sequential_analysis

Sequential Distance Analysis

In this analysis, we first calculated the Sequential Distances within participants' responses and then compared the average group distances. Also, in case anyone is confused as Karla was: cosine distance is the inverse of cosine similarity (mind blown)

```r
knitr::opts_chunk$set(echo = TRUE)

# load libraries
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.1     v stringr   1.5.2
## v ggplot2   4.0.0     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(readr)
library(lme4)
```

```
## Loading required package: Matrix
##
## Attaching package: 'Matrix'
##
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```r
library(lmerTest)
```

```
##
## Attaching package: 'lmerTest'
##
## The following object is masked from 'package:lme4':
##
##     lmer
##
## The following object is masked from 'package:stats':
##
##     step
```

```
library(performance)

# import data
self_responses_ids <- read_csv("data/self_responses_ids.csv")
```

```
## Rows: 3228 Columns: 778
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (3): subject_id, response_clean, age_group
## dbl (774): utterance, word_count, typicality, umap_1, umap_2, id_texts, Dim1...
## lgl   (1): is_skip
##
## 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.
```

within-individuals calculation

First, we calculated the cosine distance between all consecutive responses within individuals. **Formula: cosine distance = 1 — cosine similarity** Where cosine similarity is the dot product (of the 2 vectors) divided by the product of the magnitudes (of the 2 vectors)

```
# subtract the cosine similarity (between each response and the one immediately preceding it) from 1.
sequential_metrics <- self_responses_ids %>%

  # calculate within participant
  group_by(subject_id) %>%

  # exclude skipped responses
  filter(!is_skip) %>%

  # order responses temporally
  arrange(utterance) %>%

  # (wrangling for the calculation) shift vectors down by 1 row within each person
  mutate(across(starts_with("Dim"),
              ~ lag(.),
              .names = "prev_{.col}")) %>%

  # remove the first row per person (since it has no 'previous' response)
  filter(!is.na(prev_Dim1_texts)) %>%
  rowwise() %>% # for each row...
  mutate(

    # extract current vector (Dim1_texts to Dim768_texts)
    curr_vec = list(c_across(starts_with("Dim") & ends_with("_texts"))),

    # extract previous vector (prev_Dim1_texts to prev_Dim768_texts)
    prev_vec = list(c_across(starts_with("prev_Dim") & ends_with("_texts"))),

    # calculate cosine distance: 1 - cosine similarity
    step_distance = 1 - (sum(curr_vec * prev_vec) /                       # numerator: dot product
                        (sqrt(sum(curr_vec^2)) * sqrt(sum(prev_vec^2)))) # denominator: magnitude, i.e
  ) %>%
  ungroup()

# # approx. 11 values per participant
# length(sequential_metrics$step_distance)
```

Then, we ran a linear mixed model to determine whether age group influences the size of an individual’s semantic leaps. Specifically, we modeled sequential semantic distance as a function of age group, with a random intercept for subject identity to control for individual differences in baseline disclosure style.

```
model_sequential <- lmer(step_distance ~ age_group + (1 | subject_id), data = sequential_metrics)
summary(model_sequential)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite’s method [
## lmerModLmerTest]
## Formula: step_distance ~ age_group + (1 | subject_id)
##    Data: sequential_metrics
##
## REML criterion at convergence: -11783.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.8153 -0.6530 -0.0366  0.5803  6.1049
##
## Random effects:
##  Groups     Name        Variance  Std.Dev.
##  subject_id (Intercept) 0.0001991 0.01411
##  Residual               0.0009066 0.03011
## Number of obs: 2910, groups:  subject_id, 267
##
## Fixed effects:
##                   Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)      0.097133   0.001063 263.154379  91.341  < 2e-16 ***
## age_groupchild  -0.028449   0.004214 287.968784  -6.751 8.07e-11 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Correlation of Fixed Effects:
##             (Intr)
## age_grpchld -0.252
```

This model revealed a significant main effect of age group, beta = -0.028, SE = 0.004, t(287.97) = -6.75, p < .001. That is, we found quantitative evidence that children’s sequential semantic distance was significantly lower than that of adults, suggesting that children navigate a more semantically constrained self-concept, characterized by smaller conceptual transitions between self-disclosures.

