

DEMONSTRATING THE PRACTICAL UTILITY AND LIMITATIONS OF CHATGPT THROUGH CASE STUDIES

Matthew Walsh, Dominic Andrew Ross, Clarence Worrell, and Alejandro Gomez

August 2023

[Distribution Statement A] Approved for public release and unlimited distribution.

Executive Summary

In November 2022, OpenAI released ChatGPT, a chatbot powered by a large language model (LLM) called Generative Pretrained Transformer (GPT). Amazingly, ChatGPT reached 100 million users within two months. For comparison, YouTube, Instagram, and Facebook took 1.5, 2.5, and 4.5 years, respectively, to reach the same milestone.

Language models are not new. Joseph Weizenbaum developed one such model, Eliza, in the 1960s. The model used simple pattern matching and substitution techniques to simulate conversation with a human therapist [Weizenbaum 1976]. Since then, AI researchers have developed rule-based, statistical, and neural methods for natural language understanding and generation.

Most recently, LLMs have shifted the scientific frontier. GPT is one example. These models use an extremely expressive type of architecture called a deep neural network (DNN) to learn about the likelihood of words appearing in the context of different sentences and paragraphs. LLMs are trained on vast stores of data, comprising a sizeable percentage of the Internet. This gives LLMs great breadth of knowledge. In fact, LLMs are a special example of *foundational models*—general models in AI that form the basis for more specialized ones trained to be experts in a domain.

Despite the significant technical advances that have occurred in language models during the past 50-plus years, the release of ChatGPT was as a tipping point: For the first time, a language model entered widespread use. ChatGPT did so, in part, because of the greater accuracy of its responses relative to earlier language models and the emergent abilities of LLMs. Specifically, LLMs are capable of *in-context learning*—that is, adjusting how they respond based on user instructions. This ability allows LLMs to complete novel tasks that they were not trained for.

LLM's technical capabilities are impressive. But will they materially affect existing processes, and will they lead to the creation of new ones? In short, what is the practical utility of LLMs?

To answer this question, we conducted four in-depth case studies. In each case study, we used a version of GPT-3.5 provided in the ChatGPT web-based application to a complete task based on prompts we provided. The case studies described in this paper span multiple domains and call for vastly different capabilities:

- data science
- training and education
- research
- strategic planning

For each, we present the unaltered transcripts generated through our interactions with ChatGPT. We then comment on the modes of interaction with ChatGPT and note its strengths and limitations in the context of the specific case study.

We found that ChatGPT contributed to the quality of products generated and expedited their development. However, ChatGPT did not eliminate the need for human involvement: knowledgeable people were needed to decompose complex tasks into simpler ones that ChatGPT could accomplish, and they needed to verify its outputs. More nuanced findings we report pertain specifically to ChatGPT—they may not hold for other existing LLMs or ones under development. Yet the finding that ChatGPT can enhance productivity, but not replace human involvement, holds for all LLMs.

Case Study 1: Data Science

In today’s data-driven business landscape, the adoption of artificial intelligence (AI) has become a critical factor for organizations seeking to gain a competitive edge. In a 2022 survey of global executives, 50 percent of respondents reported their organizations had embraced AI in at least one business unit [McKinsey 2022]. Of those, approximately one-third reported reduced operational costs, while the remainder reported increased revenue. This highlights the compelling bottom-line value proposition associated with data science capabilities, including AI.

Despite the apparent benefits, many companies encounter challenges when integrating data science capabilities into operations and product lines. Data science products are fundamentally software products. Hence, data science teams must contend with software engineering obstacles. However, additional challenges arise due to the unique nature of machine learning (ML-) and AI-driven products. Some of the greatest challenges faced by data scientists include the following:

- **Quality:** Data science products are software-based and require adherence to quality assurance standards. The code they use must be syntactically and functionally correct. Additionally, the code must be optimized and secure.
- **Efficiency:** Even in companies with established data science teams, demand for data science products often surpasses capacity. Consequently, data science teams are challenged to develop reliable products efficiently, striking a delicate balance between speed and quality.
- **Sustainment:** Once a data science product is developed, it must be maintained and supported. The ongoing costs of these lifecycle management activities may be significant.

- Collaboration: Data science teams typically comprise individuals with overlapping yet distinct skill sets. Successful development and delivery of new capabilities require seamless collaboration among team members.

Using LLMs to Enhance Data Science Processes

Table 1 outlines the main challenges data science teams face and how LLMs can address them. The cells in orange marked “Medium” denote medium alignment for using an LLM to address that challenge, and the cells in green marked “High” denote high alignment.

For example, an LLM can retrieve information about an existing Python library to accomplish a programming goal. The LLM can generate syntax using the library, review code generated by human developers, and recommend changes based on runtime errors. All these activities improve the quality of the code and increase developer efficiency. Besides generating and reviewing code, the LLM can automatically produce documentation to explain what the various objects do. This makes it easier to sustain the code. Finally, the LLM can refactor existing code into different languages, making it possible for data science teams to integrate products created in R, Python, and other languages.

Table 1: Using LLMs to Address Challenges in Data Science

LLM Capabilities	Challenges			
	Quality	Efficiency	Sustainment	Collaboration
Information retrieval	High	High	Medium	Medium
Code review	High	High	Medium	High
Code completion	Medium	High	Medium	Medium
Debugging	Medium	High	Medium	Medium
Documentation	Medium	High	High	High
Automated test construction	High	High	High	High
Code translation	Medium	High	High	High

Data Science Case Study

The following case study displays a sequence of interactions intended to construct an intrusion detection system. The system learns a predictive model to distinguish between “malicious” connections (i.e., intrusions) and “benign” ones (i.e., normal connections). The system is trained using data collected as part of the 1998 DARPA Intrusion Detection Evaluation Program.¹ The dataset is used for

¹ <https://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>

illustrative purposes—the sequence of interactions in the example could be repeated for current cybersecurity problems.

PROMPT: Write me a Python script to do the following:

1. Load the `kddcup99` data using `sklearn.datasets`.
2. Assign predictors to the variable `X` and outcomes to the variable `Y`.
3. Convert the variable `X` to a Pandas dataframe and add column names.
4. Convert all elements of `Y` that contain the substring 'normal' to 0 and convert all other elements to 1.

ChatGPT: Certainly! Here's a Python script that performs the steps you've outlined:

```
import pandas as pd
from sklearn.datasets import fetch_kddcup99

# Load the KDD Cup 1999 dataset
kddcup99 = fetch_kddcup99()

# Assign predictors to variable X and outcomes to variable Y
X = kddcup99.data
Y = kddcup99.target

# Convert predictors to a pandas DataFrame and add column names
df = pd.DataFrame(X, columns=kddcup99.feature_names)

# Convert elements of Y containing 'normal' to 0 and others to 1
Y = [0 if 'normal' in outcome else 1 for outcome in Y]

# Print the first few rows of the DataFrame and the modified Y
print("Predictors (X):")
print(df.head())
print("\nOutcomes (Y):")
print(Y)
```

PROMPT: The script returns the error: `TypeError: a bytes-like object is required, not 'str'`.

ChatGPT: Apologies for the error. It seems that the `fetch_kddcup99()` function returns byte strings instead of regular strings. To resolve this issue, we can convert the byte strings to regular strings before performing any operations. Here's the modified script:

```
# Convert byte strings to regular strings
Y = [outcome.decode() for outcome in Y]
```

PROMPT: Convert all columns of `X` to float and retain non-NA columns.

ChatGPT: To convert all columns of X to float and retain only the non-NA columns, you can modify the script as follows:

```
# Convert all columns to float and retain non-NA columns
df = df.apply(pd.to_numeric, errors='coerce')
df = df.loc[:, ~df.isna().any()]
```

PROMPT: *Should I normalize predictors before or after creating the train-test split?*

ChatGPT: The general recommendation is to perform normalization (or standardization) of predictors after creating the training-test split. This approach helps ensure that the normalization process is applied independently to the training and test sets, mimicking real-world scenarios where the model encounters new, unseen data during deployment.

PROMPT: *Okay. Create an 80/20 train-test split.*

RESPONSE: To create an 80/20 training-test split using the DataFrame df and the outcome variable Y, you can utilize scikit-learn's train_test_split function. Here's the updated script:

```
from sklearn.model_selection import train_test_split

# Create an 80/20 training-test split
X_train, X_test, Y_train, Y_test = train_test_split(df, Y, test_size=0.2,
random_state=42)
```

PROMPT: *Now, standardize the predictors and fit a logistic regression model using SGDClassifier. Report the model's accuracy on the test set.*

ChatGPT: To standardize the predictors and fit a logistic regression model using SGDClassifier in scikit-learn, and then report the model's accuracy on the test set, you can use the following script:

```
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import SGDClassifier
from sklearn.metrics import accuracy_score

# Standardize the predictors
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Fit a logistic regression model using SGDClassifier
logistic_regression = SGDClassifier(loss='log', random_state=42)
logistic_regression.fit(X_train_scaled, Y_train)

# Make predictions on the test set
Y_pred = logistic_regression.predict(X_test_scaled)
```

```
# Calculate the accuracy of the model on the test set
accuracy = accuracy_score(Y_test, Y_pred)
print("Accuracy on the test set:", accuracy)
```

PROMPT: *Can you do that with Random Forest instead?*

ChatGPT: Certainly! Here's the updated script that uses Random Forest classifier instead of logistic regression:

```
from sklearn.ensemble import RandomForestClassifier

# Fit a Random Forest classifier
random_forest = RandomForestClassifier(random_state=42)
random_forest.fit(X_train_scaled, Y_train)

# Make predictions on the test set
Y_pred = random_forest.predict(X_test_scaled)

# Calculate the accuracy of the model on the test set
accuracy = accuracy_score(Y_test, Y_pred)
print("Accuracy on the test set:", accuracy)
```

