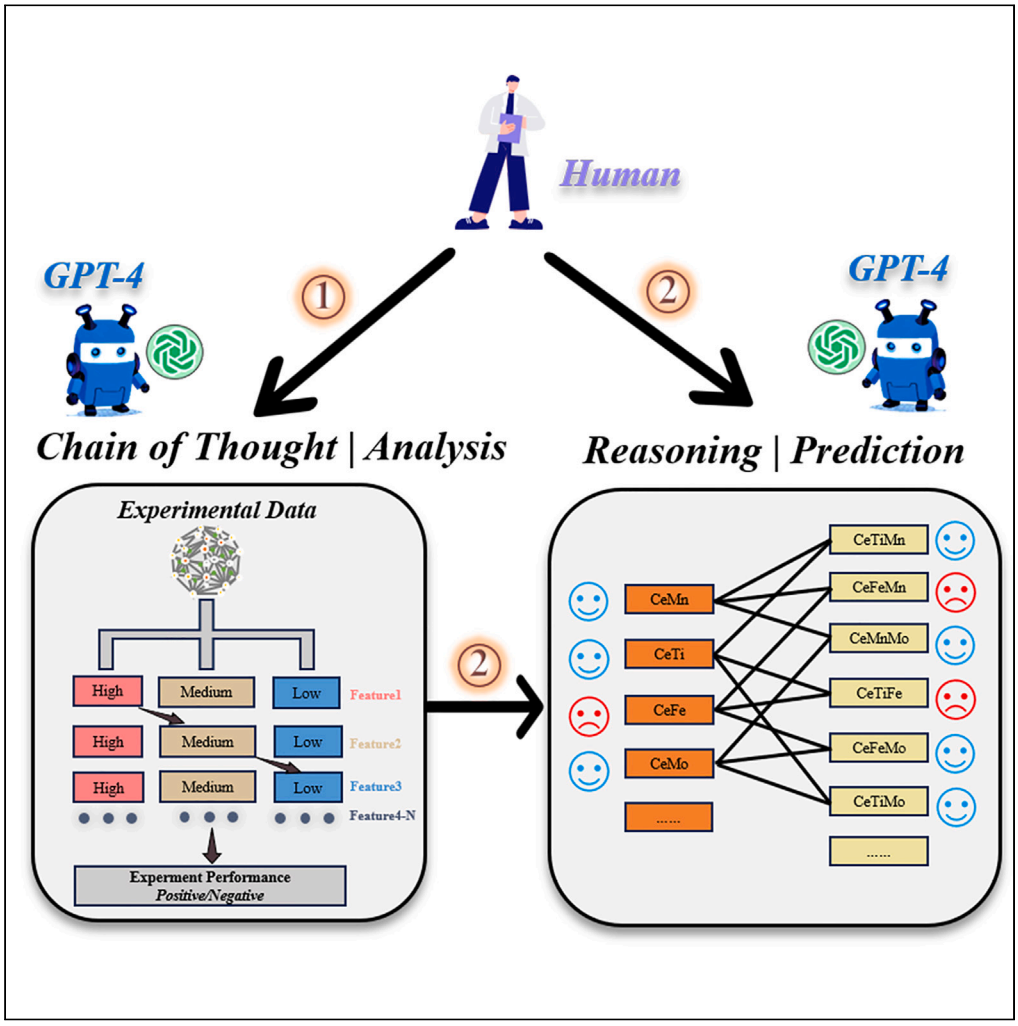
iScience

CellPress
OPEN ACCESS

Article

Analysis and prediction in SCR experiments using GPT-4 with an effective chain-of-thought prompting strategy



Muyu Lu, Fengyu Gao, Xiaolong Tang, Linjiang Chen

txiaolong@126.com (X.T.)
l.j.chen@bham.ac.uk (L.C.)

Highlights

Application of a large language model (LLM) in chemistry tasks

A new chain-of-thought prompting strategy focusing on formulating logical steps

An LLM-powered assistant that interprets, predicts, and rationalizes experimental data

Lu et al., iScience 27, 109451
April 19, 2024
https://doi.org/10.1016/j.isci.2024.109451

iScience

CellPress
OPEN ACCESS

Article

Analysis and prediction in SCR experiments using GPT-4 with an effective chain-of-thought prompting strategy

Muyu Lu,[1] Fengyu Gao,[1] Xiaolong Tang,[1,*] and Linjiang Chen[2,3,4,*]

SUMMARY

This study explores the use of large language models (LLMs) in interpreting and predicting experimental outcomes based on given experimental variables, leveraging the human-like reasoning and inference capabilities of LLMs, using selective catalytic reduction of NO_x with NH_3 as a case study. We implement the chain of thought (CoT) concept to formulate logical steps for uncovering connections within the data, introducing an "Ordered-and-Structured" CoT (OSCoT) prompting strategy. We compare the OS-CoT strategy with the more conventional "One-Pot" CoT (OPCoT) approach and with human experts. We demonstrate that GPT-4, equipped with this new OSCoT prompting strategy, outperforms the other two settings and accurately predicts experimental outcomes and provides intuitive reasoning for its predictions.

INTRODUCTION

The emergence of the latest large language models (LLMs), notably GPT-3 and GPT-4,[1,2] is transforming the landscape of human-computer interaction, revolutionizing a wide range of personal and professional tasks through advanced artificial intelligence (AI) capabilities.[3,4] These LLMs, trained on vast amounts of text data, are capable of generating human-like text, answering common-sense questions, and even performing tasks that require understanding and reasoning.[5,6] LLMs can provide textual content creation and offer personalized interactions and recommendations.[7] Their proficiency extends to tasks that require inferential reasoning, such as answering questions, solving mathematical problems,[8] and even passing bar examinations.[9] Moreover, the impact of LLMs is evident in the realm of academic research. LLMs, like GPT-4, have the potential to streamline the literature review process by efficiently summarizing vast academic resources, extracting insights from the literature, and facilitating the generation of innovative research ideas.[10] By enabling the analysis of extensive textual data, these models may uncover overlooked themes or patterns, offering fresh perspectives on existing research.[11]

