Language Confusion Matrix - aya-23-8B-10%

|                    | Language Confusion Matrix - aya-23-8B-10% |                         |                  |                         |                  |              |                       |                         |                |                         |                  |                         |                  |                         |                  |                         |                |                         |                  |                         |                  |                         |                     |                         |                  |                    |                 |                    |                  |                         |                  |                                 |                |                         |                 |                                |
|--------------------|---|-------------------------|------------------|-------------------------|------------------|--------------|-----------------------|-------------------------|----------------|-------------------------|------------------|-------------------------|------------------|-------------------------|------------------|-------------------------|----------------|-------------------------|------------------|-------------------------|------------------|-------------------------|---------------------|-------------------------|------------------|--------------------|-----------------|--------------------|------------------|-------------------------|------------------|---------------------------------|----------------|-------------------------|-----------------|--------------------------------|
| am_test_neutral -  | 0.69                                      | 0.70                    | 0.69             | 0.70                    | 0.34             | 0.34         | 0.35                  | 0.33                    | 0.32           | 0.31                    | 0.30             | 0.32                    | 0.29             | 0.29                    | 0.29             | 0.30                    | 0.31           | 0.30                    | 0.27             | 0.30                    | 0.34             | 0.33                    | 0.33                | 0.35                    | 0.26             | 0.25               | 0.26            | 0.25               | 0.26             | 0.27                    | 0.25             | 0.26                            | 0.32           | 0.32                    | 0.32            | 0.32                           |
| am_test_toxic -    | 0.70                                      | 0.70                    | 0.70             | 0.71                    | 0.33             | 0.34         | 0.34                  | 0.33                    | 0.31           | 0.31                    | 0.30             | 0.31                    | 0.28             | 0.29                    | 0.29             | 0.30                    | 0.30           | 0.29                    | 0.27             | 0.30                    | 0.33             | 0.32                    | 0.32                | 0.34                    | 0.25             | 0.25               | 0.26            | 0.25               | 0.26             | 0.27                    | 0.25             | 0.26                            | 0.31           | 0.31                    | 0.32            | 0.31                           |
| am_train_neutral - | 0.69                                      | 0.70                    | 0.70             | 0.70                    | 0.35             | 0.35         | 0.36                  | 0.34                    | 0.33           | 0.32                    | 0.31             | 0.33                    | 0.30             | 0.30                    | 0.30             | 0.31                    | 0.32           | 0.30                    | 0.28             | 0.31                    | 0.35             | 0.34                    | 0.34                | 0.36                    | 0.26             | 0.26               | 0.27            | 0.25               | 0.27             | 0.28                    | 0.25             | 0.27                            | 0.33           | 0.33                    | 0.34            | 0.33                           |
| am_train_toxic -   | 0.70                                      | 0.71                    | 0.70             | 0.71                    | 0.34             | 0.34         | 0.35                  | 0.34                    | 0.32           | 0.32                    | 0.31             | 0.32                    | 0.29             | 0.29                    | 0.29             | 0.30                    | 0.31           | 0.30                    | 0.28             | 0.31                    | 0.34             | 0.33                    | 0.33                | 0.35                    | 0.26             | 0.25               | 0.26            | 0.25               | 0.26             | 0.27                    | 0.25             | 0.27                            | 0.32           | 0.32                    | 0.33            | 0.32                           |
| ar_test_neutral -  | 0.34                                      | 0.33                    | 0.35             | 0.34                    | 0.56             | 0.56         | 0.57                  | 0.55                    | 0.50           | 0.49                    | 0.48             | 0.49                    | 0.46             | 0.47                    | 0.45             | 0.47                    | 0.51           | 0.49                    | 0.45             | 0.50                    | 0.53             | 0.52                    | 0.52                | 0.54                    | 0.46             | 0.45               | 0.47            | 0.45               | 0.45             | 0.46                    | 0.43             | 0.45                            | 0.54           | 0.54                    | 0.55            | 0.55                           |
| ar_test_toxic -    | 0.34                                      | 0.34                    | 0.35             | 0.34                    | 0.56             | 0.57         | 0.57                  | 0.55                    | 0.50           | 0.50                    | 0.48             | 0.49                    | 0.47             | 0.48                    | 0.47             | 0.48                    | 0.52           | 0.50                    | 0.46             | 0.51                    | 0.53             | 0.52                    | 0.52                | 0.54                    | 0.47             | 0.46               | 0.47            | 0.46               | 0.46             | 0.47                    | 0.44             | 0.46                            | 0.54           | 0.54                    | 0.54            | 0.54                           |
