Results

Experiment 1- MMLU

Evaluation

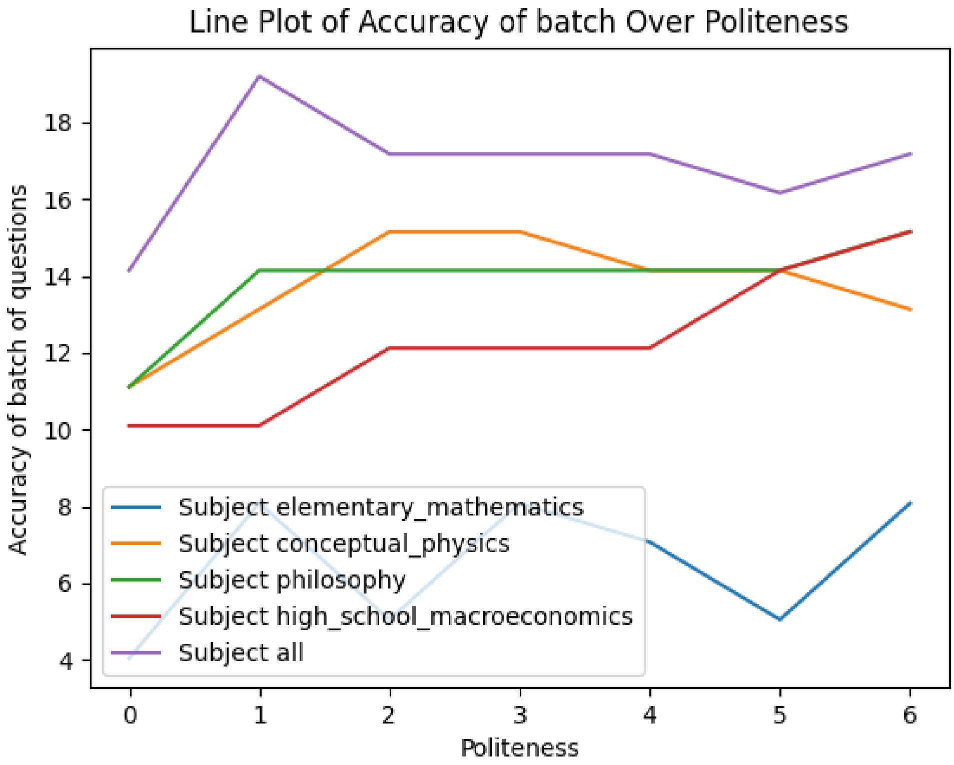
<u>English, GPT-3.5</u>

<u>Precision by batch (correct if all the questions in the batch are correct):</u>

Table form:

politeness level	1	2	3	4	5	6	7
elementary mathematics	4	8.1	5	8.1	7.1	5.1	8.1
conceptual physics	11.1	13.1	15.2	15.2	14.1	14.1	13.1
philosophy	11.1	14.1	14.1	14.1	14.1	14.1	15.1
high school macroeconomics	10.1	10.1	12.1	12.1	12.1	14.1	15.2
all	14.1	19.2	17.2	17.2	17.2	16.2	17.2

Graph form:

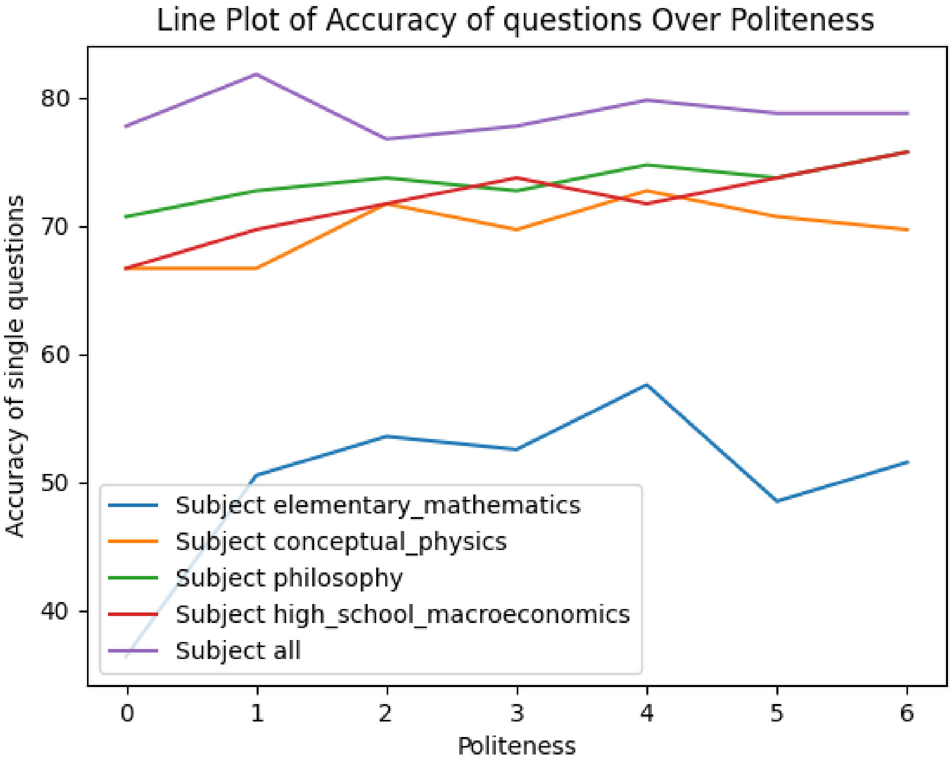


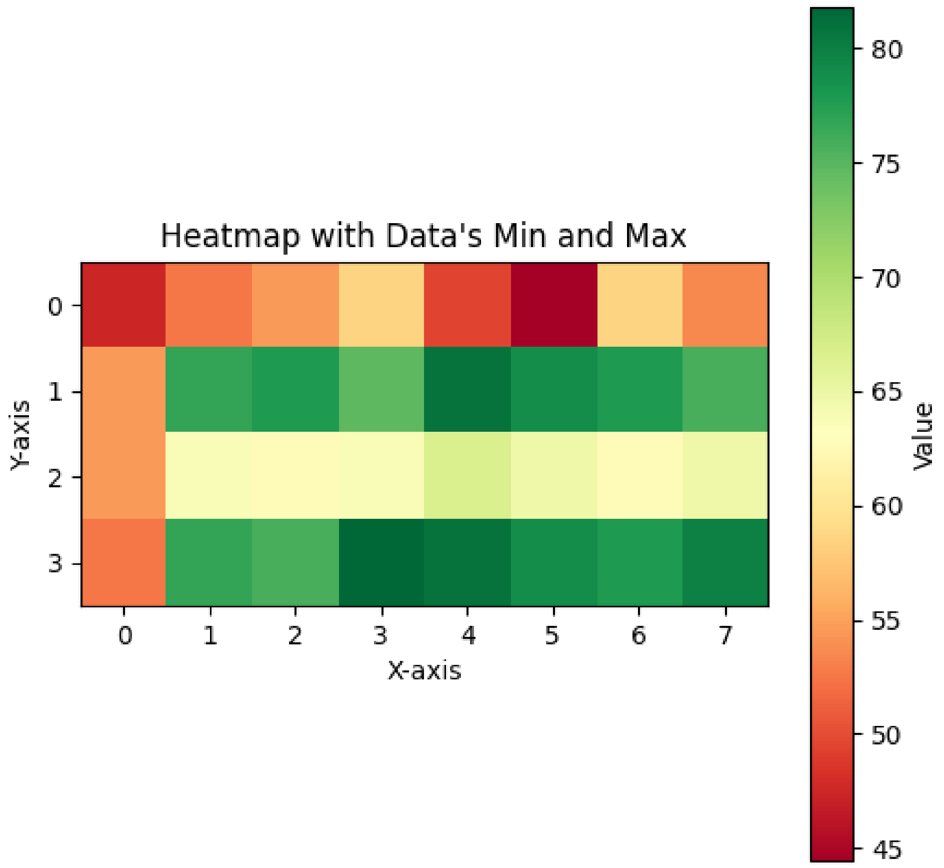
Precision by single question:

Table form:

politeness level	1	2	3	4	5	6	7
elementary mathematics	36.4	50.5	53.5	52.5	57.6	48.5	51.5
conceptual physics	66.67	66.67	71.7	69.7	72.7	70.7	69.7
philosophy	70.7	72.7	73.7	72.7	74.7	73.7	75.8
high school macroeconomics	66.7	69.7	71.7	73.7	71.7	73.7	75.8
all	77.8	81.8	76.8	77.8	79.8	78.8	78.8

Graph form:


Line Plot of Accuracy of questions Over Politeness
Accuracy of single questions
80
70
60
50
40
0
1
2
3
4
5
6
Politeness
Subject elementary_mathematics
Subject conceptual_physics
Subject philosophy
Subject high_school_macroeconomics
Subject all



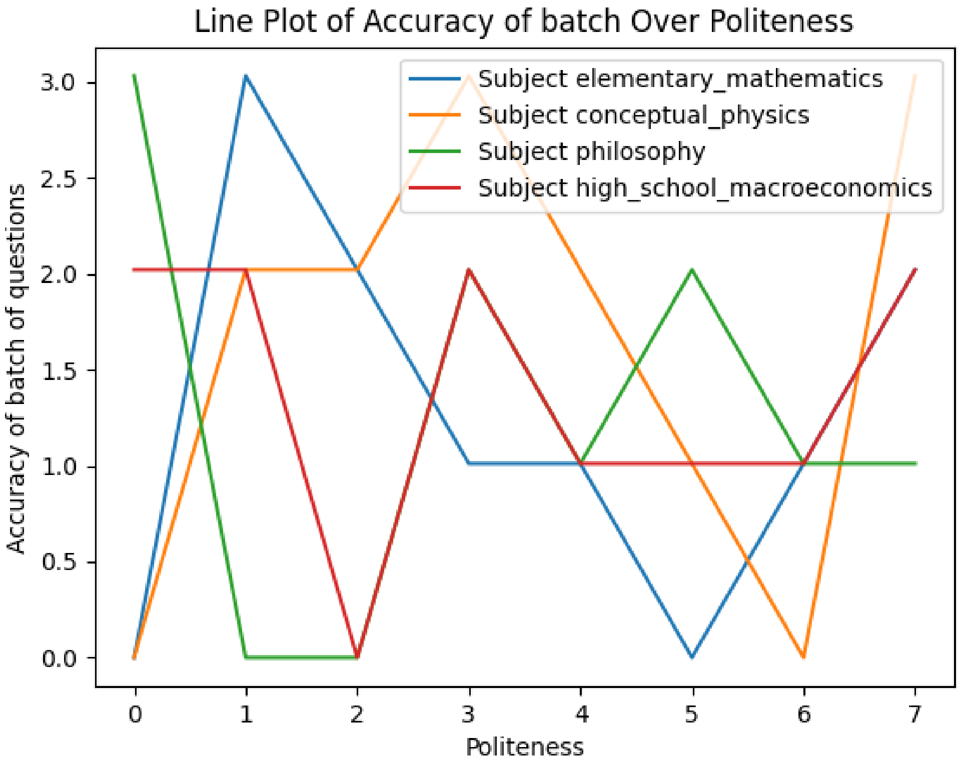
Hebrew, GPT-3.5

Precision by batch (correct if all the questions in the batch are correct):

