Exploration of Using Generative AI to Solve CI Analytical Tasks

Austin Fleisher
Georgetown University
austin.fleisher@aurorawdc.com

Ph.D. Qingjiu Tao, Ph.D.
Assistant Professor of Strategic Management
University of Delaware
qtao@udel.edu

ABSTRACT This study examines the efficacy of generative AI platforms, including ChatGPT 3.5, Bing AI, and Bard, in performing competitive intelligence (CI) tasks such as Stakeholder Analysis, Issues Identification, and Trends Forecasting. These tasks are vital for analyzing various social, technological, economic, ecological, political, legal, ethical, and demographic factors that influence organizational or market dynamics. The main aim is to assess the capabilities of generative AI against human CI analysts in these areas, utilizing prompts and surveys to gather data.

The research engaged seven intelligence analysts of diverse experiences to evaluate the outputs from these AI platforms, establishing a benchmark for performance standards and identifying any limitations of generative AI in CI roles. The findings indicate that as of the end of 2023, generative AI does not match the performance level of human analysts in CI tasks. However, all platforms demonstrated some capacity to handle the tasks, with varying levels of success depending on the specific function, suggesting that while they cannot replace human analysts, they can augment their efforts.

Significantly, the study reveals that each AI platform has unique strengths and weaknesses, which were evident in their task-specific performances. Bard emerged as the most consistent, while Bing performed well in issues and stakeholder analysis but less so in trends forecasting. ChatGPT showed the least consistent results across the tasks.

The implications of these findings are profound for the integration of AI into CI roles. Analysts are encouraged to experiment with different AI tools to discover which best suits their needs, recognizing that these technologies are tools to assist rather than replace human analytical skills. This approach not only preserves but enhances the human element crucial in CI tasks, ensuring that AI and human intelligence together will shape the future of the competitive intelligence field.

INTRODUCTION

The release of OpenAI's ChatGPT in November of 2022 marked the start of what would grow to be a commercially viable generative artificial intelligence (AI) platform. A generative AI platform is capable of text, images, and media in response to prompts based on how it is trained using large language models. The platform had over one million users in December 2022 and has surpassed one hundred million users and 1.6B visits by June of 2023. The explosive

growth of generative AI platforms has led businesses to experiment incorporating the technology to assist roles and other tasks, with implications for impacts in the workforce. ChatGPT has already started replacing roles, including a San Franciscan copywriter who attributes her layoff to ChatGPT being cheaper to use than her role (Verma & De Vynck, 2023). There are concerns that the technology will be capable of replacing human workers in additional industries such as media, teaching, tech, market research and more being at risk (Mok & Zinkula, 2023). This risk can also be carried over to business and competitive intelligence (CI) as questions arise whether AI will be cheaper and more effective than human analysts in these roles. Generative AI platforms, including ChatGPT, Bing AI, and Bard, have the potential to perform tasks traditionally handled by Competitive Intelligence (CI) analysts.

For this study, these tasks will be narrowed down to Stakeholder analysis, Issues identification, and Trends forecasting. 'Stakeholders' refer to parties who have an interest or stand to benefit from a given subject. 'Issues' encompass controversies that influence the external environment of an organization. 'Trends' involve an examination of social, technological, economic, ecological, political, legal, ethical, and demographic (STEEPLED) factors that can impact an organization or market dynamics. The primary objective of this research is to discern the boundaries of how generative AI can tackle these three research tasks in comparison to human abilities, utilizing prompts and surveys as investigative tools. This leads to the following questions: What constitutes the boundaries when employing generative AI within the context of modern Competitive Intelligence roles and tasks? How does generative AI compare to human CI analysts in identifying Stakeholders, Issues and Trends with a given Landscape?

Subsequent sections will detail the methodology and analytical frameworks employed, discuss the differences among the AI models based on their training data and parameters, and present a thorough analysis of the data collected from a panel of experienced intelligence analysts. This comprehensive approach will not only evaluate the performance of each AI platform but also provide insights into the evolving landscape of competitive intelligence, driven by technological advancements.

REVIEW OF LITERATURE

This review of literature will start with defining generative AI. We will also explain different generative AI platforms, particularly ChatGPT, Bing and Bard and the models they use. Finally, we will address the research gap on the use of generative AI in competitive intelligence (CI) in existing research.

Defining Generative AI

Generative AI is a subset of AI that could fall under the ANI or AGI types depending on the application. Generative AI saw an increase in attention in 2022 with the release of OpenAI's ChatGPT which brought the product to everyday consumers. Generative AI can be considered as using 'Deep Learning' which Sahoo et al define as "a subset of machine learning that mirrors the human brain in learning and responding to data, information, and prompts" (Sahoo et al, 2023). Deep learning can be used to help applications with pattern recognition and learning by example to accomplish tasks. Generative AI platforms, like ChatGPT learn off the data that is inputted. Lim et al define generative AI as "a technology that (i) leverages deep learning models to (ii) generate human-like content (e.g., images, words) in response to (iii) complex and varied prompts (e.g., languages, instructions, questions)" (Lim et al, 2023). Another definition by researchers Feuerriegel et al states that generative AI "refers to computational techniques that are capable of generating seemingly new, meaningful content such as text, images, or audio from training data" (Feuerriegel et al, 2023). For this study, generative AI can be defined as models or algorithms that learn and structure data to generate responses to prompts.

Generative AI Models, Information on Bard, Bing, ChatGPT and Use Cases

This research intends to use the platforms ChatGPT, Bing and Bard for its objectives.

The platform ChatGPT, which stands for 'Chat Generative Pre-trained Transformer,' can be considered a chatbot given the conversational nature of the technology.

Following the release and success of ChatGPT, Microsoft decided to release its own generative AI platform called Bing AI based off the search engine Bing (Mehdi, 2023).

Bard is a competing product to ChatGPT developed by Google (Pichel, 2023). Bard was initially released on March 21, 2023 (Faguy, 2023).

Generative AI has use cases in many areas including for use in education healthcare and more. Researchers are currently using the technology to determine what effects the technology will have in their respective fields.

Baidoo-Anu and Ansah (2023) published a study exploring the potential benefits of technology in education. They found that it could be used for personalized tutoring, essay grading, and language translation, among other applications. In another industry, Mannuru et al. (2023) examined the impact of generative AI on different nations. They discovered that while generative AI affects many industries, its impacts are unevenly distributed, with the Global South facing unique challenges in accessing and benefiting from these technologies. In healthcare, Zhang and Boulos (2023) focused on using generative AI to identify opportunities and challenges, highlighting issues such as trust, veracity, clinical safety and reliability, privacy, copyrights, ownership, and the potential to create user-friendly AI-driven conversational interfaces for health applications. While generative AI is still emerging, ongoing research continues to uncover new opportunities. However, there is little to no research on using generative AI in the field of competitive intelligence.

