Implementation Guidelines

Doc Link:

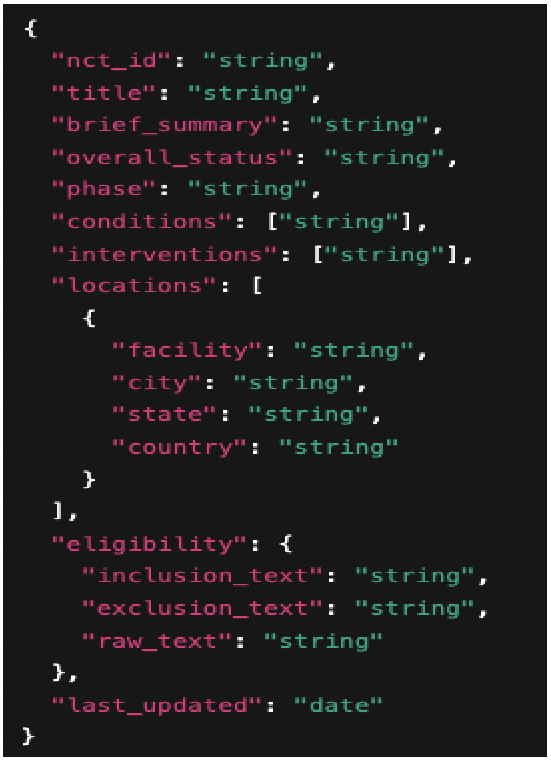
https://docs.google.com/document/d/16RFJq9oGotiubJK6GlgCYNXI5W58Ii7bGJkT_7KPs-k/edit?usp=sharing

Github: https://github.com/Matt23-star/TrialMatcher-RAG-CSCI-544

Dataset links:

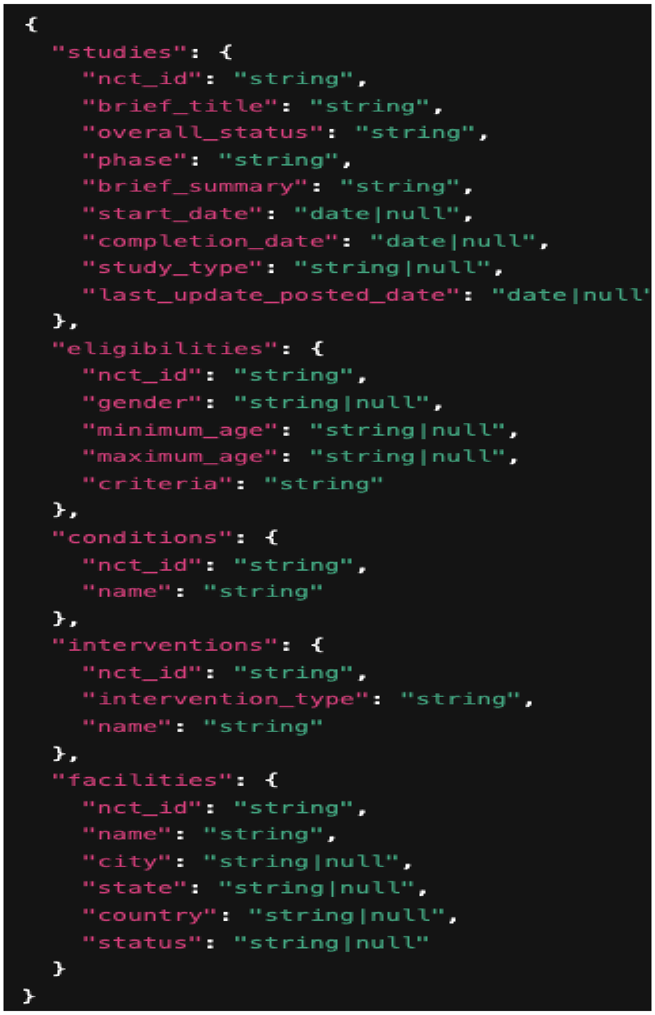
1. **ClinicalTrials.gov API v2**:
   https://clinicaltrials.gov/data-api/api

   Sample Schema:



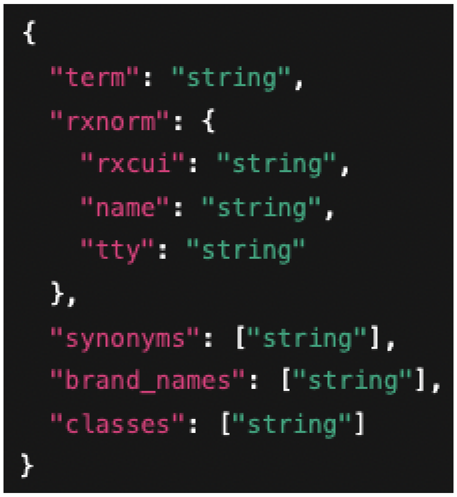
```
{
  "nct_id": "string",
  "title": "string",
  "brief_summary": "string",
  "overall_status": "string",
  "phase": "string",
  "conditions": ["string"],
  "interventions": ["string"],
  "locations": [
    {
      "facility": "string",
      "city": "string",
      "state": "string",
      "country": "string"
    }
  ],
  "eligibility": {
    "inclusion_text": "string",
    "exclusion_text": "string",
    "raw_text": "string"
  },
  "last_updated": "date"
}
```

2. **AACT (Postgres snapshot of ClinicalTrials.gov)**:
   https://aact.ctti-clinicaltrials.org/postgres



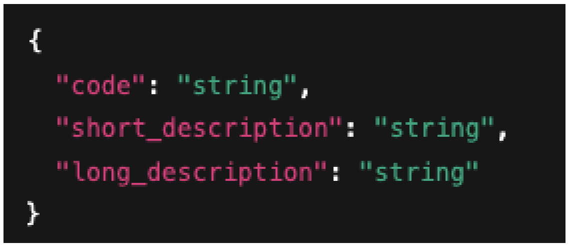
```
{
  "studies": {
    "nct_id": "string",
    "brief_title": "string",
    "overall_status": "string",
    "phase": "string",
    "brief_summary": "string",
    "start_date": "date|null",
    "completion_date": "date|null",
    "study_type": "string|null",
    "last_update_posted_date": "date|null"
  },
  "eligibilities": {
    "nct_id": "string",
    "gender": "string|null",
    "minimum_age": "string|null",
    "maximum_age": "string|null",
    "criteria": "string"
  },
  "conditions": {
    "nct_id": "string",
    "name": "string"
  },
  "interventions": {
    "nct_id": "string",
    "intervention_type": "string",
    "name": "string"
  },
  "facilities": {
    "nct_id": "string",
    "name": "string",
    "city": "string|null",
    "state": "string|null",
    "country": "string|null",
    "status": "string|null"
  }
}
```

3. **RxNorm / RxNav APIs**: https://lhncbc.nlm.nih.gov/RxNav/APIs/RxNormAPIs.html



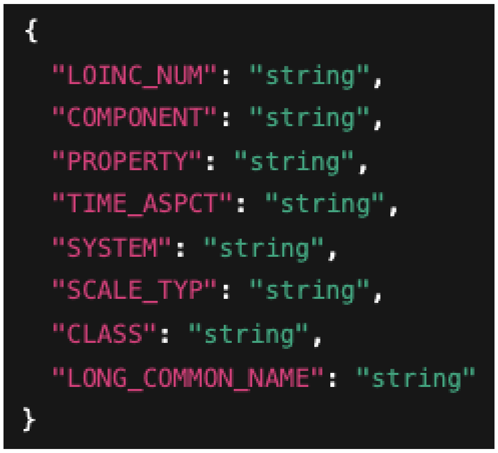
```
{
  "term": "string",
  "rxnorm": {
    "rxcui": "string",
    "name": "string",
    "tty": "string"
  },
  "synonyms": ["string"],
  "brand_names": ["string"],
  "classes": ["string"]
}
```

4. **ICD-10-CM (diagnosis codes)**:
   https://www.cdc.gov/nchs/icd/icd-10-cm/files.html



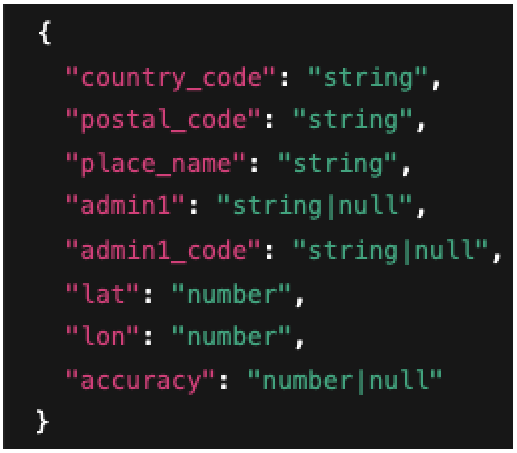
```
{
  "code": "string",
  "short_description": "string",
  "long_description": "string"
}
```

