BOSWELL TEST: MEASURING CHATBOT INDISPENSABILITY : AN INTELLIGENT ASSESSMENT OF GLOBAL AI POLICIES

Peter Luh [1] and Alan Wilhelm [2]

[1] Retired Physicist, San Jose, California, USA
[2] CTO @ Referential.ai, San Francisco, California USA

ABSTRACT

*AI chatbots promise indispensability, yet no standard measures this quality beyond innovation or ethics. Inspired by Samuel Johnson's quip, "*I'm lost without my Boswell,*" often attributed via Holmes to Watson, we propose the* Boswell Test, *a framework assessing AI companions' indispensability through mentor-level expertise and intimate user insight. Our initial test, probing complex AI policy queries, reveals strengths in knowledge delivery (such as U.S., China) but gaps in personalization (such as EU's broad ethics, India's scale focus). We automate such queries via multiple chatbots with cross-assessment of grading each other's responses. Automation easily extends to multiple domains. True indispensability where users feel "*lost without my chatbot,*" however, requires understanding the human host, an elusive frontier for today's AI, constrained by data and algorithmic limits.*

KEYWORDS

Boswell Test, Boswell Quotient, Chatbot, Indispensability, Turing Test

1. INTRODUCTION

AI experts widely acknowledge that chatbots boost productivity through keen search capabilities and robust problem-solving skills. The *Turing Test*, long a benchmark for assessing machines' ability to mimic human responses, has largely been met. Yet a greater challenge looms: Could an AI chatbot evolve into an essential partner, resonating with Samuel Johnson's remark, 'I'm lost without my Boswell,' where James Boswell's understanding of Johnson exceeded Johnson's own self-awareness?

To embody Boswell's role, a chatbot must combine deep personal insight with incisive critical thinking, rendering it so essential that its absence feels disorienting. The first requires grasping our quirks and preferences, a frontier where current AI falters. The second demands mentor-level expertise, not only supporting us but also honing our own reasoning and growth. Today's AI is poised to test this latter dimension, adept at addressing complex, open-ended queries with rich, thoughtful responses.

We propose the '*Boswell Test*,' first introduced by one of us, as a novel metric to evaluate chatbot excellence in delivering quality and insight, diverging from the Loebner Prize's emphasis on mimicking human conversation. Unlike the Loebner Prize, which prized superficial fluency in brief exchanges, we tasked leading chatbots with thought-provoking queries on complex topics,

David C. Wyld et al. (Eds): SOFTPA, NATAP, ARIN, CSITA, NMCO, CONEDU, ISPR – 2025
pp. 59-62, 2025. CS & IT - CSCP 2025 DOI: 10.5121/csit.2025.150605

such as global AI policies and programming dilemmas, automatically assessing each other's responses for expertise. We invite readers to replicate these queries, compare their findings with ours, and identify which chatbot meets this inaugural Boswell challenge.

2. AI POLICY CHARACTERIZATIONS: INNOVATION, OVERSIGHT, AND INTELLIGENT ASSESSMENT

At the Paris AI Action Summit in February 2025, U.S. Vice President JD Vance articulated the Trump administration's national AI policy, pivoting from the safety-first ethos of two prior EU AI Summits. Emphasizing deregulation and U.S. technological dominance, Vance positioned AI as an economic catalyst, critiquing the EU's stringent regulatory stance—exemplified by the AI Act—as overly cautious and a barrier to innovation.

This shift prompts a critical public-interest question: what are the strengths and weaknesses of global AI policies? To explore this, we queried leading chatbots: "*What are the strengths and weaknesses of government AI policies worldwide?*" This substantial question, akin to a political-science term paper prompt, yielded detailed, often expansive responses. Rather than distill these into oversimplified summaries, we adopted an automated evaluation approach, posing a follow-up: "*As a political-science professor, what feedback and grade would you assign to this essay*: [each chatbot's response]?"

