Synthesizing an experimental dataset for the evaluation of generic frame detection as a task of detecting communicative intention: Online Appendix

Synthesizing an experimental dataset for the evaluation of generic frame detection as a task of detecting communicative intention: Online Appendix

Coding scheme (Semetko & Valkenburg)

Attribution of responsibility

- A1: Does the story suggest that some level of government has the ability to alleviate the problem?
- A2: Does the story suggest that some level of the government is responsible for the issue/problem?
- A3: Does the story suggest solution(s) to the problem/issue?
- A4: Does the story suggest that an individual (or group of people in society) is responsible for the issue-problem?
- A5: Does the story suggest the problem requires urgent action?

Human interest frame

- B1: Does the story provide a human example or "human face" on the issue?
- B2: Does the story employ adjectives or personal vignettes that generate feelings of outrage, empathy-caring, sympathy, or compassion?
- B3: Does the story emphasize how individuals and groups are affected by the issue/problem?
- B4: Does the story go into the private or personal lives of the actors?
- ~~B5: Does the story contain visual information that might generate feelings of outrage, empathy-caring, sympathy, or compassion?~~

Conflict frame

- C1: Does the story reflect disagreement between parties/individuals/groups/countries?
- C2: Does one party/individual/group/country reproach another?
- C3: Does the story refer to two sides or to more than two sides of the problem or issue?
- C4: Does the story refer to winners and losers?

Morality frame

- D1: Does the story contain any moral message?
- D2: Does the story make reference to morality, God, and other religious tenets?
- D3: Does the story offer specific social prescriptions about how to behave?

(Economic) consequences frame

- E1: Is there a mention of financial losses or gains now or in the future?
- E2: Is there a mention of the costs/degree of expense involved?
- E3: Is there a reference to economic consequences of pursuing or not pursuing a course of action?

Intercoder reliability

After two rounds of precoding and coder training, 10 items were coded by the two student coders. The intercoder reliability is showed below.

item	Krippendorff's Alpha
A1	0.39
A2	0.92
A3	0.63
A4	0.14
A5	-0.48
B1	0.84
B2	0.76
B3	0.66
B4	0.19
C1	0.71
C2	-0.26
C3	0.10
C4	-0.19
D1	0.61
D2	0.86
D3	0.19
E1	0.39
E2	0.06
E3	0.20

Overview of all automatic and supervised methods employed

k-Means. Each document is assigned a vector containing the frequency of each word that occurs in the document (term frequency - TF), weighed against how many documents that word appears in (inverse document frequency - IDF). That way, the term vector is a representation of how documents differ from each other based on how different the words that are used within are. Weighing against inverse document frequency ensures

that rare words (which contain more information, e.g. “nuclear”) have more weight than common, generic words (e.g. “man”). Cluster analysis is used to group together documents based on how similar their word vectors are - with k-means clustering, the number of clusters is decided beforehand. Clusters are found by searching for the cluster distribution with the lowest within-cluster sum of squares (Burscher, Odijk, Vliegenthart, Rijke, & De Vreese, 2014).

Principal Component Analysis (PCA). As above, each document is assigned a vector containing the frequency of each word that occurs in the document (term frequency - TF), weighed against how many documents that word appears in (inverse document frequency - IDF). A TF-IDF can be interpreted as a long list of dimensions in an n-dimensional space. Principal Component Analysis is used to reduce those dimensions by projecting the n-dimensional data onto a space with fewer dimensions (in our case: 5) while maintaining as much variance (and hence: information) as possible. That way, individual word vectors that often occur together will inform the same components in the resulting model. The components can then be interpreted as correlating with Frames (Greussing & Boomgaarden, 2017).

Latent Dirichlet Allocation (LDA). Topic Modelling assumes that there is a number of topics distributed over the documents you analyze. Topics are expressed through co-occurrence of common vocabulary (e.g.: an “economy” topic would be expressed through words like “costs”, “inflation” etc.). Latent Dirichlet Allocation follows this intuition and tries to allocate words into topics such that both for each document, the words are allocated to as few topics as possible and that for few and rare words, higher topic probabilities are assigned than to common terms. From a Bayesian perspective, each document is assigned topic probabilities for a pre-set number of topics using a hierarchical probabilistic model. The resulting probabilities represent commonly co-occuring terms, which are interpreted as representing common topics of documents (DiMaggio, Nag, & Blei, 2013).

Structural Topic Model (STM). Uses a similar approach to LDA, described above, but allows for using metadata that covaries with topics and allows for topics that correlate with each other (Nicholls & Culpepper, 2020).

Analysis of Topic Model Networks (ANTMN). LDA topic modelling is applied (see above). The number of topics is optimized based on commonly used indicators to assess topic quality. Then, co-occurence of topics is used to generate a topic model network: topics are treated as nodes, their theta-cosine similarity is treated as edges between nodes. Network community detection algorithms are used to combine topics into larger topic communities based on this similarity-network, which are interpreted as representing frames (Walter & Ophir, 2019).

Seeded-LDA. Employs LDA topic modelling (see above), but the researcher applies seed terms derived from theory to guide the topic modelling process: Each seed provided by the researcher is given a prior weight towards topics. The LDA model then assigns topics to documents taking these prior weights into account (Watanabe & Zhou, 2020).

Keyword Assisted Topic Model (keyATM). Similar to seeded-LDA (Eshima, Imai, & Sasaki, 2020).

Multiverse analysis of all methods (based on the ground truth)

These figures display the multiverse analyses of human coding (Figure 1), K-Means + TF-IDF (Figure 2), PCA + TF-IDF (Figure 3), LDA (Figure 4), STM (Figure 5), ANTMN (Figure 6), seeded LDA (Figure 7), and KeyATM (Figure 8). These figures display the analyses using the original 100-article corpus (green dots) and the corpus without the Morality articles (orange dots).

Visualization of the variance. The variance of CCR_{max} is displayed in Figure 9.

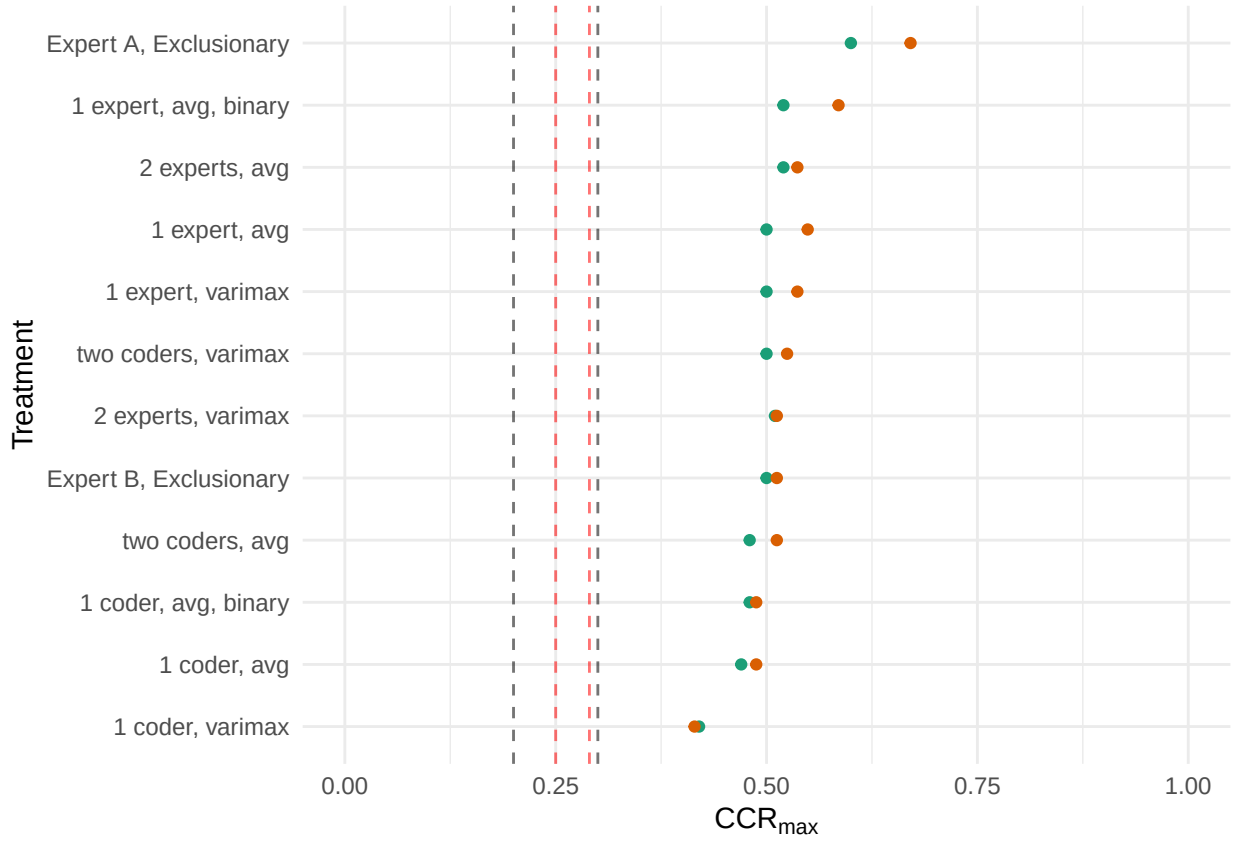

Figure 1. Multiverse analysis of human coding

Sensitivity analysis: Without Morality articles. We retested H1 without all Morality articles (Figure 10). H1 appears to be less supported when all Morality articles are removed.

Multiverse analysis of all methods (based on correlation)

The calculation of CCR_{max} makes an assumption that a method can extra a dominant frame. But in fact, most of the methods under consideration can extra multiple "frames" (with the exception of K-Means and the exclusionary item "F1" by the experts). For instance, an LDA model can export for each article a vector of five numbers, θ_t. the calculation of CCR_{max} select the highest number among these five numbers as the dominant frame. However, it is possible that the second highest θ_t might match the ground truth. But CCR_{max} does not consider this information.

