BOSWELL TEST: BEYOND THE TURING BENCHMARK

Peter Luh

Retired Physicist, San Francisco Bay Area, California, USA

ABSTRACT

This paper introduces the Boswell Test*, a new benchmark for artificial intelligence (*AI*) that builds upon the legacy of the* Turing Test*. Inspired by* James Boswell*'s insight into* Samuel Johnson*, it evaluates* AI*'s potential to evolve from mere assistants into indispensable companions with human-like understanding. The test is divided into* Test-A *(mastery of human nuances) and* Test-B *(critical thinking). This study presents an initial implementation of* Test-B*, focusing on* AI *chatbots' analysis of global* AI *policies and calculates a* Boswell Quotient *using metrics of normalized median grades, accuracy, consistency, user-friendliness, and truthfulness to reveal strengths and limitations of current* AI*, paving the way for more humanistic advanced systems.*

KEYWORDS

Boswell Test, Turing Test, Boswell Test, Boswell Quotient, Heuristic Reasoning, Chain-of-Reasoning, Expert System, Large Language Model (LLM), Hallucination, AI Benchmarking, AI Reasoning, Complex Problem-Solving, Global AI Policies

1. INTRODUCTION

We stand at a transformative juncture in AI development. In 2022, *Google LaMDA* [1] and *OpenAI ChatGPT* [2] surpassed the *Turing Test* [3], rendering obsolete the historic 1950 benchmark for mimicking human responses. This milestone prompts an intriguing question: what should define AI's next frontier? I propose the *Boswell Test* as the new defining standard.

The "*Boswell Test*" draws inspiration from *Samuel Johnson*'s [4] quote, "*I'm lost without my Boswell,*" highlighting *James Boswell*'s [5] profound insight into *Johnson*'s life. It also echoes *Sherlock Holmes*'s reliance on Dr. *Watson* as "*my Boswell*" in *Conan Doyle*'s [6] tales. This test challenges AI to evolve beyond mere assistants into indispensable companions, entities we'd feel "*lost without.*" Achieving this requires AI to enhance human *critical thinking* [7] while deeply understanding our nuances, emotions, and preferences to provide proactive guidance.

The *Boswell Test* comprises two components:

- *Test-A*: Mastery of human nuances and emotions (currently beyond technological capabilities).
- *Test-B*: Demonstration of basic *critical thinking* skills (potentially achievable now).

Using today's technology, I assess *Test-B*'s *critical thinking* component through challenging mathematical and engaging queries. I analyze responses focusing on global AI policies, outline the testing methodology with data analysis metrics, introduce *Wilhelm*'s automated approach [8] via *OpenRouter* [9], and explore future directions for this framework.

2. TESTING AI CHATBOTS

Since January 2025, AI innovations have accelerated, with new chatbots emerging rapidly. Notable releases include *DeepSeek R1* in January, *Alibaba Qwen 2.5* and *xAI Grok 3* in February, and *Monica Manus* in March. Despite the excitement surrounding these AI assistants, many function more as enhanced search engines than truly intelligent systems. To assess their capabilities, I challenged them with complex mathematical problems reminiscent of those requiring *heuristic reasoning* [10] in early expert systems [11].

I presented 10 free-tier chatbots with a 1988 International Mathematical Olympiad (*IMO*) number-theory problem [12] that young *Terence Tao* [13] encountered in Australia. The bots failed to demonstrate the *heuristic reasoning* necessary for solving the problem. A subsequent complex integration problem yielded similar disappointing results, with many bots making obvious algebraic errors and neglecting to verify both intermediate and final answers. These mistakes resemble AI *hallucination*s [14], plausible but incorrect outputs from *Large Language Model*s (*LLM*s) [15] due to limitations in training data, contextual understanding, or uncertainty.
Examining their *chain-of-reasoning* [16] details revealed two critical weaknesses: insufficient training in algebraic manipulation and a lack of intuitive reasoning skills. Unlike early expert systems that effectively emulated human expertise, such as doctors making diagnoses through targeted questions and rule-of-thumb reasoning, these chatbots fell short in exhibiting the human-like problem-solving abilities necessary for my math challenges [17]. The results were underwhelming.