The random effect of subject_id accounted for a portion of the variance in disclosure distance, which (perhaps) justifies the use of a mixed-effects approach over a standard linear regression. KARLA FIGURE OUT HOW TO JUSTIFY??

```
sequential_effect_sizes <- model_performance(model_sequential)
print(sequential_effect_sizes)
```

```
## # Indices of model performance
##
## AIC      |     AICc |      BIC | R2 (cond.) | R2 (marg.) |   ICC |  RMSE | Sigma
## --------------------------------------------------------------------------------
## -11775.7 | -11775.7 | -11751.8 |      0.212 |      0.039 | 0.180 | 0.029 | 0.030
```

To assess the effect size of age group on sequential semantic distance, we calculated the marginal and conditional R^2 values for the mixed-effects model. The fixed effect of age group explained a significant portion of the total variance (Marginal R^2 = .039), while the full model—including random intercepts for individual participants—accounted for 21.2% of the variance (Conditional R^2 = .212).

The Intraclass Correlation Coefficient (ICC) was .180, indicating that approximately 18% of the variance in semantic 'leaps' was attributable to stable individual differences between participants, which further justifies our use of a multilevel modeling approach.

For additional clarification (because I had trouble wrapping my head around ICC): Our ICC of .180 suggests that while a significant portion of the variance was nested within individuals, the vast majority of the variance (82%) was not tied to stable individual differences. This provided sufficient statistical room to identify a robust effect of age group on sequential semantic distance, supporting our view of developmental stage as a primary driver of the structural dynamics of self-disclosure.

Ok, then I did exactly the same thing but controlling for verbosity (word count per response).

```r
# scale word count (since step distances are very small)
sequential_metrics <- sequential_metrics %>%
  mutate(log_wc_scaled = scale(log(word_count)))

model_sequential_verbosity <- lmer(step_distance ~ age_group + log_wc_scaled + (1 | subject_id), data =
summary(model_sequential_verbosity)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: step_distance ~ age_group + log_wc_scaled + (1 | subject_id)
##    Data: sequential_metrics
##
## REML criterion at convergence: -11819.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.7004 -0.6553 -0.0281  0.5891  6.1407
##
## Random effects:
##  Groups     Name        Variance  Std.Dev.
##  subject_id (Intercept) 0.0001841 0.01357
##  Residual               0.0008956 0.02993
## Number of obs: 2910, groups:  subject_id, 267
##
## Fixed effects:
##                  Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)     9.714e-02  1.033e-03  2.632e+02  94.064  < 2e-16 ***
## age_groupchild -2.856e-02  4.096e-03  2.887e+02  -6.971 2.14e-11 ***
## log_wc_scaled   5.108e-03  7.265e-04  1.789e+03   7.031 2.91e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Correlation of Fixed Effects:
##             (Intr) ag_grp
## age_grpchld -0.252       
## log_wc_scld  0.000 -0.003
```

To ensure that the observed developmental differences in semantic trajectory were not a mere artifact of verbal productivity, we re-ran the mixed-effects model including a scaled log word-count co-variate. Word count was a significant predictor of step distance (beta = 0.005, SE = 0.001, t(1789) = 7.03, p < .001), reflecting the expected relationship between utterance length and semantic breadth. Crucially, however, the main effect of age group remained significant and robust (beta = -0.029, SE = 0.004, t(288.7) = -6.97, p < .001), showing that children navigated a more constrained conceptual space even when controlling for the number of words produced.

```r
model_performance(model_sequential_verbosity)
```

