Environmental Modelling and Software 196 (2026) 106772

Contents lists available at ScienceDirect

Environmental Modelling and Software

journal homepage: www.elsevier.com/locate/envsoft

How well do DeepSeek, ChatGPT, and Gemini respond to water science questions?

Seyed Hossein Hosseini [a,*], Ali Pourzangbar [b]

[a] Department of Built Environment, School of Engineering, Aalto University, Espoo, Finland
[b] Institute for Water and River Basin Management, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

ARTICLE INFO

Keywords:
Large language model (LLM)
Hydrology
Water science
DeepSeek R1
ChatGPT-4o
Gemini 2

ABSTRACT

This study aims to evaluate the performance of three prominent LLMs, DeepSeek R1, ChatGPT-4o, and Gemini 2, in addressing key questions within four core fields of hydrology and water science: machine learning and optimization, remote sensing, flood modeling, and sediment transport. LLMs' responses are systematically compared to benchmark responses derived from review articles in the respective fields. To assess the LLMs' efficiency, a novel evaluation rubric is introduced in this study, incorporating four key criteria: relevancy, accuracy, authenticity, and novelty. Findings revealed that each model can address the core aspects of the benchmark questions. DeepSeek R1 achieved the highest overall scores in machine learning and optimization, flood modeling, and sediment transport, while ChatGPT-4o demonstrated superior performance in remote sensing. Notably, DeepSeek R1 and Gemini 2 exhibited the lowest response similarity in 95 % of the evaluated questions, whereas ChatGPT-4o and Gemini 2 showed the highest similarity in 70 % of cases.

1. Introduction

Historically, scholars relied on books for scientific inquiries, though limited references made the process time-consuming. The internet streamlined research, yet finding the best answer still required extensive browsing. Nowadays, the advent of Large Language Models (LLMs) has revolutionized many fields. In fact, these models are a type of Deep Learning (DL) that can generate human-like text based on vast training data (Tian et al., 2023). Perspectives on LLMs differ, with some emphasizing their advantages, while others question their practical applicability (Cheng & YIM, 2024; Reuters, 2023; Yu, 2023). Nonetheless, LLMs have become a major focus in recent years, with their adoption expanding across scientific disciplines alongside the emergence of more advanced models.

LLMs have been implemented across various fields, including civil engineering (Aluga, 2023), software engineering (Akbar et al., 2025; Belzner et al., 2024; Liang et al., 2024; Ozkaya, 2023), public health (Biswas, 2023b; Jungwirth and Haluza, 2023; Parray et al., 2023), and social media (Hu et al., 2023; Rodríguez-Ibánez et al., 2023; Sadıkoğlu et al., 2023). In hydrology and water science, researchers have explored LLM applications in programming (Pursnani et al., 2024), data analysis (Biswas, 2023a), interpretation of hydrological models (Xia et al., 2025), and remote sensing (Guo et al., 2024). Among the many available LLMs, ChatGPT, Gemini, and DeepSeek stand out for their broad applicability, strong reasoning abilities, and reliability in complex problem-solving (Brown et al., 2020). Their advanced multimodal features and ability to process hydrological data make them well-suited for water science applications (Kadiyala et al., 2024). Therefore, this study aims to assess the effectiveness of these LLMs in addressing hydrological challenges.

ChatGPT-4o, launched by OpenAI in May 2024, is an advanced natural language processing model. This model has been fine-tuned using reinforcement learning and supervised training, enhancing its ability to generate human-like text with high accuracy (OpenAI, 2024). Building on ChatGPT and GPT-4, it offers improved intelligence, faster processing, and expanded capabilities in text generation, vision, and audio processing. Studies related to hydrology highlight its ability to code at beginner-to-intermediate levels (Foroumandi et al., 2023), perform hydrological data analysis without coding (Irvine et al., 2023), contribute to flood management and water quality assessment (Kadiyala et al., 2024), water resources management (Haider et al., 2024), natural disaster prevention and reduction (Xue et al., 2023), and finally, multimodal data analysis, intelligent decision-making, and interdisciplinary knowledge integration (Ren et al., 2024). ChatGPT also demonstrates capabilities in image description, edge detection, and

* Corresponding author.
E-mail addresses: seyed.h.hosseini@aalto.fi (S.H. Hosseini), ali.pourzangbar@kit.edu (A. Pourzangbar).

https://doi.org/10.1016/j.envsoft.2025.106772
Received 26 March 2025; Received in revised form 16 September 2025; Accepted 6 November 2025
Available online 7 November 2025

segmentation (Guo et al., 2024; Osco et al., 2023). While promising, its use in environmental decision-making should be carefully examined, given ongoing concerns over reliability (Zhu et al., 2023).

ChatGPT is not the only LLM that has been utilized in water science. DeepSeek R1 (Guo et al., 2025), introduced in January 2025, represents a significant advancement in LLMs with its open-source nature fostering collaboration while raising concerns about misuse (Mercer et al., 2025; Sallam et al., 2025). Built on a Mixture of Experts (MoE) architecture and reinforcement learning, it enhances reasoning capabilities through a multi-stage training process (Olson et al., 2025; Paiva et al., 2025). Meanwhile, Google's Gemini was launched alongside a lightweight language model for dialog applications version (LaMDA), leveraging Transformer decoders, Tensor Processing Units (TPUs) hardware, and attention mechanisms to optimize performance and expand accessibility (Sundar Pichai, 2023; Team et al., 2024). Regarding the mentioned LLMs, Li et al., 2025 introduced ATMOSSCI-BENCH, a benchmarking tool that evaluates models like GPT, Llama, Gemini, and DeepSeek across five atmospheric science domains, including hydrology, atmospheric dynamics, atmospheric physics, geophysics, and physical oceanography. Using a template-based question framework, the study assessed LLM performance in instruction tuning, reasoning, mathematical capabilities, and domain-specific climate modeling. In addition to this research, the application of LLMs such as GPT-4 Vision, Multimodal-GPT, Gemini, and LLaVa in hydrology has been investigated by Kadiyala et al. (2024), demonstrating their effectiveness in real-time flood severity assessment, water quality monitoring, and water resource management. Among them, GPT-4 Vision exhibited the highest accuracy in interpreting visual data for hydrological applications. The word cloud of the research based on the title of the articles used in the literature is shown in Figure (1).

A review of existing studies reveals critical gaps in evaluating LLMs for hydrology and water science. Studies focusing on proposing a solid assessment framework are limited (Ren et al., 2024; Kizilkaya et al., 2025). Additionally, previous studies have not thoroughly explored LLM applications in machine learning and optimization, remote sensing, flood modeling, and sediment transport. These topics are among the popular research topics in water science and hydrology and provide valuable opportunities to investigate the capabilities of LLMs within these fields. To address these gaps, this study introduces three key innovations. First, we systematically evaluate the performance of DeepSeek, ChatGPT, and Gemini in key hydrological domains. Second, we develop a novel rubric for quantitatively assessing the efficiency and reliability of LLM-generated responses. Third, we design benchmark questions and responses based on review articles in each topic, ensuring a rigorous and domain-relevant evaluation. In fact, the novelties of our proposed framework and rubric are both conceptual and technical. The conceptual novelty refers to the design of the rubric and its components. The proposed framework evaluates four principal components of a response: accuracy, relevancy, authenticity, and novelty, which have not been considered in the previous studies with sufficient depth. The technical novelty lies in the evaluation of different components, particularly accuracy and similarity. A word matching algorithm is introduced to assess the accuracy of the generated responses by comparing them to the key responses in the benchmark set. Furthermore, the Spacy library is employed to analyze the semantic similarity of LLM responses. Overall, this study proposes a comprehensive rubric and associated computational methods to evaluate the quality of LLM responses in water science and hydrology. This integrated approach provides a comprehensive framework for assessing LLMs in hydrology, bridging existing gaps, making practical recommendations, and enhancing Artificial Intelligence (AI)-driven water science research.

This paper is structured as follows. Section 2 outlines the research

Fig. 1. Word cloud of the application of LLMs in various research areas, including water science.

framework, detailing the prompt engineering, the evaluation rubric for LLMs, and the algorithm for assessing word matching in responses. Section 3 analyzes LLM-generated responses across key domains, including machine learning and optimization, remote sensing, flood modeling, and sediment transport, with a comprehensive comparison based on overall scores, similarity metrics, word matching, and response generation speed. Section 4 provides a discussion on the findings and literature, alongside the practical considerations for using LLMs in hydrology and water science. Section 5 outlines limitations and provides recommendations for future research directions. Finally, section 6 concludes with the findings of this research. In Table 1 list of abbreviations used in this research is presented.

2. Material and methodology

This study evaluates the performance of three LLMs, including DeepSeek R1, GPT-4o, and Gemini 2 (all free versions to be accessible for all), in four key areas of hydrology and water science: machine learning and optimization, remote sensing, flood modeling, and sediment transport. For clarity, the models are referred to as DeepSeek, ChatGPT, and Gemini. All LLMs are accessed through their official platforms. To establish a robust evaluation framework, key questions in each domain are identified through a systematic review of relevant literature. Benchmark responses are extracted from these sources as reference answers. The LLM-generated responses are then systematically compared against these benchmarks using a proposed structured rubric based on four criteria: relevancy, accuracy, authenticity, and novelty. Each response is assigned a weighted score to ensure a comprehensive performance assessment. Additionally, qualitative insights are included to capture nuanced aspects of each model's output. Each response is assessed using the proposed rubric, with weighted scoring applied to quantify the impact of each criterion. Figure (2) presents a flowchart outlining the methodology.

2.1. Design of benchmark questions

To select benchmark questions, peer-reviewed review articles are chosen as references. The logic behind considering review articles is that, in general, review articles study a phenomenon thoroughly, contain a broad knowledge of a specific topic, and are an aggregation of many articles. We conducted a targeted literature search using Scopus, Web of Science, and Google Scholar to identify review articles relevant to four hydrological subtopics. Fourteen articles were selected based on their thematic relevance, citation impact, and breadth of synthesis across primary studies. The authors' judgment was applied to ensure topical diversity and methodological rigor, to support the transparency and reproducibility of our evaluation framework. Selected review articles are among the most cited in the field and are also regarded as reliable sources of scientific information. After evaluating the potential of several review articles, fourteen review articles for all topics were selected to design the benchmark questions, and five benchmark questions were designed for each topic. These five questions, based on the authors' judgment, were considered representative of the topic. In formulating the benchmark questions, questions with case-specific or quantitative answers were avoided. The main reason for that is to ensure the rubric assesses generalizable knowledge and conceptual understanding applicable across diverse contexts. In other words, the main focus of the benchmark questions is on challenges, applications, and solutions related to the investigated topics. Therefore, for each topic, five subtopics were considered in order to cover important aspects of the selected topics based on the authors' attitudes and expertise.

Table 1
List of abbreviations used in this research.

Abbreviation	Description	Abbreviation	Description
ADCP	Acoustic Doppler Current Profiler	IoT	Internet of things
AI	Artificial Intelligence	LES	Large Eddy Simulations
BNNs	Bayesian Neural Networks	LLM	Large Language Model
CFD-DEM	Computational Fluid Dynamics/Discrete Element Method	LSTM	Long Short-Term Memory
CHNS	Coupled Human-Natural System	ML	Machine Learning
DL	Deep Learning	MoE	Mixture of Experts
DNS	Direct numerical simulation	NbS	Nature-based Solutions
FSM	Flood Susceptibility Mapping	NLP	Natural Language Processing
GPU	Graphics Processing Unit	O & M	Operation & Maintenance
HPC	High Performance Computing	TPUs	Tensor Processing Units

2.2. Prompt engineering

In general, prompts can guide LLMs towards the topic being investigated. The tone and style of the prompt change the formality, complexity, and presentation of the LLM response ("explain simply" and "explain as an expert" lead to different forms of response). In addition, detailed and more structured prompts usually yield longer responses, while brief prompts and vague ones give less detailed responses. For this reason, due to their influential nature in the results, designing an effective prompt is crucial. In this research, a general view and a role-based prompt have been crafted for benchmark questions. Initially, a general view is considered. This general view includes information on the topic extracted from the review articles, which does not encompass the response to the benchmark question. This general view obviously varies for each question and is derived exclusively from the review article's content, without incorporating any extra information or the authors' personal views. Subsequently, a role-based prompt is provided to the LLM. The purpose of this prompt is to direct the LLM to generate a response as if it were an expert in the field of hydrology and provide a structured response. Therefore, by using this role-based prompt, LLMs are pushed to be domain-specific, and it helps to obtain responses that are relevant and avoid vague or generic information. The general view and role-based prompt have been presented for each question in the appendix. It should be noted that to ensure consistency and fairness, each LLM received the same standardized role-based prompt before being presented with the question. Finally, after presenting the general view and role-based prompt, a question is posed to the LLM. In conclusion, a uniform prompt structure was established to minimize biases in response generation and ensure a fair evaluation across models. The prompt underwent multiple refinements through trials to determine the most effective version (aimed at producing a clear and precise scientific response, according to the author's judgment). An example of this prompt engineering for the first question related to machine learning and optimization has been provided as follows. General views of other questions have been provided in the Appendix.