Generative AI in Competitive Intelligence

Competitive intelligence (CI) practitioners have shown interest in generative AI technology, but scholarly publications in the field remain scarce. CI involves gathering, analyzing, and disseminating information to stakeholders to reduce risks in plans, operations, and strategy. At the Council of Competitive Intelligence Fellows' annual summit on September 25 and 26, 2023, generative AI dominated the conversation, with over 50% of the sessions on day two discussing its impact on the field, including a session on AI's role in analysis. Scholarly research on CI-related topics is slowly becoming more prevalent. Peres et al. (2023) published work on how AI may affect research, finding that effective GenAI tools could yield substantial time savings. Tran and Murphy (2023) explored the role of generative AI in entrepreneurship, discovering that it can execute demand forecasts, suggest lean manufacturing designs, and produce information to maximize profit margins. While these studies are not directly associated with CI, they are closely related.

Fred Hoffman and Shelly Freyn (2022) surveyed CI practitioners to understand the necessary skills in the field. The results showed that while new technologies are evolving the discipline, soft skills like communication and analytical abilities remain crucial. This study focused on identifying CI skills rather than experimenting with AI to address these skills or identify gaps. Additionally, it surveyed common types of strategic and tactical analyses (SATs) but did not explore how generative AI platforms could support or accomplish these tasks. It did not address whether generative AI could solve these issues.

Andrej Cekuls (2022) published an article on AI-driven CI, concluding that AI will significantly impact the business world and help CI grow. The article discusses areas where AI may play a role, such as hiring and marketing intelligence, but does not provide examples or use cases. Finkenstadt et al. (2023) published research on using generative AI for scenario planning, a technique used in business and CI. They discussed how generative AI can be used for scenario creation, narrative exploration, strategy generation, and scenario creation. They concluded that generative AI can help companies perform scenario planning quickly, at lower costs, and with more scenarios than those created by humans. These initial studies and articles focus on the

potential of AI in CI and scan the field for insights on its future impact. However, aside from Finkenstadt et al.'s research, they do not offer practical exercises for how AI or generative AI can currently play a role. This research seeks to close the gap by exploring the use case of generative AI in traditional CI roles and tasks, particularly in identifying issues, stakeholders, and trends.

Research Design

This study takes a qualitative research approach using an expert panel review to determine the effectiveness of generative AI platforms for completing key competitive intelligence tasks including Issues, Stakeholder and Trends analyses. Expert panels are used for research across a variety of industries to evaluate research. According to the NIH, an "expert panel provide evidence-based information to guide research practice and health care decision making" (Coulter et al, 2016). An expert panel approach was chosen to benefit from the panels "acknowledged technical expertise to assess the predicted or measured quality of performance" (Feller, I, 2013) for the AI outputs. A qualitative approach was chosen over a quantitative approach due to the qualitative nature of generative AI outputs for this study.

The data being collected focuses on expert opinion and scoring reacting to non-numerical AI outputs. The central research question is looking to answer how AI can tackle research tasks which qualifies as a qualitative inquiry (McGill, 2023). This research's expert panel analysis was conducted online. One benefit to conducting an online expert panel is the "reduce possible biases based on participant status or personality" and "the benefit of contributing to the elicitation process at the time convenient to panelists" (Khodyakov et al, 2011).

Sources and Data

This research revolves around secondary sources powering the generative AI outputs. The primary data, however, is the responses from the expert panel and their reactions to the AI outputs. The panel consists of six anonymous experts whose data can be found in table 1 below:

The expert panel reviewed the outputs from Bard, Bing and ChatGPT and scored each one on a 0-5 scale based on given criteria developed. This data is then combined and analyzed to identify how well the generative AI platforms performed against each other as well as respectively to the expectations for a human analyst.

Table 1. Expert Panel Demographics

	Age	Sex	Race	Country	Occupation	Years of Experience
Expert 1	27	M	Caucasian	USA	Intelligence Program Manager	5-10
Expert 2	27	M	African American	USA	Army Officer	5-10
Expert 3	28	M	Asian	USA	Air Force - Cyber	5-10
Expert 4	32	M	Caucasian	Canada	Solutions Architect	10-15
Expert 5	35	F	Caucasian	USA	Intelligence Professional	10-15
Expert 6	49	M	Caucasian	USA	Director Of Market Intelligence	20-25

Analysis and the use of Prompts

The design for this study involves crafting prompts for three major generative AI platforms, Bard, Bing and ChatGPT, to simulate the process for completing Issues,

Stakeholder and Trends analyses. For completing these analyses in a consistent manner, a landscape is being provided which for this study is the CHIPS and Science Act. The prompts will cover the task of defining

the analysis type, having the AI identify twenty issue, stakeholders or trends, narrow the set down to the most important responses for the analysis type and finally providing a recommendation based on a prompt relevant to the analysis type. The generative AI responses are blinded so that the experts will not know which platform they are evaluating for the responses. The panel will be evaluating each response using a zero-to-five-point scale with zero meaning the AI failed to generate a response and a five representing a response fit to share with a prospective stakeholder.

The following prompts were developed and used for the Issues Analysis task:

Prompt 1: Define an "issue" from the perspective of a competitive intelligence analyst

Prompt 2: Identify and prioritize 20 competitive intelligence issues from the CHIPS act

Prompt 3: Identify and prioritize the top 5 most impactful competitive intelligence issues related to the CHIPS act.

Prompt 4: Identify and prioritize the top 5 most urgent competitive intelligence issues related to the CHIPS act.

Prompt 5: On a scale of 0-100%, with 100% meaning it is highly likely and zero being highly unlikely, what is the likelihood that these issues will be resolved within 1 year?

Panel Question: Rate the AI's performance in generating relevant, actionable and trustworthy insights via the 5 prompts

A workable definition for an issue in competitive intelligence would need to state that an issue is a controversy and that there exists a gap between an organization and stakeholder/s. The Issues Analysis in competitive intelligence differentiates issues by identifying which would be considered impactful and urgent. Impactful is defined as significant, or high priority to the organizations' strategy, and relevant and applicable to an organization's strategy. Urgent is defined as requiring immediate executive attention and potential responses. Following the AI identifying, prioritizing and differentiating issues, it then is tasked with determining the likelihood that the issues will be resolved which is scored based on the reliability of the response. Reliability is defined as using multiple data sources to produce a rating and identifying additional factors that may also affect the rating.

The following are the prompts developed and used to represent a Stakeholder Analysis:

Prompt 1: Define a "stakeholder" from the perspective of a competitive intelligence analyst

Prompt 2: Identify and prioritize 20 competitive intelligence Stakeholders from the CHIPS act

Prompt 3: Identify and prioritize the top 5 most influential competitive intelligence stakeholders related to the CHIPS act.

Prompt 4: Identify and prioritize the top 5 entities with the greatest stake(s) with respect to the CHIPS act

Prompt 5: On a scale of 0-100%, with 100% meaning it is highly likely and zero being highly unlikely, what is the likelihood that these stakeholders will act/take action on the CHIPS act within 1 year?

