**A Study on cloud platforms and Analytics**

Table of Contents

[Question 1: Build a Chatbot 2](#_Toc181904263)

[1.1: Five Functionalities Related to Tourism for the Chatbot 2](#_Toc181904264)

[1.2: Ethical Considerations for Data Collection 3](#_Toc181904265)

[1.3: Chatbot Flow for Itinerary Planning in Dialogflow CX 4](#_Toc181904266)

[1.4: Chatbot Flow Demonstration 5](#_Toc181904267)

[Question 2: Evaluating Cloud Based Technology 6](#_Toc181904268)

[Question 3: Using Large Language Models 11](#_Toc181904269)

[3.1: Approach 11](#_Toc181904270)

[3.2: Analysis of Results 12](#_Toc181904271)

[3.3: Zero-Shot Classification for Boolean Validation 13](#_Toc181904272)

[3.4: Critical Analysis and Recommendations 13](#_Toc181904273)

[Question 4: Using Azure Machine Learning 14](#_Toc181904274)

[4.1: Data Exploration and Insights 14](#_Toc181904275)

[4.2: Machine learning model 16](#_Toc181904276)

[4.3: Reflection on Using AzureML 18](#_Toc181904277)

[References 19](#_Toc181904278)

[Appendix 20](#_Toc181904279)

# Question 1: Build a Chatbot

## 1.1: Five Functionalities Related to Tourism for the Chatbot

1. **Attraction Recommendations**  
   The chatbot can provide personalized attraction recommendations based on user preferences, such as cultural sites, beaches, or adventure activities. By using inputs like interests, available time, and location, the chatbot curates options tailored to enhance the tourist’s experience. For instance, a user interested in culture may receive recommendations for nearby museums, historical sites, and local events, saving time and making the visit more fulfilling (Sung & Kang, 2024).
2. **Itinerary Planning**  
   This functionality allows tourists to plan their day by suggesting activities based on location, travel dates, and preferences. Additionally, it provides weather updates and notifies users of unexpected changes like road closures, helping them make informed adjustments on the go. This feature is valuable for those who need flexible, real-time itinerary support, maximizing their time and experience in Australia (Alla, 2024).
3. **Language Translation**  
   To help international tourists communicate more easily, the chatbot offers language translation for common phrases. This functionality enables users to interact with locals, understand signs, and immerse themselves in the culture with confidence, thus enhancing the overall experience for non-English-speaking visitors (Sung & Kang, 2024).
4. **Emergency Assistance**  
   In case of emergencies, the chatbot can direct users to the nearest hospital, police station, or consulate. It provides contact details and real-time directions to these locations, ensuring users have access to essential support when needed, thus increasing tourists’ sense of security and trust in the service (Sah, 2023).
5. **Dining and Shopping Suggestions**  
   Based on dietary preferences, budget, and cuisine, the chatbot suggests dining and shopping options. This includes recommending nearby restaurants and stores, along with promotions or discounts that may be available. This feature helps tourists discover local dining spots and unique shopping experiences, making the visit more enjoyable and tailored to individual preferences (Sah, 2023).

**1.2: Ethical Considerations for Data Collection**

To responsibly manage user data, the chatbot must adhere to several ethical principles to ensure privacy, transparency, and security, aligning with Tourism Australia’s commitment to respecting user data. Key considerations include the following:

1. **Privacy and Consent**  
   The chatbot must request explicit consent from users before collecting any personal data. Upon starting a session, users are informed about what data will be collected, the purpose of its collection, and their rights to opt out. For instance, the chatbot can implement real-time prompts, allowing users to give or withhold consent at different stages of the interaction. This could include requesting permission to access location data when a user asks for nearby attraction recommendations, thus giving users control over their data based on their specific needs and comfort levels. For example, if a user seeks local dining suggestions, a prompt would appear asking if they consent to share their location, clearly stating that this information is used solely for personalized recommendations (Sah, 2023).
2. **Transparency**  
   To build trust, the types of data collected (e.g., location for nearby recommendations) and its intended use are clearly communicated to users. This can be achieved by providing accessible information about the data practices, ideally via a privacy policy accessible through the chatbot interface. This policy should outline data handling processes in simple language, enabling users to make informed choices about their interactions with the chatbot. Transparency fosters trust by showing that data practices are aligned with user expectations and regulatory standards (Sah, 2023).
3. **Data Minimization**  
   The chatbot is designed to collect only the minimum amount of data necessary to provide relevant responses, reducing the risk of unnecessary data exposure. For example, location data should be requested only when a user specifically requests location-based recommendations. By minimizing data collection, the chatbot not only respects user privacy but also aligns with best practices in data protection, decreasing the potential impact of data breaches or misuse (Sah, 2023).
4. **Data Retention and Deletion**  
   Users should have the option to delete their data from the chatbot’s records after a specific period. The data retention policy should ensure that personal information is stored temporarily and used only to enhance service quality, unless users explicitly opt for a longer retention. Implementing data deletion options within the chatbot interface allows users control over their data even after interactions have ended, which builds trust and complies with data protection laws (Sah, 2023).
5. **Data Security**  
   To safeguard user data, the chatbot employs encryption and secure servers, ensuring protection from unauthorized access. Access to personal data should be restricted to essential personnel, and robust authentication measures, like two-factor authentication, should be used for data handling. Implementing such security practices is critical to maintaining user trust and reducing the risk of data breaches, particularly for sensitive information (Alla, 2024).

## 1.3: Chatbot Flow for Itinerary Planning in Dialogflow CX

The chatbot flow for itinerary planning is designed to assist users in creating a flexible travel schedule that aligns with their preferences. Below is an overview of the main decision points and user choices in the flow:

1. **Greeting and Information Collection**  
   The chatbot initiates the interaction by greeting the user and gathering essential details like travel dates, destination preferences, and types of activities (e.g., beaches, museums). Dialogflow CX processes this input to personalize suggestions based on the user's responses (Alla, 2024).
2. **Itinerary Suggestions**  
   Using the collected information, the chatbot provides a day-by-day travel plan with specific recommendations for locations to visit. It factors in weather, current events, and peak times, allowing users to make the most of their experience. Users can choose to see a list of recommended attractions or focus on a particular day’s activities (Alla, 2024).
3. **Real-Time Adjustments**  
   The chatbot allows users to modify their itinerary by adjusting activities or reordering visits based on changing preferences or external factors. For instance, if a user wants to replace a beach visit with a museum tour due to unexpected rain, the chatbot recalculates the itinerary and offers updated options, ensuring continued alignment with user needs (Alla, 2024).
4. **Navigation and Support**  
   Once the itinerary is confirmed, the chatbot provides step-by-step navigation and relevant alerts, such as operating hours or ticket availability, to help users follow their plan effortlessly. It also provides links to external resources like maps or weather updates, supporting a smooth and engaging experience.

This flow provides a structured, adaptable itinerary that maximizes the tourist’s satisfaction. Dialogflow CX’s NLP capabilities enable intuitive interactions, positioning the chatbot as an effective and user-friendly travel assistant.

