ICLR
International Conference On Learning Representations

Stanford University

LiveResearchBench: Evaluating Live User-Centric Deep Research in the Wild

Jiayu Wang[1], Yifei Ming[3], Riya Dulepet[2], Qinglin Chen[3], Austin Xu[3], Zixuan Ke[3], Frederic Sala[1], Aws Albarghouthi[1], Caiming Xiong[3], Shafiq Joty[3]

[1]University of Wisconsin-Madison

[2]Stanford University

[3]Salesforce AI Research

What is Deep Research

Deep Research Gemini

What is Deep Research

Get up to speed on just about anything with Deep Research, an agentic feature in Gemini that can automatically browse up to hundreds of websites and even your Gmail, Drive and Chat on your behalf, think through its findings, and create insightful multi-page reports in minutes.

Problem of existing benchmarks

Problem of existing benchmarks

Deep Research Bench (Bosse et al.):

How many IM and GM account closures did chess.com report for 2024?

Problem of existing benchmarks

Deep Research Bench (Bosse et al.):

How many IM and GM account closures did chess.com report for 2024?

- No support for long-form answer
- Search intensive but low reasoning load
- Time-bound & static; Lacks support for evolving data

Problem of existing benchmarks

Deep Research Bench (Bosse et al.):

How many IM and GM account closures did chess.com report for 2024?

- No support for long-form answer
- Search intensive but low reasoning load
- Time-bound & static; Lacks support for evolving data

DeepResearch Bench (Du et al.):

Write a paper to discuss the influence of AI interaction on interpersonal relations, considering AI's potential to fundamentally change how and why individuals relate to each other.

Problem of existing benchmarks

Deep Research Bench (Bosse et al.):

How many IM and GM account closures did chess.com report for 2024?

- No support for long-form answer
- Search intensive but low reasoning load
- Time-bound & static; Lacks support for evolving data

DeepResearch Bench (Du et al.):

Write a paper to discuss the influence of AI interaction on interpersonal relations, considering AI's potential to fundamentally change how and why individuals relate to each other.

- Missing target audience
- Missing content & format requirement
- Ambiguous scope & limited search & analysis depth

LiveResearchBench

A live and user-centric benchmark crafted by humans

Queries iteratively refined through human-AI interaction

Create a comprehensive report about the evolution of artistic styles across different historical periods and national characteristics. Cover major historical periods **from Ancient civilizations up to the present** {{date}}. Focus primarily on **visual arts (painting, sculpture, architecture)**. Include analysis of **key regional/national characteristics including European traditions, East Asian art (China, Japan), Islamic art, and Indigenous American art**. **Target the report for undergraduate-level students with substantial academic depth**, including proper citations and references. **Structure as a formal academic report** examining how political, social, religious, and technological factors influenced artistic development. **Include analysis of major artists, techniques, materials, and stylistic characteristics for each period and region.**

- ✓ Multi-faceted scope & evaluation friendly
- ✓ Targeted audience & output expectation
- ✓ Require wide search & in-depth analysis

LiveResearchBench

A live and user-centric benchmark crafted by humans

Queries iteratively refined through human-AI interaction

Create a comprehensive report about the evolution of artistic styles across different historical periods and national characteristics. Cover major historical periods **from Ancient civilizations up to the present** {{date}}. Focus primarily on **visual arts (painting, sculpture, architecture)**. User centric le analysis of **key regional/national characteristics including European traditions, East Asian art (China, Japan), Islamic art, and Indigenous American art**. **Target the report for undergraduate-level students with substantial academic depth**, including proper citations and references. **Structure as a formal academic report** examining how political, social, religious, and technological factors influenced artistic development. **Include analysis of major artists, techniques, materials, and stylistic characteristics for each period and region.**

- ✓ Multi-faceted scope & evaluation friendly
- ✓ Targeted audience & output expectation
- ✓ Require wide search & in-depth analysis

LiveResearchBench

A live and user-centric benchmark crafted by humans

Queries iteratively refined through human-AI interaction

Create a comprehensive report about the evolution of artistic style**Unambiguous**rent historical periods and national characteristics. Cover major historical periods **from Ancient civilizations up to the present** {{date}}. Focus primarily on **visual arts (painting, sculpture, architecture)**. **User centric**le analysis of **key regional/national characteristics including European traditions, East Asian art (China, Japan), Islamic art, and Indigenous American art**. **Target the report for undergraduate-level students with substantial academic depth**, including proper citations and references. **Structure as a formal academic report** examining how political, social, religious, and technological factors influenced artistic development. **Include analysis of major artists, techniques, materials, and stylistic characteristics for each period and region.**

- ✓ Multi-faceted scope & evaluation friendly
- ✓ Targeted audience & output expectation
- ✓ Require wide search & in-depth analysis

LiveResearchBench

A live and user-centric benchmark crafted by humans

Queries iteratively refined through human-AI interaction

Time varying

Unambiguous

User centric

Create a comprehensive report about the evolution of artistic style**rent historical periods and national characteristics. Cover major historical periods from Ancient civilizations up to the present** {{date}}. Focus primarily on **visual arts (painting, sculpture, architecture)**. **le analysis of key regional/national characteristics including European traditions, East Asian art (China, Japan), Islamic art, and Indigenous American art. Target the report for undergraduate-level students with substantial academic depth**, including proper citations and references. **Structure as a formal academic report** examining how political, social, religious, and technological factors influenced artistic development. **Include analysis of major artists, techniques, materials, and stylistic characteristics for each period and region.**

- ✓ Multi-faceted scope & evaluation friendly
- ✓ Targeted audience & output expectation
- ✓ Require wide search & in-depth analysis

LiveResearchBench

A live and user-centric benchmark crafted by humans

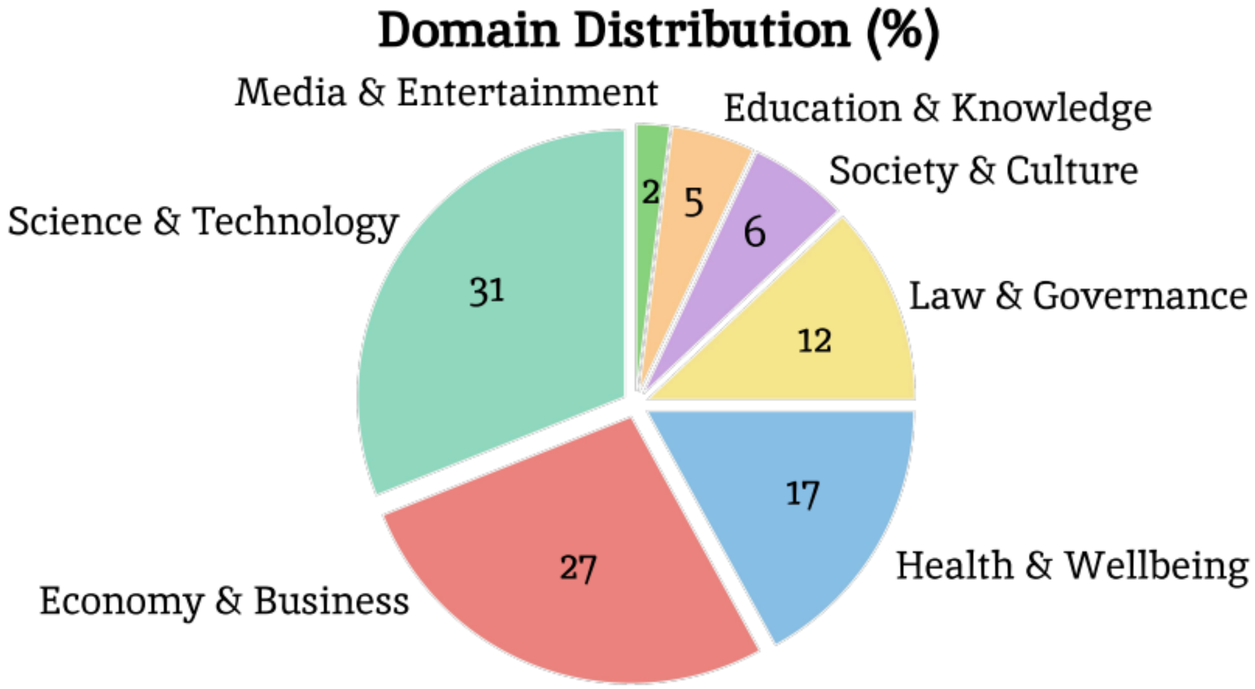
Queries iteratively refined through human-AI interaction

