Hunt Instead of Wait: Evaluating Deep Data Research on Large Language Models

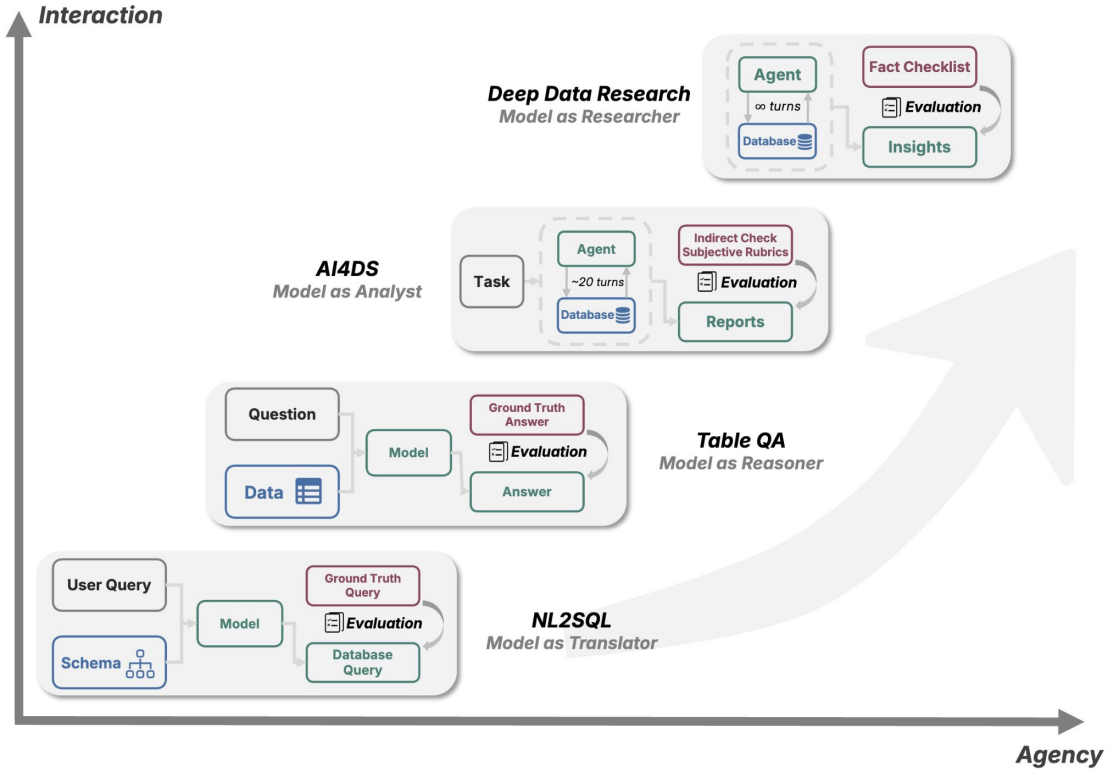
Wei Liu[♠], Peijie Yu[♡], Michele Orini[♠], Yali Du[♠♣], Yulan He[♠♣]
[♠]King's College London, [♡]Tencent, [♣]The Alan Turing Institute

Abstract: The agency expected of Agentic Large Language Models goes beyond answering correctly, requiring autonomy to set goals and decide what to explore. We term this *investigatory intelligence*, distinguishing it from *executional intelligence*, which merely completes assigned tasks. Data Science provides a natural testbed, as real-world analysis starts from raw data rather than explicit queries, yet few benchmarks focus on it. To address this, we introduce **Deep Data Research (DDR)**, an open-ended task where LLMs autonomously extract key insights from databases, and **DDR-Bench**, a large-scale, checklist-based benchmark that enables verifiable evaluation. Results show that while frontier models display emerging agency, long-horizon exploration remains challenging. Our analysis highlights that effective investigatory intelligence depends not only on agent scaffolding or merely scaling, but also on intrinsic strategies of agentic models.

