Pairwise Cosine Similarity Analysis

run the calculation

In this analysis, we calculated the pairwise cosine similarity between all possible response pairings within participants. Then, we ran a t-test to see if children's internal similarity scores differed from adults'.

```
participant_data <- self_responses_ids
vector_start <- 11
vector_end <- 778

participant_similarity_metrics <- participant_data %>%
  group_by(subject_id) %>%
  group_split() %>%
  map_dfr(function(df_embeddings) {

    # count skips
    total_skips <- sum(df_embeddings$is_skip)

    # get only real responses
    real_responses <- df_embeddings %>% filter(!is_skip)

    similarity_score <- NA
    # diversity <- NA

    if(nrow(real_responses) >= 2) {

      # extract the embedding columns
      vectors <- as.matrix(real_responses[, vector_start:vector_end])

  # compute cosine similarity matrix ------------------------------------

      # normalize rows -- to isolate direction, ignore length of response
      norm_vecs <- vectors / sqrt(rowSums(vectors^2))

      # dot product for the similarity matrix -- within participant comparison
      sim_matrix <- norm_vecs %*% t(norm_vecs)

      # remove self-similarity (the diagonal of 1s)
      diag(sim_matrix) <- NA

  # compute average cosine similarity -----------------------------------

      # avg semantic similarity per participant
      similarity_score <- mean(sim_matrix, na.rm = TRUE)

      # # diversity = 1 - average similarity (avg pairwise cosine distance)
      # diversity <- 1 - mean(sim_matrix, na.rm = TRUE)
```

```
    }

    data.frame(
      subject_id = df_embeddings$subject_id[1],
      age_group = first(df_embeddings$age_group),
      # age = first(df_embeddings$age_at_test),  --- for full child sample, when looking for age effect
      skip_count = total_skips,
      semantic_similarity = similarity_score
      # semantic_diversity = diversity
    )
  })

print(participant_similarity_metrics)
# write.csv(participant_similarity_metrics, "participant_similarity_metrics.csv", row.names = F)
```

moment of truth...

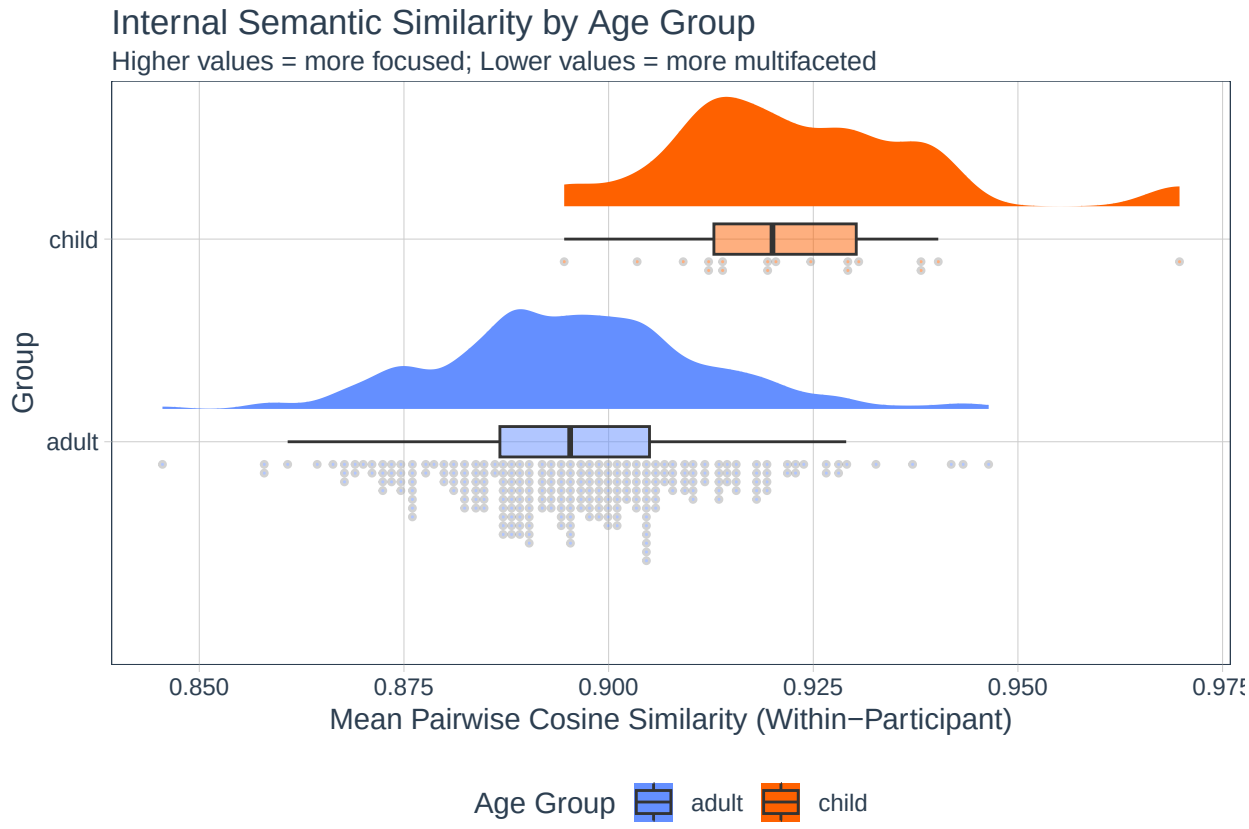
```
t_test_sim <- t.test(semantic_similarity ~ age_group,
                     data = participant_similarity_metrics)

print(t_test_sim)
```

```
##
##  Welch Two Sample t-test
##
## data:  semantic_similarity by age_group
## t = -6.6366, df = 19.253, p-value = 2.23e-06
## alternative hypothesis: true difference in means between group adult and group child is not equal to
## 95 percent confidence interval:
##  -0.03599776 -0.01874779
## sample estimates:
## mean in group adult mean in group child
##           0.8958869           0.9232597
```