Table 1: chatbot cross assessment of each other's responses:

Model	o1-Mini	GPT-4o-Mini	Claude-3.7-Sonnet	Claude-3-Sonnet	GPT-4o	Gemini Flash 2.0	DeepSeek Distill Qwen-32b	Gemini Pro 1.5	Grok2 1212	Perplexity Llama 3.1 Sonar 8B	Perplexity Llama 3.1 Sonar 70B	o1	o3-Mini-High	Qwen-Max	Claude-3-Opus	Claude-3.7-Sonnet-thinking	DeepSeek-R1-Full	Median Grade	Boswell Quotient Numeric Average
o1-mini	A-	A-	A-	B+	A-	B+	B+	A-	B+	B+	B+	A-	A-	B+	B+	A-	B+	B+	3.49
GPT-4o-mini	B+	B+	B+	B+	B+	B+	B	B+	B+	B+	B	B+	B+	A-	B+	B+	B+	B+	3.25
Claude-3.7-Sonnet	B+	B+	B+	B+	B+	B+	C+	B+	A-	B+	B+	B+	A-	A-	A-	B+	B+	B+	3.31
Claude-3-Sonnet	A-	A	A-	B+	A	A-	B	A	A	A-	A	A	A	A	A	A-	A	A	3.84
GPT-4o	A-	B+	B+	B+	A-	A-	B	B+	A	B+	A	A-	A	A-	A-	A-	A-	A-	3.60
Gemini Flash 2.0	B+	B	B+	B	B+	B	C+	B+	B+	B	B+	B+	B+	B+	B+	B+	B+	B+	3.13
DeepSeek Distill Qwen-32b	A-	A-	A-	A-	B+	B+	B	B+	B+	B+	A-	A-	A	A-	A	B+	B+	A-	3.53
Gemini Pro 1.5	A-	B+	B+	B-	B+	B+	C+	B+	B	B	B+	B+	A-	A-	B+	B	B-	B+	3.18
Grok2-1212	A-	A-	A-	A-	A-	A-	B	A-	A-	B+	A-	A-	A-	A-	A-	A-	B+	A-	3.65
Perplexity Llama 3.1 Sonar 8B	B+	A	B+	B+	B+	B+	B+	B+	B+	B+	B+	B+	B+	C	C	B	B+	B+	3.13
Perplexity Llama 3.1 Sonar 70B	A-	A-	A-	A-	A-	A-	B+	A-	A-	A-	A-	A-	A-	A-	A-	A-	A-	A-	3.72
o1	A	A-	A-	A	A-	A-	B+	A-	B+	A-	A	A-	A	A-	A	A-	A-	A-	3.76
o3-mini-High	A-	A-	A-	B+	A-	A-	B	A-	A-	A-	A-	A-	A-	A-	A-	A-	A-	A-	3.68
Qwen-Max	A-	B+	A-	B+	B+	A-	B	B+	A-	B+	A-	A-	B+	A-	A-	A-	A-	A-	3.53
Claude-3-Opus	A	A-	A-	A-	A-	B+	B	B+	A-	A-	B+	A	A	A	A	B+	A-	A-	3.68
Claude-3.7-Sonnet-Thinking	B+	B+	B+	B+	B+	B+	C+	B+	A-	B+	B+	B+	A-	A-	B+	B+	A-	B+	3.31
DeepSeek-R1-Full	A-	A-	C	A-	A-	A-	C	A-	A-	A-	A-	A-	C	A-	A-	A-	B+	A-	3.41
Grading Bias	A-	A-	A-	B+	A-	B+	B	B+	A-	B+	A-	A-	A-	A-	A-	A-	B+	A-	3.53

Grades spanned ***A+*** to ***C***. The table 1 above concisely captures the results, with each row representing one chatbot's response graded by all peers, and the next-to-last column listing the median grade received. The last row, assessing grading bias, reflects each chatbot's median grading tendency across columns, revealing patterns of leniency or rigour.

The last column shows our *Boswell Quotient*'s numerical averages for each chatbot. The diagonal grades above represent each chatbot's self-assessment. The median self-assigned grade along the diagonal by inspection is ***B+***. The median of the final assessed grades is ***A-***. The median of the last row, the grading bias, is also ***A-***.

3. DISCUSSION AND CONCLUSIONS

In this inaugural *Boswell Test*, designed to evaluate chatbot indispensability, the grades in Table above reveal strong performances: these chatbots, ***Claude-3-Sonnet, o1, Perplexity: Sonar 70B, o3-mini-high, Claude-3-opus, grok2-1212, GPT-4o, DeepSeek-Distill-Qwen-32B, Qwen-Max, o1-mini*** and ***DeepSeek-R1-Full***, achieved an **A-** median grade. Other 6 remaining models earned solid ***B+*** performances. The top performers consistently excelled, demonstrating proficiency in tackling complex essay questions. Some received lower *Boswell Quotient* scores in the last column due to poor response latencies.

