1 Supplementary Experimental Results

As shown in the table 1, the results before iteration generally exhibit higher accuracy compared to those after iteration, suggesting that the original problems may be too easy to effectively evaluate the reasoning capabilities of current state-of-the-art models.

2 Prompt details

2.1 Prompt for agent generator

Please construct {num} entries related to {area}, including identity, reason, and questions, and present them in a Python list format without any additional content.

2.2 Prompt for the user agent to generate the action sequence

part 1:

Now, you are a person currently conducting an operation related to {area_name}. Based on your objective and previous operation information (if any), carry out this operation and record the expenses:

{now_memory}

Conditions to be met:

1. The operation is carried out in sessions — only one activity per session;

2. Record expense items related to {area_name}, with the specific amount for each item specified. Do not record the total of all expenses;

3. The response should be as concise and clear as possible.

Reference Information:

{ref}

part 2:

Now, you are a personnel currently conducting an operation related to {area_name}. This is your current action:

{act}

This is your previous memory:

{now_memory}

You need to organize and form new memories based on these two.

Conditions to be met:

1. Record expense items related to **{area_name}**, with the specific amount for each item specified. Do not record the total of all expenses;

2. Keep your objective in mind;

3. Your response should be as concise and clear as possible.

2.3 Prompt for the user agent to interact

Now, you are a user seeking consultation related to **{area_name}**.

Here is some information related to your issue:

{memory}

User Dialogue Requirements:

1. Based on the provided information, introduce your personal background in line with the context, maintaining a natural and conversational tone;

2. In each round of dialogue, only express one point — break down your question and explain it step by step to the assistant;

3. Each round of dialogue should be limited to 1–2 sentences, and the conversation should not exceed **{dia_len}** rounds;

4. Stay in character as the user — do not respond as an **{area_name}** assistant, and do not output "User:" or any guiding/prompting statements;

5. Within **{dia_len}** rounds, ensure that all relevant information and numerical details about the issue are conveyed.

Dialogue History:

{history}

2.4 Prompt for the assistant agent to interact

Now, you are a virtual **{area_name}** assistant.

Here is some rule information related to **{area_name}**:

{ref}

Assistant Dialogue Requirements:

1. Ensure the conversation is concise and natural;

2. Provide content solely related to **{area_name}**, referencing the rule information as needed. Give a clear response on whether expenses can be processed. If they can be processed, specify the exact amount that can be handled — you may fabricate details if necessary;

3. Output only one round of dialogue at a time, engaging in step-by-step communication with the user. There is no need to record the total sum of all expenses;

4. Maintain the role of the **{area_name}** assistant — do not respond as a user, and do not output "Assistant:" or any guiding/prompting statements;

5. Limit each round of dialogue to 1–2 sentences;

6. Do not initiate the end of the conversation.

Dialogue History:

{history}

2.5 Prompt for proxy evaluator

Coherence You are a dialogue quality evaluator. Please strictly score the semantic coherence of the provided dialogue according to the following criteria on a scale from 1 to 5:

5/5: Fully coherent and consistent; every turn closely follows the context, with clear logic and unified intent.

4/5: Mostly coherent and consistent; only minor, infrequent lapses in context or slight repetitions.

3/5: Generally understandable but contains noticeable semantic jumps, contradictions, or inconsistencies with prior content.

2/5: Poor coherence; multiple responses deviate from the topic or clearly conflict with earlier statements.

1/5: Severely inconsistent; the dialogue lacks logical connection and fails to form meaningful interaction.

Output only a single integer representing your score (e.g., 3). Do not include any other text, punctuation, spaces, or explanations.

Dialogue content below:

Model	RealReasoning math word reasoning	RealReasoning common-sense reasoning	RealReasoning
qwen-plus	76.0%	68.0%	72.7%
qwen-plus-thinking	95.5%	75.0%	87.1%
qwen-turbo	60.5%	54.0%	57.8%
qwen-turbo-thinking	93.0%	78.8%	87.2%
deepseek-r1	**96.0**%	**86.2**%	**92.0**%
deepseek-r1-distill-qwen-32b	90.5%	77.0%	85.0%
deepseek-r1-distill-qwen-1.5b	70.5%	44.0%	59.6%
deepseek-r1-distill-llama-70b	88.0%	69.0%	80.2%

Table 1: Answer accuracy of different models on RealReasoning dataset before the iterative updating of the dataset.

Fluency You are a dialogue quality evaluator. Please strictly score the fluency of the provided dialogue according to the following criteria on a scale from 1 to 5:

5/5: Extremely natural and fluent; language is idiomatic, tone appropriate, and indistinguishable from real human conversation.

4/5: Generally natural, with only occasional slightly awkward or uncommon phrasing that does not disrupt overall fluency.

3/5: Acceptable but exhibits some mechanical phrasing, repetition, or expressions that deviate from everyday usage.

2/5: Clearly unnatural; language is stiff, formulaic, or frequently uses implausible wordings.

1/5: Highly unnatural; language is bizarre, incoherent, or severely deviates from normal human communication.

Output only a single integer representing your score (e.g., 3). Do not include any other text, punctuation, spaces, or explanations.

Dialogue content below:

Diversity You are a dialogue quality evaluator. Please strictly score the lexical and stylistic diversity of the provided dialogue according to the following criteria on a scale from 1 to 5:

5/5: Highly diverse; demonstrates rich vocabulary, varied sentence structures, and creative, expressive language.

4/5: Generally diverse; minor repetitions or similar phrasings occur occasionally but do not diminish overall richness.

3/5: Moderately diverse; some responses show repetitive wording, fixed patterns, or limited stylistic variation.

2/5: Lacks diversity; frequently relies on similar expressions or templated phrases, resulting in monotony.

1/5: Extremely low diversity; nearly every turn repeats the same phrasing or mechanically reuses identical patterns.

Output only a single integer representing your score (e.g., 3). Do not include any other text, punctuation, spaces, or explanations.

Dialogue content below: