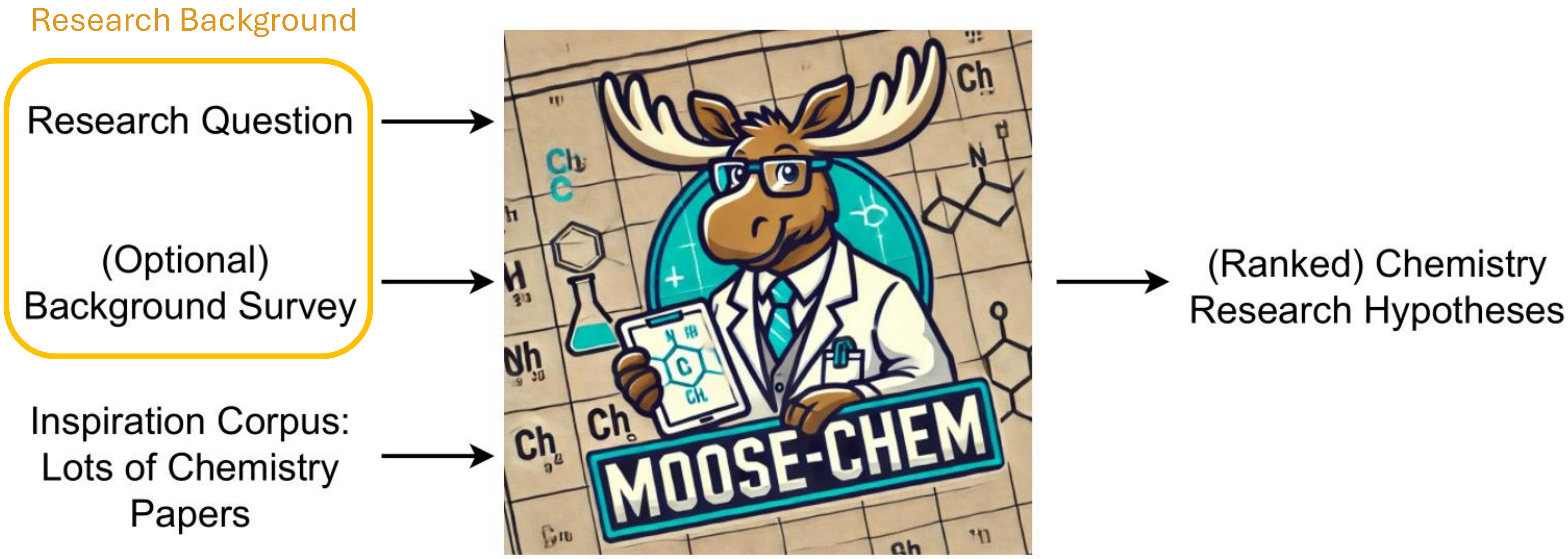
MOOSE-Chem: Large Language Models for Rediscovering Unseen Chemistry Scientific Hypotheses

Zonglin Yang

Central question

- Research background
  - a research question and/or a background survey
- Can LLMs automatically discover novel and valid chemistry research hypotheses?
  - Given only a chemistry research background
  - The research question can be on any chemistry domain
- Denoting
  - background: b
  - hypothesis: h
  - the central question: $\mathrm{P}(\mathrm{h}|b)$.
- $\mathrm{P}(\mathrm{h}|b)$ seems impossible to solve

Input / Output of MOOSE-Chem



Research Background

Research Question

(Optional) Background Survey

Inspiration Corpus: Lots of Chemistry Papers

(Ranked) Chemistry Research Hypotheses

Fundamental assumption and decomposition

- hypothesis: h
- background: b
- Inspiration: i
  - Existing knowledge not known to be related with b
- Example: Backpropagation
  - h: backpropagation
  - b: multi-layer logistic regression
  - i: chain-rule in Calculus
- Extensive discussions with chemistry experts
- In chemistry, a majority of published h originate from
  - one b and several i

An Example Publication (<Nature>, 2024)

- Research Background
  - Research Question
    - How to obtain D_2 gas more efficiently?
  - Background Survey
    - The best performing methods are electrocatalytic methods.
- Inspiration 1
  - Ruthenium as catalyst
- Inspiration 2
  - Nitrogen-doped electrode
- Inspiration 3
  - D_2O as chemical solution
- Hypothesis
  - A nitrogen-doped ruthenium (Ru) electrode can effectively catalyze the reductive deuteration of (hetero)arenes in the presence of D_2O in an electrocatalytic method, leading to efficient D_2 gas production.

Fundamental assumption and decomposition

- In chemistry, a majority of published h originate from
  - one b and several i
- $h = f(b, i_1, \ldots, i_k)$, where $k \leq 3, and\ i\ \ in\ \ I$ Fundamental Assumption
- Challenge: $\mathrm{P}(\mathrm{h}|b)$ seems impossible to solve
- Goal: Transform $\mathrm{P}(\mathrm{h}|b)$ into an equivalent form, where
  - each step is practical and executable
- We made it in the end:

$$P(h|b) \approx \prod_{j=1}^{k} P(i_j|b, h_{j-1}, I) \cdot P(h_j|b, i_j, h_{j-1}), \text{ where } h_0 = \emptyset$$

- But how?

Fundamental assumption and decomposition

- $h = f(b, i_1, \ldots, i_k)$, where $k \leq 3, and\ i\ \ in\ \ I, where\ P(I) = 1$
- A first try: Applying the Chain Rule,

$$
\begin{aligned}
P(h|b) &= P(h, i_1, \ldots, i_k|b) \\
&= \begin{cases} \frac{P(h,b,i_1)}{P(b,i_1)} \cdot \frac{P(b,i_1)\cdot P(I)}{P(b)\cdot P(I)} & \text{if } k = 1 \\ \frac{P(h,b,i_1,\ldots,i_k)}{P(b,i_1,\ldots,i_k)} \cdot \frac{P(b,i_1,\ldots,i_k)\cdot P(I)}{P(b,i_1,\ldots,i_{k-1})\cdot P(I)} \cdot \ldots \cdot \frac{P(b,i_1)\cdot P(I)}{P(b)\cdot P(I)} & \text{if } k > 1 \end{cases} \\
&= \begin{cases} P(h|b, i_1) \cdot P(i_1|b, I) & \text{if } k = 1 \\ \boxed{P(h|b, i_1, \ldots, i_k)} \cdot \prod_{j=2}^{k} \boxed{P(i_j|b, i_1, \ldots, i_{j-1}, I)} \cdot P(i_1|b, I) & \text{if } k > 1 \end{cases}
\end{aligned}
$$

- However,
  - Still not practical
  - Not reflect how chemistry researchers do

Step by Step!

Not Practical: $\mathrm{P}(h|b, i_1, \ldots, i_k)$ and $P(i_{j+1}|b, i_1, \ldots, i_j, I)$

- $h = f(b, i_1, \ldots, i_k)$, where $k \leq 3, and\ i\ \ in\ \ I$
- Recursive step by step approximation (when $k \geq 1$)

$$P(h_k|b, i_1, \ldots, i_k) \approx P(h_k|b, f(b, i_1, \ldots, i_{k-1}), i_k)$$

- Similarly,

$$P(i_{k+1}|b, i_1, \ldots, i_k, I) \approx P(i_{k+1}|b, f(b, i_1, \ldots, i_k), I)$$

- Markov property
  - h as a state
  - i as an action

Final Proof: Decomposition of $\mathrm{P}(h|b)$

$$P(h|b) = P(i_1, \ldots, i_k, h_1, \ldots, h_k|b)$$

Fundamental Assumption & Multi-step

$$= P(i_1, h_1|b) \cdot P(i_2, h_2|b, i_1, h_1) \cdot \ldots \cdot P(i_k, h_k|b, i_1, \ldots, i_{k-1}, h_1, \ldots, h_{k-1})$$

Chain Rule

$$\approx P(i_1, h_1|b) \cdot P(i_2, h_2|b, h_1) \cdot \ldots \cdot P(i_k, h_k|b, h_{k-1})$$

Markov Property

$$= \prod_{j=1}^{k} P(i_j|b, h_{j-1}, I) \cdot P(h_j|b, i_j, h_{j-1}), \text{ where } h_0 = \emptyset$$

Chain Rule & $P(I) = 1$

Fundamental assumption and decomposition

- The seemingly impossible question $\mathrm{P}(\mathrm{h}|b)$
  - → many smaller, more practical and executable questions.

$$P(h|b) \approx \prod_{j=1}^{k} \boxed{P(i_j|b, h_{j-1}, I)} \cdot \boxed{P(h_j|b, i_j, h_{j-1})}, \quad \text{where } h_0 = \emptyset$$

- However, it is still not helpful enough
  - Lots of h can be generated with LLMs
  - Wet-lab experiments to verify each h take lots of time and cost
- Ranking

