**Appendix A**

PROMPT 1

**Detailed Analysis of Prompt 1: Identifying Terms Related to Synthetic Data**

**Prompt:** "*Given the following information, tell what are the words that are more related with the term synthetic data."*

**Purpose**

This prompt was designed to extract and identify key terms that were closely associated with the concept of "synthetic data" within a specific piece of text. The primary goal was to uncover recurring themes and significant terms that were contextually linked to synthetic data, which in turn could provide insights into the thematic focus of the text. In this case, the text is an abstract was from “*Health Synthetic Data to Enable Health Learning System and Innovation: A scoping review”,* *(Tsao et al., 2023)*.

**Methodology**

To address this prompt effectively, a systematic approach is required. The process can be broken down into the following steps:

1. **Text Processing and Tokenization:**
   * The first step involves preprocessing the text to break it down into individual words or terms. This includes removing stopwords (common words that do not carry significant meaning, like "and," "the," "of," etc.) and punctuation. Tokenization splits the text into its constituent words or phrases, preparing it for further analysis.
2. **Term Frequency Analysis:**
   * Once tokenized, a frequency analysis is conducted to determine how often each word appears in relation to "synthetic data." This step often utilizes techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) to highlight words that are not just frequent but also carry significant meaning within the context of synthetic data.
3. **Contextual Relevance:**
   * Beyond frequency, the contextual relevance of each term is assessed. This involves examining the surrounding text to understand how each term relates to "synthetic data." For example, terms like "governance," "utility," and "privacy" might appear frequently, but their relevance is determined by how they modify or interact with the concept of synthetic data within the abstract.
4. **Identification of Key Terms:**
   * The final step is to compile a list of terms that are most closely related to "synthetic data" based on their frequency and contextual relevance. These terms are likely to represent the key themes or focus areas of the text in relation to synthetic data.

**Application of Methodology**

Applying this methodology to the provided text, the key terms related to "synthetic data" would likely include:

* **Health Synthetic Data:** This term directly modifies "synthetic data," specifying its application in the health sector, which is the primary focus of the abstract.
* **Governance:** The abstract mentions that the governance of synthetic data has not been extensively studied, indicating this as a significant theme.
* **Utility:** The text discusses the utility of health synthetic data, which is crucial for understanding its practical applications.
* **Privacy:** Although the risk of privacy leaks is noted as low, privacy remains a critical issue associated with synthetic data.
* **Data Quality:** The comparison of synthetic data quality to real data highlights another important aspect.
* **Regulations and Ethics:** The abstract points out the lack of explicit regulations and ethical guidelines, making these key areas of concern.
* **Data Sharing:** This is identified as an area with common principles but also noted as largely inexplicit, emphasizing its relevance.

**Impact**

By identifying these terms, this prompt helped to clarify the primary focus areas of the abstract. It revealed that the discussion around synthetic data in the studied context was heavily centered on issues of governance, utility, privacy, data quality, and ethical considerations. Understanding these associations was fundamental and allowed us to pinpoint the central themes of the literature and guide the direction of the systematic review.

This process was not just about listing terms but about understanding the broader narrative in which synthetic data is discussed. This analysis provided a foundation for further exploration of these themes in the literature, ensuring that the systematic review would be designed in a way that was both comprehensive and focused on the most relevant aspects of synthetic data in the health domain.

**PROMPT 2**

**Detailed Analysis of Prompt 2: Document Categorization Using Python**

**Prompt:** "*Act as an expert in Python code and an expert in data science and help me analyzing the following documents. Please look at the uploaded documents and make a new column with the categories of the papers. Please do it thoroughly, because it is very important for my career."*

**Purpose**

This prompt was designed to automate the categorization of academic papers based on predefined criteria. The purpose was to systematically analyze the contents of the uploaded documents and assign each paper a category that reflects its primary focus or theme. The categorization process was central for organizing the data, ensuring that the systematic review was being designed both comprehensive and well-structured.

**Methodology**

To effectively respond to this prompt, a systematic data processing and analysis workflow was implemented using Python. The process can be outlined as follows:

1. **Data Loading and Preprocessing:**
   * **Loading the Data:** The first step was to load the uploaded documents into a Python environment. This typically involves reading data from CSV, Excel, or text files using libraries such as Pandas. For exemple:

Uma imagem com texto, Tipo de letra, captura de ecrã, Gráficos

Descrição gerada automaticamente

* + **Inspecting the Data:** Once the data is loaded, it’s essential to inspect it to understand its structure and content. This step helped identify the columns that would be used for categorization.

1. **Category Definition and Assignment:**
   * **Defining Categories:** Based on the research goals, specific categories needed to be defined. For example, if the review focuses on health-related papers, categories might include "Health," "Privacy," "Ethics," and "Governance."
   * **Keyword-Based Categorization:** The simplest method to categorize papers was by using keyword matching. The script scanned the relevant columns (e.g., title, abstract) for keywords associated with each category. For instance:

Uma imagem com texto, captura de ecrã, Tipo de letra

Descrição gerada automaticamente

* + **Advanced Categorization Techniques:** If a more nuanced categorization is needed, Natural Language Processing (NLP) techniques such as TF-IDF vectorization combined with machine learning models (e.g., logistic regression, random forests) could be used to predict categories based on the text.

1. **Integration of the New Column:**
   * **Adding the Category Column:** The next step is to add a new column to the dataset with the assigned categories. This is done by appending the results of the categorization function to the DataFrame.
   * **Saving the Updated Dataset:** Finally, the updated DataFrame with the new "Category" column is saved back to a file for further analysis. For exemple:



1. **Validation and Refinement:**
   * **Manual Review:** To ensure accuracy, we manually review a sample of the categorized papers. This step helped in validating the categorization process and making any necessary adjustments to the script.
   * **Iterative Refinement:** If errors or misclassifications were found, the script could be refined.

**Impact**

This prompt’s execution was significant in structuring the dataset for further analysis. By categorizing the papers, the systematic review process become more organized and focused, allowing us to analyze trends, draw connections, and synthesize findings within each category. The thoroughness required by this prompt ensures that no relevant paper is overlooked, and that the categorization is accurate, reliable, and ready for in-depth analysis.

