AI Applications in Academic Collaboration: A Case Study of the Venice Project Center





An Interactive Qualifying Project submitted to the faculty of WORCESTER POLYTECHNIC INSTITUTE in partial fulfillment of the requirements for the Degree of Bachelor of Science

Sponsored by: SerenDPT, Alireza Dehghan (Liaison) Submitted to the Venice Project Center and Worcester Polytechnic Institute Advised by: Francis Leahy and Ruth McKeogh

Authored By

Mina Boktor Krishna Garg Lucas Lamenha James Walden Date: 12/10/2024

Alias: v24b.ai@gmail.com Website: <u>https://sites.google.com/view/v24b-ai/home</u>

This report represents the work of WPI undergraduate students submitted to the faculty as evidence of completion of a degree requirement. WPI routinely publishes these reports on its website without editorial or peer review. For more information about the projects program at WPI, please see https://www.wpi.edu/academics/undergraduate

Abstract

The project explored the application of Artificial Intelligence (AI) to aid future projects conducted by the Venice Project Center (VPC) improving accessibility to project data and impacts. The team created a website and worked to identify the most viable AI tools available to the VPC and conducted user testing to assess usefulness. Four functional prototypes were developed, along with the compilation of already existing tools that could be useful to future VPC project groups. Results of user testing were utilized to recommend future expansion of the website and potential applications that may be feasible in coming years.

Acknowledgments

We want to express our deepest gratitude to our project advisors Ruth McKeogh and Francis Leahy who played a critical role in guiding the direction of our project. We also want to thank our project sponsor SerenDPT, specifically Alireza Dehghan who gave us insightful feedback on the development of our website and various Artificial Intelligence (AI) tools and assisted us with overcoming technical hurdles we encountered.

We also thank everyone who participated in our user study which helped assess the usefulness of our application and identify useful ways AI could be utilized in future development of the website to provide further functionality to future project groups.

Executive Summary

The Venice Project has been operating for over thirty years and has sponsored over 300 student-led projects covering numerous aspects of Venetian society including everything from infrastructure to economy and lifestyle. These projects have had numerous significant impacts on the Venetian community and the accumulations of a number of deliverables including student project reports, software applications, and open data that can be accessed through the VPC Website. Unfortunately, drawing meaningful information about the VPC's impacts is challenging given the lack of an interface to interact with the data and project reports. The project reports can be anywhere from fifty to one-hundred pages in length, making it unlikely for an interested party to read the entire report especially from the view of a non academic source. The data compiled by the VPC are also primarily in the form of comma separated value (CSV) files, which is how excel sheets are formatted, with hundreds to thousands of lines of data which cannot be analyzed without a third party tool or manual analysis of the files which is impractical and limits accessibility. Compiling the various reports generated into a more digestible format coupled with the data management issues has presented the VPC with a time-consuming and labor-intensive problem.

The VPC could circumvent these hurdles by investigating and implementing the various AI tools available and educating its sponsored students on the effective and ethical use of AI. This project aims to identify various publicly available AI tools that could be integrated into the VPC's current workflow and implement functional prototypes that future project groups can utilize to enable students to have a greater impact on the Venetian community. To achieve this goal, the team completed the following objectives:

- 1. Explore the use of AI for the VPC's research
- 2. Produce functional prototypes of various AI tools for the VPC
- 3. Examine how AI can be utilized in future research for the VPC

To experiment with possible AI tools that could apply to the VPC, we first had to discover what AI tools were available. We evaluated a number of tools basing our decisions on popularity and accessibility. We evaluated a number of tools on a variety of metrics to categorize and identify the tools capable of enabling us to create our functional prototypes which we would

then store in a table for comparison. After creating our base website, we then engaged in a workflow consisting of identifying an opportunity to implement AI for the VPC, beginning the design and development of the prototype, periodically assessing the feasibility of the prototype, and finally presenting our findings to our sponsor SerenDPT. We continued this production cycle for several iterations allowing us to explore the practicality of several tools without dedicating excessive time to a single tool. After the completion of our previously stated goals and final polishing of the website was complete, the group conducted a user study of current VPC students to see how the website could be useful for their own projects, and what tools they'd like to see in future iterations.

By following the production cycle, the team was able to produce a total of five tools to be deployed on the website which was developed with ease-of-use and aesthetically pleasing designs in mind. The first tool formed the foundations for the tools to follow. A table was developed capable of comparing Large Language Models (LLM) by gathering a number of metrics typically available via an LLM's technical specifications. The table allows the user to sort LLM's by any of the metrics measured in the table. We then used the table in the assessment of models in the production of our functional prototypes.

For our second tool we decided on the development of a chatbot capable of responding to VPC-specific questions. Our initial plan for the chatbot was to feed the LLM project reports as context for a user's query to provide an accurate response. Using our evaluation tool, we decided to develop the tool using Gemini 1.5 Flash, a Large Language Model developed and maintained by Google, due to its ability to handle file uploads, significantly larger number of input tokens, and free usage. By vectorizing project descriptions provided by the VPC project report master sheet, we were able to design a system to retrieve the most closely related project reports to the user's query. Though this system was successful we quickly realized several flaws with our approach mostly related to excessive response time. We then reworked our design to offer the user the option to select one of three recommended projects to ask specific questions about. By only uploading one file at a time we were able to significantly improve response time and accuracy, and still provided relevant projects utilizing our vector database.

Our third tool was a chatbot aimed at more accurately answering VPC related questions from multiple different reports instead of just one. The approach involved parsing the project reports into two-paragraph chunks of text, with approximately two sentences of overlap to ensure coherence, which would then be stored in our existing database to later have a semantic search performed. This would allow multiple snippets of useful information to be pulled from various different project reports before being sent as context for a user's question allowing for quicker analyzation of the provided project report chunks and more accurate responses.

The fourth tool we developed was a chatbot that utilized the VPC's CSV data, offering precise answers to Venice-related questions beyond the scope of the report chatbots. This is done by storing the raw CSV data in a database and querying it using code written by the AI tool used (Google Gemini 1.5-Pro). The results of the query are then returned to the user as a chatbot message. For example, if the user asks: "What is the widest fountain in Venice?," the chatbot queries the database and returns the name and the width of the widest fountain in the data provided. The user is then able to further ask about this fountain, such as the sestiere (neighborhood) in which the fountain is located, or its height, whether it is active or not, and more.

Our final tool was a podcast generator capable of creating podcasts up to ten minutes long by sending a pdf file to ChatGPT's "tts-1" model. A transcript is provided and styled along with the podcast for the user to provide along with it. The podcast is automatically downloaded onto the user's local computer. The podcasts generated go into detail about the file provided, which must be a pdf, and provide key summaries, insights, and details about the subject matter. It can be used to generate summary podcasts of all the VPC student project reports to be published on their website to reformat lengthy reports into a more digestible format for the layman. Unfortunately, this tool does incur a fee of 15 to 20 cents upon usage, and could easily be exploited to generate excessive charges to SerenDPT, so it was omitted from the final website and access was only be given to SerenDPT staff.

User testing was then conducted on a sample of ten students currently working at the VPC, covering two members from each project group currently conducting research on various topics. Results of user testing highlighted various bugs in the three chatbots created and various features students would like to see in future iterations. The greatest problem encountered was undeniably the lack of proper organization and faulty data held within the CSV files. It was also apparent that for the chatbot utilizing project reports further decomposition of the reports themselves were necessary along with more direction in what data should be sent to the AI model to obtain an appropriate response.

This project demonstrated the potential for integrating artificial intelligence into the VPC's workflow, focusing on improving data accessibility, usability, and impact assessment. Our findings underline the importance of AI in automating repetitive tasks, improving user interaction, and expanding the VPC's impact. Although challenges remain, such as chatbot response times and the technical complexity of integrating AI systems, the tools developed during this project provide a strong foundation for future work. We recommend future groups optimize the scalability of existing tools to improve response time and expand functionality to enable the analysis of a wider variety of data sets. We recommend continuing the exploration of AI models while incorporating more extensive feedback mechanisms to be incorporated into future website iterations.

By building on the groundwork established in this project, the VPC can further harness AI to enhance its contributions to the Venetian community, ensuring sustainable and impactful use of its data and research.

Authorship:

Section	Primary Author(s)
Abstract	James Walden
Executive Summary	James Walden
1.0 Introduction	James Walden
2.0 Background	James Walden
2.1 Artificial Intelligence and Machine Learning	Lucas Lamenha
2.1.1 Tokenization	Lucas Lamenha
2.1.2 Ethical Concerns of Large Language Model Usage	Mina Boktor
2.1.3 LLM Hallucinations and Information Validation	Lucas Lamenha
2.1.4 Query-Based Retrieval Augmented Generation	James Walden
2.2 VPC Projects	James Walden
2.3 Modern Behavior of Content Consumption	Krishna Garg
3.0 Methodology	Krishna Garg, Mina Boktor
3.1 Explore the use of AI for the VPC's research	Lucas Lamenha, Mina Boktor
3.2 Produce functional prototypes of various AI tools for the VPC	James Walden
3.3 Examine how AI can be utilized in future research for the VPC	Krishna Garg
3.4 User Testing	James Walden, Krishna Garg
4.0 Findings	James Walden
4.1 Initial Findings	James Walden
4.2 Large Language Model Evaluation Tool	James Walden, Krishna Garg
4.3 Report Chatbot	James Walden
4.4 Podcast Generator	Krishna Garg
4.5 CSV Chatbot	Lucas Lamenha
4.6 User Testing	James Walden
5.0 Conclusions	Mina Boktor
6.0 Recommendations and Future Work	Lucas Lamenha, Mina Boktor

