Human Validation and Annotation Protocol

Regulatory AML/KYC Benchmark (RegBench-463)

1. Purpose and scope

This document defines the annotation protocol followed by both human evaluators during the independent validation of the Regulatory AML/KYC Benchmark (RegBench-463), described in Appendix B of the accompanying paper. The protocol mirrors, with domain-specific adaptations, the validation procedure applied to the Integrated Financial Benchmark (IFB) and documented in human_validation_guidelines_IFB.docx. It applies to the stratified validation sample of n = 150 instances (32.4% of the released benchmark), drawn proportionally to source representation (EU 48%, FATF 27%, Basel 25%).

Each instance in the validation sample presents a short regulatory excerpt (the context field), a scenario-reasoning prompt, and a gold action produced by the automated rule-matching pipeline described in Appendix B.3. Evaluators do not assess model outputs; they assess the quality of the benchmark instance itself, i.e., whether the gold action is a correct, well-specified consequence of the regulatory excerpt and whether the prompt is answerable without ambiguity.

2. Reviewers and independence

The validation protocol employs two independent human evaluators, following the same separation-of-roles rationale used for IFB validation: dataset-construction involvement is kept distinct from independent quality assessment wherever possible, to reduce construction bias in the reported agreement statistics.

- Evaluator 1 participated in the construction of RegBench-463 (PDF extraction, keyword scoring configuration, and rule-template authoring) and has professional experience in financial and regulated-IT environments.
- Evaluator 2 was not involved in the construction pipeline and reviewed each instance solely on the basis of the released prompt, context, and gold action, without access to the rule-matching logic that produced them.

Both evaluators completed the annotation independently, without communication or access to each other's judgments, before any adjudication step took place. All inter-annotator agreement statistics reported in Appendix B.5 therefore reflect genuinely independent assessments.

3. Annotation criteria

For every instance in the validation sample, each evaluator records three judgments, summarised in Table S1 below.

3.1 Correctness (binary)

Correctness asks whether the gold action is an accurate, complete, and internally consistent consequence of the regulatory excerpt shown in the context field. A gold action is coded 1 (correct) if it correctly identifies the obligation class implied by the excerpt and does not contradict any part of the source text. It is coded 0 (problematic) if the action is incorrect, materially incomplete relative to the scope of the excerpt, inconsistent with the prompt, or so generic that it fails to reflect the specific obligation described in the chunk (a known artefact of the fallback templates described in Appendix B.3, Stage 3).

3.2 Clarity (five-point ordinal scale)

Clarity assesses whether the prompt is answerable on its own terms, independent of whether the gold action is correct. The five-point scale is identical to the one used for IFB validation (Table A7), reproduced here as Table S2 for completeness.

3.3 Issue typing (categorical, mandatory when correctness = 0)

Whenever an instance is coded 0 on correctness, the evaluator selects one issue type from a fixed list and provides a short free-text comment. The issue taxonomy was derived empirically from the disagreement patterns observed during pilot annotation and is summarised in Table S3.

Code	Meaning
1 (correct)	Gold action accurately, completely, and consistently reflects the obligation described in the regulatory excerpt.
0 (problematic)	Gold action is incorrect, materially incomplete, inconsistent with the excerpt, or reduced to a non-specific category-level fallback; comment required.

Table S1. Correctness coding scheme.

Clarity	Meaning
5	Fully clear and unambiguous; the entity and required action are evident from the prompt and excerpt alone.
4	Clear, with only minor wording issues.
3	Generally answerable but contains some ambiguity (e.g., unclear which entity the obligation applies to).
2	Substantial ambiguity or lack of clarity; multiple obligations bundled in one excerpt.
1	Unclear or effectively unanswerable from the excerpt provided.

Table S2. Clarity coding scheme (identical scale to IFB validation, Table A7).

Issue type	Description
gold_too_generic	Gold action is the category-level fallback template rather than a chunk-specific obligation.

gold_too_narrow_for_chunk	Source passage covers a broader obligation than the gold action reflects.
multi_obligation_overreach	Chunk bundles multiple distinct obligations; gold action captures only one.
ambiguous_prompt	Prompt does not specify which entity (parent, branch, third party) the action applies to.
context_answer_inconsistency	Minor mismatch between a specific risk factor in the context and the generic gold action.
format_inconsistency	Scenario reads as a yes/no determination but is scored as free-text scenario_reasoning.
insufficient_context	Chunk is too short to unambiguously support the scenario classification assigned by the pipeline.

Table S3. Gold-quality issue taxonomy, derived from Appendix A.6/A.7's automated-audit categories and adapted to scenario-reasoning instances.

4. Workbook fields

Annotations are recorded in regulatory_human_llm_validation_150.xlsx, one row per validation instance, with the following fields: VAL_ID, ID, source, category, question_type, prompt, context_abridged, expected, evaluator_1_correct, evaluator_1_clarity, evaluator_1_comment, evaluator_2_correct, evaluator_2_clarity, evaluator_2_comment, llm_correct, llm_clarity, llm_justification, final_decision, gold_quality_issue_type, and adjudication_note. The evaluator_* fields are populated independently before adjudication; final_decision and adjudication_note are populated only after both evaluators have submitted their judgments.

5. LLM-as-a-judge procedure

In parallel with the two human evaluators, each validation instance was assessed by an LLM-as-a-judge procedure (Claude-3.5-Sonnet). The model received the prompt, context, and gold action and produced a binary correctness judgment, a clarity score on the same five-point scale, and a one-sentence justification. Consistent with the IFB protocol (Appendix A.8), the LLM judgment is treated strictly as a complementary, scalable quality-assurance signal and does not by itself determine the final_decision field; it enters the adjudication step on equal footing with the two human judgments via majority vote, as described in Section 6.

6. Adjudication

Following independent annotation, all instances where evaluator_1_correct and evaluator_2_correct disagree, or where both evaluators agree the instance is problematic, are reviewed during an adjudication stage. The adjudicator examines the prompt, context, gold action, and both evaluators' comments, and records final_decision (1 = retain the instance as released, 0 = flag for correction in a future release) together with an adjudication_note documenting the basis for the decision. Where the LLM judgment and the two human judgments

do not unanimously agree, final_decision follows a majority vote among the three independent assessments; in this validation round, the LLM judgment served as the tie-breaking input whenever the two human evaluators disagreed.

7. Reported statistics

Raw agreement and Cohen's κ are reported for the evaluator_1_correct / evaluator_2_correct pair. Because the validation sample is dominated by valid instances, κ is subject to the well-documented prevalence (class-imbalance) paradox: high raw agreement can coincide with only moderate κ because the expected-by-chance agreement term is itself close to the observed agreement when one category is rare. Krippendorff's α is therefore reported alongside Cohen's κ as a complementary chance-corrected reliability statistic, following the same practice adopted for IFB validation (Appendix A.8).

8. Relation to the automated quality audit

The issue taxonomy in Table S3 corresponds directly to the automated rule-matching artefacts described in Appendix B.3 (Stage 3) rather than to a generic error taxonomy: gold_too_generic flags fallback-template instances, multi_obligation_overreach and insufficient_context flag chunking boundary effects, and format_inconsistency flags the same boolean-determination schema tension documented for FinQA in Appendix A.7. This alignment allows disagreement patterns observed during human validation to be traced back to specific, correctable stages of the construction pipeline rather than treated as undifferentiated noise.