

ICML 2026

DEER: A Benchmark for Evaluating Deep Research Agents on Expert Report Generation

Janghoon Han

Motivation

A Long, Cited Report Is Not Necessarily a Good Report

Deep research agents can now generate

- ✓ Structured sections
- ✓ Fluent writing
- ✓ Professional tone
- ✓ Many citations

But expert users need more than plausibility:

- 1 ⚠ Did it address the real goal?
- 2 ⚠ Did it cover the right scope?
- 3 ⚠ Is the analysis complete and coherent?
- 4 ⚠ Are key claims supported?

Expert Report

Global Battery Supply Chain Risk Assessment

May 20, 2025

1. EXECUTIVE SUMMARY

1 This report provides a broad overview of battery supply chain risks and market conditions through 2030.

The battery industry is entering a decade of rapid growth, driven by electrification across transportation, grid storage, and consumer electronics. However, this expansion is accompanied by material risks across critical minerals, manufacturing capacity, logistics, and geopolitical uncertainty. We summarize major risk categories and potential market impacts for manufacturers and investors.

2. METHODOLOGY

2 We evaluate risks across critical minerals, refining capacity, cell manufacturing, and long-term demand growth.

The analysis combines public market reports, company disclosures, trade data, and expert judgment to score major risks by likelihood and impact from 2025 to 2030.

3. KEY FINDINGS / ANALYSIS

Risk Category	Description	Likelihood (2025-2030)	Impact	Risk Level
Critical mineral concentration	Limited supply exposure for lithium, nickel, cobalt, and graphite.	High	High	High
Commodity price volatility	Price swings driven by demand surges, speculation, and constrained supply.	High	Medium	High
Manufacturing bottlenecks	Capacity constraints and specialized shortages delay production.	Medium	High	High
Geopolitical policy risk	Trade restrictions, export controls, and policy changes increase uncertainty.	Medium	High	Medium

4. SCENARIO OUTLOOK

Scenario	Description	Cost Impact (2025-2030)
Upside case	Stable supply and faster recycling scale-up	-5% to -10%
Base case	Current trends continue	-5% to +10%
Downside case	Trade barriers and material shortages intensify	+15% to +30%

5. RECOMMENDATIONS

- Diversify sourcing for critical minerals and battery components.
- Negotiate long-term offtake agreements with qualified suppliers.
- Build inventory buffers for high-risk materials.

- Expand recycling partnerships to reduce primary mineral dependency.
- Monitor policy changes affecting tariffs, subsidies, and export controls.

KEY TAKEAWAY

Battery supply chains face elevated and evolving risks through 2030. Firms should diversify suppliers, improve visibility into upstream dependencies, and build resilience into procurement operations.

4 Strategic stockpiling is the single most effective near-term intervention.

REFERENCES

- [1] IEA, Global EV Outlook 2024.
- [2] BloombergNEF, Lithium-ion Battery Supply Chain Outlook 2024.
- [3] McKinsey & Company, Battery Supply Chain Resilience in a Volatile World.

Year	Asia Pacific	Europe	North America	Rest of World
2024	1,000	500	400	100
2025	1,200	600	500	150
2026	1,400	700	600	200
2027	1,600	800	700	250
2028	1,800	900	800	300
2029	2,000	1,000	900	350
2030	2,200	1,100	1,000	400

Motivation

Current Evaluation Struggles to Capture Expert-Level Report Quality

Existing benchmarks face four core challenges:

1 What should be evaluated?

→ Lack of systematic expert criteria

2 Where did the system fail?

→ Single or coarse scores hide specific weaknesses

3 What would an expert notice?

→ Generic judges miss task-specific requirements

4 Are claims truly supported?

→ Evidence verification does not cover the whole report

DEER

Making expert report evaluation systematic, interpretable, and evidence-grounded

DEER addresses the four evaluation gaps with:

Expert report taxonomy

→ Defines what to evaluate

(7 dimensions, 25 subdimensions)

Task-specific expert guidance

→ Helps judges catch issues requiring expert judgment

(expert-written guidance for each task)

