**Model Evaluation Results:**

The data collected in different phases of the ablation study is evaluated using RoBERTa, trained on SNLI. The 4 phases are: regular crowdsourcing, traffic signal feedback, AutoFix, and Full System availability.

|  |  |  |
| --- | --- | --- |
| **Data** | **ROBERTA (S2)** | **S22** |
| Crowdsourcing | 1 | 1.000 |
| Traffic Signal | 1 | 0.856 |
| AutoFix | 1 | 0.803 |
| Full System | 1 | 0.748 |
| Full Data (SNLI Dev set) | 0.89 | -- |

**S1 and S2**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Data | T1 | T2 | C1 | W1 | W2 | Bi T1 | Bi T2 | Word | S | Adj | Adv | N | V | Bi | Tri | C3 |
| Crowdsourcing | 6.07 | 5.55 | 11.62 | 144.93 | 0.57 | 621.01 | 0.88 | 419 | 138 | 99 | 9 | 201 | 88 | 994 | 1149 | 202 |
| Traffic Signal | 6.43 | 5.33 | 11.76 | 170.09 | 0.71 | 662.55 | 0.86 | 444 | 138 | 88 | 17 | 213 | 102 | 1027 | 1167 | 202 |
| AutoFix | 5.81 | 5.08 | 10.89 | 131.92 | 0.6 | 564.54 | 0.89 | 401 | 138 | 83 | 12 | 204 | 80 | 938 | 1041 | 202 |
| Full | 6.62 | 5.85 | 12.48 | 203.01 | 0.69 | 922.06 | 0.96 | 457 | 138 | 90 | 8 | 245 | 89 | 1055 | 1156 | 202 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Data | C4 | WO | STS | W1 | W2 | Bi T1 | Bi T2 | Word | S | Adj | Adv | N | V | Bi | Tri |
| Crowdsourcing | 0.008 | 0.074 | 0.001 | 108.08 | 0.36 | 362.27 | 0.83 | 163 | 46 | 35 | 4 | 69 | 36 | 347 | 369 |
| Traffic Signal | 0.008 | 0.07 | 0.001 | 108.54 | 0.43 | 406.45 | 0.84 | 167 | 46 | 32 | 1 | 83 | 41 | 346 | 363 |
| AutoFix | 0.007 | 0.08 | 0.001 | 98.14 | 0.34 | 477.10 | 0.89 | 164 | 46 | 37 | 6 | 73 | 37 | 350 | 360 |
| Full | 0.008 | 0.074 | 0.001 | 134.56 | 0.35 | 601.65 | 0.93 | 169 | 46 | 28 | 3 | 91 | 35 | 366 | 370 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Data | W1 | W2 | Bi T1 | Bi T2 | Word | S | Adj | Adv | N | V | Bi | Tri |
| Crowdsourcing | 126.35 | 0.36 | 531.05 | 0.90 | 181 | 46 | 43 | 1 | 89 | 38 | 390 | 407 |
| Traffic Signal | 210.84 | 0.78 | 668.42 | 0.94 | 224 | 46 | 45 | 16 | 93 | 54 | 455 | 465 |
| AutoFix | 136.15 | 0.48 | 364.03 | 0.81 | 172 | 46 | 35 | 3 | 87 | 35 | 339 | 350 |
| Full | 162.47 | 0.40 | 749.45 | 0.97 | 178 | 46 | 40 | 3 | 85 | 39 | 385 | 381 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Data | W1 | W2 | Bi T1 | Bi T2 | Word | S | Adj | Adv | N | V | Bi | Tri |
| Crowdsourcing | 144.37 | 0.39 | 492.82 | 0.87 | 194 | 46 | 45 | 3 | 103 | 35 | 401 | 410 |
| Traffic Signal | 131.17 | 0.47 | 369.08 | 0.81 | 168 | 46 | 32 | 5 | 85 | 36 | 342 | 369 |
| AutoFix | 106.91 | **0.33** | 390.213 | 0.87 | 170 | 46 | 27 | 5 | 85 | 36 | 353 | 372 |
| Full | 170.83 | 0.57 | 618.19 | 1 | 212 | 46 | 53 | 4 | 113 | 34 | 412 | 429 |

**S1 and S22**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Data | T1 | T2 | C1 | W1 | W2 | Bi T1 | Bi T2 | Word | S | Adj | Adv | N | V | Bi | Tri | C3 |
| Crowdsourcing | 6.246 | 5.58 | 11.83 | 156.32 | 0.62 | 687.91 | 0.89 | 431 | 138 | 89 | 6 | 223 | 91 | 1027 | 1145 | 202 |
| Traffic Signal | 7.39 | 5.22 | 12.62 | 248.08 | 0.81 | 919.53 | 0.88 | 503 | 136 | 109 | 14 | 117 | 237 | 1117 | 1219 | 202 |
| AutoFix | 7.08 | 5.04 | 12.13 | 237.01 | 0.72 | 886.36 | 0.91 | 489 | 138 | 89 | 20 | 247 | 104 | 1030 | 1071 | 202 |
| Full | 7.34 | 5.95 | 13.30 | 263.46 | 0.78 | 1149.68 | 1 | 507 | 138 | 111 | 12 | 247 | 113 | 1107 | 1173 | 202 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Data | C4 | WO | STS | W1 | W2 | Bi T1 | Bi T2 | Word | S | Adj | Adv | N | V | Bi | Tri |
| Crowdsourcing |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Traffic Signal |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| AutoFix | 0.10 | 0.08 | 0.001 |  |  |  |  |  |  |  |  |  |  |  |  |
| Full |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Data | W1 | W2 | Bi T1 | Bi T2 | Word | S | Adj | Adv | N | V | Bi | Tri |
| Crowdsourcing |  |  |  |  |  |  |  |  |  |  |  |  |
| Traffic Signal |  |  |  |  |  |  |  |  |  |  |  |  |
| AutoFix |  |  |  |  |  |  |  |  |  |  |  |  |
| Full |  |  |  |  |  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Data | W1 | W2 | Bi T1 | Bi T2 | Word | S | Adj | Adv | N | V | Bi | Tri |
| Crowdsourcing |  |  |  |  |  |  |  |  |  |  |  |  |
| Traffic Signal |  |  |  |  |  |  |  |  |  |  |  |  |
| AutoFix |  |  |  |  |  |  |  |  |  |  |  |  |
| Full |  |  |  |  |  |  |  |  |  |  |  |  |