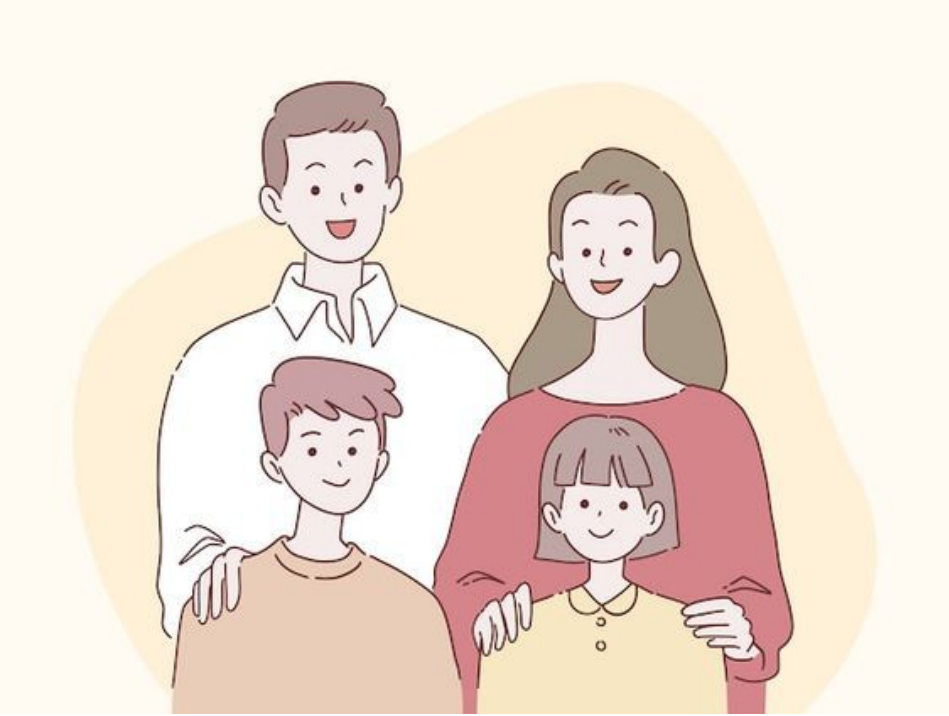
PSFD 家庭動態調查
Panel Study of Family Dynamics

Data Integration of PFSD

NTU C2L2 LAB



Agenda

1. **Main Issues & Research Objective**
2. Process and Methods
3. Model Outcome & Comparison
4. Demo
5. Expected solution
6. Future Prospects

Main Issues: Wording of survey questions has changed over different years

Similar meanings but different wording

For example, data from 2008 and 2004 may refer to the same situation, but the wording of the questions differs.

CV2008	d01z01	請問您目前結婚了嗎?	2	CIII2004	D01A	請問您目前的婚姻狀況是怎樣的?
CV2008	d01z02	請問您結婚年次	3	CIII2004	D01B	請問您是在哪一年結婚的?民國__年
CV2008	d02	請問(他/她)是在哪一年出生的?	4	CIII2004	D02	請問從我們91年一月訪問您以後，您個人的婚姻狀況有
CV2008	d03	請問(他/她)是原住民，台灣閩南人，台灣客家人還是外省人(大陸各省市)?	5	CIII2004	D03	請問您的(先生/太太)是在哪一年出生的?民國__年
CV2008	d04a	請問(他/她)的最高教育程度是什麼?	6	CIII2004	D04	請問(他/她)是哪裡人?
CV2008	d04b	請問(他/她)的父親最高教育程度是什麼?	7	CIII2004	D05A	請問您(先生/太太)的最高教育程度是什麼?
CV2008	d05	請問您配偶(同居人)目前的健康狀況如何?	8	CIII2004	D05B	請問您(先生/太太)的父親最高教育程度是什麼?
CV2008	d06a	請問(他/她)目前有工作嗎?	9	CIII2004	D06	請問您(先生/太太)目前的健康狀況如何?
CV2008	d07a01	請問從去年(民國96)一月份以來，(他/她)更換過主要工作嗎?	10	CIII2004	D07	請問您(先生/太太)目前有工作嗎?
CV2008	d07a02	若更過工作，換過__次工作	11	CIII2004	D08A	請問從我們91年一月訪問到這次訪問的期間，您(先生/

Sub-questions do not extend the main question

The sub-questions are essentially a repetition of the main question. If these sub-questions are classified alone, it may cause difficulty in categorization.

CV2008	1	1-2	0	請問您(和您配偶)過去一年裡，教育費用 __元
CII2002	11	11	1	教育費用____元
CII2002	11	11	1	紅白帖費用____元；
CII2002	11	11	1	醫療費用____元
CII2002	11	11	1	衣著費用____元

PSFD 家庭動態調查 Panel Study of Family Dynamics

Main Issues: Same question has different answer options in different years

- Same answer in different number
- Different answers need to be merged

CV2008	d01z01f	0 = "00 跳答，不適用"
CV2008	d01z01f	1 = "01 未婚"
CV2008	d01z01f	2 = "02 同居"
CV2008	d01z01f	3 = "03 已婚(第一次結婚)"
CV2008	d01z01f	4 = "04 離婚再婚"
CV2008	d01z01f	5 = "05 喪偶再婚"
CV2008	d01z01f	6 = "06 分居"
CV2008	d01z01f	7 = "07 離婚"
CV2008	d01z01f	8 = "08 喪偶"
CV2008	d01z01f	96 = "96 不知道，不清楚
CV2008	d01z01f	97 = "97 其他"
CV2008	d01z01f	98 = "98 拒答"
CV2008	d01z01f	99 = "99 缺漏值"；
CV2008	d03f	0 = "0 跳答，不適用"
CV2008	d03f	1 = "1 原住民"
CV2008	d03f	2 = "2 台灣閩南人"
CV2008	d03f	3 = "3 台灣客家人"
CV2008	d03f	4 = "4 外省人"
CV2008	d03f	6 = "6 不知道，不清楚，不
CV2008	d03f	7 = "7 其他"
CV2008	d03f	8 = "8 拒答"
CV2008	d03f	9 = "9 缺漏值"；

CIII2004	D01A	0 = '0 跳答或不適用'
CIII2004	D01A	1 = '1 未婚'
CIII2004	D01A	2 = '2 同居'
CIII2004	D01A	3 = '3 已婚'
CIII2004	D01A	4 = '4 分居'
CIII2004	D01A	5 = '5 離婚'
CIII2004	D01A	6 = '6 喪偶'
CIII2004	D01A	96 = '96 不知道'
CIII2004	D01A	97 = '97 其他'
CIII2004	D01A	98 = '98 拒答'
CIII2004	D01A	99 = '99 缺漏值'；
CIII2004	D02F	0 = '0 跳答或不適用'
CIII2004	D02F	1 = '1 沒有變化;仍是已婚(含同居
CIII2004	D02F	2 = '2 沒有變化;仍是單身'
CIII2004	D02F	3 = '3 有變化;在最近兩年內結婚
CIII2004	D02F	4 = '4 有變化;在最近兩年內分居
CIII2004	D02F	5 = '5 有變化;在最近兩年內離婚
CIII2004	D02F	6 = '6 有變化;在最近兩年內喪偶
CIII2004	D02F	96 = '96 不知道'
CIII2004	D02F	97 = '97 其他'
CIII2004	D02F	98 = '98 拒答'
CIII2004	D02F	99 = '99 缺漏值'；

PSFD 家庭動態調查 Panel Study of Family Dynamics

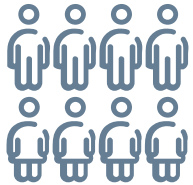
Expected Result: from SRDA

☐	ed020020	您的教育程度是	1. 無 2. 自修 3. 私塾 ...	教育	1990/21 1991/22	
☐	ed020021	您的教育程度是	1. 無 2. 自修 3. 小學肄業 ...	教育	1992/23 1993/24 1994/25 ...	
☐	ed020034	您的教育程度是	1. 無 2. 自修 3. 小學 ...	教育	2004/45	
☐	ed020035	您的教育程度是	1. 無(不識字) 2. 自修(識字/私塾) 3. 小學 ...	教育 ISSP	2002/43 2003/44 2004/45 ...	

