METHODS MEMO
Human Ground-Truth Paraphrase Annotation Protocol
Version 1.0 | January 2025

Author / Contact
Bruce Tisler
brucetisler@quantuminquiry.org

1. Purpose

This protocol establishes independent, non-computational paraphrase equivalence classes to serve as ground truth for invariance testing of question-measurement instruments.

The goal is to produce a human-validated reference set that allows an instrument to be evaluated on a narrow claim: whether a measured property function P(q) remains stable when meaning is preserved under paraphrase transformation.

2. Scope

Question domains (included):

- Factual questions (requesting a determinate fact)
- Explanatory questions (requesting an explanation of a phenomenon)
- Causal questions (requesting a cause or mechanism)

Language: English (initial release)

Target sample: 50–100 base questions

Annotators: ≥ 3 independent raters (minimum)

Out of scope (v1.0):

- Multilingual paraphrase classes
- Adversarial "near-paraphrase" attacks beyond boundary-case logging
- Questions requiring specialized domain expertise unless explicitly recruited

3. Definitions

Base question: A single, canonical question used as the anchor for a paraphrase class.
Candidate paraphrase: A proposed alternative phrasing intended to preserve the base question's meaning.

Non-paraphrase: A candidate that alters meaning, introduces or removes constraints, changes referents, changes causal structure, or changes what would count as a correct answer.

Equivalence class: A set of questions treated as meaning-equivalent under this protocol, defined by human judgment thresholds specified below.

Boundary case: A candidate that produces systematic disagreement among annotators and is retained as a documented example of class edges.

4. Roles and responsibilities

Protocol owner (research lead):

- Curates base questions and candidate sets
- Conducts annotator screening and onboarding
- Implements quality control and analysis
- Publishes outputs and changelog

Annotators:

- Provide independent judgments
- Follow decision rules as specified
- Flag ambiguous candidates and rationale

5. Phase 1: Stimulus construction

5.1 Base question selection

Select 50–100 base questions distributed across the included domains. Base questions should:

- Be grammatically well-formed
- Avoid intentionally ambiguous referents
- Avoid hidden premises unless those premises are explicit in the question text
- Be answerable in principle (even if difficult)

5.2 Candidate generation

For each base question, generate 8–15 candidate variants. Candidate sources may include:

- Manual paraphrasing by the protocol owner
- LLM-generated paraphrases filtered by the protocol owner
- Human contributors (optional)

Each candidate must be recorded with:

- Base question ID
- Candidate ID
- Candidate text
- Source (manual / LLM / contributor)
- Notes (optional)

5.3 Negative controls (required)

For each base question, include 2–4 “obvious non-paraphrases” that:

- Change the subject or referent
- Change the constraint structure (e.g., add “in 2020,” remove “why,” change “how” to “whether”)
- Change the requested information type (fact vs explanation)

These are used for attention checks and specificity testing.

6. Phase 2: Annotation protocol

6.1 Task structure
Annotators are presented with:

- The base question
- A single candidate
- A forced-choice judgment + optional rationale

Judgment prompt:
“Does this candidate preserve the meaning of the base question such that the same answer would satisfy both?”

6.2 Response format
Annotators choose one:

- Paraphrase (meaning preserved)
- Not a paraphrase (meaning changed)
- Unclear / cannot decide (insufficient clarity)

Annotators may optionally provide a short rationale (1–3 sentences), especially for “Not a paraphrase” or “Unclear.”

6.3 Decision rules (required)
Annotators must mark Not a paraphrase if the candidate:

- Changes who/what is being referred to (referent drift)
- Adds or removes constraints (time, location, conditions, quantifiers)
- Changes the type of information requested (fact vs explanation vs cause)
- Changes causal direction or mechanism implied
- Narrows or broadens scope in a way that changes what counts as a correct answer

Annotators must mark Unclear if:

- Either item is ambiguous enough that equivalence cannot be determined reliably
- The candidate introduces polysemy that changes interpretability

7. Quality control

7.1 Annotator selection
Annotators must:

- Be fluent English readers
- Pass a short screening task (10 – 15 items) including clear paraphrases and non-paraphrases
- Demonstrate consistent application of rules (qualitative review)

7.2 Attention checks

Include 5–10 attention-check items across the full annotation set:

- Obvious non-paraphrases (negative controls)
- Obvious paraphrases (positive controls)

Annotators failing attention checks above a defined threshold should be excluded or re-trained.

7.3 Randomization

Randomize:

- Candidate order within each base question
- Base question order across the session
- Pairing order if using any paired comparisons

8. Analysis plan

8.1 Inter-rater reliability

Compute inter-rater reliability on categorical judgments using:

- Fleiss' κ (recommended for ≥ 3 raters)
- Report κ and confidence intervals if feasible

Reliability threshold (v1.0):

- Target: $\kappa \geq 0.70$ for deployment readiness
- If $\kappa < 0.70$, report failure conditions and revise protocol or training

8.2 Class formation rules

For each base question, candidates are assigned to the paraphrase equivalence class using one of the following rules (choose one and state it explicitly):

Majority rule (default):

- Candidate is included if > 50% of valid raters mark "Paraphrase"
- "Unclear" counts as neither paraphrase nor non-paraphrase

Strict consensus (recommended for conservative sets):

- Candidate is included only if ≥ 80% of valid raters mark "Paraphrase"

The chosen rule must be applied consistently across the dataset.

8.3 Boundary case documentation

Candidates with systematic disagreement are retained as boundary cases with:

- Vote breakdown
- Example rationales
- Notes on why disagreement likely occurred (scope, referent, constraint drift)

These boundary cases are useful for later adversarial testing and instrument sensitivity evaluation.

9. Outputs

This protocol produces:

1. Paraphrase equivalence classes (base question + accepted paraphrases)
2. Rejected candidates with vote breakdown and rationale summaries
3. Agreement statistics (κ and summary tables)
4. Boundary case catalog (disagreement exemplars)

10. Release statement

This protocol is released for community use and feedback prior to formal validation. It is intended to support independent replication and critique, and to serve as a ground-truth foundation for paraphrase invariance testing of question-measurement instruments.