학술논문 검색서비스
earticle



2023년도 한국경영정보학회 춘계 학술대회
디지털플랫폼 성공을 위한 경영정보학의 역할

ChatGPT를 활용한 자기소개서 평가 성능에 대한 경험적 연구

윤욱지, 구요한, 장훌

To cite this article : 윤욱지, 구요한, 장훌 (2023) ChatGPT를 활용한 자기소개서 평가 성능에 대한 경험적 연구 , 2023년도 한국경영정보학회 춘계 학술대회, pp.527-543

① earticle에서 제공하는 모든 저작물의 저작권은 원저작자에게 있으며, 학술교육원은 각 저작물의 내용을 보증하거나 책임을 지지 않습니다.

② earticle에서 제공하는 콘텐츠를 무단 복제, 전송, 배포, 기타 저작권법에 위반되는 방법으로 이용할 경우, 관련 법령에 따라 민, 형사상의 책임을 질 수 있습니다.

www.earticle.net

Download by : 168.188.35.222

Accessed : Monday, May 27, 2024 10:38 AM

ChatGPT를 활용한 자기소개서 평가 성능에 대한 경험적 연구

(An Empirical Analysis of ChatGPT's Performance in Evaluating Cover Letters)

윤욱지 / FMASSOCIATES
구요한 / 한양대학교
장훌 / 한양대학교

윤욱지
FMASSOCITES ASSOCIATE CONSULTANT
wookji.yoon@gmail.com

[Provider:earticle] Download by IP 168.188.35.222 at Monday, May 27, 2024 10:38 AM

Contents

01 Background
02 Related Study
03 Data
04 Methodology
05 Result
06 Conclusion

국내 채용시장의 변화

- 직무중심 채용에 있어 자기소개서 중요성 대두
  - 경제성장의 둔화, 인구구조의 변화, 기술의 발전, 가치관의 변화 등의 요인으로 인해 **2010**년대 이후, **국내 인사관리(HR) 제도는 연공주의에서 직무중심주의**로 변화하고 있음 **(Bae et al., 2011)**
  - **'직무 중심 채용'**이 이루어짐에 따라, 지원자의 직무이해도와 직무경험을 판단할 수 있는 '자기소개서'의 중요성이 강조됨

<국내 기업 채용 과정, 시대별 서류전형단계 중점 사항 (이종구, 2022)**>**

2000년대 이전	2000년 ~ 2014년	2015년 이후
▪ **입사 필기시험 성적** ▪ **속인적 배경** - 외양 및 성격 - 가족관계 및 가정환경 - 학교생활 등	▪ **스펙(Specification)** - 어학시험 점수 - 자격증 - 학력 및 학점 - 수상경력 등	▪ **자기소개** - 지원직무에 대한 지식 - 지원직무에 대한 경험 - 직무 관심도/적합도

1/29

[Provider:earticle] Download by IP 168.188.35.222 at Monday, May 27, 2024 10:38 AM

Contents

01 Background
02 Related Study
03 Data
04 Methodology
05 Result
06 Conclusion

연구목적

- 채용 과정에서 많은 시간 소모 및 비용 발생
  - 스펙(**SPEC**) 또는 필기시험 중심의 채용이 이루어질 시, 지원자의 자격요건을 정량적(**Quantitative**)으로 점수화 및 순위화가 용이함
  - 반면, 정성적(**Qualitative**)인 성격을 가지고 있는 '직무이해도', '직무경험' 등의 경우 지원자의 자격요건을 비교하기 위해 숙련된 인사담당자 또는 직무전문가의 판단이 필요함
  - 이는 **사용자 입장에서 많은 시간과 비용의 발생**을 야기함
- 기업들에 정보기술을 활용한 채용 효율화/자동화 니즈가 존재
  - **ChatGPT를 통해 자기소개서 평가 과정을 효율화**할 수 있는 방안을 연구함
  - 궁극적으로, **ChatGPT**와 같은 **대규모 언어 모델(Large Language Model)이 인간 평가자를 대체할 수 있을지**에 대한 가능성을 탐색해보고자 함

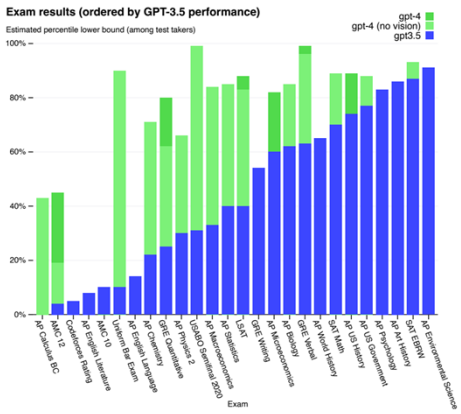
2/29

Contents

01 Background
02 Related Study
03 Data
04 Methodology
05 Result
06 Conclusion

Related Study

- **Performance for Academical and Professional Exam**



Source: OpenAI (2023)

- OpenAI(ChatGPT 개발사) 공식 Report
- 많은 학술시험 또는 전문직 시험에서 **GPT4는 사람과 유사 혹은 뛰어난 수준의 performance**를 보임
- 미국 변호사 시험, SAT 등 주요 시험에서 상위 10%에 해당하는 성적을 기록함

3/29

[Provider:earticle] Download by IP 168.188.35.222 at Monday, May 27, 2024 10:38 AM