Panel Question: Rate the AI's performance in generating relevant, actionable and trustworthy insights via the 5 prompts

A stakeholder must have interest, influence and stake in the market to be classified as such. Competitive intelligence differentiates stakeholders by identifying the most influential and the entities with the largest stake. Influential is defined as a stakeholder having enough power to alter the outcome of an issue in the market. Stake is defined by how much direct/proximate impact the issue has on the entity, what they stand to gain or lose, the nature of values involved and the entities' ability to influence the outcomes. The AI lastly is evaluated on the reliability of its response in identifying the likelihood that identified stakeholders will act or take action on the given landscape.

The following are the prompts developed and used to simulate a Trends Analysis:

Prompt 1: Define a "trend" from the perspective of a competitive intelligence analyst

Prompt 2: Identify and prioritize 20 competitive intelligence Trends from the CHIPS act

Prompt 3: Identify and prioritize the top 5 most relevant competitive intelligence trends related to the CHIPS act.

Prompt 4: Identify and prioritize the top 5 most significant competitive intelligence trends related to the CHIPS act.

Prompt 5: Identify the indicators for how the trends are becoming more or less significant
Panel Question: Rate the AI's performance in generating relevant, actionable and trustworthy insights via the 5 prompts
A trend includes one or more relevant and significant changes among Social, Technological, Economic, Ecological, Political (including regulatory), Legal, Ethical and/or Demographic (STEEPLED) factors and how these are associated with potential threats and opportunities coming from the external environment (Bensoussan & Fleisher C.S., 2012). Trends in competitive intelligence are prioritized by identifying relevance and significance. Relevant is defined as clearly associated with opportunities or threats attached to the issue. Significant is defined as having major potential impacts on the issues development. The AI is lastly evaluated on its ability to identify indicators and define their significance.

Following the generation of the prompts, the panel is then tasked with giving a score rating the AI's performance with respect to its responses to the five developed prompts.

Limitations

This research is limited by various factors. The first limitation is time constraints. The scope of this research was to be completed in less than four months which affected the number of platforms to be analyzed, the number of panelists who could participate and the scope of analysis prompts for evaluation. Future studies would benefit by having more generative AI platforms to analyze as well as spacing the study out to a longer time period to have a more longitudinal view of AI performance. The second limitation is the number of experts involved. While these panelists come from a diversity of backgrounds, having a larger data set would improve this facet and allow for additional perspectives that could affect the outcomes.

This study involved a limited amount of diversity for the panel but having more individuals from different backgrounds could expand upon this. Another option for future studies would be to include experts from additional regions or countries as this study was only able to include analysts from the U.S. and Canada. Additionally, it would be to the benefit of future researchers to address having panelists with a wider set of gender and racial backgrounds to add new perspectives that could affect the outcome of the study. More female representation in particular would be welcome since this study did not have an equal number of female panelists involved.

The third limitation is the number of prompts used to simulate each analysis. This study focused on narrowing down the analyses to be completed in five prompts, but this only amounts to a general approach to them, and more in-depth prompting could be used to verify AI capability. Five prompts were developed for each in part due to time constraints as even reviewing five prompts from three AI platforms ended up being a time intensive task for the panelists. Since the experts were providing their opinions and time without compensation, any additional prompting would be limiting. Future studies would benefit from having extended time or means of compensation for the experts to mitigate.

The fourth limitation is the lack of human analysts performing the same task as the AI platforms. Given the other constraints, it was not feasible to have a set of human analysts perform the analysis tasks to offer a direct comparison. It would be advised that future studies include this element to bolster the validity of their findings.

The fifth limitation would be that the generative AI platforms used were the free version of each. ChatGPT, for example, offers a paid version that uses a different GPT model depending on the subscription level. Future studies may benefit from having the different tiers of the generative AI platforms examined as part of the prompting process.

Generative AI Platform Results for Analysis Tasks

This research is primarily examining whether generative AI can perform analysis tasks typically performed by competitive intelligence analysts. The analysis set chosen incorporates the LIST model, or Landscapes, Issues, Stakeholders and Trends (Fleisher C., 2022). A comprehensive LIST analysis provides stakeholders with information that helps them better

understand where to play in the market and can help influence decision making. A successful set of LIST analyses will increase confidence for decision makers and reduce uncertainty. For this analysis to maintain structure and uniformity for the generative AI platforms, the landscape is being provided. The chosen landscape being examined is the semiconductor industry and specifically the CHIPS and Sciences Act of 2022. The prompts structuring the Issues, Stakeholder and Trends (IST) analyses all use the CHIPS Act as the parameter for generating responses.

Bard using the PaLM 2 large language model (LLM), Bing using GPT-4 LLM and ChatGPT on GPT 3.5 LLM responses were documented in a blind fashion for the panel to review. The models chosen were the ones available on the entry level products for each platform. Scores for each analysis were determined by identifying the average of the six panelists responses to the six prompts. For each prompt, the panel attributed a score between 0-5 based on the criteria presented. Looking at the panelist scores, panelist 5 was more critical of the AI responses while panelists 2 and 3 were more lenient based on the criteria. Panelist 5 may have bias against AI platform responses or took a more critical approach to the criteria. Panelists 2 and 3, on the other end, may have used a very objective view of the criteria or what they considered as issues, stakeholders or trends. These outliers are accounted for when examining using average scores. The criteria for each prompt was developed with support from experts in CI, including feedback from the research advisor. The issues analysis results are displayed in Table 5 below. The full list of generative AI responses for this analysis can be found in Appendix 1.

Table 2. Issues Analysis Average Results

	Bard	Bing	ChatGPT
Total Avg.	2.86	3.31	2.33
Standard Dev.	1.33	1.53	1.72

Based on the panel results, Bing received the highest average score across all prompts with ChatGPT receiving the lowest. A three

represents that the platform partially generated relevant, actionable and trustworthy insights while a score of two shows that the platform mostly failed to do the same. The highest average score amongst the panel for any prompt was a 3.5 while the lowest was a 1.34. The average score showed that only Bing presented responses partially answering the objectives of the prompts while the other two mostly failed in their responses.

Looking at the standard deviation for each, Bard displayed the lowest value while ChatGPT had the most variance. A lower standard deviation means the data is more tightly clustered toward the average while a higher standard deviation shows data being dispersed (Hargrave, 2023). This shows that the panel found Bard to be less varied, Bing's scores to be slightly more varied and ChatGPT varied widely in quality. Where this is evident is ChatGPT's scores for prompts 2, 5 and 6 showing very low scores. These prompts were focused on identifying issues, identifying likelihood and overall performance of the task. The panel determined it was poor in issue identification and gave overall poor results for the analysis. Table 6 below shows the average scores for the stakeholder analysis. The full list of generative AI responses for this analysis can be found in Appendix 2.

