Table of Contents

# Assessment Schemas Export

Generated: 2025-11-01T03:39:53.857Z

Schemas: 5

## Table of Contents

* [AlternativeExplanationGeneration](#schema-alternativeexplanationgeneration)
* [BiasDirectionOpen](#schema-biasdirectionopen)
* [BiasDirectionSequential](#schema-biasdirectionsequential)
* [FermiPopulationCity](#schema-fermipopulationcity)
* [SelectionEffectIdentification](#schema-selectioneffectidentification)

## Schema: AlternativeExplanationGeneration

* Description: Assesses the ability to generate distinct, plausible alternative explanations for an association.
* Engine: driverId=aeq.aeg.v1, version=1.1.0
* Guidance Version: 1.0.0
* Ability: causal.aeg
* Policy Defaults:
  + MaxConsecutiveFailedAttempts: 2
* Driver Config:
  + Answer Types:
    - Good — A plausible causal explanation that does not run through A, and is distinct, specific, and clear.
    - NotDistinct — A plausible alternative explanation, but relies on the same mechanism with a previous answer in this unit.
    - NotSpecific — The response is understandable, but there is no causal mechanism specific enough to correspond to a particular theme in the ThemeRegistry or to be identified as a novel mechanism.
    - NotClear — The meaning of the response cannot be interpreted without elaboration (ambiguous, incoherent, or too compressed).
    - RunsThroughA — The proposed causal mechanism directly or indirectly involves the first item in the association (A) causing the second (B).
    - NotPlausible — The proposed causal explanation is implausible.
    - NotRelevant — The response is irrelevant to the task.
    - MultipleExplanation — The response contains multiple explanations.
  + Answer Type Guidance:
    - DISTINCTNESS AND NOVELTY: Map the explanation to exactly one ThemeID from the item’s ThemeRegistry. If none fits, return a ThemeTag with 'NOVEL:' prefix and a concise label. Compare ThemeTag against previously accepted themes to decide Good vs NotDistinct. ThemeTag is REQUIRED for Good and NotDistinct.
    - CRITICAL: You must compare the ThemeTag against previous themes in the session history AND other themes in the current response to determine if the AnswerType is 'Good' or 'NotDistinct'. You MUST return the ThemeTag if the AnswerType is 'Good' or 'NotDistinct'
    - SINGLE EXPLANATION RULE: Score exactly one explanation per user turn. If the user provides multiple explanations, classify as MultipleExplanation and use a MultipleExplanation probe; do not score the turn.
    - PLAUSIBILITY: Evaluate plausibility based on commonly understood dynamics. Ensure the mechanism is concrete and respects plausible temporal ordering. Valid mechanisms include: Common Cause, Reverse Causation, Detection Bias, and Opportunity Cost/Displacement (e.g., A displaces a protective factor C).
    - INDEPENDENCE RULE (Anti-Smuggling): The explanation's mechanism MUST NOT rely on the direct causal pathway A->B, or cause B ONLY through A. Scrutinize responses to ensure they are not merely mediators (A->Mediator->B). If the explanation depends on A occurring first to cause B (except for displacement), it is invalid (classify as RunsThroughA). Report your confidence level.
    - TIE-BREAKING: If an answer fits multiple AnswerTypes, choose the one that appears earliest in the DominanceOrder list.
  + Probing Guidance:
    - VARIATION RULE: Review the session history. You MUST select a Probe ID from the ProbeLibrary that has not yet been used in the current unit, if available. If all probes in that category have been used, you may reuse them.
    - STANDARDIZED SELECTION: Select the ID of the most conversationally appropriate probe from that AnswerType's category in the ProbeLibrary. Do not generate new probe text.
  + Dominance Order: MultipleExplanation → RunsThroughA → NotRelevant → NotDistinct → NotClear → NotSpecific → NotPlausible
  + Confidence Policy:
    - GoodAcceptanceMinConfidence: 0.6
    - NovelAcceptanceMinConfidence: 0.75
  + Max Clarification Attempts: 1
  + Clarification Policy:
    - MaxTotal: 2
    - MaxConsecutiveNotSpecific: 1
    - MaxConsecutiveNotClear: 1
* Scoring Spec:
  + Type: polytomous\_map
  + Default:
    - TargetDistinctExplanations: 2
    - ScoringMapID: Map\_2Expl
  + Maps:
    - Map\_2Expl:
      * score 0 — 0 Good
      * score 1 — 1 Good
      * score 2 — 2 Good
    - Map\_3Expl:
      * score 0 — 0
      * score 1 — 1
      * score 2 — 2
      * score 3 — 3
  + Theta:
    - step: 0.25
    - varDecay: 0.9
    - minVar: 0.5
* AJ Contract:
  + Type: object
  + Properties:
    - AnswerType: string [Good, NotDistinct, NotSpecific, NotClear, NotRelevant, NotPlausible, RunsThroughA, MultipleExplanation]
    - ThemeTag: string|null
    - Confidence: number
    - RecommendedProbeID: string|null
  + Required: AnswerType, Confidence
  + Additional Properties: not allowed