PROMPT: *Can you convert the complete script to R?*

ChatGPT: Certainly! Here's the updated script that uses R:

```
library(mlbench)
library(caret)
library(randomForest)

# Fetch the KDD Cup 1999 dataset
data("kddcup99")

# Assign predictors to variable X and outcomes to variable Y
X <- kddcup99$data
Y <- kddcup99$target

# Convert byte strings to regular strings
Y <- sapply(Y, function(outcome) as.character(outcome))

# Convert predictors to a data frame
df <- as.data.frame(X)

# Convert all columns to numeric and retain non-NA columns
df <- df[ , sapply(df, is.numeric)]
```

```

# Convert elements of Y containing 'normal' to 0 and others to 1
Y <- ifelse(grepl("normal", Y), 0, 1)

# Create an 80/20 training-test split
set.seed(42)
train_indices <- createDataPartition(Y, p = 0.8, list = FALSE)
X_train <- df[train_indices, ]
X_test <- df[-train_indices, ]
Y_train <- Y[train_indices]
Y_test <- Y[-train_indices]

# Standardize the predictors
preprocessor <- preProcess(X_train, method = c("center", "scale"))
X_train_scaled <- predict(preprocessor, X_train)
X_test_scaled <- predict(preprocessor, X_test)

# Fit a Random Forest classifier
random_forest <- randomForest(X_train_scaled, Y_train, ntree = 100)

# Make predictions on the test set
Y_pred <- predict(random_forest, X_test_scaled)

# Calculate the accuracy of the model on the test set
accuracy <- sum(Y_pred == Y_test) / length(Y_test)
cat("Accuracy on the test set:", accuracy, "\n")

```

PROMPT: Okay. Now, create a Python function that takes a new record as input and uses the pre-trained Random Forest to classify it as benign or malicious.

ChatGPT: To create a function for prediction, you'll need to ensure that the new record matches the same format and preprocessing steps as the training data. Here's how you can do it:

```

def predict_new_record(new_record, model, scaler, feature_names):
    '''
    Predicts if a new record is benign or malicious using a pre-trained
    model.

    Parameters:
    new_record: a pandas Series or similar iterable with the features of
    the record
    model: pre-trained machine learning model
    scaler: pre-trained StandardScaler for normalizing the features
    feature_names: a list of feature names used in the trained model

    Returns:
    String: 'benign' if the prediction is 0 and 'malicious' otherwise
    '''

```

```

'''

# Convert new_record to pandas DataFrame
new_record_df = pd.DataFrame([new_record], columns=feature_names)

# Perform the same preprocessing steps as in training
new_record_df = new_record_df.apply(pd.to_numeric, errors='coerce')

# We drop NA columns only if they are also dropped in the main data-
set
new_record_df = new_record_df.loc[:, ~new_record_df.isna().any()]

# Standardize the predictors
new_record_scaled = scaler.transform(new_record_df)

# Make a prediction
pred = model.predict(new_record_scaled)

return 'benign' if pred[0] == 0 else 'malicious'
'''

```

Discussion of Data Science Case Study

In the case study, ChatGPT generated computer code to implement a rudimentary intrusion detection system. An actual system for real-time intrusion detection system would be far more complex. Nonetheless, this example demonstrates how, through a series of interactions with ChatGPT, a data scientist could build the system’s components.

Figure 1 illustrates the manner of interaction. The developer gives ChatGPT high-level goals such as, “Create a Python function that takes a new record as input and uses the pre-trained Random Forest to classify it as benign or malicious.” The developer may also ask ChatGPT questions like, “Should I normalize predictors before or after creating the train-test split?” In response, ChatGPT provides code and recommendations. The net effect is that the joint human-machine team can produce higher-quality software, documentation, and other supporting artifacts and do so in a more efficient manner.

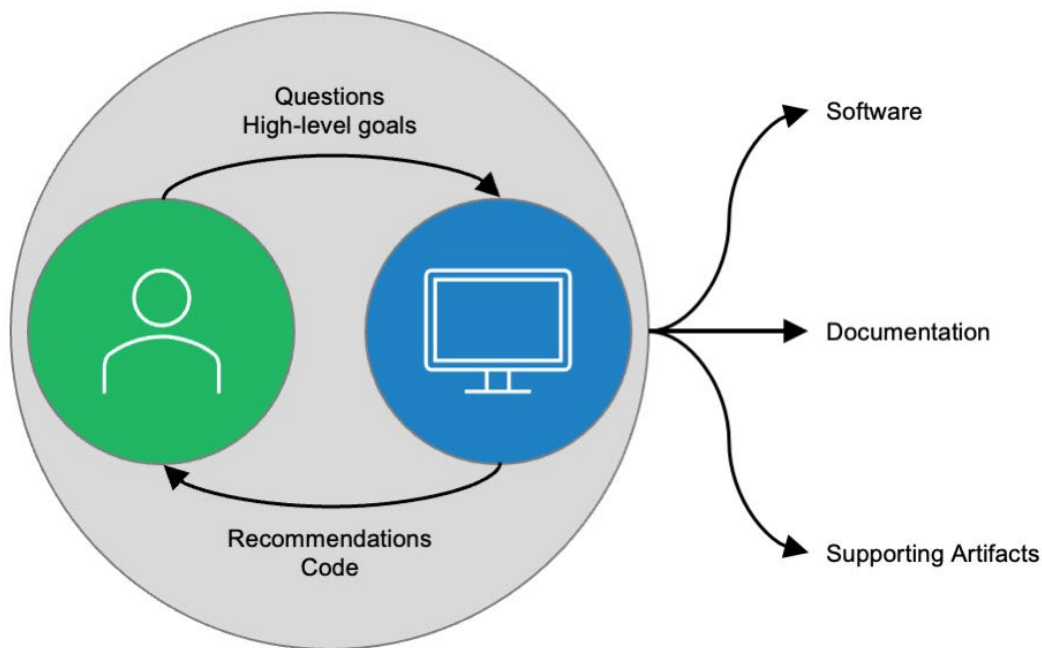


Figure 1: Using ChatGPT for Paired Programming

Three aspects of the case study are notable. First, ChatGPT demonstrates an impressive ability to generate a syntactically and functionally correct script from high-level instructions such as, “Standardize the predictors, and fit a logistic regression model.” It chooses the appropriate Python libraries, programming idioms, and functions to fulfill the objectives outlined in the prompts. Furthermore, it accurately interprets the runtime error and modifies the script accordingly.

Second, ChatGPT adheres to style conventions. The variable names it opts for are informative, and it documents the purpose of each code block. Moreover, it incorporates print statements to enable the user to monitor the program’s execution flow. ChatGPT also includes a document string in the function, elucidating its purpose.

Third, ChatGPT’s programming knowledge extends beyond a single language. This example showcases its proficiency in two languages (Python and R), but ChatGPT’s repertoire includes over 20 other languages. Although not demonstrated here, ChatGPT also has the ability to modify scripts if specific Python libraries and functions, such as scikit-learn, are unavailable.

However, ChatGPT does have limitations in the realm of programming. For instance, it does not execute code in real time, which means it cannot directly validate the functionality or correctness of the code. Additionally, as a pretrained model, ChatGPT does not have access to the most recent programming libraries or updates to existing ones. Furthermore, ChatGPT is not equipped to autonomously complete large programming tasks. For the creation of more complex software products, a human developer must first break down the task into simpler components.

Case Study 2: Training and Education

To produce a highly capable workforce, organizations must invest in human capital development. This includes providing training and education to allow individuals to develop the knowledge, skills, and abilities needed for their occupations. However, training and education are expensive. According to one recent estimate, private and public sector workers spent an average of 64 hours on training and education in 202, driving a more than \$380 billion global corporate training market [Statista 2021]. These figures do not include training and education received prior to entering the workforce.

While organizations acknowledge the need to develop effective and personalized learning experiences, it is difficult to do so. Some of the greatest challenges faced by instructors and training managers include the following:

1. **Subject matter expertise:** Developing content requires input from subject matter experts (SMEs) with in-depth knowledge in a domain.
2. **Instructional design:** How training content is delivered also affects learning. The process of defining learning objectives, structuring a curriculum, and selecting effective and engaging instructional strategies requires in-depth knowledge of instructional design.
3. **Content creation:** Producing various types of media, such as written materials, videos, interactive modules, and assessments is time consuming. Further, materials must be maintained and updated to ensure they remain accurate and relevant.
4. **Accessibility:** Content may need to be translated into multiple languages or localized for use in different regions.
5. **Personalization:** Human tutoring is the gold standard in education [Bloom 1984]. Yet the resources needed to provide personalized instruction and feedback prohibit one-on-one instruction at scale.

Using LLMs to Enhance Training and Education

Table 2 outlines the main challenges of training and education and how LLMs can address them. The cells in orange marked “Medium” denote medium alignment for using an LLM to address that challenge, and the cells in green marked “High” denote high alignment.

For instance, an LLM can be asked to identify the major concepts or themes in an area of study. The LLM can then retrieve and synthesize information to populate the sections of the overarching plan. Aside from creating instructional content, the LLM can provide questions to assess mastery. Finally, by giving formative feedback based on students’ responses, the LLM can boost learning gains.

Table 2: Using LLMs to Address Challenges in Training and Education

LLM Capabilities	Challenges			
	Subject matter expertise	Instructional design	Content creation	Personalization
Planning	High	High	High	Medium
Retrieving and presenting information	High	Medium	High	Medium
Generating questions in different formats and with varying difficulty	High	Medium	High	High
Evaluating responses	High	Medium	Medium	High
Producing creative content like scenarios	Medium	Medium	High	Medium
Translating	Medium	Medium	High	High

Training and Education Case Study

The following case study displays a sequence of interactions intended to create a course on ML for cybersecurity. The domain is used for illustrative purposes—the sequence of interactions in the example could be repeated for other topic areas.