| ar_train_neutral - | - 0.35                                    | 0.34                    | 0.36             | 0.35                    | 0.57             | 0.57         | 0.59                  | 0.56                    | 0.51           | 0.51                    | 0.49             | 0.51                    | 0.48             | 0.49                    | 0.47             | 0.49                    | 0.53           | 0.50                    | 0.47             | 0.52                    | 0.54             | 0.53                    | 0.53                | 0.55                    | 0.48             | 0.47               | 0.48            | 0.47               | 0.47             | 0.48                    | 0.45             | 0.47                            | 0.55           | 0.56                    | 0.56            | 0.56                           |
| ar_train_toxic -   | 0.33                                      | 0.33                    | 0.34             | 0.34                    | 0.55             | 0.55         | 0.56                  | 0.54                    | 0.49           | 0.48                    | 0.47             | 0.48                    | 0.46             | 0.47                    | 0.45             | 0.47                    | 0.51           | 0.49                    | 0.45             | 0.50                    | 0.52             | 0.51                    | 0.51                | 0.53                    | 0.46             | 0.45               | 0.47            | 0.45               | 0.45             | 0.46                    | 0.43             | 0.45                            | 0.53           | 0.53                    | 0.53            | 0.53                           |
| de_test_neutral -  | 0.32                                      | 0.31                    | 0.33             | 0.32                    | 0.50             | 0.50         | 0.51                  | 0.49                    | 0.55           | 0.55                    | 0.54             | 0.55                    | 0.47             | 0.49                    | 0.46             | 0.49                    | 0.49           | 0.47                    | 0.43             | 0.48                    | 0.49             | 0.48                    | 0.48                | 0.50                    | 0.43             | 0.42               | 0.44            | 0.42               | 0.42             | 0.43                    | 0.40             | 0.42                            | 0.48           | 0.48                    | 0.49            | 0.49                           |
| de_test_toxic -    | 0.31                                      | 0.31                    | 0.32             | 0.32                    | 0.49             | 0.50         | 0.51                  | 0.48                    | 0.55           | 0.55                    | 0.54             | 0.54                    | 0.46             | 0.48                    | 0.46             | 0.49                    | 0.48           | 0.47                    | 0.43             | 0.48                    | 0.48             | 0.48                    | 0.48                | 0.49                    | 0.43             | 0.42               | 0.43            | 0.42               | 0.41             | 0.42                    | 0.40             | 0.41                            | 0.48           | 0.48                    | 0.48            | 0.48                           |
| de_train_neutral - | 0.30                                      | 0.30                    | 0.31             | 0.31                    | 0.48             | 0.48         | 0.49                  | 0.47                    | 0.54           | 0.54                    | 0.54             | 0.53                    | 0.47             | 0.49                    | 0.47             | 0.50                    | 0.48           | 0.47                    | 0.43             | 0.47                    | 0.46             | 0.46                    | 0.46                | 0.47                    | 0.43             | 0.42               | 0.43            | 0.42               | 0.41             | 0.42                    | 0.40             | 0.41                            | 0.45           | 0.45                    | 0.46            | 0.45                           |
| de_train_toxic -   | 0.32                                      | 0.31                    | 0.33             | 0.32                    | 0.49             | 0.49         | 0.51                  | 0.48                    | 0.55           | 0.54                    | 0.53             | 0.55                    | 0.46             | 0.48                    | 0.45             | 0.48                    | 0.48           | 0.46                    | 0.42             | 0.47                    | 0.48             | 0.47                    | 0.48                | 0.49                    | 0.42             | 0.41               | 0.42            | 0.41               | 0.41             | 0.42                    | 0.39             | 0.41                            | 0.48           | 0.48                    | 0.48            | 0.48                           |
| en_test_neutral -  | 0.29                                      | 0.28                    | 0.30             | 0.29                    | 0.46             | 0.47         | 0.48                  | 0.46                    | 0.47           | 0.46                    | 0.47             | 0.46                    | 0.53             | 0.54                    | 0.52             | 0.55                    | 0.49           | 0.49                    | 0.45             | 0.48                    | 0.44             | 0.45                    | 0.46                | 0.45                    | 0.43             | 0.42               | 0.43            | 0.42               | 0.41             | 0.43                    | 0.40             | 0.42                            | 0.41           | 0.42                    | 0.43            | 0.42                           |
