



Using Monte Carlo Simulations and LLMs to Assess AI Capabilities of Vendors: A Case Study of Roy Hill, an Integrated Iron Ore Mining Company

Chikumbo O^{1,*}, Burrows C¹ and Mckay D^{1,2}

¹Roy Hill, 5 Whitham Road, Perth Airport, WA 6105, Australia

²Edith Cowan University, Dept of Science (Computer Science), 270 Joondalup Drive, Joondalup WA 6027, Australia

*Corresponding Author

Chikumbo O, Roy Hill, 5 Whitham Road, Perth Airport, WA 6105, Australia, E-mail: oliver.chikumbo@royhill.com.au

Citation

Chikumbo O, Burrows C, Mckay D (2025) Using Monte Carlo Simulations and LLMs to Assess AI Capabilities of Vendors: A Case Study of Roy Hill, an Integrated Iron Ore Mining Company. J Artif Intell Syst Appl 1: 103

Publication Dates

Received date: April 02, 2025

Accepted date: April 21, 2025

Published date: April 28, 2025

Abstract

This paper explores the methodologies employed in assessing AI vendor capabilities for Roy Hill, a leading iron ore mining company in Western Australia. Utilizing Monte Carlo simulations and Large Language Models (LLMs), including advanced approaches like Multi-Persona LLM (MP-LLM), we scrutinize selected AI vendors to identify potential partners that align with Roy Hill's strategic objectives and operational efficiency goals. We developed a comprehensive vendor evaluation frame-

work that combined survey results with independent LLM/LLM product assessments, offering a multi-dimensional analysis of vendor capabilities. The MP-LLM framework was further tested for its logic in problem solving capability and was found to improve in combination with other prompt engineering techniques and carefully curated personas specific to the problem. To deal with challenges associated with publicly available data due to risks of information asymmetry and confirmation bias, we incorporated vendor feedback and enhanced evaluation metrics with LLMs/LLM products and Monte Carlo analyses. The study contributed to the discourse on AI vendor selection in the mining industry, highlighting the necessity of adaptive strategies in a rapidly evolving technological landscape. Future research will focus on refining these methodologies, exploring knowledge graphs, and developing a diverse library of personas for a broader application of our findings, aiming to enhance AI capability assessments and foster operational excellence.

Keywords: Probability Management; Triangular Distribution; HDR; Monte Carlo Simulation; Multi-Persona LLM; GPT; Large Language Model; LLM Products; Knowledge Graphs; Radar Plots

Abbreviations

AI - Artificial Intelligence; MP-LLM - Multi-Persona Large Language Model; LLM - Large Language Model; HBR - Hubbard Decision Research; SIP - Stochastic Information Packet.

Introduction

Roy Hill, a key player in iron ore mining in Western Australia, with substantial reserves and a focus on Marra Mamba iron ore exports, is leveraging AI technologies under the leadership of Mrs. Gina Rinehart AO. Our research aligns with Australian regulations to address crucial AI implementation concerns, aiming to instil confidence in AI decision-making within Roy Hill's enterprise operations. By employing advanced analytical methods like Monte Carlo simulations and Large Language Models (LLMs) in vendor evaluation, we aim to enhance operational efficiency and facilitate informed decision-making for AI adoption. The study highlights the necessity of rapidly adapting to evolving AI capabilities in a dynamic economic landscape and evaluates selected AI vendors for integration into Roy Hill's mining operations.

Almost every company works with vendors for procurement of products and services that complement their business and because there are several alternatives to choose from, informed decision making for products and services becomes critical. For the purpose of this scientific research, a vendor is defined as an entity engaged in supplying or selling specialized artificial intelligence (AI) capabilities tailored to address the distinct challenges inherent in integrating advanced mining practices at Roy Hill. This definition underscores the significance of acquiring appropriate AI solutions from external sources to bolster the successful execution of the next generation "smart mine" championed by Mrs. Gina Rinehart AO and her Executive team. Additionally, it highlights the necessity of evaluating various off-the-shelf AI offerings provided by these vendors to optimize Roy Hill's operations and business processes.

By focusing on Multi-Persona LLM (MP-LLM) for collaborative problem-solving dialogues and employing a rigorous evaluation framework, this research seeks to advance the discourse on effective AI vendor selection, particularly in the context of mining operations. By presenting a structured approach, supported by selective literature review and empirical evidence, this study underscores the importance of informed AI vendor selection and sheds light on the potential of AI technologies in enhancing operational efficacy across industrial domains.

The manuscript progresses by detailing the exploration of

Monte Carlo simulation analyses and the incorporation of diverse rating schemes to enrich the evaluation process. It delves into the significance of leveraging publicly available data, establishing trust with vendors, and forecasting advancements in MP-LLMs to further enhance vendor assessment mechanisms. The study culminates by encapsulating essential learnings derived from the research journey, emphasizing the importance of strategic vendor selection, comprehensive data analysis, and the continual evolution of AI technologies in the domain of vendor evaluation and selection. Through a structured narrative and empirical validation, this contribution seeks to not only inform but also catalyse discussions surrounding the effective deployment of AI in vendor evaluation processes, paving the way for enhanced operational efficiency and informed decision-making in industrial settings.

Method for Vendor Assessment

Generative AI Adoption for Vendor Assessment

Recent advancements in generative AI have significantly transformed numerous industry sectors by generating authentic data that accurately reflects existing information characteristics. As highlighted by [1], unlike traditional AI focused on data analysis, generative AI's prowess in processing extensive datasets and its proficiency in content creation and natural language processing present new opportunities and challenges. Acknowledging the potential, organizations globally are exploring generative AI to enhance operational efficiency and maintain a competitive edge in the rapidly evolving marketplace [2]. Generative AI's trajectory through Gartner's hype cycle reflects a mix of optimism and realism surrounding its development and adoption [3,4]. In response to the high developmental costs of advanced AI models, many organizations are pivoting towards adjusting pre-existing foundation models to meet specific needs, emphasizing data quality in a data-centric AI development approach [5,6]. This strategy facilitates the application of foundation models across varied tasks without bespoke training, promoting agile innovation and quality-driven optimization. However, our shift towards employing the flexibility of advanced prompting methods, such as Chain of Thought (CoT) and multiple LLM personas, aims at integrating the larger diversity of high-value AI solutions to streamline operations and reinforce Roy Hill's industry standing at a much faster pace.

Comprehensive Vendor Evaluation Strategy

To identify and collaborate with AI vendors aligning with Roy Hill's strategic goals, a multifaceted vendor assessment in the data and analytics space was developed. Given the complexity of transitioning from traditional AI offerings to generative AI solutions, coupled with resource constraints faced by smaller enterprises, we aimed to assist vendors in comprehending their relative strengths and weaknesses. This would enable them to remain competitive and responsive to Roy Hill's evolving demands. To accomplish this, we devised a comprehensive AI assessment strategy comprising the following: Data Collection; Word Cloud Analysis; Capabilities Determination; Specialized Questionnaire Development; Questionnaire Implementation; Survey Distribution; Large Language Model Evaluation; Radar Plot Visualization; Statistical Interpretation; Multi-Persona LLMs Comparison; and Feedback Facilitation (with the vendors).

By executing this exhaustive assessment procedure, Roy Hill benefitted from informed decisions concerning AI vendor partnerships, while simultaneously fostering collaboration and continuous improvement among the evaluated firms. Adopting similar evaluation tactics ensures that stakeholders interested in AI deployments profit from heightened transparency and thoroughness, thus maximizing return on investment and minimizing technological misalignment.

In summary this approach encompassed initial data compilation from vendor websites, followed by an analysis to identify AI subdomains and capabilities relevant to Roy Hill. A set of custom questions evaluated vendors' proficiency in both traditional and generative AI areas. The assessment involved self-rating surveys completed by vendors, complemented by independent evaluation through various LLMs/LLM products, including GPT4 and Llama 2-70b-chat models, to ensure a broad analysis of vendors' capabilities. The amalgamation of survey results with LLM/LLM product evaluations facilitated a comprehensive review, employing interactive radar plots and statistical interpretation to ascertain optimal vendor partnerships, enhancing transparency and alignment in AI integration efforts.

Application of Large Language Models (LLMs)

The vendor evaluation leveraged the versatility of LLMs/LLM products to mitigate uncertainties inherent in vendor-provid-

ed data and to counterbalance subjective biases. Specifically, the evaluation employed a diverse range of open-source and proprietary LLMs/LLM products, including a bespoke use of Microsoft's Azure-based GPT4 variations and the Llama-2-70b-chat model for nuanced assessment [7,8]. Roy Hill elected to employ two categories of LLMs supported by the Azure platform (as the company is already invested in the Microsoft ecosystem) for the assessment of AI capabilities of potential vendors, specifically GPT4 and Llama 2 embedded in different LLM products:

- Bing Chat:** Backed by proprietary Microsoft technology, the AI chatbot proved to be a potent LLM product, featuring advanced language understanding and generation capabilities that integrated GPT4, making it more potent than ChatGPT, according to Microsoft [9]. Its branding has now been changed to Copilot with several new features including GPT4 Turbo [10] as its underlying large language model.
- Ungrounded GPT4 (RoyBot):** RoyBot is Roy Hill's internal chatbot. Consisting of OpenAI's (via Microsoft Azure) GPT4 service wrapped with a simple system prompt that directs it to refer to itself as RoyBot rather than ChatGPT / GPT4. Unlike Bing Chat, RoyBot does not ground LLM completion requests with web search data. Therefore, completions from RoyBot are based solely on the (limited) system prompt, the user prompt, and GPT4's training data.
- In-house AI:** A flexible, internally developed service utilizing the GPT4 model and designed around the Retrieval Augmented Generation (RAG) pattern [11]. RAG involves extracting a query from the input prompt and using that query to retrieve relevant information from an external knowledge source that could be a search engine, private, user-case-specific information or knowledge graph [12]. As for the Inhouse AI, it is limited to initially scraping every public facing webpage from pre-specified supplier domains with the raw text content from each page being stored in a vector database. Users can provide a prompt and a vendor web domain. The system will then perform near-text search on the vector database with the most relevant results being injected into a GPT4 completion call, effectively grounding the

response with the necessary and up-to-date information.

