효율적 복습을 위한 모의고사 자동제작 서비스의 UX 분석

자연어처리와 교육 기말발표
심리학과 송상록

목차

① 선행연구 및 프로젝트 동기

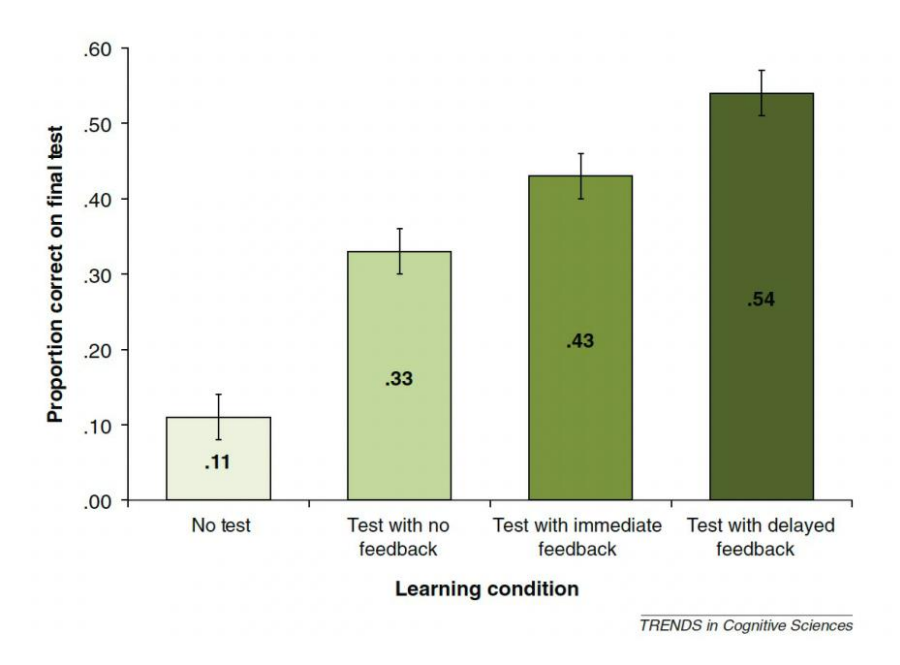
② 프로젝트 태스크 정의

③ 프로젝트 진행과정

④ 프로젝트 결과

1. 선행연구 및 프로젝트 동기

- **testing effect**: 학습한 내용을 잘 기억하는지 스스로 점검(test)하는 것은 암기에 유용하다
- Rowland(2014): 단순히 학습 내용을 다시 읽는 것(restudy)보다, test가 학습에 더 도움이 된다
- Butler et al.,(2008): test 이후 **delayed feedback**이 주어지면 학습 효과가 더 커진다



- 객관식 모의시험을 통해 testing effect를 알아본 연구
- Delayed feedback 조건의 최종시험 성적이 제일 높았음
  - no test: 모의시험을 치르지 않음
  - no feedback: 문제풀이만 하고 답은 알려주지 않음
  - immediate feedback: 한 문제를 푼 직후 바로 답을 알려줌
  - delayed feedback: **모든 문제를 다 푼 이후** 답을 알려줌

1. 선행연구 및 프로젝트 동기

- OpenAI사의 GPT 등 LLM(거대언어모형)은, 주어진 **text input** 및 지시된 **prompt**를 바탕으로 학습자를 위한 문제를 생성할 수 있다
- Lee et al.(2023): ChatGPT에 어떤 prompt를 지시해야 적절한 문제가 생성되는지를 다룬 연구
- 풀 문제가 부족해서 testing을 통한 복습이 어려운 학습자들에게 좋은 대안이 될 수 있음

(예: 별도의 평가문제집이 없는 대학 수업, 선택자 수가 부족한 고등학교 선택과목 등)

Example 1

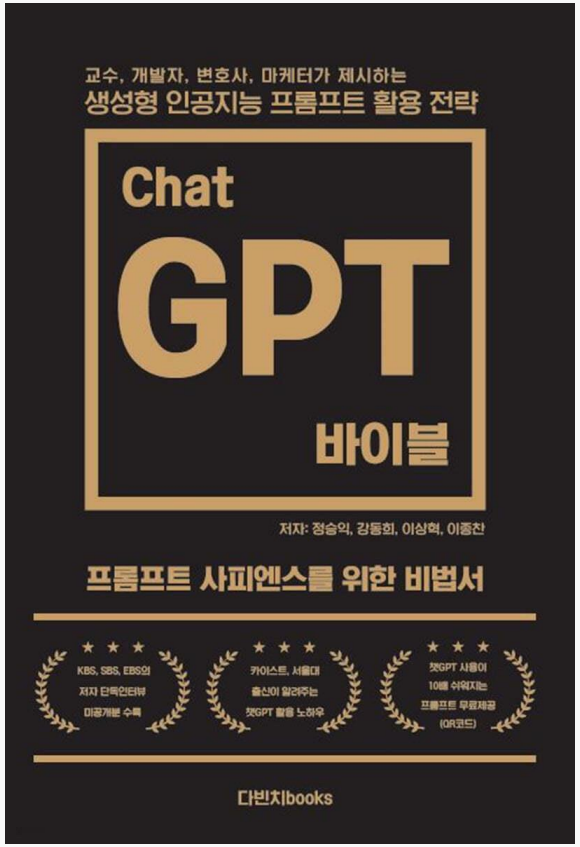
A1	Identify / y-n, alternative, t-f / multiple-choice
Prompt	Passage / make a yes–no question of identifying information that is explicitly shown in the text. After questions, give 'yes or no' choice option
Question	Is the public's reliance on rapid, intuitive, affect-driven sources of information processing a result of their limited capacity to comprehend science? Yes or no: No

Example 2

A1	Identify / y-n, alternative, t-f / multiple-choice
Prompt	Passage / Make a multiple choice, 'wh' question asking you to choose what is the topic of the passage
Question	What is the main topic discussed in the passage? A) The public's limited capacity to comprehend science B) The Earth's orbit around the Sun C) Nitrogen in the atmosphere D) Two-by-two contingency tables for medical testing E) Heuristics and polarization in the public's thinking about science-related issues

1. 선행연구 및 프로젝트 동기

- 현재 AI를 이용한 문제생성 연구는, ChatGPT를 비롯한 **대화형 인공지능** 위주로 진행됨
  - 대화가 이루어지면 시험장에서의 **현장감**을 느끼기 어려움
  - 문제를 1개씩 풀게 되고, 바로 정답을 알게 되므로 delayed feedback보다 효과가 덜한 **immediate feedback**을 받게 됨
  - ChatGPT에게 **어떤 prompt를 지시해야** 좋은 문제를 만들어 줄지 판단하기 어려움
- 해결책: LLM을 이용해 **모의시험**을 제작하는 웹사이트 제작



"GPT가 좋은 건 알겠는데 어떻게 써야 할지 모르겠어요"

2. 프로젝트 태스크 정의

- 본 웹사이트는 다음 3가지 기능을 수행할 수 있어야 함

① 수업 내용이 담긴 텍스트를 입력받고, **자동으로 객관식 문제를 생성**함 (LLM의 효과)

