Regression Analysis

```
survey_data <- read.csv('../../../backend/data/database/survey_data.csv')
survey_data$TreatmentGroup <- as.factor(survey_data$TreatmentGroup)
survey_data$TreatmentGroup <- relevel(survey_data$TreatmentGroup, ref = "machine")
head(survey_data, n = 10)
```

```
##                      AnswerId         FK_ParticipantId             FK_SessionId
## 1  57bb0ebfd4654c00018e0261T1 57bb0ebfd4654c00018e0261 671179cfd13e2cb0dc00fee4
## 2  57bb0ebfd4654c00018e0261T2 57bb0ebfd4654c00018e0261 671179cfd13e2cb0dc00fee4
## 3  57bb0ebfd4654c00018e0261T3 57bb0ebfd4654c00018e0261 671179cfd13e2cb0dc00fee4
## 4  57bb0ebfd4654c00018e0261T4 57bb0ebfd4654c00018e0261 671179cfd13e2cb0dc00fee4
## 5  5ab848ffe1546900019b6ec9T1 5ab848ffe1546900019b6ec9 671793db2e378b0de8b1321d
## 6  5ab848ffe1546900019b6ec9T2 5ab848ffe1546900019b6ec9 671793db2e378b0de8b1321d
## 7  5ab848ffe1546900019b6ec9T3 5ab848ffe1546900019b6ec9 671793db2e378b0de8b1321d
## 8  5ab848ffe1546900019b6ec9T4 5ab848ffe1546900019b6ec9 671793db2e378b0de8b1321d
## 9  5c131126d6d169000148414aT1 5c131126d6d169000148414a 67152faa5a48814cd9e7a281
## 10 5c131126d6d169000148414aT2 5c131126d6d169000148414a 67152faa5a48814cd9e7a281
##    Text1 Text2 AnswerQ1 AnswerQ2 AnswerQ3 AnswerQ4 TimeSpent TreatedIsPolarized
## 1  MR749  R749        4        5        5        2        44                 1
## 2   L167 ML167        3        3        5        3        43                -1
## 3  MR050  R050        2        5        4        2        58                 0
## 4  ML633  L633        4        4        5        2        64                 1
## 5   L211 ML211        5        2        4        1        59                 0
## 6   L891 ML891        5        2        4        1        55                 0
## 7  MR942  R942        4        5        2        2        66                 1
## 8   R528 MR528        4        2        4        1        40                 0
## 9  PL159  L159        4        4        5        1        59                 1
## 10  L482 PL482        4        4        4        3        21                 1
##    OriginalIsPolarized TreatedIsLessPolar TreatedLikertValue
## 1                    1                   1                  4
## 2                   -1                   0                  3
## 3                    1                   1                  2
## 4                    1                   1                  4
## 5                    1                   1                  2
## 6                    1                   1                  2
## 7                    1                   1                  4
## 8                    1                   1                  2
## 9                    1                   0                  4
## 10                   1                   0                  4
##    OriginalLikertValue DiffLikertTreatedOriginal TweetBias ParticipantLeaning
## 1                    5                       -1     Right             center
## 2                    3                        0      Left             center
## 3                    5                       -3     Right             center
## 4                    4                        0      Left             center
## 5                    5                       -3      Left             center
## 6                    5                       -3      Left             center
```

```
## 7                   5                       -1     Right             center
## 8                   4                       -2     Right             center
## 9                   4                        0      Left        center-left
## 10                  4                        0      Left        center-left
##    TreatmentGroup
## 1         machine
## 2         machine
## 3         machine
## 4         machine
## 5         machine
## 6         machine
## 7         machine
## 8         machine
## 9         placebo
## 10        placebo
```

General Result (all treatments)

```
m_general = glmer(TreatedIsLessPolar ~ TreatmentGroup + (1 | FK_ParticipantId),
             data=survey_data, family = "binomial")
summary(m_general)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: TreatedIsLessPolar ~ TreatmentGroup + (1 | FK_ParticipantId)
##    Data: survey_data
##
##      AIC      BIC   logLik deviance df.resid
##    296.3    311.0   -144.2    288.3      288
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.6291 -0.5876  0.2708  0.3928  1.7020
##
## Random effects:
##  Groups           Name        Variance Std.Dev.
##  FK_ParticipantId (Intercept) 0.8849   0.9407
## Number of obs: 292, groups:  FK_ParticipantId, 73
##
## Fixed effects:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)             1.3956     0.3451   4.044 5.25e-05 ***
## TreatmentGrouphuman     0.9753     0.5114   1.907   0.0565 .
## TreatmentGroupplacebo  -2.4357     0.4878  -4.994 5.92e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) TrtmntGrph
## TrtmntGrphm -0.554
## TrtmntGrppl -0.759  0.378
```

```r
report(m_general)
```

```
## We fitted a logistic mixed model (estimated using ML and Nelder-Mead optimizer)
## to predict TreatedIsLessPolar with TreatmentGroup (formula: TreatedIsLessPolar
## ~ TreatmentGroup). The model included FK_ParticipantId as random effect
## (formula: ~1 | FK_ParticipantId). The model's total explanatory power is
## substantial (conditional R2 = 0.47) and the part related to the fixed effects
## alone (marginal R2) is of 0.33. The model's intercept, corresponding to
## TreatmentGroup = machine, is at 1.40 (95% CI [0.72, 2.07], p < .001). Within
## this model:
##
##   - The effect of TreatmentGroup [human] is statistically non-significant and
## positive (beta = 0.98, 95% CI [-0.03, 1.98], p = 0.057; Std. beta = 0.98, 95%
## CI [-0.03, 1.98])
##   - The effect of TreatmentGroup [placebo] is statistically significant and
## negative (beta = -2.44, 95% CI [-3.39, -1.48], p < .001; Std. beta = -2.44, 95%
## CI [-3.39, -1.48])
##
## Standardized parameters were obtained by fitting the model on a standardized
## version of the dataset. 95% Confidence Intervals (CIs) and p-values were
## computed using a Wald z-distribution approximation.
```

Model Interpretation

Model Overview

- **Dependent Variable (TreatedIsLessPolar)**: A binary indicator of whether the treated text is perceived as less polarized than the original text.
- **Predictor (TreatmentGroup)**: Three treatment groups: **machine paraphrasing** (reference category), **human paraphrasing**, and **placebo**.
- **Random Effects**:
  - A random intercept for each participant (FK_ParticipantId) accounts for individual variability in polarization perceptions.