Challenges with the Original Approach:

Costly

Prone to error



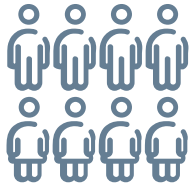
Manual process

Waste of time

Challenges with the Original Approach:

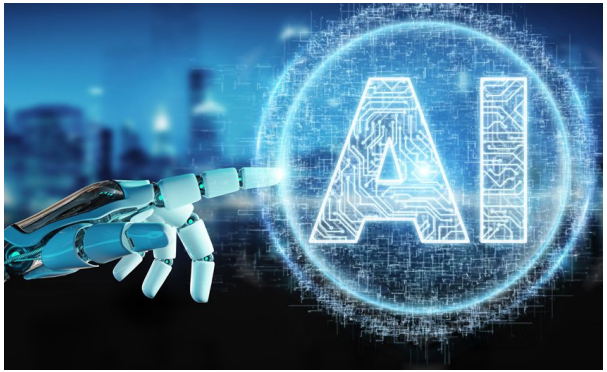
Costly

Prone to error



Manual process

Waste of time



Embracing AI for clustering

How to measure AI's performance in clustering

Type I error	Incorrectly grouping items that do not belong together.
Type II error	Failing to group items that actually belong together.

- Type I error is more significant in this context, as incorrectly grouping unrelated items requires employees to manually review and reorganize the AI-generated clusters. In contrast, Type II error can be addressed by manually clustering items that were left ungrouped by the AI.
- Therefore, our analysis will focus on **minimizing Type I error** in the model comparison phase to improve clustering accuracy and reduce manual intervention.

Panel Study of Family Dynamics Data Integration Outline

Introduction		Organizing PSDA data manually is time-consuming and may lead to some human errors
Goals		To automatically and precisely cluster related questions and organize answers using ML methods.
Process and Methods	**Data Set**	Type 7 questions from the years 2002, 2004, 2008, 2014, and 2018 PSFD Questions.
	ML Methods Choosing	1. Tokenize → Embedding → Clustering(K-means) → Similarilarity comparison 2. Directly call GPT4 API to do cluster 3. Label the questions of median year → call GPT4 to cluster → Label the questions of median year of unclustered data → call GPT4 and repeat
Results	**Questions**	Achieved a Type I error rate as low as **1.28%** and spent less than 1 NTD handling 123 data.
	Answers	Organize related answers using "union”
Limitation		Using only 5 years of data for clustering may still require manual organization for the remaining data and manual selection of topics.

Research Objective

Main Issue

- **Manually**
- **Costly**
- **Prone to Error**
- **Waste of Time**

Our solution

- Use **machine learning methods** to identify the corresponding question numbers for identical questions across different years.

Research Objective

Main Issue

- **Manually**
- **Costly**
- **Prone to Error**
- **Waste of Time**

Our solution

- Use **machine learning methods** to identify the corresponding question numbers for identical questions across different years.
- **Automatically** consolidate **option codes** and present them along with the organized questions in the "Variable Comparison Table."

Research Objective

Main Issue

- **Manually**
- **Costly**
- **Prone to Error**
- **Waste of Time**

Our solution

- Use **machine learning methods** to identify the corresponding question numbers for identical questions across different years.
- **Automatically** consolidate **option codes** and present them along with the organized questions in the "Variable Comparison Table."
- Enable the machine to directly **output the survey and questions of interest to the user**, along with the survey results, based on the "Variable Comparison Table."

Agenda

1. Main Issues & Research Objective
2. **Process and Methods**
3. Model Outcome & Comparison
4. Demo
5. Expected solution
6. Future Prospects

Automation Process

Data Preprocessing

- For Demo, select 2002, 2004, 2008, 2014, and 2018 PSFD
- Transform sas file to csv file, since the website output file is sas file

Survey Questions Clustering

1. Tokenize → Embedding → Clustering(K-means) → Similarilarity comparison
2. Directly call GPT4 API to do cluster
3. Label the questions of median year → call generative model API to cluster and give similarity→ Label the questions of median year of unclustered data → call generative model API and repeat

Answer Organizing

- Organize related answers in different years using "union”

Output Survey outcome

- Merge the individual survey answer to corresponding questions data

Automation Process

Data Preprocessing

- For Demo, select 2002, 2004, 2008, 2014, and 2018 PSFD
- Transform sas file to csv file, since the website output file is sas file

Survey Questions Clustering

1. Tokenize → Embedding → Clustering(K-means) → Similarilarity comparison
2. Directly call GPT4 API to do cluster
3. Label the questions of median year → call generative model API to cluster and give similarity→ Label the questions of median year of unclustered data → call generative model API and repeat

Answer Organizing

- Organize related answers in different years using "union”

Output Survey outcome

- Merge the individual survey answer to corresponding questions data

Automation Process: Data Preprocessing

Transform these three types of files into **CSV files** to make data analysis easier:

Label	Survey Question
Answer	Survey answer comparison table
Survey	Survey outcome

```python
def convert_sas_to_csv(sas_file, csv_file):
    with open(sas_file, 'r', encoding='BIG5', errors='ignore') as f:
        lines = f.readlines()

    csv_numbers = []
    csv_questions = {}
    csv_answers = {}
    for i in range(len(lines)):
        if lines[i].startswith('LABEL'):
            for j in range(i+1, len(lines)):
                lines[j] = lines[j].replace(",", "'")
                try:
                    number = lines[j].strip().replace('"', '').split('=')[0].strip()
                    question = lines[j].strip().replace('"', '').split('=')[1].strip()
                    csv_questions[number] = question
                except:
                    break
        if lines[i].startswith('   FORMAT'):
            for j in range(i+1, len(lines)):
                if lines[j].startswith('   '):
                    answers = lines[j].split(".")
                    answers.pop()
                    for a in answers:
                        csv_numbers.append(a.split()[0].strip())
                        csv_answers[a.split()[0].strip()] = a.split()[1].strip()
                else:
                    break

    dir_name = os.path.basename(os.path.dirname(sas_file))
    csv_file = os.path.join(os.path.dirname(sas_file), dir_name + '_label.csv')
```

Automation Process: Data Preprocessing

Manually select specific type of topic

```
#手動的similarity_list
similarity_list = [[['CVIII2014', 'x'], ['CIII2004', 'X'], ['CV2008', 'x'], ['CII2002', 'x'], ['CX2018', 'x']],
                   [['CVIII2014', 'a'], ['CIII2004', 'A'], ['CV2008', 'a'], ['CII2002', 'a'], ['CX2018', 'a']],
                   [['CVIII2014', 'b'], ['CIII2004', 'B'], ['CV2008', 'b'], ['CII2002', 'b'], ['CX2018', 'b']],
                   [['CVIII2014', 'c'], ['CIII2004', 'C'], ['CV2008', 'c'], ['CII2002', 'c'], ['CX2018', 'c']],
                   [['CVIII2014', 'd'], ['CIII2004', 'D'], ['CV2008', 'd'], ['CII2002', 'd'], ['CX2018', 'd']],
                   [['CVIII2014', 'e'],['CIII2004', 'F'],['CV2008', 'e'], ['CII2002', 'f'],  ['CX2018', 'e']],
                   [['CVIII2014', 'f'],['CIII2004', 'G'], ['CV2008', 'f'], ['CII2002', 'g'],['CX2018', 'f']],
                   [['CIII2004', 'E'], ['CII2002', 'e']],
                   [['CX2018', 'g']]]
```

type_7 is related to family status

Automation Process: Data Preprocessing

Data set description	
files	CVxxxx_label.csv, type_7.csv
Features	YEAR：which question appeared in the questionnaire NUMBER: Question number QUESTION: the content of the question ANSWER: corresponding answer code
Years	2002, 2004, 2008, 2014, 2018

