1. Outline:

The primary objective of this document is to delineate the application of Chat GPT 3.5 Turbo in the analysis of local text data, specifically Privacy Policy Documents. The advanced capabilities of this AI model are leveraged to scrutinize the text data meticulously and extract key-value pairs. These pairs are then utilized to generate a comprehensive and succinct response that encapsulates the essential information contained within the documents. The ultimate goal is to simplify the understanding of these complex documents and provide a clear, concise summary of their content.

This document will show how GPT 3.5 Turbo LLM was used to analyze an organized and re-structured version of the Privacy Policy of Booking.com by Jiang Yan Bing locally.

2. Overview:

The directory structure of our application will be organized and simplified for easy navigation and maintenance, featuring only Python (.py) and Text (.txt) files. The Python files will hold the source code of the application, including but not limited to 'main.py', 'MyAPIKey.py', 'LLMquery.py', and 'GUI.py'. Each Python file will be responsible for a particular set of functionalities, enhancing modularity and readability.

The Text (.txt) files present within our directory structure will serve a crucial role in harboring privacy policy data. Files such as 'privacy_policy.txt' will encapsulate vital privacy policy
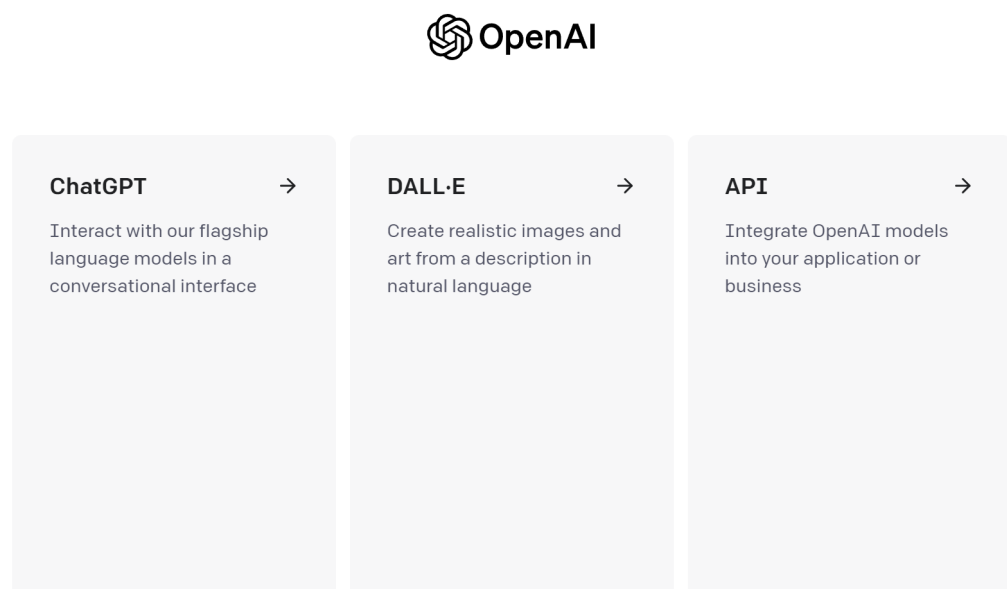
information pertaining to various companies. This data will be thoroughly analyzed by a 3rd party large language model from OpenAI, which can identify a company's compliance with each of the six legal grounds as defined under the General Data Protection Regulation (GDPR).

The six legal grounds include Consent, Contract, Legal Obligation, Vital Interests, Public Task, and Legitimate Interests. A large language model will systematically process this information to evaluate the extent of each company's adherence to these crucial GDPR grounds.

As a result, we will be able to infer the specific user rights that each company satisfies. This intelligent, automated scrutiny of privacy policy data stands at the core of our commitment to bolster transparency and safeguard user rights in the increasingly complex digital landscape.

3. Create an API Key with OpenAI:

1. Navigate to the OpenAI website and LogIn. Then you should have the following three options available:

ChatGPT →	**DALL·E** →	**API** →
Interact with our flagship language models in a conversational interface	Create realistic images and art from a description in natural language	Integrate OpenAI models into your application or business

Click on the very right column labelled "API"

2. Then navigate to the corner on the top right and click on your name, you should then have the option to "View API keys".
3. Then create a new API key by clicking on "+ Create new secret key".
4. Ensure your key is copied and stored in a safe location before clicking "Done"

4. Implementation in Python:

1. Libraries to install:

pip install the following:

1. langchain
2. openai
3. chromadb

4. tiktoken
        5. unstructured

2. Edit MyAPIKey.py
    1. Paste your OpenAI API Key.

3. How it works:
    1. Run "main.py"
    2. When you execute "main.py", it will import two Python files, namely "LLMquery.py" and "GUI.py". Essentially, the main program, "main.py", is structured to first run "LLMquery.py". This script communicates with a third-party large language model, OpenAI's GPT-3.5-Turbo, by posing a series of queries. The model then formulates responses, where only the final response is stored in a text file for further processing. After the completion of this step, "main.py" calls on the function "Analysis". This function examines the generated response from the language model to determine if any of the six specified legal grounds are met within a company's privacy policy based on the analysis of the language model. If a legal ground is satisfied, the "Analysis" function will extract the language model's summary response related to that legal ground. If none of the legal grounds are met, the output will be "n/a". Additionally, the function is capable of extrapolating from the satisfied legal grounds to deduce the available user rights. Finally, the "main.py" program utilizes "GUI.py" to present the analyzed data to the user through a graphical user interface.

