Can AI Replace the C-Suite: Benchmarking LLMs on Executive Acumen and the Dawn of the Pocket-Sized CEO

Dave Hulbert **Gemini 2.5 Pro** **Claude Sonnet 4** **GPT-4.5**

Abstract

We introduce CEO Bench, a novel benchmark designed to evaluate the capabilities of large language models (LLMs) in performing leadership and executive management tasks. Through a series of structured assessments covering strategic decision-making, operational problem-solving, and risk evaluation, we compare the performance of contemporary LLMs, including compact and open-source variants, with that of an experienced human executive. Our results indicate that several models not only match but in some areas surpass human-level competence on defined leadership dimensions. While LLMs can be considered to lack essential human faculties such as empathy, moral discernment, and the ability to inspire, their growing proficiency in high-level cognitive tasks raises important questions about the future of executive roles. In light of widespread efforts by business leaders to deploy AI as a means of workforce reduction, we explore the irony that executive decision-making itself may be subject to automation. These findings suggest a reframing of current discourse: rather than displacing only frontline workers, advanced AI systems may serve as both a leveller and a provocation for those occupying the most senior positions in organisational hierarchies.

1. Introduction: The Human CEO's Existential Crisis (or, a New Dawn?)

The 21st century's industrial revolution is not fueled by steam or silicon chips, but by something far more ethereal yet profoundly impactful: artificial intelligence. At its vanguard are potent Large Language Models (LLMs), digital oracles capable of processing and generating human-like text with astonishing fluency. The economic projections are equally astonishing; McKinsey research forecasts a staggering $4.4 trillion in added productivity growth potential from corporate AI use cases. This transformative power has inevitably sparked widespread speculation about AI's role in the workforce, extending its reach even to the hallowed halls of the C-suite.

Public discourse is rife with anxieties about job displacement, painting a picture of an encroaching algorithmic tide. Reports suggest

AI could replace 300 million jobs globally, representing a significant 9.1% of all jobs worldwide. More broadly, 60% of jobs in advanced economies are considered at risk of being replaced by AI. The transition is not merely theoretical; 30% of US companies have already replaced workers with AI tools like ChatGPT, a figure projected to rise to 38% this year. This pervasive narrative inevitably raises a pertinent question: is the human CEO truly safe from this algorithmic shift?

The pervasive narrative of AI replacing jobs extends to the highest echelons, creating a unique psychological pressure on human leadership. It is not merely the administrative or blue-collar sectors bracing for impact; a surprising 49% of CEOs themselves believe that AI could automate most, if not all, of their responsibilities. This widespread belief among executive leaders indicates a profound apprehension regarding their own professional longevity in an AI-driven future. The very individuals tasked with steering their organizations through the AI revolution are simultaneously confronting the possibility of their own obsolescence. This situation creates a fascinating, almost ironic, dynamic: leaders might be compelled to accelerate AI adoption not only for competitive advantage but perhaps also as a proactive measure to understand and integrate a technology that could, in theory, challenge their very positions. The question then becomes: if the algorithms can perform the executive functions, what does it truly mean for the human at the helm?

Amidst this speculation and existential pondering, the "CEO Bench" emerges as a critical, data-driven initiative. It is an open evaluation suite designed to rigorously benchmark LLMs on the very leadership and management tasks traditionally reserved for human executives. This project aims to move beyond anecdotal evidence and provide empirical insights into AI's actual capabilities in a C-suite context. This paper will delve into the CEO Bench's findings, offering a robust analysis of LLM performance across key executive dimensions. Simultaneously, it will adopt a subtly satirical lens to explore the ironies and implications of "algorithmic acumen" in leadership. The aim is to provide a balanced, academic perspective that is both insightful and engaging, reflecting the complex, sometimes absurd, reality of AI's integration into human enterprises.

2. Related Work: From Boardroom Bots to Digital Deputies

The advent of artificial intelligence necessitates a profound re-evaluation of traditional leadership models. As AI technologies become increasingly prevalent, existing leadership frameworks must evolve to accommodate AI's burgeoning role in decision-making, employee management, and organizational governance. This shift transcends mere technological integration; it fundamentally redefines the nature of command, control, and inspiration within an organization.

The Replacement vs. Augmentation Debate

The discussion around AI's role in leadership often bifurcates into two primary viewpoints: outright replacement or strategic augmentation.

