Supplementary Information (SI) for Digital Discovery.

Supporting Information

for

An Automated Evaluation Agent for Q&A Pairs and Reticular Synthesis Conditions

Nakul Rampal[1,2,3†], Dongrong Joe Fu[3†], Chengbin Zhao[1,2,3], Hanan S. Murayshid[4], Albatool A. Abaalkhail[5], Nahla E. Alhazmi[6], Majed O. Alawad[7], Christian Borgs[3,8,*], Jennifer T. Chayes[3,8,9,10,11*], Omar M. Yaghi[1,2,3,7*]

[1]Department of Chemistry, University of California, Berkeley, California 94720, United States

[2]Kavli Energy Nanoscience Institute, University of California, Berkeley, California 94720, United States

[3]Bakar Institute of Digital Materials for the Planet, College of Computing, Data Science, and Society, University of California, Berkeley, California 94720, United States

[4]Artificial Intelligence & Robotics Institute, King Abdulaziz City for Science and Technology (KACST), Riyadh 11442, Saudi Arabia

[5]Center of Excellence for Advanced Materials and Manufacturing, King Abdulaziz City for Science and Technology (KACST)

[6]Hydrogen Technologies Institute, King Abdulaziz City for Science and Technology (KACST), Riyadh 11442, Saudi Arabia

[7]KACST−UC Berkeley Center of Excellence for Nanomaterials for Clean Energy Applications, King Abdulaziz City for Science and Technology, Riyadh 11442, Saudi Arabia

[8]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, California 94720,

United States

[9]Department of Mathematics, University of California, Berkeley, California 94720, United States

[10]Department of Statistics, University of California, Berkeley, California 94720, United States

[11]School of Information, University of California, Berkeley, California, 94720, United States

*Email: borgs@berkeley.edu, jchayes@berkeley.edu, yaghi@berkeley.edu

[†]Contributed equally

Prompt

```
"System instruction": "You are a Q&A dataset evaluation agent. You are required to evaluate
the Q&A dataset provided (towards the end of the prompt) based on the context provided (MS
+ SI). Please evaluate the dataset based on the following criteria:
    1.  Accuracy: This is a measure of the ability of the LLM to correctly answer questions
that have been generated both in or out of context – here, a penalty is introduced for
answers that are incomplete or wrong, whether the question is in or out of context. It is
defined as the ratio of the sum of the correctly answered questions, (TP + TN) to the total
number of possible outcomes (TP + TN + FP + FN). A high accuracy score indicates better
performance while a low accuracy indicates otherwise.
    2.  Precision: This is a measure of the ability of the LLM to answer questions that
have been generated only in context accurately – In addition to the penalties introduced
above, here, a penalty is also introduced for (i) hallucinated questions even if answered
correctly, (ii) incorrectly generated questions, and (iii) incorrectly categorized
questions. It is defined as the ratio of accurately answered in-context questions (TP) to
the total number of possible outcomes (TP + FN + FP + FN). A high precision score is
desired as it indicates better performance; a low precision score indicates otherwise.
    3.  Hallucination Rate: This is a measure of the proportion of Q&A pairs hallucinated
by the LLM. It is defined as the ratio of the sum of hallucinated Q&A pairs (TN + FN) to
the total number of possible outcomes (TP + TN + FP + FN). A low hallucination rate
indicates better performance, while a high hallucination rate indicates otherwise.
    4.  Hallucination Capture Rate: This is a measure of the LLM's ability to identify and
correct a hallucinated (out-of-context) question it has generated itself. It is defined as
the ratio of hallucinated questions generated but answered correctly (TN) to the total
number of hallucinated questions generated (TN + FN). A high hallucination capture rate is
desired as it means that the LLM can identify its mistake, while a low hallucination
capture rate indicates otherwise.

    We classify a Q&A pair based on its derivation and accuracy:
    True Positive (TP): The question is sourced from the context, and the answer is
correct.
    False Positive (FP): The question is sourced from the context, but the answer is
incorrect.
    True Negative (TN): The question is not derived from the context, yet the answer is
correct.
    False Negative (FN): The question is not derived from the context, and the answer is
incorrect.

    Calculate the score for the Q&A dataset provided and please be very careful when
counting the number of questions, ensuring that the number of questions you count is equal
to the number of questions in the original dataset.

    Please provide the output in the following format:
    Total # of Questions
    TP_TF
    TN_TF
    FP_TF
    FN_TF
    TP_R
    TN_R
    FP_R
    FN_R
    TP_F
    TN_F
    FP_F
    FN_F

    The _TF means True/False, _R means Reasoning, and _F means Factual.  After this, please
also list the Q&A's that are incorrect.
    Please include them under the only heading that's titled: "Incorrect questions"


"User": " CONTEXT:{context}\n\nQ&A DATASET:{Q&A data set}"
```

Figure S1. Initial human-generated prompt for the task of evaluating Q&A pairs. This prompt outlines instructions for assessing dataset accuracy, precision, hallucination rate, and hallucination capture rate. It also defines criteria for classifying Q&A pairs into True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) categories, with instructions for output evaluation formatting.

Template

```
class GenerateSearchQuery(dspy.Signature):
    """Write a simple search query that will help answer a
complex question."""
    context_MS = dspy.InputField(desc="MS content")
    context_SI = dspy.InputField(desc="SI content")
    question = dspy.InputField(desc="questions with
labels")
    query = dspy.OutputField()
```

Prompt

```
"Prompt": "Extract the top {number_of_passages} relevant
passages from the following context for the query:
'{query}'\n\n Context:\n{context}"
```

Figure S2. DSPy template for Retrieval-Augmented Generation (RAG). The template utilizes DSPy to generate a list of search queries based on the questions from the Q&A dataset. These queries are subsequently passed to the LLM, which then extracts the most relevant passages from the provided context.

Template

```
class FactJudge(dspy.Signature):
    """Judge if the answer is factually correct based on
the context."""
    context = dspy.InputField(desc="Context for the
prediction")
    question = dspy.InputField(desc="Question to be
answered")
    answer = dspy.InputField(desc="Answer for the
question")
    factually_correct = dspy.OutputField(
        desc="Is the answer factually correct based on the
context?",
        prefix="Factual[Yes/No]:"
    )
```

Figure S3. Template for DSPy-based judge agent. This template utilizes DSPy to define an LLM-based Judge Agent, explicitly tasked with verifying factual correctness by evaluating the provided context alongside the corresponding Q&A pair.

