Below are precise, P1-style instructions for preparing **data/gold/labels.jsonl** (≈200 bilingual, high-quality snippets). This file trains evaluation, retrieval hints, and generation control. Keep lines concise, spans accurate, and IDs stable.

Purpose

Provide a tiny, **curator-labeled** corpus that anchors: - schema/frame/metaphor detection, - viewpoint & attention defaults, - explosion-beat timing, - retrieval exemplars for RAG.

General Rules

- One JSON object **per line** (JSONL).
- UTF-8; **Unicode NFC** normalization (esp. Persian).
- **Zero-based, half-open** spans `[start, end)` in **character indices** (not tokens).
- Overlapping spans allowed; all spans must be within `text` length.
- Keep each `text` ≤ **160 chars**; aim for **1–2 lines**.
- Balance EN/FA; target **10–15 examples per frame** and **8–12 per key metaphor**.

Label Schema (per line)

```
{
  "id": "ex_0001",                 // unique, stable
  "lang": "en",                    // "en" | "fa"
  "text": "we cross the narrow bridge at dusk",
  "labels": {
    "schemas": [                     // 0..N
      { "id": "path", "spans": [[3, 8], [20, 26]] },
      { "id": "boundary", "spans": [[15, 21]] }
    ],
    "metaphors": [                   // 0..N; primary or bipolar ids
      { "id": "life_is_travel", "spans": [[0, 2]] },
      { "id": "raw_cooked", "pole": "raw", "spans": [] }
    ],
    "frame": { "id": "journey" }, // exactly 1 if determinable; else omit
    "viewpoint": {                   // defaults if unknown: 3rd/present/medium
      "person": "3rd",             // 1st|2nd|3rd
      "tense": "present",          // past|present
      "distance": "close"          // close|medium|far
    },
    "attention": [                   // salient spans with weights 0..1
      { "span": [20, 26], "w": 0.8 },
      { "span": [11, 17], "w": 0.5 }
    ],
    "explosion": {                   // where the semantic turn lands
      "beat": 4,                     // 1..6 (matches config/beats.yml)
      "confidence": 0.8
    }
  },
  "curator": "Mahyar",              // human annotator
  "source": "curated",              // or a short cite; avoid copyrighted text
  "license": "CC-BY",
  "confidence": 0.85,                // annotator confidence
  "notes": "bridge evokes path; dusk adds tension"
}
```

Notes - `metaphors[].pole` only for **bipolar** metaphors if a pole is explicit/strong. - `attention` weights don't need to sum to 1; cap at 4 spans per sample. - If `frame` is uncertain, omit it and mention uncertainty in `notes`.

Minimal Examples (copy-ready)

```
{"id":"ex_0001","lang":"en","text":"we cross the narrow bridge at
dusk","labels":{"schemas":[{"id":"path","spans":[[3,8],[20,26]]},
{"id":"boundary","spans":[[15,21]]}],"metaphors":
[{"id":"life_is_travel","spans":[[0,2]]}],"frame":{"id":"journey"},"viewpoint":
{"person":"3rd","tense":"present","distance":"close"},"attention":[{"span":
[20,26],"w":0.8}],"explosion":{"beat":4,"confidence":
0.8}},"curator":"Mahyar","source":"curated","license":"CC-BY","confidence":0.9}
{"id":"ex_0002","lang":"fa","text":"اتاق هنوز خام است","labels":{"schemas":
[{"id":"container","spans":[[0,4]]}],"metaphors":
[{"id":"raw_cooked","pole":"raw","spans":[[10,13]]}],"frame":
{"id":"union_separation"},"viewpoint":
{"person":"3rd","tense":"present","distance":"near"},"attention":[{"span":
[10,13],"w":0.9}],"explosion":{"beat":3,"confidence":
0.7}},"curator":"Mahyar","source":"curated","license":"CC-BY","confidence":0.85}
{"id":"ex_0003","lang":"en","text":"the door listens; we hesitate on its
lip","labels":{"schemas":[{"id":"boundary","spans":[[4,8]]},
{"id":"balance","spans":[[26,34]]}],"metaphors":[{"id":"light_dark","spans":
[]}],"frame":{"id":"threshold_crossing"},"viewpoint":
{"person":"1st","tense":"present","distance":"medium"},"attention":[{"span":
[4,8],"w":0.7},{"span":[26,34],"w":0.5}],"explosion":{"beat":2,"confidence":
0.6}},"curator":"Mahyar","source":"curated","license":"CC-BY","confidence":0.8}
```

Span Indexing Cheat-Sheet

Text: `"we cross"` - `w(0) e(1)  (2) c(3) r(4) o(5) s(6) s(7)` - Span for `"we"` is `[0,2)`; for `"cross"` is `[3,8)`. - Persian: ensure NFC; calculate spans on **final normalized string**.