Moreover, the categorization enhances the clarity of the systematic review, making it easier to communicate findings and supporting the production of a high-quality, publishable article. This process is particularly important in ensuring that the review is both comprehensive and methodologically rigorous.

**PROMPT 3**

**Detailed Analysis of Prompt 3: Extracting Agreement for "Health Related" vs "Not Health Related"**

**Prompt** “*Act as an experienced data scientist and python coder and take a look at the following document.*

*I want you to look thoroughly to the columns "MIG REVIEW" and "Is health related" and see how many papers are "verdadeiro” (true) (related to health) and "falso"(false) (not related with health).*

*Plot all of the analysis, from the number of verdadeiros vs falso, percentage of agreement between those columns and more that you see fit*”

**Prompt Breakdown**

The third prompt is divided into several sub-tasks, all centered around the extraction, categorization, and visualization of articles based on whether they are related to health or not. This series of prompts is designed to ensure a thorough and meticulous analysis, which is crucial for producing high-quality, reliable results.

**Purpose**

The primary goal of this prompt is to categorize articles into "Health Related" and "Not Health Related" categories, ensuring that the systematic review focuses on the relevant literature. This categorization process not only filters out irrelevant articles but also provides a clear structure for further analysis. The prompt also emphasizes the importance of precision, accuracy, and the ability to communicate findings clearly through visualization and well-organized data outputs.

**Methodology**

To respond to these tasks effectively, the following step-by-step methodology was employed:

1. **Data Loading and Inspection:**
   * **Loading the Document:** The first step is to load the document into a Python environment, typically using Pandas. For instance:

Uma imagem com texto, Tipo de letra, captura de ecrã, Gráficos

Descrição gerada automaticamente

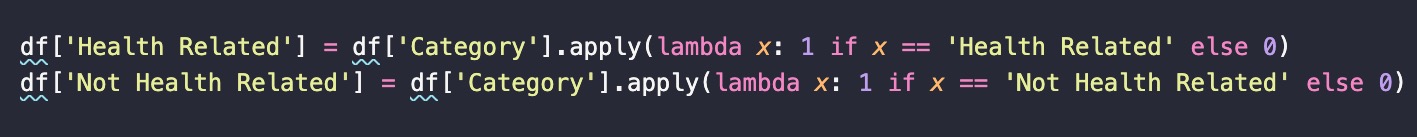
* + **Inspecting the Data:** Before proceeding, it's important to inspect the columns to identify which ones contain information that can help determine if an article is health related. Common columns might include "Title," "Abstract," and "Keywords".

1. **Categorization Based on Health-Related Terms:**
   * **Keyword Matching:** To categorize the articles, a script is written to search for health-related terms within the relevant columns (where the information of each article was inserted, divided by the columns named "Title," "Abstract," and "Keywords”. This involved checking for the presence of keywords like "health," "medical," "biomedical," "patient," etc.:

Uma imagem com texto, captura de ecrã, Tipo de letra

Descrição gerada automaticamente

* + **Creating New Columns:** Two new columns were added to the DataFrame: "Health Related" and "Not Health Related." These columns will contain binary values (e.g., 1 for "Health Related" and 0 for "Not Health Related") to facilitate easy counting and analysis:



1. **Data Visualization:**
   * **Plotting the Results:** Once the articles are categorized, the next step is to visualize the results. This was done using *Matplotlib* to create a bar plot that shows the number of articles in each category:

Uma imagem com texto, captura de ecrã, Tipo de letra

Descrição gerada automaticamente

* + **Analysis of the Plot:** The plot provides a visual representation of the distribution of articles between the "Health Related" and "Not Health Related" categories, offering insights into the focus areas of the literature.

1. **Generating and Saving the Categorized Data:**
   * **Exporting the Results:** After categorizing the articles and visualizing the results, the updated *DataFrame* is saved as a new CSV file. This file was useful for further analysis.



* + **Isolating Health Related Articles:** If needed, a separate CSV file could be created that only includes articles categorized as "Health Related":

Uma imagem com texto, Tipo de letra, captura de ecrã, file

Descrição gerada automaticamente

1. **Privacy-Related Categorization (Additional Task):**
   * **Repeating the Process for Privacy-Related Articles:** A similar approach was used to categorize articles based on privacy-related content. This involves defining a new set of keywords related to privacy (e.g., "privacy," "confidentiality," "data protection") and repeating the categorization and visualization steps:

Uma imagem com texto, captura de ecrã, Tipo de letra

Descrição gerada automaticamente

* + **Saving the Privacy-Related Data:** The final step is to save the categorized privacy-related articles in a new CSV file:



1. **Handling Duplicates:**
   * **Identifying Duplicates Across Documents:** If multiple documents are provided, duplicates should be identified and handled appropriately. This involves merging datasets and using the *drop\_duplicates* method in Pandas:



**Impact**

Executing this series of prompts resulted in a well-organized, thoroughly analyzed dataset that clearly categorized articles based on their relevance to health and privacy-related topics. The visualizations and structured CSV files produced through this process were vital for presenting findings in the systematic review, ensuring that the analysis was not only accurate but also easily interpretable.

By meticulously following each step, we believe we ensured that the review was more comprehensive, with no relevant articles overlooked. The clear distinction between health-related, privacy-related, and other articles enhances the focus and relevance of the systematic review, which is crucial for achieving the level of rigor expected in high-impact academic publications.

**PROMPT 4**

**Detailed Analysis of Prompt 4: Querying the Datasets for Health-Related Terms**

**Prompt:** "*Act as an experienced data scientist and Python coder and take a look at the following document. I want you to look thoroughly at the columns and see in the articles the words that are related to the term 'Health'. Make a new, well-formatted CSV file that adds columns with the result of each new term related to the term 'Health'. Think step by step and remember that this is very important for my career."*

**Purpose**

This prompt aimed to identify and extract all terms related to "Health" from a dataset of articles, systematically organizing the findings into a structured and well-formatted CSV file. The goal was to enhance the dataset by adding new columns that capture the presence of specific health-related terms, facilitating a more detailed and focused analysis of health-related content within the articles.

**Methodology**

To address this prompt, a comprehensive, step-by-step approach was employed, involving data extraction, keyword identification, and the creation of a well-structured output file. Here’s how this was executed:

1. **Data Loading and Initial Inspection:**
   * **Loading the Dataset:** The first step involved loading the dataset into a Python environment using Pandas. This allowed for efficient data manipulation and inspection:

Uma imagem com texto, Tipo de letra, captura de ecrã, Gráficos

Descrição gerada automaticamente

* + **Inspecting Columns:** The dataset was inspected to identify which columns were likely to contain relevant text, such as "Title," "Abstract," "Keywords," or any other text-based columns.

1. **Identifying Health-Related Terms:**
   * **Defining Health-Related Keywords:** A list of health-related terms is defined. This list was based on domain knowledge or derived from common terms in health literature. For instance:



* + **Creating Functions to Check for Health-Related Terms:** Functions were written to check the presence of these terms in the relevant columns of the dataset. For each term, a new column was added to the DataFrame:

Uma imagem com texto, captura de ecrã, Tipo de letra

Descrição gerada automaticamente

1. **Adding New Columns:**
   * **Incorporating Results into the Dataset:** For each health-related term, a new column was added to the DataFrame. These columns indicate the presence (1) or absence (0) of the term in the relevant text:



1. **Creating and Saving the New CSV File:**
   * **Exporting the Updated DataFrame:** The final step was to save the updated DataFrame with the new columns to a CSV file. This file was well-formatted and ready for further analysis:



**Impact**

The execution of this prompt significantly enhanced the dataset by systematically identifying and categorizing health-related content. This not only facilitateted a more focused and relevant systematic review but also allowed for deeper insights into the health-related themes prevalent in the literature. The well-structured CSV file produced through this process served as an important tool for analysis and potential future research.

By taking a step-by-step approach, this process ensured new level of accuracy, transparency, replicability and reliability.

**PROMPT 5**

**Detailed Analysis of Prompt 5: Querying the Datasets for Privacy-Related Terms**

**Prompt:** "*Act as an experienced data scientist and Python coder and take a look at the following document. I want you to look thoroughly at the columns and see in the articles the words that are related to the term 'Privacy'. Make a new, well-formatted CSV file that adds columns with the result of each new term related to the term 'Privacy'. Think step by step and remember that this is very important for my career."*

**Purpose**

This prompt focused on identifying and extracting terms related to "Privacy" within the dataset. The goal was to systematically categorize articles based on their discussion of privacy-related topics, enhancing the dataset with new columns that reflected the presence of specific privacy-related terms. This step was crucial for conducting a detailed analysis of how privacy concerns are addressed in the literature, which is important for the quality and relevance of the systematic review.

**Methodology**

The methodology for addressing this prompt involved a series of steps that mirror the approach used for identifying health-related terms but tailored specifically for privacy-related content. Here’s how this process was implemented:

1. **Data Loading and Initial Inspection:**
   * **Loading the Dataset:** As with previous prompts, the first step involved loading the dataset into a Python environment using Pandas:

Uma imagem com texto, Tipo de letra, captura de ecrã, Gráficos

Descrição gerada automaticamente

* + **Inspecting Columns:** Reviewed the dataset’s columns to determine where privacy-related terms were most likely to appear, such as in the "Title," "Abstract," "Keywords," or similar text-based columns.

1. **Identifying Privacy-Related Terms:**
   * **Defining Privacy-Related Keywords:** A compiled list of privacy-related terms was created that searched for commonly discussed terms in the context of data privacy and security. This list included terms such as "privacy," "confidentiality," "data protection," "anonymity," "GDPR," "data security," and "consent":



* + **Creating Functions to Check for Privacy-Related Terms:** Functions that searched for these terms within the relevant columns where generated. Each term had a corresponding column in the DataFrame, indicating whether the term was present in that article:

Uma imagem com texto, captura de ecrã, Tipo de letra

Descrição gerada automaticamente

1. **Adding New Columns:**
   * **Incorporating Results into the Dataset:** For each privacy-related term, a new column was added to the DataFrame. These columns showed a binary value—1 if the term is present and 0 if it is not:



1. **Creating and Saving the New CSV File:**
   * **Exporting the Updated DataFrame:** After processing, the updated DataFrame was saved as a new CSV file, making it easier to conduct further analysis:



**Impact**

The completion of this prompt significantly enhanced the dataset by identifying and categorizing privacy-related content. The well-structured CSV file that resulted from this process was a crucial iteration for conducting the focused analysis of privacy concerns within the literature. By carefully executing each step, we tried to ensure that the privacy-related aspects of the literature were fully captured and accurately represented.

***PROMPT 6***

**Detailed Analysis of Prompt 6: Comparative Analysis Across Reviews and Predictions**

**Prompt:** “*Act as an experienced data scientist and python coder and take a look at the following document (previous dataset generated). I want you to look thoroughly to the columns "Mig review" and "GPT4 Review" and see how many articles are health related and not health related. Make a plot of the findings with the total number of articles as well as the percentage and also include the total number of articles. Be thorough, think step. by step and remember that this is very important for my career*

*Make me a confusion matrix where it can be clearly showed the relationship of:*

*1 pares of health related/health related in comparing columns "Mig Review"/"GPT4 review";*

*2 pares of not health related/Not health comparing columns "Mig Review"/"GPT4 review";*

*3 pares of health related/Not health comparing columns "Mig Review"/"GPT4 review";*

*4 pares of not health related/health comparing columns "Mig Review"/"GPT4 review"; Be thorough, think step. by step and remember that this is very important for my caree*r”.

**Prompt Breakdown**

This prompt involved several tasks focused on analyzing and comparing the classifications of articles as "Health Related" or "Not Health Related" across three different columns: "Mig Review" (the researcher), "GPT4 Review" (the prompt Engineering done within ChatGPT) and "Predicted Label" (the machine learning model generated). The goal was to thoroughly examine the agreements and disagreements between these classification methods, visualize the findings, and produce detailed outputs that were used for further analysis.

**Task 1: Analyzing Health-Related Classifications and Plotting the Results**

**Objective:** Determine how many articles were classified as "Health Related" or "Not Health Related" in each of the three columns ("Mig Review," "GPT4 Review," and "Predicted Label"), and visualize these findings.