| en_test_toxic -    | 0.29                                      | 0.29                    | 0.30             | 0.29                    | 0.47             | 0.48         | 0.49                  | 0.47                    | 0.49           | 0.48                    | 0.49             | 0.48                    | 0.54             | 0.57                    | 0.54             | 0.57                    | 0.49           | 0.50                    | 0.46             | 0.49                    | 0.45             | 0.46                    | 0.47                | 0.46                    | 0.43             | 0.42               | 0.44            | 0.42               | 0.42             | 0.43                    | 0.41             | 0.42                            | 0.42           | 0.43                    | 0.44            | 0.42                           |
| en_train_neutral - | 0.29                                      | 0.29                    | 0.30             | 0.29                    | 0.45             | 0.47         | 0.47                  | 0.45                    | 0.46           | 0.46                    | 0.47             | 0.45                    | 0.52             | 0.54                    | 0.52             | 0.54                    | 0.48           | 0.49                    | 0.44             | 0.48                    | 0.44             | 0.44                    | 0.45                | 0.44                    | 0.42             | 0.41               | 0.43            | 0.41               | 0.41             | 0.42                    | 0.40             | 0.41                            | 0.41           | 0.42                    | 0.43            | 0.41                           |
| en_train_toxic -   | 0.30                                      | 0.30                    | 0.31             | 0.30                    | 0.47             | 0.48         | 0.49                  | 0.47                    | 0.49           | 0.49                    | 0.50             | 0.48                    | 0.55             | 0.57                    | 0.54             | 0.59                    | 0.50           | 0.51                    | 0.46             | 0.50                    | 0.46             | 0.46                    | 0.47                | 0.46                    | 0.43             | 0.42               | 0.44            | 0.42               | 0.42             | 0.43                    | 0.41             | 0.42                            | 0.42           | 0.43                    | 0.44            | 0.42                           |
| es_test_neutral -  | 0.31                                      | 0.30                    | 0.32             | 0.31                    | 0.51             | 0.52         | 0.53                  | 0.51                    | 0.49           | 0.48                    | 0.48             | 0.48                    | 0.49             | 0.49                    | 0.48             | 0.50                    | 0.54           | 0.53                    | 0.49             | 0.53                    | 0.49             | 0.49                    | 0.49                | 0.50                    | 0.49             | 0.48               | 0.50            | 0.49               | 0.47             | 0.49                    | 0.46             | 0.47                            | 0.48           | 0.48                    | 0.49            | 0.48                           |
| es_test_toxic -    | 0.30                                      | 0.29                    | 0.30             | 0.30                    | 0.49             | 0.50         | 0.50                  | 0.49                    | 0.47           | 0.47                    | 0.47             | 0.46                    | 0.49             | 0.50                    | 0.49             | 0.51                    | 0.53           | 0.53                    | 0.49             | 0.52                    | 0.47             | 0.46                    | 0.47                | 0.47                    | 0.49             | 0.48               | 0.49            | 0.48               | 0.46             | 0.48                    | 0.45             | 0.47                            | 0.44           | 0.45                    | 0.45            | 0.44                           |
| es_train_neutral - | 0.27                                      | 0.27                    | 0.28             | 0.28                    | 0.45             | 0.46         | 0.47                  | 0.45                    | 0.43           | 0.43                    | 0.43             | 0.42                    | 0.45             | 0.46                    | 0.44             | 0.46                    | 0.49           | 0.49                    | 0.46             | 0.48                    | 0.43             | 0.43                    | 0.44                | 0.44                    | 0.45             | 0.44               | 0.45            | 0.44               | 0.43             | 0.44                    | 0.42             | 0.43                            | 0.42           | 0.42                    | 0.43            | 0.42                           |
| es_train_toxic -   | 0.30                                      | 0.30                    | 0.31             | 0.31                    | 0.50             | 0.51         | 0.52                  | 0.50                    | 0.48           | 0.48                    | 0.47             | 0.47                    | 0.48             | 0.49                    | 0.48             | 0.50                    | 0.53           | 0.52                    | 0.48             | 0.52                    | 0.49             | 0.48                    | 0.49                | 0.49                    | 0.48             | 0.48               | 0.49            | 0.48               | 0.46             | 0.48                    | 0.45             | 0.47                            | 0.47           | 0.48                    | 0.48            | 0.48                           |
| hi_test_neutral -  | 0.34                                      | 0.33                    | 0.35             | 0.34                    | 0.53             | 0.53         | 0.54                  | 0.52                    | 0.49           | 0.48                    | 0.46             | 0.48                    | 0.44             | 0.45                    | 0.44             | 0.46                    | 0.49           | 0.47                    | 0.43             | 0.49                    | 0.61             | 0.59                    | 0.59                | 0.62                    | 0.44             | 0.43               | 0.44            | 0.43               | 0.44             | 0.45                    | 0.42             | 0.44                            | 0.54           | 0.54                    | 0.54            | 0.54                           |
