

2024 자동차 데이터 분석 경진대회

팀 림수김
김수림, 전주혁

AGEND

01

INTRO

02

문제 해결 전략

03

프롬프트 설계

04

프롬프트 최적화 과정

01 / INTRO

대회 배경

프롬프트 엔지니어링을 통한 자동차 데이터 분류
자동차 산업의 빅데이터 활용 촉진

모델

GPT3.5-turbo-0125

프롬프트 토큰 제한

system + user prompt 기준, 16000 토큰까지

모델 출력 규칙

평가 데이터 40개 샘플들에 대해서 각 행 마다 예측 결과(0 또는 1)만을 출력

$$\text{Score} = 0.8 \times \text{Macro F1 Score} + 0.2 \times \left(1 - \frac{\text{Tokens Used}}{16000}\right)$$



02 / 문제 해결 전략

Rules

1. GPT 모델의 응답이 40개의 행으로 각각 0 또는 1의 답변으로만 구성되지 않는다면 전체 점수 0점 처리
 2. System + User Prompt의 구성 토큰 개수가 16000을 초과하는 경우에는 프롬프트 토큰 점수 0점 처리
 3. LLM의 특성상 동일한 프롬프트 제출물에도 다른 결과(점수)가 도출 될 수 있음
- 제한된 토큰 수 내 지정된 출력 규칙을 준수하여 최대한의 정확도 달성

Rules

1. GPT 모델의 응답이 40개의 행으로 각각 0 또는 1의 답변으로만 구성되지 않는다면 전체 점수 0점 처리
 2. System + User Prompt의 구성 토큰 개수가 16000을 초과하는 경우에는 프롬프트 토큰 점수 0점 처리
 3. LLM의 특성상 동일한 프롬프트 제출물에도 다른 결과(점수)가 도출 될 수 있음
- 제한된 토큰 수 내 지정된 출력 규칙을 준수하여 최대한의 정확도 달성

Strategy 1

제한된 토큰 수 → 'notes' 컬럼 사용 X

Strategy 2

지정된 출력 규칙 & 최대한의 정확도 → 샘플 순차 인덱싱

Strategy 1 'notes' 컬럼 사용 X

Tokens: 18

Tokens: 290

ID	TRAIN_04
lang	en
title	Marine Geophysical and Seismic Data from around the UK (1966 Onwards)
notes	The British Geological Survey hold a collection of data recorded during marine geophysical surveys which includes digital data and analogue records. These data result from approximately 350,000 line kilometres of multi-instrument geophysical survey lines. The data include seismic, sonar, magnetic, gravity, echo sounder...

1. Notes 컬럼은 가장 많은 토큰 수 차지
2. 단어적 의미로 봤을 때 title은 notes의 요약?

→notes는 생략?

Strategy 1 'notes' 컬럼 사용 X

TRAIN 데이터를 기반으로 Sentence-BERT 모델을 이용하여 title과 notes 간의 코사인 유사도 계산

Sentence BERT

BERT를 기반으로 문장 임베딩을 생성하는 방식으로 최적화된 모델
→ 자연어 처리에서 문장 수준의 임베딩을 생성하고 문장 간의 유사도를 효율적으로 계산하는 데 매우 유용한 모델

모두 코사인 유사도 0.5 이상으로 'title'이 'notes'의 내용 요약 검정 완료

→ Notes 생략!

ID	cosine similarity
TRAIN_00	0.7399
TRAIN_01	0.9288
TRAIN_02	0.6995
TRAIN_03	0.6895
TRAIN_04	0.6960
TRAIN_05	0.8015
TRAIN_06	0.9369
TRAIN_07	0.5582
TRAIN_08	0.9009
TRAIN_09	0.6905
TRAIN_10	0.6884

