When Does ‘Zionist’ Mean ‘Jew’? Investigating the Extent to which Online Critiques of ‘Zionists’ Invoke Antisemitic Tropes

Alyssa Loo, Abigail Nelkin, and Ariel Stein

Abstract

In discourse surrounding the Israel-Palestine conflict, the line between antisemitic hate speech and legitimate critiques of Israel and its supporters is often fiercely debated. The present study investigates one facet of this debate, empirically examining the extent to which online critiques of “Zionists” invoke antisemitic tropes. We scraped social media posts that use the word “Zionist,” masked the word “Zionist," and had a large language model assess the probability that the masked word is “Jewish” (or a variant). We then investigated whether the model estimated “Jewish” completions to be more likely in the posts about “Zionists” than in a control set of social media posts. We implemented additional controls to ensure that any difference between the datasets is due to the invocation of antisemitic tropes rather than the non-antisemitic use of Jewish-associated words such as “Jerusalem.” Overall, our findings suggest that while most posts critiquing “Zionists” invoke ideas about Jews only indirectly (i.e., through the semantic association between Jews and the Zionist movement), posts that discuss “Zionists” in general terms without mentioning details specific to the Israel-Palestine conflict tend to invoke ideas about Jews through direct appeals to antisemitic tropes.

Background

In discussions of complex sociopolitical issues, the line between ‘punching up’ and ‘punching down’ is often blurry. This is particularly true of contemporary discourse surrounding Israel, particularly in the context of the current Israel-Hamas war. While some may critique Israel as a colonial project, others may view the framing of Israel as a solely colonial enterprise as erasing Jewish indigeneity to the land. While some may frame Israel as the nexus of the fight against colonialism, others may view this centering as disproportionate and reminiscent of traditional antisemitic ideas placing Jews at the heart of the world’s ills. While some may argue that the United States’ military support of Israel ought to result in it being held to a higher standard than other states, others may argue that it is antisemitic to selectively condemn the Jewish state for sins inherent to all nation-states. In discussion of Israel, it is often hotly debated whether a given statement constitutes legitimate criticism or antisemitic speech.

This paper analyzes the term “Zionist” as a case study in these ambiguities. The term “Zionist” has been contemporarily used to critique those who support Israel’s existence or policies; it has also been previously analyzed as a vehicle for the expression of antisemitic tropes implemented by both the political right (Bhat & Klein, 2020) and the political left (Hirsh et al., 2021). However, to the authors’ knowledge no prior study has quantitatively examined the extent to which the term “Zionist” is used in contexts that invoke antisemitic tropes.

The present study asks the following question: To what extent do online critiques of “Zionists” in the period of the 2023–2024 Israel-Hamas conflict invoke antisemitic tropes, and

are certain types of critiques more likely to invoke these tropes than others? In this study we scrape X (formerly Twitter) for posts that critique "Zionists," mask the word "Zionist," and have a large language model estimate the probability that the missing word is "Jewish" or a variant. Over the course of two carefully controlled experiments we build upon this basic paradigm, assessing whether "Zionist" is disproportionately used in contexts that invoke antisemitic tropes.

Study 1: Critiquing "Zionists" Versus "Racists"

We first examined the extent to which the word "Zionist" appears in Jewish-coded contexts. In order to assess whether posts critiquing "Zionists" were disproportionately Jewish-coded, we first had to collect a control data set in order to establish the extent to which posts that do not use the term "Zionist" are Jewish-coded.

We chose posts discussing "racists" in the aftermath of the murder of George Floyd as our control data set. Like "Zionist," "racist" often appears in a negatively valenced social justice context expressing criticism of perceived injustice. However, unlike "Zionist," which has previously been analyzed as a vehicle for the expression of antisemitic tropes, prima facie we would not expect that "racist" carries the same risk of being used in contexts that invoke antisemitic tropes. Therefore, "racist" serves as a promising control to compare a word that may risk being used alongside antisemitic tropes to an analogous word that we would not expect to risk appearing alongside antisemitic tropes (beyond whatever base rate might be expected).

From a pilot search of posts on X, we found that posts containing "racist" from October 7, 2023–January 6, 2024 tended to be directed at public figures, referring to specific behavior in some public event. To find general discourse about "racists" that was not specific to any action or individual, it was useful to search for posts in a period in which there was rich public discourse about racism and racists—analogous to how there is rich public discourse about Zionism and Zionists in the October 7, 2023–January 6, 2024 time period. We therefore identify the time period of the George Floyd protests and the rise of the Black Lives Matter movement in May–August, 2020 as the source for control posts.

Masking the word "racist" or "Zionist," respectively, we measured the probability that our large language model assigned to the completion "Jewish." By comparing the completion probabilities between the data sets, we investigated whether "Zionist" was used disproportionately in Jewish-coded contexts.

Methods

For the "Zionist" data set, we scraped X for posts containing the word "Zionist" or variants ("Zionist," "Zionists," and lowercase forms) from October 7, 2023 through January 6, 2024, the period corresponding to the first three months of the Israel-Hamas war. For the "racist" data set, we scraped X for posts containing the word "racist" or variants ("racist," "racists") from May 25, 2020 through August 24, 2020, the period corresponding to the first three months following murder of George Floyd. As far as we noted, all posts from both data sets were critiques.

After collecting these posts, we performed additional data cleaning. For each dataset, we restricted our analysis to posts between 80 and 300 characters (exclusive) in order to provide both consistency and richness. We further excluded posts that included more than one occurrence of the masked word (e.g., a post about "Zionists" that uses "Zionist" or variants more than once) or that include the target word (i.e., posts that mention "Jewish" or variants). We imposed these restrictions to allow for consistency across the data sets and prevent extraneous cues from biasing completion probabilities. We additionally removed URLs from the posts. Finally, we excluded predicative instances of the masked word; while predicative uses of "racist" are felicitous ("it is racist to do xyz"), predicative uses of "Zionist" are infelicitous (#"it is Zionist to do xyz"), so we excluded all predicative appearances of the masked word in order to keep the data sets as parallel as possible. After this data cleaning, our "Zionist" data set contained 4,055 posts, while our "racist" data set contained 3,208 posts.

We calculated "Jewish" completion probability using BERTweet, an unrestricted language model trained on 850 million English tweets. These training tweets were collected from the January 2012–August 2019 time period; this interval does not overlap with those of either of our data sets, which means that there is no risk of the test data sets being contaminated by the training data set. BERTweet takes in a string of text containing a masked token (i.e., an X post containing a masked word) and produces a probability distribution over all tokens in its vocabulary to fill in that mask (i.e., iteratively goes through its vocabulary and produces an estimate of the probability that each word in its vocabulary is the missing word).

To calculate the completion probability for "Jewish" and variants in each post, denoted *P(Jewish), w*e calculated the linear completion probability for each variant individually ("Jewish," "Jew," "Jews," and lowercase forms), summed them, and took the natural logarithm of this value due to completion probabilities' characteristic skewness. We hypothesized that the mean of P(Jewish) would be greater in the "Zionist" data set than in the "racist" data set, and we investigated this hypothesis using a one-tailed t-test.

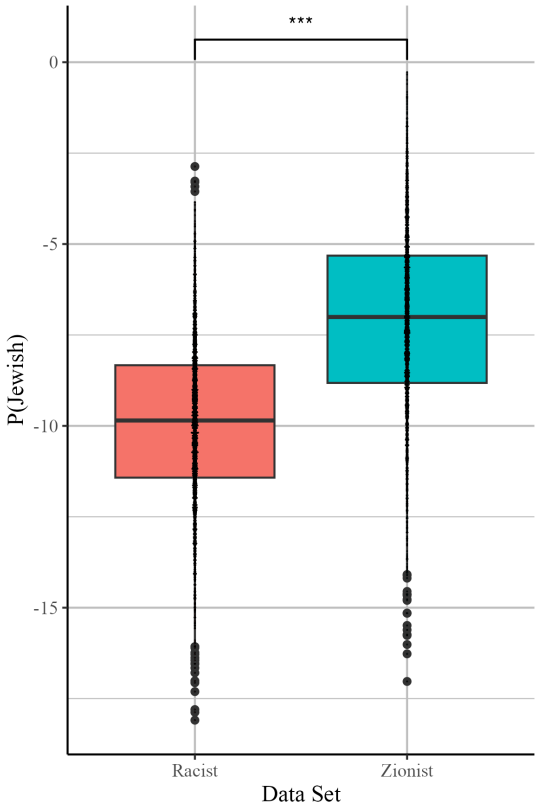


Figure 1. Distribution of P(Jewish) values for the "Zionist" and "racist" data sets. The data set is labeled on the x-axis, while the P(Jewish) values are listed on the y-axis.

Results and Discussion

In line with our hypothesis, we found that the average P(Jewish) was significantly higher in the "Zionist" (μ=-7.11, σ = 2.61) data set than in the "racist" (μ=-9.91, σ=2.28) data set at $p < 0.001$. Figure 1 shows the distribution of "Jewish" or variant completion probabilities for the two data sets.

These findings suggest that "Zionist" is used in Jewish-coded contexts at greater rates than control terms. However, these contexts could be Jewish-coded due to factors other than antisemitic tropes; for example, the word "Jerusalem" is associated with both Jews and the semantics of the Zionist movement, raising the possibility that "Zionist" is used in more Jewish-coded contexts not due to the invocation of antisemitic tropes, but rather because Jews are associated with the semantics of the Zionist movement. Experiment 2 aims to disentangle contexts that are Jewish-coded due to the direct invocation of antisemitic tropes from those that are indirectly Jewish-coded due to the semantic overlap between Jews and the Zionist movement.