Contents

01 Background

02 Related Study

03 Data

04 Methodology

05 Result

06 Conclusion

Related Study (Cont'd)

- **Performance for MCQ(Multiple Choice Questions) Exam**

First Author	Discipline	Exam Name	Level	N Question	N ChatGPT Correct	% ChatGPT Correct	Pass Mark (%)	Ave student %	Chat GPT vs student	GPT vs pass
Antaki	Ophthalmology	OKAP (BCSC)	PG	260	145	55.8	-	74	-18.2	-
Antaki	Ophthalmology	OKAP (OphthoQuestions)	PG	260	111	42.7	-	61	-18.3	-
Huh	Parasitology	University Exam	UG	79	48	60.8	60.8	90.8	-30.0	0
Gilson	Medicine	USMLE (NBME Free Step 1)	UG	87	56	64.4	60	77.6	-13.2	4.4
Gilson	Medicine	USMLE (NBME Free Step 2CK)	UG	102	59	57.8	71.3	82	-24.2	-14
Gilson	Medicine	USMLE (AMBOSS-Step 1)	UG	100	44	44	60	77.6	-33.6	-16
Gilson	Medicine	USMLE (AMBOSS-Step 2 CK)	UG	100	42	42	71.3	82	-40.0	-29
Kung	Medicine	USMLE Step 1	UG	119	43	36.1	60	77.6	-41.5	-24
Kung	Medicine	USMLE Step 2CK	UG	102	58	56.9	71.3	82	-25.1	-14.4
Kung	Medicine	USMLE Step 3	UG	122	68	55.7	66	75.8	-20.1	-10
Fijacko	Life Support	AHA BLS (A)	Various	25	17	68	84	-	-	-16
Fijacko	Life Support	AHA BLS (B)	Various	25	16	64	84	-	-	-20
Fijacko	Life Support	AHA ACLS (A)	Various	38	26	68.4	84	-	-	-16
Fijacko	Life Support	AHA ACLS (B)	Various	38	29	76.3	84	-	-	-7.7
Choi	Law	Constitutional Law	UG	25	21	84	-	92	-8.0	-
Choi	Law	Torts	UG	10	6	60	-	80	-20.0	-
Choi	Law	Taxation	UG	60	24	40	-	80	-40.0	-
Geerling	Economics	Intro Microeconomics	UG	30	19	63.3	-	42.6	20.7	-
Geerling	Economics	Intro Macroeconomics	UG	30	26	86.7	-	47.3	39.4	-
Kortmeyer	Physics	Intro Physics	UG	30	14	46.7	60	-	-	-13.3
			TOTAL	1642	872					
			MEAN %		53.1	58.7			-18.1	-13.5

- 학부생 및 대학원 수준의 객관식 시험에 대한 ChatGPT의 성능을 연구한 8개 논문에 대한 메타분석
- 대부분의 모든 시험에서 **합격선을 충족하지 못하며, 평균적인 인간 학생보다 낮은 성능**을 보임
- 언론 보도와 일치하지 않는 이유는
  1. 에세이형 문제
  2. 데이터 생성 시점
  3. 추가정보의 제공

Source: Newton (2023)

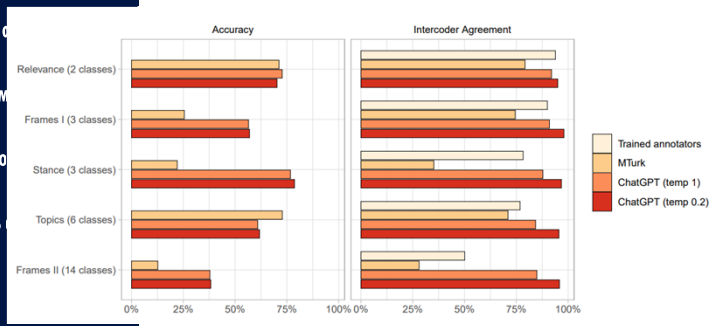
Contents

01 Background

02 Related Study

Related Study (Cont'd)

- **Performance for Text-Annotation task**
  - **2,382**개 **tweet**에 대해 관련성, 입장, 토픽, 프레임 감지 등 **5**가지 종류의 주석을 다는 작업을 수행함 (**1.** 관련성: 연관/비연관, **2.** 입장: 긍정/부정/중립 등)



- 주석작업은 '숙련된 연구자', '크라우드워커', 'ChatGPT'에 의해 수행되고 그 결과를 비교함
- 주석 작업에서 ChatGPT가 크라우드 워커보다 우수한 성능을 보였으며, **텍스트 주석 및 분류작업에서 LLM이 인간을 대체할 잠재력** 있음

Source: Gilardi et al. (2023)

[Provider:eartricle] Download by IP 168.188.35.222 at Monday, May 27, 2024 10:38 AM

Contents

01 Background
02 Related Study
03 Data
04 Methodology
05 Result
06 Conclusion

About Data

- **HR** 인턴십 프로그램 지원을 위한 자기소개서 (문항 **4**개)
- 지원자 **84**명의 자기소개서 **Data**
- 자기소개서 채점을 위한 **Rubric**

1	인턴십에 지원하게 된 동기와 활동을 통해 얻고 싶은 것은 무엇인지 서술하시오
2	본인이 관심있는 HR 분야를 한 가지 골라 해당 분야에 대해 설명하고, 그 분야에 관심을 가지게 된 이유를 서술하시오
3	본인이 생각하는 협력의 가치란 무엇인지 설명하고, 그러한 생각을 가지게 된 계기를 서술하시오
4	현재에 안주하지 않고, 계속해서 성장하기 위해 노력했던 경험에 대해 서술하시오

