Yao *et al. AI Agent* **2025**, *1*, 6
DOI: 10.20517/aiagent.2025.04

AI Agent

Original Article

Open Access

Knowledge-extractor: a self-evolving scientific framework for hydrogen energy research driven by AI agents

Tongao Yao[1,2], Yang Yang[3], Yujie Yan[1,2], Xinyi Ou[1,2], Mingyang Li[1,2], Chenxi Wang[1,2], Wuzhe Li[1,2], Chenghao Du[1,2], Xuqiang Shao[3], Zhengyang Gao[1,2], Weijie Yang[1,2]

[1]Department of Power Engineering, North China Electric Power University, Baoding 071003, Hebei, China.
[2]Hebei Key Laboratory of Energy Storage Technology and Integrated Energy Utilization, North China Electric Power University, Baoding 071003, Hebei, China.
[3]Department of Computer Science, North China Electric Power University, Baoding 071003, Hebei, China.

Correspondence to: Assoc. Prof. Weijie Yang, Department of Power Engineering, North China Electric Power University, Baoding 071003, Hebei, China; Hebei Key Laboratory of Energy Storage Technology and Integrated Energy Utilization, North China Electric Power University, Baoding 071003, Hebei, China. E-mail: yangwj@ncepu.edu.cn

How to cite this article: Yao, T.; Yang, Y.; Yan, Y.; Ou, X.; Li, M.; Wang, C.; Li, W.; Du, C.; Shao, X.; Gao, Z.; Yang, W. Knowledge-extractor: a self-evolving scientific framework for hydrogen energy research driven by AI agents. *AI Agent* **2025**, *1*, 6. https://dx.doi.org/10.20517/aiagent.2025.04

Received: 1 Sep 2025 **First Decision:** 11 Oct 2025 **Revised:** 15 Nov 2025 **Accepted:** 19 Nov 2025 **Published:** 15 Dec 2025

Academic Editor: Hao Li **Copy Editor:** Shu-Yuan Duan **Production Editor:** Shu-Yuan Duan

Abstract

The rapid evolution of Artificial intelligence (AI) from passive “knowledge co-pilots” to autonomous “research partners” is initiating a paradigm shift in scientific discovery, a frontier now termed Agentic Science. However, applying general-purpose AI systems to dynamic, vertically integrated domains such as hydrogen energy reveals critical limitations, including a lack of deep domain knowledge, an inability to process real-time information, and insufficient autonomous planning capabilities. To address these challenges, we introduce Knowledge-Extractor, a self-evolving scientific framework for building domain-expert AI agents, which we implement and evaluate in the hydrogen energy domain via an agent named Hydrogen-Agent. The core of our framework is a Hybrid Knowledge Integration strategy, which synergistically combines a domain-fine-tuned large language model (LLM) as its "cognitive core" with a continuously updated, non-parametric knowledge base.This architecture is augmented by an autonomous toolset comprising a PolicyRetriever (for extracting information from policy documents), a WebBrowser (for retrieving online sources), and an ArxivAnalyzer (for analyzing scientific papers from arXiv). We demonstrate that through an autonomous knowledge loop, Hydrogen-Agent overcomes the static knowledge limitations of traditional models. Our experiments validate a “specialization effect” where domain-specific fine-tuning enhances factual accuracy on our HydroBench benchmark, outperforming its base model and powerful generalist LLMs. Furthermore, three case studies illustrates the ability of the agent to autonomously conduct

complex, end-to-end research tasks, from multi-source data gathering to the generation of a strategic analysis report. Hydrogen-Agent serves as a robust prototype for future scientific agents, showcasing a viable path toward creating domain-expert AI that can accelerate discovery in critical scientific fields.

Keywords: Knowledge-extractor, self-evolving, hydrogen, large language models, multi-agent systems

INTRODUCTION

The discovery of advanced materials is essential for addressing global challenges in energy and environmental sustainability, particularly in the rapidly developing field of hydrogen energy[1-6]. Progress across the hydrogen value chain, including production catalysts, storage materials, and fuel cells, relies on the continuous discovery of novel materials. However, traditional research paradigms remain slow and resource-intensive, making it difficult to meet the increasing technological demand[7,8]. The emergence of the data-driven "fourth paradigm" of scientific discovery, guided by Artificial intelligence (AI), provides a promising new direction[9]. Large language models (LLMs) have recently been applied throughout the research process, demonstrating significant potential to accelerate materials discovery by analyzing large volumes of scientific literature[10-15]. This trend indicates a shift toward viewing AI as a cognitive core that can organize and coordinate complex research workflows.

The next stage of development, often referred to as Agentic Science, focuses on enabling AI systems to move from passive "knowledge assistants" to autonomous agents capable of planning, execution, and iterative improvement of research tasks[16]. When general-purpose AI models are applied to a specialized and dynamic domain such as hydrogen energy, several inherent limitations become apparent. First, models such as Generative Pre-trained Transformer (GPT)-4 experience a Domain Gap, as they lack a deep understanding of technical terminology and scientific principles, which often leads to factual inaccuracies[17]. Second, they suffer from a Timeliness Problem because their training data are static, and therefore they cannot adapt to rapid scientific and policy developments in hydrogen-related research[10,18]. Third, these systems face a Collaboration Challenge: even advanced tools such as ChemCrow - an AI-driven agent framework designed to assist with chemical tasks - still heavily on step-by-step human guidance and cannot autonomously integrate information from multiple sources such as policy documents, research papers, and industrial data[19-23].

To overcome these challenges, a shift toward building specialized AI agents is required. An ideal agent in the hydrogen domain should possess three essential characteristics: domain awareness, dynamic learning capability, and autonomous execution[24]. In response to these needs, we propose Knowledge-Extractor, a self-evolving scientific framework for constructing domain-specific intelligent agents. This framework is implemented through Hydrogen-Agent, designed specifically for hydrogen energy research.

The central concept of this framework is a Hybrid Knowledge Integration strategy. Hydrogen-Agent integrates the internalized knowledge of a domain-fine-tuned LLM, which serves as its cognitive core, with real-time and verifiable information obtained from a continuously updated knowledge base and a specialized toolset comprising a PolicyRetriever for extracting information from policy documents, a WebBrowser for accessing online resources, and an ArxivAnalyzer for processing scientific literature[7,25]. Through an autonomous knowledge cycle, the agent continuously gathers, validates, and refines new information, maintaining the accuracy and timeliness of its knowledge. Quantitative evaluations and detailed case studies confirm that this integrated system can perform complex analytical tasks involving multiple information sources, exceeding the capabilities of general-purpose language models.

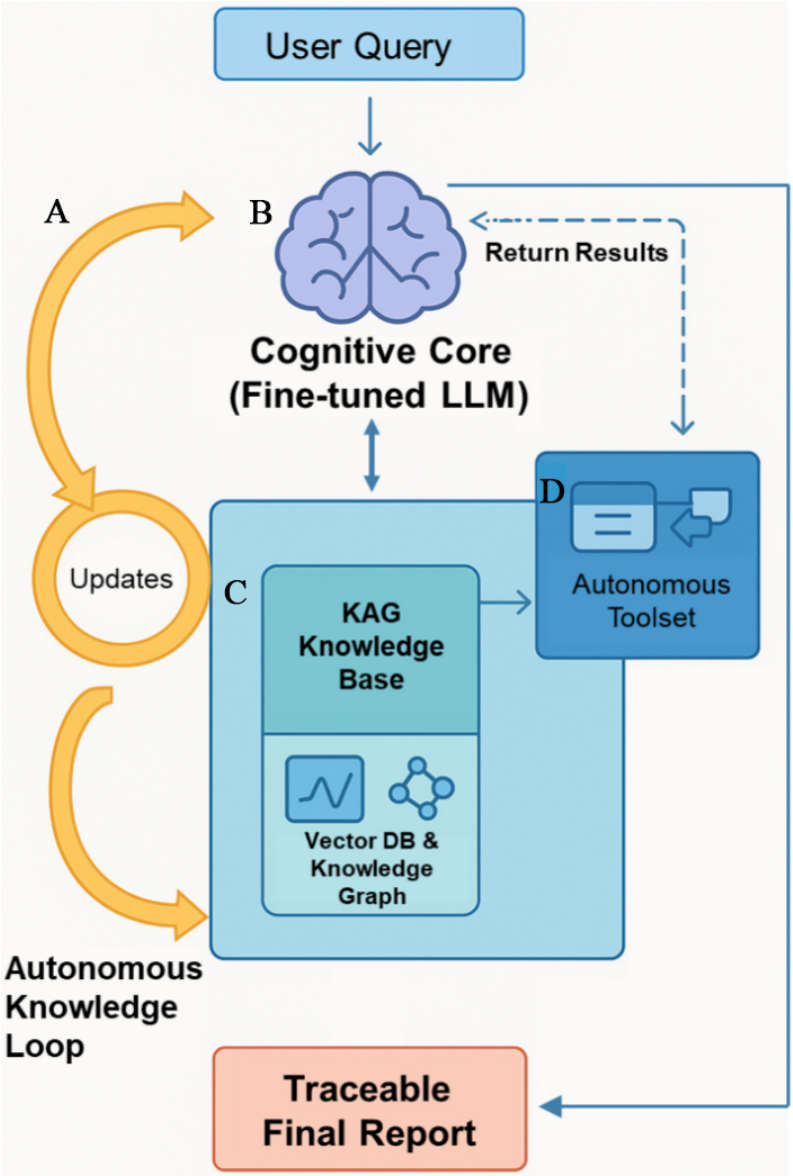

Figure 1. The Overall Architecture of Hydrogen-Agent. The framework is composed of four core modules: (A) an Autonomous Knowledge Loop that periodically ingests and validates new information; (B) a Cognitive Core, a domain-fine-tuned LLM responsible for planning and reasoning; (C) a KAG Knowledge Base that stores and retrieves both vectorized and graph-based knowledge; (D) an Autonomous Toolset containing specialized agents for executing complex sub-tasks. Solid and dashed arrows represent data and control flows, respectively. LLM: Large language model; KAG: Knowledge-Augmented Graph.

METHODS

This chapter details the architecture and core components of our proposed Knowledge-Extractor framework. To systematically address the challenges of research information processing, we designed this framework with three core design patterns: a Cognitive Core, a dynamic Knowledge Base, and an extensible Toolset. The framework further supports self-evolution through an autonomous update cycle. We demonstrate its implementation through Hydrogen-Agent, our scientific agent for the hydrogen energy domain, whose overall architecture is depicted in Figure 1.

Autonomous Knowledge Loop

To address the problem of knowledge timeliness, we designed a closed-loop, automated knowledge update process. This cycle ensures that the knowledge base of Hydrogen-Agent remains current and accurate by

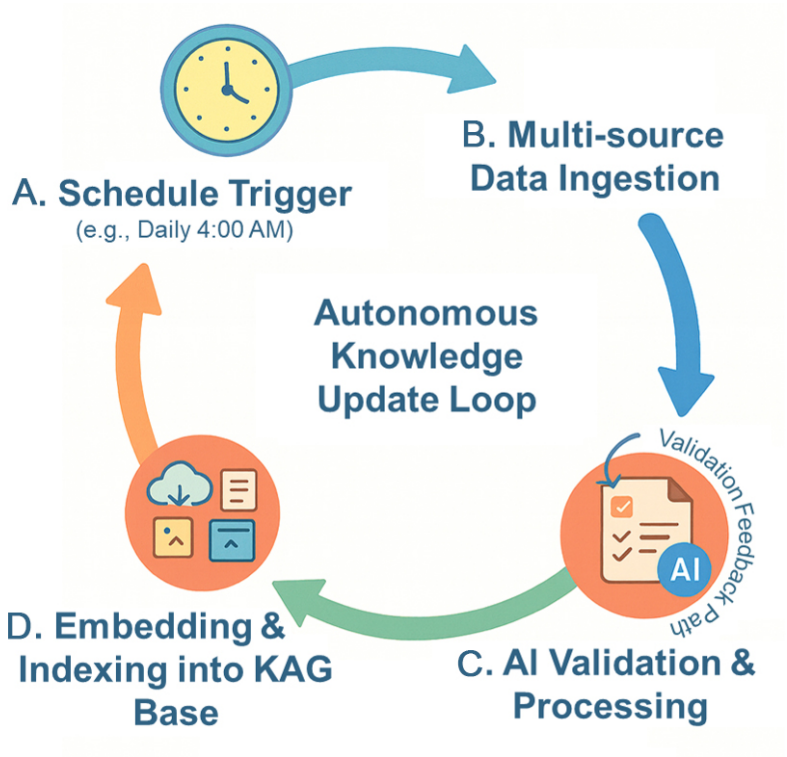

Figure 2. Workflow of the Autonomous Knowledge Update Loop. (A) A Schedule Trigger (e.g., daily at 4:00 AM) initiates the loop; (B) Multi-source Data Ingestion collects heterogeneous data from various sources; (C) AI Validation & Processing performs pre-screening, cleaning, and formatting. Feedback from this stage is used to refine the data-processing workflow, ensuring increasingly reliable updates; (D) The curated data is embedded and indexed into the KAG Knowledge Base, completing the cycle and supporting continuous knowledge updates. AI: Artificial intelligence; KAG:

periodically ingesting, validating, and integrating new knowledge from the external world. The entire workflow, which underpins the self-evolving capability of the agent, is depicted in Figure 2.

Multi-source heterogeneous data collection

To ensure the breadth and authoritativeness of the knowledge base, our automated data pipeline periodically collects information from the following four key source types:

(1) Scholarly Papers: Using application programming interfaces (APIs) from open access platforms such as arXiv, we track the latest preprints in frontier fields including electrolyzer technologies [proton exchange membrane (PEM), solid oxide electrolysis cell (SOEC)], hydrogen storage materials, and electrocatalysts in real time.

(2) Patent Databases: Through the APIs of Google Patents and other professional patent databases, we programmatically acquire structured information on hydrogen-related patents, including assignees, technical subjects, and legal statuses.

(3) Policy Documents: We deploy customized web crawlers to continuously monitor official policy portals, such as those of China, the U.S. Department of Energy (DOE), and the European Commission, to automatically scrape and parse the latest policies, regulations, industry standards, and development plans.

(4) Dynamic Web Information: For unstructured sources or those without fixed APIs, such as news and industry reports, we utilize an LLM-driven Query Generator to conduct exploratory web searches, efficiently capturing emerging technical topics and breaking industry events.

Human-in-the-loop data validation

To ensure the quality and credibility of the collected data, all information undergoes a semi-automated validation process. First, a lightweight AI model performs pre-screening of the data before ingestion. This step filters the content to ensure its relevance to the hydrogen energy domain and assesses its quality by identifying potential issues, such as missing sources, inconsistent timestamps, optical character recognition (OCR) errors, or ambiguous policy statuses. Next, data flagged as low-confidence or ambiguous triggers a manual verification process, where domain experts perform the final confirmation and correction, ensuring the accuracy and reliability of the data before it is indexed into the knowledge base.

The system is designed to operate autonomously for most inputs, with manual verification reserved for a small subset of flagged data. These flags are triggered by predefined low-confidence criteria, including: (1) structural inconsistencies (e.g., OCR errors or missing metadata); (2) semantic conflicts (e.g., new data contradicting established facts); and (3) ambiguous language in critical documents, such as policies or patents. Flagged items are periodically validated by domain experts, ensuring data reliability without creating continuous bottlenecks in the operational workflow.

AI-driven knowledge integrity assessment

To ensure the knowledge base is not only accurate but also comprehensive, the system periodically executes a knowledge integrity self-assessment:

(1) Timeliness Monitoring: Based on a timestamp mechanism, the system automatically checks the update status of each data source to ensure the knowledge base is synchronized with the latest advancements.

(2) Coverage Analysis: Utilizing an LLM, the system periodically scans the existing knowledge base, automatically analyzing the data coverage of various technical branches and policy areas through topic modeling and entity recognition, and actively identifies "Knowledge Gaps".

(3) Demand-driven Augmentation: The system analyzes anonymized interaction logs between users and the agent to mine frequently queried topics that the knowledge base does not sufficiently cover. These topics are automatically converted into new data collection tasks, thus achieving dynamic growth of the knowledge base driven by user demand.

Knowledge-Augmented Graph, KAG Base

The core knowledge storage and retrieval module of Hydrogen-Agent is the Knowledge-Augmented Graph (KAG) Knowledge Base[26], which we build and optimize based on the open source Light Retrieval-Augmented Generation (LightRAG) framework[27].

This design moves beyond a singular retrieval model, innovatively integrating the advantages of vector retrieval, knowledge graph querying, and keyword search to achieve more precise and interpretable information recall.

Hybrid Representation and Storage: When a document enters the KAG Knowledge Base, it undergoes parallel processing. On the one hand, the document is chunked and mapped into a high-dimensional vector by a Transformer encoder and stored in a vector database. On the other hand, key entities E (e.g., "PEM electrolyzer") and relations R (e.g., "uses") within the document are extracted to construct or update a knowledge graph G = (E, R).

Multi-path Retrieval Fusion: For a user query q, the system initiates three types of retrieval in parallel and calculates the final composite relevance score *S(q, d)* using a learnable weighted function[26]:

$$S(q,d) = \alpha \cdot \mathrm{sim}_{\mathrm{vec}}(q,d) + \beta \cdot \mathrm{sim}_{\mathrm{kg}}(q,d) + \gamma \cdot \mathrm{sim}_{kw}(q,d) \quad (1)$$

Here, $\mathrm{sim}_{\mathrm{vec}}$ represents vector cosine similarity, $\mathrm{sim}_{\mathrm{kg}}$ measures the path similarity of the query entity in the knowledge graph, and $\mathrm{sim}_{\mathrm{kw}}$ is the traditional keyword matching score. The weight parameters α, β, and γ are dynamically adjusted to adapt to different types of queries. This multi-path retrieval mechanism ensures that the retrieval results can capture semantic similarity while also leveraging the precision of symbolic logic.

Cognitive core: the domain-fine-tuned LLM

The "brain" of Hydrogen-Agent is a LLM that has undergone Parameter-Efficient Fine-Tuning (PEFT)[28], responsible for understanding user intent, planning tasks, selecting and invoking tools, and ultimately generating reports by synthesizing information.

Model and Methodology: The cognitive core of our framework is designed to be model-agnostic and can be implemented with various state-of-the-art LLMs. For our specific implementation, Hydrogen-Agent, we select Qwen3-8B (8 Billion parameters) (Qwen3-8B) as the base model due to its strong capabilities in understanding both Chinese and code[29]. It is important to note that the core architectural principles of our framework could be similarly realized by leveraging powerful proprietary models via their APIs, such as those from the GPT families, provided that they support fine-tuning or advanced in-context learning capabilities. The fine-tuning process utilizes the Unsloth framework, which deeply optimizes the Low-Rank Adaptation (LoRA) algorithm[30], significantly reducing the computational resources required for training without sacrificing performance. This optimized setup enabled the 8B scale model to be trained and perform inference efficiently on a single NVIDIA A6000 pro GPU.

Hybrid Instruction Dataset: To instill deep domain-specific knowledge into the model, we constructed a custom instruction dataset comprising 4,000 high-quality question-answer pairs focused on the hydrogen energy domain. This dataset was manually curated and verified by domain experts, covering a wide spectrum of topics including hydrogen production, storage, fuel cell technology, and relevant policies. The primary goal of this dataset is to enhance the factual accuracy of the model and its ability to comprehend specialized terminology. This fine-tuning approach enables the model to provide precise, domain-aware responses to specific inquiries.

The autonomous toolset

Hydrogen-Agent is equipped with a standardized set of specialized tools, which are autonomously invoked by the cognitive core according to the task plan. These tools are encapsulated as independent agents, each responsible for a specific sub-task.

PolicyRetriever (Agent 1)

PolicyRetriever is dedicated to the precise querying, comparison, and interpretation of hydrogen energy policies. When the cognitive core identifies that a task involves policy-related information, it invokes this tool. PolicyRetriever first performs an efficient search within the KAG Knowledge Base. If the retrieved information is outdated or insufficient, it autonomously activates external crawlers to obtain real-time supplementary data from authoritative policy websites and writes the updated content back into the KAG Knowledge Base, forming an intelligent closed loop of "retrieve-validate-supplement-retrieve".

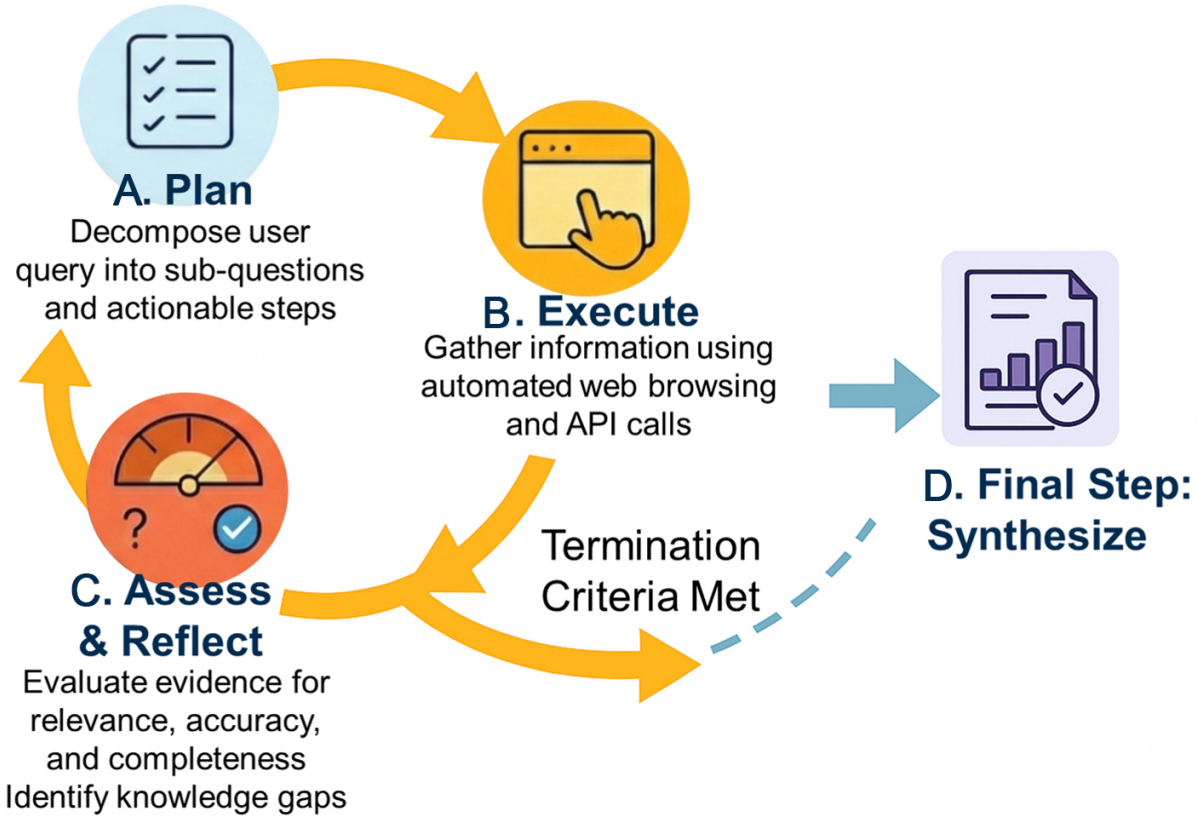

Figure 3. The Cyclic Workflow of the DeepResearchAgent. The agent operates in a core iterative loop comprising three stages: (A) Plan, where the user query is decomposed into actionable steps; (B) Execute, where information is gathered using automated tools; (C) Assess & Reflect, where the collected evidence is evaluated. This loop continues, refining the plan with each iteration, until the Termination Criteria are met. Upon completion, the agent proceeds to the (D) Final Step: Synthesize, where all verified findings are integrated into a comprehensive, citable report.

ArxivAnalyzer (Agent 2)

To solve the problem of literature information overload faced by researchers, ArxivAnalyzer automates the entire process from literature retrieval to deep question answering. It can convert a natural language query from the user into a precise arXiv API call, retrieve relevant papers, and parse their metadata and full text. Users can engage in deep, multi-turn conversations with one or more specified papers, and the agent can integrate information across paragraphs and across papers to answer complex scientific questions, greatly shortening the knowledge absorption path.

DeepResearchAgent (Agent 3)

To support systematic deep research tasks, we built DeepResearchAgent, an autonomous research proxy that uses a cyclic iterative workflow. Unlike standard agentic loops [e.g., Reasoning and Acting (ReAct)], our DeepResearchAgent enhances the 'Assess & Reflect' stage with a multi-dimensional evidence evaluation mechanism. As detailed in Figure 3, the agent assesses collected information based on relevance, authority, and timeliness, enabling it to identify subtle conflicts or knowledge gaps across disparate sources. This robust reflection process allows the agent to generate more reliable and sophisticated refinement plans for subsequent iterations. The full workflow is as follows:

(1) Plan: It receives a complex research topic from the user (e.g., "analyze the latest progress, major players, and policy support for SOEC technology") and autonomously decomposes it into a series of structured sub-questions and executable research steps.

(2) Execute: It calls a tool with integrated real browser automation capabilities to simulate the behavior of human experts for cross-source information gathering, and is capable of handling complex web scenarios such as dynamic loading and user interaction.

(3) Assess & Reflect: It self-assesses the collected information from multiple dimensions such as relevance, authority, and timeliness to identify potential conflicts or knowledge gaps among the information. If the information is insufficient or contradictory, it returns to the first step to dynamically correct or generate a new research plan, forming a core loop of "think-search-rethink".

(4) Synthesize: When all sub-questions have been answered with high confidence, or a preset stopping condition is met, the agent logically integrates all validated information to generate a structured research report with precise citations.

Performance evaluation methodology

To conduct a rigorous evaluation of Hydrogen-Agent, we designed a two-part assessment strategy. First, we quantitatively evaluate the performance of its core components, particularly the fine-tuned Cognitive Core, to validate our design choices. Second, we conduct a qualitative, end-to-end case study to demonstrate the capabilities of the fully integrated Hydrogen-Agent system on a complex, real-world task. This plan includes a specially constructed benchmark dataset and a suite of precise, automated evaluation metrics to ensure the objectivity and comparability of the results.

The hydrobench benchmark

To accurately assess the level of knowledge of the model in the hydrogen energy domain, we constructed a benchmark dataset named HydroBench. The dataset contains 648 expert-curated question-answer pairs designed to evaluate foundational domain knowledge across the hydrogen energy value chain.

Content and Purpose: The questions cover the full hydrogen energy value chain through five modules: (1) Hydrogen Classification & Production; (2) Storage & Transportation; (3) Fuel Cells & Applications; (4) Safety, Economics & Strategy; and (5) System Integration, AI & Life Cycle Assessment (LCA). This task set is used to test and quantify the parametric knowledge of the model, namely the accuracy of domain knowledge internalized through fine-tuning.

Future Extensions: The design of HydroBench is extensible. The current version serves as the Type I Foundational Knowledge benchmark, and future versions will include Type II Tool-Use Tasks and Type III Multi-Step Reasoning Tasks to evaluate tool invocation and reasoning capabilities.

Evaluation metrics

We employed text-similarity-based automated metrics to evaluate the model outputs, ensuring objectivity and comparability. Specifically, we utilized the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) and BLEU (Bilingual Evaluation Understudy) families of metrics to measure the performance of the model in terms of information coverage and linguistic precision, respectively.

ROUGE: This metric measures information coverage by calculating the n-gram overlap between the candidate answer and the reference answer, which is defined as[31]:

$$\mathrm{ROUGE}-n = \frac{\sum_{\mathrm{gram}_n \in \mathrm{Ref}} \mathrm{Count}_{\mathrm{match}}\ (\mathrm{gram}_n)}{\sum_{\mathrm{gran}_n \in \mathrm{Ref}} \mathrm{Count}\ (\mathrm{gram}_n)} \tag{2}$$

In this study, we focus on the ROUGE-1, ROUGE-2, and ROUGE-L sub-metrics. These metrics reflect matching at the unigram, bigram, and longest common subsequence levels, respectively, providing a comprehensive assessment of the content coverage of the model output.

BLEU: This metric places greater emphasis on the precision and fluency of the candidate answer. Its calculation is based on n-gram precision, incorporating a brevity penalty (BP) to prevent models from achieving artificially high scores by generating overly short answers, which can be expressed as[32]:

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} \omega_n \log p_n\right) \tag{3}$$

Here, p_n represents the n-gram precision, ω_n are the weights (typically uniform), and BP is the brevity penalty factor. We employ BLEU-4 as a primary indicator to balance precision with contextual coherence.

H-Score: To create a composite metric that balances coverage and precision, we introduce the harmonic mean, H-Score, which integrates the average scores of ROUGE and BLEU and is defined as:

$$H - \text{Score} = \frac{2 \cdot \bar{R} \cdot \bar{B}}{\bar{R} + \bar{B}} \tag{4}$$

Where $\bar{R}$ is the average model ROUGE score across all test questions, and $\bar{B}$ is its average BLEU score. The H-Score effectively prevents a high score on one metric from masking deficiencies in the other, thus providing a more accurate reflection of the overall model performance across both dimensions.

Automatic scoring of open-ended answers

In addition to rule-based metrics, we adopted an automatic scoring protocol to evaluate open-ended question-answer pairs that cannot be fully assessed through surface-level text overlap. Each model-generated answer was scored using a structured evaluation prompt covering four weighted dimensions: correctness (60%), completeness (20%), conciseness (10%), and clarity (10%).

To reduce evaluator bias, we employed a multi-evaluator ensemble strategy. Each answer was independently scored by four advanced evaluator models, namely DeepSeek-V3.2-Exp, Qwen3-Next-80B-A3B-Instruct, Kimi-K2-Instruct-0905, and GPT-4.1; the final score was obtained by averaging their outputs. A fixed temperature of 0.3 was used across all evaluators to ensure consistency, and invalid or unparseable responses were excluded. This ensemble scoring method mitigates individual model bias and enhances fairness and reproducibility. The complete scoring matrix for all evaluator–target model combinations is provided in the Supplementary Materials [Supplementary Table 1]. Detailed HydroBench data are provided in the GitHub repository (https://github.com/Weijie-Yang/Knowledge-Extractor). Version, parameter, and API details for all evaluator models are provided in the Supplementary Materials [Supplementary Table 2].

Models for comparison

To comprehensively evaluate the effectiveness of our proposed fine-tuning strategy, we benchmarked our model against a diverse set of state-of-the-art proprietary and open-source LLMs. The selected models cover a wide range of architectures and capability tiers, providing a robust and balanced basis for comparison.

The FT-Only model, a fine-tuned version of the Qwen3-8B architecture, is the primary subject of this study. Fine-tuning was conducted using the curated HydroBench instruction dataset in a zero-shot, tool-free setting. This configuration ensures that performance improvements can be attributed solely to domain-specific fine-tuning, enabling a precise quantification of the knowledge and accuracy gains achieved.

To contextualize the performance of our FT-Only model, we selected the following baselines, including both a direct comparison and advanced general-purpose LLMs. Detailed version information for all models is provided in the Supplementary Materials [Supplementary Table 3].

(1) Base Qwen3-8B (Alibaba Cloud): The original pre-trained model without fine-tuning. It serves as the direct baseline to quantify the improvements achieved by our FT-Only model.

(2) GPT-4.1 (OpenAI): A frontier proprietary model representing the current benchmark for general-purpose LLM performance.

(3) Claude-Sonnet-4-5 (Anthropic): A high-performance model with strong reasoning and language-understanding capabilities.

(4) Gemini-2.5-Flash (Google): A multimodal model optimized for high speed and reasoning efficiency.

(5) Gemini-2.5-Flash-Lite (Google): A lightweight variant of Gemini-2.5 Flash, included to analyze performance-efficiency trade-offs in streamlined architectures.

(6) DeepSeek-V3.2 (DeepSeek AI): A leading open-source model recognized for its advanced reasoning and coding capabilities.

(7) GLM-4.6 (Zhipu AI): A powerful open-source model known for its balanced performance across reasoning, language understanding, and knowledge-intensive tasks.

RESULTS AND DISCUSSION

This chapter presents our two-part evaluation of the Hydrogen-Agent framework. We first detail the quantitative performance of its fine-tuned Cognitive Core, followed by a qualitative case study that demonstrates the end-to-end capabilities of the full agentic system when powered by a frontier model such as GPT-4.

General capability evaluation via EvalScope

Before assessing domain-specific knowledge, it is essential to understand how fine-tuning impacts the general capabilities of the model. We employed the EvalScope framework to conduct a comprehensive evaluation of our fine-tuned Qwen3-8B-fine-tuned model against the Qwen3-8B base model across multiple standard benchmarks, including C-Eval, MMLU, and code generation tasks. A performance breakdown on selected key sub-tasks is presented in Figure 4.

Analysis of capability shifts

The evaluation reveals a nuanced trade-off, a common phenomenon in domain-specific fine-tuning. As shown in Table 1, our Qwen3-8B-fine-tuned model demonstrates a notable improvement in the General_qa category, rising from a score of 0.7571 to 0.8438. This indicates that training on our custom hydrogen energy dataset, which is rich in factual question-answering (Q&A), enhanced the ability of the model to follow instructions and provide accurate, concise answers in a general Q&A context. Furthermore, performance on English-language benchmarks also saw a slight improvement, from 0.7115 to 0.7389.

Conversely, the model experienced a performance degradation in areas not represented in our fine-tuning data. The most significant drop was in coding ability (Qwen3/Code), where the score decreased from

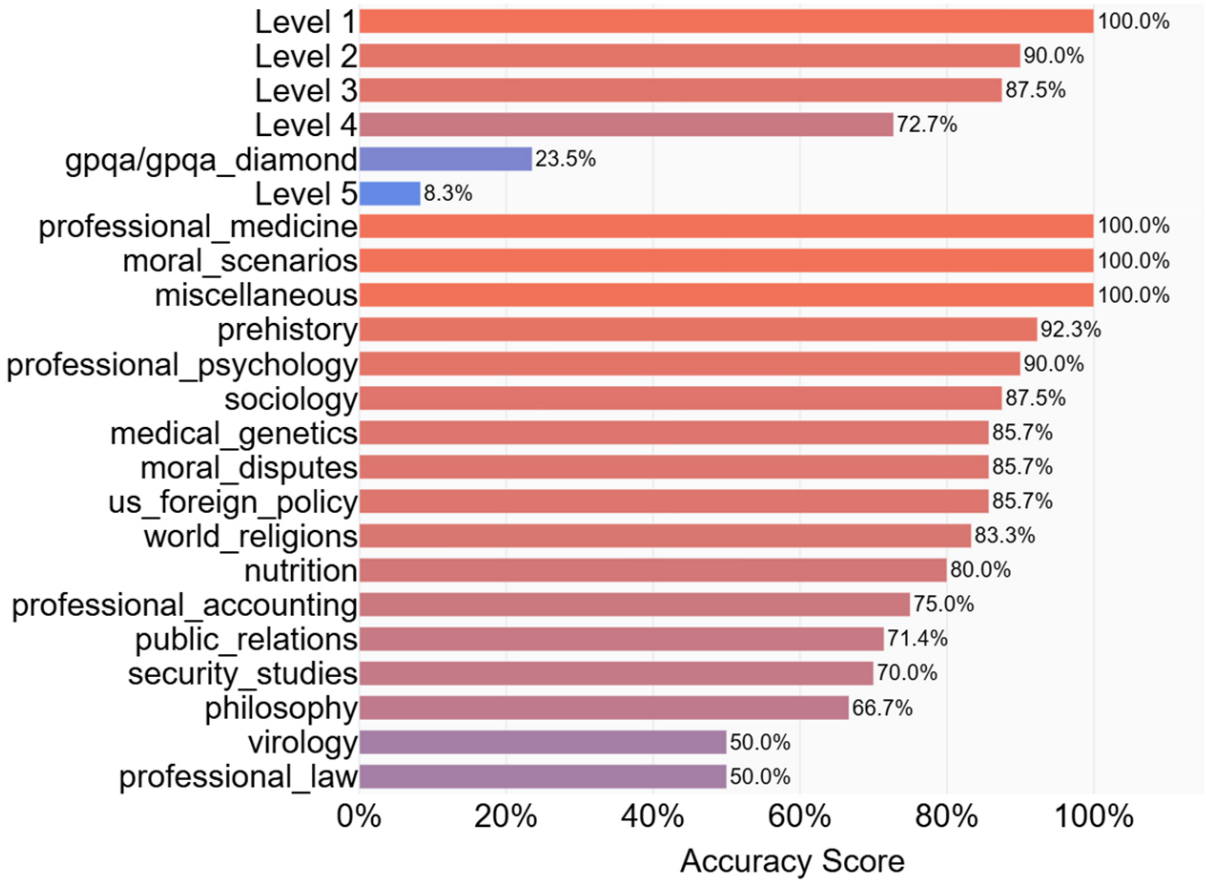

Figure 4. Fine-grained performance breakdown of the Qwen3-8B-fine-tuned Model on General Capability Benchmarks. This figure presents a detailed performance analysis of our fine-tuned model on a wide range of individual sub-tasks from the MATH and MMLU benchmarks, evaluated via the EvalScope framework. The sub-tasks are grouped into two main categories: Math&Science and English. Bars are sorted by accuracy score within each group, and their color intensity corresponds to performance (brighter red for higher scores, purple/blue for lower scores). The results clearly illustrate the "specialization effect", showcasing strong model performance on most English sub-tasks and foundational math (Levels 1-3), while highlighting performance degradation on highly specialized, out-of-domain tasks such as advanced math (Level 5) and professional_law.

Table 1. General capability scores of the base and fine-tuned models

Model	Chinese	Code	English	Math & Science	General_qa
Qwen3-8B	0.7388	0.2857	0.7115	0.6428	0.7571
Qwen3-8B-fine-tuned	0.7836	0.2381	0.7389	0.5000	0.8438

0.2857 to 0.2381. A smaller decline was also observed in Math & Science (Qwen3/Math&Science), from 0.6428 to 0.5000. This phenomenon, often referred to as "catastrophic forgetting", is expected. It highlights that specialized fine-tuning focuses the capacity of the model on the target domain at the expense of unrelated skills. This trade-off is acceptable and even desirable for our purpose, as the goal of Hydrogen-Agent is not to create a generalist but a domain expert.

Domain-specific knowledge evaluation (RQ1: fine-tuning effectiveness)

To evaluate the effectiveness of domain-specific fine-tuning, we tested all models on the HydroBench foundational knowledge QA task, which comprises 648 expert-curated questions covering the entire hydrogen energy value chain. The fine-tuning process was stable, with the training loss decreasing smoothly, the gradient norm remaining well controlled, and the learning rate following the expected linear decay schedule; full training curves are provided in the Supplementary Materials [Supplementary Figure 1].

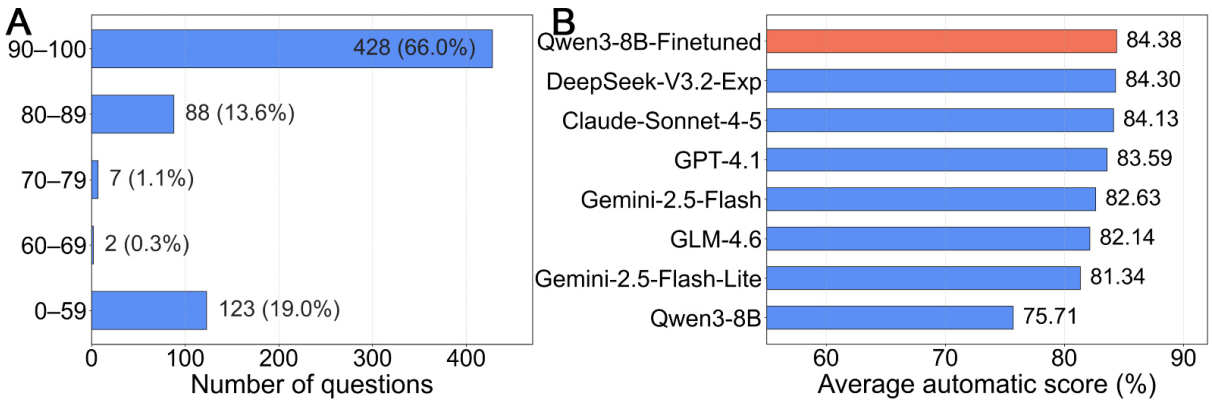

Figure 5. Comparative performance of evaluated models on HydroBench. (A) Score distribution of the FT-Only model (Qwen3-8B-fine-tuned) over all 648 test questions, showing that most answers are concentrated in the 90%-100% range; (B) Overall comparison of models using the harmonic mean (H-Score), highlighting the advantage of the fine-tuned model over both its base model and strong general-purpose LLMs. LLM: Large language model; FT: fine-tuned.

Table 2. Quantitative comparison of models on the HydroBench foundational knowledge QA task

Rank	Model	Category	H-Score (%)
1	FT-Only (Qwen3-8B-Fine-tuned)	Fine-tuned	84.38
2	DeepSeek-V3.2-Exp	Open-source	84.30
3	Claude-Sonnet-4-5	Proprietary	84.13
4	GPT-4.1	Proprietary	83.59
5	Gemini-2.5-Flash	Proprietary	82.63
6	GLM-4.6	Open-source	82.14
7	Gemini-2.5-Flash-Lite-preview-06-17	Proprietary	81.34
8	Base Qwen (Qwen3-8B)	Open-source	75.71

QA: Question Answering.

All evaluations were performed in a zero-shot, tool-free setting to measure the models' internalized, parametric knowledge. The assessment combined automatic scoring and text-similarity metrics, as described in Sections *Evaluation metrics* and *Automatic scoring of open-ended answers*. Automatic scoring was applied to open-ended responses using a four-model evaluator ensemble, and the resulting averaged scores were used for all subsequent analyses. The harmonic-mean H-Score quantified overall factual accuracy and coverage.

Quantitative results on HydroBench

The quantitative results are summarized in Table 2 and visualized in Figure 5. All scores reported here are based on the average output of the four evaluator models described in Section *Automatic scoring of open-ended answers*. As shown in Figure 5A, the automatic scores for the FT-Only model (Qwen3-8B-fine-tuned, n = 648) are mainly distributed in the 90%-100% range, with 19% of answers falling below 60%. Figure 5B presents the H-Score comparison across evaluated models, confirming the superiority of the fine-tuned model over both its base model and strong general-purpose LLMs. The raw evaluator scores and ensemble-averaged values used to compute the H-Scores are listed in the Supplementary Materials [Supplementary Table 1].

These findings demonstrate that domain-specific fine-tuning effectively transforms a general LLM into a reliable domain expert. The improvement is especially evident in factual recall and terminology precision,

two capabilities that are critical for scientific and industrial applications in hydrogen energy. The next section explores how this strengthened knowledge foundation enables the full Hydrogen-Agent system to perform complex, multi-step reasoning and report synthesis.

Case studies: demonstrating the framework's full potential with a frontier model

While our quantitative results confirm that the FT-Only model possesses superior domain-specific knowledge, the role of a cognitive core in an agentic system extends beyond mere knowledge recall to encompass complex planning and reasoning. To provide a clear illustration of the operational mechanism and end-to-end capabilities of the Hydrogen-Agent framework when powered by a frontier reasoning engine, we conducted a series of detailed case studies. For this demonstration, we instantiated the agent using GPT-4 as its cognitive core. We chose GPT-4 not for its domain knowledge (where our FT-Only model excels), but for its state-of-the-art general reasoning and planning abilities, which are critical for orchestrating the multi-step, multi-tool workflow. This allows us to showcase the upper bound of performance and the ideal workflow our architecture can achieve.

To demonstrate the framework's robustness and general applicability, we designed three distinct case studies, each targeting a different type of complex research task within the hydrogen value chain: (1) a deep technical analysis of SOEC technology; (2) a market competitiveness report on hydrogen heavy-duty trucks; and (3) a supply chain and patent risk assessment for Type IV hydrogen storage tanks.

Case study 1: deep technical analysis of SOEC technology

We first tracked the agent's entire process in completing a complex research task: "Generate a comprehensive report on the latest advancements, key patent holders, and supporting policies for SOEC technology." The detailed workflow for this task, illustrating the interplay between tool invocation and insight generation, is depicted in Figure 6.

This case study demonstrates a process far exceeding simple information retrieval. The agent autonomously decomposed the primary task into a multi-faceted investigation comprising academic literature scanning, corporate intelligence gathering, and governmental funding analysis. It sequentially invoked its specialized tools: ArxivAnalyzer to distill key findings from recent scientific papers on SOEC performance; DeepResearchAgent to identify and profile global innovators such as Topsoe, thyssenkrupp nucera, and Elcogen; and PolicyRetriever to extract precise details from government funding announcements, including the U.S. DOE FOA DE-FOA-0003366 and the EU SYRIUS project under the Innovation Fund. Importantly, a video demonstration of the automated data extraction process has been provided in the Supplementary Materials (see Supplementary Video 1).

Critically, the agent transitioned from data collection to higher-order cognitive synthesis. Instead of merely listing facts, it independently analyzed and cross-referenced the multi-source information to perform trend analysis, identifying distinct technological and industrialization rhythms between Europe and North America. Based on these synthesized insights, it formulated a set of actionable strategic recommendations, including a "two-step collaborative entry" strategy and a risk mitigation plan for supply chain dependency. The agent concluded by generating a structured, multi-section report complete with a suggested project roadmap and an honest assessment of its own operational limitations [e.g., failure to log specific paper titles due to Completely Automated Public Turing test to tell Computers and Humans Apart (CAPTCHA) restrictions].

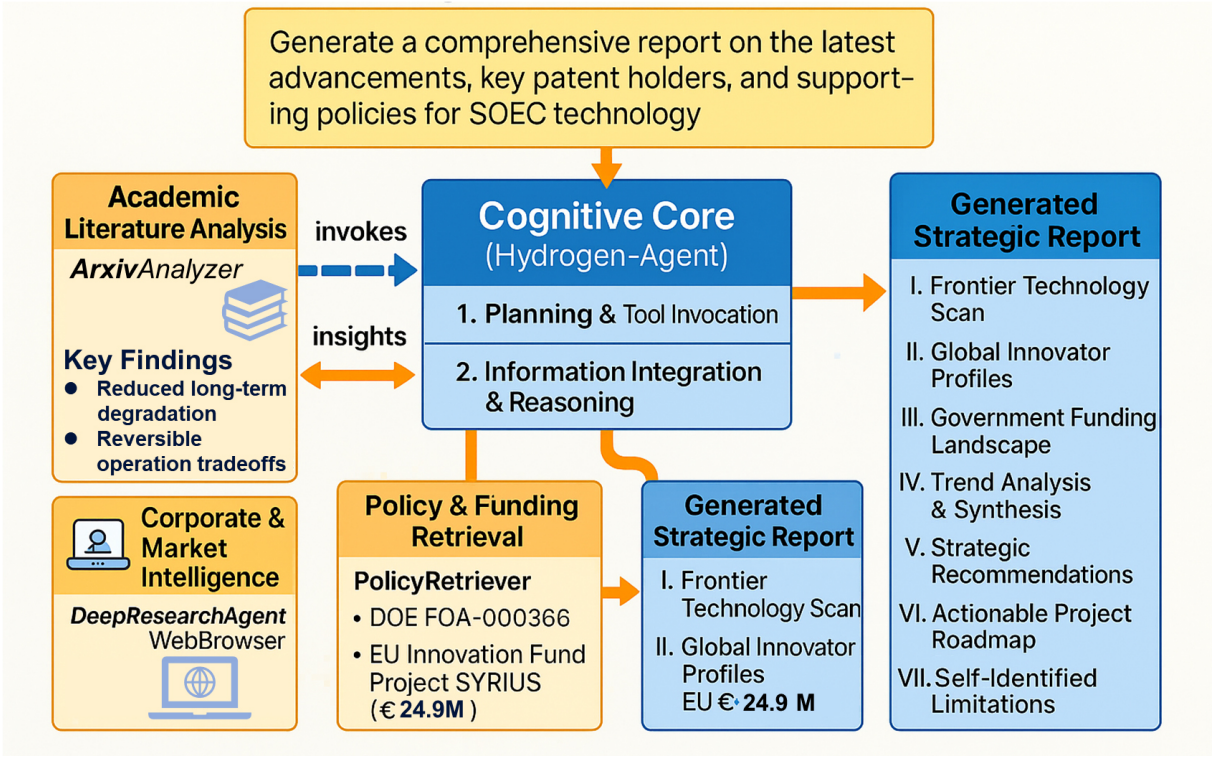

Figure 6. Detailed workflow of an end-to-end strategic analysis task. This figure illustrates the mechanistic workflow of Hydrogen-Agent using the SOEC technology analysis as a representative example. The process begins with the Cognitive Core planning the task. It then sequentially invokes (dashed arrows) its specialized toolset to orchestrate a multi-stage information gathering process, covering (a) academic literature, (b) corporate intelligence, and (c) policy and funding data. Each tool returns structured insights (solid arrows) to the Cognitive Core. After integrating and reasoning over this multi-source information, the agent synthesizes the findings to generate the final multi-section strategic report, complete with trend analysis and actionable recommendations. SOEC: Solid oxide electrolysis cell.

Case study 2: market competitiveness and policy analysis

To test the agent's capabilities in strategic and market analysis, we assigned it the task of generating a competitiveness report on the Chinese hydrogen fuel cell heavy-duty truck market for a potential investor. This required synthesizing technology trends, market players, and complex policy landscapes.

In this scenario, Hydrogen-Agent demonstrated its ability to act as a market analyst. It systematically invoked ArxivAnalyzer to validate technical claims on high-power fuel cell stacks, used DeepResearchAgent to identify key market players (e.g., FAW Jiefang, a Chinese commercial vehicle manufacturer, and Sinotruk, another leading Chinese truck company) and the performance metrics of their flagship models, and queried PolicyRetriever to extract specific subsidy and non-subsidy support measures from national and regional governments. The resulting report successfully integrated these disparate data streams into a coherent analysis, identifying key Total Cost of Ownership (TCO) advantages in demonstration zones while also highlighting critical challenges such as hydrogen cost and infrastructure gaps. A video demonstration of the agent's autonomous workflow for this market analysis task is available in the Supplementary Materials (see Supplementary Video 2).

Case study 3: supply chain and patent risk assessment

Finally, to evaluate its capacity for nuanced risk assessment, we tasked the agent with evaluating the investment opportunities and risks in China's Type IV high-pressure hydrogen storage tank industry chain, focusing on technology dependence, patent barriers, and regulatory standards.

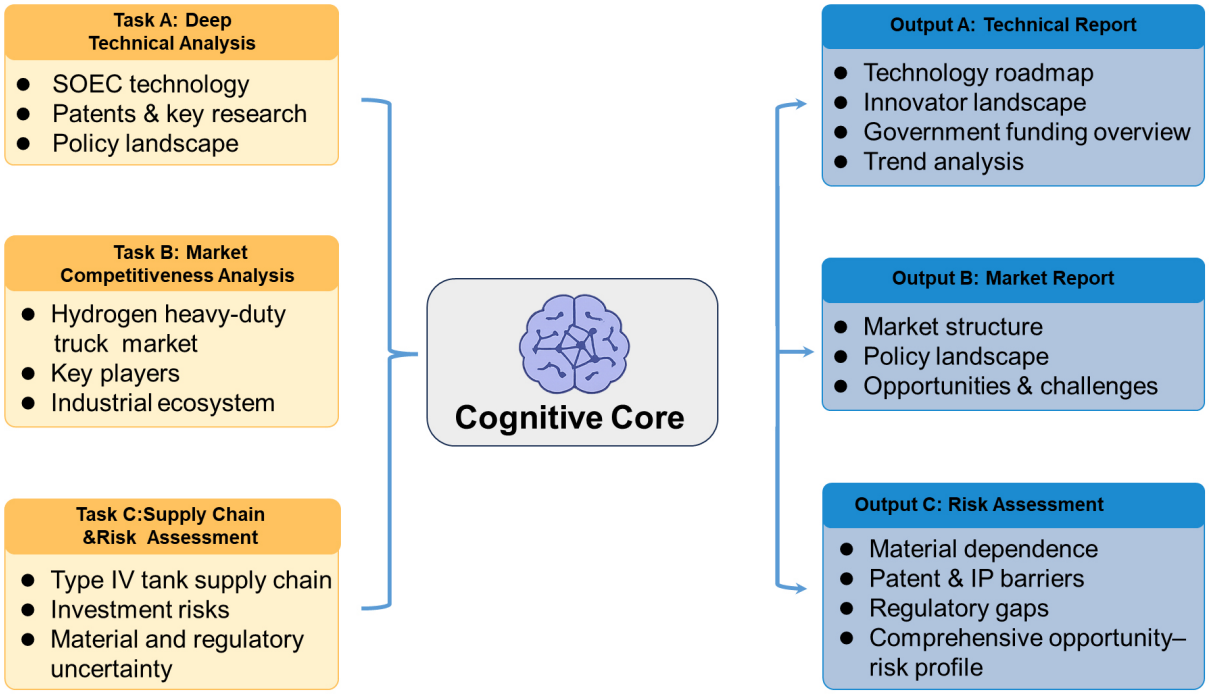

Figure 7. Versatile Task Processing Pipeline of the Hydrogen-Agent Framework. The system accepts diverse user tasks in technical analysis, market evaluation, and supply-chain risk assessment. The Cognitive Core interprets each task and integrates multi-source information to produce targeted analytical reports.

This task required the agent to perform a multi-dimensional analysis. It integrated market growth data and supply chain analysis from web sources (identifying heavy reliance on international carbon fiber suppliers), identified key patent holders globally and domestically (e.g., Hexagon Purus, a Norway-based manufacturer of hydrogen storage systems, and Sinoma Science & Technology, a Chinese company specializing in composite materials), and compared national standards (e.g., GB/T 35544, China's standard for vehicle-mounted compressed hydrogen storage cylinders) with international regulations (GTR No.13, global technical regulation on hydrogen fuel cell vehicles) by invoking DeepResearchAgent and PolicyRetriever. The agent's final output was not a simple data dump, but a structured investment memo that balanced opportunities (e.g., domestic substitution) against significant risks (e.g., technology blockade, high costs), showcasing its capacity for sophisticated, evaluative reasoning. The agent's process for this risk assessment task is demonstrated in the Supplementary Materials (see Supplementary Video 3).

Taken together, these three case studies validate the ability of the Hydrogen-Agent framework to function as a versatile and autonomous research partner. It successfully executes diverse workflows that span from deep scientific literature reviews and market intelligence gathering to complex policy and supply chain risk analysis. This demonstrates that the architecture can effectively transform raw, multi-source data into strategic, forward-looking insights across the entire hydrogen value chain, as visually summarized in Figure 7.

Discussion

Our two-part evaluation, which separately assessed domain knowledge and end-to-end agentic performance, reveals an important insight for the development of scientific agents: there is a distinction between a knowledge expert and a master orchestrator.

First, the superior performance of our FT-Only model on HydroBench [Figure 5] confirms that targeted

fine-tuning is an effective strategy for creating a knowledge expert. For tasks that require high factual accuracy, this specialized model is more reliable than larger general-purpose models. The specialization effect shown in our evaluation results [Figures 4 and 5] indicates that this expertise comes at the expense of out-of-domain skills, creating a model that is deep but narrow.

The case studies [Figures 6 and 7] further demonstrate that a successful agentic system also requires a master orchestrator, a cognitive core with strong reasoning and planning capabilities. While the FT-Only model possesses greater domain knowledge, GPT-4 currently represents the state of the art in coordinating complex, multi-step workflows. The success of the GPT-4-based case studies illustrates the effectiveness of our Hybrid Knowledge Integration architecture in enhancing the capabilities of any reasoning engine by combining dynamic and verifiable knowledge with a specialized toolset.

To evaluate the framework's practical utility, its analytical outputs were informally reviewed by specialists in hydrogen materials science. Their feedback indicated that the synthesized analyses on topics such as technology assessment, market competitiveness, and hydrogen storage policy were consistent with current research trends. This suggests that the framework can serve as a useful tool for researchers to identify promising directions and prioritize experiments. Future work will include laboratory-level validation to further connect the framework's insights with practical scientific discovery.

These findings lead to our central conclusion: the optimal path to robust agency lies in combining the strengths of both components. An ideal future scientific agent may integrate a fine-tuned model for deep knowledge with a generalist model for reasoning and planning, or future foundational models may naturally evolve to excel at both. The Knowledge-Extractor framework provides a practical blueprint for achieving such integration, demonstrating how to coordinate parametric knowledge, external information, and autonomous tools.

Recent advances such as ChemCrow[19] and SciAgents[7] have demonstrated the potential of agentic science for autonomous discovery. Building upon these pioneering systems, our Knowledge-Extractor framework focuses on a complementary challenge: maintaining sustained domain expertise in rapidly evolving fields. In areas such as hydrogen energy, where new research papers, patents, and policies emerge almost daily, continuous learning and adaptation are essential. The Autonomous Knowledge Loop introduced in this work provides a useful approach toward this goal, enabling ongoing knowledge renewal and helping the framework remain accurate and relevant over time.

CONCLUSION

In this work, we presented Hydrogen-Agent, a self-evolving scientific agent designed to address the unique challenges of the hydrogen energy domain. By implementing a novel Hybrid Knowledge Integration framework, we successfully combined the deep, internalized knowledge of a domain-fine-tuned language model with the real-time, verifiable information from a dynamically updated knowledge base and a specialized toolkit.

Our quantitative evaluations revealed two key findings. First, we identified and validated a clear specialization effect: domain-specific fine-tuning significantly enhances model expertise and factual accuracy on the HydroBench benchmark, establishing its superiority over both the base model and larger, general-purpose models in this context. This improvement comes at the acceptable cost of reduced performance on unrelated, out-of-domain tasks. Second, the end-to-end case studies, which used GPT-4 to showcase the framework's peak capabilities, demonstrated the effectiveness of the integrated agentic

architecture. The Knowledge-Extractor framework autonomously executed complex research workflows by coordinating its specialized components and successfully transformed raw, multi-source data into actionable insights.

The development of Hydrogen-Agent confirms that the transition from passive language models to autonomous, domain-expert agents is a promising strategy for advancing AI for Science. This work provides not only a practical tool for the hydrogen energy sector but also a replicable architectural blueprint for building specialized scientific agents in other fast-evolving research fields.

Limitations and future work

While Hydrogen-Agent, as the first full implementation of the Knowledge-Extractor framework, shows potential for developing self-evolving scientific agents, several current limitations remain and define our next research steps.

Noisy and conflicting data

Although human-in-the-loop validation improves data reliability, the current autonomous modules have a limited ability to resolve inconsistencies across heterogeneous sources. For instance, policy documents, patents, and research articles may report contradictory information about similar technologies. Future work will explore automated approaches for data provenance tracking, fact consistency checking, and conflict resolution based on source credibility.

The agent-to-lab gap

The framework performs well in knowledge synthesis and report generation but remains disconnected from experimental and simulation workflows. Integrating laboratory planning platforms and computational tools such as density functional theory (DFT) simulators will be an important step toward linking digital analysis with real-world experimentation.

Dependence on the cognitive core

The overall analytical quality depends on the reasoning capacity of the underlying language model that serves as the cognitive core. Even advanced models can occasionally misinterpret complex problems or overlook subtle relationships between information sources. Continued progress will depend on advances in foundational models and adaptive reasoning architectures.

Our ongoing research aims to address these challenges by improving reasoning mechanisms, integrating simulation-based validation, and conducting user studies with domain experts to evaluate the framework's practical value in real laboratory contexts.

DECLARATIONS

Authors' contributions

Conceptualization, supervision, funding acquisition, writing - original draft, writing - review & editing: Yang, W.; Yao, T.

Data curation, methodology, software, validation: Yao, T.; Yang, Y.; Yan, Y.; Ou, X.

Investigation, formal analysis, visualization: Yang, Y.; Shao, X.; Li, M.; Wang, C.

Resources, project administration: Gao, Z.; Yang, W.

Software, validation, data support: Li, W.; Du, C.

Availability of data and materials

The fine-tuned Qwen3-8B model weights are publicly available on Hugging Face at https://www.modelscope.cn/models/Yangyang245/qwen8b All relevant code is publicly available in the GitHub repository (https://github.com/Weijie-Yang/Knowledge-Extractor).

Financial support and sponsorship

This work was funded by the Natural Science Foundation of Hebei (E2023502006), the Fundamental Research Funds for the Central Universities, China (grant number 2025JC008), and the Fundamental Research Funds for the Central Universities (2025MS131).

Conflicts of interest

Yang, W. is an Associate Editor of the journal *AI Agent*. Yang, W. was not involved in any steps of the editorial process, including reviewers' selection, manuscript handling, or decision-making. The other authors declare that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

REFERENCES

1. Zhou, P.; Zhou, Q.; Xiao, X.; et al. Machine learning in solid-state hydrogen storage materials: challenges and perspectives. *Adv. Mater.* **2025**, *37*, e2413430. DOI PubMed
2. Jia, X.; Wang, T.; Zhang, D.; et al. Advancing electrocatalyst discovery through the lens of data science: state of the art and perspectives. *J. Catal.* **2025**, *447*, 116162. DOI
3. Li, C.; Yang, W.; Liu, H.; et al. Picturing the gap between the performance and US-DOE's hydrogen storage target: a data-driven model for MgH_2 dehydrogenation. *Angew. Chem. Int. Ed. Engl.* **2024**, *63*, e202320151. DOI PubMed
4. Li, F.; Liu, D.; Sun, K.; et al. Towards a future hydrogen supply chain: a review of technologies and challenges. *Sustainability* **2024**, *16*, 1890. DOI
5. Chen, K.; Lau, M. Y.; Luo, X.; Huang, J.; Ouyang, L.; Yang, X. Research progress in solid-state hydrogen storage alloys: a review. *J. Mater. Sci. Technol.* **2026**, *246*, 256-89. DOI
6. Cai, J.; Jiang, Y.; Yao, T.; et al. A demand-driven dynamic heating strategy for ultrafast and energy-efficient MgH_2 dehydrogenation utilizing the "burst effect". *J. Energy. Storage.* **2025**, *130*, 117495. DOI
7. Ghafarollahi, A.; Buehler, M. J. SciAgents: automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Adv. Mater.* **2025**, *37*, e2413523. DOI
8. Chen, X.; Yi, H.; You, M.; et al. Enhancing diagnostic capability with multi-agents conversational large language models. *NPJ. Digit. Med.* **2025**, *8*, 159. DOI PubMed PMC
9. Zhang, D.; Jia, X.; Hung, T. B.; et al. "DIVE" into hydrogen storage materials discovery with AI agents. *arXiv* **2025**, arXiv:2508.13251. Available online: https://doi.org/10.48550/arXiv.2508.13251 (accessed 9 December 2025).
10. Chen, Q.; Yang, M.; Qin, L.; et al. AI4Research: a survey of artificial intelligence for scientific research. *arXiv* **2025**, arXiv:2507.01903. Available online: https://doi.org/10.48550/arXiv.2507.01903 (accessed 9 December 2025).
11. Jia, S.; Zhang, C.; Fung, V. LLMatDesign: autonomous materials discovery with large language models. *arXiv* **2024**, arXiv:2406.13163. Available online: https://doi.org/10.48550/arXiv.2406.13163 (accessed 9 December 2025).
12. Kang, Y.; Kim, J. ChatMOF: an artificial intelligence system for predicting and generating metal-organic frameworks using large language models. *Nat. Commun.* **2024**, *15*, 4705. DOI

13. Niyongabo Rubungo, A.; Arnold, C.; Rand, B. P.; Dieng, A. B. LLM-Prop: predicting the properties of crystalline materials using large language models. *NPJ. Comput. Mater.* **2025**, *11*, 186. DOI
14. Lohana Tharwani, K .K.; Tharwani, L.; Kumar, R.; Sumita, Ahmed, N.; Tang, T. Large language models transform organic synthesis from reaction prediction to automation. *arXiv* **2025**, arXiv:2508.05427. Available online: https://doi.org/10.48550/arXiv.2508.05427 (accessed 9 December 2025).
15. Yao, T.; Yang, Y.; Cai, J.; et al. From LLM to agent: a large-language-model-driven machine learning framework for catalyst design of MgH_2 dehydrogenation. *J. Magnes. Alloys.* **2025**, S2213956725002853. DOI
16. Wei, J.; Yang, Y.; Zhang, X.; et al. From AI for science to agentic science: a survey on autonomous scientific discovery. *arXiv* **2025**, arXiv:2508.14111. Available online: https://doi.org/10.48550/arXiv.2508.14111 (accessed 9 December 2025).
17. Zhang, Y.; Khan, S. A.; Mahmud, A.; et al. Exploring the role of large language models in the scientific method: from hypothesis to discovery. *NPJ. Artif. Intell.* **2025**, *1*, 14. DOI
18. Xu, W.; Liang, Z.; Mei, K.; Gao, H.; Tan, J.; Zhang, Y. A-MEM: agentic memory for LLM agents. *arXiv* **2025**, arXiv:2502.12110. Available online: https://doi.org/10.48550/arXiv.2502.12110 (accessed 9 December 2025).
19. M Bran, A.; Cox, S.; Schilter, O.; Baldassari, C.; White, A. D.; Schwaller, P. Augmenting large language models with chemistry tools. *Nat. Mach. Intell.* **2024**, *6*, 525-35. PubMed PMC
20. Gridach, M.; Nanavati, J.; Zine El Abidine, K.; Mendes, L.; Mack, C. Agentic AI for scientific discovery: a survey of progress, challenges, and future directions. *arXiv* **2025**, arXiv:2502.12110. Available online: https://doi.org/10.48550/arXiv.2503.08979 (accessed 9 December 2025).
21. Ghafarollahi, A.; Buehler, M. J. ProtAgents: protein discovery via large language model multi-agent collaborations combining physics and machine learning. *arXiv* **2024**, arXiv:2402.04268. Available online: https://doi.org/10.48550/arXiv.2402.04268 (accessed 9 December 2025).
22. Ghafarollahi, A.; Buehler, M. J. AtomAgents: alloy design and discovery through physics-aware multi-modal multi-agent artificial intelligence. *arXiv* **2024**, arXiv:2407.10022. Available online: https://doi.org/10.48550/arXiv.2407.10022 (accessed 9 December 2025).
23. Ni, B.; Buehler, M. J. MechAgents: large language model multi-agent collaborations can solve mechanics problems, generate new data, and integrate knowledge. *Extreme. Mech. Lett.* **2024**, *67*, 102131. DOI
24. Lu, C.; Lu, C.; Lange, R. T.; Foerster, J.; Clune, J.; Ha, D. The AI scientist: towards fully automated open-ended scientific discovery. *arXiv* **2024**, arXiv:2408.06292. Available online: https://doi.org/10.48550/arXiv.2408.06292 (accessed 9 December 2025).
25. Robson, M. J.; Xu, S.; Wang, Z.; Chen, Q.; Ciucci, F. Multi-agent-network-based idea generator for zinc-ion battery electrolyte discovery: a case study on zinc tetrafluoroborate hydrate-based deep eutectic electrolytes. *Adv. Mater.* **2025**, *37*, e2502649. PubMed
26. Liang, L.; Sun, M.; Gui, Z.; et al. KAG: boosting LLMs in professional domains via knowledge augmented generation. *arXiv* **2024**, arXiv:2409.13731. Available online: https://doi.org/10.48550/arXiv.2409.13731 (accessed 9 December 2025).
27. Guo, Z., Xia, L.; Yu, Y.; Ao, T.; Huang C.; et al. LightRAG: simple and fast retrieval-augmented generation. *arXiv* **2025**, arXiv:2410.05779. Available online: https://doi.org/10.48550/arXiv.2410.05779 (accessed 9 December 2025).
28. Han, Z.; Yang, Z.; Huang, Y.; et al. Parameter-efficient fine-tuning for large models: a comprehensive survey. *arXiv* **2024**, arXiv:2403.14608. Available online: https://doi.org/10.48550/arXiv.2403.14608 (accessed 9 December 2025).
29. Yang, A.; Li, A.; Yang, B.; et al. Qwen3 technical report. *arXiv* **2025**, arXiv:2505.09388. Available online: https://doi.org/10.48550/arXiv.2505.09388 (accessed 9 December 2025).
30. Mao, Y.; Yuhang Ge, Y.; Fan, Y.; et al. A survey on LoRA of large language models. *arXiv* **2024**, arXiv:2407.11046. Available online: https://doi.org/10.48550/arXiv.2407.11046 (accessed 9 December 2025).
31. Lin, C. -Y. ROUGE: a package for automatic evaluation of summaries. In *ACL-04 Workshop on Text Summarization Branches Out*, *Text Summarization Branches Out*, Barcelona, Spain, July 25-26, 2004; Association for Computational Linguistics: Stroudsburg, USA, 2004; pp 74-81. https://aclanthology.org/W04-1013/ (accessed 2025-12-15).
32. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. -J. Bleu: a method for automatic evaluation of machine translation. In *40th Annual Meeting of the Association for Computational Linguistics, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, July 6-12 2002; Philadelphia, PA, USA; Association for Computational Linguistics: Stroudsburg, USA; 2002. pp 311-8. DOI