Guidelines for Human Validation of the IFB Dataset

Version: Training / validation protocol draft
Dataset sample: n = 300 instances
Evaluators: Two independent human validators + one LLM-as-a-judge pass
Important note: The provided simulated workbook is for workflow testing only. Final paper results must be based on real evaluator inputs.

1. Purpose of the validation

The purpose of the validation is to assess whether each sampled IFB dataset instance is suitable for model evaluation. Validators judge the quality of the gold answer, the clarity of the prompt, and the consistency between the prompt, context, and expected answer.

2. What validators should review

Field	What to check
prompt	Is the question/task clear, answerable, and not ambiguous?
context	If context is provided, does it contain enough information to answer the prompt? For FiQA/NO_CONTEXT tasks, absence of context is expected.
expected	Is the gold answer correct, complete, and in the right format?
task_type / answer_format	Does the expected answer match the intended task and output format?
source	Use only for orientation; do not assume correctness from source name alone.

3. Correctness coding

Use evaluator_1_correct or evaluator_2_correct depending on your assigned column.

Code	Meaning	Use when
1	Valid / correct	The expected answer is correct and supported by the prompt/context.
0	Problematic / incorrect	The expected answer is wrong, unsupported, ambiguous, incomplete, or incompatible with the task format.

4. Clarity coding

Use evaluator_1_clarity or evaluator_2_clarity depending on your assigned column. Clarity measures how easy it is to understand what the instance asks, not whether the model would answer correctly.

Score	Definition
1	Very unclear or unanswerable.
2	Mostly unclear; major ambiguity.
3	Partly clear; answerable but with noticeable ambiguity.
4	Clear with minor wording issues.
5	Fully clear and unambiguous.

5. Comments

Comments are optional for clearly valid instances, but mandatory for any instance marked 0. Keep comments short and actionable.

- ambiguous prompt
- insufficient context
- unclear or incomplete gold answer
- context-answer mismatch
- format inconsistency
- possible duplicate or malformed text

6. Adjudication procedure

1. Both validators complete their evaluation independently.
2. Disagreements are identified after both sheets are returned.
3. A final reviewer resolves disagreements using prompt, context, expected answer, and comments.
4. final_decision = 1 means keep; final_decision = 0 means fix or remove before final release.
   https://en.wikipedia.org/wiki/Krippendorff%27s_alpha - annotatoor agreement

7. LLM-as-a-judge pass

The LLM-as-a-judge pass is used as a third, scalable consistency check. It should not replace human validation. It can flag likely inconsistencies and provide an additional comparison point against human judgments.

Recommended judge prompt:

> *You are evaluating the quality of a financial benchmark dataset instance. Given the prompt, context, and expected answer, decide whether the expected answer is correct and whether the prompt is clear. Return: correct: 0/1; clarity: 1-5; comment: short reason.*

8. Reporting

Report at minimum: sample size, human-human agreement, LLM-human agreement, final accepted rate, issue-type distribution, and mean clarity scores. Do not report simulated values as real validation results.

Appendix: suggested spreadsheet columns

validation_id, prompt, context, expected, task_type, answer_format, source, evaluator_1_correct, evaluator_2_correct, llm_correct, evaluator_1_clarity, evaluator_2_clarity, llm_clarity, evaluator_1_comment, evaluator_2_comment, llm_comment, final_decision, gold_quality_issue_type, adjudication_note