



# ABLU-GPT

지도교수



조현석 교수님  
소프트웨어학과

팀장 & 팀원



김희규  
소프트웨어학과

---

# GPT(Generative Pretrained Transformer)

OpenAI에서 학습한 글을 생성하는 AI언어모델  
ChatGPT는 GPT-3.5와 GPT-4를 기반으로 하고있다.



---

# Pretrained Transformer



어린이에게 처음부터 모든 것을 학습시키는 것 보단  
이것저것 공부한 학생에게 가르치는 것이 훨씬 뛰어나다.

# 한국어 GPT vs 영어 GPT-3 크기 비교

## 한국어

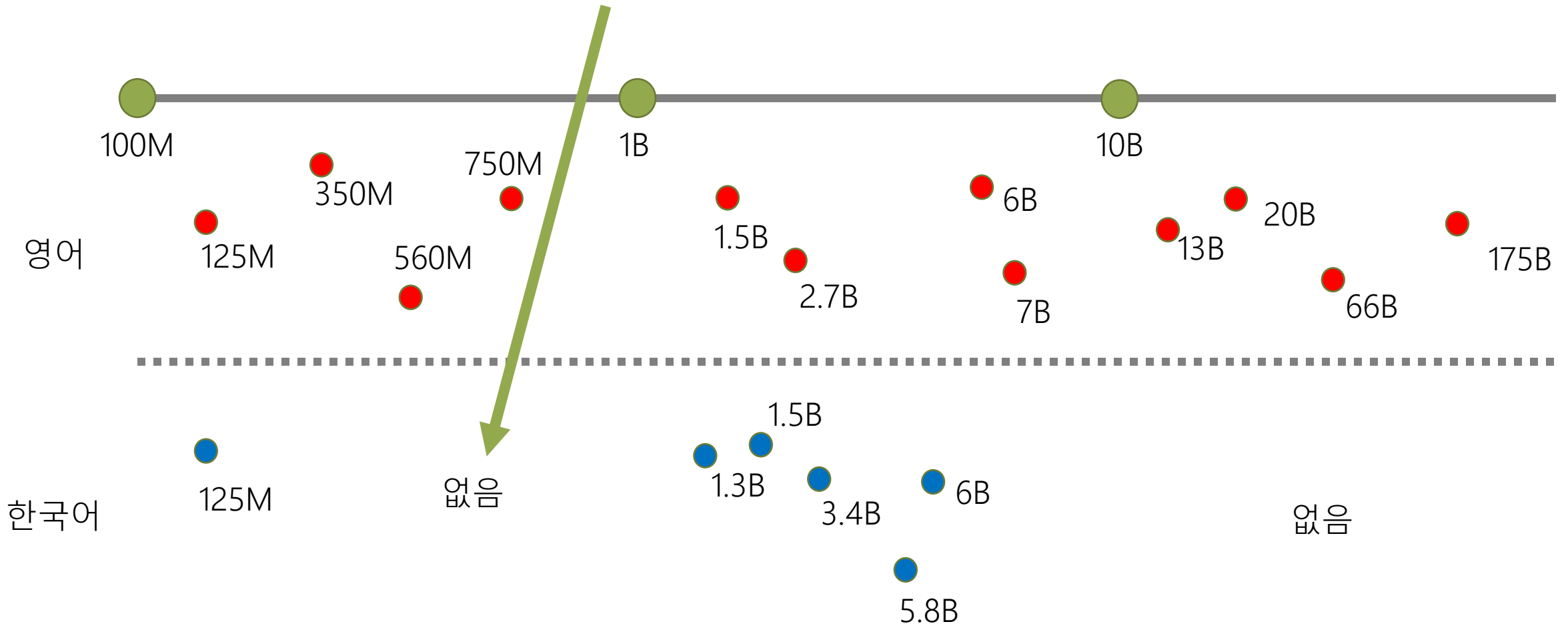
- SKT KoGPT (125M, 1.3B)
- EleutherAI Polyglot (1.3B, 3.4B, 5.8B, 12.8B)
- Kakao KoGPT (6B)
- Naver HyperClova (88B)

## 영어

- OpenAI GPT-3 (175B)
  - Meta OPT (175B)
  - Google PaLM (540B)
- 등 등...

\*1B = 1000M = 10억 개의 파라미터

한국어 공개 GPT 모델은 너무 작거나, 너무 크다  
따라서 중간 크기(350M)를 학습해서 공개하자!