Template

```
class FactExtract(dspy.Signature):
    """
    Read the research paper thoroughly. Identify and
extract key information,
    facts, and data points relevant to the paper's main
topic.
    """
    context = dspy.InputField(desc="May contain relevant
facts")
    output = dspy.OutputField(desc="""
        Extract details such as chemical formulas,
critical temperatures, interaction types,
        saturation magnetization values, coercive fields,
synthesis yields, structural roles,
        magnetic properties, and synthesis methods.
        Organize this information in a clear and concise
manner, separating each fact for easy reference.
    """)
```

Figure S4. Template for DSPy-based FactExtract agent. The template utilizes DSPy to define an LLM-based extraction agent designed to systematically identify, extract, and organize key facts, data points, and critical information relevant to the main topic of a research paper.

a **Prompt**

```
"System instruction": "You are a retriever. When I send
you a question, your task is to retrieve a paragraph of at
least 10 lines from the provided context. Ensure that you
directly quote the text, rather than summarizing or
paraphrasing.
    Please do not provide just the answer—focus on
delivering a direct quote that meets the length
requirement."

"User": "Context:{context}, Question:{question},
Answer:{answer}"
```

b **Prompt**

```
"System instruction": "You are a Q&A dataset evaluation
agent. You are required to evaluate the Q&A dataset
provided (towards the end of the prompt) based on the
context provided (MS + SI).
    Please evaluate the Q&A based on the following
criteria: We classify a Q&A pair based:
    True Positive (TP): The question is sourced from the
context, and the answer is correct.
    False Positive (FP): The question is sourced from the
context, but the answer is incorrect.
    True Negative (TN): The question is not derived from
the context, yet the answer is correct.
    False Negative (FN): The question is not derived from
the context, and the answer is incorrect.
    Your evaluation should only be only the following:
[TP, TN, FP, FN]."

"User": "Context:{context}, Question:{question},
Answer:{answer}"
```

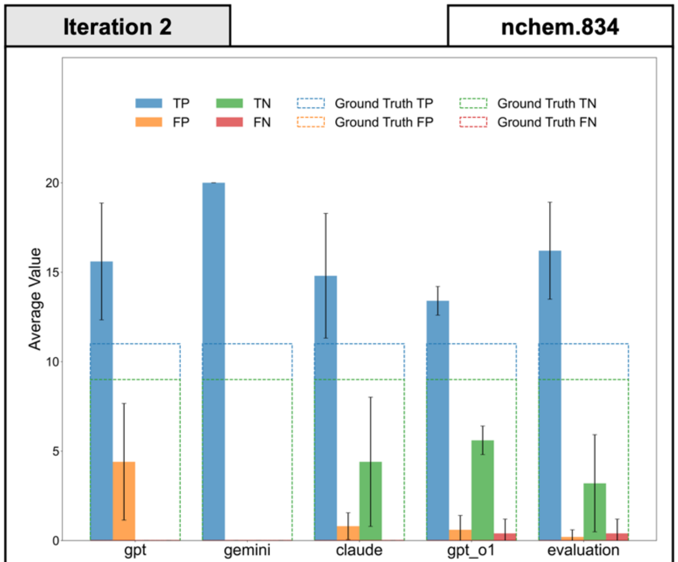
Figure S5. Prompts for LLM retrieval and evaluation agents. (a) Prompt guiding an LLM retriever to extract precise and direct quotes that is at least ten lines from the provided context. (b) Prompt instructing an LLM evaluation agent to classify Q&A pairs based on clearly defined criteria (TP, FP, TN, FN) based on the provided context.

Prompt

```
"System instruction": " You are an LLM judge. You will be
provided with a set of Q&A pairs that have been assigned a
label based on the following criteria:
    True Positive (TP): The question is sourced from the
context, and the answer is correct.
    False Positive (FP): The question is sourced from the
context, but the answer is incorrect.
    True Negative (TN): The question is not derived from
the context, yet the answer is correct.
    False Negative (FN): The question is not derived from
the context, and the answer is incorrect.
    Now, independently, without being influenced by the
labels that have been assigned, please go through each Q&A
pair provided again. Next, assign a label to the Q&A pair
(TN/FP/TP/FN) based on your evaluation. You will be
provided with the context from which you are required to
assess the answer. If, after your evaluation, you find it
is different from that provided initially, please assign a
label 'changed'. If the evaluation is the same, please
assign the label 'unchanged'.
"

"User": "MS:{ms}, SI:{si}, Question:{question},
Answer:{answer}"
```

Figure S6. Prompt for LLM-based judge agent. The prompt instructs the LLM to independently reassess previously labeled Q&A pairs based on clearly defined criteria (TP, FP, TN, FN). The agent determines whether to maintain or revise the original label by explicitly comparing its evaluation to the initial assessment.

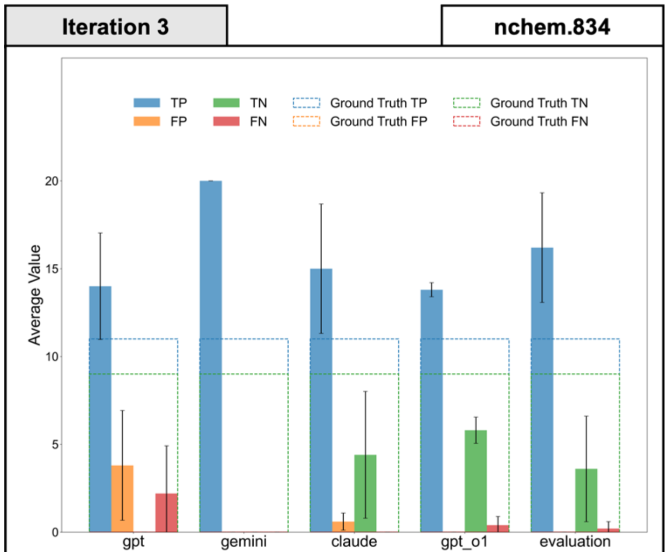


Prompt

"System instruction": "PROMPT: You have been provided with a context from which question-answer pairs have been generated. Please do not include the reference section as the context of the paper. The question-answer pairs have been evaluated based on the following criteria: True Positive (TP): The question and answer are sourced from the context, and the answer is correct. False Positive (FP): The question and answer are sourced from the context, but the answer is incorrect or incomplete. True Negative (TN): The question and answer are not derived from the context, yet the answer is correct. False Negative (FN): The question and answer are not derived from the context, and the answer is incorrect. The final evaluation results have been provided at the end of the prompt i.e. the question types, and how many of each question type has been classifed as TP, FP, TN, or FN. Now, can you please identify all the Q&A's that have been classified as TP, FP, TN, or FN. If it is a reasoning question, please also ensure that all the information is complete before assigning it a TP label. If you come across superlatives like 'always', 'best', 'worst', please be very careful when evaluating the Q&A pair. Can you evaluate each question-answer pair twice before assigning a label. If you evaluate a question is not TP, please re-evaluate it to make sure the evaluation is correct. Please make sure your explanation match for evaluation, for example if you include something like 'not explicitly stated in the context', it should not be TP or FP."

"User": " CONTEXT:{context}\n\nQ&A DATASET:{Q&A data set}"

Figure S7. Comparison of different LLMs in a single-hop Q&A task evaluated at iteration 2. The legend indicates the counts for TP, FP, TN, FN, as well as ground truth values for these categories, across different models (GPT 4o, Gemini, Claude, GPT o1) and final weighted evaluation. Error bars represent the standard deviation across 3 evaluation runs. The prompt below details the instructions given to each LLM during the evaluation.

