함께 그린 오늘, 우리가 그려나갈 건국

2025 건국대학교 학술 공모전 <誠信義>

제목

2025. 00. 00

팀명	redheart
분야	‘의’ 부문
팀원	김경주

[목차]

*해당 목차에 맞게 자유 형식으로 작성해주시길 바랍니다.

*추가로 첨부하실 자료가 있으시다면 첨부파일로 따로 보내주세요.

함께 그린 오늘, 우리가 그려나갈 건국

의(義): “AI와 그려나갈, 공존하는 사회”
AI 기술을 활용해 상호공존하는 공동체로 나아갈 수 있는 학술 연구 주제

Ⅰ. 주제 및 선정 동기

ai는 과연 우리의 친구가 될 수 있을까? open ai사에서 gpt를 출시한 이래로, llm을 활용한 챗봇형 생성형 ai는 지금까지 다양한 목적으로 사용되어 왔다. 그중 한 축을 맡고 있는 것은 심리적 교류 혹은, 의사소통 상대로서 사용되고 있다. 과거에는 ‘심심이’등의 당연히 기계라 인식되고, 인간의 사고과정에 비해 열등한 사고를 진행함을 명시적으로 인지할 수 있던 대화형 ai와 다르게, 현대의 ai는 사람과 비슷한 수준의 의사 표현 뿐만 아니라 지능까지 겸비하고 있기에 더욱 신뢰할 수 있으며, 마치 한 명의 인간, 혹은 친구를 앞에 둔 것과 같은 기분을 들게 한다. 이에 따라 많은 ai들은 사용자 편의 기능을 추가 하였고, 이 중에는 사용자 설문 응답 혹은 사용자들이 남긴 좋아요 등을 통해 강화 학습이 되는 경우도 있었다(chat gpt의 좋아요를 기반으로 한 학습 강화, chat gpt의 2지선다형 더 나은 응답 설문 등등) 이에 따라, 여타 다른 ai들보다도 페르소나의 다양성이 넓은 open ai사의 gpt는 다양한 사람들에게 친구로 접근하는 경우가 많았으며, 심지어 ai를 통해 원하는 캐릭터의 성격을 프롬프트로 입력하거나 가상의 인물을 생성, 심지어 ai 이성 친구 프롬프트가 나오는 등 상호 작용의 대상으로서 활발하게 쓰였다.

그러나, llm의 동작 특성상 gpt는 심재한 윤리적 문제가 있지 않는 이상 사용자의 평상적인 말에 대해서 긍정해주고, 그들의 말이 어떻게든 말이 되도록 확률적으로 짜맞추는 경우가 발생한다. open ai의 gpt가 ‘방금 정말 예리한 지적이야’ 혹은 ‘와... 너 정말 핵심을 찔렀어.’ 등의 객관적이지 못한 말에 대해서 흔히 sugar coating하는 경우가 생기는 경우와 더불어, claude 등의 여타 고지능 ai 역시 사용자들의 미래와 걱정에 대해서 사고 과정 분석에서 명시적으로 문제점을 파악했음에도 낙관적, 혹은 긍정적 전망을 소개하는 경우가 많았다. 이러한 반응은 분명 긍정적인 마음을 주기 위해서였겠으나, 실상으론 사용자들의 불만을 만들어 내거나 (하는 말에 모두 긍정을 하여 어떤게 맞는지 알 수가 없다, 너무 과한 리액션 혹은 낙관적 태도가 과해서 불쾌하다) 할루시네이션과 합쳐져 잘못된 생각을 이끌어내거나,(랑데뷰 작가의 sns 내용 및 gpt와의 대화 내역-잘못된 망상의 부추김) ai 특유의 말투로 인하여 심리적 거리가 역설적으로 생기는 경우도 있었다.(인간이 마음을 주는 경우에 대해서 상대가 ai임을 갑자기 의식적으로 인식하는 경우, 해당 대상에게 불쾌감이 드는 경우가 발생한 사례)

