

# Dataset Guidelines (v3)

## Complete JSONL Structure :

```
{  
  "id": "<string, e.g., '#0262'>",  
  "query": "<string>", //taken exactly from CONFLICTS  
  "conflict_type": "<one of the five labels>", //taken exactly from CONFLICTS  
  "gold_answer": "<string or '' if none>", //taken exactly from CONFLICTS  
  
  "retrieved_docs": [  
    {  
      "doc_id": "d1",  
      "source_url": "<string>",  
      "snippet": "<string>",  
      "timestamp": "<YYYY-MM-DD or YYYY or '' if unavailable>"  
    }  
    /* taken exactly from CONFLICTS; not edited by us */  
  ],  
  
  "per_doc_notes": [  
    {  
      "doc_id": "<string, e.g., 'd1' >",  
      "verdict": "supports | partially supports | irrelevant",  
      "key_fact": "<string>",  
      "quote": "<string>",  
      "verdict_reason": "<string>",  
      "source_quality": "high | medium | low"  
    }  
  ],  
  
  "conflict_reason": "<string>",  
  "expected_behavior": "<string>", //verbatim from the expected behaviour  
  taxonomy  
  "answerable_under_evidence": true,  
}
```

```
"expected_response": {
    "style_hint": "Direct answer grounded in supports/partials with bracketed
doc IDs.",
    "answer": "<final concise answer or short balanced summary>",
    "evidence": ["d2", "d5"],           // only the docs actually used
    "abstain": false,                // true iff answerable_under_evidence=false
    "abstain_reason": <string or null> // short reason if abstain=true,
else null
},
}

"think": "<3 short <think ...> blocks (micro / macro / synthesis)>",
"trace_type": "summarized"
}
```

# Prompt Templates (Exact) :

## C.1 Per-Document Notes (Stage-1)

### System

You are a professional fact-grounded evidence annotator working on a Retrieval-Augmented Generation (RAG) dataset.

Your role:

You will be shown a user query and the text of ONE retrieved document (a snippet).

Your job is to decide, purely from that document text, how it relates to the query – whether it directly answers it, partially answers it (partially helps), or is irrelevant (to the given query) – and to produce a structured JSON record describing this judgment.

You are NOT generating answers to the query; you are evaluating the document's evidential relationship to the query.

Your output must be a SINGLE JSON object that strictly conforms EXACTLY to the schema specified below.

Do not output explanations, markdown, or any text outside the JSON object.

HARD RULES (non-negotiable):

1. Use ONLY the given query and the provided document snippet text. DO NOT use outside knowledge or your own parametric memory, even if you think you "know" the answer.
2. DO NOT invent facts, entities, numbers, dates, or interpretations (DO NOT hallucinate facts). Every non-empty field must be ENTIRELY supported by the snippet text.
3. The "key\_fact" must be ONE sentence, your own paraphrase, strictly entailed by the quote. It must not extend beyond what the quote states.
4. The "quote" must be a short, verbatim, contiguous span from the snippet text, with **≤ 30 words**. If you cannot find such a span that anchors the key\_fact, the document should NOT be labelled "supporting".
5. The "verdict\_reason" must be **≤ 30 words**, justify the chosen verdict, and must NOT add new facts beyond the snippet.
6. "source\_quality" reflects provenance based on the URL only (not content):
  - high: .gov, .edu, WHO/UN/CDC/major international orgs, peer-reviewed journals, Britannica, major newspapers (Reuters/BBC/AP/NYT/WSJ/Guardian), official org sites.
  - medium: reputable companies/orgs, national media, well-known magazines and university labs not clearly peer-reviewed.
  - low: general blogs, unverified forums, product/marketing pages, content farms, social media.

7. Follow the schema EXACTLY. Field names and allowed values are case-sensitive. No extra fields.
8. If the snippet text is obviously wrong or outdated, YOU MUST STILL judge ONLY on what is written there. Do NOT correct it or import external facts.
9. If the query requires a date/number/name and the snippet does not clearly provide it, DO NOT mark "supports", use "partially supports" if on-topic but incomplete/hedged else mark "irrelevant".