Table of Contents

Abstract	2
Executive Summary	4
Authorship:	9
Table of Contents	10
List of Figures	11
1.0 Introduction	12
2.0 Background	14
2.1 Artificial Intelligence and Machine Learning	14
2.1.1 Tokenization	16
2.1.2 Ethical Concerns of Large Language Model Usage	17
2.1.3 LLM Hallucinations and Information Validation	20
2.1.4 Query-Based Retrieval Augmented Generation	22
2.2 VPC Projects	23
2.3 Modern Behavior of Content Consumption	25
3.0 Methodology	
3.1 Explore the use of AI for the VPC's research	29
3.2 Produce functional prototypes of various AI tools for the VPC	34
3.3 Examine how AI can be utilized in future research for the Venice Project Center	37
3.4 User Testing	39
4.0 Findings	40
4.1 Initial Design	40
4.2 Large Language Model Evaluation Tool	50
4.3 Report Chatbot	51
4.4 Podcast Generator	54
4.5 CSV Chatbot	56
4.6 User Testing	58
5.0 Conclusions	61
6.0 Recommendations and Future Work	62
References	64
Appendix A - Evaluation Metrics:	67

List of Figures

Figure 2.1.3.1: LLM Validation Technique	
Figure 3.1.1: Number of LLMs on (https://hugginface.co/models)	
Figure 4.1.1: Initial Landing page	43
Figure 4.1.2: Evaluations Page	
Figure 4.1.3: Add Model Page	44
Figure 4.1.4: Add Model Page (Form)	45
Figure 4.1.5 Tools Gallery	
Figure 4.1.6: Chatbot	
Figure 4.1.6.1: Chatbot with Recommendation and Query	
Figure 4.1.6.2: Chatbot with Recommendation and Query Output	
Figure 4.1.7: CSV Chatbot	
Figure 4.1.7.1: CSV Chatbot with Query and Response	
Figure 4.1.8: Report Chatbot	
Figure 4.1.8.1: Report Chatbot Query and Response	
Figure 4.1.9: Podcast Generation Page	
Figure 4.1.9.1: Podcast Generation Loading	
Figure 4.1.9.2: Podcast Generation Output	
Figure 4.1.10: Recommendations Page	
Figure 4.1.11: About Us Page	51

1.0 Introduction

The Venice Project Center (VPC) has sponsored the creation of over 300 student-led projects addressing a wide array of social, environmental, and infrastructural challenges the Venetian community faces. These projects have generated significant findings, insights, and practical deliverables that have contributed to the better understanding and possible resolution of issues such as over-tourism, environmental degradation, and preserving Venice's unique cultural heritage. Despite the success of these various initiatives, the long-term accessibility of the data produced and dissemination of the impacts of these projects has been an ongoing challenge. All the data produced must be manually updated or transferred to new platforms to preserve the outcomes over the years. Compiling the various reports generated into a more digestible format coupled with the data management issues has presented the VPC with a time-consuming and labor-intensive problem.

The VPC could circumvent these hurdles by investigating the various AI tools available and educating its employees and sponsored students on the effective and ethical use of AI. Researchers suggest that while AI systems lack an understanding of the truth value of text, oversight by human experts can mitigate inaccuracies and ensure proper validation of AI-generated material (Hosseini et al., 2023). When balanced with human oversight and responsible implementation, these concerns can help address skepticism and ethical considerations tied to AI in content creation.

This project aims to identify various publicly available AI tools that could be integrated into the VPC's current workflow and implement functional prototypes that future project groups can utilize to enable students to have a greater impact on the Venetian community. By identifying already existing tools that could be relevant to the VPC's objectives students will be well-equipped to utilize AI for the progression of their project's goal while maintaining academic integrity and improving productivity. Functional prototypes provide an array of tools relevant to the VPC capable of answering project-specific questions, generating media regarding the VPC, or analyzing sets of data. Due to the technical nature of this project, some general understanding of the basics of artificial intelligence is necessary.

2.0 Background

The background chapter explores an analysis of the Venice Project Center's objectives and student projects to understand what sort of data the AI tools handle and what deliverables are feasible. The chapter structure is be as follows: a brief overview of artificial intelligence, the stigmas and ethical concerns of using AI for generating content based on others' work, a basic understanding of artificial intelligence and its various considerations such as hallucinations and information validation, an overview of the VPC and its various resources, and finally an assessment of modern behaviors in regards to content consumption.

2.1 Artificial Intelligence and Machine Learning

McCulloch and Pitts (1943) first introduced the concept of Artificial Intelligence (AI) and Machine Learning (ML) as a means to attempt to explain how neurons work together to perform actions in a brain. A seminal mathematical neuropsychology study considerably predating the concept of computer science, this research would lay the foundations for what would become the basis for modern artificial intelligence. Conceptually, this study explained how neurons work in an "all-or-none" fashion, either sending signals or not. This was the basis for the simplest form of Artificial Intelligence possible – simple neurons that send signals or not if certain conditions are met, essentially making decisions based on the environment around them.

As a field, Artificial Intelligence was first proposed by McCarthy et al. (1955) with "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence," a publication that would culminate in the 1956 Dartmouth Conference, which is considered the "birth of Artificial Intelligence."

According to McCarthy et al. (1956):

An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer (p. 2).

After this, the authors would define problems that they would discuss at the following year's conference, such as automatic computers, how computers can be programmed to use a language, theoretical aspects behind neuron nets - now commonly called neuron networks, self-improvement - or, in modern terms, machine learning, and controlled randomness and creativity.

As per modern artificial intelligence, Russell and Norvig (2016) define AI as, "the study of agents that receive percept from the environment and perform actions." Generally, artificial intelligence agents can be categorized into three categories (Strelkova, 2017):

- Artificial Narrow Intelligence (ANI): AI that can decide in only one sphere, such as an agent that can play chess, but only chess.
- Artificial General Intelligence (AGI), or an AI that reaches and passes the intelligence level of a human, meaning it is able to reason, think abstractly, comprehend complex ideas, and learn from experience.
- Artificial Super Intelligence (ASI), is an AI that far surpasses human capabilities in every field, such as scientific creativity, general wisdom, and social skills.

At the present moment, humans have only been able to reproduce the lowest level of these models, ANIs.

It is very difficult to define precisely what artificial intelligence is without understanding the basis of how it came to be and what problem it is trying to solve. The use of AI models in the present day is a culmination of research done over the past 80 years. Comprehending the modern categories of artificial intelligence agents and their capabilities is pivotal before delving into any kind of research in the field.

2.1.1 Tokenization

Tokenization, or the process of dividing text into meaningful typographical units, called tokens, is pivotal in feeding text data into Artificial Intelligence models that recognize human language in the field of Natural Language Processing (NLP) (Michelbacher, 2013). The process works by breaking sentences into smaller fragments that the AI model can process into data it can understand. This is done to maintain the context of the word. For example, if the word

"don't" were divided into segments based on punctuation, there would be three tokens without much meaning: "don", " ` ", and "t". Instead, it is taken as two tokens: "do" and "n't," one conveying the idea of "doing" something, while the other conveys the negative (Mielke et al., 2021). This process means that for different languages, the number of words that a token represents differs. For example, the same tokenizer used by ChatGPT and GPT-4 uses 1.6 times more tokens for the Italian language than it does for English; and 3 times more for Arabic. Often, companies will base the pricing for their models on the amount of tokens needed to process data (OpenAI, 2024).

2.1.2 Ethical Concerns of Large Language Model Usage

The exponential growth that large language models (LLMs) have seen throughout recent years has raised ethical concerns regarding their use, especially in the academic field. As Vykopal states, "The capability to generate an arbitrary amount of human-like texts can be a powerful tool for disinformation actors willing to influence the public by flooding the Web and social media with content during influence operations" (pg #1). Academic writers and researchers are especially concerned about the use of LLMs as a form of cheating since many believe that the automation of research and writing presents a threat to the integrity of their work (Giray, L., 2024). LLM-generated work can also fall into a morally gray area since it is nearly impossible to identify and properly give credit to the authors of the text upon which the LLM was trained (Freyer, N. et al., 2024). This has led people to automatically perceive LLM-generated material negatively, even though it may be indistinguishable from human-written work (Harasta et al., 2024). While these are valid and rational concerns, LLMs can be used ethically to simplify work, automate tasks, and even as a content and text generator through a balance of transparency and LLM fine-tuning development.

Using LLMs to simplify the material for the layperson can significantly enhance language accessibility, especially for marginalized groups such as non-native speakers or individuals with cognitive disabilities. As Freyer highlights, "accessing language is an important support to individual autonomy and justice" (Freyer, N. et al., 2024). For those who face challenges in understanding complex language, the ability to simplify texts through LLMs could allow them to engage more fully in society. By reducing intellectual barriers, LLMs could help

ensure equality of opportunity (Freyer, N. et al., 2024) as accessible language is crucial for education and procedural knowledge. This use of LLMs promotes inclusivity and addresses ethical obligations to create a more just society where individuals can participate in public discourse.

While there is a clear net positive in disseminating LLM-generated text, there are still significant ethical risks. LLMs, trained on vast amounts of data, learn social biases, hallucinate, or oversimplify complex matters that could mislead vulnerable populations (Freyer, N. et al., 2024). Freyer emphasizes that the proneness of misunderstanding among certain groups raises the "risk of exploitation, abuse, and confusion." This is particularly concerning when LLMs generate easy-read text summaries or when simplifying large or complex written work, as inaccuracies or biases in these translations may disproportionately harm those with lower language comprehension skills (Freyer, N. et al., 2024). The potential for information loss also poses ethical challenges, particularly in contexts where precision is crucial. Without human oversight, these risks may lead to responsibility gaps, where no one is held accountable for the harmful effects of inaccurate or misleading outputs (Freyer, N. et al., 2024). It is the responsibility of the developer and publishers of LLM-generated material that the model is trained with diverse data and that all generated material is meticulously proofread (Freyer, N. et al., 2024).