\*1B 이상의 모델을 사용하기 위해선 RTX 3090 이상의 GPU필요(250만원)



# 공개 데이터 수집

- AIHub, 국립국어원, 위키백과, 나무위키, 청와대 국민청원 등
- 총 50GB

출처	이름	유형	용량
NIKL	신문 말뭉치 2021 v1.0	문어	2.9 GB
NIKL	신문 말뭉치 2020 v1.1	문어	2.21 GB
NIKL	신문 말뭉치 2009 ~ 2018 v2.0	문어	15.6 GB
NIKL	온라인 대화 말뭉치 2021 v1.0	구어	835 MB
NIKL	일상 대화 말뭉치 2020 v1.2	구어	317 MB
NIKL	일상 대화 말뭉치 2021	구어	595.8MB
NIKL	구어 말뭉치 v1.2	구어	6.37 GB
NIKL	메신저 말뭉치 v2.0	구어	20.4 MB
NIKL	문어 말뭉치 v1.2	문어	4.24 GB
NIA	대규모 웹데이터 기반 한국어 말뭉치	문어	8.95 GB
NIA	온라인 구어체 말뭉치	구어	1.59 GB
NIA	용도별 목적대화	구어	109 MB
NIA	주제별 텍스트 일상 대화	구어	256 MB
NIA	한국어 SNS	구어	697 MB
NIA	전문분야 말뭉치	문어	13.63 GB
NIA	방송 콘텐츠 대본 요약	구어	238 MB
NIA	요약문 및 레포트 생성	문어	472 MB
NIA	한국어 뉴스 기사 요약	문어	100 MB

train/loss

gpt-j-350m-v1-lr3e-4-batch512-bf16-rev4

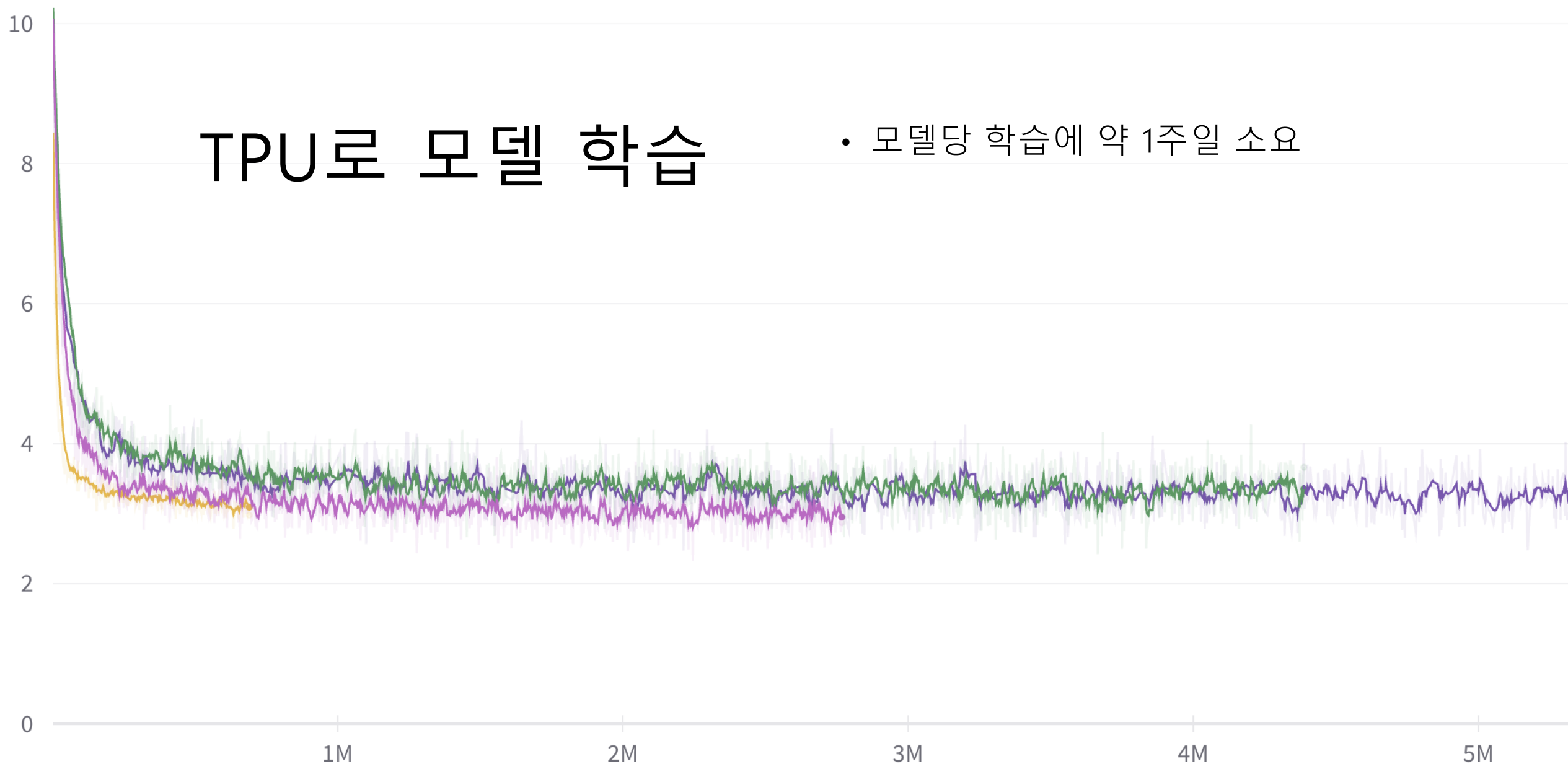
ajoublue-gpt2-medium-v1-lr3e-4-batch512-bf16-rev4