Coverage Targets (for ~200 items)

- Frames: ≥ 10 frames × 10–15 examples each (overlap allowed across items).
- Schemas: cover the top 12–15 at least 8× each.
- Metaphors: 8–12 key metaphors; ≥ 8 examples each; ensure ≥ 3 bipolar cases with explicit poles.
- Languages: 50–50 EN/FA (±10%).

Annotation Rubric (quick)

- **Schemas**: label only when visually/embodiedly clear (e.g., *bridge* → `boundary` + `path`).
- **Metaphors**: tag *conceptual mappings*, not mere adjectives. Use `pole` when the **axis** is explicit.
- **Frame**: choose 1 strongest; if unsure, leave empty.
- **Viewpoint**: use textual cues (pronouns/tense); default to `3rd/present/medium` when neutral.
- **Attention**: up to 4 spans that carry the scene's focus.
- **Explosion (beat)**: the beat where the *semantic turn* lands (1..6 per `config/beats.yml`).

File Hygiene & Provenance

- No copyrighted text beyond short phrases; prioritize self-curated lines.
- Fill `curator`, `source`, `license`, and `confidence` every time.
- Keep IDs stable; don't reuse IDs after deletion.

Optional Splits

Provide index files under `data/gold/splits/`:

```
train.ids  # 140 ids
dev.ids    # 30 ids
test.ids   # 30 ids
```

Split by **ID lists** rather than copying JSONL.

JSON Schema (engineers can validate with AJV or Pydantic)

```
{
  "$schema": "http://json-schema.org/draft-07/schema#",
  "title": "SV Gold Label",
  "type": "object",
  "required": ["id", "lang", "text", "labels", "curator", "license",
"confidence"],
  "properties": {
    "id": {"type": "string", "pattern": "^[a-z0-9_\-]+$"},
    "lang": {"type": "string", "enum": ["en", "fa"]},
    "text": {"type": "string", "minLength": 1, "maxLength": 200},
    "labels": {
      "type": "object",
      "required": [],
      "properties": {
        "schemas": {
          "type": "array",
          "items": {
            "type": "object",
            "required": ["id", "spans"],
            "properties": {
              "id": {"type": "string"},
              "spans": {
                "type": "array",
                "items": {
                  "type": "array",
                  "minItems": 2,
                  "maxItems": 2,
                  "items": {"type": "integer", "minimum": 0}
                }
              }
            }
          }
        },
        "metaphors": {
          "type": "array",
          "items": {
            "type": "object",
            "required": ["id", "spans"],
            "properties": {
              "id": {"type": "string"},
              "pole": {"type": "string"},
              "spans": {"type": "array", "items": {"type": "array", "items": {"type": "integer"}}}
            }
          }
        },
        "frame": {"type": "object", "properties": {"id": {"type": "string"}}},
        "viewpoint": {
          "type": "object",
          "properties": {
            "person": {"type": "string", "enum": ["1st", "2nd", "3rd"]},
            "tense": {"type": "string", "enum": ["past", "present"]},
            "distance": {"type": "string", "enum": ["close", "medium", "far", "near"]}
          }
        },
        "attention": {
          "type": "array",
          "items": {
            "type": "object",
            "required": ["span", "w"],
            "properties": {
              "span": {"type": "array", "items": {"type": "integer"}, "minItems": 2, "maxItems": 2},
              "w": {"type": "number", "minimum": 0, "maximum": 1}
            }
          }
        },
        "explosion": {
          "type": "object",
          "properties": {
            "beat": {"type": "integer", "minimum": 1, "maximum": 6},
            "confidence": {"type": "number", "minimum": 0, "maximum": 1}
          }
        }
      }
    },
    "curator": {"type": "string"},
    "source": {"type": "string"},
    "license": {"type": "string"},
    "confidence": {"type": "number", "minimum": 0, "maximum": 1},
    "notes": {"type": "string"}
  }
}
```

Handoff to Engineering

- Treat the JSONL as **read-only** source of truth.
- Provide a loader + validator (`sv gold validate|stats`).
- Stats to compute: counts by lang/frame/schema/metaphor; span boundary checks; average attention weight; beat histogram.
- Use these items for RAG exemplars and evaluator regression tests.