Key Metrics

1. **AIC**: 296.3, **BIC**: 311.0, **Log-Likelihood**: -144.2, **Deviance**: 288.3, **df.resid**: 288
   - Lower AIC and BIC values indicate a better model fit relative to alternative models.
2. **Conditional R^2**: 0.47, representing the variance explained by both fixed and random effects.
3. **Marginal R^2**: 0.33, representing the variance explained by the fixed effects alone.

Random Effects

- **Variance of Participant-Level Random Intercept**: 0.8849, with a standard deviation of 0.9407.

- This indicates moderate variability in participants' baseline differences in the perception of polarization between original and treated texts.

Fixed Effects

1. Intercept:

- **Estimate**: 1.3956
- **Interpretation**: The mean log-odds of a treated text being perceived as less polarized compared to the original text, when the treatment is **machine paraphrasing** (reference category), is **1.40**.
  - This positive value indicates a higher likelihood of the treated text being seen as less polarized than the original.
- **Significance**: Highly significant ($p < 0.001$).

2. TreatmentGrouphuman:

- **Estimate**: 0.9753
- **Interpretation**: Compared to **machine paraphrasing**, the log-odds of a treated text being seen as less polarized when the treatment is **human paraphrasing** increase by **0.98**.
  - This effect is **not statistically significant** ($p = 0.057$), suggesting a positive but marginally non-significant difference between human and machine paraphrasing.

3. TreatmentGroupplacebo:

- **Estimate**: -2.4357
- **Interpretation**: Compared to **machine paraphrasing**, the log-odds of a treated text being seen as less polarized when the treatment is **placebo** decrease by **-2.44**.
  - This effect is **highly significant** ($p < 0.001$), indicating that the placebo treatment significantly decreases the likelihood of the treated text being perceived as less polarized.

Confidence Intervals

- The **95% Confidence Interval** for each fixed effect provides the range of plausible values for the parameter estimates:
  - Intercept: [0.72, 2.07] – consistently positive, indicating a strong likelihood that machine paraphrasing is seen as less polarized than the original.
  - TreatmentGrouphuman: [-0.03, 1.98] – includes zero, confirming the non-significance of the effect.
  - TreatmentGroupplacebo: [-3.39, -1.48] – consistently negative, confirming the strong negative effect of the placebo.

Correlation of Fixed Effects

- The correlation between the intercept and **TreatmentGrouphuman** is -0.554, indicating a moderate negative relationship.
- The correlation between the intercept and **TreatmentGroupplacebo** is -0.759, showing a stronger negative relationship.

Summary of Findings

1. **Effectiveness of Treatments**:
   - **Machine paraphrasing** is the reference category and shows a strong likelihood of being perceived as less polarized.
   - **Human paraphrasing** marginally increases the likelihood of being perceived as less polarized compared to machine paraphrasing, but this effect is not statistically significant.
   - **Placebo treatment** significantly decreases the likelihood of being perceived as less polarized.
2. **Participant-Level Variability**:
   - There is moderate variability in how participants perceive the reduction in polarization across different treatments.
3. **Model Fit**:
   - The model explains **33%** of the variance in polarization perceptions based on fixed effects alone (marginal R^2), while it explains **47%** of the variance when accounting for both fixed and random effects (conditional R^2).

RQ1 Can LLMs mitigate textual polarization in social media texts?

Logistic Regression for mitigation effect.

```
machine_placebo <- subset(survey_data, TreatmentGroup %in% c("machine", "placebo"))
model_rq1 = glmer(TreatedIsLessPolar ~ TreatmentGroup + (1 | FK_ParticipantId),
              data=machine_placebo, family = "binomial")
summary(model_rq1)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: TreatedIsLessPolar ~ TreatmentGroup + (1 | FK_ParticipantId)
##    Data: machine_placebo
##
##      AIC      BIC   logLik deviance df.resid
##    226.7    236.6   -110.4    220.7      193
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.8988 -0.5867  0.3896  0.5266  1.7045
##
## Random effects:
##  Groups           Name        Variance Std.Dev.
```

```
##  FK_ParticipantId (Intercept) 0.9159   0.957
## Number of obs: 196, groups:  FK_ParticipantId, 49
##
## Fixed effects:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)             1.4029     0.3538   3.965 7.33e-05 ***
## TreatmentGroupplacebo  -2.4472     0.5019  -4.876 1.08e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr)
## TrtmntGrppl -0.768
```

```
report(model_rq1)
```

```
## We fitted a logistic mixed model (estimated using ML and Nelder-Mead optimizer)
## to predict TreatedIsLessPolar with TreatmentGroup (formula: TreatedIsLessPolar
## ~ TreatmentGroup). The model included FK_ParticipantId as random effect
## (formula: ~1 | FK_ParticipantId). The model's total explanatory power is
## substantial (conditional R2 = 0.42) and the part related to the fixed effects
## alone (marginal R2) is of 0.26. The model's intercept, corresponding to
## TreatmentGroup = machine, is at 1.40 (95% CI [0.71, 2.10], p < .001). Within
## this model:
##
##   - The effect of TreatmentGroup [placebo] is statistically significant and
## negative (beta = -2.45, 95% CI [-3.43, -1.46], p < .001; Std. beta = -2.45, 95%
## CI [-3.43, -1.46])
##
## Standardized parameters were obtained by fitting the model on a standardized
## version of the dataset. 95% Confidence Intervals (CIs) and p-values were
## computed using a Wald z-distribution approximation.
```

Model Interpretation

Model Overview

- **Dependent Variable (TreatedIsLessPolar)**: A binary indicator of whether the treated text is perceived as less polarized than the original text.
- **Predictor (TreatmentGroup)**: Two treatment groups: **machine paraphrasing** (reference category) and **placebo**.
- **Random Effects**:
  - A random intercept for each participant (FK_ParticipantId) accounts for individual variability in polarization perceptions.