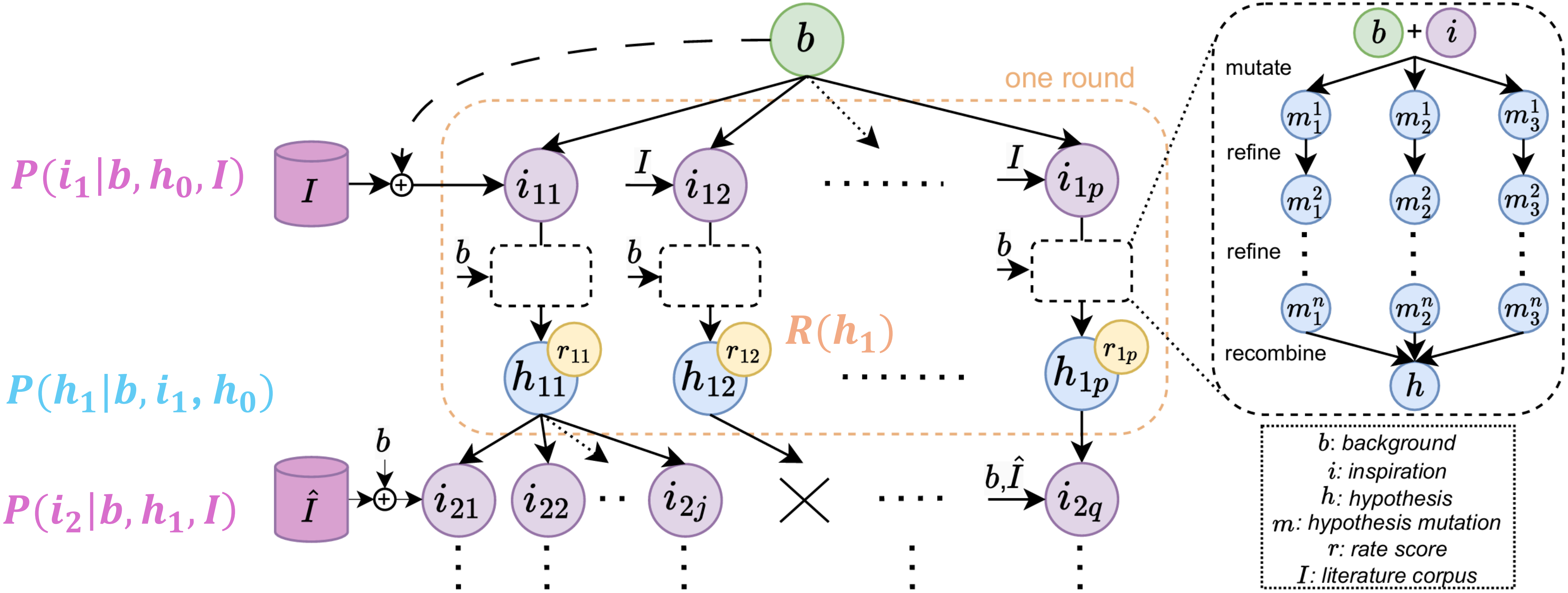
$$P(H_{\text{ranked}}) = P(H, R), \quad \text{where } H_{\text{ranked}} = \{h_1, h_2, \ldots, h_n \mid \boxed{R(h_i)} \geq R(h_{i+1}) \text{ for all } i\}$$

$$P\left(i_j \middle| b, h_{j-1}, I\right), P(h_j | b, i_j, h_{j-1}), R(h)$$

- How can we develop an LLM-based framework based on them?
- How well do LLMs perform on each of them?
- With this decomposition, how well can LLMs perform P$(h|b)$?

$$P(h|b) \approx \prod_{j=1}^{k} \boxed{P(i_j|b, h_{j-1}, I)} \cdot \boxed{P(h_j|b, i_j, h_{j-1})}, \text{ where } h_0 = \emptyset$$

$$P(H_{\text{ranked}}) = P(H, R), \text{ where } H_{\text{ranked}} = \{h_1, h_2, \ldots, h_n \mid \boxed{R(h_i)} \geq R(h_{i+1}) \text{ for all } i\}$$



b: background
i: inspiration
h: hypothesis
m: hypothesis mutation
r: rate score
I: literature corpus