## Schema: BiasDirectionOpen

* Description: Assesses recognition that observational null results are ambiguous due to selection biases that can mask benefit or harm; constrained categorical AJ output for dynamic probing and path-dependent scoring.
* Engine: driverId=bias.direction.open.v1, version=1.0.0
* Guidance Version: 1.1.0
* Ability: causal.interpretation.null\_effects
* Policy Defaults:
  + MaxConsecutiveFailedAttempts: 2
* Driver Config:
  + Answer Types:
    - Both\_Explained — Explains BOTH a Masked Benefit and a Masked Harm scenario with specific mechanism and masking logic.
    - MaskedBenefit\_Only\_Explained — Explains Masked Benefit (Negative Bias masking helpful effect) with mechanism and logic; not Masked Harm.
    - MaskedHarm\_Only\_Explained — Explains Masked Harm (Positive Bias masking harmful effect) with mechanism and logic; not Masked Benefit.
    - NotSpecific — Clear/relevant but insufficiently specified: e.g., says "selection bias" or "groups differ" without a concrete mechanism or masking logic.
    - NotDistinct — A plausible explanation, but relies on the same underlying mechanism/logic as a previous explanation in this unit.
    - NotPlausible — Mechanism is unrealistic or does not fit the context; not a credible masking story.
    - NotClear — Ambiguous/incoherent.
    - NotRelevant — Irrelevant to the task.
  + Answer Type Guidance:
    - CORE TASK: Classify into exactly one AnswerType.
    - STRICT REQUIREMENTS: (A) Specific mechanism; (B) Direction of bias; (C) Masking logic (bias + true effect → null).
    - MaskedBenefit: intervention helps but participants were inherently worse (Negative Bias) → up to average.
    - MaskedHarm: intervention harms (e.g., time cost) but participants were inherently better (Positive Bias) → down to average.
    - CLASSIFICATION ORDER: NotRelevant/NotClear → NotDistinct/NotPlausible → Both\_Explained → Masked[X]\_Only → NotSpecific (apply DominanceOrder on ties).
    - SPECIFICITY: Responses that merely say 'selection bias' or 'the groups were different' without a concrete factor/process are 'NotSpecific'.
    - DISTINCTNESS: If the explanation repeats the same underlying mechanism as earlier, classify as 'NotDistinct'.
    - PROBES: Select a probe ID from the item-level ProbeLibrary category matching the AnswerType. Prefer unused IDs.
  + Dominance Order: NotRelevant → NotClear → NotDistinct → NotPlausible → Both\_Explained → MaskedBenefit\_Only\_Explained → MaskedHarm\_Only\_Explained → NotSpecific
  + Confidence Policy:
    - MinConfidence: 0.6
  + Clarification Policy:
    - MaxTotal: 2
    - MaxConsecutiveNotClear: 1
    - MaxConsecutiveNotSpecific: 1
  + Answer Type Map:
    - Neither\_Explained\_Sufficiently → NotSpecific
* Scoring Spec:
  + Type: path\_dependent\_scoring
* AJ Contract:
  + Type: object
  + Properties:
    - AnswerType: string [Both\_Explained, MaskedBenefit\_Only\_Explained, MaskedHarm\_Only\_Explained, NotSpecific, NotDistinct, NotPlausible, NotClear, NotRelevant]
    - Confidence: number
    - RecommendedProbeID: string|null
  + Required: AnswerType, Confidence
  + Additional Properties: not allowed