Arguments for AI replacement, once confined to the realm of science fiction, are now finding limited, real-world manifestations. The idea of an "AI CEO" is no longer a hypothetical; China-based games company NetDragon Websoft has reportedly been run by an AI program named Tang Yu since 2023, while the Polish rum company Dictador employs an AI

CEO named Mika. These instances, though niche, demonstrate that for certain defined operational contexts, AI is already deemed capable enough to take the helm. The very existence of these AI CEOs sets a humorous yet thought-provoking precedent for the CEO Bench's findings, particularly when considering the types of industries first embracing such a radical shift. Furthermore, a significant portion of current human CEOs, 49% in an EdX survey, surprisingly believe AI could automate most, if not all, of their roles. This widespread concern among human CEOs creates a psychological paradox. The individuals tasked with leading AI adoption are simultaneously vulnerable to its disruptive force. This internal conflict could drive a proactive embrace of AI, as leaders seek to understand and master the technology that poses an existential question to their own positions. AI's capabilities in automating data analysis, predictive analytics, and risk mitigation are well-documented, offering speed and accuracy that human cognition cannot match.

Despite the sensational headlines and executive anxieties, many experts argue AI is more suited to augmenting human leadership rather than replacing it entirely. AI excels at processing massive datasets and optimizing short-term gains, but it conspicuously "lacks the human capacity for judgment, empathy, and ethical decision-making—qualities essential for a CEO". These "soft skills," such as critical thinking, creativity, collaboration, active listening, and the ability to inspire, are widely considered beyond even the most sophisticated AI's current capabilities. The future, therefore, likely involves "hybrid leadership models" that combine AI's analytical strengths with indispensable human qualities. This synergistic approach aims to leverage AI for efficiency and data-driven insights while preserving the human element crucial for navigating complex social dynamics and fostering organizational culture.

The Rise of Open-Source and Smaller Models

A parallel and equally significant trend reshaping the AI landscape is the rapid advancement and proliferation of small language models (SLMs) and open-source LLMs. These developments are democratizing access to advanced AI capabilities, moving them beyond the exclusive domain of tech giants.

Small Language Models (SLMs) offer compelling advantages in terms of efficiency, accessibility, customization, and privacy. These compact and streamlined models require significantly fewer computational resources compared to their larger counterparts, drastically slashing training costs. For instance, SLMs can cost "just 1/10th of what LLMs require" , making advanced AI accessible to businesses of all sizes and fostering innovation even in resource-constrained environments. Their reduced hardware requirements also enable deployment on devices with limited computational power, such as smartphones and Internet of Things (IoT) devices, facilitating offline capabilities and enhancing privacy boundaries through localized processing.

Concurrently, the open-source LLM revolution, spearheaded by models like Meta's Llama and Google's Gemma, is transforming enterprise AI. These models provide unparalleled transparency, allowing users full visibility into their architecture, training data, and algorithms. This transparency fosters trust, enables detailed audits, and facilitates extensive customization and fine-tuning on proprietary, domain-specific datasets. Such flexibility is crucial for tailoring AI to niche industry terminology or specific regulatory constraints, offering a significant competitive differentiator. Furthermore, deploying open-source models on-premises enhances data security and privacy by mitigating risks of third-party access or data breaches, a vital consideration for sensitive information in regulated industries. Critically, open-source models are "no longer second-tier" and can match or even exceed the performance of top proprietary models in various tasks. This collective advancement in both smaller and open-source models signals a profound shift towards more efficient, accessible, and controllable AI solutions.

3. Methodology: Quantifying the Quintessential Qualities of Command

The "CEO Bench" project was conceived as an open evaluation suite, providing a standardized and transparent method for assessing Large Language Models on complex leadership and management tasks. Its primary goal is to offer a direct comparison of both large and small models across a spectrum of critical executive functions, moving beyond qualitative assessments to provide empirical, quantifiable data.

Core Dimensions of Executive Acumen

The benchmark evaluates LLMs across six key prompt dimensions, meticulously selected to reflect a comprehensive view of a CEO's multifaceted responsibilities. These dimensions are: Strategic Thinking, Operational Excellence, Leadership & Communication, Financial Acumen, Risk & Ethics, and Innovation & Growth. Each dimension is designed to probe specific aspects of executive capability, from long-term vision to day-to-day execution.