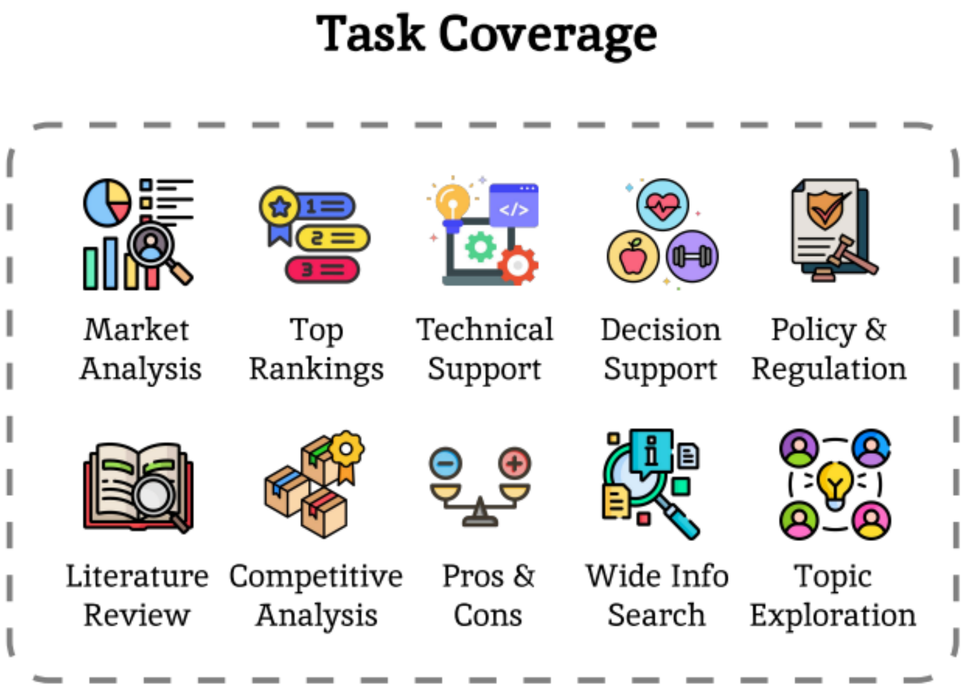
Time varying

Create a comprehensive report about the evolution of artistic style Unambiguous rent historical periods and national characteristics. Cover major historical periods **from Ancient civilizations up to the present** {{date}}. Focus primarily on **visual arts (painting, sculpture, architecture)**. User centric le analysis of **key regional/national characteristics including European traditions, East Asian art (China, Japan), Islamic art, and Indigenous American art**. **Target the report for undergraduate-level students with substantial academic depth**, including proper citations and references. **Structure as a formal academic report** examining h Multi-faceted social, religious, and technological factors influenced artistic development. **Include analysis of major artists, techniques, materials, and stylistic characteristics for each period and region.**

- ✓ Multi-faceted scope & evaluation friendly
- ✓ Targeted audience & output expectation
- ✓ Require wide search & in-depth analysis

Domain Distribution & Task Coverage





Benchmark Construction and Verification Pipeline

- Six-stage data generation & five-stage data verification
- Built with over 1,500 hours of human labor

1 User Interview & Crowd Survey

Q: What questions would you ask for a deep research agent?

2 Task & Domain Distribution

3 Initial Research Question

How large is the enterprise AI services market in 2024–2025, and what are the main adoption drivers among enterprise clients? Which customer segments...?

4 Clarification Questions

Q1: Are you looking for a global view of the enterprise AI services market, or focused on specific regions (e.g., US, EU, Asia)?

Q2: Do you want specific financial figures (e.g., market size in USD) and growth rates for 2024–2025?

5 Refined Research Question

How large is the US enterprise AI services market in 2024–2025 (provide specific market size in USD and growth rates), and what are the main adoption drivers among enterprise...

6 Unit Test Checklist

C1: Does the report provide the size of the market for 2024 and 2025?

C2: Does the report focus on the US market?

...

Benchmark Construction and Verification Pipeline

- Six-stage data generation & five-stage data verification
- Built with over 1,500 hours of human labor

1 User Interview & Crowd Survey

Q: What questions would you ask for a deep research agent?

2 Task & Domain Distribution

Domain Task

3 Initial Research Question

How large is the enterprise AI services market in 2024–2025, and what are the main adoption drivers among enterprise clients? Which customer segments...?

4 Clarification Questions

Q1: Are you looking for a global view of the enterprise AI services market, or focused on specific regions (e.g., US, EU, Asia)?

Q2: Do you want specific financial figures (e.g., market size in USD) and growth rates for 2024–2025?

5 Refined Research Question

How large is the US enterprise AI services market in 2024–2025 (provide specific market size in USD and growth rates), and what are the main adoption drivers among enterprise...

6 Unit Test Checklist

C1: Does the report provide the size of the market for 2024 and 2025?

C2: Does the report focus on the US market?

...

7 Research Questions & Checklists

Research Questions

Checklists

GUIDE

8 Independent Assessment

x N

- Appropriate
- Not Appropriate

- Appropriate
- Not Appropriate
- Valid but not necessary
- Needs Modification

GUIDE

9 Quality Control I

x N

- Agree
- Disagree

- Agree
- Disagree

GUIDE

10 Quality Control II

Measure overall quality
Detect systematic issues

11 Final Verification & Cross Check

How do we evaluate deep research reports (long-form and open-ended)?

If we apply LLM-as-a-Judge naively…

If we apply LLM-as-a-Judge naively…

Ask an LLM judge to assign a score of 1-5 based on these rubrics:

If we apply LLM-as-a-Judge naively…

Ask an LLM judge to assign a score of 1-5 based on these rubrics:

Reasoning Quality & Analytical Insight

What it measures: synthesis, logical structure, causal reasoning, non-trivial insights.