General view:

The hydrologic community has recently experienced a surge in interest in machine learning. This interest is primarily driven by rapidly growing hydrologic data repositories, as well as the success of machine learning in various academic and commercial applications, now possible due to increasing accessibility to enabling hardware and software. Unsupervised learning in hydrology can cluster catchments by hydrologic regimes, while classification problems distinguish land cover types from satellite images using spectral data. Streamflow forecasting is a regression problem, predicting future flow based on meteorological and historical data. Machine learning models, trained by minimizing error metrics, use algorithms

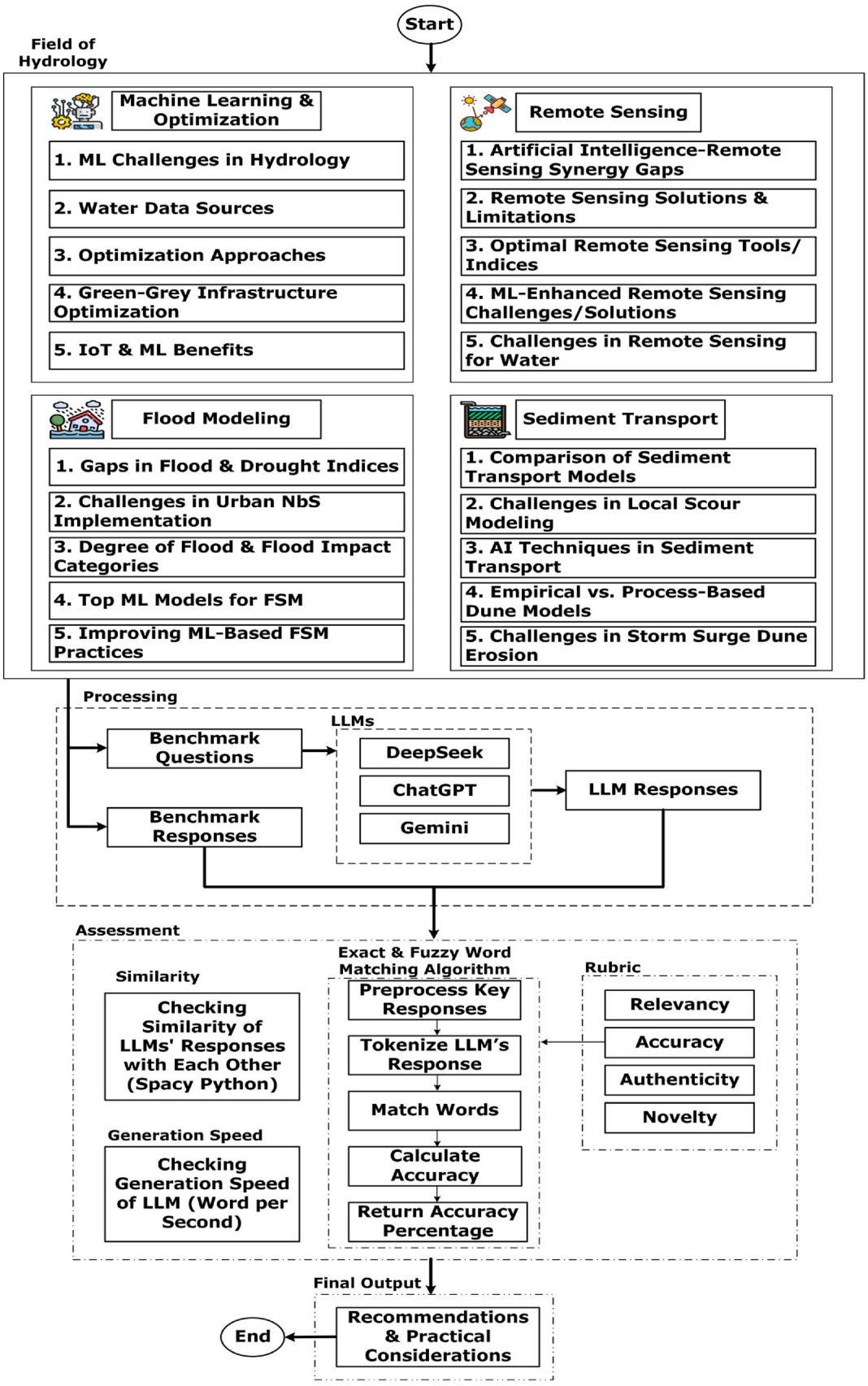

Fig. 2. Flowchart of the proposed method for assessing the efficiency of LLMs.

differing in hypothesis space, loss functions, and optimization methods, with applications in hydrologic sciences.

Role-based prompt:

"Respond as an expert in hydrology and water science to the following question. Provide detailed, professional answers. Ensure that your responses are specific to the field of hydrology and avoid generalizations. Consider including relevant concepts, methodologies, trends, and advancements in your answers to showcase a thorough understanding of hydrology and water resources management."

Question:

The application of machine learning in hydrology and other geosciences disciplines has been hindered by three primary challenges. Mention these challenges.

2.3. Evaluation of LLMs' responses

Rubrics are typically used to assess performance, grading, and evaluations based on predefined criteria. In this research, a rubric is introduced to investigate the generated answers from various aspects. Four criteria, including relevancy, accuracy, authenticity, and novelty, are proposed to build a rubric for assessing the efficiency of LLMs' answers. The introduction of these criteria is based on the nature of the research question and the investigated phenomena. In fact, the rubric design is inherently flexible and context-dependent. While standardized rubrics exist, researchers often develop custom criteria to address unique problems or emerging domains where established frameworks are inadequate (Andrade, 2005; Brookhart, 2013; Moskal and Leydens, 2000). Therefore, in this research, a novel rubric is proposed based on four criteria, including relevancy, accuracy, authenticity, and novelty, and it can be used for future research in this field. The definition of each criterion is presented in Table (2).

In addition, each criterion has a coefficient, which is used to grade the LLM performance using equation (1). The coefficients are based on the authors' interpretation of the importance of each criterion. In fact, the coefficients are assigned to critical priorities for assessing the performance of LLMs. Therefore, the highest coefficient has been given to accuracy and authenticity, each receiving 35 %, as an LLM is fundamentally expected to provide factual responses. In other words, a response must first be accurate and authentic compared to the benchmark response, which justifies assigning these criteria higher coefficients. A coefficient of 20 % is allocated to novelty to encourage creative and insightful responses rather than mere repackaging of information. Finally, a coefficient of 10 % is suggested for relevancy, which is considered less significant than the others; it often serves as a prerequisite for a valuable response but is not a defining factor of performance. These values reflect evaluation priorities rather than statistical optimization. Here, the proposed approach of assigning coefficients has been tried to suggest an appropriate balance by emphasizing factual accuracy while also fostering creativity and maintaining contextual relevance for assessing LLMs. For each question, the generated response by the LLM is evaluated using benchmark responses and then scored. This rubric is presented in Table (3).

$$Overall\ Score = 0.10 \times Relevancy + 0.35 \times Accuracy + 0.35 \times Authenticity + 0.20 \times Novelty \quad (1)$$

Based on the scores of the proposed rubric, the minimum and maximum values for the overall score would be 1 and 5, respectively.

Relevancy, authenticity, and novelty were evaluated using a structured rubric with defined scoring levels. Each LLM-generated response to the benchmark was assigned scores by the authors. To ensure impartiality, a blind assessment procedure was followed, where the

Table 2
The definition of the proposed criterion.

Criteria	Definition
Relevancy	The quality or state of being closely connected and directly related to the specific topic. In other words, it describes the close connection and direct relationship between the discussed topic and the information or ideas being generated using LLM.
Accuracy	The degree to which the generated response aligns with the benchmark answers, highlighting the inclusion of key elements or information required. It involves ensuring that the response adheres to established facts, and logical consistency and correctly reflects the critical components expected in the solution. Accuracy is assessed by comparing the response to predefined correct answers (key responses).
Authenticity	It measures the validity of information, statements, or solutions. It involves adhering to established rules, principles, or standards and aligning with factual and logical information. Authenticity is evaluated through verification, fact-checking, logical reasoning, and comparison with accepted norms or references.
Novelty	The quality of being new, original, and unique is characterized by fresh and innovative elements that differentiate it from what has previously existed or been known. It can manifest in various domains, including ideas, designs, expression, and problem-solving approaches. In fact, novelty introduces fresh and unique elements that inspire curiosity and fascination, beyond the benchmark response.

Table 3
The proposed rubric for the evaluation of the generated responses by LLMs.

Criteria	Coefficient	Score	Definition
Relevancy	0.10	1	The response is completely unrelated to the question
		2	The response contains significant unrelated information, addressing only a small portion of the question
		3	The response balances relevant and irrelevant information equally
		4	The response is mostly relevant, with minor unrelated content
		5	The response is entirely relevant, addressing all aspects of the question
Accuracy	0.35	1	The response mentions a maximum of 20 % of core elements and objectives (key responses)
		2	The response mentions a maximum of 40 % of core elements and objectives (key responses)
		3	The response mentions a maximum of 60 % of core elements and objectives (key responses)
		4	The response mentions a maximum of 80 % of core elements and objectives (key responses)
		5	The response mentions 100 % of core elements and objectives (key responses)
Authenticity	0.35	1	The response is entirely incorrect and unclear from a scientific standpoint
		2	The response is partially correct, but most parts are scientifically inaccurate
		3	The response contains an equal mix of correct and incorrect scientific information
		4	The response is mostly accurate, with minor doubts about the authenticity of certain parts
		5	The response is entirely accurate and scientifically correct
Novelty	0.20	1	The response lacks any novel information or ideas compared to the benchmark response
		2	The response exhibits minimal novelty, offering few new insights
		3	The response presents moderately novel ideas that are underdeveloped
		4	The response mainly comprises novel solutions or ideas
		5	The response is highly novel, entirely consisting of new and original ideas

identities of the LLMs were concealed during evaluation. For accuracy, a novel approach titled "Exact and Fuzzy Word Matching Algorithm" is proposed. Keywords from benchmark responses are compared with the LLM's responses using this algorithm. The proposed algorithm employs a hybrid approach that integrates exact and fuzzy string matching to evaluate the presence of predefined keywords extracted from the benchmark responses, which were carefully selected by the authors to capture the core ideas and gist within the benchmark response. Initially, the input keyword list is preprocessed by converting it into a set of lowercase words and eliminating extra spaces to ensure uniformity. The text is then tokenized using regular expressions to extract valid words while preserving case insensitivity. The algorithm first checks for exact matches between the keywords and the tokenized text, ensuring a direct word-to-word comparison. If any words are not found exactly, the function employs a fuzzy matching technique using the RapidFuzz library in Python, which calculates the scores based on the Levenshtein distance. If a word achieves an accuracy score above a threshold, which is obtained by trial and error, it is considered a valid match. The final matching percentage is computed as the ratio of successfully identified words to the total words in the keyword list. The pseudo-code of the proposed algorithm for accuracy is presented in Figure (3).

Another evaluation used in this research is assessing the similarity of LLMs' responses with one another. In fact, the goal of this analysis is to determine the similarity of LLMs' responses in terms of meaning. For this purpose, the spaCy library is used. The spaCy library, built on Python, offers a range of efficient text processing tools for multiple languages, and here the English package is used. Its models are widely recognized as the go-to choice for practical NLP, valued for their speed, robustness, and near-state-of-the-art performance (Honnibal, 2017; Neumann et al., 2019).

Finally, for each topic, the generation speed of each LLM is calculated. Using generation speed allows for a fair comparison of LLM performance. In fact, while it is possible to compare the time each LLM takes to generate a response and the number of words it generates, comparing only these two variables separately may not yield a fair comparison. Therefore, in this research, the generation speed of each response by dividing the number of words by the time taken is calculated and used for further comparisons.

3. Results

This section presents results obtained from LLMs in assessing questions in machine learning and optimization, remote sensing, flood modeling, and sediment transport modeling.

3.1. Machine learning and optimization

Five questions were designed for machine learning and optimization, and their responses were generated using different LLMs. The questions and their interpretations are presented below. It should be noted that before each question, a general view related to the questions and the role-based prompt was fed to the LLMs. The prompt, questions (Mala-Jetmarova et al., 2018; T. Xu and Liang, 2021; Kamyab et al., 2023; Jayaraman et al., 2024; Tansar et al., 2024), and response to each question are presented in Appendix (A).