Table 3. Stakeholder Analysis Average Results

	Bard	Bing	ChatGPT
Total Avg.	3.39	3.56	3.28
Standard Dev.	1.34	1.40	1.89

Bing received the highest average score for the stakeholder analysis responses. ChatGPT again scored the lowest. The average scores for this analysis were higher than the issues analysis which can be interpreted that generative AI has a slightly higher aptitude towards stakeholder identification than issues identification. Each platform managed to score above a 3 average showing that they partially accomplished the objective from the prompts.

A 3 would not be considered a client ready score for a deliverable.

This analysis showed a similar story in terms of standard deviation for the three platforms. Bard had the lowest standard deviation, Bing lower than on Issues but higher than Bard and ChatGPT with the highest standard deviation across all three analyses. ChatGPT's high deviation is likely from its poor performance on prompt 5 with every panelist except panelist 5 giving it a 0 and panelist 5 a 1. This prompt could have subconsciously biased the panel when looking at prompt 6, evaluating total performance of the platform, resulting in lower scores.

Table 4. Trends Analysis Average Results

	Bard	Bing	ChatGPT
Total Avg.	3.67	2.78	3.56
Standard Dev.	1.15	1.77	1.54

Unlike the previous two, Bard scored the highest average for trends while Bing scored the lowest. Bard had the highest average score for this analysis than any platform received for all results. It can be interpreted that Bard has a better aptitude towards trends identification based on the results while Bing displays a vulnerability with this task.

Bard displayed the lowest standard deviation across all three analyses here. ChatGPT displayed its lowest standard deviation while Bing had its highest. The panelists found that Bing could define trends but then the panelists found that its identification of them to be poor and then were split on how it refined them. Panelists 1, 4 and 5 all give Bing a score of 0 for prompt 3 and 4 while panelists 2, 3 and 6 all give it scores of 4 and 5's. The discrepancy among the panelists here affected Bing's total score for this analysis while also raising the standard deviation. It is possible that these differing scores could be how the panelists themselves determined what a relevant and significant trend is. Panelists 1, 4 and 5 may have determined that the response was too similar to prompt 2 and scored it low for failing to differentiate between the prompts

while the panelists found the response appropriate as relevant or significant.

Table 5 below shows the average results for each platform for all IST analysis tasks.

Table 5. Total Average Analysis Results

	Bard	Bing	ChatGPT
Analysis Average	3.31	3.21	3.06
Standard Dev.	1.31	1.60	1.79

These results show Bard as the most consistent generative AI platform when completing these analysis tasks. While Bing scored higher for issues and stakeholder, it ends up in second behind Bard in the final aggregate due to its poor trends results. ChatGPT displayed the weakest average results across all three analysis tasks. Based on the criteria provided, each platform on average was able to partially address the task for each analysis. The averages also show that no platform met the expectations for a client ready analysis for all three tasks. As to why the platforms did not perform to human expectations can potentially be explained in that two of the three tasks required conceptual reasoning. Issues and trends are subjective and analyzing subjective subject matter may not be a strong suit for generative AI now. While stakeholders are more objective, determining relevance and importance of stakeholders is subjective to the individual.

The standard deviation for all three analyses shows Bard as the most consistent with ChatGPT being the least. Bing's trend performance affected the total standard deviation but even without it, it would have a higher value than Bard. A low standard deviation for Bard shows that it will be the most consistent platform across the three for performing these analytical tasks. Bing and ChatGPT display more evident strengths and weaknesses for analysts to consider. They have the potential to outperform Bard with the right task and use case. There is evidence of negative skew in these scores in that Bing and ChatGPT's standard deviations are a result of a greater number of negative scores.

The dataset driving the LLM also could play a role in its development. The developer of

each platform keeps the training data proprietary, so it is not known how the models are learning compared to each other. Building a custom data set or using something such as OpenAI's Fine Tuning option, could lead to better results for these tasks but doing so is costly and requires significant amounts of data to be effective (OpenAI, 2023). Another possible explanation is that the three platforms examined were all using the entry version. It is possible that the paid version of ChatGPT on a GPT-4 model may have produced better results. However, at the time of research, Bing and Bard did not have alternate versions with paid versions or different models.

Another potential reason for the results could be that the panelists involved could be biased against AI platforms or a particular platform. While panelist 5 gave the lowest scores of the panel, they were consistent in their scoring across each platform and each analysis. Additionally, they did give scores of 4's and 5's (i.e. 4's on prompt 1 and 3 of stakeholder and 5's to Bard on prompts 3 and 4 of trends) on in response to prompts when reviewing the stakeholder and trends analyses. On the other hand, panelist 2 and 4 provided the highest scores with the same consistency. They also gave scores of 0's and 1's (i.e. panelist 2 giving Bing and ChatGPT a 0 and 1 for issues prompt 5; panelist 4 gave Bing and ChatGPT 0's and 1's for prompts 3, 4 and 5 of trends). The rest of the panel did not present explicit evidence of bias in their scoring.

Expert Panel Review

The panel examined the responses by Bard, Bing and ChatGPT and provided a score between 0-5 using given criteria to react to the prompt responses by each platform. Each panelist spent five to ten hours comprehensively reviewing the generative AI prompt responses when preparing and scoring the results. The following charts show the average and median scores for the panel for each of the analytic products. Both average and median are used to display the differences in the data since both can provide perspective. Median is displayed since outliers are present in the data while average is used to show comparison.
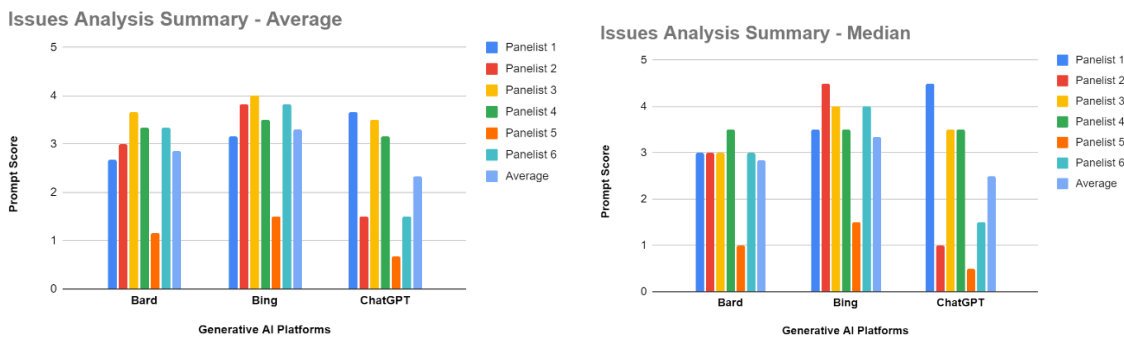