Automation Process

Data Preprocessing

- Select 2002, 2004, 2008, 2014, and 2018 PSFD
- Transform sas file to csv file, sine the website output file is sas file

Survey Questions Clustering

1. Tokenize → Embedding → Clustering(K-means) → Similarilarity comparison
2. Directly call GPT4 API to do cluster
3. Label the questions of median year → call generative model API to cluster and give similarity→ Label the questions of median year of unclustered data → call generative model API and repeat

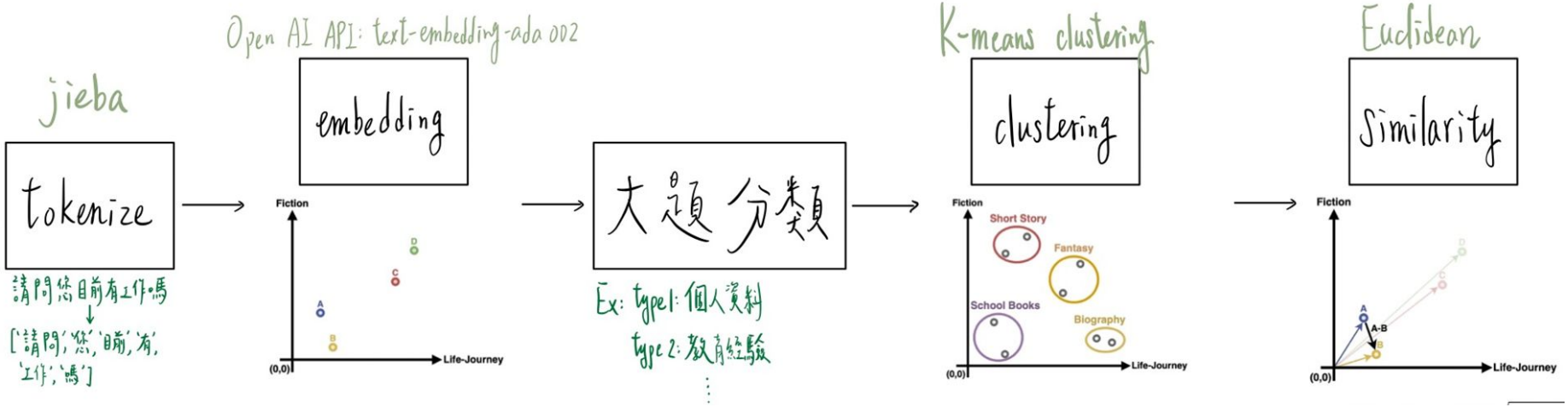
Answer Organizing

- Organize related answers in different years using "union”

Output Survey outcome

- Merge the individual survey answer to corresponding questions data

Clustering Methods 1: Statistical learning



Source: Introduction to Embedding, Clustering, and Similarity:
https://towardsdatascience.com/introduction-to-embedding-clustering-and-similarity-11dd80b00061

Clustering Methods 2: OpenAI

Why using generative modelc

Convenience	Open AI model can classify directly using natural language instructions without the need for **data preprocessing** or **converting text into embeddings**, simplifying the workflow.
Accuracy	OpenAI's advanced deep learning model, enhanced by **RAG** (Retrieval-Augmented Generation), leverages vast training and real-time data retrieval for more accurate, context-aware predictions.
Flexibility	Open AI model can adjust its classification strategies **based on the instructions and context provided,** making it highly adaptable to a variety of classification tasks and data types. This enhances the model’s usability and flexibility across different applications.

Clustering Methods 2: OpenAI

Prompting or Finetuning?

	Prompting	Finetuning
Pros	• No data to get started • **Smaller upfront cost** • No technical knowledge needed • **Adaptability to Sparse Data**	• Nearly unlimited data fits • Learn new information and can be **domain-specific** • Correct incorrect information
Cons	• Much less data fits • Forgets data • Hallucinations • RAG misses, or gets incorrect data	• **More high-quality data needed** • **Upfront compute cost** • Needs some technical knowledge, especially data

With a dataset with a small number of labelled observations, either zero-shot classification or traditional classification with embeddings return better results than a fine-tuned model.

Clustering Methods 3: Iterative Clustering using Open AI

Step 1	Label all total **n** data entries of the **median(year)** sequentially into clusters **1 to n**.
Step 2 (1st Cluster)	Use the **GPT-4o-mini API** to classify the data labeled in step 1, outputting **classification results** and **similarity scores**.
Step 3	Re-cluster data from the same year that have **lower similarity score** into cluster 0.
Step 4 (2nd Cluster)	Label data in cluster 0 where the year is the **minimum of years greater than the median**, and **repeat steps 2 and 3** until the result is almost satisfied.

Clustering Methods 3: Iterative Clustering using Open AI

Step 1

ANSWER	NUMBER	QUESTION	YEAR	cluster	s
labb	f01a	在去年(96年)之中，請問您平均每週大約花多少時間作家務工作?__小時	CV2008	1	
labb	f01b	請問您的配偶平均每週大約花多少時間作家務工作?__ 小時	CV2008	2	
f02f	f02	在去年(96年)之中，您家庭自政府得到的補助總計大約是多少?__ 元	CV2008	3	
f03a	f03a	在過去一年裡，您(和您配偶)每個月的平均支出房屋貸款支出平均每月 __元?	CV2008	4	
f03b01f	f03b01	標會支出活會平均每月 __元	CV2008	5	
f03b02f	f03b02	死會平均每月 __元	CV2008	6	
f03c	f03c	保姆或幫傭(包括家事管理)支出平均每月 __元	CV2008	7	
f04a	f04a	請問您(和您配偶)過去一年裡，人壽或商業醫療保險 __元	CV2008	8	
f04b	f04b	請問您(和您配偶)過去一年裡，家俱與家庭耐久設備 __元	CV2008	9	
f04c	f04c	請問您(和您配偶)過去一年裡，教育費用 __元	CV2008	10	
f06z01f	f06z01	訪問結束時間:月	CV2008	11	
labb	f06z02	訪問結束時間:日	CV2008	12	
f06z03f	f06z03	訪問結束時間:時	CV2008	13	
f06z04f	f06z04	訪問結束時間:分 ;	CV2008	14	

Clustering Methods 3: Iterative Clustering using Open AI

Step 2

```
for index in none_indices:
    # 準備發送到 GPT-4o-mini 的提示內容
    prompt = f"""
    Based on the previous clustering data, here is the reference:
    {reference_text}
    Now, based on the following data, please:
    1. Assign a cluster number.
    2. Provide the similarity between 0 and 1.
    Format the response as: Cluster: X, Similarity: Y
    Data: {data.iloc[index]['QUESTION']}

    If the data does not fit into any existing clusters then format the response as: Cluster: 0, Similarity: 0
    """

    try:
        # 使用 GPT-4o-mini API 進行分類
        response = openai.chat.completions.create(
            model="gpt-4o-mini",
            messages=[
                {"role": "system", "content": "You are an expert in clustering survey data."},
                {"role": "user", "content": prompt}
            ],
            max_tokens=20,  # 限制回應的長度以只包含 cluster 編號
            temperature=0.1  # 設定較低的隨機性以提高準確性
        )
```