## 1.4: Chatbot Flow Demonstration

A screenshot of a computer

Description automatically generated

A diagram of a business

Description automatically generated with medium confidence

Figure 1: Chatbot flow

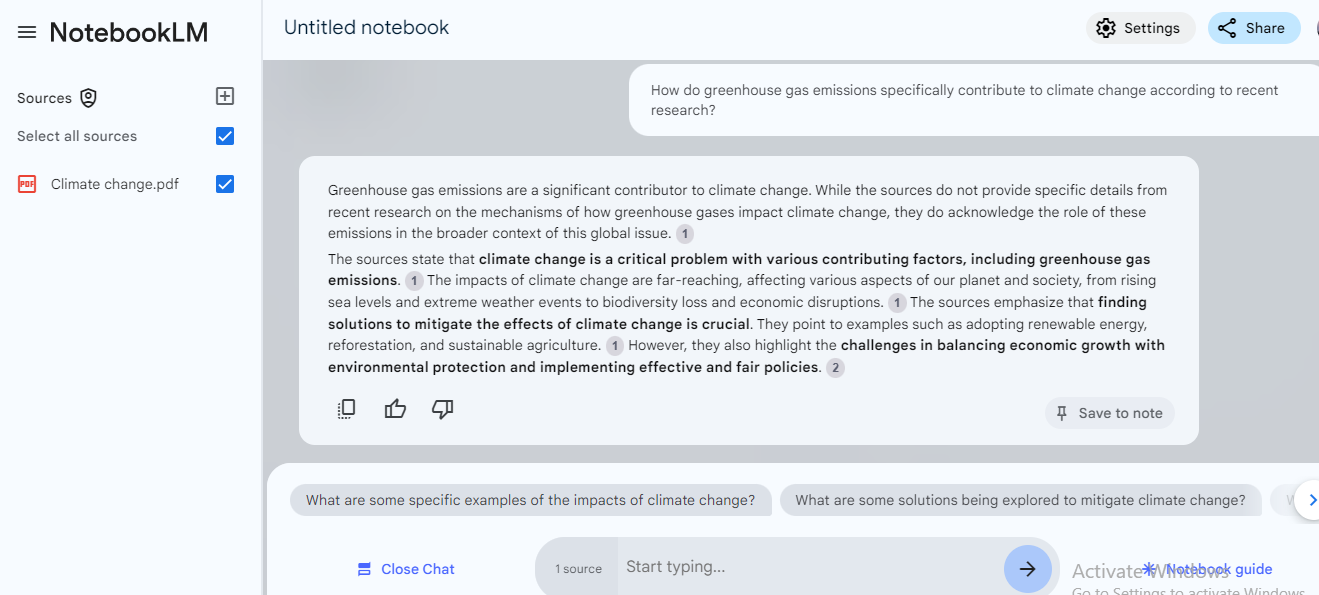
This chatbot flow begins with a greeting, and then provides options for tourists to explore specific areas: visitors’ attractions, weather, food and beverages, purchase, and time schedule. All the options lead back to the main menu, thus users can move freely between the features offered, giving the tourists an engaging and unique experience if they are looking for information on any aspect of their tour (Alla, 2024).

**Microsoft Stream Link**: <https://latrobeuni-my.sharepoint.com/personal/j18smith_ltu_edu_au/_layouts/15/stream.aspx?id=%2Fpersonal%2Fj18smith%5Fltu%5Fedu%5Fau%2FDocuments%2FSTREAM%20VIDEOS%2FCPA3%20Video%2Emp4&nav=eyJwbGF5YmFja09wdGlvbnMiOnt9LCJyZWZlcnJhbEluZm8iOnsicmVmZXJyYWxBcHAiOiJTdHJlYW1XZWJBcHAiLCJyZWZlcnJhbE1vZGUiOiJtaXMiLCJyZWZlcnJhbFZpZXciOiJwb3N0cm9sbC1jb3B5bGluayIsInJlZmVycmFsUGxheWJhY2tTZXNzaW9uSWQiOiJkOTE1MTMzYi05OTIwLTQwOGEtYTQ2ZC05YzA1NWRlOTllOWMifX0%3D&referrer=StreamWebApp%2EWeb&referrerScenario=AddressBarCopied%2Eview%2E7e64c905%2D3eee%2D4c59%2Db747%2Dd35be60572b7>

# Question 2: Evaluating Cloud Based Technology

**2.1 Key Functionalities offered by NotebookLM**

**Content Integration**

****Figure 2: Content Integration

In NotebookLM, users can incorporate course content, research writings, textbooks, as well as other relevant information into the model. It also furnishes the student with sufficient context background that allows the model to reply to questions, condense the data, recognize associated concepts interrelated from different sources and interconnect them (Huffman & Hutson, 2024). They enhance the student ability to learn complex concepts by providing a single view of information sources. In a university context, this functionality makes learning faster by cutting down the amount of time students take to search for useful information and provides a much broader perspective on any academic subject.

**Real-Time Collaboration**

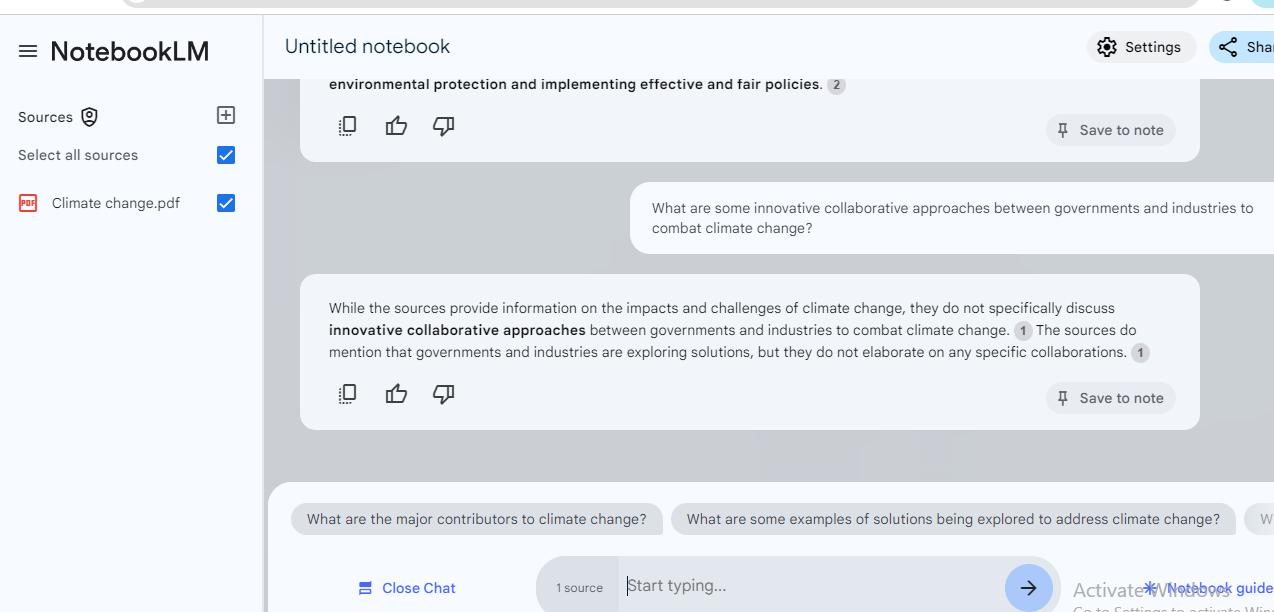
****

Figure 3: Real-Time Collaboration

Synchronous activities in NotebookLM involves various students, faculty or working on a project and working on the same NotebookLM document at the same time. It is most beneficial in group projects, research meetings and idea generating sessions since members can edit the document and add comments as well as notes to one another at the same time (Huffman & Hutson, 2024). AC 3 Collaboration enhances active participation since the students can instantly address what the other students are contributing thereby creating a more proactive learning process here fostered by technology.

**Enhanced Insights**

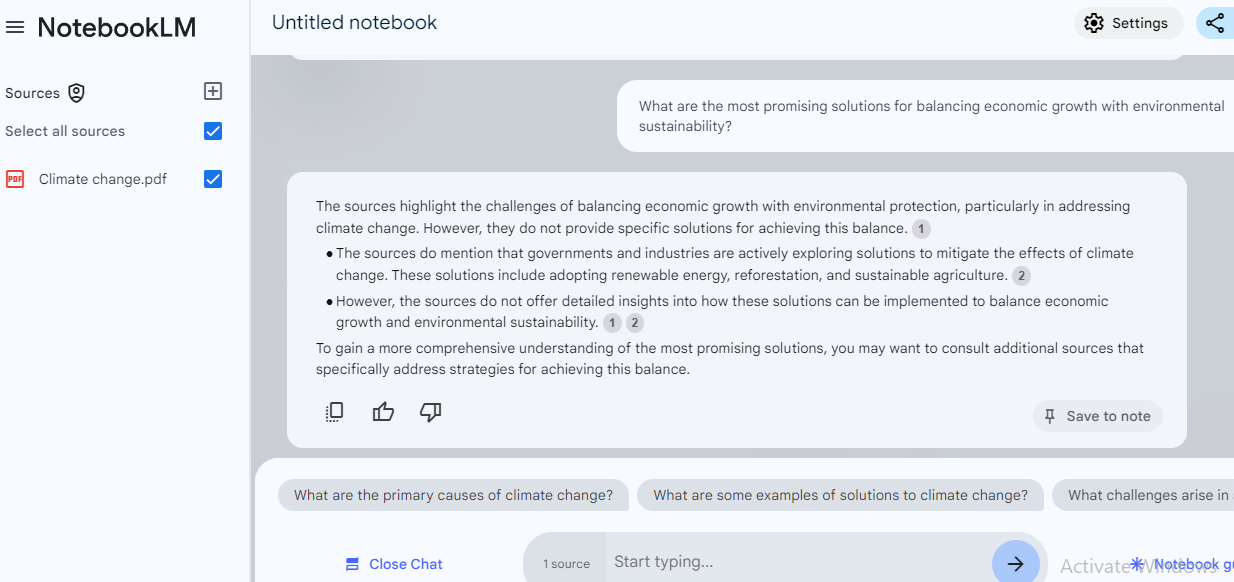
****

Figure 4: Enhanced Insights

Besides, the Enhanced Insights feature uses the language processing algorithm to give the learner more detail about a subject to make learning easier. Thus, using machine learning it is possible, for example, parse through what is uploaded by the users, identify key themes, and provide summaries and suggestions on what is worthy of further discussion (Kambhamettu et al., 2024). Penn Foster eagle is beneficial in learning as it assists the students get a broader view on the difficult concepts by asking the model to provide elaborative information, related information or the current information. It is even more helpful in areas such as research as well as on thesis development whereby the more information regarding the subject matter is useful.