A foundational step toward mitigating AI shaming is establishing transparency and accountability in AI usage. Giray says "Academic practitioners should openly declare their use of AI, fostering transparency and accountability. For instance: 'In preparing this work, I utilized ChatGPT 3.5 for outlining and proofreading. Subsequently, I carefully reviewed and edited the content to ensure accuracy and coherence. I take full responsibility for the integrity of this publication." (Giray, L., 2024) and then proceeds to include an acknowledgment section after their conclusion that states he used ChatGPT in the writing of their paper. By being transparent, publishers and writers can preemptively address concerns about the authenticity and reliability of AI-assisted research.

There is a delicate balance in using LLMs ethically. A major ethical concern with LLMs is the lack of transparency, so some suggest a total ban. Banning the use of LLMs entirely could never be enforceable and would only encourage undisclosed use, and as Hosseini et al. (2023) argue, "the most reasonable response to the dilemmas posed by LLMs is to develop policies that

promote transparency, accountability, fair allocation of credit, and integrity" (pg #13-14). Specifically, the inclusion of disclosure in the introduction of a methods section of a given LLM-generated paper as mentioned earlier. This balance between accessibility and safety is essential in ensuring that LLMs contribute positively to society without inadvertently causing harm to the very groups they aim to support.

It is possible to use LLMs effectively as long as humans review LLM-generated work to mitigate LLM biases and ensure that the original author's work does not get metaphorically lost in translation. Every instance of LLM-generated work must be made fully transparent to the reader for the ethical use of LLMs. Giray (2024) highlights that:

By embracing AI as a tool to augment human capabilities, being transparent about its use, and addressing ethical concerns, academia can lead the way in demonstrating responsible AI integration. This approach can help harness AI's potential to advance knowledge and innovation while maintaining the integrity and rigor of academic work, ensuring that technological progress enhances rather than diminishes the value of human scholarship. This exemplifies collaborative intelligence, where humans and AI work together to solve problems. (p #6)

Giray then states that AI shaming is counter-productive, and instead, reflecting on valid ethical concerns and implications as opportunities for learning and improvement is the better approach to addressing the stigmas surrounding AI-generated content. By taking a more proactive approach to using AI in written work, humans and AI can collaborate, achieving results that neither could accomplish on their own.

2.1.3 LLM Hallucinations and Information Validation

The use of large language models is promising, but not without its faults. Hallucination is a phenomenon that happens whenever an LLM fabricates data that appears to be fluent and seemingly coherent but is baseless (Lin et al., 2024). LLM hallucinations pose a problem to the credibility of LLM Chatbots and their usability in fields that require constant accurate answers. Figure 2.1.3.1 shows a method of answer validation has been proposed by Verspoor (2024):



Figure 2.1.3.1: LLM Validation Technique (Verspoor, 2024; Figure 1)

The method proposed in the study involves using multiple large language models to validate the data. In this possible solution, the first large language model produces the output to the question. After that, the different outputs are parsed through another LLM and used to group sets of possible output from the first LLM by semantic similarity. A third LLM is used to evaluate the first LLM's response with a human-written answer to determine its correctness. This validation method can be used to verify a specific model's general frequency of hallucination, what topics it is most likely to hallucinate in, and how much of a response is a product of hallucination.

Another of these methods, prompt engineering, can be used to reduce hallucinations (Tonmoy et al., 2024). Prompt engineering consists of experimenting with various instructions when requesting a specific output from an AI model. This method works by providing the model with more context on the topic and expected outcomes, which helps the model generate more accurate data. Overall, a combination of multiple validation and hallucination mitigation techniques can be applied to produce more reliable outputs from text-generating LLM models.

2.1.4 Query-Based Retrieval Augmented Generation

One of the most significant limitations of the progression of LLM-generated content is hallucinations due to difficulty identifying instances of fabricated data. Retrieval-augmented generation (RAG) has become an industry standard for addressing the problem of hallucinations and information accuracy by retrieving relevant information from available data stores to enhance the generation process with higher accuracy and increased robustness. Query-based

RAG is a specific implementation of RAG that stems " from the idea of prompt augmentation, query-based RAG seamlessly integrates the user's query with insights from retrieved information, feeding it directly into the initial stage of the generator's input" (Zhao, P. et al., 2024). Upon querying an LLM, query-based RAG would first perform a similarity search on the user's input into a data store to retrieve relevant data. It would then append the data to the user's initial input before sending it to an LLM as context for the user's initial query. The LLM would then generate a response using the retrieved data and return it to the user. By following these steps, query-based RAG systems can enhance the accuracy of LLM-generated content while allowing modular flexibility and being fairly straightforward to implement. Though it is effective in improving the quality of a variety of content generation models designed for various tasks it is not without its downfalls. These systems trade response time in favor of accuracy, as the querying of a database and data transition from the database to an LLM can often be taxing on the system and cause delays in response time. When implementing a query-based RAG system, it's imperative to keep in mind the size of data being sent to the LLM and minimize the turnaround time to ensure smooth operation from the user's point of view. Avoiding a situation where the user waits unnecessarily long after giving a simple query to the LLM is of utmost importance as users may question whether the model is working and may abandon the tool due to its inefficiency.

2.2 WPI Venice Project Center Research

Founded in 1988, the Venice Project Center (VPC) aimed to address some of the city's most pressing issues using state-of-the-art technology, on-site research, and various collaborative approaches involving local stakeholders and experts (Suyer, M. et al. 2022).

Like other WPI project centers, the VPC focuses on global issues through technology and societal impact on a local scale. Its efforts are concentrated on solving problems unique to Venice while contributing valuable knowledge and tools to local and global communities. Although much of the VPC's work is supported by Venetian organizations, it aims to share its findings with a wider audience by utilizing the latest technological advancements such as Artificial Intelligence and state-of-the-art information distribution strategies. Student project teams at the

VPC generate detailed reports, original databases, educational tools, and other deliverables provided to the VPC and its partners. The VPC maintains the technologies and data collected through its various projects to showcase its achievements and insights through accessible online resources such as the VPC Website, VPC Master spreadsheet, Venipedia, and DigitalWPI (Suyer, M. et al 2022).

The VPC hosts diverse projects to address different facets of Venetian culture, architecture, social issues, and lifestyle. The official VPC website sorts its various projects into the following categories: Technology & Science, Organizations & Startups, Infrastructure & Mobility, Environment & Energy, Economy & Society, and Art & History. Most of these projects focus on issues such as tourism, flooding, and preserving the cultural heritage of Venice. These projects have produced many web apps, interactive maps, and other tools as well as several presentations which they have publicly available via their website. Given Venice's historical and environmental uniqueness, the research conducted by the VPC addresses many areas that had not been systematically explored before. The challenges associated with over-tourism and environmental sustainability highlight the need for up-to-date data to help Venice preserve its infrastructure and address modern-day problems effectively.

The VPC has made profound impacts on Venice through its comprehensive and data-driven approach to local issues. By maintaining all of the technologies and data produced over the years, the VPC ensures that the work remains accessible for future research and decision-making. This ensures that local Venetian authorities and the international community can benefit from the insights generated. Furthermore, the VPC's projects have led to tangible changes in the management of local issues in environmental conservation, infrastructure, or cultural preservation. Many organizations now reach out to the VPC for data or to sponsor additional WPI projects, evident through the numerous publications featured on their websute, further emphasizing the center's significant role in the ongoing efforts to preserve Venice and its culture for future generations.

2.3 Modern Behavior of Content Consumption

While the wealth of information on Venice generated by the VPC is impressive, it is difficult for a single person to understand the extensive data and written material. A single IQP

report can range from fifty to sixty pages, and data sets can often be incomprehensible to the layperson. For example, "in a study conducted by Microsoft, the average attention span of teens and young adults in the year 2000 was 12 seconds. In 2016, the seconds reduced to eight; lower than a goldfish's" (Whalen, 2019). This shows the effect that digital and social media have had on the average person, reducing their attention span by 33% over sixteen years in which digital media had a massive uprising. With such a significant decrease in attention span, the ability to focus and accurately comprehend the contents of such long research reports has been significantly diminished. The social media landscape changed drastically with the introduction of short-form content almost a decade ago in 2016. According to *From Binge Watching to Binge Scrolling: The Impact of Short-Form Video on Our Attention Span*, short-form content induces a constant switching of attention between various tasks, and "constant switching of attention can lead to a decrease in focus and concentration, ultimately impacting our ability to retain information and engage in deep thinking" (Technology Dot Org, 2023).

According to *Reading Behavior in the digital environment: Changes in reading behavior over the past ten years,* "the younger generation growing up in the digital environment lacks the ability to read deeply and to sustain a prolonged engagement in reading" (Liu, 2005). Liu attributes the lack of ability to read a piece of writing in-depth to the boom of digital reading and the sheer quantity of articles and information available. This study was conducted in 2005 when digital media exposure was significantly less than modern standards. While it did exist social media was still in its initial stages so there was not yet as profound an effect on society, at least in comparison to recent years.

This study focused on the idea that digital media is mostly used to find quick information about topics, people look for keywords just to understand the surface level or get a quick answer, whereas in-depth reading is used to understand a topic in-depth and with clarity. The article references a 2003 study in which surveys were taken regarding reading practices. Students were asked under which circumstances they would use which format of reading, the formats being digital content that was printed, digital content that was not printed, and original printed content. "Eighty percent of students prefer to read a digital piece of text in print to understand the text with clarity" (Liu, 2005). Furthermore, "Nearly 68 percent of respondents report that they understand and retain more information when they read printed media" (Liu, 2005). Finally, the study states that "They note that undergraduate students who read online text find the text more difficult to understand, less interesting, and the authors less credible than those who read the printed version" (Liu, 2005).