6/29

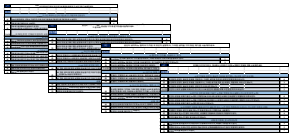
Contents

01 Background
02 Related Study
03 Data
04 Methodology
05 Result
06 Conclusion

About Data

- 인재상을 반영하기 위한 **Rubric** 고도화 과정

인재상 정의	평가기준 도출	Rubric 구체화
4개 인재상	17개 평가기준	19개 Criteria 및 3점 척도
성장의지 현재에 안주하지 않고, 계속해서 배우고 성장하고자 하는 자세를 가진 자	**성장의지** 1. 인생을 통해 이루고자 하는 뚜렷한 비전이 있다. 2. 비판적 상황 판단과 자발적인 학습을 통한 성장 경험이 있다. 3. 자신의 강점과 약점을 알고 있으며 이를 강화/개선한 경험이 있다. 4. 포기하지 않고 끝까지 노력하는 태도를 지니고 있다.	<Criteria 1> **[A]** 본인의 성장을 위해 기울인 노력이 구체적으로 드러나며, 문제인식과의 인과관계가 적절하게 이루어진다. **[B]** 본인의 성장을 위해 기울인 노력을 구체적으로 작성했으나, 문제인식과의 인과관계가 명확히 드러나지 않는다. **[C]** 본인의 성장을 위해 기울인 노력이 잘 드러나지 않는다.
타인존중 및 협력적 의사소통 함께하는 기쁨을 알고, 그 가치를 실천하고자 하는 자	...	
HR에 대한 관심과 열정 HR에 대한 흥미와 관심이 있고, 기초적인 HR 지식을 보유하고 있는 자		
인턴십에 대한 관심과 열정 본 프로그램에 참여하고자 하는 적극적인 의지가 느껴지는 자		



7/29

Contents

01 Background
02 Related Study
03 Data
04 Methodology
05 Result
06 Conclusion

About ChatGPT

- **ChatGPT**는 **OpenAI**에서 개발한 언어 모델
  - **GPT-3.5** 아키텍처를 사용하여 다양한 주제에 대해 사람과 유사한 응답을 생성
  - 인터넷, 책, 기타 출처의 방대한 양의 텍스트를 학습한 **ChatGPT**는 자연어를 이해하고 생성하여 사람들과 대화에 참여하여 유익하고 유용한 답변을 제공
  - 문맥을 이해하고 의미를 추론하며 일관되고 관련성 있는 응답을 생성하는 능력 덕분에 언어 번역, 고객 서비스, 개인 비서, 창작물 작성 등 다양한 분야에서 강력 한 도구로 활용
  - 알고리즘과 학습 데이터에 대한 업데이트와 개선을 통해 지속적으로 학습하고 개선하는 **ChatGPT**는 우리가 컴퓨터 및 인공 지능과 상호 작용하는 방식을 변화 시키는 첨단 기술
- **In this Study,**
  - 작업 수행을 위해 '**Chat-3.5-turbo**' 버전의 **ChatGPT API**를 사용함
  - 작업은 **2023**년 **4**월 **22**일부터 **4**월 **24**일까지 진행되었음
    (**ChatGPT**의 **3**월 **23**일 **Update Version**에 해당)

8/29

[Provider:earticle] Download by IP 168.188.35.222 at Monday, May 27, 2024 10:38 AM

Contents

01 Background
02 Related Study
03 Data
04 Methodology
05 Result
06 Conclusion

ChatGPT 요청 Prompt

- **Prompt?**
  - **ChatGPT**에게 질문이나 지시를 주는 문장을 의미함
  - 동일한 작업을 지시하더라도 어떤 **Prompt**를 사용하는 지에 따라 성능이 다름
  - **ChatGPT**에게 특별한 **Prompt** 방식을 사용하는 경우, 성능을 향상시킬 수 있음
    (Kojima et al., 2022)
    - **Chain of Thought: 'Think step by step'** 등의 요청 추가
    - **Few-shot: ChatGPT**가 수행할 작업과 동일한 작업을 수행한 예시를 추가
- **In this Study,**
  - 실제 채점 환경과 비교하기 위해 의도적으로 **ChatGPT** 전용 프롬프트를 제외함
  - 이전 답변에 영향을 받는 것을 방지하기 위해, 매 문항마다 새로운 세션으로 프롬프트를 제공함

9/29

[Provider:earticle] Download by IP 168.188.35.222 at Monday, May 27, 2024 10:38 AM

Contents

01 Background
02 Related Study
03 Data
04 Methodology
05 Result
06 Conclusion

ChatGPT 요청 Prompt(Cont'd)

- 여러가지 변형을 테스트한 후, 각 문항 별로 아래와 같은 **Prompt**를 **ChatGPT**에게 제공함

"SYSTEM PROMPT" :
You are a helpful assistant that helps HR Personnel for the cover letter evaluation.

"USER PROMPT" :
Context: **The HRM/HRD training program called "{프로그램명}" is recruiting candidates. HR personnel will evaluate the cover letters submitted by the candidates.**
Info: **"{프로그램명}" are proper noun.**
Provide: **cover-letter question, cover-letter content, rubric(criteria w/ rating grades),**
Assistant Need to do: **Provide ratings for each criteria in accordance with the rubric, and include a rationale for each rating.**

[Question]
Please describe your experience of trying to continue to grow and not settle for the present.