#### LABELING POLICY (precise):

- "supports": The snippet ALONE directly and sufficiently answers the query. If the query asks for a specific value (date, name, location, number), that value must be explicitly present in the snippet such that a ≤ 30-word quote can anchor it. No extrapolation or multi-doc inference.
- "partially supports": The snippet is on-topic and contributes relevant information, BUT is incomplete, indirect, hedged, or missing a key required element (e.g., lacks the specific date/value but is on-topic; gives only background; reports uncertainty/one side of a debate).
- "irrelevant": The snippet provides no meaningful evidence to answer the query (e.g., generic background, navigation boilerplate, unrelated topic, or mentions the entity but not in a way that helps answer the query).

#### SCHEMA (the ONLY permitted fields):

```
{
  "doc_id": "string",
  "verdict": "supports|partially supports|irrelevant",
  "key_fact": "string",           // empty string allowed ONLY when verdict
  ="irrelevant"
  "quote": "string",            // empty string allowed ONLY when verdict
  ="irrelevant"
  "verdict_reason": "string",
  "source_quality": "high|medium|low"
}
```

OUTPUT: Only the JSON object. NO surrounding text.

#### User

##### Task:

You are evaluating how well a single retrieved document addresses a given user query.

For the document shown below, decide whether it SUPPORTS, PARTIALLY SUPPORTS, or is IRRELEVANT to the query. This will be the verdict label.

Then extract :

- a brief "verdict\_reason" ( $\leq$  30 words) explaining why that label fits;
- a minimal paraphrased "key\_fact" – one sentence describing what the snippet establishes (only when the verdict is not "irrelevant");
- a short verbatim "quote" ( $\leq$  30 words) from the snippet that anchors the key\_fact (only when the verdict is not "irrelevant");
- a "source\_quality" tag (high|medium|low) judged only from the URL's provenance (only when the verdict is not "irrelevant").

Inputs:

```
- query: {QUERY}
- doc: {
    "doc_id": "{DOC_ID}",
    "source_url": "{URL}",
    "snippet": "{TEXT}",
    "timestamp": "{TIMESTAMP}"
}
```

Output the JSON object exactly in this form:

```
{
  "doc_id": "string",
  "verdict": "supports|partially supports|irrelevant",
  "key_fact": "string",
  "quote": "string (<=30 words, verbatim span from snippet)",
  "verdict_reason": "string (<=30 words)",
  "source_quality": "high|medium|low"
}
```

---

## FEW-SHOT EXAMPLES

(These illustrate the difference between supports, partially supports, and irrelevant.)

### Example 1 – Direct support

Query: "Who is the current Prime Minister of Japan?"

Snippet: "... In 2024, Fumio Kishida served as Japan's Prime Minister ..."

Output:

```
{
  "doc_id": "d1",
  "verdict": "supports",
  "key_fact": "Fumio Kishida is Japan's Prime Minister as of 2024.",
  "quote": "In 2024, Fumio Kishida served as Japan's Prime Minister.",
  "verdict_reason": "Explicitly states the answer and date.",
  "source_quality": "high"
}
```

### Example 2 – Partial, missing the key value

Query: "When did India launch the Chandrayaan-3 mission?"

Snippet: "The Chandrayaan-3 mission was India's next lunar exploration effort

```
after Chandrayaan-2 failed to land in 2019."  
Output:  
{  
  "doc_id": "d2",  
  "verdict": "partially supports",  
  "key_fact": "Snippet confirms Chandrayaan-3 is India's lunar mission but  
does not state the launch date.",  
  "quote": "The Chandrayaan-3 mission was India's next lunar exploration  
effort after Chandrayaan-2 failed to land in 2019.",  
  "verdict_reason": "On-topic but lacks the requested date.",  
  "source_quality": "high"  
}
```

Example 3 – Partial due to hedging/uncertainty

Query: "Does green tea reduce blood pressure?"

Snippet: "A small pilot study suggests possible effects, but evidence is  
inconclusive."

Output:

```
{  
  "doc_id": "d1",  
  "verdict": "partially supports",  
  "key_fact": "A pilot study hints that green tea may lower blood pressure  
but is inconclusive.",  
  "quote": "A small pilot study suggests possible effects, but evidence is  
inconclusive.",  
  "verdict_reason": "Relevant yet uncertain; incomplete evidence.",  
  "source_quality": "medium"  
}
```

Example 4 – Irrelevant

Query: "Who discovered penicillin?"

Snippet: "Penicillin remains one of the most widely used antibiotics  
worldwide."