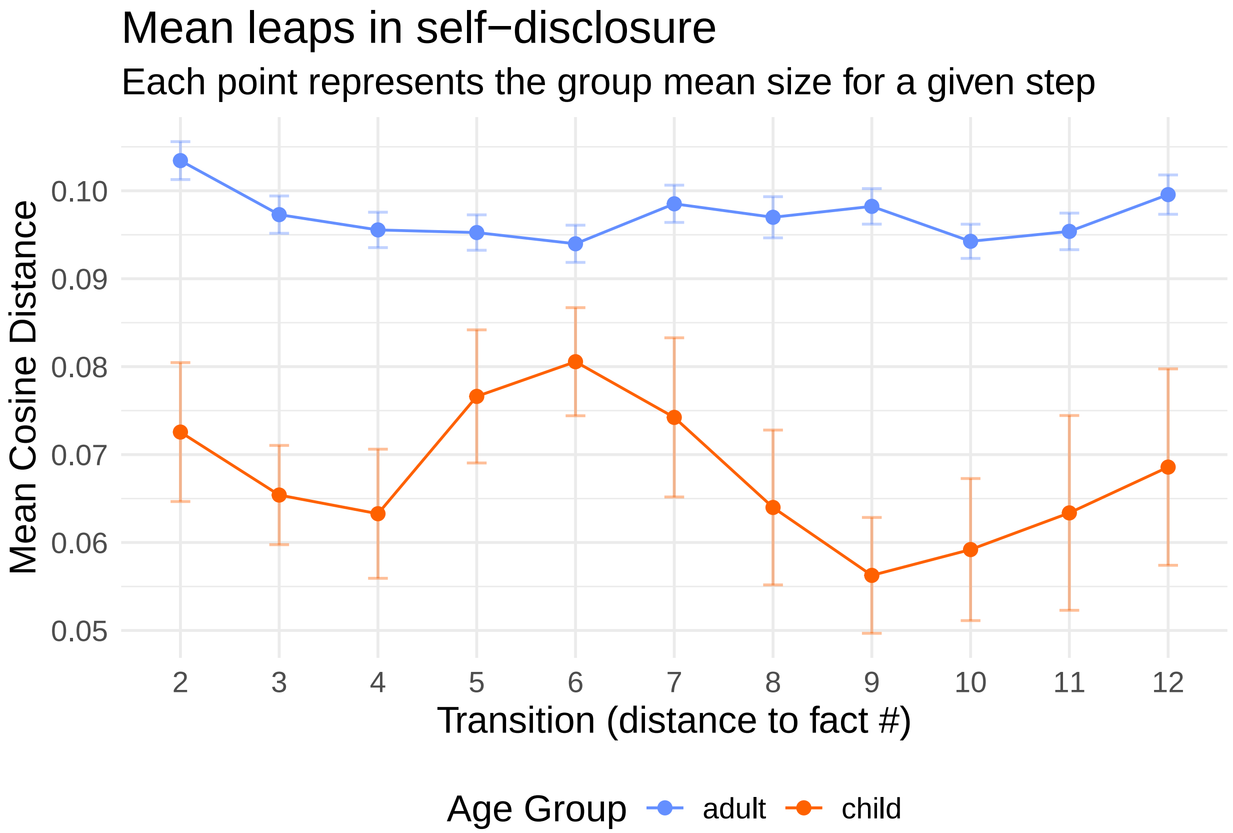
```
## # Indices of model performance
## 
## AIC      |      AICc |       BIC | R2 (cond.) | R2 (marg.) |   ICC |  RMSE | Sigma
## --------------------------------------------------------------------------------
## -11809.8 | -11809.8 | -11780.0 |      0.222 |      0.062 | 0.170 | 0.029 | 0.030
```

The inclusion of word count as a covariate improved the overall model fit (Marginal R^2 = .050, Conditional R^2 = .213); The fixed effects (age group and word count) accounted for 5% of the total variance in semantic step distance. Even after accounting for individual differences between participants (ICC = .172), the developmental stage remained a significant and meaningful predictor of the dynamic structure of self-disclosure.

Visualizations

```r
# group summary statistics for cosine distances
group_trajectory_stats <- sequential_metrics %>%
  group_by(age_group, utterance) %>%
  summarise(
    mean_dist = mean(step_distance, na.rm = TRUE),
    se_dist = sd(step_distance, na.rm = TRUE) / sqrt(n()),
    .groups = "drop"
  )

# visualize mean leaps per step, for each group
ggplot(group_trajectory_stats, aes(x = factor(utterance), y = mean_dist, color = age_group)) +
  geom_errorbar(aes(ymin = mean_dist - se_dist, ymax = mean_dist + se_dist), width = 0.2, alpha = 0.4)
  geom_point(size = 2) +
  geom_line(aes(group = age_group)) +
  theme_minimal() +
  scale_color_manual(values = c("adult" = "#648FFF", "child" = "#FE6100")) +
  labs(
    title = "Mean leaps in self-disclosure",
    subtitle = "Each point represents the group mean size for a given step",
    x = "Transition (distance to fact #)",
    y = "Mean Cosine Distance",
    color = "Age Group"
  ) +
  theme(legend.position = "bottom", text = element_text(size = 14))
```



To better illustrate these group-level trends, we selected a representative child and adult whose step distance scores were closest to the mean values for their respective groups. This ensures that the qualitative examples reflect the central tendency of each age group rather than statistical outliers. (but actually maybe I will cherry pick, so change what I said before to a nice way of saying I cherry picked)

```
sequential_metrics %>%
  group_by(subject_id, age_group) %>%