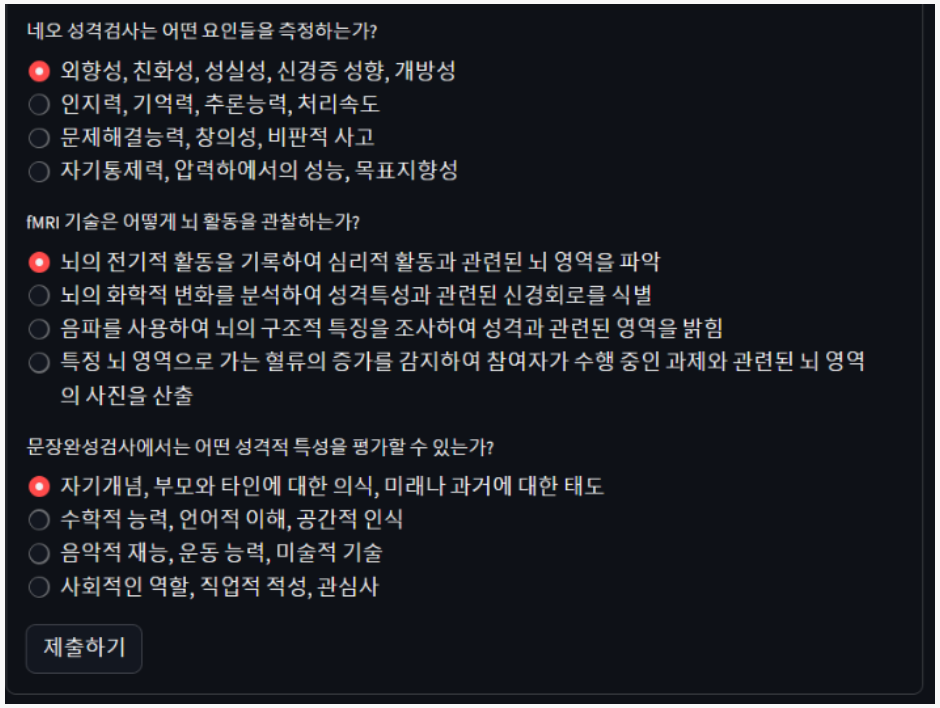
② 객관식 문제를 웹사이트상에서 **모의고사 형태**로 풀 수 있게 함 (testing)

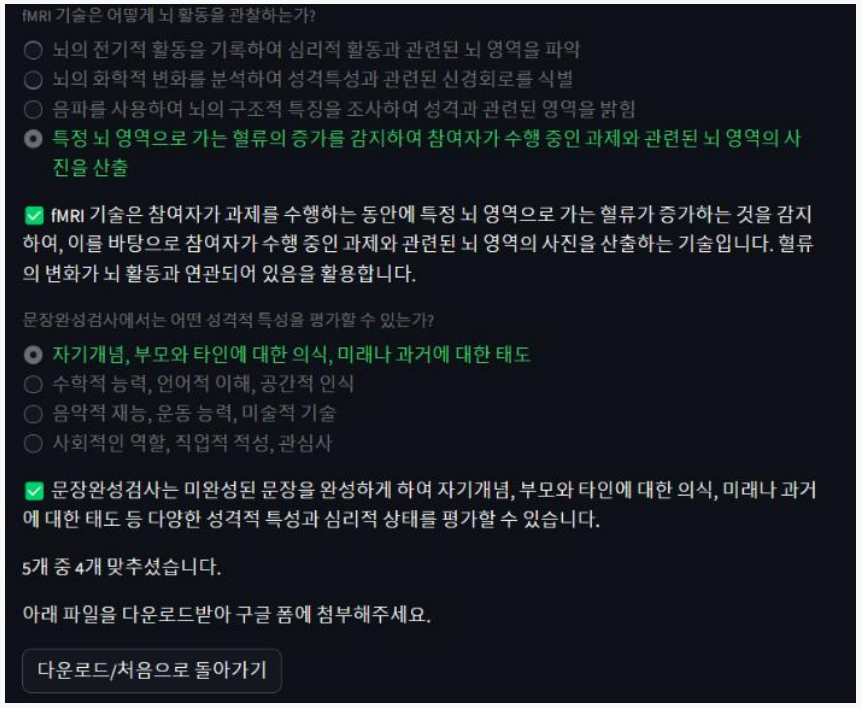
③ 문제풀이 종료 후, **채점 및 구체적 풀이**를 제공함 (delayed feedback)

- 연구 질문: LLM을 이용한 모의고사 자동제작 서비스가 제공하는 UX(사용자경험)을 분석한다.
  - **인지과학적 측면:** 본 서비스는 효율적 복습에 도움이 되는가?
  - **컴퓨터공학적 측면:** 학습효과 증진을 위해, 본 서비스에 어떠한 개선점이 필요한가?
  - **교육학적 측면:** 본 서비스는 교육 분야에 어떠한 변화를 줄 것인가?