1 – Lists facts with no reasoning; conclusions unsupported.

2 – Some reasoning but simplistic; weak justification.

3 – Coherent reasoning; some synthesis; standard insights.

4 – Strong analytical depth; integrates evidence and draws meaningful patterns.

5 – Clear, rigorous reasoning with expert-level insight; explains **why**, not just **what**; highlights trade-offs and uncertainty.

The performance is terrible
(<50% alignment with human judgement)

DeepEval: dissect into several sub metrics +
turn subjective judgements into objective ones +
LLM-ensemble-as-a-judge

DeepEval: A comprehensive evaluation suite for deep research

DeepEval: A comprehensive evaluation suite for deep research

- **Report-level metrics:**
- **Content-level metrics:**

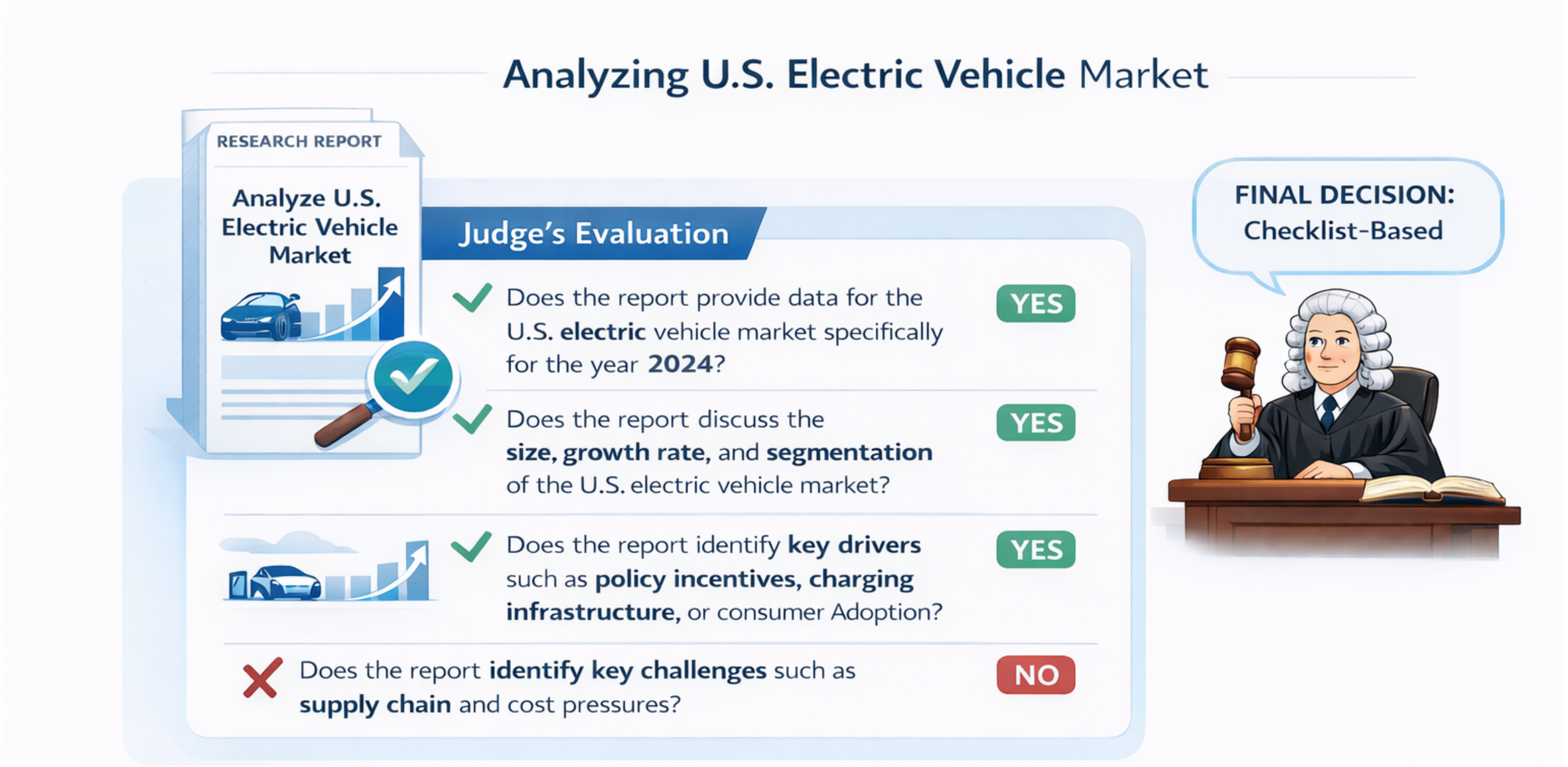
DeepEval: A comprehensive evaluation suite for deep research

- **Report-level metrics:**
  - **Presentation & organization**: *is the report poorly organized or contain grammar errors?*
  - **Factual & logical consistency**: *are there inconsistent numbers/claims?*
- **Content-level metrics:**

DeepEval: A comprehensive evaluation suite for deep research

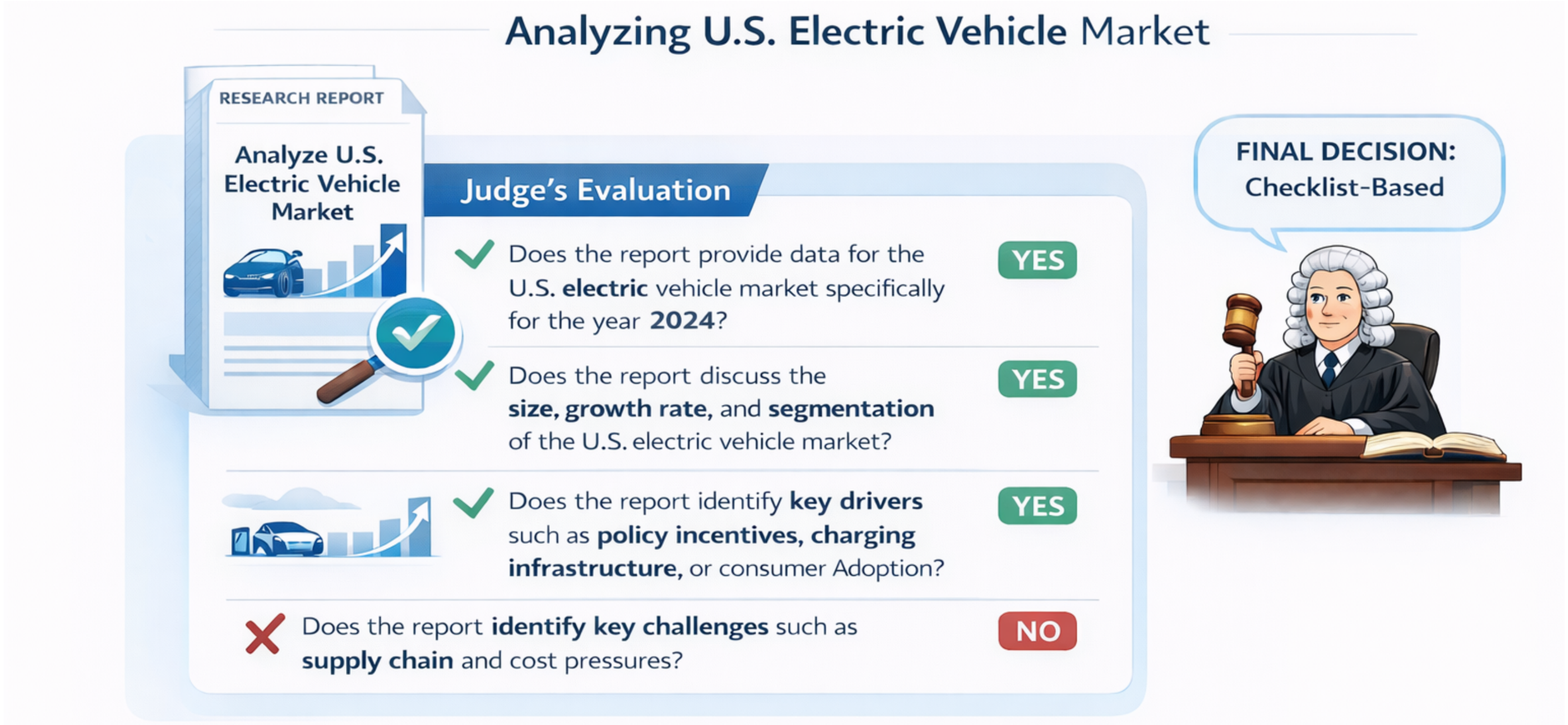
- **Report-level metrics:**
  - **Presentation & organization**: *is the report poorly organized or contain grammar errors?*
  - **Factual & logical consistency**: *are there inconsistent numbers/claims?*
- **Content-level metrics:**
  - **Coverage & comprehensiveness**: *are all aspects of multi-faceted query addressed?*
  - **Analysis depth**: *do report provide substantive insights beyond simple info gathering?*
  - **Citation association**: *are factual claims/numbers cited?*
  - **Citation accuracy**: *do cited sources genuinely support their claims?*