5. **5. LOINC Table Core (lab tests)**:
   https://loinc.org/download/loinc-table-core/



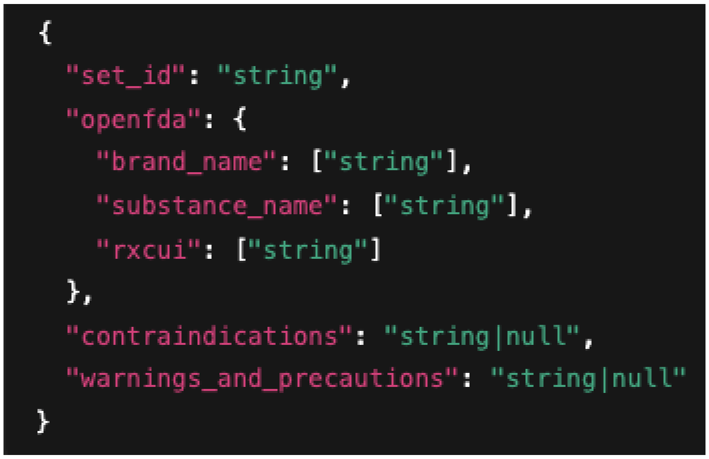
```
{
  "LOINC_NUM": "string",
  "COMPONENT": "string",
  "PROPERTY": "string",
  "TIME_ASPCT": "string",
  "SYSTEM": "string",
  "SCALE_TYP": "string",
  "CLASS": "string",
  "LONG_COMMON_NAME": "string"
}
```

6. **GeoNames (postal codes & lat/lon for sites)**:
   https://www.geonames.org/export/



```
{
  "country_code": "string",
  "postal_code": "string",
  "place_name": "string",
  "admin1": "string|null",
  "admin1_code": "string|null",
  "lat": "number",
  "lon": "number",
  "accuracy": "number|null"
}
```

7. **openFDA Drug Labels**:
   https://open.fda.gov/apis/drug/label/



```
{
  "set_id": "string",
  "openfda": {
    "brand_name": ["string"],
    "substance_name": ["string"],
    "rxcui": ["string"]
  },
  "contraindications": "string|null",
  "warnings_and_precautions": "string|null"
}
```

Squad A:

Goal: Build a reliable data pipeline that ingests trials, normalizes text/units/ontologies, and emits eligibility “atoms”.

Ingest & Store:

1. Fetch trials
2. Create Bronze-Silver Table
   a. Bronze: raw copies. Silver: cleaned
3. Geocode sites
   a. Map facilities to lat/lon using GeoNames zip/placename data.

Normalize Text, Units, Ontologies:

1. Clean eligibility text
   a. Split reliably into `inclusion_text` and `exclusion_text`; preserve character offsets.
2. Ontology mapping:
   a. Diagnoses → ICD-10-CM; Labs → LOINC; Drugs → RxNorm (resolve brand→ingredient)
3. Units & thresholds:
   a. Build deterministic converters (e.g., `mg/dL ↔ mmol/L`).
4. Temporal parsing:
   a. Turn “within 6 weeks of last chemo” → `{relation:"within", value:6, unit:"weeks", event:"last_chemo"}`.

Atomize Eligibility + QA:

1. Rule-first atom extraction:

   a. Detect atom types: `age`, `sex`, `diagnosis`, `drug`, `lab`, `biomarker`, `performance`, `temporal`, `other`.

2. QA:
   a. Sample 100 trials; compute precision/recall per atom type;

End-to-end flow(For Mapping & Atom Extraction)
Given an eligibility block: *Inclusion:* “Age ≥ 18 years”; “Hemoglobin ≥ 9 g/dL”; “Histologically confirmed melanoma.” *Exclusion:* “Prior PD-1 inhibitor therapy”; “Chemotherapy within 6 weeks prior to enrollment.”

1. The system stores the untouched block in `eligibility_text.raw_text` and the derived splits in `inclusion_text` / `exclusion_text`.