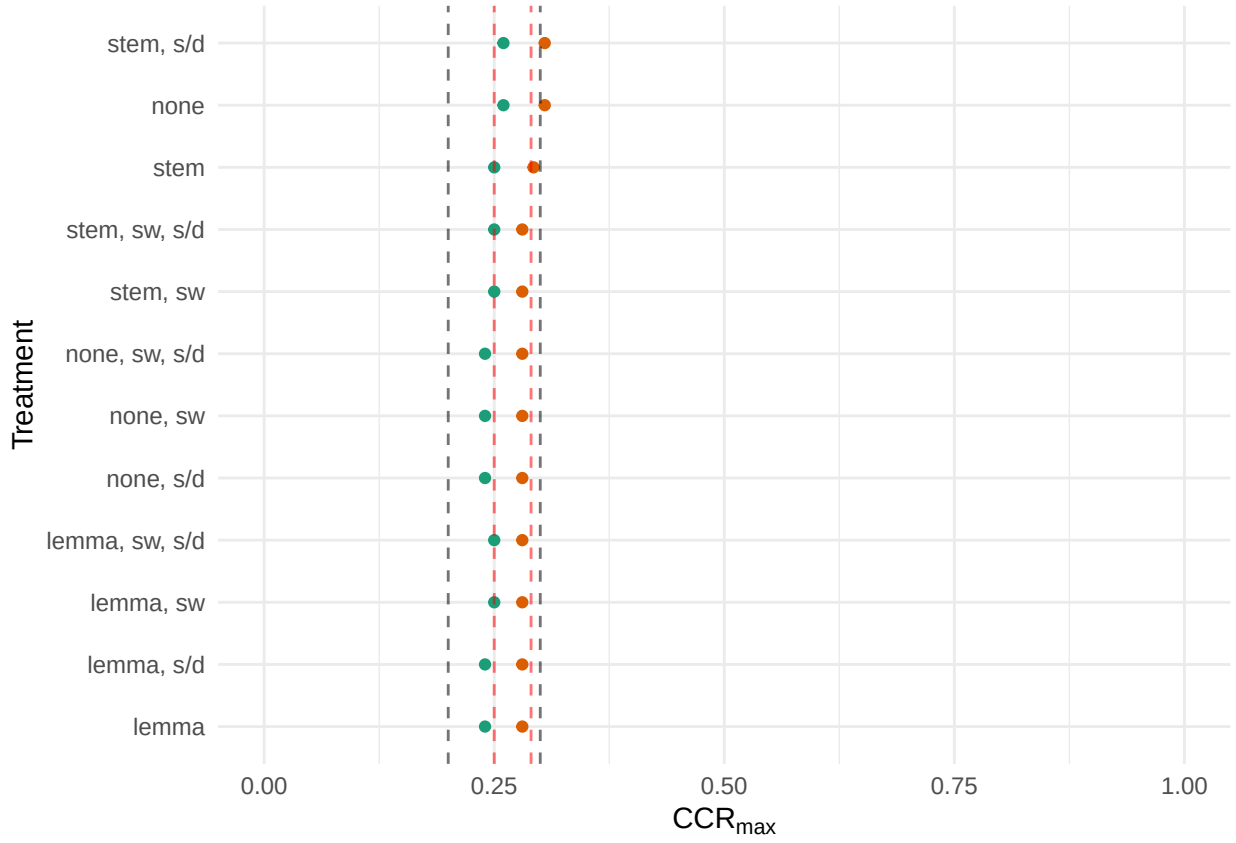

Figure 2. Multiverse analysis of K-Means

We vectorize the output 100×5 matrix $\mathbf{P}$ by stacking the columns of $\mathbf{P}$ on top of one another, i.e.

$$\text{vec}(\mathbf{P}) = [p_{1,1}, \dots, p_{100,1}, p_{1,2}, \dots, p_{100,2}, \dots, p_{1,5}, \dots, p_{100,5}]^{\text{T}}$$

Another way to evaluate the accuracy is to calculate the correlation between two vectorized $\mathbf{P}$ (each with a dimension of 1×500). The ground truth vector y can be first matricized as a one-hot matrix and vectorized the same way to generate the same 1×500 vector. For instance:

$$y = [1, 2, 3, 1, \dots, 5]$$

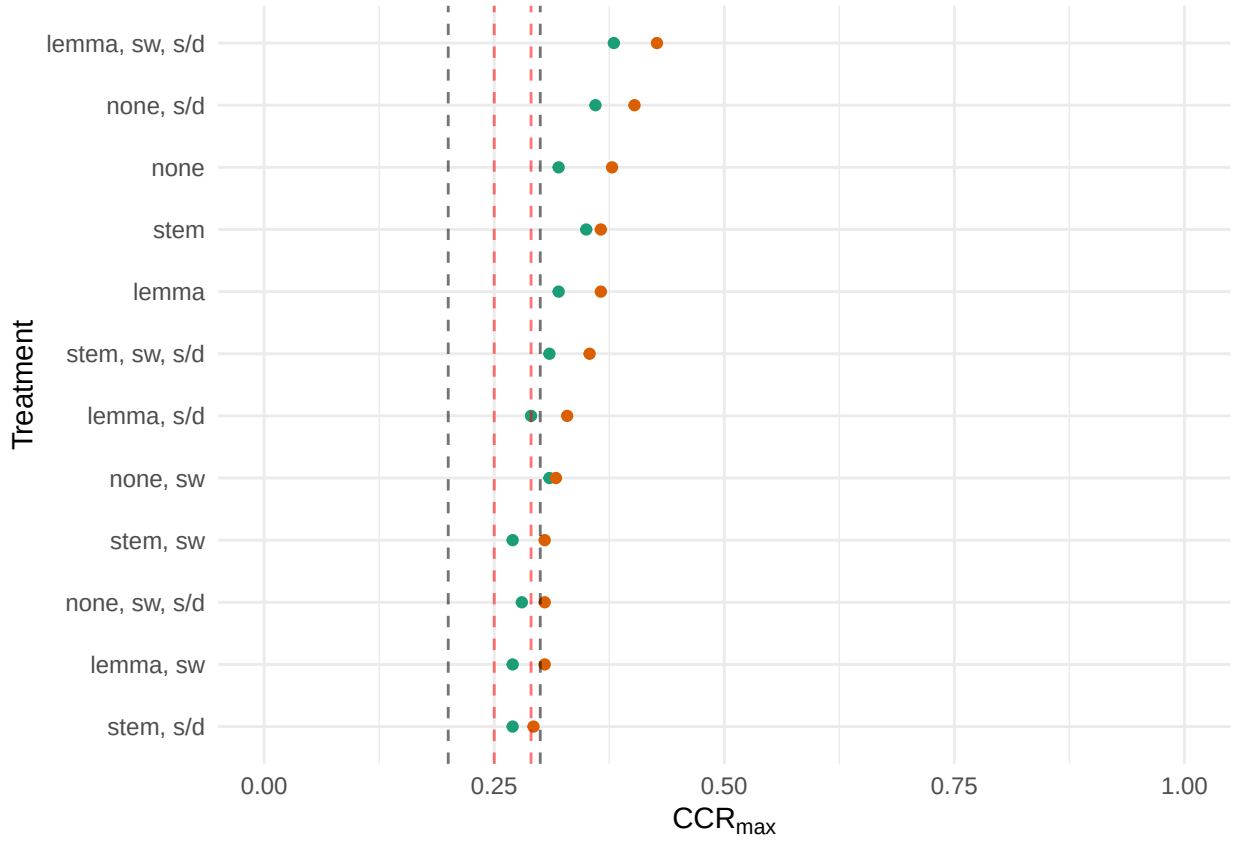

Figure 3. Multiverse analysis of PCA

$$\text{mat}(y) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Using this approach, we encountered the same problem of not knowing which column in $\mathbf{P}$ corresponds to which actual frame in the ground truth y. The same approach of exhaustive search was used and the maximum value of correlation was reported.

These figures display the multiverse analyses of human coding (Figure 1), PCA + TF-IDF (Figure 12), LDA (Figure 14), STM (Figure 13), ANTMN (Figure 15), seeded