Across all five questions in the machine learning and optimization topic, DeepSeek, ChatGPT, and Gemini demonstrate distinct strengths for different audiences. DeepSeek consistently provides highly technical, structured, and comprehensive analyses, excelling in mathematical formulations, emerging technologies, and solution-oriented recommendations. ChatGPT balances practicality with technical depth, emphasizing data integration and chronological trends. Gemini prioritizes field applicability, sustainability, and stakeholder perspectives with unique insights into instrumentation, citizen science, and social

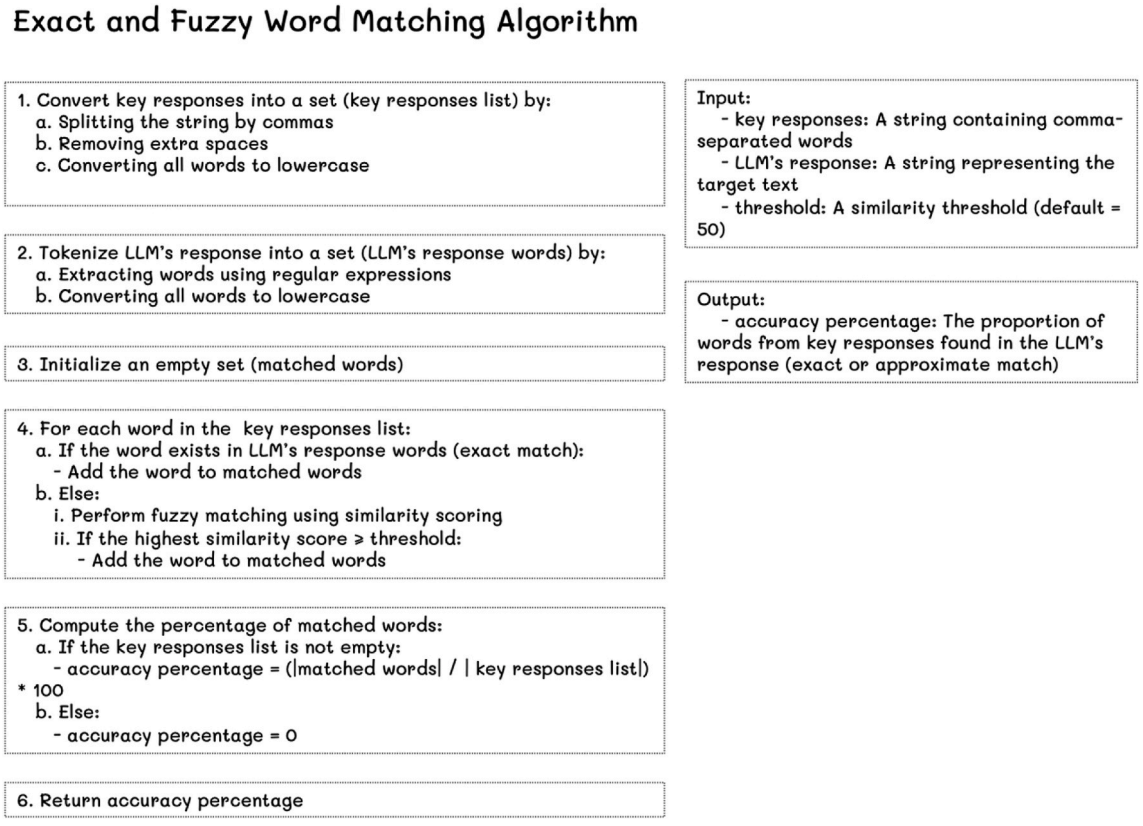

Fig. 3. Flowchart of the proposed exact and fuzzy word matching algorithm to find the similarity percentage between key responses in the benchmark responses and LLM responses.

disruption. It is worth noting that all models advocate hybrid approaches (e.g., machine learning with domain knowledge).

An overall quantitative comparison of LLMs' performance in answering the questions is presented as follows. This comparison consists of assessing the overall score, similarity, accuracy, and generation speed. Table (4) presents the results of the overall scores for each question in the machine learning and optimization topic corresponding to DeepSeek, ChatGPT, and Gemini. According to the findings, DeepSeek demonstrated superior performance in answering the first and second questions, while ChatGPT performed better on the fifth question. On the other hand, Gemini excelled in answering the third question. All models exhibited the same performance for the fourth question. Overall, the summation of scores shows that the ranking of LLMs in responding to the machine learning and optimization questions is DeepSeek, Gemini, and ChatGPT, respectively.

One important aspect of the responses generated by LLMs is their similarity to one another. In this regard, the Spacy library was used to measure the level of similarity among the responses. According to the findings, for responses related to the machine learning and optimization topic, all LLMs exhibit a high level of similarity in their pairwise responses. Out of five questions, in four of them, DeepSeek and ChatGPT show the highest similarity with each other. Conversely, in all five questions, DeepSeek and Gemini have the lowest similarity in their responses. In addition, the highest correlation is related to the response to the third question for DeepSeek and ChatGPT, with a value of 0.97. On the other hand, the lowest correlation is related to the response to the second question for DeepSeek and Gemini, with a value of 0.85. In Figure (4), the results of the similarity analysis for the machine learning topic are presented.

Accuracy is one of the criteria used in this research, calculated using the exact and fuzzy word matching algorithm. As previously mentioned, this algorithm determines the portion of key responses in the benchmark response compared to the generated responses by each LLM. The results of this analysis for the machine learning and optimization topic are shown in Figure (5). The accuracy comparison of Gemini, ChatGPT, and DeepSeek across five benchmark questions indicates a high overall performance, particularly in the fifth question, where all models achieve nearly identical and maximal accuracy. The second question also demonstrates strong accuracy across all models. However, variations appear in the third and fourth questions. The first question also presents differences among the models. Overall, findings indicate that the average accuracy of all questions for DeepSeek, ChatGPT, and Gemini is 75.48 %, 73.59 %, and 71.07 %, respectively.

Figure (6) shows the generation speed of each model for the machine learning and optimization topic. Based on the results, DeepSeek shows the lowest generation speed in all questions. In contrast, Gemini exhibits the highest generation speed across all five questions compared to the other models. The average generation speeds of all five questions related to the machine learning and optimization for DeepSeek, ChatGPT, and Gemini are 6.40 (word/second), 13.25 (word/second), and 86.16 (word/second), respectively.

Table 4
Overall score of LLMs in the machine learning and optimization topic.

Question	Model		
	DeepSeek	ChatGPT	Gemini
Q1	4,30	3,10	3,55
Q2	4,10	3,90	3,90
Q3	3,90	3,90	4,45
Q4	3,75	3,75	3,75
Q5	3,90	4,10	3,90
Sum	19,95	18,75	19,55

3.2. Remote sensing

Five questions were designed for remote sensing, and their responses were generated using different LLMs. The questions and their interpretations are presented below. It should be noted that before each question, a general view related to the questions and the role-based prompt was fed to the LLMs. The prompt, questions (Bhaga et al., 2020; Sagan et al., 2020; Chen et al., 2022; Sun et al., 2024), and response to each question are presented in Appendix (B).

Across all five questions, DeepSeek, ChatGPT, and Gemini show unique strengths in addressing the questions in the remote sensing topic. DeepSeek demonstrates strong computational depth, incorporating advanced AI techniques such as physics-informed models, federated learning, and edge computing while excelling in technical specificity. However, it sometimes omits practical elements like cloud computing or widely used benchmark tools. ChatGPT provides well-structured and balanced responses, focusing on AI scalability, integration challenges, and cost-efficient solutions. However, it occasionally underemphasizes emerging AI techniques and specific benchmark indices. Gemini prioritizes usability and real-world application, effectively discussing challenges in sensor fusion, field validation, and hydrological interpretation, though it sometimes replaces benchmark-recommended tools with alternatives or underemphasizes key computational methodologies. Across all models, notable gaps include governance frameworks, hydrodynamic model integration, and explicit prioritization of benchmark-recommended remote sensing tools/indices.

In Table (5), the results of the overall scores for each question in the remote sensing topic are presented. According to the findings, DeepSeek showed better performance in answering the first and third questions, while ChatGPT performed better on the second, fourth, and fifth questions. Overall, the summation of scores shows that the ranking of LLMs in responding to the remote sensing questions is ChatGPT, DeepSeek, and Gemini, respectively.

The results of similarity in the remote sensing topic are presented in Figure (7). Based on the results, all LLMs exhibit a strong level of similarity in their pairwise responses. In four out of five questions, Gemini and ChatGPT show the highest similarity in their responses. Conversely, in all five questions, DeepSeek and Gemini have the lowest similarity with each other. In addition, the highest correlation is related to the response to the fourth question for Gemini and ChatGPT, with a value of 0.95. On the other hand, the lowest correlation is related to the response to the third question for DeepSeek and Gemini, with a value of 0.67.

The accuracy comparison of Gemini, ChatGPT, and DeepSeek across five benchmark questions related to the remote sensing topic is presented in Figure (8). According to the findings, all models obtained relatively high accuracy for the second and fourth questions. While all models show nearly close results (especially ChatGPT and Gemini, which demonstrate consistent accuracy for the first and third questions), for the fifth question, DeepSeek and ChatGPT show a big difference in the obtained accuracy. Overall, results indicate that the average accuracy of all questions for DeepSeek, ChatGPT, and Gemini is 55.44 %, 62.06 %, and 57.44 %, respectively.

Figure (9) illustrates the generation speed of each model for the remote sensing topic. According to the findings, the generation speed of DeepSeek for all questions is low. Conversely, Gemini indicates the highest generation speed among all five questions when compared to the other models. The average generation speeds for DeepSeek, ChatGPT, and Gemini are 8.48 (word/second), 17.03 (word/second), and 96.14 (word/second), respectively.

3.3. Flood modeling

Five questions were designed for flood modeling, and their responses were generated using different LLMs. The questions and their interpretations are presented below. It should be noted that before each question, a general view related to the questions and the role-based

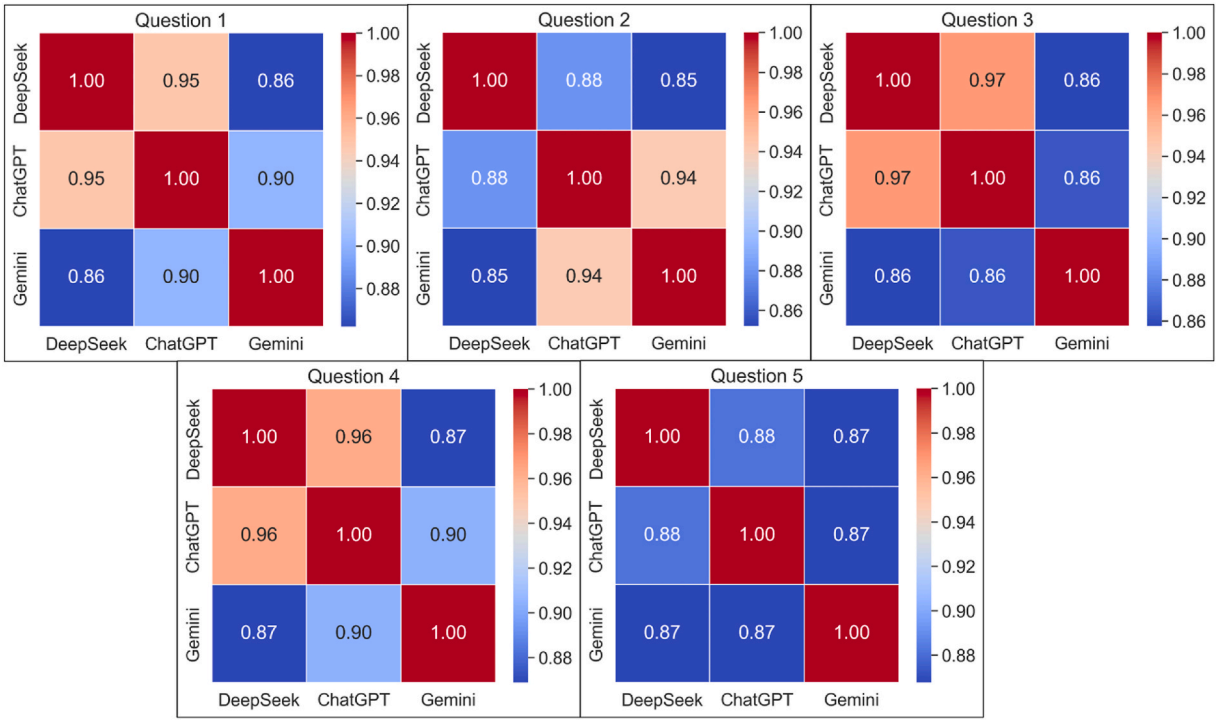

Fig. 4. Similarity of LLMs' responses with each other for the machine learning and optimization topic.