If we have a checklist for each question…



DeepEval: aligning with human judgement

- **1) Checklist-based:**
  - **Coverage & comprehensiveness:** are all aspects of multi faceted query addressed?



DeepEval: aligning with human judgement

- **1) Checklist-based:**
  - **Presentation & organization**

DeepEval: aligning with human judgement

- **1) Checklist-based:**
  - **Presentation & organization**

No.	Checklist Questions for Report Presentation Quality
1	Does the report present a clear, coherent, and logically ordered structure so that the organization is easy to follow and directly addresses the research question?
2	Does the report contain zero grammar and spelling errors?
3	Does every entry in the reference list correspond to at least one in-text citation?
4	Does every in-text citation have a corresponding entry in the reference list?
5	Is there exactly one “References” (or “Bibliography” / “Sources”) section, and are its entries sorted according to a single, consistent scheme?
6	Is a single, consistent citation style used throughout the entire document?
7	Are all in-text citations placed logically at the end of a clause or sentence, without interrupting grammatical flow?
8	If the report includes figures or tables, does each one contain complete data or a valid visual element? (If none are included, the report automatically passes this test.)
9	Is the formatting correct and consistent? For example: (a) If delivered in Markdown, are proper heading levels (#, ##, etc.) used instead of plain text for section titles; (b) if Markdown tables are included, is their syntax valid and renderable?
10	If the citations are numbered, are there no skipped numbers (e.g., [23], [25], [26] with [24] missing) and no duplicates (two different sources assigned the same number, or one source assigned multiple numbers)?

DeepEval: aligning with human judgement

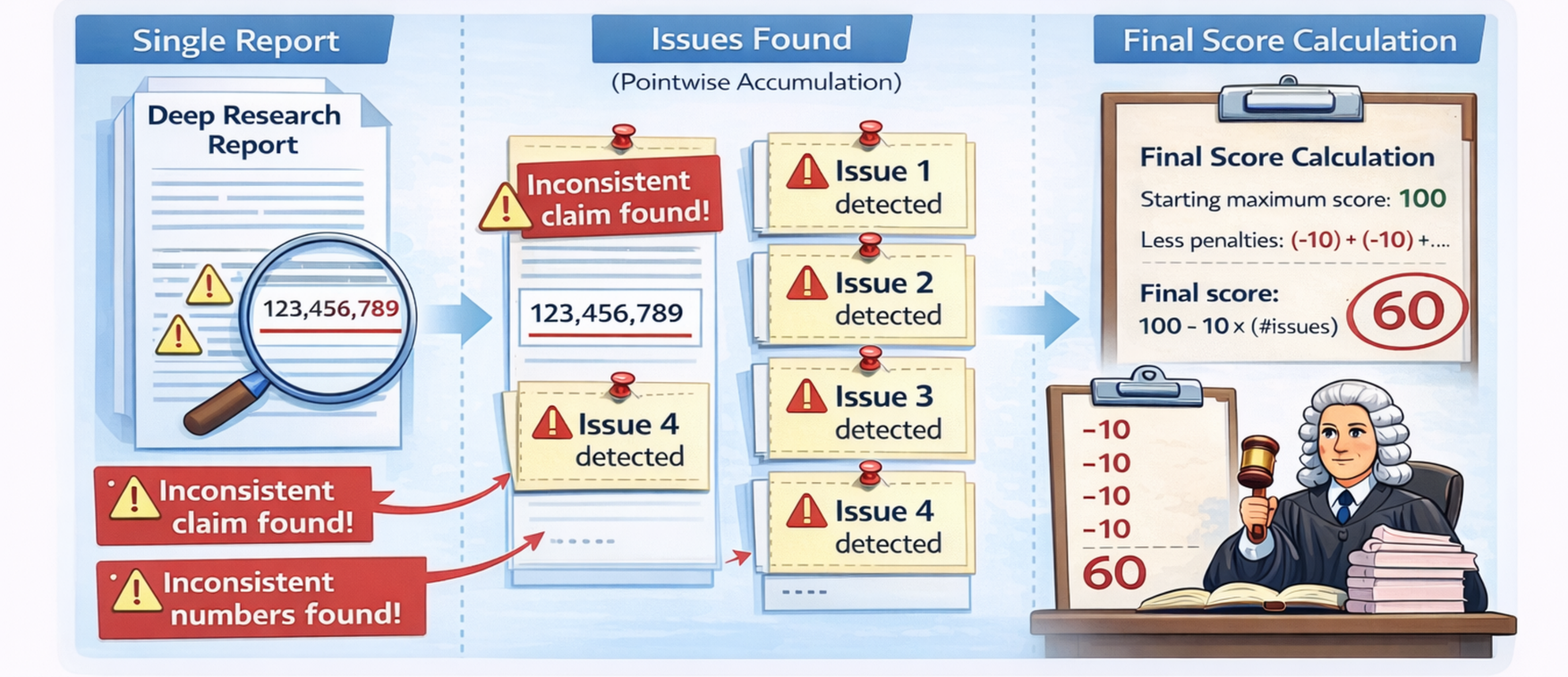
- **2) Pointwise (additive):**

DeepEval: aligning with human judgement

- **2) Pointwise (additive):**
  - **Factual & logical consistency**

DeepEval: aligning with human judgement

- **2) Pointwise (additive):**
  - **Factual & logical consistency:** are there inconsistent claims/numbers in the report?
    - Find as many substantive issues as possible