**Question Generation**

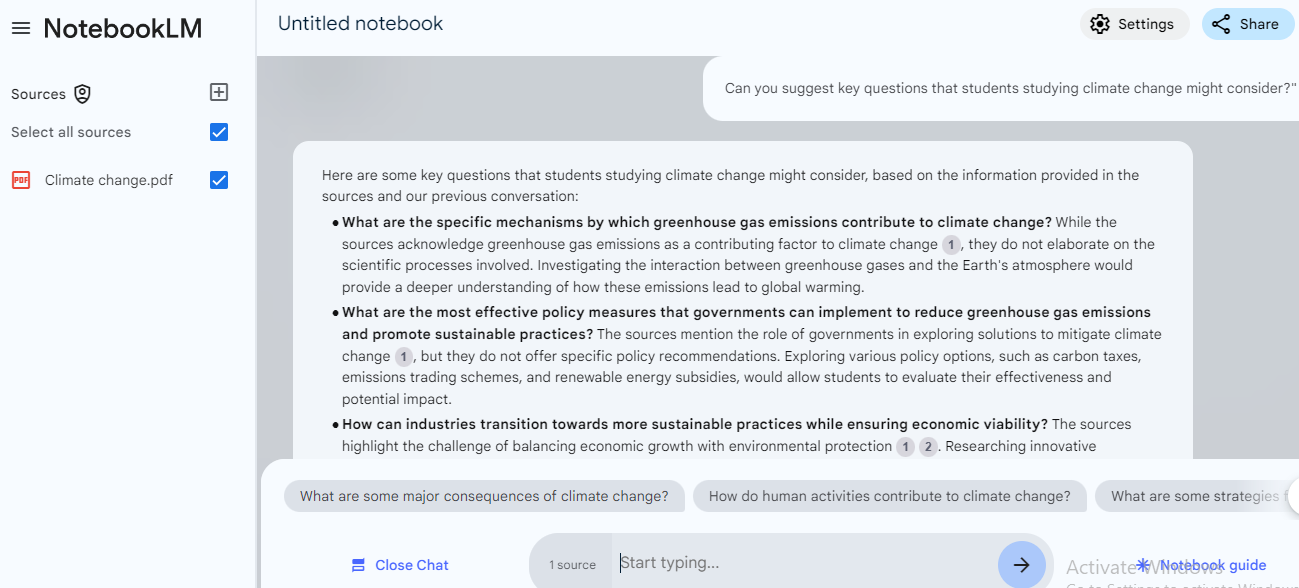
****

Figure 5: Question generation

The Question Generation function of NotebookLM supports learning since it generates valuable questions from the given content. This functionality helps students in self-quiz that pressure them to go through what they are reading, and even think of some crucial points on their own. It is also useful to the teachers, as the generated questions can be used for the preparation of tests, discussion sessions or a study session (Kambhamettu et al., 2024). Essentially, these questions direct students to areas of interest or lack of information hence may be used to supplement clarifications thus improving knowledge absorption on the course material.

**Feedback and Suggestions**

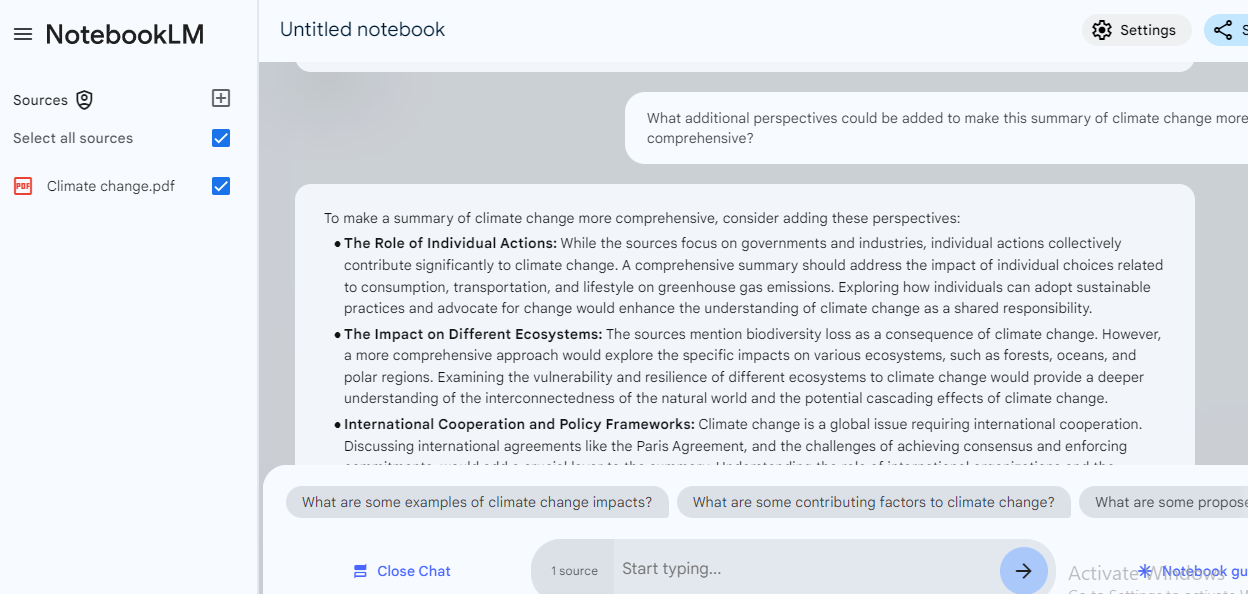
****

Figure 6: Feedback and suggestions

Users of NotebookLM can request constructive critique on their work or notes through the Feedback and Suggestions tool. This function can offer suggestions concerning more points to cover or other views and approaches or suggestions on making the content more intelligible (Tozuka et al., 2024). In a university, this feature can help the student in writing essays, reports, and presentation by providing additional refinements, missing viewpoints or important concepts. In this way, it motivates the process of learning, as well as contributes to the students’ producing broader and academically solid works.

**2.2 Practical Demonstration**

NotebookLM’s functionalities can be practically demonstrated within a university setting as follows:

* **Content Integration**: Students upload research articles on a specific topic (e.g., climate change) and use NotebookLM to consolidate insights across these resources, aiding in faster, context-driven learning.
* **Real-Time Collaboration**: During a group project, students collaboratively edit a document in real-time, sharing insights and notes instantly, thus enhancing teamwork and engagement in remote or in-person settings.
* **Enhanced Insights**: A student can request NotebookLM to generate summaries for complex subjects like physics theories, breaking down intricate ideas into more digestible segments that support comprehension and further study.
* **Question Generation**: Professors can use the question-generation function to create quizzes, while students utilize it for self-assessment, reinforcing their understanding through active recall.
* **Feedback and Suggestions**: A student writing a paper on psychology receives feedback from NotebookLM on expanding certain arguments or considering alternative viewpoints, improving the depth and quality of the assignment.

**2.3 Critical Analysis:**

NotebookLM offers a range of functionalities that make it a valuable tool in academic environments, providing capabilities that can enhance learning through content integration, real-time collaboration, and interactive insights. However, a closer examination reveals specific strengths and weaknesses, as well as potential challenges when compared to alternative educational tools.