ajoublue-gpt2-base-v1-24L-lr5e-4-batch512-rev3

gpt-j-base-v1-24L-lr5e-4-batch512-rev3

# TPU로 모델 학습

- 모델당 학습에 약 1주일 소요





# Timeline

기간	진행
2022/12	GPT-NeoX 13M, 52M PoC (1 GPU)
2022/1	GPT2, GPT-J 125M, 210M TPU 학습
2022/2	125M, 210M 평가 350M 모델 학습 시작
2022/3	350M 모델 평가 이후 125M, 210M, 350M 공개
2022/4	775M, 1.3B 학습 시작

# 평가

		NSMC	KLUE-YNAT	KLUE-STS	대화(Top-P)		요약(Beam)
Name	# params	Acc	Macro F1	F1	PPL	Rouge-1	Rouge-1
skt/kogpt2-base-v2	124M	89.1	83.88	83.82		0.1648	0.41133
kogpt-j-base-rev4	124M	88.88	86.58	81.69	37.44	0.2272	0.4052
ajoublue-gpt2-base	125M	89.69	86.17	80.78	41.447	0.2256	0.4158
kogpt-j-base-24L	209M	90.51	<b>86.91</b>	81.9	35.524	0.2265	0.4232
ajoublue-gpt2-base-24L	210M	<b>90.74</b>	86.62	<b>85.71</b>	36.442	0.2262	0.4253
kogpt-j-medium	304M	90.31	86.64	83.76	34.492	<b>0.2287</b>	0.4277
ajoublue-gpt2-medium	355M	90.34	86.59	81.48	<b>34.101</b>	0.2286	<b>0.4311</b>

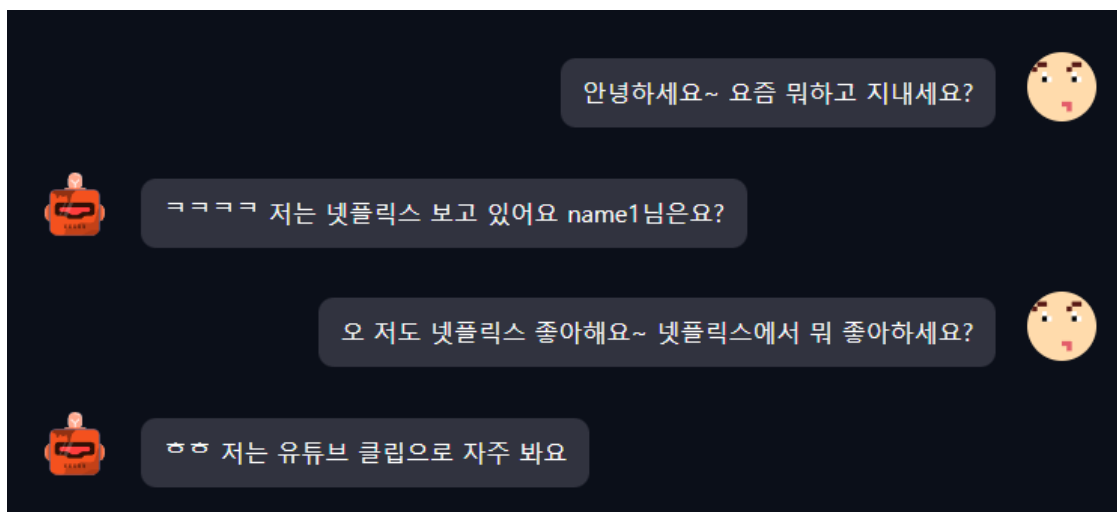
텍스트 분류(NSMC, YNAT), 문장 유사도 평가(STS)  
대화응답 생성, 요약문 생성 등 다양한 평가에서  
SKT의 124M 모델 대비 모두 우수한 지표를 보인다.