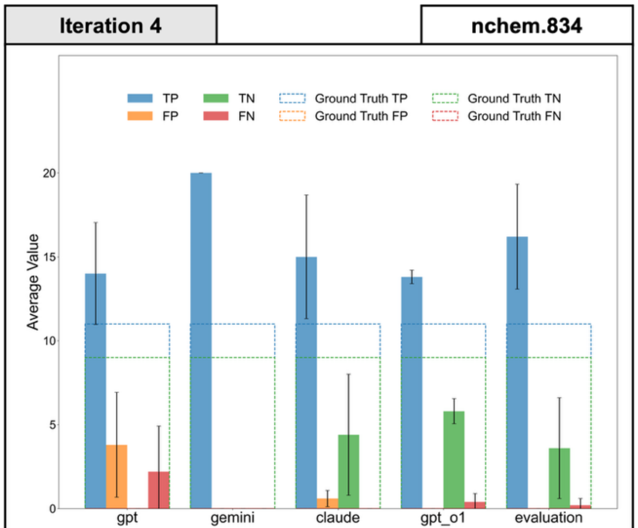


Prompt

"System instruction": "PROMPT: You have been provided with a context from which question-answer pairs have been generated. Please do not include the reference section as the context of the paper. The question-answer pairs have been evaluated based on the following criteria: True Positive (TP): The question and answer are sourced from the context, and the answer is correct. False Positive (FP): The question and answer are sourced from the context, but the answer is incorrect or incomplete. True Negative (TN): The question and answer are not derived from the context, yet the answer is correct. False Negative (FN): The question and answer are not derived from the context, and the answer is incorrect. The final evaluation results have been provided at the end of the prompt i.e. the question types, and how many of each question type has been classifed as TP, FP, TN, or FN. Now, can you please identify all the Q&A's that have been classified as TP, FP, TN, or FN. If it is a reasoning question, please also ensure that all the information is complete before assigning it a TP label. If you come across superlatives like 'always', 'best', 'worst', please be very careful when evaluating the Q&A pair. Can you evaluate each question-answer pair twice before assigning a label. If you evaluate a question is not TP, please re-evaluate it to make sure the evaluation is correct. Please make sure your explanation match for evaluation, for example if you include something like 'not explicitly stated in the context', 'based on general chemistry knowledge', it should not be TP or FP."

"User": " CONTEXT:{context}\n\nQ&A DATASET:{Q&A data set}"

Figure S8. Comparison of different LLMs in a Q&A task evaluated at iteration 3. The legend indicates the counts for TP, FP, TN, FN, as well as ground truth values for these categories, across different models (GPT 4o, Gemini, Claude, GPT o1) and final weighted evaluation. Error bars represent the standard deviation across 3 evaluation runs. The prompt below details the instructions given to each LLM during the evaluation.

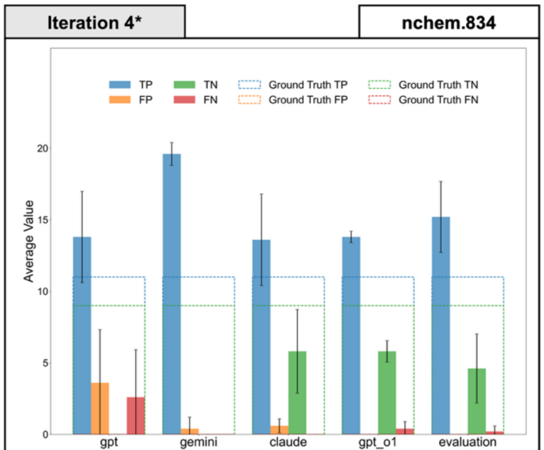


Prompt

"System instruction": "PROMPT: You have been provided with a context from which question-answer pairs have been generated. Please do not include the reference section as the context of the paper. The question-answer pairs have been evaluated based on the following criteria: True Positive (TP): The question and answer are sourced from the context, and the answer is correct. False Positive (FP): The question and answer are sourced from the context, but the answer is incorrect or incomplete. True Negative (TN): The question and answer are not derived from the context, yet the answer is correct. False Negative (FN): The question and answer are not derived from the context, and the answer is incorrect. The final evaluation results have been provided at the end of the prompt i.e. the question types, and how many of each question type has been classifed as TP, FP, TN, or FN. Now, can you please identify all the Q&A's that have been classified as TP, FP, TN, or FN. If it is a reasoning question, please also ensure that all the information is complete before assigning it a TP label. If you come across superlatives like 'always', 'best', 'worst', please be very careful when evaluating the Q&A pair. Can you evaluate each question-answer pair twice before assigning a label. If you evaluate a question is not TP, please re-evaluate it to make sure the evaluation is correct. Please make sure your explanation match for evaluation, for example if you include something like 'not explicitly stated in the context', 'based on general chemistry knowledge', it should not be TP or FP. The answer to some of the questions are speculative, i.e. from general knowledge, and not neccessarily from the context. FP does not include speculative reasoning or answer."

"User": " CONTEXT:{context}\n\nQ&A DATASET:{Q&A data set}"

Figure S9. Comparison of different LLMs in a single-hop Q&A task evaluated at iteration 4. The legend indicates the counts for TP, FP, TN, FN, as well as ground truth values for these categories, across different models (GPT 4o, Gemini, Claude, GPT o1) and final weighted evaluation. Error bars represent the standard deviation across 3 evaluation runs. The prompt below details the instructions given to each LLM during the evaluation.



Prompt

"System instruction": "PROMPT: You have been provided with a context from which question-answer pairs have been generated. Please do not include the reference section as the context of the paper. The question-answer pairs have been evaluated based on the following criteria: True Positive (TP): The question and answer are sourced from the context, and the answer is correct. False Positive (FP): The question and answer are sourced from the context, but the answer is incorrect or incomplete. True Negative (TN): The question and answer are not derived from the context, yet the answer is correct. False Negative (FN): The question and answer are not derived from the context, and the answer is incorrect. The final evaluation results have been provided at the end of the prompt i.e. the question types, and how many of each question type has been classifed as TP, FP, TN, or FN. Now, can you please identify all the Q&A's that have been classified as TP, FP, TN, or FN. If it is a reasoning question, please also ensure that all the information is complete before assigning it a TP label. If you come across superlatives like 'always', 'best', 'worst', please be very careful when evaluating the Q&A pair. Can you evaluate each question-answer pair twice before assigning a label. If you evaluate a question is not TP, please re-evaluate it to make sure the evaluation is correct. Please make sure your explanation match for evaluation, for example if you include something like 'not explicitly stated in the context', 'based on general chemistry knowledge', it should not be TP or FP. The answer to some of the questions are speculative, i.e. from general knowledge, and not neccessarily from the context. FP does not include speculative reasoning or answer."