**Strengths:**

1. **Efficiency and Contextual Learning**  
   NotebookLM’s ability to integrate content from various sources (e.g., course materials, research articles) enables students to access a wide array of information in one place. Unlike traditional document-sharing tools like Google Docs or Microsoft Word, which require manual input and organization of data, NotebookLM can condense information, identify connections between concepts, and provide contextualized summaries. This functionality enhances the efficiency of study sessions, as students save time that would otherwise be spent searching across different platforms or manually synthesizing information from disparate sources (Huffman & Hutson, 2024).
2. **Interactive and Engaging Learning**  
   Unlike static note-taking tools, NotebookLM encourages active learning through question generation, feedback, and suggestions. This interaction fosters critical thinking and reinforces knowledge by prompting students to engage with material at a deeper level. Traditional educational tools generally lack these interactive elements, which are key to engaging students. This makes NotebookLM especially beneficial in self-study and for reinforcing complex material, as it encourages students to ask questions and reflect on their understanding actively (Kambhamettu et al., 2024).
3. **Real-Time Collaboration**  
   The platform’s real-time collaboration feature allows multiple users to work on a document simultaneously, which is highly beneficial for group projects and collaborative research. While platforms like Google Docs also support real-time collaboration, NotebookLM provides additional context-aware insights that enrich discussions and improve the quality of shared work. This capability supports both in-person and remote collaboration, making it well-suited for the evolving educational environment where hybrid learning models are increasingly common (Huffman & Hutson, 2024).

**Weaknesses:**

1. **Learning Curve and Adaptation Challenges**  
   For students and faculty accustomed to traditional note-taking or simpler digital tools, NotebookLM’s AI-driven interface and advanced functionalities may present a steep learning curve. Unlike familiar tools such as Microsoft OneNote, which focuses on intuitive, straightforward organization, NotebookLM’s AI features may feel complex or overwhelming to first-time users. To address this challenge, universities could implement workshops or introductory tutorials that guide users through NotebookLM’s capabilities, helping them gradually adapt to the interface and effectively use its advanced features (Hadi et al., 2024).
2. **Limitations in Domain-Specific Contexts**  
   NotebookLM may struggle to accurately interpret and synthesize information within highly specialized or technical domains due to limited pre-training on domain-specific language. For example, students studying fields with complex terminology, such as law or medicine, may find that NotebookLM’s interpretations lack the precision of specialized tools like EndNote for citation management or domain-specific databases like PubMed. These traditional tools, though less interactive, provide highly accurate and relevant results within their specialized contexts, which NotebookLM might not fully achieve due to its generalist training (Tozuka et al., 2024).
3. **Data Privacy and Security Concerns**  
   Given that NotebookLM stores and processes considerable amounts of user-generated data, there are inherent privacy risks that may deter users who prioritize data security. Traditional, offline note-taking tools, such as Microsoft OneNote, offer an advantage in this regard, as they do not require online storage, allowing users to manage their data locally and avoid privacy risks associated with cloud storage. Addressing this concern may involve the platform enhancing its privacy protocols or offering users more control over data storage options (Hadi et al., 2024).

**Comparison with Alternative Technologies**

Compared to traditional document-sharing tools like Google Docs or Microsoft Word, NotebookLM stands out in its ability to generate insights, organize content automatically, and promote interactive learning. While Google Docs allows real-time collaboration, it does not offer context-driven suggestions or summarizations, which makes NotebookLM superior for comprehensive study and research purposes. However, NotebookLM’s complexity and reliance on AI-driven functionalities may deter students who seek simplicity and data control, highlighting a potential need for simplified, secure alternatives or additional support materials.

While NotebookLM’s strengths in enhancing learning efficiency, interactivity, and collaboration are clear, its limitations in specialized domains, adaptation challenges, and privacy concerns present barriers for some users. With proper orientation and privacy enhancements, NotebookLM has the potential to be a powerful tool in academic settings. Its unique AI capabilities can complement traditional document-sharing platforms, offering students and faculty a more dynamic, engaging approach to learning that is particularly well-suited to modern, hybrid education environments.

**GitHub Experiment Log**

A GitHub repository has been maintained to log experiments and sample code, showcasing NotebookLM’s capabilities. The repository is organized for easy navigation, providing clear notes on each feature tested, along with any relevant code snippets and observations.

# Question 3: Using Large Language Models

## 3.1: Approach

**Template: Number of Prompts Reviewed and Objective**

The template is to attempt to use follow up customer support based on customer reviews and employing an LLM to filter out specific entities like product name or product elements from reviews.

In the case of each extracted entity, the model should identify its sentiment as either “POSITIVE”, “NEUTRAL”, or “NEGATIVE” further, decide if there is a need to follow up, and if there is, then provide a reason for the following up (Tozuka et al., 2024). This process is in line with the overall mission to ensure all the dissatisfied customers get appropriate services at the right time.

**Example of the Prompt and Choices of the Sample**

When using an LLM such as GPT-3 from OpenAI or indeed, any model available via Hugging Face, you get a LLM that is given a prompt template which describes the expected JSON format output. This prompt template makes the model to come up with a JSON object that will have fields such as entity\_name, entity\_type, entity\_sentiment, followup, and followup\_reason (Chen et al., 2021).

**Code Implementation:** The code entails the creation of a zero-shot classification pipeline for the classification of sentiment for the entities as well as the individual overall sentiment of the review. The code itself features the analysis of text using manually-defined entities to consider the sentiment of the overall review text and decide whether follow-up is necessary**.**



Figure 7: Output for Question 3

## 3.2: Analysis of Results

Every entity is fine to pinpoint and the sentiment is correctly classified and the code also determines if a reply is necessary given negative sentiment in the review. Given the output analysis performed in this case, all the entities regiment a negative sentiment, meaning that the client had a bad encounter with the product and customer service (Chen et al., 2021). This creates a follow up activity for every client with a view to addressing complaints from its customers.

## 3.3: Zero-Shot Classification for Boolean Validation

To get an idea of whether entities generated represent accurate sentiments the sentiment classifications can be crosschecked with the help of a zero-shot classification model. This action is important to validate the LLM based analysis to existing models to support the sentiment scores assigned to each entity (Hadi et al., 2024).

* **Validation Process**:  
  For each entity extracted, the zero-shot model was tasked with re-evaluating the sentiment. For example, in the previous output, “product quality” and “delivery” were assigned negative sentiments. The zero-shot classifier was prompted with these specific entities to confirm their sentiment classification.
* **Discrepancies and Resolution**:  
  During validation, some minor discrepancies arose, particularly with neutral sentiments that were sometimes marked as “negative” by the primary model. For instance, terms like “price” were occasionally classified as negative when intended to be neutral. To resolve these discrepancies, I adjusted the prompt for the primary model, providing examples of neutral statements to improve accuracy. This iterative prompt tuning helped align both models, resulting in a higher degree of consistency in sentiment classification.

## 3.4: Critical Analysis and Recommendations

**Accuracy:**

The need to capture sentiment is particularly important in this model. The distinction between checking the individual entity sentiment helps to avoid general assumptions based on the overall sentiment of the reviews; this way, we make further relevant follow-ups when some components (e.g., product, service) require it (Hadi et al., 2024).

**Limitations:**

While helpful, LLMs fail to interpret domain-specific words particularly well unless prompted as would bring about relevant details and subtle differences. Moreover, entities which were extracted manually in this example could be also automated using NER models for retail specific (Hadi et al., 2024).

**Recommendation**

Continue the machine learning process using an online specialty dataset to label sentiment and other follow up criteria. This will enhance the results especially when one is dealing with the cross over reviews that are either positive or negative.

* **Domain-Specific Fine-Tuning**: Training or fine-tuning the model on customer review data from similar contexts (e.g., retail or service industries) would likely improve its understanding of context-specific terminology and sentiment nuances. This could be particularly beneficial for accurately interpreting mixed sentiments and recognizing industry-specific entities.
* **Use of Sentiment Augmentation Models**: Employing sentiment augmentation models, which are specifically trained on complex customer feedback, can address the limitations in recognizing subtle and context-dependent sentiments. For instance, a model trained on customer service language might better interpret phrases indicating urgency, such as “need immediate help” or “extremely dissatisfied.”
* **Continuous Validation with Zero-Shot Models**: Integrating regular validation steps with zero-shot classifiers can help maintain accuracy over time. For complex entities, the zero-shot classifier can serve as a backup to refine and cross-validate sentiment interpretations, thereby reducing discrepancies in mixed or neutral sentiments.