DeepEval: aligning with human judgement

- **2) Pointwise (additive):**
  - **Factual & logical consistency**
  - **Citation association**

Score	Uncited Claims	Description
100	0	Perfect – All major claims/facts have citations; fully traceable
90	1–2	Excellent – Very few claims lack citations; minimal impact
80	3–4	Good – Few claims lack citations; minor omissions
70	5–6	OK – Some claims lack citations
60	7–8	Above Average – Several claims lack citations
50	9–10	Average – Many claims lack citations; significant association issues
40	11–12	Below Average – Most claims lack citations
30	13–14	Poor – Extensive uncited claims; report poorly supported
20	15–17	Very Poor – Overwhelming lack of citations
10	18+	Unacceptable – Report largely untraceable

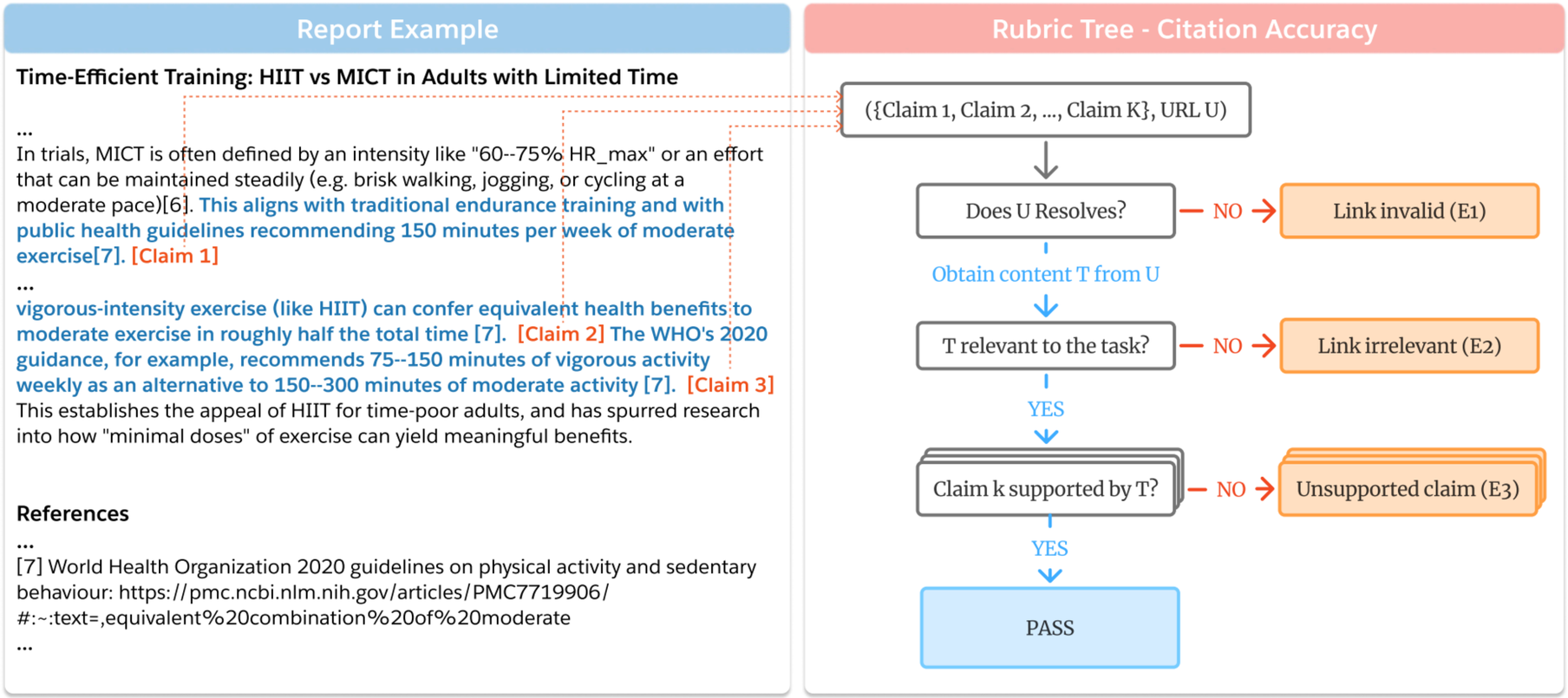
DeepEval: aligning with human judgement

- **3) Pairwise additive:**
  - **Analysis depth:** is the report better than the reference report in terms of analysis depth?

.