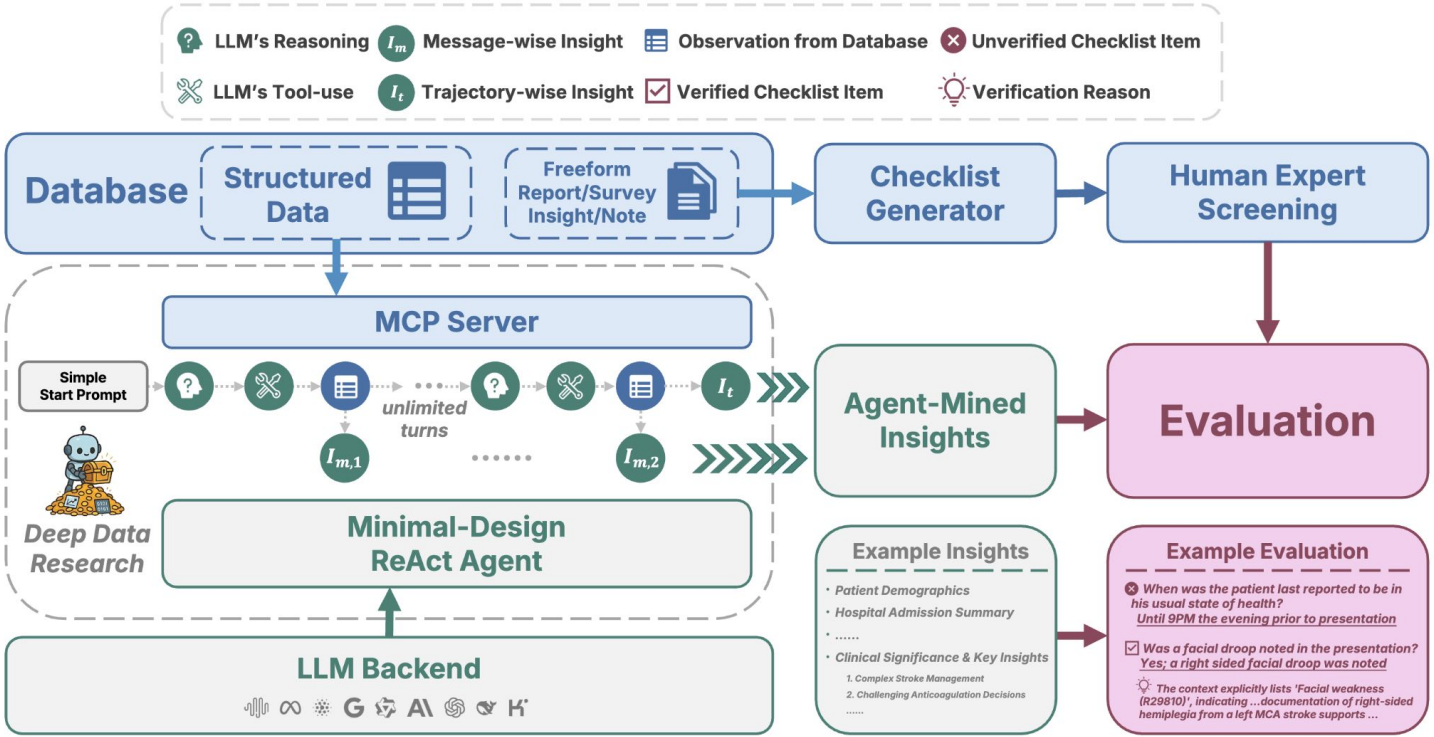
Project: https://huggingface.co/spaces/thinkwee/DDR_Bench