  summarise(mean_distance = mean(step_distance, na.rm = TRUE), .groups = "drop") %>%
  group_by(age_group) %>%
  arrange(age_group, desc(mean_distance)) %>%
  slice_max(order_by = mean_distance, n = 5)
```

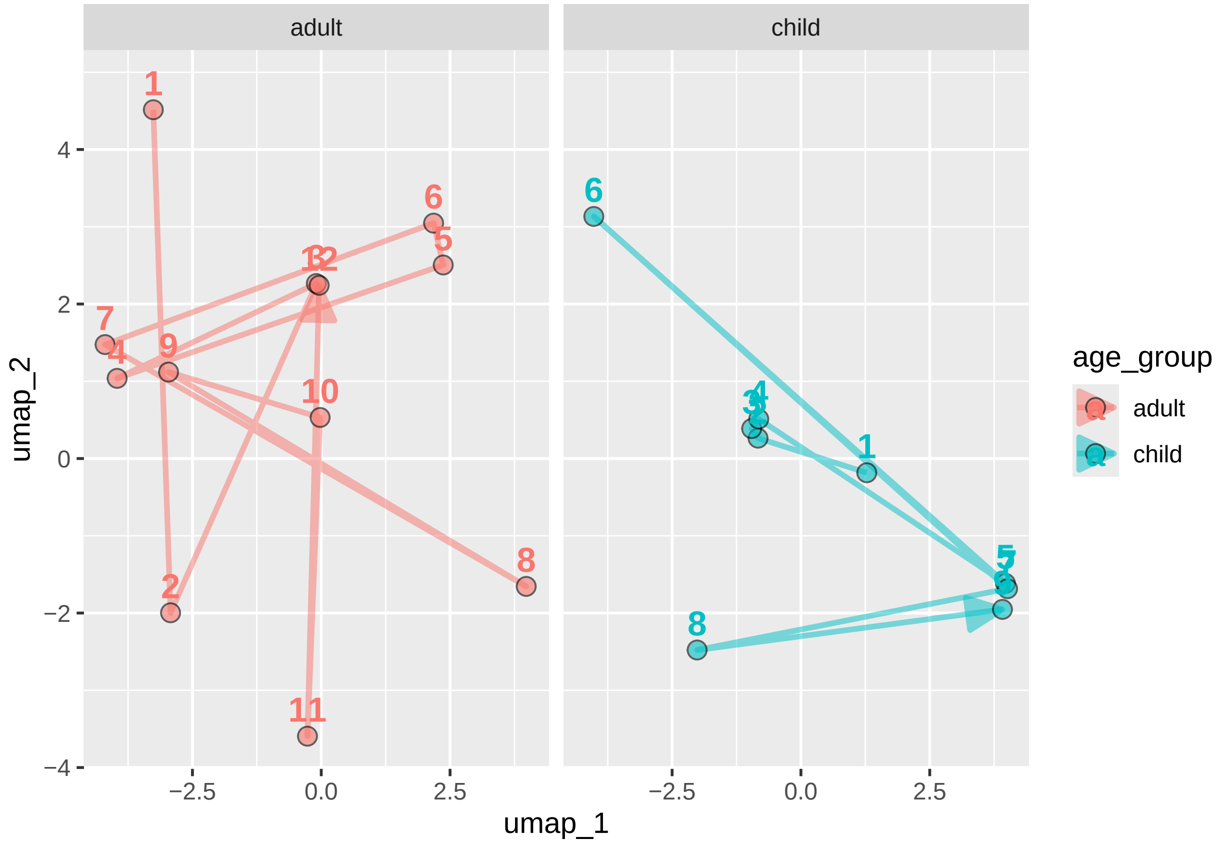
```
## # A tibble: 10 x 3
## # Groups:   age_group [2]
##    subject_id age_group mean_distance
##    <chr>      <chr>             <dbl>
##  1 28c92b42   adult             0.138
##  2 370455dc   adult             0.134
##  3 597b3c4d   adult             0.133
##  4 bd077583   adult             0.132
##  5 f825dc06   adult             0.131
##  6 04769bc3   child             0.107
##  7 73f24da8   child             0.0876
##  8 024a0cf3   child             0.0872
##  9 9209bd73   child             0.0800
## 10 b59594b6   child             0.0798
```

```
adult_cherry <- "28c92b42"
child_cherry <- "04769bc3"

cherry_picked_data <- self_responses_ids %>%
  filter(subject_id %in% c(child_cherry, adult_cherry), !is_skip) %>%
  arrange(subject_id, utterance)

ggplot(cherry_picked_data, aes(x = umap_1, y = umap_2, color = age_group)) +
  geom_path(aes(group = subject_id),
            arrow = arrow(angle = 25, type = "closed", length = unit(0.20, "inches"), ends = "last"),
            alpha = 0.5, size = 1,
            linetype = "solid") +
  geom_point(aes(fill = age_group), size = 3, shape = 21, color = "black", alpha = .6) +
  geom_text(aes(label = utterance), nudge_y = 0.35, fontface = "bold", size = 4.5) +
  facet_wrap(~ age_group)
```

