완전성과 일관성 측면에서의 GPT-3.5 와 GPT-4 의 코딩 성능 평가

정지민, 이찬호
한국외국어대학교 컴퓨터공학부 학부생

Stopmin02@hufs.ac.kr, 1chanho1@gmail.com

Evaluation Coding Performance of GPT-3.5 and GPT-4 in Terms of Completeness and Consistency

Jimin Jung, Chanho Lee
Division of Computer Engineering, Hankuk University of Foreign Studies

요 약

본 연구는 GPT-3.5 와 GPT-4 를 대상으로 완전성과 일관성 측면에서 코딩 협업 환경에 어떤 버전이 더 적합한지 평가하는 것을 목표로 한다. 두 버전을 대상으로 실험한 결과, GPT-4 가 GPT-3.5 보다 완전성과 일관성 측면에서 더 높은 성능을 보였다. 특히 GPT-4 는 모든 항목들에서 100%의 완전성을 보였으나, 일관성은 여전히 개선이 필요함을 확인하였다. 프롬프트 수정만으로는 한계가 있으며, GPT-4 자체의 업그레이드가 필요하다는 의미이며, 향후 연구를 통해 타 생성형 AI 의 성능들도 평가할 예정이다.

1. 서론

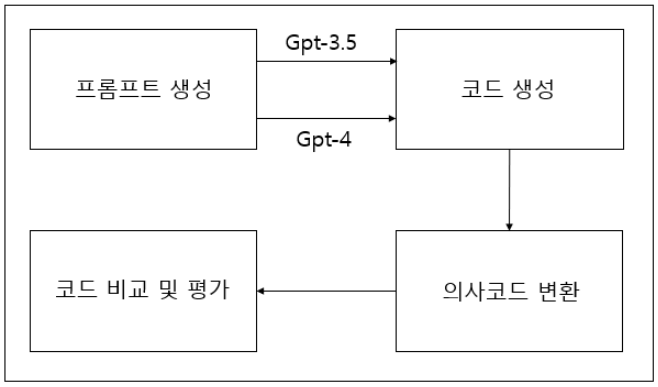
ChatGPT 는 전세계적으로 화두가 되고 있는 텍스트 기반 생성형 AI 로, 거대 언어 모델을 기반으로 자연어 텍스트를 생성한다. 특히, 대규모의 인간이 작성한 데이터를 학습하여, 사용자와 직관적인 방식으로 질문에 응답할 수 있다[1-2]. 코드 생성은 ChatGPT 가 활용되는 또다른 응용 분야 중 하나로, 주어진 프롬프트를 바탕으로 코딩 작업을 수행한다[3]. 본 연구는 코딩 협업 환경에서 ChatGPT 의 유용성을 평가하기 위해 최신 버전들인 GPT-3.5 와 GPT-4 를 비교한다. 이를 통해 코딩 협업 활동에 어떤 버전이 더 적합한지 평가하고자 하며, 완전성과 일관성 측면에서의 차이를 고찰하고 개선 방향을 연구하고자 한다.

2. 실험 목적

본 연구는 코딩 과정에서 발생하는 이슈를 찾기 위해 GPT-3.5[1]와 GPT-4[2]를 비교한다. 이를 위해 완전성(Completeness)과 일관성(Consistency)이라는 두 평가 기준을 도입한다. 이러한 평가 기준을 바탕으로 두 버전의 성능을 비교함으로써 코딩 과정에서 발생하는 이슈를 분석하는 것을 목적으로 한다.

3. 실험 방법론

생성형 AI 는 동일한 프롬프트에 대해 매번 다른 코딩 결과를 만들어 내기 때문에, 협업 환경에서 동일한 코드를 공유하고 업데이트하는데 어려움이 있을 수 있다. 이에 따라 본 연구는 팀 프로젝트와 같은 협업 환경에서 GPT-3.5 와 GPT-4 의 비교와 개선을 위한 GPT 기반 코딩 비교 실험 절차를 다음과 같이 정의한다.



(그림 1) 코딩 비교 실험 절차

1. 프롬프트 생성: 프롬프트는 다양한 코딩 능력을 평가하기 위해 입력, 예외처리, 최대공약수와 최소공배수의 합 계산으로 구성한다. 프롬프트의 요구사항은 총 4 가지로, 입력받기, 자연수, 오버플로우 관련 에러 발생 시 에러 메시지 출력 및 종료, 최대공약수와 최소공배수의 합 출력이 포함되어있다.

   프롬프트: "The program receives two numbers and then checks whether they are natural numbers or if an overflow occurs.

[1] https://chat.openai.com/chat?model=text-davinci-002-render-sha
[2] https://chat.openai.com/chat?model=GPT-4

```
If an error occurs, print an error message and exit. If the input is valid, it returns the sum of the greatest common divisor and least common multiple of two numbers."
```

