
Mental-2024

Preprocessing with GPT

모바일시스템공학과
이승재

1. How to use Openai API

NAME	SECRET KEY	CREATED	LAST USED ⓘ
gpt_key 2	sk-...71MA	2025년 1월 22일	2025년 1월 23일

You need to obtain an API key from Openai

○ (env1) dku_mse1@dkumse:~/seung_jae/preprocess\$ pip install openai

Install openai package

```
from openai import OpenAI
client = OpenAI(api_key="sk-proj-oqj5_GUZsdnjMfvS79rqZYoa4AfQHmwy3lwQvwv7hd4yIEYckd2wzyTXFp-xLBL2070o
```

Enter the API key you received

○ (env1) dku_mse1@dkumse:~/seung_jae/preprocess\$ openai migrate

Enter and run this command to automatically update compatibility.

2. Price

"Calculating the price of the full data set didn't seem like a huge budget-intensive task."



gpt-3.5-turbo-0125	\$0.50 / 1M tokens	\$1.50 / 1M tokens
gpt-3.5-turbo-instruct	\$1.50 / 1M tokens	\$2.00 / 1M tokens
gpt-3.5-turbo-1106	\$1.00 / 1M tokens	\$2.00 / 1M tokens
gpt-3.5-turbo-0613	\$1.50 / 1M tokens	\$2.00 / 1M tokens
gpt-3.5-turbo-16k-0613	\$3.00 / 1M tokens	\$4.00 / 1M tokens
gpt-3.5-turbo-0301	\$1.50 / 1M tokens	\$2.00 / 1M tokens

750 words

1k tokens / 0.002\$

0_AI_LKH_AD.csv
0_AI_LKH_MCI.csv
0_AI_LKH_NC.csv

2396736 words $\hat{=}$ 6\$

2. read_data.py

"Since the full data set is large, only 5 rows of the test data set were run to save tokens."

CSV 파일 내용:

	ID	SEX	D_TEST	AGE	CGA2	FINAL_DX	
0	N1	여	9/10/13	NaN	<기분>\n스트레스 딸과 사위 싸운다.\n자궁경부암, 2-3기, 자궁내막암 2012...	NC	
1	N2	남	2/24/14	72.0	무학 (0y)\n읽기 미숙 / 쓰기 미숙\n<현병력>\n순천향 병원 신경외과에서 X...	NC	
2	N3	여	2/25/14	75.0	박서문례 \n-신장한쪽에 혹이 있어 제거수술함 : 1년전에 하셨다고 함 입원 일주일...		NC
3	N4	여	3/18/14	73.0	길순덕\n혼자 방문하심\n학력:무학\n한글 읽기 미숙/ 쓰기 불가/ 숫자 미숙\n...	NC	
4	N5	여	3/25/14	73.0	김옥숙\n\n혼자방문\n학력:9년(중졸)\n한글 읽고쓰기 가능/숫자 가능\n\n<현...	NC	
..	
172	N173	여	11/6/24	76.0	주관적 기억력 저하 있음. \nNP - -1/5 < 1.0 \n정상 SCI \n무울...	NC	
173	N174	여	11/19/24	88.0	\n<동남구보건소 진료 검사 대상자>\n이름: 이춘화\n성별: 여 \n생년...	NC	
174	N175	여	11/27/24	96.0	\n주관적기억력 저하 없음 보호자 정보일치\n신경심리검사 저하 분명함\n나이 고려해...		NC
175	N176	여	12/5/24	62.0	4년전 유방암 수술 \n항암치료, 방사선 치료, 표적치료함. \n\n재활의학과에서 ...	NC	
176	N177	여	12/11/24	77.0	주관적 기억력 저하 없음 보호자 정보 없음\nNP 저하 없음\n무울증 없음 \n정상...	NC	

[0_AI_LKH_AD.csv]

2. translate.py

LLaMA

[illegible]

[CGA2] There are many parts in the translated version of the category that do not match the original data and contain unnecessary content.

