1. 10K

step 1: Firm-quarters from Compustat/CRSP with datadate between 2010 through 2024

step 2: Less: non-positive assets or missing net income, shares outstanding, price, or book value

step 3: Less: share codes other than 10 and 11

step 4: Less: missing CIK from WRDS linking table

step 5: Less: missing 10Q/K and PR from SEC Analytics Suite

step 6: Less: Filing or earnings date > 90 days after period-end

step 7: Less: length < 0 for RF in 10K, length < 151 for RF in 10Q, or length < 101 for MD&A

step 8: Less: top/bottom 1% of length, by fiscal year-quarter

step 9: Less: datadates during 1 Sept – 30 Nov 2022

2. 10Q

step 1: Firm-quarters from Compustat/CRSP with datadate between 2010 through 2024

step 2: Less: non-positive assets or missing net income, shares outstanding, price, or book value

step 3: Less: share codes other than 10 and 11

step 4: Less: missing CIK from WRDS linking table

step 5: Less: missing 10Q/K and PR from SEC Analytics Suite

step 6: Less: Filing or earnings date > 90 days after period-end

step 7: Less: length < 0 for RF in 10K, length < 151 for RF in 10Q, or length < 101 for MD&A

step 8: Less: top/bottom 1% of length, by fiscal year-quarter

step 9: Less: datadates during 1 Sept – 30 Nov 2022

3. 8K

step 1: Firm-quarters from Compustat/CRSP with datadate between 2010 through 2024

step 2: Less: non-positive assets or missing net income, shares outstanding, price, or book value

step 3: Less: share codes other than 10 and 11

step 4: Less: missing CIK from WRDS linking table

step 5: Less: missing 10Q/K and PR from SEC Analytics Suite, CC from Capital IQ

step 6: Less: Filing or earnings date > 90 days after period-end

step 7: Less: length < 0 for RF in 10K, length < 151 for RF in 10Q, or length < 101 for MD&A

step 8: Less: top/bottom 1% of length, by fiscal year-quarter

step 9: Less: datadates during 1 Sept – 30 Nov 2022

4. Conference Calls

step 5: Less: missing 10Q/K and PR from SEC Analytics Suite, CC from Capital IQ

step 6: Less: Filing or earnings date > 90 days after period-end

step 7: Less: length < 0 for RF in 10K, length < 151 for RF in 10Q, or length < 101 for MD&A

step 8: Less: top/bottom 1% of length, by fiscal year-quarter

```
link_cc_length_win = link_cc_length %>% group_by(fyearq, fqtr) %>% mutate(bottom = quantile(number_of_words, 0.01, na.rm = TRUE),
                                                                          top = quantile(number_of_words, 0.99, na.rm = TRUE)) %>%
                                                                  filter(number_of_words >= bottom, number_of_words <= top) %>% ungroup() %>%
                                                                  dplyr::select(-bottom, -top)

# nrow(link_cc_length_win) # 152,489 (155,655 - 152,489 = 3,166)
# length(unique(link_cc_length_win$GVKEY)) # 5,476 (5,482 - 5,476 = 6)

saveRDS(link_cc_length_win, paste0(path, "1_data_processing/data/sample_selection/final_sample/sample_cc_skip_last_step.RDS"))
```

step 9: Less: datadates during 1 Sept – 30 Nov 2022

```
linked_cc_last_step = link_cc_length_win %>% filter(datadate > "2022-11-30" | datadate < "2022-09-01")

# nrow(linked_cc_last_step) # 149,690 (152,489 - 149,690 = 2,799)
# length(unique(linked_cc_last_step$GVKEY)) # 5,471 (5,476 - 5,471 = 5)

saveRDS(linked_cc_last_step, paste0(path, "1_data_processing/data/sample_selection/final_sample/sample_cc.RDS"))
```