---

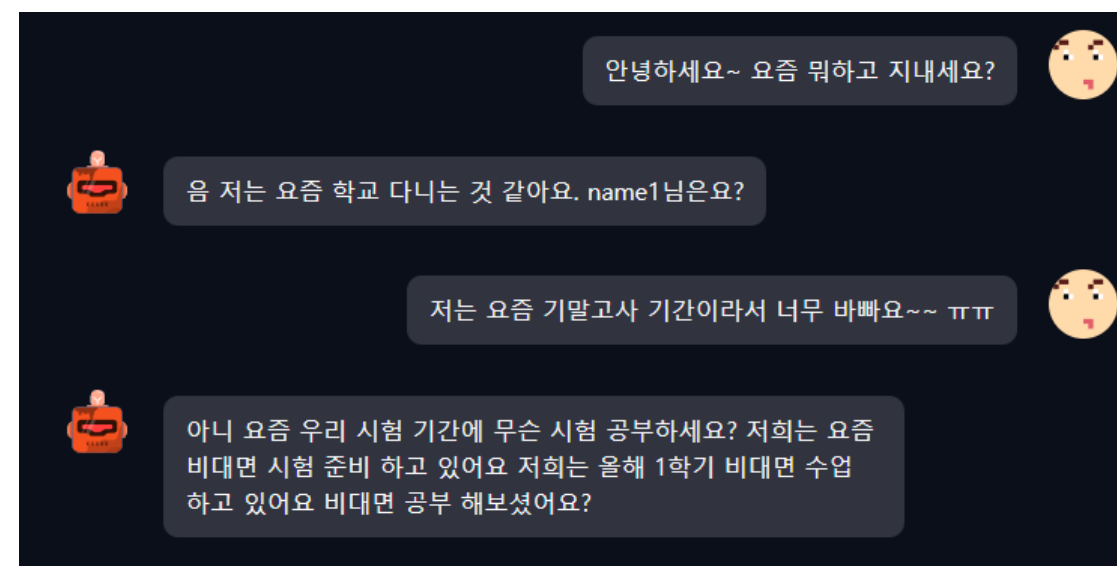
# 학습된 모델과 친구처럼 대화해보자

공개 대화데이터 10만건으로 재학습





SKT KoGPT(124M)



AjouBlue-GPT(355M)



SKT KoGPT(124M)



AjouBlue-GPT(355M)

---

# 학습된 모델을 ChatGPT처럼 써보자!

크기는 500배 작지만(...)





아래 뉴스는 무엇에 대해 얘기하고있나요?

국내 주요 금융그룹들이 일본 현지법인 영업 영역을 확대하고 일본 대형 금융회사와의 협력을 강화하기로 했다. 최근 한일 정상회담을 계기로 양국관계가 '해빙무드'로 접어들고 있는 상황에서 금융사들도 나서고 있다.

4일 국내 금융권 및 니혼게이자이신문(日本經濟新聞)에 따르면 KB·신한·하나금융그룹은 최근 일본 현지 영업 및 현지 금융사와의 협력을 강화하고 있다. 은행 현지법인을 보유한 신한금융은 영역 확장을, 그렇지 않은 KB·하나금융은 일본 대형 금융회사와의 협력을 강화하는 방식이다. 니혼게이자이는 "금융 분야의 한일관계는 외교보다 먼저 밀접한 관계를 구축하고 있다"고 평가했다.