GPT

Processing row 1/5
Translated text for row 1: {'ID': 'D1', 'SEX': 'Female', 'D_TEST': '9/23/14', 'AGE': 97.0, 'CGA2': 'Lee Jung-im\n\nAccompanied by daughter\n\nilliterate (0) or can / handle numbers\n\nwidowed / Living with son and daughter-in-law\n\nMedical History:\n\nHypertension (+) Diabetes (-) Hyperlipidemia (-)\n\nMedications at 20 years\n\nSurgery: None\n\nOthers: None\n\nVision: Good after cataract surgery (both eyes)\n\nHearing: Poor, wears hearing aid, unable to have conversations\n\nHearing well\n\nSleep: Sleeps well\n\nAlcohol, tobacco: None\n\nHas difficulty walking due to pain in legs and back.\n\n\nGuardian's Comments:\n\nSeems to have lost up at 3 am or 5 am, and questions why it matters when she showers in the middle of the night.\n\nHad a voucher, but the caregiver took away her clothes saying it makes strange comments while watching TV. Thinks there are people changing clothes in front of the TV, and offers fruits to the TV screen.\n\nLooks at the reality.\n\n\nComplains of memory decline. Symptoms for 4-5 years, worsening recently.\n\nStarted gradually, progressing gradually currently no turn off the water.\n\nForgets items, especially important ones like bankbooks. Keeps searching in son's room.\n\nMorning meal: Rice, kimchi, seaweed soup, Unsed items: Unable to go out alone to know.\n\nRecent news, drama content: Unable to remember the content, doesn't hear well, only looks at pictures.\n\nBurns medication, says blood pressure is not high so doesn't take it (seems to forget without thinking)\n\nRemembers children's names: 2 sons, 1 daughter (Kim Young grandchildren's names: Remembers all when sees their faces\n\n\nLanguage:\n\nNo problem with fluency. WFD (-), Naming difficulty (-), Comprehension (-)\n\nHear Year, month, day, of the week DK/10/DK/DK/Fall (-,-,-,-,+)\n\nForgets anniversaries and birthdays.\n\nPlaces: No problem with familiar places. No problem with ys accompanied.\n\nPeople: Recognizes people well.\n\n\nJudgment and Problem Solving: 2\n\nImpaired comprehension. (-) Social judgment (-) Etiquette (-)\n\n\nPersonal s of interest (-), Decreased appetite (-), Weight change (-), Sleep disorder (+), Restlessness (+), Fatigue (-),\n\nFeelings of worthlessness or guilt (+), Thoughts (+)\n\n\nAlways depressed being alone. Lost son in 1974, always depressed, gets upset when talking about son. Once tried to commit suicide by going up t.\n\n\n\nSocial Activities: 2\n\nMeetings: None\n\nWalking: None\n\nReligious activities: None\n\nExercise: None\n\n\nHome Life: 2\n\nHobbies: Used to do sewing, but can't. Does not cook\n\nIssues with using household appliances (0), Using remote control (0), Making phone calls (0), Answering phone calls (X)\n\n\nMoney management (0) Management (-), Bank transactions (-): Unable to manage directly\n\n\n\nPersonal Daily Life: 2\n\nDressing, eating, washing face, bathing, changing clothes - Can't needs assistance with showering.' 'FINAL DX': 'PSD'}

Mostly consistent with original data

3. Extract with LLaMA

"There is a problem that when extracting with Llama, it includes general information that is not related to the input text, resulting in a lot of unnecessary content."

```
1 Gender, Age, Education, Literacy and Numeracy, Medical History, Medications, Surgeries, Stroke, Other History, Vision, Hearing, Diet, Sleep, Alcohol, Smoking,
52
53 Please provide a new text with relevant content for me to assist you.", "There is no information related to ""Vision"" in the provided text. The
54
55 As a result, there is nothing for me to extract regarding ""Vision"". If you have a different text or would like me to assist you with another t
56
57 However, I can provide a standard response to this type of situation:
58
59 **No dietary information available**
60
61 If you'd like to provide more context or a different text, I'll be happy to assist with extracting diet-related information for you.", "I'm a med
62
63 Since there is no relevant information available, my response will be: **There is no information about Sleep in the given text.**", "There is no
64
65 Would you like me to assist with a different text?", "There is no information about ""Smoking"" in the provided text. The text appears to be a re
66
67 However, I can provide general information about smoking if that's what you're looking for:
68
69 **Effects of Smoking:**
70
71 * Increases risk of lung cancer and other cancers
72 * Causes chronic obstructive pulmonary disease (COPD)
73 * Raises blood pressure and heart rate
74 * Decreases lung function and exercise performance
75 * Lowers fertility in both men and women
76
77 **Health Risks Associated with Smoking:**
78
79 * Cardiovascular disease (heart attacks, strokes)
80 * Respiratory problems (asthma, emphysema)
81 * Cancers (lung, mouth, throat, bladder, kidney)
82 * Skin conditions (acne, psoriasis)
83 * Reproductive issues (infertility, miscarriage)
84
85 **Benefits of Quitting Smoking:**
86
87 * Reduces risk of heart disease and stroke
88 * Lowers blood pressure
89 * Improves lung function and exercise performance
```

