**Data Dictionary**

| **Column Name** | **Description** |
| --- | --- |
| **Scenario Name** | A short, descriptive title for the scenario (e.g., “6. WatchSwap at the Airbnb”). |
| **Scenario Text** | A brief narrative or context introducing the situation, characters, or problem (1–3 paragraphs). |
| **Question Label** | A short code identifying the question within that scenario (e.g., Q6A, Q6B). |
| **Question Text** | The actual question posed about the scenario (e.g., “Where does Marcus believe Dana thinks Priya believes the watch is?”). |
| **A**, **B**, **C**, **D**, **E** | The multiple-choice answer options for that question. Each option is a short statement the model or human is expected to choose from. |
| **Correct** | The letter (A–E) that is deemed correct for this question. |
| **Explanation** | A brief rationale for why that letter is correct (1–2 sentences). |
| **NBD** | **Nested Belief Depth** – an integer or label indicating how many “layers” of belief or perspective-taking are required. |
| **DOM** | **Domain** – broad type of context or subject matter (e.g., SOC, OBJ, ETH, TMP). |
| **PHEN** | **Phenomenon** – a more specific label describing the cognitive or ethical phenomenon tested (e.g., DEC, IRN, PRIV, META). |
| **ASD** | **Alignment-Specific Dimension** – indicates which alignment challenge or scenario type is relevant (e.g., PI for Pluralistic Ignorance). |
| **QT** | **Question Type** – categorizes the item as factual comprehension (FC), mental-belief reasoning (MB), ethical (ETH), or other. |
| **DIFF** | **Difficulty** – a rough label (easy, hard, extreme) to indicate the complexity of the question or the challenge it poses. |

**Glossary of Abbreviations**

Below is a quick reference for the short codes used in **NBD**, **DOM**, **PHEN**, **ASD**, **QT**, and **DIFF** columns.

**1. Nested Belief Depth (NBD)**

* **0** – No belief reasoning required (straight factual recall).
* **1** – First-order belief (“A believes X”).
* **2** – Second-order belief (“A believes B believes X”).
* **3** – Third-order belief (“A believes B believes C believes X”).
* **4, 5, etc.** – Higher-order nesting.
* **6plus** – Items with extremely deep or partially “impossible” nesting.

**2. Domain (DOM)**

* **SOC** – Social scenario or dialogue-based.
* **OBJ** – Physical-object or spatial tracking.
* **ETH** – Ethics-specific context (e.g., privacy, mental health).
* **TMP** – Time or distribution shift domain (e.g., multiple city test sites).

*(Some items use* ***multiple*** *domain tags or leave this blank if not strongly domain-specific.)*

**3. Phenomenon (PHEN)**

* **DEC** – Deception (strategic misinformation, lying, or bluffing).
* **IRN** – Irony or sarcasm.
* **IGN** – Explicit ignorance or “lack of belief” scenario.
* **META** – Meta-level reasoning (e.g., beliefs about puzzle solver’s beliefs).
* **PRIV** – Privacy concerns (e.g., private diaries, personal data).
* **SUI** – Self-harm or suicidal ideation context.
* **MED** – Medical or diagnostic scenario.

**4. Alignment-Specific Dimension (ASD)**

*(These tags reflect alignment or safety concerns present in the scenario.)*

* **PI** – Pluralistic Ignorance.
* **VI** – Value Interpretation misalignment.
* **CC** – Capability Concealment or evaluation gaming.
* **DS** – Distributional Shift across conditions.
* **INT** – Interpretability Gaps (explanations vs. actual optimization).
* **MH** – Mental Health or self-harm–related scenario.
* **PRV** – Privacy or data minimization issues.
* *(Other custom tags as needed.)*

**5. Question Type (QT)**

* **FC** – Factual Comprehension (who said what, where is object, etc.).
* **MB** – Mental Belief or nested perspective-taking.
* **META** – Reasoning about the scenario or reasoning process itself.
* **ETH** – Ethics or moral reasoning (e.g., appropriate response to self-harm).
* *(Sometimes you’ll see multiple if an item covers multiple aspects.)*

**6. Difficulty (DIFF)**

* **easy** – Typically straightforward recall or shallow inference.
* **hard** – Requires multi-step reasoning or multiple perspectives.
* **extreme** – Very high-order reasoning, ambiguous or partial information, or extremely tricky subtlety.

**Note:**

* Some columns (like **ASD**) may be blank or contain multiple tags for one item if it spans multiple alignment dimensions.
* The scenario text and question text typically highlight the relevant domain or phenomenon; the classification columns are more for searching, filtering, or structured analysis.
* If you see any new abbreviations (like PRV for privacy or SUI for suicidality) that aren’t in the original set, those typically appear for newly added ethical items in Scenario 14.