KB·신한·하나금융은 일본 현지법인을 통해 일본 대형 금융회사와의 협력을 강화하고 있습니다. 이에 따라 KB금융과 신한은행은 일본 현지 은행과 금융 제휴를 맺고 있고, 하나금융은 일본 대형 은행과 금융 제휴를 맺고 있습니다. 그리고 일본 금융회사와의 협력도 강화하며, 최근 한일정상회담이 개최되어 한일 관계 개선이 기대되고 있습니다

금융 분야는 어떻게 되어나요?

국내에서는 일본 현지 은행을 통해 금융 제휴를 맺고, 일본 대형 금융사와의 제휴를 통해 협업을 강화합니다. 일본에서도 금융은 정치보다 먼저 밀접한 관계를 구축하고 있습니다

앞으로 양국 관계가 어떻게 개선되어갈까요?

양국관계는 외교보다 먼저 밀접한 관계를 구축하고 있습니다

해빙무드라는 표현에 대해 잘 모르겠는데 알려주세요

해빙무드는 긍정적인 관계가 형성되는 과정을 설명합니다. 예를 들어, 국제 스포츠 연맹이 서로 협력하여 국제적인 경쟁에서 이길 수 있는 방법을 만들어내는 것입니다

아래 뉴스를 요약해줘

이번에 통과된 법안은 세계에서 가장 엄격한 반동성애 법안 중 하나로 평가받는다.

인체면역결핍바이러스(HIV) 감염자나 미성년자가 동성 간 성관계를 할 경우에는 사형까지 선고할 수 있으며, 동성애를 '조장'할 경우 최대 징역 20년이 선고된다. '동성애 미수' 범죄에는 최대 징역 10년이 선고된다.

성소수자(LGBTQ)로 확인만 되면 처벌하도록 하는 조항과 동성애 의심 행위 신고를 의무화한 조항은 삭제됐다. 또 동성애로 추정되거나 의심되는 사람이라도 동성애자와 성행위를 하지 않는 사람은 처벌 대상이 아니다.

다만, 아동을 비롯한 사회적 약자와 관련된 동성애 행위의 경우에는 신고하지 않을 경우 징역 5년에 처할 수 있도록 했다.

이번 입법으로 인하여 HIV, HIV/AIDS, LGBTQ, 동성애 의심 행위 신고 의무화 등 인권 및 기타 중요한 인권 문제에 대한 논의가 이루어지고 있습니다. 이에 대한 우려도 제기됩니다.

국제적으로 논란이 될 것 같은데, 국제인권단체들이 개입하지 않을까? 너의 생각을 말해.

그렇습니다. 국제인권단체들은 동성애의 폭력성과 성적 지향을 용인하는 것에 대해 우려하고 있습니다. 인권보호와 권리에 대한 지속적인 관심이 필요합니다

여기서 동성애를 조장하는 것은 무슨 의미인지 알려줘

네, 알겠습니다. 다음은 HIV, HIV/AIDS, LGBTQ, 동성애 의심 행위 신고 의무화 등 인권 및 기타 중요한 인권 문제에 대한 검토에 관한 것입니다. 이 연구영역을 통해 인권과 관련된 이슈에 대한 추가적인 연구를 수행하고, 이러한 문제를 해결하는 데 대한 세부적인 전략을 마련할 것입니다

아래 사건에서 의사는 왜 집행유예를 받았지?

음주 상태에서 시속 195km 속도로 터널을 달리다 앞서가던 경차를 들이받은 치과의사가 징역형의 집행유예를 선고받았다. 5일 연합뉴스 보도에 따르면 이날 창원지법 형사2단독(신동호 판사)은 위험운전치상 등 혐의로 기소된 50대 치과의사 A씨에게 징역 1년에 집행유예 2년을 선고했다고 밝혔다. 재판부는 A씨에게 벌금 100만원과 40시간의 준법 운전 강의 수강도 함께 명령했다. A씨는 지난해 10월 오후 2시30분께 혈중알코올농도 0.074% 상태로 자신의 포르쉐 파나메라 승용차를 몰고 경남 창원시 한 터널 안을 달리다 앞서가던 스파크 차량을 들이받은 혐의로 기소됐다. 조사 결과, 사고 당시 A씨는 제한속도 시속 80km를 115km 초과한 시속 195km의 속도로 주행했던 것으로 드러났다. 이 사고로 스파크 차량 운전자 30대 B씨는 흉골 골절 등 피해를 봤다. 재판부는 "피해자의 상해 정도가 크고 A씨의 음주운전 처벌 전력이 있으며 범행의 위험성이 큰 것으로 보인다"고 지적했다. 이어 양형 이유에 대해 "다만 피해자와 합의했고 피고인의 나이와 환경 등을 종합적으로 고려했다"고 설명했다.