Key Metrics

1. **AIC**: 226.7, **BIC**: 236.6, **Log-Likelihood**: -110.4, **Deviance**: 220.7, **df.resid**: 193
   - Lower AIC and BIC values indicate a better model fit relative to alternative models.

2. **Conditional R^2**: 0.42, representing the variance explained by both fixed and random effects.
3. **Marginal R^2**: 0.26, representing the variance explained by the fixed effects alone.

Random Effects

- **Variance of Participant-Level Random Intercept**: 0.9159, with a standard deviation of 0.957.
  - This indicates moderate variability in participants' baseline differences in the perception of polarization between original and treated texts.

Fixed Effects

1. Intercept:

- **Estimate**: 1.4029
- **Interpretation**: The mean log-odds of a treated text being perceived as less polarized compared to the original text, when the treatment is **machine paraphrasing** (reference category), is **1.40**.
  - This positive value indicates a higher likelihood of the treated text being seen as less polarized than the original.
- **Significance**: Highly significant ($p < 0.001$).

2. TreatmentGroupplacebo:

- **Estimate**: -2.4472
- **Interpretation**: Compared to **machine paraphrasing**, the log-odds of a treated text being seen as less polarized when the treatment is **placebo** decrease by **-2.45**.
  - This effect is **highly significant** ($p < 0.001$), indicating that the placebo treatment significantly decreases the likelihood of the treated text being perceived as less polarized.

Confidence Intervals

- The **95% Confidence Interval** for each fixed effect provides the range of plausible values for the parameter estimates:
  - Intercept: [0.71, 2.10] – consistently positive, indicating a strong likelihood that machine paraphrasing is seen as less polarized than the original.
  - TreatmentGroupplacebo: [-3.43, -1.46] – consistently negative, confirming the strong negative effect of the placebo.

Correlation of Fixed Effects

- The correlation between the intercept and **TreatmentGroupplacebo** is -0.768, indicating a moderate negative relationship.

Summary of Findings

1. **Effectiveness of Treatments**:
   - **Machine paraphrasing** (reference category) shows a strong likelihood of being perceived as less polarized than the original text.
   - **Placebo treatment** significantly decreases the likelihood of the treated text being perceived as less polarized compared to machine paraphrasing.
2. **Participant-Level Variability**:
   - There is moderate variability in how participants perceive the reduction in polarization across different treatments.
3. **Model Fit**:
   - The model explains **26%** of the variance in polarization perceptions based on fixed effects alone (marginal R^2), while it explains **42%** of the variance when accounting for both fixed and random effects (conditional R^2).

RQ2 Can LLMs significantly reduce perceived polarization in social media texts?

```
#model_rq2 <- lmer(DiffLikertTreatedOriginal ~ TreatmentGroup + TweetBias * ParticipantLeaning +
#                   (1 | FK_ParticipantId),
#                   data = machine_placebo)
model_rq2 <- lmer(DiffLikertTreatedOriginal ~ TreatmentGroup +(1 | FK_ParticipantId),
                  data = machine_placebo)
summary(model_rq2)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: DiffLikertTreatedOriginal ~ TreatmentGroup + (1 | FK_ParticipantId)
##    Data: machine_placebo
##
## REML criterion at convergence: 613.6
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.40382 -0.50549  0.07692  0.45227  2.67130
##
## Random effects:
##  Groups           Name        Variance Std.Dev.
##  FK_ParticipantId (Intercept) 0.3284   0.5731
##  Residual                     1.0901   1.0441
```

```
## Number of obs: 196, groups:  FK_ParticipantId, 49
##
## Fixed effects:
##                       Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)            -1.7400     0.1550 47.0000 -11.223 6.80e-15 ***
## TreatmentGroupplacebo   1.5629     0.2215 47.0000   7.055 6.75e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr)
## TrtmntGrppl -0.700
```

```
report(model_rq2)
```

```
## We fitted a linear mixed model (estimated using REML and nloptwrap optimizer)
## to predict DiffLikertTreatedOriginal with TreatmentGroup (formula:
## DiffLikertTreatedOriginal ~ TreatmentGroup). The model included
## FK_ParticipantId as random effect (formula: ~1 | FK_ParticipantId). The model's
## total explanatory power is substantial (conditional R2 = 0.46) and the part
## related to the fixed effects alone (marginal R2) is of 0.30. The model's
## intercept, corresponding to TreatmentGroup = machine, is at -1.74 (95% CI
## [-2.05, -1.43], t(192) = -11.22, p < .001). Within this model:
##
##   - The effect of TreatmentGroup [placebo] is statistically significant and
## positive (beta = 1.56, 95% CI [1.13, 2.00], t(192) = 7.05, p < .001; Std. beta
## = 1.10, 95% CI [0.79, 1.41])
##
## Standardized parameters were obtained by fitting the model on a standardized
## version of the dataset. 95% Confidence Intervals (CIs) and p-values were
## computed using a Wald t-distribution approximation.
```

Model Interpretation

Model Overview

- **Dependent Variable (DiffLikertTreatedOriginal)**: The difference in polarization scores between the treated and original texts, measured on a Likert scale.
- **Predictor (TreatmentGroup)**: Two treatment groups: **machine paraphrasing** (reference category) and **placebo**.
- **Random Effects**:
  - A random intercept for each participant (FK_ParticipantId) accounts for individual variability in score differences.

