%	0.4%	0.0%	0.0%

In comparing the raw and post-stratified human data (group 1 versus 3), we observe a drop of 3.5% in those who favor the Oxford comma in Q1. We see a large drop of 7% in familiarity with

the Oxford comma in Q2, as well as a 3.3% increase in the interpretation of data as singular in Q4. There is a slight growth in those who care about the "data" debate in Q6; otherwise, responses stay in similar ranges. This could perhaps be a reduction of education bias, which we observed in our analysis of the original demographics. In any case, it seems that demographics had a heavier impact on actual deployment of "proper" grammar usage (i.e., Q1 and Q4) than on perceptions of the importance of grammar.

Moving to the LLM responses (group 2 versus 4), first notice that we were not able to replicate the 100% Oxford comma usage in Q1 during post-stratification because the logistic regression models failed to converge (resulting in 0.0% estimates) because the training data contained only a single class. This confirms mode collapse in the LLM's initial responses; the same effect can be observed in Q4. Similarly, when comparing group 2 versus 5, we would expect to see some demographic variance. However, the LLM retained 100% usage of both the Oxford comma and data as a singular noun. We conclude that while post-stratification may be helpful in the human context, it is less helpful when there is no actual variation across simulated demographics. Our helpful LLM agent is likely incapable (at this time) of emulating demographic-based responses.

One final comparison worth making is across the human and LLM groups overall. In Q7, up to 65% of humans said grammar is "very important." Meanwhile, LLMs responded with this same sentiment up to 93.7% of the time (92% when post-stratified). When asked about how much one cares about the Oxford comma in Q3, humans responded with a diverse spread, whereas LLM responses bunched around "Some." It appears that LLM agents structurally value correct grammar but avoid taking strong, opinionated stances.

Note: Post-stratification was performed against this US Census query from the 2024 vintage of the ACS 1-Year Estimates Public Use Microdata Sample.

2 [Proposal] *The "Digital Closet": Detecting the Gap Between Search Behavior and Public Queer Identity across US Regions*

2.1 Motivation and Research Question

In an increasingly polarized political state, the presence of LGBTQIA+ individuals is often tied to left-leaning, or liberal, cities in the United States. Conversely, right-wing regions are perceived to tout "traditional family ideals" inherent to the nuclear family, leaving no space for same-sex relationships or the socialization of queer identities. However, recent "heat map" analyses of Grindr (a dating app that caters to MSM individuals) usage reveal high active user concentrations at notable right-wing gatherings, such as the Republican National Convention and any of the incumbent president's rallies.

This observation, though rudimentary, motivates our study on the extent to which regional geography (and, inherently, each region's views on LGBTQIA+ identities) suppresses the public self-identification of queer individuals compared to their private digital inquiries. By identifying the discrepancy between the frequency of queer-coded search queries and the density of individuals who openly identify as LGBTQIA+ in regionally administered surveys, we can quantify areas where there is a mismatch of queer individuals (identified by their web searches) and the reported proportion of queer individuals based on existing studies. Identifying this delta can help provide resources to areas where individuals may need additional support coming to terms with their queer identity, as well as to potentially evidence the claim that queerness is apolitical, and rather a biological inevitability.

2.2 Methodology

Using the full log of search queries for a popular search engine, we adapt the approach of White et al. (2012), who demonstrated that search logs can reveal hidden signals (drug interactions) that traditional clinical reporting misses. In the context of our study, the hidden signal in question is the aggregate frequency of searches related to queer identity, coming out resources, and same-sex interest.

This signal would be aggregated by region and then normalized against the total search volume to create a density index for digital queer identities. This index could then be compared to existing surveys (e.g., Williams Institute, Gallup). We are interested in regions where there is a large residual between "known" queer volume and "observed" queer volume via the search indices we generate. A high delta would indicate a significant population that is active in the digital queer space but is absent from published demographic records.

2.3 Benefits and Limitations

The primary advantage of using search data over traditional surveys is the reduction of social desirability bias, where individuals in high-stigma environments are less likely to answer survey questions about sexual orientation honestly due to a fear of exposure or internalized stigma. Instead, search queries act as a form of passive observation, capturing intent in a private, low-stakes environment.

Conversely, there are two primary limitations to our methodology. Similar to what White et al. encountered, searches such as "coming out support" could be performed by a queer individual, concerned parents, or educators. Distinguishing between identity-driven versus curiosity-driven

searches becomes complex, though perhaps the need is lessened with more refined and pointed search terms. More importantly, the usage of this data raises numerous ethical and privacy concerns. Even with aggregated data, the ability to detect and quantify a marginalized population feels like a double-edged sword: the wielder must be aware of potential malice through inadvertent use of this information for surveillance or discriminatory policy.

3 Disclaimer on Usage of AI

As an avid user of R, the transition to Python for this assignment was aided by Google Gemini Pro's coding capabilities. Particularly, Gemini was helpful in developing `ggplot2`-esque visualizations and properly formatting print statements for easy interpretation after running each Python script. It was also a huge timesaver in converting the tabular results across the five groups of interest into a properly formatted LaTeX table for this report.