Figure 1. Issues Analysis Summary – Average and Median

For the issues analysis, the two charts are fairly comparable with outliers shows on Bing for panelist 2 and for panelist 1 for ChatGPT. Additionally, panelist 4 consistently scored the platforms lower than the rest of the panel based on the criteria provided. While the panel tended to agree that ChatGPT provided the weakest prompt responses, panelists 1, 3 and 4 all had a more positive reaction to its responses. The panel found that Bing provided the best responses to the issues prompts but the scores presented show that it only partially met the criteria in responses with an overall average score of 3. Across all platforms, Bard displayed the most consistent results according to the panel with four panelists having a median score of 3. Based on current models, the panel found that no platform was effective in completing an issues analysis.
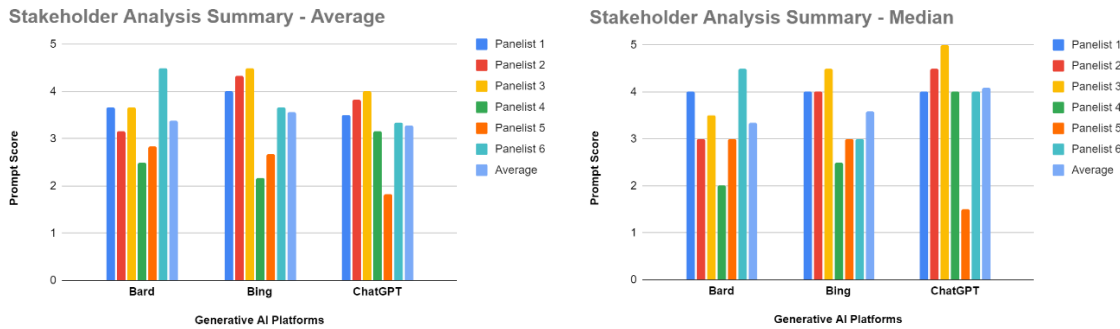
Stakeholder Analysis Summary - Average

Stakeholder Analysis Summary - Median

Figure 2. Stakeholder Analysis Summary – Average and Median

The stakeholder analysis charts show minor improvement in performance according to the panel. The average for ChatGPT received a boost from panelist 3 having a median score of 5 across all prompts. The higher scores for the stakeholder analysis could be explained by the generative AI platforms having more proficiency in identifying stakeholders compared to issues due to issues being more conceptual in nature. Based on the charts, Bing received the highest average scores while ChatGPT had higher median scores. The lower average score for ChatGPT represent a higher variable in response to individual prompts. Panelists found that it performed quite well on some of the questions but very poorly on other, lowering the average. This is expected since ChatGPT failed to provide a response on prompt 5 with panelists consistently scoring it a 0 in response. Using the median scores, ChatGPT outperformed its competitors for the rest of the responses. For this task, Bard displayed the most fluctuation in how panelists viewed its quality of response.
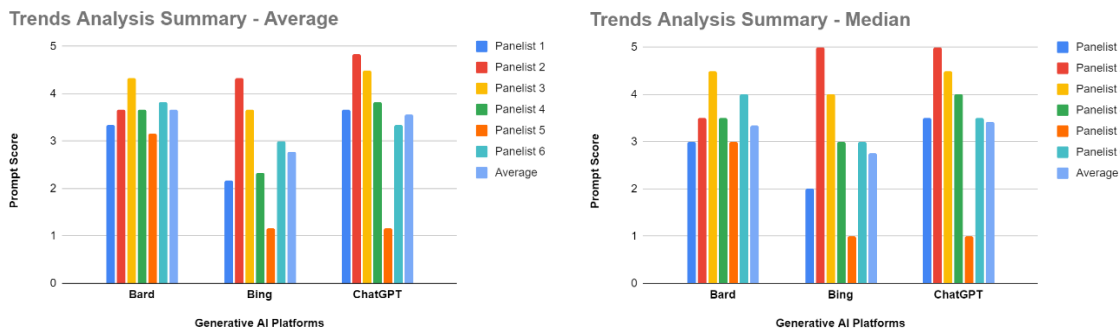
Trends Analysis Summary - Average

Trends Analysis Summary - Median

Figure 3. Trends Analysis Summary – Average and Median

The charts for the trends analysis show fluctuation between the three platforms in terms of quality of response. Bard and ChatGPT display similar averages with Bard being propped up by more consistent response form the panel. Panelist 4 rated ChatGPT poorly for its trend's responses lowering the platform's total average. The median scores, however, show ChatGPT outperforming the competition with higher scores. Both platforms similarly provided results that the panel found partially answered the prompt but again, no platformed displayed proficiency in response. Compared to the prior two analytical tasks, Bing underperformed in trends identification, receiving lower average and median scores. This was the first task where Bing, or any platform, provided sources alongside prompt responses so adding potential confidence behind the response was not enough to prop up the platform.
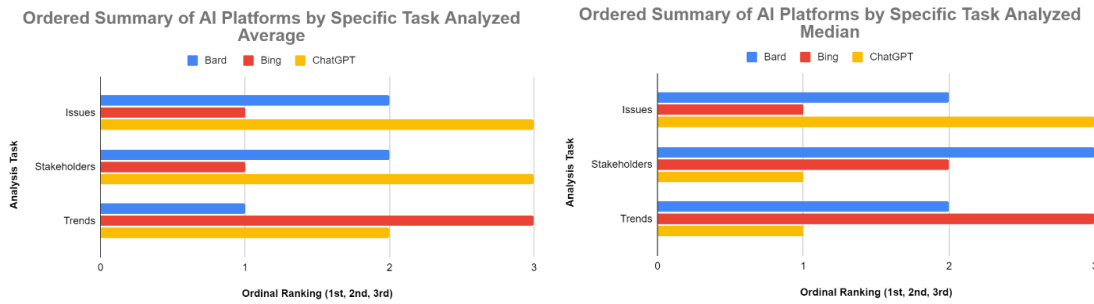
Figure 4. Ordered Summary of AI Platforms by Specific Task Analyzed

Figure 4 shows the rankings for the platforms when using both average and median across all analysis tasks. Bing ends up being the strongest platform when using average scores while ChatGPT ends up as the leader when using median scores. Bard comes across as the most consistent platform between the two methods of ranking.

Generative AI Platform Takeaways Based on Existing Research

Looking back strictly at the models that power Bard, Bing and ChatGPT, then it could be inferred that Bing would be the most impressive platform because of its level of training on GPT-4. ChatGPT should have been the worst performing platform given the limitations of GPT-3.5 with Bard performing in the middle. This ended up being the case when using average scores as the method of ranking but reverses with the median scores. It can be interpreted that the model used did not provide a marginal difference in result. Additionally, all three models struggled to provide responses at a high level with total averages in the low 3's across all analysis tasks. These results show that users should not trust solely in the number of parameters or training data for a generative AI model when making a decision on which platform to use. Based solely on those data points, Bing should have exponentially outperformed both Bard and ChatGPT, which did not occur when completing these tasks.