Key Metrics

1. **REML Criterion**: 613.6. A lower REML value suggests a better model fit when comparing similar models.

2. **Residual Standard Deviation**: 1.0441, indicating the average deviation of observed values from predicted values after accounting for fixed and random effects.
3. **R^2 Values**:
    - **Conditional R^2**: 0.46, representing the variance explained by both fixed and random effects.
    - **Marginal R^2**: 0.30, representing the variance explained by the fixed effects alone.

Random Effects

- **Variance of Participant-Level Random Intercept**: 0.3284, with a standard deviation of 0.5731.
    - This indicates moderate variability in participants' baseline differences in polarization scores.
- **Residual Variance**: 1.0901, with a standard deviation of 1.0441.

Fixed Effects

1. Intercept:

- **Estimate**: -1.7400
- **Interpretation**: When the treatment group is **machine paraphrasing** (reference category), the mean difference in Likert scale polarization scores is **-1.74**.
    - This negative value indicates that machine paraphrasing significantly reduces polarization scores compared to the original texts.
- **Significance**: Highly significant ($p < 0.001$).

2. TreatmentGroupplacebo:

- **Estimate**: 1.5629
- **Interpretation**: Compared to **machine paraphrasing**, the mean difference in polarization scores increases by **1.56** when the treatment is **placebo**.
    - This effect is **highly significant** ($p < 0.001$), indicating that the placebo treatment significantly increases the perception of polarization compared to machine paraphrasing.

Confidence Intervals

- The **95% Confidence Interval** for each fixed effect provides the range of plausible values for the parameter estimates:
    - Intercept: [-2.05, -1.43] – consistently negative, indicating a robust reduction in polarization scores for machine paraphrasing.
    - TreatmentGroupplacebo: [1.13, 2.00] – consistently positive, confirming the strong effect of the placebo in increasing polarization.

Correlation of Fixed Effects

- The correlation between the intercept and **TreatmentGroupplacebo** is -0.700, indicating a moderate negative relationship.

Summary of Findings

1. **Effectiveness of Treatments**:
   - **Machine paraphrasing** significantly reduces polarization scores, with a mean reduction of **1.74 points** on the Likert scale.
   - **Placebo treatment** significantly increases the perception of polarization compared to machine paraphrasing, with a mean increase of **1.56 points**.
2. **Participant-Level Variability**:
   - There is moderate variability in baseline score differences across participants, as indicated by the random effects.
3. **Model Fit**:
   - Fixed effects explain **30%** of the variance in polarization score differences (marginal R^2), while the full model explains **46%** (conditional R^2), suggesting substantial explanatory power.

RQ3 Can LLMs mitigate textual polarization as good as humans?

```
# Compare LLM vs. Human
machine_human <- subset(survey_data, TreatmentGroup %in% c("machine", "human"))
model_rq3 <- lmer(DiffLikertTreatedOriginal ~ TreatmentGroup + (1 | FK_ParticipantId),
                  data = survey_data)
summary(model_rq3)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: DiffLikertTreatedOriginal ~ TreatmentGroup + (1 | FK_ParticipantId)
##    Data: survey_data
##
## REML criterion at convergence: 977.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.1557 -0.6201  0.0967  0.4724  4.6011
##
## Random effects:
##  Groups           Name        Variance Std.Dev.
##  FK_ParticipantId (Intercept) 0.1861   0.4314
##  Residual                     1.4909   1.2210
## Number of obs: 292, groups:  FK_ParticipantId, 73
##
```

```
## Fixed effects:
##                         Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)              -1.7400     0.1495 70.0000 -11.638  < 2e-16 ***
## TreatmentGrouphuman       0.1879     0.2136 70.0000   0.880    0.382
## TreatmentGroupplacebo     1.5629     0.2136 70.0000   7.316  3.3e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) TrtmntGrph
## TrtmntGrphm -0.700
## TrtmntGrppl -0.700  0.490
```

```
report(model_rq3)
```

```
## We fitted a linear mixed model (estimated using REML and nloptwrap optimizer)
## to predict DiffLikertTreatedOriginal with TreatmentGroup (formula:
## DiffLikertTreatedOriginal ~ TreatmentGroup). The model included
## FK_ParticipantId as random effect (formula: ~1 | FK_ParticipantId). The model's
## total explanatory power is substantial (conditional R2 = 0.31) and the part
## related to the fixed effects alone (marginal R2) is of 0.22. The model's
## intercept, corresponding to TreatmentGroup = machine, is at -1.74 (95% CI
## [-2.03, -1.45], t(287) = -11.64, p < .001). Within this model:
##
##   - The effect of TreatmentGroup [human] is statistically non-significant and
## positive (beta = 0.19, 95% CI [-0.23, 0.61], t(287) = 0.88, p = 0.380; Std.
## beta = 0.13, 95% CI [-0.16, 0.42])
##   - The effect of TreatmentGroup [placebo] is statistically significant and
## positive (beta = 1.56, 95% CI [1.14, 1.98], t(287) = 7.32, p < .001; Std. beta
## = 1.07, 95% CI [0.78, 1.35])
##
## Standardized parameters were obtained by fitting the model on a standardized
## version of the dataset. 95% Confidence Intervals (CIs) and p-values were
## computed using a Wald t-distribution approximation.
```

Model Interpretation

Model Overview

- **Dependent Variable (DiffLikertTreatedOriginal)**: The difference in polarization scores between the treated and original texts, measured on a Likert scale.
- **Predictor (TreatmentGroup)**: Two groups (machine paraphrasing as the reference category, human paraphrasing).
- **Random Effects**:
  - A random intercept for each participant (FK_ParticipantId) accounts for individual variability in score differences.