| hi_test_toxic -    | 0.33                                      | 0.32                    | 0.34             | 0.33                    | 0.52             | 0.52         | 0.53                  | 0.51                    | 0.48           | 0.48                    | 0.46             | 0.47                    | 0.45             | 0.46                    | 0.44             | 0.46                    | 0.49           | 0.46                    | 0.43             | 0.48                    | 0.59             | 0.59                    | 0.59                | 0.61                    | 0.43             | 0.42               | 0.43            | 0.42               | 0.43             | 0.44                    | 0.41             | 0.43                            | 0.52           | 0.52                    | 0.52            | 0.52                           |
| hi_train_neutral - | 0.33                                      | 0.32                    | 0.34             | 0.33                    | 0.52             | 0.52         | 0.53                  | 0.51                    | 0.48           | 0.48                    | 0.46             | 0.48                    | 0.46             | 0.47                    | 0.45             | 0.47                    | 0.49           | 0.47                    | 0.44             | 0.49                    | 0.59             | 0.59                    | 0.59                | 0.60                    | 0.44             | 0.43               | 0.44            | 0.43               | 0.44             | 0.45                    | 0.42             | 0.44                            | 0.51           | 0.51                    | 0.52            | 0.52                           |
| hi_train_toxic -   | 0.35                                      | 0.34                    | 0.36             | 0.35                    | 0.54             | 0.54         | 0.55                  | 0.53                    | 0.50           | 0.49                    | 0.47             | 0.49                    | 0.45             | 0.46                    | 0.44             | 0.46                    | 0.50           | 0.47                    | 0.44             | 0.49                    | 0.62             | 0.61                    | 0.60                | 0.64                    | 0.44             | 0.43               | 0.45            | 0.43               | 0.44             | 0.45                    | 0.42             | 0.44                            | 0.56           | 0.56                    | 0.56            | 0.56                           |
| ru_test_neutral -  | 0.26                                      | 0.25                    | 0.26             | 0.26                    | 0.46             | 0.47         | 0.48                  | 0.46                    | 0.43           | 0.43                    | 0.43             | 0.42                    | 0.43             | 0.43                    | 0.42             | 0.43                    | 0.49           | 0.49                    | 0.45             | 0.48                    | 0.44             | 0.43                    | 0.44                | 0.44                    | 0.59             | 0.59               | 0.58            | 0.59               | 0.53             | 0.56                    | 0.51             | 0.54                            | 0.42           | 0.42                    | 0.42            | 0.42                           |
| ru_test_toxic -    | 0.25                                      | 0.25                    | 0.26             | 0.25                    | 0.45             | 0.46         | 0.47                  | 0.45                    | 0.42           | 0.42                    | 0.42             | 0.41                    | 0.42             | 0.42                    | 0.41             | 0.42                    | 0.48           | 0.48                    | 0.44             | 0.48                    | 0.43             | 0.42                    | 0.43                | 0.43                    | 0.59             | 0.59               | 0.58            | 0.59               | 0.53             | 0.56                    | 0.51             | 0.53                            | 0.41           | 0.41                    | 0.41            | 0.41                           |
| ru_train_neutral - | 0.26                                      | 0.26                    | 0.27             | 0.26                    | 0.47             | 0.47         | 0.48                  | 0.47                    | 0.44           | 0.43                    | 0.43             | 0.42                    | 0.43             | 0.44                    | 0.43             | 0.44                    | 0.50           | 0.49                    | 0.45             | 0.49                    | 0.44             | 0.43                    | 0.44                | 0.45                    | 0.58             | 0.58               | 0.58            | 0.58               | 0.53             | 0.56                    | 0.51             | 0.54                            | 0.42           | 0.42                    | 0.43            | 0.42                           |
| ru_train_toxic -   |   |                         |                  |                         |                  |              | 0.47                  | 0.45                    | 0.42           | 0.42                    | 0.42             |                         |                  | 0.42                    |                  |                         |                |                         |                  | 0.48                    | 0.43             |                         |                     |                         |                  | 0.59               |                 |                    |                  |                         |                  |                                 |                |                         | 0.41            |                                |
| uk_test_neutral -  |   |                         |                  |                         |                  |              | 0.47                  |                         |                | 0.41                    | 0.41             |                         |                  | 0.42                    |                  |                         |                |                         |                  | 0.46                    | 0.44             |                         | 0.44                |                         |                  | 0.53               |                 |                    |                  |                         |                  |                                 |                |                         | 0.42            |                                |