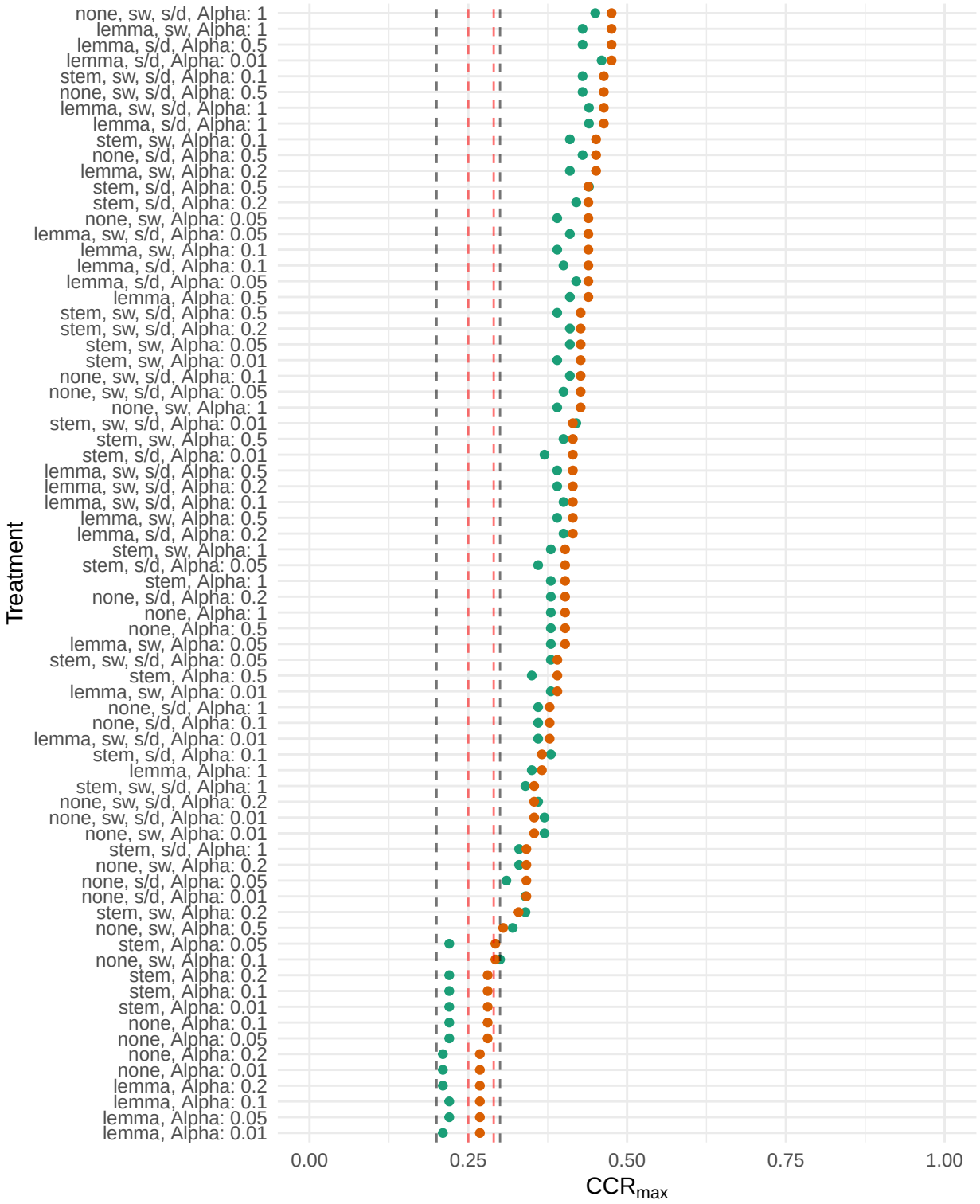

Figure 4. Multiverse analysis of LDA

Figure 5. Multiverse analysis of STM

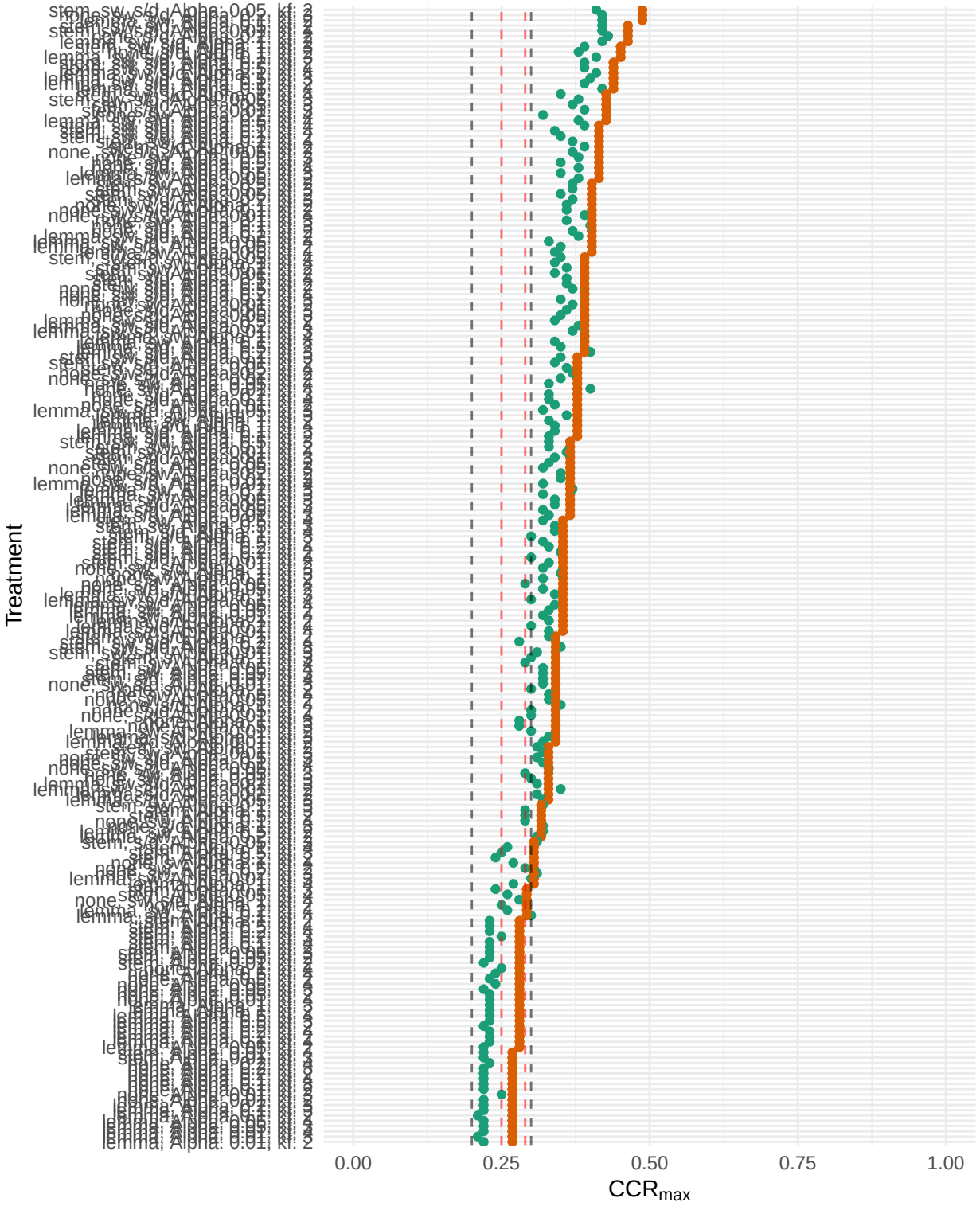

Figure 6. Multiverse analysis of ANTMN

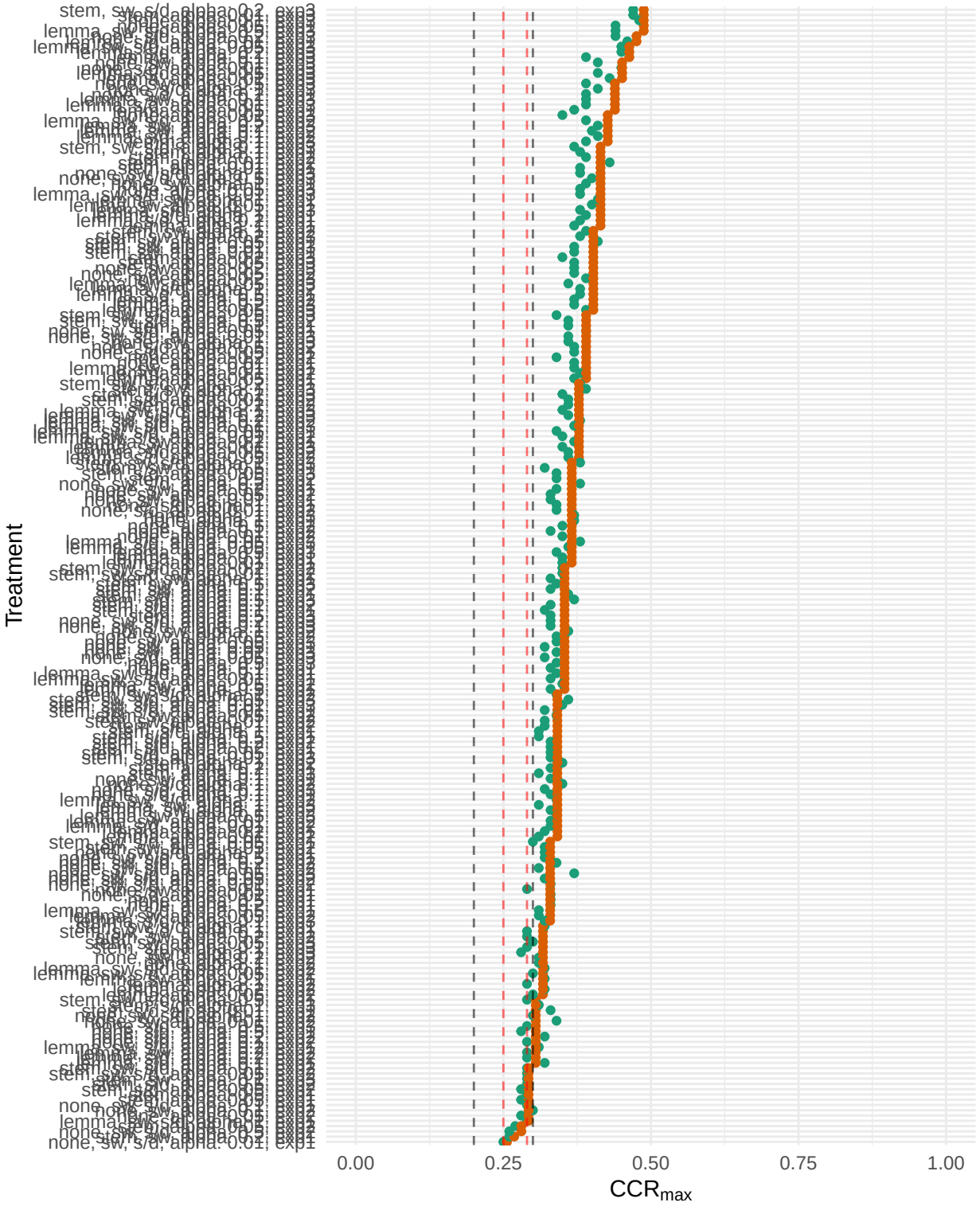

Figure 7. Multiverse analysis of seeded-LDA

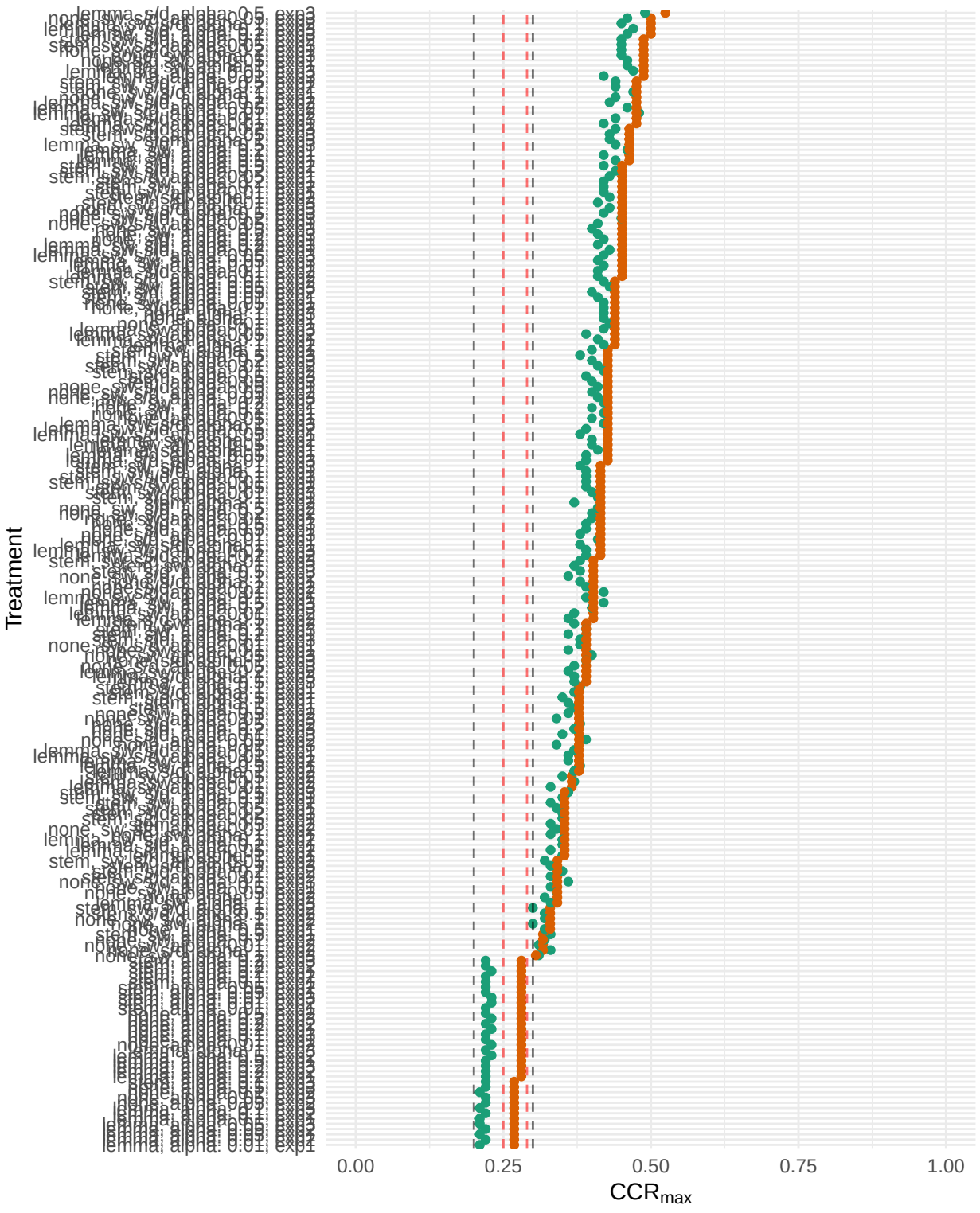

Figure 8. Multiverse analysis of keyATM

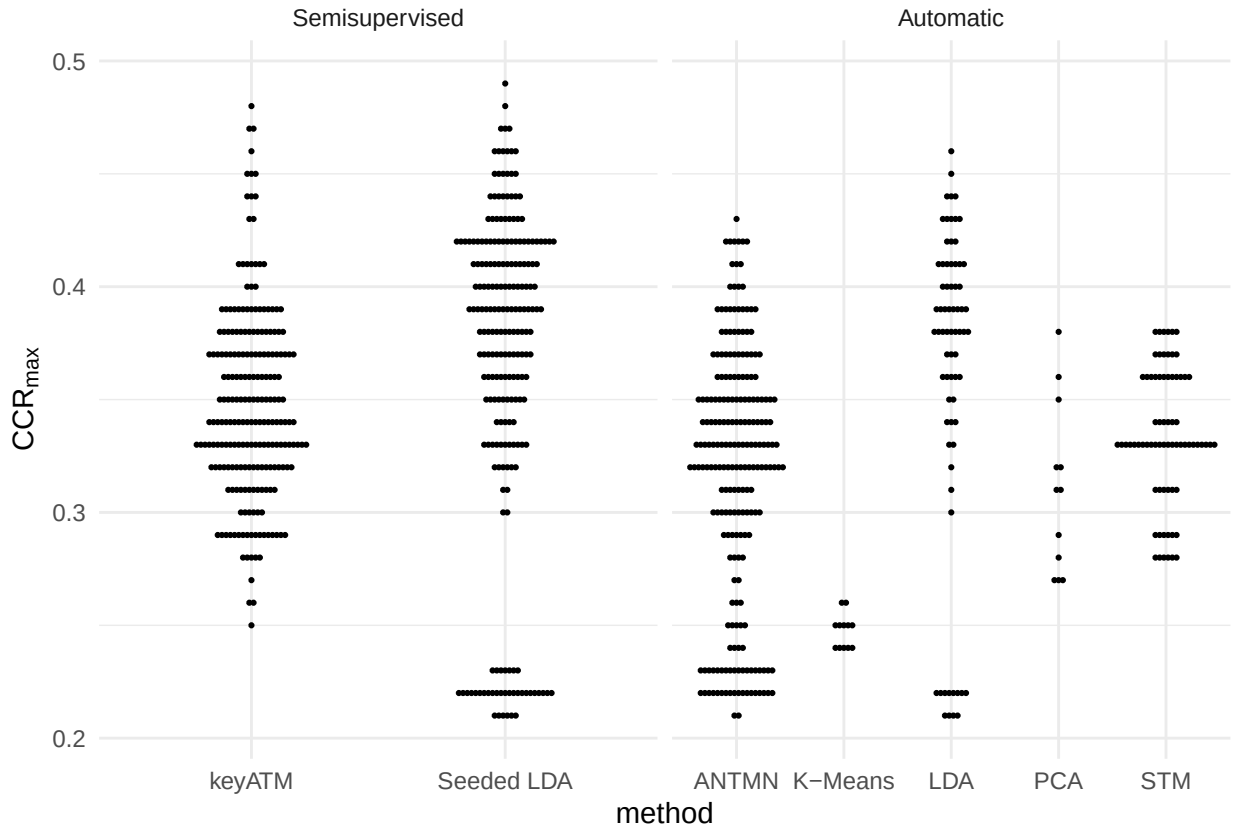

Figure 9. Variance of the best case correct classification rates

LDA (Figure 16), and KeyATM (Figure 17). These figures display the analyses using the ground truth (green dots) and the expert coding (orange dots, averaged of the two experts).

Sensitivity analysis: With correlation, using ground truth. We retested H1 with correlation and ground truth (Figure 18). H1 appears to be similarly supported.

Sensitivity analysis: With correlation, using expert coding. We retested H1 with correlation and expert coding (Figure 19). H1 appears to be similarly supported. Please note that this is not a task of detecting communicative intention.

Comparing confidence level of correct and incorrect expert coding

We modeled the correctness of expert coding ("F1" is equal to the ground truth) and confidence level ("F2"), while adjusting for individual differences between the two experts

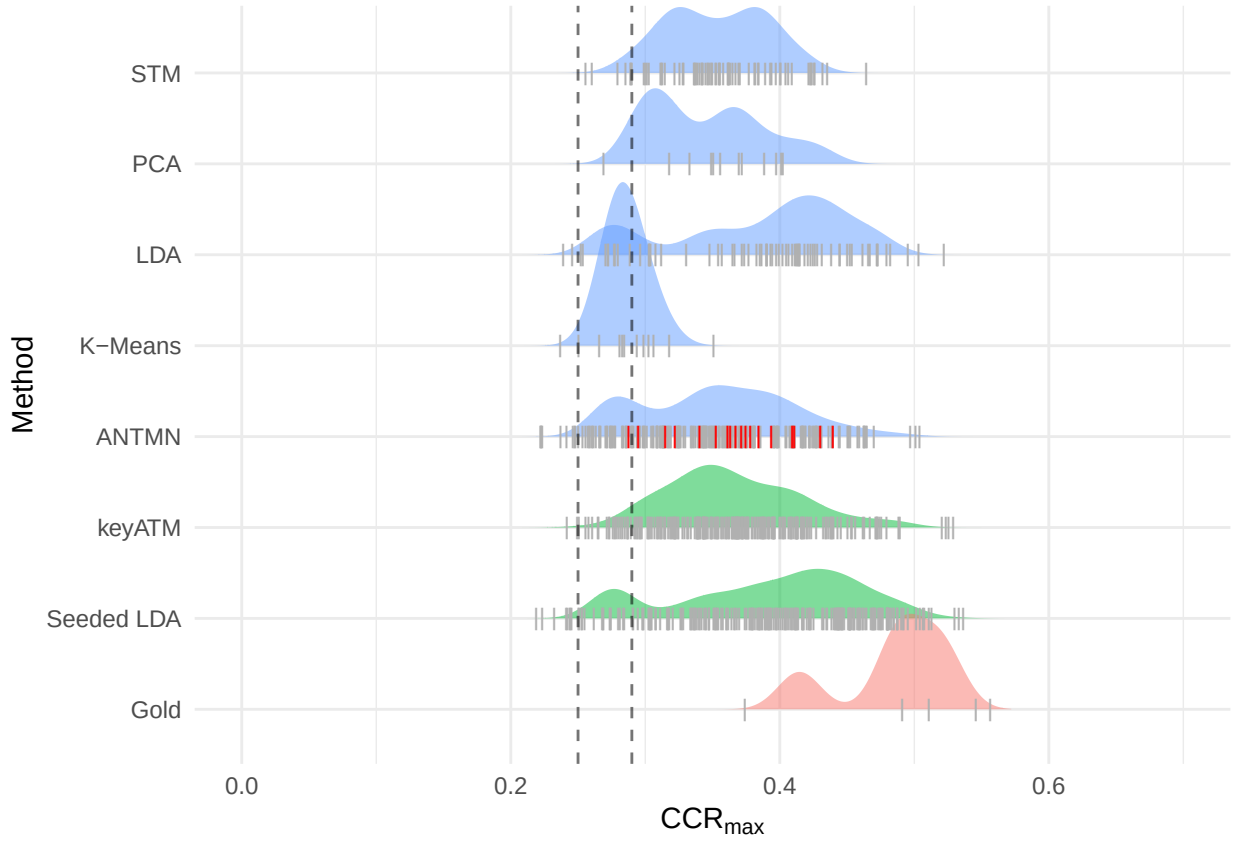

Figure 10. Distribution of best-case correct classification rates by methods (without Morality articles)

using Bayesian multilevel logistic regression analysis. The following is the robust conditional effect plot. There is no evidence to suggest that there is a trend. Therefore, experts can either confidently give correct and incorrect coding.

Simulation of increasing sample size

In this analysis, we simulated the possible outcome of increasing the sample size on the multiverse analysis.