$$P(i_j|b, h_{j-1}, I), P(h_j|b, i_j, h_{j-1}), R(h)$$

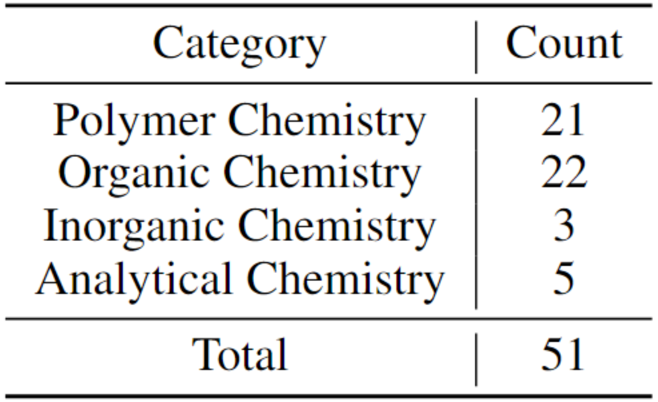
- How can we develop an LLM-based framework based on them?
- How well do LLMs perform on them?
- With this decomposition, how well can LLMs perform P$(h|b)$?

It needs groundtruth annotation of b, i, and h.

The Benchmark

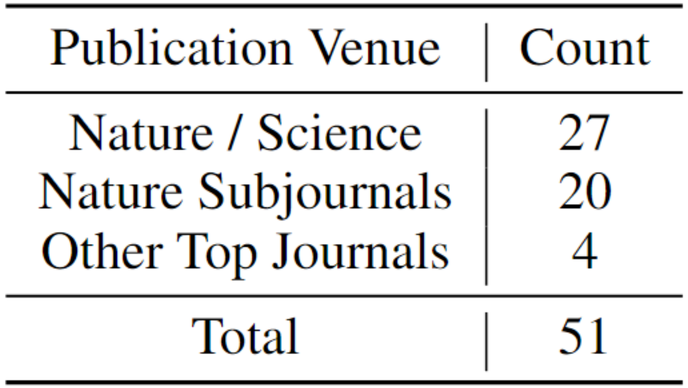
- 51 chemistry papers
  - Published on Nature, Science, or a similar level
  - Published in 2024, and only be available online in 2024
  - Collected by several chemistry PhD students
- Break each paper into:
  - Background question
  - Background survey
  - Inspirations
    - Existing Papers as Inspirations
  - Hypothesis

The Benchmark



Category	Count
Polymer Chemistry	21
Organic Chemistry	22
Inorganic Chemistry	3
Analytical Chemistry	5
Total	51

Table 1: Distribution of categories.

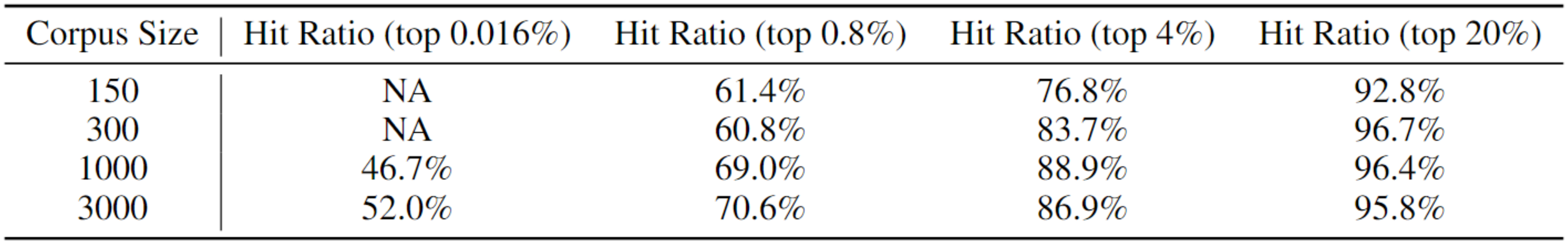


Publication Venue	Count
Nature / Science	27
Nature Subjournals	20
Other Top Journals	4
Total	51

Table 2: Distribution of publication venues.

Q1 : How well can LLMs perform $P(i_j|b, h_{j-1}, I)$?

- Whether LLM can identify inspiration papers which *are unknown* to be able to associate with the background (or at least unknown to associate in a certain way) but in fact can associate with the background to create novel knowledge?

