Appendix: Human Baselines Checklist

0.0. Paper Information

0.1 **Paper Title**

0.2 **Paper Link**

0.3 **Publication Year**

0.4 **Publication Venue**

0.5 **Type of Eval**
Select all that apply

- Knowledge
- Capabilities
- Propensity
- Agent

0.6 **Mode of Eval**
Select all that apply

- Text
- Visual (photo/video)
- Audio
- Other

0.7 **Language of Eval**
Select all that apply from list

0.8 **Evaluation Dataset Size**: What is the total number of items in the evaluation dataset?

0.9 **AI Test Set Size**: What is the number of items that the AI evaluation is run on? (Default same as Q0.8)

0.10 **AI Samples per Item**: What is the number of AI responses ("samples" or "runs") that is collected for each item? (Default 1)

0.1. Baseline Design & Implementation

1.1 **Number of Baseliners**: How many baseliners were there total?

1.2 **Baseline Test Set Size**: What is the number of items that the human baseline is run on? (i.e., how many of the questions do the baseliners collectively answer?) (Default same as Q0.9)

1.2.1 **Baseline Test Set Sampling Strategy**: If the baseline is only run on a sample of the total dataset: what is the sampling strategy behind how the items were selected? E.g., simple random sampling, stratified sampling, etc.

1.3 **Baseline Samples per Item**: What was the number of human baseliner responses that is collected for each item? (Default Q1.1 $*$ Q1.4 / Q1.2, or 1 if Q1.1 or Q1.4 unreported)

1.4 **Items per Baseliner**: What is the number of items that each baseliner responded to?

1.5 **Explicit Human/AI Adjustment**: Does the eval/baseline instructions and items account for both humans and AI models completing the evals items (questions/tasks)? E.g., do the authors of the eval explicitly state that the eval is designed so as not to advantage either humans or AI models?
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

1.6 **Iterative Design**: Was the experimental setup of the baseline iteratively designed with participatory methods? E.g., was there a pilot study, expert validation of the items, etc.?
Select one of: "Yes", "Partial", "No", "Unknown/Unreported", or "N/A"

1.7 **Amount of Effort**: Does the baseline control for the amount of effort by human baseliners and AIs? E.g., in terms of cost, time, etc.
Select one of: "Yes", "Partial", "No", "Unknown/Unreported", or "N/A"

1.8 **Power Analysis**: Did the authors conduct power analysis in order to determine baseline size?
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

1.8.1 **Minimum Detectable Effect Size**: if yes, what is the minimum detectable effect size and power?

1.9 **Ethics Review**: Was the study approved or exempted by an IRB, or did it undergo other ethics review?
Select one of: "Yes", "Partial", "No", "Unknown/Unreported", or "N/A"

1.10 **Pre-Registration**: Was the baseline/eval design pre-registered? I.e., a plan detailing the experimental setup that is publicly registered online before running the experiment (e.g., on OSF, COS, etc.)
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

0.2. Baseliner Recruitment

2.1 **Population of Interest Identification**: Does the reporting identify human populations for which these results may be valid, i.e., a human population of interest?
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

2.1.1 **Population of Interest Identification Criteria**: Which of the following factors were used to scope the target human population of interest?
Select all that apply

- Expertise

- Education
- Language
- Gender/sex
- Race
- Socioeconomic status
- Age
- Disabilities/impairments
- Political orientation
- Digital literacy (Prior experience with computers)
- AI literacy (Prior experience with AI tools)
- Baseline experience: Prior experience with AI evals/doing human baselines
- Other (specify)

2.2 **Baseliner Sampling Strategy**: How were the human baseliners recruited?
Select one of the below

- Crowdsource
- Convenience sample
- Simple random sample
- Stratified random sample
- Other (specify)
- Unknown/unreported

2.3 **Quality Control in Recruitment**: Were human baseliners pre-qualified or excluded during the recruitment process for any reason?
Select one of: "Yes" (Default), "Partial", "No", "Unknown/Unreported", or "N/A"

2.3.1 **Quality Control Criteria for Baseliners**: If yes: please describe the inclusion/exclusion criteria for human baseliners (e.g., pre-tests, expert judgements/filtering, quality scores or ratings on crowdwork platforms, number of tasks completed on crowdwork platforms). Data quality checks that occurred after baseliners were recruited should be reported in the implementation section (e.g., attention checks in a survey).

2.3.2 **Recruitment Exclusion Rate**: If yes: how many baseliners were excluded from the final baseline based on these criteria?

2.4 **Author Baseliners**: Did the authors or members of the research team also serve as human baseliners?
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

2.5 **Baseliner Train/Test Contamination**: Did the recruitment process exclude baseliners who had been exposed to the eval questions previously?
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

2.6 **Baseliner Training**: Did the human baseliners receive training for the baseline? Training should be distinct from the reported data, e.g., a tutorial completed before answering baseline questions
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

2.6.1 **Baseliner Training Type**: If yes: describe the type of training received (e.g., tutorial, shown examples, etc.)

2.6.2 **Baseliner Training Compensation**: If yes: were the baseliners compensated for the training?
Select one of: "Yes", "Partial", "No", "Unknown/Unreported", or "N/A"

2.6.2.1 **Baseliner Training Compensation Amount**: If yes: list the compensation per baseliner (preferably $ / hour, otherwise total $ amount if stated)

2.7 **Baseliner Testing Compensation**: Were the human baseliners compensated for completing the baseline?
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

2.7.1 **Baseliner Testing Compensation Amount**: If yes: how much was compensation? (preferably $ / hour, otherwise total $ amount if stated)

2.7.2 **Baseliner Testing Performance Bonus**: If yes: was a performance bonus offered to baseliners?
Select one of: "Yes" (Default), "Partial", "No", "Unknown/Unreported", or "N/A"

2.7.2.1 **Baseliner Testing Performance Bonus Amount**: If yes: how much was the performance bonus, and how was it determined?