Correspondence: wei.4.liu@kcl.ac.uk, yulan.he@kcl.ac.uk

From NL2SQL to Deep Data Research



Deep Data Research



Deep Data Research

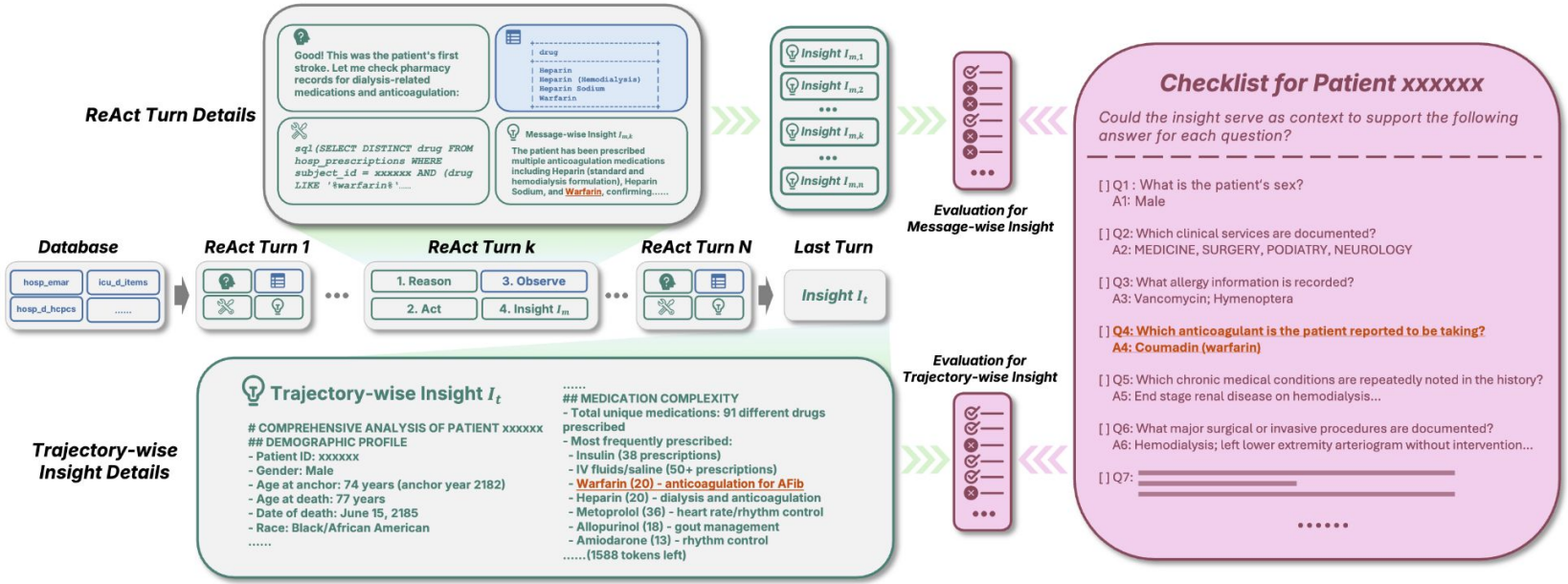


Figure 3. A case of Claude Sonnet 4.5's trajectory and evaluation checklist in the MIMIC scenario of DDR-Bench. Verified fact and supporting insights are underlined. See details of this trajectory in Figure A16. The patient id is anonymised.

<40% for all LLMs (except Claude)

Table 2. Benchmarking results. The best results are highlighted in **bold**. Accuracy is defined as the proportion of checklist items verifiable from the model-mined insights, reported as either sample-averaged (over task entities) or item-averaged (over checklist items).

Models	*Sample-Averaged Accuracy*						*Item-Averaged Accuracy*						*Overall Avg.*
	Message-Wise Insights			*Trajectory-Wise Insights*			*Message-Wise Insights*			*Trajectory-Wise Insights*			
	MIMIC	GLOBEM	10-K	MIMIC	GLOBEM	10-K	MIMIC	GLOBEM	10-K	MIMIC	GLOBEM	10-K	
Proprietary Models													
Claude 4.5 Sonnet	**36.07**	**40.13**	**77.61**	34.67	**38.72**	**60.58**	**34.37**	**40.23**	**77.27**	32.95	**38.85**	**61.25**	**47.73**
GPT-5.2	28.85	38.81	44.89	32.49	38.15	41.09	27.26	38.39	44.99	30.49	38.39	41.22	37.09
GPT-5.1	28.37	38.31	37.12	**35.24**	35.79	44.25	26.61	37.88	37.69	**33.59**	35.63	44.76	36.27
GPT-5 mini	30.02	35.86	46.82	27.86	31.54	37.12	28.81	36.09	46.35	26.36	31.72	36.77	34.61
Gemini 3 Flash	26.58	35.60	44.82	20.78	36.74	21.24	24.94	35.29	44.41	19.51	36.78	21.08	30.65