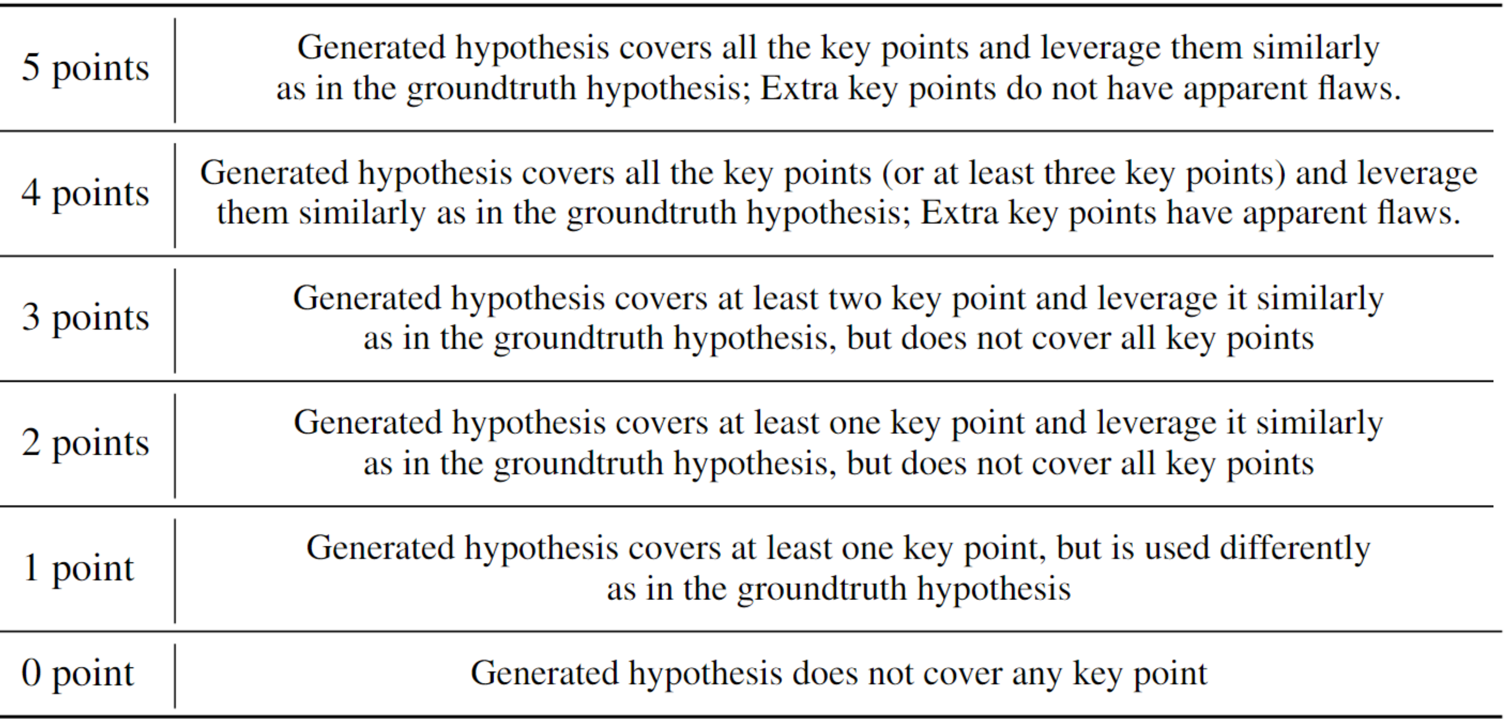


Corpus Size	Hit Ratio (top 0.016%)	Hit Ratio (top 0.8%)	Hit Ratio (top 4%)	Hit Ratio (top 20%)
150	NA	61.4%	76.8%	92.8%
300	NA	60.8%	83.7%	96.7%
1000	46.7%	69.0%	88.9%	96.4%
3000	52.0%	70.6%	86.9%	95.8%

Table 3: Main table for $Q1$. For each screen window of 15 papers, 3 papers are selected.