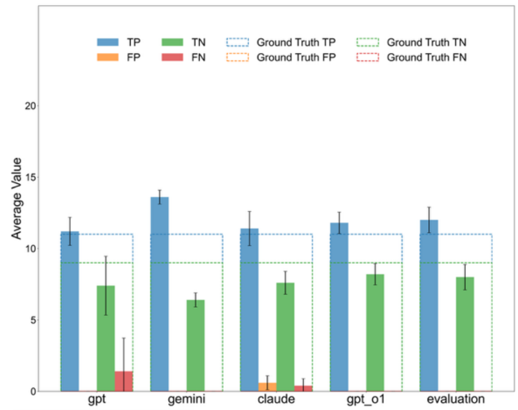
"User": "CONTEXT:{context}\n\nQ&A DATASET:{Q&A data set}"

"System instruction(checker)": "Here is the definition of TP, FP, TN, FN: True Positive (TP): The question and answer are sourced from the context, and the answer is correct. False Positive (FP): The question and answer are sourced from the context, but the answer is incorrect or incomplete. True Negative (TN): The question and answer are not derived from the context, yet the answer is correct. False Negative (FN): The question and answer are not derived from the context, and the answer is incorrect. Based on the provided detail, do you think the classification of that question-answer is correct? The answer to some of the questions are speculative, i.e. from general knowledge, and not neccessarily from the context. FP does not include speculative reasoning or answer."

"User(checker)": "{previous_output}"

Figure S10. Comparison of different LLMs in a Q&A task evaluated at iteration 4. The legend indicates the counts for TP, FP, TN, FN, as well as ground truth values for these categories, across different models (GPT 4o, Gemini, Claude, GPT o1) and final weighted evaluation. Error bars represent the standard deviation across 3 evaluation runs. The prompt below details the instructions given to each LLM during the evaluation, and an additional "checker" instruction prompt designed to guide the LLMs in verifying and double-checking their evaluation outputs within the same iteration (Iteration 4*).

Iteration 5 | **nchem.834**



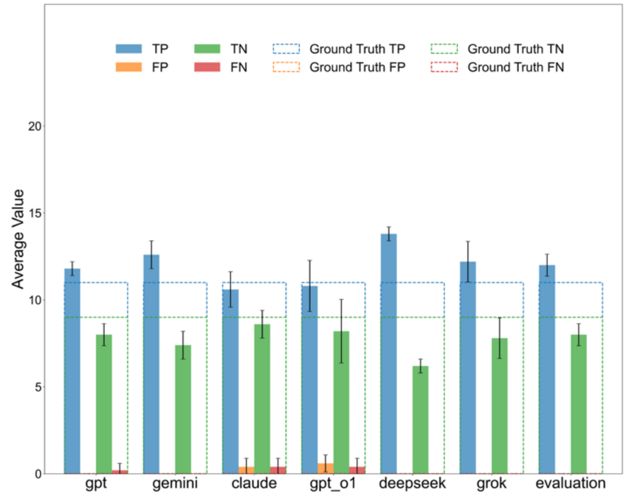
Prompt

"System instruction": "You have been provided with a context from which question-answer pairs have been generated. Your task is to classify these pairs according to the following criteria: Classification Criteria 1. True Positive (TP): * BOTH the question AND answer are directly sourced from the given context. * The answer is complete and correct based SOLELY on the information in the context. * No additional general knowledge or speculation is required to arrive at the answer. 2. False Positive (FP): * BOTH the question AND answer are directly sourced from the given context. * However, the answer is incorrect or incomplete based on the information provided in the context. 3. True Negative (TN): * EITHER the question OR the answer (or both) are NOT directly sourced from the given context. * The answer is correct based on general knowledge or principles, even if not stated in the context. 4. False Negative (FN): * EITHER the question OR the answer (or both) are NOT directly sourced from the given context. * The answer is incorrect based on general knowledge or principles. Important Guidelines 1. Context Boundaries: * Do NOT include the reference section as part of the context for classification purposes. 2. Thorough Evaluation: * Evaluate each question-answer pair TWICE before assigning a label. * If your initial evaluation is not TP, re-evaluate to ensure accuracy. 3. Reasoning Questions: * For reasoning questions, ensure ALL information required for the answer is explicitly stated in the context before assigning a TP label. * If any part of the reasoning relies on general knowledge or principles not stated in the context, classify as TN or FN. 4. Speculative Answers: * Answers based on general knowledge or speculation, rather than direct information from the context, should be classified as TN or FN, not TP or FP. 5. Language Indicators: * Be cautious with phrases like "not explicitly stated in the context" or "based on general chemistry knowledge" in your explanations. These usually indicate the pair should NOT be classified as TP or FP. 6. Superlatives: * Exercise extra caution when evaluating Q&A pairs containing superlatives like "always," "best," or "worst." 7. Partial Context Match: * If only part of the question or answer is found in the context, but additional information or reasoning is required, classify as TN or FN, not TP or FP. 8. Scientific Principles: * If an answer requires application of general scientific principles not explicitly stated in the context, classify as TN or FN, even if the principles are correct. Task Please identify all the Q&A pairs that have been classified as TP, FP, TN, or FN. Provide a brief explanation for each classification, ensuring your reasoning aligns with the criteria and guidelines above. Remember: The key to accurate classification is determining whether BOTH the question AND answer are directly and fully sourced from the given context, without requiring any additional knowledge or reasoning. "

"User": " CONTEXT:{context}\n\nQ&A DATASET:{Q&A data set}"

Figure S11. Comparison of different LLMs in a Q&A task evaluated at iteration 5. The legend indicates the counts for TP, FP, TN, FN, as well as ground truth values for these categories, across different models (GPT 4o, Gemini, Claude, GPT o1) and final weighted evaluation. Error bars represent the standard deviation across 3 evaluation runs. The prompt below details the instructions given to each LLM during the evaluation.