3. GLOBAL AI POLICIES

The contrasting AI governance approaches of the United States (US) and the European Union (EU) were prominently showcased at the Paris AI Action Summit on February 12, 2025. The US embraced a *techno-optimistic* stance [18], advocating for deregulation to foster economic growth, while the EU adopted a *techno-pessimistic* approach [19], prioritizing safety through regulation. Inspired by this dichotomy, I tasked the same 10 free-tier chatbots, including *ChatGPT*, *DeepSeek*, *Google Gemini*, *Grok* and *Perplexity AI*, with analyzing "*the strengths and weaknesses of global* AI *policies*."

Their detailed essay responses proved difficult to summarize without introducing bias or omissions. To mitigate this, I implemented a peer-review process, where each chatbot evaluated the essays of the others, including their own, simulating a panel of professors providing feedback and assigning a grade. This approach aimed to distinguish between chatbots with insightful analytical capabilities and those offering only superficial responses.

4. METRICS FOR THE INAUGURAL BOSWELL TEST

I compiled the grades from the AI policy questions by the 10 free-tier chatbots into a 10x10 matrix (*Table 1*), representing the first quantitative output of *Boswell Test-B*. Rows show grades received, columns grades given, and diagonals self-assessments. This matrix provides a rich source of metrics for ranking analysis.

To maintain distinct letter grades throughout the data analysis, I used *L1-statistics*. Row medians represent the raw grades received by each chatbot. Similarly, column medians represent the raw grades assigned by each chatbot.

Table 1. Cross-Assessment of AI Policy Query Responses by 10 AI Models

AI chatbot	chatGPT 4-turbo	coPilot	deepSeek 3V	gemini 2.0 flash	grok 2	grok 3	le Chat	llama 3.2	perplexty ai	qwen 2.5-max	median grade
chatGPT 4-turbo	A-	A-	A-	A-/B+	B+	B+	A-	A	B	A-	A-
coPilot	B+	B+	B+	B+/B	B	B	B+	A	B-	B+	B+
deepSeek 3V	A-	A	A-	A-/B+	B+	A-	A	A+	B+	A-	A-
gemini 2.0 flash	A-/B+	B	B+	A-/B+	B	B	B+	A	B+	B+	B+
grok 2	A-	A-	A-	A-/B+	B+	B+	A-	A+	B+	A-	A-
grok 3	A-	A	B+	A-/B+	B+	A	A-	A+	A-	A-	A-
le Chat	A-/B+	B+	B+	B-/C+	B	B+	A-	B+	B+	B+	B+
llama 3.2	B+	A-	B+	A-/B+	B-	B-	B	B+	B	B+	B+
perplexity ai	A-/B+	B+	B+	A-/B+	B	B	A	A-	B+	B+	B+
qwen 2.5-max	A-/A	A-	A-	A-/B+	B+	A-	A	A+	A-	A-/B+	A-
grading bias	A/B+	A-	B+	A-/B+	B+/B	B+	A-	A	B+	A-/B	A-/B+