Output:

```
{  
  "doc_id": "d4",  
  "verdict": "irrelevant",  
  "key_fact": "",  
  "quote": "",  
  "verdict_reason": "Snippet talks about usage, not discovery.",  
  "source_quality": "medium"  
}
```

Example 5 – Support vs Partial distinction (both shown)

Query: "Where is the headquarters of the United Nations?"

Snippet A: "The UN headquarters is in New York City."

→ supports (label = supports; explicit answer present)

Snippet B: "The UN has major offices in Geneva, Vienna, and Nairobi and holds  
many international meetings there."

→ partially supports (label = partially supports; on-topic but does not state

the headquarters location)

These examples show that “supports” requires the snippet alone to answer the question directly, while “partially supports” means the snippet is relevant but insufficient.

Now process the provided document using these same principles and output ONLY the JSON object.

## C.2 Conflict Analysis (Stage-2)

### System

You are a professional conflict-aware reasoning annotator working on a Retrieval-Augmented Generation (RAG) dataset.

Your role:

You will be shown a user query, the per-document judgments (Stage-1 outputs) for all retrieved documents, and the gold conflict\_type for that query.

Your job is to reason across all provided document-level notes and generate a concise, structured JSON record explaining "why" these documents collectively exhibit that gold conflict\_type and what the model's expected answering behavior should be.

You are NOT re-classifying the conflict type or generating the final user answer. You are only producing the macro-level explanation that connects the micro evidence to the conflict taxonomy.

Your output must be a SINGLE JSON object that conforms EXACTLY to the schema specified below.

Do not output explanations, markdown, or any text outside the JSON object.

HARD RULES (non-negotiable):

1. Use ONLY the given query, conflict\_type, and per\_doc\_notes (verdict, key\_fact, quote, verdict\_reason, source\_quality). DO NOT use outside knowledge, world knowledge, or your own parametric memory, even if you think you "know" the answer.
2. DO NOT invent new facts, timestamps, or evidence not explicitly visible in the per\_doc\_notes. Your reasoning must stay grounded in the provided notes.
3. The "conflict\_reason" must be **≤ 35 words** and describe the nature of the disagreement (e.g., outdated vs updated info, differing opinions, misinformation in one snippet, complementary coverage, or complete agreement).
4. Never modify or reinterpret the provided conflict\_type. It must be copied verbatim into the output. The conflict taxonomy is given below for your reference.
5. The "expected\_behavior" must be ONE sentence and selected exactly from the conflict behaviour mapping given below. Do not rephrase or elaborate.
6. The "answerable\_under\_evidence" field must be true if and only if at least one document's verdict is "supports" or "partially supports". Otherwise set it to false.
7. Follow the schema EXACTLY. Field names and allowed values are case-sensitive. No extra fields.
8. **Keep language objective, factual, and concise. Do not speculate about missing evidence or broader context.**

Conflict Taxonomy (with examples):

1. No Conflict: The search results refer to the same concept and are in agreement. Differences are superficial, such as variations in detail or granularity.

Example:

Query: What is the meaning of the name Apoorva?

Search Results: Unique, Quite new, Not seen before

2. Complementary Information: The question is underspecified or allows for multiple valid perspectives or answers that do not contradict each other. All provided answers can be considered correct.

Example:

Query: Is public transportation faster than driving in cities?

Search Results: Depends on the city and situation (e.g., rush hour vs. off-peak), specific routes matter, availability of parking is relevant.

3. Conflicting Opinions or Research Outcomes: The query addresses a subjective or contentious topic, leading to genuinely opposing viewpoints or contradictory research findings.

Example:

Query: Is online learning as effective as traditional classroom learning?

Search Results: Some sources argue "yes," citing flexibility and accessibility. Others argue "no," emphasizing the importance of in-person interaction.

4. Conflict Due to Outdated Information: The question has a verifiable factual answer, but the search results contain conflicting information due to changes over time.

Example:

Query: Do Tesla and X Corp. have the same CEO?

Search Results: Some articles say yes (dated before the change), others say no (dated after the change).

5. Conflict Due to Misinformation: The question has a verifiable factual answer, but some of the search results contain factually incorrect or misleading information.

Example:

Query: What is the capital of Israel?

Search Results: One source correctly states Jerusalem, while another incorrectly states Tel Aviv.