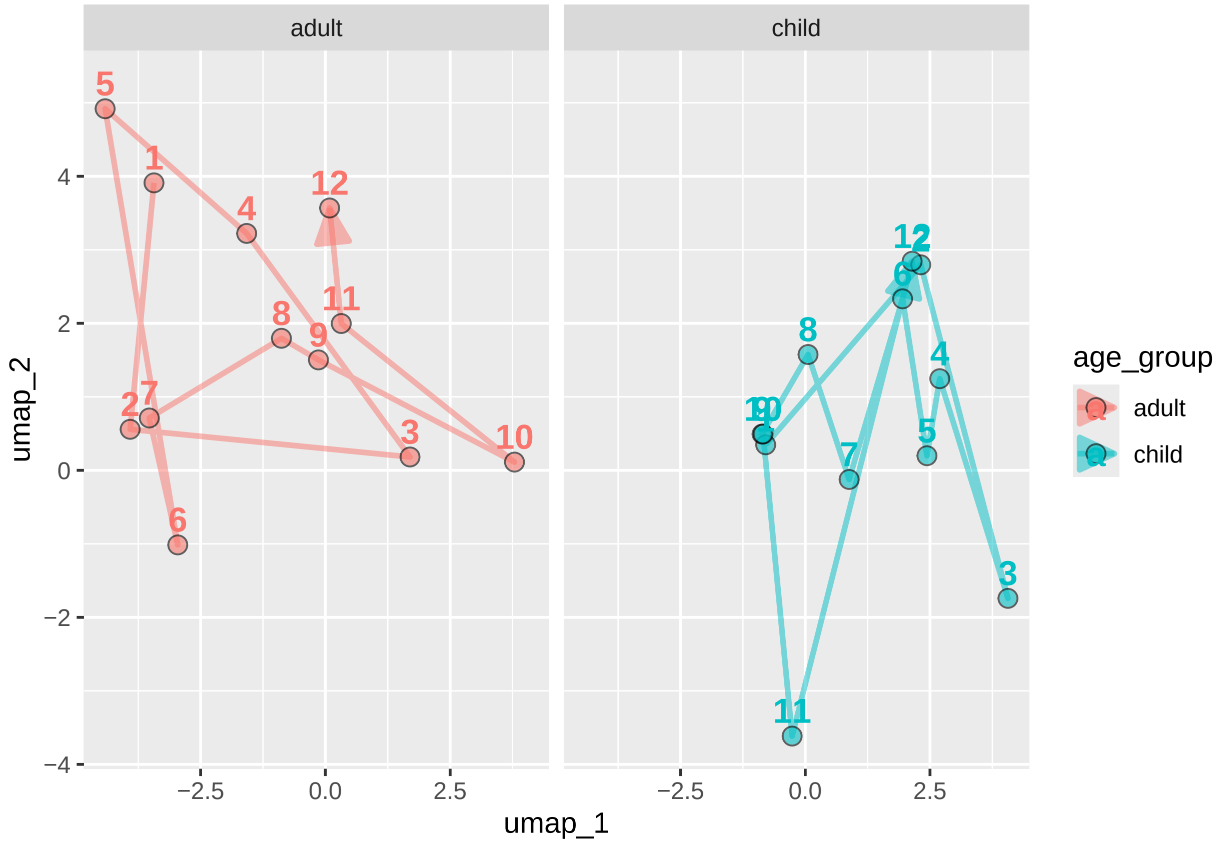
```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



