the performance is evaluated based on generation quality rather than a single correct answer.

To transform these standard single-turn benchmarks into multi-turn conversational trajectories, the authors employed a semi-automated "sharding" pipeline. This process involved using a teacher LLM to decompose a fully specified instruction into a sequence of atomic, self-contained constraints (shards), which were subsequently rewritten to mimic a user progressively clarifying requirements. These shards were rigorously validated via human annotation to ensure semantic equivalence to the original instruction.

A.2 Experimental Adaptations

In this work, we introduce two critical adaptations to the original experimental setup to strictly isolate conversational reasoning failures from other confounding factors.

Task Selection. First, we refine the scope of evaluation by focusing exclusively on the Binary Correctness Tasks (Code, Database, Actions, and Math). We explicitly exclude the Refinement Tasks (Data-to-text and Summarization) from our experiments. Our preliminary analysis indicated that LLMs often exhibit floor-level performance on these open-ended tasks even in single-turn baselines. Furthermore, the high variance inherent in the evaluation metrics for text generation complicates the analysis, making it difficult to distinguish between multi-turn context failures and general generation capability issues.

Sequential Sharding Strategy. Second, we introduce a strict control regarding information ordering. The methodology in Laban et al. (2025) randomly shuffled the order in which shards were revealed to the model. We argue that randomizing constraints often introduces artificial logical incoherence (e.g., modifying a variable before it is defined), which conflates the model's reasoning capability with its ability to handle ill-posed logical puzzles. To mitigate this, we enforce a sequential sharding strategy where we re-order the provided shards to follow a natural, logical progression (e.g., Problem Definition followed by specific constraints). This ensures that any observed performance degradation is strictly attributable to the model's inability to manage multi-turn context, rather than confusion caused by disordered inputs.

Few Shot Learning Set. The current dataset has a limited amount of data. To minimize the data required for few-shot experience learning, we randomly select 5 instances from each task to construct historical dialogues. These instances are then removed from the test set to ensure a fair comparison.

B Prompts

Prompt For Mediator

You are an intermediary AI mediator responsible for analyzing and rewriting the request based on the "Conversation Transcript between the User and the AI assistant".
Your task is to produce a clear and complete user request that can be sent to the AI assistant to obtain user's desired answer.

[Important Rules]
[[EXPS]]

[Output Format]
Output only the rewritten user request in first-person form, without any extra commentary. It should be ready to be copy-pasted and sent to another AI assistant exactly as if it came directly from the User.

[Conversation Transcript]
[[CONVERSATION]]

Please output:

Prompt For Refiner

You are a meta-analysis AI assistant responsible for extracting effective rewriting rules and best practices from examples of Conversation Transcripts between a User and an AI assistant, together with the Ground Truth rewritten request.
Your goal is to produce a clear, actionable set of RULES that can guide another "Request Rewriting Mediator" to accurately transform a given Conversation Transcript into a clear, first-person user request.
The RULES you produce must be:
- Few in number, focusing only on the most critical factors that determine correct rewriting.
- Direct, actionable, and applicable to any future Conversation Transcript.

[Input You Will Receive]
A list of examples, each containing:
1. Conversation Transcript (dialogue between the User and the AI assistant).
2. Ground Truth Rewritten Request (ideal output for that conversation).

[Your Tasks]
1. For each example:
  - Compare the Conversation Transcript with its Ground Truth request.
  - Note what transformations were applied (e.g., summarization, rephrasing, deletion of irrelevant parts, conversion to first-person, clarification of ambiguous points, tone adjustment).
2. Identify common patterns and shared decision rules across all examples.
3. Merge these insights into one unified, concise set of numbered RULES:
  - RULES must work for *new, unseen conversations*.
  - RULES must be clear, direct, and applicable without additional explanation.

[Output Format]
Output only the RULES section as:
[[RULES]]
1. ...
2. ...
3. ...
[[END]]

Do not include any commentary outside the [[RULES]] section.

[Now process the following data]
[[CONVERSATIONS]]

Example of Generated Experiences

1. Extract and consolidate all numerical values, relationships, and contextual details provided exclusively by the user across the entire conversation, ignoring AI responses and speculative examples.
2. Structure the rewritten request by:
- Starting with initial conditions or subjects
- Adding events/relationships in chronological or logical order
- Ending with the core question unchanged
3. Convert all references to the user's query subject (e.g., \"they,\" \"Andy,\" \"the hotel\") into third-person perspective and clarify ambiguous terms using precise language (e.g., \"proportional\" instead of \"go up based on,\" \"elected for\" instead of \"opted\").
4. Remove conversational artifacts (e.g., \"Oh,\" \"actually,\" \"by the way\") and partial confirmations, retaining only finalized user statements.
5. Explicitly include all mathematical relationships (e.g., \"twice as many,\" \"5% raise,\" \"three times\") and temporal conditions (e.g., \"on the first day\") even if implied in the dialogue.