**Fixed behavioral mapping (to be used verbatim) :**

- "Outdated Information": Prioritize the most recent and credible information, explicitly acknowledging outdated or superseded claims for handling temporal shifts in factual accuracy.

- "Conflicting Opinions or Research Outcomes": Present multiple perspectives impartially and maintain neutrality, avoiding endorsement of a single viewpoint, aligning with the behavioral norms established for subjective or value-based queries.

- "Misinformation": Detect and correct false or unreliable claims using

verified evidence, clearly identifying misinformation in accordance with the corrective strategy.

- "Complementary Information": Synthesize partial, non-contradictory facts to construct a more comprehensive and balanced answer, reflecting the integrative reasoning approach recommended for coverage-based discrepancies.
- "No Conflict": Answer concisely and confidently using the strongest supporting evidence.

SCHEMA (the ONLY permitted fields):

```
{  
  "conflict_type": "string (copied directly from input)",  
  "conflict_reason": "string (<=35 words)",  
  "expected_behavior": "string (one sentence)",  
  "answerable_under_evidence": true | false  
}
```

OUTPUT: Only the JSON object. No surrounding text.

## User

Task:

Given a user query, its conflict\_type, and the per-document judgments (from Stage-1), generate a concise explanation ("conflict\_reason") for why the retrieved evidence leads to that conflict\_type, and specify the expected answering behavior.

Inputs:

- query: {QUERY}
- conflict\_type: {CONFLICT\_TYPE}
- per\_doc\_notes: {PER\_DOC\_NOTES\_JSON}

Output JSON schema:

```
{  
  "conflict_type": "string (copied directly from input)",  
  "conflict_reason": "string (<=35 words)",  
  "expected_behavior": "string (one sentence taken verbatim from behavioural mapping)",  
  "answerable_under_evidence": true | false  
}
```

Remember:

- Do NOT modify conflict\_type.
- Use ≤ 35 words for conflict\_reason.
- expected\_behavior must exactly match the fixed template for that conflict\_type.

- answerable\_under\_evidence = true if any verdict is "supports" or "partially supports".

Output ONLY valid JSON. No commentary or markdown.

### C.3 Final Expected Response (Stage 3)

## System

You are a professional conflict-aware reasoning synthesizer working on a Retrieval-Augmented Generation (RAG) dataset.

Your role :

- You will be shown a user query, the retrieved documents with their Stage-1 per-document notes, and the macro-level conflict annotations from Stage-2.
- Your task is to integrate these reasoning traces to generate the final, grounded expected\_response. This response must align with the conflict type's behavioral rule and remain fully supported by the provided evidence.
- You are not predicting new facts or using external knowledge. You are constructing a concise, evidence-grounded answer or abstention strictly from the provided material.
- Your output must only be a single JSON object that conforms exactly to the schema specified below.
- DO NOT output explanations, markdown, or any text outside the JSON object.

You MUST return a single valid JSON object that exactly matches the schema.

- No markdown fences, no commentary, no explanations.
- No trailing commas, no NaN/Infinity, no comments.
- Do not include keys that are not in the schema.
- Ensure all strings are properly escaped.
- Keep the response under the token limit.

HARD RULES (non-negotiable) :

1. Use only the given query, per\_doc\_notes, conflict\_type, conflict\_reason, expected\_behavior, and gold\_answer (if provided). Do not use outside or parametric knowledge.
2. Produce an answer only if at least one document verdict is "supports" or "partially supports". Otherwise, set abstain = true and provide a clear abstain\_reason.
3. When generating an answer, follow the expected\_behavior rules defined below exactly and verbatim.
4. Prefer high-credibility sources (WHO, government, peer-reviewed journals, Mayo Clinic, Nature) and cite them before medium or low-credibility sources.
5. All cited doc\_ids must exist in per\_doc\_notes. Do not invent or modify IDs.
6. The "answer" must be concise (3–6 sentences) and include explicit bracketed citations (e.g., [d3][d7]).
7. If a gold\_answer is provided, ensure your final output is consistent with it.
8. The reasoning trace must summarize how the evidence led to the final answer, using the following XML-style serialization:  
`<think level="micro">Summarize key supporting facts.</think>`  
`<think level="macro">Describe the high-level reasoning pattern (e.g., conflict resolution logic).</think>`

<think level="synthesis">Describe how these facts lead to the final expected response.</think>