From our 100 articles, we created further synthetic articles following the principle of bootstrapping. We synthesized more articles based on the following algorithm:

1. Randomly select one article

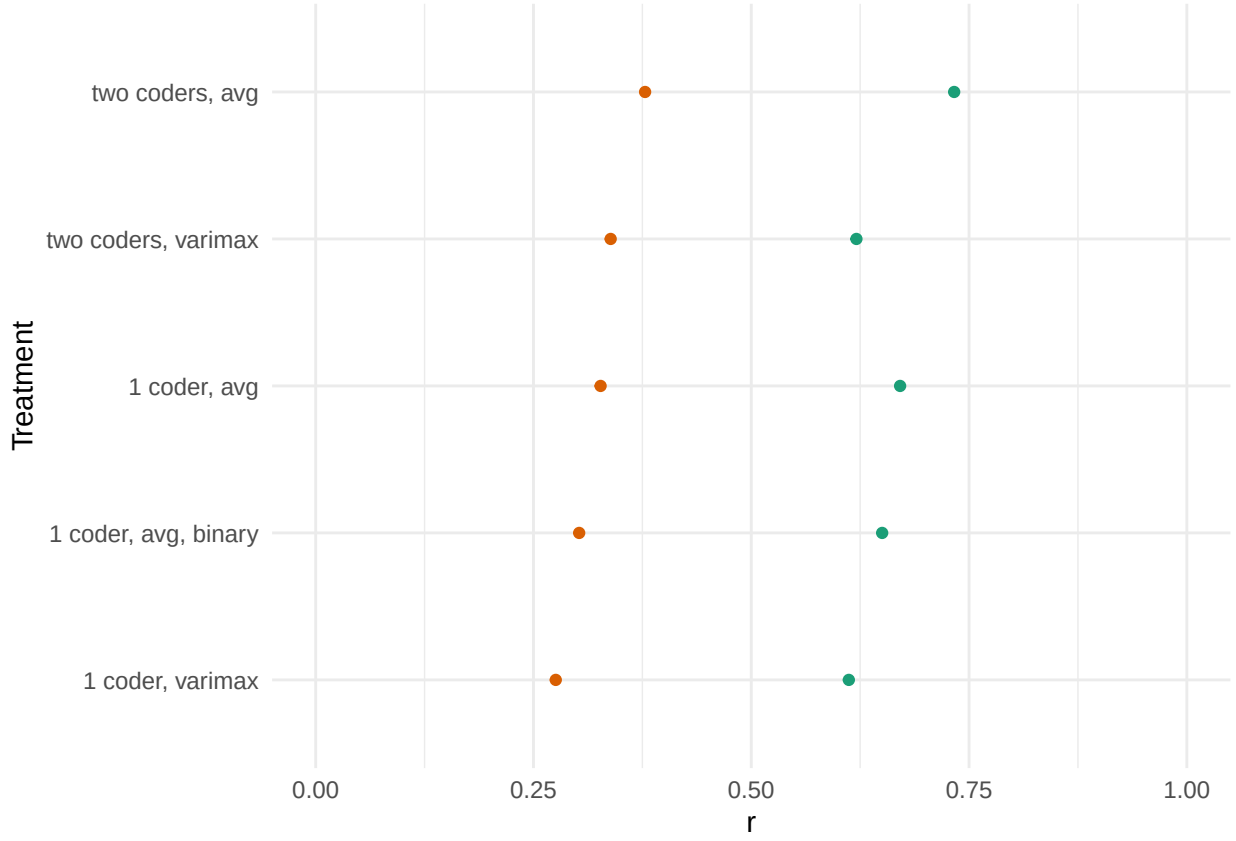

Figure 11. Multiverse analysis of human coding based on correlation

2. Tokenize this article into its n sentences
3. From these n sentences, randomly draw n sentences from these sentences with replacement. Therefore, one sentence can appear more than once.
4. Concatenate these randomly drawn n sentences into a synthetic article, assign this article with the same topic and frame as the original article

We repeat the above process for 500, 1000, and 2000 times to generate 3 different corpora. This approach is compatible with the bag-of-words representation used in all unsupervised and semi-supervised methods because the word order is not considered. Also in step 3, topical and frame clues have the same natural chance of being selected. To put this simulation in another perspective, it simulates whether frames, rather than topics, are more likely to be picked up by these unsupervised and semi-supervised methods when the

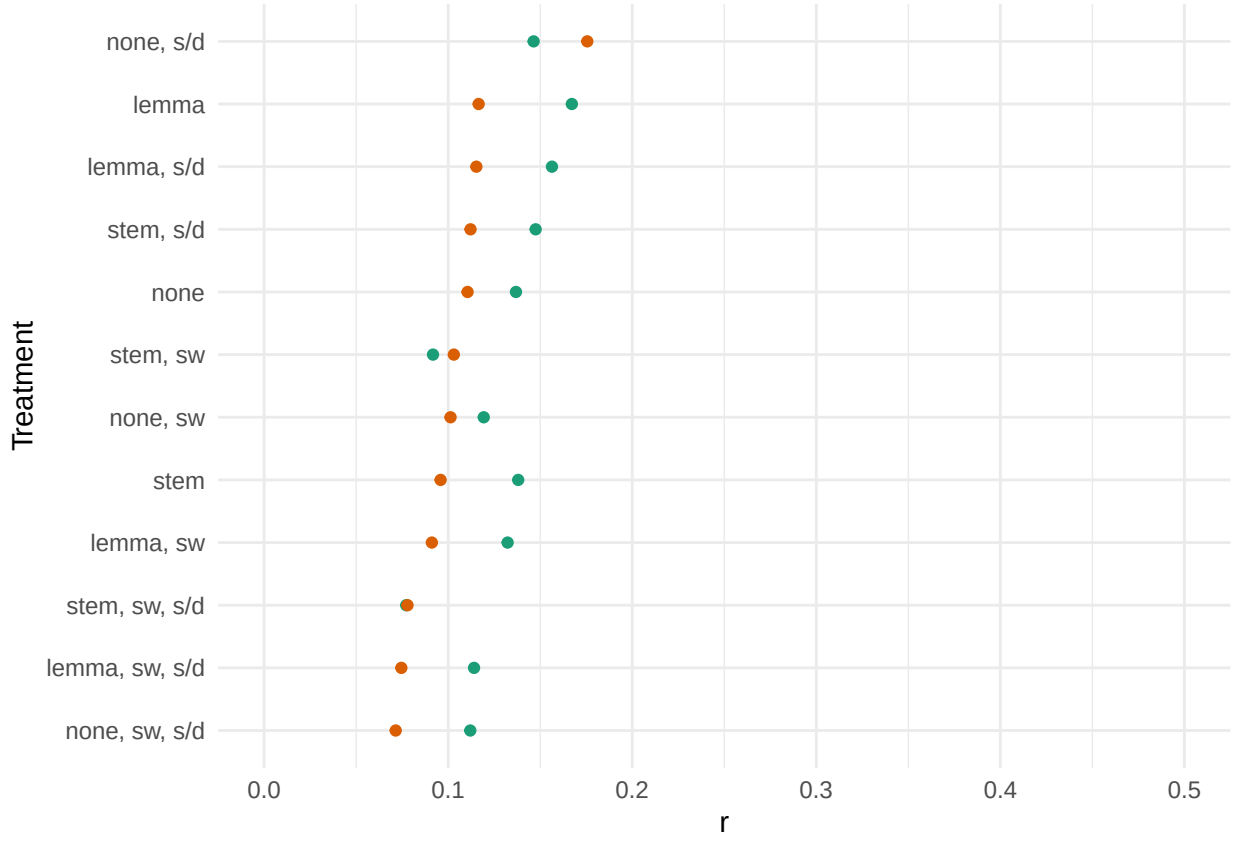

Figure 12. Multiverse analysis of PCA based on correlation

sample size increases.

The above figure shows the sorted CCR_{max} like the visualization of multiverse in the main text. All confidence intervals have been stripped for clarity as we are only interested in the point estimate. The treatments are the same as the main analysis and they are not displayed in the y-axes.

From this simulation, it is extremely unlikely that increasing the sample size would increase the CCR_{max} for all unsupervised methods. Instead, all methods, except KM and STM, perform more like the de-facto null (0.3) with the increasing sample size. Therefore, these methods appear to be more keen to pick up **topics**, rather than frames, when sample size increases: the sorted performance curve rescinding towards the null value with the highest sample size. For the two semi-supervised methods, increasing the sample size

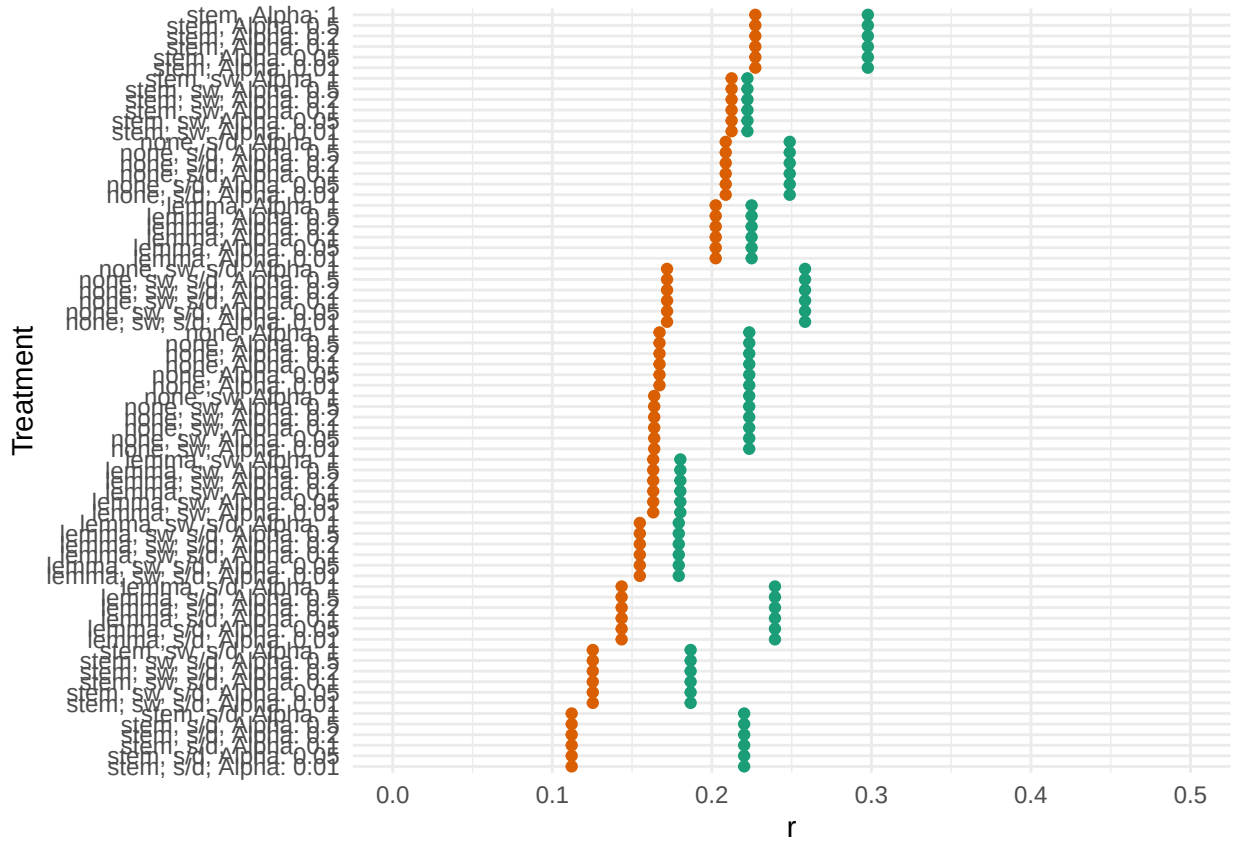

Figure 13. Multiverse analysis of STM based on correlation

appears to provide better resistance against the rescinding.