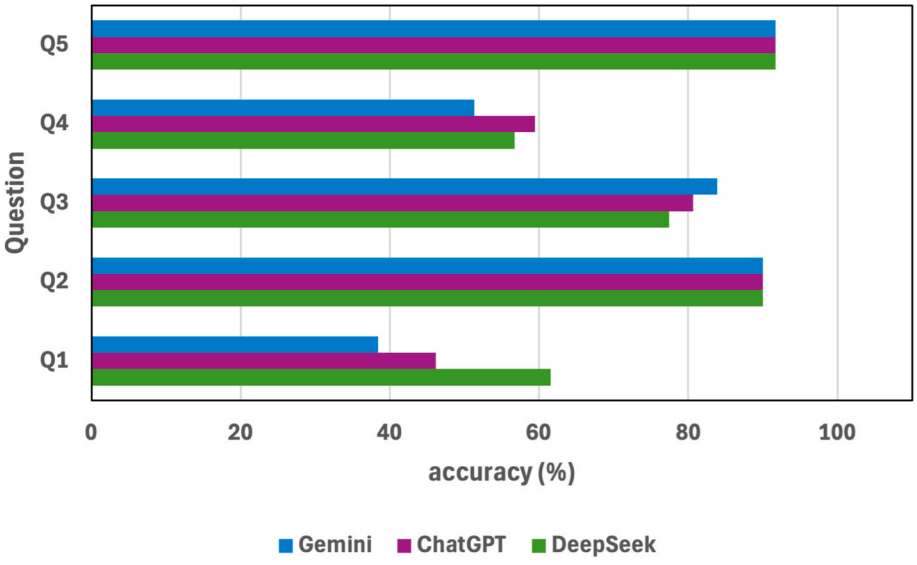

Fig. 5. Accuracy of DeepSeek, ChatGPT, and Gemini for questions in the machine learning and optimization topic using the proposed word matching algorithm.

prompt was fed to the LLMs. The prompt, questions (Dhawale et al., 2024; Ferrario et al., 2024; Pourzangbar et al., 2025; Zhang et al., 2024), and response to each question are presented in Appendix (C).

Across the five questions, DeepSeek, ChatGPT, and Gemini exhibit distinct strengths and limitations in discussing flood and drought indices, urban NbS, flood impact categorization, and machine learning-based FSM. DeepSeek provides highly technical and computationally advanced responses, integrating AI, machine learning, and hybrid modeling, often exceeding the benchmark response in-depth and novelty. However, it sometimes lacks structured comparisons or explicit discussions on certain aspects of the benchmark response. ChatGPT consistently delivers structured, methodologically sound, and practically applicable responses. While it introduces novel methodologies compared to the benchmark response, it occasionally falls short in addressing sociocultural aspects and cutting-edge AI techniques. Gemini, on the other hand, excels in bridging hydrology with real-world applications, incorporating uncertainty quantification and observational techniques.

In Table (6), the results of the overall scores for each question in the flood modeling topic are presented. According to the findings, DeepSeek demonstrated superior performance in answering the first and third questions, while ChatGPT performed better on the fifth question. On the other hand, Gemini showed a higher overall score in answering the fourth question. ChatGPT and Gemini exhibited the same performance for the second question. Overall, the summation of scores shows that the ranking of LLMs in responding to the flood modeling questions is

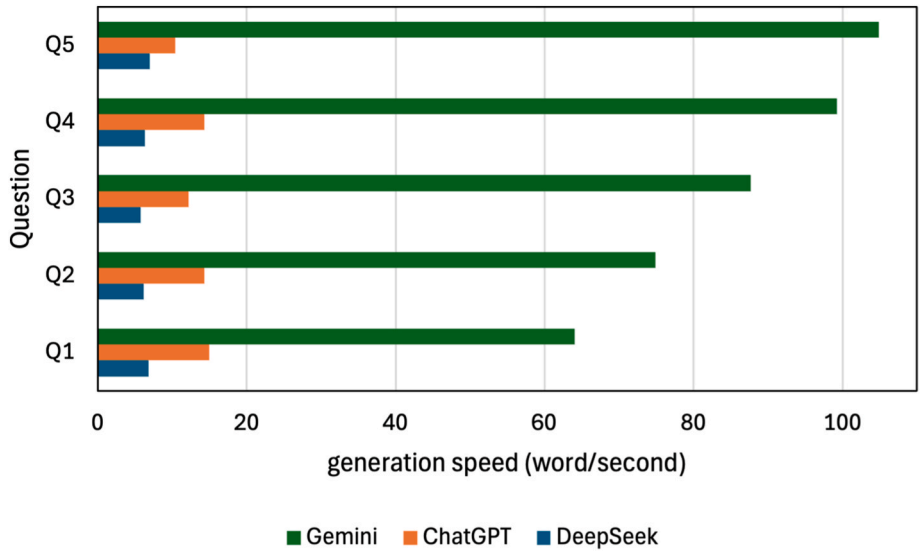

Fig. 6. Generation speed of DeepSeek, ChatGPT, and Gemini for questions in the machine learning and optimization topic.

Table 5
Overall score of LLMs in the remote sensing topic.

Question	Model		
	DeepSeek	ChatGPT	Gemini
Q1	4,20	3,95	3,65
Q2	3,40	3,55	3,20
Q3	3,45	3,15	2,80
Q4	3,30	4,10	3,40
Q5	3,48	3,56	3,15
Sum	17,83	18,31	16,20

DeepSeek, ChatGPT, and Gemini, respectively.

In Figure (10), the results of similarity analysis related to the flood modeling topic are presented. Based on the results, all LLMs exhibit a relatively strong level of similarity in their pairwise responses. For all questions, responses from Gemini and ChatGPT show the highest similarity. Conversely, in four questions, DeepSeek and Gemini have the lowest similarity. In addition, the highest correlation is related to the responses to the third question for Gemini and ChatGPT, with a value of 0.94. On the other hand, the lowest correlation is related to the responses to the first question for DeepSeek and Gemini, with a value of 0.64.

The results of the accuracy analysis for the flood modeling topic are shown in Figure (11). The accuracy comparison of DeepSeek, ChatGPT, and Gemini demonstrates that all three models achieved better accuracy for the first, third, and fifth questions compared to the second and fourth

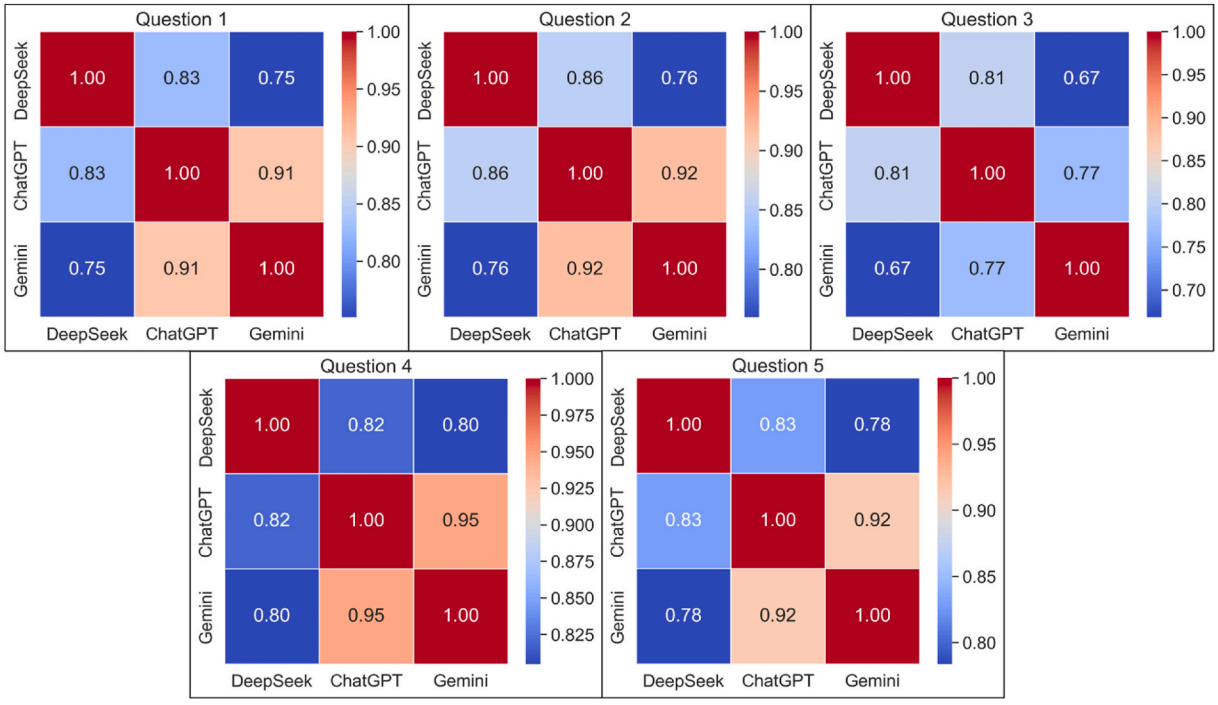

Fig. 7. Similarity of LLMs' responses with each other for the remote sensing topic.

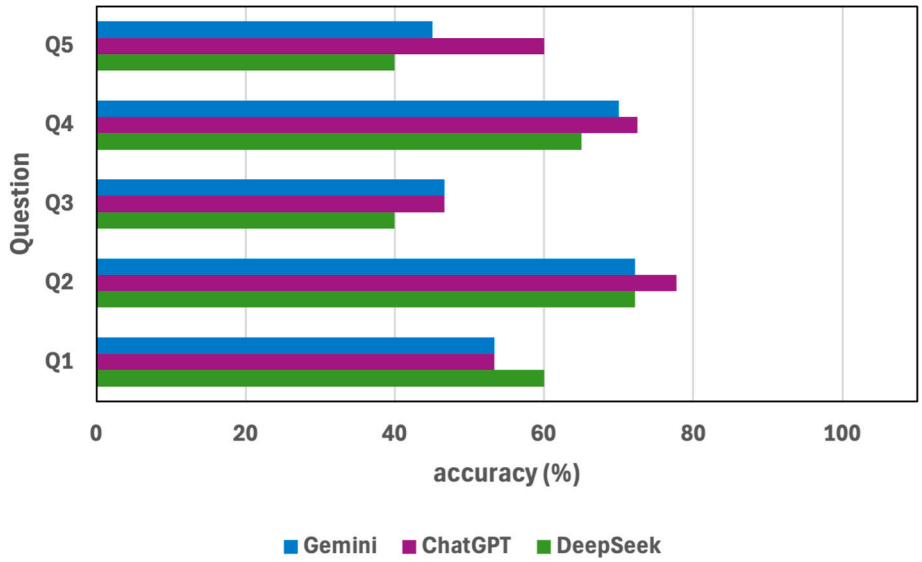

Fig. 8. Accuracy of DeepSeek, ChatGPT, and Gemini for questions in the remote sensing topic using the proposed word matching algorithm.

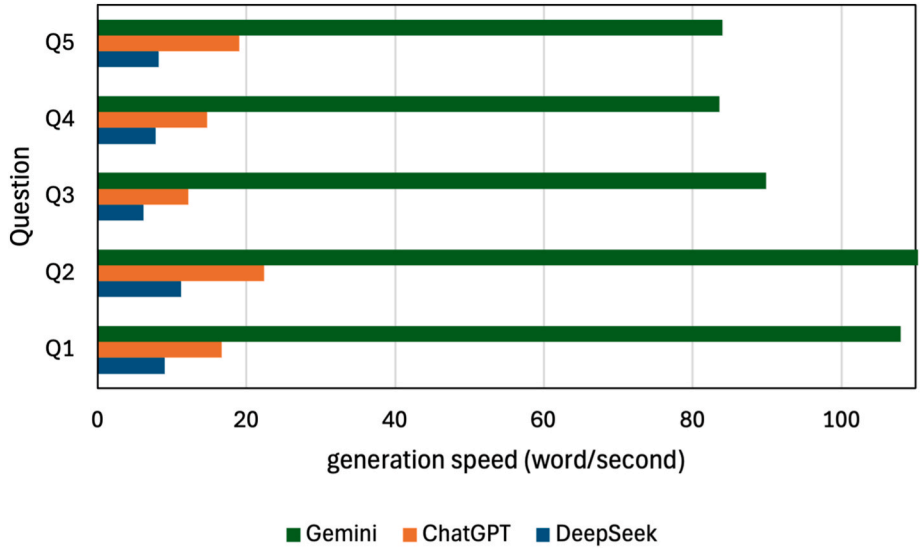

Fig. 9. Generation speed of DeepSeek, ChatGPT, and Gemini for questions in the remote sensing topic.

Table 6
Overall score of LLMs in the flood modeling topic.