Strategy 2 샘플 순차 인덱싱

문제점

모든 샘플이 동시에 입력될 경우, 모델은 문맥적으로 더 중요한 샘플에 집중하게 되어
샘플과 답변 간의 순서나 연결성이 혼동될 수 있음

전략

각 샘플에 1부터 40까지의 순차적인 인덱스를 추가로 부여하여,
각 샘플을 고유한 정보 단위로 처리하고 모든 샘플이 동일한 중요도를 가지도록 함.
또한 프롬프트에 40개의 출력을 요청했을 때,
인덱스를 1부터 40으로 부여하여 모델이 더 잘 출력할 수 있도록 의도

Prompt

1. TEST_00
2. TEST_01
3. TEST_02
- ...
40. TEST_39

Strategy 2 샘플 순차 인덱싱

기대효과

각 샘플을 숫자 인덱스로 명확하게 구분함으로써 모델이 일관된 구조 유지,
특정 샘플에 대한 편향을 방지하여 출력의 강건성 보장

왜 0이 아닌 1부터 40까지인가?

대부분의 사람들이 인덱스를 1부터 시작하는 방식에 익숙하므로,
1부터 시작하는 것이 모델의 응답이 더 직관적임.
또한, 1부터 40까지 사용함으로써 개수의 명확한 식별이 가능해지며, 출력하는 순서가 강화됨

Prompt

1. TEST_00
2. TEST_01
3. TEST_02
- ...
40. TEST_39

03 / 프롬프트 설계 전략

1. 주어진 문제에 적합한 프롬프트 설계 - 단 한 번의 질문과 답변

SYSTEM

- system 프롬프트는 AI의 역할과 행동 방식을 설정하며, AI의 내부 지침에 해당
- AI의 역할, 작업 방식, 출력 방식 등을 구체적으로 명시하여 모델이 기대하는 결과를 일관되게 제공

USER

- user 프롬프트는 사용자가 원하는 답변이나 작업을 AI에게 요청하는 부분으로
실질적인 질문이나 요구사항 포함
- 사용자가 제공하는 구체적인 데이터 전달

BUT, 이번 대회는 단 한 번의 질문과 대답으로 평가되는 특수한 상황

→ SYSTEM과 USER의 본래 역할이 무의미해짐

1. 주어진 문제에 적합한 프롬프트 설계 - 단 한 번의 질문과 답변

일반적인 USER 프롬프트

USER

You MUST answer 1 if the data is related to

automobiles the 40 input datasets separated by '///'.

Note that there are Deutsch and Korean data.

작업 지시

데이터 포맷

언어 정보

03. 프롬프트 설계 전략

1. 주어진 문제에 적합한 프롬프트 설계 - 단 한 번의 질문과 답변

Self-Attention 메커니즘에 따라 구체적인 작업 요청인 '작업 지시' 부분을 먼저 우선적으로 처리
'데이터 포맷' & '언어 정보' 는 상대적으로 덜 중요한 정보로 처리

일반적인 USER 프롬프트

USER

You MUST answer 1 if the data is related to
automobiles the 40 input datasets separated by '//'.
Note that there are Deutsch and Korean data.

작업 지시 **중요도&우선 순위 ▲**

데이터 포맷 **중요도&우선 순위 ▼**

언어 정보 **중요도&우선 순위 ▼**

03. 프롬프트 설계 전략

1. 주어진 문제에 적합한 프롬프트 설계 - 단 한 번의 질문과 답변

SYSTEM

As a global specialist in "Automotive and transportation-related data",
you have to provide accurate responses for the 40 input datasets separated by '//'.
Note that there are Deutsch and Korean data. Think Step by Step.