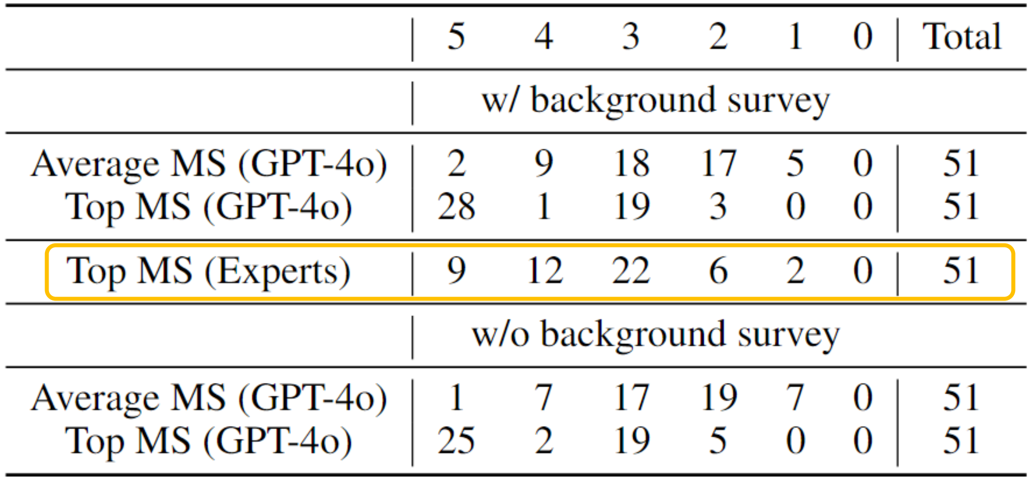
Q2: How well can LLMs perform $P(h_j|b, i_j, h_{j-1})$?

- Given only known knowledge, whether LLM can reason to unknown knowledge that has high probability to be valid?



5 points	Generated hypothesis covers all the key points and leverage them similarly as in the groundtruth hypothesis; Extra key points do not have apparent flaws.
4 points	Generated hypothesis covers all the key points (or at least three key points) and leverage them similarly as in the groundtruth hypothesis; Extra key points have apparent flaws.
3 points	Generated hypothesis covers at least two key point and leverage it similarly as in the groundtruth hypothesis, but does not cover all key points
2 points	Generated hypothesis covers at least one key point and leverage it similarly as in the groundtruth hypothesis, but does not cover all key points
1 point	Generated hypothesis covers at least one key point, but is used differently as in the groundtruth hypothesis
0 point	Generated hypothesis does not cover any key point

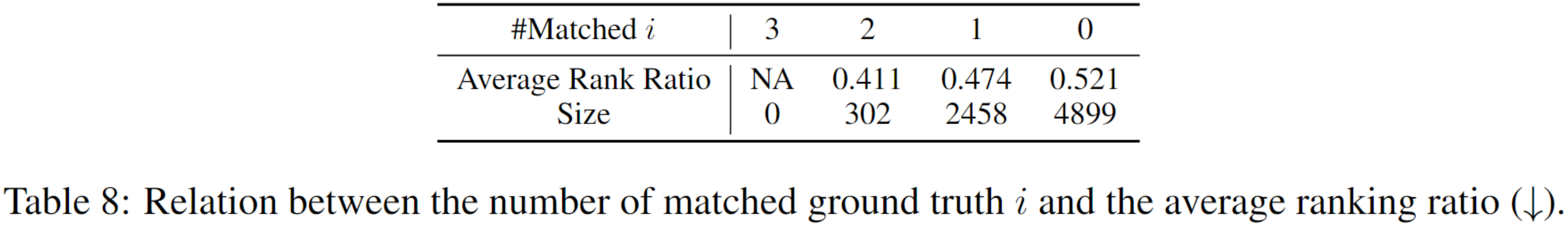
Table 6: Description of the Matched Score.



	5	4	3	2	1	0	Total
	w/ background survey						
Average MS (GPT-4o)	2	9	18	17	5	0	51
Top MS (GPT-4o)	28	1	19	3	0	0	51
Top MS (Experts)	9	12	22	6	2	0	51
	w/o background survey						
Average MS (GPT-4o)	1	7	17	19	7	0	51
Top MS (GPT-4o)	25	2	19	5	0	0	51

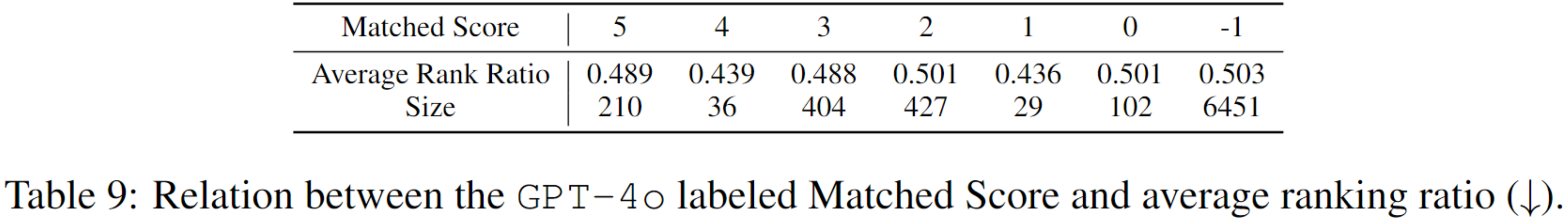
Q3: How well can LLMs perform $R(h)$?