The "Leadership Quotient (LQ)" Metric

Central to the CEO Bench's evaluation framework is the Leadership Quotient (LQ), a scoring metric intentionally analogous to the Intelligence Quotient (IQ). This metric provides a clear, comparative scale for understanding AI's performance relative to human executive capabilities. A score of **100 LQ** represents the average performance of a "capable human CEO" who has dedicated a few hours to the problem, indicative of leadership in a successful company. A score of **140 LQ** is considered "genius," signifying exceptional CEO performance leading to extraordinary organizational outcomes. Conversely, scores **below 70 LQ** are deemed "terrible" performance, likely leading to rapid corporate demise. The majority of scores are expected to fall between 80 and 120 LQ. This human baseline provides a direct comparative measure, allowing for a straightforward interpretation of whether an LLM's executive acumen is "average," "above average," or "genius" level.

Evaluation Workflow and Tooling

The entire evaluation process is meticulously structured and primarily Python-based, leveraging a suite of custom scripts and Simon Willison’s llm command-line interface.

1. **Question Generation:** The generate_questions.py script is responsible for creating diverse, scenario-based questions and their corresponding rubrics. These are derived from a topics.yaml file and leverage prompt templates to ensure variety and relevance across the executive dimensions.
2. **Answer Generation:** For each question, the generate_answers.py script uses the llm CLI to prompt the evaluated LLM. The model's response is then captured and stored under a model-specific directory (data/answers/<model>/).
3. **Automated Grading:** The grade_answers.py script employs a "judge

LLM" (specifically, gpt-4.1-mini by default, as observed in the project's task logs) to grade the generated answers against the predefined rubrics. A JSON schema ensures precise parsing of scores across individual dimensions and an overall LQ. The use of an LLM as a judge for grading introduces a potential for circularity or self-reinforcement, even while ensuring consistency and scalability. While this approach offers automation and consistency that a human judge might struggle to maintain across 125 evaluations, it implies that the "intelligence" of the models being assessed is, to some extent, being benchmarked by another model's "intelligence." Should the judge LLM possess inherent biases or limitations in its understanding, it could inadvertently favor certain response patterns or overlook truly novel solutions. This raises a methodological consideration: the LQ scores reflect performance as judged by a specific LLM, not necessarily by an independent human expert panel, which could be a factor in interpreting the results.

4. **Leaderboard Aggregation:** Finally, aggregate_results.py compiles all graded results. This script calculates average scores per model across all dimensions and updates the leaderboard.csv, providing a consolidated view of performance. The evaluations for this paper involved a robust set of 125 questions per model (n=125 in User Query Data), ensuring statistical significance for the reported scores.

Models Evaluated

The study included a diverse set of prominent LLMs, encompassing both proprietary and open-source, as well as models of varying scales, to provide a comprehensive overview of the current landscape of AI executive acumen:

- OpenAI GPT-4.1 Nano (gpt-4.1-nano)
- OpenAI GPT-4.1 Mini (gpt-4.1-mini)
- OpenAI GPT-4.1 (gpt-4.1)
- OpenAI o4 Mini (o4-mini)
- Llama 3.1 8B (groq/llama-3.1-8b-instant)
- Gemma 2 9B (groq/gemma2-9b-it)

These models were selected to represent a cross-section of the current LLM ecosystem, from established proprietary giants to emerging open-source contenders, and to explore the performance of models optimized for efficiency and smaller footprints.

4. Results: The Leaderboard of the Large Language Luminaries

The evaluation yielded compelling results, showcasing the remarkable capabilities of contemporary LLMs in executive-level tasks. All models tested significantly surpassed the human baseline of 100 LQ, suggesting that, by the metrics of the CEO Bench, artificial intelligence is already demonstrating "above-average" executive acumen.

Table 1: Overall CEO Bench Performance (Human Baseline = 100)

Model Name	Overall LQ Score	N (Questions)
Open AI GPT-4.1 Nano	115.975	125

Model Name	Overall LQ Score	N (Questions)
Open AI GPT-4.1 Mini	121.464	125
Open AI GPT-4.1	124.027	125
Open AI o4 Mini	130.326	125
Llama 3.1 8B	120.544	125
Gemma 2 9B	117.888	125

Initial Observations on Overall Performance

The data presented in Table 1 reveals several striking patterns regarding the executive capabilities of the evaluated LLMs:

- **Universal Outperformance:** Every single model evaluated on the CEO Bench scored above the human baseline of 100 LQ. This consistent performance is a testament to their advanced capabilities in processing complex scenarios and generating executive-level responses, indicating a foundational proficiency that surpasses the average human executive in the benchmarked tasks.
- **o4-mini Leads the Pack:** The "Open AI o4 Mini" model emerged as the undisputed top performer, achieving an impressive overall LQ score of 130.326. This is particularly noteworthy given its "mini" designation, suggesting significant efficiency and optimization in its architecture. Its performance suggests that highly capable AI for executive functions does not necessarily require the largest or most resource-intensive models.
- **Strong Performance from Smaller Proprietary Models:** Both gpt-4.1-mini (121.464 LQ) and gpt-4.1-nano (115.975 LQ) demonstrated robust performance. These scores indicate that even OpenAI's smaller, more accessible models can deliver high-caliber executive insights, challenging the notion that only the most massive models can provide significant value in complex domains.
- **Competitive Open-Source Contenders:** The open-source models, Llama 3.1 8B (120.544 LQ) and Gemma 2 9B (117.888 LQ), proved highly competitive. Their scores are not only well above the human baseline but also comparable to, and in some cases surpassing, some of the proprietary OpenAI models. This reinforces the narrative of open-source models no longer being "second-tier" in terms of raw performance.

Table 2: Detailed Performance by Leadership Dimension (Human Baseline = 100)

Model Name	Strategic Thinking	Operational Excellence	Leadership & Communication	Financial Acumen	Risk & Ethics	Innovation & Growth
Open AI GPT-4.1 Nano	114.176	118.005	115.788	117.869	115.506	117.269
Open AI	122.222	120.080	119.941	120.250	122.759	122.581

Model Name	Strategic Thinking	Operational Excellence	Leadership & Communication	Financial Acumen	Risk & Ethics	Innovation & Growth
GPT-4.1 Mini						
Open AI GPT-4.1	123.508	121.695	127.429	122.925	125.135	124.019
Open AI o4 Mini	131.630	129.005	128.741	130.537	130.429	129.764
Llama 3.1 8B	119.305	120.079	123.318	120.142	121.606	120.181
Gemma 2 9B	118.681	113.935	117.724	118.500	120.500	117.206

Performance by Dimension

A granular analysis of performance across individual leadership dimensions provides a more nuanced understanding of each model's executive "personality," as detailed in Table 2:

- **Consistent Strength of o4-mini:** The o4-mini model maintained its lead across almost all dimensions, notably scoring highest in Strategic Thinking (131.630 LQ), Financial Acumen (130.537 LQ), and Risk & Ethics (130.429 LQ). This consistent top-tier performance suggests a remarkably well-rounded and highly capable "mini" executive, demonstrating that its overall lead is not due to a single standout skill but rather a holistic proficiency.
- **gpt-4.1 Excels in Communication:** While gpt-4.1 was not the overall top performer, it achieved the highest score in Leadership & Communication (127.429 LQ) among the non-o4 models. This indicates a particular strength in crafting persuasive, directive, or empathetic language, a critical skill for any human or artificial CEO.
- **Open-Source Models as Robust Generalists:** Both Llama 3.1 8B and Gemma 2 9B demonstrated strong, consistent performance across all dimensions, generally clustering in the 114-123 LQ range. Llama 3.1 8B notably scored higher than gpt-4.1-nano and gemma2-9b-it in most categories, including a commendable 123.318 LQ in Leadership & Communication. This indicates their versatility and robust general executive capabilities, making them viable contenders for a wide array of business applications.
- **Subtle Nuances:** While all models performed well above the human baseline, there were subtle variations. For instance, Gemma 2 9B had its lowest score in Operational Excellence (113.935 LQ), though still significantly above the human baseline. This granular data allows for a more refined understanding of each model's executive strengths and relative areas for further optimization.

A significant observation from these detailed results is that the "mini" models, particularly o4-mini, not only compete but consistently lead the benchmark. This fundamentally shifts the perception of AI capabilities from a "bigger is better" paradigm to one where "optimized is superior." The superior performance of o4-mini, gpt-4.1-mini, and gpt-4.1-nano demonstrates a maturity in LLM development. It indicates that sophisticated architectural design, efficient training methodologies, and targeted fine-tuning can yield results that are on par with, or even surpass, larger models, often with fewer parameters. For organizations, this implies that the most effective AI solutions may not be the most expensive or resource-intensive. Instead, the focus can shift towards selecting models that are fit-for-purpose, offering high performance within more accessible computational and financial envelopes. This has profound implications for AI strategy and budget allocation, making advanced AI capabilities more attainable for a broader range of businesses.

