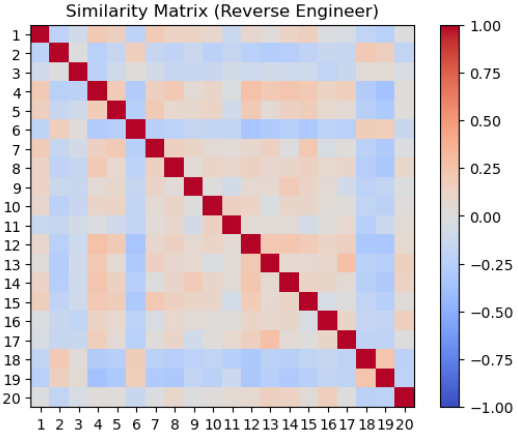
Meeting — Summary LLM

Bas Donkers, Dennis Fok, Finn-Ole Höner

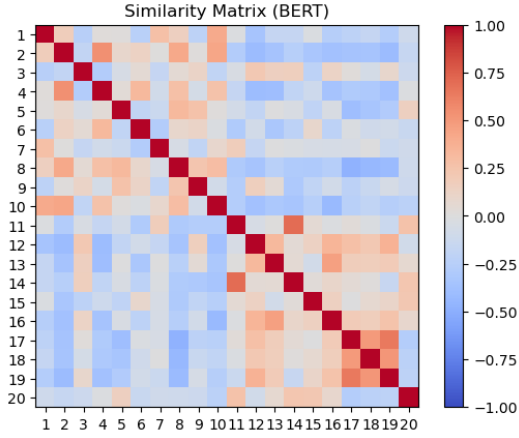
March 14, 2024

Updates

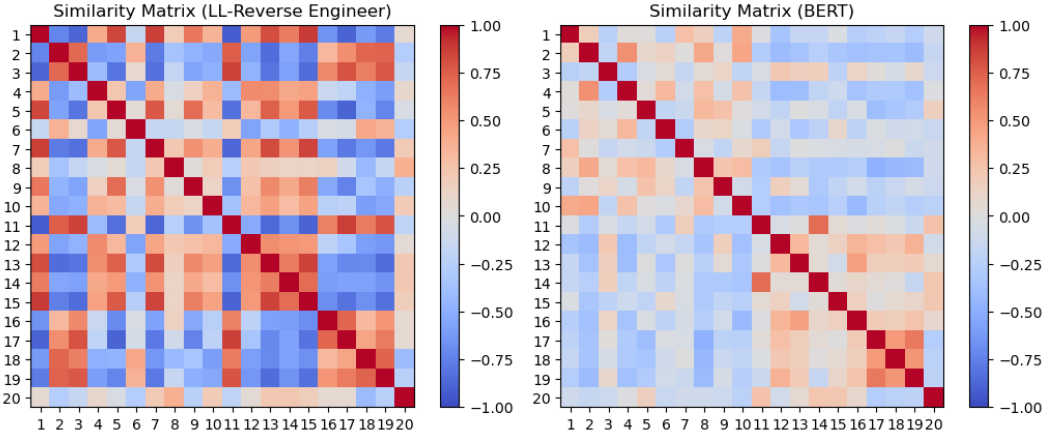
- Grant proposal handed in
- SURF Research Cloud Workshop
  - Pieter Meijndert, Yuliia Orlova
  - Small SURF Grant (quick, easy to get, uncomplicated); up to 5k GPU hours; once per calendar year
- (Start thinking about thesis comittee)
- Improving implementation of optimization still work in progress


Similarity Matrix (Reverse Engineer)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