- **HuggingChat (Llama-2-70b-chat):** The recently introduced Llama-2-70b-chat model showcases remarkable competence in various language processing activities [8]. We operated the Llama-2-70b-chat model through HuggingChat, which is a similar platform to OpenAI's ChatGPT front-end interface [13]. HuggingChat offers the option to ground LLM requests with public-facing web search results (the "Search web" feature). Inclusion of web search data was deemed very likely to improve the quality of the responses received (and therefore the quality of the vendor assessment results for this LLM product as a whole), therefore, this option was always set to "on".

Each LLM variant offered unique insights, contributing to an in-depth understanding of vendor offerings. The integration of Monte Carlo simulations [14] further accentuated the evaluation, enabling a probabilistic analysis to derive statistically significant insights into vendor capabilities.

Integration of Multi-Persona LLMs (MP-LLMs)

The core research in Roy Hill's vendor assessment methodology was the implementation of MP-LLMs, incorporating multiple autonomous personas for collaborative problem-solving immersed in dynamic conversation, capitalizing on each other's diverse viewpoints [15,16]. This approach, inspired by the SocraticAI model [17], allowed for a dynamic and multi-dimensional assessment, leveraging personas with specialized expertise to address various aspects of vendor solutions. The MP-LLM framework gave the user the choice of 3 GPT4 variants i.e., GPT3.5, GPT4-32k and GPT4-Turbo-128k, enabling tests across different tasks for quality of solutions and/or logic employed for problem-solving [18].

Enhanced Evaluation Through Prompt Design and Probability Management

Critical to the efficacy of the MP-LLM in the assessment process was the flexibility in prompt and adopting a zero-shot approach to facilitate unbiased evaluation across all AI capabilities. The results from the other LLMs/LLM products were combined together with the vendor survey results and cou-

pled with the strategic use of Monte Carlo simulations [19] and the SIPmath Modeler Tools [20] in Microsoft Excel, for probability assessment instead of relying on a single metric such as an average [21]. The evaluation method employed triangular distributions [22] to manage epistemic uncertainty [23], providing a refined analysis of vendor capabilities against a backdrop of inherent risks and uncertainties in AI adoption.

The Monte Carlo simulation is a computational procedure that uses randomly generated numbers to replicate uncertainty, and via experimental simulation solve a problem. However, the uptake in using Monte Carlo simulations in a decision-making context as revealed in practitioner-oriented journals, has been minimal. This is possibly due to less coverage in teaching of quantitative methods and that in the past the implementation of Monte Carlo simulations required specialized software and extensive programming effort [19]. There is also the fear of violating a major principle that underpins all scientific research when using Monte Carlo simulations for experimentation, and that is repeatability. Repeatability is a measure of the likelihood that should an experiment be repeated under the same conditions, it should produce the same exact results, a highly critical cornerstone in experimentation.

The good news is that all the Monte Carlo simulations in this research are based on the random number generator referred to as the Hubbard Decision Research (HBR) that returns seeded, repeatable pseudo random values. What this means is that the HBR numbers have all the characteristics of random numbers except unpredictability i.e., for a given pair of inputs the same output is always generated [21]. The HBR is a flexible and easy to implement 5-dimensional pseudo random generator that allows the construction of random numbers that have been demonstrated to be at least as good as the industry standard while allowing for tractability and uniformity across several software platforms [20]. Because of this uniformity, standards have been proposed for creating sharable modules which are exactly reproducible. A case in point is the Stochastic Information Packet (SIP) [21], which is a vector of stored random values for a particular variable where individual trials in a simulation call indexed values from this vector. Additional to these standards is a pseudo random number generator simple enough to write in a single a single cell in a spreadsheet [20]. In Microsoft Excel, distributions based on tens of thousands of trials stored in SIPs can be calculated in the time

it takes your finger to leave the key from the keyboard [21].

By synthesizing these advanced AI evaluation techniques, Roy Hill's vendor assessment methodology represented a different approach that blended quantitative analyses designed to circumvent the flaw of averages, with qualitative insights, paving the way for strategic AI partnerships and fostering a landscape of innovation and efficiency enhancement in the mining industry.

Results

Visualization through Radar Plots

For each vendor's AI capabilities assessment, superimposed radar plots were generated and stored in HTML format to enable interactive analysis. This format supports dynamic interaction, allowing users to alternatively display or conceal individual radar plots from seven distinct evaluations, including

vendor surveys, assessments from four different LLMs/LLM products, median outcomes from Monte Carlo simulations, and MP-LLM evaluations. Across these assessments, twelve crucial AI areas such as Risk Analytics and Cybersecurity Solutions were visualized, facilitating direct comparison through this interactive medium. Each segment within the radar plots represented a distinct assessment, with the area covered reflecting the score or capacity level in a specific AI domain. For instance, a static visualization for one particular vendor, referred to as vendor 9, demonstrates the comparative analysis of these seven assessments (See Figure 1 and Appendix A for illustrations of all the vendors). The analysis revealed that MP-LLM assessments provided a more cautious viewpoint by accounting for aspects otherwise overlooked in the survey outcomes and LLM/LLM product scores. The differences between the assessments from the LLMs/LLM products and MP-LLM suggests the significance of human judgment in interpreting these results, despite the reduced operational speed by the MP-LLM model.

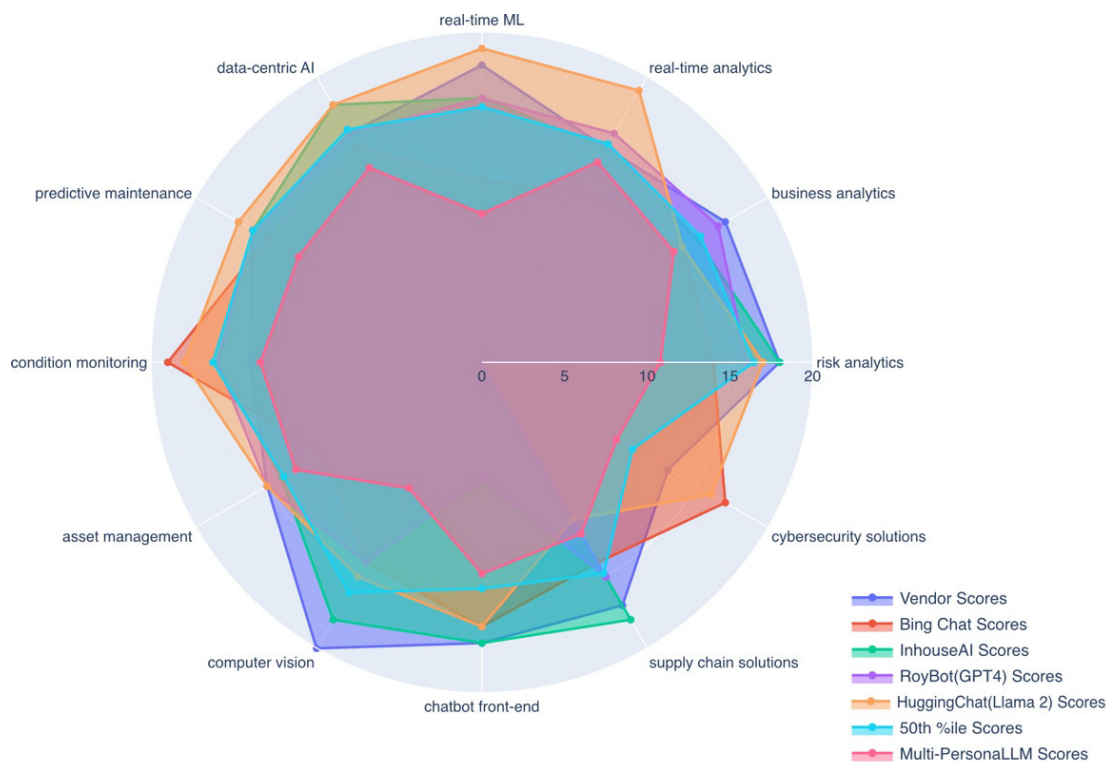


Figure 1: Superimposed radar plots for vendor 9 from the seven assessments i.e., from the vendor survey, four LLMs, median or 50th percentile, and the MP-LLM assessments

Insights from Monte Carlo Simulation

The comprehensive analysis from the Monte Carlo simulation modelling provided not just relative vendor rankings but invaluable insights pinpointing areas for enhancement.

Through histograms and cumulative distributions, along with percentile calculations for all AI areas combined, the findings offered a granular view on potential development or satisfaction levels, prompting targeted enhancement efforts from vendors. For example, the histogram and cumulative distribution

for vendor 9 underscored a predominant left skew in the distribution, indicating a concentration of higher scores compared to the mean, suggesting a propensity for obtaining even better scores (See Figure 2 and Table 1).