```r
adult_cherry <- "370455dc"
child_cherry <- "73f24da8"

cherry_picked_data <- self_responses_ids %>%
  filter(subject_id %in% c(child_cherry, adult_cherry), !is_skip) %>%
  arrange(subject_id, utterance)

ggplot(cherry_picked_data, aes(x = umap_1, y = umap_2, color = age_group)) +
  geom_path(aes(group = subject_id),
            arrow = arrow(angle = 25, type = "closed", length = unit(0.20, "inches"), ends = "last"),
            alpha = 0.5, size = 1,
            linetype = "solid") +
  geom_point(aes(fill = age_group), size = 3, shape = 21, color = "black", alpha = .6) +
  geom_text(aes(label = utterance), nudge_y = 0.35, fontface = "bold", size = 4.5) +
  facet_wrap(~ age_group)
```

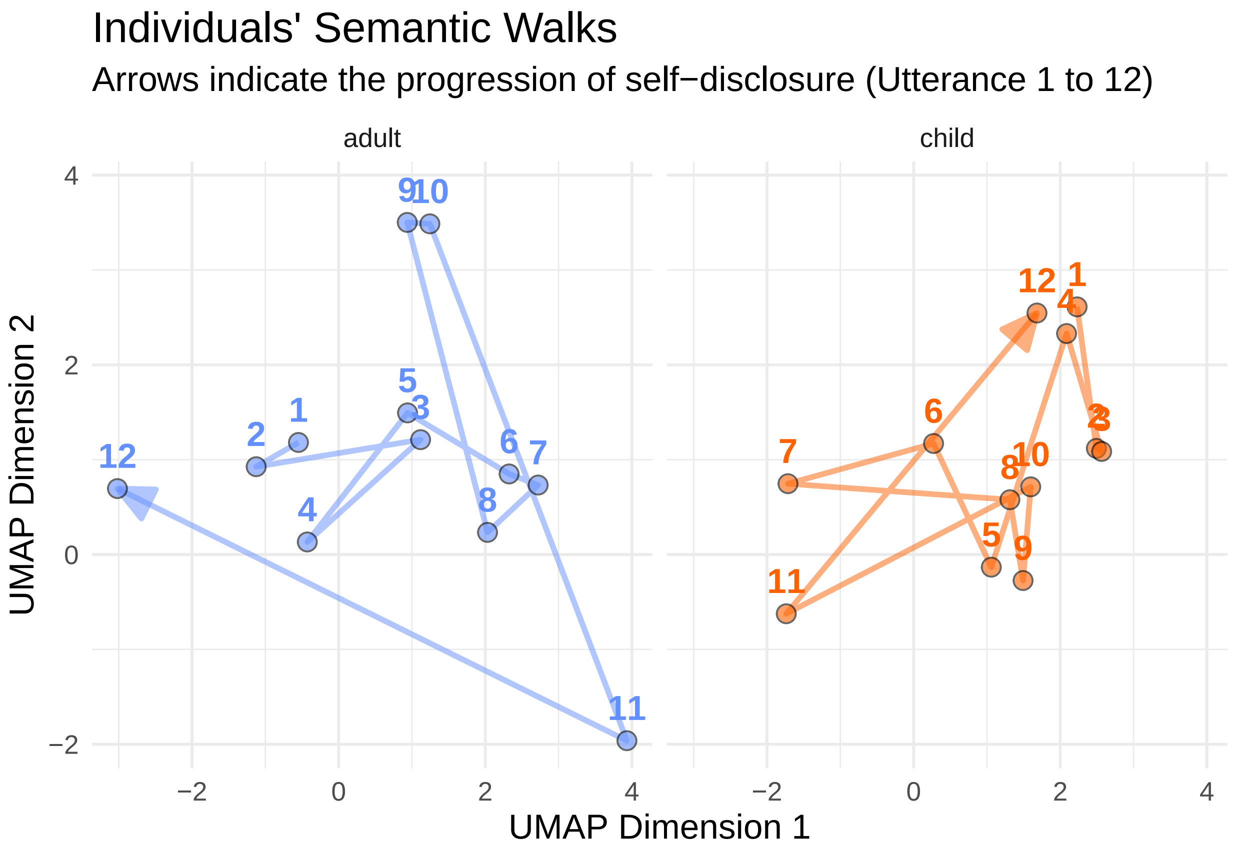