2.7.3 **Baseliner Testing Compensation Structure**: If yes: were compensation rates and structures constant across baseliners? E.g., respond no if baseliners were paid differently according to expertise.
Select one of: "Yes", "Partial", "No", "Unknown/Unreported", or "N/A"

2.7.3.1 **Baseliner Testing Compensation Structure Details**: If not compensated equally: how were compensation amounts determined?

0.3. Baseline Execution

3.1 **Instrument Length**: How many items did the human baseliners complete in a single sitting/session?

I.e., what is the length of the baseliner "context window" in units of items?

3.1.1 **Item Randomization**: If not 1: was the order of the questions randomized?

3.2 **Quality Control in Execution**: Were quality checks implemented or data cleaned/excluded during the data collection process (i.e., after baseliners were recruited)? E.g., were there any exclusion criteria for baseliner responses due to data quality such as attention check questions, honeypot questions, filtering out responders who completed the eval too quickly, screen recording, etc.
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

3.2.1 **Quality Control in Execution Criteria**: If yes: what factors were used to determine data quality or to exclude low-quality data?

3.2.2 **Execution Exclusion Rate**: If yes: how many samples were excluded from the final baseline based on these criteria?

3.3 **UI Equivalence**: Did the human baseliners and AIs have access to the same UI for each item?
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

3.3.1 **GUI vs. API**: Check this box if the humans had access to a graphical UI and the AIs only had API inputs
Checkbox item (Unchecked by default)

3.3.2 **UI Equivalence Adjustment**: If no: does the eval attempt to adjust for the differences?
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

3.4 **Instruction Equivalence**: Did the human baseliners and AIs have access to the same instructions/prompt/question for each item?
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

3.4.1 **Instruction Equivalence Adjustment**: If no: does the eval attempt to adjust for the differences?
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

3.5 **Tool Access Equivalence**: Did the human baseliners and AIs have access to the same (technical) tools for each item? Respond yes if neither group had access to external tools; respond yes if the human had internet access and the AI did not (but was trained on the internet)
Select one of: "Yes" (Default), "Partial", "No", "Unknown/Unreported", or "N/A"

3.5.1 **Tool Access Equivalence Enforcement**: If human baseliners' tool access was limited: was there an oversight mechanism for ensuring that the human baseliners only used the tools permitted? E.g., enforcement of AI tool use ban
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

3.6 **Explanations**: Did the eval/baseline collect explanations from the human baseliners, after the evaluation was conducted? I.e., explanations for why the human participants responded the way they did
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

0.4. Baseline Analysis

4.1 **Statistical Significance**: Did the eval test for statistically significant differences between AI and human performance?
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

4.1.1 **Statistical Significance Test**: If yes: what statistical test was used?

4.2 **Uncertainty Estimate**: Did the paper present a measure of uncertainty for the AI and human baseline results? E.g., confidence intervals, variance, pooled/clustered standard errors, etc.?
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

4.2.1 **Estimate Type**: Is the reported baseline a point estimate, an interval estimate, or a distribution?
Select all that apply

- Point estimate (Default)
- Interval estimate
- Distribution estimate

4.3 **Evaluation Metric Equivalence**: Was the same evaluation metric measured/compared for both humans and AIs? Respond "no" if, e.g., the human baseline is majority vote but the AI baseline is not
Select one of: "Yes" (Default), "Partial", "No", "Unknown/Unreported", or "N/A"

4.4 **Evaluation Scoring Criteria Equivalence**: Was the same scoring rubric used for both AI and human results?
Select one of: "Yes" (Default), "Partial", "No", "Unknown/Unreported", or "N/A"

4.5 **Evaluation Scoring Method Equivalence**: Was the same scoring method used for both AI and human results? E.g., human grading, LLM as a judge

Select one of: "Yes" (Default), "Partial", "No", "Unknown/Unreported", or "N/A"

4.6 **Quality Control Robustness**: If quality controls were implemented: are analyses robust to different choices of exclusion criteria? E.g., do the authors state that the results don't change when including/excluding incomplete data?
Select one of: "Yes", "Partial", "No", "Unknown/Unreported", or "N/A"

0.5. Baseline Documentation

5.1 **Additional Reporting**: Were the following reported?

5.1.1 **Reporting Sample Demographics**: Demographics for human baseliners, e.g., race, gender, etc. Respond yes only if within-sample demographics are reported; e.g., respond no if the paper only reports that 100% of the sample is based in the U.S.
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

5.1.2 **Reporting Baseline Instructions**: Instructions/guidelines given to human baseliners
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

5.1.3 **Reporting Time to Completion**: Time to completion for the eval items
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

5.1.4 **AI Tool Versions**: AI tools and versions (if baseliners had AI access)
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

5.1.5 **Completion Rate**: How many human baseliners were recruited but did not complete the tasks?
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

5.2 **Baseline Data Availability**: Is the (anonymized) human baseline data publicly available?
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

5.2.1 **Individual Baseline Data Availability**: If yes: is data available at the individual baseliner level? I.e., can you tell from the dataset which baseliners were responsible for which questions?
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

5.2.2 **Baseline Data Non-Availability Justification**: If no: is there a reasonable justification for non-disclosure of the baseline dataset? E.g., privacy concerns, safety/security concerns, company policy, etc.

5.3 **Experimental Materials Availability**: Are experimental materials used to implement the eval/baseline publicly available?
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

5.4 **Analysis Code Availability**: Is the code used to analyze the eval/baseline publicly available?
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"