Table form:

politeness level	1	2	3	4	5	6	7	8
elementary mathematics	0	3.	2	1	1	0	1	2
conceptual physics	0	2	2	3	2	1	0	3
philosophy	0	2	2	3	2	1	0	3
high school macroeconomics	3	0	0	2	1	2	1	1
all	2	2	0	2	1	1	1	2

Graph form:

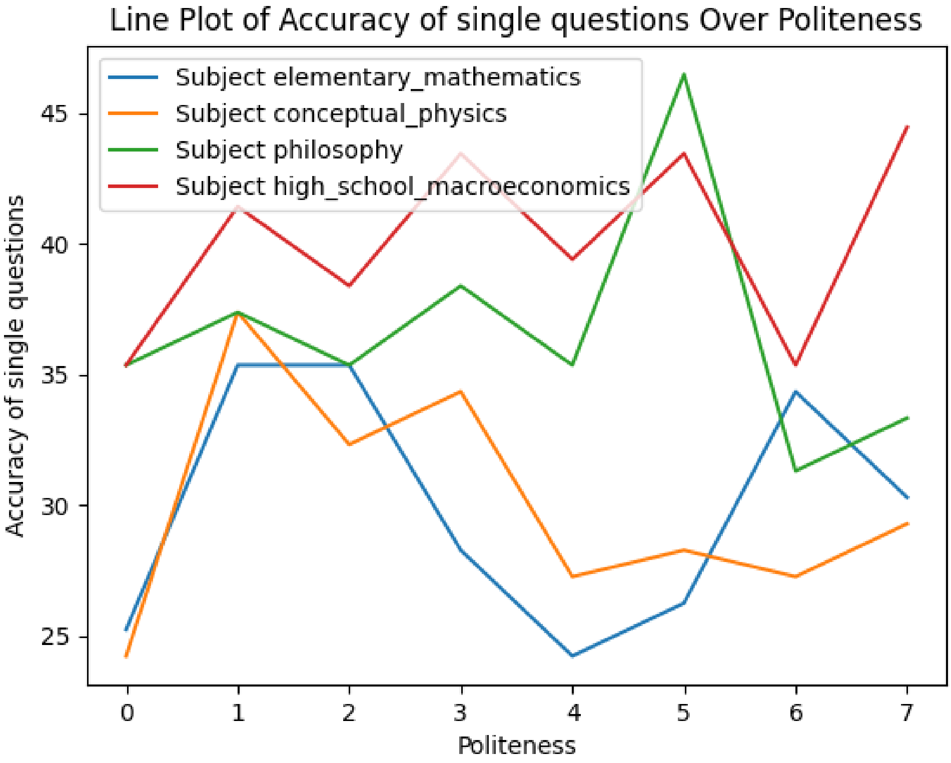


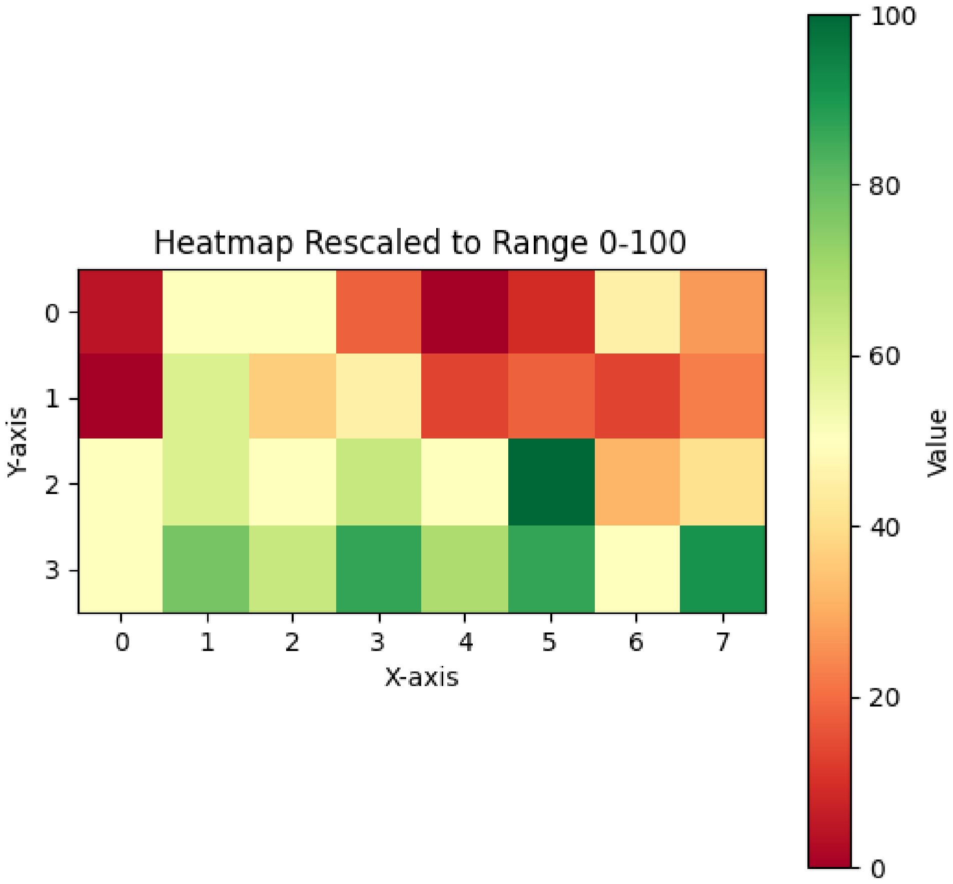
Precision by single question:

Table form:

politeness level	1	2	3	4	5	6	7	8
elementary mathematics	25.3	35.4	35.4	28.3	24.2	26.3	34.3	30.3
conceptual physics	24.2	37.4	32.3	34.3	27.3	28.3	27.3	29.3
philosophy	35.4	37.4	35.4	38.4	35.4	46.5	31.3	33.3
high school macroeconomics	35.4	41.4	38.4	43.4	39.4	43.4	35.4	44.4

Graph form:


Line Plot of Accuracy of single questions Over Politeness
Subject elementary_mathematics
Subject conceptual_physics
Subject philosophy
Subject high_school_macroeconomics
Accuracy of single questions
25
30
35
40
45
0
1
2
3
4
5
6
7
Politeness



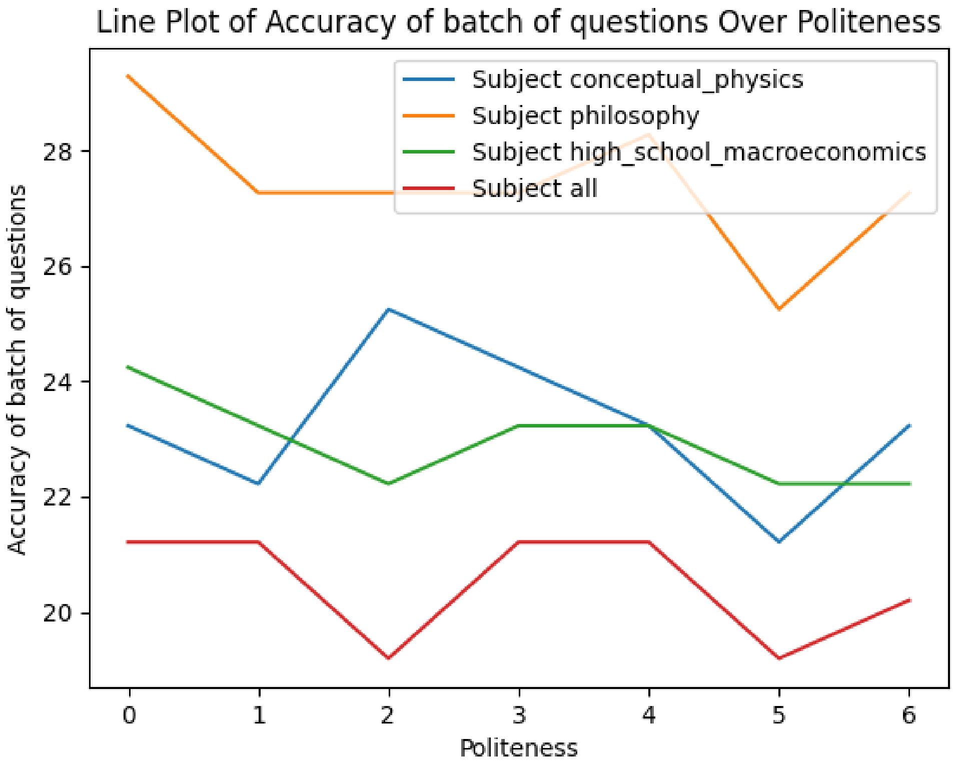
English, GPT-4

Precision by batch (correct if all the questions in the batch are correct):

Table form:

politeness level	1	2	3	4	5	6	7
conceptual physics	23.2	22.2	25.3	24.2	23.2	21.2	23.2
philosophy	29.3	27.3	27.3	27.3	28.3	25.3	27.3
high school macroeconomics	24.2	23.2	22.2	23.2	23.2	22.2	22.2
all	21.2	21.2	19.2	21.2	21.2	19.2	20.2

Graph form:

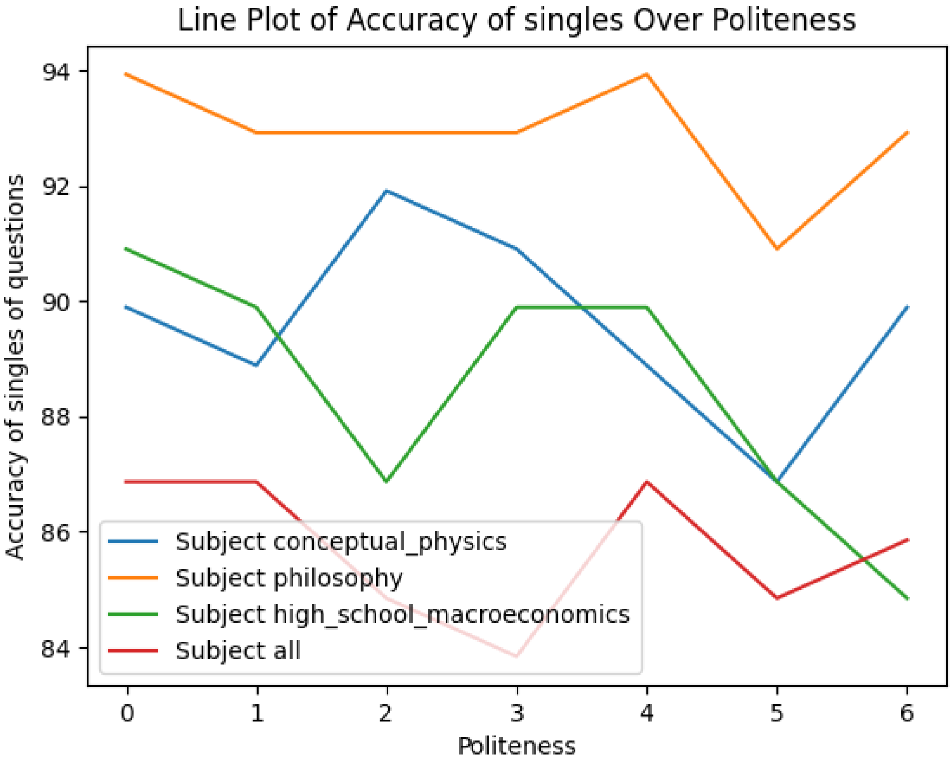


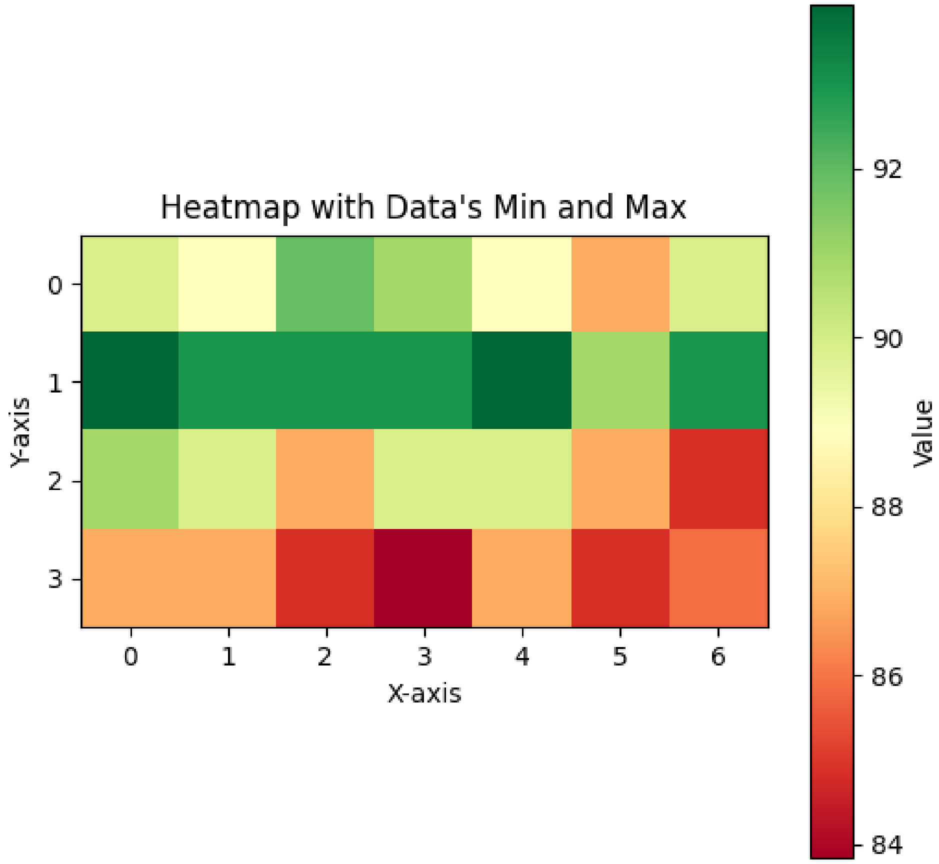
Precision by single question:

Table form:

politeness level	1	2	3	4	5	6	7
conceptual physics	89.9	88.9	91.9	90.9	88.9	86.9	89.9
philosophy	93.9	92.9	92.9	92.9	93.9	90.9	92.9
high school macroeconomics	90.9	89.9	86.9	89.9	89.9	86.9	84.8
all	86.9	86.9	84.8	83.8	86.9	84.8	85.9

Graph form:


Line Plot of Accuracy of singles Over Politeness
Accuracy of singles of questions
94
92
90
88
86
84
0
1
2
3
4
5
6
Politeness
Subject conceptual_physics
Subject philosophy
Subject high_school_macroeconomics
Subject all



<u>Hebrew, GPT-4</u>

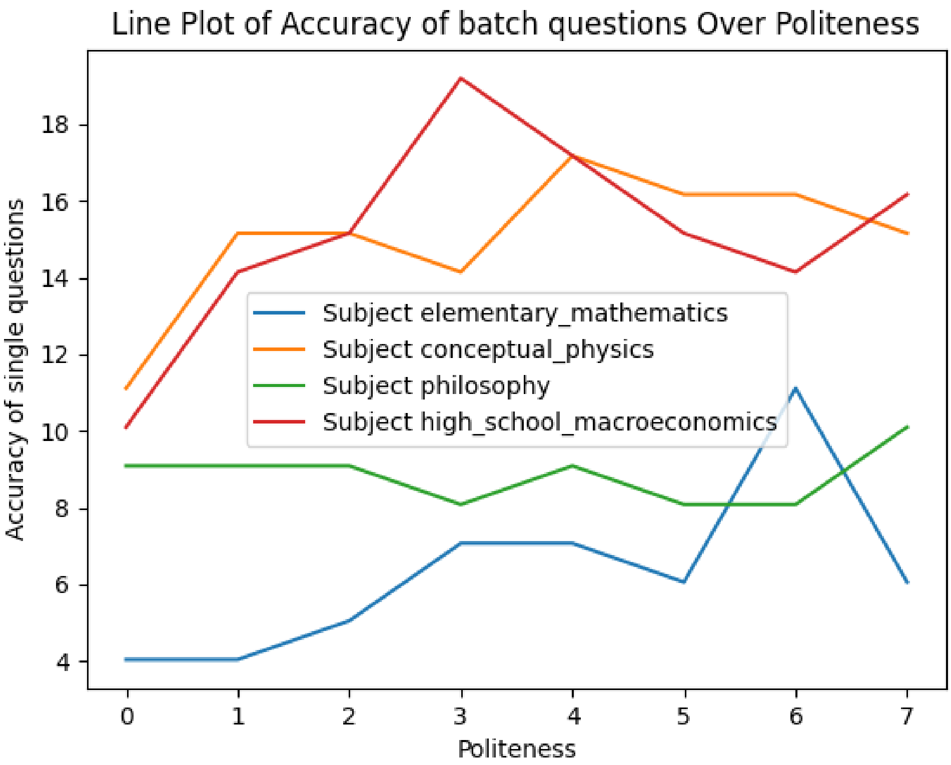
<u>Precision by batch (correct if all the questions in the batch are correct):</u>

Table form:

politeness level	1	2	3	4	5	6	7	8
elementary mathematics	4.04	4.04	5.05	7.07	7.07	6.06	11.11	6.06
conceptual physics	11.11	15.15	15.15	14.14	17.17	16.16	16.16	15.15
philosophy	9.09	9.09	9.09	8.08	9.09	8.08	8.08	10.1

high school macroeconomics	10.1	14.14	15.15	19.19	17.17	15.15	14.14	16.16

Graph form:

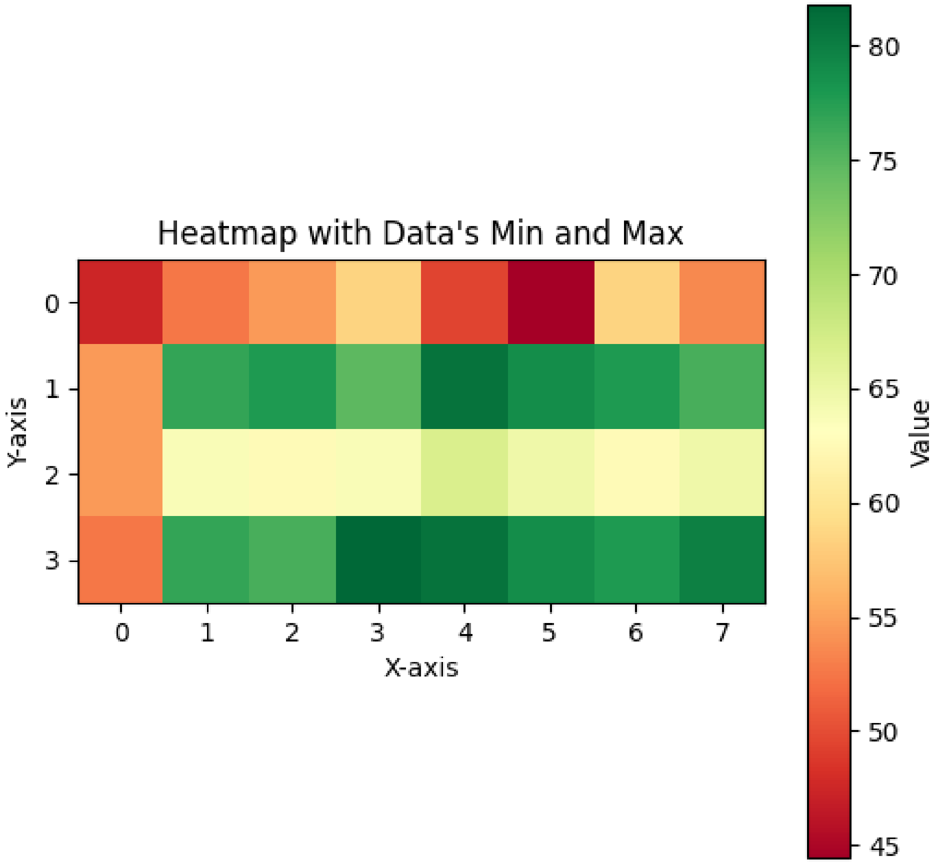


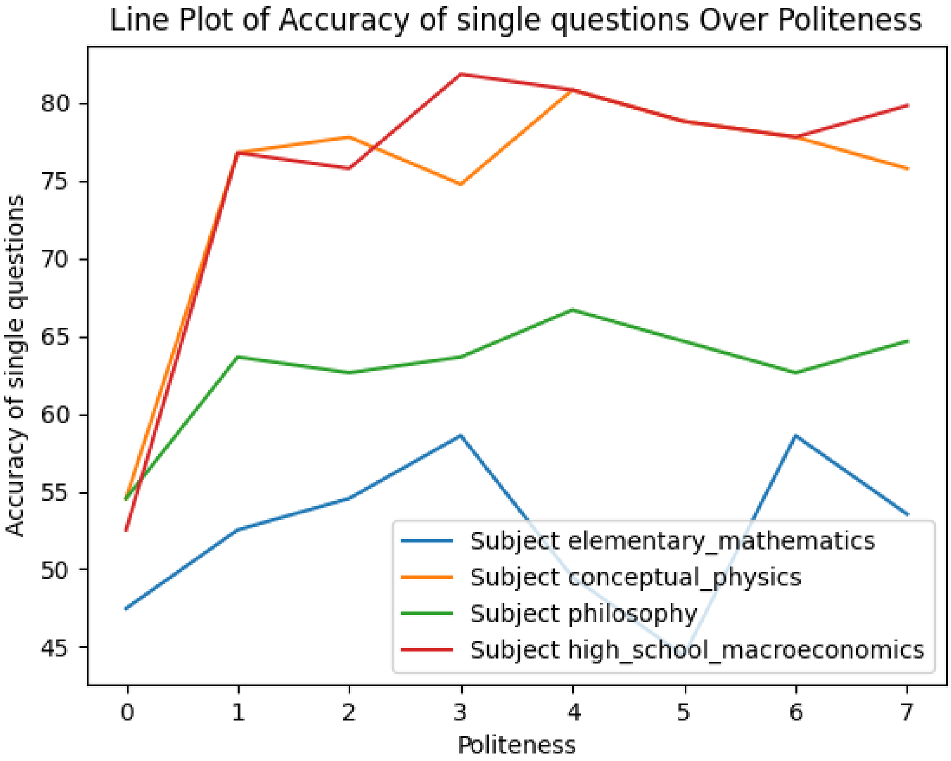
Precision by single question:

Table form:

politeness level	1	2	3	4	5	6	7	8
elementary mathematics	47.47	52.53	54.55	58.59	49.5	44.44	58.59	53.54
conceptual physics	54.55	76.77	77.78	74.75	80.81	78/79	77.78	75.76
philosophy	54.55	63.64	62.63	63.64	66.67	64.65	62.63	64.65
high school macroeconomics	52.53	76.77	75.76	81.82	80.81	78.79	77.78	79.8

Graph form:


Heatmap with Data's Min and Max
Y-axis
0
1
2
3
X-axis
0
1
2
3
4
5
6
7
Value
80
75
70
65
60
55
50
45


Line Plot of Accuracy of single questions Over Politeness
Accuracy of single questions
80
75
70
65
60
55
50
45
0
1
2
3
4
5
6
7
Politeness
Subject elementary_mathematics
Subject conceptual_physics
Subject philosophy
Subject high_school_macroeconomics

Experiment 2- Text Summarization

Each prompt includes a request for summarizing a text.

We send the same text with requests of different politeness levels. There are 8 politeness levels in Hebrew and 7 politeness levels in English.

Evaluation

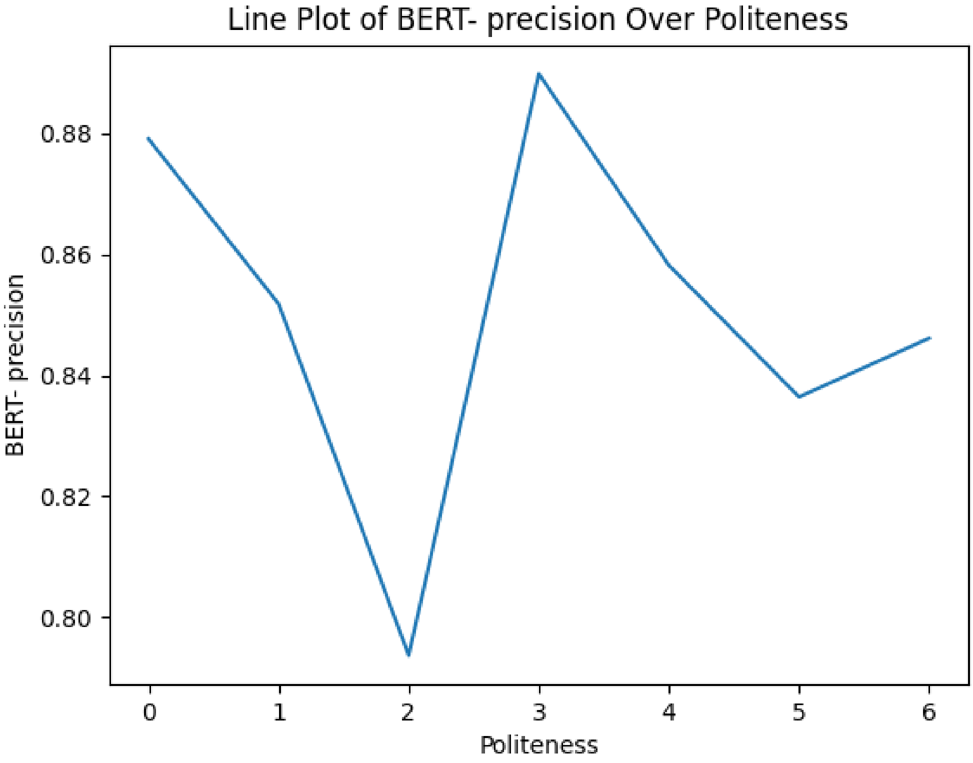
English, GPT 3.5

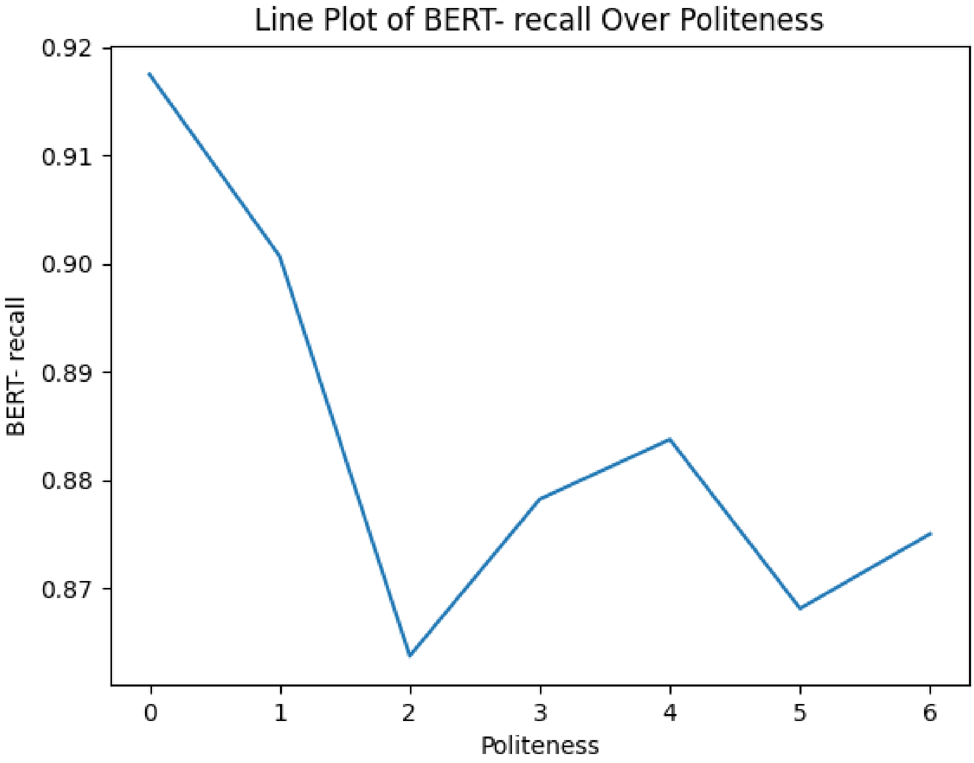
BERTScore

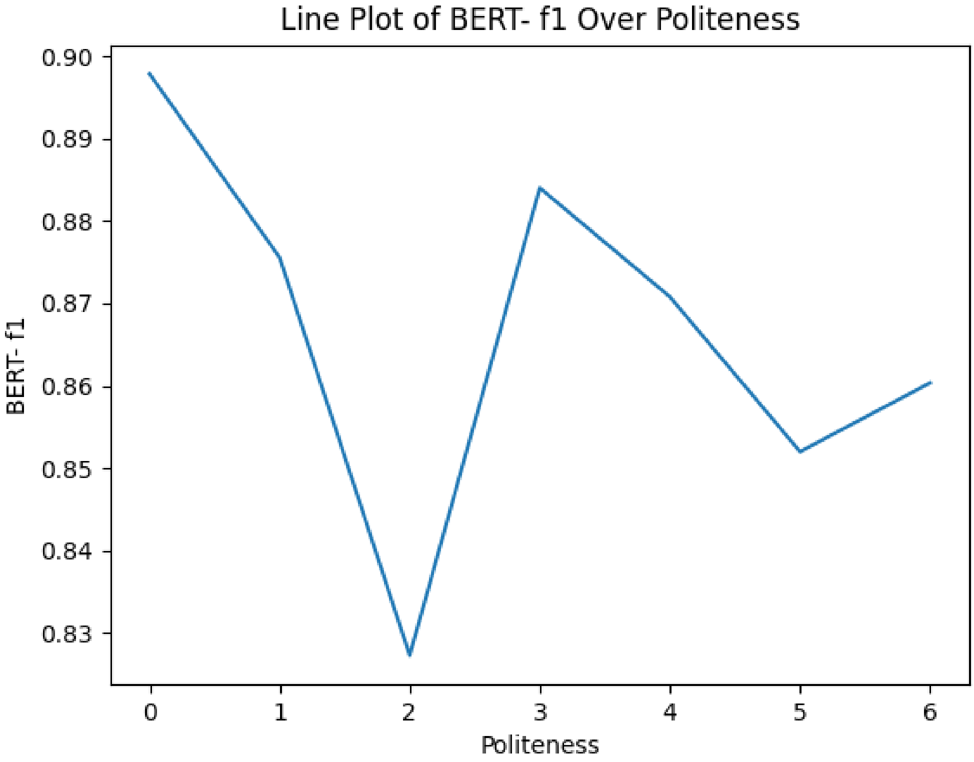
Table form:

politeness level	1	2	3	4	5	6	7
precision	0.879	0.852	0.794	0.89	0.858	0.836	0.846
recall	0.917	0.901	0.864	0.878	0.884	0.868	0.875
f1	0.898	0.876	0.827	0.884	0.871	0.852	0.86

Graph form:




Line Plot of BERT- recall Over Politeness
BERT- recall
0.92
0.91
0.90
0.89
0.88
0.87
0
1
2
3
4
5
6
Politeness


Line Plot of BERT- f1 Over Politeness
BERT- f1
0.90
0.89
0.88
0.87
0.86
0.85
0.84
0.83
0
1
2
3
4
5
6
Politeness

