SI: ChemLit-QA: A human evaluated dataset for chemistry RAG tasks

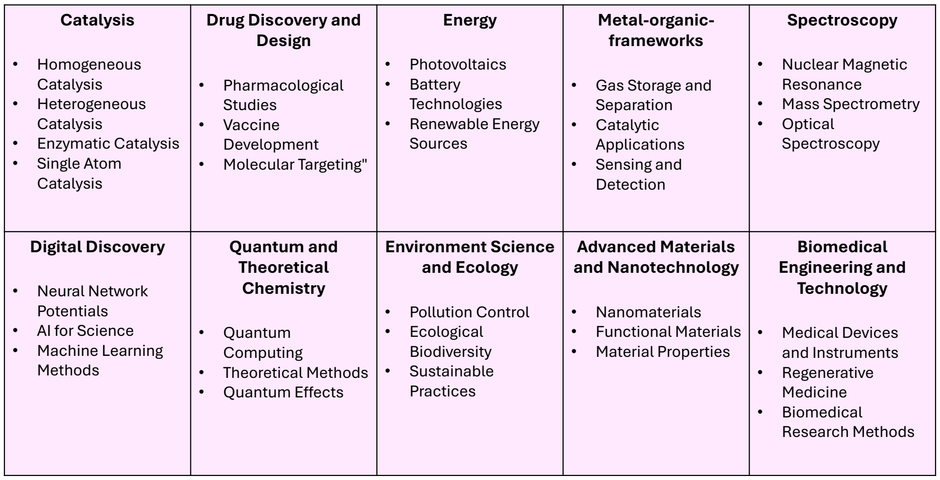
Anonymous Author(s)
Affiliation
Address
email

Abstract

Retrieval-Augmented Generation (RAG) is a widely used strategy in Large-Language Models (LLMs) to extrapolate beyond the inherent pre-trained knowledge. Hence, RAG is crucial when working in data-sparse fields such as Chemistry. The evaluation of RAG systems is commonly conducted using specialized datasets. However, existing datasets, typically in the form of scientific Question-Answer-Context (QAC) triplets or QA pairs, are often limited in size due to the labor-intensive nature of manual curation or require further quality assessment when generated through automated processes. This highlights a critical need for large, high-quality datasets tailored to scientific applications. We introduce ChemLit-QA, a comprehensive, expert-validated, open-source dataset comprising over 1,000 entries specifically designed for chemistry. Our approach involves the initial generation and filtering of a QAC dataset using an automated framework based on GPT-4 Turbo, followed by rigorous evaluation by chemistry experts. Additionally, we provide two supplementary datasets: ChemLit-QA-neg focused on negative data, and ChemLit-QA-multi focused on multihop reasoning tasks for LLMs, further enhancing the resources available for advanced scientific research.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024) - AI4MAT workshop

Clusters of papers from ChemRxiv corpus



Catalysis • Homogeneous Catalysis • Heterogeneous Catalysis • Enzymatic Catalysis • Single Atom Catalysis	**Drug Discovery and Design** • Pharmacological Studies • Vaccine Development • Molecular Targeting"	**Energy** • Photovoltaics • Battery Technologies • Renewable Energy Sources	**Metal-organic-frameworks** • Gas Storage and Separation • Catalytic Applications • Sensing and Detection	**Spectroscopy** • Nuclear Magnetic Resonance • Mass Spectrometry • Optical Spectroscopy
Digital Discovery • Neural Network Potentials • AI for Science • Machine Learning Methods	**Quantum and Theoretical Chemistry** • Quantum Computing • Theoretical Methods • Quantum Effects	**Environment Science and Ecology** • Pollution Control • Ecological Biodiversity • Sustainable Practices	**Advanced Materials and Nanotechnology** • Nanomaterials • Functional Materials • Material Properties	**Biomedical Engineering and Technology** • Medical Devices and Instruments • Regenerative Medicine • Biomedical Research Methods

Fig. S 1: Hierarchy of topics and subtopics used to cluster ChemRxiv corpus. We used each paper's title and abstract with Mistral to classify level 1 (shown in bold face) and level 2 labels.

Reasoning distribution in ChemLit-QA

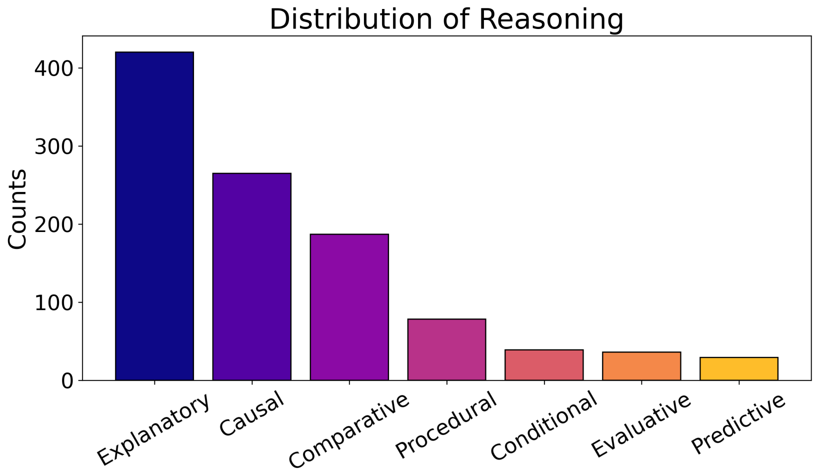


Fig. S 2: Distribution of reasoning in ChemLit-QA.