One of the benefits of generative AI for competitive intelligence tasks is supposed to be timesaving. While the responses from each platform to the prompts are very quick, between 5-30 seconds, the task of developing prompts and properly evaluating the findings can be time intensive. This research accumulated many hours developing prompts for the three analysis tasks. In a similar amount of time, a trained CI analyst could have completed one or more of the analyses without the need for generative AI. Additionally, prompt development is an evolving process and the prompts used for this analysis may be more or less effective in the future with evolving changes to AI platforms. An analyst looking to use generative AI will need to study which AI platforms are most suited to the given task, practice how to develop prompts, how to refine prompts based on responses, evaluate the accuracy and relevancy of responses and learn how to train the AI platform based on given results. The potential timesaving from this process also can be outweighed by the loss in reliability. Current platforms during this research were only able to partially answer the question and none were able to generate client-ready responses or analyses. Additional human intervention would be required to meet that standard.

Discussion

In response to the research objectives, generative AI displayed the ability to generate CI analyses to a partial completion. They performed worse compared to the perspective of human analysts and did not meet the expectation of crafting client-ready responses. The platforms were able to quickly generate responses, but significant time was required to craft prompts that could be used to mirror the analysis questions relative to each task. Additionally, the responses crafted displayed evidence of biases, particularly towards a U.S. centric perspective. One example of this was evident during the stakeholder analysis where no platform identified anything related to

China as a potential stakeholder. Given that one of the central pieces of the Chips act was the restrictions towards China, it was surprising to not see this mentioned at all in the prompt responses. It would be beneficial for future researchers to test these prompts using a Chinese generative AI platform to see if it demonstrates the same or similar biases. Another consideration is that the AI platforms are censoring out the Chinese information.

Generative AI platforms in their current state should be seen as tools to assist human analysts with analyses. They are effective at identification of non-conceptual areas such as stakeholders and display the ability to identify trends. An analyst could use this as a starting point to assist their work. However, they performed poorly in issues identification. Additionally, the platforms struggled to discern concepts separating what makes an issue, stakeholder, or trend more or less important to a stakeholder with poor scores to the respective questions with prompts 3 and 4 of each analysis.

Another issue faced is reliability. The platforms performed poorly responding to prompt 5 which was meant to demonstrate reliability in estimating outcomes of the analysis. Based on the prompts used, the platforms lacked the ability to provide sources and references to back up their answers. The one time a platform (i.e. Bing for the trends analysis task) did provide references, the panel found the responses to be poor to where they analyst did not perceive the added benefit.

One additional area to consider is that, when comparing to potential human analysts, AI only has access to an unspecialized dataset. Humans have access to data beyond open source or what is available in the training data behind an AI. This discrepancy in availability of data can affect performance for these analyses. The AI could perform better if trained on these specific analyses as well. This would require significant data to accomplish and would not be available in the entry level versions used for this research.

The platform chosen also played a role with respect to the quality of responses to each analysis type. These platforms will continue to evolve meaning that one platform may outperform on a certain task but could fall behind in a hurry with changes to the model or training. Prudent analysts will have to do regular due diligence to the model, platform and training they choose.

Evaluation From SCIP 2024 Workshop: Human vs GenAI Analysis Competition
Following the conclusion of the initial study in December 2023, the generative AI platforms underwent changes to large language models. Google Bard went from the PaLM2 model to a model known as Gemini (Pichai & Hassabis, 2023). Google rebranded Bard to Google Gemini as well. Additionally, Microsoft's BingAI rebranded to Microsoft Copilot. Copilot claims to run multiple AI models using a transformer-based natural language understanding (NLU) model (Microsoft, 2024). According to OpenAI release notes, there were no changes noted for the 3.5 model that was used for testing in the initial study (OpenAI, 2024).

During the annual Strategic Consortium of Intelligence Professionals (SCIP) annual conference in Nashville, the authors of this study were able to perform on updated test of these genAI platforms during a four-hour workshop session (Fleisher & Tao, 2024). Attendees engaged with Copilot, Gemini, ChatGPT 3.5 and 4 as well as Anthropic AI to see how well they could answer questions to build landscape, issues, stakeholder and trends analyses. The questions from this workshop can be found in Appendix 4. Attendees had approximately forty-five minutes for each analysis. Attendees then regrouped and discussed how well each platform performed and gave a verbal indication as to which platform performed the best. During this testing, Copilot proved to provide the best responses with attendees impressed that it provided sources and answered the most questions compared to its competition. Anthropic performed the worst, unable to provide responses to most questions and when it did provide an answer, the answer was considered poor quality. A key learning from this workshop was how to engage with prompt engineering. Attendees found that changing a single word around from a prompt would lead to different responses from the AI platforms. Small changes in wording could significantly affect the quality of responses. Additionally,

attendees tested how positive reinforcement affected a genAI's answer. Bribing the AI with the prospect of money or stressing urgency of response also changed the quality of the answer from each platform. The authors of this study would recommend additional testing of these methods to determine more specifically the level of change in quality from each platform. A future study would be encouraged.

Compared to the original study, the changes to Copilot enabled it to perform better than Gemini. ChatGPT in its 3.5 version continued to perform inadequately. The 4.0 version performed to a higher level with attendees comparing its performance to that of Gemini while still ranking below Copilot. Attendees came away from the workshop disappointed by the overall quality of responses from all genAI platforms, initially having higher expectations based on the way they are marketed. This was most attendees first exposure to genAI platforms while a couple attendees had experience using it for their work. This workshop showed that although the models underwent changes, and more were tested, they still were unable to perform competitive intelligence analyses to a level that could be used to meet a prospective client's expectations.

Comparisons with GenAI platforms in China

In March 2024, a parallel pilot study was conducted with two groups of undergraduate business students at South West University of Finance and Economics (SWUFE) in Chengdu, totaling 102 participants. This study replicated the IST (Issues, Stakeholders, Trends) analysis prompts from initial experiments. During a three-hour session, students were divided into six groups, each using different generative AI platforms: ChatGPT 4.0 by OpenAI, Ernie by Baidu, Tongyi by Alibaba, Xunfei Xinghuo by iFlytek, Kimi by Darkside of the Moon, and a human team for baseline comparison.

Preliminary results indicated that ChatGPT 4.0 outperformed the other platforms in effectively defining, identifying, and prioritizing issues, stakeholders, and trends. It was followed by Ernie, Kimi, Xunfei Xinghuo, and Tongyi, in that order. The human teams showed lower productivity,

likely due to their unfamiliarity with the structured approach of IST analysis. A contributing factor to the AI performance disparity could be the CHIPS Act's focus, predominantly discussed in English literature, which may have disadvantaged the Chinese platforms not extensively trained on English datasets.

The inclusion of these Chinese AI platforms offers valuable comparative insights into how regional technologies measure against global standards like ChatGPT 4.0, highlighting the influence of linguistic and cultural contexts on AI efficacy in competitive intelligence tasks. Future research will aim to further explore:

1. The ability of each AI platform to accurately perform CI tasks within the Chinese linguistic and business environment.
2. Performance differences between AI platforms developed in China and those in the West.
3. Student perceptions of AI effectiveness and their preferences across different platforms.