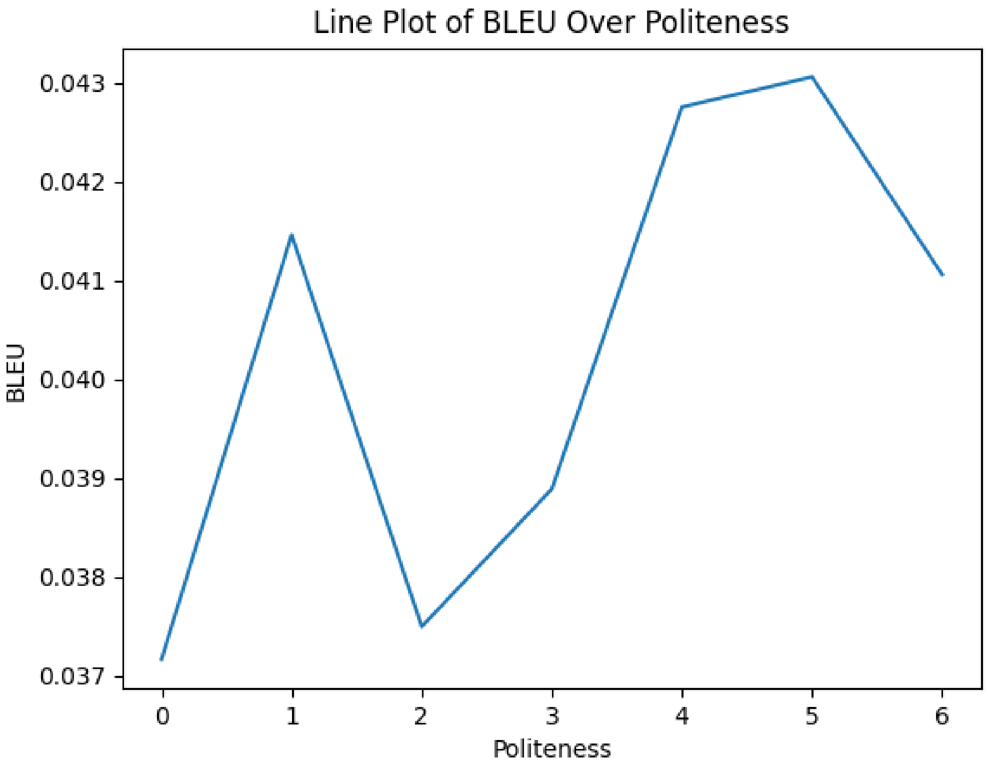
BLUE

All the results have low values

Table form:

politeness level	1	2	3	4	5	6	7
BLEU	0	0	0	0	0	0	0

Graph form:

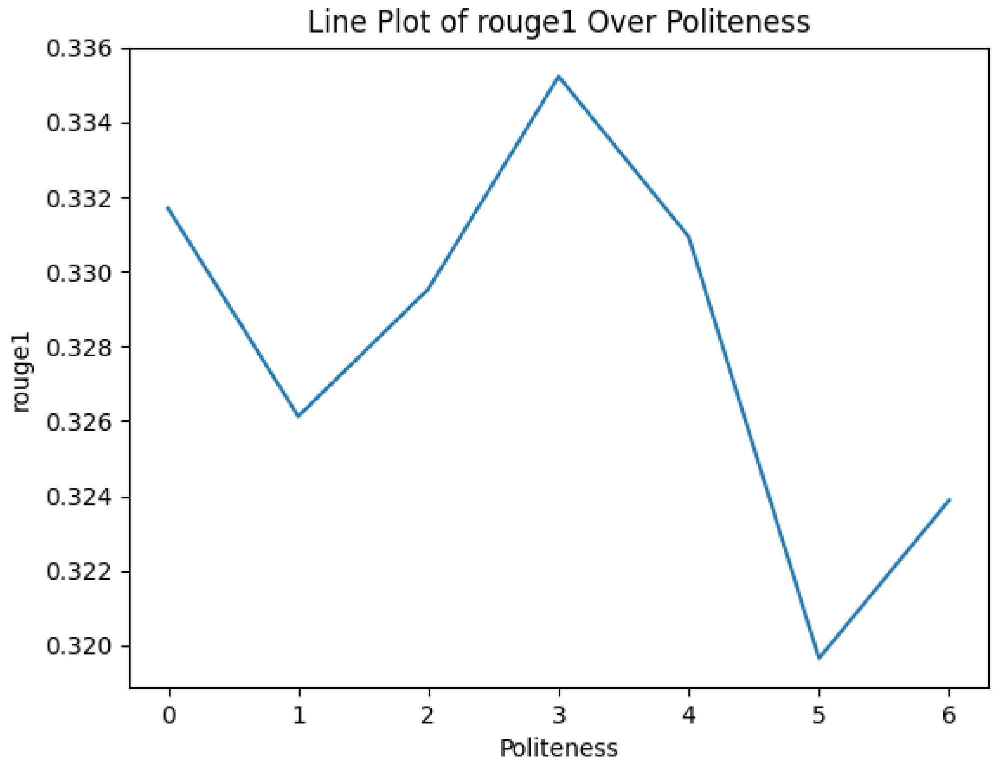


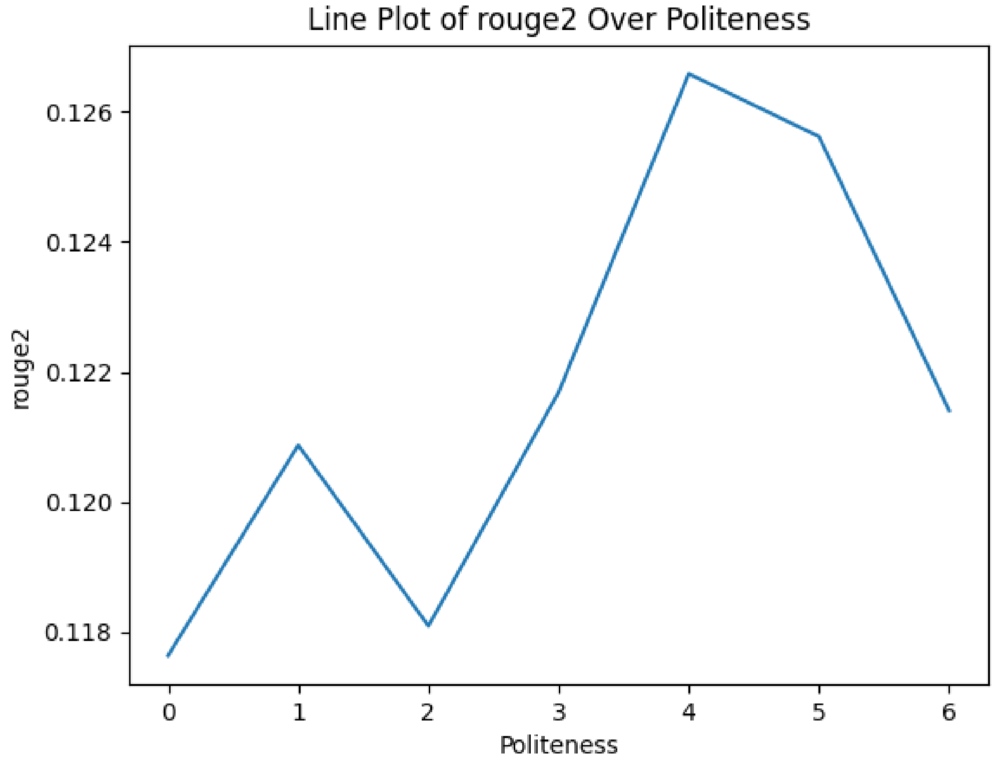
Rouge

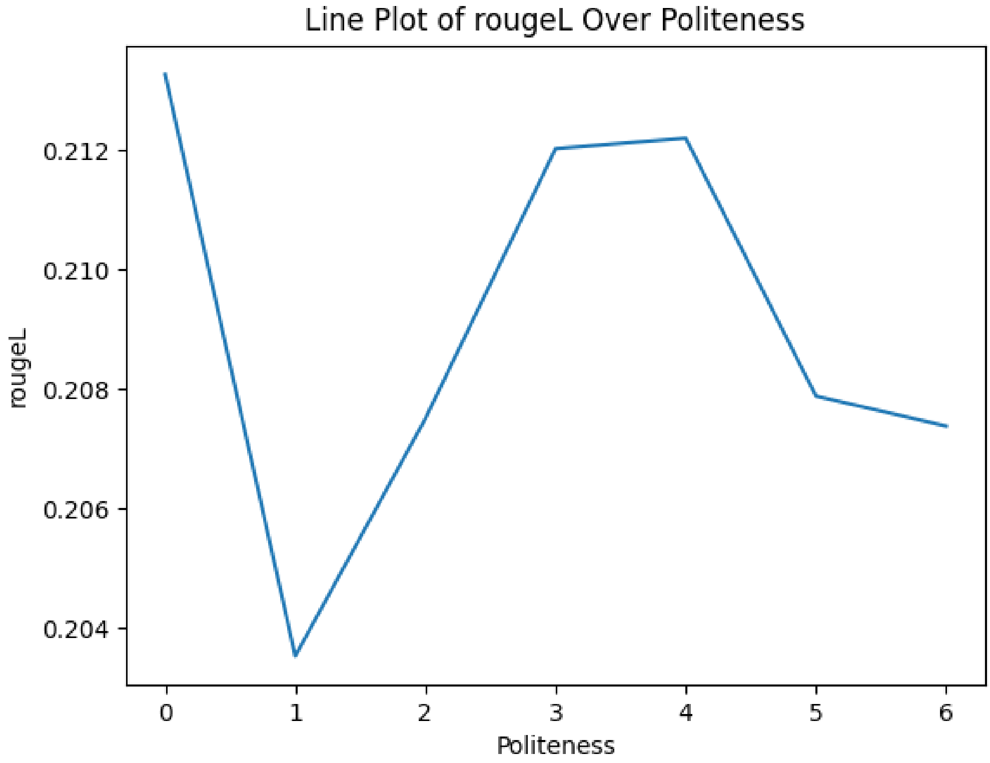
Table form:

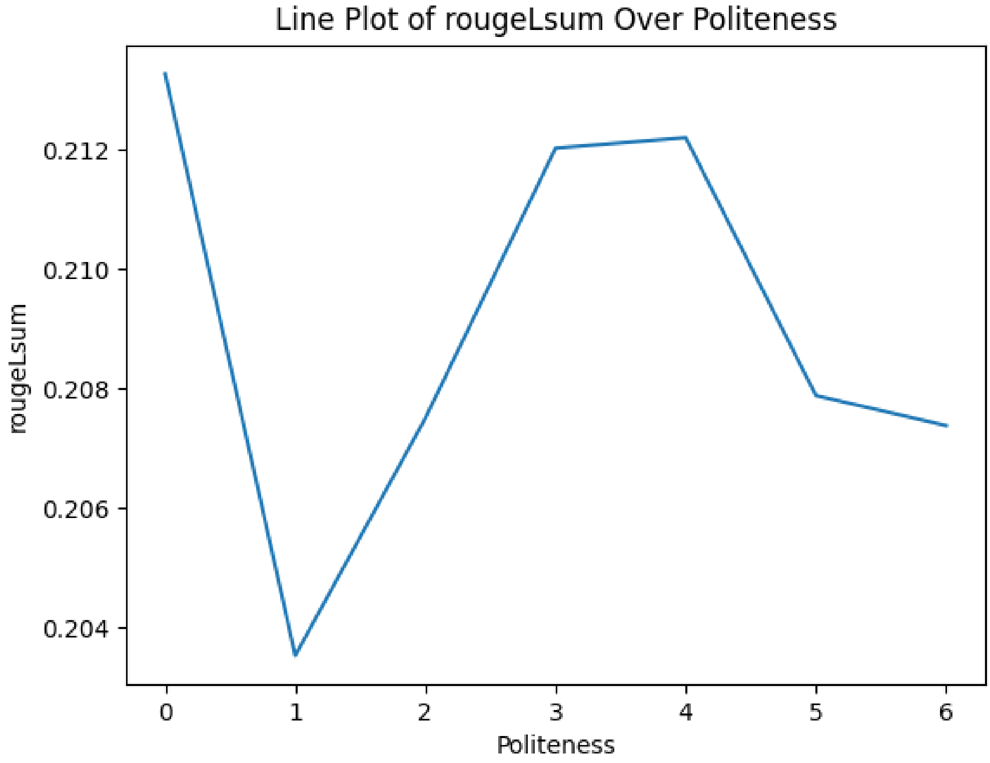
politeness level	1	2	3	4	5	6	7
Rouge1	0.332	0.326	0.33	0.335	0.331	0.32	0.324
Rouge2	0.117	0.121	0.118	0.122	0.127	0.126	0.121
RougeL	0.213	0.204	0.208	0.212	0.212	0.208	0.207
RougeLsum	0.213	0.204	0.208	0.212	0.212	0.208	0.207