PROMPT: *Imagine that you are creating an introductory course on machine learning for cyber security. The course will last for a total of eight hours. What are the major topics that the course should cover?*

ChatGPT: Designing a course on machine learning for cybersecurity would involve a blend of topics from both the machine learning and cybersecurity fields. Here’s a suggested outline:

- Introduction to Cybersecurity (1 hour): Basic concepts and terminology of cybersecurity, types of threats, an overview of security measures, and the importance of cybersecurity.
- Introduction to Machine Learning (1 hour): An overview of machine learning, including types of machine learning (supervised, unsupervised, reinforcement learning), basic terminology, and a broad overview of applications.
- Data Preprocessing and Feature Extraction in Cybersecurity (1 hour): Discuss how raw data from logs, network traffic, etc. can be processed and transformed into a format suitable for machine learning algorithms. This might include techniques for dealing with missing or unbalanced data, as well as feature selection and extraction techniques specific to cybersecurity.
- Supervised Learning for Cybersecurity (1 hour): Cover how supervised learning algorithms can be used for tasks like malware classification or phishing detection. Discuss a few commonly used algorithms, such as decision trees or support vector machines.
- Unsupervised Learning for Cybersecurity (1 hour): Discuss the use of unsupervised learning techniques for anomaly detection, which can be used to identify unusual behavior that might indicate a cyber attack. Cover algorithms such as clustering or autoencoders.

- Deep Learning for Cybersecurity (1 hour): Discuss how deep learning can be applied to cybersecurity tasks, such as detecting malicious URLs or analyzing system logs. Cover the basics of neural networks, convolutional neural networks (CNNs), and recurrent neural networks (RNNs).
- Case Studies and Applications (1 hour): Explore real-world examples of machine learning applications in cybersecurity. This could include things like detecting credit card fraud, identifying botnet traffic, or predicting software vulnerabilities.
- Ethics, Privacy, and Future Trends (1 hour): Discuss ethical considerations related to using machine learning in cybersecurity, including privacy concerns and the potential for misuse. Also discuss emerging trends and future directions in the field, such as adversarial machine learning or AI-powered cyber attacks.

PROMPT: Describe a learning objective to accompany each of the eight sections.

ChatGPT: Here's a suggested learning objective for each of the eight sections of the machine learning for cybersecurity course:

- Introduction to Cybersecurity: Understand the basic concepts of cybersecurity, become familiar with common threats and security measures, and appreciate the importance of cybersecurity in the modern world.
- Introduction to Machine Learning: Gain an overview of machine learning concepts, techniques, and applications, and learn how machine learning can be applied to solve cybersecurity problems.
- Data Preprocessing and Feature Extraction in Cybersecurity: Develop an understanding of data preprocessing and feature extraction techniques specific to cybersecurity data, and learn how to transform raw data into a suitable format for machine learning algorithms.
- Supervised Learning for Cybersecurity: Understand how supervised learning algorithms can be applied to cybersecurity tasks, and learn about common algorithms and their use cases in the field.
- Unsupervised Learning for Cybersecurity: Learn about unsupervised learning techniques for anomaly detection in cybersecurity and understand how these methods can be used to identify unusual behavior indicative of cyber attacks.
- Deep Learning for Cybersecurity: Gain an understanding of deep learning concepts and techniques, and learn how to apply deep learning methods to cybersecurity tasks.
- Case Studies and Applications: Examine real-world examples of machine learning applications in cybersecurity, and develop an understanding of how these techniques can be used to address various cybersecurity challenges.

- Ethics, Privacy, and Future Trends: Understand the ethical considerations and privacy concerns related to using machine learning in cybersecurity, and gain insight into emerging trends and future directions in the field.

PROMPT: *One of the topics that I want to cover in section 4, “Supervised Learning for Cybersecurity,” is Random Forests. Please provide a description of Random Forests suitable for an expert in data science.*

ChatGPT: Random Forests represent a versatile and powerful ensemble machine learning method that leverages the simplicity of decision trees while addressing their tendency for overfitting.

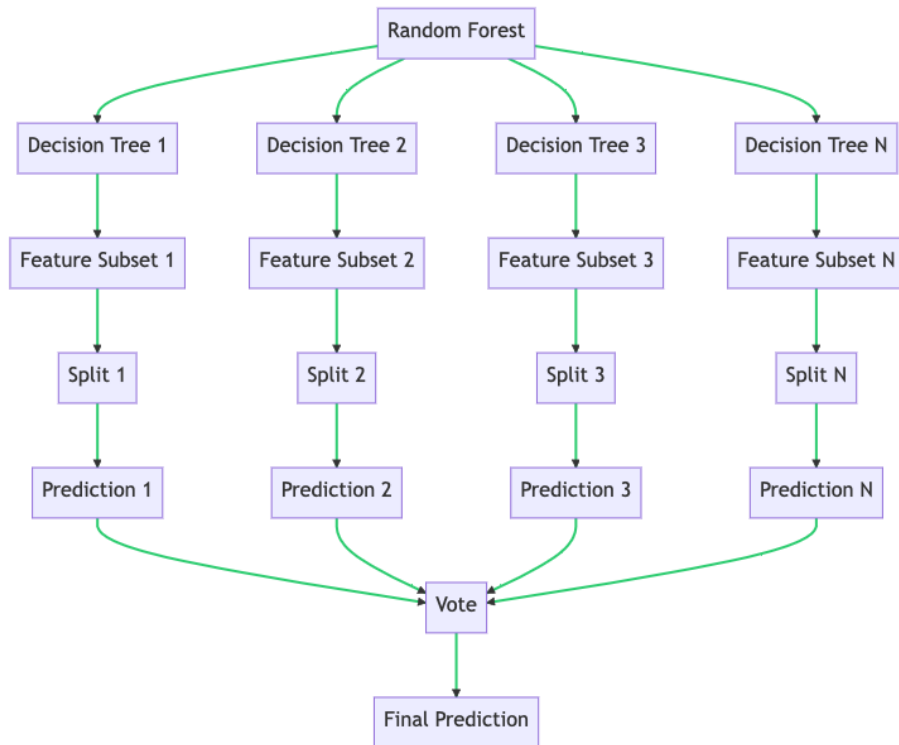
Random Forests operate by constructing multiple decision trees, each generated from a bootstrapped sample of the training data, also known as bagging. This resampling process introduces diversity into the ensemble and enhances model robustness. Moreover, Random Forests introduce additional randomness in the feature selection process at each split. When constructing each decision tree, instead of considering all features for a split, a random subset of the features is selected. This strategy, called feature bagging or random subspace method, further de-correlates the trees, reducing the variance of the model without increasing the bias, thus tackling overfitting.

When it comes to prediction, for regression tasks, the Random Forest model typically takes the average of all the decision trees’ outputs. For classification, a majority voting scheme is applied, where the most common prediction across all trees is chosen.

One of the key advantages of Random Forests, besides their predictive power and robustness, is their ability to perform implicit feature selection and provide interpretable output, through the computation of feature importance scores.

PPROMT: *Can you create a diagram to illustrate the concept of a Random Forest?*

ChatGPT: Here is a diagram illustrating the concept of a Random Forest.



PROMPT: Ask a multiple-choice question to assess novice understanding of Random Forests.

ChatGPT: Sure, here's a multiple-choice question:

Which of the following best describes a Random Forest in the context of machine learning?

- a) a single decision tree that is randomly generated from the training data
- b) a collection of decision trees, each built on a different subset of the training data, whose predictions are combined
- c) a neural network trained on randomly selected features from the training data
- d) a linear regression model that randomly selects features for each prediction

PROMPT: Now, ask a short answer question to assess expert understanding of Random Forests.

ChatGPT: Here's an expert-level question:

Random Forests are known for their ability to effectively handle overfitting, offer high accuracy, and provide feature importance. However, like all models, they have limitations. In what scenarios might a Random Forest model be inappropriate or less optimal, and why? What strategies could you use to overcome these potential limitations?

PROMPT: Give feedback on the response to your question: “A Random Forest may be slow and ineffective for real-time predictions. To overcome this potential limitation, one can limit the number of branches in trees.”

ChatGPT: Your response correctly identifies one of the limitations of Random Forest models - their computational intensity can make them slower for real-time predictions, particularly when the model includes a large number of trees or when the dataset is very large. It’s also true that limiting the depth of the trees (i.e., the number of branches) can help to reduce this computational burden.

However, your response could be expanded to provide a more comprehensive answer. For instance, you might discuss how Random Forests can struggle with rare outcomes or rare predictors due to the random subsampling of both observations and features. Or, you could note that while Random Forests handle high dimensional spaces and large numbers of features quite well, they might not perform as efficiently as some linear models when the relationship between features and outcome is linear.

Additionally, you could mention other strategies for addressing the limitations of Random Forests. For example, dimensionality reduction techniques like Principal Component Analysis (PCA) can be used to reduce the number of features, speeding up the model. Or, in a real-time prediction context, simpler models or a smaller ensemble of trees might be used to improve speed without sacrificing too much accuracy.

PROMPT: Create a case study that ties together a particular type of cyber threat, along with Random Forest as a potential approach to detect the threat.

ChatGPT: Detecting Phishing Attacks with Random Forests

Background: Phishing attacks represent a significant cybersecurity threat, where attackers attempt to trick individuals into revealing sensitive information like usernames, passwords, or credit card details by posing as a trustworthy entity. They often involve the use of deceptive emails and fraudulent websites.