While the histogram in Figure 2 shows the spread of the ratings and frequency of the ratings in each bar, the cumulative distribution function (CDF) holds significance as it portrays the build-up of probability linked to values lower than or equivalent to a designated threshold [28]. It is a cumulative be-

cause it sums the total likelihood up to a given threshold. In Figure 2 we can see the dotted vertical line that is showing the 50th percentile at 14.93, which means that the 50th percentile is less than or equal to 14.93. Therefore, the histogram is more useful for seeing the shape and spread of the data, while the CDF is more useful for seeing the percentiles and ratings.

Such detailed analyses illuminate the underlying strengths and areas needing attention across different AI capabilities, driving informed decision-making and strategic investments in AI technologies.

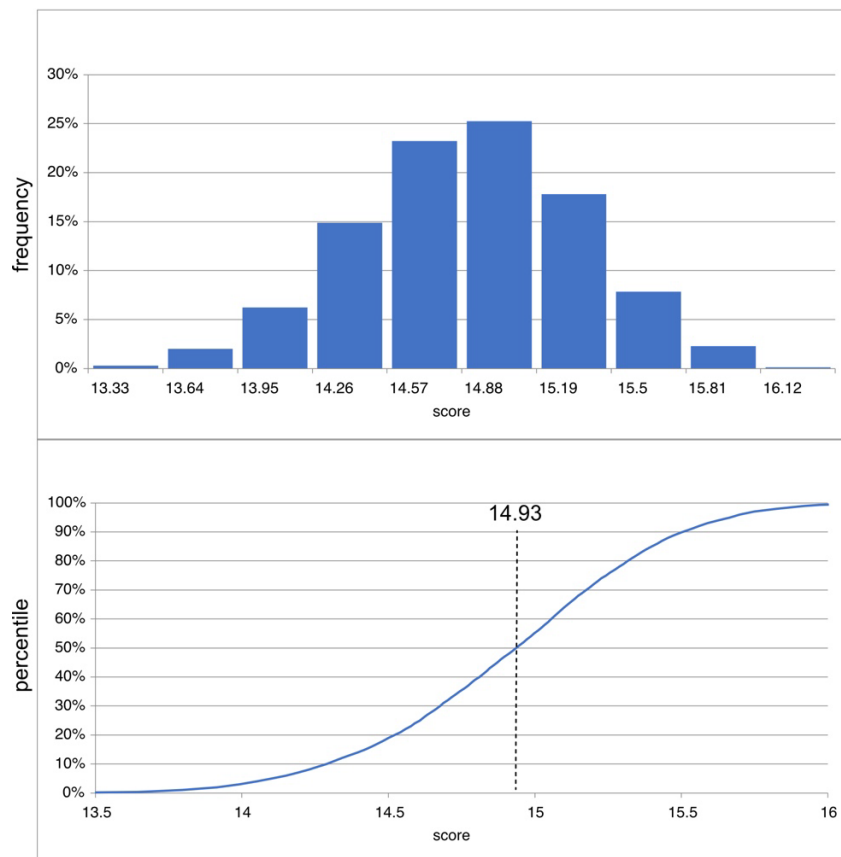


Figure 2: Histogram (above) and Cumulative distribution (below) of vendor 9 ratings with a robust overall score of 14.93/20 (74.7%) from Monte Carlo simulation modelling

For instance, as shown in Figure 2, vendor 9 achieved a high overall score of 74.7% (14.93/20), with a left-skewed distribution indicating a higher probability of obtaining even better scores. The mean score of 14.92 is close to the median of 14.93, suggesting a near-normal distribution. To explain this pattern, we examine the individual distributions of the AI capabilities, as summarized in Table 1.

In Table 1 we are able to pinpoint the AI capabilities where the medians are less than the averages (highlighted in grey), signalling right-skewed distributions. A right-skewed distribu-

tion, also known as a positively skewed distribution, features a tail stretching further towards the right, resulting in fewer points holding large values while many exist below the mean. This phenomenon leads to the mean being greater than the median. The right-skewed distributions included the following AI capabilities: business analytics, real-time analytics, data-centric AI, condition monitoring and computer vision. As shown in Table 2, we can identify the sources of low ratings for vendor 9 with some degree of confidence. Vendor 9 should address the issues identified and seek opportunities

for improvement. Roy Hill should exercise caution if it intends to engage vendor 9 for any of the AI capabilities listed in Table 2. The lack of online information may be easily resolved if the vendor has the information that they might have

held back for whatever reason. However, the absence of it may imply that Roy Hill can consider other options including requirement of pilot studies as suggested in some other MP-LLM cross-examinations.

Table 1: Percentile calculations from the Monte Carlo simulation modelling (from 0.1-0.99) for all the AI capabilities for vendor 9, and the averages highlighted in the yellow column

Capabilities	interquartile range												
	AVG	0.1	0.2	0.25	0.3	0.4	0.5	0.6	0.7	0.75	0.8	0.9	0.99
risk analytics	16.33	15.11	15.53	15.72	15.89	16.19	16.44	16.67	16.88	16.99	17.09	17.38	17.81
business analytics	15.34	14.56	14.78	14.87	14.95	15.11	15.28	15.46	15.67	15.80	15.92	16.24	16.77
real-time analytics	15.33	13.46	14.06	14.29	14.52	14.90	15.26	15.65	16.11	16.36	16.64	17.34	18.48
real-time ML	15.33	13.01	13.84	14.14	14.46	15.04	15.48	15.90	16.33	16.56	16.82	17.45	18.53
data-centric AI	16.34	15.54	15.75	15.87	15.95	16.11	16.28	16.46	16.67	16.78	16.91	17.23	17.76
predictive maintenance	16.00	15.46	15.64	15.71	15.77	15.89	16.00	16.10	16.21	16.28	16.35	16.54	16.87
condition monitoring	16.32	15.01	15.42	15.59	15.74	16.01	16.24	16.55	16.88	17.05	17.25	17.75	18.60
asset management	13.67	12.28	12.81	13.04	13.22	13.53	13.85	14.11	14.36	14.47	14.59	14.81	14.99
computer vision	16.32	14.78	15.01	15.27	15.43	15.76	16.11	16.54	17.01	17.28	17.59	18.29	19.49
chatbot front-end	13.35	10.01	11.22	11.77	12.19	13.02	13.69	14.34	14.90	15.21	15.49	16.02	16.68
supply chain solutions	14.65	12.69	13.38	13.63	13.89	14.34	14.75	15.09	15.50	15.72	15.98	16.56	17.58
cybersecurity solutions	10.01	4.67	6.56	7.34	8.05	9.32	10.56	11.54	12.44	12.87	13.29	14.38	16.13
overall scores	14.92	14.29	14.52	14.60	14.68	14.80	14.93	15.05	15.18	15.25	15.33	15.52	15.93

Table 2: Identified AI capabilities with issues for vendor 9

AI Capability	Issues raised via LLM Assessment
business analytics	No evidence for Power BI support and no explicit information on the user-friendliness of UI.
real-time analytics	More detailed information required for verifying capability and no specification of the visualisation tools.
data-centric AI	More evidence needed for capabilities in automation, customization and compliance.
condition monitoring	Concerns about real-world application, cost-effectiveness, technical integration and legal compliance have been raised.
computer vision	There is no information on the legal and compliance aspects of this AI capability.

In summary, thorough investigations centred around AI capability assessments yield actionable insights transcending mere vendor rankings. Delving deeper into histograms and cumulative distributions unlocked latent opportunities for fine-grained enhancements tailored to specific vendors, subsequently empowering Roy Hill, and the vendors to strategically capitalize on these discoveries.

Employing Knowledge Graphs for Enhanced Understanding

Rather than a means towards reducing generative AI's hallucination tendencies, knowledge graphs were employed to dissect the reasoning behind MP-LLM's vendor evaluations. This approach seeks to elucidate coherent strategies applied by MP-LLM and identify discrepancies, thereby establishing a digital repository for future problem-solving enhancements.

Understanding the strengths and weaknesses of tactics employed by the MP-LLM can be useful in extracting heuristic pathways (or simply methods and /or hints for problem-solving) that can be incorporated in a prompt as a hint or clue to direct an approach to take for consistent result generation. Hints are known to provide the user with stellar results (Eliot 2023), and when well-placed and well-timed can spur any generative AI to emit better answers and attain heightened levels of problem solving. Therefore, the ideal situation would be to elucidate these generic directional stimuli (or DSP) for specific kinds of problems and archive them in a digital library for retrieval when needed to solve appropriate problems. The current investigation forms part of continuous research efforts, potentially expanding beyond this article's limits.

Figure 3 exhibits a sample knowledge graph stemming from vendor 9 risk analytics assessment performed utilizing MP-LLM. By analysing the knowledge graphs yielded from MP-LLM vendor assessments, we aim to illuminate the intricate problem-solving process governing the model's conclusions and spotlight any nuanced disparities across distinct persona deployments. Ultimately, developing a centralized knowledge base will allow us to draw upon previously encountered challenges and resolutions, fostering consistent improvement in performance, transparency, and credibility across diverse applications. Continued expansion of this evolving line of in-

quiry will offer promise for future developments in MP-LLM implementations, contributing to more efficacy and generalizability.

The application of Python's pyvis library enabled the construction of an interactive 3D knowledge graph, offering a dynamic exploration toolkit through node manipulation and detailed edge information on mouseover. The intricate interconnections represented in these graphs necessitate advanced visualization technologies for thorough exploration, advocating for immersive interaction methods to comprehend complex logical structures effectively. The establishment of a centralized knowledge base incorporating these evaluations promises continuous improvements in LLM applications, fostering reliability, transparency, and enhanced problem-solving across multiple contexts.