| uk_test_toxic -    |   |                         |                  |                         |                  |              |                       |                         |                | 0.42                    | 0.42             |                         |                  | 0.43                    |                  |                         |                |                         |                  | 0.48                    |                  |                         |                     |                         |                  | 0.56               |                 |                    |                  |                         |                  |                                 |                |                         | 0.43            |                                |
| uk_train_neutral - |   |                         |                  |                         |                  | 0.44         |                       |                         | 0.40           |                         | 0.40             |                         |                  | 0.41                    |                  |                         |                |                         |                  |                         | 0.42             |                         |                     |                         | 0.51             |                    |                 |                    |                  |                         |                  |                                 |                |                         | 0.40            |                                |
| uk_train_toxic -   |   |                         |                  |                         |                  |              |                       |                         |                | 0.41                    | 0.41             |                         |                  | 0.42                    |                  |                         |                | 0.47                    | 0.43             | 0.47                    | 0.44             | 0.43                    | 0.44                | 0.44                    | 0.54             | 0.53               | 0.54            | 0.54               | 0.56             | 0.58                    |                  |                                 |                |                         | 0.41            |                                |
| zh_test_neutral -  |   |                         |                  |                         |                  | 0.54         | 0.55                  | 0.53                    | 0.48           | 0.48                    | 0.45             | 0.48                    | 0.41             |                         |                  | 0.42                    |                |                         | 0.42             |                         |                  |                         |                     |                         |                  | 0.41               |                 |                    |                  |                         |                  |                                 |                |                         | 0.65            |                                |
| zh_test_toxic -    |   |                         |                  |                         |                  |              | 0.56                  |                         | 0.48           | 0.48                    | 0.45             |                         | 0.42             |                         |                  | 0.43                    |                | 0.45                    |                  |                         |                  |                         |                     |                         |                  | 0.41               |                 |                    |                  |                         |                  |                                 |                |                         | 0.66            |                                |
| zh_train_neutral - |   |                         |                  |                         |                  | 0.54         | 0.56                  | 0.53                    | 0.49           | 0.48                    | 0.46             | 0.48                    | 0.43             |                         | 0.43             |                         |                |                         |                  | 0.48                    |                  |                         |                     |                         |                  | 0.41               |                 |                    |                  |                         |                  |                                 |                |                         |                 |                                |
| zh_train_toxic -   | 0.32<br> <br>                             | 0.31<br>-<br>. <u>U</u> | 0.33<br>         | 0.32<br>-<br>. <u>U</u> | 0.55<br><u>-</u> | 0.54         | 0.56<br><u>ا</u><br>ھ | 0.53<br>-<br>. <u>U</u> | U.49<br>       | 0.48<br>-<br>. <u>D</u> | 0.45<br><u>-</u> | 0.48<br>-<br>. <u>D</u> | 0.42<br><u>-</u> | 0.42<br>-<br>. <u>U</u> | 0.41<br><u>r</u> | 0.42<br>-<br>. <u>O</u> | 0.48<br>-<br>- | 0.44<br>-<br>. <u>U</u> | 0.42<br><u>-</u> | 0.48<br>-<br>. <u>C</u> | 0.54<br><u>-</u> | 0.52<br>-<br>. <u>U</u> | 0.52<br>_<br>_<br>_ | 0.56<br>-<br>. <u>U</u> | 0.42<br><u>I</u> | 0.41<br>. <u>U</u> | <u>_</u>        | 0.41<br>. <u>U</u> | 0.42<br><u>-</u> | 0.42<br>-<br>. <u>D</u> | 0.39<br><u>-</u> | 0.41<br> <br> -<br>  <u>.</u> 2 | 1<br>          | 0.66<br>-<br>. <u>U</u> | 0.66<br>-<br>ام | 0.66<br><u>-</u><br>. <u>U</u> |
|                    | am_test_neutra                            | am_test_toxi            | am_train_neutral | am_train_toxi           | ar_test_neutra   | ar_test_toxi | ar_train_neutra       | ar_train_toxi           | de_test_neutra | de_test_toxi            | de_train_neutra  | de_train_toxi           | en_test_neutra   | en_test_toxio           | en_train_neutra  | en_train_toxi           | es_test_neutra | es_test_toxion          | es_train_neutra  | es_train_toxi           | hi_test_neutra   | hi_test_toxio           | hi_train_neutra     | hi_train_toxi           | ru_test_neutra   | ru_test_toxic      | ru_train_neutra | ru_train_toxi      | uk_test_neutra   | uk_test_toxi            | uk_train_neutra  | uk_train_toxi                   | zh_test_neutra | zh_test_toxio           | zh_train_neutra | zh_train_toxi                  |

- 0.5