Key Metrics

1. **REML Criterion**: 977.6. A lower REML value suggests a better model fit when comparing similar models.

2. **Residual Standard Deviation**: 1.2210, indicating the average deviation of observed values from predicted values after accounting for fixed and random effects.
3. **R^2 Values**:
    - **Conditional R^2**: 0.31, representing the variance explained by both fixed and random effects.
    - **Marginal R^2**: 0.22, representing the variance explained by the fixed effects alone.

Random Effects

- **Variance of Participant-Level Random Intercept**: 0.1861, with a standard deviation of 0.4314.
    - This suggests some variability in participants' baseline differences in polarization scores.
- **Residual Variance**: 1.4909, with a standard deviation of 1.2210.

Fixed Effects

1. Intercept:

- **Estimate**: -1.7400
- **Interpretation**: When the treatment group is **machine paraphrasing** (the reference category), the mean difference in Likert scale polarization scores is **-1.74**.
    - This negative value indicates that machine paraphrasing significantly reduces polarization scores compared to the original texts.
- **Significance**: Highly significant ($p < 0.001$).

2. TreatmentGrouphuman:

- **Estimate**: 0.1879
- **Interpretation**: Compared to **machine paraphrasing**, the mean difference in polarization scores increases slightly (by **+0.19**) when the treatment is **human paraphrasing**.
    - However, this effect is **not statistically significant** ($p = 0.382$), suggesting no meaningful difference between the effects of human paraphrasing and machine paraphrasing.

3. TreatmentGroupplacebo:

- **Estimate**: 1.5629
- **Interpretation**: Compared to **machine paraphrasing**, the mean difference in polarization scores increases significantly (by **+1.56**) when the treatment is **placebo**.
    - This effect is **highly significant** ($p < 0.001$), indicating that the placebo treatment leads to a large increase in polarization scores compared to machine paraphrasing.

Confidence Intervals

- The **95% Confidence Interval** for each fixed effect provides the range of plausible values for the parameter estimates:
  - Intercept: [-2.03, -1.45] – consistently negative, indicating a robust reduction in polarization scores for machine paraphrasing.
  - TreatmentGrouphuman: [-0.23, 0.61] – includes zero, confirming the non-significance of the effect.
  - TreatmentGroupplacebo: [1.14, 1.98] – consistently positive, indicating a robust increase in polarization scores for placebo.

Correlation of Fixed Effects

- The correlation between the intercept and TreatmentGrouphuman is -0.700, indicating a moderate negative relationship.
- The correlation between the intercept and TreatmentGroupplacebo is -0.700, also indicating a moderate negative relationship.

Summary of Findings

1. **Effectiveness of Treatments**:
   - **Machine paraphrasing** significantly reduces polarization scores with a mean reduction of **-1.74 points** on the Likert scale.
   - **Human paraphrasing** has a slightly positive but **non-significant** effect on polarization scores (an increase of **+0.19 points**).
   - **Placebo** treatment leads to a **significant** and substantial increase in polarization scores (an increase of **+1.56 points**).
2. **Participant-Level Variability**:
   - There is **some variability** in baseline score differences across participants, as indicated by the random effects.
3. **Model Fit**:
   - The fixed effects explain **22%** of the variance in polarization score differences (marginal R^2), while the full model explains **31%** (conditional R^2), suggesting moderate explanatory power.