LLMs have started showing the potential to revolutionize chemistry by accelerating research and discovery in collaboration with human chemists. For instance, GPT-4 has been used in the discovery of new metal–organic frameworks (MOFs) through a cooperative workflow with human experts. This synergy enabled the discovery of a series of MOFs, each synthesized using unique strategies and conditions.[12] In the broader landscape, LLM-empowered AI tools and agents are making strides in organic synthesis, drug discovery, and materials design. ChemCrow,[13] a GPT-4-based chemistry tool, exemplifies this. It streamlines reasoning for various chemical tasks, from drug design to synthesis. By sequentially prompting GPT-4, ChemCrow guides the model through the task, aligning actions with the end goal. This tool not only aids expert chemists but also simplifies access to chemical knowledge for novices. Moreover, task-specific fine-tuning of GPT-3 has been shown to yield highly effective and predictive models for a range of chemistry machine-learning (ML) tasks, often surpassing the performance of dedicated ML models specifically developed for these tasks.[14]

In this study, we hypothesize that GPT-4's language understanding capabilities, when combined with its strengths in pattern recognition and inferential reasoning, might enable effective analysis and interpretation of knowledge specific to a research topic or scientific domain. Our focus is on structured data rather than texts directly presented in scientific publications; in essence, we are not evaluating GPT-4's text mining capabilities. Instead, we aim to assess GPT-4's proficiency in recognizing patterns within structured data, which allows it to discern and capitalize on underlying trends. Such pattern recognition is invaluable when analyzing experimental variables and their associated outcomes. Furthermore, we seek to determine if GPT-4's ability for inferential reasoning enables it to make well-founded predictions based on the provided information and to assess the robustness of its rationale behind those predictions.

[1]School of Energy and Environmental Engineering, University of Science and Technology Beijing, Beijing 100083, P.R. China
[2]School of Chemistry and School of Computer Science, University of Birmingham, Birmingham B15 2TT, UK
[3]Key Laboratory of Precision and Intelligent Chemistry, University of Science and Technology of China, Hefei, Anhui 230026, China
[4]Lead contact
*Correspondence: txiaolong@126.com (X.T.), l.j.chen@bham.ac.uk (L.C.)
https://doi.org/10.1016/j.isci.2024.109451

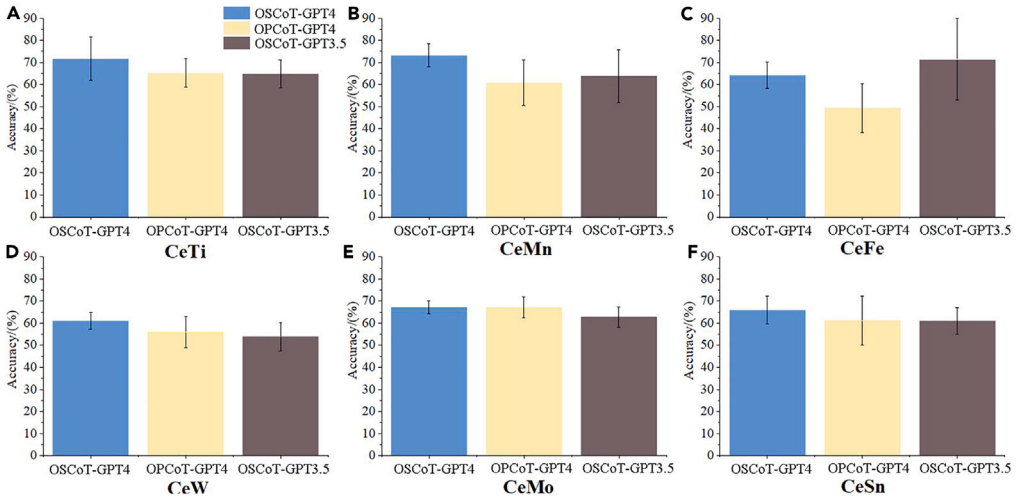


Figure 1. Comparison of the three CoT-GPT combinations in predicting experimental NO_x conversion outcomes for the six binary CeM_1 composites samples

For all composite types (A–F), five independent runs were conducted, each using one of the five batches of 48 samples for training. Prediction accuracies were evaluated using 50 samples randomly selected from the dataset, which were distinct from the training samples. The average, maximum, and minimum prediction accuracy values for each composite type were determined across the five runs.

RESULTS

Binary CeM_1 metal-oxide composites

To prepare representative training data, rational selection by custom search (Table S2) from the dataset was conducted five runs, each generating a batch of 48 samples that covered six types of binary CeM_1 metal-oxide catalysts (M_1 = Ti, Mn, W, Sn, Mo, or Fe). The selected samples were then integrated into **UM1** and **UM2**, respectively, to generate OP-and OS-CoT, using both GPT-3.5 and GPT-4. We evaluated the effectiveness of **UM1** and **UM2** in generating CoTs from GPT-3.5 and GPT-4 against three common metrics for assessing the performance of LLMs: '*disobedience*', '*helpfulness*', and '*honesty*'. Detailed analyses are shown in Figures S4 and S5. Notably, we observed that GPT-3.5 malfunctioned when using **UM2** in trying to generate OPCoT, leading to its exclusion from comparison. We examined the optimal number of input tokens for GPT-3.5-turbo-16k, as illustrated in Figure S6, revealing that it began to malfunction when burdened with more than 3,000 tokens in the OP approach. Finally, three combinations—OPCoT-GPT4, OSCoT-GPT4, and OSCoT-GPT3.5—were incorporated into **UM3** to infer the experimental performance of CeM_1 metal-oxide composites samples in five runs.

OP-and OS-CoTs were used to guide GPT-3.5 and GPT-4 to infer the experimental performance of each of the CeM_1 samples. As depicted in Figures 1A–1F, the average prediction accuracies of OSCoT-GPT4 for the six different binary composites reached 71.6%, 74%, 64.2%, 60%, 67.6%, and 65.6%, respectively. The maximum prediction accuracies for them reached 82%, 85%, 69%, 67%, 71%, and 73%, respectively. OSCoT-GPT4 consistently outperformed both OPCoT-GPT4 and OSCoT-GPT3.5 with the only exceptions of the maximum and average accuracy values for CeFe, for which other two combinations were more effective. Notably, the minimum accuracy values of OSCoT-GPT4 were the higher than the other two CoT-GPT combinations for all six CeM_1 samples.

