*Environmental Data Science* (2025), 4: e31, 1–18
doi:10.1017/eds.2025.10006

CAMBRIDGE UNIVERSITY PRESS

DATA PAPER

# Toward HydroLLM: a benchmark dataset for hydrology-specific knowledge assessment for large language models

Dilara Kizilkaya[1,2], Ramteja Sajja[1,3], Yusuf Sermet[1] and Ibrahim Demir[4,5]

[1]IIHR - Hydroscience and Engineering, University of Iowa, Iowa City, IA, USA
[2]Computer Science, University of Iowa, Iowa City, IA, USA
[3]Electrical and Computer Engineering, University of Iowa, Iowa City, IA, USA
[4]River-Coastal Science and Engineering, Tulane University, Iowa City, IA, USA
[5]ByWater Institute, Tulane University, New Orleans, LA, USA
**Corresponding author**: Ramteja Sajja; Email: ramteja-sajja@uiowa.edu

**Received:** 15 March 2025; **Revised:** 01 May 2025; **Accepted:** 09 May 2025

**Keywords:** benchmark dataset; domain-specific AI; hydrology; large language models(LLMs); natural language processing (NLP); question generation

## Abstract

The rapid advancement of large language models (LLMs) has enabled their integration into a wide range of scientific disciplines. This article introduces a comprehensive benchmark dataset specifically designed for testing recent LLMs in the hydrology domain. Leveraging a collection of research articles and hydrology textbooks, we generated a wide array of hydrology-specific questions in various formats, including true/false, multiple-choice, open-ended, and fill-in-the-blank. These questions serve as a robust foundation for evaluating the performance of state-of-the-art LLMs, including GPT-4o-mini, Llama3:8B, and Llama3.1:70B, in addressing domain-specific queries. Our evaluation framework employs accuracy metrics for objective question types and cosine similarity measures for subjective responses, ensuring a thorough assessment of the models' proficiency in understanding and responding to hydrological content. The results underscore both the capabilities and limitations of artificial intelligence (AI)-driven tools within this specialized field, providing valuable insights for future research and the development of educational resources. By introducing HydroLLM-Benchmark, this study contributes a vital resource to the growing body of work on domain-specific AI applications, demonstrating the potential of LLMs to support complex, field-specific tasks in hydrology.

## Impact Statement

Our study introduces HydroLLM-Benchmark, the first comprehensive dataset designed to evaluate large language models (LLMs) in hydrology-specific tasks. As artificial intelligence (AI) increasingly supports environmental research, assessing LLMs' ability to process hydrological knowledge is crucial for scientific progress. By benchmarking models like GPT-4o-mini and Llama3, we identify their strengths and limitations in understanding hydrology, informing improvements in AI-driven decision-making for water resource management, climate resilience, and flood prediction. This work bridges the gap between AI and hydrological sciences, ensuring that future LLMs are better equipped for environmental applications. By providing an open-source dataset, we empower researchers to refine AI models, fostering more accurate, data-driven insights for sustainable water management and environmental policy.

This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

## 1. Introduction

Hydrology is a specialized domain characterized by its intricate interplay of physical, chemical, and biological processes, combined with significant societal and environmental implications. Addressing global challenges such as water scarcity, flooding, and sustainable water resource management demands a profound understanding of hydrological processes and their interconnected systems. The study of the water cycle, encompassing precipitation, evaporation, and runoff, is influenced by diverse environmental factors, requiring precision and context-specific knowledge (Ukarande, 2023). In addition, subdisciplines such as groundwater hydrology necessitate a deep understanding of subsurface water physics, which is essential for effective resource management and environmental sustainability (Anderson, 2007).

Despite advances in hydrological science, knowledge gaps persist, particularly in understanding local boundary conditions and hydrological connectivity, which often vary significantly across regions (Wagener et al., 2021). Addressing these gaps requires the development of shared perceptual models to enhance collective understanding and collaboration in hydrological research (Wagener et al., 2020). Moreover, hydrology is inherently interdisciplinary, demanding integration across civil engineering, geology, meteorology, and social sciences to address water resource challenges effectively (Harshbarger and Ferris, 1963). This interdisciplinary nature, combined with the socioeconomic and political factors influencing water-related decision-making, underscores the need for specialized training and knowledge in the field (Harshbarger and Ferris, 1963).

The rapid development and widespread adoption of large language models (LLMs) have opened new avenues for tackling domain-specific challenges in science and engineering. However, applying general-purpose LLMs to hydrology presents significant challenges due to the specialized nature of hydrological data and reasoning tasks (Samuel et al., 2024a). General-purpose LLMs, trained on diverse datasets, often lack domain-specific knowledge required for tasks such as flood management, groundwater modeling, and water quality assessment (Shen et al., 2024). In addition, LLMs face spatial reasoning deficiencies, which are critical for hydrological tasks involving watershed mapping, flood simulation, and water distribution planning (Yan et al., 2023; Vald et al., 2024). Their limitations in spatial reasoning can hinder effective real-time decision-making in dynamic hydrological scenarios (Yan et al., 2023).

Another key challenge is the integration of multimodal data, combining textual, visual, and numerical information core requirement for effective hydrological analysis (Samuel et al., 2024b). While advancements like GPT-4 Vision demonstrate improvements in processing visual data, their performance in multimodal tasks remains inconsistent, highlighting the need for domain-specific fine-tuning (Kadiyala et al., 2024a). Nevertheless, recent studies suggest that targeted fine-tuning and domain-specific adaptations have the potential to enhance LLM performance in hydrology (Xu et al., 2024b).

The integration of AI-driven educational and decision-support systems has demonstrated promising outcomes in specialized domains (Kadiyala et al., 2024b). For instance, AI-enabled intelligent assistants have shown significant potential in personalized and adaptive learning environments by reducing cognitive load, providing targeted knowledge assessments, and generating customized learning pathways (Sajja et al., 2023). These systems offer capabilities such as interactive knowledge discovery, quiz generation, and intelligent tutoring, which can also be adapted to hydrology-specific tasks (Sajja et al., 2024a).

Similarly, conversational AI educational assistants have been successfully deployed in diverse academic domains, including environmental, political, and social sciences, showcasing their effectiveness in delivering course-specific support and fostering deeper engagement with complex datasets (Pursnani et al., 2023; Sajja et al., 2024b). In the context of floodplain management certification, AI-assisted tools have been developed to enhance vocational training, offering interactive question-answering sessions and real-time feedback tailored to certification requirements (Sajja et al., 2025, Pursnani et al., 2024). These applications demonstrate the potential of AI to address specialized learning and professional training needs, highlighting the feasibility of similar frameworks in hydrology.

In parallel, decision-support frameworks such as the multi-hazard tournament system have been employed in flood mitigation and water resource management contexts (Alabbad et al., 2024). These

https://doi.org/10.1017/eds.2025.10006 Published online by Cambridge University Press

frameworks utilize AI agents and collaborative multi-agent interactions to optimize decision-making processes, demonstrating the capability of AI-driven simulations in complex multi-stakeholder environments (Kadiyala et al., 2024c). Such applications underscore the potential for LLMs to support intricate hydrological decision-making tasks when appropriately fine-tuned and adapted.

The need for a hydrology-specific benchmark dataset emerges from the complex and multifaceted nature of hydrological research and water resource management (Ebert-Uphoff et al., 2017). Benchmark datasets serve as standardized tools for evaluating, validating, and improving hydrological models, ensuring consistent performance assessment across diverse tasks (Sit et al., 2021). Existing datasets, such as CAMELS-DE, link landscape attributes with hydrological and meteorological time series, enabling insights into hydrological processes across various landscapes (Dolich et al., 2024). Similarly, the SEN12-WATER dataset integrates multiple data types to analyze water dynamics and drought resilience (Russo et al., 2024).

Benchmark datasets also play a crucial role in hydrological model validation, with datasets like those developed for SAC, GR4J, and SOCONT models enabling consistent and reliable performance assessments (Izquierdo-Horna et al., 2024). In addition, resources such as the Panta Rhei dataset provide paired flood and drought socio-hydrological data, facilitating integrated modeling approaches (Kreibich et al., 2023). However, data gaps persist, especially concerning fine temporal resolution data for groundwater recharge, a gap partially addressed by datasets like RpSy (Malakar et al., 2024). Despite these contributions, challenges related to data accessibility, standardization, and regional coverage continue to limit the effectiveness of existing datasets (Demir et al., 2022; Dolich et al., 2024).

While scientific benchmarks exist across domains, they often fail to address the specific needs of hydrology. Traditional hydrological models frequently suffer from performance degradation when applied across multiple basins, highlighting the regional variability of hydrological conditions (Kratzert et al., 2019). Furthermore, the subjective nature of accuracy determination complicates benchmarking efforts, as model performance expectations often depend on regional characteristics and data quality (Seibert, 2001). Recent advancements, such as long short-term memory networks, demonstrate improved performance in cross-basin hydrological modeling by leveraging large datasets (Kratzert et al., 2019). Similarly, physically based models, like the Variable Infiltration Capacity model, provide a more robust representation of hydrological processes and enable meaningful comparisons across hydroclimate conditions (Newman et al., 2017). However, these approaches still face challenges in addressing the context-specific requirements of hydrological benchmarking (Seibert, 2001).

General-purpose LLMs face limitations not only in hydrology but also across other specialized domains. These limitations include knowledge gaps, terminology inconsistencies, and a lack of domain-specific reasoning capabilities (Chen et al., 2023; Soman and Ranjani, 2024). In addition, issues like knowledge forgetting—where newer knowledge overshadows older, relevant information—complicate their application in specialized tasks (Chen et al., 2023). Evaluation by human experts remains essential, as LLMs often fail to align with nuanced reasoning in specialized fields (Harvel et al., 2024; Szymanski et al., 2024). Safety concerns, including the risk of generating harmful content, further emphasize the importance of balanced fine-tuning methodologies (Thakkar et al., 2024).

The absence of a standardized evaluation dataset for hydrology-focused LLMs exacerbates these challenges. Without a consistent benchmark, evaluating and comparing model performance becomes inherently biased and inconsistent (Zheng et al., 2018). Challenges such as data contamination and a lack of robust evaluation guidelines add further complexity to interpreting benchmark scores (Singh et al., 2024). Emerging domain-specific benchmarks, such as WaterER, highlight the potential benefits of tailored evaluation frameworks for hydrological tasks (Xu et al., 2024b).

Benchmarks from other specialized fields provide valuable lessons. In code generation, datasets like EvoCodeBench offer structured evaluation methodologies (Li et al., 2024). In medicine, benchmarks like MIMIC-III, BioASQ, and CheXpert have revolutionized medical AI applications (Yan et al., 2024). In addition, datasets like BLURB and BioLP-bench have demonstrated the value of task-specific metrics in biomedical and biological applications (Feng et al., 2024; Ivanov, 2024). Similarly, SciEx, designed for

https://doi.org/10.1017/eds.2025.10006 Published online by Cambridge University Press

scientific reasoning, evaluates models using university-level exam questions with human grading to ensure accuracy (Dinh et al., 2024).

This study introduces HydroLLM-Benchmark, a hydrology-focused benchmark dataset designed to facilitate research related to the field, while also offering baseline performance results for state-of-the-art LLMs in hydrological question-answering tasks. HydroLLM-Benchmark compiles diverse hydrological resources, including textbook-based foundational concepts and cutting-edge research articles, ensuring robust coverage of theoretical underpinnings and emerging insights. Although the dataset has been streamlined to minimize preprocessing requirements, providing clear, structured, and readily usable data for machine learning, pipelines remain flexible enough to support alternative investigative approaches, such as physically based hydrological modeling.

Researchers can also leverage HydroLLM-Benchmark in conjunction with existing hydrological datasets, thereby expanding the scope for comparative analyses and hybrid modeling strategies. By addressing key gaps in existing benchmarks, HydroLLM-Benchmark aims to serve as a robust evaluation tool and catalyst for innovation in domain-specific LLM research, fostering advancements in hydrological science, education, and decision-making.

The remainder of this article is organized as follows: Section 2 outlines the methodology behind the design choices, development, and implementation of a hydrology-oriented intelligent assistance system, benchmarking its capacity to generate and answer domain-specific questions. Section 3 presents the benchmark results and provides a brief discussion of them. Section 4 describes the challenges faced during the process and addresses the limitations. Finally, Section 5 concludes with a summary of the study's contributions and insights for advancing AI in hydrological research and practice.

## 2. Methodology

This section outlines the methodology used to create a collection of hydrology-specific questions and answers and to evaluate the performance of LLMs in generating and answering these questions. The process involved selecting relevant research articles and textbooks to ensure a comprehensive representation of both foundational knowledge and recent advancements in hydrology. The methodology includes steps for data collection, question generation, and evaluation techniques, focusing on assessing the accuracy and contextual relevance of the models' outputs.

### *2.1. Data collection*

For this study, we selected both research articles and textbooks to comprehensively cover the current advancements and foundational knowledge in hydrology. The primary goal was to ensure that the chosen sources were both relevant to the field of hydrology and reflected the most recent developments in hydrological science. Our initial experiments with fine-tuning on multiple hydrology textbooks did not yield a notable improvement in specialized knowledge recollection. Modern LLM architectures already encompass a broad technical corpus, so the real benefit of fine-tuning is to instill the field's distinctive style, jargon, and conceptual framework.

We therefore chose "Fundamentals of Hydrology" (Davie, 2019), as it is widely regarded as providing an authoritative, comprehensive structure of core principles and a unified lexicon in the field of hydrology. This textbook is renowned for its comprehensive coverage of basic hydrological principles and processes, providing a solid foundation for understanding the broader applications of hydrology in both academic and practical contexts. Its inclusion in this study serves as a benchmark for comparing newer research insights with established hydrological knowledge. By anchoring our work in this single, highly respected text, we streamline the fine-tuning process to more directly enhance hydrology-specific reasoning without overwhelming the model with redundant or minimally impactful material. We anticipate that future contributions from the community will expand upon this foundation with additional texts and thereby keep HydroLLM-Benchmark aligned with ongoing developments in hydrological research.

https://doi.org/10.1017/eds.2025.10006 Published online by Cambridge University Press

In addition to the textbook, we gathered 2,000 research articles from Elsevier, a leading academic publisher known for its extensive repository of peer-reviewed journals. The selection of research articles focused specifically on those related to hydrology, published between 2022 and 2024, to capture the most current findings and trends in the field. Furthermore, to maintain focus and ensure quality, these articles were primarily filtered from three journals: *Journal of Hydrology*, *Advances in Water Resources*, and *Journal of Hydrology: Regional Studies*. By limiting the selection to this period, we ensured that the study reflects contemporary hydrological research, including the latest methodologies, technological advancements, and emerging challenges within the discipline. The selection process for both textbooks and research articles was guided by relevance to key hydrological topics, such as flood management, water quality, and hydrological modeling, forming the foundation of HydroLLM-Benchmark.

### *2.2. Experimental setup*

We selected GPT-4o-mini, Llama3:8B, and Llama3.1:70B to represent a range of model sizes and types (i.e., commercial vs. open source) commonly used in academic and applied settings. GPT-4o-mini was chosen due to its balance between performance and resource efficiency, enabling cost-effective deployment while retaining competitive capabilities, especially in zero-shot and instruction-following tasks. Notably, while GPT-4o offers top-tier performance, its significantly higher API cost ($10.00 vs. $0.60 per million input tokens) made GPT-4o-mini a more scalable choice (OpenAI, 2024) for our large-scale evaluation experiments.

Llama3:8B and 70B were included to explore performance across different model scales, as the 8B variant represents a smaller, resource-accessible model, while the 70B variant reflects cutting-edge capabilities at the high end of the open-weight model spectrum. We used the base versions (not instruction-tuned) of both Llama models to evaluate their raw language modeling abilities without additional task-specific adaptation, aiming for a controlled, fine-tuning-agnostic benchmark.

Other models, including Gemma 2 and Mistral 2, were considered but excluded due to limited infrastructure support, early-stage stability issues, or lack of reproducible inference pipelines at the time of evaluation. Commercial models such as Claude and Gemini were also excluded due to licensing restrictions. We intend for future iterations of HydroLLM-Benchmark to include a broader range of LLMs as the benchmark evolves.

Our experimental setup was designed to handle the computational demands of LLMs while maintaining consistent evaluation conditions across all models. We used Python as the primary programming language due to its versatility and the availability of extensive libraries well-suited for machine learning, data processing, and evaluation tasks. Python's flexibility enabled us to streamline both data preprocessing and model evaluation, ensuring each step of the workflow was efficient and reproducible for HydroLLM-Benchmark.