3. 프로젝트 진행과정

- Langchain 및 OpenAI의 GPT-4 API: Prompt Engineering을 통해 객관식 문제 자동생성 기능 구현
- Streamlit: 웹사이트 제작 및 배포





소스코드: github.com/freud-sensei/test_maker

3. 프로젝트 진행과정

- Langchain을 통해 OpenAI의 GPT-4에 적절한 명령을 내려 웹사이트 구현

```
class Questionmaker(BaseModel):
    questions: List[SingleQuestion] = Field(description="List of questions")

template = '''
You are a study assistant which must generate unique {num_questions} multiple-choice questions for
students.\n
By using [INPUT_DATA], make multiple choice questions in a structured format.\n
You must make one correct answer, and a maximum of four incorrect answers.\n
The language must be in {language}.\n
Tips: You must return {num_questions} questions, not more and not less.

[INPUT_DATA]:\n
{input_data}
'''
```

prompt (input값을 이용해 문제를 생성해줘)

GPT-4의 문제생성

output schema (딕셔너리와 유사)

```
class SingleQuestion(BaseModel):
    question: str = Field(description="Multiple choice questions")
    correct: str = Field(description = "The correct answer")
    incorrect: List[str] = Field(description = "List of incorrect answers")
    explanation: str = Field(description="Explanation on why the answer is correct for each
    question")
```

3. 프로젝트 진행과정

- 04.03 ~ 04.14: 대학생 10명을 대상으로 **사용성 평가** (실험 아님) 진행

(영어교육, 중국어교육, 윤리교육, 물리교육, 화학교육, 심리학, 지리학, 연합전공 정보문화학, 의학(2명))

- 본인 전공과목의 학습자료 1개를 입력해 생성된 문제를 풀어본 뒤, 아래의 세 질문에 응답
- 학습자료는 1주 분량의 강의자료/교재 혹은 논문 1편으로 설정

<사용성평가 참여자 모집>

안녕하세요! 서울대학교 [자연어처리와 교육] 수업 프로젝트로 "효율적 복습을 위한 모의고사 자동제작 서비스 제작"을 진행중인 심리학과 송상록입니다.

본 서비스는 여러분의 학습자료를 입력받아 객관식 문제를 자동으로 생성합니다. 서비스를 사용해 보시고 후기를 남겨주실 분들을 모집 중입니다.

세부 사항은 아래 사진을 참고하시길 바라며, 참여 희망자는 https://open.kakao.com/o/s6OTNvjg 로 연락 바랍니다.

3. 프로젝트 진행과정

- 객관적 평가 (실험 연구)가 아닌 주관적 평가 (사용성 평가)를 선택한 이유?
- **피실험자 모집의 어려움**: 10명도 간신히 모은 것임. 적은 표본 대상으로는 무의미한 통계분석을 하는 것보단, 질적 분석을 진행하는 것이 더 의미있을 것.
- **자유로운 학습자료 선택**: 본 프로젝트에선 참여자들이 자신이 원하는 학습자료를 선택할 수 있고, 각자 다른 문제를 풀게 됨. 이에 따라 평균을 이용한 객관적 비교는 어려움.
- **사용자경험 분석**: 사용자가 서비스를 사용하면서 느꼈던 장단점은 단순한 점수 평가로 분석할 수 있는 것이 아님. 인공지능의 실제 성능과, 사용자가 체감하는 성능 간엔 차이가 존재함.

4. 프로젝트 결과

(인지과학적 측면) 인공지능이 제작한 문제를 푸는 것이 학습에 도움이 된다고 생각하시나요?

얼마나 잘 문제를 만들어낼지 궁금했는데, 생각보다 자세하게 문제를 만들고 해설도 좋아서 놀랐습니다. 특히 간단한 난이도의 문제만 만들어낼 줄 알고 쉽게 풀었는데, 나름의 실수를 유도하는 문제들도 섞여 있어 문제의 질과 양 모두를 보장할 수 있는 프로그램이라는 생각이 들었습니다.

(보급해주시면 안될까요 ..? ㅎㅎㅎ)

전반적으로 핵심 내용을 선별하여 문제를 제작하여 주었기 때문에 도움이 된다고 생각한다. 특히 암기 과목의 경우 어떤 부분의 내용을 제대로 암기 하였고 어떤 부분에 대해서 보충이 필요한지 알 수 있기 때문에 해당 측면에서 학습에 큰 도움을 줄 수 있을 것이다.