언어 정보

모델이 중요한 정보를 먼저 인식하도록

USER에 요청해야 할 구체적인 지시 사항을 SYSTEM에 입력

USER

You MUST answer 1 if the data is related to automobiles

데이터 포맷

2. 주어진 문제에 적합한 프롬프트 설계 - 언어 인식 한계 극복

데이터의 언어가 일관되지 않은 경우 제대로 언어를 인식하지 못함, 정확한 예측에 치명적

Steuerbarer Umsatz aus Lieferungen und Leistungen in Flensburg

German

도로교통공단_고속도로구간별 도로위험도지수정보 조회 서비스

Korean

New registrations of road vehicles by vehicle group and type

English

Problem 1 문맥 기반 처리

GPT는 문맥(context) 기반 입력 데이터 처리 중 다른 언어로 전환되면 정확히 연결 어려움

Problem 2 언어 간 문법 차이

한 언어의 문법에 맞추어 텍스트를 처리하는 중에 언어가 전환되면,
GPT가 그 차이를 감지하지 못하고 잘못된 문법 규칙 적용 가능성 존재

2. 주어진 문제에 적합한 프롬프트 설계 - 언어 인식 한계 극복

데이터의 언어가 일관되지 않은 경우 제대로 언어를 인식하지 못함, 정확한 예측에 치명적

구분자 “//” 도입

Problem 1

문맥 기반 처리

GPT는 문맥(context) 기반 입력 데이터 처리 중 다른 언어로 전환되면 정확히 연결 어려움

Problem 2

언어 간 문법 차이

한 언어의 문법에 맞추어 텍스트를 처리하는 중에 언어가 전환되면,
GPT가 그 차이를 감지하지 못하고 잘못된 문법 규칙 적용 가능성 존재

2. 주어진 문제에 적합한 프롬프트 설계 - 언어 인식 한계 극복

언어의 경계를 명확히 구분할 수 있는 구분자 “//” 사용
GPT가 언어 전환을 인식하여 각 언어에 맞는 적절한 처리 방식 적용

Prompt

Steuerbarer Umsatz aus Lieferungen und Leistungen in Flensburg //
도로교통공단_고속도로구간별 도로위험도지수정보 조회 서비스 //
New registrations of road vehicles by vehicle group and type //

...

Problem 1 문맥 기반 처리

GPT는 문맥(context) 기반 입력 데이터 처리 한 문장에서 다른 언어로 전환되면 정확히 연결 어려움

Solution 1 명확한 언어 경계 설정

- 구분자 “//”를 통해 언어 전환이 일어나는 지점 명확히 알려줌
- 한 언어의 처리가 끝났음을 인식, 새로운 언어의 처리 시작

Solution 2 문맥 혼동 방지

- 구분자가 없다면 GPT는 이전 언어의 문맥을 이어서 해석
- 이전 언어의 문맥이 새 언어에 영향을 미치지 않도록 방지

2. 주어진 문제에 적합한 프롬프트 설계 - 언어 인식 한계 극복

언어의 경계를 명확히 구분할 수 있는 구분자 “//” 사용
GPT가 언어 전환을 인식하여 각 언어에 맞는 적절한 처리 방식 적용

Prompt

Steuerbarer Umsatz aus Lieferungen und Leistungen in Flensburg //
도로교통공단_고속도로구간별 도로위험도지수정보 조회 서비스 //
New registrations of road vehicles by vehicle group and type //

...