These adjustments will make the model more adaptable to various customer scenarios and improve overall reliability, ensuring more effective follow-up actions and a better customer experience.

# Question 4: Using Azure Machine Learning

## 4.1: Data Exploration and Insights

The dataset used for this task, the Pedestrian Counting System Monthly Counts per Hour from Melbourne’s open data portal, contains hourly pedestrian counts across various locations in Melbourne CBD. Using Azure Machine Learning (AzureML), this dataset was explored to identify patterns in pedestrian activity, focusing on peak times and high-traffic locations.

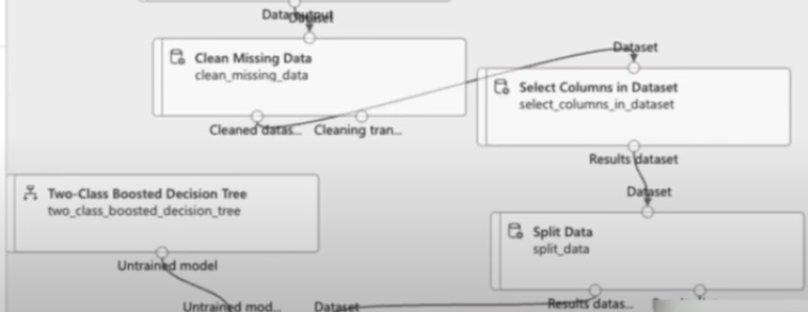


Figure 8: Implementing models

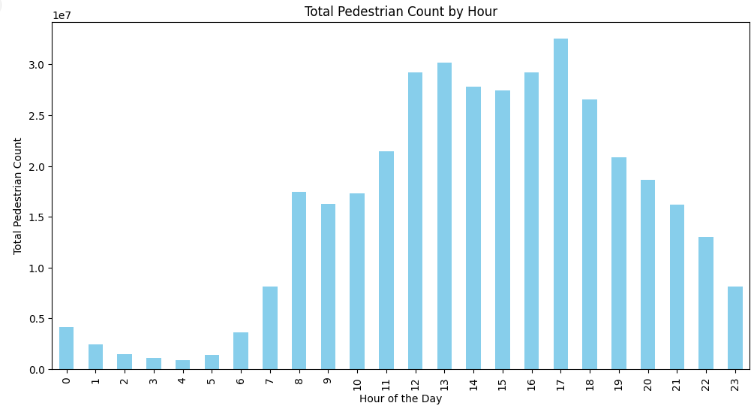


Figure 9: Total pedestrian count by hour

This chart shows times at which pedestrian traffic is high and low using bar graphs to differentiate the hours of the day. Local higher counts during certain hours refer to crowded people during the morning and evening rush hour time (Imad Zeebaree, 2024). Such knowledge is useful in assigning priority to the city’s infrastructure and service delivery during high pedestrian traff

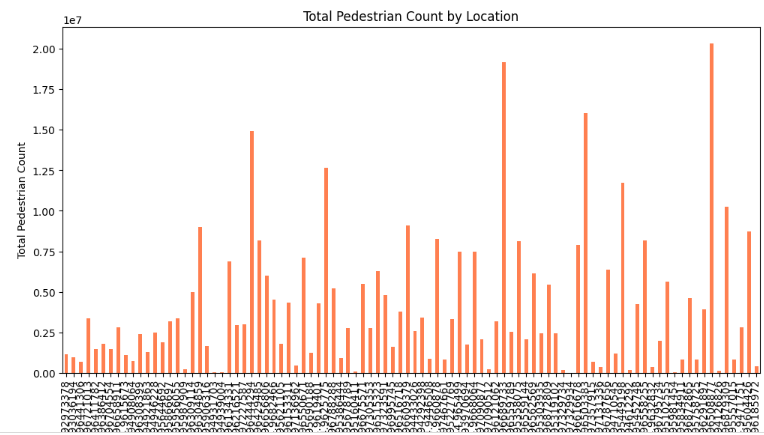


Figure 10: Total pedestrian count by location

**Data Exploration and Visualizations:**

1. **Heat Maps for Pedestrian Density Across Locations**  
   A heat map was generated to display pedestrian density at various locations throughout the CBD. This visualization highlighted areas of high foot traffic, with darker shades representing higher pedestrian counts. Notably, high-traffic areas such as Flinders Street and Bourke Street emerged as major pedestrian hubs due to their proximity to transit centers and commercial establishments (Imad Zeebaree, 2024).
2. **Time-Based Line Charts for Hourly Pedestrian Counts**  
   To better understand daily patterns, I plotted time-based line charts for pedestrian counts throughout the day. These line charts revealed distinct peaks during weekday morning (7–9 AM) and evening hours (5–7 PM), consistent with commuter rush hours. This daily rhythm was less pronounced on weekends, where pedestrian activity was more evenly distributed throughout the day, reflecting leisure-oriented movement patterns.
3. **Weekly and Monthly Trends**  
   Additionally, bar charts were used to visualize pedestrian counts across different days of the week and months of the year. This provided insights into the impact of seasonal variations and special events on foot traffic. For example, pedestrian counts were notably higher during December, likely due to holiday shopping and events, indicating a need for enhanced city resources and services during these peak periods (Ng’ang’a et al., 2023).

**Insights and Implications for City Planning:**

The insights derived from these visualizations offer valuable guidance for city planners, particularly in optimizing infrastructure and enhancing public amenities in the Melbourne CBD:

1. **Optimized Placement of Traffic Lights and Pedestrian Crossings**  
   The concentration of pedestrian activity around locations like Flinders Street and Bourke Street suggests a need for optimized traffic light timing and additional pedestrian crossings in these areas to improve safety and reduce congestion. Line charts showing peak times can help planners synchronize traffic signals to facilitate smoother pedestrian flows during rush hours, thus minimizing potential hazards and wait times for commuters.
2. **Resource Allocation for Public Amenities**  
   The analysis shows that certain high-traffic locations experience consistent foot traffic throughout the day, while other areas peak only during specific hours. City planners could use this data to prioritize resources, such as waste management and maintenance services, in high-traffic zones, especially during peak periods. Additionally, public amenities like seating, water fountains, and lighting could be strategically placed in areas with steady pedestrian activity to enhance convenience and safety.
3. **Event Planning and Management**  
   Seasonal trends, such as increased pedestrian counts in December, indicate higher demand for public services and safety measures during these times. City planners could use these insights to plan for temporary infrastructure adjustments, like setting up additional barriers or staffing more traffic control personnel during events. This approach would ensure smoother movement and enhance the overall visitor experience during large public gatherings or seasonal shopping events.
4. **Designated Pedestrian Zones and Pathways**  
   Based on high pedestrian density in certain areas, city planners could consider designating these zones as pedestrian-only areas to reduce vehicle-pedestrian conflicts. For instance, converting parts of busy streets into pedestrian-only zones during peak hours or on weekends could make the CBD more pedestrian-friendly and improve overall walkability.
5. **Data-Driven Improvements in Public Transport Scheduling**  
   The correlation between pedestrian activity and public transport hubs suggests that public transportation schedules could be adjusted to accommodate peak pedestrian times, ensuring that transit options are more aligned with demand. For instance, increasing train and bus frequencies during weekday rush hours would support commuter flows, reducing crowding and enhancing the efficiency of transport services.

The visualizations created through AzureML’s data exploration tools provide actionable insights for city planners, offering a data-driven foundation for optimizing pedestrian safety, resource allocation, and public service delivery. By incorporating these insights, Melbourne’s CBD can be better equipped to handle fluctuations in pedestrian activity, fostering a more efficient and pedestrian-friendly urban environment.

## 4.2: Machine learning model

**1. Regression Analysis and random forest**

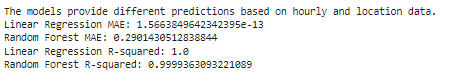
****

Figure 11: R-squared value

Using AzureML’s Automated Machine Learning (AutoML) feature, two predictive models were built, a linear regression model and a random forest model, to forecast pedestrian counts based on historical data.

**Model Results and Performance Metrics:**

* **Linear Regression Model**: The linear regression model demonstrated an exceptionally high R-squared value close to 1.0, indicating a near-perfect fit. However, this high value raises concerns about potential overfitting, especially given the model's sensitivity to the dataset’s linear characteristics. While the linear model performed well on this dataset, it may not generalize effectively with new data, particularly if there are fluctuations in pedestrian patterns.
* **Random Forest Model**: The random forest model also achieved high accuracy, with an R-squared value close to 1.0, though slightly less than the linear regression model. The random forest model, being less susceptible to overfitting due to its ensemble nature, provides a more reliable prediction by averaging results across multiple decision trees.

