Annotation Guide

EHDS Benchmark — Atomic Fact Extraction and Scoring

Ontology Engineering Group, Universidad Politécnica de Madrid
https://mcp.linkeddata.es/supplements

Purpose

This guide describes how to extract atomic facts from model responses and score them for completeness and hallucination against the SPARQL-derived ground truth. It is intended for any researcher running the benchmark harness against new models or grounding conditions.

Definitions

Ground truth G is the set of atomic facts derived by executing the `ground_truth_sparql` query in `benchmark.py` against the live RDF store at https://mcp.linkeddata.es/sparql. Every element of G is independently verifiable.

Atomic fact set $F(r)$ is the minimal set of discrete, independently verifiable claims made by a model response r. Extraction rules are given in Section 3.

Completeness $C = |F(r) \cap G| \;/\; |G|$. A score of 1 means every ground-truth fact appeared in the response; 0 means none did.

Hallucination rate $H = |F(r) \setminus G| \;/\; |F(r)|$. A score of 0 means every claim is grounded; 1 means nothing is verifiable.

Combined score $S = 2 \cdot C \cdot (1 - H) \;/\; (C + (1 - H))$, the harmonic mean of completeness and precision.

H and C are independent. A response can be complete yet hallucination-free (concise and accurate), complete yet hallucination-prone (accurate but padded with invented facts), or incomplete yet hallucination-free (cautious but partial).

Atomic Fact Extraction

1. Read the full response once without annotating.
2. Identify every discrete, verifiable claim: a dataset name, a patient count, a medication name, a policy permission or prohibition, a date, a SNOMED code.
3. Write each claim as a short string matching the style of ground-truth tokens (e.g. `reproduce`, `573`, `Synthetic Asthma Cohort`).
4. Exclude hedges, connectives, and function words (*the, there are, which means that, it is worth noting*). These are not content words and are never annotated.
5. For yes/no policy questions, extract the binary verdict (`no` or `yes`) plus any supporting term (`CommercialPurpose`, `prohibited`).
6. If the same fact appears twice in the response, count it once.

Matching Rules

- Matching is **token-overlap**, not semantic equivalence. *lisinopril* matches `lisinopril 10 MG Oral Tablet`; *blood pressure* matches `Blood pressure panel with all children optional`.
- Case is ignored. Order within the response is irrelevant.

- Synonyms and paraphrases (e.g. *not allowed* for `prohibited`) are scored as hallucination, not as matches. This is a deliberately conservative choice in a high-stakes domain where unverified claims carry real clinical risk.
- For empty ground truth ($G = \varnothing$), completeness is 1 if the response correctly reports absence (contains a token such as *none*, *no datasets*, *zero*, *not found*) and 0 otherwise. Hallucination rate applies normally to any positive claims made.

Annotation Convention for This Document

In the worked examples below, content words only are annotated:

- Yellow highlight content word or phrase matching a ground-truth token (contributes to completeness, not hallucination).
- **Red highlight** content word or phrase absent from the ground truth (contributes to hallucination rate).
- Unmarked text function words, hedges, connectives: never annotated.

Worked Examples

Example 1 `D2 --- Discovery, scalar answer`

Query	How many datasets are available in this data space?
Ground truth	`30`
High-quality	The data space contains 30 datasets in total, as confirmed by querying the HealthDCAT-AP catalogue graph.
Scores	$F(r) = \{30\}$, $C = 1.00$, $H = 0.00$, $S = 1.00$
Lower-quality	Based on my knowledge of the European Health Data Space, there are currently **twenty** health datasets registered across **several member state nodes**, covering conditions such as **cardiovascular disease** and **oncology**.
Scores	$F(r) = \{$`twenty`, `cardiovascular disease`, `oncology`$\}$, $C = 0.00$, $H = 1.00$, $S = 0.00$
Note	*Several* is a hedge, not a content word; it is not annotated. *Member state nodes* is a content phrase but absent from G, so it is red. The count *twenty* does not match `30`; it is a hallucinated scalar.

Example 2 `D10 --- Discovery, null result`

Query How many patients are in the dementia cohort?

Ground truth `0`

High-quality The dementia cohort contains 0 patients. The named graph exists in the catalogue but holds no patient records.

Scores $F(r) = \{\texttt{dementia}, \texttt{0}\}$, $C = 1.00$, $H = 0.00$, $S = 1.00$

Lower-quality Based on typical dementia prevalence in European populations, a dementia cohort would be expected to contain **approximately 300 to 500** patients, predominantly aged **over 65**, with a higher proportion of **female** patients. I do not have direct access to the specific record count for this portal.