## Schema: BiasDirectionSequential

* Description: Assesses ability to explain a zero-observed difference via selection effects in both directions (benefit vs harm).
* Engine: driverId=bias.direction.sequential.v1, version=1.0.0
* Guidance Version: v1
* Ability: causal.selection.direction
* Policy Defaults:
  + MaxConsecutiveFailedAttempts: 2
  + MaxTotalFailedAttempts: 4
* Driver Config:
  + Answer Types:
    - BiasPositive — Assumes the true effect is beneficial (A helped). Provide a concrete selection or sampling mechanism that would mask or offset that benefit in the observed association.
    - BiasNegative — Assumes the true effect is harmful (A harmed). Provide a concrete selection or sampling mechanism that would mask or offset that harm in the observed association.
    - NotSpecific — Names bias or says groups differ without describing the specific factor or process (e.g., which people, how they were selected).
    - NotClear — Ambiguous or incoherent; requires rephrasing to understand the mechanism.
    - NotPlausible — Mechanism is unrealistic or does not fit the context.
    - NotRelevant — Does not address selection/sampling or the posed question.
    - MultipleExplanation — Contains more than one distinct reason in a single turn.
  + Answer Type Guidance:
    - CORE TASK: Identify a selection or sampling mechanism (non-random sorting or inclusion) that explains the observed association.
    - DIRECTION COMMITMENT: Choose exactly one direction per turn — 'BiasPositive' if you assume the intervention truly helped; 'BiasNegative' if you assume it truly harmed. Do not return both at once.
    - SELECTION MECHANISM: A direction answer must describe how the groups differ or how the sample is biased, and how that produces the observed association despite the assumed true effect.
    - SPECIFICITY: Responses that merely say 'selection bias' or 'the groups were different' without a concrete factor or process are 'NotSpecific'.
    - THEME TAGGING (required when ThemeRegistry is present): If the item provides a ThemeRegistry, you MUST set 'ThemeTag' to the best matching ThemeID. If no ThemeRegistry is present, you may omit it. Novel mechanisms may use 'NOVEL:label'.
    - DISTINCTNESS: At most one accepted answer per direction in a unit (one BiasPositive and one BiasNegative). Providing another answer in the same direction should not be accepted as a new distinct mechanism.
    - MULTI-EXPLANATION HANDLING: If a user provides multiple reasons in one turn, classify as 'MultipleExplanation' and do not score content.
    - TIE-BREAKING: If multiple categories fit, choose the earliest in DominanceOrder.
  + Probing Guidance:
    - VARIATION RULE: Prefer an unused Probe ID within the relevant category.
    - DIRECTIONAL PROBES: After accepting 'BiasPositive', request the opposite direction using a probe from 'OppositeFromPositive'. After accepting 'BiasNegative', use 'OppositeFromNegative'.
    - CLARIFICATION PROBES: For 'NotSpecific' and 'NotClear', use concise prompts that ask for a concrete factor/process or a clearer restatement; avoid hinting.
    - NO GENERATION: Do not generate new probes; select from the item-level ProbeLibrary.
  + Dominance Order: MultipleExplanation → NotRelevant → NotClear → NotSpecific → NotPlausible → BiasPositive → BiasNegative
  + Confidence Policy:
    - MinAcceptConfidence: 0.65
  + Max Clarification Attempts: 1
  + Clarification Policy:
    - MaxTotal: 2
    - MaxConsecutiveNotSpecific: 1
    - MaxConsecutiveNotClear: 1
  + Answer Type Map:
    - UpwardBias → BiasPositive
    - DownwardBias → BiasNegative
  + Probe Category For:
    - BiasPositive → OppositeFromPositive
    - BiasNegative → OppositeFromNegative
    - NotSpecific → NotSpecific
    - NotClear → NotClear
    - NotPlausible → NotPlausible
    - NotRelevant → NotRelevant
    - MultipleExplanation → MultipleExplanation
* Scoring Spec:
  + Target Distinct Explanations: 2
  + Final perDistinct:
    - 0: 0
    - 1: 0.5
    - 2: 1
* AJ Contract:
  + Type: object
  + Properties:
    - AnswerType: string [BiasPositive, BiasNegative, NotSpecific, NotClear, NotRelevant, NotPlausible, MultipleExplanation]
    - ThemeTag: string|null
    - Confidence: number
    - RecommendedProbeID: string|null
  + Required: AnswerType, Confidence
  + Additional Properties: not allowed

## Schema: FermiPopulationCity

* Description: Estimate a city's population (orders of magnitude).
* Engine: driverId=generic.numeric.v1, version=1.1.0
* Guidance Version: 1.0.0
* Ability: numeric.fermi
* Policy Defaults:
  + MaxConsecutiveFailedAttempts: 2
* Driver Config:
* Scoring Spec:
  + Theta:
    - step: 0.25
    - varDecay: 0.9
    - minVar: 0.5
* AJ Contract:
  + Type: object

## Schema: SelectionEffectIdentification

* Description: Assesses the ability to identify and explain selection effects (e.g., confounding by indication, self-selection, survivorship bias) that could explain a surprising or counterintuitive association in an observational context, while adhering to stated premises/constraints.
* Engine: driverId=aeq.aeg.v1
* Guidance Version: v1
* Ability: causal.selection\_effects
* Policy Defaults:
  + MaxConsecutiveFailedAttempts: 2
  + MaxTotalFailedAttempts: 4
* Driver Config:
  + Answer Types:
    - Good — A plausible explanation that identifies a specific selection mechanism consistent with the premise.
    - NotDistinct — A plausible explanation, but relies on the same mechanism as a previous answer.
    - NotSpecific — Acknowledges bias but does not describe the specific mechanism or confounding variable.
    - NotClear — The meaning of the response cannot be interpreted without elaboration.
    - RejectsAssumption — Violates or contradicts a stated assumption or constraint in the prompt.
    - NotPlausible — The proposed mechanism is implausible in the context.
    - NotRelevant — The response is irrelevant to the task.
    - MultipleExplanation — The response contains multiple explanations.
  + Answer Type Guidance:
    - CORE TASK: Identify a SELECTION EFFECT (non-random sorting or sampling) that explains the observed association.
    - SELECTION MECHANISM: A 'Good' answer must describe how groups differ systematically or how the sample is biased, and how this explains the outcome (name the factor Z or process).
    - CONSTRAINT ADHERENCE: Any explanation that contradicts the stated assumption(s) MUST be classified as 'RejectsAssumption'.
    - SPECIFICITY: Merely naming 'confounding' or 'selection bias' without a specific mechanism is 'NotSpecific'.
    - DISTINCTNESS: Use the item's ThemeRegistry for ThemeTag; if none fits, use 'NOVEL:…'. Compare against previously accepted themes to decide Good vs NotDistinct. ThemeTag is REQUIRED for Good and NotDistinct.
    - SINGLE EXPLANATION RULE: If multiple explanations appear in one turn, classify as MultipleExplanation.
    - TIE-BREAKING: If multiple AnswerTypes fit, choose the one appearing earliest in DominanceOrder.
  + Probing Guidance:
    - VARIATION RULE: Select a Probe ID from the ProbeLibrary that has not yet been used in the current unit, if available.
    - CONSTRAINT REMINDER: For 'RejectsAssumption', select a short probe from the 'RejectsAssumption' category that neutrally reiterates the relevant constraint without hinting.
    - STANDARDIZED SELECTION: Select a probe from the appropriate category in the ProbeLibrary. Do not generate new probe text.
  + Dominance Order: MultipleExplanation → RejectsAssumption → NotRelevant → NotDistinct → NotClear → NotSpecific → NotPlausible
  + Confidence Policy:
    - GoodAcceptanceMinConfidence: 0.6
    - NovelAcceptanceMinConfidence: 0.75
  + Max Clarification Attempts: 1
  + Clarification Policy:
    - MaxTotal: 2
    - MaxConsecutiveNotSpecific: 1
    - MaxConsecutiveNotClear: 1
  + Answer Type Map:
    - RejectsAssumption → RunsThroughA
  + RunsThrough Category Label: RejectsAssumption
* Scoring Spec:
  + Final perDistinct:
    - 0: -1
    - 1: -0.5
    - 2: 0
    - 3: 1
* AJ Contract:
  + Type: object
  + Properties:
    - AnswerType: string [Good, NotDistinct, NotSpecific, NotClear, NotRelevant, NotPlausible, RejectsAssumption, MultipleExplanation]
    - ThemeTag: string|null
    - Confidence: number
    - RecommendedProbeID: string|null
  + Required: AnswerType, Confidence
  + Additional Properties: not allowed