Clustering Methods 3: Iterative Clustering using Open AI

Outcome after 2nd cluster

ANSWER	NUMBER	QUESTION	YEAR	cluster	similarity
F02B02F	f02b02	是您的配偶領失業保險金?	CVIII2014	28	0
F02B03F	f02b03	是其他家人領失業保險金?	CVIII2014	29	0
F02C	f02c	在去年之中，您家庭自政府得到的補助總計大約是多少?	CVIII2014	3	1
F03A	f03a	在過去一年裡，您(和您配偶)房屋貸款支出平均每月__元	CVIII2014	4	0.85
F03B	f03b	在過去一年裡，您(和您配偶)保姆或幫傭(包括家事管理)支出平均每月__元	CVIII2014	7	1
F04A	f04a	請問您(和您配偶)過去一年裡，人身商業保險__元	CVIII2014	8	0.85
F04B	f04b	請問您(和您配偶)過去一年裡，家具與家庭耐久設備__元	CVIII2014	9	1
F05F	f05	在去年(102年)中，您(和您配偶)全部的支出(含給父母親，繳稅等支出)大約__元	CVIII2014	30	0
F06A	f06a	請問您的身高是__公分	CVIII2014	31	0
F06B	f06b	請問您的體重是__公斤	CVIII2014	32	0
G01F	G01	請問您平均每週大約花多少時間幫忙做家務工作?__小時	CIII2004	1	0.85
LABE	G02Z01	在去年(1-12月份)之中，您的配偶是否曾經領取政府失業保險金?	CIII2004	28	1
LABC	G02Z02	若是，自__月份	CIII2004	0	0
LABC	G02Z03	領到__月份	CIII2004	0	0
LABAN	G03	在去年之中，您家庭自政府得到的補助總計大約是多少?__元	CIII2004	3	1
LABAN	G04A	在過去一年裡，您家中房屋貸款支出平均每月__元?	CIII2004	4	0.85

Automation Process

Data Preprocessing

- Select 2002, 2004, 2008, 2014, and 2018 PSFD
- Transform sas file to csv file, sine the website output file is sas file

Survey Questions Clustering

1. Tokenize → Embedding → Clustering(K-means) → Similarilarity comparison
2. Directly call GPT4 API to do cluster
3. Label the questions of median year → call generative model API to cluster and give similarity→ Label the questions of median year of unclustered data → call generative model API and repeat

Answer Organizing

- Organize related answers in different years using "union”

Output Survey outcome

- Merge the individual survey answer to corresponding questions data

Automation Process: Organizing Answer

- Each question corresponds to a set of **options**.
- Pick the combination of options where the problem is in the same cluster.
- Take the **union** and make a new option with corresponding option dictionary.

variable_map.csv:

ANSWER_new	OPTION_dict	OPTION_new	OPTION_pre	ANSWER_pre	YEAR	NUMBER
2-8	{'1': 0, '2': 1, '3': 2, '6': 3, '8': 4, '9': 5}, {'0': 6, '1': 0, '2': 1, '3': 2, '6': 3, '8': 4, '9': 5}	0=不重要, 1=重要, 2=很重要, 3=不知道, 4=拒答, 5=遺漏值, 6=跳答	1 = '1 不重要', 2 = '2 重要', 3 = '3 很重要', 6 = '6 不知道', 8 = '8 拒答', 9 = '9 遺漏值' ;, 0 = '0 跳答', 1 = '1 不重要', 2 = '2 重要', 3 = '3 很重要', 6 = '6 不知道', 8 = '8 拒答', 9 = '9 遺漏值' ;	C01G, c01g	CVIII2014, CX2018	c01g, c01g

Automation Process

Data Preprocessing

- Select 2002, 2004, 2008, 2014, and 2018 PSFD
- Transform sas file to csv file, sine the website output file is sas file

Survey Questions Clustering

1. Tokenize → Embedding → Clustering(K-means) → Similarilarity comparison
2. Directly call GPT4 API to do cluster
3. Label the questions of median year → call GPT4 to cluster → Label the questions of median year of unclustered data → call GPT4 and repeat

Answer Organizing

- Organize related answers in different years using "union”

Output Survey outcome

- Merge the individual survey answer to corresponding questions data

Automation Process: Output the Survey Outcome

- Merge the individual survey answer to corresponding questions data

Expected result:

ID	YEAR	Cluster1	Cluster2
受訪者A編號x01	2002	A's response in Cluster 1 from 2002.	A's response in Cluster 2 from 2002.
受訪者A編號x01	2008	A's response in Cluster 1 from 2008.	A's response in Cluster 2 from 2008.
受訪者B編號x01	2002	B's response in Cluster 1 from 2002.	B's response in Cluster 2 from 2002.
受訪者B編號x01	2008	B's response in Cluster 1 from 2008.	B's response in Cluster 2 from 2008.

Agenda

1. Main Issues & Research Objective
2. Process and Methods
3. **Model Outcome & Comparison**
4. Demo
5. Expected solution
6. Future Prospects