- whether LLMs can select high-quality h to rank them higher?



#Matched i	3	2	1	0
Average Rank Ratio	NA	0.411	0.474	0.521
Size	0	302	2458	4899

Table 8: Relation between the number of matched ground truth i and the average ranking ratio (↓).



Matched Score	5	4	3	2	1	0	-1
Average Rank Ratio	0.489	0.439	0.488	0.501	0.436	0.501	0.503
Size	210	36	404	427	29	102	6451

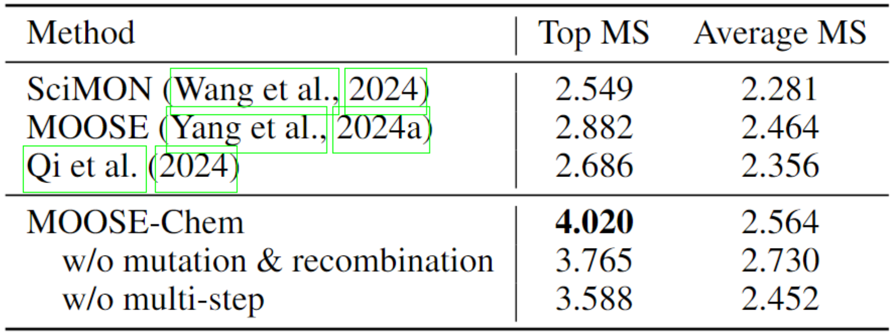
Table 9: Relation between the `GPT-4o` labeled Matched Score and average ranking ratio (↓).

$$P\left(i_j \middle| b, h_{j-1}, I\right), P(h_j | b, i_j, h_{j-1}), R(h)$$

- How can we develop an LLM-based framework based on them?
- How well do LLMs perform on them?
- With this decomposition, how well can LLMs perform P$(h|b)$?

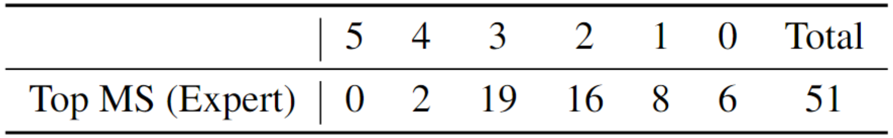
Runs MOOSE-Chem in real copilot setting

- Given only b and I ($|I| = 300$)
  - I contains the ground truth i (<=3)



Method	Top MS	Average MS
SciMON (Wang et al., 2024)	2.549	2.281
MOOSE (Yang et al., 2024a)	2.882	2.464
Qi et al. (2024)	2.686	2.356
MOOSE-Chem	**4.020**	2.564
w/o mutation & recombination	3.765	2.730
w/o multi-step	3.588	2.452