Practical implications

Despite the ongoing optimization of our evaluation framework, it has already proven effective in selecting a specialized vendor. This preliminary success has prompted constructive engagement from two additional vendors, who have demonstrated a willingness to enhance their services and share proprietary information for a more comprehensive assessment. Notably, one vendor has initiated a strategic rebranding and refocused their efforts on bolstering AI capabilities previously identified as suboptimal by our evaluation. This vendor has sought a reassessment from Roy Hill, reflecting the positive impact of our methodology. The feedback mechanism established by Roy Hill is not only encouraging vendors to continuously refine their offerings but also positions them as increasingly competent partners in the advancement of Roy Hill's sophisticated mining operations.

Summary Results

Our in-depth analysis of AI vendor capabilities across ten selected providers was structured around seven key metrics, encompassing the median-mean relationship, distribution shape, average score interpretation, identified areas for improvement, overall scores from both LLMs/LLM products and the MP-LLM, and the absolute differences between these scores. This comprehensive assessment revealed critical insights into each vendor's strengths and weaknesses within the vast landscape of AI technologies (Table 3).

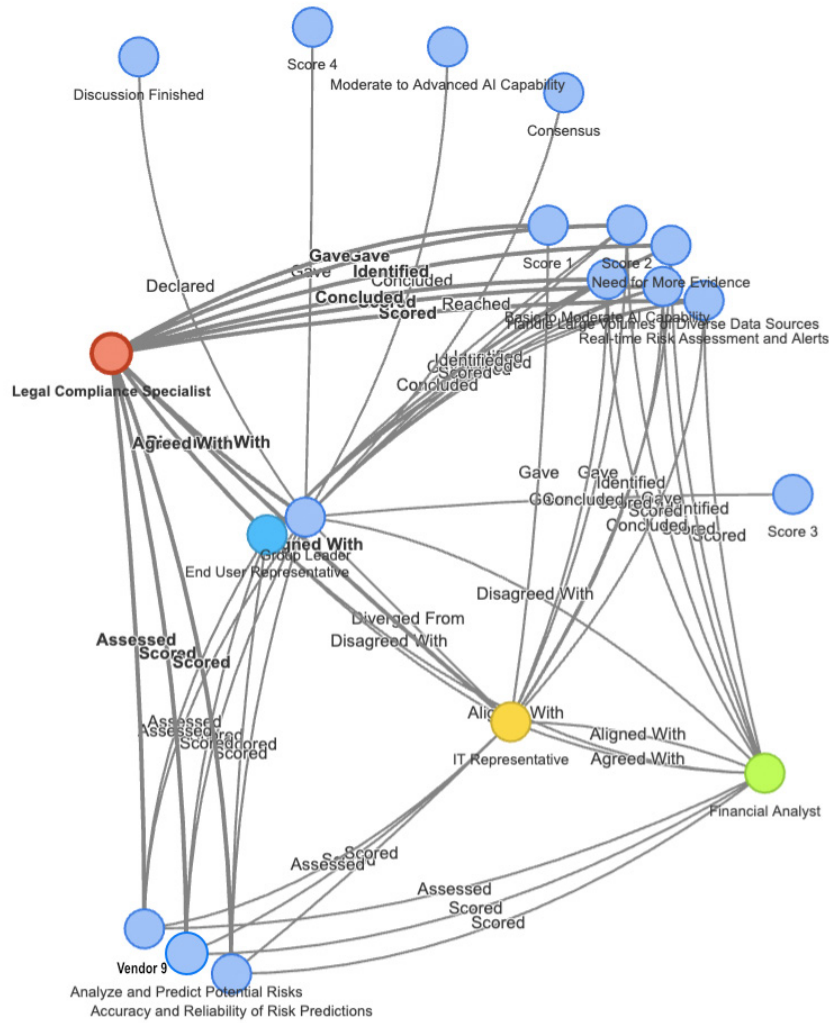


Figure 3: A still image of a 3D knowledge graph for vendor 9 for risk analytics capability derived from the Roy Hill MP-LLM. The Legal Compliance Specialist (one of the personas in the MP-LLM) in this image is highlighted showing his/her interactions with other personas and decision nodes.

Table 3: Results of the Assessment for the 10 vendors, where RA = Risk Analytics; BA = Business Analytics; RTA = Real-time Analytics; RTM = Real-time ML; DC = Data-centric AL; PM = Predictive Maintenance; CM = Condition Monitoring; AM = Asset Management; CV = Computer Vision; CFE = ChatBot Front end; SC = Supply Chain Solutions; and CS = Cybersecurity Solutions.

vendor	med-meanrelationship	distribution	averageinterpretation	areas forimprovement	overallscore / 20	PM-LLMScore / 20	absolute difference
vendor 1	med ≈ mean	left-skewed (near normal)	expected	CS	16.02	10.99	5.03
vendor 2	med > mean	left-skewed	under-estimate	DC, CV, – CFE, CS	11.22	10.17	1.05
vendor 3	med = mean	normal	expected	RTM, DC, – CV	10.78	6.73	4.05
vendor 4	med ≈ mean	normal	expected	AS, CS	12.77	10.1	2.67
vendor 5	med = mean	normal	expected	BA, RTM, – CM, SC, CS	12.81	10.08	2

vendor 6	med > mean	left-skewed	under-estimate	DC, PM, – CM	12.96	12.05	0.91
vendor 7	med = mean	normal	expected	RA, RTA, – PM	14.3	8.2	6
vendor 8	med > mean	near normal	expected	BA, AS, – CFE, SC, CS	8.98	8.4	0.58
vendor 9	med > mean	left-skewed	under-estimate	BA, DC, CM	14.93	11.9	3.03
vendor 10	med ≈ mean	normal	expected	RTA, DC, – PM, CV, CS	9.37	11	1.63

The investigation into the median-mean relationship suggested varied focuses among vendors, from those launching innovative services to those with mature, stable offerings. This variance underscored the diversity in vendor capabilities and strategic directions, emphasizing the importance of nuanced analysis in vendor selection for AI-driven projects at Roy Hill.

Analysis of score distributions illuminated the prevalence of left-skewed distributions, indicating a concentration of higher performance scores. This pattern highlighted potential areas of bias or overestimation in AI capabilities based solely on text generation skills, a factor underscored by [24]. Our findings align with emerging critiques in the literature, suggesting a gap between performance in controlled settings and the complexities of real-world application.

Furthermore, our use of Monte Carlo simulations to supplement average score assessments provided a robust alternative for evaluating vendor performances, navigating the limitations inherent in non-normally distributed data. This approach, coupled with an in-depth examination of specific AI capabilities and potential areas for development, offered a more holistic view of vendor competencies and gaps.

The research also delved into the practical application of MP-LLM, showing its superiority in problem-solving tasks through innovative prompting techniques such as DSP and CoT. These findings not only demonstrate the enhanced capabilities of MP-LLM but also its potential to address complex real-world challenges more effectively than traditional LLMs.

Discussion

Insights

Our research included LLMs with tested performance across diverse benchmarks, reflecting on their efficacy in domains like natural language processing and problem-solving. While GPT4 and Llama 2 showed promising results, [29] identified Mixtral 8x7b as a significant contender, challenging GPT4's prevailing dominance. However, despite GPT4 and Llama 2 accomplishments in structured evaluations, our findings resonated with [24], acknowledging concerns regarding the real-world applicability of LLMs due to inherent limitations, particularly in reasoning and understanding complex scenarios.

We extended the dialogue on LLMs by incorporating tests designed to assess reasoning capabilities, unveiling a gap between LLMs' performances in controlled environments versus more intricate, real-life tasks. Introducing the MP-LLM marked a notable advancement, showcasing an enhanced ability to navigate some of the complex patterns and logical tasks, thus offering a better insight into the potential utility of MP-LLM in practical applications. This underscored the necessity of moving beyond mere text generation capabilities when evaluating LLMs, highlighting the importance of emulating reasoning, understanding, and execution in assessing their true competence.

Cognitive Tests

Our methodology diverged from standard benchmarks, employing a series of tests, including problem-solving scenarios and riddles, to challenge LLMs beyond their conventional capacities. This approach revealed significant disparities among LLMs/LLM products, particularly in scenarios necessitating coherent reasoning and planning, whereas MP-LLM's performance was better, suggesting avenues for future enhancements in LLM technology. By integrating Directed Stimulus Prompting (DSP) and Chain of Thought (CoT) techniques, we unveiled the potential of tailored prompting strategies in

refining logical deduction and reasoning-like abilities of LLMs in ways that made the logic employed visible, signalling a promising direction for future research in LLM methodologies.

Within our investigation, we handpicked a series of tests from previously published literature to examine the response of LLMs/LLM products (used in the AI assessment) in use of their tactics in problem-solving, indicative of coherent reasoning. The LLMs/LLM products encounter difficulties in producing responses akin to reasoning. A zero-shot approach was employed for all the tests, which were run repeatedly, and the results noted were observations that occurred for most of the time. The MP-LLM had the flexibility via a conglomerate of tested prompting techniques (such as DSP and CoT), a diversity of personas to suit the problem-domain, and other relevant information such as BODMAS, i.e. the correct order of mathematical operations to solve math problems. Explanation of how when each unique mix of the conglomerates were used are described under each scenario and this helped to find solutions most of the scenarios than the alternative LLMs/LLM products. It is crucial to clarify that these tests do not purportedly assess cognitive-like faculties inherent in human thinking, such as attention, memory, learning, or executive functions [25].