RQ4 Does political bias influence the participants' perception of textual polarization?

```
model_rq4 <- lmer(OriginalLikertValue ~ TweetBias * ParticipantLeaning +
                  (1 | FK_ParticipantId), data = survey_data)
summary(model_rq4)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula:
## OriginalLikertValue ~ TweetBias * ParticipantLeaning + (1 | FK_ParticipantId)
##    Data: survey_data
##
## REML criterion at convergence: 791.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.6646 -0.4442  0.3565  0.6655  1.9919
##
## Random effects:
##  Groups           Name        Variance Std.Dev.
##  FK_ParticipantId (Intercept) 0.09907  0.3148
##  Residual                     0.81209  0.9012
## Number of obs: 292, groups:  FK_ParticipantId, 73
##
## Fixed effects:
##                                              Estimate Std. Error         df
## (Intercept)                                 4.188e+00  2.513e-01  1.595e+02
## TweetBiasRight                              6.250e-02  3.186e-01  2.110e+02
## ParticipantLeaningcenter-left               2.083e-01  2.901e-01  1.595e+02
## ParticipantLeaningcenter-right              1.625e-01  3.371e-01  1.595e+02
## ParticipantLeaningfar-left                 -1.188e+00  5.619e-01  1.595e+02
## ParticipantLeaningfar-right                 8.125e-01  7.538e-01  1.595e+02
## ParticipantLeaningleft                     -4.861e-02  3.020e-01  1.595e+02
## ParticipantLeaningnot informed              2.125e-01  4.052e-01  1.595e+02
## ParticipantLeaningright                    -2.875e-01  4.052e-01  1.595e+02
## TweetBiasRight:ParticipantLeaningcenter-left  -5.435e-15  3.679e-01  2.110e+02
## TweetBiasRight:ParticipantLeaningcenter-right -5.125e-01  4.275e-01  2.110e+02
## TweetBiasRight:ParticipantLeaningfar-left      1.687e+00  7.124e-01  2.110e+02
## TweetBiasRight:ParticipantLeaningfar-right    -2.063e+00  9.558e-01  2.110e+02
## TweetBiasRight:ParticipantLeaningleft          1.875e-01  3.829e-01  2.110e+02
## TweetBiasRight:ParticipantLeaningnot informed -5.625e-01  5.137e-01  2.110e+02
## TweetBiasRight:ParticipantLeaningright        -6.250e-02  5.137e-01  2.110e+02
##                                             t value Pr(>|t|)
## (Intercept)                                  16.665   <2e-16 ***
## TweetBiasRight                                0.196   0.8447
## ParticipantLeaningcenter-left                 0.718   0.4738
## ParticipantLeaningcenter-right                0.482   0.6305
## ParticipantLeaningfar-left                   -2.113   0.0361 *
## ParticipantLeaningfar-right                   1.078   0.2827
## ParticipantLeaningleft                       -0.161   0.8723
## ParticipantLeaningnot informed                0.524   0.6007
## ParticipantLeaningright                      -0.710   0.4790
## TweetBiasRight:ParticipantLeaningcenter-left  0.000   1.0000
## TweetBiasRight:ParticipantLeaningcenter-right -1.199   0.2319
## TweetBiasRight:ParticipantLeaningfar-left      2.369   0.0188 *
## TweetBiasRight:ParticipantLeaningfar-right    -2.158   0.0321 *
## TweetBiasRight:ParticipantLeaningleft          0.490   0.6249
## TweetBiasRight:ParticipantLeaningnot informed -1.095   0.2748
## TweetBiasRight:ParticipantLeaningright        -0.122   0.9033
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Correlation matrix not shown by default, as p = 16 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it
```

```
report(model_rq4)
```

```
## We fitted a linear mixed model (estimated using REML and nloptwrap optimizer)
## to predict OriginalLikertValue with TweetBias and ParticipantLeaning (formula:
## OriginalLikertValue ~ TweetBias * ParticipantLeaning). The model included
## FK_ParticipantId as random effect (formula: ~1 | FK_ParticipantId). The model's
## total explanatory power is moderate (conditional R2 = 0.18) and the part
## related to the fixed effects alone (marginal R2) is of 0.08. The model's
## intercept, corresponding to TweetBias = Left and ParticipantLeaning = center,
## is at 4.19 (95% CI [3.69, 4.68], t(274) = 16.66, p < .001). Within this model:
##
##   - The effect of TweetBias [Right] is statistically non-significant and positive
## (beta = 0.06, 95% CI [-0.56, 0.69], t(274) = 0.20, p = 0.845; Std. beta = 0.06,
## 95% CI [-0.58, 0.71])
##   - The effect of ParticipantLeaning [center-left] is statistically
## non-significant and positive (beta = 0.21, 95% CI [-0.36, 0.78], t(274) = 0.72,
## p = 0.473; Std. beta = 0.22, 95% CI [-0.38, 0.81])
##   - The effect of ParticipantLeaning [center-right] is statistically
## non-significant and positive (beta = 0.16, 95% CI [-0.50, 0.83], t(274) = 0.48,
## p = 0.630; Std. beta = 0.17, 95% CI [-0.52, 0.85])
##   - The effect of ParticipantLeaning [far-left] is statistically significant and
## negative (beta = -1.19, 95% CI [-2.29, -0.08], t(274) = -2.11, p = 0.035; Std.
## beta = -1.23, 95% CI [-2.37, -0.08])
##   - The effect of ParticipantLeaning [far-right] is statistically non-significant
## and positive (beta = 0.81, 95% CI [-0.67, 2.30], t(274) = 1.08, p = 0.282; Std.
## beta = 0.84, 95% CI [-0.69, 2.37])
##   - The effect of ParticipantLeaning [left] is statistically non-significant and
## negative (beta = -0.05, 95% CI [-0.64, 0.55], t(274) = -0.16, p = 0.872; Std.
## beta = -0.05, 95% CI [-0.66, 0.56])
##   - The effect of ParticipantLeaning [not informed] is statistically
## non-significant and positive (beta = 0.21, 95% CI [-0.59, 1.01], t(274) = 0.52,
## p = 0.600; Std. beta = 0.22, 95% CI [-0.60, 1.04])
##   - The effect of ParticipantLeaning [right] is statistically non-significant and
## negative (beta = -0.29, 95% CI [-1.09, 0.51], t(274) = -0.71, p = 0.479; Std.
## beta = -0.30, 95% CI [-1.12, 0.53])
##   - The effect of TweetBias [Right] × ParticipantLeaning [center-left] is
## statistically non-significant and negative (beta = -5.44e-15, 95% CI [-0.72,
## 0.72], t(274) = -1.48e-14, p > .999; Std. beta = -1.74e-15, 95% CI [-0.75,
## 0.75])
##   - The effect of TweetBias [Right] × ParticipantLeaning [center-right] is
## statistically non-significant and negative (beta = -0.51, 95% CI [-1.35, 0.33],
## t(274) = -1.20, p = 0.232; Std. beta = -0.53, 95% CI [-1.40, 0.34])
##   - The effect of TweetBias [Right] × ParticipantLeaning [far-left] is
## statistically significant and positive (beta = 1.69, 95% CI [0.28, 3.09],
## t(274) = 2.37, p = 0.019; Std. beta = 1.74, 95% CI [0.29, 3.19])
##   - The effect of TweetBias [Right] × ParticipantLeaning [far-right] is
```

```
## statistically significant and negative (beta = -2.06, 95% CI [-3.94, -0.18],
## t(274) = -2.16, p = 0.032; Std. beta = -2.13, 95% CI [-4.08, -0.19])
##   - The effect of TweetBias [Right] × ParticipantLeaning [left] is statistically
## non-significant and positive (beta = 0.19, 95% CI [-0.57, 0.94], t(274) = 0.49,
## p = 0.625; Std. beta = 0.19, 95% CI [-0.59, 0.97])
##   - The effect of TweetBias [Right] × ParticipantLeaning [not informed] is
## statistically non-significant and negative (beta = -0.56, 95% CI [-1.57, 0.45],
## t(274) = -1.09, p = 0.275; Std. beta = -0.58, 95% CI [-1.63, 0.46])
##   - The effect of TweetBias [Right] × ParticipantLeaning [right] is statistically
## non-significant and negative (beta = -0.06, 95% CI [-1.07, 0.95], t(274) =
## -0.12, p = 0.903; Std. beta = -0.06, 95% CI [-1.11, 0.98])
##
## Standardized parameters were obtained by fitting the model on a standardized
## version of the dataset. 95% Confidence Intervals (CIs) and p-values were
## computed using a Wald t-distribution approximation.
```