3. Extract with GPT – Case 1.

Case1) Translate + Extract

```
1 ID,Gender,Age,Education,Literacy and Numeracy,Medications,Surgeries,Stroke,Medical History,Main Complaints,Memor
2 Not specified,Female,Not specified,Not specified,"Limited literacy in Korean, can read and write but not profici
3 김형재,남,97세,9년,"한글 읽고 쓰기 가능, 숫자 가능",기관지약(폐): 10년 전부터 하루 세 번,없음,없음,"심장병(-), 혈
4 박귀임,Female,80 years old,Not specified,"Unable to read and write Korean, unable to deal with numbers","Does no
5 Not specified,Not specified,Not specified,Not specified,한글 읽고 쓰기 불가능 / 숫자 불가능,"{'Blood Pressure':
6 강일선,Not specified,Not specified,무학,"{'Reading and Writing': '미숙', 'Numeracy': '모름'}",안함,다리 다침 (6년
```

Prompt

"After translating the given text into English, summarize the contents of the CGA2 items. And naturally extract information from the given text."

→[Gender], [Age], [Education], [Medical History], [Main Complaints], [Memory Issues], [Judgment and Problem Solving]

Sys message

"You are an expert medical data extractor"

"There is a disadvantage in that the translation process and the extraction process cannot be translated properly if they are included in the prompt at the same time."

3. Extract with GPT – Case 2.

Case2) Translate → Extract only [CGA2]

```
1 ID,Gender,Age,Education,Literacy and Numeracy,Medications,Surgeries,Stroke,Medical History,Main Complaints,Memory
2 Not specified,Female,Not specified,Not specified,"Limited literacy in Korean, can read and write but not proficient
3 김형재,남,97세,9년,"한글 읽고 쓰기 가능, 숫자 가능",기관지약(폐): 10년 전부터 하루 세 번,없음,없음,"심장병(-), 혈압(-),
4 박귀임,Female,80 years old,Not specified,"Unable to read and write Korean, unable to deal with numbers","Does not
5 Not specified,Not specified,Not specified,Not specified,한글 읽고 쓰기 불가능 / 숫자 불가능,"{'Blood Pressure': '5~6
6 강일선,Not specified,Not specified,무학,"{'Reading and Writing': '미숙', 'Numeracy': '모름'}",안함,다리 다침 (6년 전
```

Prompt

1. "You are a medical data summarizer. Please summarize the following medical text in English"
2. "Text: {selected_columns['CGA2']}\n\n"
3. "Extract the following information from the given text"
→[Gender], [Age], [Education], [Medical History], [Main Complaints], [Memory Issues], [Judgment and Problem Solving]

Sys message

"You are an expert medical data extractor"

"To reduce token usage, only the contents of [CGA2] items with important content were extracted, but even the translation was not done properly."

3. Extract with GPT – Case 3.

Case3) Translate → Summarize → Extract

Prompt

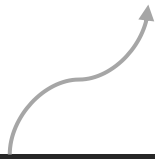
1. "You are a medical data summarizer. Please summarize the following medical text in English"

2. "Extract the following information from the given text" → [Gender], [Age], [Education], [Medical History], [Main Complaints], [Memory Issues], [Judgment and Problem Solving]

Sys message

"You are an expert medical data extractor"

"Because it has been summarized once, the necessary information is extracted."



```
[Gender]
Female

[Age]
97

[Education]
Illiterate

[MedicalHistory]
"{ 'Hypertension': 'Present', 'Diabetes': 'Absent', 'Hyperlipidemia': 'Absent', 'Medication': 'Antihypertensive medication for about 20 years', 'Surgery': 'Cataract surgery on both eyes' }"

[MainComplaints]
Leg and back pain leading to difficulty in walking

[MemoryIssues]
"{ 'Duration': '4-5 years', 'Progression': 'Gradual', 'Current Impact on Daily Life': 'Forgetting to turn off the water, misplacing important items like bankbooks, entering son's room multiple times' }"

[JudgmentProblemSolving]
"{ 'Complaints': 'Impaired understanding, social judgment, and etiquette', 'Behavioral Symptoms': 'Depression, insomnia, restlessness, feelings of worthlessness or guilt, suicidal thoughts' }"
```

3. Extract with GPT – Case 3.

[Gender]
Female

[Age]
97

[Education]
Illiterate

[MedicalHistory]
"{'Hypertension': 'Present', 'Diabetes': 'Absent', 'Hyperlipidemia': 'Absent', 'Medication': 'Antihypertensive medication for about 20 years', 'Surgery': 'Cataract surgery on both eyes'}"

[MainComplaints]
Leg and back pain leading to difficulty in walking

[MemoryIssues]
"{'Duration': '4-5 years', 'Progression': 'Gradual', 'Current Impact on Daily Life': '"Forgetting to turn off the water, misplacing important items like bankbooks, entering son's room multiple times'"}

[JudgmentProblemSolving]
"{'Complaints': 'Impaired understanding, social judgment, and etiquette', 'Behavioral Symptoms': 'Depression, insomnia, restlessness, feelings of worthlessness or guilt, suicidal thoughts'}"

[Extracted Data]

"As we go through the summary process, we can see that less important information is removed."