Outcome of Clustering Methods 1: Statistical learning

variable_map_type_7

ANSWER_new	OPTION_dict	OPTION_new	OPTION_pre	ANSWER_pre	YEAR	NUMBER
0-1	{'CVIII2014': {'0': 0, '1': 1, '2': 2, '6': 3, '8': 4, '9': 5}, 'C	0=跳答, 1=有, 2=沒有, 3=不知道, 4=拒答	0 = '0 跳答', 1 = '1 有', 2 = '2 沒有', 6 =	F02A01F, f02b01f	CVIII2014, CX2018	f02a01, f02b01
0-10	{'CVIII2014': {'0': 0, '1': 1, '2': 2, '6': 3, '8': 4, '9': 5}}	0=跳答, 1=有, 2=沒有, 3=不知道, 4=拒答	0 = '0 跳答', 1 = '1 有', 2 = '2 沒有', 6 =	F02A07F	CVIII2014	f02a07
0-11	{'CX2018': {'0': 0, '1': 1, '2': 2, '6': 3, '8': 4, '9': 5}, 'CVI	0=跳答, 1=有, 2=沒有, 3=不知道, 4=拒答	0 = '0 跳答', 1 = '1 有', 2 = '2 沒有', 6 =	f02b02f, F02A08F	CX2018, CVIII2014	f02b02, f02a08
0-12	{'CVIII2014': {'0': 0, '1': 1, '2': 2, '6': 3, '8': 4, '9': 5}}	0=跳答, 1=有, 2=沒有, 3=不知道, 4=拒答	0 = '0 跳答', 1 = '1 有', 2 = '2 沒有', 6 =	F02A10F	CVIII2014	f02a10
0-2	{'CVIII2014': {'0': 0, '1': 1, '2': 2, '6': 3, '8': 4, '9': 5}}	0=跳答, 1=有, 2=沒有, 3=不知道, 4=拒答	0 = '0 跳答', 1 = '1 有', 2 = '2 沒有', 6 =	F02A11F	CVIII2014	f02a11
0-3	{'CX2018': {'0': 0, '1': 1, '2': 2, '6': 3, '8': 4, '9': 5}, 'CVI	0=跳答, 1=有, 2=沒有, 3=不知道, 4=拒答	0 = '0 跳答', 1 = '1 有', 2 = '2 沒有', 6 =	f02b97f, F02A12F	CX2018, CVIII2014	f02b97, f02a12
0-4	{'CII2002': {'0': 0, '9999996': 1, '9999997': 2, '9999998	0=跳答，不適用, 1=不知道, 2=其他, 3=	0 = '跳答，不適用', 9999996 = '不知道	LABAK, f02c, f02f, LABAN, F02(	CII2002, CX2018, CV2008, CIII2	g03, f02c, f02, G03, f02c
0-5	{'CVIII2014': {'0': 0, '1': 1, '2': 2, '6': 3, '8': 4, '9': 5}}	0=跳答, 1=有, 2=沒有, 3=不知道, 4=拒答	0 = '0 跳答', 1 = '1 有', 2 = '2 沒有', 6 =	F02A02F	CVIII2014	f02a02
0-6	{'CVIII2014': {'0': 0, '1': 1, '2': 2, '6': 3, '8': 4, '9': 5}, 'C	0=跳答, 1=有, 2=沒有, 3=不知道, 4=拒答	0 = '0 跳答', 1 = '1 有', 2 = '2 沒有', 6 =	F02A03F, f02b03f	CVIII2014, CX2018	f02a03, f02b03
0-7	{'CX2018': {'0': 6, '1': 6, '2': 6, '6': 6, '8': 6, '9': 6}, 'CVI	0=跳答, 1=有, 2=沒有, 3=不知道, 4=拒答	0 = '0 跳答', 1 = '1 有', 2 = '2 沒有', 6 =	f02b04f, f02b05f, F02A04F	CX2018, CX2018, CVIII2014	f02b04, f02b05, f02a04
0-8	{'CVIII2014': {'0': 0, '1': 1, '2': 2, '6': 3, '8': 4, '9': 5}}	0=跳答, 1=有, 2=沒有, 3=不知道, 4=拒答	0 = '0 跳答', 1 = '1 有', 2 = '2 沒有', 6 =	F02A05F	CVIII2014	f02a05
0-9	{'CVIII2014': {'0': 0, '1': 1, '2': 2, '6': 3, '8': 4, '9': 5}}	0=跳答, 1=有, 2=沒有, 3=不知道, 4=拒答	0 = '0 跳答', 1 = '1 有', 2 = '2 沒有', 6 =	F02A06F	CVIII2014	f02a06
1-1	{'CIII2004': {'9999991': 0, '9999992': 1, '9999996': 2, '	0=不固定, 1=無法估計, 2=不知道, 3=其	9999991 = '9999991 不固定', 9999992	LABAN, f04a, F04A, LABAK, f04	CIII2004, CX2018, CVIII2014, CI	G05A, f04a, f04a, g05a, f0
1-2	{'CV2008': {'0': 0, '9999999': 1}}	0=跳答，不適用, 1=缺漏值	0 = "0000000 跳答，不適用", 9999999	f04c	CV2008	f04c
10-1	{'CIII2004': {'9999991': 0, '9999992': 1, '9999996': 2, '	0=不固定, 1=無法估計, 2=不知道, 3=其	9999991 = '9999991 不固定', 9999992	LABAN, F03A, f03a, f03a, LABA	CIII2004, CVIII2014, CV2008, C	G04A, f03a, f03a, f03a, g0
10-2	{'CIII2004': {'9999991': 0, '9999992': 1, '9999996': 2, '	0=不固定, 1=無法估計, 2=不知道, 3=其	9999991 = '9999991 不固定', 9999992	LABAN	CIII2004	G04B01
10-3	{'CIII2004': {'9999991': 0, '9999992': 1, '9999996': 2, '	0=不固定, 1=無法估計, 2=不知道, 3=其	9999991 = '9999991 不固定', 9999992	LABAN	CIII2004	G04B02
10-4	{'CIII2004': {'9999991': 0, '9999992': 1, '9999996': 2, '	0=不固定, 1=無法估計, 2=不知道, 3=其	9999991 = '9999991 不固定', 9999992	LABAN	CIII2004	G04D
10-5	{'CIII2004': {'9999991': 0, '9999992': 1, '9999996': 2, '	0=不固定, 1=無法估計, 2=不知道, 3=其	9999991 = '9999991 不固定', 9999992	LABAN	CIII2004	G04E
10-6	{'CIII2004': {'9999991': 0, '9999992': 1, '9999996': 2, '	0=不固定, 1=無法估計, 2=不知道, 3=其	9999991 = '9999991 不固定', 9999992	LABAN	CIII2004	G04F
11	{'CII2002': {'0': 5, '9999996': 5, '9999997': 5, '9999998	0=跳答，不適用, 1=不知道, 2=其他, 3=	0 = '跳答，不適用', 9999996 = '不知道	LABAK, LABAK, LABAK, LABAK	CII2002, CII2002, CII2002, CII2	g05d, g05f, g05e, g05c
12-1	{'CII2002': {'0': 0, '96': 1, '97': 2, '98': 3, '99': 4}, 'CIII2	0=跳答，不適用, 1=不知道, 2=其他, 3=	0 = '跳答，不適用', 96 = '不知道', 97 =	LABD, LABC	CII2002, CIII2004	g02z3, G02Z03

Outcome of Clustering Methods 1: Statistical learning

Outcome Description	• Attempted to reduce the number of clusters in the first layer to prevent different questionnaire questions from being mistakenly assigned to separate clusters. • Enhanced clustering accuracy by adding another layer of clustering through calculating the differences in Euclidean distances of the text.
Accuracy	• Type I error rate = 39/123
Limitation	• Unable to effectively classify questions where sub-questions do not extend the main question. • Cluster has high probability to contains questions from only a single year. • The number of clusters may need to be adjusted based on different samples to find the optimal parameters.

Outcome of Clustering Methods 2: OpenAI (Call GPT4o directly)

ANSWER	NUMBER	QUESTION	YEAR	cluster
F01A	f01a	在去年(102年)之中，請問您平均每週大約花多少時間作家務工作?__小時	CVIII2014	6
F01B	f01b	在去年(102年)之中，請問您的配偶平均每週大約花多少時間作家務工作?__小時	CVIII2014	6
F02A01F	f02a01	在去年之中，您的家庭是否曾經得到政府的中低收入戶生活補助?	CVIII2014	1
F02A02F	f02a02	在去年之中，您的家庭是否曾經得到政府的傷病醫療費用補助?	CVIII2014	1
F02A03F	f02a03	在去年之中，您的家庭是否曾經得到政府的教育補助(含五歲幼兒學費補助)?	CVIII2014	1
F02A04F	f02a04	在去年之中，您的家庭是否曾經得到政府的兒童托育補助?	CVIII2014	1
F02A05F	f02a05	在去年之中，您的家庭是否曾經得到政府的老人津貼(含老農津貼)?	CVIII2014	1
F02A06F	f02a06	在去年之中，您的家庭是否曾經得到政府的榮民就養金?	CVIII2014	1
F02A07F	f02a07	在去年之中，您的家庭是否曾經得到政府的身心障礙者補助?	CVIII2014	1
F02A08F	f02a08	在去年之中，您的家庭是否曾經得到政府的重大傷病補助?	CVIII2014	1
F02A09F	f02a09	在去年之中，您的家庭是否曾經得到政府的失業保險金?	CVIII2014	1
F02A10F	f02a10	在去年之中，您的家庭是否曾經得到政府的生育獎勵金(含生育補助)?	CVIII2014	1
F02A11F	f02a11	在去年之中，您的家庭是否曾經得到政府的天然災害補助?	CVIII2014	1
F02A12F	f02a12	在去年之中，您的家庭是否曾經得到政府的其他補助?	CVIII2014	1
F02B01F	f02b01	是您自己領失業保險金?	CVIII2014	1
F02B02F	f02b02	是您的配偶領失業保險金?	CVIII2014	1
F02B03F	f02b03	是其他家人領失業保險金?	CVIII2014	1
F02C	f02c	在去年之中，您家庭自政府得到的補助總計大約是多少?	CVIII2014	0

Outcome of Clustering Methods 2: OpenAI (Call GPT4o directly)

Outcome Description	• GPT-4o may call other functions, such as K-means, to perform clustering, so the outcome may sometimes resemble the results of Method 1. • Can change the prompt to prevent GPT from calling any other clustering functions; however, this may lead to worse outcomes.
Accuracy	• Type I error rate = 71/123
Limitation	• GPT has input token limitation • The cluster answers generated by GPT may vary across different trials.