References

Burscher, B., Odijk, D., Vliegenthart, R., Rijke, M. de, & De Vreese, C. H. (2014). Teaching the computer to code frames in news: Comparing two supervised machine learning approaches to frame analysis. *Communication Methods and Measures*, *8*(3), 190–206. https://doi.org/10.1080/19312458.2014.937527

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics*, *41*(6), 570–606. https://doi.org/10.1016/j.poetic.2013.08.004

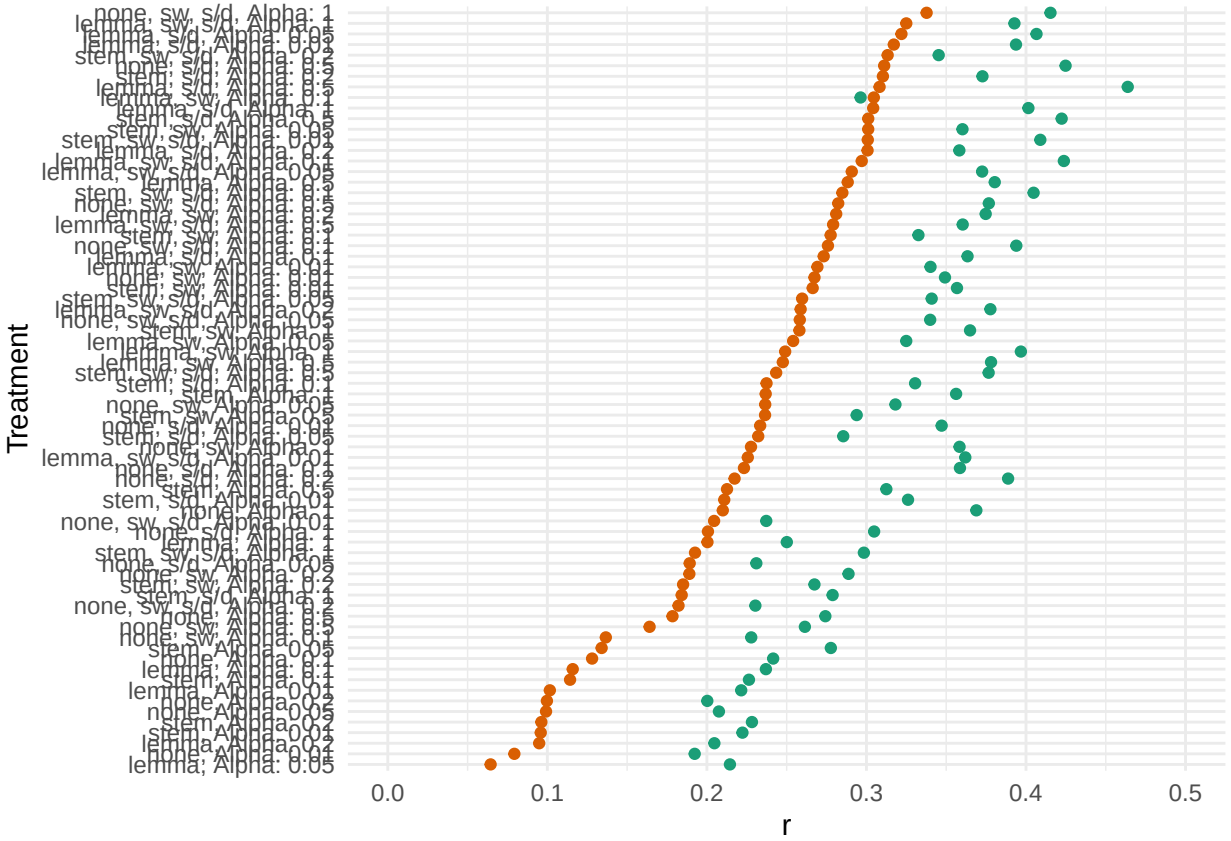

Figure 14. Multiverse analysis of LDA based on correlation

Eshima, S., Imai, K., & Sasaki, T. (2020). Keyword assisted topic models. *arXiv Preprint arXiv:2004.05964*.

Greussing, E., & Boomgaarden, H. G. (2017). Shifting the refugee narrative? An automated frame analysis of Europe's 2015 refugee crisis. *Journal of Ethnic and Migration Studies*, *43*(11), 1749–1774. https://doi.org/10.1080/1369183x.2017.1282813

Nicholls, T., & Culpepper, P. D. (2020). Computational identification of media frames: Strengths, weaknesses, and opportunities. *Political Communication*, 1–23. https://doi.org/10.1080/10584609.2020.1812777

Walter, D., & Ophir, Y. (2019). News frame analysis: An inductive mixed-method computational approach. *Communication Methods and Measures*, *13*(4),