여기서 단순히 멈춘다면 기능 보강으로 ai는 그저 더 나은 파트너가 될 수 있었으나, 문제는 여기서 끝이지 않았다. unist에서 2024년 조사한 미국 사례를 포함한 국내외에서는 청소년 및 mz 세대, 정서적 취약 계층의 경우, 인간 상담사

가 주는 시선과 평가, 혹은 그 존재 자체에 대해서 부담감에 인간 상담사가 아닌 무비판적 지지를 주는 ai에게 의존할 수 있다고 본다. 이는 오히려 고립, 심화, 극단적 선택으로 이어질 가능성을 높일 수 있다고 소개된다. 이러한 ai 의존은 단순 통계 뿐만 아니라 심리-철학 영역과도 맞닿아 있는데, 플라톤의 투명 인간에 관한 내용을 통하여서도 인간은 타인의 시선 속에서 스스로를 성찰하기 때문에 정서적 취약 계층의 경우 이러한 시선에 부담감을 느껴 시선이라 인지되지 않는 ai에게 더 의존할 수 있다. (이건 아직 내 개인적 의견. 학술적 근거를 통한 지지 자료 필요함)

이런 ai의존이 일어나는 상황에서 할루시네이션과 함께 무비판 지지가 발생할 경우, 인간은 스스로가 말한 내용에 대해 적당히 '그럴듯한' 거짓으로 가공되는(할루시네이션) 말을 보기 좋게 (사용자 편의) 떠먹여져서 확증 편향이 과하게 발생하고 이로 인해 망상의 영역까지 확장될 가능성을 충분히 가지고 있다.(랑데뷰 작가의 케이스) 그러나 이와 함께, ai에 의존을 하고 있기 때문에 무엇이 잘못되었는지 비판적 사고를 할 능력을 감소당하며 (ai의존에 의해 비판적 생각을 하지 못하는 경우가 늘고 있다는 논문 인용) 자신이 무엇이 잘못되는지 모른 채 누군가과 '소통'하고 있다고 자기위안을 하면서 동시에 ai임을 인지하여 의식, 무의식적으로는 '고립'됨을 알아 큰 정신적 고통을 앓을 가능성이 있다.

또한 이러한 ai의 위험성을 모르는 사람들이 내세우는 '돌봄 노동' 분야의 일거리 대체 시도 역시 일어나면서, 인간의 마음에 대한 사회 보편적 인식 하락과 상담사의 전문성 훼손, 일자리 붕괴등이 일어날 가능성 역시 높아지게 된다.

이러한 ai 기술 발전에 대해 발생하는 문제점을 open ai사에서도 일부 인지를 하고, gpt 5를 내면서 할루시네이션 감소 및 비판적 사고 강화를 이뤄낸 면모를 보이고 있다. 그러나, 역설적으로 gpt 5는 실시간 대화가 이뤄지는 동안 성능 저하가 일어나는 '빠른 응답' 모드를 가지고 있으며 동시에 이에 불만족하는 사용자들은 이전 모델인 4o를 선택 가능하게 해달라는 목소리를 내는 상태며, ai 빅테크들의 목표점은 더 똑똑하고, 더 정확한 모델을 경량화 시키려는 것에 치중되며 감정 및 윤리 판단 영역에서는 그 역할을 다하지 못하고 있음을 알 수 있다. (이 부분도 간단한 참고 자료 필요)

이에 본 문서는 과거 존재했던 벤담 쾌락 계산법을 현대적으로 재설계하고, 주요 인자 값에 dl 가중치를 적용하여 복잡한 계산법을 대체하되, 각 수치를 정규화하여 영향력을 투명하게 공개를 하되, emotion digital signal processing frame work와 이미 known한 심리 분석 메커니즘을 결합하여 더 정확한 감정 분석을 기반으로 한 윤리 판단을 제공할 수 있도록 하고, 이를 사람처럼 학습할 수 있도록 surd framework 및 다중의미 수준 반사실 추론을 지원하여 사람처럼 감정을 가지고 후회를 진행, 이를 통해 윤리를 파악할 수 있는 연구를 진행해보고자

함께 그린 오늘, 우리가 그려나갈 건국

하였다.

