2024.11.15

산학협력프로젝트2 발표

성균관대학교 소프트웨어학과 박소정, 박윤지, 이우린

목차

OCR

개요

중간발표

100개의 수학 문제 중
38개의 문제에서 에러 발생
→ 이중에서 16문제가 OCR 관련 문제

1. 새로운 OCR 사용

중간 발표 이후로 추천받은 OCR 사용(surya)
→ tesseract에 비해 눈에 띄는 장점이 없다고 판단하여 기존 코드를 개선

2. latex-문제 분리

2010개의 데이터를 새로 받은 이후,
600개의 문제를 테스트해보았을 때 latex-문제를 분리하는 과정에서
127개의 문제에서 latex-문제가 분리가 안 됨

3. 새로운 문제 유형

기존에는 없던 문제 유형이 생김(범분수 등)

OCR

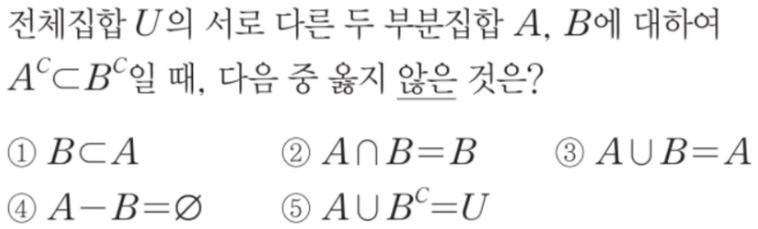
유형별 에러 설명

항목	설명	문제 수
특수기호	다양한 특수기호 관련 에러 - subset 기호를 subseteq 기호로 인식 - %를 읽지 못함	11
부호	극한 문제에서 양의 무한대/음의 무한대와 같은 작은 부호를 인식하지 못함	11
선분	latex 기호 \overline을 읽지 못함	9
분수	분수를 인지하지 못함 ex. frac{e}{2}의 경우 e2로 인식	4
소수	소수점 아래에 세 자리수 이상 나오게 되면 소수점을 인식하지 못함 ex. 0.0012 → 00012	2
루트	루트의 범위를 잘못 설정 ex. sqrt{ax^2+bx}+c를 sqrt{ax^2+bx+c}로 인식	1

OCR

에러 예시

Input



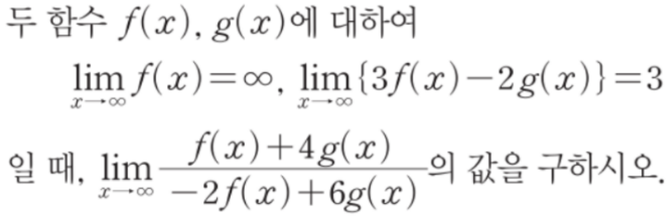
Output

{
"File": "/content/drive/MyDrive/2024_project/img/img/question_40.png",
"Question": "문제]\n전체집합 Latex_1]의 서로 다른 두 부분집합 Latex_2]에 대하여 Latex_3]일 때, 다음 중 옳지 않은 것은?\n\n선지]\n1. Latex_4]\n2. Latex_5]\n3. Latex_6]\n4. Latex_7]\n5. Latex_8]",
"Latex_1": "\\(U \\)",
"Latex_2": "\\(A, B \\)",
"Latex_3": "\\(A^C \\subseteq B^C \\)",
"Latex_4": "\\(B \\subseteq A \\)",
"Latex_5": "\\(A \\cap B = B \\)",
"Latex_6": "\\(A \\cup B = A \\)",
"Latex_7": "\\(A - B = \\emptyset \\)",
"Latex_8": "\\(A \\cup B^C = U \\)"
},

OCR

에러 예시

Input



Output

{
"File": "/content/drive/MyDrive/2024_project/img/img/question_928.png",
"Question": "[문제]\n두 함수 [Latex_1]에 대하여\n\n\\[\n\\lim_{{x \\to -\\infty}} f(x) = \\infty, \\quad \\lim_{{x \\to -\\infty}} \\{3f(x) - 2g(x)\\} = 3\n\\]\n\n일 때, [Latex_2]의 값을 구하시오.",
"Latex_1": "\\(f(x), g(x) \\)",
"Latex_2": "\\(\\lim_{{x \\to -\\infty}} \\frac{f(x) + 4g(x)}{-2f(x) + 6g(x)}\\)"
},

Parsing

개요

1. 분수 분리 → 분수 및 부등식 분리로 확대(제일 안됐던 유형들)

2. 분리한 다음, 각 유형 별로 전용 프롬프트를 통과.