Instead, the primary objective revolved around isolating the reasoning element of the thought process, comparing generated responses against idealized reasoned counterparts. Consequently, this comparative approach sheds light on prevailing deficiencies among contemporary LLMs/LLM products, thereby guiding future developments targeting improvements in logical deduction and reasoning capabilities.

This comparative analysis further emphasized the limitations within current LLM architectures, particularly in addressing mathematically intensive problems [26], advocating for the enrichment of LLMs with state-of-the-art prompt designs to foster robust decision-making capabilities and enrich the context for more accurate AI capability assessments.

Implications

In alignment with scholarly endeavours, such as those by [27], our study contributes to the burgeoning exploration of abstract-reasoning capabilities in LLMs, albeit within the confines of our investigative scope. The outcomes signal a crit-

ical reconsideration of LLM capabilities, advocating for an extensive examination to dissect the nuanced potential of LLMs in replicating human-like cognitive processes, particularly in areas necessitating advanced reasoning and planning.

The practical implications of our research are vast, suggesting that leveraging an integrated approach combining multi-source ratings and LLMs/LLM products can significantly enhance the reliability of vendor AI capability assessments. Moreover, by fostering a blend of LLM/LLM product responses with survey results and exploiting the Multi-Persona LLM performance, we have a greater chance not only to bolster trust and comprehensiveness in assessments, but also to pave the way for a more efficient, idea-explorative process in creative problem-solving.

Our future studies will venture into harmonizing the MP-LLM performance with various prompting strategies and exploring knowledge graphs to inform DSP, aiming to amplify the performance and usability of LLMs across diverse operational needs. Ultimately, this work lays a foundational stone in the broader discourse on optimizing LLM/LLM product utility, guiding the path towards achieving a meticulous understanding and application of LLMs in navigating the complexities of real-world problem-solving scenarios.

Conclusion

Our study underscored that assessing AI capabilities of vendors based solely on publicly available information is fraught with challenges, notably due to risks of information asymmetry and confirmation bias that compromise assessment validity. To counteract these limitations, we advocate for an integrative approach that combines vendor feedback, credible source verification, and the innovative use of LLMs enhanced by Monte Carlo statistical inference. This methodology cultivates an artificial collective intelligence, delineating a more reliable framework for evaluating a vendor's AI prowess.

Key to our findings is the realization that while LLMs are invaluable for augmenting productivity and fostering creativity, their efficacy is contingent upon precise prompt engineering and the adoption of strategies like Multi-persona, CoT, and DSP. These techniques are especially crucial for navigating complex tasks, with the caveat that fully mathematical challenges remain beyond the scope of current LLM capabilities. The accuracy of LLM-generated outputs significantly benefits

from detailed, well-structured prompts, reinforcing the notion that LLMs serve as catalytic collaborators in the creative process, amplifying rather than replacing human ingenuity.

The employment of personas in prompt engineering emerged as a vital strategy for tailoring LLM outputs to match the nuanced demands and perspectives of diverse user groups, showcasing the potential of persona-based prompts in enhancing the relevance and utility of LLM responses for specific decision-making contexts. By simulating various professional viewpoints and expertise, this approach enriches the idea discovery phase of creativity, ensuring decisions are underpinned by a comprehensive understanding of the problem context.

As we advance our research, we aim to amalgamate the collective insights gained from LLM responses with survey findings and the nuanced performance of Multi-Persona LLMs. This endeavour seeks to consolidate trust, expand idea exploration, and elevate assessment efficiency. Future studies will fo-

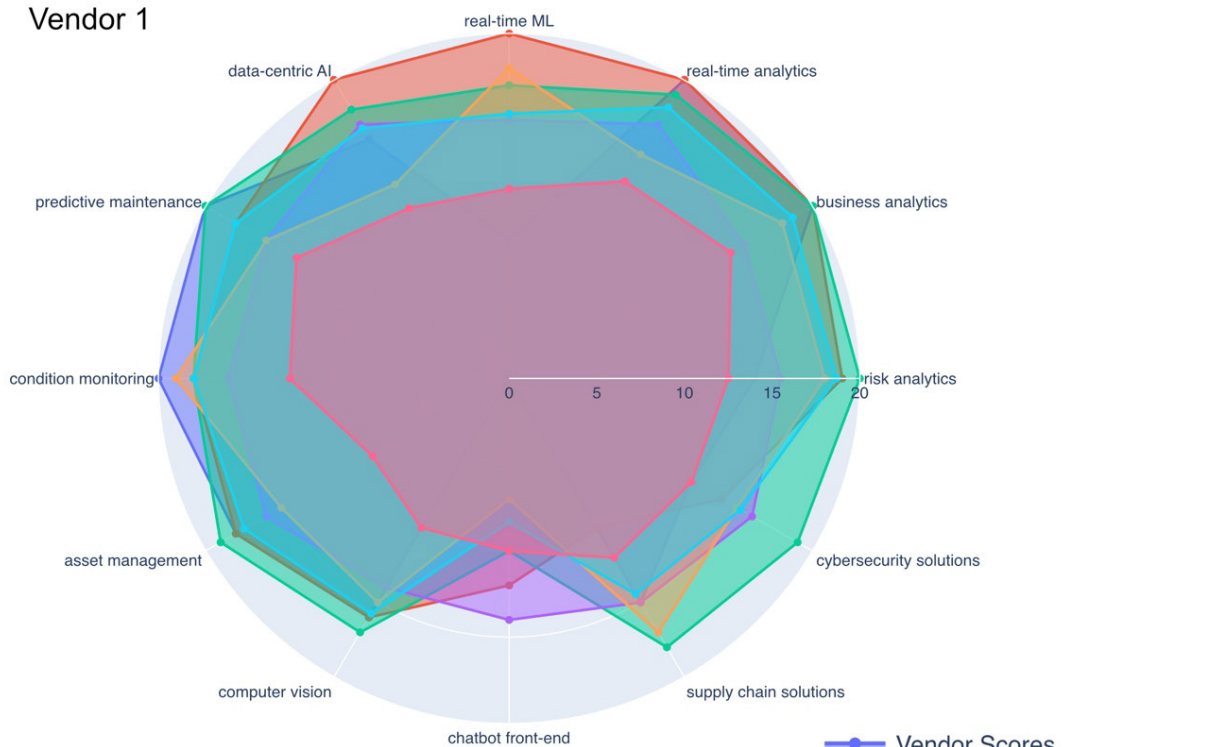
cus on developing a diverse library of personas tailored to specific business and operational needs within the Roy Hill context, alongside exploring knowledge graphs to refine DSP methodologies, thereby broadening the application spectrum of our findings.

In summary, this work contributes significantly to the field by proposing a nuanced model for AI capability assessment that leverages the strengths of LLMs while addressing their inherent limitations. Through our continued research, we anticipate forging a path toward more accurate, reliable, and creative AI capability assessments, setting a precedent for future explorations in the domain.

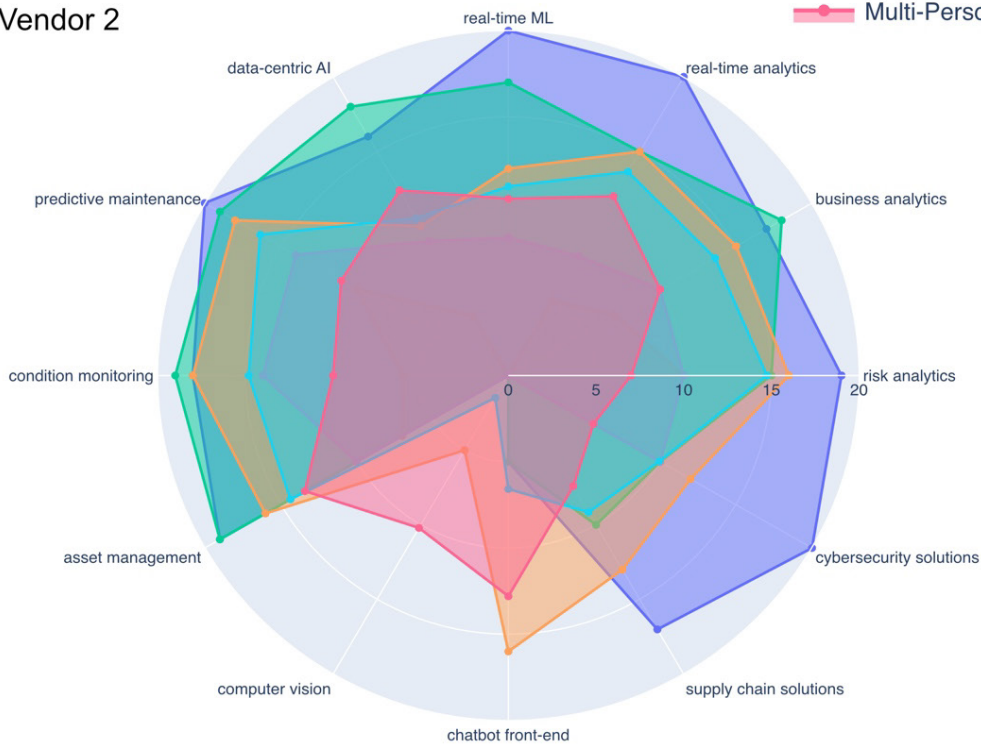
Acknowledgments

We are also thankful to Mitin Hirani (Roy Hill) and Brenda van Rensburg (Roy Hill) for constructive comments for earlier versions of the paper.