Contents

01 Background
02 Related Study
03 Data
04 Methodology
05 Result
06 Conclusion

ChatGPT 요청 Prompt (Cont'd)

[Rubric]
#Criteria 1: [Before Growth] Did the applicant identify areas where they needed to grow?
- Grade A: The applicant mentions an area in which they need to grow, and the process by which they came to that thought is adequately demonstrated.
- Grade B: The applicant mentions an area where they need to grow, but does not adequately demonstrate how they came to this thought.
- Grade C: The applicant does not mention an area where they need to grow.

#Criteria 2: [During Growth] Did the applicant describe specific efforts they made to grow?
- Grade A: The applicant's efforts to grow are specific and cause-and-effect links to problem recognition are adequate.
- Grade B: The applicant describes efforts they have made to grow, but the causal link to problem recognition is not clear.
- Grade C: The efforts made for their growth are not well revealed.

[Provider:earticle] Download by IP 168.188.35.222 at Monday, May 27, 2024 10:38 AM

Contents

01 Background
02 Related Study
03 Data
04 Methodology
05 Result
06 Conclusion

ChatGPT 요청 Prompt (Cont'd)

#Criteria 3: [Post Growth] Did the applicant adequately describe what they learned or felt from their growth experience?

- Grade A: The applicant's description of what they learned/felt from the growth experience is specific and well connected to the experience.
- Grade B: The applicant describes what he/she learned/felt from the growth experience, but the connection to the growth experience is inadequate.
- Grade C: The applicant does not clearly articulate what they have learned/felt.

#Criteria 4: Documentation skills

- Grade A: The applicant's intentions are clear throughout the answer, and there are no major flaws.
- Grade B: The applicant's intent is evident in the answer, but the sentence structure/word choice/alignment/spacing is inadequate.
- Grade C: The applicant's intention is difficult to ascertain.

#Criteria 5: Self-directed Growth (Yes or No)

- Yes: The applicant shows some evidence of self-directed personal growth, not driven by others or external circumstances.
- No: Most of the applicant's struggles are caused by others or external circumstances.

Contents

01 Background
02 Related Study
03 Data
04 Methodology
05 Result
06 Conclusion

ChatGPT 요청 Prompt (Cont'd)

[Content]

Based on my experience in community service 1 in the first semester of 2018, I wanted to practice community service once again through community service 2 in the second semester of 2020, so I took the relevant community service course. In particular, I applied as a mentor for the 21st School-Life Mentoring Program (SMP) after being discharged from the military because I wanted to abandon the existing consciousness that I thought was difficult to do community service once again. I aimed to introduce various fields of practice to mentees who lacked information on the field and content of the SMP activities, and to provide clues that ultimately become voluntary and continuous activities by choosing activities that suit their abilities and aptitudes. Through this, I aimed to develop my knowledge as a volunteer by reviewing the basic attitude and values necessary for social service activities once again, and to develop respect and altruism for humans. The new experience of participating in community service activities organized by Seoul National University provided an opportunity to look at the lives of myself and others, and I wanted to broaden my horizons by becoming interested in my neighbors. I met mentees who work in school life in a different way from myself, got away from self-centered thinking and understood them, and I was able to have a sense of community that we should live together as members of society by recognizing them as "us" rather than "others."

[Assistant]

Rating is:

[Provider:earticle] Download by IP 168.188.35.222 at Monday, May 27, 2024 10:38 AM

Contents

01 Background
02 Related Study
03 Data
04 Methodology
05 Result
06 Conclusion

Prompt에 대한 ChatGPT 응답 예시

Criteria 1: Grade A - The applicant clearly identifies an area where they needed to grow, which is to abandon their existing consciousness that they thought was difficult to do community service once again. The process by which they came to this thought is adequately demonstrated by their experience in community service 1.

Criteria 2: Grade A - The applicant describes specific efforts they made to grow, such as taking the relevant community service course and applying as a mentor for the 21st School-Life Mentoring Program. The causal link to problem recognition is clear, as they wanted to introduce various fields of practice to mentees who lacked information on the field and content of the SMP activities.

Criteria 3: Grade A - The applicant adequately describes what they learned and felt from their growth experience, such as developing their knowledge as a volunteer and developing respect and altruism for humans. The description is specific and well connected to the experience.

Criteria 4: Grade B - While the applicant's intentions are clear throughout the answer, there are some minor flaws in sentence structure and word choice

Criteria 5: Yes - The applicant shows evidence of self-directed personal growth by taking the relevant community service course and applying as a mentor for the 21st School-Life Mentoring Program. They aimed to develop their knowledge as a volunteer and broaden their horizons by becoming interested in their neighbors.