3. 파싱 단계는 현재는 오류가 없이 동작하고 있으나,
계속 취약점을 찾으면서 보완중에 있음.


```
{
  "File": "./img/question_1432.png",
  "Question": "문제\n삼각형 [Latex_1]의 세 내각의 크기를 각각 [Latex_
을 있는 대로 고르시오\n\n보기\n\nㄱ [Latex_3]\n\nㄴ [Latex_4]\n\nㄷ [
[Latex_7]는 예각삼각형이다",
  "Latex_1": "\\( ABC \\)",
  "Latex_2": "\\( A, B, C \\)",
  "Latex_3": "\\(\\cos \\frac{A}{2} = \\sin \\left( \\frac{B+C}{2} \\right)\\)",
  "Latex_4": "\\(\\tan (B+C) = -\\frac{1}{\\tan A}\\)",
  "Latex_5": "\\(\\tan A + \\tan (B+C) = 0\\)",
  "Latex_6": "\\(\\cos (B+C) > 0\\)",
  "Latex_7": "\\( ABC \\)"
}
```

```
{
  "Latex_1": {
    "Origin": "\\( ABC \\)"
  },
  "Latex_2": {
    "Origin": "\\( A, B, C \\)"
  },
  "Latex_3": {
    "Origin": "\\(\\cos [FRAC 1] = \\sin \\left( [FRAC 2] \\right)\\)",
    "Frac_list": {
      "[FRAC 1]": "\\frac{A}{2}",
      "[FRAC 2]": "\\frac{B+C}{2}"
    }
```

Parsing

step2

분리된 식들은 각각의 전용 프롬프트를 통해 음독

분수
- Original : '\\frac{743}{234}'
- Components : '743', '234'
- Rearranged : '234', '743
- Translated : '이백삼십사', '칠백사십삼'
- Merged : 이백삼십사 분의 칠백사십삼

부등식
- Original : '-2 \\ leq x \\leq [FRAC 2]'
- Components : '-2', '\\leq', 'x', '\\leq', '[FRAC 2]'
- Translated : '마이너스 이', '\\leq', '엑스', '\\leq', '[FRAC 2]'
- Merged : 마이너스 이는 엑스보다 작거나 같고, 엑스는 [FRAC 2]보다 작거나 같다.

RAG

사용이유

1. 실제로 오류가 발생한 문제들을 볼 때 특정유형에서 많이 나왔다기 보단 여러유형에서 골고루 나왔음.

2. 모든 오류에 대해 프롬프트 Few-Shot을 넣는것을 비효율적이기에, 특정한 예시를 적재적소에서 가져올 수 있는 RAG를 활용하고자 함

RAG

과정

1. excel 파일로 에러 few-shot db 구축
   db 에 총 70개 에러 퓨샷 (라텍스 input > 음독 output) 존재

2. 문제 에서 각 라텍스랑 db input이랑 유사도 계산하여
   코사인 유사도 0.8 이상의 context (input + output) 출력