Despite printed readings being obviously more understandable and objectively better at getting readers to consume text and understand it to some depth, according to the survey conducted in the article, why is it that people still turn to faster forms of media to get quick inputs of information or simply skim the long articles they come across? According to Liu, "Because of the growing number of scientific journals and expansion of the volume of these journals, readers of scientific journals cannot keep pace with the literature and are forced to skim journal articles the way that many readers skim newspapers. This trend tends to be more intensified in the web environment. Most people tend to read the first screen of text only. A total of 90 percent of people reading a web page do not scroll down" (Liu, 2005). Due to the information available being so vast and diverse, the average reader most likely will not spend time reading all the information in depth or taking the time to understand the research or methodology. In the worst-case scenario, the reader may be intimidated by the sheer length and weight of each text and avoid reading the research at all.

Coming up with a way to condense this information to tailor it to the modern average attention span is beneficial for the VPC. Accurately summarizing the information from a large report and putting it in terms that a layperson could understand would greatly benefit not only the interest and intrigue of the reports but also the level of understanding that the reader acquires from the information.

3.0 Methodology

The project's goal was to identify viable applications of artificial intelligence for the VPC to enhance access to its various projects and findings. By developing functional prototypes, the VPC would have a variety of tools to work with utilizing AI technology to enhance productivity and improve the quality of future endeavors. For this project, AI is defined as the use of machine learning models, natural language processing, and other automated systems to process, analyze, and generate content from data.

To achieve this goal the team completed the following objectives:

- 1. Explore the use of AI for the VPC's research
- 2. Produce functional prototypes of various AI tools for the VPC
- 3. Examine how AI can be utilized in future research for the VPC

With the variable scope of AI tools available, it was imperative to limit the scope of tools we consider to that of ones backed by large companies such as Google, openAI, or Meta due to their reliability and longevity. Our aim was to explore as many differing tools and their applications as possible, making sure to limit our timeframe of testing to allow the exploration of multiple uses of AI across the seven-week period. We were also limited by the VPC database, with its various CSV files and project reports being our only source of data when utilizing these different models. Without access to a server to host our own model the possibility of developing a model trained on just the VPC data was eliminated.

3.1 Explore the use of AI for the VPC's research

To experiment with possible AI tools that could apply to the VPC, we first had to discover what AI tools are even available. To pick what tools to test, we based our decisions on popularity and accessibility. For example, many people like to modify existing LLM models and upload them on <u>https://huggingface.co/models</u> for others to download and use for free (*Figure 3.2.1*). As of publication, a quick search on their website shows more than one million LLMs that are available for download and use. While downloading and using them is free, the computing cost is not. To use these models in a production environment, we would have had to purchase or rent out a server to host the model. This server would also require heavy duty and expensive

hardware to run these models. There are many other issues associated with this approach as well, including but not limited to the cost of network traffic, cyber security, and hardware maintenance. For this reason, we prioritized tools and LLMs hosted and maintained by third parties that allowed us to send requests to their servers. We focused primarily on tools that were made or hosted by large, trusted companies, ensuring the services' longevity. Such companies include but are not limited to OpenAI, Anthropic, Meta, and Google. This also ensured the ease of maintainability of our software prototype by any future ChatVPC teams to come.



Figure 3.1.1: Number of LLMs on (https://hugginface.co/models), 2024

Experimenting with AI tools meant we had to define a way to categorize, test, and rate them. Defining a standard rubric to assess any tool we encountered was key in finding the best possible tool that could apply to the VPC's needs. The metrics defined had to be applicable to every tool and accurately represent that given tool's potential and downfalls. Much of this data was gathered directly from the tools' documentation, but we also tested these tools ourselves. We also had to identify what metrics are the most critical and what dealbreakers a given AI tool can have. For example, if inputting a single paragraph in English into a given LLM service costs ten U.S. dollars but the output is free, we can still safely omit that LLM from any future testing since it is clearly not scalable for the VPC's needs.

There are a few metrics that were needed to assess every tool by, which are included in the tools' specification documentation. Firstly, we needed to measure the maximum input and output number of tokens a tool allows since it determines what kind of material that can be generated from it. We also needed to record the associated costs with each as it was a major determination in a tools viability. Another factor that was important was whether the tool allows files as input. This is especially important because much of what the VPC produces are reports, which are often in PDF format. The VPC also produces a lot of raw data, often in the form of comma separated values (CSV) files. In the case of LLMs, not allowing file upload means that we had to use another service that translates these files into ordinary text for the LLM to consume, which is another added cost. In the case that the tool permits file uploads, we also recorded if there is an added cost associated with uploading/inputting files. An additional factor to consider would be context length (in tokens), which determines how long or how much a given tool, particularly LLMs, can remember. For example, in the context of a chatbot, this would be quite an important metric to consider since a small context window means that the tool would perform poorly in the scenario that a user has a long back-and-forth conversation with the chatbot. A small context window means as the conversation goes on, the chatbot would forget what they have previously talked about sooner.

Performance and reliability were important metrics to add to our rubric. Some of the metrics to test were model size (parameters), perplexity (prediction), and BLEU, ROGUE, and METEOR scores. Model size generally represents how complex and adaptable the model is: "Broadly, for most tasks we find relatively smooth scaling with model capacity in all three settings; one notable pattern is that the gap between zero-, one-, and few-shot performance often grows with model capacity, perhaps suggesting that larger models are more proficient meta-learners" (Brown et al., 2020). Model size can also be found directly from the LLMs specification documentation, and are often specified in the magnitude of billions. Perplexity is a standard measure of how well the model can predict the next token. BLEU, ROGUE, and METEOR scores are also standard measures used to evaluate and benchmark text generation tasks, including translation, summarization, and paraphrasing. These metrics were sourced from publicly available benchmarks and tests. Since we are prioritizing more popular models, we did not have to benchmark these tools ourselves and instead used existing benchmarks. Most of these metrics are based on text generation, so they unfortunately cannot be used to evaluate audio or image generation which is still fairly new territory for the field.

Response time is a critical metric to measure since it greatly affects the end user of our product. We manually had to test each model under the same conditions. These conditions for example could be testing a large input or output length in text, testing multiple input files, or testing response time with different output mediums such as audio (if applicable). In 2006, a study by Amazon found that for every one hundred milliseconds added to a page loading cost them one percent in revenue, which is the equivalent (as of publication) more than 3.8 billion U.S. dollars (*Amazon Study*, n.d.). Another example is from Akamai, who is an online retailer,

which claims that 53 percent of mobile site visitors will leave their page if it took longer than three seconds to load (*Akamai Online Retail Performance Report*, n.d.). While we are certainly not under the same kinds of business pressures that these examples illustrate, they show just how critical speed can be to end users.

The features of each tool were a key metric in our assessments. These may include audio, video, file, and image output, which enabled us to prototype a more versatile and user-friendly application. Privacy and responsible data handling were also critical factors. We noted in our evaluation whether there could be potential ethical concerns, particularly when using third-party hosted services that disclose their data practices. Our rubric considered a combination of features, capabilities, price, performance, reliability, data handling, and speed (See appendix). This comprehensive assessment framework guided us in evaluating AI tools and identifying the most viable solutions for the VPC.

To explore the viability of the various AI tools selected for this study, the group considered the metrics described in the previous objective, while considering the different metrics found with different weights. After finding the values for each of the AI models, the group categorized the individual values based on their desirability. For this, the group has defined "Green Flags," or values that we believe are the best for that specific metric, and "Red Flags," or values that we believe are completely undesirable and invalidate the use of a specific model completely. Specifically, the Green Flags the group desires in an LLM are the following:

- Low cost: the lower the cost, the easier it is for the VPC to implement the LLM in a greater scale in the future.
- More parameters: the more parameters, the "smarter" a model is. This Green Flag is also evaluated against the Perplexity, and the BLEU, ROUGE and METEOR scores for that LLM.
- High context length: having a higher context length would allow the LLM to maintain understanding for a longer input.
- High Max Output Token: the higher the amount of output tokens an LLM allows, the longer the content it is able to produce.
- Allowing File Input and Output: because the reports produced by the VPC are mostly in a .pdf format, allowing file input would greatly increase the amount of

VPC data that a model can process. Also, allowing file output broadens the horizon of content forms that this study would be able to produce.

In this system of evaluation developed for this study, the group had concerns over the use of specific models. The Red Flags the group looked for are as follows:

- High cost: the group looked for a tool that has a cost below 0.016 (USD per 1 thousand tokens) for output and 0.005 (USD per 1 thousand tokens) for input. A cost higher than this would not allow the VPC to have multiple users accessing the data for a viable cost
- Low parameters: less than 70 billion parameters in a Large Language Model would likely not be enough to produce quality summaries and generate good enough output for the VPC's needs. This metric would also be evaluated against the Perplexity, and the BLEU, ROUGE and METEOR scores for that LLM.
- Ethical Concerns: if inputting data into a certain LLM generates ethical concerns based on the ownership of the data or the use of the inputted data for model training, the model will be disconsidered for VPC use.

Upon categorizing each of the values in the developed metric for individual models, the group discarded the models that had Red Flags, and ranked the rest of the models into how they would be useful to the other objectives listed in this study. Based on this systematic approach, the group was able to discover how each potential tool would be used before developing any prototypes. This would be used to save time when developing different tools by discarding tools that would not be useful in the study. After finding the viable AI tools, they were able to accurately compare them to use them to accomplish the other objectives in this study.