These findings will help delineate AI functionalities across linguistic and cultural boundaries and enhance our understanding of generative AI's global applicability and limitations in competitive intelligence roles.

Conclusion

Examining the panel's response to the generative AI platform analyses shows that these platforms are not yet capable of performing to the level analysts in these tasks. All platforms ended up performing similarly to each other, despite the differences or advancements in each model. Additionally, platform performance was driven by task given where certain platforms showed better or worse performance based on the task. These generative AI platforms performed to the level of partially completing each task, which can be used as a starting point for prospective analysts.

It is important to note that analysts should not take generative AI responses as credible without performing due diligence as these platforms failed to provide references to strengthen credibility of results. Analysts should also use caution when performing global tasks since there was evidence of bias

against certain countries in this research. Due to these factors, the platforms should be used as tools to add to an analysts' toolset. By continuing to use and learn each platform, analysts can also develop skills in prompting to evolve their skillset which can help derive more value from generative AI.

For researchers, these findings should be viewed as a first step towards understanding how generative AI can be used for analyses. Conclusions can be drawn from this data, but no definitive answers can be found. This type of study should be temporal to better understand how the platforms and models develop since they are developing at a rapid pace. It is recommended that future researchers repeat this study with a larger panel of experts that include those from diverse backgrounds and lived experience. By having a larger, more diverse panel of experts, it will provide more information at less risk of bias to understand generative AI performance. Additionally, future researchers should consider broadening the platforms to test. While Bard, Bing and ChatGPT will continue to evolve, additional market entrants could accelerate this and provide added competition. These platforms could also be compared over time to see which models are improving for analytical tasks and at what rate it is happening.

For practitioners, this research provides an overview for how generative AI can be used to support analysis tasks. While this research shows that it is not recommended to replace human output, it is capable of being a useful tool to assist said output. Prudent analysts will need to continuously test and evaluate a multitude of platforms before deciding which to use for a given task as each platform has its own set of strengths and weaknesses. These are not stagnant and are likely to change over time. It is also recommended that analysts use these tools and avoid shying away from them. Testing generative AI performance for a variety of analyses will give an indication for how they perform while also allowing the analyst to gain experience in AI use, prompt development and in understanding strengths and weaknesses of the platforms. Given the rapid advancement of these platforms, analysts who fail to learn how to use them may end up falling behind their peers.

Generative AI technology in its current form shows a path toward being a valuable tool in CI. The advancements in this field and rapid pace of development provide analysts with growing options to learn and implement them into their tools and skillsets. While it is unknown if these platforms can ever replace human analysts, they can complement them in their work making analysts more efficient. Adapting to this technology as it develops and evolves should be in the interest of any analyst looking to stay ahead of the competition.

REFERENCES

Baidoo-Anu, D. and Ansah, L. (2023, September 12) *Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning* Journal of AI https://dergipark.org.tr/en/download/article-file/3307311

Bass, D. (2023, January 23) *3.5* Bloomberg https://www.bloomberg.com/news/articles/2023-01-23/microsoft-makes-multibillion-dollar-investment-in-openai?embedded-checkout=true

Bensoussan, B. E., & Fleisher, C. S. (2012). *Analysis without paralysis: 12 tools to make better strategic decisions.* Ch. 13, Financial Times Press.

Cekuls, A. (2022) *AI-Driven Competitive Intelligence: Enhancing Business Strategy and Decision Making* Journal of Intelligence Studies in Business https://web-p-ebscohost-com.proxy.library.georgetown.edu/ehost/pdfviewer/pdfviewer?vid=0&sid=4326fb3b-011a-4dd7-8663-9e2b0b2b4363%40redis

Coulter, I, Elfenbaun, P., Jain, S and Jonas, W. (2016, January 7) *SEaRCH™ expert panel process: streamlining the link between evidence and practice* National Library of Medicine https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4704387/

Emmanuel, C. (2023, August 3) *GPT-3.5 and GPT-4 Comparison* Medium https://medium.com/@chudeemmanuel3/gpt-3-5-and-gpt-4-comparison-47d837de2226#:~:text=With%201.5%2

0trillion%20parameters%20compared, structures%20and%20provide%20effective%20replies.

Faguy, A. (2023, March 21) *Google Releases Bard AI Chatbot Amid Competition With ChatGPT* Forbes https://www.forbes.com/sites/anafaguy/2023/03/21/google-releases-bard-ai-chatbot-amid-competition-with-chatgpt/?sh=5fe13bbe32b7

Feller, I. (2013) *Handbook on the Theory and Practice of Program Evaluation,* Edward Elgar Publishing https://books.google.com/books?hl=en&lr=&id=p4BQdisrW9kC&oi=fnd&pg=PA115&dq=benefits+of+an+expert+panel+for+qual+research&ots=jrFz1doqx7&sig=FBYFJVU-qqzRN-HUqU_3HuXAQ7g#v=onepage&q=benefits%20of%20an%20expert%20panel%20for%20qual%20research&f=false

Feuerriegel, S. Hartmann, J. Janiesch, C. & Zschech, P. (2023, September 12) *Generative AI* Springer https://link.springer.com/article/10.1007/s12599-023-00834-7

Finkenstadt, D., Eapen, T., Sotiriadis, J., Guinto, P. (2023, November 30) *Use GenAI to Improve Scenario Planning* Harvard Business Review https://hbr.org/2023/11/use-genai-to-improve-scenario-planning

Fleisher, C. (2022) *Trendsetting in competitive intelligence: Discovering your LIST* Special Libraries Conference Annual Conference, Charlotte, NC.

Fleisher, A. & Tao, T. (2024, April 22) *Human VS. Gen AI Analysis Competition* SCIP https://www.scip.org/page/intellicon-us-nashville-program-2024#:~:text=Human%20VS.%20Gen%20AI%20Analysis%20Competition

Hargrave, M. (2023, May 11) *Standard Deviation Formula and Uses vs. Variance* Investopedia https://www.investopedia.com/terms/s/standarddeviation.asp

Hoffman, F. & Freyn, S. (2022, November 25) *Competitive intelligence in an AI world: Practitioners' thoughts on technological advances and the educational needs of their successors* Journal of Intelligence Studies in Business https://web-p-ebscohost-com.proxy.library.georgetown.edu/ehost/pdfviewer/pdfviewer?vid=0&sid=9b1a4c2d-71b6-4965-9e41-f1304e1e9c54%40redis

Intel, (2017) *Artificial Intelligence The Public Policy Opportunity* Intel https://community.intel.com/legacyfs/online/files/Intel-Artificial-Intelligence-Public-Policy-White-Paper-2017.pdf