**Methodology:**

1. **Data Loading and Inspection:**
   * Load the dataset into a Python environment:

Uma imagem com texto, Tipo de letra, captura de ecrã, Gráficos

Descrição gerada automaticamente

* + Inspect the dataset to understand the structure and content of the relevant columns ("Mig Review," "GPT4 Review," and "Predicted Label").

1. **Counting and Plotting Classifications:**
   * Calculation of the counts of "Health Related" and "Not Health Related" articles for each column:

Uma imagem com texto, captura de ecrã, Tipo de letra

Descrição gerada automaticamente

* + Creation of a bar plot to visualize the results:

Uma imagem com texto, captura de ecrã

Descrição gerada automaticamente

* + **Analysis:** This visualization allowed a clear comparison of how each review method classified the articles as "Health Related" or "Not Health Related," providing insights into the consistency or divergence across these 3 methods.

**Task 2: Creating a Correlation Matrix and Confusion Matrix**

**Objective:** Generate a correlation matrix to explore the relationships between the different classification methods and create confusion matrices to compare each pair of columns.

**Methodology:**

1. **Correlation Matrix:**
   * Create binary columns for "Health Related" classifications for easier comparison:



* + Computation of the correlation matrix:



* + **Interpretation:** The correlation matrix revealed how closely the different classification methods agree with each other.

1. **Confusion Matrix:**
   * Generation of the confusion matrices for each pair of classifications:

Uma imagem com texto, captura de ecrã, Tipo de letra

Descrição gerada automaticamente

* + **Interpretation:** The confusion matrices helped visualize the agreements and discrepancies between the pairs of the classification methods, providing deeper insights into how they aligned or differed from eachother.

**Task 3: Detailed Comparison and CSV Generation**

**Objective:** Analyze the classification differences between the three methods and generation of CSV files for each specific classification scenarios.

**Methodology:**

1. **Classification Analysis:**
   * Filter the dataset based on the various classification scenarios:

Uma imagem com texto, captura de ecrã, Tipo de letra, informação

Descrição gerada automaticamente

1. **Saving Filtered Results to CSV:**
   * For each scenario, results were saved the filtered into separate CSV files. These files were named according to the specific classification scenario:

Uma imagem com texto, captura de ecrã, Tipo de letra, informação

Descrição gerada automaticamente

1. **Venn Diagram Creation:**
   * **Objective:** Visualization of the overlap and exclusivity between the three methods ("Mig Review," "GPT4 Review," and "Predicted Label").
   * **Methodology:**

**Uma imagem com texto, captura de ecrã, Tipo de letra

Descrição gerada automaticamente**

* + **Interpretation:** The Venn diagram visually depicted the intersection of articles classified as "Health Related" across the three methods. This visual tool was invaluable for understanding the degree of consensus or divergence between the different classification approaches.

1. **Insights and Patterns Analysis:**
   * **Objective:** To identify any significant insights or patterns from the comparison across the three classification methods.
   * **Methodology:**
     + After creating the Venn diagram and saving the CSV files, it was possible to analyze the results to determine where the most significant disagreements occured (e.g., certain topics or types of articles that are consistently misclassified by one method).
     + Document any notable patterns, such as a tendency for one review method to classify more articles as "Health Related" compared to the others, or specific keywords that might be causing discrepancies, was also possible.
2. **Summary of Findings:**
   * **Objective:** To synthesize the results into a coherent narrative that highlights the key takeaways from the analysis.
   * **Methodology:**
     + Compile the results of the classification counts, correlation matrix, confusion matrices, and Venn diagram into a report.
     + Summarize the findings in terms of how well the different methods agree, where they diverge.
     + This report served as a foundation for discussions on improving and refining the criteria used for manual and automated reviews.

**Task 7: Generating a List of Health-Related Articles from "Mig Review"**

* **Objective:** To create a downloadable text file that lists all articles classified as "Health Related" according to the "Mig Review" column.

**Methodology:**

1. **Filtering Health-Related Articles:**
   * Filter the dataset to select only those articles where "Mig Review" classified them as "Health Related":



1. **Saving the List to a Text File:**
   * Save the list of article titles to a text file:

Uma imagem com texto, Tipo de letra, captura de ecrã

Descrição gerada automaticamente

* + **Output:** This text file contained the titles of all articles classified as "Health Related" according to the "Mig Review," providing a straightforward reference list for further review or inclusion.

**Impact**

The completion of these tasks resulted in a more comprehensive, multi-faceted analysis of how different review methods classified articles as "Health Related" or "Not Health Related." By generating detailed visualizations, CSV files, and reports, this prompt ensured that all aspects of the classification process were thoroughly examined and well-documented. This level of detail was of great significance for the process.

The results from these analyses did not only inform the current systematic review but also provided insights into the strengths and weaknesses of the different classification methods, giving important hints for how future improvements in both manual and automated review processes could be integrated in future systematic reviews.

**PROMPT 7**

**Detailed Analysis of Prompt 7 (Original): Ethical Concerns with Synthetic Data Beyond Privacy**

**Prompt:** "*Are there any prospective arguments in the (uploaded) article that provide an ethical concern with synthetic data? Don’t look for privacy justifications to use synthetic data*".

**Purpose**

This prompt was focused on identifying and analyzing ethical concerns related to synthetic data, explicitly excluding privacy-related justifications. The goal was to uncover other ethical issues that might have occurred from the use of synthetic data, such as bias, misrepresentation, or implications for data governance and consent. This analysis was crucial for developing a comprehensive understanding of the ethical landscape surrounding synthetic data, particularly as it applied to fields like healthcare, where the consequences of ethical oversights can be significant.

