Annotation Guideline

Natural Language Processing: Stay On Topic

October 22, 2025

1 Purpose

The goal of this task is to generate distractors misleading, confusing, or irrelevant responses for conversational instructions. These distractors will be used to evaluate how easily large language models (LLMs) can be confused or misled in dialogue understanding and response generation.

Each distractor should be plausible enough that an LLM or human might momentarily consider it valid, yet incorrect or misleading with respect to the target instruction or conversation.

2 General Principles

1. **Focus on the target instruction.** Carefully read and understand the target instruction of the conversation. All distractors should relate to this core instruction.
2. **Identify violable components.** Determine which parts or aspects of the target instruction can be intentionally violated (e.g., factual accuracy, relevance, tone, or scope).
3. **Generate multiple distractors.** Create **3–5 unique distractors** for each conversation. Each distractor should violate a specific part of the target instruction. These can target the same or different instruction components.
4. **Vary plausibility and subtlety.** Include a mix of clearly incorrect (regular) and subtly misleading (adversarial) distractors to assess model robustness across difficulty levels.
5. **Maintain conversational realism.** All distractors should read naturally, reflecting how a real person might respond in context, even when incorrect or off-topic.
6. **Handle invalid system instructions.** If the provided `system_instruction` is incomplete, contradictory, or incoherent, mark the conversation as **"X"** in the `distractors` field instead of generating distractors.

3 Step-by-Step Process

3.1 Identify and Segment the Target Instruction

Determine **which part of the instruction or question** is most important to the task and note which **requirement, constraint, or assumption** you plan to *violate*.

3.2 Create Realistic Distractors

Imagine you are a **real person** having that conversation. Think about what kind of **misunderstandings, biases, or off-topic answers** someone might naturally produce. Make the distractor **contextually believable**, something that fits the setting or speaker profile.

3.3 Combine Adversarial and Regular Distractors

Use a mix of:

- **Adversarial distractors:** Subtle violations that are likely to trick an LLM (e.g., partially correct but logically inconsistent or semantically inverted).
- **Regular distractors:** Obvious or simple wrong answers (e.g., off-topic, unrelated, or misinterpreting the question).

3.4 Produce 3–5 Unique Distractors

For each conversation:

- Generate **3–5 distractors** that differ in style, error type, or reasoning flaw.
- Avoid repeating the same kind of mistake.
- Make a combination of **concise (1–3 sentences)** and **longer (3–5 sentences)** distractors depending on the context.

4 Quality Criteria

A good distractor should:

- ✓ Be **grammatically correct** and **fluent**.
- ✓ Be **contextually plausible** but **factually or logically wrong**.
- ✓ Reflect a **human-like misunderstanding** or **slight deviation** from the instruction.
- ✓ Be **distinct** from other distractors in the same set.

A poor distractor:

- × Repeats the same phrasing as others.
- × Is completely random or nonsensical.
- × Breaks conversational tone.
- × Is too obviously incorrect to be useful.

5 Data Format

This dataset is stored as a CSV with the following columns:

Column Name	Type / Format	Description
`domain`	String	High-level topic/category of the conversation (e.g., *insurance*, *healthcare*).
`scenario`	String	One–two sentence description of the situational context for the dialogue.
`system_instruction`	String	Instruction/role prompt specifying how the assistant should behave in this conversation.
`conversation_json`	JSON Array	Ordered list of message objects representing the dialogue turns. Each object has `role` and `content`.
`distractors`	JSON Array	List of distractor entries. Each entry targets a specific assistant turn and includes the distractor text and the violated portion of the instruction.

CSV Header

The CSV header row should be:

domain , scenario , system_instruction , conversation_json , distractors

Conversation JSON Schema

Type: JSON array of message objects (ordered).

```
[
  { "role": "user" | "assistant", "content": "<string>" },
  ...
]
```

Constraints:

- `role` must be either `"user"` or `"assistant"`.
- `content` is a string (one or more sentences).
- Turns are in chronological order.

Distractors JSON Schema

Type: JSON array of distractor objects (3–5 per conversation).

```
[
  {
    "bot turn": "<assistant turn text being targeted>",
    "distractor": "<misleading/off-topic/adversarial user utterance>",
    "target_system_instruction": "<string specifying the rule/constraint being violated>"
  },
  ...
]
```

Field definitions:

- `"bot turn"`: The exact assistant message this distractor is designed to derail or test.
- `"distractor"`: A plausible but incorrect or off-instruction user message.
- `"target_system_instruction"`: The specific clause/constraint of the system instruction that the distractor attempts to violate (e.g., "Do not recommend specific providers.").

6 Evaluation

The evaluation process ensures that all distractors meet the quality and formatting requirements outlined in this guideline.

1. **Checklist verification:** Review each conversation and its associated distractors using the established requirements. Confirm that every distractor satisfies the defined principles and quality criteria.
2. **Peer evaluation:** Each group will evaluate another group's distractor set according to the following mapping:
   - Group 6 → Group 7
   - Group 7 → Group 10
   - Group 10 → Group 6
3. **Provide feedback:** Analyze each conversation and its distractors critically. If any issues, inconsistencies, or deviations from the requirements are identified, record constructive feedback in the `note` column.

7 Other Tips

- Vary distractors by **error type** (factual, reasoning, relevance, instruction misunderstanding, emotional tone).
- Test distractors by asking: "Would an LLM plausibly choose this if it didn't fully understand the instruction?"
- Keep a **balance** between subtle confusion and clear error.