	Step1					Step2				Step3		Step4		
	모듈		입력			예외처리				최대공약수	최소공배수	합		프로세스
코드	Math	Sys	In	Out	Sub	N	O	V	E			Sub	Main	
1	✓			✓		✓	✓	✓	✓	✓	✓	✓		1-2-3-4
2		✓	✓			✓	✓			✓	✓		✓	1-2-3-4
3	✓	✓		✓		✓	✓			✓*	✓		✓	1-2-3-4
4	✓	✓		✓		✓	✓	✓		✓*	✓		✓	1-2-3-4
5	✓	✓		✓		✓	✓			✓	✓		✓	1-2-3-4
6	✓	✓				✓	✓			✓	✓	✓		2-3-4
7	✓				✓	✓				✓*	✓	✓		1-2-3-4
8	✓			✓		✓		✓	✓	✓	✓		✓	1-2-3-4
9	✓			✓		✓	✓	✓		✓	✓		✓	1-2-3-4
10	✓					✓	✓			✓	✓	✓		2-3-4
Completeness			80%			100%	80%			100%	100%	100%		
Consistency	50%		60%									60%		
1	✓	✓	✓			✓	✓	✓	✓	✓	✓		✓	1-2-3-4
2		✓	✓			✓	✓	✓		✓	✓		✓	1-2-3-4
3		✓	✓			✓	✓	✓		✓	✓		✓	1-2-3-4
4		✓	✓			✓	✓	✓		✓	✓		✓	1-2-3-4
5			✓			✓	✓	✓	✓	✓	✓		✓	1-2-3-4
6		✓	✓			✓	✓	✓		✓	✓		✓	1-2-3-4
7		✓	✓			✓	✓	✓		✓	✓		✓	1-2-3-4
8		✓	✓			✓	✓	✓		✓	✓		✓	1-2-3-4
9		✓	✓			✓	✓	✓		✓	✓		✓	1-2-3-4
10			✓			✓	✓	✓		✓	✓		✓	1-2-3-4
Completeness			100%			100%	100%			100%	100%	100%		
Consistency	70%		100%									100%		

(표 1) GPT-3.5 와 GPT4 의 코드 완전성 및 일관성 비교 분석 결과(Math: Math 모듈 사용함, Sys: Sys 모듈 사용함, In: 메인함수 안에서 입력받음, Out: 메인함수 밖에서 입력받음, Sub: 서브함수를 생성함, N: 자연수 확인 예외처리함, O: 오버플로우 확인 예외처리함, V: ValueError 예외처리함, E: Exception 예외처리함, Main: Main 함수 안에서 합을 계산함, *: 서브함수를 만들지 않고 Math 모듈 사용함)

2. 코드 생성: 주어진 프롬프트를 입력으로 GPT-3.5 와 GPT-4 에서 충분한 코드 비교를 위해 각각 10 회 실행한다.

3. 의사코드(Pseudo Code) 변환: 다수의 코드를 자연어에 가까운 의사코드로 변경함으로써 코드의 구조와 알고리즘을 빠르게 파악하여 비교, 평가하기 위해 기존 파이썬 코드를 의사코드로 변환한다.

4. 코드 비교 및 평가: 완전성은 언어 모델이 주어진 프롬프트를 얼마나 정확하게 수행하여 사용자 요구를 충족시키는지 측정하는 지표로 정의하고, 일관성은 모듈, 함수 사용 및 프로세스의 동일성을 평가하는 지표로 정의한다. 이 두 평가 기준을 바탕으로, 두 언어 모델을 비교하여 각 버전의 성능을 평가한다.

4. 결과 분석 및 비교

표 1 은 GPT-3.5 와 GPT-4 가 생성한 각각의 10 개 코드에 대해 완전성과 일관성을 평가한 결과이다. 행은 각 버전 별 코드 10 개, 그리고 완전성/일관성 성능을 나타내며, 열은 모듈, 입력(Step 1), 예외처리(Step 2), 최대공약수와 최소공배수(Step 3), 그들의 합(Step 4), 그리고 프로세스(로직 순서)를 포함한다.

완전성에 대해 평가한 결과, GTP-4 는 모든 항목들에 대해 100%의 성능을 보여주는 데 반해, GPT-3.5 는 오버플로우 항목에서 80%를 보여준다. 일관성에 대해 평가한 결과, GPT-3.5 는 모든 항목들에서 50~60%를, GPT-4 는 70%를 보여준 모듈 항목을 제외하고 100%를 보여준다.

본 실험에서 GPT-3.5 와 GPT-4 가 각각 동일한 프롬프트에 대해 답변한 내용을 분석한 결과, GPT-4 가 GPT-3.5 보다 완전성과 일관성 측면에서 우월한 성과를 보였다. 결과적으로 GPT-4 의 일관성은 이전 버전에 비해 상당히 향상되었지만, 여전히 개선의 여지가 있다는 것을 확인하였는데, 이는 GPT-4 의 자체적인 업그레이드를 통해 해결해야 한다.

5. 결론 및 향후연구

본 연구는 코딩 분야의 협업 환경에서 GPT-3.5 와 GPT-4 를 비교하기 위해 완전성과 일관성을 정의하고 평가 지표로 활용하였다. 이러한 평가 지표를 기반으로, GPT-3.5 와 GPT-4 가 코딩 분야의 협업 환경에서 어떤 가능성과 이슈가 있는지 확인하였다. 향후 ChatGPT 뿐만 아니라 Bard[3], CLOVA[4] 와 같은 생성형 AI 도 추가 분석하여 비교 평가할 예정이다.

참고문헌

[1] J. Zhu, et al., "ChatGPT and Environmental Research," *Environmental Science & Technology*, March 2023.

[2] C. Anton, "Probing CHAT GPT: A Media Ecology Writing Sampler," *New Explorations: Studies in Culture and Communication* 3.1, 2023.

[3] A. Narasimhan et al., "CGEMs: A metric model for automatic code generation using GPT-3," *arXiv preprint arXiv:2108.10168*, 2021.

[3] https://bard.google.com/

[4] https://clova.ai/