combine 10-K and 10-Q

```
item_1A_10k = readRDS(paste0(path, "1_data_processing/data/sample_selection/final_sample/sample_10k_item_1A.RDS"))
item_1A_10q = readRDS(paste0(path, "1_data_processing/data/sample_selection/final_sample/sample_10q_item_1A.RDS"))
item_7_10k = readRDS(paste0(path, "1_data_processing/data/sample_selection/final_sample/sample_10k_item_7.RDS"))
item_7_10q = readRDS(paste0(path, "1_data_processing/data/sample_selection/final_sample/sample_10q_item_7.RDS"))
cc = readRDS(paste0(path, "1_data_processing/data/sample_selection/final_sample/sample_cc.RDS"))
pr = readRDS(paste0(path, "1_data_processing/data/sample_selection/final_sample/sample_pr.RDS"))
```

```
item_1A_10k_skip = readRDS(paste0(path, "1_data_processing/data/sample_selection/final_sample/sample_10k_item_1A_skip_last_step.RDS"))
item_1A_10q_skip = readRDS(paste0(path, "1_data_processing/data/sample_selection/final_sample/sample_10q_item_1A_skip_last_step.RDS"))
item_7_10k_skip = readRDS(paste0(path, "1_data_processing/data/sample_selection/final_sample/sample_10k_item_7_skip_last_step.RDS"))
item_7_10q_skip = readRDS(paste0(path, "1_data_processing/data/sample_selection/final_sample/sample_10q_item_7_skip_last_step.RDS"))
cc_skip = readRDS(paste0(path, "1_data_processing/data/sample_selection/final_sample/sample_cc_skip_last_step.RDS"))
pr_skip = readRDS(paste0(path, "1_data_processing/data/sample_selection/final_sample/sample_pr_skip_last_step.RDS"))
```

```
item_1A = rbind(item_1A_10k %>% dplyr::select(id:cik) %>% mutate(fqtr = 4),
                item_1A_10q %>% dplyr::select(id:fqtr)) %>% dplyr::select(id, GVKEY, LPERMNO, datadate, fqtr, rdate, fdate, fname, cik)

# nrow(item_1A) # 105,795
# length(unique(item_1A$GVKEY)) # 7,046

saveRDS(item_1A, paste0(path, "1_data_processing/data/sample_selection/final_sample/sample_item_1A.RDS"))
```

```
item_7 = rbind(item_7_10k %>% dplyr::select(id:cik) %>% mutate(fqtr = 4),
               item_7_10q %>% dplyr::select(id:fqtr)) %>% dplyr::select(id, GVKEY, LPERMNO, datadate, fqtr, rdate, fdate, fname, cik)

# nrow(item_7) # 199,456
# length(unique(item_7$GVKEY)) # 7,310

saveRDS(item_7, paste0(path, "1_data_processing/data/sample_selection/final_sample/sample_item_7.RDS"))
```

```
cc = cc %>% dplyr::select(id, GVKEY, LPERMNO, datadate, fqtr, companyid, keydevid, mostimportantdateutc)

# nrow(cc) # 149,690
# length(unique(cc$GVKEY)) # 5,471

saveRDS(cc, paste0(path, "1_data_processing/data/sample_selection/final_sample/sample_cc.RDS"))
```

```
pr = pr %>% dplyr::select(id, GVKEY, LPERMNO, datadate, fqtr, rdate, fdate, fname, cik)

# nrow(pr) # 189,647
# length(unique(pr$GVKEY)) # 7,053

saveRDS(pr, paste0(path, "1_data_processing/data/sample_selection/final_sample/sample_pr.RDS"))
```

```
item_1A_skip = rbind(item_1A_10k_skip %>% dplyr::select(id:cik) %>% mutate(fqtr = 4),
                     item_1A_10q_skip %>% dplyr::select(id:cik, fqtr)) %>% dplyr::select(id, GVKEY, LPERMNO, datadate, fqtr, rdate, fdate, fname, cik)

# nrow(item_1A_skip) # 107,584
# length(unique(item_1A_skip$GVKEY)) # 7,049

saveRDS(item_1A_skip, paste0(path, "1_data_processing/data/sample_selection/final_sample/sample_item_1A_skip_last_step.RDS"))
```

```
item_7_skip = rbind(item_7_10k_skip %>% dplyr::select(id:cik) %>% mutate(fqtr = 4),
                    item_7_10q_skip %>% dplyr::select(id:cik, fqtr)) %>% dplyr::select(id, GVKEY, LPERMNO, datadate, fqtr, rdate, fdate, fname, cik)

# nrow(item_7_skip) # 203,203
# length(unique(item_7_skip$GVKEY)) # 7,311

saveRDS(item_7_skip, paste0(path, "1_data_processing/data/sample_selection/final_sample/sample_item_7_skip_last_step.RDS"))
```

```
cc_skip = cc_skip %>% dplyr::select(id, GVKEY, LPERMNO, datadate, fqtr, companyid, keydevid, mostimportantdateutc)

# nrow(cc_skip) # 152,489
# length(unique(cc_skip$GVKEY)) # 5,476

saveRDS(cc_skip, paste0(path, "1_data_processing/data/sample_selection/final_sample/sample_cc_skip_last_step.RDS"))
```

```
pr_skip = pr_skip %>% dplyr::select(id, GVKEY, LPERMNO, datadate, fqtr, rdate, fdate, fname, cik)

# nrow(pr_skip) # 193,194
# length(unique(pr_skip$GVKEY)) # 7,053

saveRDS(pr_skip, paste0(path, "1_data_processing/data/sample_selection/final_sample/sample_pr_skip_last_step.RDS"))
```

```
.end_time <- Sys.time()
cat("===== END TIME =====\n")
```

```
## ===== END TIME =====
```

```
print(.end_time)
```

```
## [1] "2026-02-02 21:21:16 PST"
```

```
cat("Elapsed:", difftime(.end_time, .start_time, units = "secs"), "seconds\n")
```

```
## Elapsed: 513.816 seconds
```