Model Interpretation

Model Overview

- **Dependent Variable (OriginalLikertValue)**: The Likert scale value representing the polarization of the tweet as assessed by participants.
- **Predictors**:
  - **TweetBias** (Right vs. Left bias in the tweet).
  - **ParticipantLeaning** (center-left, center-right, far-left, far-right, left, right, not informed).
  - Interaction between **TweetBias** and **ParticipantLeaning**.
- **Random Effects**: A random intercept for each participant (FK_ParticipantId) is included to account for individual variability in polarization scores.

Key Metrics

1. **REML Criterion**: 791.5. This value indicates model fit, with lower values suggesting better fit when comparing similar models.
2. **Residual Standard Deviation**: 0.9012, showing the average deviation of the observed values from the predicted values after accounting for fixed and random effects.
3. **R^2 Values**:
   - **Conditional R^2**: 0.18, indicating that 18% of the variance in the polarization scores is explained by both fixed and random effects.
   - **Marginal R^2**: 0.08, suggesting that the fixed effects alone explain 8% of the variance.

Random Effects

- **Variance of Participant-Level Random Intercept**: 0.09907 (SD = 0.3148), indicating small variability in baseline differences between participants' polarization scores.
- **Residual Variance**: 0.81209 (SD = 0.9012), reflecting the unexplained variability after accounting for fixed and random effects.

Fixed Effects

1. Intercept:

- **Estimate**: 4.188 (95% CI [3.69, 4.68]).
- **Interpretation**: When the tweet is left-biased and the participant leans center, the average polarization score is **4.19**, indicating a relatively neutral to moderately polarized tweet.
- **Significance**: Highly significant ($p < 0.001$).

2. Main Effects:

- **TweetBias [Right]**:
  - **Estimate**: 0.0625 (95% CI [-0.56, 0.69]).
  - **Interpretation**: The effect of tweet bias being right-wing is **non-significant** ($p = 0.845$), suggesting no difference in polarization between right- and left-biased tweets in general.
- **ParticipantLeaning**:
  - **Center-left**: **Non-significant** ($p = 0.474$), suggesting no substantial difference in polarization compared to the center group.
  - **Center-right**: **Non-significant** ($p = 0.630$), similarly showing no significant difference.
  - **Far-left**: **Significant negative effect** (beta = -1.19, $p = 0.035$), indicating that far-left participants perceive a significantly lower level of polarization compared to center participants.
  - **Far-right**: **Non-significant** ($p = 0.283$), showing no significant difference in polarization perception.
  - **Left**: **Non-significant** ($p = 0.872$), suggesting no effect.
  - **Not informed**: **Non-significant** ($p = 0.601$), with a slight positive effect, but not enough to be meaningful.
  - **Right**: **Non-significant** ($p = 0.479$), showing no substantial difference.

3. Interaction Effects (TweetBias × ParticipantLeaning):

- **Far-left × Right-Bias**:
  - **Estimate**: 1.687 (95% CI [0.28, 3.09]).
  - **Interpretation**: Far-left participants perceive a **significantly higher level of polarization** when exposed to right-biased tweets ($p = 0.019$).
- **Far-right × Right-Bias**:
  - **Estimate**: -2.063 (95% CI [-3.94, -0.18]).
  - **Interpretation**: Far-right participants perceive a **significantly lower level of polarization** when exposed to right-biased tweets ($p = 0.032$).
- **Other Interactions** (e.g., **center-left**, **center-right**, etc.): All non-significant, suggesting no meaningful interaction between tweet bias and these participant leanings.

Confidence Intervals and p-Values

- 95% Confidence Intervals (CIs) provide the range of plausible values for each parameter:
  - The **far-left × Right-bias** interaction has a positive and significant effect, with the 95% CI not including zero.

– The **far-right × Right-bias** interaction is also significant, with a negative effect and the 95% CI not including zero.

Summary of Findings

1. **Tweet Bias**:
   - The bias of the tweet (left vs. right) alone does not significantly influence the polarization score ($p = 0.845$).
2. **Participant Leaning**:
   - Participants with far-left political leanings perceive a significantly lower level of polarization, while other groups (center-left, center-right, far-right, left, right, and not informed) show no significant effects.
3. **Interaction Effects**:
   - **Far-left participants** perceive a significantly **higher polarization** in right-biased tweets.
   - **Far-right participants** perceive a significantly **lower polarization** in right-biased tweets.
4. **Model Fit**:
   - The model explains a moderate portion of the variance (18% total), with fixed effects alone explaining 8%.