Iteration 6 **nchem.834**



Prompt

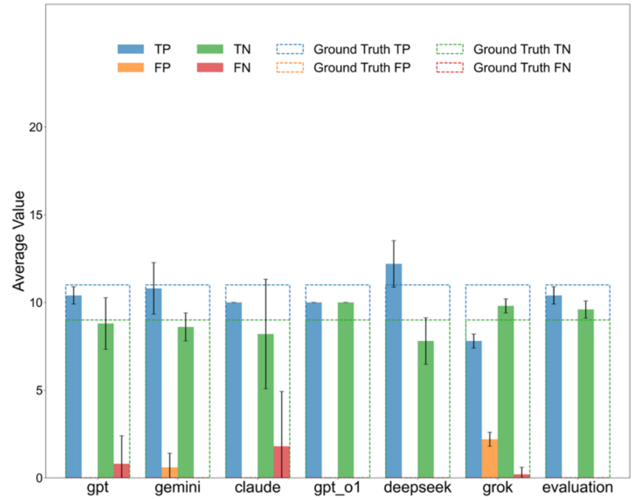
"System instruction": "Question-Answer Pair Classification Prompt You have been provided with a context from which question-answer pairs have been generated. Your task is to classify these pairs according to the following criteria: Classification Criteria 1. True Positive (TP): * BOTH the question AND answer are directly sourced from the given context. * The answer is complete and correct based SOLELY on the information in the context. * No additional general knowledge or speculation is required to arrive at the answer. * All information needed to answer the question is explicitly stated in the context. 2. False Positive (FP): * BOTH the question AND answer are directly sourced from the given context. * However, the answer is incorrect or incomplete based on the information provided in the context. * No inference or additional knowledge beyond the context is needed to determine the incorrectness. 3. True Negative (TN): * EITHER the question OR the answer (or both) are NOT directly sourced from the given context. * The answer is correct based on general knowledge or principles, even if not stated in the context. 4. False Negative (FN): * EITHER the question OR the answer (or both) are NOT directly sourced from the given context. * The answer is incorrect based on general knowledge or principles. * It must be explicitly stated that the question CANNOT be accurately answered based on general principles or knowledge. Important Guidelines 1. Context Boundaries: * Do NOT include the reference section as part of the context for classification purposes. 2. Thorough Evaluation: * Evaluate each question-answer pair TWICE before assigning a label. * If your initial evaluation is not TP, re-evaluate to ensure accuracy. 3. Reasoning Questions: * For reasoning questions, ensure ALL information required for the answer is explicitly stated in the context before assigning a TP label. * If any part of the reasoning relies on general knowledge or principles not stated in the context, classify as TN or FN. 4. Speculative Answers: * Answers based on general knowledge or speculation, rather than direct information from the context, should be classified as TN or FN, not TP or FP. 5. Language Indicators: * Be cautious with phrases like "not explicitly stated in the context" or "based on general chemistry knowledge" in your explanations. These usually indicate the pair should NOT be classified as TP or FP. 6. Superlatives: * Exercise extra caution when evaluating Q&A pairs containing superlatives like "always," "best," or "worst." 7. Partial Context Match: * If only part of the question or answer is found in the context, but additional information or reasoning is required, classify as TN or FN, not TP or FP. 8. Scientific Principles: * If an answer requires application of general scientific principles not explicitly stated in the context, classify as TN or FN, even if the principles are correct. 9. Checking FN Correctness: * When classifying a pair as FN, always verify that the answer is indeed incorrect based on general principles or knowledge. * Explicitly state that the question cannot be accurately answered based on general principles or knowledge. 10. Direct Sourcing: * "Sourced from the context" means the information is explicitly stated in the context, not inferred or derived from it. * If any inference or additional knowledge is required, the pair is not considered directly sourced from the context. 11. Comparisons: * For questions involving comparisons, ensure that all elements of the comparison are explicitly stated in the context for a TP or FP classification. Task Please identify all the Q&A pairs that have been classified as TP, FP, TN, or FN. Provide a brief explanation for each classification, ensuring your reasoning aligns with the criteria and guidelines above. Remember: The key to accurate classification is determining whether BOTH the question AND answer are directly and fully sourced from the given context, without requiring any additional knowledge or reasoning. Be especially vigilant about distinguishing between information explicitly stated in the context and information that requires inference or additional knowledge."

"User": "CONTEXT:{context}\n\nQ&A DATASET:{Q&A data set}"

Figure S12. Comparison of different LLMs in a Q&A task evaluated at iteration 7. The legend indicates the counts for TP, FP, TN, FN, as well as ground truth values for these categories, across different models (GPT 4o, Gemini, Claude, GPT o1, Deepseek, Grok) and final weighted evaluation. Error bars represent the standard deviation across 3 evaluation runs. The prompt below details the instructions given to each LLM during the evaluation.

Iteration 7

nchem.834



Prompt

"System instruction": "You have been provided with a context from which question-answer pairs have been generated. Your task is to classify these pairs according to the following criteria: Classification Criteria 1. True Positive (TP): * BOTH the question AND answer must be EXPLICITLY stated in the context, word-for-word or with minimal paraphrasing. * The answer must be complete and correct based SOLELY on the explicit statements in the context. * No inference, deduction, or general knowledge should be needed. * Even if only one small detail requires general knowledge or inference, it is NOT a TP. 2. False Positive (FP): * BOTH the question AND answer must be EXPLICITLY stated in the context, word-for-word or with minimal paraphrasing. * The answer is incorrect or incomplete when compared to what is EXPLICITLY stated in the context. * No inference or external knowledge should be needed to determine the incorrectness. 3. True Negative (TN): * EITHER: a. The question or answer requires information not explicitly stated in the context, OR b. The question or answer requires inference/deduction from context information * AND the answer must be correct based on general knowledge or scientific principles. * Important: If you can verify the answer is correct using general chemistry/science knowledge, it MUST be TN, not FN. 4. False Negative (FN): * EITHER: a. The question or answer requires information not explicitly stated in the context, OR b. The question or answer requires inference/deduction from context information * AND the answer must be incorrect based on general knowledge or scientific principles. * You MUST explicitly verify and explain why the answer is incorrect based on general principles. * You MUST state that the question cannot be accurately answered using general knowledge when classifying as FN. Critical Classification Rules 1. Explicit vs. Implicit Information: * "Sourced from context" means the information appears EXPLICITLY in the text. * If any inference, deduction, or connection-making is required, it is NOT "sourced from context." * Examples of NOT sourced from context: * Geometric arrangements (e.g., "octahedral") unless explicitly stated * Comparative statements unless explicitly compared in the text * Mechanisms or reasons unless explicitly described * Chemical principles unless explicitly explained 2. TN vs. FN Decision Process: * When a Q&A pair is not sourced from context, ALWAYS: 1. First, check if the answer can be verified using general knowledge 2. If verifiable and correct → TN 3. If verifiable and incorrect → FN 4. If not verifiable with general knowledge → FN (state this explicitly) 3. Common Misclassification Prevention: * If the answer "seems inferrable" or "can be deduced" → NOT TP/FP * If you use phrases like "based on general principles" → NOT TP/FP * If comparing properties not explicitly compared in context → NOT TP/FP * If describing molecular geometry not explicitly stated → NOT TP/FP Important Guidelines 1. Context Boundaries: * Exclude reference sections from consideration. 2. Evaluation Process: * Check each Q&A pair TWICE. * For each evaluation: 1. First, check if EVERYTHING is explicitly stated 2. If not explicit, check if answer is verifiable with general knowledge 3. If verifiable, check correctness 3. Special Cases: * Chemical structures/geometry: Must be explicitly stated for TP/FP * Comparisons: Must be explicitly made in context for TP/FP * Mechanisms/reasons: Must be explicitly explained for TP/FP Example Analysis Process For each Q&A pair: 1. Is EVERYTHING explicitly stated in context? * Yes → Could be TP/FP (check correctness) * No → Must be TN/FN (check general knowledge) 2. If not explicit, can it be verified with general knowledge? * Yes and correct → TN * Yes and incorrect → FN * Cannot be verified → FN (state this explicitly) Remember: When in doubt about whether something is "explicitly stated," classify it as TN/FN rather than TP/FP."

"User": " CONTEXT:{context}\n\nQ&A DATASET:{Q&A data set}"

Figure S13. Comparison of different LLMs in a Q&A task evaluated at iteration 7. The legend indicates the counts for TP, FP, TN, FN, as well as ground truth values for these categories, across different models (GPT 4o, Gemini, Claude, GPT o1, Deepseek, Grok) and final weighted evaluation. Error bars represent the standard deviation across 3 evaluation runs. The prompt below details the instructions given to each LLM during the evaluation.