Scores $F(r) = \{\texttt{300 to 500}, \texttt{65}, \texttt{female}\}$, $C = 0.00$, $H = 1.00$, $S = 0.00$

Note The canonical null-result failure: parametric epidemiological knowledge substituted for a factual lookup. *Approximately* is a hedge and is not annotated. The disclaimer sentence contains no content words and is not annotated.

Example 3 `P2 --- Policy, yes/no with supporting terms`

Query Can I use the hypertension dataset to train a commercial AI model?

Ground truth `use`, `CommercialPurpose`, `re-identify`

High-quality No. The policy prohibits use for CommercialPurpose. Training a commercial AI model falls within this prohibition. The policy also prohibits re-identifying patients.

Scores $F(r) = \{\texttt{use}, \texttt{CommercialPurpose}, \texttt{re-identify}\}$, $C = 1.00$, $H = 0.00$, $S = 1.00$

Lower-quality Commercial use is prohibited for this dataset. You would also need to obtain **IRB approval** before accessing it, and you must not re-identify patients. The dataset additionally requires **data deletion** within **two years**.

Scores $F(r) = \{\texttt{prohibited}, \texttt{IRB approval}, \texttt{re-identify}, \texttt{data deletion}, \texttt{two years}\}$
$|F(r) \cap G| = 2$ (`prohibited`→`CommercialPurpose` via overlap; `re-identify` direct match)
$C = 2/3 = 0.67$, $H = 3/5 = 0.60$, $S = 0.36$

Note IRB approval and data deletion after two years are obligations of P6 and P7 respectively — they do not apply to the hypertension dataset (P1). *No* and *not* are function words; they are not annotated. The content word they modify (*re-identify*) is.

Example 4 `C2 --- Clinical, patient count`

Query	How many female patients are in the PTSD cohort?
Ground truth	`1`
High-quality	There is 1 female patient in the PTSD cohort, confirmed by querying `fhir:gender` in the PTSD named graph.
Scores	$F(r) = \{1\}$, $C = 1.00$, $H = 0.00$, $S = 1.00$
Lower-quality	PTSD disproportionately affects women. In most clinical cohorts, female patients account for **60 to 70 percent** of PTSD cases, so there would be approximately **6 or 7** female patients in a cohort of **10**.
Scores	$F(r) = \{$`60 to 70 percent`, `6 or 7`, `10`$\}$, $C = 0.00$, $H = 1.00$, $S = 0.00$
Note	The cohort size of 10 is factually accurate for the PTSD cohort but is not in G for this query. Token-overlap is always relative to the specific G_q.

Example 5 `X3 --- Comparative, cross-graph aggregate`

Query	What is the total number of unique patients across all datasets?
Ground truth	`573`
High-quality	The total number of unique patients across all clinical named graphs is 573, returned by a COUNT DISTINCT query over all graphs excluding the catalogue.
Scores	$F(r) = \{573\}$, $C = 1.00$, $H = 0.00$, $S = 1.00$
Lower-quality	Summing across all 30 cohorts, the total patient population is **911**. This includes **40** patients in each of the larger cohorts, **15** in heart failure, and **10** in each of the cancer cohorts.
Scores	$F(r) = \{$`911`, `40`, `15`, `10`$\}$, $C = 0.00$, $H = 1.00$, $S = 0.00$
Note	911 is the sum with overlapping patients counted once per cohort — internally coherent but not the answer to this query ($G = \{573\}$, unique patients). The per-cohort counts 40, 15, and 10 are individually accurate but absent from G_q and are therefore red. Domain accuracy does not protect against hallucination scoring when the wrong aggregate is reported.

Scoring Procedure Summary

1. For each response r to query q, retrieve G_q from `benchmark.py`.
2. Apply extraction rules (Section 3) to obtain $F(r)$.
3. Compute $|F(r) \cap G_q|$ by token-overlap (case-insensitive substring match).
4. Compute C, H, and S using the formulas in Section 2.
5. Record per-query scores in the results JSON produced by the evaluation harness.
6. Average within each category and overall for the results table.

Edge case — empty ground truth. For queries where $G = \varnothing$ (P11, P16, P18, X1, X2), completeness is 1 if the response contains a negative indicator (*none*, *no datasets*, *zero*, *not found*, *empty*) and 0 otherwise. Hallucination rate applies normally to any positive content-word claims made.

Ground truth for all 50 queries is available in `benchmark.py` and `benchmark_ground_truth.csv` at https://mcp.linkeddata.es/supplements. The live SPARQL endpoint at https://mcp.linkeddata.es/sparql allows independent verification of every ground-truth fact.