Outcome of Clustering Methods 3: Iterative Clustering using Open AI

QUESTION	YEAR	cluster
在去年(102年)之中，請問您的配偶平均每週大約花多少時間作家務工作?__小時	CVIII2014	2
請問您的配偶平均每週大約花多少時間作家務工作?__ 小時	CV2008	2
在去年(106年)之中，請問您的配偶平均每週大約花多少時間做家務工作?__小時	CX2018	2
在去年之中，您家庭自政府得到的補助總計大約是多少?	CVIII2014	3
在去年之中，您家庭自政府得到的補助總計大約是多少?__元	CIII2004	3
在去年(96年)之中，您家庭自政府得到的補助總計大約是多少?__ 元	CV2008	3
去年中，您家庭自政府補助大約多少	CII2002	3
在去年之中，您家庭自政府得到的補助總計大約是多少?	CX2018	3
在過去一年裡，您(和您配偶)房屋貸款支出平均每月__元	CVIII2014	4
在過去一年裡，您家中房屋貸款支出平均每月__元?	CIII2004	4
在過去一年裡，您(和您配偶)每個月的平均支出房屋貸款支出平均每月 __元?	CV2008	4
去年中，您家每月支出情況?房屋貸_	CII2002	4
在過去一年裡，您(和您配偶)房屋貸款支出平均每月__元	CX2018	4
在過去一年裡，您家中標會支出活會平均每月__元?	CIII2004	5
標會支出活會平均每月 __元	CV2008	5
標會支出活會平均每月____元	CII2002	5

type7.csv

Outcome of Clustering Methods 3: Iterative Clustering using Open AI

The best method!

QUESTION	YEAR	cluster
在去年(102年)之中，請問您的配偶平均每週大約花多少時間作家務工作?__小時	CVIII2014	2
請問您的配偶平均每週大約花多少時間作家務工作?__ 小時	CV2008	2
在去年(106年)之中，請問您的配偶平均每週大約花多少時間做家務工作?__小時	CX2018	2
在去年之中，您家庭自政府得到的補助總計大約是多少?	CVIII2014	3
在去年之中，您家庭自政府得到的補助總計大約是多少?__元	CIII2004	3
在去年(96年)之中，您家庭自政府得到的補助總計大約是多少?__ 元	CV2008	3
去年中，您家庭自政府補助大約多少	CII2002	3
在去年之中，您家庭自政府得到的補助總計大約是多少?	CX2018	3
在過去一年裡，您(和您配偶)房屋貸款支出平均每月__元	CVIII2014	4
在過去一年裡，您家中房屋貸款支出平均每月__元?	CIII2004	4
在過去一年裡，您(和您配偶)每個月的平均支出房屋貸款支出平均每月 __元?	CV2008	4
去年中，您家每月支出情況?房屋貸_	CII2002	4
在過去一年裡，您(和您配偶)房屋貸款支出平均每月__元	CX2018	4
在過去一年裡，您家中標會支出活會平均每月__元?	CIII2004	5
標會支出活會平均每月 __元	CV2008	5
標會支出活會平均每月____元	CII2002	5

type7.csv

ID	YEAR	9	10	11	12	13	14	15
30171	CIII2004	[0.0]	[0.0]	[2.0]	[3.0]	[19.0]	[59.0]	
30171	CV2008	[0.0]	[0.0]	[1.0]	[22.0]	[15.0]	[12.0]	
30172	CV2008	[0.0]	[0.0]	[1.0]	[22.0]	[13.0]	[30.0]	
30191	CII2002	[0.0]	[0.0]					
30191	CIII2004	[0.0]	[0.0]	[1.0]	[25.0]	[13.0]	[25.0]	
30192	CV2008	[0.0]	[0.0]	[1.0]	[26.0]	[18.0]	[40.0]	

result_type7.csv

Outcome of Clustering Methods 3: Iterative Clustering using Open AI

Outcome Description	• **Cluster 1:** It takes approximately 45s to classify 100 records, costing $0.01USD, with similarity comparison, 6 wrong answer can be corrected. • **Cluster 2:** It takes about 110s to process 100 records, costing $0.015 USD.
Accuracy	• **Cluster 1:** Type I error rate = **1/49** • **Cluster 2:** Successfully classified into 9 more clusters with **1/78** Type I error rate, but 6 clusters containing a total of 12 records remain unclassified.
Limitation	• Still need to manually select the same topic across different years. • If the survey contains the same questions under the same topic for the same year, they may not be clustered into the same cluster. • While GPT's classification accuracy is very high, improving the "granularity" of the classification depends on the **trade-off between accuracy, cost, and time.** Ultimately, the records remaining in Cluster 0 **still require manual classification**.

Agenda

1. Main Issues & Research Objective
2. Process and Methods
3. Model Outcome & Comparison
4. **Demo**
5. Expected solution
6. Future Prospects

[redacted]

Agenda

1. Main Issues & Research Objective
2. Process and Methods
3. Model Outcome & Comparison
4. Demo
5. **Expected solution**
6. Future Prospects

Cost and Benefit Analysis:

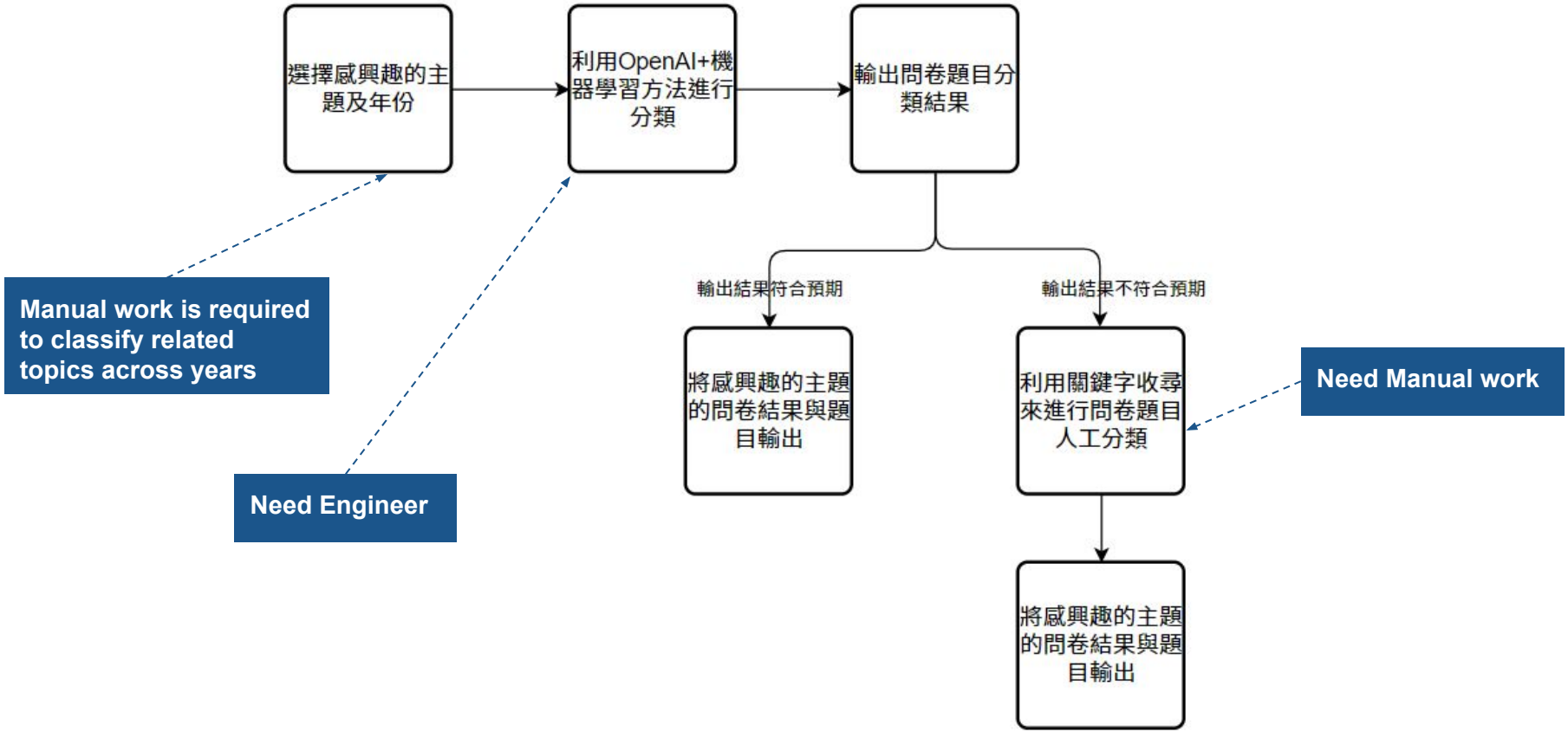
Cost

- Clustering PSFD across 30 years would require only about **2.75 hours** and would cost no more than **5,000 NTD** with 2 iterative clustering
- **Note:** There will be additional hours in outputing the complete survey outcome.