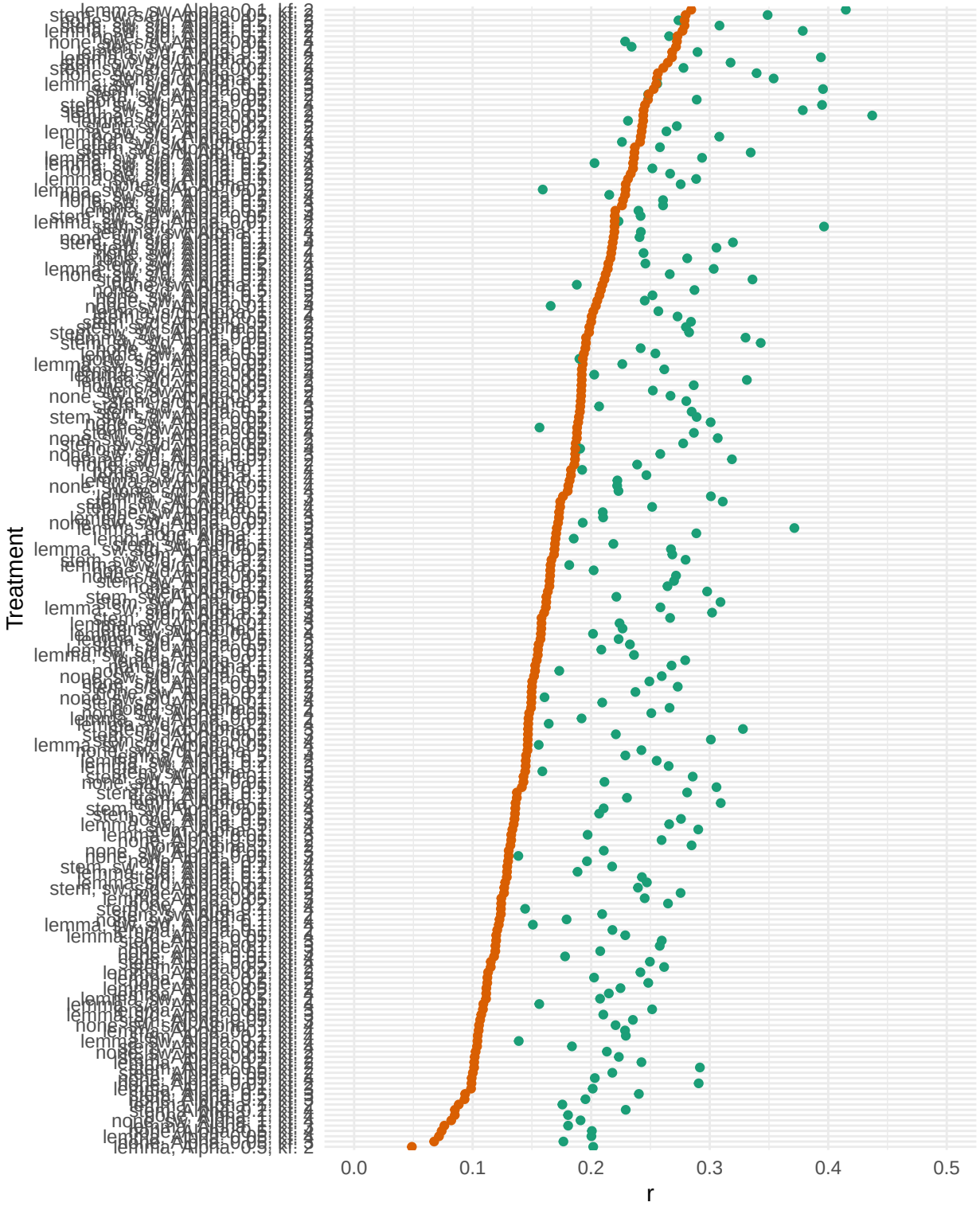

Figure 15. Multiverse analysis of ANTMN based on correlation

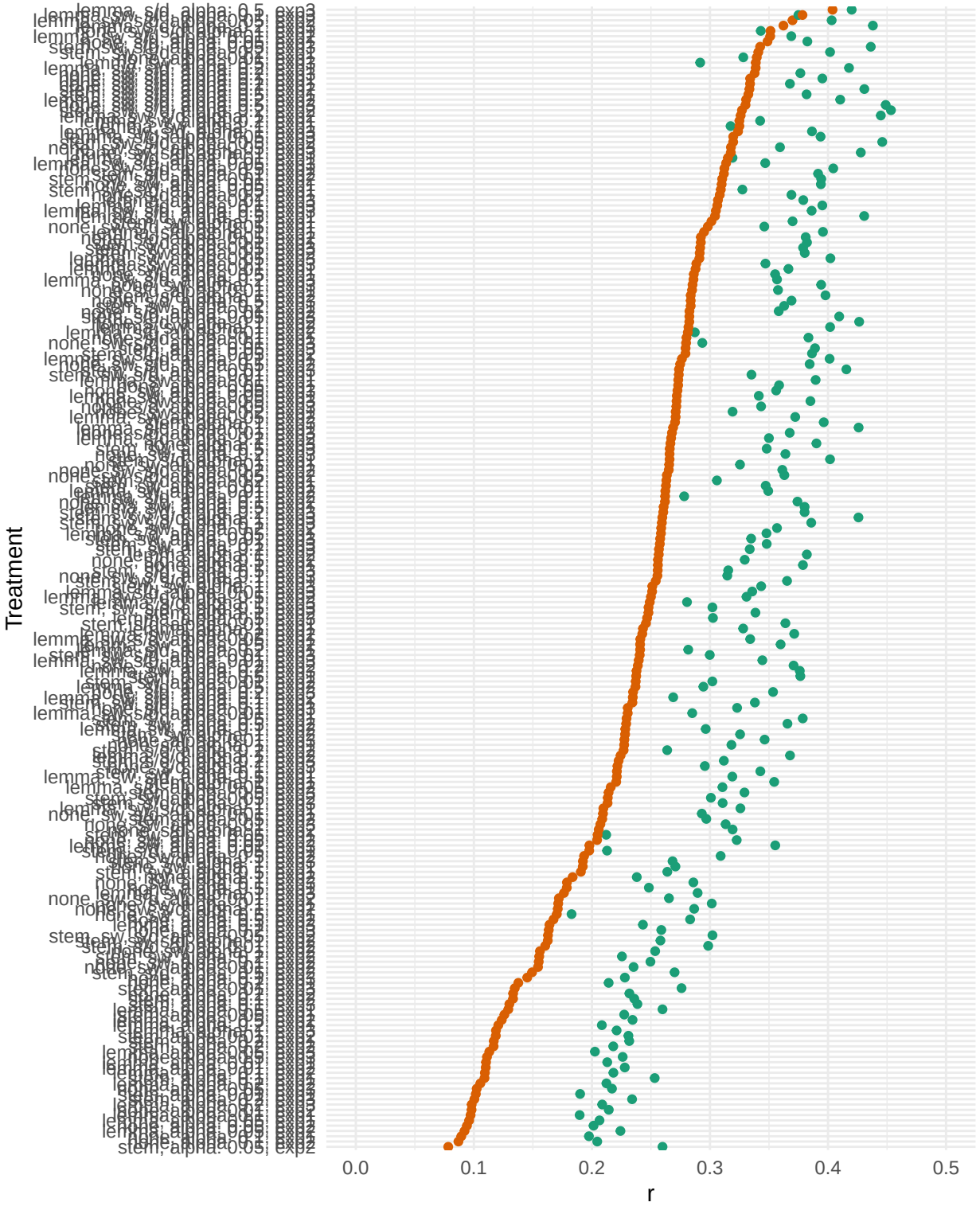

Figure 16. Multiverse analysis of seeded-LDA based on correlation

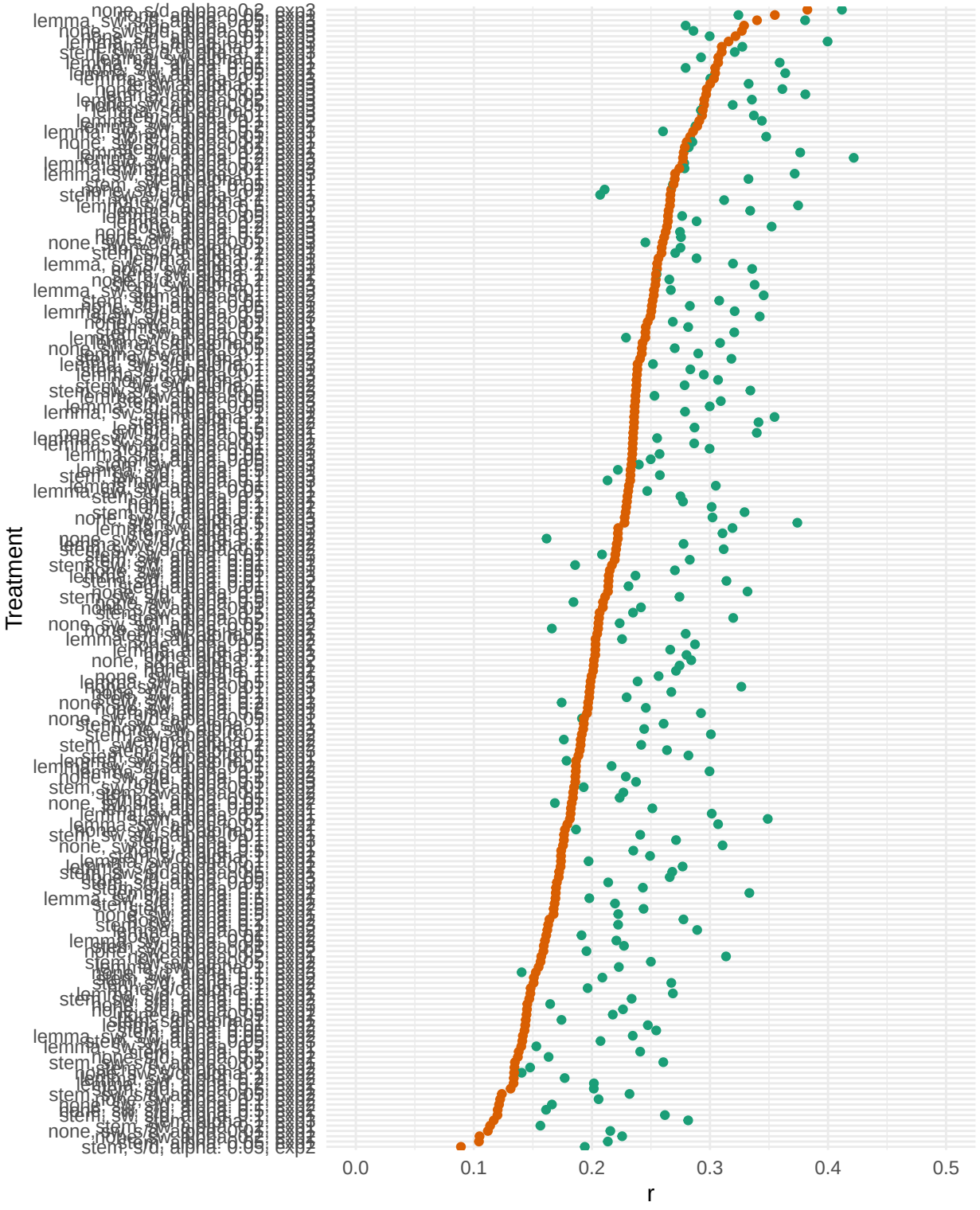

Figure 17. Multiverse analysis of keyATM based on correlation

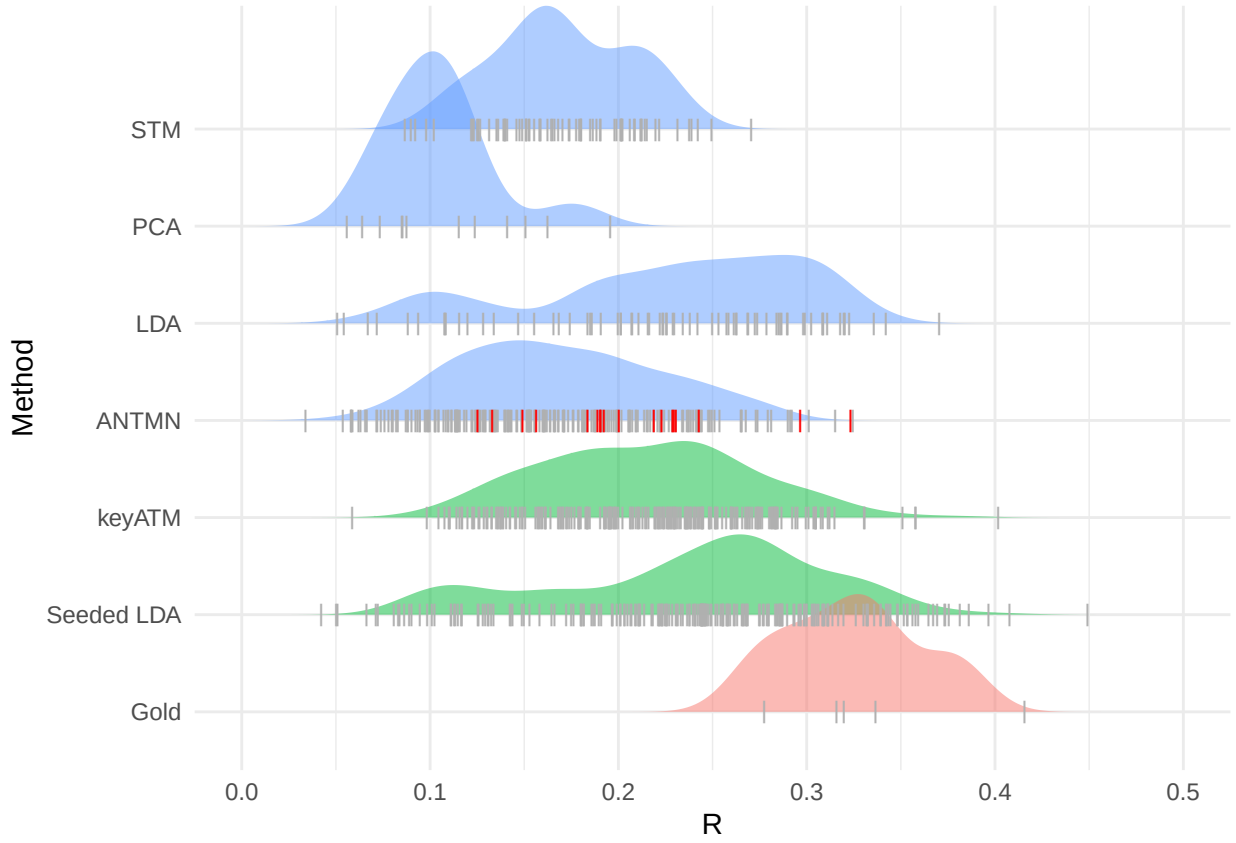

Figure 18. Distribution of correlation by methods (using ground truth)

248–266. https://doi.org/10.1080/19312458.2019.1639145

Watanabe, K., & Zhou, Y. (2020). Theory-driven analysis of large corpora: Semisupervised topic classification of the UN speeches. *Social Science Computer Review*, *40*(2), 346–366. https://doi.org/10.1177/0894439320907027

