Risk 1: Unstable Medical Evaluation

LLJ judgements can change with prompt wording, judge model, repeated runs, grading scale, input format, or reference availability.

Define task-specific decision rules, derive prompts from the human codebook, and validate across prompts, runs, and judge models before scaling.

Protocol Steps: 1 4

Risk 2: Opaque Scores Hide Clinically Meaningful Errors

Aggregate LLJ scores can hide omissions, unsupported claims, attributions failures, unsafe reasoning, or other clinically meaningful errors.

Use criterion-level judgements, require evidence or source grounding when possible, and pair agreement scores with clinical error analysis and reporting.

Protocol Steps: 3 6

Risk 3: Weak Validity Under Clinical Uncertainty

High LLJ-clinician agreement can hide limited clinician agreement, ambiguity, or cases where clinician-aligned does not mean clinically correct.

Build an expert-annotated validation set, report dimension-level agreement and disagreement, and adjudicate clinically meaningful uncertainty.

Protocol Steps: 2

Risk 4: Downstream Safety and Data Contamination

Evaluator errors can distort model selection, deployment confidence, procurement decisions, or synthetic-data filtering.

Sample high-risk outputs, validate for the intended downstream use, classify clinical error types, and disclose the validation scope and artifacts.

Protocol Steps: 4 5 6