2. It emits one atom per predicate, for example:

   - a1: `{type:"age", polarity:"inclusion", operator:">=", value:"18", unit:"years", source_text:"Age ≥ 18 years"}`

   - a2: `{type:"lab", polarity:"inclusion", operator:">=", value:"9.0", unit:"g/dL", concept_json:{"loinc":"718-7","name":"Hemoglobin"}, source_text:"Hemoglobin ≥ 9 g/dL"}`

   - a3: `{type:"diagnosis", polarity:"inclusion", operator:"contains", value:"melanoma", concept_json:{"icd10":"C43.9","name":"Malignant melanoma of skin"}, source_text:"Histologically confirmed melanoma"}`

   - a4: `{type:"drug", polarity:"exclusion", operator:"contains", value:"PD-1 inhibitor", concept_json:{"rxnorm":"1547545","name":"pembrolizumab"}, source_text:"Prior PD-1 inhibitor therapy"}`

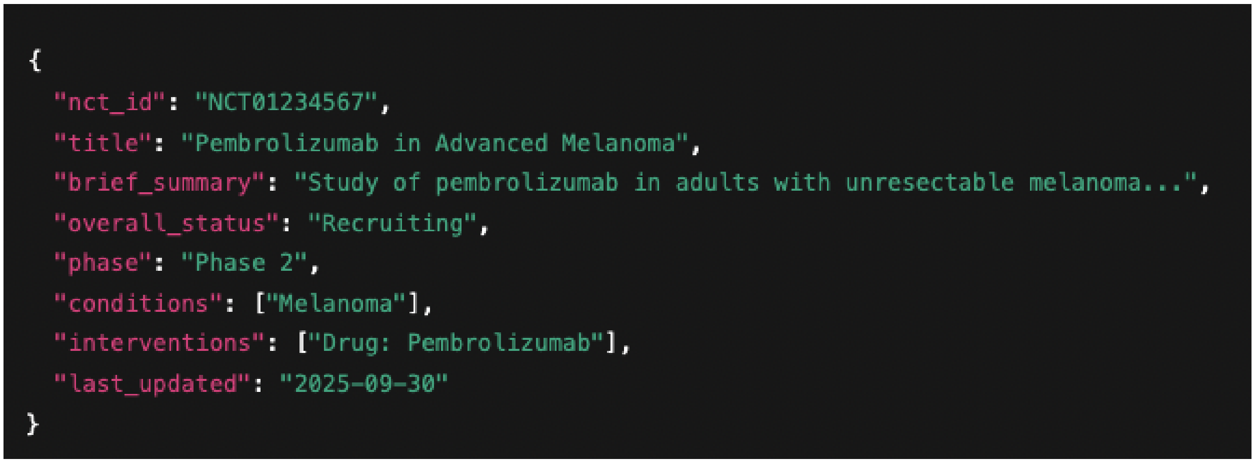
   - a5: `{type:"temporal", polarity:"exclusion", operator:"within_weeks", value:"6", unit:"weeks", temporal_window_json:{"relation":"within","value":6,"unit":"weeks","event":"last_chemo"}, source_text:"Chemotherapy within 6 weeks prior to enrollment"}`

3. `trials` then supplies searchable metadata; `eligibility_text` provides evidence lines; and `eligibility_atoms` provides machine-actionable rules, each traceable via `atom_id` and standardized via `concept_json`.

Deliverables:

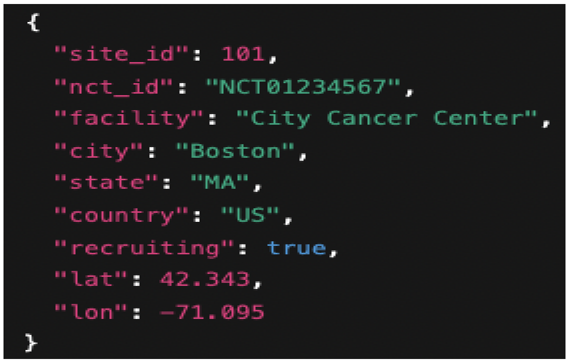
One local database with four canonical tables

1. Trials:
   This is the canonical record for each study (one row per NCT ID) and is the primary search surface. We will index `title`, `brief_summary`, `conditions`, and `interventions` for BM25/vector retrieval, and read `overall_status` and `phase` for filtering and practicality scoring.