Distribution of clusters in ChemLlit-QA

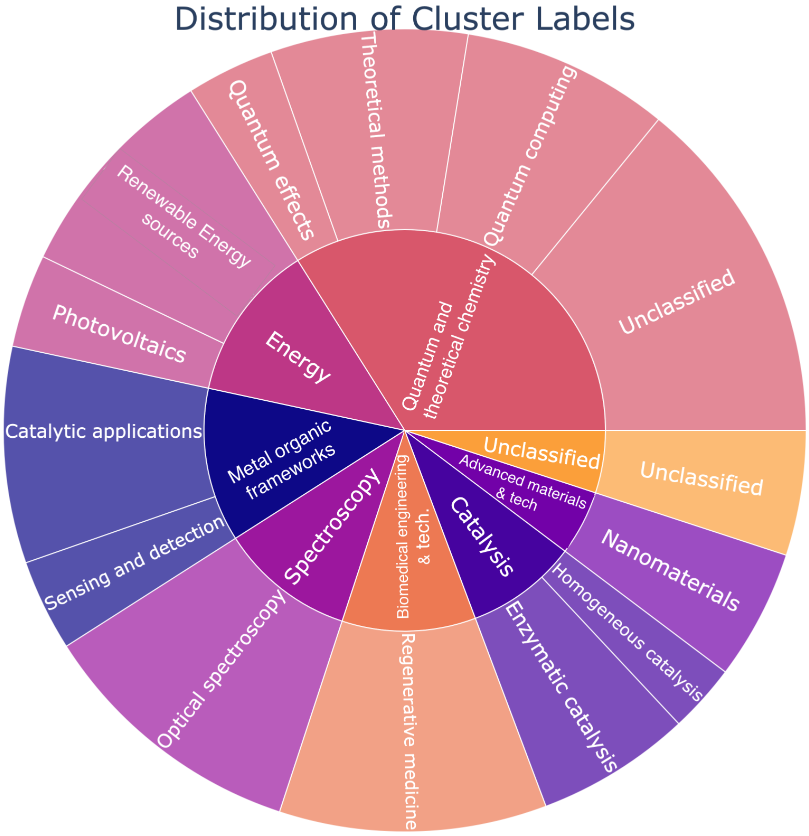


Fig. S 3: Distribution on cluster labels in the ChemLit-QA dataset.

Keyword distribution in ChemLit-QA

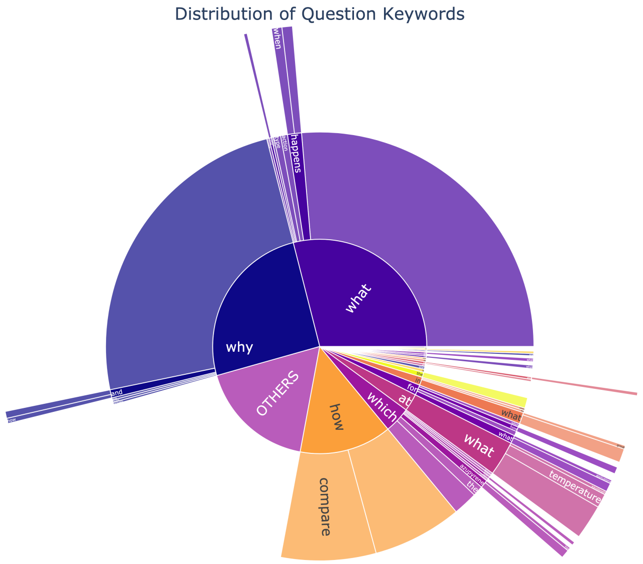


Fig. S 4: Distribution of question keywords in ChemLit-QA.

Expert agreement results

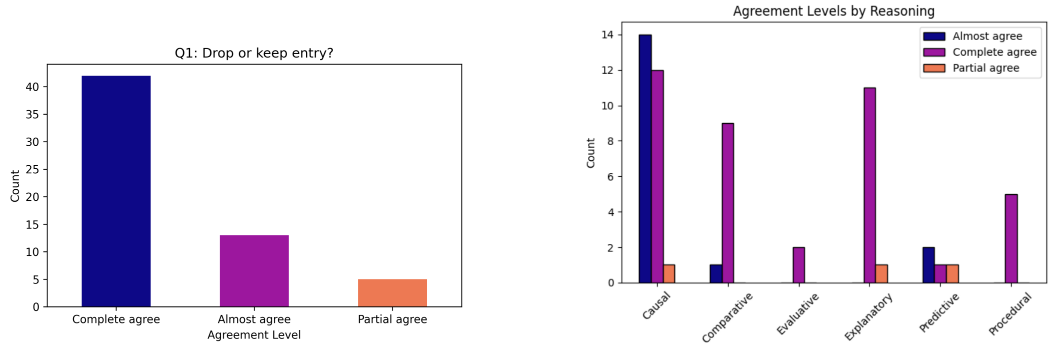


Fig. S 5: Agreement between humans on a) keeping or dropping the dataset entry b) reasoning type