단순한 사실 관계를 파악하는 문제를 풀 땐 인공지능이 제작한 문제를 푸는 것이 학습에 큰 도움이 된다고 생각한다. 특히, 학습을 할 때 연습문제 등을 풀어봄으로써 자신이 알고 있는 지식을 확인하고, 오개념이 있는지 파악할 수 있기에 이러한 부분에서 인공지능이 제작한 문제를 활용하면 학습에 도움이 될 것이다.

- 학습자료의 내용 중 핵심내용이 무엇인지 이해할 수 있었음
- 어떤 부분을 정확히 암기했고, 보충이 필요하거나 오개념이 있는 부분을 파악할 수 있음
- 만약 실험연구가 가능했다면, 단순 복습과 AI 제작 문제를 이용한 testing 간 학습 효과를 비교하는 것도 의미있었을 것

4. 프로젝트 결과

(컴퓨터공학적 측면) 본 서비스를 개선하기 위해 어떤 점을 추가하거나 변경하고 싶으신가요?

시험 전 테스트를 위해 개인적으로 사용하는 학생들이 생길 것 같습니다. 또한 학원 등지에서 학습지/교재 제작 시 사용하는 등 상업적으로도 가치가 있을 것 같습니다.(개인적으로 영어 교과서와 수특을 바탕으로 단어문제, 빈칸 뚫기, 순서 문제 등을 양산형으로 만들어내 제공하던 서비스가 떠오르네요.) 다만 이 서비스를 통해 만들어진 문제들은 사실 확인 정도에서 그치기 때문에 암기가 최종 목표가 아닌 과목들을 공부하는 데에 있어서는 무리가 있을 것 같습니다.

선지에 대한 개선이 필요하다고 생각이 된다. 문제를 풀 때 정답 선지만큼 오답 선지 역시 학습에 매우 필수적인 요소 중 하나인데, 관련 서비스의 경우 일부 문제에서 오답 선지와 정답 선지가 지나치게 구분할 수 있었다. (해당 내용을 정확히 기억하지 못해 정답을 모른다고 하더라도 나머지 오답이 지나치게 말이 안 되거나 추상적이기 때문에 역으로 정답을 찾게 되는 현상이 발생하였음)

답안을 채점하고 오답이 나올 때 바로 답이 나오는 것이 아니라 설정에 따라 어떤 문제가 틀렸는지만 표시해주고 다시 풀 수 있는 기회가 있었으면 좋겠습니다. 또는 틀린 문제의 개념에 대해서 다시 새로운 문제를 생성해주는 것도 좋을 것 같습니다.

- 논리적 추론이 필요한 문제를 생성할 수 있게끔, prompt engineering 연구를 진행해야 함
- 정오답 선지를 구별하기 어렵게 하거나, 새로운 문제를 통해 틀린 문제를 복습할 수 있는 기능 등,

개발자 역시 testing effect를 강화할 수 있는 기능을 고민해 보아야 함

4. 프로젝트 결과

(교육학적 측면) 본 서비스가 보급된다면, 교육 분야에 어떠한 변화가 생길 것 같나요?

긍정적 : 학생들이 굳이 문제집을 구매하지 않고도 다양한 문제에 대해 접할 수 있게 될 것 같습니다.
부정적 : 업로드 한 pdf에 오류가 있을 경우 오류의 고착화가 심해질 수 있다고 생각합니다. 불법 복제 pdf 사용이 늘어날 수 있다고 생각합니다.

긍정적: 기사 등 문제은행 출제방식으로 진행되는 시험의 경우 새로운 문제를 제작하는데 유용할 것입니다. 동시에 입력된 자료가 오류투성이라면 오류가 가득한 문제가 만들어지기 때문에,정확한 교재를 만들기 위하여 연구자들이 더 노력할 것 같아 학문의 발전이 기대됩니다.

