Table 1: Calibration Performance Comparison: Ranking and Probability Metrics (Bold = Best per Metric per Dataset)

| Dataset | Model | Method | MRR (↑) | H@1 (↑) | H@10 (↑) | ECE (↓) | Brier (↓) | ACE (↓) |
|---|---|---|---|---|---|---|---|---|
| WN18RR | MC Dropout | Uncalibrated | 0.4353 | 0.3952 | 0.5093 | 0.2344 | 0.1997 | 0.4213 |
| | | Temp Scaling | 0.4350 | 0.3952 | 0.5110 | 0.1977 | 0.1780 | 0.3457 |
| | | Platt Scaling | 0.4356 | 0.3958 | 0.5080 | 0.0713 | 0.0980 | 0.1562 |
| | | Isotonic | 0.1736 | 0.0549 | 0.4100 | 0.1287 | 0.1239 | 0.3370 |
| | Deep Ensemble | Uncalibrated | 0.4559 | 0.4097 | 0.5391 | 0.2337 | 0.1959 | 0.4122 |
| | | Temp Scaling | **0.4559** | **0.4097** | **0.5391** | 0.2049 | 0.1752 | 0.3963 |
| | | Platt Scaling | 0.4470 | 0.3826 | 0.5389 | **0.0141** | **0.0835** | **0.0626** |
| | | Isotonic | 0.3680 | 0.2589 | 0.5246 | 0.1226 | 0.000 | 0.000 |
| FB15K-237 | MC Dropout | Uncalibrated | 0.2530 | 0.1624 | 0.4400 | 0.2518 | 0.2253 | 0.3958 |
| | | Temp Scaling | 0.2523 | 0.1607 | 0.4425 | 0.2136 | 0.1614 | 0.2481 |
| | | Platt Scaling | 0.2529 | 0.1616 | 0.4409 | 0.0311 | 0.0739 | 0.0568 |
| | | Isotonic | 0.2143 | 0.1183 | 0.3835 | 0.1732 | 0.1148 | 0.3249 |
| | Deep Ensemble | Uncalibrated | 0.2864 | 0.1877 | 0.4897 | 0.1930 | 0.1709 | 0.4007 |
| | | Temp Scaling | 0.2864 | 0.1877 | 0.4897 | 0.1843 | 0.1310 | 0.2787 |
| | | Platt Scaling | **0.2864** | **0.1877** | **0.4897** | **0.0248** | **0.0589** | **0.0603** |
| | | Isotonic | 0.2521 | 0.1438 | 0.4463 | 0.1580 | 0.1007 | 0.3068 |