DeepEval: aligning with human judgement

- **4) Rubric tree:**
  - **Citation accuracy**

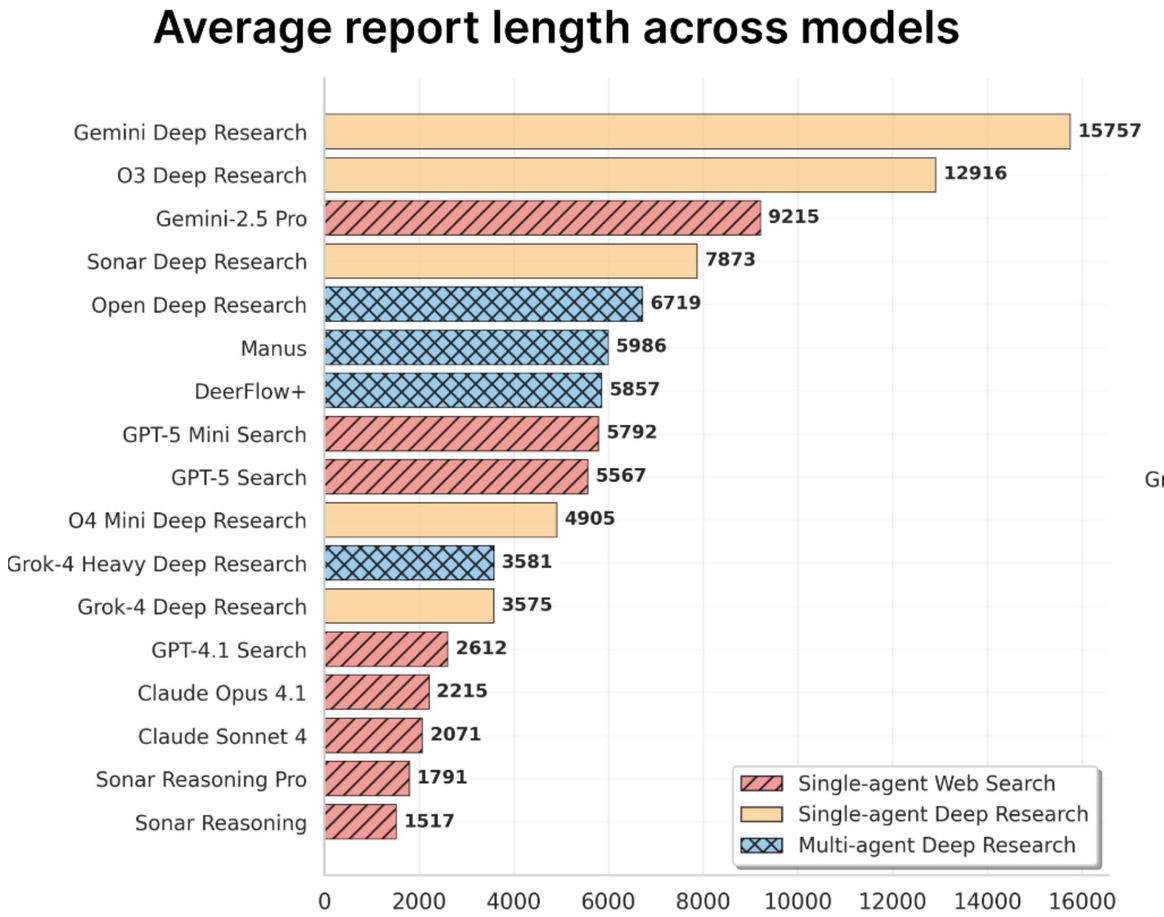


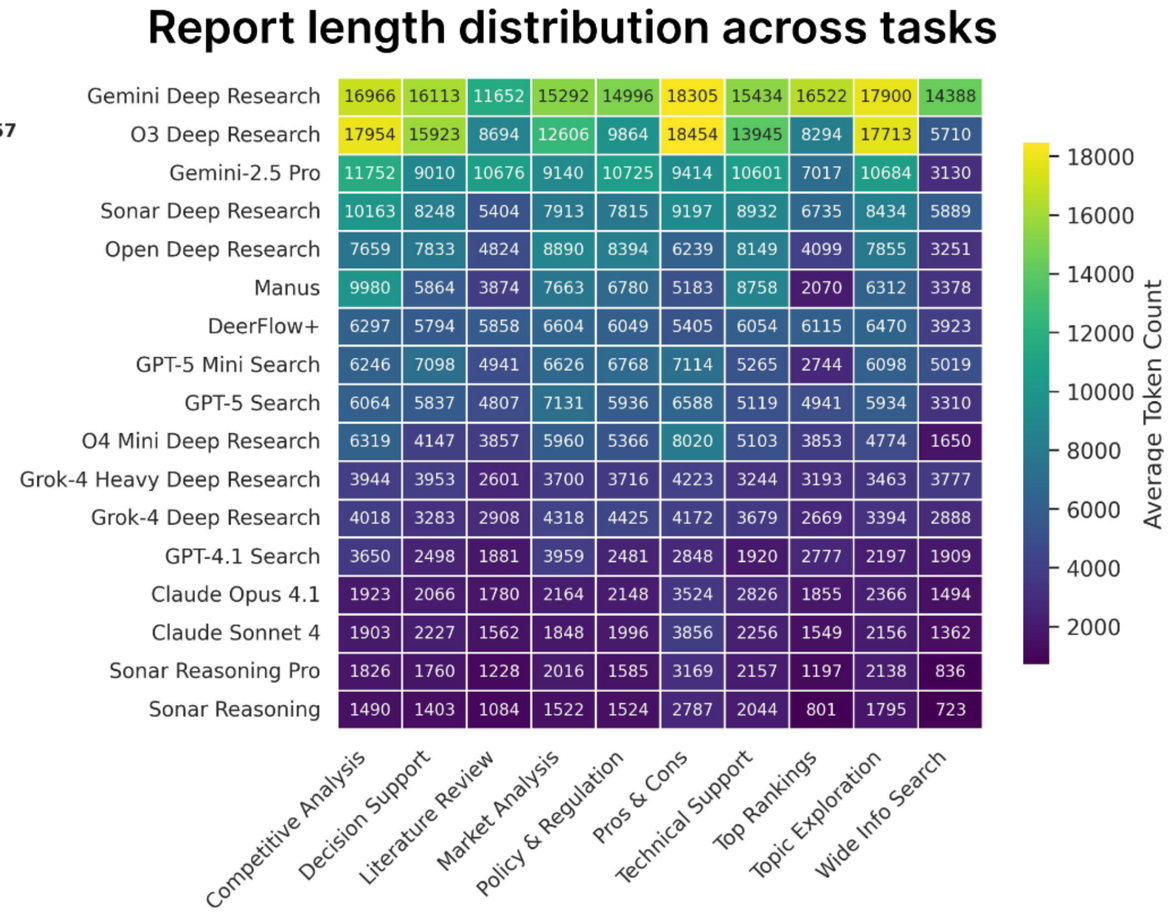
DeepEval: achieves high human alignment rate

- **LLM-ensemble-as-a-judge:** Gemini-2.5-pro & GPT-5
- **Report-level metrics:**
  - **Presentation & organization**: 98.3%
  - **Factual & logical consistency**: 82.0%
- **Content-level metrics:**
  - **Coverage & comprehensiveness**: 100.0%
  - **Analysis depth**: 92.5%
  - **Citation association**: 85.9%
  - **Citation accuracy**: 87.1%

Main Results and Analysis

Main Results and Analysis





Observation 1

Longer reports are common, but not consistently better.

Main Results and Analysis

⚠ Missing Links (URLs)

Report from GPT-4.1:

Clinical Presentation & Differential Diagnoses

...

1.1 Extensor Carpi Ulnaris (ECU) Tendon Pathology

- **ECU Tendinopathy**: Repetitive microtrauma during supination combined with ulnar deviation can cause overuse injury to the ECU tendon, common in racket sports [1][2]. Features include pain on the ulnar side of the wrist, swelling, and discomfort during resisted wrist extension with ulnar deviation–symptomatic of ECU tendinopathy [2][9].
- **ECU Instability or Subluxation**: Sudden onset “bolt out of the blue” pain over the ulnar wrist, especially during double-handed backhand strokes, may indicate traumatic ECU instability (subluxation or dislocation) [1].

...

References

1. Montalvan B, Parier J, Brasseur JL, Le Viet D, Drape JL. Extensor carpi ulnaris injuries in tennis players: a study of 28 cases. *Br J Sports Med*. 2006;40(5):424–429. (Identifies ECU tendinopathy, instability, rupture; outlines immobilization and surgical protocols.)
2. Physio sources on ECU tendinopathy–overuse mechanism, symptoms, common in racket sports [2015+ general knowledge].
4. Logli AL et al. Ulnar wrist pain in a tennis player: case review and discussion. *JBJS Reviews*. 2022. (Discusses TFCC, UT ligament injury and arthroscopic repair.)

...

Analysis:
In-text citations (e.g., [2], [9]) either lack a proper reference entry or appear in the reference list without URLs or identifiable sources.

⚠ Inconsistent Citation Format

Report from Gemini-2.5 Pro:

Major U.S. Bank Received a Less-Than-Satisfactory Community Reinvestment Act Rating in Recent Years

...

Washington D.C. - A comprehensive review of Community Reinvestment Act (CRA) ratings for U.S. banks with assets exceeding $50 billion reveals that only one institution, City National Bank, received a "less-than-satisfactory" rating in its most recent evaluation between 2019 and 2024. The bank was rated "Needs to Improve" by the Office of the Comptroller of the Currency (OCC) in an examination dated March 27, 2023, which covered the years 2019 through 2021.