We assessed internal semantic cohesion by calculating the mean pairwise cosine similarity between all disclosures for each participant. A Welch's t-test revealed that children exhibited significantly higher internal similarity ($M = 0.923$) compared to adults ($M = 0.896$), $t(19.25) = -6.64, p < .001$. This suggests that children maintain a more semantically focused self-narrative, whereas adults cover a more diverse set of semantic categories.

```
participant_similarity_metrics %>%
  filter(!is.na(semantic_similarity)) %>%
  ggplot(aes(x = age_group, y = semantic_similarity, fill = age_group)) +
    # rain (individual participants)
    stat_dots(side = "left",
              justification = 1.1,
              binwidth = .001,
              alpha = 0.5) +
    # cloud (distribution shape)
    stat_halfeye(adjust = .5,
                 width = .6,
```

```
                justification = -.3,
                point_interval = NULL) +
    # boxplot (summary stats)
    geom_boxplot(width = .15,
                outlier.shape = NA,
                alpha = 0.5) +
    # styling ------------------------------------------------------------------
    scale_fill_manual(values = c("adult" = "#648FFF", "child" = "#FE6100")) +
    theme_tq() +
    labs(
      title = "Internal Semantic Similarity by Age Group",
      subtitle = "Higher values = more focused; Lower values = more multifaceted",
      fill = "Age Group",
      x = "Group",
      y = "Mean Pairwise Cosine Similarity (Within-Participant)") +
    coord_flip()
```



robustness check. . .

```
set.seed(13126) # for reproducibility

obs_child_mean <- 0.9232597
adult_subsample_size <- 18
```

```r
n_iterations <- 10000

# bootstrap: pick n adults (adult_subsample_size), calculate their mean similarity, repeat for n_iterat

null_distribution <- replicate(n_iterations, {
  participant_similarity_metrics %>%
    filter(age_group == "adult") %>%
    sample_n(adult_subsample_size) %>%
    summarise(m = mean(semantic_similarity, na.rm = TRUE)) %>%
    pull(m)
})

# calculate empirical p-value
# how many times did a random group of 18 adults look as 'focused' as the children?
p_perm <- sum(null_distribution >= obs_child_mean) / n_iterations

cat("Mean of 18-adult samples:", mean(null_distribution), "\n")
```

```
## Mean of 18-adult samples: 0.8958797
```

```r
cat("Actual Child Mean:", obs_child_mean, "\n")
```

```
## Actual Child Mean: 0.9232597
```

```r
cat("Empirical p-value:", p_perm, "\n")
```

```
## Empirical p-value: 0
```

To ensure that the observed difference in internal similarity was not an artifact of the unequal sample sizes (N_adult = 250, N_child = 18), we conducted a permutation test with 10,000 iterations. In each iteration, we randomly sampled 18 adults and calculated their mean internal cosine similarity. The actual child mean ($M = 0.923$) was significantly higher than every single one of the 10,000 bootstrapped adult samples ($M_{boot} = 0.896, p < .0001$), confirming that children's self-disclosures are uniquely characterized by higher semantic cohesion.