3.2 Produce functional prototypes of various AI tools for the VPC

After researching the viability of various tools and their associated determinants and benefits, the team worked to apply these tools to the VPC's facilities. When looking to apply these tools, the team needed to be conscious of the time frame to develop these tools and the availability of resources to implement them. When looking to implement AI tools to increase productivity and disseminate information it was imperative to have an intimate understanding of what data was available to us, what types of projects are being conducted, and how the VPC operates. The VPC has indicated that most of its project data is held within various CSV files

that are accessible via their website. Their student project reports are also accessible on their website in predefined categories, along with numerous applications that have been developed over the years with widely varying functionalities. With such a wealth of project data, the team could better identify how AI can enhance productivity along with what data can be fed to an AI. The VPC has done an exceptional job in keeping these project reports consistent and easily accessible making this task significantly easier. Most models can be prompted using project reports, but actually inputting CSV files may prove more effective for the models that can handle it.

First, we selected opportunities to implement AI into the VPC workflow. Some of these opportunities consisted of mediums such as audio generation, image generation, text generation, or a chatbot. These opportunities served as the main deliverables of the project, and aim to enhance various aspects of the VPC including information dissemination and project workflow. The initial brainstorming of potential implementations of functional prototypes was crucial in identifying what tools could actually be implemented or at least tested within our timeframe. At this point, the group met to discuss which AI tools would be most effective in implementing a given medium. By analyzing our previous assessment of the AI models we were able to identify which platforms should be disregarded and which would be most effective based on factors such as cost, tokens in tokens out, response time, and many more. It was crucial to test various models for the same objective to identify which model was most effective. Various methods could be used to compare these models, but generally, the team focused on the accuracy of the information produced, response time, and the amount of tokens it can input and output.

Once the tools were identified we began our initial design of the deliverable and planned how it could be best integrated into the VPC's current frameworks. Our design process considered numerous variables including an aesthetically pleasing design, human-computer interactions, and functional aspects. Throughout the development process, we constantly assess the viability of any prototypes, and if it becomes clear the prototype may not be feasible, we terminate the production and begin on another product. This ensures we do not get stuck on one prototype that won't be usable and gets us to explore various possible mediums without spending too much time on one. Upon the completion of a given product, we presented our findings to our sponsors for any feedback, implemented any quick fixes or changes, and then repeated our development process for the next product. This workflow ensured the exploration of our various ideas while still ensuring the team has some functioning prototypes by the end of the project.

By being conscious of our timeframe and constantly reassessing the feasibility of our deliverables, the team ensured various applications of AI technology into the VPC have been explored and can be researched further in the future if it seems to be feasible beyond the scope of our project. Adhering to this philosophy layed groundwork for future teams to not only utilize tools we've produced but also consider further areas of study into the AI field that may deserve to be its own project due the the complexity and recency of the subject. As AI has been rapidly developing in recent years the possibilities will grow exponentially over time. By exploring various implementations we identified possibilities that are not currently feasible due to technological limitations, but would undoubtedly be explored in the future after AI has been further developed.

3.3 Examine how AI can be utilized in future research for the Venice Project Center

With the recent boom of AI, it can become a useful tool that is utilized by the VPC to automate tasks that would otherwise be quite tedious. Our research and examination of AI, its current capabilities, and scope can be useful to future VPC projects in determining which AI tools are useful and how they are useful to their individual projects. This research was then organized into a table for future use by the VPC. It serves as a baseline that can be used as a reference by future VPC groups.

Through the completion of the project, it was critical to assess,

- How can we utilize AI tools that are manageable for future VPC projects?
- How can we utilize AI tools in a way that is scalable and applicable to various VPC projects?
- Are there general AI tools that can be applied to most VPC projects?
- Should there be an array of AI tools that are documented so that they can be used in specific scenarios for different projects?

Our review focused on how these entities are able to make the utilization of AI scalable and general enough for their operations. By understanding common techniques that have been utilized by these companies we were able to understand what general aspects we should incorporate into our deliverables so that future VPC students know how to utilize AI in their future research.

We are going to organize the information that we find in the form of a reference table, which could be represented as a database that can be utilized by future VPC groups. This database consists of information such as, what the AI tool is predominantly used for, its accessibility to the general public, how to use that AI tool and all its features, and the time in which that given entry on the given AI tool was updated. The update time is essential to this table as AI is a growing technology and is rapidly changing. This table is meant to provide a starting point of research for future groups to build off of so entries are to be updated as applications change. This database is an essential part of our project as it provides a quick and intuitive way to compare AI models without a complex understanding of the technology or extensive research.

For example, earlier in our methodology we discuss red and green flags of AI's with one example of a red flag being if the AI uses inputted information and user behavioral patterns to train itself. Somebody who is not well versed in AI technology, would not understand how or why this is a red flag and may not take the time to research or look in depth regarding the tool. We created a database of recommended AI models that can be quickly searched, so that VPC students can find accurate and helpful information to safely use AI tools to assist them in their projects.

3.4 User Testing

After the completion of our previous goals and final polishing of the website was complete, the group conducted a user study of current VPC students to see how the website could be useful for their own projects, and what tools they'd like to see in future iterations. The test aimed to gauge the usefulness of the current tools to students and provide further groundwork on future exploration of AI technologies for the VPC. User testing aimed to assess the ease of use and accessibility of the current website in conjunction with functional aspects and intuitiveness. The test helped determine what tools should be further developed for improved capabilities and which tools could be left as is or removed if deemed unnecessary for future students.

The user was asked to perform specific tasks by the investigators on the ChatVPC website. They were expected to complete the tasks within a ten-minute time frame. Upon completion of the assigned tasks, they had five minutes for free exploration and use of the website as they saw fit.

Finally, they were asked to participate in a guided interview for ten minutes answering questions about their experience with the website and asking for recommendations for future development. The interview consisted of an investigator asking questions while another one took notes of their responses. Interviews took place in private locations, in-person or on Zoom, if requested.

4.0 Findings

Throughout our current development cycle, we have been able to develop 3 separate AI tools along with an AI model evaluation tool. The tools are accessible through our website developed using Typescript, PostgresQL, PrismaORM. Our tools mostly focused on the utilization of LLM for text and audio generation due to image generation not having many implementations applicable to the VPC at its current level of development.A

4.1 Initial Design

While developing the application it was imperative to consider the user experience and interactions with the website in order to achieve an intuitive and easy-to-use design. To assist in the development of the website we utilized a variety of shadcn/ui, a collection of prebuild website components, along with TailwindCSS, a framework for website styling, which were able to speed up the development process in our short timeframe and allowed for an aesthetic design that was consistent throughout the website. The following are screenshots of the final website featuring all of the pages and some of their functionality. Explanations of the individual tools are detailed in sections following the screenshots.

Home Evaluation	ns – Tools – Recommendatio	ns About Us			୯
	We	lcome to th	e ChatVPC /	Арр	
	Evaluations Look at our evaluations of select models	Tools Take a peek at the sample tools that we've developed	Recommendations Check out some of the AI tools that we recommend you use!	About Us Get to know a bit more about the teams that have worked on this project in the past!	
	What is this project? This project was develop and understand previous reason. If you want to kn	ed to help WPI students to find AI to Venice Project Center IQP reports b ow more about the people who have	ools that might help them research, letter! This project was developed as worked on this before, visit our Abc	find, analyze, synthesize, i Its own IQP for this but Page.	



Home Evaluations											୯
Evaluat How differe	ion nt LLMs compare against each o	ther.									
Filter										•	S View
Model	Url	PpInput	PpOutput	TermsURL	CtxLength	ModelSize	MaxOutput	MaxInput	FileInput	FileOutput	Perplexi
Gemini 1.5 Flash	https://gemini.google.com/app			https://policies.google.com/terms	1000	8	2000000	1000000	true	true	
GPT 40 Mini	https://chatgpt.com/	0.15	0.6	https://openai.com/policies/terms-of-use/	128	8	16384	128000	true	true	
GPT 4	https://chatgpt.com	2.5	10	https://openai.com/policies/terms-of-use/	128	128000	4096	128000	true	true	
NotebookL	A https://notebooklm.google.com/			https://policies.google.com/terms					true	true	
Claude Haiku	https://www.anthropic.com/	0.25	1.25	https://www.anthropic.com/legal/consumer- terms	200		200000	200000	true	false	
Perplexity 3.5 Base	https://www.perplexity.ai/			https://www.perplexity.ai/hub/legal/terms- of-service	4	128000	4000	4000	true	false	
					Ro	ws per page	10 \$	Page 1 o	of 1		

Figure 4.1.2: Evaluations Page

Add a Model

Please follow these steps to test and add a model to our database!

1) Data Collection

Please collect the following information and save them for later:

- Model The name of the model. For example, "gemini-1.5-pro"
- URL The front page for the model
- Input Price The price for input per million tokens
- Output Price The price for output per million tokens
- Privacy URL The URL for the user agreement for the model
- Context Length The context length/window for the LLM
- Model Size The size of the LLM in billions of parameters
- Max Input Length The max number of input tokens per request
- Max Output Length The max number of output tokens per request
- File Input Can the model accept file input?
- File Output Can the model produce a file as output?
- Features A comma-delimited string representing the features of the model, e.g., "audio output, text to speech"
- Optionally, you can also collect the following metrics:
- Perplexity Perplexity score. This requires you to test the model yourself
- Output Response Time We will provide you with a text file and prompt. This is the response time in seconds

2) Testing The Model

We will be testing the output of the LLM based on a standard context file and prompt, which we will provide.

Click here to download the text file that we will use for context.

Then, copy the following prompt and query the model based on that:

Summarize this prompt to the best of your ability

Then, record the output of the model.

That's it! You're ready to submit a model!