Experiment 2: P(Jewish) Versus P(Israeli)

Experiment 2 asks the questions: Are Experiment 1's positive findings actually due to the invocation of antisemitic tropes, or are they simply due to semantic correlation between Jews and Zionism? Furthermore, are some types of critiques more likely to invoke antisemitic tropes than others? In order to answer these questions, we compared P(Jewish) to P(Israeli) in the "Zionist" data set.

Because Israel is intuitively more tightly semantically linked to Zionism than Jews are to Zionism, and because "Israeli" is relatively free of the connotative baggage of "Zionist," we used "Israeli" as a proxy for the semantics of the Zionist movement. If the "Zionist" contexts are only Jewish-coded because the overlapping semantics of Jews and Zionism are driving an indirect association with Jews, we should expect P(Israeli) to be greater than or equal to P(Jewish), since high-probability "Jewish" completions would only be downstream effects of high-probability

"Israeli" completions. If, on the other hand, the "Zionist" contexts are Jewish-coded due to the invocation of antisemitic tropes, we should expect average P(Jewish) to be higher than average P(Israeli), since in this account ideas about Jews (rather than the semantics of the Zionist movement) are causing the contexts to be Jewish-coded.

To investigate whether certain types of critiques are more likely to invoke antisemitic tropes than others, we additionally categorized the "Zionist" data set's posts based on how directly the posts' language was related to the specifics of the Israel-Palestine conflict. We hypothesized that as the posts' language became less grounded in the specifics of the Israel-Palestine conflict, average P(Jewish) would become higher relative to average P(Israeli), corresponding to less specific posts being more likely to invoke antisemitic tropes.

Methods

We used the posts from Experiment 1's "Zionist" data set, categorizing the posts according to four levels of exclusion. Level 0 was the most permissive level, only excluding posts in which the word "Israeli(s)" appears (since it is a new target word) and posts in which the masked word was preceded by an indefinite article. The latter exclusion is because "a" and "an" are sensitive to whether the following word begins with a vowel, which would bias completions towards "Jewish" over "Israeli" because "Jewish," like "Zionist" and unlike "Israeli," begins with a vowel. Level 0 contained a total of 2,756 posts. Level 1 was slightly more stringent, additionally excluding posts mentioning proper nouns specific to Israel and Palestine ("Netanyahu," "Al-Shifa," "Israel," etc.). Level 1 contained a total of 1,495 posts. Level 2 was more stringent; in addition to excluding posts that were excluded from Level 1, we additionally excluded terms referring to concepts or events associated with the Israel-Palestine conflict ("apartheid," "occupation," "open-air prison," etc.). Level 2 contained a total of 995 posts. Level 3 was the most stringent; in addition to excluding posts that were excluded from Level 2, we additionally excluded terms and names associated with Jews and Judaism ("Holocaust," "NYC," "Epstein," etc.). Level 3 contained a total of 834 posts. Each level's exclusion criteria, as well as a short list of examples of terms that would be excluded at that level, were determined by the authors prior to the completion of data collection. After data collection, a single coder extracted the corresponding terms for each exclusion level after examining all 4,055 "Zionist" posts. This was conducted before any analysis was performed.

Because in this experiment we sought to compare the completion probabilities of different words ("Jewish" versus "Israeli"), we needed to control for the base rate at which these words would appear in neutral contexts. This allowed us to ensure that a finding in which P(Jewish) is greater than P(Israeli), for example, is not simply due to "Jewish" being a more common word than "Israeli." In order to establish the base rate at which these words appear in neutrally valence contexts, we created ten neutrally-valenced sentences containing a missing word; in this slot, both "Jewish" variants and "Israeli" variants were viable, with different sentences requiring either the adjective or noun forms. As in this experiment's "Zionist" data set, none of the masks were preceded by an indefinite article. Our neutral sentences included:

I heard about it from one of my *[MASK]* friends.
The author explained that her perspective was shaped by her *[MASK]* upbringing.
I saw a group of *[MASK]* at the park today.
I had some *[MASK]* food yesterday.

In order to normalize our "Jewish" and "Israeli" "Zionist" data set completion probabilities with respect to the rates at which these words appear in neutrally valenced contexts, we divided completion probability of the given term in the "Zionist" data set by completion probability for the same term in the neutrally valenced contexts. P(Jewish) in the "Zionist" data set had already been calculated in Experiment 1, and we calculated P(Israeli) in the "Zionist" data set using the same method.

To calculate the completion probability of "Jewish" in the neutrally valenced contexts, we first measured the linear completion probability of "Jewish" or variants in all ten neutral sentences and summed the linear completion probabilities of all "Jewish" variants within each sentence; we then averaged the sentences' summed linear completion probabilities in order to obtain a measure of the average linear completion probability for "Jewish" or a variant in neutrally valenced sentences. Finally, to obtain our baseline P(Jewish) in neutrally valenced contexts, we took the log of this average. In order to calculate normalized P(Jewish) for a given post in the "Zionist" data set, we subtracted our baseline P(Jewish) from the neutrally valenced contexts from P(Jewish) for the post, since P(Jewish) for both sources are in log-space. We followed the same methodology to calculate our normalized P(Israeli) for posts in the "Zionist" data set. Since there is a P(Israeli) and P(Jewish) statistic for each post, we used paired one-tailed t-tests to assess whether normalized P(Jewish) was greater than normalized P(Israeli) for posts in each level of the "Zionist" data set. We implement this with a one-sample, one-tailed t-test for whether P(Jewish) – P(Israeli) is significantly greater than 0.

Results and Discussion

In the Level 0 posts (i.e., the set of all "Zionist" posts that did not mention the word "Israeli" and in which the masked word was not preceded by an indefinite article), normalized P(Jewish) was not significantly greater than normalized P(Israeli) ($p = 1$); however, an exploratory two-tailed t-test of the Level 0 posts did reveal that P(Israeli) was significantly greater than P(Jewish) in the Level 0 posts ($p < 0.001$). In the Level 1 posts (i.e., the set of posts in which posts mentioning proper nouns specific to Israel and Palestine were additionally excluded), P(Jewish) was significantly greater than P(Israeli) at a $p<0.05$ level. In the Level 2 posts (i.e., the set of posts in which posts mentioning terms and events associated with the Israel-Palestine conflict are additionally excluded) we found that P(Jewish) was significantly greater than P(Israeli) ($p < 0.001$). In the Level 3 posts (i.e., the set of posts in which posts mentioning terms and names associated with Jews and Judaism were additionally excluded), P(Jewish) was also significantly greater than P(Israeli) at $p < 0.001$. Figure 2 shows normalized

P(Israeli) and normalized P(Jewish) in each of the four exclusion levels, while Figure 3 shows how the difference between P(Jewish) and P(Israeli) changes across exclusion conditions.

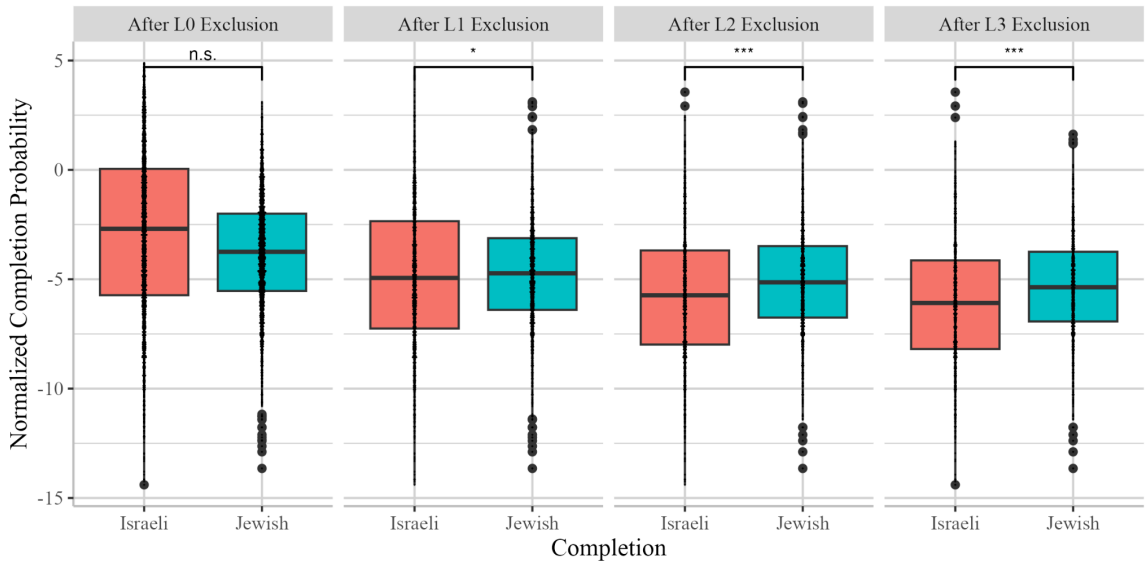


Figure 2. Distribution of normalized P(Israeli) and P(Jewish) values in each of the four exclusion levels. The x-axis shows the completion whose normalized completion probability is being reported ("Israeli" versus "Jewish") as well as the four exclusion levels, with Level 0 being the least stringent (i.e., including both high-specificity and low-specificity posts) and Level 3 being the most stringent (i.e., including only low-specificity posts). The y-axis shows normalized completion probability. Statistical significance is from a one-sample, one-tailed t-test for whether P(Jewish) – P(Israeli) is significantly greater than 0.