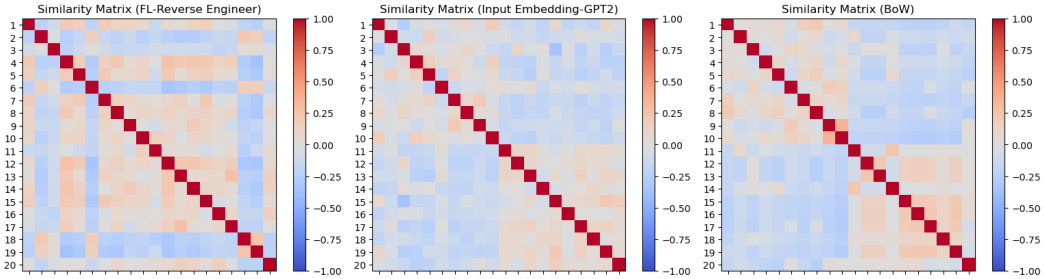
(a)


Similarity Matrix (BERT)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

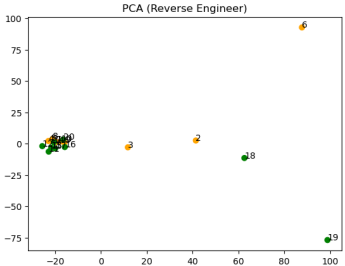
(b)


Similarity Matrix (LL-Reverse Engineer)
Similarity Matrix (BERT)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

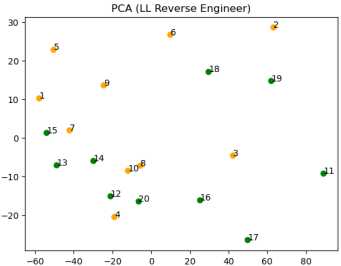
(a) **(b)**


Similarity Matrix (FL-Reverse Engineer)
Similarity Matrix (Input Embedding-GPT2)
Similarity Matrix (BoW)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
1.00 0.75 0.50 0.25 0.00 −0.25 −0.50 −0.75 −1.00

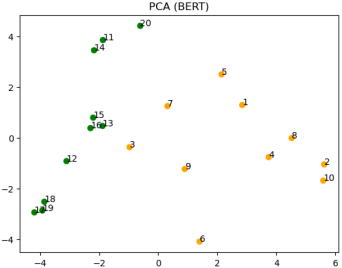
(a) **(b)** **(c)**



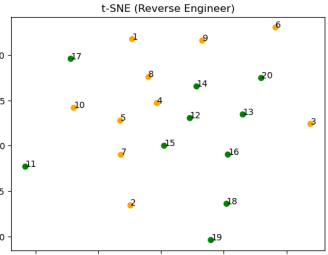
(a)



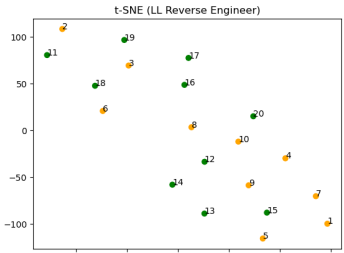
(b)



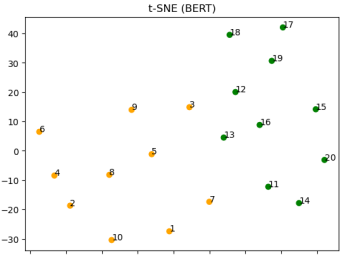
(c)



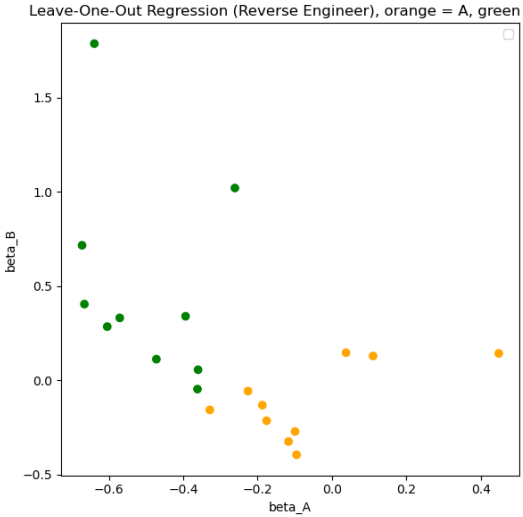
(a)