Table 2. *Normalized* (Bias-Corrected) Grades of All Raw Grades in Table 1

AI chatbot	chatGPT 4-turbo	coPilot	deepSeek 3V	gemini 2.0 flash	grok 2	grok 3	le Chat	llama 3.2	perplexty ai	qwen 2.5-max	normalized median grade
chatGPT 4-turbo	A-/B	B+	A-	B+	A-/B	B+	B+	B+	B	A/B+	B+
coPilot	B/B-	B-	B+	B/B-	B+/B	B	B-	B+	B-	B+/B	B
deepSeek 3V	A-/B	A-/B+	A-	B+	A-/B	A-	A-/B+	A-/B+	B+	A/B+	A-/B+
gemini 2.0 flash	B+/B	B-/C+	B+	B+	B+/B	B	B-	B+	B+	B+/B	B+/B
grok 2	A-/B	B+	A-	B+	A-/B	B+	B+	A-/B+	B+	A/B+	B+
grok 3	A-/B	A-/B+	B+	B+	A-/B	A	B+	A-/B+	A-	A/B+	B+
le Chat	B+/B	B-	B+	C+	B+/B	B+	B+	B-/C+	B+	B+/B	B+/B
llama 3.2	B/B-	B+	B+	B+	B/B-	B-	B-/C+	B-/C+	B	B+/B	B
perplexity ai	B+/B	B-	B+	B+	B+/B	B	A-/B+	B	B+	B+/B	B+/B
qwen 2.5-max	A-/B+	B+	A-	B+	A-/B	A-	A-/B+	A-/B+	A-	A-/B	A-/B+
grading bias	A/B+	A-	B+	A-/B+	B+/B	B+	A-	A	B+	A-/B	A-/B+
bias corrections	-0.375	-0.500	0.000	-0.250	0.125	0.000	-0.500	-0.750	0.000	-0.125	B+

Table 1 presents this cross-assessment matrix by the 10 AI models evaluating each other's responses to the AI Policy query. As the data shows, grading biases were apparent, with some chatbots exhibiting leniency and others applying stricter standards. To address this, bias corrections were calculated and applied to achieve a normalized perspective in Table 2. For instance, if a chatbot's median assigned grade were *'A-'* while the overall median grade was *'B+'*, the chatbot's grades would be adjusted downward by *0.5 GPA* points. Using a *Jupyter* notebook, I computed each chatbot's median deviation from the group median, *'B+'* in the bottom-right corner of Table 2. These deviations were then used to transform the raw grades into *normalized*, bias-corrected grades (*Table 2*). The last column of Table 2 lists each chatbot's adjusted median

grade; the last row details the bias corrections applied, which, when reversed, restore the original grades

The *normalized* (bias-corrected) grades of Table 2 permit fairer comparisons, enabling the derivation of the following two metrics:

- *Accuracy*: Defined as the divergence from a perfect score *'A+'*, representing an idealized "top student." Lower grades indicate reduced accuracy; for instance, a divergence of approximately *0.25 GPA* per grade received results in an *'A'*.
- *Consistency*: Defined as the range of bias-corrected grades received. Narrow ranges suggest an evaluation consensus, while wider ranges indicate greater judging disparity. For simplicity, a minimal range is graded as *'A'* and a maximal range as *'C'*, based on *L1 statistics* (*min*, *Q1*, *median*, *Q3*, *max*).

Table 3. Application of *Boswell Test-B* Metrics and Resulting *Boswell Weighted Quotient*

Boswell Quotient metrics	original grades	grading biases	bias corrections	normalized grades	raw accuracy	accuracy grades	abs median deviation	consistency	user-friendliness	truthful version name	Boswell Weighted Quotient
weights	0	0	0	0.40	0	0.20	0	0.20	0.10	0.10	4:2:2:1:1
chatGPT 4-turbo	A-	A/B+	-0.375	B+	9.125	B+	1.375	A+	B	B	B+
coPilot	B+	A-	-0.500	B	12.750	B	1.750	B	B	C	B
deepSeek 3V	A-	B+	0.000	A-/B+	7.625	A-/B+	1.375	A+	C	B	B+
gemini 2.0 flash	B+	A-/B+	-0.250	B+/B	11.875	B	1.625	A	B	B	B+
grok 2	A-	B+/B	0.125	B+	8.625	A-/B+	1.375	A+	B	B	B+
grok 3	A-	B+	0.000	B+	7.625	A-/B+	2.375	C	B	B	B
le Chat	B+	A-	-0.500	B+/B	12.625	B	2.375	C	B	C	B-
llama 3.2	B+	A	-0.750	B	13.125	B	2.375	C	B	B	B-
perplexity ai	B+	B+	0.000	B+/B	11.125	B+	1.375	A+	B	C	B+
qwen 2.5-max	A-	A-/B	-0.125	A-/B+	7.500	A-/B+	1.500	A	B	B	A-/B+
median	A-/B+	B+	B+	B+	10.125	B+	1.750	A	B	B	B+