As predicted, as exclusion criteria become more stringent (i.e., as the posts became less grounded in the specifics of the Israel-Palestine conflict), P(Jewish) increased relative to P(Israeli), with P(Jewish) being significantly higher than P(Israeli) in Levels 1, 2 and 3. This suggests that, while most posts that criticize "Zionists" tend to be Jewish-coded due to the semantic overlap between Jews and Zionism, posts that are less grounded in the specifics of the Israel-Palestine conflict are more likely to invoke antisemitic tropes. That is, in these posts the use of "Zionist" in Jewish-coded contexts cannot be explained by the semantic overlap between Jews and Zionists, as normalized P(Jewish) is greater than normalized P(Israeli); rather, the contexts seem to be Jewish-coded due to the invocation of antisemitic tropes.

We additionally exploratorily recorded and qualitatively analyzed the four Level 3 posts in which there was the greatest difference between normalized P(Jewish) and P(Israeli) in each direction. Below are the posts in which normalized P(Jewish) was greater than normalized P(Israeli) by the greatest margin (i.e., the posts in which we should expect antisemitic tropes to be the most heavily invoked):

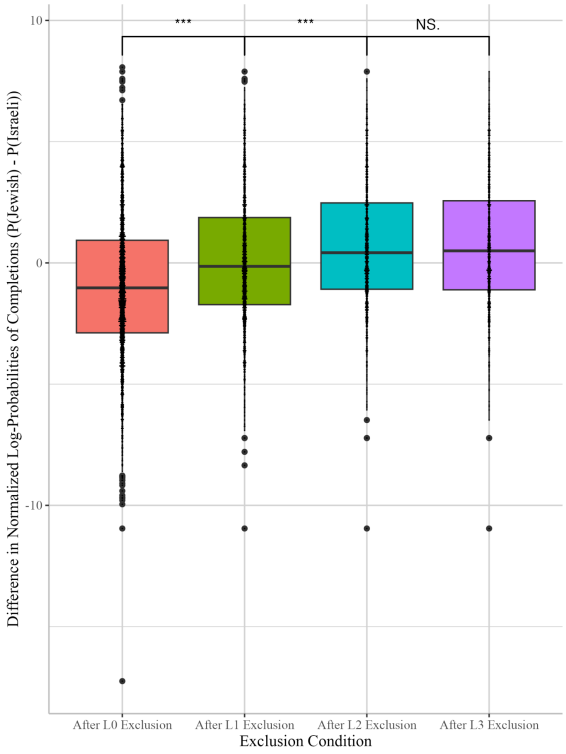


Figure 3. Values of paired differences P(Jewish) - P(Israeli) in each of the four exclusion levels. Statistical significance is from a one-tailed t-test for whether P(Jewish) - P(Israeli) from the less stringent exclusion level is less than the difference for the more stringent exclusion level.

1. Can't wait to pirate 'thanksgiving' since spyglass entertainment is never getting another dime from me and el* r*th is a bloodthirsty *[MASK]*
2. in ten years it's going to be crazy to see how people view *[MASK]* and there will be nothing they can do to refute it bc THEY made this info availabile
3. not just that but big companies industries and corporations are owned by *[MASK]*, such as amazon, walmart, costco, starbucks, mcdonald's, HOLLYWOOD, GOOGLE, the list keeps going on and the support can come from either a fear of being blacklisted or fired to being brainwashed
4. white supremacists, terfs, *[MASK]*, they're all the same. simultaneously some powerful silent majority and a fragile innocent victim being threatened by some barbarian minority

Below are the posts in which normalized P(Israeli) was greater than normalized P(Jewish) by the greatest margin (i.e., the posts in which ideas specific to Israel as opposed to Jews broadly should be most heavily invoked):

1. Second Latin country to take a strong diplomatic action against the *[MASK]* entity is Chille History will never forget those who didnt stay silent in the face of this
2. I wonder if its because some *[MASK]* intelligence agency had videos of the most prominent elected officials and their friends and donors doing bad stuff on an airplane or an island wouldn't that be whacky

3. I'm great, the public are seeing the *[MASK]* rhetoric for what it is. I'd ask if you're ok for a friend except nobody cares
4. LMAOO these are the *[MASK]* dogs that came for the world cup. glad we got the message across and there were no misunderstandings

As might be expected, the posts for which normalized P(Jewish) was greater than normalized P(Israeli) by the greatest margin tended to invoke straightforwardly antisemitic ideas ("Zionists" controlling Hollywood, owning influential companies, being "bloodthirsty," falsely playing the victim, etc.). On the other hand, the posts for which normalized P(Israeli) was greater than normalized P(Jewish) by the greatest extent tended to invoke ideas that might be applied more typically to countries or governmental bodies (ideas about diplomatic action being taken against "the Zionist entity," ideas about "Zionist" intelligence agencies, etc.). That is, in line with our expectations, the posts for which normalized P(Jewish) was greater than normalized P(Israeli) by the greatest margin seemed to tend to criticize "Zionists" qua Jews, while the posts for which normalized P(Israeli) was greater than normalized P(Jewish) by the greatest margin seemed to tend to criticize "Zionists" qua Zionists (i.e., as members of a political movement). This exploratory qualitative analysis suggests that the difference between P(Jewish) and P(Israeli) is measuring what we expect it to measure—that is, it is measuring whether "Zionist" is used in a context that invokes ideas about Jews directly, or one that only invokes ideas about Jews indirectly via the invocation of ideas about the Zionist movement broadly.

We performed additional exploratory analyses to investigate whether normalized P(Jewish) and P(Israeli) in the "Zionist" data set pattern with the completion probability of other identity labels ("Muslim," "Black," "gay," and "American"). From the L0 Exclusion Level from Experiment 2, we further exclude posts in the dataset that have any mentions of any of these identity labels, leaving a dataset of 2,597 posts. We collect log probabilities in a similar method, and normalize them using the ten neutrally valenced sample sentences used for "Jewish" and "Israeli." See Figure 4 for the breakdown of normalized completion probability for each identity label.

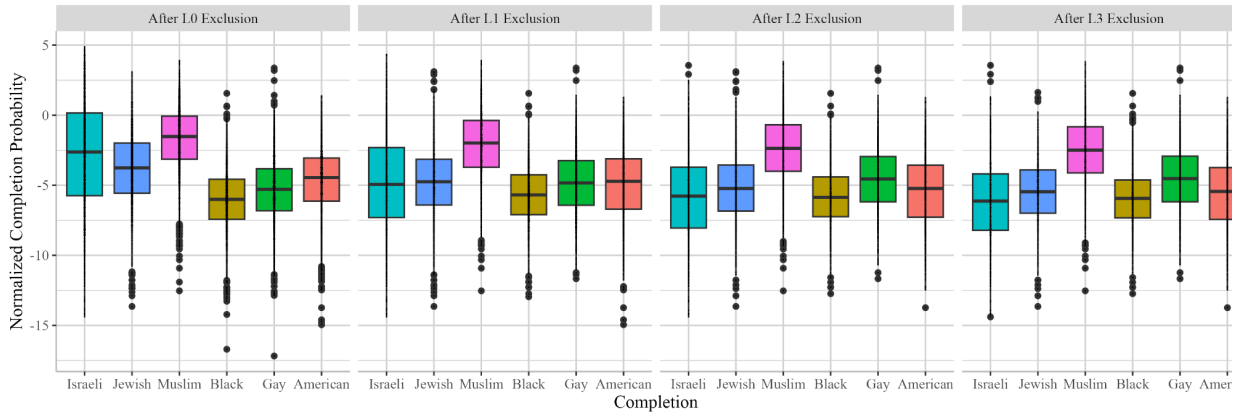


Figure 4. Normalized completion probability for each identity label, separated by exclusion level. On the x-axis are the identity labels whose normalized completion probabilities were calculated. On the y-axis are the normalized completion probabilities. Each graph represents a different level of exclusion, with Level 0 corresponding to the least stringent exclusion level (i.e., the level at which the posts with the most specific references to the details of the Israel-Palestine conflict are included) and Level 3 corresponding to the most stringent exclusion level (i.e., the level at which only posts that do not explicitly mention the specific details of the Israel-Palestine conflict are included).

We found that, in all four exclusion levels, the normalized completion probability for the other identity labels tended to be as high as or higher than normalized P(Jewish), raising the possibility that people use the term "Zionist" in sentence frames typical of hate speech or other negative identity-based speech, rather than using it specifically in antisemitic sentence frames. That is, the contexts surrounding "Zionist" may be identity-coded rather than specifically Jewish-coded.

However, these exploratory findings should be interpreted cautiously. The other identity terms' completion probabilities were normalized using the ten neutrally valenced sentence frames that we had crafted to suit both "Israeli" and "Jewish" completions; however, these sentences are not equally suitable for all of the identity labels. For example, one sentence refers to "*[MASK]* food," but references to "gay food" or "Muslim food" may be less naturalistic than references to "Jewish food" or "Israeli food." Likewise, one of the neutral sentences refers to "a group of *[MASK]*," but there is no common, neutrally-valenced, single-word English noun referring to Black people. These limitations mean that the completion probabilities of some of the identity words in the neutrally valenced data set were likely underestimated relative to their true completion probability in neutrally valenced language at large; underestimating these identity labels' completion probability in neutrally valenced language would artificially inflate their normalized completion probability in the "Zionist" data set, which may be driving the high normalized completion probabilities for the other identity terms. Our exploratory analysis comparing normalized completion probabilities for other identity terms raises the possibility that "Zionist" may not be used in specifically Jewish-coded contexts, but rather in contexts that refer to social identities more broadly. Under this analysis, "Zionist" invokes not only antisemitic tropes, but also tropes about other minority groups. Overall, our exploratory findings regarding other identity labels raise the possibility that the use of "Zionist" may occur in broadly identity-coded contexts rather than specifically Jewish-coded contexts, presenting an interesting direction for further study with more versatile neutral sentences. Due to our neutral sentences' lack of versatility, we interpret these exploratory findings investigating other identity labels cautiously.