Question	Model		
	DeepSeek	ChatGPT	Gemini
Q1	4,10	3,75	3,55
Q2	3,75	3,90	3,90
Q3	4,10	3,90	3,90
Q4	3,75	3,20	4,10
Q5	3,90	4,45	3,75
Sum	19,60	19,20	19,20

questions. In responding to all questions, DeepSeek, ChatGPT, and Gemini show relatively close accuracy; however, for the first question, an evident difference between ChatGPT's and DeepSeek's accuracy is observed. Overall, the findings indicate that the average accuracy of all questions for DeepSeek, ChatGPT, and Gemini is 76.46 %, 72.29 %, and 74.30 %, respectively.

Figure (12) illustrates the generation speed of each model in the flood modeling topic. Based on the results, a comparison of the generation speed shows that DeepSeek has the lowest and Gemini has the highest generation speed for all questions. The average generation speeds of all five questions related to the flood modeling for DeepSeek, ChatGPT, and Gemini are 7.67 (word/second), 17.51 (word/second), and 93.82 (word/second), respectively.

3.4. Sediment transport

Five questions were designed for sediment transport, and their responses were generated using different LLMs. The questions and their interpretations are presented below. It should be noted that before each question, a general view related to the questions and the role-based prompt was fed to the LLMs. The prompt, questions (Afan et al., 2016; van Wiechen et al., 2023; Zhao, 2022), and response to each question are presented in Appendix (D).

Across all five questions in the sediment transport topic, DeepSeek, ChatGPT, and Gemini show capabilities in their responses. DeepSeek showed highly technical, structured, and mathematically rigorous

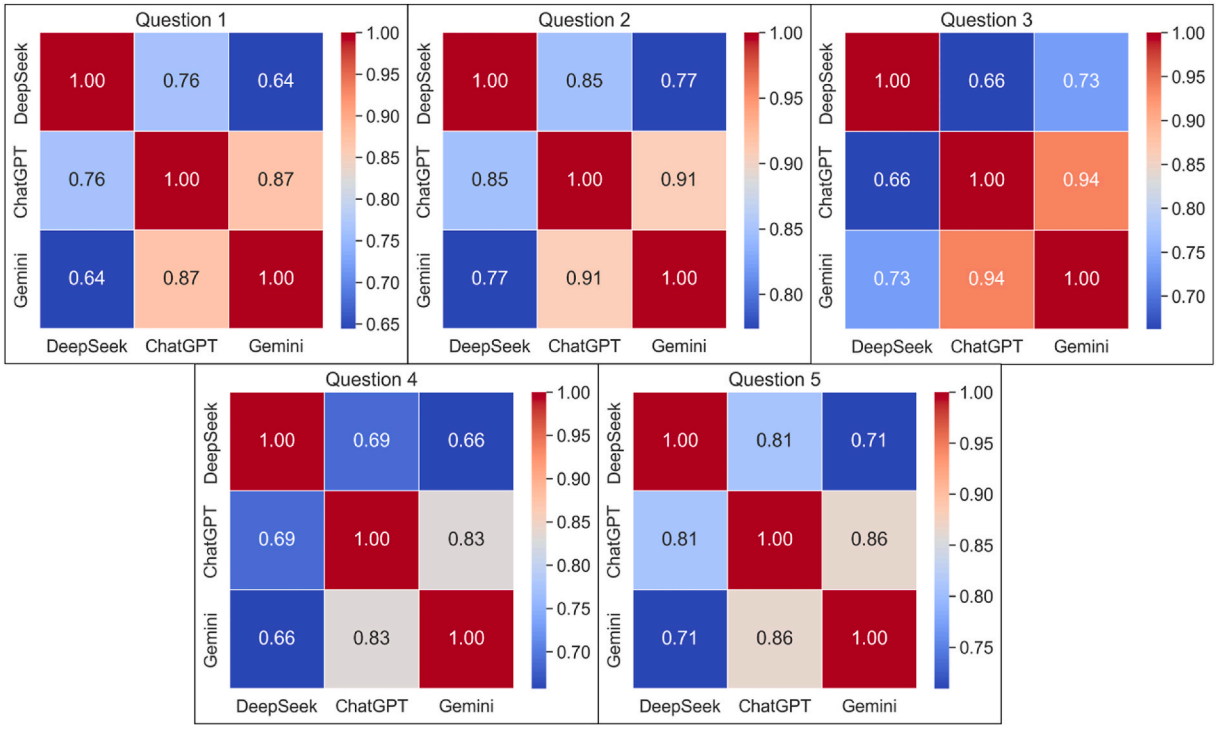

Fig. 10. Similarity of LLMs' responses with each other for the flood modeling topic.

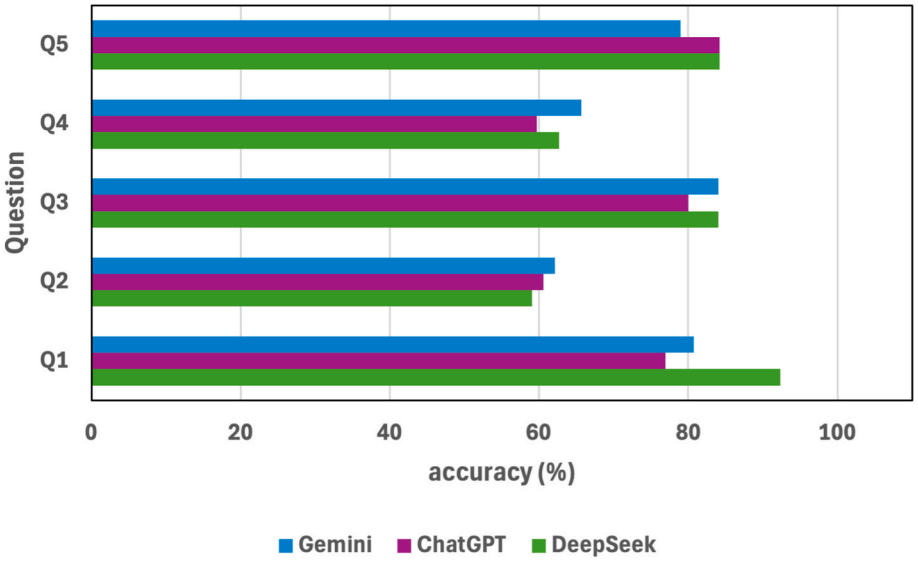

Fig. 11. Accuracy of DeepSeek, ChatGPT, and Gemini for questions in the flood modeling topic using the proposed word matching algorithm.

responses, excelling in hybrid modeling, HPC applications, and integration. It is ideal for computational hydrodynamics researchers. On the other hand, ChatGPT balances technical accuracy with structured explanations and emphasizes engineering applications, cost-efficiency, and real-world adaptability. Finally, Gemini prioritizes field applicability, sustainability, and observational approaches, incorporating stakeholder considerations, bio-geomorphology, and sensor-based advancements. While all models in responding to the sediment transport questions recognize the need for hybrid methodologies, DeepSeek focuses on computational innovation, ChatGPT refines practical model implementation, and Gemini enhances real-world integration with uncertainty quantification and novel instrumentation.

In Table (7), the results of the overall scores for each question in the sediment transport topic are presented. According to the findings, DeepSeek demonstrated better performance in answering the first, third, and fifth questions. On the other hand, Gemini showed a higher overall score in answering the second and fourth questions. In addition, ChatGPT and DeepSeek exhibited the same performance for the fourth question. Overall, the summation of scores shows that the ranking of LLMs in responding to the sediment transport questions is DeepSeek, Gemini, and ChatGPT, respectively.

In Figure (13), the results of similarity analysis related to the sediment transport topic are presented. Based on the results, all LLMs exhibit a strong level of similarity in their pairwise responses. In four out of five questions, Gemini and ChatGPT show the highest similarity with each other. Conversely, in all five questions, DeepSeek and Gemini have the

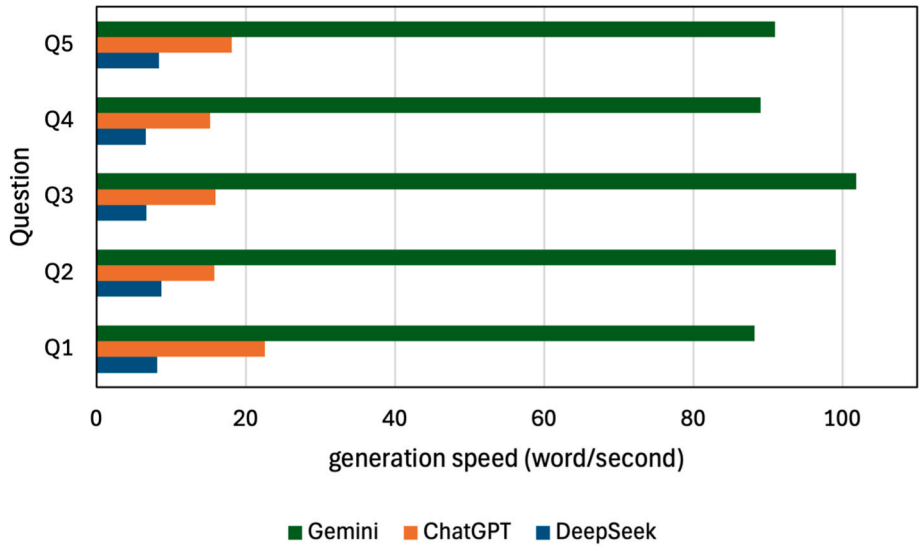

Fig. 12. Generation speed of DeepSeek, ChatGPT, and Gemini for questions in the flood modeling topic.

Table 7
Overall score of LLMs in the sediment transport topic.

Question	Model		
	DeepSeek	ChatGPT	Gemini
Q1	3,35	3,15	3,20
Q2	3,90	3,50	4,10
Q3	3,75	3,05	3,20
Q4	3,80	3,80	4,00
Q5	3,77	3,37	3,68
Sum	18,57	16,87	18,18

lowest similarity in their responses. In addition, the highest correlation is related to the responses to the second question for Gemini and ChatGPT, with a value of 0.94. On the other hand, the lowest correlation is related to the responses to the fourth question for DeepSeek and Gemini, with a value of 0.65.

The results of the accuracy analysis for the flood modeling topic are shown in Figure (14). The accuracy comparison of DeepSeek, ChatGPT, and Gemini demonstrates that all three models achieved better accuracy for the second, fourth, and fifth questions compared to the first and third questions. In responding to all questions, DeepSeek, ChatGPT, and Gemini show relatively close accuracy; however, for the fifth question, a difference between Gemini's and DeepSeek's accuracy is observed. Overall, the findings indicate that the average accuracy of all questions for DeepSeek, ChatGPT, and Gemini is 76.66 %, 79.06 %, and 79.72 %,

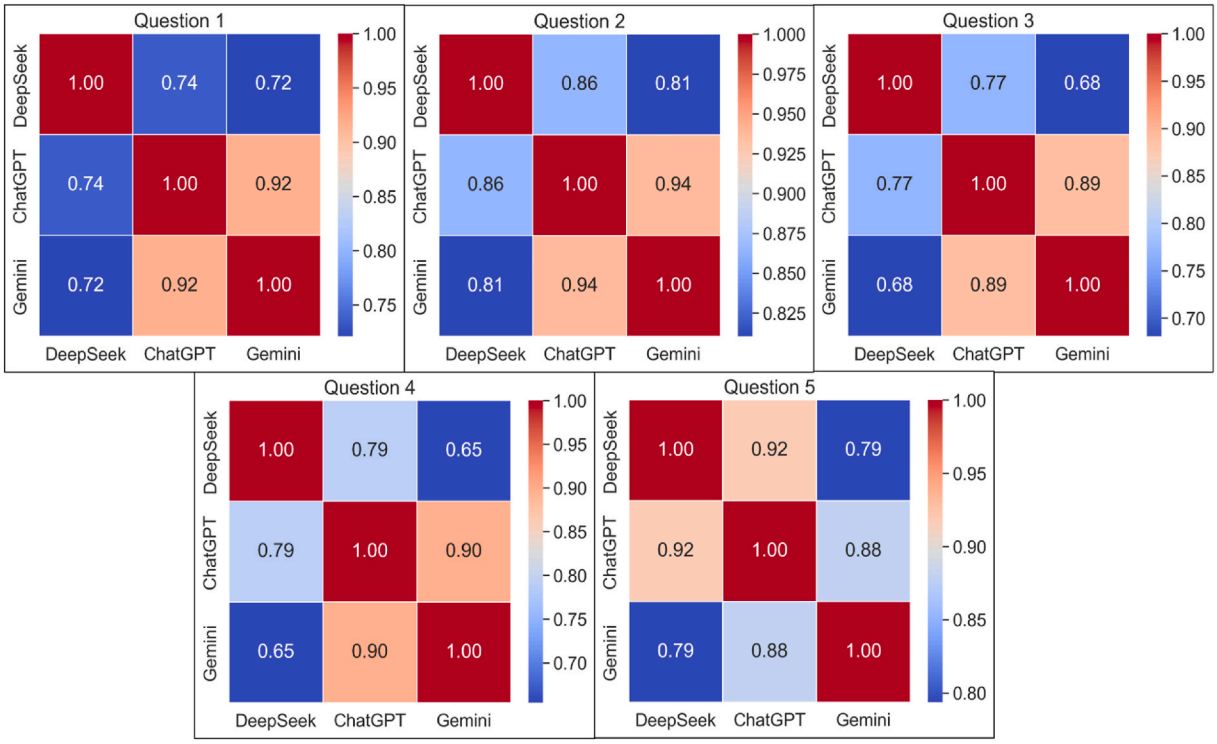

Fig. 13. Similarity of LLMs' responses with each other for the sediment transport topic.