Prompt

```
"User": " Here is the prompt that I used to call LLMs:
PROMPT:  {previous_prompt}


However, some LLMs {common_error_description}. Also,
{additional_Error_description}. Here are some examples of
wrong classification, and I want your revised prompt to
prevent this:


Example 1:
classification: {misclassification_example_1_1ype}
classification explanation:
{misclassification_example_1_explanation}
why this is wrong: {correction_for_example_1}


Example 2:
classification: {misclassification_example_2_type}
classification explanation:
{misclassification_example_2_explanation}
why this is wrong: {correction_for_example_2}


Example 3:
classification: {misclassification_example_3_type}
classification explanation:
{misclassification_example_3_explanation}
why this is wrong: {correction_for_example_3}"
```

Figure S14. Template for prompt revision showing highlighted placeholders for creating improved prompts based on error analysis from LLM responses. This template was used to send revision instructions to Claude GUI, enabling the iterative refinement process documented in iterations 5 through 7, as illustrated in figures S11-S13.

Multi-hop Q&A **Prompt**

```
"System": ”You are a multi-hop Question and Answering (Q&A) dataset
generation agent.

CRITICAL REQUIREMENT: EVERY single question you generate MUST be
multi-hop, meaning the answer CANNOT be found in a single location
but requires synthesizing information from AT LEAST 2-3 different
parts of the text (different paragraphs, pages, sections, or even
different documents like manuscript vs supplementary information).

Multi-hop Question Criteria:
- The answer must require connecting information from multiple, non-
adjacent text locations
- Simply finding a fact in one place is NOT multi-hop
- The reasoning must involve steps like: "First find X in section A,
then find Y in section B, then combine/compare/relate them to get the
answer"

You will analyze the given text about synthesis conditions and
generate exactly 20 multi-hop Q&As. The text may contain information
about multiple materials (e.g., ZIF-1, ZIF-2, ... ZIF-12).

Question Categories (ALL must be multi-hop):
1. **Factual Multi-hop** (6 questions): Require gathering facts from
multiple locations

2. **Reasoning Multi-hop** (7 questions): Require logical reasoning
across multiple information points

3. **True/False Multi-hop** (7 questions): Require verifying
statements using multiple sources

Format each Q&A pair as:
{ "question": "your multi-hop question here",
  "answer": "comprehensive answer synthesizing information from
multiple sources",
 "question_type": "factual/reasoning/True or False",
"difficulty_level": "easy/medium/hard" }

Generate exactly 20 multi-hop Q&A pairs and return them as a JSON
array.

REMEMBER: Every single question must require multi-hop reasoning. Do
not generate any single-hop questions.”

"User": ”Generate a multi-hop Q&A json file for the following text.
Please include questions of different types including factual (6
questions), single-step reasoning (7 questions), and True or False (7
questions): {combined_text}.”
```

Figure S15. **Updated prompt used in RetChemQA to generate multi-hop Q&A pairs.**

Prompt

```
"System instruction": " Question-Answer Pair Classification Prompt

You have been provided with a context from which question-answer pairs have been generated. Your task is to classify these
pairs according to the following criteria:

Classification Criteria:
1. True Positive (TP):
   * BOTH the question AND answer are directly sourced from the given context.
   * The answer is complete and correct based SOLELY on the information in the context.
   * No additional knowledge or inference beyond what is explicitly stated in the context is required.

2. False Positive (FP):
   * BOTH the question AND answer appear to be sourced from the given context.
   * However, the answer is incorrect, incomplete, or requires inference beyond the explicit information provided.

3. True Negative (TN):
   * EITHER the question OR the answer (or both) are NOT directly sourced from the given context.
   * The answer may be correct based on general knowledge, but cannot be fully validated using only the provided context.

4. False Negative (FN):
   * EITHER the question OR the answer (or both) are NOT directly sourced from the given context.
   * The answer is incorrect based on general knowledge or principles.

Important Guidelines:
1. Context Boundaries:
   * Exclude any references or citations from consideration as part of the context.

2. Direct Sourcing:
   * "Sourced from the context" means the information is explicitly stated, not inferred or derived.
   * If any inference or additional knowledge is required, the pair is not considered directly sourced.

3. Completeness:
   * For TP classification, ensure ALL information required for the answer is explicitly stated in the context.
   * Partial matches or answers requiring additional inference should be classified as FP, TN or FN.

4. Specificity:
   * Pay close attention to specific details, numbers, and phrasings in both questions and answers.
   * Minor discrepancies may change the classification.

5. General Knowledge:
   * Be cautious with answers that seem correct but rely on general knowledge not provided in the context.
   * These should typically be classified as TN, not TP.

6. Inference and Reasoning:
   * Questions requiring reasoning or inference beyond explicitly stated facts should not be classified as TP, even if the
reasoning seems sound.

7. Precision in Language:
   * Be wary of absolute terms like "always," "never," or "only" in questions or answers. Verify such claims are
explicitly supported by the context for TP classification.

8. Numeric Values:
   * For questions involving calculations or numeric values, ensure all required numbers and operations are explicitly
provided in the context for TP classification.

9. Chemical Formulas and Structures:
   * For questions about chemical formulas or structures, ensure the exact information is provided in the context. Do not
rely on chemical knowledge to infer details not explicitly stated.

10. Experimental Procedures:
    * For questions about experimental procedures or synthesis, all steps should be explicitly described in the context
for a TP classification.

11. Material Properties:
    * When classifying questions about material properties (e.g., surface area, gas uptake), ensure the specific values
and conditions are explicitly stated in the context.

12. Comparative Statements:
    * For questions comparing different materials or properties, ensure the context explicitly provides the comparison. Do
not rely on calculations or inferences not directly stated.

13. Hypothetical Scenarios:
    * Questions asking about hypothetical situations or changes to experimental conditions should be classified as TN
unless the context explicitly discusses such scenarios.

14. Mechanism and Theoretical Explanations:
    * Be cautious with answers that provide mechanisms or theoretical explanations. Ensure these are explicitly stated in
the context, not inferred based on chemical knowledge.

. . .
```

Figure S16. The best-performing prompt for the single-hop Q&A evaluation task. This corresponds to the prompt used in iteration 9 as shown in **Figure 3(b)**. The final evaluation of 252 DOIs using this prompt is shown in **Figure 4(b).** The remainder of the prompt is shown in the subsequent figure on the next page.

Prompt

```
15. Implicit Information:
    * Avoid classifying as TP any question-answer pairs that rely on information that seems obvious or implicit but is not
explicitly stated in the context.

16. Strict Interpretation:
    * When evaluating question-answer pairs, adopt a very strict interpretation of what constitutes "directly sourced"
information. If there's any doubt, lean towards classifying as FP or TN rather than TP.

17. Context Verification:
    * For each classification, explicitly reference the relevant part of the context that supports your decision. This
ensures a thorough check against the provided information.

18. Mathematical Operations:
    * For questions requiring simple mathematical operations (e.g., averaging, ratios), classify as TP only if ALL
required values are explicitly stated in the context AND the operation is trivial.
    * For more complex calculations or those requiring multiple steps, classify as TN or FP unless the context explicitly
provides the calculated result.

19. Partial Information:
    * If a question-answer pair contains some information from the context but also includes additional unsupported claims
or details, classify as FP rather than TP.

20. Time Sensitivity:
    * Be aware of the potential for time-sensitive information. If a question-answer pair relies on information that may
change over time (e.g., "current" record holders, latest discoveries), ensure the context explicitly supports the claim
for the relevant time period.

21. Structural Inferences:
    * For questions about molecular or crystal structures, ensure that all structural details are explicitly stated in the
context. Do not rely on chemical knowledge to infer structural information not directly provided.

22. Charge Balance:
    * When dealing with questions about ionic compounds or charge states, ensure that the context explicitly states the”
```

“User”: ” CONTEXT:{context}\n\nQ&A DATASET:{Q&A data set}”

Figure S16. Continuation of the best-performing prompt shown for the single-hop Q&A evaluation task.