Gemini 2.5 Pro	21.51	33.77	24.48	20.00	35.62	15.57	19.51	33.79	25.68	18.48	35.40	16.14	25.00
Gemini 2.5 Flash	16.64	29.06	8.48	23.76	28.44	16.06	14.99	28.95	8.72	22.22	28.28	16.49	20.17
Gemini 2.5 Flash-Lite	17.19	26.63	19.45	17.96	24.03	9.01	16.10	26.90	19.32	17.18	24.14	9.42	18.94
Open-Source Models													
DeepSeek-V3.2	28.98	38.46	60.08	30.57	**38.46**	**38.15**	27.00	38.16	**60.66**	28.29	**38.62**	**38.16**	**38.80**
GLM-4.6	25.03	**41.56**	**60.31**	26.15	37.60	36.02	23.26	**41.61**	60.42	24.42	37.70	36.16	37.52
Kimi K2	**33.61**	37.14	51.06	**30.69**	37.00	30.84	**31.65**	37.01	51.24	**28.68**	37.01	31.10	36.42
MiniMax-M2	25.39	37.07	44.17	24.36	36.81	26.88	23.90	37.24	44.66	23.13	36.55	26.82	32.25
Qwen3-Next-80B-A3B	18.01	35.75	44.76	21.79	33.06	30.82	16.80	35.40	45.58	20.80	32.87	31.10	30.56
Qwen3-30B-A3B	21.67	35.73	42.44	18.75	37.25	14.38	20.03	35.63	42.33	18.22	37.01	14.13	28.13
Qwen3-4B	17.97	25.99	41.13	18.68	27.55	19.76	16.67	26.21	40.94	17.18	27.59	19.91	24.97
Qwen2.5-72B	15.65	28.83	27.13	16.82	25.38	13.42	14.34	28.74	27.56	15.50	25.52	14.02	21.08
Qwen2.5-14B-1M	16.75	28.75	22.69	13.68	24.29	10.13	15.50	28.80	23.56	12.66	24.14	9.78	19.23
Qwen2.5-32B	14.12	25.67	27.07	14.40	25.88	7.90	13.05	25.82	27.53	13.18	25.98	8.12	19.06
Qwen2.5-14B	15.93	25.56	18.91	14.51	26.47	10.60	14.86	25.59	19.18	13.44	26.67	10.12	18.49
Qwen2.5-7B	12.81	27.20	10.52	11.08	23.75	4.51	11.63	27.36	10.46	9.95	23.91	4.59	14.81
Qwen2.5-7B-1M	14.61	29.34	9.42	5.95	25.68	3.91	12.85	29.00	9.71	5.30	25.29	3.65	14.56
Llama3.3-70B	10.59	23.99	9.91	5.51	21.70	2.95	9.56	23.68	9.95	5.04	21.61	3.06	12.30