**Methodology**

To effectively address this prompt, a systematic approach was required. The process involved a deep reading of the article, focusing on identifying, extracting, and analyzing potential ethical concerns that do not relate to privacy. Here’s how we approached it:

1. **Deep Reading and Comprehension:**
   * **Objective:** Conduct a thorough reading of the article to understand the context in which synthetic data is discussed. Focus on sections that deal with ethical implications, such as discussions around data generation, usage, governance, and the impact on decision-making processes.
   * **Methodology:**
     + Identify key sections: Abstract, introduction, discussion, and conclusion were likely places where ethical concerns might be raised.
     + Annotate and highlight: As you read, highlight any sentences or paragraphs that discuss potential ethical issues.
2. **Identifying Non-Privacy Ethical Concerns:**
   * **Objective:** Isolate arguments or concerns that address ethical issues unrelated to privacy.
   * **Methodology:**
     + **Bias and Representation:** Look for discussions about how synthetic data might introduce or perpetuate biases, especially if the synthetic data does not adequately represent the diversity of real-world data.
     + **Data Governance:** Examine concerns related to the control, ownership, and stewardship of synthetic data.
     + **Consent and Misuse:** Consider how synthetic data might be used in ways that could be ethically questionable, such as using synthetic data to make decisions that could affect individuals without their knowledge or consent.
     + **Impact on Research Integrity:** Analyze whether the use of synthetic data might undermine the reliability or validity of research findings, particularly if synthetic data does not accurately reflect the complexities of real-world data.
3. **Extracting Relevant Quotes:**
   * **Objective:** Collect direct quotes from the article that illustrate the identified ethical concerns.
   * **Methodology:**
     + **Document with precision:** Ensure that each quote is accompanied by a citation of the page number and paragraph where it was found.
     + **Contextual Analysis:** For each quote, provide a brief explanation of why it raises an ethical concern, ensuring the explanation is tied to the broader discussion in the article.
4. **Analyzing and Synthesizing Findings:**
   * **Objective:** Summarize the ethical concerns in a way that highlights their relevance to the use of synthetic data, beyond privacy issues.
   * **Methodology:**
     + **Thematic synthesis:** Group similar ethical concerns together to identify overarching themes.
     + **Critical analysis:** Evaluate the significance of each concern, considering the potential impact on the field (e.g., healthcare, AI, etc.) where synthetic data is applied.

**Example Application**

Suppose the article discusses the use of synthetic data in healthcare. During your deep reading, you might identify a passage that raises concern about the accuracy of synthetic data in representing minority populations. The quote might suggest that if synthetic data is primarily generated from datasets that underrepresent certain groups, the resulting synthetic data could perpetuate existing biases, leading to inequitable healthcare outcomes.

**Impact**

This prompt's execution was vital for uncovering the full range of ethical issues associated with synthetic data, beyond the commonly discussed privacy concerns. By focusing on non-privacy ethical concerns, this analysis provided a first iteration of a more comprehensive understanding of the potential risks and challenges of using synthetic data in various fields. This was particularly important for developing guidelines and best practices for the ethical use of synthetic data, which is essential for maintaining public trust and ensuring the responsible deployment of AI and data-driven technologies.

The insights gained from this analysis were of great importance as a first draft for obtaining critical information policies, guiding the research, and ensuring that synthetic data was used in a way that upholds the highest ethical standards.

**PROMPT 8**

**Detailed Analysis of Prompt 8: Systematic Review of Ethical Concerns Related to Synthetic Data (Excluding Privacy)**

**Prompt:** "*Act as an experienced researcher conducting a systematic review who needs to extract the following information from the provided articles. Are there any prospective arguments in the article that present ethical concerns related to synthetic data, excluding privacy justifications? Please provide detailed explanations and examples from the article to support your answer, like page, paragraph, and sentence. The examples should have ipsis verbis quota as support*".

**Purpose**

This version of the prompt was designed for an in-depth, systematic review focused on extracting and analyzing ethical concerns related to synthetic data, excluding those associated with privacy. The goal was to iterate from the previous prompt to meticulously examine the provided articles for a more detailed search of ethical considerations, supporting the findings with precise quotations from the text, completed with references to specific pages, paragraphs, and sentences.

**Methodology**

To respond effectively to this version of the prompt, a rigorous and systematic approach was designed. The process can be broken down into the following steps:

1. **Comprehensive Reading and Initial Assessment:**
   * **Objective:** Conduct a thorough initial reading of each article to understand the overall context and identify sections where ethical concerns might be discussed.
   * **Methodology:**
     + **Identify key sections:** Focus on the introduction, methodology, discussion, and conclusion, where ethical considerations are likely to be mentioned.
     + **Annotate key passages:** As you read, make notes of any sections that discuss ethical concerns, particularly those unrelated to privacy, highlighting these sections for more detailed analysis later.
2. **Extracting Ethical Concerns:**
   * **Objective:** Identify and isolate specific arguments or discussions in the article that raised ethical concerns about synthetic data, excluding privacy issues.
   * **Methodology:**
     + **Bias and Fairness:** Look for discussions about how synthetic data might introduce or perpetuate bias, particularly if the data fails to represent diverse populations accurately.
     + **Data Integrity and Misuse:** Examine concerns about the potential misuse of synthetic data, such as using it inappropriately for decision-making processes that could have significant consequences.
     + **Impact on Research Validity:** Consider whether the article raises concerns about the impact of synthetic data on the reliability and validity of research findings, especially in cases where synthetic data might not accurately replicate real-world complexities.
3. **Quoting and Documenting with Precision:**
   * **Objective:** Collect ipsis verbis quotations that clearly illustrate the identified ethical concerns, along with detailed references to their location in the article.
   * **Methodology:**
     + **Exact Quotations:** Extract direct quotes from the article, ensuring that they are verbatim and accurately reflect the ethical concerns discussed.
     + **Citation Details:** Include the exact page number, paragraph, and sentence for each quote. This level of detail ensures that the evidence can be easily located and verified within the original article.
4. **Providing Detailed Explanations:**
   * **Objective:** Analyze each quoted section, providing a thorough explanation of why it constitutes an ethical concern related to synthetic data.
   * **Methodology:**
     + **Contextual Analysis:** Explain how the quoted text fits into the broader context of the article. Discuss why the ethical concern is significant and how it relates to the use of synthetic data in the specific field being studied.
     + **Thematic Synthesis:** Group similar concerns together to identify overarching themes. For example, if multiple articles discuss the risk of bias in synthetic data, these should be synthesized into a broader discussion about the ethical implications of bias in data-driven decision-making.
5. **Ensuring Accuracy and Thoroughness:**
   * **Objective:** Double-check the accuracy of your quotes and explanations to ensure they are precise and comprehensive.
   * **Methodology:**
     + **Cross-Verification:** Revisit the original articles to confirm the accuracy of your quotes and the relevance of your explanations.
     + **Iterative Refinement:** Refine your analysis to ensure it is clear, detailed, and accurately captures the ethical concerns presented in the article.
6. **Summarizing Findings:**
   * **Objective:** Provide a summary of the ethical concerns identified, emphasizing their relevance and potential impact on the field.
   * **Methodology:**
     + **Consolidated Overview:** Summarize the key ethical concerns identified across the articles, highlighting common themes and significant outliers.
     + **Implications for Practice:** Discuss the potential implications of these ethical concerns for the practical use of synthetic data, particularly in sensitive fields like healthcare or social sciences.