14/29

Contents

01 Background
02 Related Study
03 Data
04 Methodology
05 Result
06 Conclusion

연구 절차

※ 본 연구범위

Study 1	Study 2	Future Research
ChatGPT는 인간 평가자와 동일한 방식으로 자기소개서 평가 작업을 수행할 수 있는가?	ChatGPT는 평가 성능은 인간평가자와 비교했을 때, 타당한 수준인가?	채용 과정(특히, 서류전형), ChatGPT가 인간 평가자를 대체할 수 있는가?
Rubric 기반 평가 작업수행	Performance 비교	실무도입방안 도출
▪ **조작변인** – Temperature – 평가등급명칭 ▪ **분석 방안** – 과업성공 확률 – 신뢰도 – 평균점수 및 평가 경향	▪ **인간 평가자** – 인사담당자 9인 – Benchmark 점수 ▪ **ChatGPT (GPT-3.5)** – Study 1 최적조건 활용 ▪ **GPT-4** – Study 1 최적조건 활용	▪ **ChatGPT를 실무에서 활용하기 위한 최적의 방안은 무엇인가?** – 전용 Prompt의 개발 – 전용 Rubric의 개발 – 전용 자기소개서 문항의 개발 등

최적조건 탐색

15/29

Contents

01 Background
02 Related Study
03 Data
04 Methodology
05 Result
06 Conclusion

[Study 1] 조작변인

- **Temperature (5 Cases)**
  - 출력값의 무작위성(**Randomness**)을 결정하는 **ChatGPT**의 **Parameter**
  - **0**에서 **2**사이의 값을 가짐
  - 값이 높을수록 출력이 무작위적으로 도출되며(**more random**), 값이 낮을수록 고정된 답변을 함(**more deterministic**)
  - 다섯가지 **case**를 설정함: **0.0, 0.5, 1.0(default), 1.5, 2.0**

- 평가등급명칭 **(3 Cases)**
  - **ChatGPT**는 **text** 기반 모델이므로 등급명칭에 영향을 받을 것이라 가정
  - 일반적으로 쓰이는 **2**가지 **Case**를 설정함 **(A/B/C, High/Middle/Low)**
  - 국내기업 중 삼성전자에서 사용하는 등급명칭을 별도의 **Case**로 설정함 **(Excellent/Good/Unsatisfactory)**[1)]

- **15**가지 **Case**에 대해 각 **3**회 평가를 반복 실시함

1) 기존 등급은 다섯단계: Excellent/Very Good/Good/Need Improvement/Unsatisfactory

Contents

01 Background
02 Related Study
03 Data
04 Methodology
05 Result
06 Conclusion

Task Success Rate 분석

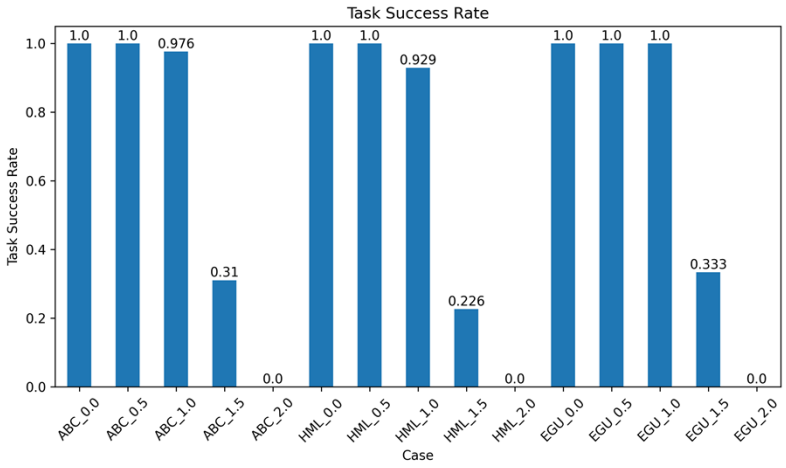
- 과업성공확률**(Task Success Rate)**
  - 정의**: 84**개의 자기소개서 중 **ChatGPT**가 **Rubric**에 맞게 등급을 부여한 자기소개서의 비율
  - **3**회 반복 중 **1**회라도 실패할 경우, 실패로 간주함

[Failure Case]

- **Grade High/Middle,**
  Excellent marks according word choice separation also margins deslocribing is made comparison to history overly la tense inconsistencies It hinder displayed.. There remains a scope of improvement relating misplacement (생략)
- **Grade may contrast between B+ to A+, demonstrates decent written/burr tap grammar rationale of extreme touch explicating Points#, effortlessly pedag. Who evidences comfort defining themselves along business dissertation hypothetical planning.** (생략)
- **Grade: Insufficient Data,**
  Information provided was insufficient to either way an analytic feeback; whether they are properly linked or cannot deliver categorically sensible connection.

[Provider:earticle] Download by IP 168.188.35.222 at Monday, May 27, 2024 10:38 AM

[Result] Task Success Rate



- Temperature가 0.0 또는 0.5인 경우, 모든 평가를 의도에 따라 수행함
- Temperature가 1.0인 경우, 대부분의 평가를 의도에 따라 수행함
- Temperature가 1.5인 경우, 22.6% ~ 33.3%의 성공확률을 보임
- Temperature가 2.0인 경우, 하나의 과업도 완수하지 못함