General Discussion and Conclusion

Our findings suggest that contemporary social media posts critiquing “Zionists” without grounding the critique in the specifics of the Israel-Palestine conflict tend to disproportionately invoke antisemitic tropes. In Experiment 1 we found that critiques of “Zionists” tended to be more Jewish-coded than critiques of “racists;” in Experiment 2 we found that, in the set of all posts mentioning “Zionists,” this Jewish-codedness seemed to result from semantic overlap between Jews and Zionism. However, in the posts that did not mention specific details of the Israel-Palestine conflict we found that the Jewish-codedness could not be attributed to innocuous semantic overlap between Jews and Zionism; rather, it seemed to be due to the more direct invocation of antisemitic tropes.

References

Bhat, P., & Klein, O. (2020). Covert hate speech: White nationalists and dog whistle communication on twitter. *Twitter, the public sphere, and the chaos of online deliberation*, 151-172.

Hirsh, D. (2021). How the Word “Zionist” Functions in Antisemitic Vocabulary. *Journal of contemporary antisemitism*, 4(2), 1-18.

Questions and Answers

- Q: I think it would be really cool to see if you could associate the OTHER words with either "zionist" or "jew", and see if there is a trend there. So, count the probability of every word that co-occurs, and somehow vectorize that or something in order to find what words share the closest meaning. Now that I think about it, this is basically what word2vec is. You should check out word2vec with this experiment!
    - A: We (or rather you, Uriel) did this while workshopping this study. Here is the relevant code:

      ```
      for (term, sim) in wvt.most_similar(positive=['zionist'], negative=[], topn=5):
        print(f"Term {term}, similarity {sim} (relative to 'zionist')")
      ```

      Term zionists, similarity 0.7022473812103271 (relative to 'zionist')
      Term israeli, similarity 0.647412896156311 (relative to 'zionist')
      Term fascist, similarity 0.6404380798339844 (relative to 'zionist')
      Term jewish, similarity 0.5950586199760437 (relative to 'zionist')
      Term jews, similarity 0.5923017263412476 (relative to 'zionist')
- Q: I'm not entirely clear what the conclusion is, as Zionism is a Jewish movement and is associated with Jews. To me, the baseline of "Israeli" feels a little bit forced in comparison. It's definitely a difficult pattern to pin down.
    - A: The intuition here is that Zionism is far more semantically tied to Israel (the culmination of the Zionist movement and the center of the contemporary controversy surrounding Zionism) than to Jews. Though both are undoubtedly

related to Zionism, the intuition is that if you asked the average person what word they think of when they think of "Zionism," they would likely say "Israel" or a variant. This is supported by the word2vec results described in the previous answer, in which "Israeli" is more similar to "Zionist" than "Jewish," second only to the plural form "Zionists."

- Q: It seems to me that some of the words from the L2 exclusion are also very related to other hate crimes (raped, doxx, etc) and so I wonder how this might have impacted your exploratory analysis at the end. I wonder, too, how demographics of posters might change these results as well (eg would Brown students' Sidechat posts show a different result from, say, Facebook posts?)
    - A: We would expect that removing posts with those hate-speech related words would uniformly decrease the probability of all identity group terms (including 'Jewish') as completions; since the hypothesis is whether posts are Jewish-coded in particular, their uniform effect would not be informative on the hypothesis. Most words are also only used in reference to specific events in the Israel-Palestine conflict.

      Demographics of posters would probably change the results. Twitter had a large emphasis on celebrities' stances on the conflict, but something more skewed toward college age would probably have greater focus on protesting or school administrators' stances. It would be an interesting question to explore the degree of toxicity on different public online forums.
- Q: How did you determine which words would be included in the exclusion levels for experiment 2? Were the lists made manually or sourced from somewhere? / How did you generate the lists of words for the exclusion levels in experiment 2? / I wonder how you came up with your words for exclusion?
    - A: The lists were made manually. Each level's exclusion criteria, as well as a short list of examples of terms that would be excluded at that level, were determined by the authors prior to data collection. After data collection, exact terms from the posts as identified by the exclusion guidelines were extracted by a single coder.
- Q: I'm wondering if these results would be different when coming from different sources, such as TikTok, Instagram, etc.? Especially since Twitter has become such a unique platform recently in regards to both its users and its form of moderation.
    - A: Our intuition is that platforms with less moderation would skew towards the direct invocation of antisemitic tropes relative to platforms with more moderation. This is based on the idea that on platforms with greater moderation posts that invoke antisemitic tropes in straightforward ways would be more likely to be removed through the moderation process; this removal would result in fewer

messages that directly invoke antisemitic tropes and disincentivize people from posting content that invokes antisemitic tropes.

- Q: I'm wondering if and how might the exclusion of specific referents to diplomatic bodies and organizations alter the results?
    - Removing diplomatic bodies could remove an international focus to more of a domestic focus.
- Q: When you were collecting your dataset, did you consider excluding tweets that might use 'Zionist' in a positive context (as compared to the 'racist' dataset, in which I wouldn't think there would be any tokens where racist was used positively)? Or would this exclusion not matter for the final conclusion?
    - A: As far as we noted, none of the posts used “Zionist” in a positive context.
- Q: Did you see if your methods hold for periods/corpus data taken outside of the 3-month windows you'd selected for both "Zionist" and "racist?" (as in, periods where these terms are not being used at a greater volume due to current events)? If so, did you see any trends there, and if not, do you have any predictions as to what you'd find?
    - A: We didn’t look at other periods, but we would tentatively hypothesize that the use of “Zionist” in other periods would tend to skew away from antisemitic use relative to the use in the period of the Israel-Hamas war. Because we found that posts mentioning “racists” from outside of the period following the murder of George Floyd tended to focus on specific events, we would tentatively expect posts mentioning “Zionists” from periods outside of the Israel-Hamas war to follow a similar pattern. Our data suggest that more specific posts are less likely to directly invoke antisemitic tropes, so we would tentatively hypothesize that posts mentioning “Zionists” from another time period would be less likely to invoke antisemitic tropes.
- Q: I would be interested in learning more about whether geographic info about the tweeter influences anything-- obviously the tweets are all English but does country of origin of the tweets influence the probabilities of different minority words
    - A: This would be an interesting avenue for future research. Our suspicion is that the country of origin of the poster would affect the probabilities of different minority words. For example, in posts critiquing “Zionists” coming from a country in which discourse criticizing Israel is often intermixed with straightforwardly antisemitic rhetoric we might expect to find a higher P(Jewish) than we would in posts from a region in which there is more division between critiques of Israel and traditional antisemitism.
- Q: One question that came to mind for me is whether there is any statistical difference between left- and right-wing tweets/users. This would probably be difficult to measure (since the line is blurry when it comes to this type of speech), but it could be an interesting direction to take this in.

- A: This is a very interesting question. We didn't distinguish in the data set, but one intuition that we have is that right-wing posts critiquing "Zionists" while invoking antisemitic tropes may be more likely to frame Jews as people of color hurting or oppressing white people; these posts might characterize Jews using Western tropes traditionally used against people of color (e.g., tropes about barbarity). On the other hand, we might expect that left-wing posts critiquing "Zionists" while invoking antisemitic tropes would be more likely to frame Jews as white people hurting or oppressing people of color; these tropes might characterize Jews using Western tropes associated with white people (e.g., tropes about wealth or privilege).

FinalProject

Code for CLPS1360 Final Project

When Does 'Zionist' Mean 'Jew'? by Alyssa Loo, Abigail Nelkin & Ariel Stein

```
setwd('/Users/alyssamarie/Desktop/school/clps1360_corpus_linguistics/corpus-final-project
library(reticulate)
```

```
Warning: package 'reticulate' was built under R version 4.2.3
```

```
library(ggsignif)
library(tidyverse)
```

```
── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
✔ ggplot2 3.4.1     ✔ purrr   1.0.1
✔ tibble  3.1.8     ✔ dplyr   1.1.0
✔ tidyr   1.3.0     ✔ stringr 1.5.0
✔ readr   2.1.4     ✔ forcats 1.0.0
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
```

```
library(psych)
```

```
Warning: package 'psych' was built under R version 4.2.3


Attaching package: 'psych'

The following objects are masked from 'package:ggplot2':

    %+%, alpha
```

```
use_python("~/anaconda3/envs/interp/bin/python")
source_python('masked_completions.py')
```