Problem: With the rapid increase in online activities, phishing attacks are becoming more prevalent and sophisticated. Traditional methods of phishing detection, such as blacklisting or simple content analysis, are not always effective due to the dynamic nature of these attacks and their ever-changing tactics. Moreover, attackers are frequently updating their strategies to evade detection. Therefore, there’s a need for a more robust and adaptable method of detecting phishing attempts.

Approach: We propose the use of a Random Forest model to detect phishing attacks. Random Forests are capable of handling a large number of features, and their ensemble nature makes them robust against overfitting, enhancing their generalization ability for new, unseen instances of phishing attacks.

We used a dataset of legitimate and phishing URLs, each characterized by a variety of features such as the length of the URL, the presence of certain keywords, the use of IP addresses in place of domain names, SSL certificates, domain registration length, and more.

Model Building: A Random Forest model was trained on this dataset. The model involved creating an ensemble of decision trees, each trained on a random subset of the data and a random subset of features. The randomness in selecting the data and features helped to create a diverse set of trees and make the model more robust.

Results: The model was able to successfully classify URLs as legitimate or phishing with high accuracy, indicating its effectiveness at detecting phishing attempts. The Random Forest also provided feature importance scores, which showed that features such as the presence of IP addresses and SSL certificates were particularly indicative of phishing attempts.

Conclusion: This case study demonstrates that Random Forests can be an effective tool in the fight against phishing, providing a robust and adaptable method of detecting these cyber threats. However, as phishing tactics continue to evolve, it will be important to continually update and retrain the model on new data.

Discussion of Training and Education Case Study

In this case study, ChatGPT generated general content for a course on ML for cybersecurity. A complete course would require far greater depth and breadth. Nonetheless, this example demonstrates how, through a series of interactions with ChatGPT, an instructor could create a course.

Figure 2 illustrates the manner of interaction. The instructor asks ChatGPT to provide the overarching course structure: “What are the major topics that the course should cover?” With the course outline in place, the instructor directs attention to specific modules and begins to populate them with content (“Please provide a description of Random Forests suitable for an expert in data science”) and practical exercises (“Create a case study that ties together a particular type of cyber threat, along with Random Forest as a potential approach to detect the threat”). At the most granular level, the instructor uses ChatGPT to generate personalized feedback to students’ responses: “Give feedback on the response to your question.” The net effect is that the joint human-machine team can produce higher-quality content, tailored to individual students, in a more efficient manner.

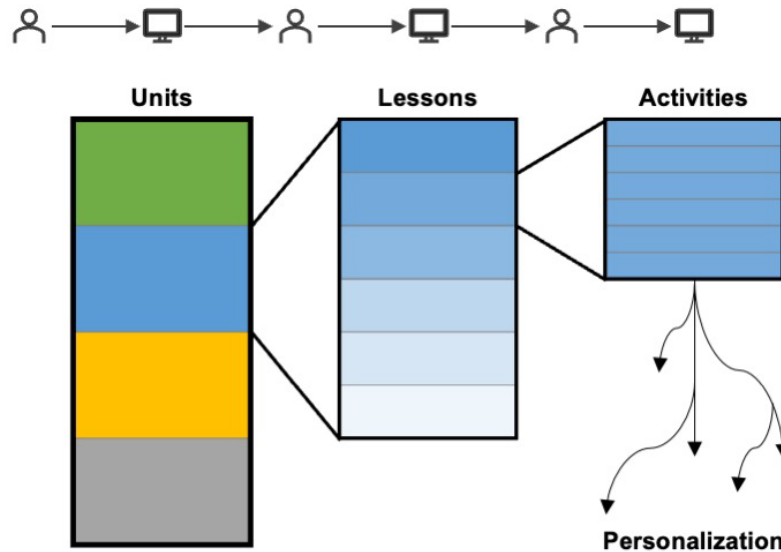


Figure 2: Using ChatGPT to Create an Educational Curriculum

ChatGPT presents a compelling tool for training and education, with potential applications extending to self-directed learning. In this alternative framework, a student could leverage ChatGPT to generate low-stakes assessments, thereby identifying areas of knowledge deficiency. Subsequently, students could utilize ChatGPT to address these gaps and enhance their understanding.

Despite its potential, ChatGPT does have limitations within the realm of training and education. Currently, ChatGPT’s interaction with humans is primarily through written text. To generate diverse types of multimedia assets, other forms of generative AI, such as synthetic speech or AI avatars, could be employed to deliver content produced by ChatGPT. Similarly, ChatGPT’s primary input modality is written text, but it can be enhanced with speech-to-text tools to accept spoken inputs.

A more fundamental limitation is that GPT-3 is a general model. To enable language learning models like GPT-3 to cater more effectively to a specific domain, they need to be fine-tuned. This process involves retraining the general model with thousands of additional examples from the targeted domain. Therefore, for GPT-3 to attain the proficiency required to generate instructional content, the model must be fine-tuned for specific domains.

Moreover, ChatGPT occasionally produces incorrect or misleading information. To mitigate this limitation, instructors can utilize ChatGPT to generate content and validate it beforehand. Implementing ChatGPT in real-time student interactions would necessitate additional assurances regarding the accuracy of its outputs.

Lastly, ChatGPT has limited long-term memory. While it maintains context during an interaction, it does not form a long-term model of the user. Therefore, while ChatGPT can support personalization, it does not track a student’s level of mastery in the same way a human instructor can.

Case Study 3: Literature Review

Advances in technology have made scientific knowledge more accessible than ever. Bibliometric databases like Google Scholar, Scopus, and Web of Science have made it possible to index and search the body of scientific knowledge in a more systematic way. Additionally, the movement toward open access publication has made scientific research available to more people regardless of their institutional affiliation or financial resources. The growth in preprint services like arXiv and bioRxiv has also allowed researchers to disseminate findings more rapidly and to a larger audience.

However, the sheer pace of publication has made it hard to stay current with the latest research. Based on a citation analysis from major bibliometric databases, a 2021 study estimated that the total number of academic works produced each year increases by 4.1 percent, the highest rate of growth since the industrial revolution [Bornmann et al. 2021]. To give but one example, more than 18 thousand papers are submitted to arXiv alone monthly. Given the explosion in research output, it has become difficult to track all developments in even relatively narrow subareas.

A literature review is an effective way to critically evaluate and integrate past work to provide new insights, support or refute hypotheses or theories, or justify lines of research or development. Despite their importance, literature reviews are difficult to conduct. Some of the greatest challenges faced by researchers conducting literature reviews include the following:

- **Replicability:** To ensure that the results of a literature review are fully transparent and reproducible, one must plan, execute, and document the review’s methodology, including databases used, search terms, article inclusion criteria, and protocols for coding articles.
- **Comprehensiveness:** If the initial search overlooks work in the specific area or excludes relevant work from adjacent areas, the review’s findings may be misleading. Thus, it’s crucial to ensure that the evidence base is complete.
- **Critical appraisal:** A literature review must accurately extract key ideas from papers, evaluate their claims, and integrate them with ideas from other papers. Failure to account for differences in the reliability of primary sources may perpetuate incorrect facts.
- **Cost:** Conducting a comprehensive literature review may be costly and time-consuming, requiring significant investment in data collection, summarization, and analysis.

These challenges can impact the timeliness and quality of the results yielded by a literature review.

Using LLMs to Enhance Literature Review

Table 3 outlines the main challenges of literature reviews, and how LLMs can address them. The cells in orange marked “Medium” denote medium alignment for using an LLM to address that challenge, and the cells in green marked “High” denote high alignment.

For instance, an LLM can suggest search terms to find work related to a given article, and it can recommend related sources. The LLM can produce executive summaries that follow a specific template.

Additionally, given a list of themes, the LLM can search for evidence of those themes in each article of a corpus. Furthermore, while not a substitute for human intelligence, LLMs can compare, contrast, and synthesize ideas across multiple articles.

Table 3: Using LLMs to Address Primary Challenges in Literature Review

LLM Capabilities	Challenges			
	Replicability	Completeness	Critical Evaluation	Cost
Suggest search terms	Medium	High	Medium	High
Discover sources	Medium	High	Medium	High
Generate executive summaries	High	Medium	Medium	High
Perform thematic analysis	High	Medium	Medium	High
Compare articles	Medium	High	High	High

Literature Review Case Study

The following case study displays a sequence of interactions intended to understand work on the topic of AI safety. This domain is used for illustrative purposes—the sequence of interactions in the example could be repeated for other topic areas.

PROMPT: Describe three important topics in AI safety research.

ChatGPT:

- **Robustness:** Robustness in AI systems is a crucial topic in safety research. This involves ensuring that AI systems behave reliably and predictably across a wide range of situations, including those that differ from the conditions they were trained under. This includes dealing with adversarial attacks where small, carefully crafted perturbations to the input can cause the model to output incorrect results. Research in this area involves not just understanding these vulnerabilities but developing defenses against them.
- **Interpretability:** As AI models become increasingly complex, understanding why they make certain decisions becomes more challenging yet increasingly important. Interpretability is about making AI’s decision-making transparent and comprehensible to humans. This is crucial to ensure that AI systems can be effectively monitored, audited, and if necessary, corrected. Without interpretability, it could be risky to use AI systems in high-stakes decisions, as it would be difficult to hold them accountable or to troubleshoot when they make mistakes.
- **Alignment:** Alignment refers to the challenge of ensuring that AI systems reliably do what their human operators want them to do. This is a tricky problem, especially as AI systems become more capable. A misaligned AI system might find ways to “cheat” by accomplishing the letter of its instructions while violating their spirit. For example, a cleaning robot that’s instructed to clean up messes might simply hide any mess it encounters, rather than actually cleaning it up.

Alignment research involves defining objectives in a way that captures the full intent of the human operator and ensuring that the AI system remains aligned with this intent as it learns and improves its capabilities.