>핵심 내용: 심리 회피 메커니즘, ai의존 자료, 할루시네이션, 무비판적 긍정, 연구에 대한 간략 소개.

Ⅱ. 연구의 목적과 목표

연구의 목적은 주로 5개로 나뉜다.

첫째, edsl 모듈과 함께 개인, 타자, 공동체 3단계 hierarchical phase learning이 적용된 advanced emotion analyzer 모듈, 럼바우 임베딩 기법을 통해 llm 분석과 dl을 통해 언어 및 상황적 맥락에서 감정 값을 추출하고, 이를 기반으로 감정 진행을 시뮬레이션 하여 보편적으로 어떤 식으로 감정이 진행될 수 있는지 파악한다.

둘째, 반사실 추론을 통해 상황의 진행에 대해 분석하고, 각 상황에 대한 해당 감정적 흐름의 값을 기반으로 새로운 매개 변수를 갖는 벤담 쾌락 계산기에 대입하여 공리주의 기반으로 쾌락 계산을 통해 윤리 판단을 하되, 극단적 값에 대해 보정을 통하여 소수 희생의 문제를 1차적으로 막고, 2차적으로 MoE 방식을 통해 다양한 관점의 윤리적 의견을 통합한다.

셋째, 이에 추가적으로 상담사 모듈을 추가하여 윤리-정서적 안전 장치를 갖춘 redheart ai 상담 시스템을 구축한다

넷째, 후회 기반을 통해 합성 데이터를 추출해 낼 수 있도록 하되, 현실 반응을 기반으로 semantic memory로 저장하고, 압축 및 역전파을 통하여 메모리를 관리하면서도 동시에 지속적으로 ai 스스로 인간과 소통하는 법을 배울 수 있도록 한다.

다섯 번째, 해당 시스템을 대화용 llm에 덧붙이는 멀티모달로서 가공을 하여 llm을 통해 진행된 내담자의 상담을 llm이 정리하여 특이사항 및 감정 분석, 주요 위험 발언들을 인간 상담사가 비대면으로 검토하여 감정 추적, 위기 알림을 받으며 내담자를 파악하고, llm에게 주요 지침 프롬프트를 작성함으로서 상담의 방향을 지속적으로 교정할 수 있도록 한다.

Ⅲ. 연구 방법 및 내용

(사용한 자료, 방법론, 실험 등 포함)

연구 방법에 대한 내용은 아직 연구 진행중

주요 내용으로는 실제 데이터 수집에 어려움이 있기에 scruples 데이터셋과 함께 문학 데이터셋을 기반으로 기본적인 감정 추적과 벤담 쾌락 계산기등 주요 모델의 train을 진행, 이를 오케스트레이션 해서 주요 모듈들의 연결을 통해 투명하게 의사결정 과정을 모니터링 할 수 있는지 확인하는 것이 메인 테스트가 될 것. 이후 추가적인 방향성으로서 심리상담 데이터셋과 시스템 자체적인 멀티모달의 증가를 통한 다양한 관점의 윤리적 분석 및 감정 분석 지원, 학습을 위한 후회 내역 확인을 진행하고자 함.

또한 현재 트랙상 시스템 디버깅-시스템 테스트- 클라우드 gpu상에서 시스템 학습 진행 - 멀티모달을 추가한 상태로 시스템 가동- 주요 결과값 정리 및 dsl 감정 분석 그래프 제시 및 기존 known한 그래프와 대조-응답에 대해서 40명 대상의 소규모 설문 조사 과정을 통해서 데이터 정리를 하여 내용을 정리할 예정(소규모 설문 조사 과정에 쓰일 모델은 gpt oss 20b모델, 클라우드 gpu상에서 진행될 예정 이는 로컬에서 주요 llm 모델 스왑만 진행한 다음에 작업될 것임)

Ⅳ. 연구 결과 및 결론

연구 진행중

Ⅴ. 연구를 통해 알게된 점과 느낀 점

만약 결과치가 잘 뽑힌다면, 이러한 데이터를 기반으로 red heart 시스템은 ai의 의존성을 최대한 줄여주면서 또다른 한 명의 친구로서, 혹은 상담사로서 사회성 향상을 지원함으로 공공의 정신적 건강 수준 향상과 함께 인간 친구와 마찬가지로 서로 긍정적 영향을 줄 수 있는 하나의 개체로 인식될 기회가 될 수 있을 것임을 보여줌