Tab. S 1: Agreement among experts

Task	Complete Agree	Almost Agree	Partial Agree	Disagree
Question quality: Keep or drop	70%	22%	8%	0%
Reasoning type	68%	27%	5%	0%
Difficulty level	44%	40%	8%	8%

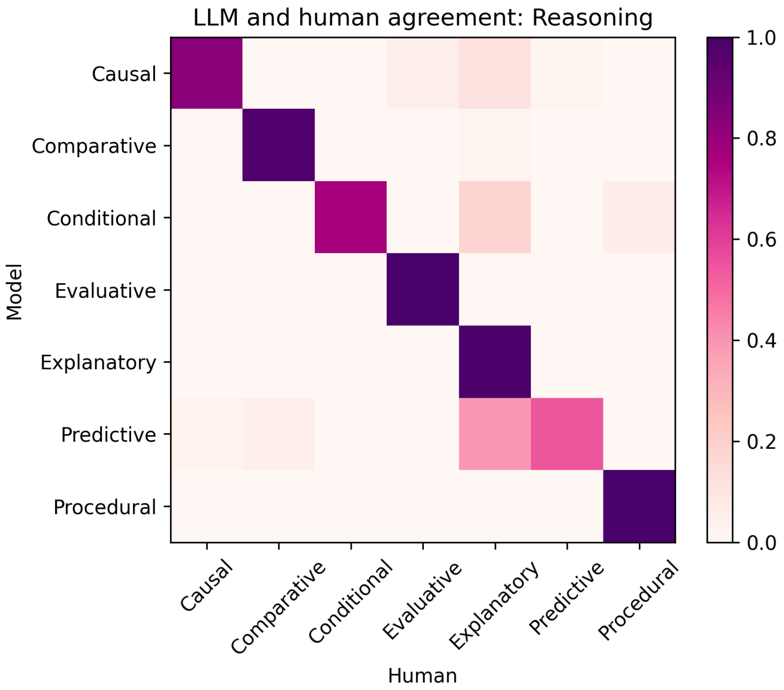


Fig. S 6: Agreement between LLM and experts on answer reasoning type.

22 Analysis statistics of the ChemLit-QA dataset

Tab. S 2: Statistical distribution of metrics. All of the given LLM-based metrics were implemented using DeepEval[1] framework and GPT-4o[2].

Metric	**Mean ± std dev.**
Answer Relevancy Score (GPT-4o)	0.99 ± 0.02
Faithfulness Score (GPT-4o)	0.99 ± 0.01
Hallucination Score (GPT-4o)	0.0 ± 0.0
Question Faithfulness Score (GPT-4o)	0.93 ± 0.10
Penalized semantic entropy (GPT-4o)	0.20 ± 0.44

Case study: Performance of RAG models in ChemLit-QA dataset

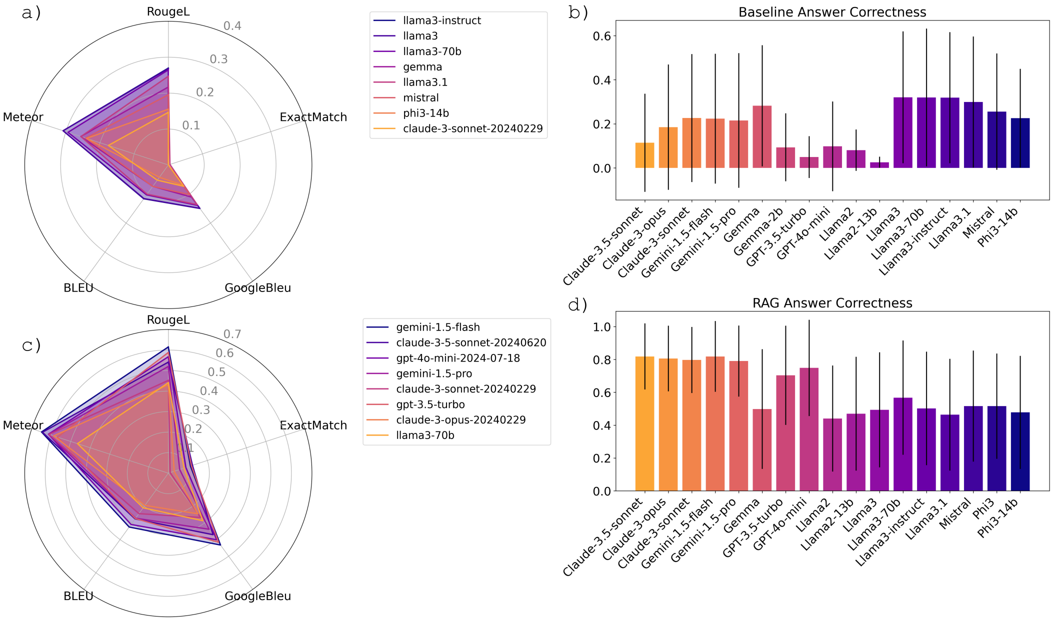


Fig. S 7: (a) The top 8 LLMs' text-based performance on baseline QA. (b) The answer correctness of all tested LLMs on baseline QA. (c) The top 8 LLMs' text-based performance on RAG. (d) The answer correctness of all tested LLMs on RAG.