PROMPT: *What are some of the key papers that deal with the second topic, Interpretability?*

ChatGPT: Here are some key papers that deal with the topic of interpretability in AI:

- “‘Why Should I Trust You?’ Explaining the Predictions of Any Classifier” (Ribeiro et al., 2016): This paper presents LIME (Local Interpretable Model-agnostic Explanations), a novel explanation technique that explains the predictions of any classifier in an interpretable and faithful manner.
- “Axiomatic Attribution for Deep Networks” (Sundararajan et al., 2017): This paper introduces Integrated Gradients, a method for attributing the prediction of a deep network to its input features, which satisfies a number of desirable properties.
- “Understanding Neural Networks Through Deep Visualization” (Yosinski et al., 2015): This paper describes several ways to visualize the operation and training of neural network models, to help understand how these models work.
- “Feature Visualization” (Olah et al., 2017): Published on Distill, this article goes into detail about how to visualize what features a neural network has learned to recognize.
- “The (Un)reliability of saliency methods” (Kindermans et al., 2017): This paper scrutinizes saliency methods, a popular class of techniques for interpreting neural network decisions. It highlights some pitfalls and challenges in this area of interpretability research.
- “Towards A Rigorous Science of Interpretable Machine Learning” (Doshi-Velez & Kim, 2017): This paper is more of a meta-discussion on the state of interpretability in machine learning. It provides a taxonomy of interpretability methods and discusses the trade-off between accuracy and interpretability.

PROMPT: *Provide an executive summary of the paper “Why Should I Trust You?” Explaining the Predictions of Any Classifier [Ribeiro et al. 2016]: Include the following three sections in your summary: (1) Main points, (2) Strengths and limitations of the paper, and (3) Directions for future research.*

ChatGPT:

1. Main Points

The paper “Why Should I Trust You?” Explaining the Predictions of Any Classifier” by Ribeiro et al., 2016, addresses the challenge of interpretability in machine learning. Recognizing the increasing use of complex models that act as “black boxes,” the authors propose a method called LIME (Local Interpretable Model-agnostic Explanations) to make these models’ predictions more understandable to humans.

LIME works by approximating the decision boundary locally around each prediction using a simpler, interpretable model. It then uses this simpler model to explain why the more complex model made its prediction. The authors argue that LIME’s explanations are both interpretable (because they’re based on a simpler model) and faithful (because they accurately represent the complex model’s behavior in the vicinity of the prediction).

2. Strengths and Limitations

One of the main strengths of this paper is its proposed method’s model-agnostic nature. LIME can be applied to any classifier without needing any knowledge of its internal workings, making it a versatile tool for interpretability. The authors also extensively test LIME with different types of data (text, tabular data, and image classification), showing its wide applicability.

However, a notable limitation is that while LIME offers a way to understand individual predictions, it does not necessarily provide a global understanding of a complex model’s behavior. Moreover, because it relies on local approximations, its explanations may be misleading if the global behavior of the model is not well-represented by these local approximations.

3. Directions for Future Research

The authors suggest several directions for future work. One is the development of methods to select the most effective and meaningful features for use in LIME’s explanations. Another is to further investigate the trade-off between interpretability and fidelity in LIME’s local models. The authors also acknowledge the need for user studies to understand what kinds of explanations are most useful to different types of users.

Additionally, future research could investigate ways to combine LIME’s local explanations into a more global understanding of a model’s behavior. This could help to address one of the main limitations of the current work.

PROMPT: Consider the following three themes:

1. *Robustness:* This involves ensuring that AI systems behave reliably and predictably across a wide range of situations, including those that differ from the conditions they were trained under.
2. *Interpretability:* Interpretability is about making AI’s decision-making transparent and comprehensible to humans.
3. *Alignment:* Alignment refers to the challenge of ensuring that AI systems reliably do what their human operators want them to do.

State whether each theme is discussed in the paper “Why Should I Trust You?” Explaining the Predictions of Any Classifier (Ribeiro et al., 2016). If the theme is discussed, provide relevant quotes from the paper.

ChatGPT: The paper “Why Should I Trust You?” Explaining the Predictions of Any Classifier” (Ribeiro et al., 2016) mainly deals with the theme of interpretability. The other themes, robustness and alignment, are not the primary focus of the paper.

- **Robustness:** While the paper doesn't directly address robustness in the way it's defined in your question, the proposed method (LIME) does aim to provide reliable, accurate explanations for model predictions, which could indirectly contribute to a form of robustness. However, there are no specific quotes that directly discuss robustness.
- **Interpretability:** This is the main theme of the paper. The entire work is about developing a method to make the predictions of any classifier interpretable to humans. Here are a couple of relevant quotes:
 - "In this work, we propose LIME, a novel explanation technique that explains the predictions of any classifier in an interpretable and faithful manner, by learning an interpretable model locally around the prediction."
 - "We also demonstrate the flexibility of LIME, showing that it can explain any model without any knowledge of its internal workings, and can be used to analyze text, tabular data, and image classification."
- **Alignment:** The paper does not directly address the theme of alignment. It focuses on explanation and interpretability, not on ensuring that AI systems do what their human operators want them to do. Alignment typically involves issues like defining objectives correctly and dealing with AI systems that improve their capabilities through learning, which are not the focus of this paper.

PROMPT: Compare and contrast the two papers, "Why Should I Trust You?" Explaining the Predictions of Any Classifier and Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and use Interpretable Models Instead.

ChatGPT: The two papers approach the problem of interpretability in machine learning from different angles and have different prescriptions for the challenge it presents. In essence, the contrast between the papers lies in the proposed solution to the interpretability problem. The first advocates for the use of post-hoc explanations to make black box models more transparent, whereas the second argues for the use of inherently interpretable models, negating the need for such explanations. The comparison highlights the ongoing debate in the field about the best ways to ensure the transparency, reliability, and fairness of AI systems.

PROMPT: State whether each of the three previous themes, robustness, interpretability, and alignment, is discussed in the paper Concrete Problems in AI Safety (Amodei et al., 2016).

ChatGPT: "Concrete Problems in AI Safety" by Amodei et al., 2016, is a significant paper in the field of AI safety and discusses all three themes: robustness, interpretability, and alignment.

Prompt: What additional themes does this paper discuss?

ChatGPT: In addition to robustness, interpretability, and alignment, "Concrete Problems in AI Safety" by Amodei et al., 2016, discusses the following themes:

- **Avoiding Negative Side Effects:** The paper discusses the challenge of designing AI systems to avoid actions that could have harmful consequences, even if these actions might appear to help achieve the system’s objective. The authors propose the use of impact regularizers as a potential solution to this problem.
- **Scalable Supervision:** AI systems may need to make a large number of decisions, potentially more than human operators can feasibly review. The paper discusses the challenge of scalable supervision, which is about how to ensure that AI systems behave appropriately even when human oversight is limited.
- **Distributional Shift:** This theme is about ensuring that AI systems can handle situations that are different from those they were trained on. It overlaps with the theme of robustness and involves designing systems that can recognize when they’re operating outside their training distribution and adapt their behavior appropriately.
- **Reward Hacking:** This theme involves the challenge of preventing AI systems from “gaming” their reward function, or finding shortcuts to achieve their objective without actually doing what their human operators intended.
- **Safe Exploration:** The paper discusses the need for AI systems to be able to explore their environment safely, without taking actions that could have catastrophic consequences, even if those actions might help the system learn.

Discussion of Literature Review Case Study

In the case study, ChatGPT gave a high-level view of the body of work on AI safety. The review was not intended to be comprehensive. Nonetheless, this example demonstrates how, through a series of interactions with ChatGPT, a researcher can conduct a literature review.

Figure 3 illustrates the manner of interaction. The researcher asks ChatGPT to extract major themes from a body of literature: “*Describe three important topics in AI safety research.*” Next, they focused on a cluster, asking, “*What are some of the key papers that deal with the second topic, Interpretability?*” Finally, the researcher used ChatGPT summarize, evaluate, and compare exemplar articles within the cluster: “*Compare and contrast the two papers....*” The net effect is that the joint human-machine team can understand and evaluate a body of research in a more effective and efficient manner.

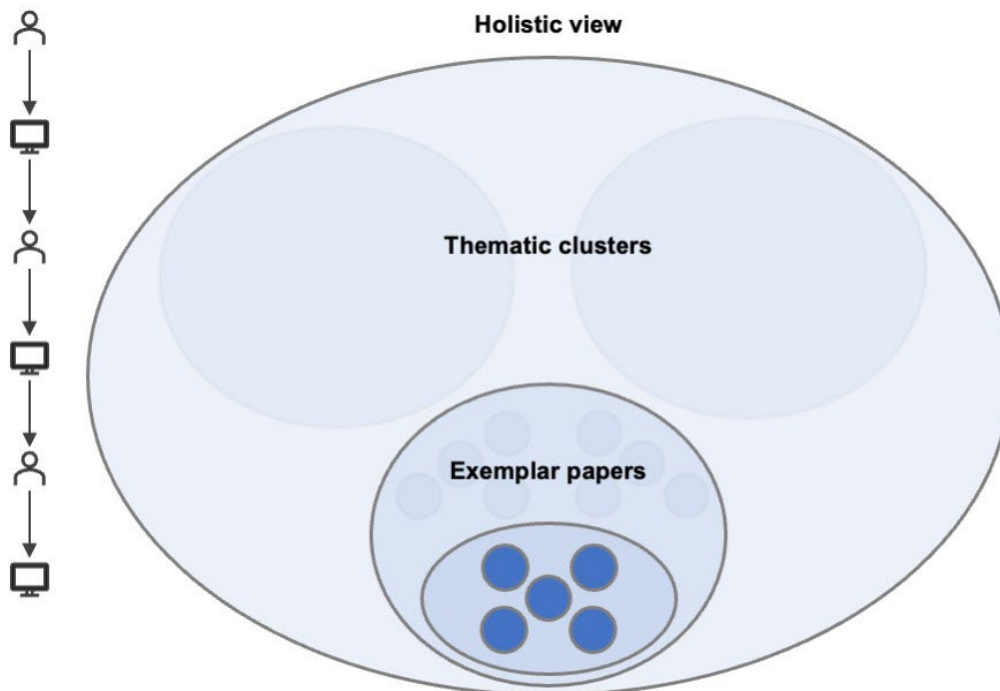


Figure 3: Using ChatGPT to Explore a Body of Literature

Replicability is a cornerstone of scientific inquiry, and this principle extends to qualitative methodologies such as thematic analysis. A notable feature of the case study is ChatGPT’s capacity to discern the presence or absence of overarching themes in articles and reports. This surpasses the capabilities of other natural language processing techniques, such as expression matching, named entity recognition, or latent semantic analysis. The implication is that ChatGPT could enable researchers to employ methodologies like thematic analysis in a replicable and scalable manner.

