

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

Manuscript Author:

Tongao Yao^{1,2}, Yang Yang³, 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 Shao³, Zhengyang Gao^{1,2}, Weijie Yang^{1,2}

¹Department of Power Engineering, North China Electric Power University, Baoding 071003, Hebei, China.

²Hebei Key Laboratory of Energy Storage Technology and Integrated Energy Utilization, North China Electric Power University, Baoding 071003, Hebei, China.

³Department of Computer Science, North China Electric Power University, Baoding 071003, Hebei, China.

Correspondence to: 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

Supplementary Tables

Supplementary Table S1. Multi-Evaluator Scores for All Models (Raw + Averaged)

Evaluator → / Model ↓	Claude- Sonnet- 4.5	Gemini- 2.5- Flash- Lite	GPT- 4.1	Qwen- 3-8B	DeepSee k-V3.2	Gemini- 2.5-Flash	Qwen3-8B- finetuned	GLM- 4.6
DeepSeek- V3.2-Exp	90.73	88.25	92.32	77.82	92.81	92.01	91.69	88.90
GPT-4.1	92.05	88.38	91.14	84.98	92.03	90.96	91.09	87.64

Kimi-K2	92.02	86.06	90.87	81.14	93.07	90.55	92.63	88.60
Qwen3-Next-80B	61.72	62.68	60.03	58.88	59.31	57.00	62.12	63.41
Sum	336.52	325.37	334.35	302.82	337.22	330.52	337.53	328.56
Average H-Score	84.13	81.34	83.59	75.71	84.30	82.63	84.38	82.14

Supplementary Table 2. Model and version information for baseline models

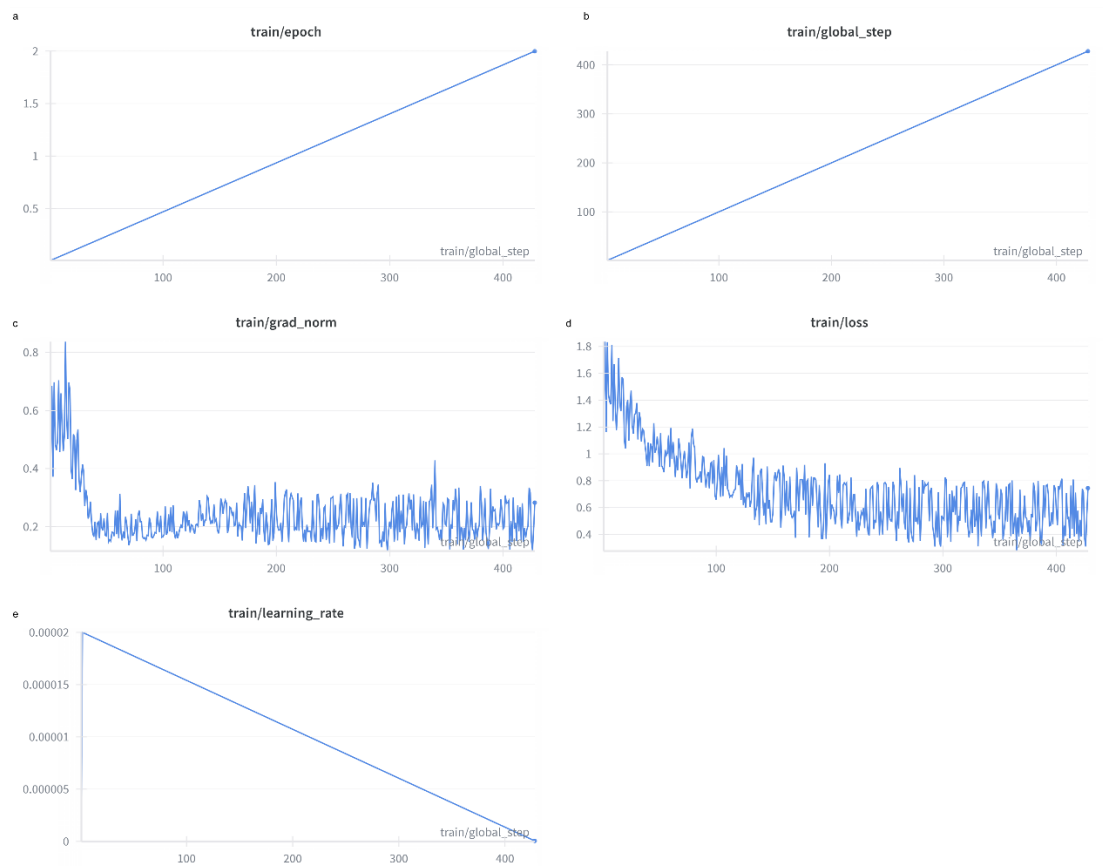
Model Name	Category	Model Provider	Version / Release Tag
FT-Only (Qwen3-8B-finetuned)	Fine-tuned (ours)	Alibaba Qwen Team	Qwen3-8B base + our HydroBench fine-tuning
DeepSeek-V3.2-Exp	Open-source	DeepSeek AI	V3.2 Experimental
Claude-Sonnet-4-5	Proprietary	Anthropic	Sonnet 4.5 (2025 release)
GPT-4.1	Proprietary	OpenAI	GPT-4.1 (2025 frontier version)
Gemini-2.5-Flash	Proprietary	Google DeepMind	Flash 2.5 (2025)
GLM-4.6	Open-source	Zhipu AI	GLM-4.6 (2025)
Gemini-2.5-Flash-Lite-preview-06-17	Proprietary	Google DeepMind	preview-06-17
Base Qwen (Qwen3-8B)	Open-source	Alibaba Qwen Team	Qwen3-8B base model

Supplementary Table 3. Evaluator models used in the multi-model ensemble scoring protocol described in Section *Automatic scoring of open-ended answers*

Model Name	Parameters	Open / Open-Weight	Usage	API Endpoint
------------	------------	--------------------	-------	--------------

DeepSeek-V3.2-Exp	~ 236B (Mixture-of-Experts)	Open-weight	Evaluator model (E1)	DeepSeek-ai/DeepSeek-V3.2-Exp
Qwen3-Next-80B-A3B-Instruct	80B	Open-weight	Evaluator model (E2)	Qwen/Qwen3-Next-80B-A3B-Instruct
Kimi-K2-Instruct-0905	undisclosed (frontier-class)	Proprietary	Evaluator model (E3)	MoonshotAI/Kimi-K2-Instruct-0905
GPT-4.1	undisclosed (frontier-scale)	Proprietary	Evaluator model (E4)	OpenAI/gpt-4.1

Supplementary Figures



Supplementary Figure 1. Training dynamics of the Qwen3-8B-finetuned model. a) epoch progression, b) global training steps, c) gradient norm, d) Training loss, e) learning-rate schedule.