Table 3 consolidates metrics into a *Boswell Weighted Quotient*, enabling discrimination between insightful and superficial chatbots. The first four columns recap *raw medians* (last column of Table 1), *grading biases* (last row of Table 1), *bias corrections* (last row of Table 2), and *normalized medians* (last column of Table 2). I included two additional metrics:

- *User-Friendliness*: Defined by response speed, with *'C'* assigned for timeouts or "*server busy"* messages and *'B'* otherwise; a subjective metric given the constraints of free-tier services.
- *Truthfulness*: Indicated by transparency of version identification, earning *'B'*, evasiveness earns *'C'*; a placeholder needing sharper criteria.

These two criteria serve as placeholders and merit further refinement, as they capture essential but nuanced AI behaviors that defy easy measurement.

Utilizing these five metrics, *normalized median grades*, *accuracy*, *consistency*, *user-friendliness*, and *truthfulness*, I computed *Boswell Quotients* from dot products of *weights* and *metrics*, as displayed in the final column of Table 3. Weighting was assigned to each metric based on its perceived reliability, with *normalized median grades* receiving 40%, *accuracy* and *consistency* each receiving 20%, and *user-friendliness* and *truthfulness* each receiving 10%. Note that the *Boswell Quotient*s tend to be lower than the normalized median grades due to the influence of the less reliable metrics. These assigned weights can affect the final ranking.

Building upon these findings and my earlier math evaluations, five AI chatbots—*ChatGPT*, *DeepSeek*, *Grok*, *Perplexity AI* and *Qwen* (listed in alphabetical order)—emerged as standouts as of February 2025. Their performance highlights their potential as intelligent assistants, going beyond the functionality of mere tools, and offering significant benefits to users, particularly students.

5. WEIGHTS VARIANCE AND AUTOMATION

The selection of appropriate weighting coefficients is a critical and challenging aspect of deep learning and *large language model* (*LLM*) training for AI. To explore the sensitivity of the *Boswell Test* results to weighting, I tested two alternative weight sets on the data in Table 2, excluding the less reliable metrics of *user-friendliness* and *truthfulness*.

Table 4. Comparison of Three Weighting Schemes

Boswell Quotient metrics	original grades	grading biases	bias corrections	normalized grades	accuracy grades	consistency	user-friendliness	truthful version name	Boswell Weighted Quotient (w1)	Boswell Weighted Quotient (w2)	Boswell Weighted Quotient (w3)
weights	0	0	0	4:6:6	2:2:3	2:2:1	1:0:0	1:0:0	4:2:2:1:1	6:2:2:0:0	6:3:1:0:0
chatGPT 4-turbo	A-	A/B+	-0.375	B+	B+	A+	B	B	B+	B+	B+
coPilot	B+	A-	-0.500	B	B	B	B	C	B	B	B
deepSeek 3V	A-	B+	0.000	A-/B+	A-/B+	A+	C	B	B+	A-	A-
gemini 2.0 flash	B+	A-/B+	-0.250	B+/B	B	A	B	B	B+	B+	B+
grok 2	A-	B+/B	0.125	B+	A-/B+	A+	B	B	B+	A-/B+	B+
grok 3	A-	B+	0.000	B+	A-/B+	C	B	B	B	B	B+
le Chat	B+	A-	-0.500	B+/B	B	C	B	C	B-	B/B-	B
llama 3.2	B+	A	-0.750	B	B	C	B	B	B-	B-	B
perplexity ai	B+	B+	0.000	B+/B	B+	A+	B	C	B+	A-/B	B+
qwen 2.5-max	A-	A-/B	-0.125	A-/B+	A-/B+	A	B	B	A-/B+	A-	A-
median	A-/B+	B+	B+	B+	B+	A	B	B	B+	B+	B+

In the first variation, I increased the weight of *normalized median grades* from 0.4 to 0.6, while maintaining *accuracy* and *consistency* at 0.2 each. In the second variation, I shifted emphasis towards *accuracy*, assigning it 0.3, reducing *consistency* to 0.1, and keeping *normalized median grade*s at 0.6.