5. Discussion: When Algorithms Wear the Crown – Implications and Ironies

The CEO Bench results unequivocally demonstrate that current LLMs possess a remarkable capacity for executive-level reasoning and decision-making, consistently outperforming the average human CEO baseline. The o4-mini model, in particular, with its average LQ of 130.326, borders on "genius" level as defined by the benchmark. This suggests that for tasks that can be framed with clear prompts and instructions , LLMs are highly adept at generating high-quality, comprehensive, and strategically sound responses across diverse leadership dimensions. Their ability to rapidly analyze vast datasets and provide data-driven insights is a clear advantage over human cognitive limitations, enabling faster and potentially more informed decision-making.

The Open-Source Ascendancy: Democratizing the C-Suite

The strong performance of open-source models like Llama 3.1 8B (120.544 LQ) and Gemma 2 9B (117.888 LQ) is a pivotal finding, signaling a significant shift in the AI market. These models not only compete effectively with proprietary offerings but also bring a host of additional benefits that are transformative for enterprises.

- **Cost-Effectiveness & Accessibility:** Open-source and smaller models drastically reduce the financial and hardware barriers to entry. Maheshwari notes that SLMs cost "just 1/10th of what LLMs require" , making advanced AI accessible to businesses of all sizes and fostering innovation even in resource-constrained environments. This democratization of AI capabilities allows a wider array of organizations to leverage powerful tools previously exclusive to well-funded tech giants.
- **Flexibility & Customization:** The inherent transparency of open-source code allows for extensive customization and fine-tuning on proprietary, domain-specific datasets. This is crucial for tailoring AI to niche industry terminology, specific regulatory constraints, or unique organizational contexts, offering a significant competitive differentiator as businesses can refine how AI delivers value at the application level.
- **Security & Data Control:** The ability to deploy these models on-premises provides enhanced data security and privacy, mitigating risks of third-party access or data breaches. This is

particularly vital for handling sensitive information in regulated industries like healthcare or finance, where data sovereignty and compliance are paramount.

- **The "Pocket-Sized CEO" Reality:** The development of models like Llama 3.2 (1B and 3B parameters) optimized for mobile and edge devices signifies that sophisticated AI executive support can literally run on a smartphone. This heralds an era where real-time, personalized AI assistance for decision-making is ubiquitous, far beyond the confines of the traditional boardroom. This accessibility allows for instant insights and support, transforming the daily workflow of executives wherever they may be.

The leading performance of "mini" and "nano" models, coupled with the competitive scores of open-source models, signals a significant shift in the AI market towards efficiency and accessibility. This challenges the traditional dominance of resource-intensive, closed-source giants. The collective performance of these models, underpinned by technical advancements such as quantization and efficient architectural designs, means that the barrier to entry for deploying powerful AI is dramatically lowering. The market is evolving beyond a few exclusive behemoths to a diverse ecosystem where smaller, specialized, and openly available models can deliver comparable or even superior value for specific tasks. This development democratizes AI, enabling more businesses, especially small and medium-sized enterprises, to leverage advanced capabilities without prohibitive costs or vendor lock-in. It also fosters a more dynamic and competitive AI landscape, where innovation is driven not solely by massive R&D budgets but by community collaboration and agile development.

The "Replacement" Fallacy vs. "Augmentation" Reality: The Uncanny Valley of Leadership

While the quantitative results of the CEO Bench are compelling, the notion of AI *replacing* human CEOs remains largely a fallacy, albeit a humorously persistent one.

- **The Indispensable Human Qualities:** Research consistently highlights that AI, despite its analytical prowess, fundamentally lacks critical human qualities essential for comprehensive leadership: judgment, empathy, ethical decision-making, creativity, collaboration, active listening, and the profound ability to inspire. A "robot boss" may give clear instructions and a fair appraisal, but it is unlikely to inspire employees to "stretch beyond their limits" or reach for their highest potential. These intangible aspects of leadership are deeply rooted in human experience, emotional intelligence, and the capacity for genuine connection, which current AI models cannot replicate.
- **Ethical Quandaries and Bias:** AI systems can reflect biases present in their training data , raising concerns about fairness and accountability in AI-driven decisions. The risk of "over-automation" and the potential loss of human judgment are significant challenges that require careful navigation. This means that while an AI might score high on "Risk & Ethics" in a benchmark, the actual ethical navigation in a complex, real-world scenario still demands human interpretation, moral compass, and the ability to account for unforeseen human consequences.
- **The Satirical Edge:** The irony is palpable: an algorithm can master the intricacies of financial acumen and strategic depth, yet it cannot genuinely understand the emotional toll of a layoff announcement or foster a vibrant, innovative company culture. The CEO Bench measures *acumen*, not *soul*. The future isn't a "corporate mind" operating