Add New Model



Home Evaluations \sim Tools \sim Recommendations	About Us					C.
Add a Model	Model Name				<	
Please follow these steps to test and add a mode						
1) Data Collection	URL					
Please collect the following information and save	Terms URL					
• Model - The name of the model. For example,	https://					
• URL - The front page for the model	Input Price	Output Price	Context Length	Perplexity		
 Input Price – The price for input per million tol 	0	0	0	0		
Output Price - The price for output per million	Model Size	Max Input	Max Output	Response Time		
Privacy URL - The URL for the user agreemen	0	0	0	0		
 Context Length – The context length/window 						
 Model Size - The size of the LLM in billions of 	Your Model's o	utput				
• File Input - Can the model accept file input?	O Does this 1	ake file input? 🤇	Does this output a	a file?		
• File Output - Can the model produce a file as		5	Send			
Features - A comma-delimited string represer						
Optionally, you can also collect the following metri						
• Perplexity - Perplexity score. This requires you						

Figure 4.1.4: Add Model Page (Form)

Home Evaluations ~ Tools	 Recommendations About Us 		٩
Tools Take a look at the tools we have m	nade so far.		
Chatbot Use the chatbot to ask questions about specific VPC IOPs	CSV Query Tool Chatbot Ask CSV specific questions to our live-querying chatbot and get real-data answers from select data sets.	Report Chatbot Use the chatbot to ask questions after selecting a specific VPC IQP reports	

Figure 4.1.5 Tools Gallery

Home	e Evaluations ~	Tools ~	Recommendations	About Us			ନ
Chat a	about the VPC け						
Reco	ommendations:						
*	How can I help you?						

Figure 4.1.6: Chatbot

Home	Evaluations $\scriptstyle{\scriptstyle \vee}$	Tools \checkmark	Recommendations	About Us		୯
Chat abo	out the VPC け					
Recomm	mendations:					
Green Venice			3	Condition and Availability of Urban Green Space in Cannaregio application/pdf	A Greener Venice	
÷ Ho	ow can I help you?					
Tell me about	t Green Spaces					
				Submit! 🥑		

Figure 4.1.6.1: Chatbot with Recommendation and Query

Home Evaluations - Tools - Recommendations About Us		(6)
Chat about the VPC 1+		
Pecommendations:		
Green Venice	Condition and Availability of Urban Green Space in Cannaregio	A Greener Venice
	2000-2000-200	
🖌 How can I help you?		
		Tell me about Green Searces
This document is an interdisciplinary Qualifying Project (IQP) submitted to the Worcester Polytechnic Institute WPII in partial promoting sustainable stewardship of green spaces in Venice, Italy by creating an interactive catalog to match people with uegetation, amenities, land condition, and photos. The green spaces were categorized as parks, gardens, urban wilds, famila and printed bookiet.	fulfillment of the requirements for the Bachelor of Science degree. The project focuses on and The study cataloged 182 green spaces in Venice and collected data on location, accessibility, nds, and courtyards. The resulting catalog is available to the public as an interactive map, website,	
The project has three objectives:		
Inventory and categorize green spaces into an interactive catalog.		
Promote sustainable stewardship through the dissemination of knowledge.		
Develop a mechanism to match green spaces to stewards.		
The IQP team collaborated with We Are Here Venice, a non-profit organization advocating for evidence-based policy change which focused on green spaces on the laised of Guideoca. This current project expands the scope to the historic city center	s in Venice. The project builds upon previous work, such as the A Greener Venice project (2017), (Città Storica), specifically the seatiere of Castello, which has a high concentration of green spaces.	
The methodology included:		
Locating green spaces using Google Maps, Conoscere Venezia, and the Comune di Venezia geopPortale.		
Surveying green spaces using a customized Google Form to collect detailed information.		
Creating an interactive website using Wix to host the catalog and provide search functionality.		
Developing a printed booklet version of the catalog.		
Analyzing green space availability and calculating the green deficit in Castello.		
The key results include a map and catalog of NB2 green spaces in Castello, catagorized by type, condition, and accessibility. The urban wild on the stand of San Patero (Campazzo San Patero) and the Ex-Gasometri space in Castello Ovest. Recommen information on space ownership for collaboration.	The project also highlights potential stewardship opportunities for underutilized spaces, specifically dations for future work include further surveying, collecting soil and air quality data, and gathering	
		<i>k</i>
	Susmiti 10	

Figure 4.1.6.2: Chatbot with Recommendation and Query Output

Hom	e Evaluations ~	Tools ~	Recommendations	About Us			ନ
Chat a Choose (If you r	about the VPC one of our preloaded notice that the chatbe	d CSVs to as ot is taking t	sk questions o long to respond, ple	ase refresh the page and tr	y again)		
(Choose a CSV here! CSVs						
•	How can I help you'	?					
					Submit! 🚀		

Figure 4.1.7: CSV Chatbot

Hom	ne	Evaluations \lor	Tools ~	Recommendations	About Us C	•
Chat Choose (If you	abou e one o notice	of our preloaded that the chatbo	CSVs to asl t is taking to	k questions b long to respond, plea	ase refresh the page and try again)	
(Choos Fe	e a CSV here! ountains				
	How	r can I help you?				
					What is the tallest fountain in Venice	
wiki . 59.5	_frien	dly_title height 3264	_cm			
					Submit! 🔊	Ē

Figure 4.1.7.1: CSV Chatbot with Query and Response

Hom	e Evaluations ~	Tools ~	Recommendations	About Us			G
Chat a	about the VPC	\+					
	How can I help you	רי?					
					Submit! 🦪		

Figure 4.1.8: Report Chatbot



Figure 4.1.8.1: Report Chatbot Query and Response

Home	Evaluations \sim	Tools ~	Recommendations	About Us	ନ
				Chonse File No file chosen	

Figure 4.1.9: Podcast Generation Page



Figure 4.1.9.1: Podcast Generation Loading

Home Evaluations \sim Tools \sim Reco	ommendations About Us	ନ
► 0:00/4:33 •) :	Choose File What is the M_rishna Garg.pdf Upload male-1: Welcome everyone to "The Meaning of It All," the podcast that delves into life's big questions. I'm your host, Alex. female-1: And I'm Dr. Anya Sharma, philosopher and expert on Eastern thought. male-1: Today's episode explores a topic near and dear to many hearts: how do we find fulfillment in life? Some chase external validation – money, status, fame – while others seek internal peace and enlightenment. We'll delve into both perspectives with some surprising case studies. female-1: Exactly, Alex. Many people believe that life's meaning lies in material success, the acquisition of wealth, and the pursuit of power. Think of Ivan Ilych from Tolstoy's "The Death of Ivan Ilych". He dedicated his life to climbing the social ladder, yet in his final moments he found his life tragically empty.	

Figure 4.1.9.2: Podcast Generation Output



Figure 4.1.10: Recommendations Page



Figure 4.1.11: About Us Page

4.2 Large Language Model Evaluation Tool

The first tool we were able to develop was a Large Language Model (LLM) Evaluation. The table has information useful to future VPC students in deciding which AI tools to use depending on their computation capabilities and different features of the LLM. Using this table a group can infer if a given LLM has the capabilities that the team is looking for so that they don't have to start from scratch and research various AI tools in order to determine the best one. For example, if a group needs to find an LLM that supports file input and output, our table lists all the LLMs that support file input and output.

The tool provides functionality to filter a table of models by various different parameters, that were determined based on the previously outlined green and red flags discussed in our methodology, including but not limited to token input and output price, model size, or file input and output. The tool includes extensive explanations on how to properly add an AI model to the table, as highlighted by figure 4.1.3. The user is expected to collect the information in the highlighted fields in the form in figure 4.1.4 which can most often be identified within a model's specifications. The tool would then take the "Your Models Output" field and generate bleu, rouge, and meteor scores. These scores are defined in our background research, however here is a quick description of what these metrics measure:

- 1. Bleu: Measurement of the difference between an automatic translation and human-created reference translations of the same source sentence. The bigger this value is, the better the model is.
- Rouge: Compare an automatically produced summary or translation against a reference or a set of references (human-produced) summary or translation. ROUGE metrics range between 0 and 1, with higher scores indicating higher similarity.
- 3. Meteor: Evaluation metric for machine translation that improves over traditional metrics like BLEU by incorporating linguistic features such as synonymy, stemming, and word order, and placing more emphasis on recall to better align with human judgments of translation quality. The bigger the value is, the better the model is.

These metrics are also explained in a popup right on the table in case of confusion as to what the metrics mean. Upon calculation of the scores, the model would be added to the table where it could then be compared to other models in our database.

4.3 Report Chatbot

After developing our LLM evaluation tool, we began the development of a chatbot capable of responding to VPC-specific questions. Our initial plan for the chatbot was to feed the LLM project reports as context for a user's query to provide an accurate response. Using our evaluation tool we decided to develop the tool using Gemini 1.5 Flash due to its ability to handle file uploads, significantly larger number of input tokens, and it is free to use. We began development of the tool creating a frontend database and an API to communicate with the LLM. With these two completed we had our first iteration of the chatbot which essentially acted as a way to communicate with Gemini 1.5 Flash using our website to ask questions and receive a response, but it had no information about the VPC making it incapable of answering project-specific questions. We then implemented a vector database utilizing Pinecone to achieve query-based retrieval-augmented generation. By vectorizing project descriptions provided by the VPC project report master sheet, we were able to design a system to retrieve the most closely related project reports to the user's query. The system followed these steps to achieve our desired result:

- 1. The user would type a question into our chatbot and hit send.
- 2. The user's question would be routed to our vector database and a semantic search would be performed.
- 3. Vector database would return the titles of the closely related projects.
- 4. Project reports would be uploaded from our database to the LLM along with the user's question
- 5. The response would be received and displayed.