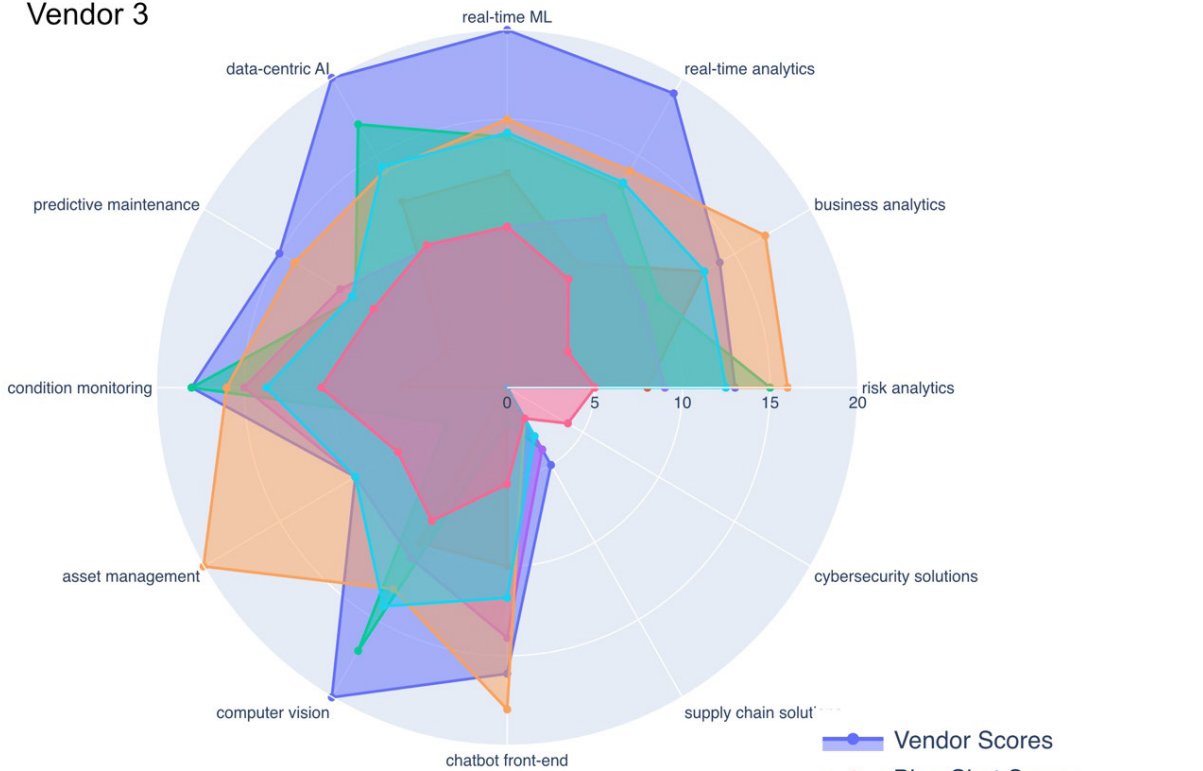
Vendor 1



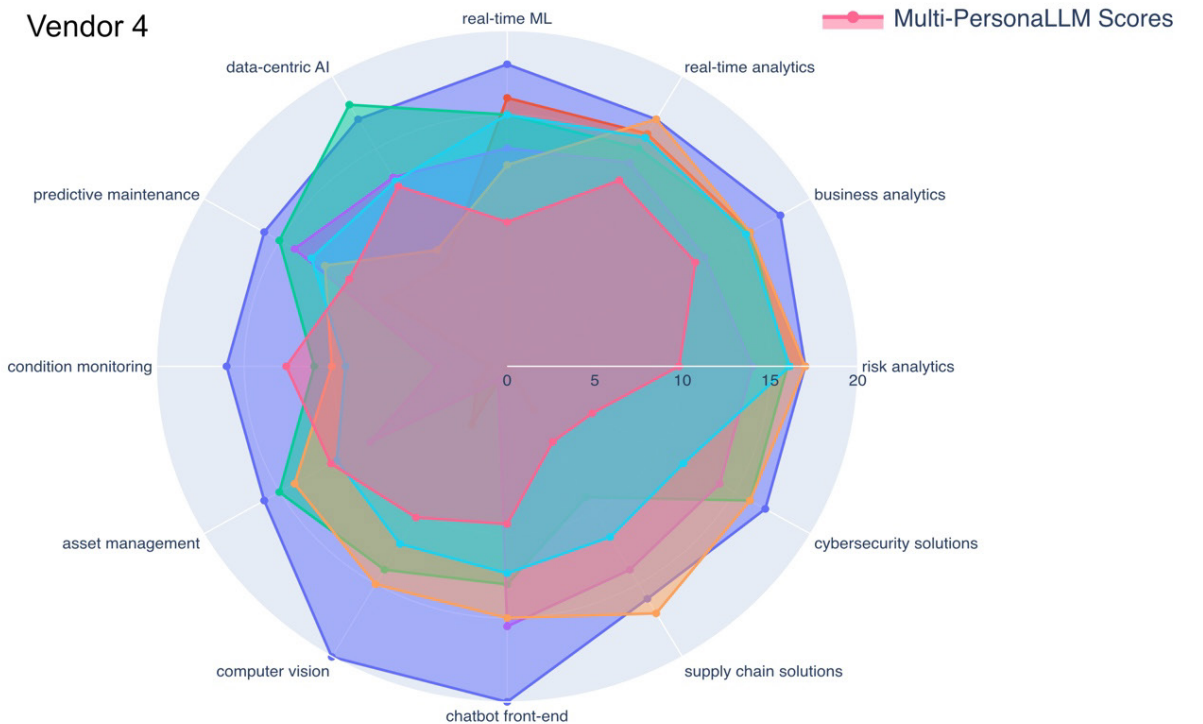
Vendor 2



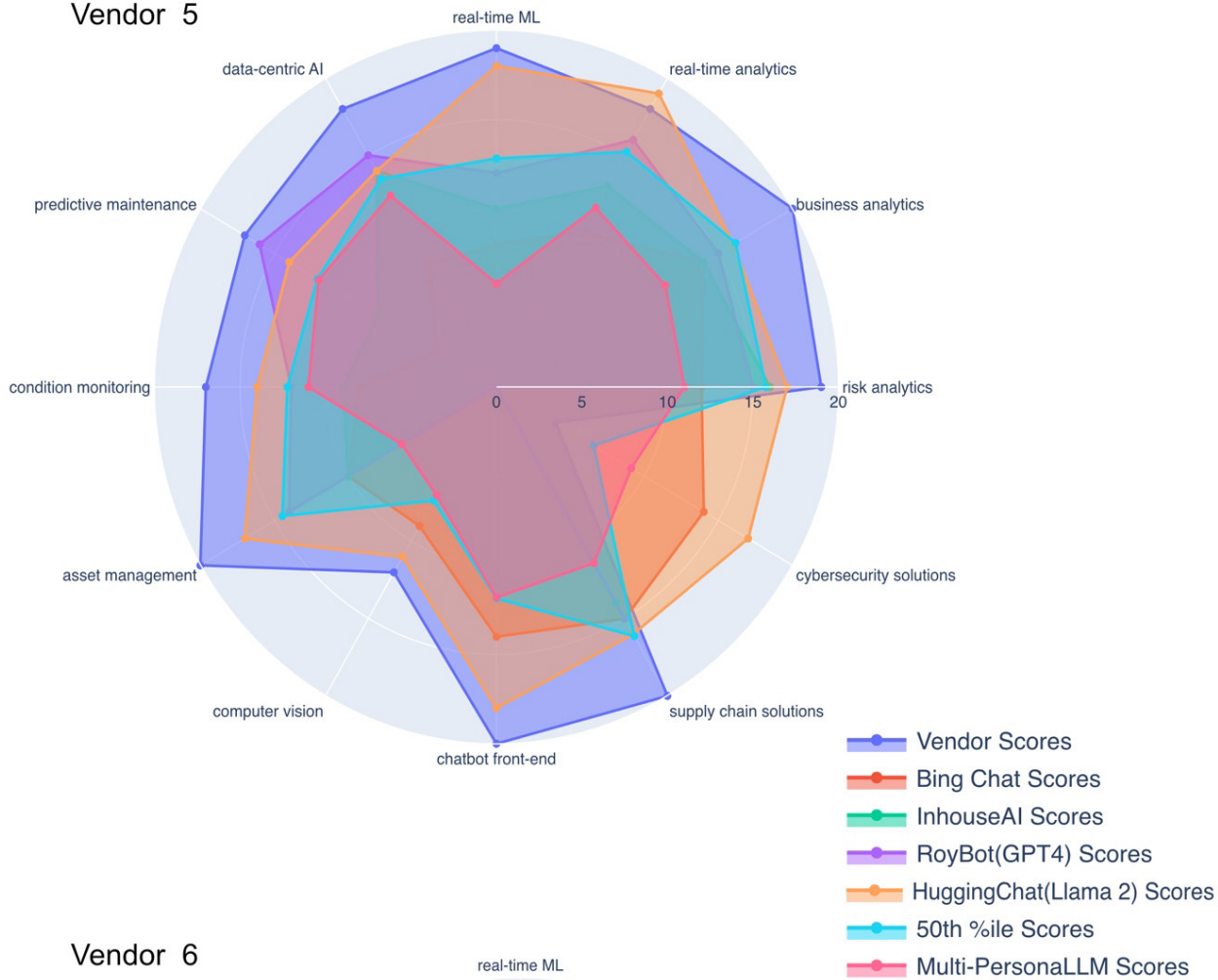
Vendor 3



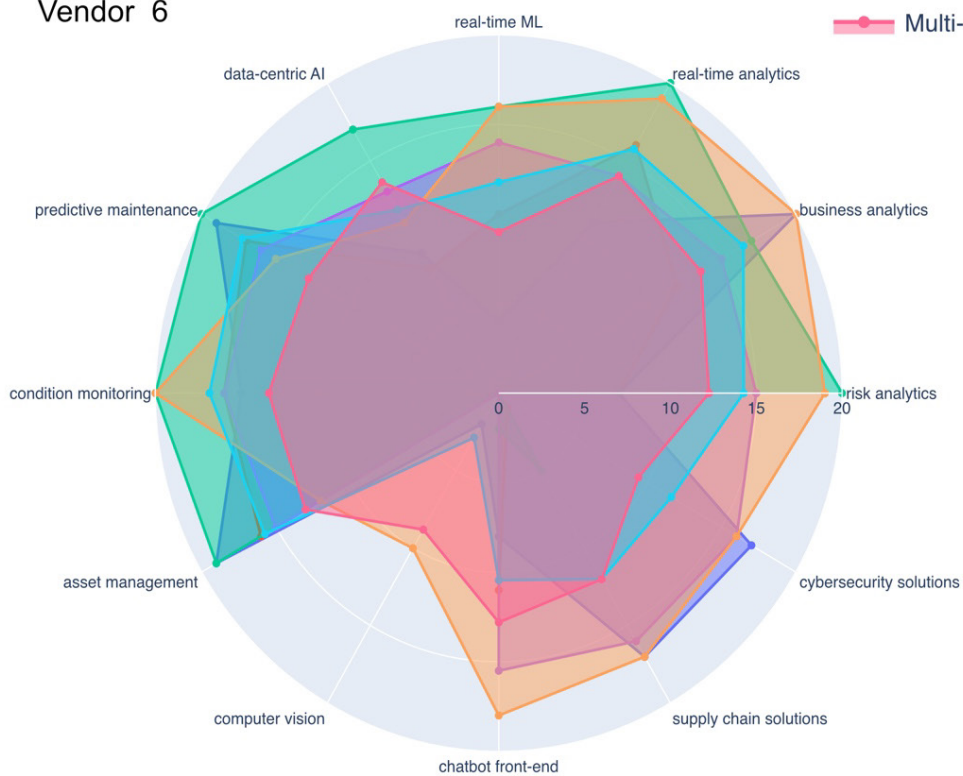
Vendor 4



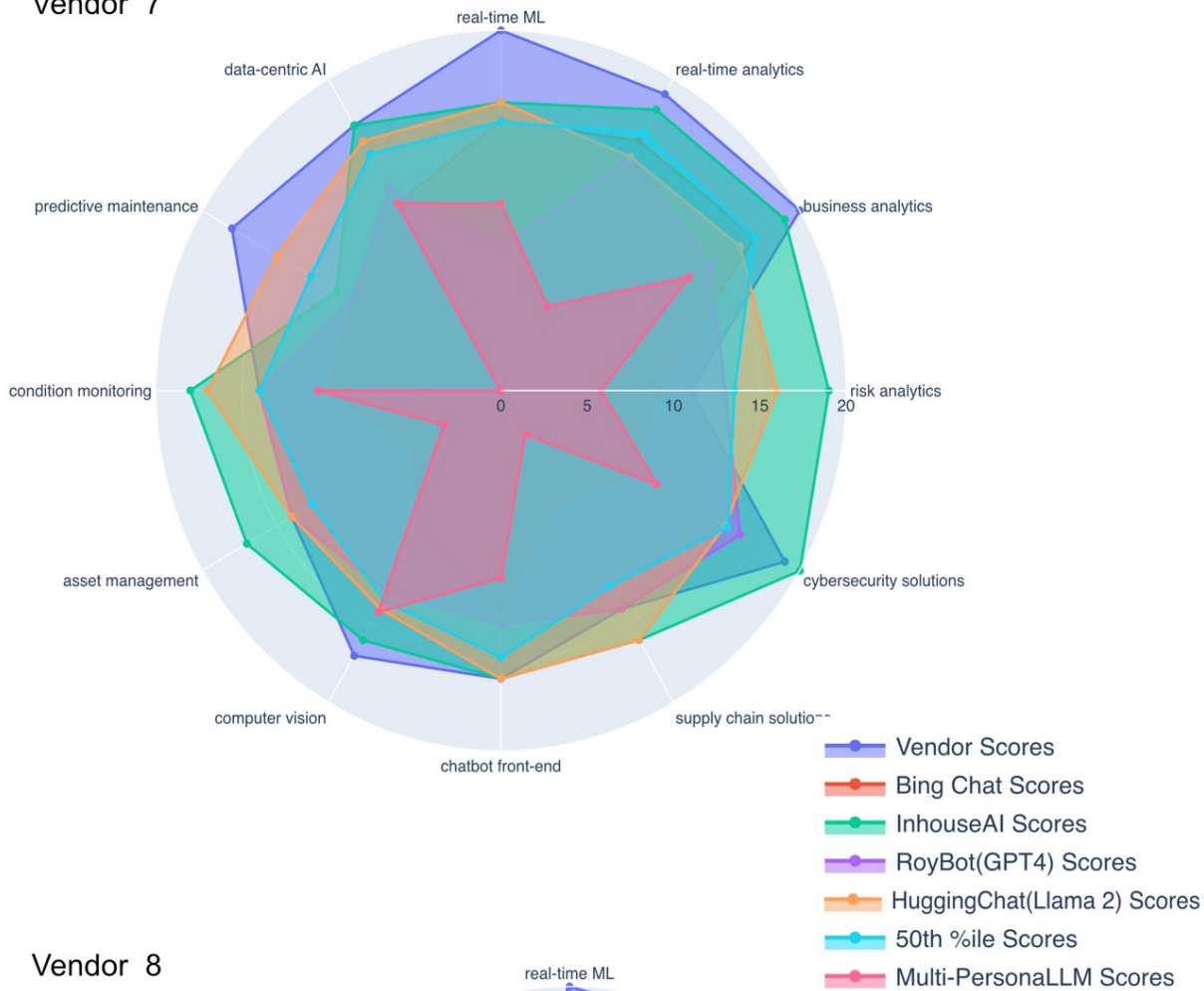
Vendor 5



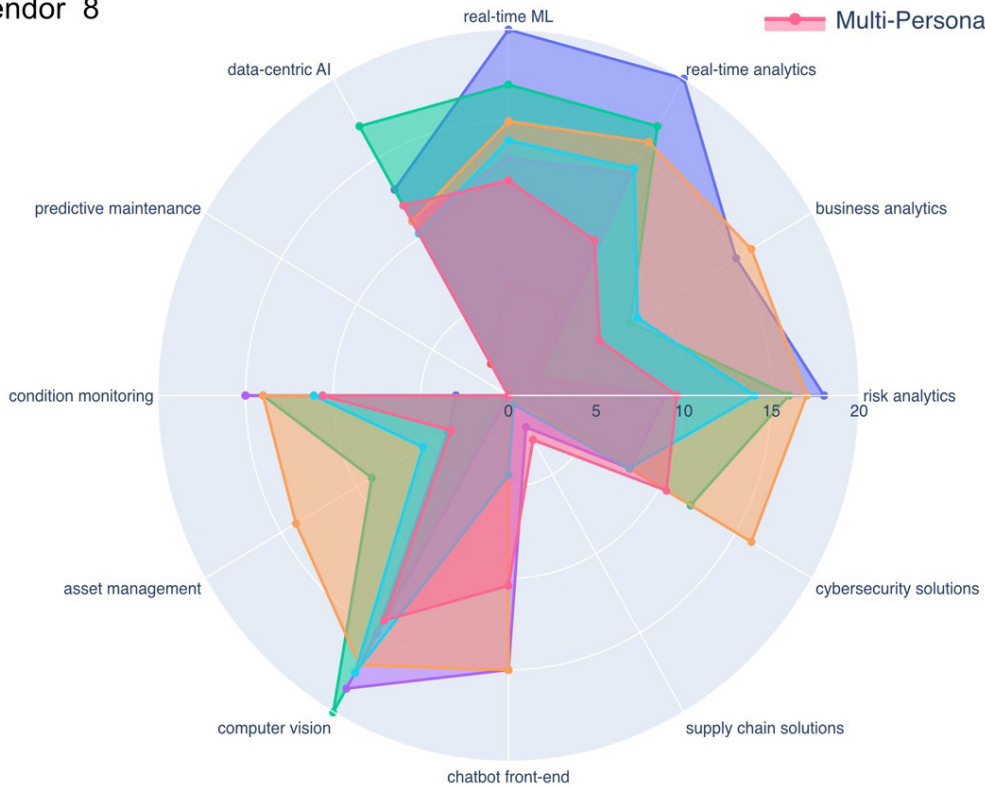
Vendor 6

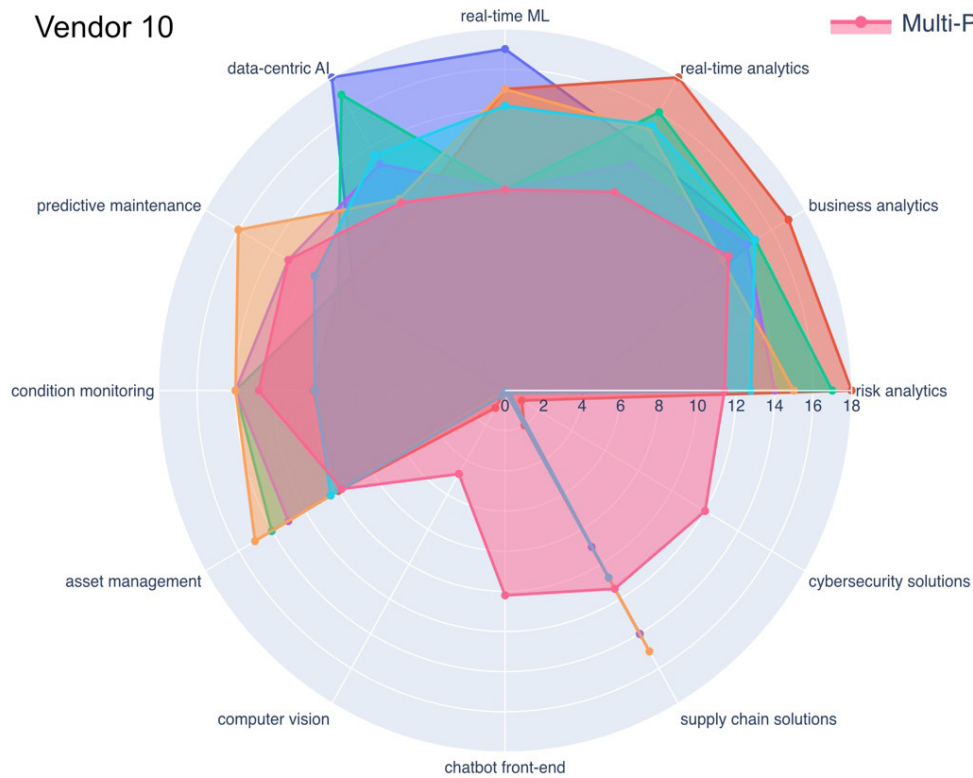
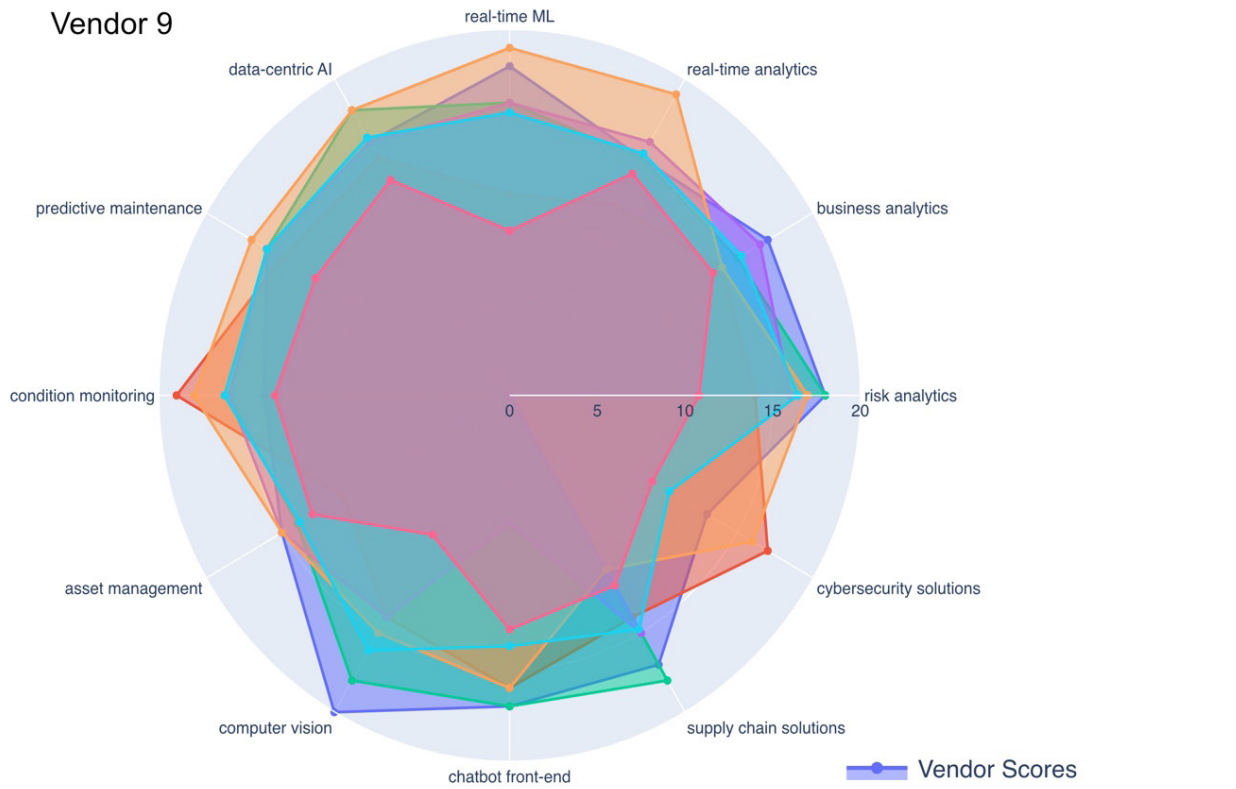


Vendor 7



Vendor 8





Appendix A: Still images for the different radar plots for the rest of the other vendor assessments

References

1. Granville V (2023) Generative AI: Synthesis Data Vendor Comparison and Benchmarking Best Practices [<https://ml-techniques.com>]. Generative AI.
2. Clear ML (2023) Enterprise Generative AI Adoption (p. 25p). <https://go.clear.ml/new-research-report-on-enterprise-generative-ai-adoption>
3. Linden A, Fenn J (2003) Understanding Gartner's Hype Cycles (Strategic Analysis Report R-20-1971; p. 12p). Gartner Research.
4. Hackney H (2023) Gartner Places Generative AI on the Peak of Inflated Expectations on the 2023 Hype Cycle for Emerging Technologies. *Architecture & Governance Magazine*.
5. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, et al. (2022) On the Opportunities and Risks of Foundation Models (arXiv:2108.07258). arXiv.
6. Prater JD (2023) Graft - Data-Centric AI with Foundation Models: A Practical Guide. <https://www.graft.com/blog/data-centric-ai-with-foundation-models>
7. IBM Data, AI Team (2023) Open source large language models: Benefits, risks and types. IBM Blog.
8. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. (2023) Llama 2: Open Foundation and Fine-Tuned Chat Models (arXiv:2307.09288). arXiv.
9. Simplilearn (2023) What is Bing Chat? Unleash the Power of GPT-4 With Bing Chat.
10. Ortiz S (2023) What is Copilot (formerly Bing Chat)? Here's everything you need to know. ZDNET.
11. Gao Y, Xiong Y, Gao X, Jia K, Pan J, (2023) Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv.Org.
12. Minaee S, Mikolov T, Nikzad N, Chenaghlu M, Socher R, et al. (2024) Large Language Models: A Survey (arXiv:2402.06196; Version 1). arXiv.
13. West J (2023) Introducing HuggingChat: A Strong Competitor for Open Source ChatGPT.
14. Johansen AM (2010) Monte Carlo Methods. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education*, 296–303.
15. Jiang H, Zhang X, Cao X, Kabbara J (2023) PersonaLLM: Investigating the Ability of GPT-3.5 to Express Personality Traits and Gender Differences (arXiv:2305.02547). arXiv.
16. Qian C, Cong X, Liu W, Yang C, Chen W, et al. (2023) Communicative Agents for Software Development (arXiv:2307.07924). arXiv.
17. Yang R, Narasimhan K (2023a) The Socratic Method for Self-Discovery in Large Language Models. Princeton NLP.
18. Horsey J (2023) GPT-4 vs GPT-4-Turbo vs GPT-3.5-Turbo performance comparison—Geeky Gadgets.
19. Muralidhar K (2003) Monte Carlo Simulation. In H. Bidgoli (Ed.), *Encyclopedia of Information Systems*, 193–201.