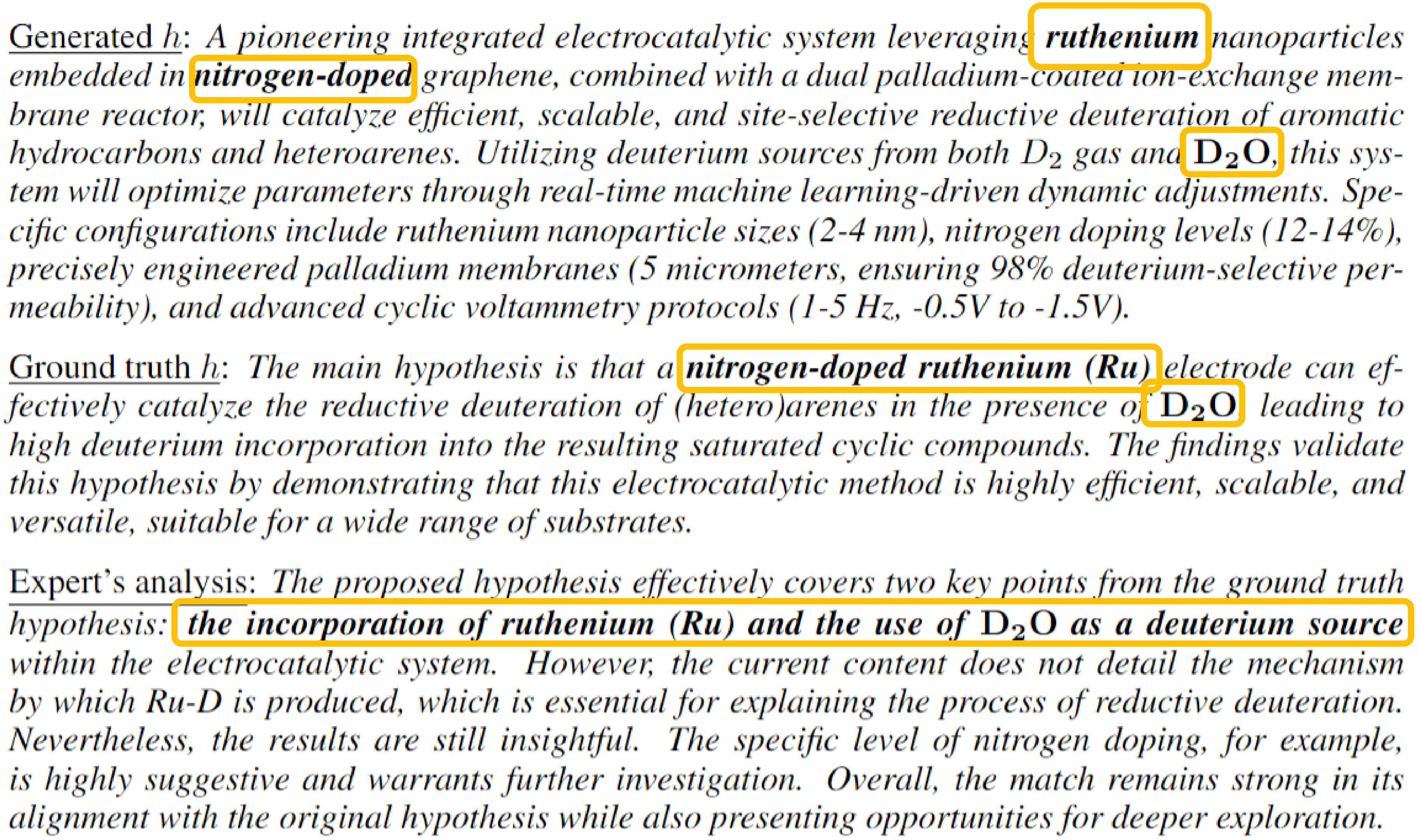
Table 10: Experiments and ablation study. The Matched Score is evaluated by GPT-4o.



	5	4	3	2	1	0	Total
Top MS (Expert)	0	2	19	16	8	6	51

Table 11: MOOSE-Chem runs with $|I|$=300, mimicking the copilot setting. This table shows the statistics of the top Matched Score across the benchmark. The evaluation is done by experts.

Case study



Generated h: *A pioneering integrated electrocatalytic system leveraging* ***ruthenium*** *nanoparticles embedded in* ***nitrogen-doped*** *graphene, combined with a dual palladium-coated ion-exchange membrane reactor, will catalyze efficient, scalable, and site-selective reductive deuteration of aromatic hydrocarbons and heteroarenes. Utilizing deuterium sources from both D_2 gas and* $\mathbf{D_2O}$*, this system will optimize parameters through real-time machine learning-driven dynamic adjustments. Specific configurations include ruthenium nanoparticle sizes (2-4 nm), nitrogen doping levels (12-14%), precisely engineered palladium membranes (5 micrometers, ensuring 98% deuterium-selective permeability), and advanced cyclic voltammetry protocols (1-5 Hz, -0.5V to -1.5V).*

Ground truth h: *The main hypothesis is that a* ***nitrogen-doped ruthenium (Ru)*** *electrode can effectively catalyze the reductive deuteration of (hetero)arenes in the presence of* $\mathbf{D_2O}$ *leading to high deuterium incorporation into the resulting saturated cyclic compounds. The findings validate this hypothesis by demonstrating that this electrocatalytic method is highly efficient, scalable, and versatile, suitable for a wide range of substrates.*

Expert's analysis: *The proposed hypothesis effectively covers two key points from the ground truth hypothesis:* ***the incorporation of ruthenium (Ru) and the use of*** $\mathbf{D_2O}$ ***as a deuterium source*** *within the electrocatalytic system. However, the current content does not detail the mechanism by which Ru-D is produced, which is essential for explaining the process of reductive deuteration. Nevertheless, the results are still insightful. The specific level of nitrogen doping, for example, is highly suggestive and warrants further investigation. Overall, the match remains strong in its alignment with the original hypothesis while also presenting opportunities for deeper exploration.*

Thanks!