Predicting catalysis outcomes poses a significant challenge due to the complexity and multi-step nature of the process. The reasoning route, generated by the OS prompting strategy (Figures 2 and 3), is particularly valuable in this context as it facilitates structured problem-solving. By breaking down the intricate task of catalysis prediction into smaller, logical steps, OSCoT mimics human reasoning, leading to improved understanding and interpretation of the problem. This structured reasoning not only makes the GPT-4's thought process more transparent but also enhances trust in its outputs. Additionally, implementing OSCoT-GPT4 in predicting catalysis can serve as an enhanced form of training, encouraging models to develop a deeper level of understanding and process information in a more nuanced, human-like manner.

Extrapolation to ternary CeM_1M_2 metal-oxide composites

Next, we assessed the performance of the GPT-4, using OSCoT, in predicting outcomes for ternary composites of metal oxides by learning from experiments involving only binary composites. Specifically, we trained a GPT-4 on experiments involving CeM_1 and CeM_2 (M_1, M_2 = Ti, Mn, W, Sn, Mo, or Fe; $M_1 \neq M_2$), following the same OSCoT-GPT4 training procedure as described above. We then evaluated this GPT-4's prediction performance for experiments involving the corresponding ternary CeM_1M_2 composites. This process was independently repeated five times, each instance yielding a unique OSCoT-GPT4. Each time was trained with a distinct batch of 48 experiments for CeM_1 and another for CeM_2, both batches being rationally selected.

We can partially attribute the observed extrapolation performance of the GPT-4 Assistant to its use of associations between specific metals and experimental variables in making predictions. Specifically, the GPT-4 Assistant appears to have constructed a knowledge graph from the

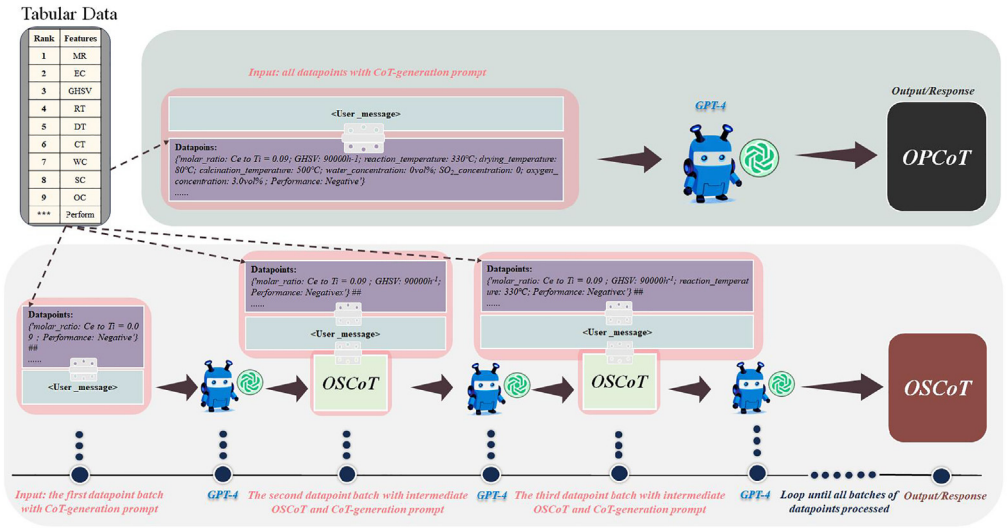


Figure 2. Illustration of CoT generation using the "One-Pot" and "Ordered-and-Structured" approaches

In the OP method, all data points from the table are simultaneously processed to form a single CoT (termed OPCoT). Conversely, in the OS method, table data points are batched according to feature rank hierarchy, with each batch sequentially giving rise to intermediate CoTs. Each CoT incrementally builds upon the logic of the preceding one, representing a progressive development of understanding. The OSCoT materializes through iterative processing of all data point batches. The small chain icon represents the integration of messages, indicating, for example, that a connection between user_message and data points, along with the intermediate OSCoT, integrates these elements into the corresponding user message, thereby solidifying the foundation for subsequent OSCoT iterations. Table S1 details the full names of the abbreviations.

binary training data, linking certain combinations of input variables to either high or low catalytic performances. When making predictions for ternary metal-oxide composites, it considered such associations. However, what remains unclear from the current data and results is the way the GPT-4 Assistant prioritizes these associations, particularly when they conflict. This uncertainty might partly account for the GPT-4 Assistant's reduced predictive performance with ternary systems, where little is known about the interplay and interactions among various metals—other than individual metals each with Ce—as represented on the knowledge graph derived from binary systems. To gain some interpretability, one could consider methods related to information geometry, which offer a structured and mathematical framework to comprehend how information is processed and represented within AI models.

Figure 4 presents results for eight ternary CeM_1M_2 composites, divided into two groups: $CeMnM_2$ (M_2 = W, Ti, Sn, or Fe) and $CeTiM_2$ (M_2 = W, Sn, Mo, or Fe). We focused on these eight, out of the 15 possible permutations of CeM_1M_2, considering the prevalence of their corresponding binary counterparts in the dataset. For each specific CeM_1M_2 composite, 50 experiments involving it were used to evaluate the prediction performance of the GPT-4 trained on the corresponding binary systems. Stoichiometry, as the molar ratio between the metal oxides (binary or ternary), was a variable in all cases.

Figure 4A shows the results of the five independent runs for each of the eight ternary systems. Notably, the GPT-4, all using the OSCoT prompting strategy, exhibited significantly different levels of prediction performance for the various ternary systems. For CeMnFe, CeMnTi, and CeTiFe, the OSCoT-GPT4 consistently yielded high prediction accuracies for the 50 ternary systems, after analyzing only the corresponding binary counterparts (48 CeM_1 + 48 CeM_2). In contrast, the OSCoTs-GPT4 intended for predicting CeTiSn and CeTiMo demonstrated poor performance. For CeMnW, they generated in the five independent runs showed markedly varied prediction performances. Figure 4B provides a statistical summary of the run-specific accuracies for each ternary system. Upon detailed analysis of the various OSCoTs-GPT4 with respect to the systems in Figure 4A, we could attribute the differing levels of prediction performance for the ternary systems to the varying prediction performances of the OSCoTs-GPT4 for the binary systems they were trained to analyze. For instance, the OSCoTs-GPT4 intended for CeMnFe exhibited high prediction performances for CeMn and CeFe, while the ones for CeTiMo demonstrated poor performance for CeTi and CeMo.