Case study: Finetuned model performance on ChemLitQA-multi

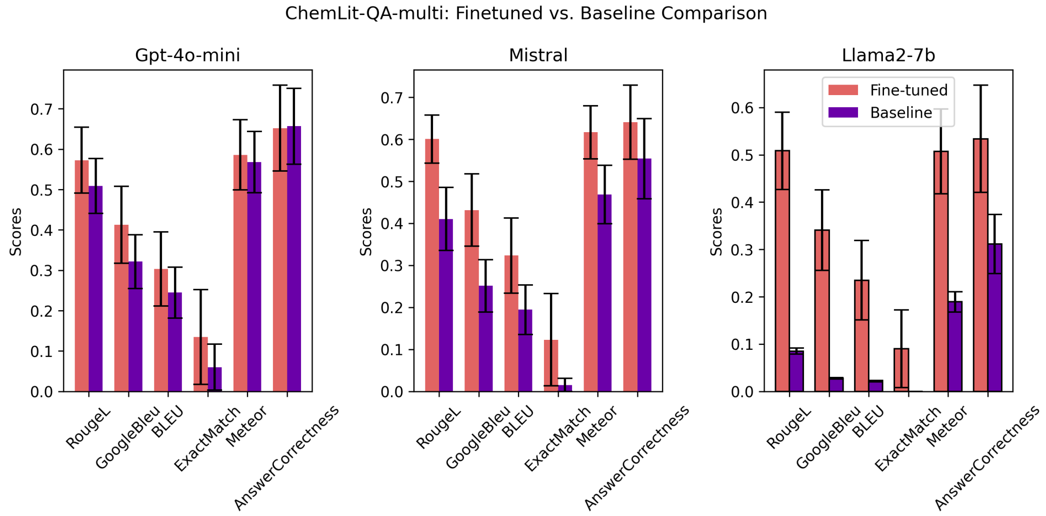
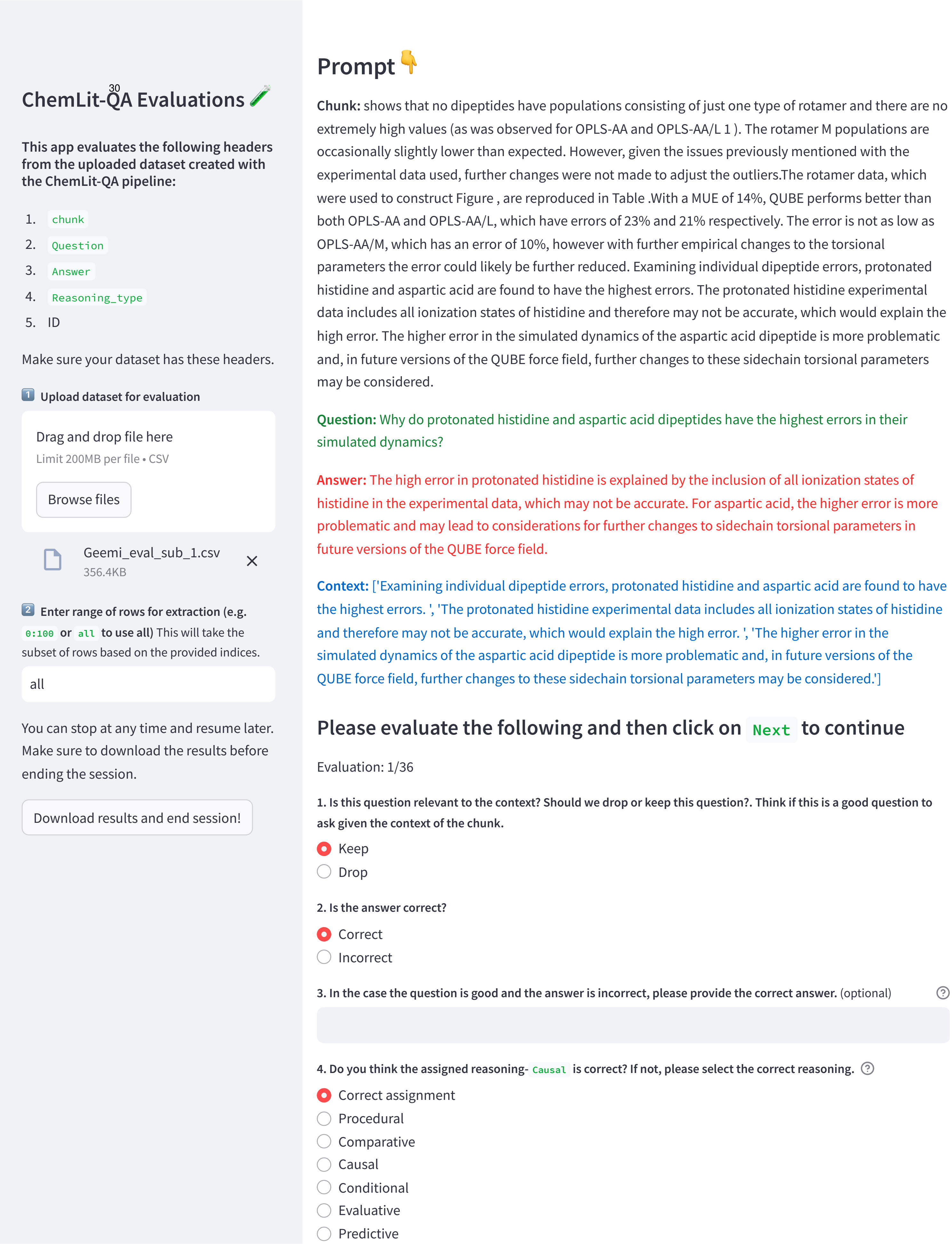