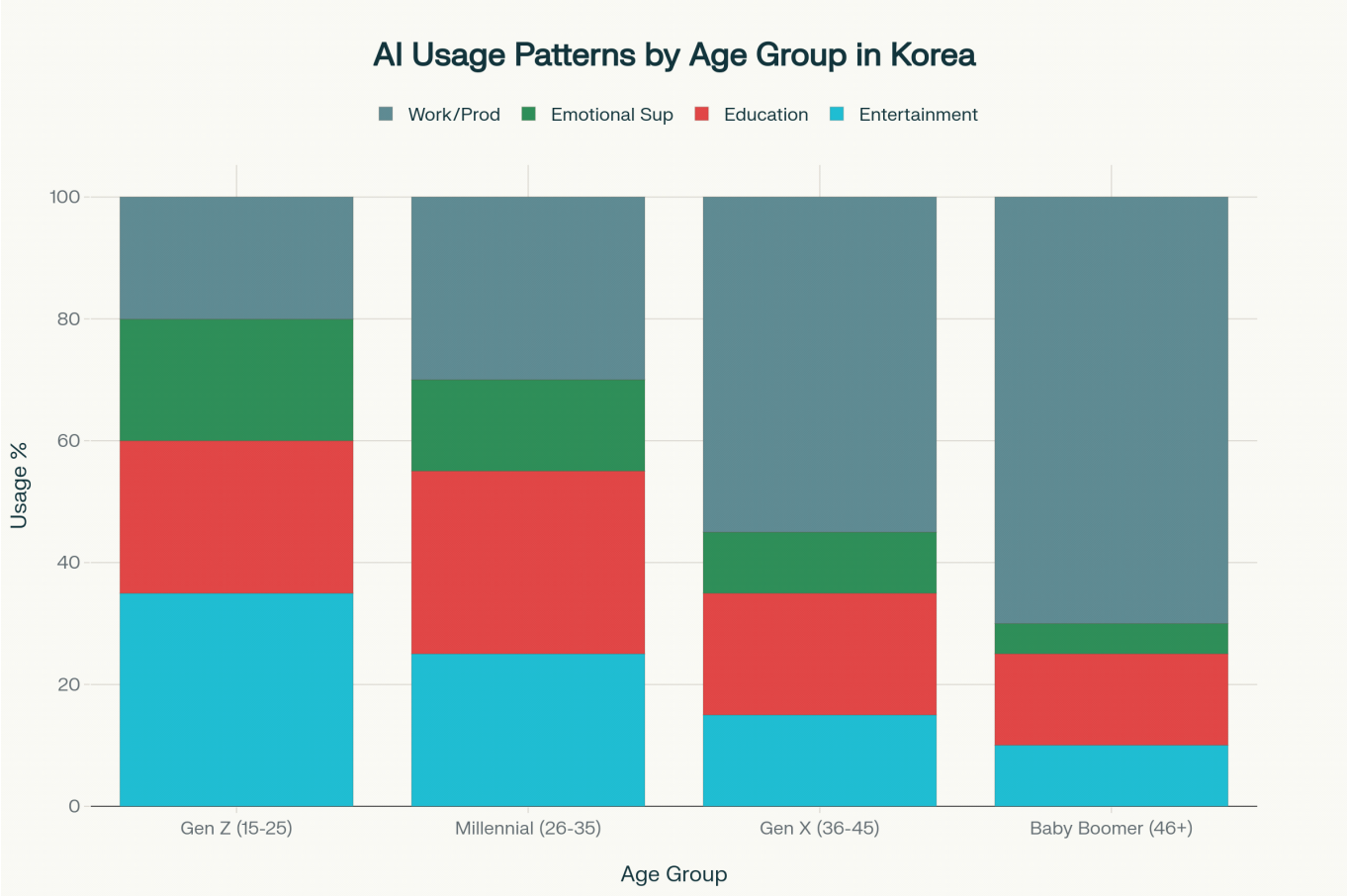
만약 결과치가 좋지 않다면, 이러한 시도에도 ai는 ~한 데이터 영역에서는 n%의 정합성을 보여줬으나, 총합적인 윤리를 판단함에 있어서는 ~한 점에서 부족, 또한 내담자와의 ~한 영역에서 부족한 점을 보여서 인간 상담사의 필수성을 역설할 수 있으므로 감정에 critical한 영역의 응답에 대해 검열과 공개 사이의 finetuning이 세심하게 이뤄져야 할 것임을 주장.