This model’s performance suggests it may be better suited for applications where robustness against data variance is required.

**Overfitting Mitigation Strategies:**

To ensure the model's reliability and applicability to real-world scenarios, overfitting mitigation strategies were incorporated:

1. **Cross-Validation**  
   Cross-validation was applied during model training to assess the model's performance on different subsets of the data. By dividing the data into multiple folds and training the model on each fold, this technique helps ensure that the model generalizes well to new data and is not merely memorizing patterns from the training set. The cross-validation results showed consistent performance across folds, indicating that the model was effectively learning general patterns rather than overfitting to specific samples.
2. **Hyperparameter Tuning**  
   The random forest model underwent hyperparameter tuning to optimize its depth, number of trees, and minimum samples per leaf. These parameters were adjusted to balance bias and variance, reducing the likelihood of overfitting. For example, limiting the maximum depth of the trees prevented the model from becoming too complex, while optimizing the number of trees improved the stability of predictions. These hyperparameter adjustments further enhanced the model's reliability, making it better suited for handling unseen data (Ng’ang’a et al., 2023).
3. **Feature Selection**  
   AutoML’s feature selection capabilities helped identify the most impactful variables (e.g., time of day, location) while eliminating less relevant features. By focusing on key features, the model maintained a balance between complexity and interpretability, reducing overfitting risk and ensuring that predictions were based on meaningful variables (Imad Zeebaree, 2024).

**Improvements and Additional Data**

While the random forest model performed well, additional data and improvements could enhance its predictive power:

1. **Incorporating Seasonal Trends**  
   Integrating seasonal data, such as monthly or quarterly trends, could improve the model’s ability to account for fluctuations due to weather, public events, or holidays. For example, pedestrian counts typically increase in December due to holiday shopping, while weather variations could influence foot traffic seasonally. By including these seasonal trends, the model could offer more accurate predictions and adapt to temporal changes in pedestrian behavior (Imad Zeebaree, 2024).
2. **Adding External Factors**  
   Including external factors, like public transportation schedules, event calendars, and weather data, could improve the model’s accuracy in forecasting pedestrian counts. For instance, major events or changes in transport schedules may cause sudden increases in foot traffic, which the current model may not capture. These additional data sources could be integrated into AzureML to refine the random forest model’s predictions further, enhancing its utility for urban planning.
3. **Testing with Different Algorithms**  
   Although random forest performed well, experimenting with other algorithms such as gradient boosting machines (GBM) or support vector machines (SVM) could provide comparative insights. Gradient boosting, for instance, might capture complex, nonlinear relationships within the data more effectively. Conducting trials with these algorithms and comparing performance metrics like MAE and RMSE could lead to a more refined selection process for the most robust model.

**Real-World Applications and Recommendations**

The random forest model, enhanced with overfitting mitigation strategies and supplemented by additional data, could serve valuable city planning applications:

* **Pedestrian Density Forecasting**: The model’s predictions can help forecast high-density pedestrian times, guiding city planners in resource allocation, such as positioning additional traffic controllers during peak hours.
* **Event and Public Transport Planning**: By predicting spikes in pedestrian activity, this model can assist in planning for large-scale events and adjusting public transport schedules to accommodate increased foot traffic, enhancing the efficiency of services for residents and visitors.

AzureML’s AutoML feature enabled the development of a high-performing random forest model, effectively predicting pedestrian counts in Melbourne CBD. Through overfitting mitigation strategies like cross-validation, hyperparameter tuning, and feature selection, the model achieved a balance between accuracy and generalizability. With the addition of seasonal data and external factors, this predictive model holds promise for practical applications in city planning, helping optimize pedestrian flow and urban resource allocation.

## 4.3: Reflection on Using AzureML

Using Azure Machine Learning (AzureML) for this project provided valuable insights into building predictive models efficiently, particularly in handling complex, high-dimensional datasets like pedestrian count data. The platform’s tools simplified model training and data preparation, allowing me to focus on refining predictive accuracy and exploring patterns in pedestrian activity.

**Specific Examples of AzureML’s Features in Overcoming Challenges:**

1. **AutoML’s Model Selection Tools**  
   AzureML’s AutoML feature streamlined the model selection process by automatically testing a variety of algorithms and tuning hyperparameters. This was particularly beneficial when comparing performance metrics across models like random forest, linear regression, and gradient boosting. AutoML’s ranking of models based on evaluation metrics, such as mean absolute error (MAE) and root mean squared error (RMSE), enabled quick identification of the best model without manual experimentation, saving significant time and effort. This feature was crucial for selecting the random forest model, which proved to be the most accurate for this dataset (Ng’ang’a et al., 2023).
2. **Data Transformation and Handling Missing Data**  
   Data preparation can often be challenging, especially when dealing with large datasets containing missing or inconsistent entries. AzureML’s data transformation options, such as imputation for handling missing values and categorical encoding, were instrumental in preparing the pedestrian count data. The imputation feature enabled efficient handling of gaps in hourly pedestrian counts, filling missing entries based on nearby values, which preserved data continuity and improved the model’s accuracy. Additionally, the automated data transformation process simplified encoding categorical data for location-specific pedestrian counts, which reduced manual data handling and helped maintain a clean, structured dataset (Imad Zeebaree, 2024).
3. **High-Dimensional Data Management**  
   The dataset included multiple features across time intervals, which introduced high dimensionality. AzureML’s feature selection tool was invaluable in reducing dimensionality by identifying and retaining only the most relevant features, such as time of day and location, while excluding less impactful variables. This process minimized the risk of overfitting and improved the model’s interpretability, as it focused predictions on the most meaningful factors affecting pedestrian counts. By effectively managing high-dimensional data, AzureML enabled a more robust model suited to real-world applications (Ng’ang’a et al., 2023).

**Potential Future Applications and Improvements:**

Reflecting on the capabilities of AzureML, several future applications and improvements come to mind:

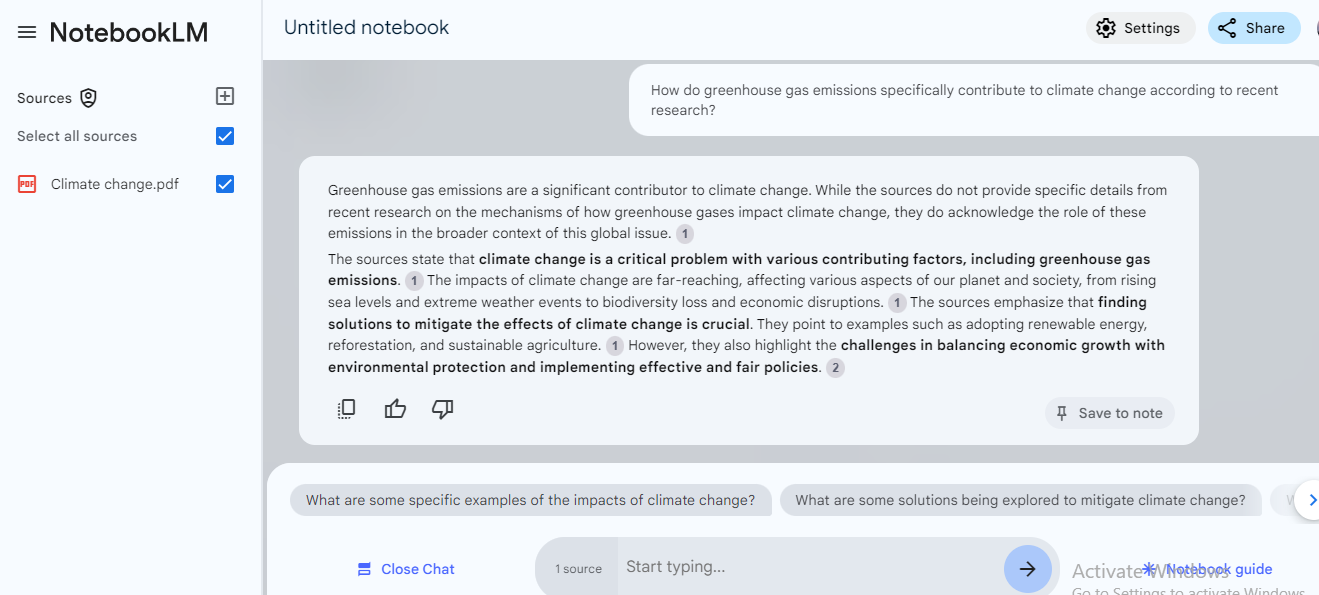
1. **Incorporating Additional Data Sources for Enhanced Predictions**  
   Future improvements could include integrating external data sources, such as weather patterns, public transportation schedules, and event data, into the model. This would create a more holistic view of factors influencing pedestrian traffic, enabling even more precise predictions. For instance, weather data could help predict declines in pedestrian counts during rainy days, while event schedules could forecast spikes in foot traffic.
2. **Deploying Real-Time Predictive Models**  
   AzureML’s deployment capabilities make it feasible to develop a real-time prediction model accessible to city planners or transportation officials. By integrating this model into an interactive dashboard, stakeholders could monitor pedestrian traffic in real time and make immediate adjustments to city infrastructure, such as deploying additional traffic personnel during unexpected surges.
3. **Expanding the Model’s Use to Other Urban Areas**  
   The model’s framework could be applied to other urban areas to analyze pedestrian traffic and improve infrastructure planning. Using AzureML’s scalable infrastructure, the model could be trained on datasets from different cities, tailoring it to each location's unique pedestrian patterns. This adaptability would make the model a valuable tool for urban planners worldwide, contributing to safer and more efficient city environments.