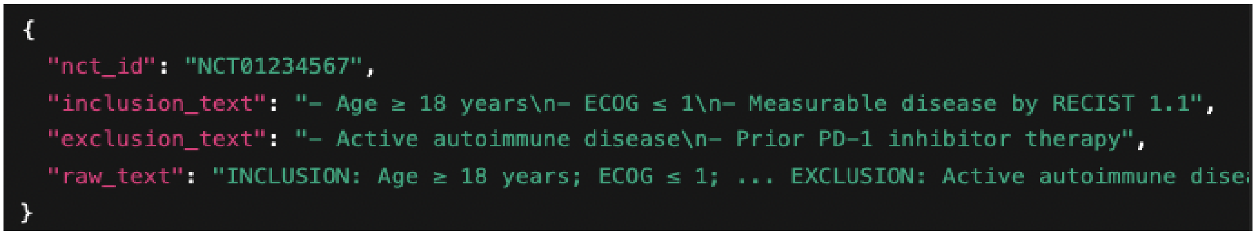
```
{
  "nct_id": "NCT01234567",
  "title": "Pembrolizumab in Advanced Melanoma",
  "brief_summary": "Study of pembrolizumab in adults with unresectable melanoma...",
  "overall_status": "Recruiting",
  "phase": "Phase 2",
  "conditions": ["Melanoma"],
  "interventions": ["Drug: Pembrolizumab"],
  "last_updated": "2025-09-30"
}
```

2. Sites:
   Each row represents a study location with optional latitude/longitude and a simple recruiting flag. It powers the “practicality” features: computing the minimum distance from the patient to any recruiting site and enabling geographic filters (country/state). The final ranking blends this distance-derived score with relevance so nearby, actively recruiting trials rank higher.



```
{
  "site_id": 101,
  "nct_id": "NCT01234567",
  "facility": "City Cancer Center",
  "city": "Boston",
  "state": "MA",
  "country": "US",
  "recruiting": true,
  "lat": 42.343,
  "lon": -71.095
}
```

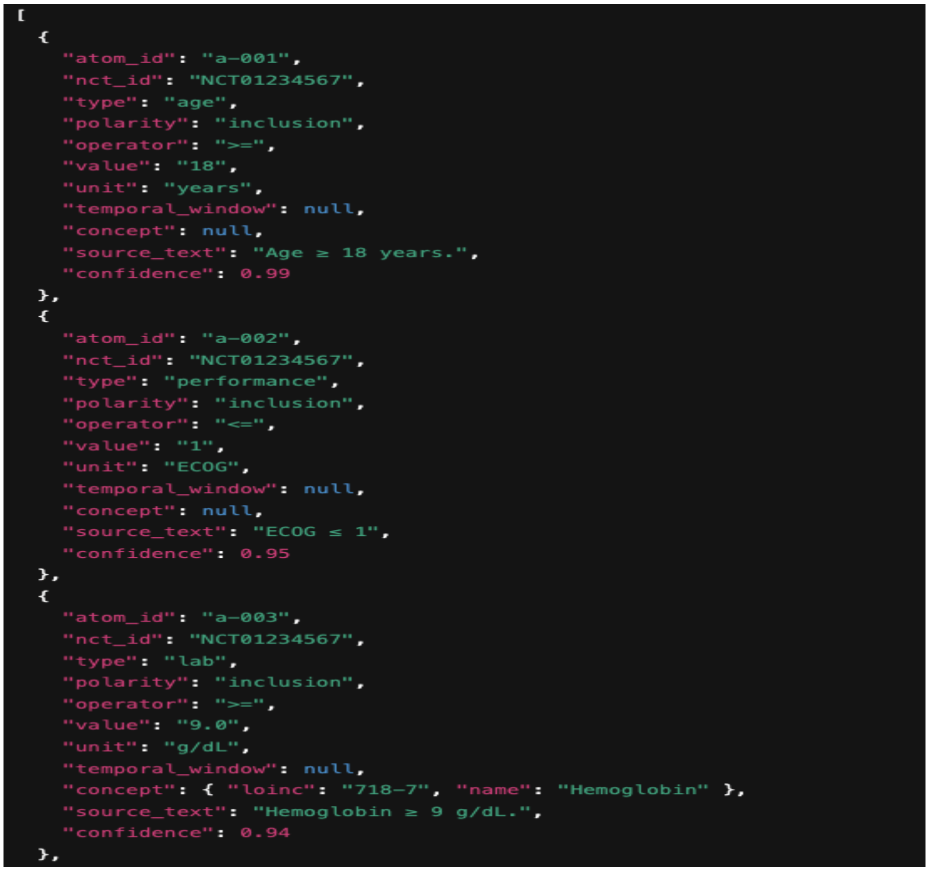
3. Eligibility Text:
   This stores the cleaned criteria prose, split into `inclusion_text` and `exclusion_text` (plus the original `raw_text`). It boosts retrieval recall when indexed (many key terms only appear in criteria), and it's the source of the "evidence lines". It also serves as the input for the atomization step, ensuring every structured atom can be traced back to the exact sentence clinicians would read.