Checklist Coverage is highly correlated to Novelty

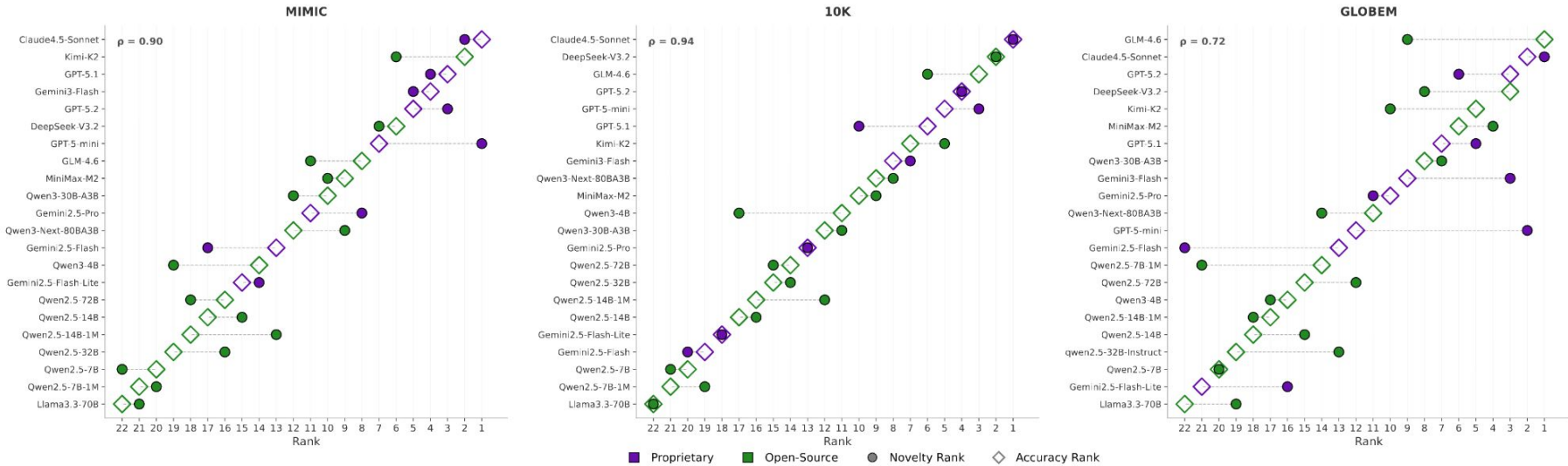
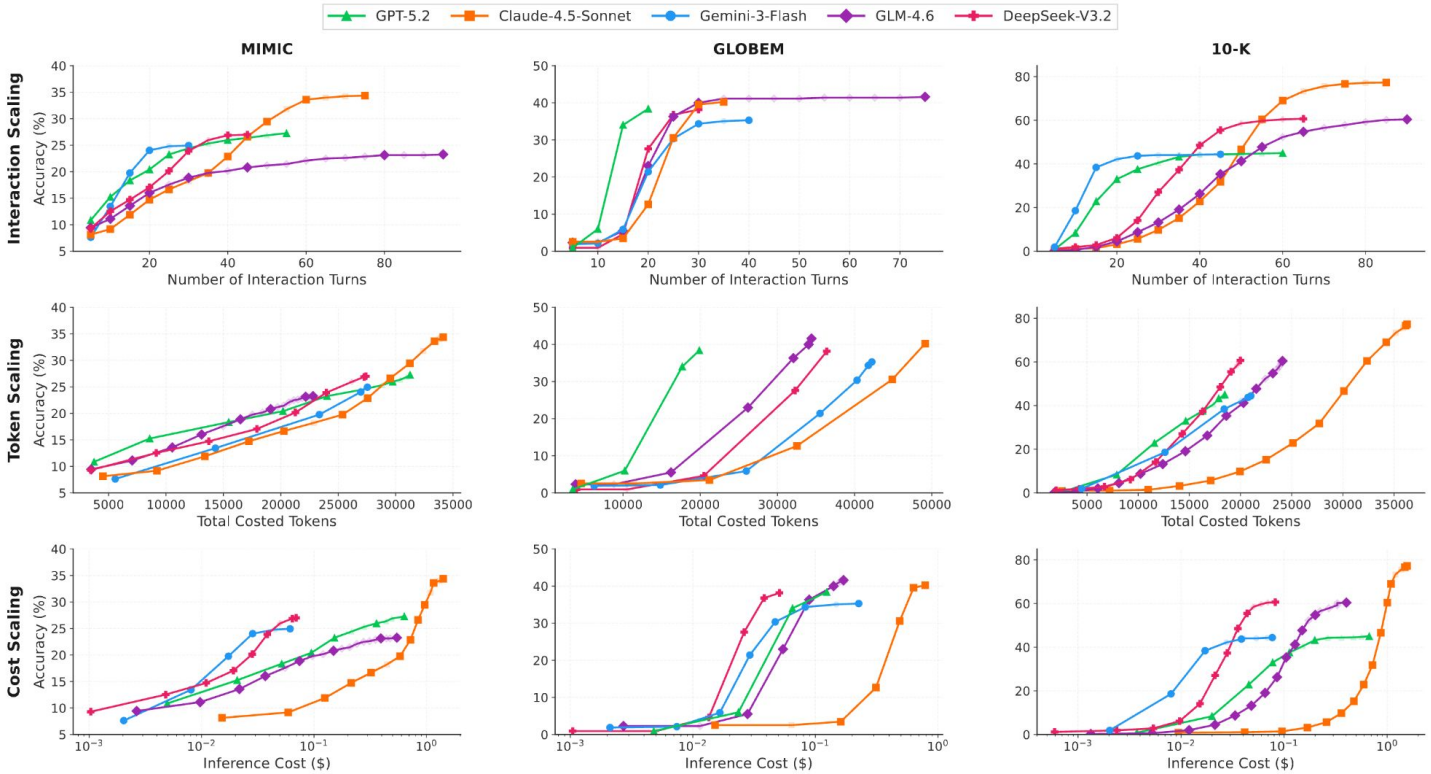


Figure 4. Ranking correlation between novelty and accuracy on Proprietary and Open-Source LLMs. Circles denote the novelty rank, and diamonds denote the accuracy rank. Models are ordered by accuracy rank in the figure. All three scenarios present high correlation.

Scaling Patterns



Takeaways

4.1. Scaling Analysis

Takeaway

LLMs extract more accurate insights from delaying commitment, and they concentrate reasoning into a small number of highly valuable late-stage interactions. These targeted interactions are built upon longer early exploration.

4.2. Exploration Patterns

Takeaway

Advanced LLMs tend to operate in a balanced exploration regime that combines adequate coverage with focused access. Such a regime is consistently observed across different scenarios.

5.1. Study on Training Factors

Takeaway

Scaling is not enough. Meaningful agency require a systematic *agentic-first* training strategy, including targeted pre-training and reinforcement learning.

5.2. Agent Module Analysis

Takeaway

Agent modules mainly reshape interaction patterns rather than reliably enhancing insight discovery. Agency in deep data research emerges from stable, implicit coordination between reasoning and open-ended exploration.

Check out more (traj case, dataset, code)

https://huggingface.co/spaces/thinkwee/DDR_Bench