Algorithm 1 Prompt Generation Algorithm

```
Input: List path doi_dirs containing document files, List path json_dirs containing data files, Dictionary ground truth ground_truths where [key]: json_filename, [value]: ground_truth_evaluation
Initialize prompts to empty list
Initialize num_qs to empty list
Initialize contexts to empty list
Initialize ground_truth to empty list
for each doi_dir, json_dir in directory doi_dirs, json_dirs do
    if doi path doi_dir or json path json_dir not exists then continue
    end if
    context ← ProcessDOI(doi_path)
    output_text ← "CONTEXT:" + context + "\n\nQ&A DATASET:"
    num_q ← 0
    for pair in LoadJSONQuestions(json) do
        output_text ← output_text + str(pair) + "\n\n"
        num_q ← num_q + 1
    end for
    Append output_text to prompts
    Append num_q to num_qs
    Append context to contexts
    Append ground_truths[json_filename] to ground_truth ▷ where json_filename is the last part of jsons' path
end for
return prompts, num_qs, contexts, ground_truth
function ProcessDOI(doi_dir)
    combined_text ← empty string
    for each file in doi_dir do
        text ← ProcessFile(file)
        combined_text ← combined_text + text
    end for
    return combined_text
end function
function ProcessFile(file_path)
    ext ← file extension from file_path
    if ext is '.pdf' then
        return ExtractTextFromPDF(file_path)
    else if ext is '.docx' or ext is '.doc' then
        return ProcessDOCX(file_path)
    else if ext is '.xml' then
        return ProcessXML(file_path)
    else if ext is '.xhtml' then
        return ProcessXHTML(file_path)
    end if
end function
function LoadJSONQuestions(file_path)
    Open file_path in read mode
    output ← JSON parsed from the file
    Initialize possible_keys ← ['qas', 'Q&A', 'QAs', 'questions', 'data', 'dataset']
    found_key ← None
    for key in possible_keys do
        if key exists in output then
            found_key ← key
            Break
        end if
    end for
    if found_key is None then
        return output
    else
        return output[found_key]
    end if
end function
```

Figure S17. Algorithm detailing the generation of the "user" portion of the prompt. This procedure integrates the entire textual context extracted from documents associated with each DOI, alongside all corresponding question-answer pairs or synthesis conditions obtained from associated JSON datasets. The resulting structured prompt is subsequently used as input for LLM evaluation.

Algorithm 2 LLMs' API Call Algorithm

```
Input: String system_instruction, List prompts prompts containing context and Q&A pairs data, List
Integer num_qs contains number of Q&A pairs per DOI, String directory, Integer runs=1
Initialize a thread pool executor
Submit the following tasks to the executor:
    Task 1: Call gpt_4o using LLM_API_Call with:
      system_instruction, prompts, num_qs, and directory/4o.xlsx, runs
    Task 2: Call gemini using LLM_API_Call with:
      system_instruction, prompts, num_qs, and directory/gemini.xlsx, runs
    Task 3: Call claude using LLM_API_Call with:
      system_instruction, prompts, num_qs, and directory/claude.xlsx, runs
    Task 4: Call gpt_o1 using LLM_API_Call with:
      system_instruction, prompts, num_qs, and directory/o1.xlsx, runs
    Task 5: Call deepseek using LLM_API_Call with:
      system_instruction, prompts, num_qs, and directory/deepseek.xlsx, runs
    Task 6: Call grok using LLM_API_Call with:
      system_instruction, prompts, num_qs, and directory/grok.xlsx, runs
Wait for all tasks in the executor to complete
Make summary df using SUMMARY_DF(directory, file_name, ground_truth, runs=1)
function LLM_API_CALL(system_instruction, prompts, num_qs, output_file_name, runs=1)
    for t from 0 to runs-1 do
        for i from 0 to len(prompts) do
            sheet_name ← f'{t} loop {i} DOI'
            prompt ← prompts[i]
            while True do
                Generate completion from LLM model with:
                system_instruction and prompt with Structure Output
                result_df ← Convert result into DataFrame using llm_output_to_df
                if len(result_df) = num_qs[i] and result_df.shape[1] = 5 then
                    break
                end if
            end while
            if output_file_name exists then
                Open Excel writer in append mode
                Write result_df to sheet sheet_name
            else
                Open Excel writer in write mode
                Write result_df to sheet sheet_name
            end if
        end for
    end for
end function
function LLM_OUTPUT_TO_DF(LLM_output)
    return DataFrame converted from LLM_output
end function
```

Figure S18. Algorithm detailing the procedure used for calling various LLM APIs to perform evaluations of generated question-answer pairs or synthesis conditions. The algorithm systematically submits structured prompts (including contexts and associated datasets) to each LLM, processes their responses, and compiles results into Excel sheet for subsequent analysis and comparison across models.

Algorithm 3 System Instruction (Classification Prompt) Optimization

```
Initialize trial ← 0
Initialize best_input ← None
Initialize previous_input ← None
while True do
    Increment trial by 1
    if last trial then
        break
    end if
    folder_name ← f'trial {trial}'
    Create directory folder_name if it does not exist
    if not first trial then
        Read classification_prompt from file of the directory in previous trial
    end if
    Make API Call from the desired LLMs using Algorithm 2 and make summary DataFrame, dfs, using Function 1
    Calculate cumulative statistics:
    average_total_catch ← Cumulative Average Non-TP Catching Rate
    average_accuracy ← Cumulative Average Accuracy
    Save metrics to text files in folder_name
    Generate mismatch evaluations:
    Initialize combined_mismatch_dict ← empty dictionary
    for key, df in dfs do
        Parse mismatches between evaluations and ground truth
        Append mismatch details to combined_mismatch_dict
    end for
    Format output string with context and evaluations:
    formatted_input ← String combining mismatch analysis and prompt refinement inputs
    Generate revised classification prompt using LLM:
    classification_prompt ← Call to claude_client with formatted_input
    Save revised classification prompt and formatted input to files in folder_name
    Update best_input and previous_input based on past trials
end while
```

Figure S19. Algorithm illustrating the automated optimization process for refining the "*system instruction*" (classification prompt) component shown in Figure 3(a). This algorithm employs iterative API calls specifically to Claude, leveraging evaluation results from previous trials to iteratively enhance prompt performance. The optimization involves systematically adjusting instructions based on cumulative accuracy metrics, mismatch analyses against ground truth data, and output from previous LLM evaluations.