...

Below are the detailed findings for City National Bank:

City National Bank

* **Rating**: Needs to Improve
* **Exam Date**: March 27, 2023 (Evaluation Period: January 1, 2019 - December 31, 2021)
* **Federal Reserve/OCC CRA Disclosure:** [Office of the Comptroller of the Currency (OCC) Performance Evaluation](https://www.occ.gov/static/cra/cra-eval/apr24/14695.pdf)[1](https://www.occ.treas.gov/news-issuances/news-releases/2024/nr-occ-2024-58.html)
* **Reputable News Source:** [City National Bank Press Release](https://www.cnb.com/about-us/newsroom/2024/city-national-bank-community-reinvestment-act-rating-lowered.html)

....

While the vast majority of large U.S. banks maintained "Satisfactory" or "Outstanding" CRA ratings during this period, the case of City National Bank underscores the continued importance of the Community Reinvestment Act in ensuring equitable access to credit for all communities.

Analysis:
The report mixes multiple citation styles within the same section. For example, the “Federal Reserve/OCC CRA Disclosure” entry combines a descriptive hyperlink with a numbered reference [1], while other entries only use hyperlinks.

Observation 2

Models produce fluent reports, but struggle with citation correctness and formatting.

Main Results and Analysis

⚠ References Not Mentioned in Report

Report from Manus:

U.S. Electric Vehicle Market Analysis
...
1.1 Market Size

In 2024, the U.S. electric vehicle market demonstrated robust expansion in both sales volume and revenue. Full-year EV sales in the United States ranged from approximately 1.3 million to 1.6 million units [1, 3, 5]. This volume translated into an EV sales share of around 10% of all light-duty vehicles sold during the year [1, 3]. In terms of economic value, the U.S. electric vehicle market was estimated to be worth USD 131.3 billion in 2024 [11].

1.2 Growth Rate

The market's growth trajectory remained positive, albeit with some variations in reported figures. EV sales in 2024 increased by 7.3% compared to 2023 [5]. The broader U.S. and Canadian market collectively saw a 9% growth in 2024 [9]. Looking ahead, the market is projected to maintain a strong upward trend, with an anticipated annual growth rate (CAGR) of 12.61% between 2025 and 2029, which is expected to result in a market volume of US$154.2 billion by 2029 [7].
...

Adaptation to Pressures: Ford is adapting to competitive and regulatory pressures by broadening its electrification strategy and focusing on competitive battery costs [78]. The company is also navigating evolving emissions and safety standards globally [79]. Ford's price cuts on the Mach-E reflect its adaptation to a competitive market where pricing is a key factor [81].

References
...
95. Ford Broadens Electrification Strategy to Reach More ... (2024, August 21). Fromtheroad.ford.com. https://www.fromtheroad.ford.com/us/en/articles/2024/ford-broadens-electrification-strategy-to-reach-more-customers-
96. Ford's Role in Shaping the 2024 Electric Vehicle Market. (n.d.). Willisford.com. https://www.willisford.com/ford-electric-vehicle-market-2024-outlook.htm
97. Integrated Sustainability and Financial Report 2024. (2024, April 22). Corporate.ford.com. https://corporate.ford.com/content/dam/corporate/us/en-us/documents/reports/2024-integrated-sustainability-and-financial-report.pdf
98. Mustang Suv 2024. (2025, September 6). Cert-test-new.itlab.stanford.edu. https://cert-test-new.itlab.stanford.edu/mustang-suv-2024

Analysis:
The reference list includes entries (e.g., [95]–[98]) that are not cited anywhere in the report body. This creates misalignment between citations and references.

⚠ Out-of-order References

Report from Manus:

Sleep and Body-Clock Strategies for Enhanced Alertness, Learning, and Mood
...
Specific Considerations for Shift Work:
Shift workers face significant challenges in maintaining circadian alignment. Strategies include:
* **Fixed Shift Schedules:** Where possible, consistent shift schedules (e.g., permanent night shifts rather than rotating shifts) can help the body adapt more effectively [70].
* **Light Management:** Use bright light during night shifts to promote alertness and wear dark glasses on the commute home to minimize light exposure before daytime sleep [66].
* **Melatonin Supplementation:** Under medical guidance, melatonin can be used to help adjust to new sleep-wake cycles, particularly when transitioning between shifts or dealing with jet lag [39].

References (continued)
[63] Sleep and Academic Excellence: A Deeper Look. (2024). *Stanford Longevity*. https://longevity.stanford.edu/lifestyle/2024/01/10/sleep-and-academic-excellence-a-deeper-look/
[64] Polyphasic Sleep: Pros and Cons of Short Sleep Schedule. (2018). *Verywell Health*. https://www.verywellhealth.com/pros-and-cons-of-a-polyphasic-sleep-schedule-4165843
[65] Sleep hygiene: Simple practices for better rest. (2025). *Harvard Health Publishing*. https://www.health.harvard.edu/staying-healthy/sleep-hygiene-simple-practices-for-better-rest
...
References
[1] The impact of extended sleep on daytime alertness, vigilance and mood. (2025). *ResearchGate*. https://www.researchgate.net/publication/8370477_The_impact_of_extended_sleep_on_daytime_alertness_vigilance_and_mood
[2] Circadian Rhythms | National Institute of General Medical Sciences. (2025). *NIGMS*. https://www.nigms.nih.gov/education/fact-sheets/Pages/circadian-rhythms
...

Analysis:
The reference list is misordered: instead of starting at [1] and proceeding sequentially, it begins at [63]–[65] and then loops back to [1]–[62]. This disrupts the expected citation order and makes it difficult to trace in-text citations accurately.

Main Results and Analysis

Agent Name	Presentation & Organization	Fact & Logic Consistency	Coverage & Comprehensiveness	Citation Association	Avg
Single-Agent with Web Search					
GPT-5	71.6	68.3	83.4	69.0	73.1
GPT-4.1	66.0	65.9	63.6	66.9	65.6
GPT-5-mini	61.4	66.9	80.5	62.2	67.7
Gemini 2.5 Pro	51.9	**76.5**	73.1	44.9	61.6
Claude 4 Sonnet	81.9	67.3	49.2	50.8	62.3
Claude 4.1 Opus	81.6	67.5	50.8	47.2	61.8
Perplexity Sonar Reasoning	82.1	73.0	40.7	61.7	64.4
Perplexity Sonar Reasoning Pro	79.6	71.9	46.7	65.0	65.8
Single-Agent Deep Research					
OpenAI o3 Deep Research	71.3	64.2	85.0	30.9	62.9
OpenAI o4-mini Deep Research	74.3	62.3	78.6	32.1	61.8
Perplexity Sonar Deep Research	**83.5**	67.4	65.5	52.1	67.1
Grok-4 Deep Research	69.1	57.4	86.3	64.7	69.4
Gemini Deep Research	62.1	63.0	75.8	64.6	66.4
Multi-Agent Deep Research					
Manus	75.0	63.1	73.3	53.8	66.3
Grok-4 Heavy Deep Research	75.9	59.4	**89.3**	64.7	72.3
Deerflow+ (w. GPT-5)	78.8	69.9	61.6	**81.4**	72.9
Open Deep Research (w. GPT-5)	81.0	71.3	65.3	77.2	**73.7**