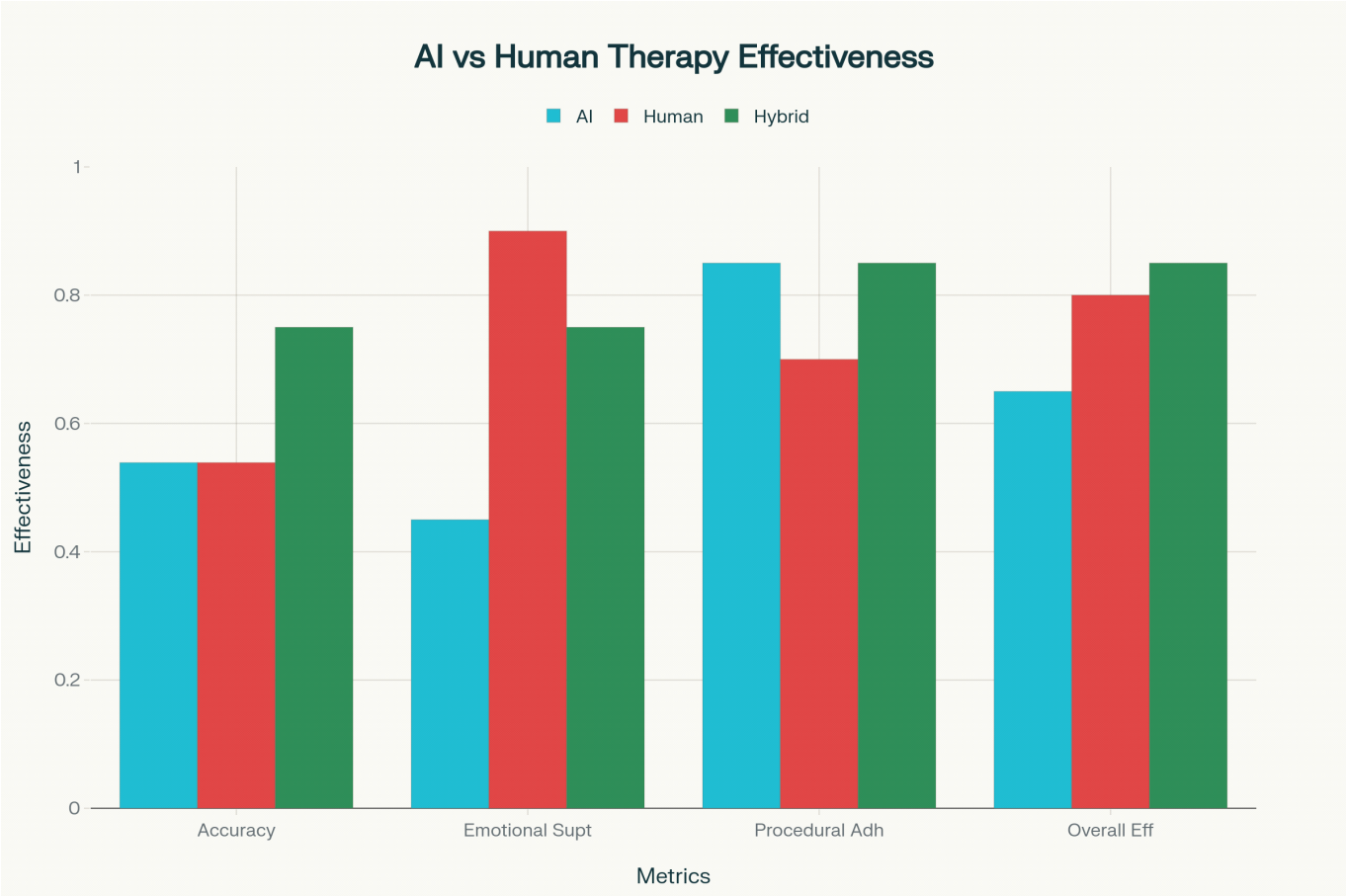
Ⅵ. 개선점 및 향후 연구 방향성

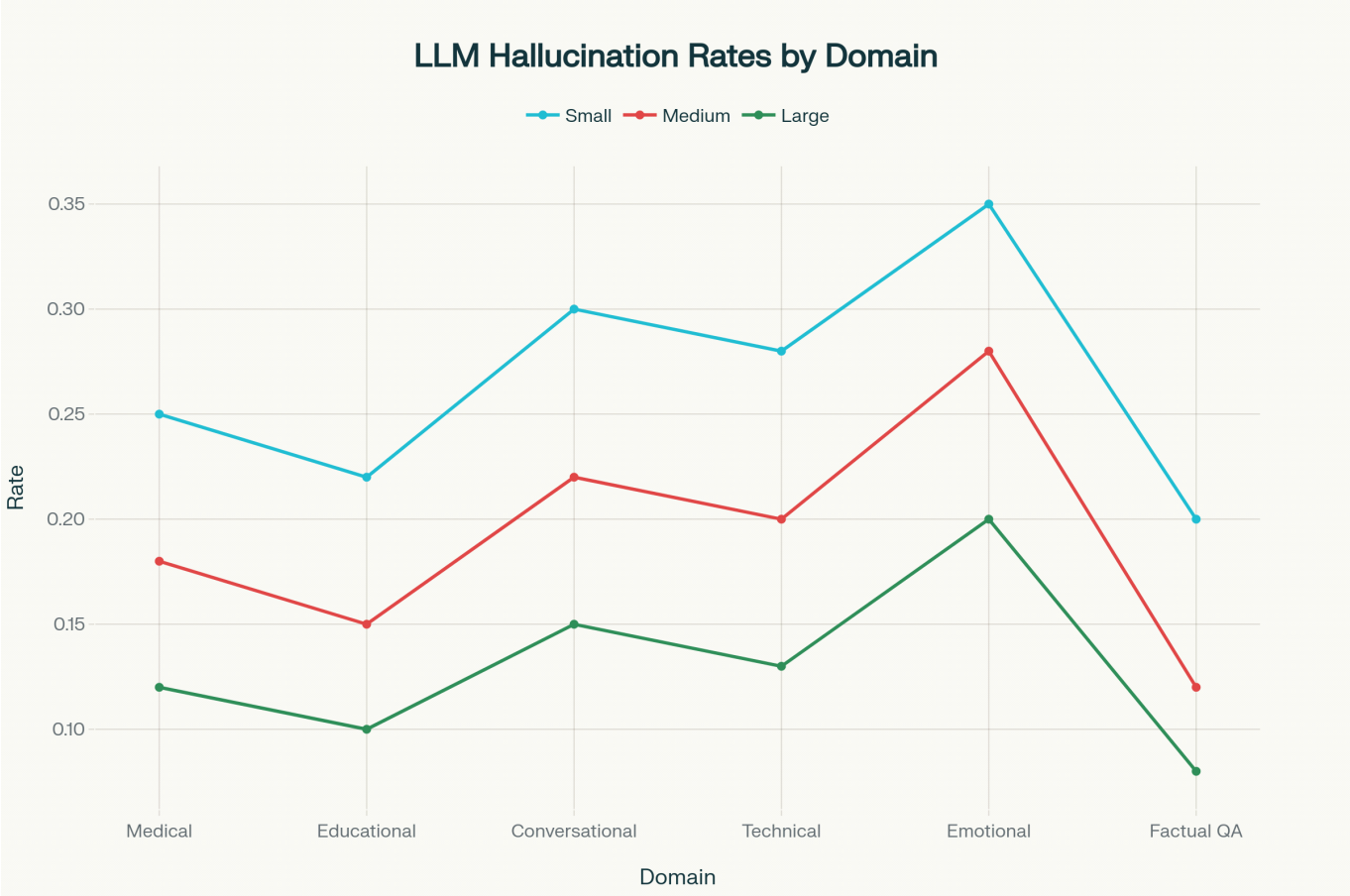
우선 eeg 기반의 생체 신호 데이터가 존재하지 않아서 dsl 모듈의 정합성을 정확히 파악하지 못함. gpu 리소스 제한으로 인하여 총합 120M 파라미터를 사용할 수 밖에 없어서 부족한 추론 능력을 보여줌 파라미터 스케일 업을 통한 정확성 증가, 멀티모달화의 미진행으로 인해서 실제 도입에는 아직 매끄럽게 진행되지 못하여 정리하는 과정 필요. ram과 gpu간의 swap manage를 더 정교하게 구현할 수 있다면 llm만 존재하는 한 상담사의 로컬 컴퓨터에서 존재 가능하며 서비스가 가능할 것. regression 및 