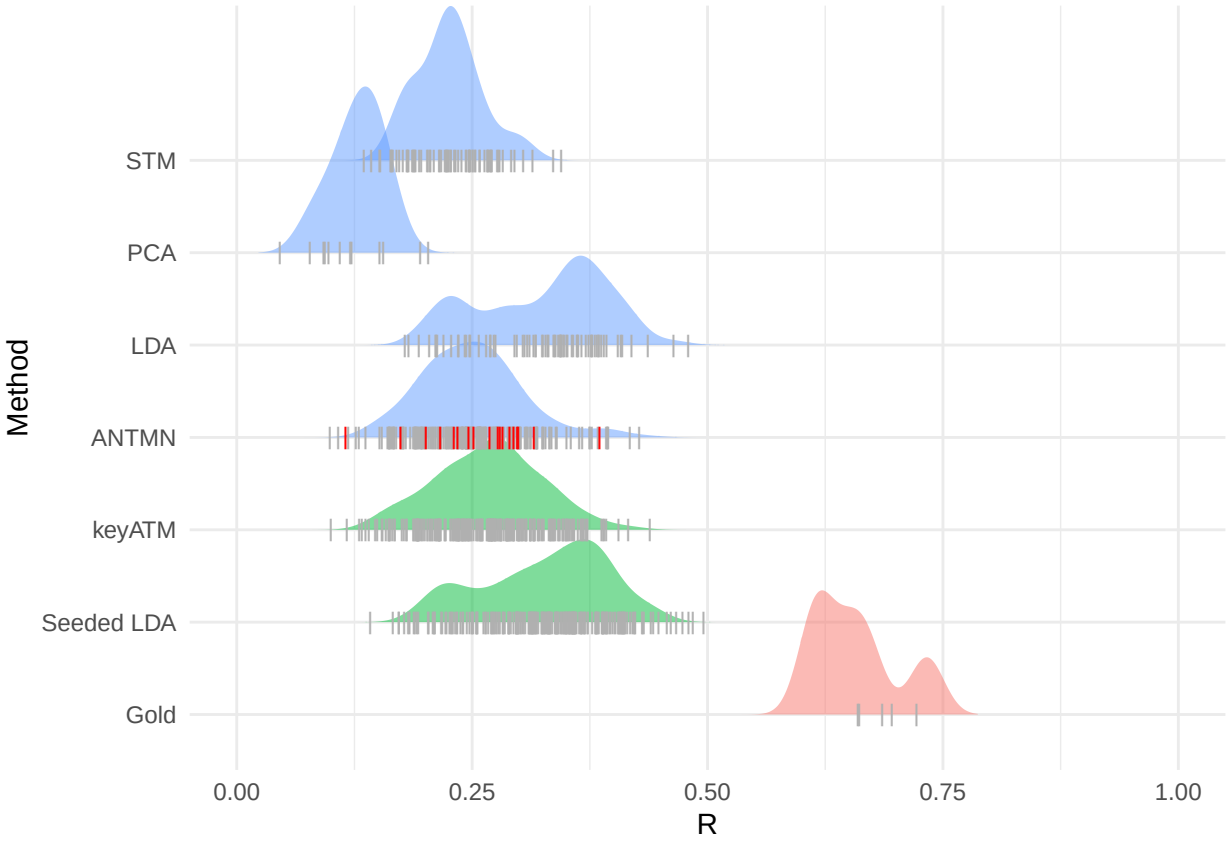

Figure 19. Distribution of correlation by methods (using expert coding)

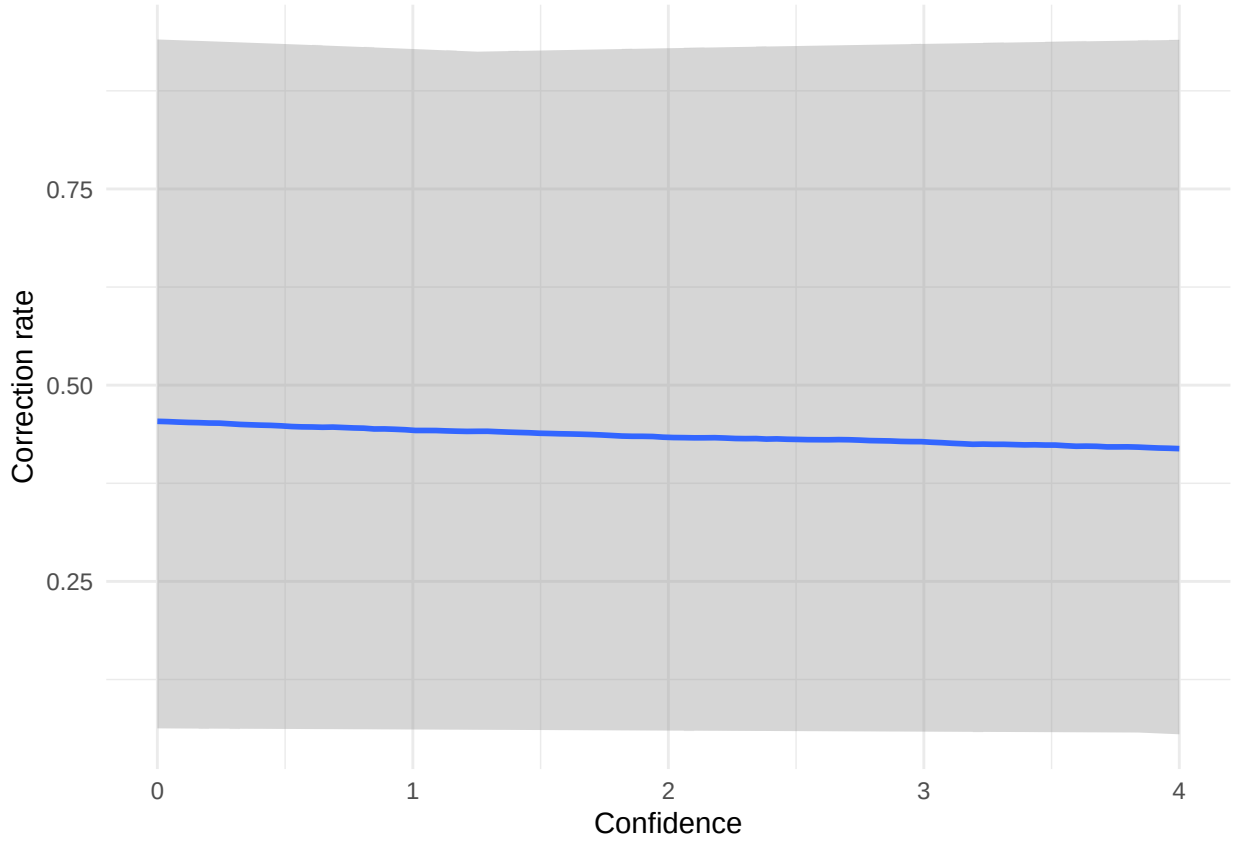

Figure 20. Robust conditional effects from the Bayesian model on the relationship between correction rate of expert coding

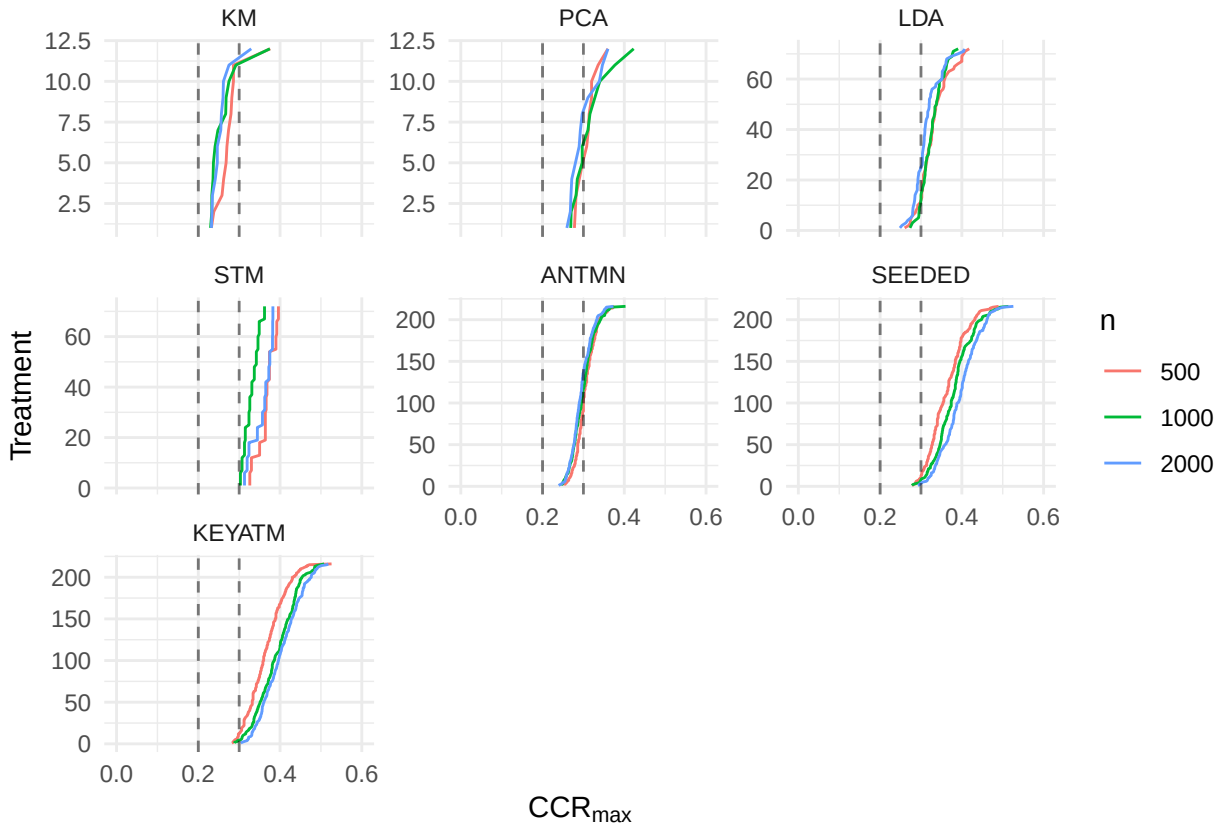

Figure 21. Multiverse analyses with different sample sizes: 500, 1000, 2000