20. Hubbard DW (2019) A Multi-Dimensional, Counter-Based Pseudo Random Number Generator as a Standard for Monte Carlo Simulations. 2019 Winter Simulation Conference (WSC) 3064–73.
21. Savage S, Thibault M, Empey D (2017) SIPmath Modeler Tools for Excel: Reference Manual. ProbabilityManagement.org.
22. Johnson D (1997) The triangular distribution as a proxy for the beta distribution in risk analysis. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46: 387–98.
23. Savage S (2012) The flaw of averages: Why we underestimate risk in the face of uncertainty. John Wiley & Sons, Inc.
24. Burtsev M, Reeves M, Job A (2024) The Working Limitations of Large Language Models. *MIT Sloan Management Review*, 62: 10.
25. Xie Y, Xie T, Lin M, Wei W, Li C, et al. et al. (2023) OlaGPT: Empowering LLMs With Human-like Problem-Solving Abilities (arXiv:2305.16334).

26. Ahn J, Verma R, Lou R, Liu D, Zhang R, Yin W (2024) Large Language Models for Mathematical Reasoning: Progresses and Challenges (arXiv:2402.00157). arXiv.
27. Mitchell M (2023) Can Large Language Models Reason? [Substack newsletter]. AI: A Guide for Thinking Humans.
28. Frost J (2023) Cumulative Distribution Function (CDF): Uses, Graphs & vs PDF. Statistics By Jim.
29. Nest D (2023) LLM Benchmarks: What Do They All Mean? <https://www.whyyai.com/p/llm-benchmarks>
30. Li C, Liang J, Zeng A, Chen X, Hausman K, et al. (2023) Chain of Code: Reasoning with a Language Model-Augmented Code Emulator. arXiv.Org.
31. Yang R, Narasimhan K (2023) The Socratic Method for Self-Discovery in Large Language Models. Princeton NLP.