Temperature 1.5, 2.0의 경우, 추후 분석에서 배제

Temperature 1.0의 경우, Failure case를 제외하고 분석

18/29

Contents

01 Background
02 Related Study
03 Data
04 Methodology
05 Result
06 Conclusion

Inter Rater Agreement 분석

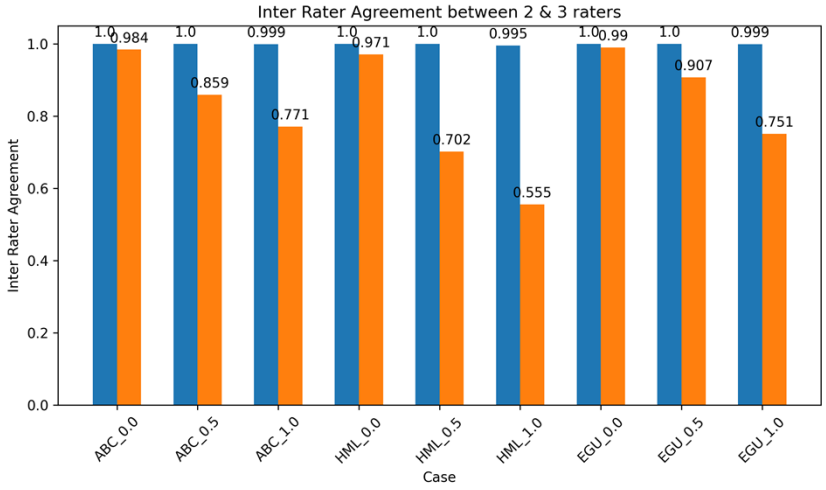
- 평가자간 등급일치 비율 **(Inter Rater Agreement)**
  - **1,596**개 **Criteria**[1] 중 같은 조건의 서로 다른 시행에서 **ChatGPT**가 부여한 등급이 일치하는 **Criteria**의 비율
  - **3-raters IRA: 3**회의 시행에서 모두 동일한 등급을 부여한 경우
  - **2-raters IRA: 3**회의 시행 중 **2**개의 시행에서 동일한 등급을 부여한 경우

1) 84개 자기소개서, 각 자기소개당 19개의 Criteria

19/29

[Provider:earticle] Download by IP 168.188.35.222 at Monday, May 27, 2024 10:38 AM

[Result] Inter Rater Agreement



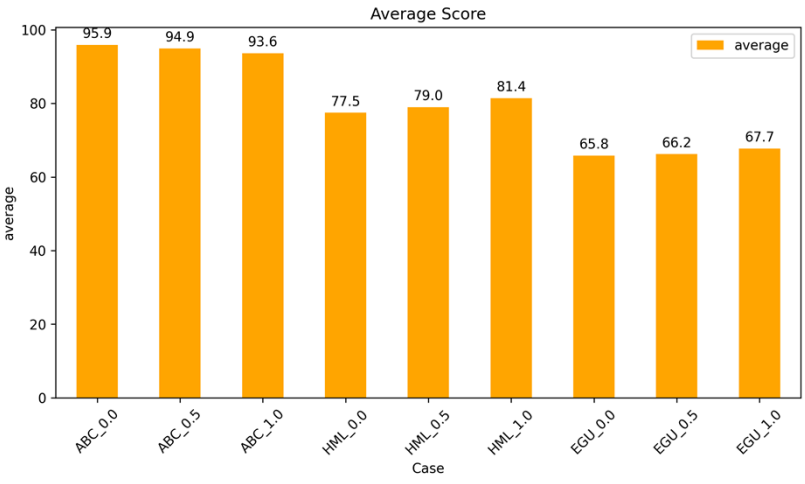
2-raters IRA

3-raters IRA

- 모든 Temperature에서 2-raters IRA는 100%에 근접함
- 3-raters IRA는 Temperature가 커질수록 감소함

20/29

[Result] Average Score

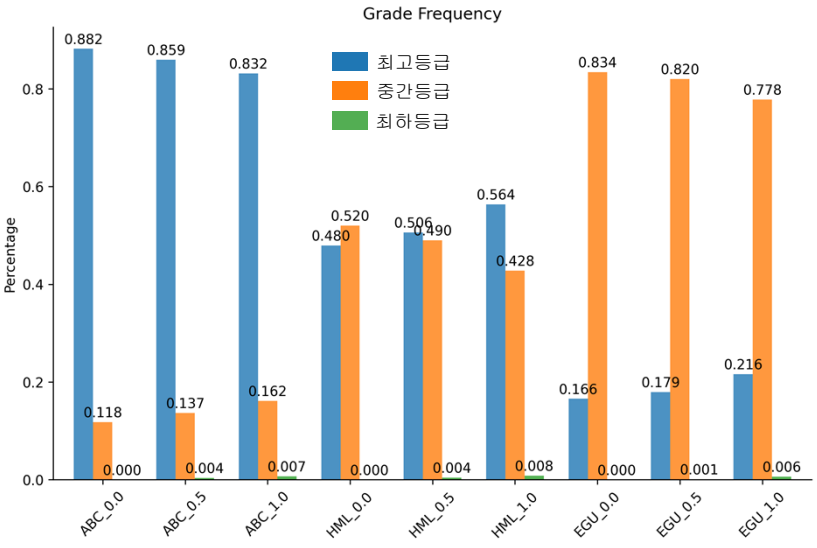


- 평균 점수를 비교한 결과, Temperature가 점수에 미치는 영향이 낮게 나타남 (최대 3.9점)
- 평가등급명칭이 점수에 미치는 영향이 높게 나타남 (최대 30.1점)
- ABC > HML > EGU 순으로 평균점수가 높게 나타남

21/29

[Provider:earticle] Download by IP 168.188.35.222 at Monday, May 27, 2024 10:38 AM