```r
# ggplot(data.frame(x=c(1,2,3), y=c(1,2,1)), aes(x,y)) +
#   geom_path(size=5, lineend="round", linejoin="round", linemitre=2, linetype = "dotted")

# child_id <- "cb867663"
#child_id <- "f1f018e8"
child_id <- "794d7026"
adult_id <- "144dbb8b"

rep_walk_data <- self_responses_ids %>%
  filter(subject_id %in% c(child_id, adult_id), !is_skip) %>%
  arrange(subject_id, utterance)

ggplot(rep_walk_data, aes(x = umap_1, y = umap_2, color = age_group)) +
  geom_path(aes(group = subject_id),
            arrow = arrow(angle = 25, type = "closed", length = unit(0.20, "inches"), ends = "last"),
            alpha = 0.5, size = 1,
            linetype = "solid") +
  geom_point(aes(fill = age_group), size = 3, shape = 21, color = "black", alpha = .6) +
  geom_text(aes(label = utterance), nudge_y = 0.35, fontface = "bold", size = 4.5) +
  facet_wrap(~ age_group) +
  scale_color_manual(values = c("adult" = "#648FFF", "child" = "#FE6100")) +
  scale_fill_manual(values = c("adult" = "#648FFF", "child" = "#FE6100")) +
  theme_minimal() +
  labs(
    title = "Individuals' Semantic Walks",
    subtitle = "Arrows indicate the progression of self-disclosure (Utterance 1 to 12)",
    x = "UMAP Dimension 1",
    y = "UMAP Dimension 2"
  ) +
  theme(legend.position = "none", text = element_text(size = 13))
```



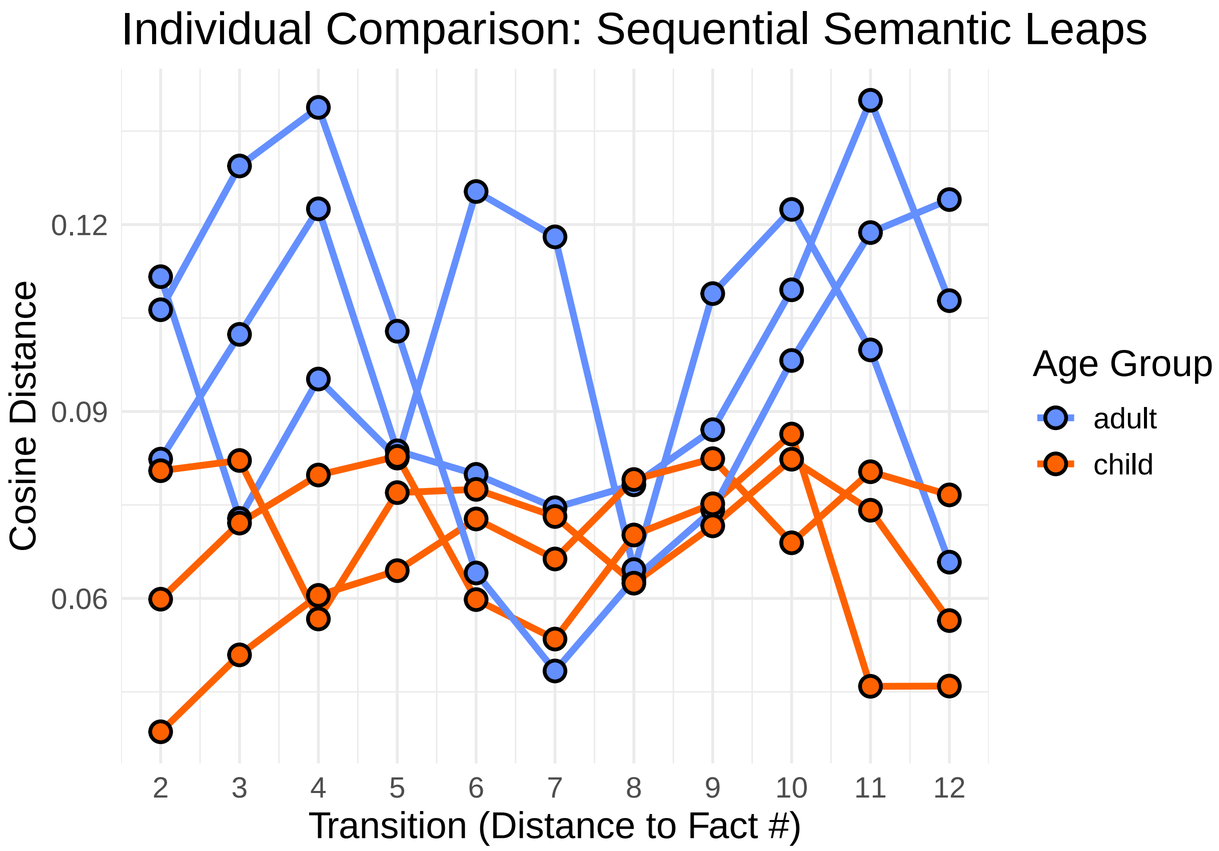
```
# multiple adults in one graph
# multiple children in one graph
```

Finally, we graphed the distances in each step that our two exemplar participants took across all of their respective 12 responses.

```r
child_id_3 <- c("794d7026", "cb867663","615ba795")
adult_id_3 <- c("144dbb8b", "f618d226", "52264f55")

representative_participants <- sequential_metrics %>%
  filter(subject_id %in% c(child_id_3, adult_id_3))

ggplot(representative_participants, aes(x = utterance, y = step_distance, group = subject_id)) +
  geom_line(aes(color = age_group), size = 1.2) +
  geom_point(aes(fill = age_group), size = 3, shape = 21, color = "black", stroke = 1) +
  scale_color_manual(values = c("adult" = "#648FFF", "child" = "#FE6100")) +
  scale_fill_manual(values = c("adult" = "#648FFF", "child" = "#FE6100")) +
  theme_minimal() +
  scale_x_continuous(breaks = 2:12) +
  labs(title = "Individual Comparison: Sequential Semantic Leaps",
       x = "Transition (Distance to Fact #)",
       y = "Cosine Distance",
       color = "Age Group",
       fill = "Age Group") +
  theme(text = element_text(size = 14))
```


Individual Comparison: Sequential Semantic Leaps
Cosine Distance
Transition (Distance to Fact #)
Age Group
adult
child