Comparison with human experts

To gauge the performance of the various combinations of CoTs-GPT, as described in the preceding sections, we conducted a survey involving four human experts to assess their performance on the same prediction tasks. All four experts were postgraduate research students specializing in NH_3-SCR catalysis and had experience in synthesizing metal-oxide composites and measuring their performances for NH_3-SCR catalysis. First, we asked Experts 1–3 to predict experimental outcomes—i.e., whether the NO_x conversion would be above (Positive) or below (Negative) 95%—for 50 experiments involving binary metal-oxide composites; in Figure 5A, these results are designated as 'Without Training'. In these 50 experiments, the type of CeM_1, its stoichiometry, synthesis conditions, and catalysis reaction conditions were all variables, though certain experiments shared the same values for certain variables. Expert 1 performed the best, attaining a

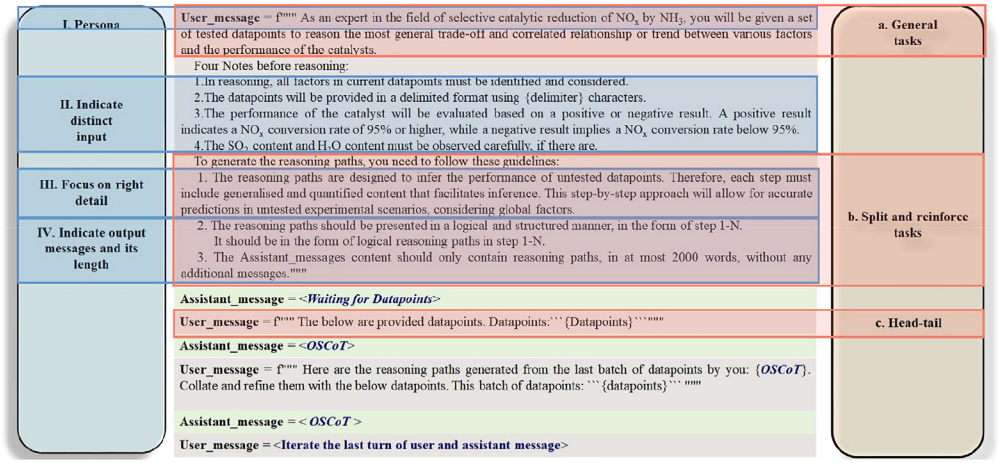


Figure 3. Illustration of structured prompting tactics used to direct the reasoning process for OSCoT generation within user message 1 (UM1)

I: Establishing the persona of the expert expected to analyze the data; II: Providing detailed instructions for input data, including the format and specific observations to note; III: Emphasizing the importance of focusing on relevant details; IV: Dictating the desired output, its structure, and word limit.

(A and B) Decomposing the complex task into general, manageable tasks, and splitting and reinforcing them to ensure thorough analysis.

(C) Putting the analyzed datapoints in the rear of the messages. In Python, the f-string format employs curly braces {} to insert the content within variables into the string. Here, we use angle brackets <> as separators for generated sentences or paragraphs. Additionally, the variable delimiter adopted here is "##".

prediction accuracy of 66%, while Experts 2 and 3 attained prediction accuracies of 51% and 56%, respectively, only marginally better than random guess.

Subsequently, after providing Experts 1–3 with the correct answers for these 50 experiments, they were given another set of 50 different experiments to predict. Their performances, denoted as 'After Training' in Figure 5A, interestingly showed no gains; in fact, Expert 1's performance appreciably worsened. Post-prediction, Experts 1–3 were interviewed and asked to summarize their rationales. As shown in Figure 5C, all three Experts applied relatively simple rules, considering no more than a couple of experimental variables. Their strategies were as follows: (1) focusing on a small set of experiments to identify correlations between the experimental variables and the outcomes or (2) trying to identify a few key experimental variables that significantly influenced the outcomes when altered. While both strategies appear sensible and intuitive, they highlight the challenges humans encounter when analyzing multi-variable problems in sizable datasets. Specifically, correlations significant for a small dataset may not apply to a larger dataset. Similarly, factors that seem significant across the entire dataset may not aid in individual predictions, as the combination and balance of different factors can play an equally, or even more, important role.

Similar observations and conclusions were drawn when Experts 1–3 were asked to make predictions for experiments involving ternary metal-oxide composites (Figure 5B). The variation in prediction performances among the different Experts, as well as the differences in their performance across the various tasks, seem to suggest a strong element of guessing. This is not entirely unexpected, considering the challenge of retaining and processing information from tens of experiments and then applying any discerned rules and correlations to an entirely new set of experiments. An additional expert, Expert 4, who possessed similar experience in the research topic as Experts 1–3, was explicitly instructed to consider multiple experimental variables when approaching the prediction tasks (Figure 5C). Expert 4 was provided with the same sets of experiments for training, as well as the same sets for prediction, as were given to Experts 1–3. Among the four experts, Expert 4 performed the most poorly for both binary and ternary systems.

All four experts' performances were inferior to those of the GPT-4 using OSCoT, as discussed above (Figures 1 and 4). There are several factors that could have contributed to this. The sampling of just four human experts is far from adequate for establishing a comprehensive baseline of human performance on the prediction tasks. Nonetheless, LLMs like GPT-4 may outperform humans in predicting outcomes for complex, multi-variable chemistry experiments for several reasons. First, LLMs possess remarkable data processing capacity, allowing them to analyze and utilize information from lots of experiments simultaneously, a task that is challenging for humans. Second, LLMs consistently apply rules and patterns across datasets without experiencing cognitive fatigue or bias, in contrast to humans who might get overwhelmed by the volume or complexity of the data. Furthermore, LLMs are not susceptible to cognitive biases that can affect human analysis and conclusions. Their ability to detect subtle patterns in diverse and extensive datasets enables them to make more accurate predictions in intricate scenarios. Finally, LLMs benefit from rapid iterative learning, adapting and improving at a pace faster than the typical learning curve of human experts. Overall, the immense data processing capabilities, consistent and objective analysis, and rapid learning and pattern detection of LLMs make them well-equipped for complex tasks such as predicting outcomes in NH_3-SCR catalysis experiments.