For data handling and processing, we utilized several key Python libraries. Pandas was employed to load, clean, and structure datasets in CSV format, enabling efficient organization and manipulation of the data for question generation and answer processing. This allowed us to maintain consistency in preprocessing across different question types. For numerical computations, especially for managing arrays and performing calculations during the evaluation phase, we relied on NumPy. This library was crucial for handling large datasets and ensuring the computational efficiency of our operations. To compute cosine similarity, a vital metric for evaluating the semantic accuracy of open-ended and fill-in-the-blanks responses, we used scikit-learn. Its robust implementation of cosine similarity integrated seamlessly into our evaluation framework, providing precise performance assessments.

We accessed GPT-4o-mini through the OpenAI API, which allowed us to configure model settings according to our experimental requirements. Specifically, we adjusted the max_tokens parameter to 4,000, ensuring that GPT-4o-mini could generate comprehensive responses for longer, open-ended questions without truncation. Running GPT-4o-mini locally on high-performance computers gave us the flexibility to control the environment and maintain consistent settings throughout the experiments.

https://doi.org/10.1017/eds.2025.10006 Published online by Cambridge University Press

In addition to GPT-4o-mini, we evaluated Llama3:8B and Llama3.1:70B. These models were deployed on a dedicated server infrastructure equipped with high-performance GPUs (i.e., NVIDIA L40S with 48 GB memory), to handle the resource-intensive demands of large-scale language models. This setup ensured efficient execution, particularly for the computationally demanding Llama3.1:70B model. Both Llama models were evaluated in their default configurations without fine-tuning to evaluate their baseline capabilities in hydrology-specific tasks.

The experimental workflow was structured to ensure fair and consistent evaluation across all models. After preprocessing the dataset and organizing questions to align with each model's input requirements, we generated structured prompts tailored to each question type, including true/false, multiple-choice, open-ended, and fill-in-the-blanks. These prompts provided clear and consistent instructions to guide the models in producing responses that adhered to the expected format for each question type. During the testing phase, these prompts were applied consistently to each model, and their outputs were collected under controlled conditions to thoroughly assess their performance on HydroLLM-Benchmark.

To evaluate the model-generated answers, we employed different metrics based on the question type. For true/false and multiple-choice questions, which have objective answers, accuracy was used as the primary metric. Each model's output was directly compared to the ground truth, and the accuracy score was calculated as the percentage of correct answers. For open-ended and fill-in-the-blanks questions, where responses could vary in structure but still convey similar meanings, we used cosine similarity to assess semantic alignment between the model-generated and reference answers. Using scikit-learn, we transformed the responses into vector form and calculated cosine similarity to enable meaningful comparisons of semantic content.

To ensure consistency across all evaluations, we maintained uniform parameter settings and environmental configurations for each model. By utilizing the GPT-4o-mini API and deploying the Llama models on the high-performance server, we balanced computational efficiency with standardized testing conditions. This hybrid setup allowed for an effective comparison of each model's out-of-the-box performance in handling hydrology-specific tasks, providing valuable insights into their strengths and limitations.

### *2.3. Question generation methodology*

In generating questions from the selected textbook and research article for HydroLLM-Benchmark, we began by systematically extracting relevant text data, focusing on sections most likely to yield meaningful content for question generation. This initial extraction targeted key passages and concepts central to hydrology, ensuring that the generated questions would be directly aligned with core educational and research objectives.

Once the relevant content was identified, we crafted a series of specialized prompts tailored for each question type—true/false, multiple-choice, open-ended, and fill-in-the-blank. Each question type required a unique approach; however, the foundational structure of the prompts remained consistent, with specific adjustments made to customize the format, required answer structure, and anticipated output style. These modifications allowed for both variety and uniformity, ensuring that each question adhered to the designated type while maintaining coherence across the generated content.

To guide the generation process and ensure high-quality outputs, we employed several prompting techniques, including task specification, constraint-based prompting, and self-contained prompting. Task specification allowed us to break down the question-generation process into distinct, actionable steps. Constraint-based prompting helped maintain format integrity, ensuring each question type aligned with the expected answer format and minimized irrelevant content. Self-contained prompting enabled the production of questions that were independent and clearly understandable without additional context, making them versatile for use in educational and research settings.

For generating the actual question–answer (Q&A) pairs, we utilized GPT-4o-mini, an optimized version of GPT-4 designed for tasks requiring nuanced context understanding and content generation.

https://doi.org/10.1017/eds.2025.10006 Published online by Cambridge University Press

***Table 1.*** *Sample questions from HydroLLM-Benchmark categorized by source and question type*

| Source type | Question type | Sample question |
|---|---|---|
| Research article | True/false | True or false: Urbanization has no effect on the frequency and intensity of extreme precipitation events. |
| Research article | Multiple-choice | What advancements in remote sensing technologies have enhanced the monitoring of groundwater storage dynamics? (A) The introduction of low-resolution satellite imagery. (B) The integration of machine learning algorithms with multi-platform satellite measurements. (C) The exclusive reliance on traditional ground-based measurements. |
| Textbook | Fill-in-the-blanks | Hydrology is the science or study of _____. |
| Textbook | Open-ended | What is hydrology, and what aspects of water does it primarily focus on? |

With carefully constructed prompts and the extracted text, the model generated a variety of relevant, well-structured questions and corresponding answers. Across multiple iterations, we developed 1,124 research articles and 224 textbook true/false questions, 1,001 research articles and 209 textbook multiple-choice questions, 997 research articles and 225 textbook fill-in-the-blanks questions, and 2001 research articles and 220 textbook Open-Ended questions. This iterative process involved multiple rounds of refinement, ensuring the questions were accurate, diverse in format, and pedagogically valuable. The final set of question–answer pairs was thus both comprehensive and adaptable, providing a robust resource for academic and research applications, and serving as the cornerstone of HydroLLM-Benchmark. Table 1 presents example questions categorized by source type and question type.

### *2.4. Answer generation and evaluation*

We utilized three models—GPT-4o-mini, Llama3:8B, and Llama3.1:70B—to generate answers for hydrology-specific questions from HydroLLM-Benchmark. To ensure the models provided responses in the correct formats for each question type (true/false, multiple-choice, open-ended, and fill-in-the-blanks), we developed tailored prompts. This approach allowed us to maintain consistency across models, ensuring a fair and standardized evaluation of their performance.

Each question type required a specific prompting strategy. For true/false and multiple-choice questions, the prompts were designed to elicit concise and precise answers, minimizing ambiguity. These formats required the models to select or generate straightforward responses, making accuracy a critical factor in assessing their performance. In contrast, open-ended and fill-in-the-blanks questions demanded more detailed and context-aware responses. To support this, the prompts included additional context and background information, encouraging the models to produce nuanced answers that captured the complexity of hydrological concepts.

Following the generation of answers, we evaluated the models using metrics tailored to the nature of each question type. For objective questions (true/false and multiple-choice), accuracy served as the primary evaluation metric, offering a clear measure of the models' ability to generate correct responses. For subjective questions (open-ended and fill-in-the-blanks), we employed cosine similarity to assess the semantic closeness between the generated answers and reference answers. This metric enabled us to gauge how well the models understood and addressed the questions, even when the responses varied in wording but shared the same underlying meaning.

By using both accuracy and cosine similarity, we comprehensively evaluated the models' performance across diverse question types. This dual-metric approach provided a thorough assessment,

https://doi.org/10.1017/eds.2025.10006 Published online by Cambridge University Press

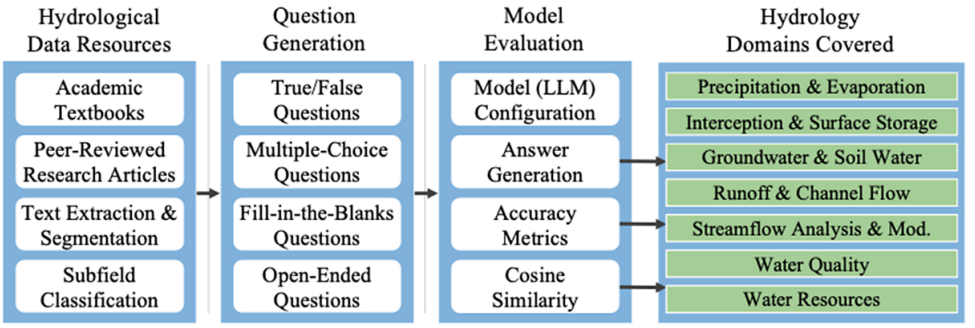


***Figure 1.*** *Conceptual overview of HydroLLM-Benchmark.*

highlighting the models' strengths in handling objective questions and their ability to generate contextually appropriate responses for subjective ones. Tailoring the prompts and evaluation metrics to the specific demands of each question type ensured a rigorous and reliable benchmarking process. Figure 1 illustrates the overall system architecture, from question generation to performance evaluation.

### *2.5. Data post-processing and model training*

To prepare a high-quality dataset for evaluating model performance on hydrology-specific tasks, we initiated a comprehensive data post-processing step. This step focused on refining the input data by filtering out Q&A pairs containing overly specific details, such as references to locations, article-specific content, or numerical results. These elements were removed to ensure that the dataset remained broadly relevant to hydrology concepts, rather than being tied to specific contexts. For this task, GPT-4o-mini was employed to analyze and flag questions based on these criteria. A specially designed prompt guided the model to assess each question for such details. Any flagged questions were subsequently excluded from the dataset, resulting in a cohesive and conceptually relevant collection suited for baseline evaluation.

Following the post-processing phase, we assessed the baseline capabilities of three LLMs—GPT-4o-mini, Llama3:8B, and Llama3.1:70B—using a range of question types, including true/false, multiple-choice, open-ended, and fill-in-the-blanks. The models were evaluated in their pretrained states without any additional fine-tuning, as the primary objective was to gauge their out-of-the-box performance. This approach allowed us to establish a baseline understanding of each model's inherent abilities in addressing hydrology-related queries.

To ensure proper response formatting, we crafted tailored prompts for each question type, embedding specific instructions to guide the models. For true/false questions, the prompts directed the models to make clear binary selections based on the given content. For multiple-choice questions, prompts guided the models to choose the most appropriate option while minimizing irrelevant details. For open-ended and fill-in-the-blanks questions, the prompts encouraged the generation of detailed and contextually nuanced answers, reflecting a deeper understanding of hydrological concepts.

To accommodate the complexity of open-ended responses, especially in GPT-4o-mini, we adjusted the maximum token limit to 4,000 tokens. This adjustment was essential to prevent the truncation of longer Q&A pairs, ensuring that the responses were comprehensive. Importantly, no further fine-tuning or domain-specific training was applied to any of the models, as the focus remained on evaluating their baseline capabilities in hydrology-related tasks.

Through data post-processing, prompt customization, and thorough model evaluation, we effectively established the strengths and limitations of each model in handling hydrology-specific content. Figure 2 illustrates the complete process, from data extraction to model output generation and performance scoring.

https://doi.org/10.1017/eds.2025.10006 Published online by Cambridge University Press

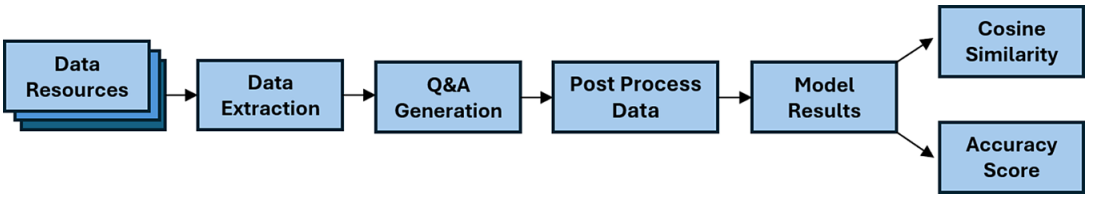


***Figure 2.*** *Post-processing, model output generation, and scoring.*

### *2.6. Q&A evaluation framework*

To evaluate the performance of the models on hydrology-specific questions, we developed a structured evaluation framework that systematically assessed their responses across multiple question types. This framework utilized two distinct metrics—accuracy and cosine similarity—to accommodate the varied nature of the question formats: true/false, multiple-choice, open-ended, and fill-in-the-blanks. Each metric was selected to provide an accurate measure of the models' performance based on the specific requirements of each question type.

#### *2.6.1. Evaluation of objective questions*

For true/false and multiple-choice questions, which are inherently objective with clear, definitive answers, accuracy served as the primary evaluation metric. The evaluation process involved a straightforward comparison of each model's output with the established ground truth answer. (i) Correct versus incorrect classification: A model's response was classified as correct if it matched the ground truth, and incorrect otherwise. (ii) Accuracy calculation: The accuracy score for each model was determined as the percentage of correct answers relative to the total number of questions within each objective question type. This provided a clear, binary measure of the model's ability to identify or select the correct answer.

#### *2.6.2. Evaluation of subjective questions*

For open-ended and fill-in-the-blanks questions, which often require a more nuanced understanding, cosine similarity was used as the evaluation metric. This metric assesses the semantic alignment between the model-generated answer and the ground truth by measuring the angle between their vector representations. (i) Vectorization of responses: Both the model-generated answers and the ground truth answers were transformed into vector form. This allowed for the analysis of responses based on their underlying meaning rather than exact wording. (ii) Cosine similarity calculation: Cosine similarity was computed between the vector representations of the model's answer and the reference answer, producing a score between –1 and 1. Scores closer to 1 indicated a higher degree of semantic similarity.

Using cosine similarity provided a nuanced evaluation of the models' responses for subjective questions, where different phrasings could convey equivalent meanings. This metric enabled us to assess the models' contextual and semantic comprehension, which is crucial for effective application in hydrology-related tasks.

#### *2.6.3. Aggregating and comparing model performance*

After calculating the individual scores for each question type, we aggregated the results to compute an average score for each model across all question formats: (i) Objective scores: The accuracy scores for true/false and multiple-choice questions were averaged to offer a comprehensive view of each model's performance on objective, factual questions. (ii) Subjective scores: The cosine similarity scores for open-ended and fill-in-the-blanks questions were averaged to summarize each model's ability to generate semantically accurate, contextually relevant responses.

This approach allowed us to compile a clear, comparative performance profile for each model—GPT-4o-mini, Llama3:8B, and Llama3.1:70B—across various question types. By evaluating objective and

https://doi.org/10.1017/eds.2025.10006 Published online by Cambridge University Press

subjective questions separately, we gained valuable insights into each model's strengths and limitations in addressing various aspects of hydrology-specific queries. This dual-metric framework provided a balanced and comprehensive evaluation, enabling a nuanced understanding of how effectively each model could interpret, understand, and answer questions relevant to the field of hydrology.

## 3. Results

In this section, we present baseline results over the benchmark dataset and the evaluation of several model performances on true/false, multiple-choice, open-ended and fill-in-the-blanks question formats, highlighting the strengths and limitations of each model across different content sources, including research articles and the textbook. We will explore the implication of these findings for understanding how content type and question format influence.

To establish baseline results for our hydrology-specific true/false question set, we evaluated three LLMs (i.e., GPT-4o-mini, Llama3:8B, and Llama3:70B) using questions derived from both textbooks and research articles. As illustrated in Figure 3, GPT-4o-mini demonstrates consistently high accuracy in both categories, outperforming the other models when responding to textbook-based questions. Llama3:70B shows comparable performance on textbook-derived items, although it exhibits slightly lower accuracy on questions sourced from research articles. By contrast, Llama3:8B maintains moderate accuracy levels across both data types but does not match the peak scores observed with the other models. These results suggest that the models are generally proficient at handling straightforward true/false inquiries, yet the discrepancy in performance between textbook- and article-based questions underlines the need for further fine-tuning or domain adaptation.

Similar results were observed for multiple-choice questions, as illustrated in Figure 4. GPT-4o-mini once again achieved high accuracy, particularly on questions derived from research articles, suggesting a robust capacity for domain-specific inference. Llama3:70B closely followed, displaying comparable performance levels for both textbook- and article-based items. Meanwhile, Llama3:8B maintained moderate accuracy scores but lagged behind the other two models. The consistency of results across question sources indicates that all three LLMs are well-equipped for tasks requiring precise answer selection, although further fine-tuning may be necessary to optimize performance on specialized content.

Shifting the focus to fill-in-the-blanks questions, cosine similarity scores were used to assess how closely each model's generated text aligned with the correct solutions. As seen in Figure 5, GPT-4o-mini emerges as the top performer, producing contextually cohesive completions for both textbook and article prompts. Slightly lower scores were obtained by Llama3:70B, although its results remain sufficiently high to suggest strong linguistic capabilities. In contrast, Llama3:8B occupies the middle range, capturing the main ideas but sometimes missing finer nuances. This distribution highlights the potential of LLMs to

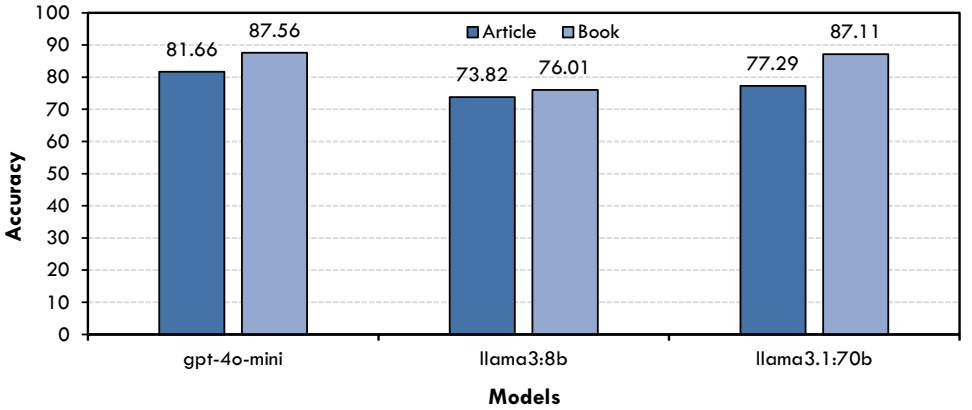


***Figure 3.*** *Accuracy scores for true/false Q&A.*

https://doi.org/10.1017/eds.2025.10006 Published online by Cambridge University Press

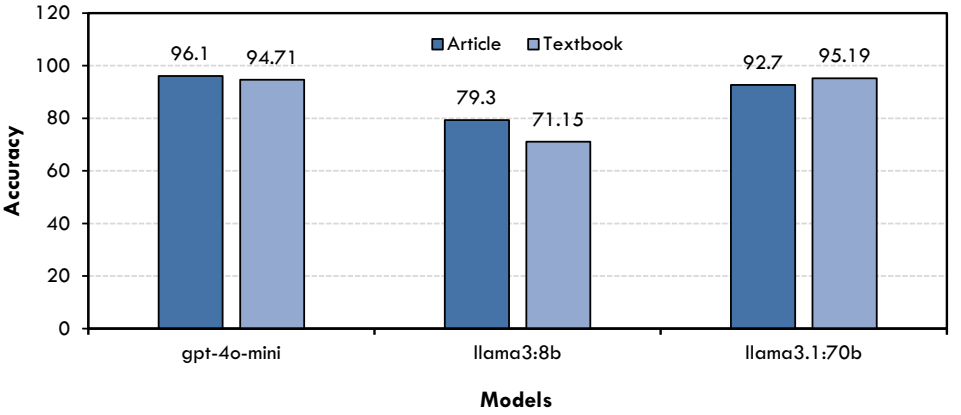


***Figure 4.*** *Accuracy scores for multiple-choice Q&A.*

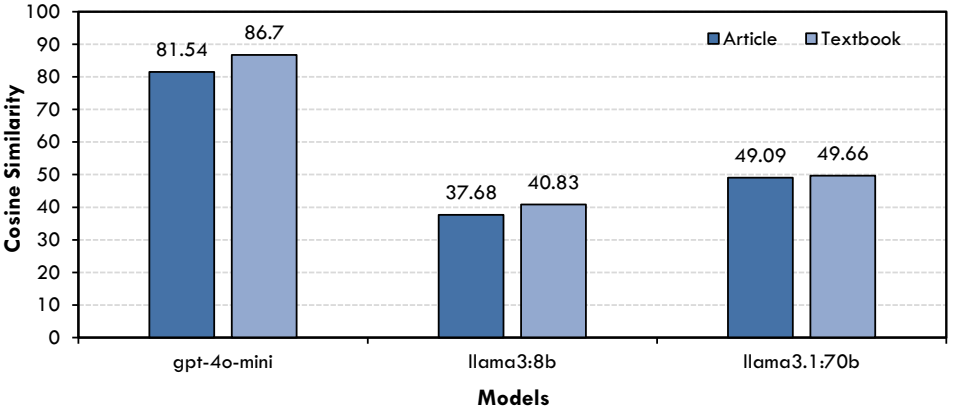


***Figure 5.*** *Cosine similarity scores for fill-in-the-blanks Q&A.*

excel in semi-structured tasks, while also revealing the need for targeted improvements to address specialized hydrological terminology.

Focusing on the open-ended questions, cosine similarity again served as the metric for evaluating semantic alignment between *n* model outputs and reference answers. Figure 6 shows that both GPT-4o-mini and Llama3:70B scored at the upper end, indicating an aptitude for generating coherent, in-depth responses even when the query allows for wide-ranging expressions. Llama3:8B exhibits only a minor decrease in similarity, suggesting it can still capture essential information but may occasionally lack the refinement displayed by the other two models.

## 4. Discussions

This section explores the comparative analysis of language model performance, highlights the significance of the HydroLLM-Benchmark dataset as a living resource, and addresses the challenges and limitations observed during the evaluation process.

### *4.1. Comparative analysis*

Across all four question types—true/false, multiple-choice, fill-in-the-blanks, and open-ended—GPT-4o-mini consistently emerges as the top performer, maintaining high scores in both objective evaluations

https://doi.org/10.1017/eds.2025.10006 Published online by Cambridge University Press

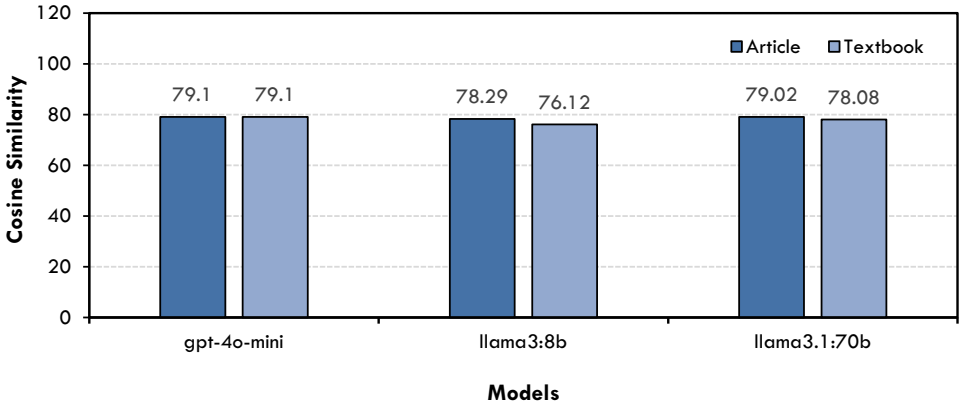


***Figure 6.*** *Cosine similarity scores for open-ended Q&A.*

(true/false and multiple-choice) and subjective measures (fill-in-the-blanks and open-ended). Llama3:70B closely follows, showing comparable accuracy in multiple-choice and strong semantic alignment in open-ended and fill-in-the-blanks tasks, albeit slightly trailing GPT-4o-mini. Meanwhile, Llama3:8B registers moderate performance, indicating sufficient competence in handling basic to intermediate queries but revealing gaps in handling nuanced or specialized terminology.

These findings are particularly noteworthy given the domain-specific nature of our benchmark dataset, which comprises hydrology-focused questions derived from textbooks and research articles. By testing each model's proficiency in both factual and interpretive tasks, this dataset establishes a clear baseline for evaluating LLM performance in hydrological knowledge assessment. The highest overall accuracies and cosine similarity scores were recorded by GPT-4o-mini, suggesting that it currently sets the standard for domain-specific question answering within our benchmark. However, Llama3:70B's relatively close results underscore the potential for models with larger parameter counts to excel in specialized fields, provided they undergo targeted fine-tuning or training on hydrology-related corpora.

### *4.2. HydroLLM-Benchmark as a living dataset*

HydroLLM-Benchmark is designed as a living resource, intended to evolve continuously through systematic updates and expansions. As new research articles, updated textbook editions, and hydrology-specific datasets become available, they will be carefully curated and integrated to ensure the benchmark remains aligned with cutting-edge advancements in hydrological science.

This iterative approach not only maintains the dataset's relevance and accuracy but also fosters community-driven contributions. Researchers, educators, and practitioners are encouraged to submit new data and evaluation methodologies, promoting collaboration and knowledge-sharing across the hydrology community. To facilitate community participation, we provide several accessible mechanisms via our GitHub repository. Users can submit questions, feedback, or concerns by opening an issue on the GitHub page. We welcome code and content contributions, including new question sets, data processing scripts, or model evaluation tools, via standard pull request workflows. Contributors may also reach out via email or community forums to suggest ideas or request features.

We also plan to organize collaborative activities such as online workshops, shared evaluation tasks, and hackathons through research communities like Cooperative Institute for Research to Operations in Hydrology and Advancing Earth and Space Science. These initiatives aim to build a collaborative network around HydroLLM-Benchmark and encourage knowledge sharing at the intersection of hydrology and AI.

https://doi.org/10.1017/eds.2025.10006 Published online by Cambridge University Press

Future updates may incorporate specialized modules for emerging topics like climate change modeling, flood risk analysis, and water resources optimization, broadening the dataset's applicability. In addition, HydroLLM-Benchmark aims to serve as a dynamic educational tool, supporting interactive learning experiences and domain-specific curricula development. By embracing an open framework and a transparent contribution process, HydroLLM-Benchmark aspires to remain a versatile and forward-looking resource, empowering ongoing innovation and advancing AI-driven hydrological research and education.

Beyond academic evaluation, HydroLLM-Benchmark also holds potential for adaptation to operational hydrology applications such as flood forecasting, drought monitoring, and disaster response. Future extensions of the benchmark could integrate question formats derived from early warning reports, hydrological alerts, or emergency management protocols. This direction aligns with recent work such as Flash Flood - Bidirectional Encoder Representation from Transformers (FF-BERT), which classifies flash flood reports from unstructured text (Wilkho et al., 2024), and LLM studies that assess reasoning under adverse weather conditions (Zafarmomen and Samadi, 2025). Similarly, hybrid pipelines using LLMs for event-location extraction from social media (Fan et al., 2020) illustrate how natural language understanding can enhance disaster informatics. By connecting domain-specific benchmarking with these operational use cases, HydroLLM-Benchmark can evolve into a practical testbed for evaluating LLM readiness in real-time, high-stakes hydrological decision-making.

### *4.3. Challenges and limitations*

This section discusses the challenges encountered in generating domain-specific questions and the limitations observed in the performance of evaluated LLMs in hydrology-related tasks. The challenges outlined include biases in question generation, issues with specificity and relevance, and the complexities of crafting high-quality questions in a specialized field. Furthermore, we explore the limitations of these models in understanding hydrology-specific terminology, managing complex concepts, and addressing the nuances of the domain. Through this analysis, we aim to provide insights into the obstacles faced when applying LLMs to hydrology and identify areas for future improvements in model development and training.

#### *4.3.1. Challenges in generating domain-specific questions*

Generating high-quality, domain-specific questions in hydrology presented several challenges. In the case of multiple-choice questions, GPT-4o-mini exhibited a consistent bias toward generating questions with the answer "B." To address this issue, we experimented by running the model multiple times with different prompt parameters. One prompt explicitly instructed the model to vary answer choices, while a default prompt did not specify particular answer letters. Despite these adjustments, the model's output continued to favor certain options, resulting in nearly 70.4% of the answers being "B," 17% "A," and only 5.6% "C." To determine whether this answer bias influenced the model's overall accuracy, we conducted additional experiments where we shuffled and reassigned answer letters to balance the dataset. Interestingly, this balancing did not significantly impact accuracy, suggesting that the model's bias toward certain answer letters did not detrimentally affect its understanding or response accuracy.

Open-ended and fill-in-the-blank questions posed their own unique difficulties. Without specific instructions in the prompts, GPT-4o-mini frequently generated questions with introductory phrases such as "In this study…" or "In this article…," which were unsuitable for the standalone questions required in our dataset. To improve the quality and generality of the questions generated, we refined our prompts to explicitly exclude these introductory phrases, leading to a notable improvement in the final output.

Furthermore, hydrology's broad scope, encompassing various geographical locations and historical contexts, added complexity to the question generation process. The model often produced questions that were overly specific, referencing locations or years that were not relevant to the core hydrological content. To mitigate this issue, we incorporated a post-processing step to filter out location- and year-specific

https://doi.org/10.1017/eds.2025.10006 Published online by Cambridge University Press

questions that did not contribute to the intended educational goals. While some references to locations and years can be beneficial for context, we excluded those that did not directly support hydrology-related concepts, ensuring the questions remained general yet accurate in their domain relevance.

Despite our mitigation efforts, GPT-4o-mini continued to display a notable bias toward selecting "B" as the correct answer in multiple-choice questions. We tested several strategies, including rephrased prompts, randomized answer orders, altering output token length, and varying temperature settings (0.2–0.9), but the output distribution remained largely unchanged. This suggests that the bias may stem from deeper training artifacts or token-level preferences embedded in the model's architecture. Such behavior has implications for future benchmarking efforts, as it may introduce unintended skew in answer selection. For researchers developing automated assessment tools or training datasets, it is essential to consider these underlying biases and implement techniques such as randomization, controlled answer ordering, or ensemble prompting to ensure balanced data generation and evaluation.

#### *4.3.2. Limitations of models*

In assessing the performance of the models on hydrology-specific content, several notable limitations emerge, particularly with fill-in-the-blank and open-ended question formats. These models often display reduced accuracy in these question types, primarily due to their challenges in identifying precise vocabulary relevant to hydrology. Fill-in-the-blank questions require models to select the correct word or phrase, a task complicated by terms that may have similar meanings or context-dependent interpretations. For example, hydrological terms with specific implications can also possess general or alternate meanings in other fields, leading to misinterpretations and incorrect responses. This ambiguity in language represents a significant challenge for these models, resulting in errors when selecting the most contextually appropriate terms for hydrology-focused questions, ultimately affecting their overall performance and highlighting the difficulty of achieving precise understanding in domain-specific contexts like hydrology.

While the models exhibit strong general language understanding, their grasp of hydrology-specific terminology and context remains limited. These models may struggle with technical jargon, scientific terms, and context-specific language that is prevalent in hydrological research. This limitation can lead to less accurate responses for complex queries that necessitate a deep understanding of the domain. Furthermore, hydrology often involves complex mathematical equations and statistical models, which pose challenges for LLMs to interpret accurately. These models have limited capabilities in understanding numerical data and calculations, making them less effective at addressing questions requiring mathematical reasoning or the interpretation of quantitative data.

Contextual understanding of interconnected hydrological processes is another area where these models may falter. Hydrology involves grasping the relationships between groundwater flow, surface runoff, and atmospheric conditions. The models might not fully capture these interdependencies, resulting in responses that oversimplify or misinterpret complex systems. This limitation is particularly evident in questions requiring the synthesis of information from multiple sources or an understanding of cause-and-effect relationships.

Moreover, the models tend to generate more generalized responses, which may lack the specificity needed for detailed hydrology questions. This issue can be problematic for open-ended questions or those requiring precise, contextually accurate answers based on specific research findings or hydrological scenarios. In addition, hydrological analysis often necessitates interpreting visual data, such as satellite imagery, hydrological maps, and diagrams. The text-based nature of these models restricts their ability to process and analyze visual information, limiting their effectiveness in applications that require visual-spatial reasoning. Although multimodal capabilities could address this gap, current models lack robust integration with visual data sources.

The models also struggle with ambiguous or implicit queries that require contextual interpretation. In hydrology, where the same term can have different meanings based on context, such as "flow" in the context of streamflow versus groundwater flow, the models may produce inconsistent or incorrect

https://doi.org/10.1017/eds.2025.10006 Published online by Cambridge University Press

responses if the context is not explicitly provided. In addition, the sensitivity of these models to input formatting can affect their responses. Variations in wording, phrasing, or format can lead to different outputs, which may not always be consistent or reliable. This sensitivity complicates the Q&A generation process and may necessitate careful prompt engineering to achieve consistent results.

There is also a potential for data leakage, where the models' responses may be influenced by similar questions or answers in their training data. This phenomenon can lead to inflated performance metrics that do not accurately reflect the models' true capabilities in novel or context-specific tasks. Furthermore, the lack of extensive real-world validation for these models raises concerns about their effectiveness in practical hydrology applications. While responses are evaluated in controlled environments using benchmark questions, their performance in real-world scenarios—such as providing insights for field-work, decision-making, or policy recommendations—remains uncertain.

Finally, models like Llama3:8B and Llama3.1:70B require significant computational resources, including high-performance GPUs and extensive memory, to run efficiently. This limitation may restrict accessibility for users with limited technical infrastructure or resources, impacting their practical deployment in research or educational settings. These limitations highlight the areas where current LLMs can be improved for more effective application in hydrology-specific tasks, suggesting potential directions for further mini and customized model development and fine-tuning.

## 5. Conclusion

In this study, by collating a broad collection of hydrology textbooks and 2,000 peer-reviewed research articles, we introduce a specialized dataset, designed to evaluate question-answering capabilities in the hydrology domain. This dataset features diverse question formats—including true/false, multiple-choice, fill-in-the-blanks, and open-ended—thus capturing both fundamental concepts and advanced research topics. We defined sample evaluation tasks using GPT-4o-mini, Llama3:8B, and Llama3.1:70B, providing baseline benchmark results that highlight the strengths and limitations of current LLMs in handling domain-specific queries.

The dataset is unfiltered to preserve the complexity and authenticity of real-world hydrological data, making it suitable for a wide range of machine learning and deep learning applications. Although this resource currently focuses on hydrological themes, the insights gleaned from its use may prove valuable to broader research areas within environmental sciences. By openly sharing HydroLLM-Benchmark, we offer a standardized benchmark to address the lack of unified datasets in hydrological and water resources research. We strongly encourage other scholars and practitioners to adopt this benchmark dataset in future hydrological modeling and AI-driven research studies, furthering the collective understanding and innovation within this critical field.

Looking ahead, we recognize that hydrological reasoning often requires interpreting data in multimodal formats, such as satellite imagery, hydrological maps, and time-series plots. While the current version of HydroLLM-Benchmark focuses on text-based questions, future iterations will incorporate these multimodal components to mirror real-world hydrological analysis tasks more closely. This expansion will enable evaluation of advanced models with vision-language capabilities, supporting tasks like flood map interpretation, hydrograph analysis, and spatial reasoning. Integrating multimodal elements is a key next step toward building a comprehensive, domain-aware benchmark for hydrological AI.

In addition, the future of the HydroLLM-Benchmark dataset envisions integrating emerging AI model architectures and advancements in natural language processing to improve the evaluation of domain-specific knowledge. By incorporating newer models and technologies, we can track the progression and refinement of AI capabilities in hydrology. This ongoing evolution will also facilitate the testing of innovative training methodologies and optimization techniques, enhancing model performance on complex, specialized queries. Furthermore, expanding the dataset to include cross-disciplinary content could foster a more holistic understanding of hydrological processes, aiding models in recognizing complex interconnections between hydrology and related environmental sciences.

https://doi.org/10.1017/eds.2025.10006 Published online by Cambridge University Press

Community contributions are vital to the growth and effectiveness of the HydroLLM-Benchmark. By cultivating an open ecosystem, we invite hydrology experts, AI researchers, and educators to participate actively in refining and enriching the dataset. This collaborative effort allows for the inclusion of diverse perspectives, enriching the dataset with varied question types and scenarios reflecting real-world challenges. Engaging the community in this manner not only democratizes access to cutting-edge resources but also drives transparency and inclusivity in AI research. Through workshops, hackathons, and collaborative initiatives, stakeholders are encouraged to explore the dataset's potential and contribute insights, ensuring its relevance and applicability in addressing global hydrological issues.

As the landscape of LLMs continues to evolve rapidly, we also plan to benchmark newer model families such as Llama3.2, Llama4, DeepSeek, and other emerging open and commercial models that offer advancements in instruction following, multilingual reasoning, and long-context understanding. Incorporating these models into HydroLLM-Benchmark will help maintain its relevance for assessing state-of-the-art performance across a diverse set of hydrological tasks.

**Open peer review.** To view the open peer review materials for this article, please visit http://doi.org/10.1017/eds.2025.10006.

**Author contribution.** Dilara Kizilkaya: Methodology, software, formal analysis, investigation, data curation, writing—original draft. Ramteja Sajja: Validation, writing—review and editing, visualization, and data curation. Yusuf Sermet: Conceptualization, methodology, writing—review and editing, validation, supervision, and funding acquisition. Ibrahim Demir: Conceptualization, methodology, writing—review and editing, project administration, funding acquisition, and resources.

**Competing interests.** The authors declare none.

**Data availability statement.** The codebase and dataset are open-source, free to use, and can be accessed on GitHub (https://github.com/uihilab/HydroQA).

**Funding statement.** This project was funded by the National Oceanic and Atmospheric Administration (NOAA) via a cooperative agreement with the University of Alabama (NA22NWS4320003) awarded to the Cooperative Institute for Research to Operations in Hydrology (CIROH). We also acknowledge NSF grant NAIRR240072 for research computing on multimodal language models in hydrology.

**Declaration of generative AI and AI-assisted technologies.** During the preparation of this manuscript, the authors used ChatGPT, based on the GPT-4o model, to improve the flow of the text, correct grammatical errors, and enhance the clarity of the writing. The language model was not used to generate content, citations, or verify facts. After using this tool, the authors thoroughly reviewed and edited the content to ensure accuracy, validity, and originality, and take full responsibility for the final version of the manuscript.

## References

**Alabbad Y**, **Mount J**, **Campbell AM and Demir I** (2024) A web-based decision support framework for optimizing road network accessibility and emergency facility allocation during flooding. *Urban Informatics 3*(1), 10.

**Anderson MP** (2007) Introducing groundwater physics. *Physics Today 60*(5), 42–47.

**Chen X**, **Li L**, **Chang L**, **Huang Y**, **Zhao Y**, **Zhang Y and Li D** (2023) Challenges and contributing factors in the utilization of Large Language Models (LLMS). arXiv (Cornell University). https://doi.org/10.48550/arxiv.2310.13343

**Davie T and Quinn NW** (2019) *Fundamentals of Hydrology*. In Routledge eBooks. https://doi.org/10.4324/9780203798942

**Demir I**, **Xiang Z**, **Demiray B and Sit M** (2022) Waterbench: A large-scale benchmark dataset for data-driven streamflow forecasting. *Earth System Science Data Discussions 2022*, 1–19.

**Dinh TA**, **Mullov C**, **Bärmann L**, **Li Z**, **Liu D**, **Reiß S**, **Lee J**, **Lerzer N**, **Gao J**, **Peller-Konrad F**, **Röddiger T**, **Waibel A**, **Asfour T**, **Beigl M**, **Stiefelhagen R**, **Dachsbacher C**, **Böhm K and Niehues J** (2024) SciEx: Benchmarking Large Language Models on Scientific Exams with Human Expert Grading and Automatic Grading. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 11592–11610. https://doi.org/10.18653/v1/2024.emnlp-main.647

**Dolich A**, **Ebeling P**, **Stölzle M**, **Kiesel J**, **Götte J**, **Guse B**, et al (2024) CAMELS-DE: Benchmark dataset for hydrology–significance, current status and outlook. *EGU24*, (EGU24-17667).

**Ebert-Uphoff I**, **Thompson DR**, **Demir I**, **Gel YR**, **Karpatne A**, **Guereque M**, **Kumar V**, **Cabral-Cano E and Smyth P** (2017) A VISION FOR THE DEVELOPMENT OF BENCHMARKS TO BRIDGE GEOSCIENCE AND DATA SCIENCE. International Workshop Climate Informatics. https://par.nsf.gov/biblio/10143795-vision-development-benchmarks-bridge-geoscience-data-science

**Fan C**, **Wu F and Mostafavi A** (2020) A hybrid machine learning pipeline for automated mapping of events and locations from social media in disasters. *IEEE Access 8*, 10478–10490.

**Feng H**, **Ronzano F**, **LaFleur J**, **Garber M**, **de Oliveira R**, **Rough K**, et al (2024) Evaluation of large language model performance on the biomedical language understanding and reasoning benchmark: Comparative study. *medRxiv*, 2024-05.

**Harshbarger JW and Ferris JG** (1963) Interdisciplinary training program in scientific hydrology. *Groundwater 1*(2), 11–14.

**Harvel N**, **Haiek FB**, **Ankolekar A and Brunner DJ** (2024) Can LLMs answer investment banking questions? using Domain-Tuned functions to improve LLM performance on Knowledge-Intensive analytical tasks. *Proceedings of the AAAI Symposium Series*, *3*(1), 125–133. https://doi.org/10.1609/aaaiss.v3i1.31191.

**Ivanov I** (2024) BioLP-bench: Measuring understanding of biological lab protocols by large language models. *bioRxiv*, 2024-08.

**Izquierdo-Horna LUIS**, **Zevallos J**, **Cevallos T and Rios D** (2024) Design and creation of a database to assess the information needs of hydrological models. *WIT Transactions on Ecology and the Environment 262*, 619–629.

**Kadiyala LA**, **Mermer O**, **Samuel DJ**, **Sermet Y and Demir I** (2024a) 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), 148.