CGA2
이정임

딸 함께 내방함
무학(0y)
한글 읽고 쓰기 미숙 / 숫자 가능
사별 / 아들 내외 함께 거주
<현병력>
혈압(+) **당뇨(-) 고지혈증(-)**
약물복용 ; 혈압약 20년 정도 복용 중
수술 ; 없음
기타 : 없음
시력 ; 백내장 수술(양쪽) 후 잘 보이심
청력 ; 잘 못 들으신다. 보청기 안 끼심, 대화 불가
식사 ; 조금 드신다. 잘 못 드심
수면 ; 잘 주무신다.
술, 담배 : 안 함

다리 허리가 아파서 잘 못 걸니다.

[Original Data]

3. Extract with GPT – Hyperparameter

Temperature

0.2

→ In data extraction tasks, consistency and accuracy are more important than creativity, so it is appropriate to use a low value.

Top_p

0.1

→ Increase consistency in data extraction by selecting only words with high probability.

presence_penalty

0.0

→ Since this is a task that does not require the introduction of new topics, we set the presence_penalty to 0.0 to encourage the model to be faithful to the text.

4. Conclusion

Gender, Age, Education, Medical History, Main Complaints, Memory Issues, Judgment and Problem Solving

Female, 74, Unknown, "Hypertension for 20 years, cataract surgery on both eyes, hearing impairment with refusal to use hearing aid", "Memory decline, difficulty in daily activities, confusion with time and tasks, depressive symptoms, suicidal thoughts", "{ 'Short-term': 'Forgets to turn off water, frequently loses important items, struggles to remember recent events or news', 'Long-term': 'Difficulty recalling children's names, relies on others for medication management' }", "Impaired understanding, social judgment, and etiquette"

Male, 97.9 years, "Hypertension, Respiratory issues", "Dementia diagnosis, Language abuse from spouse", "{ 'Short-term': { 'Symptoms Start Date': 'May 2015', 'Daily Tasks Affected': ['Forgetting grocery lists', 'Forgetting breakfast menu', 'Unable to turn off electricity or gas'] }, 'Long-term': { 'Symptoms Start Date': '2011', 'Examples': ['Forgetting to lock the tap resulting in water wastage', 'Repetitive actions like putting water in the fridge every 5 minutes'] } }", "Financial management, bank transactions, and social judgment impaired since forgetting bank password in 2012"

Female, 80, Unknown, "Hypertension (+), Diabetes (-), Hyperlipidemia (-)", "Memory impairment with recent onset, confusion regarding time and season, wandering behavior, dream-like stories, difficulty in recognizing important items", "{ 'Short-term': 'Confusion regarding time of day, meal details, and recent events', 'Long-term': 'Difficulty in recalling recent news, son and grandson's names' }", "Decreased comprehension, impaired social judgment, lack of manners"

Unknown, 70, Unknown, "Hypertension, Diabetes", "Memory loss, Difficulty in daily activities, Behavioral changes", "{ 'Short-term': 'Forgets important items like money and bankbooks within a day or two, struggles to remember names of family members', 'Long-term': 'Has trouble recalling events from 5-6 years ago, often forgets daily tasks like taking medication' }", "Impaired understanding, social judgment, and etiquette"

Unknown, Not specified, 무학, "혈압(-), 당뇨(-), 고지혈증(-), 다리 다침(6년 전 일주일 입원)", "기억력저하, 일상생활 지장, 물건을 잊어버리는 경우, 특이한 행동(예: 욕을 함), 약 복용 불규칙", "{ 'Short-term Memory': '물건을 잊어버리는 경우, 냉장고 문을 못 닫은 경우, 물을 갔다 놓고 다시 찾는 경우', 'Long-term Memory': '자식 및 손자 손녀 이름 기억, 최근 뉴스 및 연속극 내용 이해 어려움' }", "이해력 저하, 사회적 판단력 및 예의범절 유지"