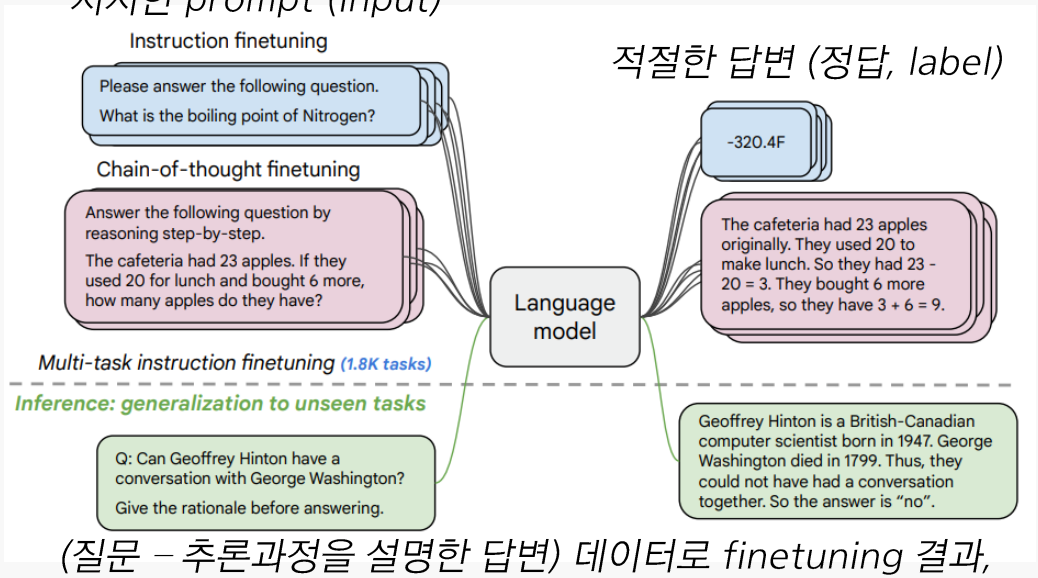
단순하게 사실을 암기해야 하는 학습을 할 때 본 서비스를 활용할 수 있을 것 같은데, 기존에는 인간이 직접 문제를 출제했다면, 앞으로는 인공지능이 저차원 수준의 문제를 출제할 수 있을 것이다.
이에 따라, 교육 전문가들은 문제 출제 분야에 힘쓰기 보다는 인공지능이 출제한 문제를 검수하는 분야에 힘을 쓰게 될 것이다.

- 기존 문제집 및 사교육으로 발생하는 비용을 줄일 수 있으리란 기대
- 인공지능의 도입 이후에도 기존 교육 전문가의 역할은 중요함: 환각 현상을 막기 위해 보다 정확한 교육자료를 제작하고, 인공지능의 활용과정 속 장점 및 문제점을 정확히 진단해야 함
- “인공지능이 세상을 지배”하는 것이 아닌, “인공지능을 잘 활용하는 인간이 세상을 지배”한다!

4. 프로젝트 결과

생각해 본 개선방법: GPT-4 Fine-tuning (추가 훈련) 진행 후 문제 생성

- 본 서비스의 문제점: 내용 일치, 암기 위주의 문제만 생성하고 추론 문제를 생성하지 못함
- 암기과목인 경우엔 괜찮지만, 수능 문제처럼 사고력을 요하거나, 수학/과학 문제처럼 계산이 필요한 문제는 생성되지 않음



(질문 – 추론과정을 설명한 답변) 데이터로 finetuning 결과, 새로운 질문을 물어봤을 때도 추론과정을 이용해 답변함

Chung et al., (2022)

Fine-tuning

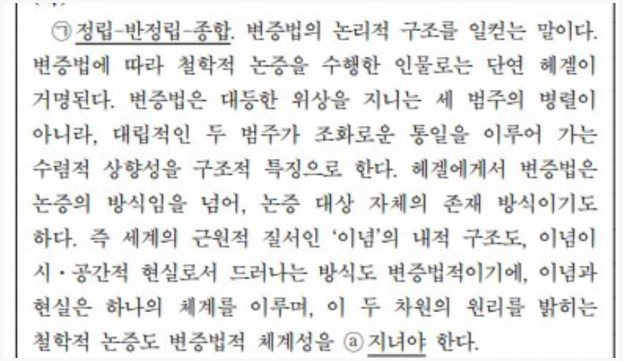
- 인공지능의 훈련은, (input – label) 쌍의 데이터를 학습하면서 이루어짐
- fine-tuning은 이 과정을, 새로운 모형이 아닌 기존의 모형에 추가적으로 진행하는 것
- 성능 좋은 기존 모형을 그대로 가져와서 새로운 작업에 적용할 수 있음