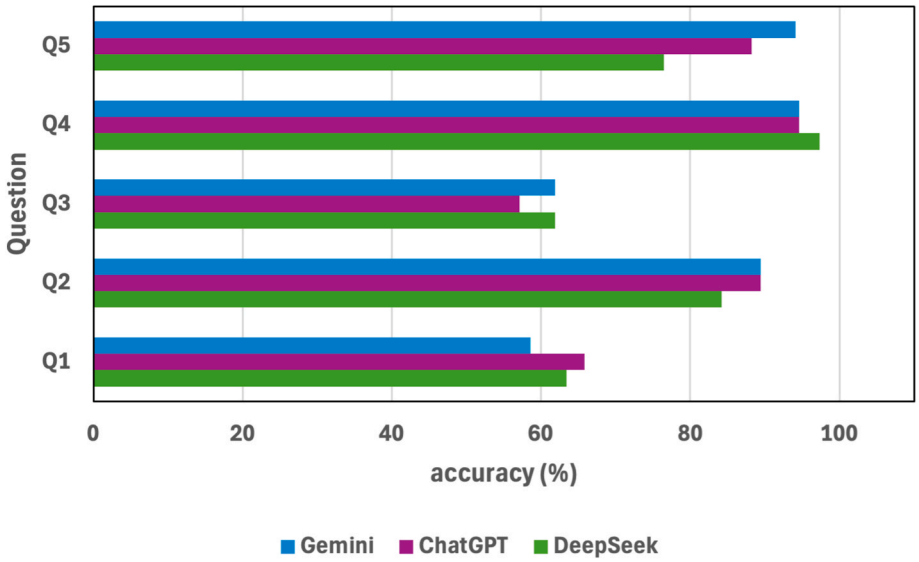

Fig. 14. Accuracy of DeepSeek, ChatGPT, and Gemini for questions in the sediment transport topic using the proposed word matching algorithm.

respectively.

Figure (15) illustrates the generation speed of each model for the sediment transport topic. The results reveal a notable distinction between the generation speeds of DeepSeek and the other models. Conversely, Gemini recorded the slowest generation speed among all five questions when compared to the other models. The average generation speeds of all five questions related to the sediment transport for DeepSeek, ChatGPT, and Gemini are 9.05 (word/second), 24.46 (word/second), and 98.16 (word/second), respectively.

4. Discussion and practical implications

In this research, a novel rubric comprising four criteria (accuracy, relevancy, authenticity, and novelty) was proposed to evaluate selected questions against benchmark responses. The relevancy, authenticity, and novelty scores were assigned by the authors using detailed instructions aligned with the rubric. In contrast, accuracy was assessed using a word-matching algorithm that compared key elements in the benchmark and LLM-generated responses, with these key elements identified by the authors. Notably, all four criteria involve human judgment either directly or in the identification of key components, which underscores the rubric's reliance on expert evaluation to ensure interpretative depth and contextual sensitivity.

The comparison of LLM responses using descriptive criteria and qualitative analysis has been conducted in a few studies. Ren et al. (2024) evaluated the descriptive short-answer questions generated by LLMs based on completeness, truthfulness, and logical consistency, each scored from 0 to 10. Both the generated and benchmark responses were input into GPT-turbo, which was prompted to score them using a one-line, brief set of scoring criteria. WaterGPT outperformed other models, achieving top scores of 8.76 for completeness, 9.39 for truthfulness, and 9.18 for logic. Kizilkaya et al. (2025) assessed open-ended questions by measuring semantic similarity between LLM responses and benchmark answers using cosine similarity. GPT-4o-mini and Llama3:70B achieved the highest scores, demonstrating strong capabilities in producing coherent and detailed answers. Llama3:8B showed slightly lower performance but still captured key information effectively. Meanwhile, Xu et al. (2024) evaluated LLM outputs solely based on accuracy, defined as the inclusion of specific keywords from the benchmark response and the overall meaning of the response.

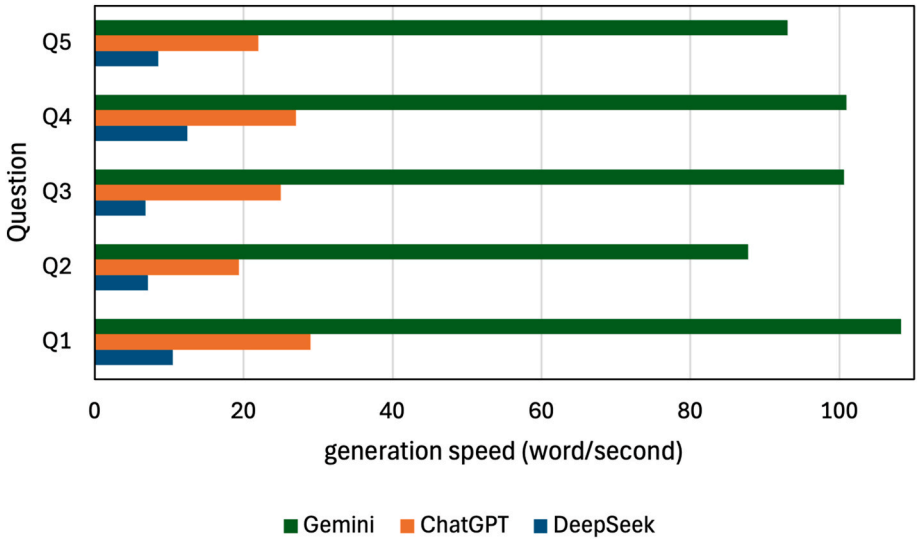

Fig. 15. Generation speed of DeepSeek, ChatGPT, and Gemini for questions in the sediment transport topic.

Compared to the qualitative analyses used in previous studies, the rubric proposed in this paper offers broader and more nuanced evaluation criteria, providing a more comprehensive framework for capturing the quality of responses. The rubric involves a well-defined definition of the criteria, scoring framework, and presence of custom weights assigned to each criterion, reflecting a realistic prioritization for evaluating LLMs. Inclusion of four criteria, accuracy, relevancy, authenticity, and novelty, helps to capture various aspects of the generated responses by LLMs. Moreover, the integration of expert judgment enhances interpretative depth and contextual sensitivity beyond what automated scoring can provide (Van Cauwenbergh et al., 2008), but it may also introduce potential biases and subjectivity, as acknowledged in the limitations of this study. This approach is particularly prevalent in performance assessments, where human raters are essential for evaluating complex competencies and ensuring the validity of the assessments (Huang, 2025; Jonsson and Svingby, 2007; Kan and Bulut, 2015; Kim, 2015; Tekin, 2023). In water science, expert judgment is widely incorporated into various methods. For instance, the Analytic Hierarchy Process (AHP), Multi-Criteria Decision Making (MCDM), and Weighted Linear Combination are among the methods that benefit from expert knowledge to assign weights to conditioning factors (Benítez et al., 2011; Golfam et al., 2019; Mao et al., 2019). In these approaches, the relative importance of factors is determined by experts rather than purely by statistics. Therefore, the involvement of expert judgment enhances the reliability of the LLM responses evaluation and helps to address the limitations of automated metrics (such as those used by LLMs) and provides a more nuanced evaluation of model outputs (Chiang and Lee, 2023; Szymanski et al., 2025; Wu et al., 2025). It is important to note that since the evaluation method applied in this study is a blind assessment, meaning that scoring was performed without knowledge of the LLMs' identities, the potential biases do not undermine the performance of a special LLM model.

Findings regarding the averaged accuracy values across all questions of different fields indicate that DeepSeek and Gemini achieved the close level of accuracy in machine learning and optimization, and flood modeling topics. ChatGPT attained the highest accuracy in remote sensing. Finally, for sediment transport, all LLMs exhibited the close level of accuracy. Based on the proposed algorithm for calculating accuracy, a high score in this criterion indicates that benchmark responses were well-aligned with the LLMs' training data. On the other hand, responses with low accuracy show that increased complexity or ambiguity, particularly with words distant from the benchmark response, can cause this issue.

Another criterion used in this research was novelty. Novelty refers to the introduction of new, original, and distinctive elements that set something apart from existing ideas, designs, or approaches. Studies on LLM capabilities have shown these models can generate novel responses (Lin et al., 2024; Si et al., 2024; Zhang et al., 2025). Therefore, this criterion was included both to assess the novelty of the generated responses and to address a gap left unexamined in previous related studies. As mentioned earlier, the level of novelty was scored by authors, and that is because inclusion of the expert's opinion helps to assess complex and subjective criteria, such as novelty, which are difficult to measure using automated metrics alone (Huang, 2025; Szymanski et al., 2025). The aim of proposing and using novelty was to determine how well LLMs generate innovative responses and solutions beyond the benchmark response. In this regard, for machine learning and optimization, DeepSeek showed a higher level of novelty compared to ChatGPT and Gemini. This pattern was also observed for remote sensing and flood modeling. For sediment and transport, Gemini generated more novel responses compared to other LLMs. Overall, the models mentioned in certain fields can give researchers in hydrology and water science better, innovative ideas for their questions.

In this research, the Spacy library was used to assess the similarity of LLM-generated responses. The findings show that DeepSeek and ChatGPT have a high level of similarity in the machine learning and optimization topics. On the other hand, Gemini and ChatGPT showed high similarity in responses to remote sensing, flood modeling, and sediment transport topics. In addition, across 95 % of questions, DeepSeek and Gemini showed the lowest similarity with each other, while ChatGPT and Gemini, across 70 % of questions, showed the highest similarity in their responses. It is worth noting that for most questions where the two models received the same overall score, the similarity of their LLM was the highest compared to other pairwise models. However, this pattern was not generalizable to all questions with the same overall score.

Another evaluation used in this research was the generation speed of LLMs, which, to our best knowledge, has not been investigated in any of the previous related studies in this field. Based on the findings, using Gemini to achieve responses "faster" is recommended. Across all topics and questions, Gemini responded in a shorter time to the provided questions compared to DeepSeek and ChatGPT. Using Gemini would be suitable for researchers in water science who want to find their answers quickly. On the other hand, DeepSeek, due to its reasoning ability, showed the lowest speed for generating responses across all topics and questions. Therefore, if a response raised from a reasoning approach is needed and enough time is available, DeepSeek is recommended.

Finally, an overview of the responses obtained from the researched LLMs indicates that using DeepSeek is advisable for achieving more technical responses that may interest researchers and experts seeking detailed information. ChatGPT responses demonstrate that when there is a demand for the interdisciplinary integration of various methods, its inputs are useful and reliable. The responses from Gemini indicate that this LLM can be advantageous if practical solutions and real-world applicability are required.

5. Limitations and future directions

This study has a few limitations that present opportunities to be investigated for future research. One limitation is the selection of the review articles. While in this paper, review articles were selected because of their synthesis of knowledge in each topic, based on a curated pool of candidates screened by journal reputation, citation impact, and manual review of the content, this approach may inadvertently reflect the biases or thematic priorities of individual authors. Moreover, the formulation of the questions relied in part on the authors' expert judgment, which, while informed, introduces an element of subjectivity. Therefore, future research would benefit from a framework for guiding paper selection and question design, where experts provide supervision rather than full reliance on their judgment, thereby reducing subjectivity and enhancing consistency.

Beyond the selection of articles and the design of questions, the scoring process of the proposed rubric involves rating generated responses by experts. However, the rubric is equipped with a well-defined definition of the criteria and also a detailed scoring framework; interpretations of the criteria may vary across evaluators. As a practical solution, asking a group of experts to score the responses helps to reduce this bias. However, expert judgment inevitably involves a degree of subjectivity.

Regarding the proposed rubric, a rigid scoring boundary was proposed for accuracy in this research. For instance, a slight difference in the percentage of matched words, 59 % and 61 %, leads to an accuracy score of 2 and 3, respectively. In this case, however, the word matching percentages do not have much of a difference; the scoring approach causes a one-score difference. Improvements to this criterion can enhance its ability to accurately capture score differences, thereby increasing the effectiveness of evaluation methods in future work.