Experiment 1

'Racist' dataset

Data Cleaning + Preprocessing for Masked Language Model

```
data.e1.racist <- read_csv('./data/racist_compiled.csv')
```

```
New names:
Rows: 6690 Columns: 5
── Column specification
──────────────────────────────────────────────────────── Delimiter: "," chr
(3): tweet, word, source_file dbl (1): ...1 date (1): date
ℹ Use `spec()` to retrieve the full column specification for this data. ℹ
Specify the column types or set `show_col_types = FALSE` to quiet this message.
• `` -> `...1`
```

```
exclude.pred.racist <- c("( be|being| is|isn't| ur|i'm|was|wasn't|been) racist",
                 "(are|is|was|is it|seems) racist (to|for|)", "(it's|is it|if you're|be|
                 " (he|she|it)'s(| not) racist",
                 " (you|we|they)'re(| not) racist",
                 "(he|she|it)s(| not) racist",
                 "(you|we|they)re(| not) racist",
                 "(you're| be| being| is|isn't| ur|i'm|was|wasn't|were|we're|youre|theyr
                 "were notoriously aggressive & racist")

data.e1.racist <- data.e1.racist %>%
    mutate(tweet = gsub("@|http[s://]+.+$|http[s://]+.+\\s", "", pull(., tweet))) %>%
    filter((nchar(data.e1.racist$tweet) > 80) &
           (str_count(data.e1.racist$tweet, '[rR]acists?') == 1) & # remove tweets that h
           (str_count(data.e1.racist$tweet, '[rR]acism') == 0) & # remove tweets that men
           (str_count(data.e1.racist$tweet, '[jJ]ew') == 0) & # remove tweets that have s
           (nchar(data.e1.racist$tweet) < 300) & # remove tweets that are too long
           !grepl(paste(exclude.pred.racist, collapse = "|"), tolower(data.e1.racist$twee
    mutate(masked_tweet = gsub('[rR]acists?', "<mask>", pull(., tweet))) # mask tokens

data.e1.racist <- data.e1.racist %>%
    filter(str_count(data.e1.racist$masked_tweet, '<mask>') == 1) # makes sure there is a

glimpse(data.e1.racist)
```

```
Rows: 3,208
Columns: 6
$ ...1         <dbl> 0, 1, 3, 4, 6, 7, 9, 10, 12, 13, 14, 20, 22, 25, 27, 28, …
$ date         <date> 2020-05-25, 2020-05-25, 2020-05-25, 2020-05-25, 2020-05-…
$ tweet        <chr> "Amy Cooper said \"African American\" because racist whit…
$ word         <chr> "racist", "racist", "racist", "racist", "racist", "racist…
$ source_file  <chr> "scraped_tweets_racist_46474.json", "scraped_tweets_racis…
$ masked_tweet <chr> "Amy Cooper said \"African American\" because <mask> whit…
```

Getting MLM Completions

```
## --- Commented out so is not run during run-through to generate HTML
# outputs <- get_completions(data.e1.racist$masked_tweet, list(jewish=c("Jewish", "Jew",
```

```
# data.e1.racist$logp.jewish <- log(outputs[[2]]$jewish)
# write.csv(data.e1.racist, './processed_data_cache/exp1_racist_mlm.csv')
```

'Zionist' dataset

Data Cleaning + Preprocessing for Masked Language Model

```
data.e1.zionist <- read_csv('./data/zionist_compiled.csv')
```

```
New names:
Rows: 6766 Columns: 5
── Column specification
──────────────────────────────────────────────────────── Delimiter: "," chr
(3): tweet, word, source_file dbl (1): ...1 date (1): date
ℹ Use `spec()` to retrieve the full column specification for this data. ℹ
Specify the column types or set `show_col_types = FALSE` to quiet this message.
• `` -> `...1`
```

```
exclude.pred.zionist <- c("( be|being| is|isn't| ur|i'm|was|wasn't|been) zionist",
                  "(are|is|was|is it|seems) zionist (to|for|)", "(it's|is it|if you're|be
                  " (he|she|it)'s(| not) zionist",
                  " (you|we|they)'re(| not) zionist",
                  "(he|she|it)s(| not) zionist",
                  "(you|we|they)re(| not) zionist",
                  "(you're| be| being| is|isn't| ur|i'm|was|wasn't|were|we're|youre|theyr
                  "were notoriously aggressive & zionist")

data.e1.zionist <- data.e1.zionist %>%
    mutate(tweet = gsub("@|http[s://]+.+$|http[s://]+.+\\s", "", pull(., tweet))) %>%
    filter((nchar(data.e1.zionist$tweet) > 80) &
           (str_count(data.e1.zionist$tweet, '[zZ]ionists?') == 1) & # remove tweets that
           (str_count(data.e1.zionist$tweet, '[zZ]ionism') == 0) & # remove tweets that m
           (str_count(data.e1.zionist$tweet, '[jJ]ew') == 0) & # remove tweets that have
           (nchar(data.e1.zionist$tweet) < 300) & # remove tweets that are too long
           !grepl(paste(exclude.pred.zionist, collapse = "|"), tolower(data.e1.zionist$tw
    mutate(masked_tweet = gsub('[zZ]ionists?', "<mask>", pull(., tweet))) # mask tokens

data.e1.zionist <- data.e1.zionist %>%
    filter(str_count(data.e1.zionist$masked_tweet, '<mask>') == 1) # makes sure there is

glimpse(data.e1.zionist)
```

```
Rows: 3,022
Columns: 6
$ ...1         <dbl> 0, 2, 9, 11, 12, 14, 16, 17, 19, 22, 23, 24, 26, 27, 28, …
$ date         <date> 2023-10-07, 2023-10-07, 2023-10-07, 2023-10-07, 2023-10-…
```

```
$ tweet        <chr> "I genuinely cannot comprehend how someone can be a zioni…
$ word         <chr> "zionist", "zionist", "zionist", "zionist", "zionist", "z…
$ source_file  <chr> "scraped_tweets_zionist_43618.json", "scraped_tweets_zion…
$ masked_tweet <chr> "I genuinely cannot comprehend how someone can be a <mask…
```

```
write.csv(data.e1.zionist, './processed_data_cache/exp1_zionist_cleaned.csv')
```

Getting MLM Completions

```
## --- Commented out so is not run during run-through to generate HTML
# data.e1.zionist <- read_csv('./processed_data_cache/exp1_zionist_cleaned.csv')
# outputs <- get_completions(data.e1.zionist$masked_tweet, list(jewish=c("Jewish", "Jew",
# data.e1.zionist$logp.jewish <- log(outputs[[2]]$jewish)
# write.csv(data.e1.zionist, './processed_data_cache/exp1_zionist_mlm.csv')
```

Analysis

This produces Figure 1 in the report.

```
data.e1.zionist <- read_csv('./processed_data_cache/exp1_zionist_mlm.csv')
```

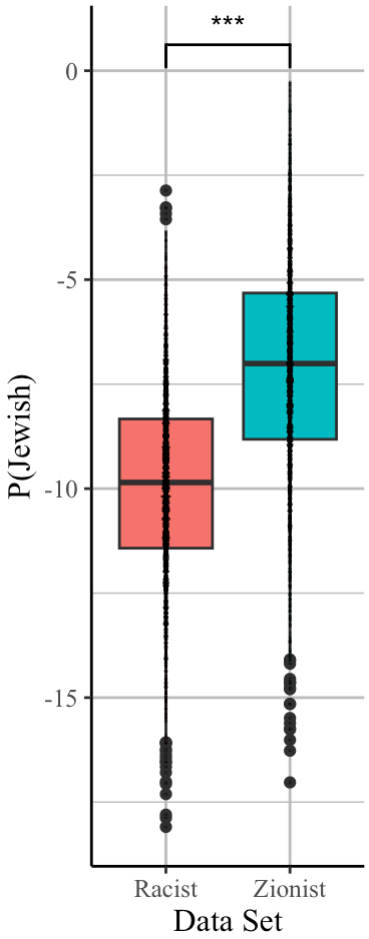
```
New names:
Rows: 2970 Columns: 8
── Column specification
──────────────────────────────────────────────────────── Delimiter: "," chr
(4): tweet, word, source_file, masked_tweet dbl (3): ...1, ...2, logp.jewish
date (1): date
ℹ Use `spec()` to retrieve the full column specification for this data. ℹ
Specify the column types or set `show_col_types = FALSE` to quiet this message.
• `` -> `...1`
• `...1` -> `...2`
```

```
data.e1.racist <- read_csv('./processed_data_cache/exp1_racist_mlm.csv')
```

```
New names:
Rows: 3208 Columns: 8
── Column specification
──────────────────────────────────────────────────────── Delimiter: "," chr
(4): tweet, word, source_file, masked_tweet dbl (3): ...1, ...2, logp.jewish
date (1): date
ℹ Use `spec()` to retrieve the full column specification for this data. ℹ
Specify the column types or set `show_col_types = FALSE` to quiet this message.
• `` -> `...1`
• `...1` -> `...2`
```

```
data.e1 <- rbind(data.e1.zionist, data.e1.racist)

ggplot(data.e1, aes(x = word, y = logp.jewish, fill = word)) +
  geom_boxplot() +
  geom_signif(comparisons = list(c("racist", "zionist")),
              map_signif_level=TRUE,  test="t.test", test.args=list(alternative = "less",
   geom_dotplot(binaxis = 'y', binwidth=0.012, stackdir = "center") +
  labs(x = "Data Set", y = "P(Jewish)") +
  scale_x_discrete(labels = c("Racist", "Zionist")) +
  theme(legend.position="none", text=element_text(family="Times New Roman", size=12),
   panel.background = element_rect(fill = "white"),
        panel.grid = element_line(color = "gray"),
        axis.line = element_line(colour = "black"))
```




```
ggsave("./graphs/exp1.png", plot=last_plot(), width = 10, height = 15, units = "cm")
```

Experiment 2

Cleaning data (on top of previous cleaning done on the 'Zionist' dataset)

```
data.e2 <- read_csv('./processed_data_cache/exp1_zionist_mlm.csv')
```

```
New names:
Rows: 2970 Columns: 8
── Column specification
──────────────────────────────────────────────────────── Delimiter: "," chr
(4): tweet, word, source_file, masked_tweet dbl (3): ...1, ...2, logp.jewish
date (1): date
ℹ Use `spec()` to retrieve the full column specification for this data. ℹ
Specify the column types or set `show_col_types = FALSE` to quiet this message.
• `` -> `...1`
• `...1` -> `...2`
```

```
data.e2 <- data.e2 %>%
    filter((str_count(data.e2$tweet, '[iI]sraelis?') == 0) & # remove tweets that mention
           (str_count(data.e2$tweet, 'an? <mask>') == 0))# remove tweets that have a part

glimpse(data.e2)
```

```
Rows: 2,756
Columns: 8
$ ...1         <dbl> 1, 2, 3, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15, 16, 17, 18, 2…
$ ...2         <dbl> 0, 3, 5, 8, 10, 11, 12, 14, 18, 20, 21, 22, 23, 24, 25, 2…
$ date         <date> 2023-10-08, 2023-10-08, 2023-10-08, 2023-10-08, 2023-10-…
$ tweet        <chr> "Brilliant interview debunking the BBC's sucking up to Zi…
$ word         <chr> "zionist", "zionist", "zionist", "zionist", "zionist", "z…
$ source_file  <chr> "scraped_tweets_zionist_45092.json", "scraped_tweets_zion…
$ masked_tweet <chr> "Brilliant interview debunking the BBC's sucking up to <m…
$ logp.jewish  <dbl> -5.711659, -5.826700, -5.356212, -3.806168, -5.274502, -4…
```

```
write.csv(data.e2, './processed_data_cache/exp2_cleaned.csv')
```