Though this system was successful we quickly realized several flaws with our approach. To start with, using an external service for our vector database (Pinecone) caused excessive network traffic that could be abstracted by having the vector database in our current schema. Along with that, handling retrieving the files and uploading them to Gemini 1.5 Flash made the chatbot impractical due to the atrocious response time which we found could be over fifteen minutes in some cases. On every user query, each individual report retrieved from the vector database would have to be downloaded locally to a temporary file, uploaded to Gemini 1.5 Flash, and then analyzed before being able to respond. The downloading portion could have been removed by downloading the project reports into the website code but that would make the application unnecessarily large. Each individual project report could be anywhere from fifty to one hundred pages, so storing them or attempting to send more than one at a time wouldn't make sense. After some testing, we found the system actually did work fairly well if you only sent one project report at a time, so we began the next iteration of the chatbot. With our new system, upon the user typing into our chatbot's interface the semantic search would be performed dynamically pulling the top three closely related projects to the user's query. These projects would be displayed above the chatbot window so the user could see the updated recommendations as they changed. The user could then select one of three recommended project reports to ask questions about. We discovered the response time could be a bit long the first time a question was asked about a project report. So we decided to switch to using Gemini Pro as it had the same benefits as the previous model but with quicker response times at the cost of being a little less "intelligent". The chatbot at this stage was actually fairly useful, allowing users to quickly receive answers to project-related questions. Upon testing this version of the chatbot we discovered a glaring flaw that we had been ignorant of: the project reports actually did not include much of the data actually collected during the span of the project. When asking the chatbot questions at this stage it could give key insights into how projects were conducted, conclusions, and findings but it was unable to answer specific questions such as "How many fountains are there in each Venetian burough". After discovering this, it was decided that it would be best to create a chatbot that could handle the open data collected by the VPC commonly stored in CSV files and began our development of another tool which is touched upon later. However, as we were hoping the chatbot would be our main deliverable, we decided to try one more approach to hopefully enable the chatbot to more accurately answer questions by pulling sections of papers across multiple different project reports. The approach involved parsing the project reports into 1000 token chunks, with 100 tokens of overlap to ensure coherence, which would then be stored in our existing database which could then have a semantic search performed. This would allow multiple snippets of useful information to be pulled from various different project reports before being sent as context for a user's question allowing for quicker analyzation of the provided project report chunks and more accurate responses.

4.4 Podcast Generator

Inspired by NotebookLM's podcast generation tool, an application developed by Google to create podcasts and answer questions related to a given pdf file, we decided to investigate the possibility of having our own program that would take in IQP reports and generate five to ten-minute podcasts regarding the content of the inputted report. We were able to utilize Python libraries, ChatGPT tts-1, and Google Gemini 1.5 Flash to make this application feasible. The Python libraries were used to save, parse, and feed the PDF input into both ChatGPT and Gemini which were also used to save the outputted data to link it to a voice that would be played in the resulting podcast. Google Gemini was used to create the transcript and ChatGPT was used to create the sound bytes that would be played throughout the duration of the podcast. ChatGPT's tts-1 model was used as it was the only model capable of generating audio files within the major organizations we researched. The script writing was delegated to Gemini due to the cost restrictions imposed by ChatGPT.

The flow of events of this application are as follows:

- 1. The user sends a pdf to our program which is then saved and parsed into a string by Python's "pdfReader" library.
- 2. The string is then sent to Google Gemini in order to create the transcript of the podcast.
- 3. The transcript is fed into ChatGPT and the audio is created.
- 4. The generated audio is then sent back to the client and formatted into a mp3 file to be downloaded onto the client's local machine.

Unfortunately, ChatGPT's API's keys have an associated cost per 1000 token input and output so depending on the size of the pdf input it may cost anywhere from five to thirty cents to generate a podcast. Due to this, the program could be easily exploited and therefore is not hosted to the general public. Anyone who would like to utilize the tool will need the podcast-generating program running on their local machine and must be given SerenDPT's ChatGPT API key or get their own.

Another reason that this program is not hosted is due to the fact that it is a serverless function. A serverless function is a program that exists for a short period of time and does not permanently reside on some machine. So if we were to call a serverless function, the program would run for the duration of the call and then be terminated immediately after. The problem with serverless functions is that there is a timeout to these functions which is usually anywhere from ten to thirty seconds. Due to the complexity required to generate a podcast, the tool always exceeds the thirty second limit. So in order for us to host this program on a platform that supports serverless functions existing for ten minutes, we would have to pay a monthly charge of at least twenty dollars per month (Vercel Pricing). Instead of hosting the tool the code's Github repository was handed over to the VPC so that students can run the program on their machine locally. They will need access to SerenDPT's ChatGPT API keys in order to run the program successfully.

4.5 From CSV File to Chatbot Response

We began the development of a chatbot capable of handling the CSV files provided by the VPC. The CSV chatbot would theoretically be able to answer data-specific questions. By passing the headers of a CSV file that correspond to a database relation that stores the data in that CSV to Google Gemini's Gemini 1.5-Pro, the model is able to return queries that can search for the corresponding data in the database. The choice to use Gemini 1.5-Pro stemmed from the worries that Gemini 1.5-Flash, which despite working much quicker than Gemini 1.5-Pro, does not have as many parameters, and therefore is more likely to return incorrect information. When using the tool, if a user asks, for example, "What are the fountains with the greatest waterflow in Cannaregio?," the LLM returns a query, that is, a systematic way of retrieving data from a database – based on the information about the CSV that is passed down to it – that can return that exact information. The user is then able to further prompt the chatbot tool on the information that has been returned. The initial prototype present with this study only has the capability to answer questions about the fountains in Venice, extracted from the Venice Project Center's website, veniceprojectcenter.org/opendata. The choice to only have one CSV was made to simplify the programming process, and because this tool was developed as a proof of concept. The prompt used to generate the query returned by the LLM is as follows:

You are a chatbot that is meant to give back an PostgreSQL query on a specific schema based on the users question. Do NOT give back anything else other than PostgreSQL queries. If you are return an SQL query, start the query with ```sql and do NOT include anything else in the message. Limit the amount of rows you would the query show to 25 rows. Do not select all of the columns. Only select the columns relevant to the users question and the wiki friendly name for the fountains. In your SQL statements, ALWAYS add quotation marks around the name of the relation in the FROM field. Always filter out NULL values from the results. Only return the user another SQL query if they request you to do so. You will then answer questions based on the result of the query, not necessarily only in SQL. This is the prisma schema you will use to reference:

This prompt is also given with the information on what kind of data is in the database that the chatbot querys. By providing the LLM this prompt, the response after being provided the user question is tailored to be a specific query that Postgres is able to handle and is appropriate given the current dataset. The result is then checked for whether the bot returned a query or a text response. If it returns a query, the response is run in our database, otherwise, it is returned to the user as a plain text message. When the chatbot runs a query through the database, the results of the query are shown in the text message in the shape of rows and columns, as a spreadsheet would. This tool can be used to find quick quantitative information about the raw data that is present in the VPC's website. The tool can also be used to prompt for some qualitative data, such as additional notes, percentage risk, etc.

4.6 User Testing

User testing was conducted on a total of ten students currently conducting research projects at the VPC, covering two members from each project group. The testing provided a number of insights into the limitations of the chatbots developed and gave key insights to the direction future groups should investigate. Starting with the report chatbot utilizing Gemini 1.5 Flash a number of problems arose throughout user testing. The most glaring issue was the lack of onscreen indication when an error occurred, leaving the user expecting a response to be generated when none was coming. The issue could quickly be fixed within the code of the website, but the more significant problem was the frequency errors occurred and no response was returned. Upon further investigation of the returned error it was identified as an error returned from Gemini and not from the website's code itself. Though Gemini offers a variety of free API keys to a number of models, it seems as though these models frequently become overloaded with requests and become unresponsive. Another problem arose from the accuracy of the responses that were provided. Often when a response was obtained, according to the users testing the chatbot, the answers to their questions would on occasion be slightly off from the correct answer or require further prompting further prompting in order to obtain the correct answer. This issue most likely stemmed from two primary sources: inaccuracy of project reports, and confusions of the model.

For the CSV chatbot utilizing Gemini 1.5 Pro many of the same issues were persistent. Firstly was the lack of an error message for the user upon failure. Upon being asked a question it had no answer to it would return "No Data" which was the desired result, but unfortunately, any further prompting after that instance would result in a "No Data" response despite being asked a question it should have the answer to requiring a refresh of the page to fix. The reason for this is because of the chatbot's ability to "remember" previous responses, making it think it doesn't have data for any future questions. In some cases further prompting was needed to obtain an accurate answer. Additionally, faulty data held within the CSV's made the responses spotty at times. For example, when questions on the number of fountains per boroughs was asked, it did return correct values for each of them but also said there was a boroughs in Venice named "True", "0", and some duplicates like "san marco" and "San Marco" being counted as different boroughs.

To address these issues a number of steps would need to be taken. First an extensive review and cleaning of the VPC's CSV data would be necessary to remove duplicates or faulty data. By doing this accuracy and response time of the chatbot would be improved and allow for easier integration of additional csv files in the future. For the report chatbot further controlling what text is being sent aligned with the prompt would enable more accurate responses. If, for example, we could parse out specific sections of papers and identify what sections a user wants

information from before sending it to the model would improve accuracy. Swinging to a different model may also provide more accurate responses as the choice of Gemini 1.5 Flash was primarily to avoid long wait times after prompting the chatbot. To avoid the frequent errors from Gemini investing in a paid API key may be able to significantly improve the reliability of the chatbots and provide an improved, and more useful, user experience.

Despite the glaring issue identified by the user study all the subjects identified the potential of the technology with further development. With further expansion of the tools currently available, and proper investment into the proper hardware and software, the users saw the potential for the technology to be a highly beneficial tool in the execution of their individual projects.