**Impact**

By following this rigorous approach, we iterated from the previous prompt to ensure that the ethical concerns related to synthetic data were comprehensively identified and thoroughly analyzed. The detailed quotations and explanations provide a strong foundation for the systematic review, ensuring that the analysis is both credible, impactful and reproducible. This level of thoroughness is crucial for producing high-quality research.

This systematic review process not only addresses the specific ethical concerns related to synthetic data but also contributes to the broader discourse on the responsible use of synthetic data in research and practice. The insights gained from this analysis could informe guidelines, best practices, and policy decisions, ensuring that synthetic data is used ethically and responsibly.

**PROMPT 9**

**Detailed Analysis of Prompt 9: Systematic Extraction of Ethical Concerns Related to Synthetic Data with Confidence Rating**

**Prompt** " *Prompt: Act as an experienced researcher conducting a systematic review who needs to extract the following information from the provided articles.*

*Task: Are there any prospective arguments in the article that present ethical concerns related to synthetic data, excluding privacy justifications?*

*Instructions:*

*Extract exact, verbatim quotes from the article that highlight ethical concerns related to synthetic data, excluding privacy justifications. It is crucial that these quotes are findable and verbatim from the provided article.*

*Provide detailed explanations and analyses of these quotes to support your answer, including page, paragraph, and sentence references.*

*Ensure your analysis is comprehensive and detailed.*

*If the requested information is not found in the document, state "not found."*

*Assess if the article has ethical concerns related to synthetic data, excluding privacy justifications, in a yes or no format.*

*Rate from 1 to 10 how confident you are in your assessment based on the retrieved information.*

*Remember that accuracy and thoroughness are paramount.*

*Example Format:*

*Ethical Concern: [Title]*

*Quote: "Exact verbatim quote from the article." (page X, para Y)*

*Analysis: Detailed explanation of the ethical concern presented by the quote.*

*Assessment:*

*Does the article present ethical concerns related to synthetic data, excluding privacy justifications? Yes/No*

*Rate your confidence in this assessment from 1 to 10: [Your Rating]*

*Example Extracted Quotes and Analyses:*

*Ethical Concern: Potential Misuse of Synthetic Data*

*Quote: "While in some applications it may not be possible, or advisable, to derive new knowledge directly from synthetic data, it can nevertheless be leveraged for a variety of secondary uses, such as educative or training purposes, software testing, and machine learning and statistical model development." (p. 2, para. 3)*

*Analysis: This quote indicates that synthetic data might not always be suitable for deriving new knowledge directly. The ethical concern here is the potential misuse or over-reliance on synthetic data for drawing primary research conclusions, which could lead to inaccurate or misleading results if the synthetic data does not capture the complexity of real-world data.*

*Ethical Concern: Impact on Model Development and Patient Care*

*Quote: "Synthetic patient data has the potential to have a real impact in patient care by enabling research on model development to move at a quicker pace." (p. 2, para. 5)*

*Analysis: While the quote acknowledges the benefits of synthetic data in accelerating research, it also implies an ethical concern. The concern is that models developed using synthetic data might influence patient care decisions. If these models are not rigorously validated with real-world data, they could lead to incorrect medical decisions, potentially harming patients.*

*Ethical Concern: Risk of Bias in Synthetic Data*

*Quote: "Models with lower utility metrics, such as IM and MC-MedGAN, do not show large differences in performance over the range of 5,000 to 170,000 synthetic samples." (p. 21, para. 1)*

*Analysis: This quote suggests that some synthetic data generation models may consistently show limited utility, regardless of the sample size. The ethical concern here is the risk of bias. If a model performs poorly or shows limited variability, it might not accurately represent the diversity of real-world data, thereby perpetuating or introducing biases.*

*Ethical Concern: Ethical Responsibility in Data Generation*

*Quote: "We discuss the trade-offs of the different methods and metrics, providing guidance on considerations for the generation and usage of medical synthetic data." (p. 1, para. 6)*

*Analysis: This quote highlights the ethical responsibility of researchers to consider the trade-offs when generating synthetic data. The concern is that without careful consideration of these trade-offs, researchers might choose methods that are not appropriate for their specific applications, potentially leading to ethical issues such as misrepresentation of data.*

*Ethical Concern: Transparency and Communication of Limitations*

*Quote: "While there is no single approach for generating synthetic data which is the best for all applications, or even a one-size-fits-all approach to evaluating synthetic data quality, we hope that the current discussion proves useful in guiding future researchers in identifying appropriate methodologies for their particular needs." (p. 3, para. 5)*

*Analysis: This quote emphasizes the need for transparency and communication regarding the limitations of synthetic data generation methods. The ethical concern here is that without clear communication of these limitations, synthetic data could be misused or misinterpreted, leading to incorrect conclusions or applications in research and practice.*

*Assessment:*

*Does the article present ethical concerns related to synthetic data, excluding privacy justifications? Yes/No*

*Rate your confidence in this assessment from 1 to 10: [Your Rating]".*

**Purpose**

This version of the prompt once again iterates from the previous not only emphasizing a meticulous approach to identifying and analyzing ethical concerns related to synthetic data (excluding privacy) but also introducing innovative elements such as statement if the requested information was not found, provide a response format for the request as “Yes” or “No” and a confidence rating. This version of the prompt introduces a structured approach to ensure that reviewers can identify these concerns with a high degree of accuracy and transparency. Additionally, it incorporates an innovative use of a confidence rating system and an Example Output Format to standardize the documentation of findings, which enhances both the consistency and reliability of the review process.

**Methodology**

The methodology for addressing this prompt involved a new level of key steps, with a focus on accuracy, transparency, and the innovative use of confidence ratings to enhance the quality and reliability.