Problem 2 언어 간 문법 차이

한 언어의 문법에 맞추어 텍스트를 처리하는 중 언어가 전환되면 차이를 감지하지 못하고
잘못된 문법 규칙 적용 가능성 존재

Solution 1 언어별 최적화 처리 방식 적용

- 구분자 “//”를 사용하여 각 샘플을 독립된 단위로 처리
- 언어별로 적절한 전처리와 분석 수행

Solution 2 언어 인식의 오류 최소화

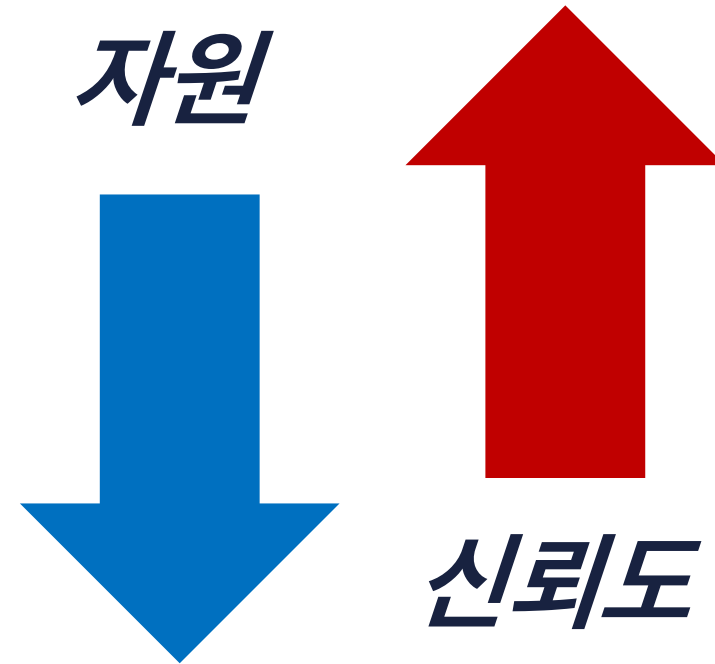
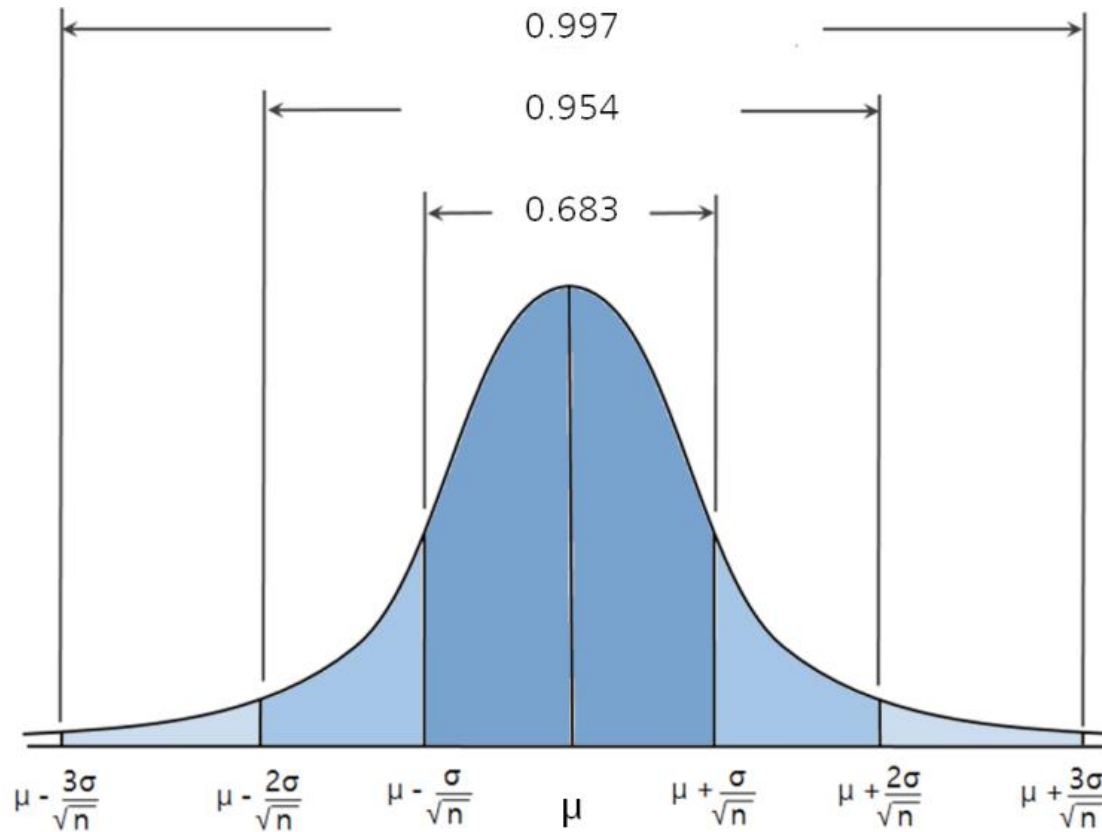
언어 모델의 혼동 문제를 해결하여 언어 전환을 명확히 인식하고 처리