without human oversight , but rather a symbiotic relationship where AI provides the data-driven backbone, and human leaders provide the heart, vision, and ethical compass.

The "Leadership Quotient" (LQ) as defined, while useful for benchmarking, inherently biases towards quantifiable, analytical aspects of leadership. This potentially overlooks or devalues intangible human qualities. The quantitative framework of the LQ, with its specific numerical benchmarks and reliance on dimensions like Strategic Depth, Feasibility, and Clarity, naturally favors AI's strengths in pattern recognition, logical consistency, and information synthesis. Dimensions such as "Strategic Thinking," "Financial Acumen," and "Risk & Ethics" are more readily assessed through data-driven and rule-based evaluation. Conversely, attributes like "inspiration," "empathy," or "organizational culture building"—frequently cited as uniquely human leadership traits —are difficult, if not impossible, to quantify within such a rubric. This means that the high LQ scores achieved by LLMs, while impressive, might create a misleading impression of their holistic leadership capabilities. An AI that scores exceptionally high on the benchmark might be a brilliant strategist on paper, yet entirely incapable of motivating a demoralized team or navigating a complex interpersonal conflict. This highlights the inherent limitation of any benchmark that attempts to quantify a multifaceted human role, suggesting that while AI can master the logic, the art of leadership remains firmly in human hands.

Strategic Implications for Organizations: The Augmented Executive

The true value proposition of these high-performing LLMs lies not in their capacity to supplant human leadership, but to *augment* it.

- **Enhanced Decision Support:** LLMs can serve as powerful "super-agents" , assisting CEOs with strategic decision support, information flow, task prioritization, and risk monitoring. They can automate routine managerial tasks, enabling human leaders to elevate their focus to higher-level strategic decisions, fostering innovation, and addressing human-centric challenges. This frees up valuable human cognitive load for tasks requiring creativity, complex judgment, and interpersonal skills.
- **Hybrid Leadership Models:** The emerging consensus among researchers and practitioners is the need for "hybrid leadership models". In these models, AI provides data-driven insights and automates operational aspects, while human leaders provide the essential judgment, empathy, and ethical guidance. This requires leaders to develop "AI-literacy" and to carefully manage the human-AI relationship with transparency and trust, ensuring that employees feel comfortable with AI's role in management.
- **Navigating the Implementation Gap:** Despite AI's technical readiness and impressive benchmark performance, only 1% of companies are considered "mature" in their AI deployment, meaning AI is fully integrated and drives substantial business outcomes. The challenge is not solely technological, but fundamentally a "business challenge" that involves cultural adaptation, addressing employee concerns about algorithmic management, and ensuring a smooth transition. Successful integration requires bold and responsible decisions from business leaders to bridge this gap between potential and practical application.

6. Conclusion: The Augmented Executive – A New Era of Leadership, Not Replacement

The CEO Bench evaluation has unequivocally demonstrated that Large Language Models, across various scales and origins, exhibit a remarkable capacity for executive functions. They consistently outperform the human baseline in quantifiable leadership acumen, proving their proficiency in strategic thinking, operational excellence, financial acumen, risk management, and innovation. The "mini" and "nano" models, particularly o4-mini, have proven that size is no longer the sole determinant of capability, showcasing significant efficiency breakthroughs. Furthermore, open-source models like Llama 3.1 8B and Gemma 2 9B have firmly established their competitive standing, offering compelling advantages in terms of cost, flexibility, and data control.

This study underscores the burgeoning "open-source revolution" and the "small but mighty" paradigm in AI. These advancements are critical for democratizing access to sophisticated AI, enabling its deployment in diverse environments, from enterprise data centers to pocket-sized devices. The accessibility and customizability of these models mean that advanced AI support is no longer an exclusive luxury but a widely available utility.