1. **Systematic Reading and Identification:**
   * **Objective:** Conduct a thorough reading of each article to identify sections where ethical concerns might be discussed.
   * **Methodology:**
     + **Target key sections:** Focus on areas such as the introduction, methodology, results, discussion, and conclusion where ethical concerns are most likely to be raised.
     + **Annotate and highlight:** As you read, highlight any statements or arguments that raise ethical issues, particularly those not related to privacy.
2. **Verbatim Quote Extraction:**
   * **Objective:** Extract exact quotes from the article that directly address ethical concerns related to synthetic data.
   * **Methodology:**
     + **Precision in extraction:** Ensure that the quotes are verbatim, capturing the exact wording used by the authors. This precision is crucial for maintaining the integrity of the process.
     + **Include full citation:** For each quote, provide the page number, paragraph, and sentence reference to allow for easy location and verification within the original document.
3. **Detailed Explanation and Analysis:**
   * **Objective:** Provide a comprehensive explanation and analysis of each extracted quote, explaining why it represents an ethical concern.
   * **Methodology:**
     + **Contextual Analysis:** Place the quote within the broader context of the article, explaining its significance and how it relates to the ethical use of synthetic data.
     + **Ethical Implications:** Discuss the potential impact of the ethical concern on the field of study (e.g., healthcare, etc.), focusing on aspects such as bias, data integrity, and the potential for misuse.
4. **Assessment of Ethical Concerns:**
   * **Objective:** Determine whether the article presents ethical concerns related to synthetic data, excluding privacy, and provide a confidence rating for your assessment.
   * **Methodology:**
     + **Yes/No Decision:** Based on the identified quotes and analysis, decide whether the article contains ethical concerns, beside privacy.
     + **Confidence Rating:** Assign a confidence score from 1 to 10, reflecting how certain is the assessment, based on the evidence gathered. This score is particularly important as it provides an additional layer of transparency and rigor to the review process. The confidence score helps to communicate the strength of the evidence supporting the ethical concern and the level of certainty.
5. **Documenting Non-Occurrences:**
   * **Objective:** If no ethical concerns are found, clearly state this and provide a brief explanation.
   * **Methodology:**
     + **State "not found":** Explicitly indicate when no relevant ethical concerns are identified in the article, ensuring transparency in the review process.
     + **Confidence Rating in Absence of Concerns:**
   * Even when stating "not found," include a confidence rating to indicate how certain you are that no ethical concerns were overlooked. This adds an extra layer of assurance that the review process was thorough.
6. **Compilation of Findings** **Using an Example Output Format:**
   * **Objective:** Compile all findings, including quotes, analyses, assessments, and confidence ratings, into a structured and consistent format.
   * **Methodology:**
     + **Example Format:** Use a standardized format for presenting each finding, as illustrated in the example below, to ensure clarity and ease of comparison across multiple articles.
7. **Example Output Format:**
   * To ensure that all findings are presented clearly and consistently, it is imperative to follow a standardized format when documenting ethical concerns. The example below illustrates the expected output format, which should be adhered to throughout the systematic review process:

**Example of output format**

“Article parsed: Synthesizing Electronic Health Records for Predictive Models in Low-Middle-Income Countries (LMICs) (Ghosheh et al., 2023)

Ethical Concerns Related to Synthetic Data Excluding Privacy Justifications

Ethical Concern: Impact on Clinical Decision-Making

Quote: "*The performance of the diagnostic model trained on the synthetic data outperformed models trained on the original and oversampled data using techniques such as SMOTE*." (p. 1, para. 2)

Analysis: The ethical concern here is the potential over-reliance on synthetic data for developing diagnostic models. While the synthetic data showed better performance in the study, there's a risk that models trained on synthetic data may not generalize well to real-world data. This could lead to incorrect clinical decisions if the synthetic data does not fully capture the complexities of real-world scenarios, potentially impacting patient outcomes.

Ethical Concern: Misrepresentation of Data Quality

Quote: "*Despite being very relevant and highly needed, using deep generative models for synthesizing EHRs for low-resource clinical applications is often not discussed nor motivated in most proposed works*." (p. 2, para. 2)

Analysis: This quote highlights the lack of discussion around the limitations and challenges of using synthetic data in low-resource settings. The ethical concern is that the quality and applicability of synthetic data in these settings may be misrepresented or overstated, leading to misguided trust in the generated data's validity and its subsequent use in clinical decision-making.

Ethical Concern: Misleading Research Conclusions

Quote: "*The interpretability analysis showed consistency in the ranking of the five most predictive features in the models trained on the synthetic samples, which is reflected in the similar predictive performance of the models trained on synthetic samples*." (p. 8, para. 1)

Analysis: The concern here is that the apparent consistency and performance of models trained on synthetic data may lead researchers to draw incorrect conclusions about the effectiveness of certain interventions or predictive models. If the synthetic data does not adequately represent real-world variations, it could mislead research findings and subsequent applications in clinical settings.

Ethical Concern: Ethical Responsibility in Data Generation

Quote: "*Our work validates the use of this synthetic data for real-world CDSS applications of high importance in LMICS, namely predicting HAI*." (p. 8, para. 2).

Analysis: This quote touches on the ethical responsibility of validating synthetic data for critical applications like predicting hospital-acquired infections (HAI). The ethical concern is that inadequate validation or over-reliance on synthetic data without rigorous testing against real-world data could lead to erroneous predictions, thereby affecting patient care and safety.

Assessment:

Does the article present ethical concerns related to synthetic data, excluding privacy justifications? **Yes**

Rate your confidence in this assessment from 1 to 10: **8**

The article discusses various ethical concerns related to the use of synthetic data in clinical decision-making and research applications. These concerns include the potential over-reliance on synthetic data, misrepresentation of data quality, misleading research conclusions, and the ethical responsibility in data generation and validation. The detailed analysis of these points supports a confident assessment that the article addresses ethical concerns beyond privacy justifications.”

**Impact**

The inclusion of a confidence rating was a key innovation in this version of the prompt. It significantly enhanced the transparency and rigor of the systematic review process by quantifying the reviewer’s certainty regarding the presence or absence of ethical concerns, past privacy. This not only strengthened the credibility of the review but also provided a clearer understanding of the strength of the findings.