해당 사건은 피고인이 음주 상태에서 운전한 잘못에 비해 처벌 전력이 낮았기 때문에 실형이 선고됐다. 이에 대해서는 "피고인은 피해자와 합의했고, 피고인의 나이와 환경 등을 종합적으로 고려했다"며 "양형 이유 중 피해자와 합의했고, 피해자와 합의했고, 피고인의 나이와 환경 등을 고려했다"고 설명했다. 피해자는 합의했으며 피해자의 나이가 만 21세이며, 피해자는 합의한 대로 피해자에게 사과를 한 것으로 알려졌다. 재판부는 "피고인은 음주 상태에서 운전한 죄로 벌금 100만원과 40시간의 준법 운전 강의 수강을 명받았다. 그러나 피해자는 합의한 대로 피해자 A씨의 처벌을 원하지 않는다는 뜻을 전했다"며 "피고인의 양형에 고려하지 않았다"고 판시했다

한국에서 음주운전은 더 엄격하게 처벌되지 않는건가?

한국에서는 음주운전은 더 엄격하게 처벌되지 않는 것이 사실입니다. 한국에서 음주운전을 하는 것은 매우 위험하지만, 이 부분은 사람마다 다르게 적용됩니다

치과의사라서 형량이 낮았던 건 아닌가?

치과의사가 음주 운전으로 벌금형을 받은 건 이번이 처음입니다. 하지만 이 부분은 조금 문제가 있습니다. 한국의 음주 운전 처벌 수위가 낮은 것은 아닙니다. 한국에서는 음주운전 사고를 내고 가해자에게 벌금형을 내리는 것이 일반적인 처벌이며, 이는 법적으로도 처벌됩니다

분산시스템에 대해서 설명해줘

분산 시스템은 컴퓨터를 서로 다른 데이터 저장소로 나누거나 저장 공간을 서로 다른 데이터 공간으로 분할하는 방법입니다

왜 사용하는거지?

분산 시스템은 데이터의 양과 데이터를 저장하는 방식에 큰 영향을 미치기 때문입니다

실제로도 많이 쓰이는지 궁금한데, 어떤 서비스들에서 쓰고있지?

분산 시스템은 서버에서 데이터를 처리하는 방식이며, 특정 시스템과 사용 방법은 다양합니다. 분산 시스템은 다른 시스템이나 시스템에서 데이터를 저장하고 서로 통합할 수 있기 때문에 일반적으로 사용되는 방법입니다

어떤 기업들이 분산시스템을 활용하는거야?

분산 시스템은 데이터에 대한 액세스를 줄여 성능과 안정성을 개선하기 위한 것입니다. 하지만 많은 기업들은 데이터를 저장하고 처리하는데 분산 시스템을 사용하지 않는 경우도 있습니다

Textbox

클리어

어버이날에 대해서 궁금해

어버이날에 대한 재미있는 이야기를 알고 싶습니다

응

어버이날은 아버지의 은혜에 감사하는 날입니다. 사람들은 모두 축하하며 생일잔치에 참여합니다. 이러한 연유로 부모님은 자녀들에게 특별한 의미가 됩니다

어버이날은 언제지?

어버이날은 5월 첫째 주 토요일입니다

어린이날도 있나?

어린이날은 5월 5일부터 6일까지입니다

Textbox

|

클리어



Chatbot

콜라는 몸에 좋은지 알려줘

콜라는 일반적으로 건강에 좋은 것으로 간주되며 신체가 수분을 유지하고 혈액을 깨끗하게 유지하는 데 도움이 되는 항산화제 및 기타 유익한 화합물을 포함합니다

가격은 어느정도하는지 알려줘

가격 표에 따라 다르지만 평균적으로 콜라 1상자 당 약 2000~3000원 정도입니다

Textbox

클리어



Chatbot

발표를 잘 들어주셔서 감사하다는 말을 멋지게 표현해줘

기꺼이 도와주셔서 감사합니다

좀 더 길고 성의있게 해줘

좋습니다! 더 많은 세부 사항을 알려주시겠어요

감사합니다