3. context 참고하여 question에 대한 음독을 실행하도록 프롬프트 구성

A	B
input	output
"1": "₩₩(₩n[FRAC 1] = [FRAC 2] + [FRAC 3] + [FRAC 4] + ₩₩cd	"Latex_1": "₩₩(₩n칠 분의 육은 십 분의 엑스일 더하기 십의 제곱 분의
"1": "₩₩(₩₩left(-x^2 y^3 ₩₩right)^4 = x^8 y^{12}₩₩)" }	"Latex_1": "괄호 열고 마이너스 엑스제곱 와이세제곱 괄호 닫고의 사제
"1": "₩₩(₩₩left(-2a^2 ₩₩right)^3 = -8a^6₩₩)" }	"Latex_2": "괄호 열고 마이너스 이에이제곱 괄호 닫고의 세제곱은 마이
"1": "₩₩(₩₩left(3a^2 b^3 ₩₩right)^2 = 3a^4 b^6₩₩)" }	"Latex_3": "괄호 열고 삼에이제곱 비세제곱 괄호 닫고의 제곱은 삼에이
"1": "₩₩(₩₩left(-x^2 y z^3 ₩₩right)^5 = -x^{10} y^5 z^{15}₩₩	"Latex_4": "괄호 열고 마이너스 엑스제곱 와이 제트의 세제곱 괄호 닫고
"1": "₩₩(₩₩left([FRAC 1] x y^2 ₩₩right)^3 = [FRAC 2] x^3 y^6	"Latex_5": "괄호 열고 오 분의 일 엑스 와이의 제곱 괄호 닫고의 세제곱
"1": "₩₩(₩n₩₩left([FRAC 1] ₩₩right)^3 = -[FRAC 2]₩₩)",	"Latex_1": "괄호 열고 엑스세제곱 와이제곱 분의 에이제곱제곱 괄호 닫
"1": "₩₩(₩₩langle x ₩₩rangle₩₩)" }	"Latex_1": "특수기호 엑스"
"1": "₩₩(x₩₩)" }	"Latex_2": "엑스"
"1": "₩₩(₩₩langle x ₩₩rangle^2 - ₩₩langle x ₩₩rangle - 6 = (	"Latex_3": "특수기호의 제곱 빼기 특수기호 엑스 빼기 육은 영"
"1": "₩₩(x₩₩)" }	"Latex_4": "엑스"
"1": "₩₩(N(m, ₩₩sigma^2)₩₩)" }	"Latex_1": "엔 괄호 엠 컴마 시그마제곱 괄호"
"1": "₩₩(P(|Z| ₩₩leq 1.96) = 0.95₩₩)" }	"Latex_10": "피 괄호 절대값 제트는 일 점 구육보다 작거나 같음 괄호는
"1": "₩₩(P(|Z| ₩₩leq 2.58) = 0.99₩₩)" }	"Latex_11": "피 괄호 절대값 제트는 이 점 오팔보다 작거나 같음 괄호는
"1": "₩₩(₩n₩₩lim_{{x ₩₩to 0}} [FRAC 1] ₩n₩₩)", "frac_list"	"Latex_1": "리미트 엑스가 영으로 갈 때 루트 엑스 더하기 사 빼기 이 분
"1": "₩₩(₩₩overline{OC_1} : ₩₩overline{OD_1} = 3 : 4₩₩)" }	"Latex_11": "선분 오씨일 대 선분 오디일은 삼 대 사"
"1": "₩₩(₩₩overline{P_1Q_1} = ₩₩overline{A_1Q_1}₩₩)" }	"Latex_14": "선분 피일큐일은 선분 에이일큐일"
"1": "₩₩(₩₩lim_{{n ₩₩to ₩₩infty}} S_n₩₩)" }	"Latex_35": "리미트 엔이 무한대로 갈 때 에스엔"
"1": "₩₩(₩₩overline{PB} = ₩₩overline{QC}₩₩)" }	"Latex_9": "선분 피비는 선분 큐씨"
"1": "₩₩(₩n₩₩overline{AP} : ₩₩overline{PP'} = 5 : 6, ₩₩quad	"Latex_13": "선분 에이피 대 선분 피피 프라임은 오 대 육, 그리고 선분

```
# ChatPromptTemplate 설정
template = """
Context: {context}
Question: {question}
Answer: 당신은 LateX를 한국어로 음독하는 역할을 수행합니다.
위의 Context를 참고하여 Question을 올바르게 한글로 음독해 주세요.
context는 참고만 하고 음독하지 말고, question만을 올바르게 한글로 음독해 주세요.
모든 영어 및 수식은 한국어로 음독이 되도록 한글로 음독해 주세요.
"""
```

RAG

보완점

함수 $y=3\tan\left(2x+\frac{\pi}{2}\right)+1$에 대한 다음 설명 중 (가)~(라)에 알맞은 것을 구하시오.

- 주기는 (가)이다.
- $y=3\tan 2x$의 그래프를 x축의 방향으로 (나)만큼, y축의 방향으로 (다)만큼 평행이동한 것이다.
- 점근선의 방정식은 (라)이다.