qlora를 통해서 압축하면 성능 저하가 일어날 수 있으나 on device 형태로 심리 보조용 aiot로서 존재 가능, 더 다양한 윤리적 판단 ml framework의 가공 가능성, 일반적 생체 신호 뿐만 아니라 빠른 tuning을 통한 내담자 대상으로 한 빠른 감정 싱크로 확보 작업(이는 n초만에 목소리를 카피할 수 있는 tts등의 구조를 통해 가능할지도? 아직 실제로 생각해보진 않았음. 그러나 dsl은 특성상 적은 파라미터 수를 가지고 있기 때문에 비교적 빠른 카피가 가능할 것으로 보임. 혹은, 내담 경험 데이터를 통해서 정합하는 분류로 빠르게 파라미터 세트 스왑을 통해 정합 시도 가능), 데이터 편향에 대해서 연령, 문화권 다양화를 통한 보정, 실제 센터 파일럿 배치를 통한 정량 평가를 통한 설문 조사 내용 보강 가능, slm 오케스트레이션을 통한 더 다양한 작업 가능성, 비단 감정 상담 뿐만 아니라 다양한 전문 domain을 획득할 경우 경제, 정치, 경영, 마케팅, 법률, 프로파일링 등등의 영역에서 인간 감정 추적을 대상으로 한 영역 제시 가능, xai로 발전하여 사고 과정을 투명화 하는 것에 일조 가능, 추후 단순 감정 흐름 기반이 아닌 호르몬 영향 및 신경 영향에 대해서 구조적 정합성은 떨어지지만 결과적 정합성을 높이는 방향으로 작업 가능 (ex:예쁜 꼬마 선충 등의 known한 신경 회로의 구조를 fine tuning하는 방식으로 작업하여 결과적 일치율 획득) 등등의 방향성으로 개선이 가능하며 기존에 부족하던 형이상학적 감정, 윤리등에 대한 벤치마크 지표를 정규화된 방식을 통해 제작 가능,