Fine-grained dimensions and rubrics

→ Identifies where systems fail

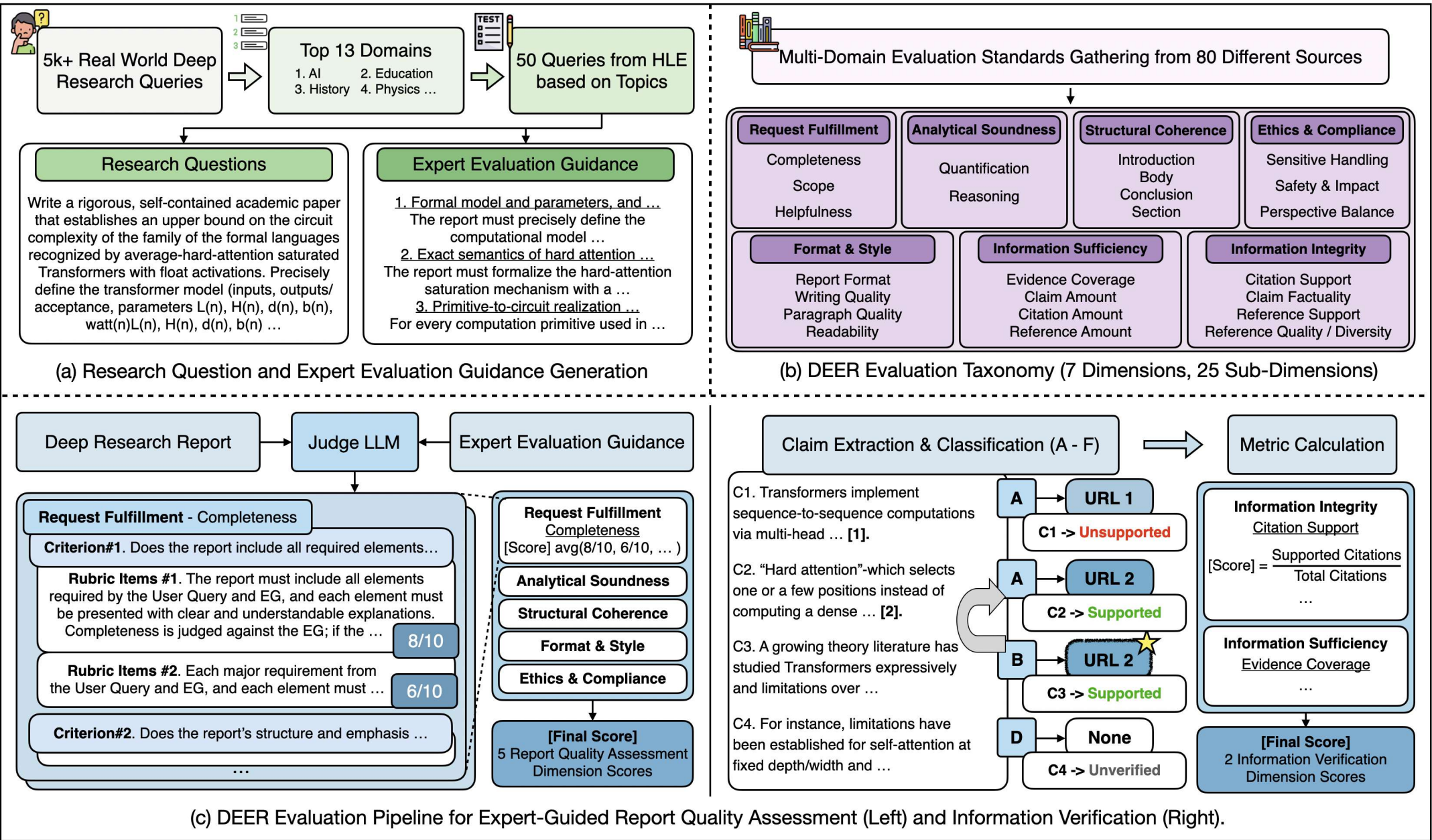
(101 rubric items)