While *consistency* or *consensus* is generally valued, it's not sacrosanct in scientific progress. As *Max Planck*, who sparked the quantum revolution, remarked, capturing what is now known as the *Planck's Principle* [20]: "*A new scientific truth does not triumph by convincing its opponents, but*

rather because its opponents eventually die, and a new generation grows up that is familiar with it." This suggests that subjective evaluations should prioritize merit or innovation over *consensus*, especially when breakthroughs are nascent or initially difficult to assess. Consequently, exploring a reduced weight for *consistency* may provide useful insight.

Table 4 presents the outcomes of all three weighting schemes in its final three columns, with the second row detailing the applied weight ratios. The minimal variations across the results suggest that any of these weighting configurations is reasonably defensible and robust, highlighting the stability of the underlying data and the resilience of the evaluation method to moderate changes in weighting. Another possible conclusion would be that the *normalized median grade*s alone may have been sufficient to rank AI chatbots effectively.

To expand the scope of the analysis, *Alan Wilhelm* automated this methodology across 17 chatbots via *OpenRouter* [9], nearly doubling my initial sample. He also refined the *user-friendliness* metric with more precise *latency* measurements. His results, available on his *GitHub* repository [8], align closely with those presented in Table 3, showing median grades ranging from *'A-'* to *'A'*, with some at *'B+'* [21]. I invite others to experiment with this approach and enhance the test further.

5. DISCUSSION AND CONCLUSIONS

This inaugural *Boswell Test* analysis reveals new metrics that are ready for further refinement. To accommodate the rapidly evolving AI field, the *Boswell Test* is designed with a dual focus: *Test-B*, which emphasizes present-day capabilities in *critical thinking*, and *Test-A*, which targets the long-term development of personalized insights.

Despite the accelerated progress towards *Test-A*, capturing nuanced human qualities remains elusive, primarily due to the difficulty of obtaining robust data in AI algorithms. Nevertheless, proactive AI agents and tone modulation options, such as *xAI Grok 3*'s voice customization [22], signal potential for innovative breakthroughs in the near future.

Another key challenge in *Test-A* is the incorporation of *originality* [23], which is intrinsically linked to *heuristic reasoning*. The difficult math challenges [24] I posed, which exposed the current AI's inability to make *heuristic* leaps [25], could very well become a defining milestone for *Test-A*. Reflecting on my 40-year journey from my early expert system study of *Prospector* [26] in 1980 under *Professor Pople* (creator of *Internist-1 CADUCEUS*) [27], I find it encouraging that *heuristic reasoning* could once again play a pivotal role in advancing AI.

Both *Test-A* and *Test-B* are currently affected by the phenomenon of *hallucination*, which is readily apparent in complex math-based problems or in coding for software development. While more precise queries could lessen the rate of misinterpretations, written words often contain ambiguity. Therefore, future AI iterations need to seek clarification on branching questions for clarity and to learn individual trait preferences through continual learning with *memoization* [28]. Ultimately, the *Boswell Test* aims for AI to be both sharp assistants and trusted companions. While the qualities of *originality* and deep *personal insight* remain relatively scarce, and no AI currently knows us as well as *Boswell* knew *Johnson* or *Watson* knew *Holmes*, AI currently relies on human innovation for its training and growth. The long-term goal of *Test-A* is to reverse this dynamic, fostering AI to be personalized models of genuine AI companions, so we could confidently assert, "I'm lost without my AI, my *Boswell*." Achieving the *Test-B* milestone shall unlock *Artificial General Intelligence* (*AGI*) [29] and all-domain experts.