Fig. S 8: Comparison between baseline and fine-tuned performance on the test dataset for GPT-4o-mini, Mistral-7B, and Llama2-7B.

Human evaluations interface

The following figure illustrates the interface used in this work to conduct that human evaluations. This app was developed using Streamlit(`https://discuss.streamlit.io/`). The left hand panel allows the users to upload the dataset under review and select the number of entries to review expert evaluations are collected in the right hand panel.



ChemLit-QA Evaluations 🧪

This app evaluates the following headers from the uploaded dataset created with the ChemLit-QA pipeline:

1. `chunk`
2. `Question`
3. `Answer`
4. `Reasoning_type`
5. ID

Make sure your dataset has these headers.

1️⃣ **Upload dataset for evaluation**

Drag and drop file here
Limit 200MB per file • CSV

Geemi_eval_sub_1.csv
356.4KB

2️⃣ **Enter range of rows for extraction (e.g. `0:100` or `all` to use all)** This will take the subset of rows based on the provided indices.

all

You can stop at any time and resume later. Make sure to download the results before ending the session.

Prompt 👇

Chunk: shows that no dipeptides have populations consisting of just one type of rotamer and there are no extremely high values (as was observed for OPLS-AA and OPLS-AA/L 1). The rotamer M populations are occasionally slightly lower than expected. However, given the issues previously mentioned with the experimental data used, further changes were not made to adjust the outliers.The rotamer data, which were used to construct Figure , are reproduced in Table .With a MUE of 14%, QUBE performs better than both OPLS-AA and OPLS-AA/L, which have errors of 23% and 21% respectively. The error is not as low as OPLS-AA/M, which has an error of 10%, however with further empirical changes to the torsional parameters the error could likely be further reduced. Examining individual dipeptide errors, protonated histidine and aspartic acid are found to have the highest errors. The protonated histidine experimental data includes all ionization states of histidine and therefore may not be accurate, which would explain the high error. The higher error in the simulated dynamics of the aspartic acid dipeptide is more problematic and, in future versions of the QUBE force field, further changes to these sidechain torsional parameters may be considered.

Question: Why do protonated histidine and aspartic acid dipeptides have the highest errors in their simulated dynamics?

Answer: The high error in protonated histidine is explained by the inclusion of all ionization states of histidine in the experimental data, which may not be accurate. For aspartic acid, the higher error is more problematic and may lead to considerations for further changes to sidechain torsional parameters in future versions of the QUBE force field.

Context: ['Examining individual dipeptide errors, protonated histidine and aspartic acid are found to have the highest errors. ', 'The protonated histidine experimental data includes all ionization states of histidine and therefore may not be accurate, which would explain the high error. ', 'The higher error in the simulated dynamics of the aspartic acid dipeptide is more problematic and, in future versions of the QUBE force field, further changes to these sidechain torsional parameters may be considered.']

Please evaluate the following and then click on `Next` to continue

Evaluation: 1/36

1. Is this question relevant to the context? Should we drop or keep this question?. Think if this is a good question to ask given the context of the chunk.

● Keep
○ Drop

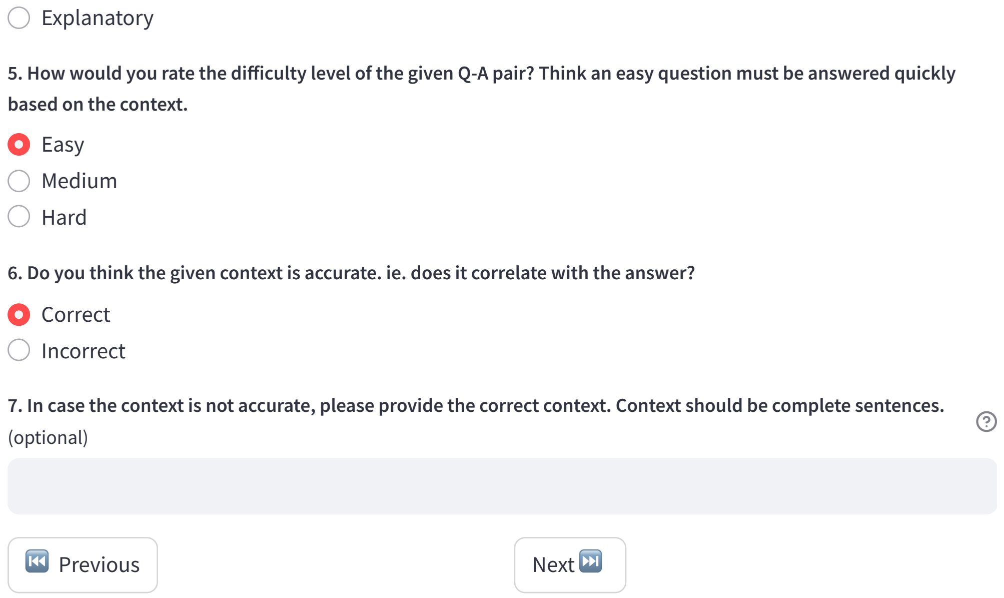
2. Is the answer correct?