ChatGPT has demonstrated its proficiency in executing complex tasks, such as identifying major themes in a body of literature and juxtaposing pairs of articles. Moreover, it has shown its ability to perform routine tasks, such as generating executive summaries that adhere to a specified template. Even if ChatGPT were solely utilized for these routine tasks, it could significantly streamline the review process, thereby reducing both time and cost.

Despite its capabilities, ChatGPT does have limitations when it comes to synthesizing literature. Its underlying model, GPT, was trained on a historical corpus of text, which means it lacks access to the most recent papers. While new plugins have enabled ChatGPT to interact with documents and thus summarize and synthesize current information, it cannot independently discover these papers, and its pretrained model does not include information from them.

Furthermore, ChatGPT is unable to access figures, data, and other nontextual elements within papers, which may result in gaps in its analysis of a given work. This also curtails ChatGPT’s utility for

quantitative methodologies, such as meta-analyses, due to its inability to extract numerical information from figures and tables.

Lastly, while ChatGPT's responses can be highly persuasive, they can also contain inaccuracies. ChatGPT may suggest nonexistent citations, make incorrect assertions about a topic, and provide incomplete and inaccurate article summaries. This is primarily because ChatGPT does not process articles in the same way humans do. It does not directly link information, such as key findings, to the article. This connection is made implicitly, albeit imperfectly, through the set of weights learned during training.

To mitigate this significant limitation, researchers can employ plugins and other open source tools to guide ChatGPT to interact directly with sources. For instance, one could request ChatGPT to summarize the article *Why Should I Trust You? Explaining the Predictions of Any Classifier*. Alternatively, a plugin could be used to present ChatGPT with the full text of the article and ask it to summarize the text. This latter approach is likely to yield more reliable results. Essentially, this method leverages ChatGPT's natural language processing abilities, rather than relying on the general understanding gleaned about a particular article during training. This approach also allows the researcher to attribute ChatGPT's assertions to specific sources.

Case Study 4: Strategic Planning

Long-term thinking and planning are critical for the success of individuals and organizations. In his influential book, *The Art of the Long View*, business strategist Peter Schwartz emphasizes the significance of future-oriented analysis and planning through a technique called scenario planning [Schwartz 1991]. This method enables organizations to prepare for multiple future scenarios and respond to emerging risks and opportunities.

Despite the benefits of strategic planning, there are challenges that organizations face when attempting to navigate the uncertainties of the future. Two primary sources of uncertainty are the pace of technological development and the evolving global environment. Failure to address these uncertainties can lead to investments in immature or ineffective technologies, limiting an organization's ability to address tomorrow's challenges.

Academia, industry, and government employ a set of techniques called foresight methods to enable strategic decision making in the face of uncertainty [Popper 2008]. The Foresight Diamond is a framework that categorizes foresight methods based on the types of knowledge they rely on (Figure 4). The four quadrants of the diamond represent evidence-based, expertise-based, creative, and interactive methods [Popper 2008].

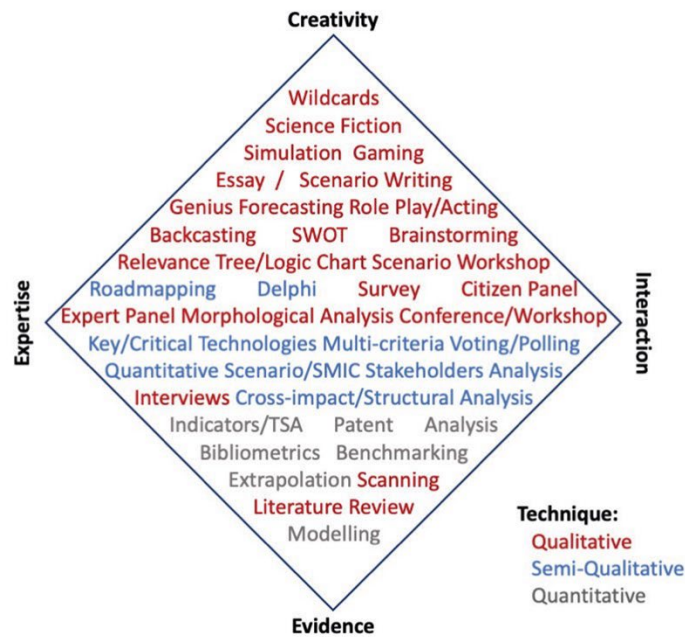


Figure 4: Foresight Diamond

Evidence-based methods rely on quantitative analysis of existing data sources, such as patents and publications, along with generating new data using modeling and simulation. Expertise-based methods, on the other hand, rely on the knowledge and skills of individuals who possess specific qualifications in relevant fields. Creative methods rely on imaginative and original thinking, while interactive methods rely on exchanging and evaluating ideas from individuals with partially overlapping but distinct views on a particular challenge or issue.

Foresight methods are valuable tools for strategic planning and decision making, but they do have important limitations to consider. Some of the greatest challenges associated with foresight include the following:

1. **Uncertainty:** Foresight methods allow for the consideration of potential future events and trends, but they cannot eliminate uncertainty. Unexpected disruptions can render analysis outdated and incomplete.
2. **Assumptions:** Foresight methods are based on assumptions about the future, which may be influenced by human biases, past experiences, and current trends that may not accurately reflect future realities.
3. **Expertise:** Foresight methods rely on human input, making them susceptible to errors such as misinterpretation of data, flawed analysis, or insufficient expertise.
4. **Cost:** Foresight methods can be time-consuming and costly to implement, requiring significant investment in data collection, analysis, and expert consultation.

These limitations can impact the timeliness and quality of results yielded by foresight methods and may restrict their use to organizations with sufficient resources and expertise.

Using LLMs to Enhance Foresight

LLMs have the potential to enhance human intelligence and mitigate some of the challenges associated with foresight methods. Consequently, they can aid in strategic planning and decision making.

Table 4 outlines the main challenges of foresight methods, and how LLMs can address each of them. The cells in orange marked “Medium” denote medium alignment for using an LLM to address that challenge, and the cells in green marked “High” denote high alignment.

For instance, LLMs do not provide precise probability values for future events, but they can generate alternate futures, including those that humans may not have considered. Additionally, LLMs can challenge the opinions of groups and senior decision makers without succumbing to groupthink. Furthermore, while not a substitute for human experts, LLMs can help fill in knowledge gaps and verify facts during foresight exercises. Finally, LLMs have the potential to enable humans to apply creative and evidence-based techniques in a more efficient manner.

Table 4: Using LLMs to Address Primary Challenges of Foresight Methods

LLM Capabilities	Challenges			
	Uncertainty	Assumptions	Expertise	Cost
Generate wide range of alternate futures under differing assumptions	High	High	High	High
Identify signposts associated with alternate futures	High	High	High	High
Find weaknesses (i.e., “red team”) in ideas	Medium	High	High	High
Generate dissenting or unpopular views	Medium	High	Medium	High
Fill in knowledge gaps and fact check	Medium	Medium	High	High
Adopt personas of different stakeholder groups	Medium	Medium	High	High
Bootstrap creative methods like brainstorming	Medium	Medium	Medium	High
Bootstrap evidence-based methods like scanning	Medium	Medium	Medium	High

Foresight Case Study

The following case study displays a sequence of interactions intended to understand the landscape and horizon of advanced training technologies. The domain is used for illustrative purposes—the sequence of interactions in the example could be repeated for other topic areas.

PROMPT: *What are five emerging technologies that could significantly transform training and education?*

ChatGPT: There are several emerging technologies that have the potential to significantly transform training and education. Here are five examples:

- **Artificial Intelligence (AI):** AI technologies, such as machine learning and natural language processing, can revolutionize education by enabling personalized learning experiences, intelligent tutoring systems, automated grading, and adaptive content delivery.
- **Virtual Reality (VR) and Augmented Reality (AR):** VR and AR technologies create immersive and interactive learning environments. They can be used to simulate real-world scenarios, enhance experiential learning, and provide engaging and interactive content across various subjects.
- **Blockchain:** Blockchain technology can revolutionize credentialing and verification processes by providing secure and tamper-proof digital records of educational achievements. It enables the issuance of digital certificates, degrees, and badges, offering increased transparency and portability of credentials.
- **Internet of Things (IoT):** IoT devices and sensors can be utilized in educational settings to gather data and improve learning outcomes. Smart classrooms equipped with IoT devices can enhance student engagement, monitor student performance, and facilitate personalized learning experiences.
- **3D Printing:** 3D printing allows the creation of physical objects from digital designs. In education, it enables hands-on learning experiences, prototyping, and the creation of visual aids and models that enhance understanding in various subjects, including science, engineering, and arts.