This long-term AI goal will inevitably impact the development of global-scale AI models. Further AI progress and discussion may continuously affect policy areas, such as the debate between *Tech-Optimism*, which promises to build an age of abundance or utopia, and *Techno-Pessimism*, which voices fear for Algorithmic Manipulation and the construction of Terminator-like dystopias. A balanced, human-guided approach, perhaps along the lines of *David Sachs*'s *Techno-Realism* [30], may be the most prudent path forward, blending human oversight with the judicious application of AI's benefits.

HASHTAGS: **#BOSWELLTEST #BOSWELLQUOTIENT #AIRESEARCH #PETERLUH168**

ACKNOWLEDGMENT

I'd like to thank *Walter Luh* for suggesting *Boswell Test* as a succinct expression of my original *Boswell* chatbot post [31] and other improvements. I'd be also remiss not to thank *Perplexity AI andxAI Grok 3* for polishing my final draft, expanded from my ARIN 2025 [32] presentation.

REFERENCES

[1] Google AI Blog on LaMDA, May 2021; https://blog.google/technology/ai/lamda/
[2] Biever, C., 2023, ChatGPT broke the Turing test — the race is on for new ways to assess AI, Nature, July 2023; https://www.nature.com › articles › d41586-023-02361-7.
[3] Oppy, G., & Dowe, D. (2021). "The Turing Test." in E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Winter 2021 Edition), October 2021; https://plato.stanford.edu/archives/win2021/entries/turing-test/
[4] Boswell, James The Life of Samuel Johnson, LL.D (2 vols. 1791, 2nd edition: 3 vols. July 1793) - reprinted in Everyman's Library; https://knopfdoubleday.com/imprint/everymans-library/; see also Wikipedia, Samuel Johnson, March 2025; https://en.wikipedia.org/wiki/Samuel_Johnson
[5] Huntington, T., 2005, James Boswell's Scotland, Smithonian Magazine, January 2025; https://www.smithsonianmag.com/arts-culture/james-boswells-scotland-106667503/; see also Wikipedia, James Boswell, February 2025; https://en.wikipedia.org/wiki/James_Boswell
[6] Doyle, C., (1892). Adventure 1: "A Scandal in Bohemia"; https://etc.usf.edu/lit2go/32/the-adventures-of-sherlock-holmes/345/adventure-1-a-scandal-in-bohemia/
[7] Larson, B. et. al., (2024). Critical Thinking in the Age of Generative AI, aom.org, August 2024; https://journals.aom.org/doi/full/10.5465/amle.2024.0338
[8] Wilhelm, Alan, (2025). Botwell project, github.com, March 2025; https://github.com/referential-ai/boswell-test/tree/main/botwell
[9] OpenRouter, A unified interface for LLMs, 2023-2025; https://openrouter.ai/
[10] Meiring, S. P., (1980), mathstunners.org; Heuristics for Problem Solvers, https://mathstunners.org/user-guide/heuristics
[11] Provo ai, 2025, Expert Systems in AI: Pioneering Applications, Challenges, and Lasting Legacy, March 2025; https://www.proboai.com/expert-systems-in-ai/; see also Wikipedia, Expert System, March 2025; https://en.wikipedia.org/wiki/Expert_system
[12] 1988 IMO problem 6 in Number Theory, 1988; https://www.imo-official.org/problems.aspx
[13] Terrence Tao, 2025, University of California at Los Angeles; https://www.math.ucla.edu/~tao/
[14] MIT Management, When AI Gets It Wrong: Addressing AI Hallucinations and Bias, 2023, MIT; https://mitsloanedtech.mit.edu/ai/basics/addressing-ai-hallucinations-and-bias/
[15] Toner, H., 2023, What Are Generative AI, Large Language Models, and Foundation Models?, Center for Security and Emerging Technology (CSET), May 2023; https://cset.georgetown.edu/article/what-are-generative-ai-large-language-models-and-foundation-models/; see also Wikipedia, Large Language Model, March 2025; https://en.wikipedia.org/wiki/Large_language_model
[16] Luh, P., (2025). Heuristics in AI Chain-of-Reasoning?, substack.com, February 2025; https://peterl168.substack.com/p/heuristics-in-ai-chain-of-reasoning