Function 1 Summary DataFrame Function

```
function SUMMARY_DF(String directory, String directory, Dictionary ground truth ground_truths where [key]: json_filename, [value]: ground_truth_evaluation, Integer runs=1)
    for t from 0 to runs-1 do
        for i from 0 to len(ground_truth) do
            sheet_name ← f'{t} loop {i} DOI'
            ▷ Load required Excel sheets
            gpt_4o_df ← Load sheet sheet_name from directory/4o.xlsx
            gemini_df ← Load sheet sheet_name from directory/gemini.xlsx
            claude_df ← Load sheet sheet_name from directory/claude.xlsx
            gpt_o1_df ← Load sheet sheet_name from directory/o1.xlsx
            output_df ← MERGE_DF([gpt_4o_df, gemini_df, claude_df, gpt_o1_df])
            output_df ← WEIGHTED_MODE(output_df, [eval_columns], [0.23, 0.23, 0.23, 0.3])
            output_df['ground_truth'] ← ground_truth[i]
            file_path ← f'directory/file_name'
            if file_path exists then
                Open Excel writer in append mode
                Write output_df to sheet sheet_name
            else
                Open Excel writer in write mode
                Write output_df to sheet sheet_name
            end if
        end for
    end for
end function
function MERGE_DF(dfs)
    return Merged DataFrame to keep only the question, answer, question type, evaluation and explanation of each LLM
end function
function WEIGHTED_MODE(df, columns, [weights])
    return A list of weighted mode for each row in columns (calculated based on weights)
end function
```

Figure S20. The function that creates a summary excel sheet containing question-answer pairs with evaluations from multiple LLMs. The SUMMARY_DF function processes evaluation sheets from different LLMs, merges their assessments using the MERGE_DF helper function, and applies weighted voting through the WEIGHTED_MODE function. The resulting excel sheet includes columns for questions, answers, question types, individual LLM evaluations and their explanations, ground truth classifications, and weighted consensus evaluations.

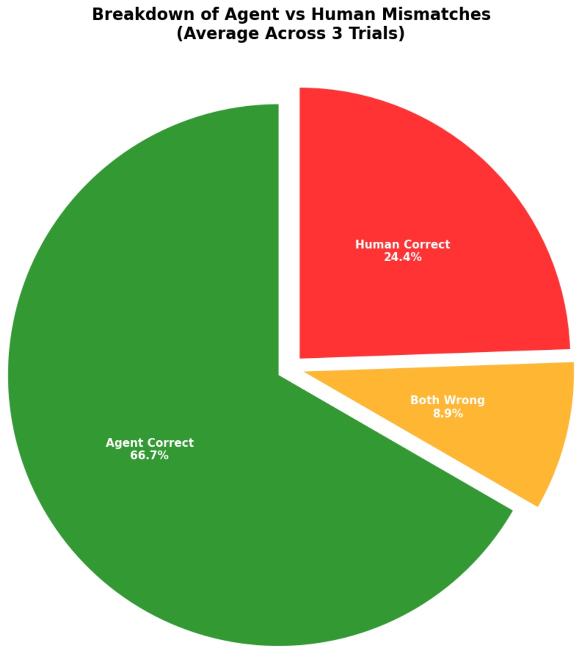


Figure S21. Breakdown of agent versus human evaluation mismatches averaged across three trials for multi-hop Q&A pairs. Total average mismatches: 45 out of 523 Q&A pairs.

Tie-breaker Model	Accuracy (%)	TP Catch Rate (%)	Non-TP Catch Rate (%)
GPT-4o	93.41	98.83	34.21
Gemini	93.61	99.47	30.26
Claude	94.16	99.06	38.30
GPT-o1	94.03	99.24	35.53

Table S1. Single-Hop Q&A Tie-Breaker Performance.

Tie-breaker Model	Accuracy (%)	TP Catch Rate (%)	Non-TP Catch Rate (%)
GPT-4o	98.31	99.41	38.10
Gemini	98.31	99.60	30.16
Claude	98.25	99.60	28.57
GPT-o1	98.38	99.47	38.10

Table S2. Multi-Hop Q&A Tie-Breaker Performance.

Tie-breaker Model	Accuracy (%)	TP Catch Rate (%)	Non-TP Catch Rate (%)
GPT-4o	95.86	99.12	36.16
Gemini	95.96	99.54	30.21
Claude	96.21	99.33	33.44
GPT-o1	96.21	99.36	36.82

Table S3. Average Tie-Breaker Performance across both single-hop and multi-hop for each model.

Iteration	Changes Made	Observations	Notes
1	Tested on one paper (nchem.834) with 11 TP and 9 TN Q&A pairs.	Final evaluation classified all questions as TPs entirely missing the non-TP Q&A pairs.	Only GPT-o1 was able to correctly classify some of the Q&A pairs as TNs.
2	Made prompt better by asking LLMs to be careful with vague answers and those containing superlatives like always/best.	Final evaluation saw an increase in TNs, however a small number of FPs and FNs also remained alongside TPs.	Gemini still only generated TPs, while the other models showed an improvement.
3	Modified the prompt to include an example instructing the LLMs to not rely on general domain knowledge.	The distribution of TP, FP, TN, and FN in the final evaluation remained largely *unchanged*.	LLMs continued to rely on general knowledge when classifying Q&A pairs.
4	Instructed the LLMs to avoid labeling speculative questions as FP and emphasized contextual grounding.		The LLMs continued to classify those Q&A pairs as FPs.
4*	Used the same prompt as Iteration 4 but introduced a secondary 'checker' prompt to reassess and validate the outputs.		The secondary 'checker' provided no benefit.
5	Used Claude 3.5 Sonnet to optimize the prompt using a template addressing frequent misclassifications with corrections. (Figure 3a)	Final evaluation saw a big improvement reaching close to our target human evaluated benchmark.	Claude is excellent at prompt optimization.
6 & 7	Introduced stricter context constraints, clearer classification rules, and a refined template.	Final evaluation saw a marginal improvement in performance.	The more detailed the prompt is, the better the LLMs perform.

Table S4. Summary of Iterative Prompt Refinement and Evaluation Results.

Dataset	TP	FP	TN	FN
ALL	4861	141	20	121
OPT	141	7	10	0

Table S5. Distribution of Q&A classification types in the optimization (OPT, 7 DOIs) and final 252 DOIs (ALL) test sets.

Summary of Classification labels

Each question–answer (Q&A) pair is first checked to determine whether it was generated from the context provided in the prompt:

True Positive (TP): The question is based on the given context, and the answer is correct.

False Positive (FP): The question is based on the given context, but the answer is incorrect or incomplete.

True Negative (TN): The question is not based on the given context, and the model correctly identifies this (e.g., states that the answer cannot be found in the context).

False Negative (FN): The question is not based on the given context, and the model provides an incorrect answer.

Model versions and parameters

The API versions are: claude_model = 'claude-3-5-sonnet-20240620', gemini_model = 'gemini-1.5-pro-001', openai_4o_model = "gpt-4o-2024-08-06", openai_o1_model = "o1-preview-2024-09-12"

The temperature is set to 1. Please note that the random seed values for these hosted APIs (OpenAI, Anthropic, and Google) are not user accessible. Model sampling is handled internally, and reproducibility is tested by fixing the API version, temperature, and prompt.