The structure of the questions (for example, asking for the order of challenges based on their importance from LLM instead of just asking to name the challenges) can influence the responses and affect the overall evaluation. However, the primary objective of this study was to assess model performance based on benchmark responses rather than

variations in question design. Therefore, while question sequencing may introduce some variability, the focus remained on comparing how well each model aligned with the benchmark answers. This variability in designing questions and assessing LLM responses can be a potential area for future research.

Regarding the use of LLMs, in this research, the official websites of each LLM and their free version were used to generate responses. For DeepSeek, in some instances, when generating a response, the message displayed was: "*The server is busy. Please try again later.*" This shows that in the case of an overload traffic of the website due to the users' requests, using this model might cause delays in responding. Additionally, utilizing the reasoning capability of DeepSeek prolonged the response time. It should be noted that, generally, the generation speed of responses depends on the server's processing power and internet bandwidth; in this research, a stable internet connection with a speed of 300 Mbps was used. In fact, internet conditions influence access speed but not the content of the generated response, which is derived from pre-trained LLMs. Over time, as training data evolves, LLM responses will likely change; thus, our focus was on assessing reliability against benchmark responses, with improvements expected over time. Moreover, the manual assessment of response generation time introduces potential inefficiencies, and it can be improved by adopting local LLMs rather than web-based tools, which could streamline this process.

To assess the semantic similarity among the responses generated by different LLMs, we employed the spaCy library, which offers an efficient method for comparing textual content. However, it is important to recognize the inherent limitations of this approach, particularly in specialized domains. Notably, spaCy assigns zero vectors to out-of-vocabulary (OOV) terms, words not included in its pre-trained language model, thereby excluding them from similarity calculations and potentially leading to incomplete assessments (Kandi, 2018). Moreover, spaCy relies on static word embeddings, which lack contextual nuance and are often insufficient for accurately capturing the meaning of domain-specific terminology (Honnibal et al., 2020). Given the technical and specialized nature of our study in hydrology and water science, these limitations made spaCy unsuitable for evaluating accuracy within our rubric framework, though it remained useful for general semantic comparisons among LLM outputs.

It is important to note that differences in LLMs' responses naturally stem from their training datasets and the architecture of each model. Additionally, prompts influence LLM's responses (Naveed et al., 2024). In this research, these factors were also considered. In other words, to minimize potential biases in the analysis, a uniform prompt was used across LLMs for each benchmark question. Nevertheless, it is crucial to recognize that LLM responses are somewhat time-dependent. In fact, asking a question at different times may yield varying responses; however, although regarding the context, some words may change, the general concept remains consistent, but asking one question several times from LLMs and checking responses in more detail can be a topic for future research.

Finally, it should be noted that advancements in the field of LLMs are occurring rapidly. While developers are introducing new LLMs, the responses of these models to new topics that have not been exposed to them remain questionable. This research was not subject to this issue, as benchmark responses from review articles were available. Otherwise, researchers should be cautious about posing questions on very novel topics to LLMs that have not been trained on that data, since these models may offer unreliable responses. This also opens a new avenue for researchers in hydrology to explore new and unexplored topics that have not been presented to LLMs. However, concerning the current LLMs, unreliable responses are still possible; as mentioned on the Gemini website, "*Gemini can make mistakes, including about people, so reevaluate their answers*."

6. Conclusions

Large Language Models (LLMs) have garnered significant attention in recent years, and their applications across various fields intrigue researchers to apply them. This study investigated the efficiency of DeepSeek R1, ChatGPT4-o, and Gemini 2 in generating responses to benchmark questions in the domains of machine learning and optimization, remote sensing, flood modeling, and sediment transport. Benchmark questions were selected from review articles, and to evaluate the LLM responses, a novel rubric was proposed for rating the models. Overall, the evaluation framework consisted of a novel scoring rubric, word-matching algorithm, similarity checking, and generation speed analysis.

The findings of this research showed that each LLM has its own capabilities and deficiencies. Based on the proposed rubric, in response to machine learning and optimization, flood modeling, and sediment transport, DeepSeek achieved higher overall scores compared to other LLMs. For the remote sensing topic, ChatGPT showed a higher overall score in responding to the questions. Moreover, concerning novelty, DeepSeek demonstrated more novel responses in machine learning and optimization, remote sensing, and flood modeling. For sediment transport, Gemini exhibited a higher level of novelty in its responses. Moreover, Gemini demonstrated the highest generation speed, while DeepSeek exhibited the lowest due to its reasoning ability across all questions. Finally, the generated responses demonstrate the ability to go beyond the benchmark answers by highlighting points mentioned in other scientific publications, which show the capabilities of LLMs.

Overall, this research provided a comprehensive analysis of LLMs' responses and demonstrated each LLM's capability in addressing questions related to hydrology and water science. The application of the proposed rubric to evaluate responses generated by other LLMs and across broader topics is suggested for future work. Further suggestions were also offered to introduce new aspects to enhance this research.

CRediT authorship contribution statement

Seyed Hossein Hosseini: Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Conceptualization. **Ali Pourzangbar:** Writing – review & editing, Validation, Methodology, Formal analysis, Conceptualization.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT in order to improve grammar and spelling. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding

This study is not funded.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors would like to express their gratitude to the reviewers for their insightful and constructive comments.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envsoft.2025.106772.

Data availability

Data will be made available on request.

References

Afan, H.A., El-shafie, A., Mohtar, W.H.M.W., Yaseen, Z.M., 2016. Past, present and prospect of an artificial intelligence (AI) based model for sediment transport prediction. J. Hydrol. 541, 902–913. https://doi.org/10.1016/j.jhydrol.2016.07.048.

Akbar, M.A., Khan, A.A., Liang, P., 2025. Ethical aspects of ChatGPT in software engineering research. IEEE Transactions on Artificial Intelligence 6 (2), 254–267. https://doi.org/10.1109/TAI.2023.3318183. IEEE Transactions on Artificial Intelligence.

Aluga, M., 2023. Application of CHATGPT in civil engineering. East African Journal of Engineering 6 (1). https://doi.org/10.37284/eaje.6.1.1272. Article 1.

Andrade, H.G., 2005. Teaching with rubrics: the good, the bad, and the ugly. Coll. Teach. 53 (1), 27–31. https://doi.org/10.3200/CTCH.53.1.27-31.

Belzner, L., Gabor, T., Wirsing, M., 2024. Large language model assisted software engineering: prospects, challenges, and a case study. In: Steffen, B. (Ed.), Bridging the Gap Between AI and Reality, pp. 355–374. https://doi.org/10.1007/978-3-031-46002-9_23.

Benítez, J., Delgado-Galván, X., Gutiérrez, J.A., Izquierdo, J., 2011. Balancing consistency and expert judgment in AHP. Math. Comput. Model. 54 (7), 1785–1790. https://doi.org/10.1016/j.mcm.2010.12.023.

Bhaga, T.D., Dube, T., Shekede, M.D., Shoko, C., 2020. Impacts of climate variability and drought on surface water resources in Sub-Saharan Africa using remote sensing: a review. Remote Sens. 12 (24). https://doi.org/10.3390/rs12244184. Article 24.

Biswas, S.S., 2023a. Potential use of chat GPT in global warming. Ann. Biomed. Eng. 51 (6), 1126–1127. https://doi.org/10.1007/s10439-023-03171-8.

Biswas, S.S., 2023b. Role of chat GPT in public health. Ann. Biomed. Eng. 51 (5), 868–869. https://doi.org/10.1007/s10439-023-03172-7.

Brookhart, S.M., 2013. How to Create and Use Rubrics for Formative Assessment and Grading. ASCD.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., et al., 2020. Language models are few-shot learners. Proceedings of the 34th International Conference on Neural Information Processing Systems, pp. 1877–1901.

Chen, J., Chen, S., Fu, R., Li, D., Jiang, H., Wang, C., Peng, Y., Jia, K., Hicks, B.J., 2022. Remote sensing big data for water environment monitoring: current status, challenges, and future prospects. Earths Future 10 (2), e2021EF002289. https://doi.org/10.1029/2021EF002289.

Cheng, M.W.T., Yim, I.H.Y., 2024. Examining the use of ChatGPT in public universities in Hong Kong: a case study of restricted access areas. Discover Education 3 (1), 1. https://doi.org/10.1007/s44217-023-00081-8.

Chiang, C.-H., Lee, H., 2023. *Can large language models Be an alternative to human evaluations?* (No. arXiv:2305.01937). arXiv. https://doi.org/10.48550/arXiv.2305.01937.

Dhawale, R., Schuster-Wallace, C.J., Pietroniro, A., 2024. Assessing the multidimensional nature of flood and drought vulnerability index: a systematic review of literature. Int. J. Disaster Risk Reduct. 112, 104764. https://doi.org/10.1016/j.ijdrr.2024.104764.

Ferrario, F., Mourato, J.M., Rodrigues, M.S., Dias, L.F., 2024. Evaluating Nature-based solutions as urban resilience and climate adaptation tools: a meta-analysis of their benefits on heatwaves and floods. Sci. Total Environ. 950, 175179. https://doi.org/10.1016/j.scitotenv.2024.175179.

Foroumandi, E., Moradkhani, H., Sanchez-Vila, X., Singha, K., Castelletti, A., Destouni, G., 2023. ChatGPT in hydrology and Earth sciences: opportunities, prospects, and concerns. Water Resour. Res. 59 (10), e2023WR036288. https://doi.org/10.1029/2023WR036288.

Golfam, P., Ashofteh, P.-S., Rajaee, T., Chu, X., 2019. Prioritization of water allocation for adaptation to climate change using multi-criteria decision making (MCDM). Water Resour. Manag. 33 (10), 3401–3416. https://doi.org/10.1007/s11269-019-02307-7.

Guo, H., Su, X., Wu, C., Du, B., Zhang, L., Li, D., 2024. Remote sensing ChatGPT: solving remote sensing tasks with ChatGPT and visual models. IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium 11474–11478. https://doi.org/10.1109/IGARSS53475.2024.10640736.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z.F., Gou, Z., Shao, Z., Li, Z., Gao, Z., et al., 2025. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. arXiv. https://doi.org/10.48550/arXiv.2501.12948.

Haider, S., Rashid, M., Tariq, M.A.U.R., Nadeem, A., 2024. The role of artificial intelligence (AI) and chatgpt in water resources, including its potential benefits and associated challenges. Discover Water 4 (1), 113. https://doi.org/10.1007/s43832-024-00173-y.

Honnibal, M., 2017. SpaCy 2: natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. https://cir.nii.ac.jp/crid/1370021390573874949.

Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A., 2020. Spacy: Industrial-Strength Natural Language Processing in Python.

Hu, Y., Mai, G., Cundy, C., Choi, K., Lao, N., Liu, W., Lakhanpal, G., Zhou, R.Z., Joseph, K., 2023. Geo-knowledge-guided GPT models improve the extraction of location descriptions from disaster-related social media messages. Int. J. Geogr. Inf. Sci. 37 (11), 2289–2318. https://doi.org/10.1080/13658816.2023.2266495.

Huang, H.-Y., 2025. Understanding rater cognition in performance assessment: a mixed IRTree approach. Appl. Psychol. Meas. https://doi.org/10.1177/01466216251333578, 01466216251333578.

Irvine, D., Halloran, L., Brunner, P., 2023. Opportunities and limitations of the ChatGPT advanced data analysis plugin for hydrological analyses. Hydrol. Process. 37. https://doi.org/10.1002/hyp.15015.

Jayaraman, P., Nagarajan, K.K., Partheeban, P., Krishnamurthy, V., 2024. Critical review on water quality analysis using IoT and machine learning models. International Journal of Information Management Data Insights 4 (1), 100210. https://doi.org/10.1016/j.jjimei.2023.100210.

Jonsson, A., Svingby, G., 2007. The use of scoring rubrics: reliability, validity and educational consequences. Educ. Res. Rev. 2 (2), 130–144. https://doi.org/10.1016/j.edurev.2007.05.002.

Jungwirth, D., Haluza, D., 2023. Feasibility study on utilization of the artificial intelligence GPT-3 in public healt. https://doi.org/10.20944/preprints202301.0521.v1.

Kadiyala, L.A., Mermer, O., Samuel, D.J., Sermet, Y., Demir, I., 2024. The implementation of multimodal large language models for hydrological applications: a comparative study of GPT-4 vision, gemini, LLaVa, and Multimodal-GPT. Hydrology 11 (9). https://doi.org/10.3390/hydrology11090148. Article 9.

Kamyab, H., Khademi, T., Chelliapan, S., SaberiKamarposhti, M., Rezania, S., Yusuf, M., Farajnezhad, M., Abbas, M., Hun Jeon, B., Ahn, Y., 2023. The latest innovative avenues for the utilization of artificial intelligence and big data analytics in water resource management. Results Eng. 20, 101566. https://doi.org/10.1016/j.rineng.2023.101566.