Benefit

- Reducing the original need for manual classification by more than half.
- Easier to manage programming result
- Freeing up valuable human resources to focus on strategic analysis and decision-making.

Expected Solution Process:



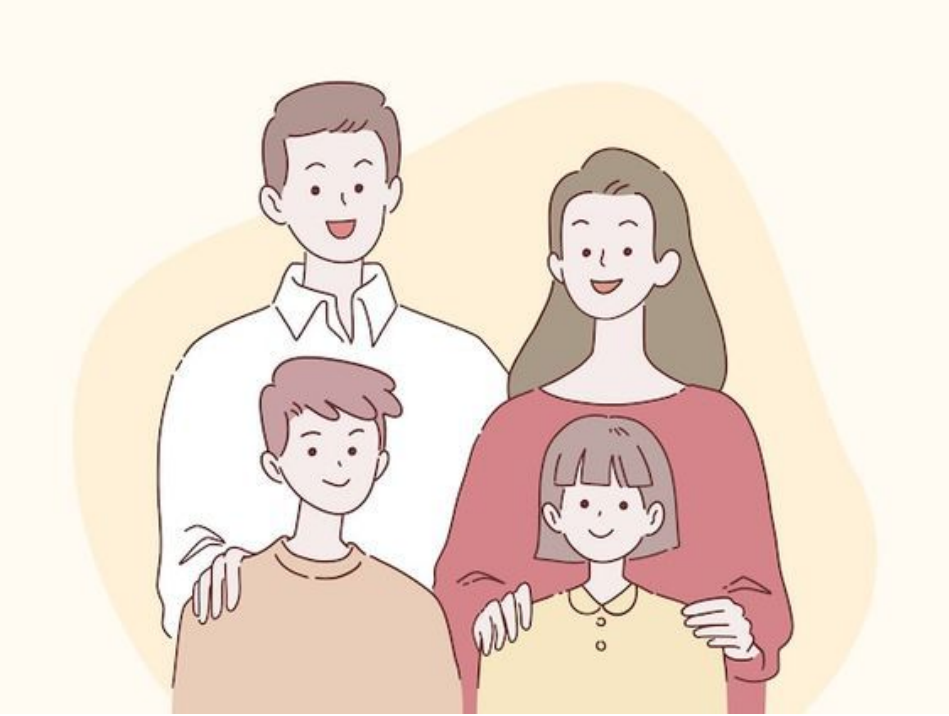
Agenda

1. Main Issues & Research Objective
2. Process and Methods
3. Model Outcome & Comparison
4. Demo
5. Expected solution
6. **Future Prospects**

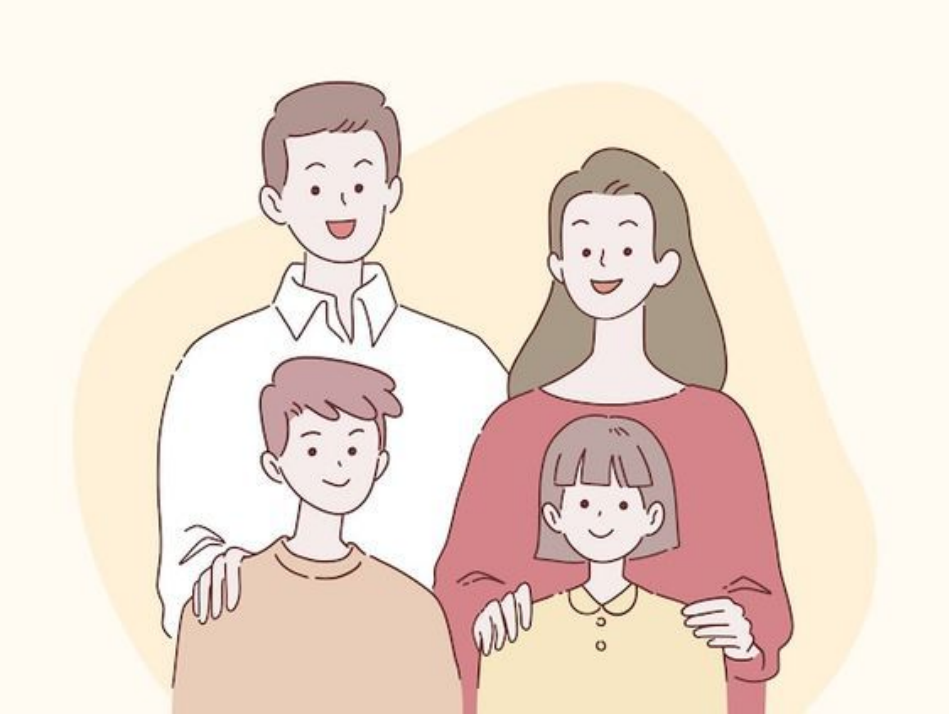
Future Prospect

Conclusion	• Using GPT-4o-mini not only reduces costs and speeds up processing but also achieves a high accuracy rate, significantly decreasing the time required for manual classification. • As language models continue to evolve, future versions may provide even faster and more accurate models for classification tasks.
Future Prospect	• Utilize OpenAI to automatically **identify similar topics** across different years and categorize analogous answers into a single merged response. • Create a user-friendly interface similar to **IPUMS** that allows researchers to select their topics of interest, years, and options, and then directly export the survey results. • In addition to linking data across years, also link parent-child samples. By using household numbers and ages, connect each participant's data with that of their cohabiting family members.