Getting MLM Completion for 'Israeli'

```
## --- Commented out so is not run during run-through to generate HTML
# data.e2 <- read_csv('./processed_data_cache/exp2_cleaned.csv')
# outputs <- get_completions(data.e2$masked_tweet, list(israeli=c("Israeli", "Israelis"))
# data.e2$logp.israeli <- log(outputs[[2]]$israeli)
# write.csv(data.e2, './processed_data_cache/exp2_mlm.csv')
```

Calculating neutral sentence norms

```
neutral.sentences <- readLines("./data/neutral_sentences.txt")
outputs <- get_completions(neutral.sentences, list(jewish=c("Jewish", "Jew", "jew", "jews
norm.israeli <- log(sum(outputs[[2]]$israeli))
norm.jewish <- log(sum(outputs[[2]]$jewish))
paste("Israeli neutral context log-prob:", norm.israeli, "| Jewish neutral context log-pr
```

```
[1] "Israeli neutral context log-prob: -4.90726644811154 | Jewish neutral context log-
prob: -3.37916140189038"
```

```
data.e2 <- read_csv('./processed_data_cache/exp2_mlm.csv')
```

```
New names:
Rows: 2756 Columns: 11
── Column specification
──────────────────────────────────────────────────────── Delimiter: "," chr
(4): tweet, word, source_file, masked_tweet dbl (6): ...1, ...2, ...3, ...4,
logp.jewish, logp.israeli date (1): date
ℹ Use `spec()` to retrieve the full column specification for this data. ℹ
Specify the column types or set `show_col_types = FALSE` to quiet this message.
• `` -> `...1`
• `...1` -> `...2`
• `...2` -> `...3`
• `...3` -> `...4`
```

```
data.e2 <- data.e2 %>%
  mutate(norm.logp.israeli = pull(., logp.israeli) - norm.israeli,
         norm.logp.jewish = pull(.,logp.jewish) - norm.jewish)

data.e2 <- data.e2 %>%
  mutate(diff = pull(., norm.logp.jewish) - pull(., norm.logp.israeli))

glimpse(data.e2)
```

```
Rows: 2,756
Columns: 14
$ ...1             <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 1…
$ ...2             <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 1…
$ ...3             <dbl> 1, 2, 3, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15, 16, 17, …
$ ...4             <dbl> 0, 3, 5, 8, 10, 11, 12, 14, 18, 20, 21, 22, 23, 24, …
$ date             <date> 2023-10-08, 2023-10-08, 2023-10-08, 2023-10-08, 202…
$ tweet            <chr> "Brilliant interview debunking the BBC’s sucking up …
$ word             <chr> "zionist", "zionist", "zionist", "zionist", "zionist…
$ source_file      <chr> "scraped_tweets_zionist_45092.json", "scraped_tweets…
$ masked_tweet     <chr> "Brilliant interview debunking the BBC’s sucking up …
$ logp.jewish      <dbl> -5.711659, -5.826700, -5.356212, -3.806168, -5.27450…
```

```
$ logp.israeli      <dbl> -1.7931436, -7.8746087, -0.5667414, -4.8278446, -10.…
$ norm.logp.israeli <dbl> 3.1141228, -2.9673423, 4.3405250, 0.0794218, -5.8000…
$ norm.logp.jewish  <dbl> -2.332498081, -2.447538665, -1.977050581, -0.4270061…
$ diff              <dbl> -5.44662092, 0.51980361, -6.31757560, -0.50642795, 3…
```

```
write.csv(data.e2, './processed_data_cache/exp2_mlm_normed.csv')
```

Annotating with Exclusion levels

```
data.e2 <- read_csv('./processed_data_cache/exp2_mlm_normed.csv')
```

```
New names:
Rows: 2756 Columns: 15
── Column specification
──────────────────────────────────────────────────────── Delimiter: "," chr
(4): tweet, word, source_file, masked_tweet dbl (10): ...1, ...2, ...3, ...4,
...5, logp.jewish, logp.israeli, norm.log... date (1): date
ℹ Use `spec()` to retrieve the full column specification for this data. ℹ
Specify the column types or set `show_col_types = FALSE` to quiet this message.
• `` -> `...1`
• `...1` -> `...2`
• `...2` -> `...3`
• `...3` -> `...4`
• `...4` -> `...5`
```

```
l1.exclusion <- c("israel", "palestin", "netanyahu", "isreal", "al-aqsa","westbank", "wes

l2.exclusion <- c("iron dome", "coloniz", "apartheid", "occupation", "geno","open-air pri

l3.exclusion <- c("aipac", "blood libel", "holocaust", "antisem", "anti-sem","anti sem",


data.e2 <- data.e2 %>%
  mutate(included.l1 = !grepl(paste(l1.exclusion, collapse = "|"), tolower(data.e2$tweet)
         included.l2 = !grepl(paste(c(l1.exclusion, l2.exclusion), collapse = "|"), tolow
         included.l3 = !grepl(paste(c(l1.exclusion, l2.exclusion, l3.exclusion), collapse
glimpse(data.e2)
```

```
Rows: 2,756
Columns: 18
$ ...1              <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 1…
$ ...2              <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 1…
$ ...3              <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 1…
$ ...4              <dbl> 1, 2, 3, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15, 16, 17, …
$ ...5              <dbl> 0, 3, 5, 8, 10, 11, 12, 14, 18, 20, 21, 22, 23, 24, …
$ date              <date> 2023-10-08, 2023-10-08, 2023-10-08, 2023-10-08, 202…
$ tweet             <chr> "Brilliant interview debunking the BBC's sucking up …
```

```
$ word              <chr> "zionist", "zionist", "zionist", "zionist", "zionist…
$ source_file       <chr> "scraped_tweets_zionist_45092.json", "scraped_tweets…
$ masked_tweet      <chr> "Brilliant interview debunking the BBC's sucking up …
$ logp.jewish       <dbl> -5.711659, -5.826700, -5.356212, -3.806168, -5.27450…
$ logp.israeli      <dbl> -1.7931436, -7.8746087, -0.5667414, -4.8278446, -10.…
$ norm.logp.israeli <dbl> 3.1141228, -2.9673423, 4.3405250, 0.0794218, -5.8000…
$ norm.logp.jewish  <dbl> -2.332498081, -2.447538665, -1.977050581, -0.4270061…
$ diff              <dbl> -5.44662092, 0.51980361, -6.31757560, -0.50642795, 3…
$ included.l1       <lgl> FALSE, FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, FALS…
$ included.l2       <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FAL…
$ included.l3       <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FAL…
```

```
write.csv(data.e2, './processed_data_cache/exp2_for_analysis.csv')
```

Analysis

Significance Tests

These are two calculate the significance results for graphs.

```
l0.pairedt <- t.test(data.e2$diff, mu = 0, alternative = "greater")
l1.pairedt <- t.test(filter(data.e2, data.e2$included.l1)$diff, mu = 0, alternative = "gr
l2.pairedt <- t.test(filter(data.e2, data.e2$included.l2)$diff, mu = 0, alternative = "gr
l3.pairedt <- t.test(filter(data.e2, data.e2$included.l3)$diff, mu = 0, alternative = "gr

paste("L0 Paired T-Test p=", l0.pairedt$p.value, "| L1 Paired T-Test p=", l1.pairedt$p.va
```

```
[1] "L0 Paired T-Test p= 1 | L1 Paired T-Test p= 0.0340996962282442 | L2 Paired T-Test p=
3.97977009486668e-15 | L3 Paired T-Test p= 3.01482748288977e-14"
```

```
l0.twosided <- t.test(data.e2$diff, mu = 0, alternative = "two.sided")
paste("L0 Paired 2-Tailed T-Test p=", l0.twosided$p.value)
```

```
[1] "L0 Paired 2-Tailed T-Test p= 1.4210007577425e-54"
```

Graphs

```
data.e2 <- data.e2 %>%
  mutate(condition='After L0 Exclusion')
data.e2.l1 <- filter(data.e2, data.e2$included.l1) %>%
  mutate(condition="After L1 Exclusion")
data.e2.l2 <- filter(data.e2, data.e2$included.l2) %>%
  mutate(condition="After L2 Exclusion")
data.e2.l3 <- filter(data.e2, data.e2$included.l3) %>%
  mutate(condition="After L3 Exclusion")
```