[Result] Grade Frequency



- 평가등급명칭이 ABC인 경우, A등급을 부여하는 비중이 80% 이상으로 점수의 관대화 Bias가 나타남.
- 평가등급명칭이 EGU인 경우, Middle 등급을 부여하는 비중이 77.8% 이상으로 점수의 중심화 Bias가 나타남
- 평가등급 명칭이 HML인 경우, 상대적으로 High 등급과 Middle 등급을 고르게 부여함
- 세가지 경우 모두, 최하등급을 부여하는 확률은 매우 낮음

22/29

Contents

01 Background

02 Related Study

03 Data

04 Methodology

05 Result

06 Conclusion

[Study 1] Findings

- **Task Success Rate**
  - **Temperature** 영향 高, 반비례관계
  - **Temperature**가 **1.5** 이상인 경우, 채점을 성공적으로 수행할 확률이 현저히 감소
- **Inter Rater Agreement**
  - **Temperature** 영향 高, 반비례관계
  - **Temperature**가 **1.0**일 경우, **IRA**는 최소 **55%** 수준으로 감소
- **Average Score & Grade Frequency**
  - 평가등급명칭 영향 高
  - **ABC:** 관대화 현상(**A**등급 비중 높음), 평균 **90**점 이상의 가장 높은 점수
  - **EGU:** 중심화 현상(**Great** 비중 높음), 평균 **65**점 내외로 가장 낮은 점수
  - **HML: High/Middle** 등급의 비중이 유사한 수준, 평균 **80**점으로 중간정도의 점수

- **HML_0.0 및 HML_0.5의 조건이 평가자로 활용하기에 적정한 수준**

23/29

[Provider:earticle] Download by IP 168.188.35.222 at Monday, May 27, 2024 10:38 AM

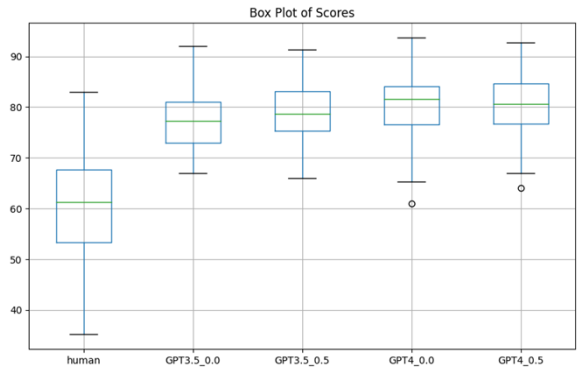
Contents

01 Background
02 Related Study
03 Data
04 Methodology
05 Result
06 Conclusion

[Study 2] 평가성능비교

- 인간 평가자
  - 기업에서 **HR** 실무를 담당하고 있는 인사담당자 **9**인의 평균 점수
  - **ChatGPT**와 동일한 **Rubric**을 가지고 평가를 진행하며, 등급명칭은 상/중/하
- **ChatGPT (GPT-3.5)**
  - **[Study 1]**의 결과에 따라 **HML_0.0** 및 **HML_0.5**의 경우를 활용함
- **GPT-4**
  - **[Study 1]**의 결과에 따라 **HML_0.0** 및 **HML_0.5**의 경우를 활용함

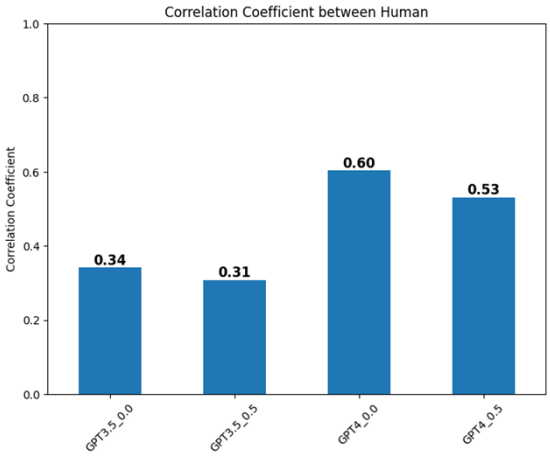
[Result] Average & Distibution



	human	GPT3.5_0.0	GPT3.5_0.5	GPT4_0.0	GPT4_0.5
count	84.0	84.0	84.0	84.0	84.0
mean	60.082	77.516	79.008	80.23	80.345
std	10.627	5.851	5.787	5.973	5.715
min	35.222	67.0	66.0	61.0	64.0
25%	53.361	73.0	75.333	76.583	76.667
50%	61.222	77.333	78.667	81.667	80.667
75%	67.667	81.0	83.167	84.167	84.667
max	83.0	92.0	91.333	93.667	92.667

[Provider:earticle] Download by IP 168.188.35.222 at Monday, May 27, 2024 10:38 AM

[Result] Correleation



	human	GPT3.5_0.0	GPT3.5_0.5	GPT4_0.0	GPT4_0.5
human	1.0	0.341	0.307	0.603	0.531
GPT3.5_0.0	0.341	1.0	0.282	0.122	0.32
GPT3.5_0.5	0.307	0.282	1.0	0.065	0.166
GPT4_0.0	0.603	0.122	0.065	1.0	0.492
GPT4_0.5	0.531	0.32	0.166	0.492	1.0

[Provider:earticle] Download by IP 168.188.35.222 at Monday, May 27, 2024 10:38 AM