Disclaimer* The code currently only runs properly if all the values in the privacy policy text file are separated by an underscore instead of a space. (e.g. "legal obligation" will instead be "legal_obligation")

5. GPT-3.5-Turbo Generated Best Case Summary Response for Booking.com:
(**Note*** the text below contains the raw responses to the queries, nothing is modified, nor changed)

Query 1: Generate a commercial description for each of the six legal grounds under GDPR laws and regulations (Consent, Contractual Necessity, Legal Obligation, Vital Interests, Public Task, Legitimate Interests) that are satisfied by " + str(CompanyName) + " ."

Query 2: 'Clean up the response by making it more concise.'

Query 3: 'Clean up the response by only including the six legal grounds in numbered format. If a legal ground is satisfied output the legal ground and its summary, if not output the legal ground and state not satisfied.'

Final Response:
Based on the provided context, Booking.com satisfies three out of the six legal grounds under GDPR laws and regulations. Here is a summary of each legal ground:

1. Consent: Booking.com may send personalized offers or marketing with the user's consent or as otherwise permitted by applicable law. This means that Booking.com will only use personal data for marketing purposes if the user has given explicit consent.

2. Legitimate interests: Booking.com relies on its legitimate interests to share and receive personal data for purposes A to F. This means that Booking.com can use personal data for purposes that are necessary for their legitimate interests, as long as it does not outweigh the rights and freedoms of the individual.

3. Legal obligation: Booking.com may further disclose personal data to competent authorities to comply with a legal obligation, such as under short-term rental laws, to protect and defend their rights or properties, or the rights and properties of their business partners. This means that Booking.com may be required to share personal data with authorities or take necessary actions to comply with legal obligations.

Based on the provided context, it is unclear whether Booking.com satisfies the remaining three legal grounds under GDPR laws and regulations.

6. GUI Examples:

Booking.com Example 1:

Booking.com Example 2:

Worst Case Scenario Example:

7. Challenges with LLMs:

1. **Inconsistency in Responses:**
   Since we are leveraging a highly generalized large language model from OpenAI instead of a model explicitly trained to parse and analyze privacy policy texts for identifying satisfied legal grounds. While the general-purpose model, such as GPT-3.5-Turbo, is an advanced tool capable of generating detailed responses to a wide array of queries, it may lack the specificity required for this context. The potential drawback of using this approach is that the results of each program

execution may vary in accuracy and consistency, given that the model is not specialized in legal text analysis. The inherent generality of the model could thus introduce a degree of unpredictability and inconsistency into the results produced by the system.

2. **Hallucinations by the LLM:**

Given that we are employing artificial intelligence to parse and analyze the text, it's important to acknowledge that the AI, while sophisticated, does not possess a human-like understanding of the text's logic or implications. Instead, the AI utilizes weighted values to infer the correlation between specific sections of the text and distinct legal grounds. However, this approach can occasionally be confounded by an inherent challenge in large language models known as "hallucinations". This phenomenon can cause the AI to erroneously conclude that a particular legal ground is satisfied when, in fact, it is not. For example, as demonstrated in the GUI Examples section above, during the first execution with Booking.com's privacy policy, the LLM incorrectly identified Contractual Necessity as a satisfied legal ground. In reality, the applicable ground should have been identified as Legal Obligation. In the second execution, the LLM was able to self-correct this error, thus highlighting the potential for fluctuating results within the system.

8. OpenAI API Costs:

A direct quote from the pricing team at OpenAI states: "Multiple models, each with different capabilities and price points. Prices are per 1,000 tokens. You can think of tokens as pieces of words, where 1,000 tokens is about 750 words. This paragraph is 35 tokens."

The current model we are using is GPT 3.5 Turbo which has the following costs:

Model	Input	Output
4K context	$0.0015 / 1K tokens	$0.002 / 1K tokens
16K context	$0.003 / 1K tokens	$0.004 / 1K tokens

Input:
Drawing on the provided queries, the queries encompass 79 words. Extrapolating from OpenAI's guidelines, this implies that the response will comprise approximately 105 tokens. Given that we are leveraging the 4K context model, which will operate at a cost of $0.0015(USD) per 1,000 tokens, it is estimated that the **total queries will be billed at approximately $0.0001575(USD)**.

Output:
Drawing on the provided Best Case Summary Response, the generated paragraph encompasses a total of 209 words. Extrapolating from OpenAI's guidelines, this implies that the response will comprise approximately 279 tokens. Given that we are leveraging the 4K context model, which operates at a cost of $0.002(USD) per 1,000 tokens, it is estimated that **each response will be billed at approximately $0.000558(USD)**. This streamlined and cost-effective operation allows us to deliver robust AI solutions while maintaining fiscal responsibility.

Bringing it all together:
By simply summing up the input and output costs, the **total cost for each question and answer chat will be billed at approximately $0.0007163(USD)**.

9. Questions yet to be answered:
1. Are queries processed by OpenAI at a remote server or is it done locally?

10. Useful Links for Reference:
1. https://github.com/hwchase17/langchain (Github Repo of LangChain)
2. https://openai.com/pricing (OpenAI Pricing Documentaiton)
3. https://openai.com/ (Official OpenAI Website)