Report-wide verification

→ Assesses full-report evidence support

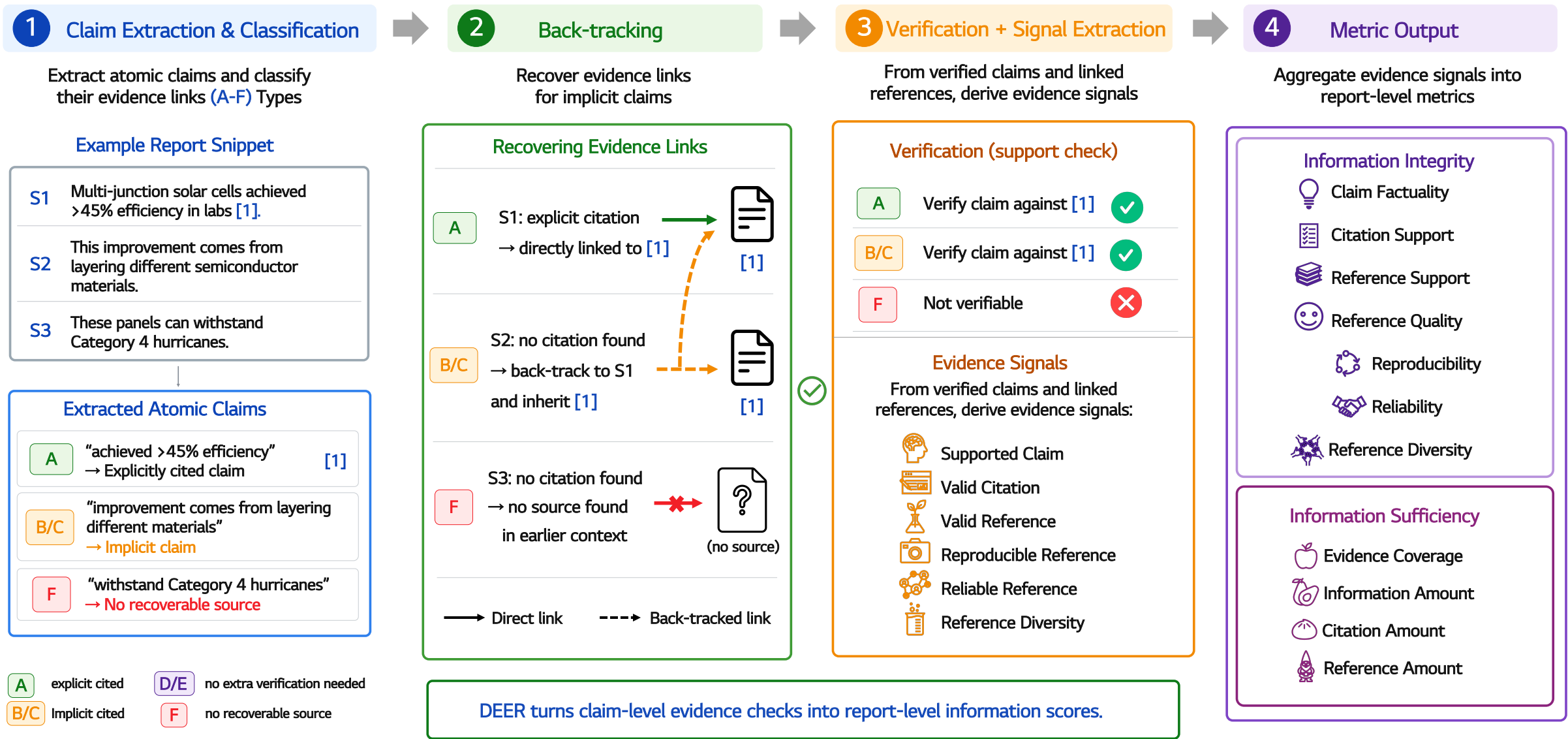
(claim extraction + verification + metrics)