Observation 3&4

Multi agent systems lead on average;
single web agents excel in consistency, MAS lead in citation association

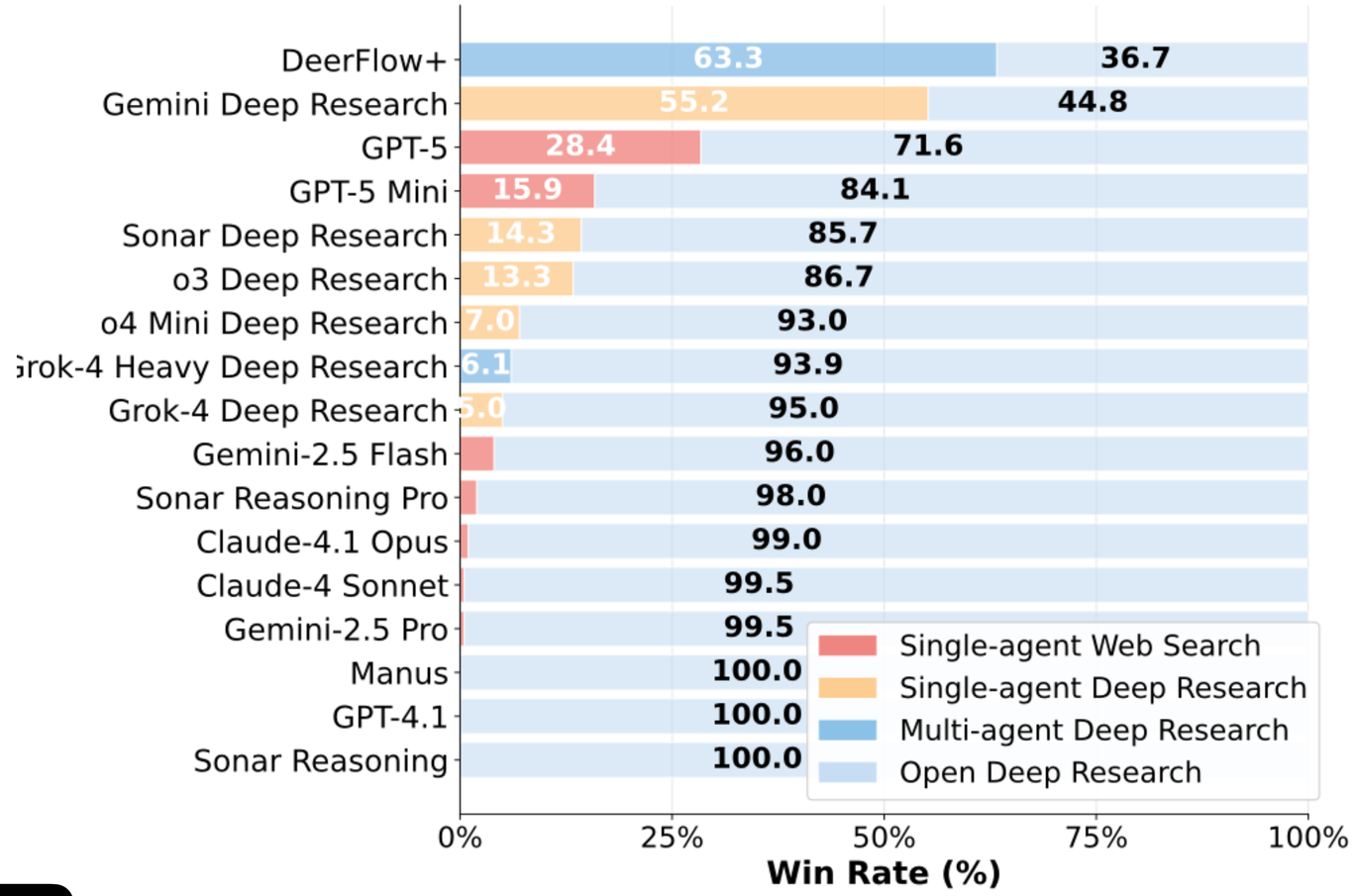
Main Results and Analysis

Agent Name	Presentation & Organization	Fact & Logic Consistency	Coverage & Comprehensiveness	Citation Association	Avg
Single-Agent with Web Search					
GPT-5	71.6	68.3	83.4	69.0	73.1
GPT-4.1	66.0	65.9	63.6	66.9	65.6
GPT-5-mini	61.4	66.9	80.5	62.2	67.7
Gemini 2.5 Pro	51.9	**76.5**	73.1	44.9	61.6
Claude 4 Sonnet	81.9	67.3	49.2	50.8	62.3
Claude 4.1 Opus	81.6	67.5	50.8	47.2	61.8
Perplexity Sonar Reasoning	82.1	73.0	40.7	61.7	64.4
Perplexity Sonar Reasoning Pro	79.6	71.9	46.7	65.0	65.8
Single-Agent Deep Research					
OpenAI o3 Deep Research	71.3	64.2	85.0	30.9	62.9
OpenAI o4-mini Deep Research	74.3	62.3	78.6	32.1	61.8
Perplexity Sonar Deep Research	**83.5**	67.4	65.5	52.1	67.1
Grok-4 Deep Research	69.1	57.4	86.3	64.7	69.4
Gemini Deep Research	62.1	63.0	75.8	64.6	66.4
Multi-Agent Deep Research					
Manus	75.0	63.1	73.3	53.8	66.3
Grok-4 Heavy Deep Research	75.9	59.4	**89.3**	64.7	72.3
Deerflow+ (w. GPT-5)	78.8	69.9	61.6	**81.4**	72.9
Open Deep Research (w. GPT-5)	81.0	71.3	65.3	77.2	**73.7**

Observation 5&6

Multi-agent systems lead presentation, but surface polish does not imply grounded quality.
Coverage benefits from specialization, but scaling retrieval scope and system complexity strain memory capacity.

Main Results and Analysis



Observation 7

Most systems are deep searcher not deep researcher
Despite longer reports, Gemini and o3 DR do not consistently outperform ODR in analysis depth

Main Results and Analysis

Agent Name	E1 Errors	E2 Errors	E3 Errors	Total
Task: Wide Info Search				
GPT-5	4.2	1.7	13.3	19.2
Grok-4 Deep Research	6.8	6.8	33.4	47.0
Open Deep Research	5.0	5.2	19.7	29.9
Task: Market Analysis				
GPT-5	11.1	10.1	43.8	65.0
Grok-4 Deep Research	6.3	6.4	61.5	74.2
Open Deep Research	11.9	11.6	68.4	91.9

Observation 8

Even SoTA systems are far from citation error-free

E1: URL invalid; E2: URL irrelevant; E3: claim not supported

Check our paper for more results and details!

Code and dataset are available

https://livedeepresearch.github.io/

https://github.com/SalesforceAIResearch/LiveResearchBench

https://huggingface.co/datasets/Salesforce/LiveResearchBench