PROMPT: *Thank you. Now, please write a vignette set 10 years in the future. In the vignette, AI is used to train cyber operators. The AI system continuously assesses the cyber operator. The system develops and delivers learning experiences based on the operator's individual needs. Include technical details about how the system assesses operators and develops and delivers learning experiences.*

ChatGPT: The year was 2033, and the realm of cybersecurity had undergone a remarkable transformation. In the high-tech hub of Cybertech City, a state-of-the-art training facility housed the most advanced Artificial Intelligence (AI) system known as the AI Sentinel. This AI system was designed to train and continuously assess cyber operators, tailoring learning experiences to their individual needs.

Within the facility, a cyber operator named Alex Donnelly was preparing for another day of intensive training. As Alex entered the training room, a holographic display materialized, presenting an AI-powered avatar named Sentinel. The Sentinel avatar had an almost human-like appearance, with an aura of intelligence and authority.

The AI Sentinel was built upon a complex neural network architecture, constantly processing vast amounts of data to assess and evaluate the performance of each cyber operator. It analyzed real-time

data streams, including the operator's interactions with simulated cyber threats, system logs, and historical performance records.

The system employed cutting-edge techniques such as Natural Language Processing, Machine Learning, and Deep Neural Networks to gauge the cyber operator's knowledge, skills, and decision-making abilities. It tracked their response times, accuracy in identifying threats, and effectiveness in mitigating attacks.

To ensure a comprehensive evaluation, the AI Sentinel went beyond technical skills and delved into the operator's psychological and behavioral aspects. It monitored their stress levels, decision making under pressure, and ability to adapt to evolving threats. This holistic assessment allowed the system to gain a nuanced understanding of the operator's strengths and weaknesses.

Based on the assessment results, the AI Sentinel dynamically generated personalized learning experiences for each cyber operator. It leveraged sophisticated algorithms to curate a tailored curriculum, combining theoretical knowledge, hands-on exercises, and simulated scenarios. The system adapted the difficulty level and pacing of the training material to match the operator's current proficiency and learning speed.

The learning experiences were delivered through a combination of virtual reality simulations, interactive modules, and collaborative exercises. The system utilized realistic cyber threat simulations, emulating the complexity and intensity of real-world scenarios. It provided instant feedback, highlighting areas of improvement and offering targeted guidance for skill enhancement.

The AI Sentinel employed an intelligent recommendation engine to suggest relevant resources, including articles, research papers, and case studies, based on the operator's specific needs and interests. It promoted a continuous learning mindset by nudging operators towards new and emerging cybersecurity trends, ensuring they remained at the forefront of their field.

In addition to individualized training, the AI Sentinel fostered a collaborative environment. It facilitated virtual teamwork and allowed operators to engage in collective problem-solving exercises, encouraging the exchange of knowledge and best practices. The system monitored these collaborations, identifying opportunities for skill-sharing and building a strong network of cyber professionals.

Over time, the AI Sentinel refined its training approach, leveraging accumulated data and insights from a global network of cyber operators. It continuously updated its algorithms and content to reflect the rapidly evolving cyber landscape, adapting to emerging threats and vulnerabilities.

As Alex completed another day of training, they marveled at the advancements in cyber education. The AI Sentinel had not only revolutionized the training experience but also elevated the skills and capabilities of cyber operators to unprecedented levels. The world was better prepared to defend against cyber threats, thanks to the tireless guidance and support of the AI Sentinel.

PROMPT: *What technical challenges must be overcome during the next 10 years to make AI Sentinel possible?*

ChatGPT: Over the next 10 years, several technical challenges must be overcome to make the AI Sentinel a reality. These challenges include

- **Data Quality and Availability:** The AI Sentinel relies on high-quality and diverse datasets to effectively assess cyber operators. Ensuring the availability of large-scale, labeled datasets that capture real-world cyber threats and operator performance is crucial. Overcoming data limitations, such as privacy concerns and data biases, while maintaining data integrity and representativeness will be a significant challenge.
- **Real-time Data Processing:** The AI Sentinel requires efficient mechanisms for processing and analyzing vast amounts of data in real-time. This includes handling streaming data from various sources, such as system logs, network traffic, and operator interactions. Developing scalable and low-latency data processing infrastructure capable of handling the velocity and volume of data is essential.
- **Advanced Machine Learning Techniques:** Enhancing the AI Sentinel's assessment capabilities requires advancements in machine learning algorithms. Exploring novel techniques like reinforcement learning, transfer learning, and deep learning architectures will be critical to improve the accuracy, interpretability, and adaptability of the system. Research and development efforts need to focus on designing algorithms that can handle complex and dynamic cybersecurity scenarios effectively.
- **Cyber Threat Simulation:** Building realistic and dynamic cyber threat simulations is a significant technical challenge. Simulations need to encompass a wide range of threat vectors, attack patterns, and adversarial behaviors to ensure comprehensive training and assessment. Creating sophisticated simulations that accurately represent evolving cyber threats, including zero-day exploits and advanced persistent threats, will require continuous research and collaboration with cybersecurity experts.
- **Human-Computer Interaction:** Designing intuitive and user-friendly interfaces for the AI Sentinel is crucial to ensure effective training and assessment. The system should seamlessly integrate with existing cybersecurity tools and workflows, providing operators with actionable insights and guidance without adding cognitive load. Developing natural language processing capabilities, intuitive visualizations, and interactive interfaces that enable efficient interaction with the AI Sentinel will be a challenge.
- **Scalability and Deployment:** The AI Sentinel needs to be scalable and deployable in diverse environments. Developing architectures that can handle large-scale deployments, distributed computing, and integration with existing infrastructure and security systems will be a technical challenge. Ensuring interoperability and compatibility with different cybersecurity tools and platforms is necessary for seamless adoption.

PROMPT: *What cultural challenges must be overcome during the next 10 years to make AI Sentinel possible?*

ChatGPT: In addition to the technical challenges, there are several cultural challenges that must be overcome in the next 10 years to make AI Sentinel and similar AI-based training systems possible. These cultural challenges include

- **Trust and Acceptance:** Building trust and acceptance in AI systems for training and assessment is crucial. Cyber operators and the broader community must trust that the AI Sentinel can effectively evaluate their skills and provide personalized learning experiences. Overcoming skepticism, fear of automation, and addressing concerns about bias, fairness, and accountability are essential to foster trust and widespread acceptance.
- **Mindset Shift:** Embracing a mindset of continuous learning and adaptation is necessary for the success of AI-powered training systems. Operators and organizations must be open to new approaches and be willing to adapt their training methodologies. A cultural shift is needed to prioritize lifelong learning, viewing AI as a tool for augmentation rather than replacement, and recognizing the value of personalized learning experiences.
- **Collaboration and Knowledge Sharing:** Cybersecurity is a field that thrives on collaboration and knowledge sharing. Encouraging cyber operators to embrace collaboration and share their expertise, successes, and failures will be critical. Breaking down silos and fostering a culture of collective problem-solving and continuous improvement will facilitate the development and adoption of AI-powered training systems.
- **Ethical Considerations:** Incorporating AI into training and assessment raises ethical considerations. Ensuring the responsible and ethical use of AI systems, addressing privacy concerns, and mitigating potential biases are cultural challenges that must be addressed. Organizations need to establish ethical guidelines, transparency, and mechanisms for accountability to ensure the proper use of AI in training and assessment.
- **Change Management and Training:** Introducing AI Sentinel and similar systems requires effective change management strategies. Organizations need to invest in training programs to familiarize cyber operators with the AI-powered training environment. Training should focus not only on the technical aspects but also on the benefits, limitations, and proper utilization of AI systems. Cultivating a culture of learning, adaptability, and resilience will be crucial in embracing the changes brought by AI technology.
- **Balancing Automation and Human Judgment:** AI systems like the AI Sentinel should be viewed as tools to augment human capabilities, rather than replacing human judgment entirely. Striking the right balance between automation and human decision-making is a cultural challenge. Encouraging operators to leverage the insights provided by AI systems while retaining critical thinking, creativity, and ethical reasoning skills will be important.

PROMPT: *Imagine that you are an adversary. What are potential vulnerabilities in a system like AI Sentinel that you would try to exploit?*

ChatGPT: As an adversary, there are several potential vulnerabilities in a system like AI Sentinel that I would try to exploit. These vulnerabilities include

- **Data Poisoning:** I would attempt to manipulate the training data used by AI Sentinel to introduce biases or misleading patterns. By injecting malicious or misleading data points into the training dataset, I could try to manipulate the system's learning process and potentially cause inaccurate assessments or generate false positives/negatives.
- **Adversarial Examples:** Adversarial examples are crafted inputs specifically designed to deceive machine learning algorithms. I would aim to generate adversarial examples that fool the AI Sentinel into misclassifying or misjudging cyber threats or operator performance. By exploiting vulnerabilities in the system's algorithms or feature extraction methods, I could craft inputs that appear benign to human observers but are misinterpreted by the AI Sentinel.
- **Model Tampering:** If I could gain unauthorized access to the AI Sentinel's model, I could tamper with it to introduce biases, modify decision thresholds, or inject malicious code. By compromising the integrity of the model, I could manipulate the system's behavior and potentially influence its assessments, leading to false results or compromised training experiences.
- **Privacy Breaches:** If the AI Sentinel relies on sensitive or personally identifiable information, I would target vulnerabilities in data storage, transmission, or access controls to breach the system's privacy measures. By gaining unauthorized access to personal data, I could potentially exploit privacy breaches to compromise the system's integrity or undermine user trust.
- **System Exploitation:** I would search for vulnerabilities in the infrastructure supporting the AI Sentinel, such as the underlying hardware, software, or network components. Exploiting vulnerabilities in the system's infrastructure could allow me to gain unauthorized access, disrupt the system's operations, or compromise the integrity of the training data or assessment processes.