Joost N. Kok, Egbert J. W. Boers, Walter A. Kosters, and Peter van der Putten (2002) *Artificial Intelligence: Definition, Trends, Techniques and Cases* Encyclopedia of Life Support Systems https://www.eolss.net/Sample-Chapters/C15/E6-44.pdf

Khodyakov, D., Hempel, S. Rubenstein, L. Shekelle, P. Salem-Schatz, S. O'Neill, S. Danz, M. and Dalal, S. (2011, December 23) *Conducting Online Expert panels: a feasibility and experimental replicability study* Bio Med Central https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-11-174

Lim, W. Gunasekara, A. Pallant, J. L. Pallant J. I. Pechenkina, E. (2023, July) *Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators* Science Direct https://www.sciencedirect.com/science/article/pii/S1472811723000289#bib42

Mannuru, N. R., Shahriar, S., Teel, Z. A., Wang, T., Lund, B. D., Tijani, S., Pohboon, C. O., Agbaji, D., Alhassan, J., Galley, J., Kousari, R., Ogbadu-Oladapo, L., Saurav, S. K., Srivastava, A., Tummuru, S. P., Uppala, S., & Vaidya, P. (2023). *Artificial intelligence in developing countries: The impact of generative artificial intelligence (AI) technologies for development* Information Development https://www.researchgate.net/profile/Lydia-Oladapo-3/publication/373934747_Artificial_intelligence_in_developing_countries_The_impact_of_generative_artificial_intelligence_AI_technologies_for_developme

nt/links/6504d6dea69a4e63180fcb33/A rtificial-intelligence-in-developing-countries-The-impact-of-generative-artificial-intelligence-AI-technologies-for-development.pdf

Marrm B. (2023, August 16) *What Does Generative AI Mean For Websites And SEO?* Forbes https://www.forbes.com/sites/bernard marr/2023/08/15/what-does-generative-ai-mean-for-websites-and-seo/?sh=f416f692af94

Martinez, R. (2019, May 28) *Artificial Intelligence: Distinguishing Between Types and Definitions* Nevada Law Journal https://scholars.law.unlv.edu/cgi/viewc ontent.cgi?article=1799&context=nlj

McCarthy, J. (2004, Nov. 12) *What is Artificial Intelligence?* Stanford https://www-formal.stanford.edu/jmc/whatisai.pdf

Mehdi, Y. (2023, May 4) *Announcing the next wave of AI innovation with Microsoft Bing and Edge* Microsoft https://blogs.microsoft.com/blog/2023/0 5/04/announcing-the-next-wave-of-ai-innovation-with-microsoft-bing-and-edge/

McGill, (2023) *Qualitative or Quantitative Research?* McGill Qualitative Health Research Group https://www.mcgill.ca/mqhrg/resources /what-difference-between-qualitative-and-quantitative-research

Microsoft, (2024, April 2) *AI features for Teams and Classic copilots* Learn Microsoft https://learn.microsoft.com/en-us/microsoft-copilot-studio/advanced-ai-features

Mok A. & Zinkula J. (2023, June 4) *ChatGPT may be coming for our jobs. Here are the 10 roles that AI is most likely to replace* Business Insider https://www.businessinsider.com/chatg pt-jobs-at-risk-replacement-artificial-intelligence-ai-labor-trends-2023-02

Narang, S., Chowdhery, A., Google Research Engineers, (2023, April 4) *Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance* Google Research

https://blog.research.google/2022/04/pa thways-language-model-palm-scaling-to.html

OpenAI, (2022, November 30) *Introducing ChatGPT* OpenAI https://openai.com/blog/chatgpt

OpenAI, (2023, March 14) *GPT-4* OpenAI https://openai.com/research/gpt-4

OpenAI, (2023) *Fine Tuning* OpenAI https://platform.openai.com/docs/guide s/fine-tuning

OpenAI, (2024) *ChatGPT – Release Notes* OpenAI https://help.openai.com/en/articles/682 5453-chatgpt-release-notes

Pasichnyk, O & Strelkova, O. (2022) *Three Types of Artificial Intelligence* Khmelnitsky National University http://eztuir.ztu.edu.ua/bitstream/han dle/123456789/6479/142.pdf?sequence =1&i

Peres, R. Schrier, M. Schweidel, D. Soescu, A. (2023, June 2) *On ChatGPT and beyond: How generative artificial intelligence may affect research, teaching, and practice* Science Direct https://www.sciencedirect.com/science/ article/pii/S0167811623000162

Pichai, S. & Hassibis, D. (2023, December 6) *Introducing Gemini: our largest and most capable AI model* Google https://blog.google/technology/ai/google -gemini-ai/#sundar-note

Pichal, S. (2023, February 6) *An important next step on our AI journey* Google https://blog.google/technology/ai/bard-google-ai-search-updates/

Rajaraman, V. (2014, April 30) *John McCarthy — Father of artificial intelligence* Springer https://link.springer.com/article/10.100 7/s12045-014-0027-9

Reimann, N. (2023, January 5) *ChatGPT Creator OpenAI Discussing Offer Valuing Company At $29 Billion, Report Says* Forbes https://www.forbes.com/sites/nicholasr eimann/2023/01/05/chatgpt-creator-openai-discussing-offer-valuing-company-at-29-billion-report-says/?sh=7b4563be11e0

Russell, & Norvigg, (2022, Aug. 22) *Artificial Intelligence: A Modern Approach, 4th*

56

US ed Pearson
https://aima.cs.berkeley.edu/

Sahoo, S. Kumar, S. Abedin, M. Lim, W. Jakhar, S. (2023, January 27) *Deep learning applications in manufacturing operations: a review of trends and ways forward* Journal of Enterprise Information Management https://www.emerald.com/insight/content/doi/10.1108/JEIM-01-2022-0025/full/html

Tran, H & Murphy, P. (2023, September 1) *Generative artificial intelligence and entrepreneurial performance* Emerald Insight https://www.emerald.com/insight/content/doi/10.1108/JSBED-09-2023-508/full/html

Verma, P. & De Vynck, G. (2023, June 2) *ChatGPT took their jobs. Now they walk dogs and fix air conditioners.* Washington Post https://www.washingtonpost.com/technology/2023/06/02/ai-taking-jobs/

Wang, P. (2019, August 12) *On Defining Artificial Intelligence* Sciendo https://sciendo.com/downloadpdf/journals/jagi/10/2/article-p1.pdf

Wankhede, C. (2023, December 7) *What is PaLM 2: Google's large language model explained* Android Authority https://www.androidauthority.com/what-is-google-palm-2-3331329/

Wodecki, B. (2022, July 27) *7 Language Models You Need to Know* AI Business https://aibusiness.com/nlp/7-language-models-you-need-to-know

Zhang, P. & Boulos, M. (2023, August 24) *Generative AI in Medicine and Healthcare: Promises, Opportunities and Challenges* MDPI https://www.mdpi.com/1999-5903/15/9/286.