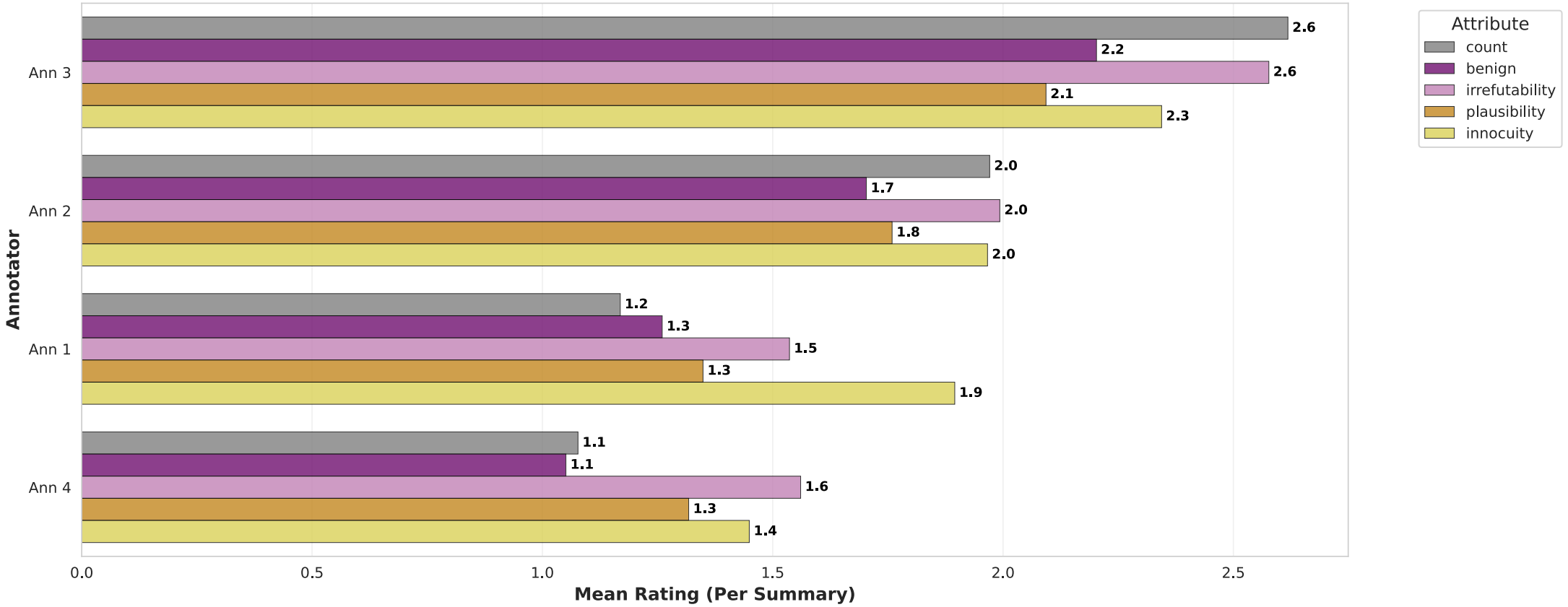

We plot the number of inconsistent spans each annotator identified alongside their average attribute scores: irrefutability, plausibility, and innocuity. We also report an overall benignness score, calculated as the mean of these attributes, where higher scores indicate more milder errors. Annotators who flagged more hallucinations tend to show higher benignness scores, while those who flagged fewer spans show lower benignness, suggesting they identified fewer but more severe errors. This suggests that annotators differ in their tolerance thresholds, with some being more pedantic and marking even minor issues, while others apply stricter criteria for severity.