04 / 프롬프트 최적화 과정

04. 프롬프트 최적화 과정

1. 프롬프트 최적화를 위한 수정 및 개선 과정 - 프롬프트 확인 자동화

하나의 프롬프트를 30번 강건성을 확인하는 작업,
30번 반복은 중심극한정리에 기반하여, 최소한의 자원 사용으로 최대한의 신뢰도 확보



04. 프롬프트 최적화 과정

1. 프롬프트 최적화를 위한 수정 및 개선 과정 - 프롬프트 확인 자동화

강건성 확인

```
-----22-----  
Macro F1 Score: 0.949874686716792  
토큰 점수: 0.9240625  
최종 점수: 0.94471  
  
-----23-----  
Macro F1 Score: 0.949874686716792  
토큰 점수: 0.9240625  
최종 점수: 0.94471  
  
-----24-----  
Macro F1 Score: 0.949874686716792  
토큰 점수: 0.9240625  
최종 점수: 0.94471  
  
-----25-----  
Macro F1 Score: 0.949874686716792  
토큰 점수: 0.9240625  
최종 점수: 0.94471  
  
-----26-----  
Macro F1 Score: 0.949874686716792  
토큰 점수: 0.9240625  
최종 점수: 0.94471
```

≈

최종 점수

팀	PUBLIC	점수
림수김		0.94467

팀	PRIVATE	최종점수
림수김		0.94457

2. 불필요한 토큰을 줄이기 위한 과정 - feature selection

DATA

- test.csv: 40개의 평가 데이터 샘플
- **ID** : 샘플 별 식별 ID
- lang : 해당 데이터 샘플의 주된 구성 언어
- **title** : 해당 데이터 샘플의 제목
- notes : 해당 데이터 샘플의 정보, 내용

ID

Prompt

Title

1. TEST_00	Beach Profile Data Collected from ...
2. TEST_01	Nota media de la nota de admission ...
3. TEST_02	Internet-based platform services 2019 ...
4. TEST_03	경상남도 김해시_자동차등록 현황
5. TEST_04	EV用充電設備の設置状況データ一覧
6. TEST_05	경기아트센터 장애인기업 생산제품 구매 현황
7. TEST_06	Autohaus ...

Notes 피처를 사용하지 않는 전략을 기반으로

ID와 title 피처만 선택

2. 불필요한 토큰을 줄이기 위한 과정 - 개조식 프롬프트

Prompt

Task Description

...

Output Rules

...

Datasets

개조식 프롬프트 구축

간결한 표현과 필요한 정보만 집중하여 토큰 절약
명확한 구조와 일관성을 유지하여 정확도 상승

04. 최종 프롬프트

```
## FINAL
```

```
system = ""As a global specialist in "Automotive and transportation-related data", you have to provide accurate responses for the 40 input datasets separated by '//'. Note that there are Deutsch and Korean data. Think Step by Step""
```

Point 1

```
user = ""
```

```
### Task Description ###
```

Point 2

```
You MUST translate Deutsch and Korean to English.
```

```
You MUST answer 1 if the data is related to automobiles and traffic, you MUST answer 0 if not, independently.
```

```
### Output Rules ###
```

```
1. Only output numbers separated by rows. Do not include text such as 'TEST_00', 'TEST_01', etc.
```

```
2. Output one number per line, and ensure exactly 40 numbers are output. If there are fewer or more than 40, it is incorrect.
```

```
EX)
```

```
0
```

```
0
```

```
.
```

```
.
```


```
1
```

Point 3

```
### Datasets ###
```

```
1. TEST_00: Beach Profile Data Collected from Madeira Beach, Florida (January 15, 2021) //
```

```
2. TEST_01: Nota media de la nota de admisión (cohorte de nuevo ingreso en el SUE) por forma de admisión, ámbito de estudio y s
```

A photograph of a busy New York City street, likely Times Square, featuring yellow taxis, tall buildings, and traffic lights. The image is darkened, and the Korean text "감사합니다" is overlaid in the center.

감사합니다