DISCUSSION

This study reveals that, by employing an effective OSCoT prompting strategy, GPT-4 achieved notable prediction accuracies regarding the performance of binary CeM_1 metal-oxide composite samples. The average prediction accuracies ranged from 60% to 74%, with peaks

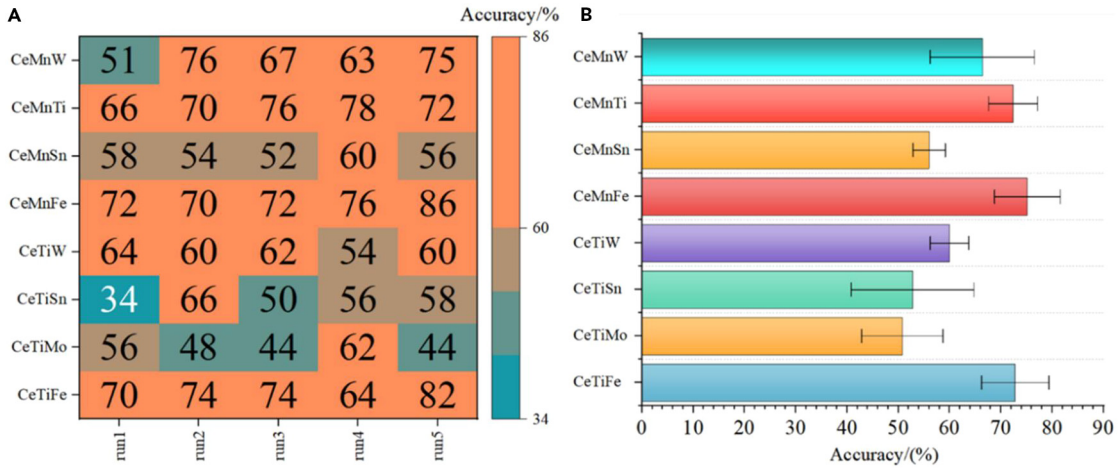


Figure 4. Prediction accuracies for different ternary CeM_1M_2 composites by GPT Assistants after analyzing only the corresponding binary counterparts

(A) Run-specific accuracy results from five independent runs for each ternary case.

(B) Statistical summaries of the corresponding results in (A), with bars indicating the average of the accuracy values from the five runs and error bars representing the standard deviation of these accuracy values.

reaching up to 85%, outperforming human experts. These results were made possible by the OSCoT strategy's ability to break down intricate problems into sequential, manageable steps, enhancing the model's understanding and interpretation of complex tabular data. Extending the generated OSCoT by the binary CeM_1 samples to reason the performance of ternary CeM_1M_2 samples, we observed a varied predictive performance, with some composites like CeMnFe and CeTiFe showing high accuracy, illustrating GPT-4's ability to extrapolate from binary to ternary systems. These findings demonstrate the intricate relationship between variables and causal outcomes in our curated dataset, solidifying the deductive connections implied by our table data through the application of a reasoning CoT.

The comparative analysis with human experts has further highlighted the advantages of employing LLMs in chemistry research. In contrast to human capabilities, LLMs like GPT-4, despite facing context length limitations, are significantly less affected by cognitive biases and are not easily overwhelmed by the immense volume and intricacy of tabular data. They demonstrate exceptional data processing prowess, reflecting some level of reasoning ability, a task that poses a considerable challenge for human experts. Moreover, their uniform application of rules and identification of patterns within table data, combined with their rapid learning and adaptability, equip them to discern subtle correlations amidst varied datasets. This leads to more precise predictions in complex research scenarios.

In this study, we investigated the application of GPT-4, coupled with our proposed OSCoT prompting strategy, to analyze experimental data concerning the variables and outcomes of a specific catalysis. We illustrated GPT-4's adeptness at unraveling intricate, multi-variable correlations within the catalysis. Broadly speaking, an experimental workflow may include several stages: synthesis, characterization, testing, data analysis, iteration of these processes, and/or others. Recent literature examples of human-LLM collaboration in chemical research have underscored the potential of AI to assist and enhance various facets of experimental chemistry research.[12,15–19]

Looking ahead, advanced LLM techniques like Retrieval Augmented Generation (RAG)[20] will continue to enhance human-LLM collaboration on user-defined tasks. For instance, RAG facilitates the seamless incorporation of user-specified datasets, allowing for efficient access to knowledge relevant to the user's queries. LLM implementation frameworks, such as LangChain, simplify the development of RAG-based chatbots.[21] These enabling techniques and their ongoing improvements will promote broader, more effective, and deeper integration of LLMs into chemistry research. They hold the promise of transforming various research tasks, including, but not limited to, the automation of labor-intensive activities like literature mining, interpretation of experimental results, and directing robotic operations.

Limitations of the study

This study investigated the capabilities of LLMs in analyzing experimental data and making predictions on related experiments, in comparison with human chemists during the post-analysis phase of chemical research. Specifically, it introduced an efficient prompt engineering technique named OSCoT for tabular data. Despite its contributions, the study has certain limitations. The method was evaluated using the state-of-the-art GPT-series models and has not been extended to other less advanced models. Furthermore, the concept of chain of thought was utilized to aid in the interpretation of tabular data. However, there remains scope for development in this area, as tabular data are inherently more complex to understand than plain text.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE

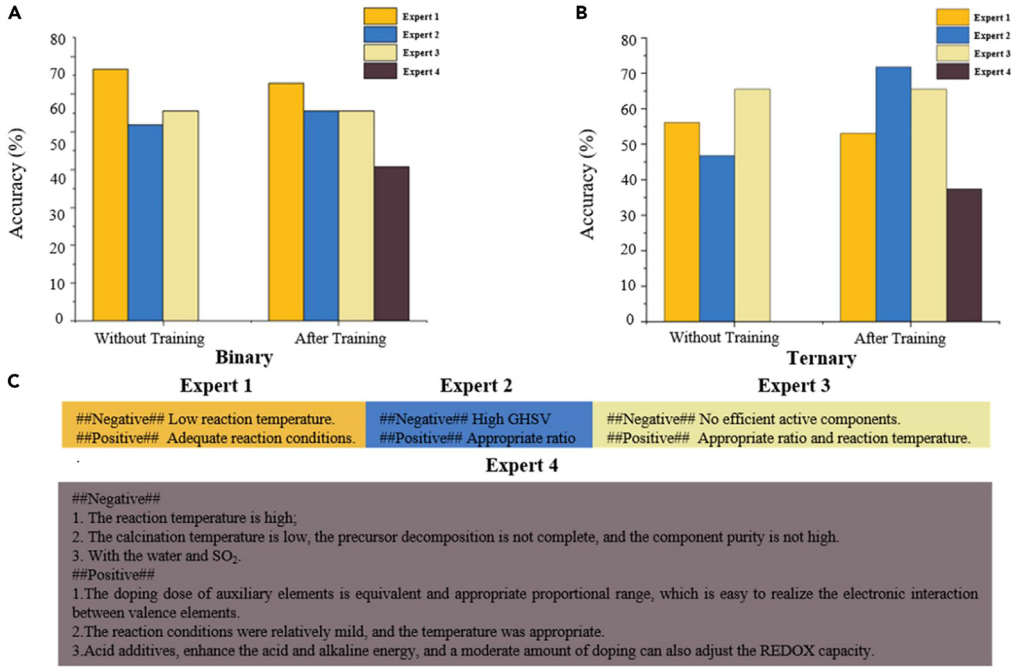


Figure 5. Prediction performances of human experts

(A and B) Prediction of experimental outcomes—specifically, NO_x conversion rates above (positive) or below (negative) 95%—of experiments involving binary (A) or ternary (B) metal-oxide composites.

(C) Summary of the experts' rationales for making their predictions.

- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHODS DETAILS
  - The chemistry context and a case study of NH_3-SCR catalysis
  - Background knowledge of chain-of-thought (CoT) prompting
  - One-pot vs. ordered-and-structured CoT prompting
  - Prompts for the generation and application of CoTs

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.109451.

ACKNOWLEDGMENTS

We gratefully acknowledge funding from the National Natural Science Foundation of China (U20A20130), the Fundamental Research Funds for the Central Universities (06500152), and the "Interdisciplinary Program for Young Teachers" of University of Science and Technology Beijing (FRF-IDRY-21-017). M.L. thanks the China Scholarship Council for a visiting research studentship (No. 202206460056).

AUTHOR CONTRIBUTIONS

L.C. and X.T. conceived and supervised the project. M.L. carried out all the computational work, with support from L.C. and F.G. All authors interpreted the results. L.C. and M.L. prepared the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: November 6, 2023
Revised: January 26, 2024

Accepted: March 6, 2024
Published: March 7, 2024

REFERENCES

1. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., and Askell, A. (2020). Language models are few-shot learners. Adv. Neural Inf. Process. Syst. *33*, 1877–1901.
2. OpenAI. (2023). GPT-4 Technical Report. Preprint at arXiv. https://doi.org/10.48550/arXiv.2303.08774.
3. Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P.S., and Sun, L. (2023). A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. Preprint at arXiv. https://doi.org/10.48550/arXiv.2303.04226.
4. Roumeliotis, K.I., and Tselikas, N.D. (2023). ChatGPT and Open-AI Models: A Preliminary Review. Future Internet *15*, 192.
5. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., and Zhou, D. (2022). Chain of thought prompting elicits reasoning in large language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2201.11903.
6. Liu, H., Ning, R., Teng, Z., Liu, J., Zhou, Q., and Zhang, Y. (2023). Evaluating the logical reasoning ability of chatgpt and gpt-4. Preprint at arXiv. https://doi.org/10.48550/arXiv.2304.03439.
7. Fraiwan, M., and Khasawneh, N. (2023). A Review of ChatGPT Applications in Education, Marketing, Software Engineering, and Healthcare: Benefits, Drawbacks, and Research Directions. Preprint at arXiv. https://doi.org/10.48550/arXiv.2305.00237.
8. OpenAI. (2023). Improving mathematical reasoning with process supervision. https://openai.com/research/improving-mathematical-reasoning-with-process-supervision.
9. Katz, D.M., Bommarito, M.J., Gao, S., and Arredondo, P. (2023). Gpt-4 passes the bar exam. Available at SSRN 4389233.
10. Zhu, J.-J., Jiang, J., Yang, M., and Ren, Z.J. (2023). ChatGPT and environmental research. Environ. Sci. Technol. *57*, 17667–17670.
11. Cheng, L., Li, X., and Bing, L. (2023). Is GPT-4 a Good Data Analyst?. Preprint at arXiv. https://doi.org/10.48550/arXiv.2305.15038.
12. Zheng, Z., Zhang, O., Borgs, C., Chayes, J.T., and Yaghi, O.M. (2023). ChatGPT Chemistry Assistant for Text Mining and Prediction of MOF Synthesis. Preprint at arXiv. https://doi.org/10.1021/jacs.3c05819.
13. Bran, A.M., Cox, S., White, A.D., and Schwaller, P. (2023). ChemCrow: Augmenting large-language models with chemistry tools. Preprint at arXiv. https://doi.org/10.48550/arXiv.2304.05376.
14. Xie, Z., Evangelopoulos, X., Omar, Ö., Troisi, A., Cooper, A.I., and Chen, L. (2023). Fine-Tuning GPT-3 for Machine Learning Electronic and Functional Properties of Organic Molecules.
15. Zheng, Z., Rong, Z., Rampal, N., Borgs, C., Chayes, J.T., and Yaghi, O.M. (2023). A GPT-4 Reticular Chemist for Guiding MOF Discovery. Angew. Chem. Int. Ed. *62*, e202311983.
16. Zheng, Z., Zhang, O., Nguyen, H.L., Rampal, N., Alawadhi, A.H., Rong, Z., Head-Gordon, T., Borgs, C., Chayes, J.T., and Yaghi, O.M. (2023). ChatGPT Research Group for Optimizing the Crystallinity of MOFs and COFs. ACS Cent. Sci. *9*, 2161–2170. https://doi.org/10.1021/acscentsci.3c01087.
17. Zheng, Z., Alawadhi, A.H., Chheda, S., Neumann, S.E., Rampal, N., Liu, S., Nguyen, H.L., Lin, Y.-h., Rong, Z., Siepmann, J.I., et al. (2023). Shaping the Water-Harvesting Behavior of Metal–Organic Frameworks Aided by Fine-Tuned GPT Models. J. Am. Chem. Soc. *145*, 28284–28295. https://doi.org/10.1021/jacs.3c12086.
18. Boiko, D.A., MacKnight, R., Kline, B., and Gomes, G. (2023). Autonomous chemical research with large language models. Nature *624*, 570–578.
19. Yoshikawa, N., Skreta, M., Darvish, K., Arellano-Rubach, S., Ji, Z., Bjørn Kristensen, L., Li, A.Z., Zhao, Y., Xu, H., and Kuramshin, A. (2023). Large language models for chemistry robotics. Aut. Robots, 1–30.
20. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., and Rocktäschel, T. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Adv. Neural Inf. Process. Syst. *33*, 9459–9474.
21. Langchain. https://python.langchain.com/docs/modules/data_connection/.
22. Lu, M., Gao, F., Tan, Y., Yi, H., Gui, Y., Xu, Y., Wang, Y., Zhou, Y., Tang, X., and Chen, L. (2024). Knowledge-Driven Experimental Discovery of Ce-Based Metal Oxide Composites for Selective Catalytic Reduction of NO_x with NH_3 through Interpretable Machine Learning. ACS Appl. Mater. Interfaces *16*, 3593–3604. https://doi.org/10.1021/acsami.3c18490.
23. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., and Iwasawa, Y. (2022). Large language models are zero-shot reasoners. Adv. Neural Inf. Process. Syst. *35*, 22199–22213.
24. Zhang, Z., Zhang, A., Li, M., and Smola, A. (2022). Automatic chain of thought prompting in large language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2210.03493.
25. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2203.11171.
26. Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., and Le, Q. (2022). Least-to-most prompting enables complex reasoning in large language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2205.10625.
27. Shinn, N., Labash, B., and Gopinath, A. (2023). Reflexion: an autonomous agent with dynamic memory and self-reflection. Preprint at arXiv. https://doi.org/10.48550/arXiv.2303.11366.
28. Liu, N.F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. (2023). Lost in the Middle: How Language Models Use Long Contexts. Preprint at arXiv. https://doi.org/10.48550/arXiv.2307.03172.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Raw MS data	This paper	https://github.com/MUYU-LU/OSCoT
Software and algorithms		
Python (version 3.9.7)	Python Software Foundation	https://www.python.org/
Pandas (version 1.3.4)	Python package	https://pandas.pydata.org/
Openai (version 0.0.27.8)	Python package	https://openai.com/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Linjiang Chen (l.j.chen@bham.ac.uk).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The authors declare that data supporting the findings of this study are available if requested. All python codes used in this study are available at https://github.com/MUYU-LU/OSCoT.