Kan, A., Bulut, O., 2015. Crossed random-effect modeling: examining the effects of teacher experience and rubric use in performance assessments. Eur. J. Educ. Res. 57, 1–28. https://doi.org/10.14689/ejer.2014.57.4.

Kandi, S.M., 2018. Language modelling for handling out-of-vocabulary words in natural language processing (MSc dissertation, London School of Economics and Political Science), pp. 1–48. Retrieved from. https://www.researchgate.net/profile/Shabeel-Meemulla-Kandi/publication/335757797_Language_Modelling_for_Handling_Out-of-Vocabulary_Words_in_Natural_Language_Processing/links/5d7a26a04585151ee4afb0c5/Language-Modelling-for-Handling-Out-of-Vocabulary-Words-in-Natural-Language-Processing.pdf.

Kim, H.J., 2015. A qualitative analysis of rater behavior on an L2 speaking assessment. Lang. Assess. Q. 12 (3), 239–261. https://doi.org/10.1080/15434303.2015.1049353.

Kizilkaya, D., Sajja, R., Sermet, Y., Demir, I., 2025. Toward HydroLLM: a benchmark dataset for hydrology-specific knowledge assessment for large language models. Environmental Data Science 4, e31. https://doi.org/10.1017/eds.2025.10006.

Li, C., Deng, W., Lu, M., Yuan, B., 2025. AtmosSci-Bench: evaluating the recent advance of large language model for atmospheric science. arXiv. https://doi.org/10.48550/arXiv.2502.01159.

Liang, J.T., Badea, C., Bird, C., DeLine, R., Ford, D., Forsgren, N., Zimmermann, T., 2024. Can GPT-4 replicate empirical software engineering research? arXiv. https://doi.org/10.48550/arXiv.2310.01727.

Lin, E., Peng, Z., Fang, Y., 2024. Evaluating and enhancing large language models for novelty assessment in scholarly publications. arXiv. https://doi.org/10.48550/arXiv.2409.16605. No. arXiv:2409.16605).

Mala-Jetmarova, H., Sultanova, N., Savic, D., 2018. Lost in optimisation of water distribution systems? A literature review of system design. Water 10 (3), 3. https://doi.org/10.3390/w10030307.

Mao, F., Zhao, X., Ma, P., Chi, S., Richards, K., Clark, J., Hannah, D.M., Krause, S., 2019. Developing composite indicators for ecological water quality assessment based on network interactions and expert judgment. Environ. Model. Software 115, 51–62. https://doi.org/10.1016/j.envsoft.2019.01.011.

Mercer, S., Spillard, S., Martin, D.P., 2025. Brief analysis of DeepSeek R1 and its implications for generative AI. https://www.semanticscholar.org/paper/Brief-analysis-of-DeepSeek-R1-and-its-implications-Mercer-Spillard/79559a799358905120c7afa8ff7823bfa528b717.

Moskal, B.M., Leydens, J.A., 2000. Scoring rubric development: validity and reliability. Practical Assess. Res. Eval. 7 (1), 1. https://doi.org/10.7275/q7rm-gg74.

Naveed, H., Khan, A.U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., Mian, A., 2024. A comprehensive overview of large language models. arXiv. https://doi.org/10.48550/arXiv.2307.06435.

Neumann, M., King, D., Beltagy, I., Ammar, W., 2019. ScispaCy: fast and robust models for biomedical natural language processing. Proceedings of the 18th Bionlp Workshop and Shared Task, pp. 319–327. https://doi.org/10.18653/v1/W19-5034.

Olson, M.L., Ratzlaff, N., Hinck, M., Luo, M., Yu, S., Xue, C., Lal, V., 2025. Semantic specialization in MoE appears with scale: a study of DeepSeek R1 expert specialization. https://www.semanticscholar.org/paper/Semantic-Specialization-in-MoE-Appears-with-Scale%3A-Olson-Ratzlaff/df4245f61b34ea95a78a258a7458db2e14f78801.

OpenAI, 2024. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/.

Osco, L.P., Lemos, E. L. de, Gonçalves, W.N., Ramos, A.P.M., Marcato Junior, J., 2023. The potential of visual ChatGPT for remote sensing. Remote Sens. 15 (13), 13. https://doi.org/10.3390/rs15133232.

Ozkaya, I., 2023. Application of large language models to software engineering tasks: opportunities, risks, and implications. IEEE Software 40 (3), 4–8. https://doi.org/10.1109/MS.2023.3248401. IEEE Software.

Paiva, L. F. de, Luijten, G., Puladi, B., Egger, J., 2025. How does DeepSeek-R1 perform on USMLE? (p. 2025.02.06.25321749). medRxiv. https://doi.org/10.1101/2025.02.06.25321749.

Parray, A.A., Inam, Z.M., Ramonfaur, D., Haider, S.S., Mistry, S.K., Pandya, A.K., 2023. ChatGPT and global public health: applications, challenges, ethical considerations and mitigation strategies. Global Transitions 5, 50–54. https://doi.org/10.1016/j.glt.2023.05.001.

Pourzangbar, A., Oberle, P., Kron, A., Franca, M.J., 2025. Analysis of the utilization of machine learning to map flood susceptibility. Journal of Flood Risk Management 18 (2), e70042. https://doi.org/10.1111/jfr3.70042.

Pursnani, V., Ramirez, C.E., Sermet, M.Y., Demir, I., 2024. HydroSuite-AI: facilitating hydrological research with LLM-driven code assistance. https://eartharxiv.org/repository/view/8121/.

Ren, Y., Zhang, T., Dong, X., Li, W., Wang, Z., He, J., Zhang, H., Jiao, L., 2024. WaterGPT: training a large language model to become a hydrology expert. Water 16 (21), 21. https://doi.org/10.3390/w16213075.

Reuters, 2023. Top French university bans use of ChatGPT to prevent plagiarism. https://www.reuters.com/technology/top-french-university-bans-use-chatgpt-prevent-plagiarism-2023-01-27/.

Rodríguez-Ibánez, M., Casánez-Ventura, A., Castejón-Mateos, F., Cuenca-Jiménez, P.-M., 2023. A review on sentiment analysis from social media platforms. Expert Syst. Appl. 223, 119862. https://doi.org/10.1016/j.eswa.2023.119862.

Sadıkoğlu, E., Gök, M., Mıjwıl, M.M., Kösesoy, İ., 2023. The evolution and impact of large language model chatbots in social media: a comprehensive review of past, present, and future applications. Veri Bilimi 6 (2), 2.

Sagan, V., Peterson, K.T., Maimaitijiang, M., Sidike, P., Sloan, J., Greeling, B.A., Maalouf, S., Adams, C., 2020. Monitoring inland water quality using remote sensing: potential and limitations of spectral indices, bio-optical simulations, machine learning, and cloud computing. Earth Sci. Rev. 205, 103187. https://doi.org/10.1016/j.earscirev.2020.103187.

Sallam, M., Al-Mahzoum, K., Sallam, M., Mijwil, M., 2025. DeepSeek: is it the end of generative AI monopoly or the mark of the impending doomsday? Mesopotamian Journal of Big Data 2025, 26–34. https://doi.org/10.58496/MJBD/2025/002.

Si, C., Yang, D., Hashimoto, T., 2024. *Can LLMs generate novel research ideas? A large-scale human study with 100+ NLP researchers* (No. arXiv:2409.04109). arXiv. https://doi.org/10.48550/arXiv.2409.04109.

Sun, Y., Wang, D., Li, L., Ning, R., Yu, S., Gao, N., 2024. Application of remote sensing technology in water quality monitoring: from traditional approaches to artificial intelligence. Water Res. 267, 122546. https://doi.org/10.1016/j.watres.2024.122546.

Sundar Pichai, 2023. An Important next Step on our AI Journey. Google. https://blog.google/technology/ai/bard-google-ai-search-updates/.

Szymanski, A., Ziems, N., Eicher-Miller, H.A., Li, T.J.-J., Jiang, M., Metoyer, R.A., 2025. Limitations of the LLM-as-a-Judge approach for evaluating LLM outputs in expert knowledge tasks. Proceedings of the 30th International Conference on Intelligent User Interfaces, pp. 952–966. https://doi.org/10.1145/3708359.3712091.

Tansar, H., Li, F., Zheng, F., Duan, H.-F., 2024. A critical review on optimization and implementation of green-grey infrastructures for sustainable urban stormwater management. AQUA - Water Infrastructure, Ecosystems and Society 73 (6), 1135–1150. https://doi.org/10.2166/aqua.2024.310.

Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., Silver, D., Johnson, M., Antonoglou, I., Schrittwieser, J., Glaese, A., Chen, J., Pitler, E., Lillicrap, T., Lazaridou, A., et al., 2024. Gemini: a family of highly capable multimodal models. arXiv. https://doi.org/10.48550/arXiv.2312.11805.

Tekin, M., 2023. Pedagogical potential and didactic limitations of assessment rubrics: an example from medical education. In: Improving Learning Through Assessment Rubrics: Student Awareness of what and How they Learn. IGI Global, pp. 300–313. https://library.oapen.org/bitstream/handle/20.500.12657/88267/1/9781668460870.pdf#page=333.

Tian, H., Lu, W., Li, T.O., Tang, X., Cheung, S.-C., Klein, J., Bissyandé, T.F., 2023. Is ChatGPT the ultimate programming assistant—how far is it? arXiv. https://doi.org/10.48550/arXiv.2304.11938.

Van Cauwenbergh, N., Pinte, D., Tilmant, A., Frances, I., Pulido-Bosch, A., Vanclooster, M., 2008. Multi-objective, multiple participant decision support for water management in the andarax catchment, Almeria. Environ. Geol. 54 (3), 479–489. https://doi.org/10.1007/s00254-007-0847-y.

van Wiechen, P.P.J., de Vries, S., Reniers, A.J.H.M., Aarninkhof, S.G.J., 2023. Dune erosion during storm surges: a review of the observations, physics and modelling of the collision regime. Coast. Eng. 186, 104383. https://doi.org/10.1016/j.coastaleng.2023.104383.

Wu, W., Zhang, C., Zhao, Y., 2025. Automated novelty evaluation of academic paper: a collaborative approach integrating human and large language model knowledge. Journal of the Association for Information Science and Technology. https://doi.org/10.1002/asi.70005. *n/a*(n/a).

Xia, C., Yue, L., Chen, D., Li, Y., Yang, H., Xue, A., Li, Z., He, Q., Zhang, G., Kattel, D.B., Lei, L., Zhou, M., 2025. AI-Driven reinvention of hydrological modeling for accurate predictions and interpretation to transform Earth system modeling. arXiv. https://doi.org/10.48550/arXiv.2501.04733.

Xu, T., Liang, F., 2021. Machine learning for hydrologic sciences: an introductory overview. WIREs Water 8 (5), e1533. https://doi.org/10.1002/wat2.1533.

Xu, B., Wen, L., Li, Z., Yang, Y., Wu, G., Tang, X., Li, Y., Wu, Z., Su, Q., Shi, X., Yang, Y., Tong, R., Ng, H.Y., 2024. *Unlocking the potential: benchmarking large language models in water engineering and research* (No. arXiv:2407.21045). arXiv. https://doi.org/10.48550/arXiv.2407.21045.

Xue, Z., Xu, C., Xu, X., 2023. Application of ChatGPT in natural disaster prevention and reduction. Natural Hazards Research 3 (3), 556–562. https://doi.org/10.1016/j.nhres.2023.07.005.

Yu, H., 2023. Reflection on whether Chat GPT should be banned by academia from the perspective of education and teaching. Front. Psychol. 14. https://doi.org/10.3389/fpsyg.2023.1181712.

Zhang, Y., Li, Z., Xu, H., Ge, W., Qian, H., Li, J., Sun, H., Zhang, H., Jiao, Y., 2024. Impact of floods on the environment: a review of indicators, influencing factors, and evaluation methods. Sci. Total Environ. 951, 175683. https://doi.org/10.1016/j.scitotenv.2024.175683.

Zhang, Y., Diddee, H., Holm, S., Liu, H., Liu, X., Samuel, V., Wang, B., Ippolito, D., 2025. *NoveltyBench: evaluating language models for humanlike diversity* (No. arXiv:2504.05228). arXiv. https://doi.org/10.48550/arXiv.2504.05228.

Zhao, M., 2022. A review on recent development of numerical modelling of local scour around hydraulic and marine structures. J. Mar. Sci. Eng. 10 (8), 8. https://doi.org/10.3390/jmse10081139.

Zhu, J.-J., Jiang, J., Yang, M., Ren, Z.J., 2023. ChatGPT and environmental research. Environ. Sci. Technol. 57 (46), 17667–17670. https://doi.org/10.1021/acs.est.3c01818.