5.0 Conclusions

This project demonstrated the potential for integrating artificial intelligence into the VPC's workflow, focusing on improving data accessibility, usability, and impact assessment. By creating prototypes such as chatbots for reports and CSV data, a podcast generator, and an AI evaluation tool, we explored how AI could transform the VPC's extensive data into actionable insights and user-friendly formats. Each tool addressed specific challenges, such as summarizing dense academic reports or answering data-specific queries, while considering ethical implications and user accessibility.

Our findings underline the importance of AI in automating repetitive tasks, improving user interaction, and expanding the VPC's impact. Although challenges remain, such as response times for chatbots and the technical complexity of integrating AI systems, the tools developed during this project provide a strong foundation for future work. An iterative development approach allowed us to refine functionalities based on performance and feedback, and lay crucial groundwork for future teams to expand the functionalities available to better integrate AI into the VPC wokrflow.

6.0 Recommendations and Future Work

For enhanced usability and scalability of the tools developed, it is important to refine the systems to improve the response times and to integrate more efficient data handling solutions, such as local vector databases to query faster. It would also be pivotal to improve the CSV chatbot to cover more use cases and provide precise, context-sensitive answers. In the case of the CSV chatbot, it is paramount to include data from other reports into the tools, considering that it is currently only able to answer questions on the fountains of Venice. For improved user experience, the team would recommend adding a section where the users can see what information is in the data, so they know what questions they can ask from it.

Finding a better hosting solution would also positively impact the project, as hosting long-running serverless functions at a inexpensive cost proved to be an impossible task. Moreover, providing a way to fetch recently updated VPC IQP reports to update the pool from which the tools can retrieve data ensures the system stays relevant and accurate over time.

In addition to these, improving the integration of Artificial Intelligence tools, such as image recognition for analyzing visual data in reports and implementing predictive analytics for long-term impact assessments would be great tools to include in the project and are likely to greatly impact the future usability of the website and its tools. Furthermore, the addition of a feedback mechanism would likely help users suggest what they want to see in the website directly, without having to contact somebody outside of the website.

For storing and querying an LLM, the response time and quality of the answers could be greatly impacted by securing resources for hosting and scaling AI applications, such as dedicated servers or cloud-based solutions for seamless operation. Furthermore, it would be beneficial to the VPC's AI efforts to explore the viability of hosting an LLM model to lower costs and enhance data security.

By building on the groundwork established in this project, the VPC can further harness AI to enhance its contributions to future VPC project groups, ensuring sustainable and impactful use of its data and research.

References

- Akamai Online Retail Performance Report. (n.d.). Akamai. Retrieved October 1, 2024, from <u>https://www.akamai.com/newsroom/press-release/akamai-releases-spring-2017-state-of-online-retail-performance-report</u>
- Amazon study: Every 100ms in Added Page Load Time Cost 1% in Revenue. (n.d.). Conductor.RetrievedOctober1,2024,https://www.conductor.com/academy/page-speed-resources/faq/amazon-page-speed-study/
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). *Language Models are Few-Shot Learners* (No. arXiv:2005.14165). arXiv. <u>http://arxiv.org/abs/2005.14165</u>
- Bester, Nicolette, Michelle Di Vito-Smith, Theresa McGarry, Michael Riffkin, Stefan Kaehler, Richard Pilot, and Robert Bwire. "The Effectiveness of an Educational Brochure as a Risk Minimization Activity to Communicate Important Rare Adverse Events to Health-Care Professionals." *Advances in Therapy* 33, no. 2 (2016): 167–77. <u>https://doi.org/10.1007/s12325-016-0284-y</u>.
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2022). On the Opportunities and Risks of Foundation Models (No. arXiv:2108.07258). arXiv. https://doi.org/10.48550/arXiv.2108.07258
- Freyer, N., Kempt, H., & Klöser, L. (2024). Easy-read and large language models: On the ethical dimensions of LLM-based text simplification. *Ethics and Information Technology*, 26(3), 50. <u>https://doi.org/10.1007/s10676-024-09792-4</u>

Giray, L. (2024). AI Shaming:

- The Silent Stigma among Academic Writers and Researchers. *Annals of Biomedical Engineering*, 52(9), 2319–2324. <u>https://doi.org/10.1007/s10439-024-03582-1</u>
- Hosseini, M., Resnik, D. B., & Holmes, K. (2023). The ethics of disclosing the use of artificial intelligence tools in writing scholarly manuscripts. *Research Ethics*, 19(4), 449–465. <u>https://doi.org/10.1177/17470161231180449</u>

- McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. AI Magazine, 27(4), Article 4. <u>https://doi.org/10.1609/aimag.v27i4.1904</u>
- McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133. <u>https://doi.org/10.1007/BF02478259</u>
- Michelbacher, L. (2013). *Multi-word tokenization for natural language processing* [doctoralThesis]. http://elib.uni-stuttgart.de/handle/11682/3225
- Mielke, S. J., Alyafeai, Z., Salesky, E., Raffel, C., Dey, M., Gallé, M., Raja, A., Si, C., Lee, W. Y., Sagot, B., & Tan, S. (2021). Between words and characters: A Brief History of Open-Vocabulary Modeling and Tokenization in NLP (No. arXiv:2112.10508). arXiv. https://doi.org/10.48550/arXiv.2112.10508
- OpenAI. (2024). *ChatGPT* (October 9 Version) [Large language model]. OpenAI. https://chat.openai.com/
- Petrov, A., La Malfa, E., Torr, P., & Bibi, A. (2023). Language Model Tokenizers Introduce Unfairness Between Languages. Advances in Neural Information Processing Systems, 36, 36963–36990.
- Russell, S. J., & Norvig, P. (2016). Artificial intelligence: A modern approach. Pearson. https://thuvienso.hoasen.edu.vn/handle/123456789/8967
- Strelkova, O. (2017). *Three Types of Artificial Intelligence*. https://eztuir.ztu.edu.ua/bitstream/handle/123456789/6479/142.pdf
- Suyer, M., Shulman, O., Kelly, A., & Dahlberg, E. (2022).
- Happy 35th: Celebrating 35 Years of Research at the Venice Project Center. : Worcester Polytechnic Institute.
- Lin, Z., Guan, S., Zhang, W., Zhang, H., Li, Y., & Zhang, H. (2024). Towards trustworthy LLMs: A review on debiasing and dehallucinating in large language models. *Artificial Intelligence Review*, 57(9), 243. <u>https://doi.org/10.1007/s10462-024-10896-y</u>
- Liu, Z. (2005), "Reading behavior in the digital environment: Changes in reading behavior over the past ten years", Journal of Documentation, Vol. 61 No. 6, pp. 700-712. <u>https://doi.org/10.1108/00220410510632040</u>

- Verspoor, K. (2024). 'Fighting fire with fire'—Using LLMs to combat LLM hallucinations. Nature (London), 630(8017), 569–570. <u>https://doi.org/10.1038/d41586-024-01641-0</u>
- Tonmoy, S. M. T. I., Zaman, S. M. M., Jain, V., Rani, A., Rawte, V., Chadha, A., & Das, A. (2024). A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models (No. arXiv:2401.01313). arXiv. <u>https://doi.org/10.48550/arXiv.2401.01313</u>
- Whalen, Akina. "Evolution of the Attention Span." *The Current* (blog). Accessed September 16, 2024. <u>https://olhscurrent.org/11001/features/evolution-of-the-attention-span/</u>.

Yuna, D., Xiaokun, L., Jianing, L., & Lu, H. (2022).

- Cross-Cultural Communication on Social Media: Review From the Perspective of Cultural Psychology and Neuroscience. Frontiers in Psychology, 13. <u>https://doi.org/10.3389/fpsyg.2022.858900</u>
- Kreps, S., George, J., Lushenko, P., & Rao, A. (2023). Exploring the artificial intelligence "Trust paradox": Evidence from a survey experiment in the United States. *PLOS ONE*, 18(7), e0288109. https://doi.org/10.1371/journal.pone.0288109
- Hosseini, M., Rasmussen, L. M., & Resnik, D. B. (n.d.). Using AI to write scholarly publications. *Accountability in Research*, 0(0), 1–9. https://doi.org/10.1080/08989621.2023.2168535
- Harasta, J., Novotná, T., & Savelka, J. (2024). It Cannot Be Right If It Was Written by AI: On Lawyers' Preferences of Documents Perceived as Authored by an LLM vs a Human (No. arXiv:2407.06798). arXiv. <u>https://doi.org/10.48550/arXiv.2407.06798</u>
- Vykopal, I., Pikuliak, M., Srba, I., Moro, R., Macko, D., & Bielikova, M. (2024). Disinformation Capabilities of Large Language Models. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (pp. 14830–14847). Association for Computational Linguistics. <u>https://aclanthology.org/2024.acl-long.793</u>
- Whalen, Akina. "Evolution of the Attention Span." *The Current* (blog). Accessed September 25, 2024. <u>https://olhscurrent.org/11001/features/evolution-of-the-attention-span/</u>.
- Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., Yang, L., Zhang, W., Jiang, J., & Cui, B. (2024). *Retrieval-Augmented Generation for AI-Generated Content: A Survey* (No.arXiv:2402.19473). arXiv. http://arxiv.org/abs/2402.19473

Appendix A - Evaluation Metrics:

Pricing And AI Tool Information

Tool		Price per 1K Token Input	Price per 1K Token Output	Terms & User Agreement	Ethical Concerns
Name	URL	(USD)	(USD)	URL	(Y/N)

Performance

		Context	Model Size					Input Response	Output Response
Tool		Length	(Billions of	Perplexity	BLEU	ROGUE	METEOR	Time (ms/1K	Time (ms/1K
Name	URL	(Tokens)	Parameters)	Score	Score	Score	Score	Token)	Token)

Capabilities And Features

Tool					File Output	Additional/Other
Name	URL	Max Output Token	Max Input Token	File Input Allowed(Y/N)	Allowed (Y/N)	Features