METHODS DETAILS

The chemistry context and a case study of NH_3-SCR catalysis

Specifically, we assess the effectiveness of prompting GPT-4 to analyze and interpret experimental tabular data pertaining to a clearly defined chemistry problem. This knowledge involves variables from experiments—such as conditions and parameters—and their corresponding outcomes, all derived from the existing body of literature. The structured experimental data, akin to a JSON format, characterised by keys with corresponding values in either text or numerical form, is rendered to GPT-4. GPT-4 is then prompted to interpret this data, predict the expected experimental outcome based on specified features sets, and explain its prediction. This is informed by our recent success in discovering Ce-based metal-oxide composites for selective catalytic reduction of nitrogen oxides with ammonia,[22] facilitated by interpretable machine learning (ML). In that investigation, utilizing SHapley Additive exPlanations (SHAP) methods to interpret the Extreme Gradient Boosting (XGB) ensemble model trained by 5654 samples detailed various aspects and corresponding NO_x conversion collected from literature brings underscoring influence of variables like reaction temperature and metal elements. The feature importance of these features can be seen as orders of analysis on features, thus embodying deductive relationship in the reasonings paths.

For this purpose, we curated a dataset from existing literature, comprising 1838 unique experiments on NH_3-SCR using Ce-based metal-oxide composites. This dataset encompasses experimental variables such as the catalyst composition (binary and ternary Ce-based metal-oxide composites, CeM_1 and CeM_1M_2 (M_1, M_2 = metal element; $M_1 \neq M_2$), synthesis parameters, reaction conditions, and experimental NO_x conversion outcomes. We adopted a threshold of 95% NO_x conversion to categorize outcomes as "Positive" or "Negative". We designed to evolve variables in the tabular data within the dataset curated here in the reasoning chain to verify whether the GPT-4 can better interpret data and make predictions. In specific, we engaged GPT-4 with the tabular data using Chain-of-Thought (CoT) in an ordered and batchwise and at-once manner. The context length of memory of GPT-4 is much longer than human and therefore we evaluated CoT prompting strategies against the performances of human chemists in predicting for the same experiments. These were critically assessed against our domain expertise and the prevailing consensus in the subject field, offering a comprehensive evaluation of GPT-4's and the chemists' predictive and reasoning capabilities.

Background knowledge of chain-of-thought (CoT) prompting

LLMs are adept at conducting "zero-shot" learning, leveraging knowledge from their extensive training datasets to respond to queries without needing specific prior examples. However, these models often encounter challenges in complex tasks that require advanced reasoning and planning. To overcome these hurdles, various strategies such as "few-shot" prompting and other advanced techniques have been introduced to bolster their capabilities.[23] The "Chain of Thought" (CoT) strategy,[5] in particular, has demonstrated potential in

enabling LLMs to process complex tasks through rational and logical reasoning. It breaks down intricate tasks into manageable, sequential steps. Few-shot CoTs can be used to assist LLMs in tasks that demand a consistent and logical progression, such as common-sense reasoning. In the realm of zero-shot learning, heuristic prompts like "*Let's think step by step*" have been shown to effectively encourage LLMs to 'think aloud', thereby enhancing their problem-solving capabilities.[23] Further advancements include Automatic CoT (Auto-CoT)[24] and self-consistency.[25] Auto-CoT simplifies the process of generating question sampling or reasoning paths, while self-consistency aims to improve the reliability and coherence of LLM outputs. These advancements mark significant progress in enhancing the performance and practicality of LLMs.