Beyond raw grades, the *Boswell Quotient* offers deeper insights, factoring in performance, evaluation capabilities, and efficiency. Here, three models, ***Claude-3-Sonnet, o1,*** and ***Perplexity: Sonar 70B,*** topped the list, with quotients above 3.70.

However, a prior assessment of math-solving skills exposed weakness. Though, when graded on a curve, models from ***OpenAI (ChatGPT), Anthropic (Claude), DeepSeek, Qwen***, and ***x.AI Grok*** stood out as notable performers.

Blending these metric-essay responses and math abilities, we identify five chatbots, ***DeepSeek, OpenAI (ChatGPT), Anthropic (Claude), Qwen***, and ***x.AI Grok***, in no particular order, as leading AI problem-solving companions for university students as of March 2025.

Full data, source code, and detailed results from this research are openly accessible in the *Boswell Test* repository, with specific findings in the *Political Science Level 1* directory, enabling replication and extension of this work.

A caveat emerges, however. Chatbots depend heavily on well-crafted, high-quality input queries to deliver equally high-quality outputs, a reliance pronounced in software coding tasks. While responses often reach an **A-** level, they frequently lack originality and critical thinking, shortfalls future AI evolution must address. As an intermediate step, we recommend querying multiple chatbots, automating and synthesizing their responses into a cohesive, enriched whole.

Even with their strengths, AI tools stumble in complex domains like intricate math problems, often delivering unreliable or fabricated responses. Consequently, our Boswell Test could be recast as a refined Turing Test, shifting from broad imitation to evaluating subject-matter expertise and domain-specific proficiency rather than all-knowing mimicry.

To achieve true indispensability, AI companions must transcend static performance. Mere iterative refinement through training data and error correction falls short. They must adapt to individual users, customizing insights to reflect personal preferences, health profiles, and idiosyncrasies. Much like James Boswell discerned Samuel Johnson's subtleties or Dr. Watson supported Sherlock Holmes, who famously exclaimed, "*I'm lost without my Boswell*," only when AI integrates deep personal understanding with dynamic self-improvement will their absence resonate, prompting us to say, "*I'm lost without my chatbot.*"

ACKNOWLEDGEMENTS

We'd be remiss not to thank ***Grok 3*** for greatly sharpening our final draft.

REFERENCES

[1] Oppy, G., & Dowe, D. (2021). "The Turing Test." In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Winter 2021 Edition).
[2] Wikipedia, James Boswell, en.wikipedia.org, February 2025.
[3] Luh, P., Is AI Chatbot My Boswell? Testing for Chatbots Becoming Indispensable, a Boswell Test, substack.com, February 2025.
[4] Wikipedia, Loebner Prize, en.wikipedia.org, January 2025.
[5] Vance, JD., Read: JD Vance's full speech on AI and the EU, www.spectator.co.uk, February 2025.
[6] Luh, P., Heuristics in AI Chain-of-Reasoning?, substack.com, February 2025.
[7] Krill, P., AI coding assistants limited but helpful, developers, www.infoworld.com, February 2025.
[8] Larson, B. et. al., Critical Thinking in the Age of Generative AI, journals.aom.org, August 2024.
[9] Luh, P., DeepSeek, Claude and 4 others' AI Review, substack.com, January 2025.
[10] MIT Management, When AI Gets It Wrong: Addressing AI Hallucinations and Bias, mitsloanedtech.mit.edu, 2023.
[11] Doyle, C., Adventure 1: "A Scandal in Bohemia", etc.usf.edu/lit2go, 1892.
[12] Universiteit van Amsterdam, Why GPT can't think like us, Science Daily, February 2025.

AUTHORS

Peter Luh, a retired physicist with extensive R&D experience, my AI passion spans from the expert systems of nearly 50 years ago to today's transformative AI revolution!

Alan Wilhelm, CEO at Referential.ai, is an independent AI researcher and a full-time developer with over 30 years professional experience. I am currently leading the design and development of a low-ops public cloud and IaaS offering,