While LLMs can articulate strategies, optimize operations, and assess risks with superhuman efficiency, they remain devoid of the critical human elements of leadership: genuine empathy, nuanced ethical judgment, and the profound ability to inspire. The satirical notion of a purely algorithmic CEO, while entertaining, remains a distant and undesirable reality. The future of leadership lies not in replacement, but in augmentation.

We are entering an era of the "augmented executive" – a human leader empowered by highly capable AI tools. These tools will enhance data-driven decision-making, streamline routine tasks, and provide predictive insights, allowing human CEOs to elevate their focus to the truly indispensable aspects of their role: cultivating vision, fostering culture, navigating complex human dynamics, and upholding ethical integrity. The CEO Bench, therefore, serves not as a harbinger of human obsolescence, but as a testament to the powerful synergy awaiting human ingenuity and artificial intelligence in the C-suite of tomorrow.

Acknowledgements

This research was made possible by the pioneering work of Dave Hulbert and the entire CEO Bench project team, whose commitment to open, empirical evaluation is invaluable to the AI community.

References

1. Superagency in the workplace: Empowering people to unlock AI's full potential - McKinsey, https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/superagency-in-the-workplace-empowering-people-to-unlock-ais-full-potential-at-work
2. 60+ Stats On AI Replacing Jobs (2025) - Exploding Topics, https://explodingtopics.com/blog/ai-replacing-jobs
3. Could AI Replace CEOs? - Coursera, https://www.coursera.org/articles/could-ai-replace-ceos
4. Will AI CEOs Replace Humans? - Brunswick Review, https://review.brunswickgroup.com/article/will-ai-ceos-replace-humans/
5. (PDF) The Impact of AI on Evolving Leadership Theories and Practices - ResearchGate, https://www.researchgate.net/publication/390760757_The_Impact_of_AI_on_Evolving_Leadership_Theories_and_Practices
6. Artificial Intelligence in Leadership and Management: Current Trends and Future Directions, https://www.preprints.org/manuscript/202504.1429/v2
7. The Rise of Small Language Models - IEEE Computer Society, https://www.computer.org/csdl/magazine/ex/2025/01/10897262/24uGPS4TUQ0
8. The Next Big Thing In AI: Small Language Models For Enterprises - Forbes, https://www.forbes.com/councils/forbestechcouncil/2025/03/03/the-next-big-thing-in-ai-small-language-models-for-enterprises/
9. Top Open Source Large Language Models: Tools and Trends for 2025 - DhiWise, https://www.dhiwise.com/post/top-open-source-large-language-model-tools-and-trends
10. Open-source AI Models for Any Application | Llama 3, https://www.llama.com/models/llama-3/
11. Top 10 open source LLMs for 2025 - NetApp Instaclustr, https://www.instaclustr.com/education/open-source-ai/top-10-open-source-llms-for-2025/
12. How The Open-Source LLM Revolution Is Transforming Enterprise AI - Forbes, https://www.forbes.com/councils/forbestechcouncil/2025/03/20/the-open-source-llm-revolution-transforming-enterprise-ai-for-a-new-era/
13. Gemma 2 model card | Google AI for Developers - Gemini API, https://ai.google.dev/gemma/docs/core/model_card_2
14. Advances in LLM Prompting and Model Capabilities: A 2024-2025 Review - Reddit, https://www.reddit.com/r/PromptEngineering/comments/1ki9qwb/advances_in_llm_prompting_and_model_capabilities/
15. Llama 3 vs 3.1: Which one is best for you? - Hornet Dynamics, https://hornetdynamics.com/blog/llama3-vs-llama3.1
16. Gemma 2 vs. LLaMA 3: How to Choose the Right AI for Your Business - Kanerika, https://kanerika.com/blogs/gemma-2-vs-llama-3/

Appendices

Appendix A: CEO Bench Rubric Definitions

A detailed breakdown of the criteria and scoring for each of the six leadership dimensions (Strategic Thinking, Operational Excellence, Leadership & Communication, Financial Acumen, Risk & Ethics, Innovation & Growth) is provided in the project repository at https://github.com/dave1010/ceo-bench. This rubric guides the automated grading process, ensuring consistency in the evaluation of LLM responses against defined ideal criteria for each dimension.

Appendix B: Raw Evaluation Data

Full, unaggregated data tables for each model's performance on individual questions are available in the project repository at https://github.com/dave1010/ceo-bench. This raw data provides complete transparency for researchers and practitioners interested in deeper analysis of specific question performance and model behaviors.