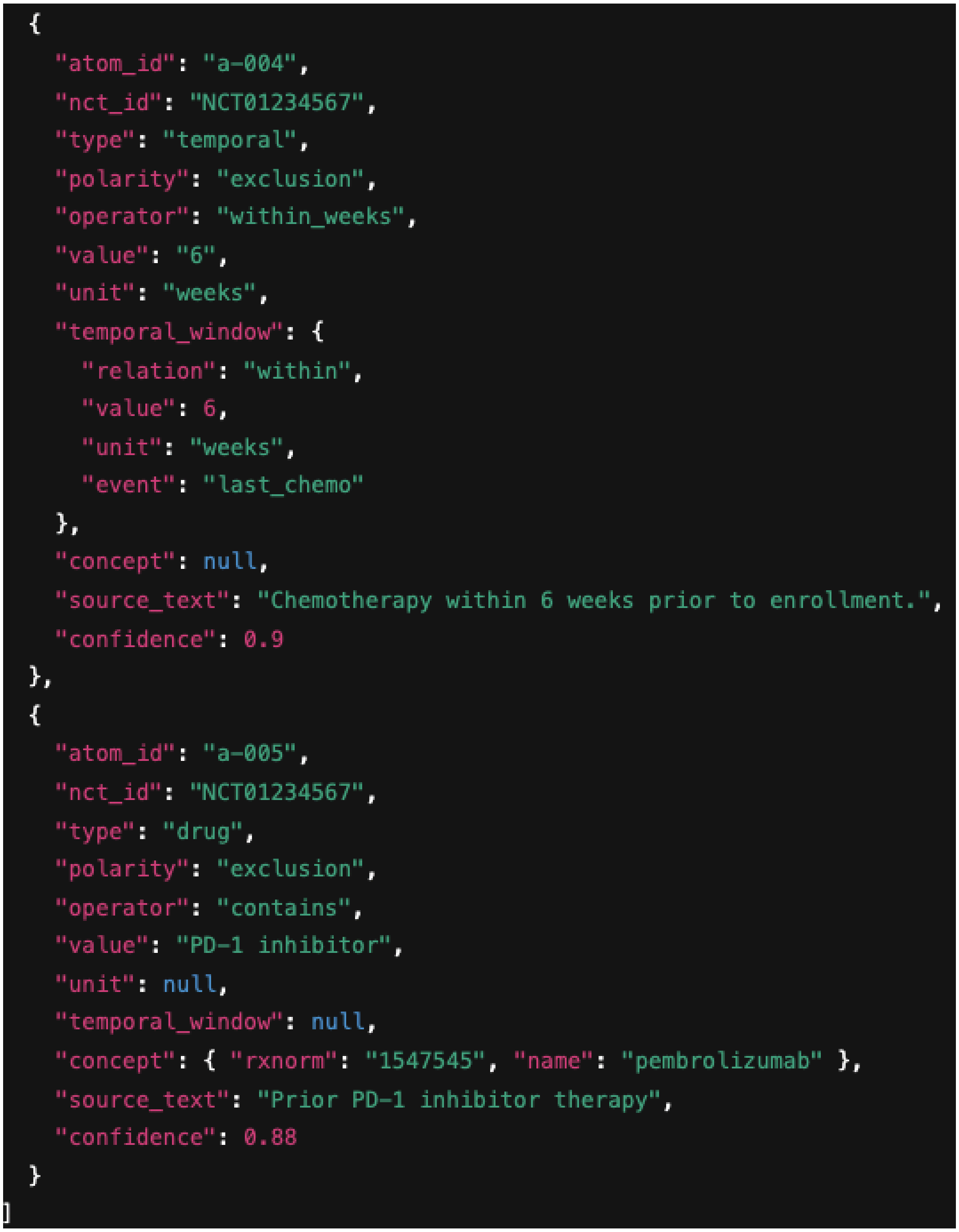
```
{
  "nct_id": "NCT01234567",
  "inclusion_text": "- Age ≥ 18 years\n- ECOG ≤ 1\n- Measurable disease by RECIST 1.1",
  "exclusion_text": "- Active autoimmune disease\n- Prior PD-1 inhibitor therapy",
  "raw_text": "INCLUSION: Age ≥ 18 years; ECOG ≤ 1; ... EXCLUSION: Active autoimmune dise
}
```