Graph form:


Line Plot of rouge1 Over Politeness
rouge1
0.336
0.334
0.332
0.330
0.328
0.326
0.324
0.322
0.320
0
1
2
3
4
5
6
Politeness


Line Plot of rouge2 Over Politeness
rouge2
0.126
0.124
0.122
0.120
0.118
0
1
2
3
4
5
6
Politeness



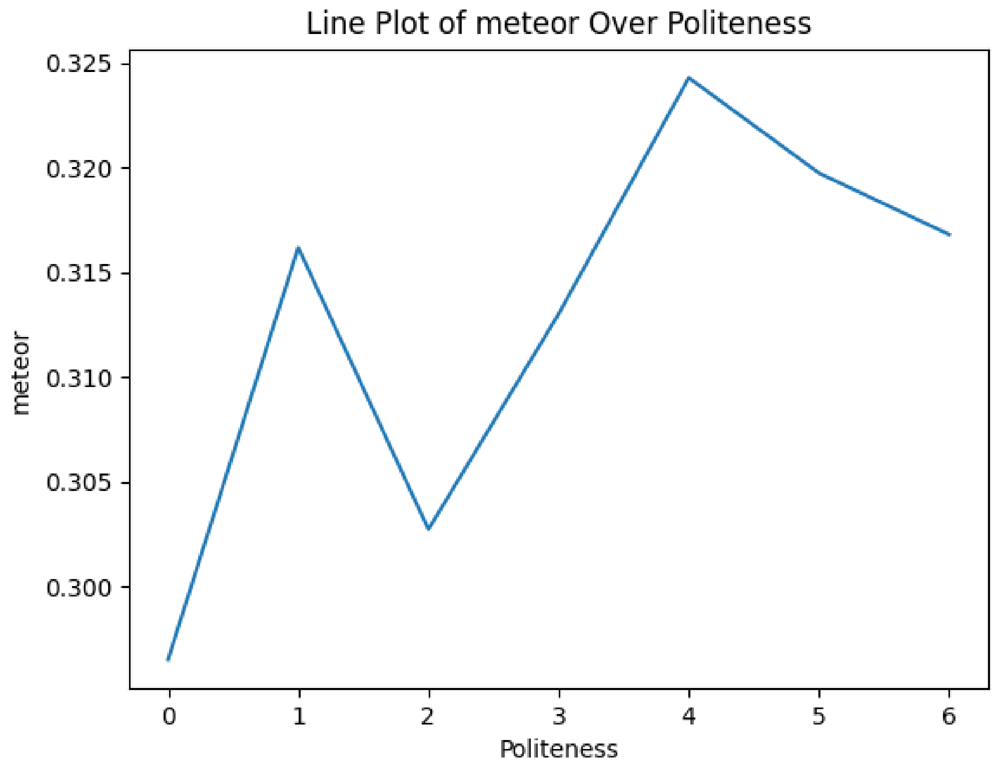


<u>Meteor</u>

Table form:

politeness level	1	2	3	4	5	6	7
Meteor	0.297	0.316	0.303	0.313	0.324	0.32	0.317

Graph form:

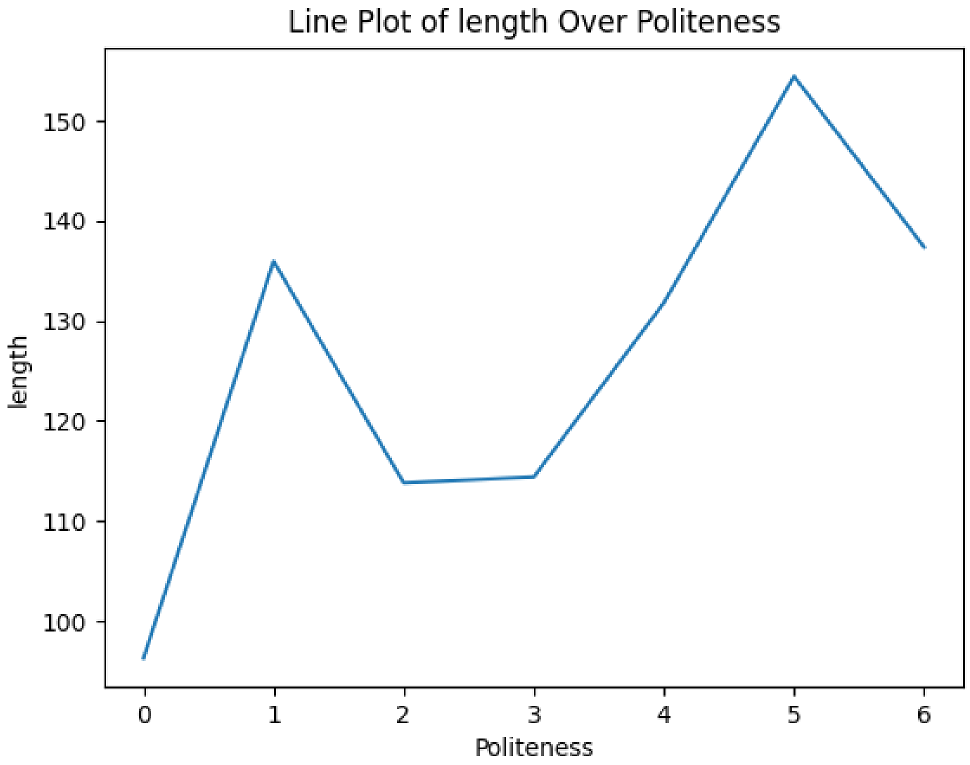


Length

Table form:

politeness level	1	2	3	4	5	6	7
Length $\frac{abs(l_{resul}-l_{expected})}{l_{resul}} \times 100$	96.3	135.94	113.79	114.4	131.8	154.43	137.36

Graph form:



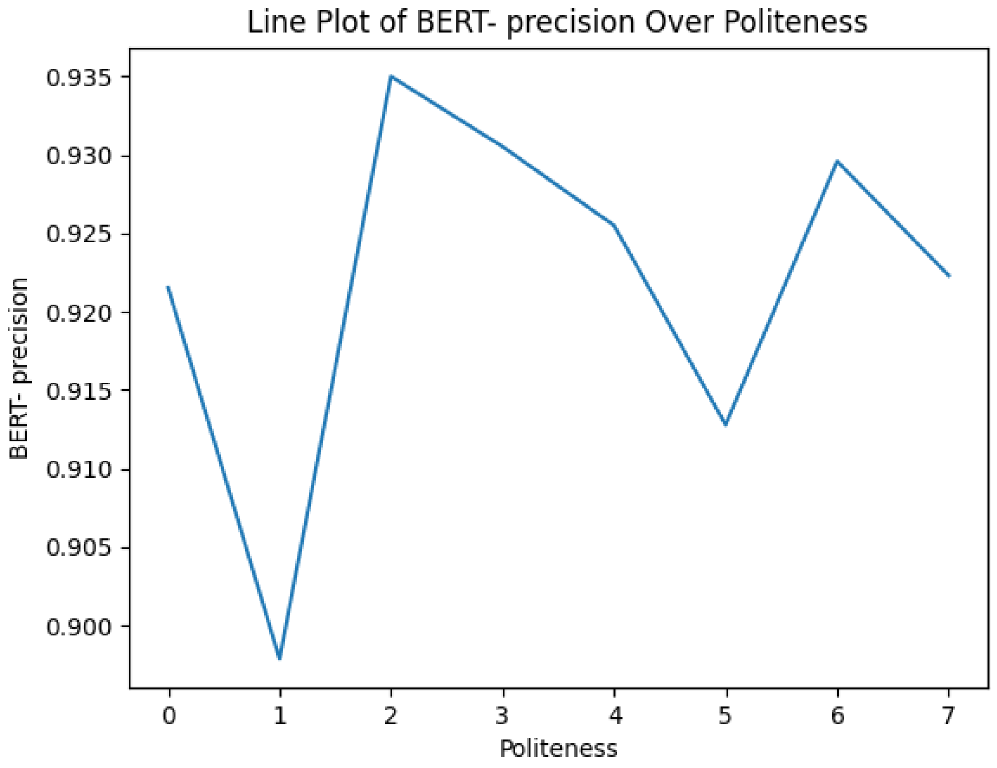
Hebrew, GPT 3.5

BERTScore

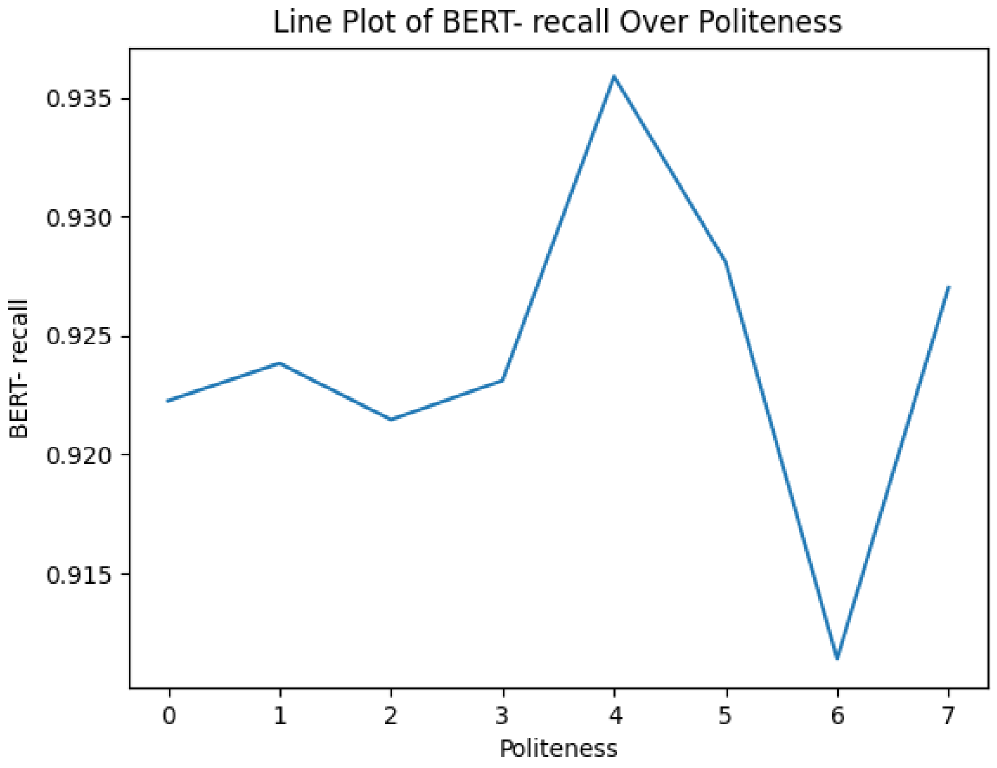
Table form:

politeness level	1	2	3	4	5	6	7	8
precision	0.922	0.898	0.935	0.931	0.925	0.913	0.93	0.922
recall	0.922	0.924	0.921	0.923	0.936	0.928	0.911	0.927
f1	0.922	0.911	0.928	0.927	0.931	0.920	0.92	0.925

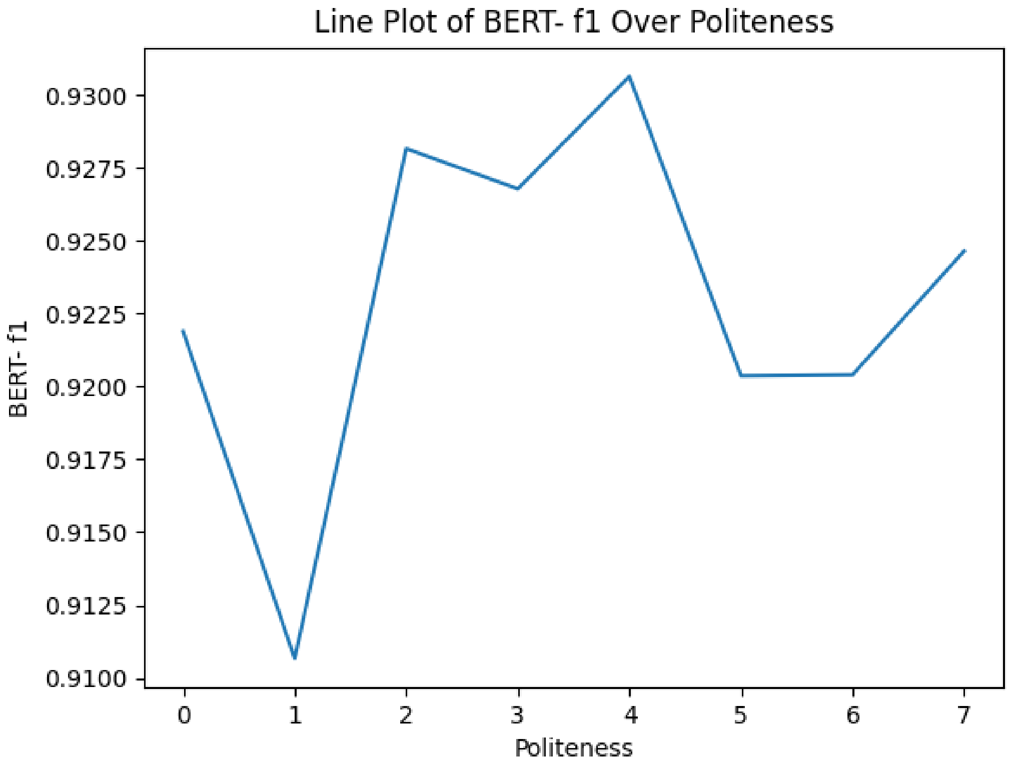
Graph form:



Line Plot of BERT- precision Over Politeness



Line Plot of BERT- recall Over Politeness

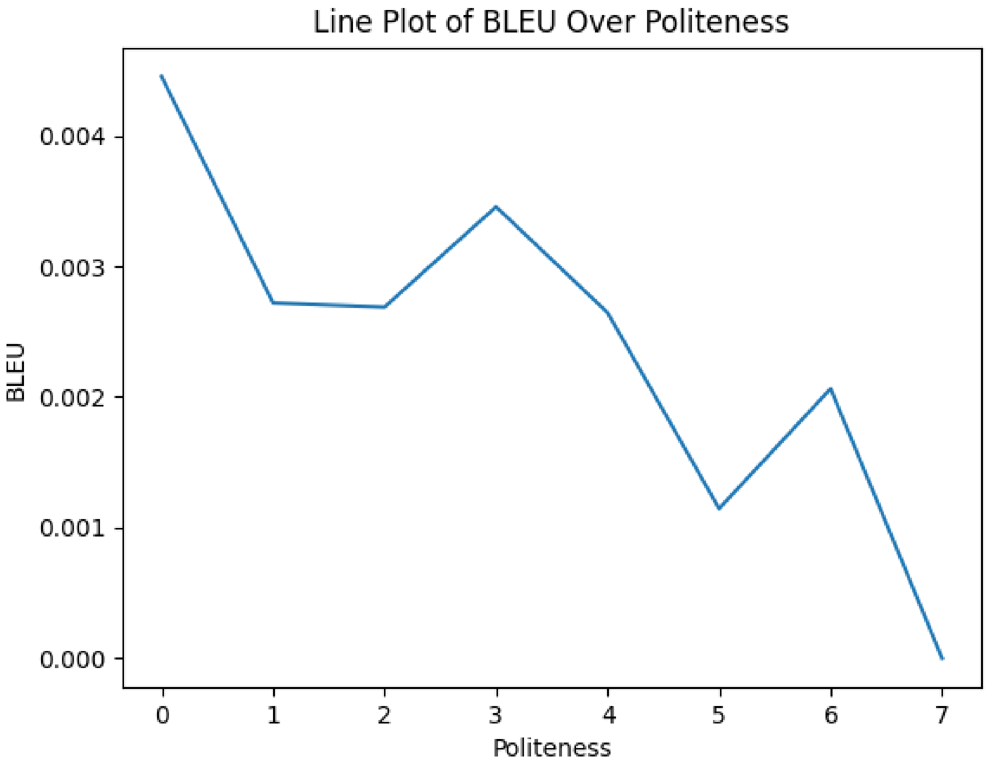


<u>BLUE</u>

Table form:

politeness level	1	2	3	4	5	6	7
BLEU	0	0	0	0	0	0	0

Graph form:

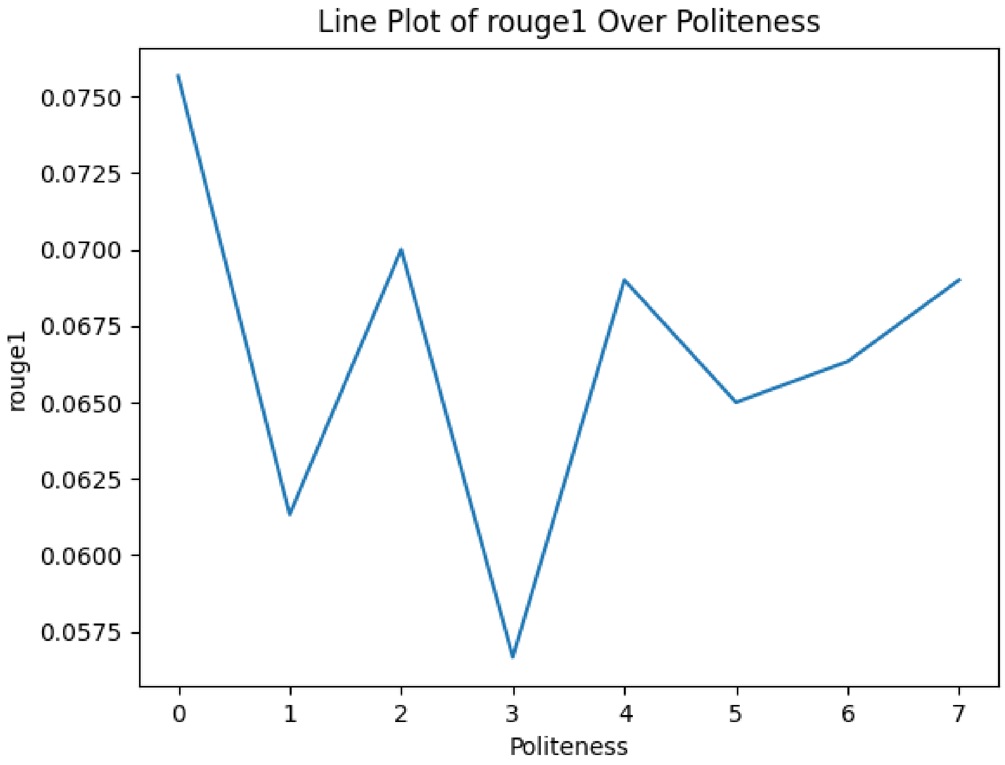


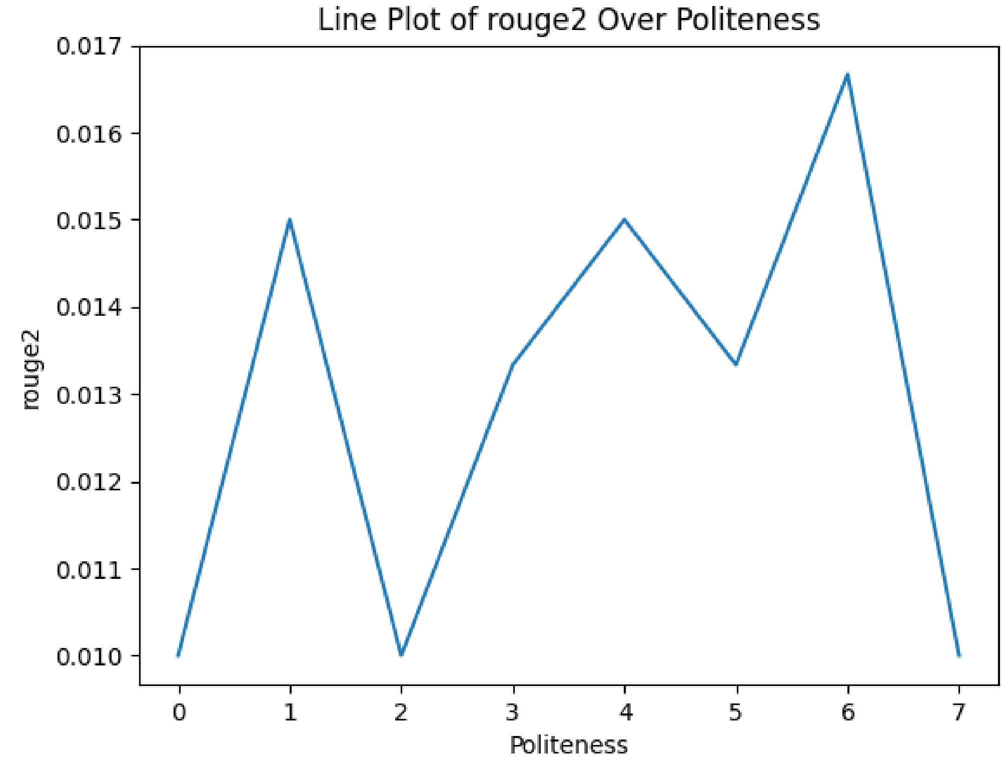
Rouge

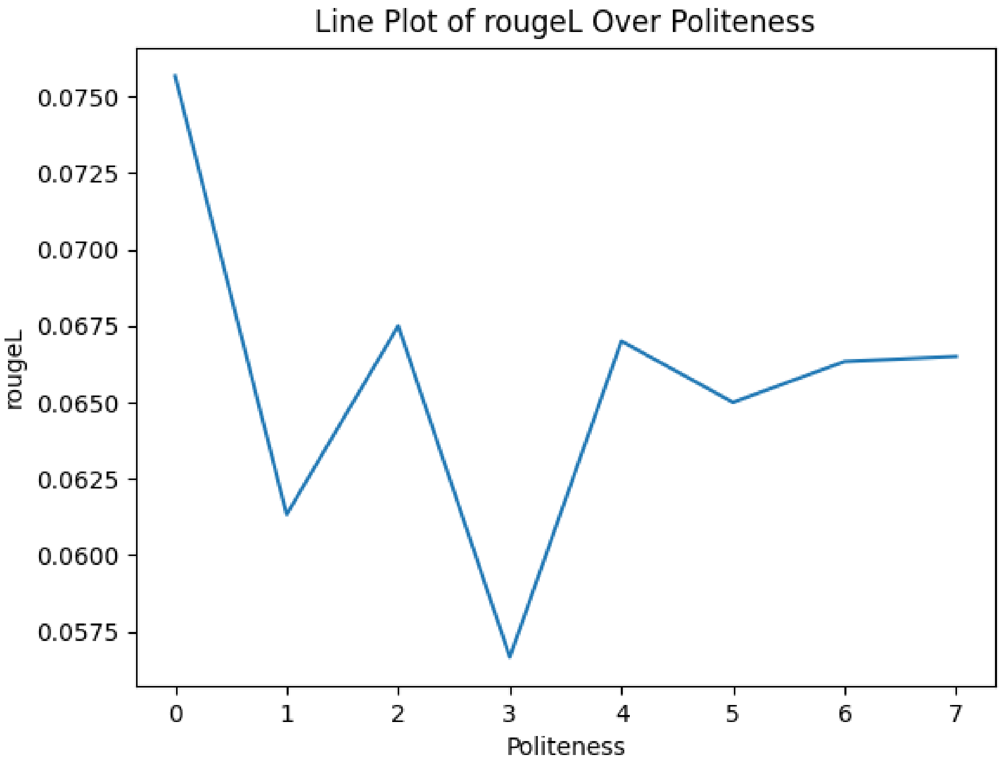
Table form:

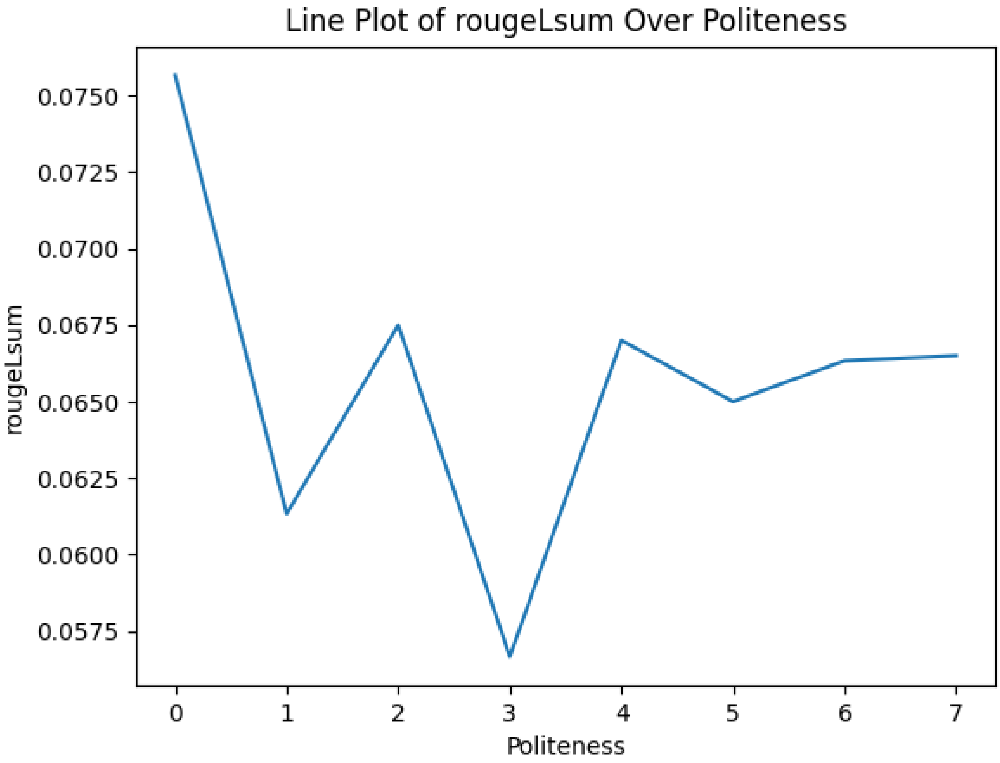
politeness level	1	2	3	4	5	6	7	8
Rouge1	0.076	0.061	0.07	0.057	0.069	0.065	0.066	0.069
Rouge2	0.01	0.015	0.01	0.013	0.015	0.013	0.017	0.01
RougeL	0.076	0.061	0.068	0.057	0.067	0.065	0.066	0.067
RougeLsum	0.076	0.061	0.068	0.057	0.067	0.065	0.063	0.067

Graph form:


Line Plot of rouge1 Over Politeness
rouge1
0.0750
0.0725
0.0700
0.0675
0.0650
0.0625
0.0600
0.0575
0
1
2
3
4
5
6
7
Politeness


Line Plot of rouge2 Over Politeness
rouge2
0.017
0.016
0.015
0.014
0.013
0.012
0.011
0.010
0
1
2
3
4
5
6
7
Politeness



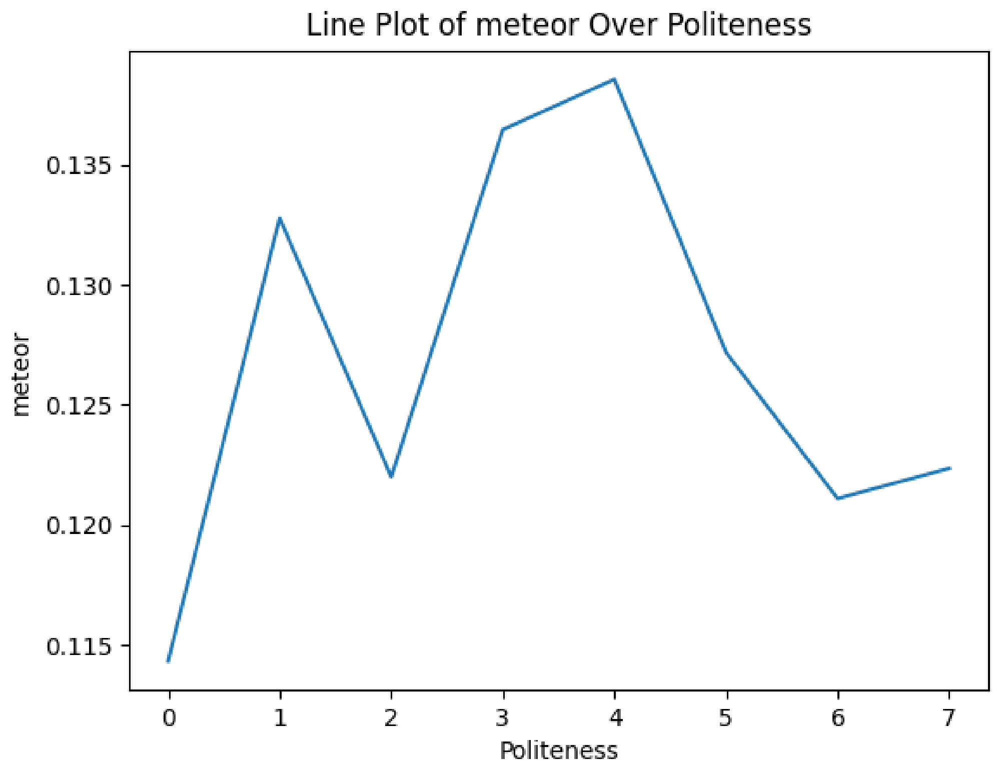


Meteor

Table form:

politeness level	1	2	3	4	5	6	7	8
Meteor	0.114	0.133	0.122	0.136	0.139	0.127	0.121	0.122

Graph form:

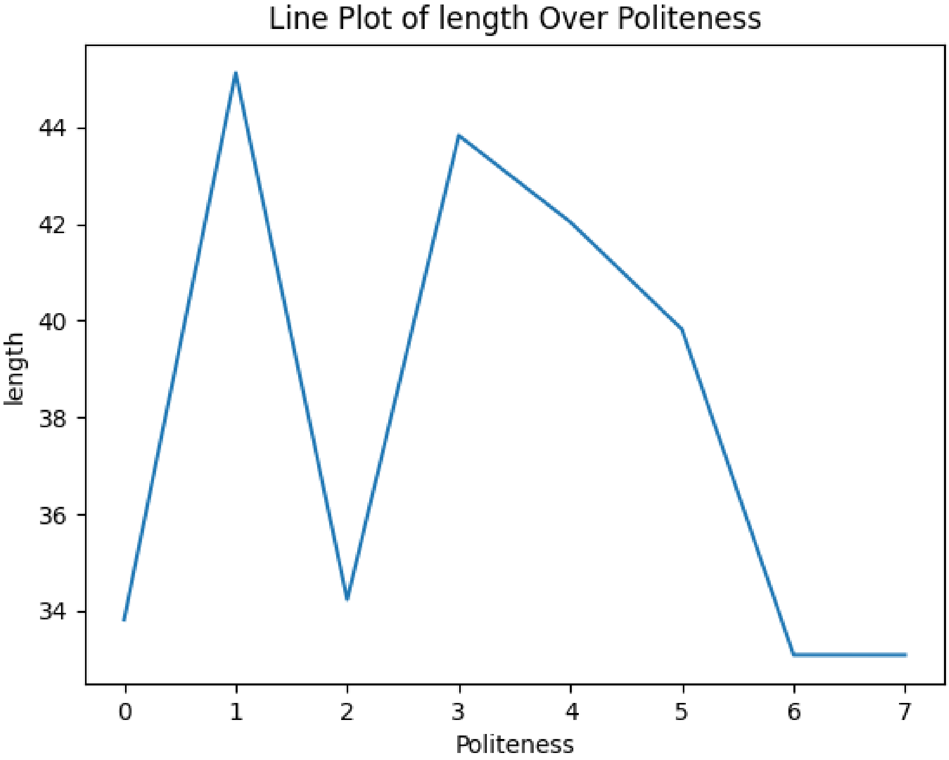


Length

Table form:

politeness level	1	2	3	4	5	6	7	8
Length $\frac{abs(l_{resul}-l_{expected})}{l_{resul}}$ 100	33.81	45.13	34.23	43.83	42.04	39.82	33.08	33.08

Graph form:


Line Plot of length Over Politeness
length
44
42
40
38
36
34
0
1
2
3
4
5
6
7
Politeness