4. 프로젝트 결과

- 미리 **(프롬프트+학습자료 텍스트 – 질문)** 쌍을 만들어 두고, 이를 GPT-4에 파인튜닝해볼 수 있음
- 수능 문제 등 기존에 출제된 양질의 문제를 이용해 파인튜닝한다면, GPT-4가 기존 내용일치 문제 말고도 논리적 사고가 필요한 문제도 생성할 수 있게끔 학습할 수 있음
- 한계: 새로운 모델을 만드는 것보단 덜하지만, 파인튜닝 역시 큰 비용이 소모됨. 개인이 하기엔 어려움.

프롬프트: 아래 글을 이용해서,
보기가 5개인 객관식 문제를 만들어

학습자료 텍스트:



㉠정립-반정립-종합. 변증법의 논리적 구조를 일컫는 말이다. 변증법에 따라 철학적 논증을 수행한 인물로는 단연 헤겔이 거명된다. 변증법은 대등한 위상을 지니는 세 범주의 병렬이 아니라, 대립적인 두 범주가 조화로운 통일을 이루어 가는 수렴적 상향성을 구조적 특징으로 한다. 헤겔에게서 변증법은 논증의 방식임을 넘어, 논증 대상 자체의 존재 방식이기도 하다. 즉 세계의 근원적 질서인 '이념'의 내적 구조도, 이념이 시·공간적 현실로서 드러나는 방식도 변증법적이기에, 이념과 현실은 하나의 체계를 이루며, 이 두 차원의 원리를 밝히는 철학적 논증도 변증법적 체계성을 ⓐ지녀야 한다.

(input)

질문:

5. (가)에서 알 수 있는 헤겔의 생각으로 적절하지 않은 것은?

① 예술·종교·철학 간에는 인식 내용의 동일성과 인식 형식의 상이성이 존재한다.
② 세계의 근원적 질서와 시·공간적 현실은 하나의 변증법적 체계를 이룬다.
③ 절대정신의 세 가지 형태는 지성의 세 가지 형식이 인식하는 대상이다.
④ 변증법은 철학적 논증의 방법이자 논증 대상의 존재 방식이다.
⑤ 절대정신의 내용은 본질적으로 논리적이고 이성적인 것이다.

(label)

참고문헌 (1)

- 길애경. (2023). [화제] 챗GPT에 “특수상대성이론 문제 내줘” 했더니. 대덕. https://www.hellodd.com/news/articleView.html?idxno=99598
- Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. Memory & Cognition, 36(3), 604–616. https://doi.org/10.3758/mc.36.3.604
- Chung, H. W., Hou, L., Longpre, S., Barret Zoph, Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Narang, S., Mishra, G., Adams Wei Yu, & Zhao, V. (2022). Scaling Instruction-Finetuned Language Models. https://doi.org/10.48550/arxiv.2210.11416
- Lee, U., Jung, H., Jeon, Y., Sohn, Y., Hwang, W., Moon, J., & Kim, H. (2023). Few-shot is enough: Exploring CHATGPT prompt engineering method for automatic question generation in English education. Education and Information Technologies. https://doi.org/10.1007/s10639-023-12249-8
- Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. Psychological Bulletin, 140(6), 1432–1463. https://doi.org/10.1037/a0037559

참고문헌 (2)

- 서지영. (2024). 랭체인으로 LLM 기반의 AI 서비스 개발하기. 길벗.
- Introduction Langchain. (2024). https://python.langchain.com/docs/get_started/introduction
- Streamlit docs. (2024). https://docs.streamlit.io/
- **프로젝트 소스코드**: https://github.com/freud-sensei/test_maker

질문과 답변