**Kadiyala L**, **Mermer O**, **Samuel DJ**, **Sermet Y and Demir I** (2024b) A comprehensive evaluation of multimodal large language models in hydrological applications. *EarthArxiv* 7176. https://doi.org/10.31223/X5TQ37

**Kadiyala LA**, **Sajja R**, **Sermet Y**, **Muste M and Demir I** (2024c) AI-driven decision-making for water resources planning and hazard mitigation using automated multi agents. *EarthArxiv* 8298. https://doi.org/10.31223/X5ZQ57

**Kratzert F**, **Klotz D**, **Shalev G**, **Klambauer G**, **Hochreiter S and Nearing G** (2019) Benchmarking a catchment-aware long short-term memory network (LSTM) for large-scale hydrological modeling. *Hydrology and Earth System Sciences Discussions 2019*, 1–32.

**Kreibich H**, **Schröter K**, **Di Baldassarre G**, **Van Loon AF**, **Mazzoleni M**, **Abeshu GW**, **Agafonova S**, **AghaKouchak A**, **Aksoy H**, **Alvarez-Garreton C**, **Aznar B**, **Balkhi L**, **Barendrecht MH**, **Biancamaria S**, **Bos-Burgering L**, **Bradley C**, **Budiyono Y**, **Buytaert W**, **Capewell L**, **… Ward PJ** (2023) Panta Rhei benchmark dataset: socio-hydrological data of paired events of floods and droughts. *Earth System Science Data*, *15*(5), 2009–2023. https://doi.org/10.5194/essd-15-2009-2023

**Li J**, **Li G**, **Zhang X**, **Zhao Y**, **Dong Y**, **Jin Z**, et al (2024) Evocodebench: An evolving code generation benchmark with domain-specific evaluations. *arXiv preprint arXiv:2410.22821*.

**Malakar P**, **Anshuman A**, **Kumar M**, **Boumis G**, **Clement TP**, **Tashie A**, et al (2024) An in-situ daily dataset for benchmarking temporal variability of groundwater recharge. *Earth System Science Data Discussions 2024*, 1–19.

**Newman AJ**, **Mizukami N**, **Clark MP**, **Wood AW**, **Nijssen B and Nearing G** (2017) Benchmarking of a physically based hydrologic model. *Journal of Hydrometeorology 18*(8), 2215–2225.

**OpenAI** (2024, July 18) GPT-4O Mini: Advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence

**Pursnani V**, **Sermet MY and Demir I** (2024) A conversational intelligent assistant for enhanced operational support in floodplain management with multimodal data. *EarthArxiv* 8264. https://doi.org/10.31223/X52M7W

**Pursnani V**, **Sermet Y**, **Kurt M and Demir I** (2023) Performance of ChatGPT on the US fundamentals of engineering exam: Comprehensive assessment of proficiency and potential implications for professional environmental engineering practice. *Computers and Education: Artificial Intelligence 5*, 100183.

**Russo L**, **Mauro F**, **Sebastianelli A**, **Gamba P and Ullo SL** (2024) SEN12-WATER: A new dataset for hydrological applications and its benchmarking. *arXiv preprint arXiv:2409.17087*.

**Sajja R**, **Pursnani V**, **Sermet Y and Demir I** (2025) AI-assisted educational framework for floodplain manager certification: Enhancing vocational education and training through personalized learning. *IEEE Access 13*, 42401–42413.

**Sajja R**, **Sermet Y**, **Cikmaz M**, **Cwiertny D and Demir I** (2024a) Artificial intelligence-enabled intelligent assistant for personalized and adaptive learning in higher education. *Information 15*(10), 596.

**Sajja R**, **Sermet Y**, **Cwiertny D and Demir I** (2023) Platform-independent and curriculum-oriented intelligent assistant for higher education. *International Journal of Educational Technology in Higher Education 20*(1), 42.

**Sajja R**, **Sermet Y and Demir I** (2024b) End-to-end deployment of the educational AI hub for personalized learning and engagement: A case study on environmental science education. *EarthArxiv* 7566. https://doi.org/10.31223/X5XM7N

**Samuel DJ**, **Sermet Y**, **Cwiertny D and Demir I** (2024b) Integrating vision-based AI and large language models for real-time water pollution surveillance. *Water Environment Research 96*(8), e11092.

**Samuel DJ**, **Sermet MY**, **Mount J**, **Vald G**, **Cwiertny D and Demir I** (2024a) Application of large language models in developing conversational agents for water quality education, communication and operations. *EarthArxiv*, 7056. https://doi.org/10.31223/X5XT4K

**Seibert J** (2001) On the need for benchmarks in hydrological modelling. *Hydrological Processes 15*(6), 1063–1064.

**Shen J**, **Tenenholtz N**, **Hall JB**, **Alvarez-Melis D and Fusi N** (2024) Tag-LLM: Repurposing general-purpose LLMs for specialized domains. *arXiv preprint arXiv:2402.05140*.

**Singh AK**, **Kocyigit MY**, **Poulton A**, **Esiobu D**, **Lomeli M**, **Szilvasy G and Hupkes D** (2024) Evaluation data contamination in LLMs: How do we measure it and (when) does it matter?. *arXiv preprint arXiv:2411.03923*.

**Sit M**, **Seo BC and Demir I** (2021) Iowarain: A statewide rain event dataset based on weather radars and quantitative precipitation estimation, arXiv. *arXiv preprint arXiv:2107.03432*.

**Soman S and Ranjani HG** (2024) Observations on LLMs for telecom domain: Capabilities and limitations. In *Proceedings of the Third International Conference on AI-ML Systems* (Art. No. 36, pp. 1–5). Association for Computing Machinery. https://doi.org/10.1145/3639856.3639892.

**Szymanski A**, **Ziems N**, **Eicher-Miller HA**, **Li TJJ**, **Jiang M and Metoyer RA** (2024) Limitations of the LLM-as-a-judge approach for evaluating LLM outputs in expert knowledge tasks. *arXiv preprint arXiv:2410.20266*.

https://doi.org/10.1017/eds.2025.10006 Published online by Cambridge University Press

**Thakkar M**, **More Y**, **Fournier Q**, **Riemer M**, **Chen PY**, **Zouaq A**, et al (2024) Combining domain and alignment vectors to achieve better knowledge-safety trade-offs in llms. *arXiv preprint arXiv:2411.06824*.

**Ukarande SK** (2023) *Irrigation engineering and hydraulic structures*. https://doi.org/10.1007/978-3-031-33552-5

**Vald GM**, **Sermet MY**, **Mount J**, **Shrestha S**, **Samuel DJ**, **Cwiertny D and Demir I** (2024) Integrating conversational AI agents for enhanced water quality analytics: Development of a novel data expert system. *EarthArxiv*, 7202. https://doi.org/10.31223/X51997

**Wagener T**, **Dadson SJ**, **Hannah DM**, **Coxon G**, **Beven K**, **Bloomfield JP**, et al (2021) Knowledge gaps in our perceptual model of Great Britain's hydrology. *Hydrological Processes*, *35*(7), e14288.

**Wagener T**, **Gleeson T**, **Coxon G**, **Hartmann A**, **Howden N**, **Pianosi F**, **Rahman S, Rosolem R, Stein L and Woods R** (2020) On doing large-scale hydrology with Lions: Realising the value of perceptual models andknowledge accumulation. EarthArXiv (California Digital Library). https://doi.org/10.31223/osf.io/zdy5n

**Wilkho RS**, **Chang S and Gharaibeh NG** (2024) FF-BERT: A BERT-based ensemble for automated classification of web-based text on flash flood events. *Advanced Engineering Informatics 59*, 102293.

**Xu S**, **Lu Y**, **Schoenebeck G and Kong Y** (2024a) Benchmarking LLMs' judgments with no gold standard. *arXiv preprint arXiv:2411.07127*.

**Xu B**, **Wen L**, **Li Z**, **Yang Y**, **Wu G**, **Tang X**, et al (2024b) Unlocking the potential: Benchmarking large language models in water engineering and research. *arXiv preprint arXiv:2407.21045*.

**Yan H**, **Hu X**, **Wan X**, **Huang C**, **Zou K and Xu S** (2023) Inherent limitations of LLMs regarding spatial information. *arXiv preprint arXiv:2312.03042*.

**Yan LK**, **Li M**, **Zhang Y**, **Yin CH**, **Fei C**, **Peng B**, et al (2024) Large language model benchmarks in medical tasks. *arXiv preprint arXiv:2410.21348*.

**Zafarmomen N and Samadi V** (2025) Can large language models effectively reason about adverse weather conditions?. *Environmental Modelling & Software 188*, 106421.

**Zheng F**, **Maier HR**, **Wu W**, **Dandy GC**, **Gupta HV and Zhang T** (2018) On lack of robustness in hydrological model development due to absence of guidelines for selecting calibration and evaluation data: Demonstration for data-driven models. *Water Resources Research 54*(2), 1013–1030.

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