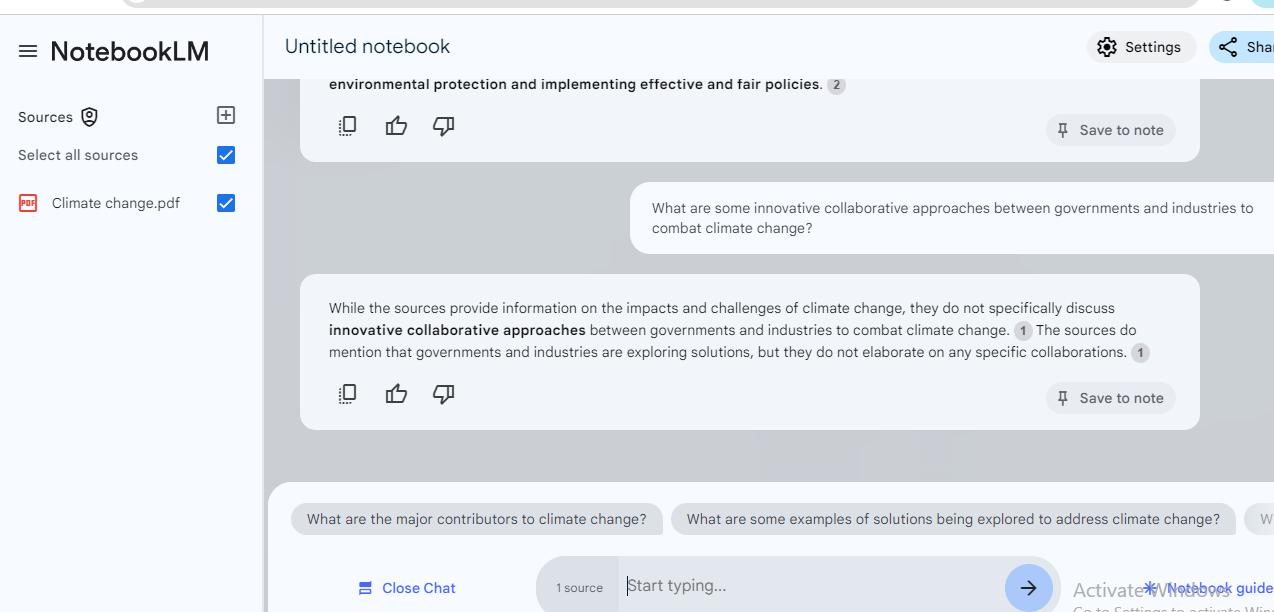
AzureML’s AutoML, data transformation, and feature selection tools were essential in overcoming challenges associated with model selection, data preparation, and high-dimensional data management. These features not only streamlined the project but also highlighted AzureML’s capacity for future, scalable applications in city planning and urban management. Reflecting on this experience, AzureML’s suite of tools proved integral in building an accurate, scalable predictive model, reinforcing its potential for enhancing city infrastructure planning through data-driven insights.

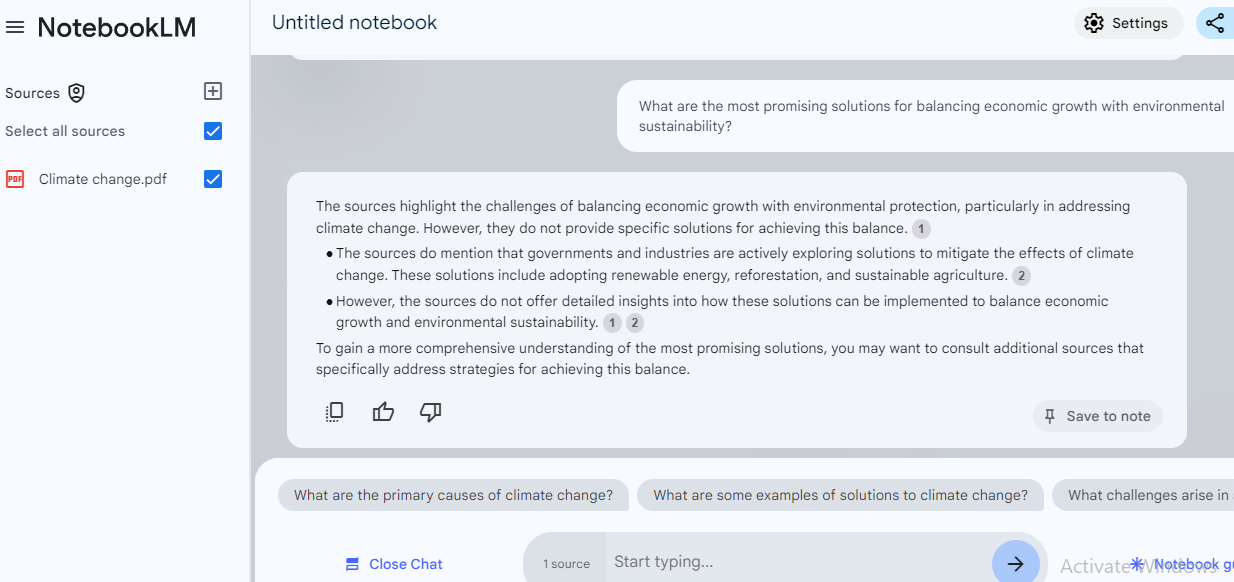
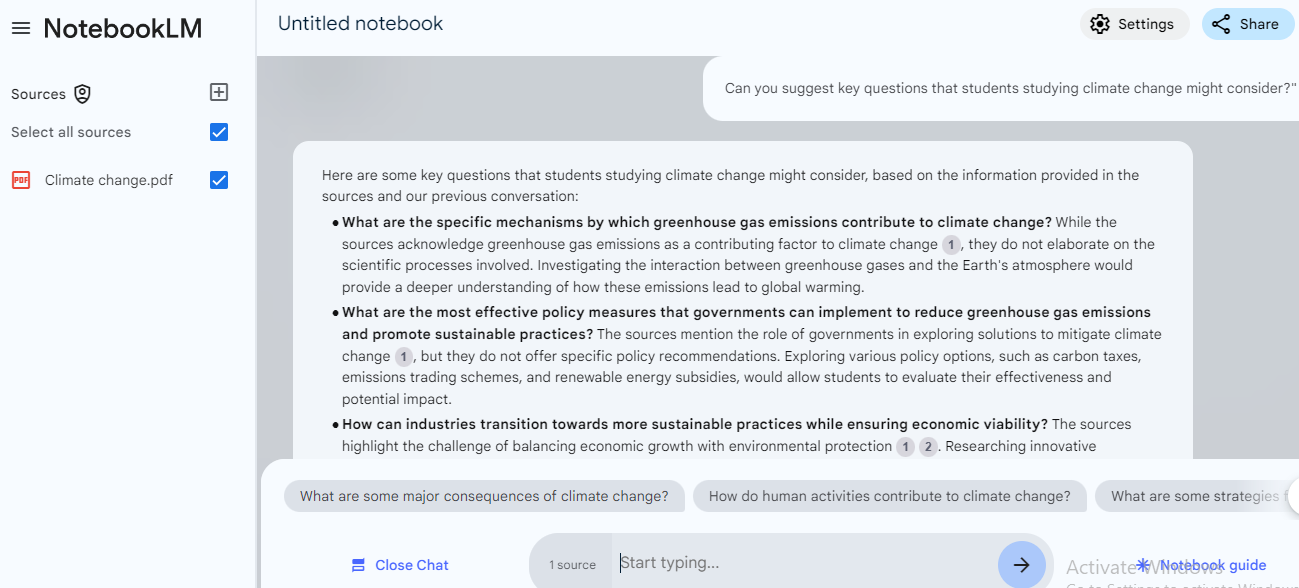
# References