(b)



(c)


Leave-One-Out Regression (Reverse Engineer), orange = A, green = B
beta_B
beta_A
1.5
1.0
0.5
0.0
−0.5
−0.6
−0.4
−0.2
0.0
0.2
0.4

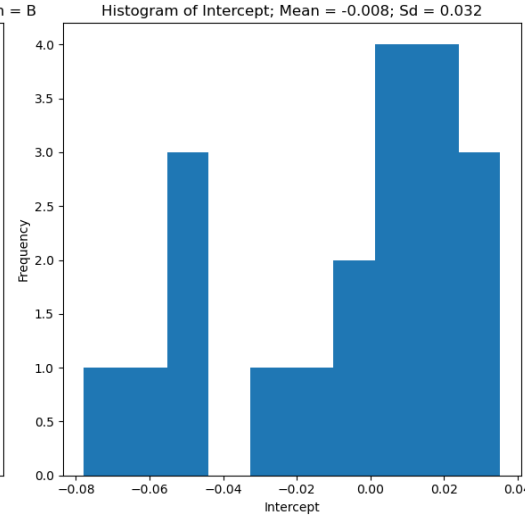

Histogram of Intercept; Mean = -0.008; Sd = 0.032
Frequency
Intercept
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.08
−0.06
−0.04
−0.02
0.00
0.02
0.04

Figure 6

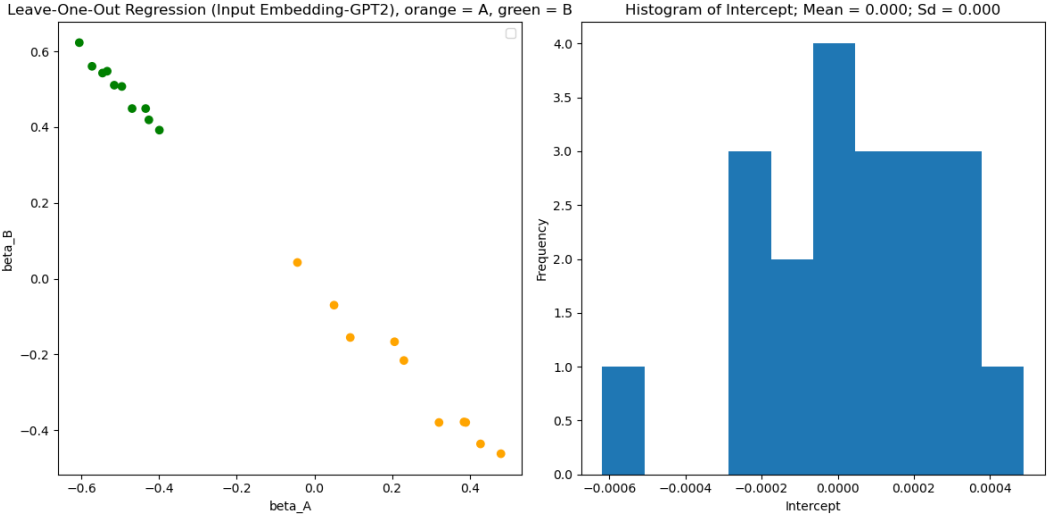
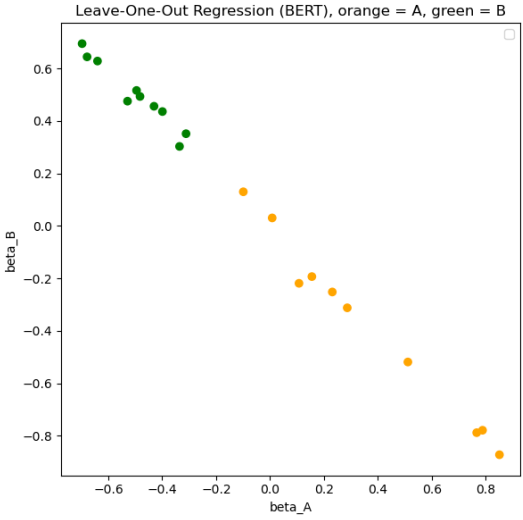