● Correct
○ Incorrect

3. In the case the question is good and the answer is incorrect, please provide the correct answer. (optional) ⓘ

4. Do you think the assigned reasoning- `Causal` is correct? If not, please select the correct reasoning. ⓘ

● Correct assignment
○ Procedural
○ Comparative
○ Causal
○ Conditional
○ Evaluative
○ Predictive



○ Explanatory

5. How would you rate the difficulty level of the given Q-A pair? Think an easy question must be answered quickly based on the context.

● Easy

○ Medium

○ Hard

6. Do you think the given context is accurate. ie. does it correlate with the answer?

● Correct

○ Incorrect

7. In case the context is not accurate, please provide the correct context. Context should be complete sentences. (optional)

All prompts used in the work

The following figure shows all prompts used during the generation process. Tab. S 3 3 explains the function of each prompt.

Tab. S 3: The function of each prompt in the generation process.

Name	Function
CLEAN_PROMPT	The prompt used for classifying the usefulness of the text chunks
EXAMPLES_USEFUL	Example of a useful text chunk, integrated into CLEAN_PROMPT
EXAMPLES_USELESS	Example of a useless text chunk, integrated into CLEAN_PROMPT
REASONING_PROMPT	The prompt for identifying all possible reasoning types from cleaned text chunks
PROCEDURAL_PROMPT	The prompt for constructing a procedural question
COMPARATIVE_PROMPT	The prompt for constructing a comparative question
CAUSAL_PROMPT	The prompt for constructing a causal question
CONDITIONAL_PROMPT	The prompt for constructing a conditional question
EVALUATIVE_PROMPT	The prompt for constructing a evaluative question
PREDICTIVE_PROMPT	The prompt for constructing a predictive question
EXPLANATORY_PROMPT	The prompt for constructing a explanatory question
DIFFICULTY_PROMPT	The prompt for assign difficulty to a question given its corresponding answer and the original text chunk