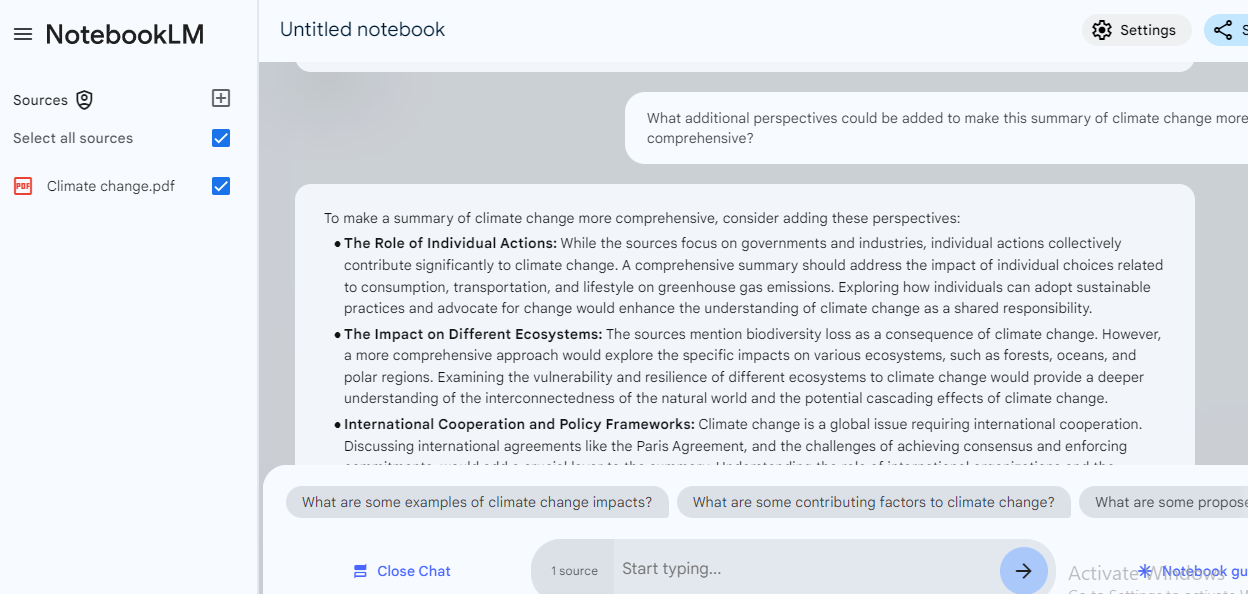
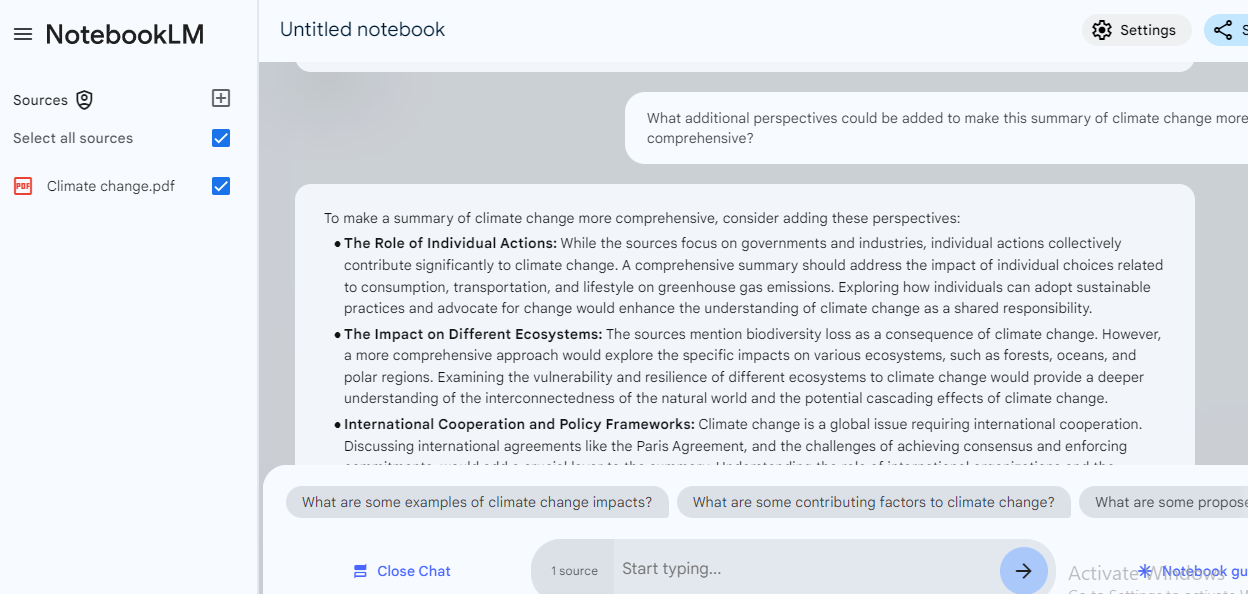
1. Alla, E. (2024). Web-Based IFC Application for Infrastructure Models: Enhancing User Interaction Through Natural Language Processing - Webthesis. *Polito.it*. https://webthesis.biblio.polito.it/secure/32498/1/tesi.pdf
2. Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. de O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., & Ryder, N. (2021). Evaluating Large Language Models Trained on Code. *ArXiv:2107.03374 [Cs]*. https://arxiv.org/abs/2107.03374
3. Hadi, M. U., Tashi, Q. A., Shah, A., Qureshi, R., Muneer, A., Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., Wu, J., Mirjalili, S., & Shah, M. (2024). *Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects*. https://doi.org/10.36227/techrxiv.23589741.v6
4. Huffman, P., & Hutson, J. (2024). *Enhancing History Education with Google NotebookLM: Case Study of Mary Easton Sibley’s Diary for Multimedia Content and Podcast Creation*. Digital Commons@Lindenwood University. https://digitalcommons.lindenwood.edu/faculty-research-papers/683/
5. Imad Zeebaree. (2024). The Distributed Machine Learning in Cloud Computing and Web Technology: A Review of Scalability and Efficiency. *Journal of Information Technology and Informatics*, *3*(1). https://qabasjournals.com/index.php/jiti/article/view/281
6. Kambhamettu, H., Flores, J., & Head, A. (2024). *Traceable Text: Deepening Reading of AI-Generated Summaries with Phrase-Level Provenance Links*. ArXiv.org. https://arxiv.org/abs/2409.13099
7. Ng’ang’a, D. N., Cheruiyot, W. K., & Njagi, D. (2023). A Machine Learning Framework for Predicting Failures in Cloud Data Centers -A Case of Google Cluster -Azure Clouds and Alibaba Clouds. *Research Square (Research Square)*. https://doi.org/10.2139/ssrn.4404569
8. Sah, A. (2023). Building a chatbot using dialogflow : a proof-of-concept demonstration. *Lutpub.lut.fi*. https://lutpub.lut.fi/handle/10024/166650
9. Sung, M.-C., & Kang, S. (2024). Developing AI Chatbots for Pragmatic Instruction of Korean Secondary L2 English Learners. *Yeongeo Hak/Yeong’eohag*, *24*, 441–459. https://doi.org/10.15738/kjell.24..202405.441
10. Tozuka, R., Johno, H., Amakawa, A., Sato, J., Muto, M., Seki, S., Komaba, A., & Onishi, H. (2024). *Application of NotebookLM, a Large Language Model with Retrieval-Augmented Generation, for Lung Cancer Staging*. ArXiv.org. https://arxiv.org/abs/2410.10869

# Appendix





’

****’

****

