|  |  |
| --- | --- |
| Dataset | Krippendorff's α |
| WorldTree v2 | 0.83 (0.81, 0.85) |
| CommonsenseQA | 0.71 (0.68, 0.73) |
| OpenBookQA | 0.65 (0.63, 0.68) |
| StrategyQA | 0.34 (0.3, 0.39) |
| MedMCQA | 0.31 (0.27, 0.34) |
| MedQA | 0.22 (0.19, 0.25) |