Leave-One-Out Regression (Input Embedding-GPT2), orange = A, green = B
beta_B
beta_A
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.4
−0.2
0.0
0.2
0.4
Histogram of Intercept; Mean = 0.000; Sd = 0.000
Frequency
Intercept
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.0006
−0.0004
−0.0002
0.0000
0.0002
0.0004

Figure 7


Leave-One-Out Regression (BERT), orange = A, green = B
beta_B
beta_A
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8

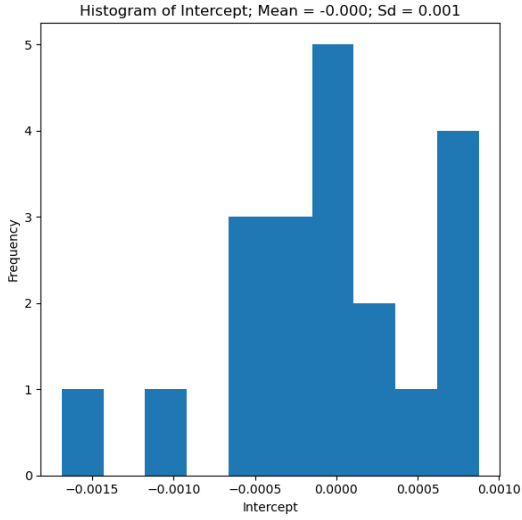
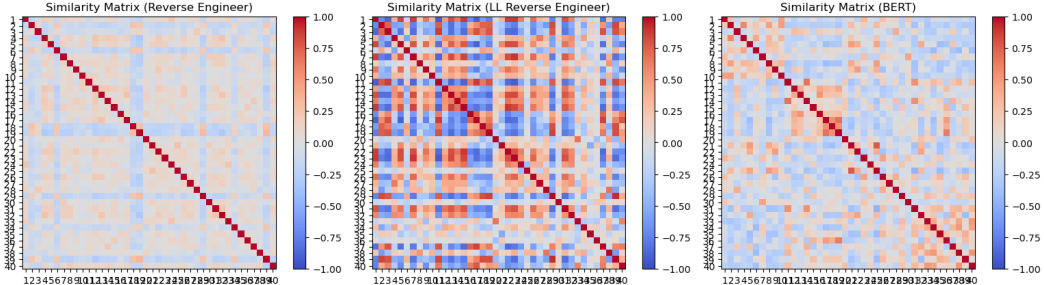
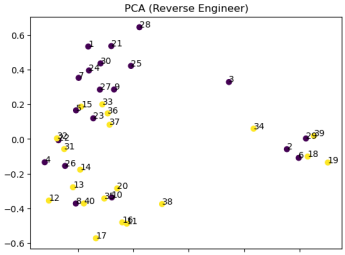

Histogram of Intercept; Mean = -0.000; Sd = 0.001
Frequency
Intercept
5
4
3
2
1
0
−0.0015
−0.0010
−0.0005
0.0000
0.0005
0.0010

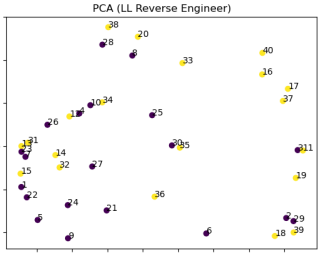
Figure 8


Similarity Matrix (Reverse Engineer)
Similarity Matrix (LL Reverse Engineer)
Similarity Matrix (BERT)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

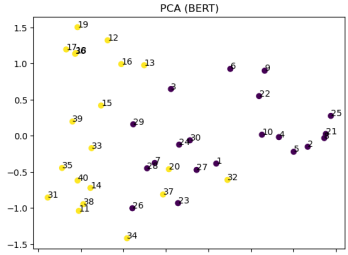
(a) **(b)** **(c)**



(a)



(b)



(c)