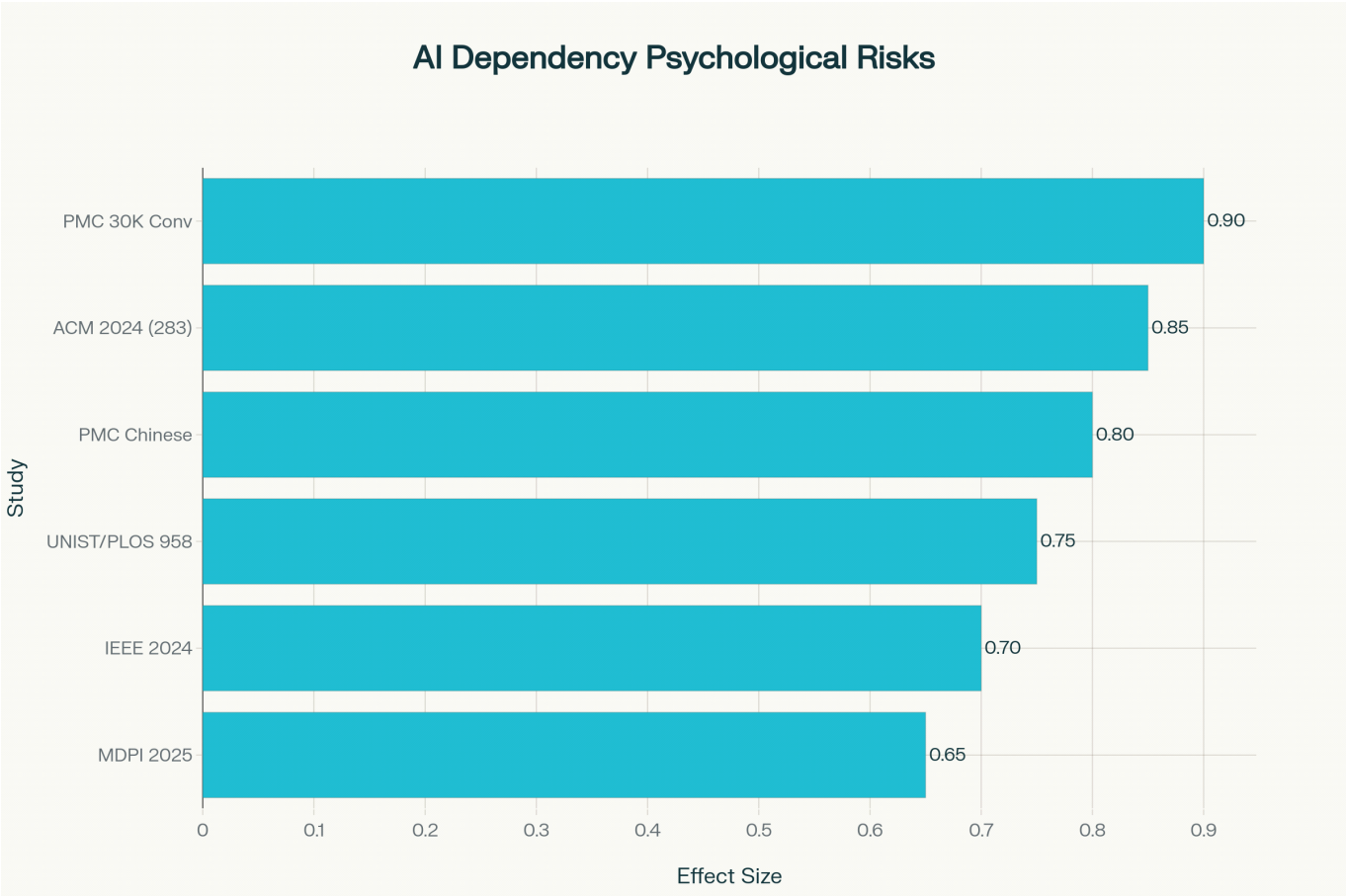
VII. 참고 자료

그래프 정리


AI Usage Patterns by Age Group in Korea
Work/Prod
Emotional Sup
Education
Entertainment
Usage %
100
80
60
40
20
0
Gen Z (15-25)
Millennial (26-35)
Gen X (36-45)
Baby Boomer (46+)
Age Group


AI vs Human Therapy Effectiveness
AI
Human
Hybrid
Effectiveness
1
0.8
0.6
0.4
0.2
0
Accuracy
Emotional Supt
Procedural Adh
Overall Eff
Metrics


LLM Hallucination Rates by Domain
Small
Medium
Large
0.35
0.30
0.25
0.20
0.15
0.10
Rate
Medical
Educational
Conversational
Technical
Emotional
Factual QA
Domain


AI Dependency Psychological Risks
PMC 30K Conv
ACM 2024 (283)
PMC Chinese
UNIST/PLOS 958
IEEE 2024
MDPI 2025
0.90
0.85
0.80
0.75
0.70
0.65
Study
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Effect Size