1. 분수, 부등식의 역할 분리

파싱 파트인 분수, 부등식은 완전히 역할 분리 필요
> 프롬프트 퓨샷으로 해결

2. 같은 짧은 문자열의 유사도 비교

A, (가) 와 같은 짧은 문자열의 유사도 비교 어려움 (좌측 그림 참고)
> 검색에서 배제 시키는 방향으로

Merge

개요

1. 분리한 분수, 부등식 및 latex를 원래 자리에 재배치하는 과정.

2. 먼저 분리한 분수, 부등식을 원래 latex에 집어 넣고, 그 latex를 다시 문제에 집어넣는다.

3. 단순히 삽입하기만 하면 어색할 수 있기에, 결과물을 LLM에 넣어서 부드럽게 하여 최종 텍스트를 도출

Merge

예시

어느 대학에 입학한 신입생들의 수능 점수는 평균이 400점, 표준편차가 80점인 정규분포를 따른다고 한다. 이 대학 신입생 중 임의추출한 64명의 수능 점수의 평균을 $\overline{X}$라 할 때, $\mathrm{P}(\overline{X} \geq k) = 0.3085$를 만족시키는 상수 k의 값을 위의 표준정규분포표를 이용하여 구하시오.

z	$\mathrm{P}(0 \leq Z \leq z)$
0.5	0.1915
1.0	0.3413
1.5	0.4332


```
{
    "Latex_1": {
        "Origin": "엑스 바"
    },
    "Latex_2": {
        "Origin": "[엑스 바는 케이보다 크거나 같다.]일 때 피는 영 점 삼공팔오",
        "Inequal_list": {
            "[INEQUAL 1]": "엑스 바는 케이보다 크거나 같다."
        }
    },
    "Latex_3": {
        "Origin": "케이"
    }
}
```



```
{
    "Latex_1": {
        "Origin": "엑스 바"
    },
    "Latex_2": {
        "Origin": "[INEQUAL 1]일 때 피는 영 점 삼공팔오",
        "Inequal_list": {
            "[INEQUAL 1]": "엑스 바는 케이보다 크거나 같다."
        }
    },
    "Latex_3": {
        "Origin": "케이"
    }
}
```


어느 대학에 입학한 신입생들의 수능 점수는 평균이 400점 표준편차가 80점인 정규분포를 따른다고 한다 이 대학 신입생 중 임의추출한 64명의 수능 점수의 평균을 [Latex_1]라 할 때 [Latex_2]를 만족시키는 상수 [Latex_3]의 값을 위의 표준정규분포표를 이용하여 구하시오

어느 대학에 입학한 신입생들의 수능 점수는 평균이 400점, 표준편차가 80점인 정규분포를 따른다고 합니다. 이 대학 신입생 중 임의로 추출한 64명의 수능 점수의 평균을 엑스 바라 할 때, 엑스 바가 케이보다 크거나 같을 때 피가 영 점 삼공팔오를 만족시키는 상수 케이의 값을 위의 표준정규분포표를 이용하여 구하시오.

Merge

문제점

1. 현재 Merge 단계에서는 AI가 문제를 요약을 해버리거나, 문제의 답을 계산해서 내놓는 등의 문제가 발생하고 있음.

2. 이를 해결하기 위해 문제 Image 또는 OCR결과를 같이 넣어서, input을 오류를 줄이려는 방향으로 진행 중에 있음.

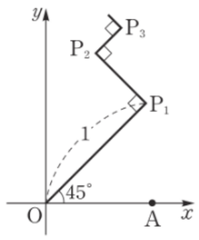
오른쪽 그림과 같이 점 P_n이

$\overline{OP_1}=1$, $\overline{P_1P_2}=\frac{1}{2}\overline{OP_1}$,

$\overline{P_2P_3}=\frac{1}{2}\overline{P_1P_2}$, …,

$\angle AOP_1=45^\circ$,

$\angle OP_1P_2=\angle P_1P_2P_3=\cdots=90^\circ$



를 만족시킬 때, 점 P_n이 한없이 가까워지는 점 (a, b)에 대하여 $\frac{b}{a}$의 값을 구하시오.

(단, O는 원점이고 A는 x축 위의 점이다.)

문제에서 주어진 조건들을 고려하여 점 피 엔이 한없이 가까워지는 점 (에이, 비)에 대해 에이 분의 비의 값을 구하시오. 단, 오가 원점이고 에이는 엑스축 위의 점입니다.