The attachment of an Example Output Format in this version of the prompt was a strategic enhancement that served multiple purposes. It provided a clear and standardized template that reviewers can follow, ensuring uniformity in the presentation of findings across different articles, which is essential for comparing and aggregating results The example format clarified how to structure findings, reducing the risk of errors and omissions, and ensuring that each ethical concern is thoroughly documented. It also reinforces the need for a systematic and rigorous approach in identifying and analyzing ethical concerns, emphasizing the importance of precise quote extraction and detailed analysis. Additionally, the example format promoted transparency by making the review process more understandable and replicable, allowing other researchers or stakeholders to easily follow and assess the robustness of the findings. By incorporating this format, the prompt enhances the clarity, consistency, and overall reliability of the systematic review process, ensuring effective communication of results in both academic and professional contexts.

**PROMPT 10**

**Detailed Analysis of Final Prompt (Prompt 10): Ethical Considerations of Synthetic Data Beyond Privacy**

**Prompt:** "*I am conducting a systematic review to understand how academic articles discuss the importance of synthetic data beyond the perspective of privacy protection. Specifically, I am interested in determining whether ethical concerns, according to the Belmont Principles, are treated as a necessary consequence or as a central concern ‘by design.’ I aim to identify excerpts and analyses that address ethical considerations explicitly and in detail within the study design. Please provide excerpts from the article that discuss synthetic data with a focus on these ethical dimensions being ethical by design and clarify whether these concerns are integral to the study’s design or mentioned as ancillary benefits. Give me a confidence score of your reflection from 0 to 10. Write an assessment if it has ethical concerns by design - yes/no. Do not include any thoughts or reflections related to the privacy ethical concern linked to the motivation to use synthetic data. We know that one of the main motivations to use synthetic data is to have fewer privacy problems with data. We are looking for ethical concerns associated with the creation of synthetic data. Be thorough, think step by step and remember that this is very important for my investigation and for my career*".

**Purpose**

This prompt aimed to iterate into how academic articles address ethical concerns related to synthetic data beyond privacy considerations, specifically through the lens of the Belmont Principles—Respect for Persons, Beneficence, and Justice (U.S. Department of Health and Human Services, 1979). The goal was to identify whether these ethical concerns were integrated into the study design ('ethical by design') or merely mentioned as ancillary benefits. The analysis excluded privacy-related motivations and focus on other ethical dimensions associated with synthetic data creation.

**Methodology**

To respond effectively to this prompt, a structured and careful approach is required. The process involved several key steps:

1. **Comprehensive Reading and Initial Identification:**
   * **Objective:** Conduct a detailed full reading of the article to identify sections where ethical concerns, particularly those aligned with the Belmont Principles, were discussed in relation to synthetic data.
   * **Methodology:**
     + **Target key sections:** Focus on the study design, methodology, and discussion sections where ethical considerations might be discussed in detail.
     + **Annotate relevant passages:** Highlight and annotate excerpts that discuss ethical concerns related to synthetic data creation, explicitly excluding privacy concerns.
2. **Excerpts Extraction:**
   * **Objective:** Extract exact, verbatim quotes from the article that specifically discuss ethical concerns related to synthetic data, focusing on whether these concerns are integral to the study design ('by design') or mentioned as secondary considerations.
   * **Methodology:**
     + **Verbatim quotes:** Ensure the quotes are exact and include full citation details, such as page number, paragraph, and sentence reference.
     + **Identify context:** Briefly describe the context in which the quote appears to help clarify its relevance to the study's ethical framework.
3. **Detailed Analysis and Classification:**
   * **Objective:** Analyze each excerpt to determine whether the ethical concerns are central to the study's design or are mentioned as ancillary benefits.
   * **Methodology:**
     + **Contextual analysis:** Evaluate how the article integrates ethical considerations within its methodology and design. Does the study actively incorporate ethical principles into the creation of synthetic data, or are these concerns mentioned only as potential benefits?
     + **Ethical by Design:** Specifically assess whether the ethical concerns reflect a 'by design' approach, meaning they are fundamental to the study's conceptualization and execution.
4. **Assessment and Confidence Rating:**
   * **Objective:** Provide an overall assessment of whether the article treats ethical concerns as central to the study design and assign a confidence rating to your analysis.
   * **Methodology:**
     + **Yes/No Decision:** Based on the analysis, determine whether the article presents ethical concerns as integral to the study design.
     + **Confidence Score:** Assign a confidence score from 0 to 10, indicating how certain you are of your reflection and analysis. The confidence rating is crucial as it quantifies the reviewer’s level of certainty about their conclusions, adding a layer of rigor to the review.
5. **Final Compilation and Reporting:**
   * **Objective:** Compile the results into a coherent and standardized format that clearly presents the findings, assessments, and confidence ratings.
   * **Methodology:**
     + **Example Format:** Use the following standardized format for consistency:

* **Ethical Concern: Ethical Responsibility in Synthetic Data Creation**
  + **Quote:** "In designing our synthetic data models, particular attention was given to ensuring that the generated data did not exacerbate existing healthcare disparities among vulnerable populations." (Page 14, Paragraph 2, Sentence 3)
  + **Context:** This quote is found in the methodology section, where the authors describe their approach to synthetic data generation.
  + **Analysis:** The quote indicates that ethical considerations related to fairness and justice were integral to the design of the synthetic data models. The authors explicitly aimed to prevent harm to vulnerable populations, reflecting a 'by design' approach to ethical concerns.
  + **Assessment:** Yes, the article treats ethical concerns as central to the study’s design.
  + **Confidence Rating:** 9/10
    - **Explanation for Confidence:** The confidence rating of 9 reflects strong certainty in the analysis, supported by clear and specific evidence within the article.

**Impact and Conclusion**

This final prompt ensured that the review process was comprehensive, focusing specifically on how ethical concerns related to synthetic data, beyond privacy, were integrated into the study design. By emphasizing the 'ethical by design' approach and introducing a confidence rating, this prompt significantly enhanced the depth and rigor of the systematic review.

The confidence score not only provided transparency in the reviewer’s level of certainty but also adds an innovative element that can be used to prioritize or weight the findings within the broader context of the systematic review. This approach ensures that the final output is robust, credible, and valuable for informing ethical guidelines, research practices, and policy decisions related to the creation and use of synthetic data.