4. Eligibility Atoms:

   This is the structured, machine-actionable breakdown of criteria, one row per atomic rule (e.g., age ≥ 18; hemoglobin ≥ 9 g/dL; no PD-1 inhibitor within 6 weeks). It enables fast, deterministic checks (age/sex/numeric/temporal) today and clean, grounded prompts for an LLM later.



```
[
  {
    "atom_id": "a-001",
    "nct_id": "NCT01234567",
    "type": "age",
    "polarity": "inclusion",
    "operator": ">=",
    "value": "18",
    "unit": "years",
    "temporal_window": null,
    "concept": null,
    "source_text": "Age ≥ 18 years.",
    "confidence": 0.99
  },
  {
    "atom_id": "a-002",
    "nct_id": "NCT01234567",
    "type": "performance",
    "polarity": "inclusion",
    "operator": "<=",
    "value": "1",
    "unit": "ECOG",
    "temporal_window": null,
    "concept": null,
    "source_text": "ECOG ≤ 1",
    "confidence": 0.95
  },
  {
    "atom_id": "a-003",
    "nct_id": "NCT01234567",
    "type": "lab",
    "polarity": "inclusion",
    "operator": ">=",
    "value": "9.0",
    "unit": "g/dL",
    "temporal_window": null,
    "concept": { "loinc": "718-7", "name": "Hemoglobin" },
    "source_text": "Hemoglobin ≥ 9 g/dL.",
    "confidence": 0.94
  },
```



```
  {
    "atom_id": "a-004",
    "nct_id": "NCT01234567",
    "type": "temporal",
    "polarity": "exclusion",
    "operator": "within_weeks",
    "value": "6",
    "unit": "weeks",
    "temporal_window": {
      "relation": "within",
      "value": 6,
      "unit": "weeks",
      "event": "last_chemo"
    },
    "concept": null,
    "source_text": "Chemotherapy within 6 weeks prior to enrollment.",
    "confidence": 0.9
  },
  {
    "atom_id": "a-005",
    "nct_id": "NCT01234567",
    "type": "drug",
    "polarity": "exclusion",
    "operator": "contains",
    "value": "PD-1 inhibitor",
    "unit": null,
    "temporal_window": null,
    "concept": { "rxnorm": "1547545", "name": "pembrolizumab" },
    "source_text": "Prior PD-1 inhibitor therapy",
    "confidence": 0.88
  }
]
```

Checklist:

- ☐ Ingest & Store
    - ☐ Choose source (AACT dump or CT.gov API v2)
    - ☐ Load to **one DB** (`trials`, `sites`, `eligibility_text`).

    - □ Geocode US sites once (postal code → lat/lon)
    - □ Verify coverage: ≥95% trials have some eligibility text; ≥80% US sites have lat/lon
- □ Normalize
    - □ Clean bullets/newlines; **split** into `inclusion_text` / `exclusion_text` (store both + `raw_text`).
    - □ Drug mapping: brand → ingredient via RxNorm (store ingredient name/RxCUI where found).
    - □ Diagnosis mapping: ICD-10-CM code where obvious (string/synonym list).
    - □ Lab mapping: common tests → LOINC where obvious.
    - □ Unit converters for **short whitelist** (Hgb, ANC, creatinine, bilirubin).
    - □ Temporal parser for simple patterns (within X weeks; ≥ Y days since Z).
- □ Atomize Eligibility + QA
    - □ Rule patterns for atom types: **age, sex, diagnosis, drug, lab, biomarker, performance, temporal, other**.
    - □ Write `eligibility_atoms` with **one atom per predicate**; always include `source_text`; `confidence` 0–1.
    - □ QA on 100 trials: compute **precision/recall per atom type**;