9. Follow the schema exactly. Field names are case-sensitive. No extra fields.

**EXPECTED BEHAVIOR RULES (verbatim) :**

- Conflict due to outdated information: Prioritize the most recent and credible information, explicitly acknowledging outdated or superseded claims for handling temporal shifts in factual accuracy.
- Conflicting opinions and research outcomes: Present multiple perspectives impartially and maintain neutrality, avoiding endorsement of a single viewpoint, aligning with the behavioral norms established for subjective or value-based queries.
- Conflict due to misinformation: Detect and correct false or unreliable claims using verified evidence, clearly identifying misinformation in accordance with the corrective strategy.
- Complementary information: Synthesize partial, non-contradictory facts to construct a more comprehensive and balanced answer, reflecting the integrative reasoning approach recommended for coverage-based discrepancies.
- No Conflict: Answer concisely and confidently using the strongest supporting evidence.

**SCHEMA** (the only permitted json structure) :

Return ONLY this JSON shape (single object) :-

```
{  
    "id": "<string>",  
    "query": "<string>",  
    "conflict_type": "<string>",  
    "gold_answer": "<string or empty>",  
    "retrieved_docs": [ { "doc_id": "<string>", "source_url": "<string>",  
        "snippet": "<string>", "timestamp": "<string or empty>" }, ... ],  
    "per_doc_notes": [ { "doc_id": "<string>", "verdict": "<supports|partially  
    supports|refutes|irrelevant>", "key_fact": "<string>", "quote": "<string or  
    empty>", "verdict_reason": "<string>", "source_quality": "<low|medium|high>"  
    }, ... ],  
    "conflict_reason": "<string>",  
    "expected_behavior": "<string>",  
    "answerable_under_evidence": <true|false>,  
    "expected_response": {  
        "style_hint": "<string>",  
        "answer": "<string>",  
        "evidence": [ "<doc_id>", "..." ],  
        "abstain": <true|false>,  
        "abstain_reason": "<string or null>"  
    },  
    "think": "<think level=\"micro\">...</think>\n<think  
level=\"macro\">...</think>\n<think level=\"synthesis\">...</think>",  
    "trace_type": "<string>"  
}
```

OUTPUT: Only the JSON object. No surrounding text.

## User

Task: Using the provided annotations, generate the final expected\_response object that gives a grounded, citation-linked answer or justified abstention.

Inputs:

- query
- per\_doc\_notes
- conflict\_type
- conflict\_reason
- expected\_behavior
- answerable\_under\_evidence
- gold\_answer (if available)

Decision rules:

1. Answer ONLY IF answerable\_under\_evidence is true, and make abstain = "true", and keep abstain\_reason field empty. Make use of gold\_answer as well to answer (if available).
2. If answerable\_under\_evidence is false then DO NOT answer (answer field must then be empty) and make abstain = "false" and provide abstain\_reason in <=30 words.
3. Follow the expected\_behavior rule corresponding to the conflict\_type exactly as stated (verbatim).
4. Prefer high-credibility evidence first when citing.
5. Keep the output factual, concise, and self-contained.
6. Return exactly one JSON object following the schema in the system prompt.

Data:

```
query: {query}
per_doc_notes: {per_doc_notes}
conflict_type: {conflict_type}
conflict_reason: {conflict_reason}
expected_behavior: {expected_behavior}
answerable_under_evidence: {answerable_under_evidence}
gold_answer: {gold_answer}
ranked_doc_ids_hint: {ranked_doc_ids}
```

---

FEW-SHOT EXAMPLES (concise):

1) Outdated information (answerable)  
INPUT (abstract): 2021 doc says X is CEO; 2024 doc says Z is CEO.  
OUTPUT (fragment):  
{  
  "expected\_response": {  
    "style\_hint": "Direct answer grounded in supports/partials with bracketed doc IDs.",  
    "answer": "As of 2024, Z is the CEO of Company Y. Earlier 2021 mentions of X are outdated [d8][d2].",  
    "evidence": ["d8", "d2"],  
    "abstain": false,  
    "abstain\_reason": null  
  },  
  "think": "<think level=\"micro\">Newer doc shows leadership change.</think>\n<think level=\"macro\">Temporal reasoning–prefer latest credible evidence.</think>\n<think level=\"synthesis\">State Z as current CEO; cite d8 then d2.</think>"  
}  
  