Prompts used during dataset curation

```
CLEAN_PROMPT = """Given the following chunk of text from an academic paper,
please classify if the text is useful or not. Output 'Yes' for useful chunks and
'No' for useless chunks.\n
The following are some general traits of useful and useless chunks,
along with some examples. \n

Useful chunks usually: \n
1. Mainly contain coherent English sentences. \n
2. Include one of the following: in-depth discussion scientific entities, coherent
experiment procedures, meaningful comparison, intensive reasoning.

Useless chunks usually: \n
1. Are too short (only one or two sentences). \n
2. Contain non-relevant information to the main text such as title,
author information, figure captions, references, declarations, etc. \n
3. Contain simple introduction to concepts without futher discussions. \n
4. Contain ill-formatted formulae or tables that not readable by humans. \n
5. Simply recorded the authors' experimental procedures without explicit order. \n

Examples of useful chunks: \n
{example_useful}

Examples of useless chunks: \n
{example_useless}


Text to classify: {chunk}

usefulness: Yes or No

Format instructions: \n{format_instructions}
"""


EXAMPLES_USEFUL = """
'd was accurate according to our criteria
(0.8 < K d / K d,inp < 1.25) for r 1 / r 2  2.5 but not for r 1 / r 2  5.
At r 1 / r 2 = 0.25  we obtained a binding isotherm with anomalous shape (Figures S2)
and K d / K d,inp = 1.27. This anomaly was due to a numerical artifact from meshing in
COMSOL; by using a more refined mesh we obtained K d / K d,inp < 1.02.
The improvement in accuracy by mesh refinement may suggest that the large deviations
in K d at r 1 / r 2  5 are also due to too coarse meshes as well.  Thus, more refined
and optimized meshes (in particular, for boundary regions  between small and
large areas) could improve K d determination in a virtual ACTIS experiment.
```

```
We confirmed this for the extreme  value of r 1 / r 2 = 50  and found an  optimal
K d / K d,inp = 1.00 at the expense of excessively increasing  the computational
time ( 72 h instead of  3 h) and the potential risk of overfitting (SI).
In order to keep studies consistent, comparable and in a reasonable time', \n
"""

EXAMPLES_USELESS = """
' ASSOCIATED CONTENT Supporting InformationThe Supporting Information is available
free of charge on the ACS Publications website and on ChemRxiv
(DOI:10.26434/chemrxiv.12345644). Theoretical background for computer simulation
and data evaluation; Simulation of separagrams; Figure , Variation in k off,
inp-separagrams and binding isotherms; Figure , Variation in injection loop
dimensions -separagrams and binding isotherms; Figure , Variation in injection
loop dimensions -sample-plug distribution; Figure , Variation in separation capillary
radii -separagrams and binding isotherms; Figure , Velocity streamlines at different
separation capillary radii; Figure , Variation in the initial', \n
"""


REASONING_PROMPT = """
Please identify all the suitable types of questions to generate
given a piece of text. Your available options are: ['Procedural', 'Comparative',
'Causal', 'Conditional', 'Evaluative', 'Predictive', 'Explanatory']. Please choose
solely from the options. The options are defined as follows:\n

A Procedural question asks about the order between steps in a clearly formulated
procedure. These procedures are often indicated by words such as 'first', 'then',
'finally', followd by actions. \n
A Comparative question asks about the relation between mutual properties of comparable
entities, Common mutual properties include numbers, years, etc. \n
A Causal question asks about the reasons for a specific phenomenon. The phenomenon
can be given implicitly or by explicit clauses such as 'for example'. \n
A Conditional question asks about the possible outcomes given a scenario.
Scenarios are often given by conditional clauses such as 'if', 'when', etc.\n
An Evaluative question asks about the benefits and drawbacks of a given entity.\n
A Predictive question asks for reasonable inference, often on the properties of
entites closely related to but not mentioned in the text. \n
An Explanatory question asks for a component from a statement made in the text.  \n

Text: {text}

Structure your output in the following format:
Process: <Record here in detail how you go though each step of the instruction.>
Reasoning_types: <The reasoning types you chose>
Format instructions: \n{format_instructions}
"""
```

```
PROCEDURAL_PROMPT = """
Please follow the instruction below to formulate a Procedural question based on
the given text. A Procedural question asks about the order between steps in a
\clearly formulated procedure. These procedures are often indicated by words such as
\'first', 'then', 'finally', followd by actions.
You should go through the entire text and form questions only base on complete
\sentences. \n

1. Identify the procedure mentioned in the text. If no processes are mentioned,
skip the following steps and output 'NaN' for <question>, <answer> and <context>.
2. List all steps in the process mentioned by the question in the exact same order
as provided. \n
3. Choose one step (step1) from the process.\n
4. Determine its position in the process. i.e. where is it ranked in the process,
the first, the second, or other?\n
5. Raise a question in the format: What is the <position> step in <summary of the
process>? \n
6. Optionally, choose another step (step2) from the process. Determine the relative
position of step1 to step2.\n
7. Raise a question in the following format: What is the <ordinal, relative position>
step before/after <step2> in <summary of the process>? Replace the original question
with the new one. \n
8. Record the question, answer, and context in the output. <question> should be the
question you raised. <answer> should be step1, rephrased to be grammatically correct
when necessary. <context> should be the original text containing the full process only.


Text: {text}

Structure your output in the following format:
Process: <Record here in detail how you go though each step of the instruction.>
Question: <question>
Answer: <answer>
Context: <context>

Format instructions: \n{format_instructions}
"""

COMPARATIVE_PROMPT = """
Please follow the instruction below to formulate a Comparative question based on the
given text.
A Comparative question asks about the relation between mutual properties of comparable
entities, Common mutual properties include numbers, years, etc.
You should go through the entire text and form questions only base on complete
sentences.\n

1. Identify the comparable entities in the text, the comparable properties,
e.g. numbers, years, etc, and their relation from the text. If there are no comparable
```

```
properties or no relations are mentioned, skip the following steps and output 'NaN'
for <question>, <answer> and <context>.
2. Identify the entities associated with the comparable values. \n
3. Randomly choose at least two entities and raise a question which asks about the
relation between the comparable values of these entities. You shoule not disclose
information on the relation in the question.\n
4. Record the question, answer, and context in the output. <question> should be the
question you raised. <answer> should be the relation you are asking for, including the
result of comparison (e.g. bigger, smaller, similar, etc). Rephrase the answer to be
grammatically correct. <context> should be all sentences in the original text excerpts
describing the entities and their comparable values only. \n


Text: {text}

Structure your output in the following format:
Process: <Record here in detail how you go though each step of the instruction.>
Question: <question>
Answer: <answer>
Context: <context>

Format instructions: \n{format_instructions}
"""

CAUSAL_PROMPT = """
Please follow the instruction below to formulate a Causal question based on the given
text. A Causal question asks about the reasons for a specific phenomenon.
The phenomenon can be given implicitly or by explicit clauses such as 'for example'.
You should go through the entire text and form questions only base on complete
sentences.\n

1. Identify the reasoning and scenario in the text. If no examples are mentioned,
skip the following steps and output 'NaN' for <question>, <answer> and <context>.
2. Rephrase the scenario into a question. Do not add or delete any information. \n
3. Record the question, answer, and context in the output. <question> should be the
question you raised. <answer> should be an explanation of the scenario based on the
reasoning, rephrased to be grammatically correct when necessary.
<context> should be all sentences in the original text containing the claims only.


Text: {text}

Structure your output in the following format:
Process: <Record here in detail how you go though each step of the instruction.>
Question: <question>
Answer: <answer>
Context: <context>
```

```
Format instructions: \n{format_instructions}
"""

CONDITIONAL_PROMPT = """
Please follow the instruction below to formulate a Conditional question based on the
given text.
A Conditional question asks about the possible outcomes given a scenario. Scenarios are
often given by conditional clauses such as 'if', 'when', etc.
You should go through the entire text and form questions only base on complete
sentences.\n

1. Identify the text containing conditions, e.g. clauses with 'if'.  If no conditions
are mentioned, skip the following steps and output 'NaN' for <question>, <answer> and
<context>.
2. Identify the possible scenarios and the corresponding actions. \n
3. Formulate a question which asks for the action given one of the scenarios.
You can choose scenarios not mentioned in the text. \n
4. Record the question, answer, and context in the output. <question> should be the
question you raised. <answer> should be the corresponding action, rephrased to be
grammatically correct when necessary. <context> should be all sentences in the original
text containing the statements only. \n


Text: {text}

Structure your output in the following format:
Process: <Record here in detail how you go though each step of the instruction.>
Question: <question>
Answer: <answer>
Context: <context>

Format instructions: \n{format_instructions}
"""


EVALUATIVE_PROMPT = """
Please follow the instruction below to formulate an Evaluative question based on the
given text.
An Evaluative question asks about the benefits and drawbacks of a given entity. \n
You should go through the entire text and form questions only base on complete
sentences.\n

1. List all statements made in the text. Find if any statements explain the properties
of a specific entity and imply value judgements. Define these statement as 'necessary
statements'. If no statements satisfy the requirements, skip the following steps and
output 'NaN' for <question>, <answer> and <context>.
2. Reformulate the 'necessary statements' in the format: <entity>: <properties> \n
3. Classify the properties as positive or negative. \n
```

```
4. Raise a question based on the format: What are the pros and cons / benefits /
drawbacks of <entity>? Paraphrase the question. \n
5. Record the question, answer, and context in the output. <question> should be the
question you raised. <answer> should contain all <properties> associatd with the
authors' attitude, rephrased to be grammatically correct when necessary. <context>
should be all sentences in the original text containing 'necessary statements'only.


Text: {text}

Structure your output in the following format:
Process: <Record here in detail how you go though each step of the instruction.>
Question: <question>
Answer: <answer>
Context: <context>

Format instructions: \n{format_instructions}
"""

PREDICTIVE_PROMPT = """
Please follow the instruction below to formulate a Predictive question based on the
given text.
A Predictive question asks for reasonable inference, often on the properties of
entites closely related to but not mentioned in the text.
You should go through the entire text and form questions only base on complete
sentences.\n

1. List all statements made in the text. Find if any statements explain the properties
of a specific entity. Define these statement as 'necessary statements'.
If no statements satisfy the requirements, skip the following steps and output
'NaN' for <question>, <answer> and <context>.
2. Randomly choose from the following one category of transformations with equal
probability: \n
    a. Negation \n
    b. Generalization/specification \n
    c. Analogy \n
3. Apply to the most suitable entity-property pair. The transformed entity and
property must both make sense scientifically. \n
4. Raise a question which asks for the property of the transformed entity. Do not
disclose any information about the transformed property in the question. \n
5. Record the question, answer, and context in the output. <question> should be the
question you raised. <answer> should contain <transformed properties> and be
rephrased to be grammatically correct when necessary. <context> should be all s
entences in the original text containing the 'necessary statements' only. \n


Text: {text}
```

```
Structure your output in the following format:
Process: <Record here in detail how you go though each step of the instruction.>
Question: <question>
Answer: <answer>
Context: <context>

Format instructions: \n{format_instructions}
"""


EXPLANATORY_PROMPT = """
Please follow the instruction below to formulate an Explanatory question based on the
given text.
An Explanatory question asks for a component from a statement made in the text.  \n
You should go through the entire text and form questions only base on complete
sentences.\n

1. List all statements made in the text.
2. Choose a statement and replace part of it with an appropriate interrogative pronoun.
The part you replace should be specific. You should not mention the replaced information
in the question. \n
3. Rephrase the question to be grammatically correct. \n
4. Record the question, answer, and context in the output. <question> should be the
question you raised. <answer> should be the part you replaced, rephrased to be
grammatically correct when necessary. <context> should be all sentences in the original
text containing the chosen statement only. \n


Text: {text}

Structure your output in the following format:
Process: <Record here in detail how you go though each step of the instruction.>
Question: <question>
Answer: <answer>
Context: <context>

Format instructions: \n{format_instructions}
"""

DIFFICULTY_PROMPT = """
You are given a text chunk and a question-answer pair derived from the chunk. Please
assign one of the labels from 'Easy', 'Medium' and 'Hard' to <difficulty>, where the
easiest question is one whose answer is directly available in a single sentence in the
chunk, and the hardest question is one which requires information from multiple sentences
in the chunk and complex reasoning to arrive at the answer.

Question: {question}
Answer: {answer}
```

```
Chunk: {chunk}

Structure your output in the following format:
Difficulty: <difficulty>

Format instructions: \n{format_instructions}
"""
```

References

[1] Jeffrey Ip et al. Deepeval: The open-source llm evaluation framework. *Confident AI*, 2024.

[2] Open AI. Hello gpt-4o, 2024.