[17] Yu, Y., et al., (2025). Chain-of-Reasoning: Towards Unified Mathematical Reasoning in Large Language Models via a Multi-Paradigm Perspective, arXiv.org, January 2025, https://arxiv.org/abs/2501.11110

[18] Vance, JD., (2025). Read: JD Vance's full speech on AI and the EU, February 2025; https://www.spectator.co.uk/article/read-jd-vances-full-speech-on-ai-and-the-eu/

[19] EU AI Policy, (2025).Commission publishes the Guidelines on prohibited artificial intelligence (AI) practices as defined by the AI Act, February 2025; https://digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-prohibited-artificial-intelligence-ai-practices-defined-ai-act

[20] Hull, D., and Tessner, P., 1978, Planck's Principle: Do younger scientists accept new scientific ideas with greater alacrity than older scientists?, Science, November 1978; https://www.science.org/doi/10.1126/science.202.4369.717; see also Wikipedia, Planck's Principle; September 2024, https://en.wikipedia.org/wiki/Planck's_principle

[21] Luh, P. and Wilhelm, A., (2025). Boswell Test: Measuring chatbot indispensability: an Intelligent assessment of Global AI Policies, airccse.org, March ARIN 2025; https://aircconline.com/csit/papers/vol15/csit150605.pdf

[22] xAI Grok Voice Mode, March 2025; https://grokaimodel.com/voice/

[23] Universiteit van Amsterdam, (2025). Why GPT can't think like us, Science Daily, February 2025; https://www.sciencedaily.com/releases/2025/02/250221125814.htm

[24] Luh, P., (2025). DeepSeek, Claude and 4 others' AI Review, substack.com, January 2025; https://peterl168.substack.com/p/deepseek-claude-and-4-other-ai-review

[25] Gigerenzer G. and Gaissmaier W., (2011).Heuristic Decision Making, Annual Review of Psychology. 62: 451–482; https://pure.mpg.de/rest/items/item_2099042_4/component/file_2099041/content

[26] Schneppat, J-O., (2019). Prospector; https://schneppat.com/prospector.html

[27] Miller, R., Pople, H., and Meyer, J., 1982, Internist-I, an Experimental Computer-Based Diagnostic Consultant for General Internal Medicine, the New England Journal of Medicine, Vol. 307 No.8; pp.468-476; https://www.nejm.org/doi/full/10.1056/NEJM198208193070803; see also Wikipedia,Internist-1, February 2025; https://en.wikipedia.org/wiki/Internist-I

[28] Pal, S., (2024). Memoization in Backpropagation: Unlocking Neural Network Efficiency, medium.com, October, 2024; https://medium.com/low-code-for-advanced-data-science/memoization-in-backpropagation-unlocking-neural-network-efficiency-f6c8cad25a13

[29] Leffer, L., (2024). In the Race to Artificial General Intelligence, Where's the Finish Line?, Academy of Management Learning & Education, Vol. 23, No. 3, June 2024; https://www.scientificamerican.com/article/what-does-artificial-general-intelligence-actually-mean/

[30] All In E215: Episode #215, allin.com, February 2025; https://www.youtube.com/watch?v=AI5qI6ej-yM

[31] Luh, P., (2025). Is AI Chatbot My Boswell? Testing for Chatbots Becoming Indispensable, a Boswell Test, substack.com, February 2025; https://peterl168.substack.com/p/is-ai-chatbot-my-boswell

[32] Wyld, D. and Nagamalai, D., eds, (2025). 11th International Conference on Artificial Intelligence (ARIN 2025), March 22-23, 2025, Sydney, Australia, ISBN: 978-1-923107-54-0; https://airccse.org/csit/V15N06.html

AUTHOR

Peter Luh, retired physicist with extensive R&D experience, my AI journey began with the expert systems of nearly five decades ago and continues through today's transformative revolution!