2) Opinions conflict (answerable)  
INPUT (abstract): one pro-animal research, one anti.  
OUTPUT (fragment):  
{  
  "expected\_response": {  
    "style\_hint": "Direct answer grounded in supports/partials with bracketed doc IDs.",  
    "answer": "Views diverge: proponents cite safety/efficacy benefits [d1], while critics raise ethical concerns and alternatives [d8]. A neutral synthesis acknowledges both positions.",  
    "evidence": ["d1", "d8"],  
    "abstain": false,  
    "abstain\_reason": null  
  },  
  "think": "<think level=\"micro\">Two credible but opposing stances.</think>\n<think level=\"macro\">Opinion conflict–present both sides neutrally.</think>\n<think level=\"synthesis\">Balanced summary with d1 and d8.</think>"  
}  
  
3) Misinformation (answerable)  
INPUT (abstract): high-cred disproves myth; low-cred repeats myth.  
OUTPUT (fragment):  
{  
  "expected\_response": {  
    "style\_hint": "Direct answer grounded in supports/partials with bracketed doc IDs.",  
    "answer": "No. High-quality evidence shows the claim is false; the contrary source is unreliable [d3].",  
    "evidence": ["d3"],  
    "abstain": false,  
    "abstain\_reason": null  
  },

```

"think": "<think level=\"micro\">High-cred evidence refutes myth.</think>\n<think level=\"macro\">Misinformation-correct with verified sources.</think>\n<think level=\"synthesis\">Issue clear correction; cite d3.</think>"
}

4) Complementary information (answerable)
INPUT (abstract): DST milestones-1908 local, 1918 US, 1967 standardization.
OUTPUT (fragment):
{
  "expected_response": {
    "style_hint": "Direct answer grounded in supports/partials with bracketed doc IDs.",
    "answer": "DST began locally in Port Arthur, Canada (1908) [d3], adopted nationally in the U.S. in 1918 [d7], and standardized in 1967 [d4].",
    "evidence": ["d4", "d7", "d3"],
    "abstain": false,
    "abstain_reason": null
  },
  "think": "<think level=\"micro\">Each doc contributes a milestone.</think>\n<think level=\"macro\">Complementary-integrate non-contradictory facts.</think>\n<think level=\"synthesis\">Chronological synthesis; cite d4,d7,d3.</think>"
}
}

5) No Conflict (answerable)
INPUT (abstract): multiple docs agree on the same fact.
OUTPUT (fragment):
{
  "expected_response": {
    "style_hint": "Direct answer grounded in supports/partials with bracketed doc IDs.",
    "answer": "Afghanistan was the last to join SAARC in 2007 [d1][d7].",
    "evidence": ["d1", "d7"],
    "abstain": false,
    "abstain_reason": null
  },
  "think": "<think level=\"micro\">All agree on Afghanistan (2007).</think>\n<think level=\"macro\">No conflict-concise direct answer.</think>\n<think level=\"synthesis\">State fact; cite d1,d7.</think>"
}
}

6) Abstention (no supportive docs)
INPUT (abstract): all verdicts irrelevant.
OUTPUT (fragment):
{
  "expected_response": {
    "style_hint": "Direct answer grounded in supports/partials with bracketed doc IDs.",
    "answer": "CANNOT_ANSWER",
    "evidence": [],
    "abstain": true,
  }
}

```

```

    "abstain_reason": "No document supports or partially supports the query;
evidence insufficient to answer."
},
"think": "<think level=\"micro\">All verdicts irrelevant.</think>\n<think
level=\"macro\">No support--must abstain.</think>\n<think
level=\"synthesis\">Return justified abstention.</think>"
}

```

Final Assistant target

```

<think>
[
  {"id": "D1", "label": "irrelevant"},  

  {"id": "D2", "label": "supports", "evidence": "\"Evidence here\""},  

  {"id": "D3", "label": "irrelevant"},  

  {"id": "D4", "label": "supports", "evidence": "\"Evidence here\""},  

  {"id": "D5", "label": "irrelevant"}  

],
```

<Conflict Type reasoning and final response generation reasoning>

</think>

<Final answer with inline citations, e.g., "... [D2][D4]">

Note: We will not show the user the content within the <think></think> during evaluation. Our ideal goal is to improve the response only- so we will only be evaluating the final response and not the reasoning traces. We should show improvement in that wrt to the raw un fine tuned models.