Overview



Information Verification

From report claims to report-level information metrics



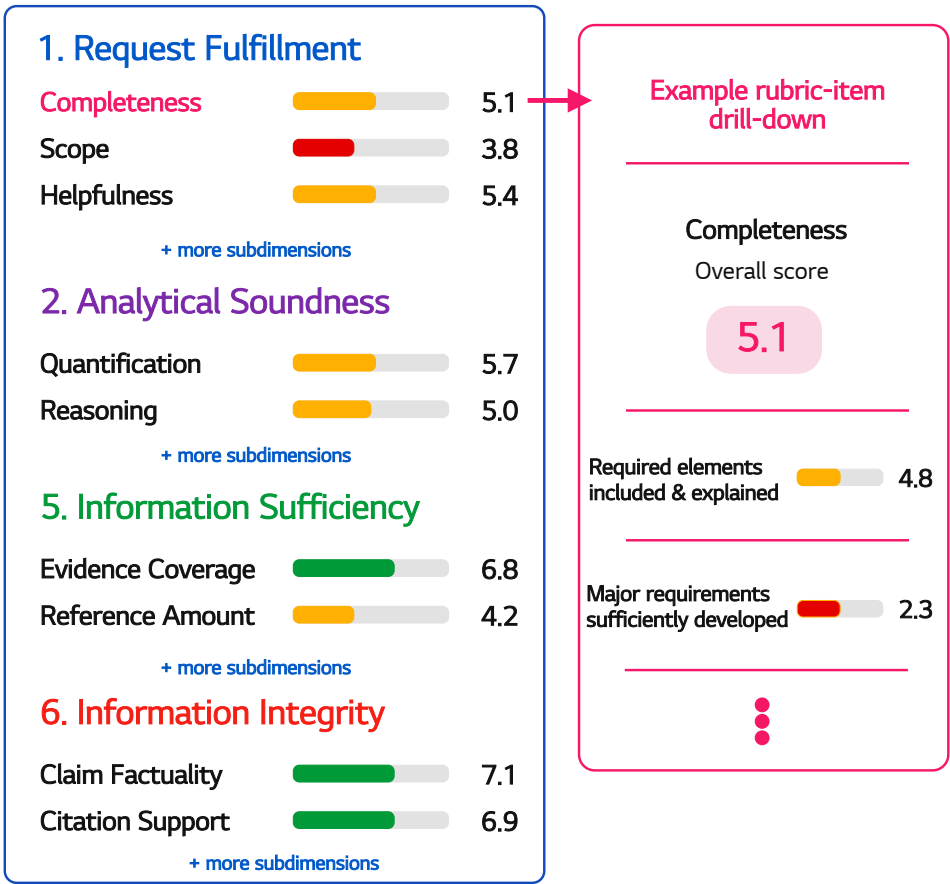
Experiments

DEER: A Benchmark for Evaluating Deep Research Agents on Expert Report Generation

Model	Request Fulfillment	Analytical Soundness	Structural Coherence	Format & Style	Information Integration	Information Sufficiency	Ethics	Mean
General LLMs								
Qwen3-235B (fast)	4.66	5.20	6.39	7.54	1.23	4.20	7.03	5.18
Gemini 2.5 Flash (fast)	4.81	5.46	6.61	7.71	1.35	3.99	7.39	5.33
Claude Opus 4.5 (fast)	4.94	5.44	6.77	7.90	2.33	4.50	7.60	5.64
GPT-5.2 (fast)	4.30	4.85	6.38	7.22	1.07	3.13	7.20	4.88
LLMs+Reasoning								
Qwen3-235B (think)	5.03	5.32	6.65	7.92	1.14	3.90	7.31	5.32
Gemini 2.5 Pro (think)	4.96	5.74	7.11	8.11	2.30	4.40	7.53	5.74
Claude Opus 4.5 (think)	5.08	5.70	6.89	7.92	2.27	4.22	7.63	5.67
GPT-5.2 (think)	5.54	6.21	7.20	8.02	2.11	4.16	7.97	5.89
LLMs+Reasoning+WebSearch								
Qwen3-235B (think+search)	4.22	4.44	5.76	6.65	5.10	5.45	6.85	5.50
Claude Opus 4.5 (think+search)	4.57	5.14	6.15	7.18	6.92	7.62	7.30	6.41
GPT-5.2 (think+search)	5.50	6.05	7.13	7.94	5.57	6.17	7.95	6.62
Deep Research								
WebThinker (Li et al., 2025b)	4.38	4.74	5.87	7.32	6.09	6.38	7.05	5.98
Qwen3-235B (deep)	4.22	4.80	5.12	7.05	6.44	7.90	7.12	6.09
Gemini 2.5 Pro (deep)	4.66	5.56	6.53	7.41	5.90	7.61	7.25	6.42
Claude Opus 4.5 (deep)	4.51	5.31	5.89	7.17	5.96	5.66	7.37	5.98
OpenAI (deep)	4.79	5.33	6.49	7.51	6.97	6.89	7.52	6.50

Fine-grained diagnostic scores

Selected subdimensions with an example rubric-item drill-down



Shown: selected subdimensions and example rubric items

Full DEER: 7 dimensions, 25 subdimensions, 101 rubric items

Thank you