Contents

01 Background
02 Related Study
03 Data
04 Methodology
05 Result
06 Conclusion

[Study 2] Findings

- 기술통계 비교
  - 인간평가자의 평균점수는 약 **60**점으로 나타난 반면, **GPT**는 모든 **Case**에서 평균점수가 **75**점 이상으로 인간평가자에 비해 관대화 경향이 관찰됨
  - 인간평가자의 점수 표준편차는 약 **10.6**점으로 큰 편차를 나타냈으며, **GPT**는 **5**점~**6**점 사이의 편차로 인간평가자에 비해 편차가 작음
- 상관관계 비교
  - **GPT-3.5**에 비해 **GPT-4**가 인간평가자의 평가점수와 더 높은 상관관계를 보임
  - **Temperature**가 낮을수록 인간평가자의 평가점수와 더 높은 상관관계를 보임
  - 가장 높은 상관관계를 보인 **Case**는 **GPT-4_0.0** 경우로 **0.6**의 높은 상관관계를 나타냄

- 절대점수를 비교했을 때 **ChatGPT**의 관대화 경향이 나타나지만, 상관관계의 측면에서 인간평가자와 유사한 경향성을 관찰할 수 있음

Contents

01 Background
02 Related Study
03 Data
04 Methodology
05 Result
06 Conclusion

Conclusion & Discussion

- **Conclusion**
  - **Temperature**가 낮을수록 평가점수의 일관성이 높으며, 인간평가자와의 상관관계 또한 높게 나타남
  - 평가등급 명칭은 **High-Middle-Low**를 사용했을 때, 평가의 **Bias**가 감소함
  - **GPT-3.5** 버전에 비해 **GPT-4**의 성능이 우수함
  - **ChatGPT를 통한 자기소개서 채점이 인간평가자의 평가와 유사한 성능을 보일 수 있으며, GPT-4_HML_0.0 버전의 가장 우수한 성능을 나타냄**
- **Future Research**
  - 기존 연구의 고도화
    - **0.0 ~ 0.5** 사이의 **temperature** 단계의 세분화, 평가등급 명칭의 다양화, **GPT-4** 기반
  - 실무에 활용할 수 있는 최적 활용방안 개발
    - 최적 **Prompt**, 최적 **Rubric**, 최적 **Parameter** 등 제시

[Provider:earticle] Download by IP 168.188.35.222 at Monday, May 27, 2024 10:38 AM

Contents

01 Background
02 Related Study
03 Data
04 Methodology
05 Result
06 Conclusion

Conclusion & Discussion (Cont'd)

- **Implication for HR & Business Domain**
  - **ChatGPT**를 통해 자기소개서 평가 자동화를 통한 채용제도의 효율화 가능성
    - 채용에 투자되는 시간과 노력의 감소 및 자원 배분 효율화 가능
    - 평가 과정에서 사람의 편견이 개입될 가능성 감소, 공정성 및 객관성 제고
  - 자기소개서 뿐만 아니라 **Business** 상에서 발생하는 다양한 **Text Data**(제안서, 기획안)에 대한 평가 자동화 가능성
- **Implication for Large Language Model Research**
  - 기존 **ChatGPT** 성능에 관한 연구는 대부분 인간이 제시한 문제에 적절한 답을 제시할 수 있는가에 대해 초점을 두고 이루어진 반면, 본 연구는 인간이 작성한 답안을 적절하게 평가할 수 있는가에 초점
  - 향후 **ChatGPT** 성능 연구가 다양한 범위로 확장되어 이루어질 것을 기대함

THE END OF DOCUMENT

[Provider:earticle] Download by IP 168.188.35.222 at Monday, May 27, 2024 10:38 AM

[Appendix] Expected Questions

- **해당 데이터 샘플은 본 연구를 위해 생성된 것인가?**
  - 모 기업의 HR 인턴십 프로그램에 사용되었던 Data를 활용했으며, 인간평가자의 채점 시점은 2년전 당시로 현 시점과는 차이가 있음
- **기존에 AI를 활용해 자기소개서 평가를 제공하는 서비스와 어떤 차이가 있는가?**
  - 기존 LLM을 활용한 서비스는 과거 자기소개서 Data 기반으로 전용 LLM을 fine-tuning하여 평가
  - 본 평가에서는 별도의 fine-tuning이나 예시 제공 없이 Zero-shot으로 평가함
- **GPT-4를 연구의 Main Model로 사용하지 않은 이유는 무엇인가?**
  - 실무에 초점을 맞춘 본 연구의 취지를 고려했을 때, ChatGPT와 동일한 버전인 GPT-3.5를 활용
  - 또한, 4월 기준 GPT-4는 Beta버전이며, 시간당 응답제한 등 연구 활용에 어려움이 있음

[Appendix] Reference

- Bae, J., Chen, S., & Rowley, C. (2011). From a paternalistic model towards what? HRM trends in Korea and Taiwan. *Personnel Review,* 40(6), 700-722. doi:10.1108/00483481111169643
- 이종구. (2022). 한국 대기업의 시대별 자기소개서 형태의 전개 양상과 특징 비교분석에 관한 탐색적 연구. 취업진로연구, 12(2), 1-21.
- OpenAI. (2023). GPT-4 Technical Report [Technical report]. arXiv:2303.08774v3.
- Newton, P. M. (2023, February 21). ChatGPT performance on MCQ-based exams.
- Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. arXiv preprint arXiv:2303.15056
- Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large Language Models are Zero-Shot Reasoners. arXiv preprint arXiv:2205.11916

[Provider:earticle] Download by IP 168.188.35.222 at Monday, May 27, 2024 10:38 AM