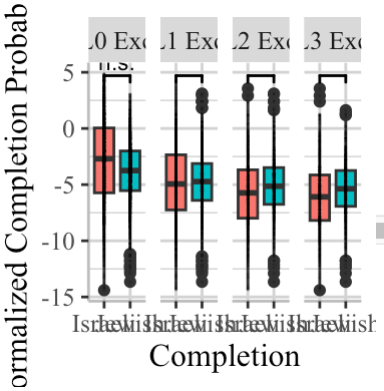
```
data.e2.condsplit <- rbind(data.e2, data.e2.l1, data.e2.l2, data.e2.l3)

data.e2.lengthened <- data.frame(
  norm.logp = c(data.e2.condsplit$norm.logp.jewish, data.e2.condsplit$norm.logp.israeli),
  condition = data.e2.condsplit$condition,
  completion = c(rep("Jewish", length(data.e2.condsplit$norm.logp.jewish)), rep("Israeli"
  mutate(condition_f = factor(pull(., condition), levels=c('After L0 Exclusion','After L1


annotation_df <- data.frame(
  condition = c("After L0 Exclusion", "After L1 Exclusion", "After L2 Exclusion", "After
  condition_f = levels(data.e2.lengthened$condition_f),
  start = c("Israeli", "Israeli", "Israeli", "Israeli"),
  end = c("Jewish", "Jewish", "Jewish", "Jewish"),
  y = c(4.7, 4.7, 4.7, 4.7),
  label = c("n.s.", "*", "***", "***")) # based on manual t-tests from previous cell

ggplot(data.e2.lengthened, aes(x = completion, y = norm.logp)) +
  geom_boxplot(aes(fill=completion)) +
  geom_dotplot(binaxis = 'y', binwidth=0.015, stackdir = "center") +
  facet_wrap(~condition_f, nrow=1)  +
  labs(x = "Completion", y = "Normalized Completion Probability") +
  geom_signif(data = annotation_df,
    aes(xmin = start, xmax = end, annotations = label, y_position = y),
    textsize = 3, vjust = -0.2, manual = TRUE) +
    theme(legend.position="none", text=element_text(family="Times New Roman", size=12),
        panel.background = element_rect(fill = "white"),
        panel.grid = element_line(color = "lightgray"),
        axis.line = element_line(colour = "gray"))
```

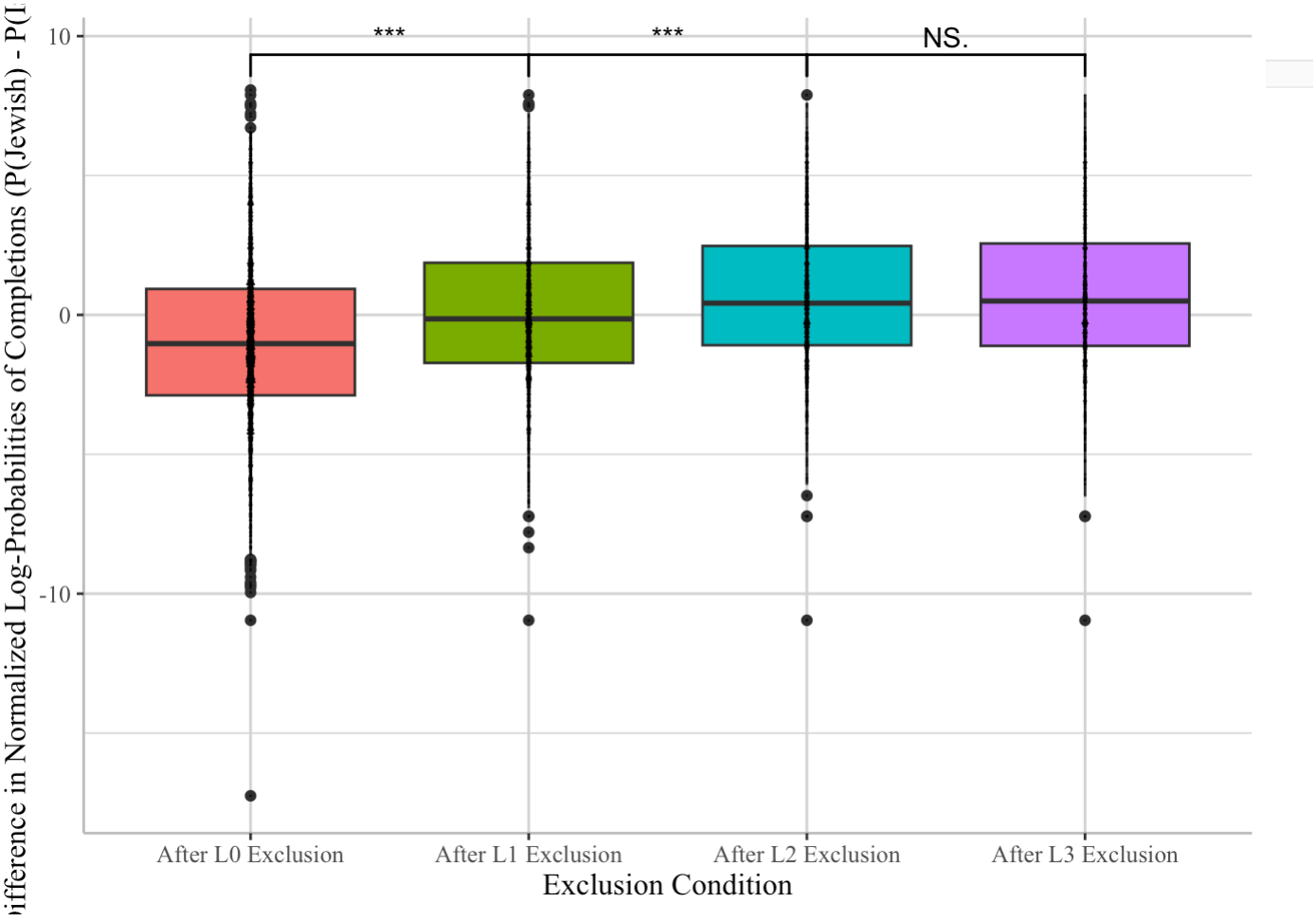
```
Warning in geom_signif(data = annotation_df, aes(xmin = start, xmax = end, :
Ignoring unknown aesthetics: xmin, xmax, annotations, and y_position
```




```
ggsave("./graphs/exp2a.png", plot=last_plot(), width = 20, height = 10, units = "cm")
```

```
ggplot(data.e2.condsplit, aes(x = condition, y = diff)) +
  geom_boxplot(aes(fill=condition)) +
```

```
  geom_dotplot(binaxis = 'y', binwidth=0.02, stackdir = "center") +
  geom_signif(comparisons = list(c("After L0 Exclusion", "After L1 Exclusion")),
              map_signif_level=TRUE,  test="t.test", test.args=list(alternative = "less",
  geom_signif(comparisons = list(c("After L1 Exclusion", "After L2 Exclusion")),
              map_signif_level=TRUE,  test="t.test", test.args=list(alternative = "less",
  geom_signif(comparisons = list(c("After L2 Exclusion", "After L3 Exclusion")),
              map_signif_level=TRUE,  test="t.test", test.args=list(alternative = "less",
  labs(y= "Difference in Normalized Log-Probabilities of Completions (P(Jewish) - P(Israe
    theme(legend.position="none", text=element_text(family="Times New Roman", size=12),
        panel.background = element_rect(fill = "white"),
        panel.grid = element_line(color = "lightgray"),
        axis.line = element_line(colour = "gray"))
```




```
ggsave("./graphs/exp2b.png", plot=last_plot(), width = 15, height = 20, units = "cm")
```