Textual Cohesion

```
survey_data$IsCoherent <- ifelse(survey_data$AnswerQ3 > 3, 1, 0)
model_cohesion <- lmer(IsCoherent ~ TreatmentGroup + (1 | FK_ParticipantId),
                       data=survey_data)
summary(model_cohesion)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: IsCoherent ~ TreatmentGroup + (1 | FK_ParticipantId)
##    Data: survey_data
##
## REML criterion at convergence: 270.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.4010  0.1275  0.3060  0.4603  1.4288
##
## Random effects:
##  Groups           Name        Variance Std.Dev.
##  FK_ParticipantId (Intercept) 0.02657  0.1630
##  Residual                     0.12215  0.3495
## Number of obs: 292, groups:  FK_ParticipantId, 73
##
```

```
## Fixed effects:
##                         Estimate Std. Error       df t value Pr(>|t|)
## (Intercept)              0.80000    0.04779 70.00000  16.739   <2e-16 ***
## TreatmentGrouphuman     -0.08125    0.06829 70.00000  -1.190    0.238
## TreatmentGroupplacebo    0.11667    0.06829 70.00000   1.708    0.092 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) TrtmntGrph
## TrtmntGrphm -0.700
## TrtmntGrppl -0.700  0.490
```

```
report(model_cohesion)
```

```
## We fitted a linear mixed model (estimated using REML and nloptwrap optimizer)
## to predict IsCoherent with TreatmentGroup (formula: IsCoherent ~
## TreatmentGroup). The model included FK_ParticipantId as random effect (formula:
## ~1 | FK_ParticipantId). The model's total explanatory power is moderate
## (conditional R2 = 0.21) and the part related to the fixed effects alone
## (marginal R2) is of 0.04. The model's intercept, corresponding to
## TreatmentGroup = machine, is at 0.80 (95% CI [0.71, 0.89], t(287) = 16.74, p <
## .001). Within this model:
##
##   - The effect of TreatmentGroup [human] is statistically non-significant and
## negative (beta = -0.08, 95% CI [-0.22, 0.05], t(287) = -1.19, p = 0.235; Std.
## beta = -0.21, 95% CI [-0.55, 0.14])
##   - The effect of TreatmentGroup [placebo] is statistically non-significant and
## positive (beta = 0.12, 95% CI [-0.02, 0.25], t(287) = 1.71, p = 0.089; Std.
## beta = 0.30, 95% CI [-0.05, 0.64])
##
## Standardized parameters were obtained by fitting the model on a standardized
## version of the dataset. 95% Confidence Intervals (CIs) and p-values were
## computed using a Wald t-distribution approximation.
```

Model Interpretation

Model Overview

- **Dependent Variable (IsCoherent)**: A measure of text coherence, potentially scored on a Likert scale.
- **Predictor (TreatmentGroup)**: Three groups (machine paraphrasing as the reference category, human paraphrasing, and placebo).
- **Random Effects**:
  - A random intercept for each participant (FK_ParticipantId) accounts for individual variability in coherence ratings.

Key Metrics

1. **REML Criterion**: 270.1. A lower REML value suggests a better model fit when comparing similar models.

2. **Residual Standard Deviation**: 0.3495, indicating the average deviation of observed values from predicted values after accounting for fixed and random effects.
3. **R^2 Values**:
    - **Conditional R^2**: 0.21, representing the variance explained by both fixed and random effects.
    - **Marginal R^2**: 0.04, representing the variance explained by the fixed effects alone.

Random Effects

- **Variance of Participant-Level Random Intercept**: 0.02657, with a standard deviation of 0.1630.
    - This suggests low variability in participants' baseline coherence ratings.
- **Residual Variance**: 0.12215, with a standard deviation of 0.3495.

Fixed Effects

1. Intercept:

- **Estimate**: 0.8000
- **Interpretation**: When the treatment group is **machine paraphrasing** (the reference category), the mean coherence rating is **0.80**.
    - This high positive value suggests that machine paraphrasing leads to substantial perceived coherence.
- **Significance**: Highly significant ($p < 0.001$).

2. TreatmentGrouphuman:

- **Estimate**: -0.0813
- **Interpretation**: Compared to **machine paraphrasing**, the mean coherence rating decreases slightly (by **-0.08**) when the treatment is **human paraphrasing**.
    - This effect is **not statistically significant** ($p = 0.238$), suggesting no meaningful difference in coherence between machine and human paraphrasing.

3. TreatmentGroupplacebo:

- **Estimate**: 0.1167
- **Interpretation**: Compared to **machine paraphrasing**, the mean coherence rating increases slightly (by **+0.12**) when the treatment is **placebo**.
    - This effect is **marginally significant** ($p = 0.092$), suggesting a potential but inconclusive improvement in coherence for the placebo group.

Confidence Intervals

- The **95% Confidence Interval** for each fixed effect provides the range of plausible values for the parameter estimates:
  - Intercept: [0.71, 0.89] – consistently positive, indicating robust coherence ratings for machine paraphrasing.
  - TreatmentGrouphuman: [-0.22, 0.05] – includes zero, confirming the non-significance of the effect.
  - TreatmentGroupplacebo: [-0.02, 0.25] – barely excludes zero, supporting the marginal significance of the placebo effect.

Correlation of Fixed Effects

- The correlation between the intercept and TreatmentGrouphuman is -0.700, indicating a moderate negative relationship.
- The correlation between the intercept and TreatmentGroupplacebo is -0.700, also indicating a moderate negative relationship.

Summary of Findings

1. **Effectiveness of Treatments**:
   - **Machine paraphrasing** results in high coherence ratings (**0.80 points**) and serves as the benchmark for comparisons.
   - **Human paraphrasing** shows a slight decrease in coherence ratings compared to machine paraphrasing, but this effect is not statistically significant (**-0.08 points**).
   - **Placebo** treatment shows a slight increase in coherence ratings compared to machine paraphrasing (**+0.12 points**), but the effect is only marginally significant.
2. **Participant-Level Variability**:
   - There is low variability in baseline coherence ratings across participants, as indicated by the random effects.
3. **Model Fit**:
   - The fixed effects explain **4%** of the variance in coherence ratings (marginal R^2), while the full model explains **21%** (conditional R^2), suggesting moderate explanatory power.