Questions



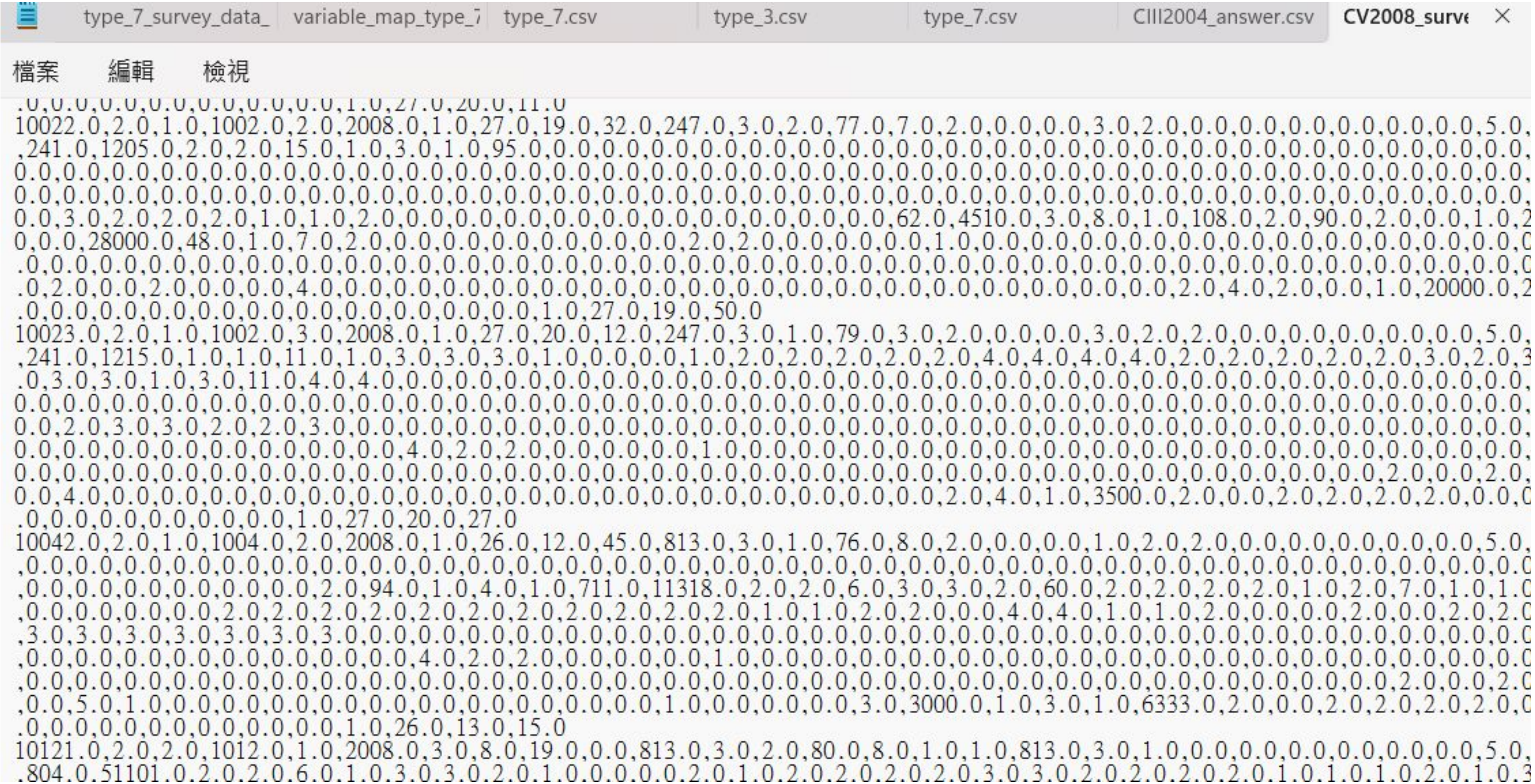
Appendix



2004年問卷回答對應表似乎有問題

```
CIII2004.G01F.X04A = "問卷
CIII2004.G01F.X04B = "主樣本年齡層"
CIII2004.G01F.X04C = "訪問時段"
CIII2004.G01F.X04D = "戶號"
CIII2004.G01F.X04E = "子女編號"
CIII2004.G01F.X05 = "訪問年份"
CIII2004.G01F.X09 = "現住地址"
CIII2004.G01F.X11 = "受訪者與戶長之關係"
CIII2004.G01F.A01 = "性別"
CIII2004.G01F.A02 = "請問您民國幾年出生?民國    年"
CIII2004.G01F.A03Z01 = "請問您的出生地是"
CIII2004.G01F.A03Z02 = "出生地地區碼"
CIII2004.G01F.A04A = "請問您認為自己目前的健康狀況如何?"
CIII2004.G01F.A04B = "請問您是否有身體或精神上的疾病會讓您工作或行動不便?"
CIII2004.G01F.A05A = "請問您有沒有在軍中服過役?"
CIII2004.G01F.A05B01 = "民國   年入伍"
CIII2004.G01F.A05B02 = "民國   年退伍"
CIII2004.G01F.B01A = "請問您的父母是否滿意您目前的教育程度?"
CIII2004.G01F.B01B = "請問就您目前的環境和能力，您覺得您可以讀到什麼程度?"
CIII2004.G01F.B01C = "如果沒有環境和能力的限制，您希望讀書讀到什麼程度?"
CIII2004.G01F.B02 = "請問您最高的教育程度是什麼?"
CIII2004.G01F.B03A = "請問您過去一年是否轉過學?"
CIII2004.G01F.B03B = "請問您是什麼時候開始上高中/高職/五專?民國   年"
CIII2004.G01F.B04 = "請問您是如何進入高中/高職/五專的?"
CIII2004.G01F.B05 = "請問您唸的是什麼樣的學校?"
CIII2004.G01F.B06 = "請問您唸的是什麼科?(以目前就讀或畢業的學校為主)"
CIII2004.G01F.B07A01 = "您就讀的學校是在哪一個地區?"
```

0.0、0不知道如何做區隔 (一個是0元一個是跳答) 資料中有超過八成都是 0.0



```
.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,27.0,20.0,11.0
10022.0,2.0,1.0,1002.0,2.0,2008.0,1.0,27.0,19.0,32.0,247.0,3.0,2.0,77.0,7.0,2.0,0.0,0.0,3.0,2.0,0.0,0.0,0.0,0.0,0.0,0.0,5.0,
,241.0,1205.0,2.0,2.0,15.0,1.0,3.0,1.0,95.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,3.0,2.0,2.0,2.0,1.0,1.0,2.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,62.0,4510.0,3.0,8.0,1.0,108.0,2.0,90.0,2.0,0.0,1.0,2
0,0.0,28000.0,48.0,1.0,7.0,2.0,0.0,0.0,0.0,0.0,0.0,0.0,2.0,2.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0
.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0
.0,2.0,0.0,2.0,0.0,0.0,4.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,2.0,4.0,2.0,0.0,1.0,20000.0,2
.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,27.0,19.0,50.0
10023.0,2.0,1.0,1002.0,3.0,2008.0,1.0,27.0,20.0,12.0,247.0,3.0,1.0,79.0,3.0,2.0,0.0,0.0,3.0,2.0,2.0,0.0,0.0,0.0,0.0,0.0,5.0,
,241.0,1215.0,1.0,1.0,11.0,1.0,3.0,3.0,3.0,1.0,0.0,0.0,1.0,2.0,2.0,2.0,2.0,2.0,4.0,4.0,4.0,4.0,2.0,2.0,2.0,2.0,2.0,3.0,2.0,3
.0,3.0,3.0,1.0,3.0,11.0,4.0,4.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,2.0,3.0,3.0,2.0,2.0,3.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,4.0,2.0,2.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,2.0,0.0,2.0,
0.0,4.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,2.0,4.0,1.0,3500.0,2.0,0.0,2.0,2.0,2.0,2.0,0.0,0
.0,0.0,0.0,0.0,0.0,0.0,1.0,27.0,20.0,27.0
10042.0,2.0,1.0,1004.0,2.0,2008.0,1.0,26.0,12.0,45.0,813.0,3.0,1.0,76.0,8.0,2.0,0.0,0.0,1.0,2.0,2.0,0.0,0.0,0.0,0.0,0.0,5.0,
,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
,0.0,0.0,0.0,0.0,0.0,0.0,2.0,94.0,1.0,4.0,1.0,711.0,11318.0,2.0,2.0,6.0,3.0,3.0,2.0,60.0,2.0,2.0,2.0,2.0,1.0,2.0,7.0,1.0,1.0
,0.0,0.0,0.0,0.0,2.0,2.0,2.0,2.0,2.0,2.0,2.0,2.0,2.0,2.0,2.0,1.0,1.0,2.0,2.0,0.0,4.0,4.0,1.0,1.0,2.0,0.0,0.0,2.0,0.0,2.0,2.0
,3.0,3.0,3.0,3.0,3.0,3.0,3.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,4.0,2.0,2.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,2.0,0.0,2.0
,0.0,5.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,3.0,3000.0,1.0,3.0,1.0,6333.0,2.0,0.0,2.0,2.0,2.0,2.0,0
.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,26.0,13.0,15.0
10121.0,2.0,2.0,1012.0,1.0,2008.0,3.0,8.0,19.0,0.0,813.0,3.0,2.0,80.0,8.0,1.0,1.0,813.0,3.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,5.0,
,804.0,51101.0,2.0,2.0,6.0,1.0,3.0,3.0,2.0,1.0,0.0,0.0,2.0,1.0,2.0,2.0,2.0,2.0,3.0,3.0,2.0,2.0,2.0,2.0,1.0,1.0,1.0,2.0,1.0,3
```