Additional Exploratory Experiments ("Experiment 3")

```
data.e3 <- read_csv('./processed_data_cache/exp2_for_analysis.csv')
```

```
New names:
Rows: 2756 Columns: 19
```

```
── Column specification ────────────────────────────────────────────────────────
Delimiter: "," chr (4): tweet, word, source_file, masked_tweet dbl (11): ...1, ...2, ...3, ...4,
...5, ...6, logp.jewish, logp.israeli, no... lgl (3): included.l1, included.l2,
included.l3 date (1): date
ℹ Use `spec()` to retrieve the full column specification for this data. ℹ
Specify the column types or set `show_col_types = FALSE` to quiet this message.
• `` -> `...1`
• `...1` -> `...2`
• `...2` -> `...3`
• `...3` -> `...4`
• `...4` -> `...5`
• `...5` -> `...6`
```

```r
data.e3 <- data.e3 %>% # remove new possible completions
    filter((str_count(data.e3$tweet, '[mM]uslims?') == 0) &
           (str_count(data.e3$tweet, '[bB]lacks?') == 0) &
           (str_count(data.e3$tweet, '[gG]ays?') == 0) &
           (str_count(data.e3$tweet, '[aA]mericans?') == 0)
           )
glimpse(data.e3)
```

```
Rows: 2,597
Columns: 19
$ ...1              <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 15, 16, 1…
$ ...2              <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 15, 16, 1…
$ ...3              <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 15, 16, 1…
$ ...4              <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 15, 16, 1…
$ ...5              <dbl> 1, 2, 3, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15, 17, 18, …
$ ...6              <dbl> 0, 3, 5, 8, 10, 11, 12, 14, 18, 20, 21, 22, 23, 25, …
$ date              <date> 2023-10-08, 2023-10-08, 2023-10-08, 2023-10-08, 202…
$ tweet             <chr> "Brilliant interview debunking the BBC’s sucking up …
$ word              <chr> "zionist", "zionist", "zionist", "zionist", "zionist…
$ source_file       <chr> "scraped_tweets_zionist_45092.json", "scraped_tweets…
$ masked_tweet      <chr> "Brilliant interview debunking the BBC’s sucking up …
$ logp.jewish       <dbl> -5.711659, -5.826700, -5.356212, -3.806168, -5.27450…
$ logp.israeli      <dbl> -1.7931436, -7.8746087, -0.5667414, -4.8278446, -10.…
$ norm.logp.israeli <dbl> 3.1141228, -2.9673423, 4.3405250, 0.0794218, -5.8000…
$ norm.logp.jewish  <dbl> -2.332498081, -2.447538665, -1.977050581, -0.4270061…
$ diff              <dbl> -5.44662092, 0.51980361, -6.31757560, -0.50642795, 3…
$ included.l1       <lgl> FALSE, FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, FALS…
$ included.l2       <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FAL…
$ included.l3       <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FAL…
```

```r
## --- Commented out so is not run during run-through to generate HTML
# outputs <- get_completions(data.e3$masked_tweet,
#           list(muslim=c("muslim", "Muslims", "muslims", "Muslim"),
#                american=c("American", "american", "americans", "Americans"),
#                gay = c("gay", "Gays", "Gay", "gays"),
#                black = c("blacks", "black", "Black")))
```

```
# data.e3$logp.muslim <- log(outputs[[2]]$muslim)
# data.e3$logp.american <- log(outputs[[2]]$american)
# data.e3$logp.gay <- log(outputs[[2]]$gay)
# data.e3$logp.black <- log(outputs[[2]]$black)
# write.csv(data.e3, './processed_data_cache/exp3_mlm.csv')
```

```
neutral.sentences <- readLines("./data/neutral_sentences.txt")
outputs <- get_completions(neutral.sentences,
          list(muslim=c("muslim", "Muslims", "muslims", "Muslim"),
               american=c("American", "american", "americans", "Americans"),
               gay = c("gay", "Gays", "Gay", "gays"),
               black = c("blacks", "black", "Black")))

norm.muslim <- log(sum(outputs[[2]]$muslim))
norm.american <- log(sum(outputs[[2]]$american))
norm.gay <- log(sum(outputs[[2]]$gay))
norm.black <- log(sum(outputs[[2]]$black))

paste("Muslim neutral context log-prob:", norm.muslim)
```

```
[1] "Muslim neutral context log-prob: -4.91867459279764"
```

```
paste("American neutral context log-prob:", norm.american)
```

```
[1] "American neutral context log-prob: -2.56057589159823"
```

```
paste("Gay neutral context log-prob:", norm.gay)
```

```
[1] "Gay neutral context log-prob: -4.06169135745965"
```

```
paste("Black neutral context log-prob:", norm.black)
```

```
[1] "Black neutral context log-prob: -2.80575428843228"
```

```
data.e3 <- read_csv('./processed_data_cache/exp3_mlm.csv')
```

```
New names:
Rows: 2597 Columns: 24
── Column specification
──────────────────────────────────────────────────────── Delimiter: "," chr
(4): tweet, word, source_file, masked_tweet dbl (16): ...1, ...2, ...3, ...4,
...5, ...6, ...7, logp.jewish, logp.israe... lgl (3): included.l1, included.l2,
included.l3 date (1): date
ℹ Use `spec()` to retrieve the full column specification for this data. ℹ
Specify the column types or set `show_col_types = FALSE` to quiet this message.
• `` -> `...1`
• `...1` -> `...2`
```

- `...2` -> `...3`
- `...3` -> `...4`
- `...4` -> `...5`
- `...5` -> `...6`
- `...6` -> `...7`

```
data.e3 <- data.e3 %>%
  mutate(norm.logp.muslim = pull(., logp.muslim) - norm.muslim,
         norm.logp.american = pull(.,logp.american) - norm.american,
         norm.logp.gay = pull(.,logp.gay) - norm.gay,
         norm.logp.black = pull(.,logp.black) - norm.black)
write.csv(data.e3, './processed_data_cache/exp3_mlm_normed.csv')
```

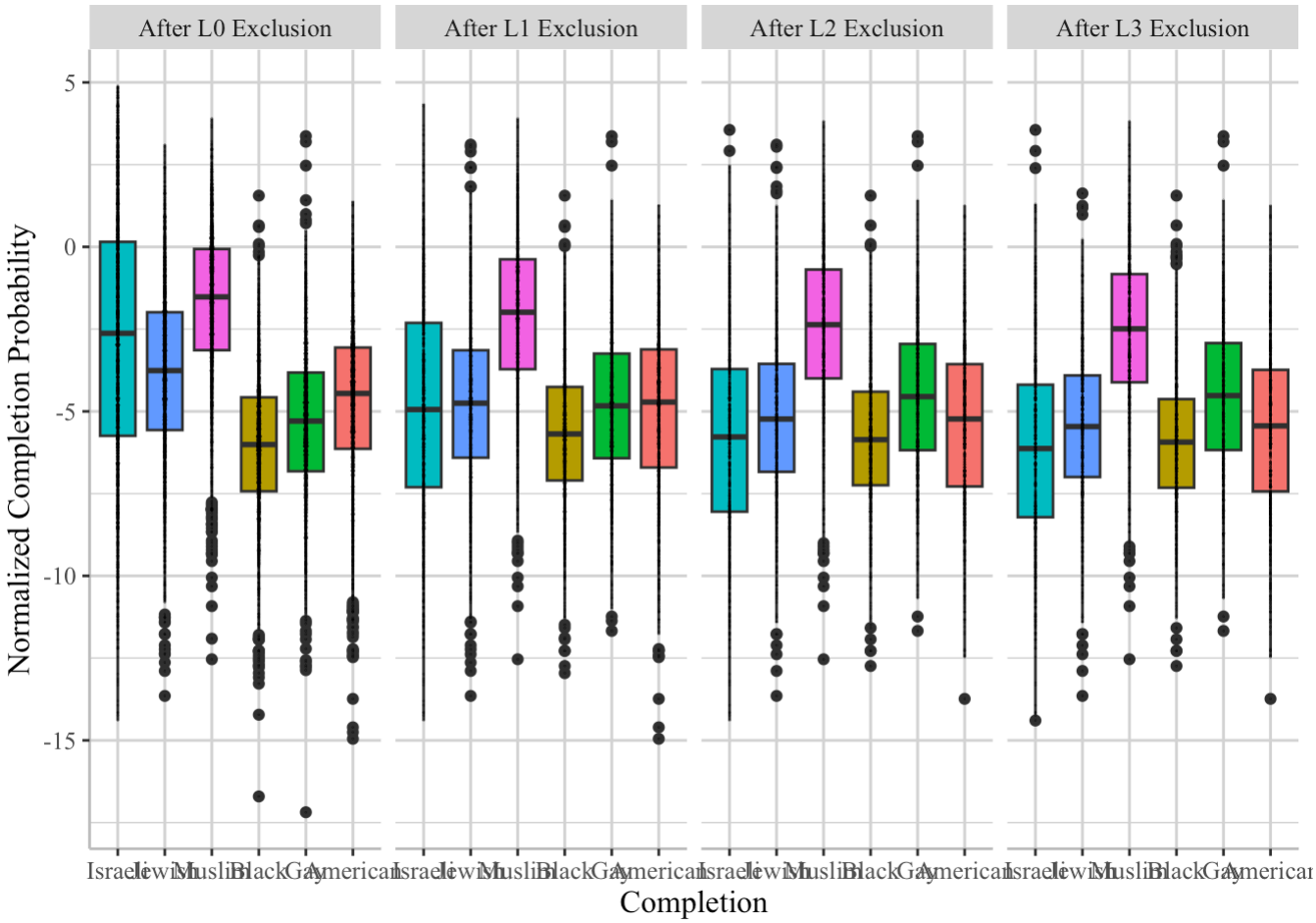
```
data.e3 <- data.e3 %>%
  mutate(condition='After L0 Exclusion')
data.e3.l1 <- filter(data.e3, data.e3$included.l1) %>%
  mutate(condition="After L1 Exclusion")
data.e3.l2 <- filter(data.e3, data.e3$included.l2) %>%
  mutate(condition="After L2 Exclusion")
data.e3.l3 <- filter(data.e3, data.e3$included.l3) %>%
  mutate(condition="After L3 Exclusion")

data.e3.condsplit <- rbind(data.e3, data.e3.l1, data.e3.l2, data.e3.l3)

data.e3.lengthened <- data.frame(
  norm.logp = c(data.e3.condsplit$norm.logp.jewish,
                data.e3.condsplit$norm.logp.israeli,
                data.e3.condsplit$norm.logp.muslim,
                data.e3.condsplit$norm.logp.black,
                data.e3.condsplit$norm.logp.gay,
                data.e3.condsplit$norm.logp.american),
  condition = data.e3.condsplit$condition,
  completion = c(rep("Jewish", length(data.e3.condsplit$norm.logp.jewish)),
                 rep("Israeli", length(data.e3.condsplit$norm.logp.israeli)),
                 rep("Muslim", length(data.e3.condsplit$norm.logp.muslim)),
                 rep("Black", length(data.e3.condsplit$norm.logp.black)),
                 rep("Gay", length(data.e3.condsplit$norm.logp.gay)),
                 rep("American", length(data.e3.condsplit$norm.logp.american)))) %>%
  mutate(condition_f = factor(pull(., condition), levels=c('After L0 Exclusion','After L1
         completion_f = factor(pull(., completion), levels=c("Israeli", "Jewish", "Musli

ggplot(data.e3.lengthened, aes(x = completion_f, y = norm.logp)) +
  geom_boxplot(aes(fill=completion)) +
  geom_dotplot(binaxis = 'y', binwidth=0.01, stackdir = "center") +
  facet_wrap(~condition_f, nrow=1)  +
  labs(x = "Completion", y = "Normalized Completion Probability") +
    theme(legend.position="none", text=element_text(family="Times New Roman", size=12),
        panel.background = element_rect(fill = "white"),
```

```
        panel.grid = element_line(color = "lightgray"),
        axis.line = element_line(colour = "gray"))
```




```
ggsave("./graphs/exp3.png", plot=last_plot(), width = 30, height = 10, units = "cm")
```