The application of CoT lies in the interpreted feature importance in the tabular data that embodies clear deductive relationships between experimental parameters and causal catalytic performance. GPT-4-powered analysis is expected to reveal patterns that signify the varying levels of influence of different experimental factors in NH_3-SCR catalysis. In this analysis, each inference step within the CoT reasoning paths is analogized to the evaluation of a specific experimental variable in catalysis, providing a structured approach to understanding the data. The results of this analysis are presented as CoTs, making the reasoning process and conclusions transparent. Our aim is to develop targeted prompting strategies for GPT-4 to enhance its ability to recognize patterns in structured data and produce outputs formatted as CoTs. These strategies are designed to guide GPT-4 in systematically identifying and interpreting the statistical relationship in the tabular data.

One-pot vs. ordered-and-structured CoT prompting

LLMs often struggle with processing excessive context, necessitating careful planning in the provision of context for enhanced performance.[26] Additionally, systematic review or self-reflection can significantly contribute to overall effectiveness.[27] Bearing these considerations in mind, the full features of experiments in tabular form was divided into batches. In this study, the categorization process is determined by feature importance derived from machine learning interpretation. Indeed, the sequence of batch input can also be influenced by feature importance, as identified through heuristic knowledge or statistical techniques such as correlation analysis. The goal was to ensure that, within each batch, only one experimental parameter varied significantly, while all other parameters remained nearly constant across the samples. For example, in batch X, all data points (i.e., experiments) featured almost identical synthesis and catalysis conditions but varied in the compositions of the catalysts. These batches, each emphasizing a single varying experimental parameter, were then sequentially presented to GPT-4. This approach augmented GPT-4's analysis depth for individual experimental parameters, ultimately facilitating the generation of more optimal CoTs.

The batchwise, sequential approach is hereafter termed the "Ordered-and-Structured" (OS)-CoT prompting strategy. It was compared with a "One-Pot" (OP) prompting strategy, where all data points were inputted simultaneously. The implementation of both CoT prompting strategies is depicted in Figure 2. The OPCoT approach creates a single CoT in at once by using all available data points. In contrast, the OSCoT approach sequentially generates multiple CoTs. As each new batch of data points is introduced, a fresh CoT is created, incorporating the reasoning from the preceding CoT developed in the previous batch. Consequently, the CoTs formed by the OSCoT approach, each building upon the insights of the last.

Prompts for the generation and application of CoTs

In the web-based ChatGPT user interface, chat sessions are facilitated through user-initiated messages. In this study, we focused not on the ChatGPT interface but rather on using OpenAI's application programming interface (API). In the context of the API, it is necessary to format messages as dialogues involving typically two participant roles: the User (or the API Client) and the Assistant (the AI Model). The User is the entity, either a person or another system, that sends requests to the API. These requests, or user inputs, are what the AI model responds to. The Assistant, on the other hand, is the role played by the AI, such as GPT-4, generating responses based on the User's input. Here, we evaluate and compare the performances of both GPT-4 and GPT-3.5 (GPT-3.5-turbo-16k). Notably, we have set the ''temperature'' hyperparameter, which influences the model's prediction randomness, to zero to ensure the most consistent output as possible for the same input.

The effectiveness of LLM-generated responses heavily relies on the quality of task-directive prompts within in-context learning across multiple conversations. Our approach involves integrating various tactics to generate querying prompts. Figure 3 illustrates the user messages designated as **UM1**, which were used for OSCoT generation. Figure S1 depicts the user messages for OPCoT generation, designated as **UM2**. Figures S2 and S3 demonstrate the utilization of CoTs to predict the experimental performance of CeM_1 and CeM_1M_2 samples, respectively, with user messages designated as **UM3** and **UM4**. Additionally, given the token limitations inherent in GPT models, where both assistant (i.e., GPT) and user messages contribute to the token input capacity, careful selection of data samples is crucial. The protocol for the rational selection process and the computation of token utilization are detailed in the Supporting Information.

Before eliciting CoTs, careful configuration of in-context content is necessary to guide the cognitive processes of LLMs. The initial interaction with **UM1**, as shown in Figure 3, demonstrates the use of distinct tactics for establishing this setup, identical to those employed for OPCoT generation using **UM2**, as presented in Figure S1. Tactics **I-IV** were utilized for the prompts providing "**Clear Instructions**." GPT-4, initially agnostic to roles, can adopt a specific role when prompted. Consequently, we assigned the role of an NH_3-SCR catalysis expert to GPT-4 for target-oriented tasks using the 'persona' tactic (**I**). This was followed by clarifying the input content to enhance GPT-4's focus on data points (**II**). Subsequently, we emphasized generating CoTs with consistent, generalized, and quantitative characteristics to establish a foundation for subsequent inferences (**III**). Next, we eliminated unnecessary information and requested reformatted output to streamline the process and enhance output quality (**IV**).

Tactics **a-c** serve the role of "**Decomposing Complex Tasks**." In particular, we outlined the overall task (**a**) and then broke it down into several simpler component parts to facilitate the generation of CoTs (**b**). The subsequent user message implemented the 'head-tail' tactic

(**c**), placing the most crucial prompts at the beginning and end of the message. This tactic was employed to append data points at the end during each iteration of CoT generation, as LLMs tend to 'lose focus in the middle' with lengthy contextual input.[28]

For using CoTs to infer experimental performance, the initial interaction of **UM3** or **UM4** (as depicted in Figures S2 and S3, respectively) in the front-user message for context-aware adaptation employed tactics **I**, **II**, **IV**, and **a**. These tactics were used to establish the role of the NH_3-SCR expert, specify input material, define output's structure and format, and outline the general objective. Data points and CoTs were placed at the end of messages as part of tactic **c**. Additionally, we employed a cautious tactic termed "**allocating thinking**" (**d**), crucial for conducting comprehensive analysis prior to making decisions. This tactic enhances the use of "**loud thinking**," considering all reasoning paths of CoTs. It also moderates the pace of text completion and prevents exceeding the rate limits, which could lead to execution errors if the server is queried too quickly.