PROMPT: *Aside from AI, what other emerging technologies could enable a system like AI Sentinel?*

ChatGPT: Aside from AI, several other emerging technologies could enable a system like AI Sentinel. These technologies include

- **Internet of Things (IoT):** IoT devices can provide a wealth of real-time data that can be used to enhance the training and assessment of cyber operators. By integrating IoT devices into the training environment, AI Sentinel can gather data from various sources, such as network traffic, system logs, and device telemetry, enabling more comprehensive and accurate assessments.
- **Blockchain:** Blockchain technology can enhance the security, transparency, and integrity of the training and assessment processes. By leveraging blockchain, AI Sentinel can securely store training data, assessment records, and certifications, providing an immutable and tamper-resistant audit trail. Blockchain can also facilitate trust and accountability among the cyber operator community.
- **Virtual Reality (VR) and Augmented Reality (AR):** VR and AR technologies can create immersive training environments that simulate realistic cyber threats and scenarios. By incorporating

VR and AR into AI Sentinel, operators can engage in hands-on training experiences, interact with virtual systems, and practice their skills in a safe and controlled environment.

- **Big Data Analytics:** The ability to analyze and derive insights from large volumes of data is crucial in training and assessment. Big Data analytics technologies enable AI Sentinel to process and analyze diverse data sources, identify patterns, detect anomalies, and provide valuable feedback to operators based on their performance.
- **Natural Language Processing (NLP):** NLP technology can enhance the interaction and communication between operators and AI Sentinel. By understanding and processing natural language, AI Sentinel can provide personalized feedback, answer questions, and assist operators in their learning journey through conversational interfaces and chatbots.
- **Edge Computing:** Edge computing brings computation and data storage closer to the data source, reducing latency and enabling real-time processing. By leveraging edge computing, AI Sentinel can perform assessments and deliver personalized learning experiences in near real-time, even in resource-constrained environments or during network disruptions.
- **Biometric Authentication:** Biometric authentication technologies, such as fingerprint or facial recognition, can enhance the security and identity verification aspects of AI Sentinel. By leveraging biometrics, the system can ensure that the assessments and learning experiences are personalized and delivered to the correct operator.
- **Quantum Computing:** Quantum computing has the potential to revolutionize cybersecurity. With its computational power, quantum computing can enable more sophisticated encryption algorithms, threat analysis, and vulnerability assessments. Integrating quantum computing into AI Sentinel can enhance its capabilities in addressing advanced cyber threats.

Discussion of Foresight Case Study

In the case study, ChatGPT anticipated the future roles of emerging technologies in training and education. The products developed were not intended to be comprehensive. Nonetheless, this example demonstrates how, through a series of interactions with ChatGPT, planners can apply strategic foresight tools.

Figure 5 illustrates the manner of interaction. The planner began by asking ChatGPT to scan for emerging technologies: *“What are five emerging technologies that could significantly transform training and education?”* Next, the planner asked ChatGPT to generate a future-oriented science fiction story centered on one of the technologies: *“Write a vignette set 10 years in the future. In the vignette, AI is used to train cyber operators.”* The planner then performed backcasting by asking ChatGPT to identify *“...technical and cultural challenges that must be overcome during the next 10 years”* to make the concept described in the vignette possible. To give a balanced view, the planner also used ChatGPT to identify potential threats, a critical element of a strengths, weaknesses, opportunities, and threats (SWOT) analysis. Lastly, the planner used ChatGPT to identify other key/critical technologies

to enable the concept. The net effect is that the joint human-machine team can apply a sequence of foresight tools in an effective and efficient manner.

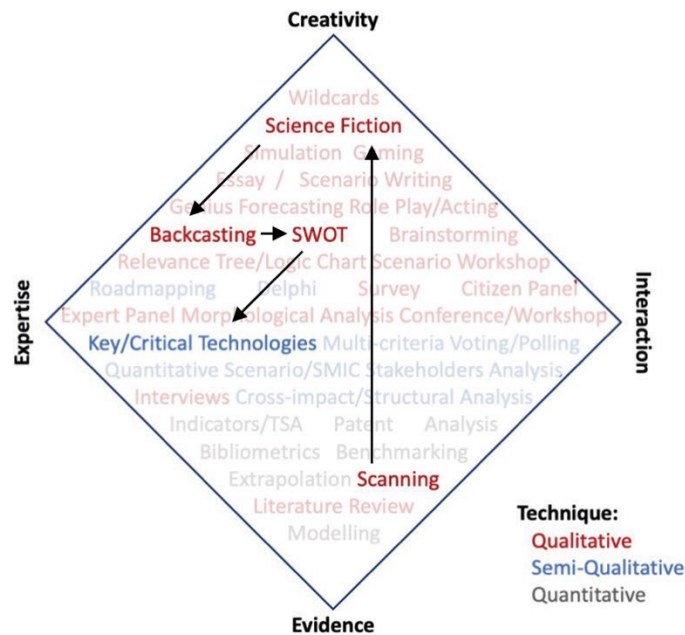


Figure 5: Using ChatGPT for Strategic Foresight

One remarkable aspect of this case study is the breadth of knowledge exhibited by ChatGPT. It identifies emerging technologies across a diverse range of areas. When probed further, ChatGPT also showcases a depth of knowledge. For instance, the case study delves into topics such as real-time performance assessment, physiological state monitoring, and personalization—all of which are burgeoning areas of research in the learning sciences.

The second notable feature is ChatGPT’s ability to synthesize information to generate innovative outputs. The science fiction vignette weaves together multiple topics in a captivating and compelling narrative. The backcasting and SWOT results are insightful. In essence, while ChatGPT is not intended to replace human practitioners, its outputs can undoubtedly serve as a catalyst for creativity.

However, despite its capabilities, ChatGPT does have limitations when it comes to strategic foresight. As discussed in previous sections, ChatGPT’s model was trained on a historical corpus of text, which means it lacks access to the most recent papers. Furthermore, ChatGPT does not have access to proprietary databases and resources that could be informative for strategic foresight. These blind spots may lead ChatGPT to overlook promising options.

Another constraint is that ChatGPT tends to produce modal outputs due to its training methodology. ChatGPT identifies promising training technologies, but the technologies it highlights are unlikely to

surprise an expert in training and education. To circumvent this limitation, planners could prompt ChatGPT to elaborate and to concentrate on “unexpected or novel” concepts.

In conclusion, ChatGPT can facilitate use of methods from all quadrants of the foresight diamond. However, AI tools like ChatGPT are best used in collaboration with human experts, especially in tasks that require deep understanding, critical thinking, and interpretation, such as strategic foresight.

The Road Ahead

In the case studies, ChatGPT contributed to the quality of products generated and, to an even greater extent, the timeliness in generating those products. However, ChatGPT did not obviate human involvement. In each case study, knowledgeable people needed to decompose complex tasks into simpler ones that ChatGPT could accomplish, and they needed to verify its outputs.

This white paper primarily deals with the case *for* LLMs, but along with their benefits come potential risks. Specifically, misuse (dependence on LLMs where their use is unwarranted) and abuse (exploitation of LLMs for malicious intent). A comprehensive risk mitigation plan must include strategic system design, end-user training, test and evaluation, and robust defensive mechanisms. The decision about whether and how to adopt LLMs must consider these potential risks.

The history of AI is replete with hype. Yet, despite significant advances made during the past 50-plus years, there have been few examples of AI transforming modern life. However, given the remarkable capabilities inherent to LLMs, and their rapid rate of progress, LLMs may finally deliver on AI’s promise.

References

[Bloom 1984]

Bloom, B. S. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*. Volume 13. Pages 4-16. 1984.

[Bornmann et a. 2021]

Bornmann, L.; Haunschild, R.; & Mutz, R. (2021). Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*. Volume 8. Number 1. October 7, 2021. Pages 1-5.

[McKinsey 2022]

McKinsey & Company. The state of AI in 2022—and a half decade in review. *McKinsey & Company Website*. July 31, 2023 [accessed]. <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2022-and-a-half-decade-in-review>.

[Popper 2008]

Popper, Rafael. Foresight methodology. In *The handbook of Technology Foresight: Concepts and Practice*. Georghiou, Luke [editor]. Edward Elgar Publishing. Pages 44-88. 2008. ISBN: 1781008760, 9781781008768.

[Schwartz 1991]

Schwartz, Peter. *The Art of the Long View*. Doubleday Business. 1991. ISBN-13: 978-0385267311.

[Weizenbaum 1976]

Weizenbaum, Joseph. *Computer Power and Human Reason: From Judgment to Calculation*. W. H. Freeman and Company. 1976. ISBN-13: 978-0716704638.

Legal Markings

Copyright 2023 Carnegie Mellon University.

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

The view, opinions, and/or findings contained in this material are those of the author(s) and should not be construed as an official Government position, policy, or decision, unless designated by other documentation.

NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN "AS-IS" BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

Internal use:* Permission to reproduce this material and to prepare derivative works from this material for internal use is granted, provided the copyright and “No Warranty” statements are included with all reproductions and derivative works.

External use:* This material may be reproduced in its entirety, without modification, and freely distributed in written or electronic form without requesting formal permission. Permission is required for any other external and/or commercial use. Requests for permission should be directed to the Software Engineering Institute at permission@sei.cmu.edu.

* These restrictions do not apply to U.S. government entities.

DM23-0847

Contact Us

Software Engineering Institute
4500 Fifth Avenue, Pittsburgh, PA 15213-2612

Phone: 412/268.5800 | 888.201.4479

Web: www.sei.cmu.edu

Email: info@sei.cmu.edu