언어(Language) + 상식(Commonsense)

고려대학교 인공지능학과
이상근

인공지능의 역사

4,500,000,000

vs.

3,000,000

vs.

200,000

vs.

70

vs.

12

vs.

7

4,500,000,000 Earth

vs.

3,000,000 Humans

vs.

200,000 Homo Sapiens

vs.

70 AI

vs.

12 Deep Learning

vs.

7 Transformer

4,500,000,000 Earth

vs.

3,000,000 Humans

vs.

200,000 Homo Sapiens

인지혁명(언어) 70,000

vs.

70 AI

vs.

12 Deep Learning

vs.

7 Transformer

인류의 지식혁명

인공지능 현주소 LLM (e.g. ChatGPT)

- 15세기 인쇄술 이후, 최대의 지식혁명

인공지능 현주소 AI for Science in the Era of LLM

- 15세기 인쇄술 이후, 최대의 지식혁명

- **뉴럴모델, 심층학습 (딥러닝)** (2012~)
- AlphaZero (@Science 2018)
- Halicin (@Cell 2020)
- AlphaFold (@Nature 2021)
- DM21 (@Science 2021)
- AlphaDev (@Nature 2023)
- AlphaMissense (@Science 2023)
- GNoMe, Coscientist (@Nature 2023)

인공지능 현주소 AI for Science in the Era of LLM

- 15세기 인쇄술 이후, 최대의 지식혁명

- **뉴럴모델, 심층학습 (딥러닝)** (2012~)
- AlphaZero (@Science 2018)
- Halicin (@Cell 2020)
- AlphaFold (@Nature 2021)
- DM21 (@Science 2021)
- AlphaDev (@Nature 2023)
- AlphaMissense (@Science 2023)
- GNoMe, Coscientist (@Nature 2023)

인간이 이해하거나 설명할 수 없는 발견

인공지능 현주소 AI for Science in the Era of LLM

- 15세기 인쇄술 이후, 최대의 지식혁명

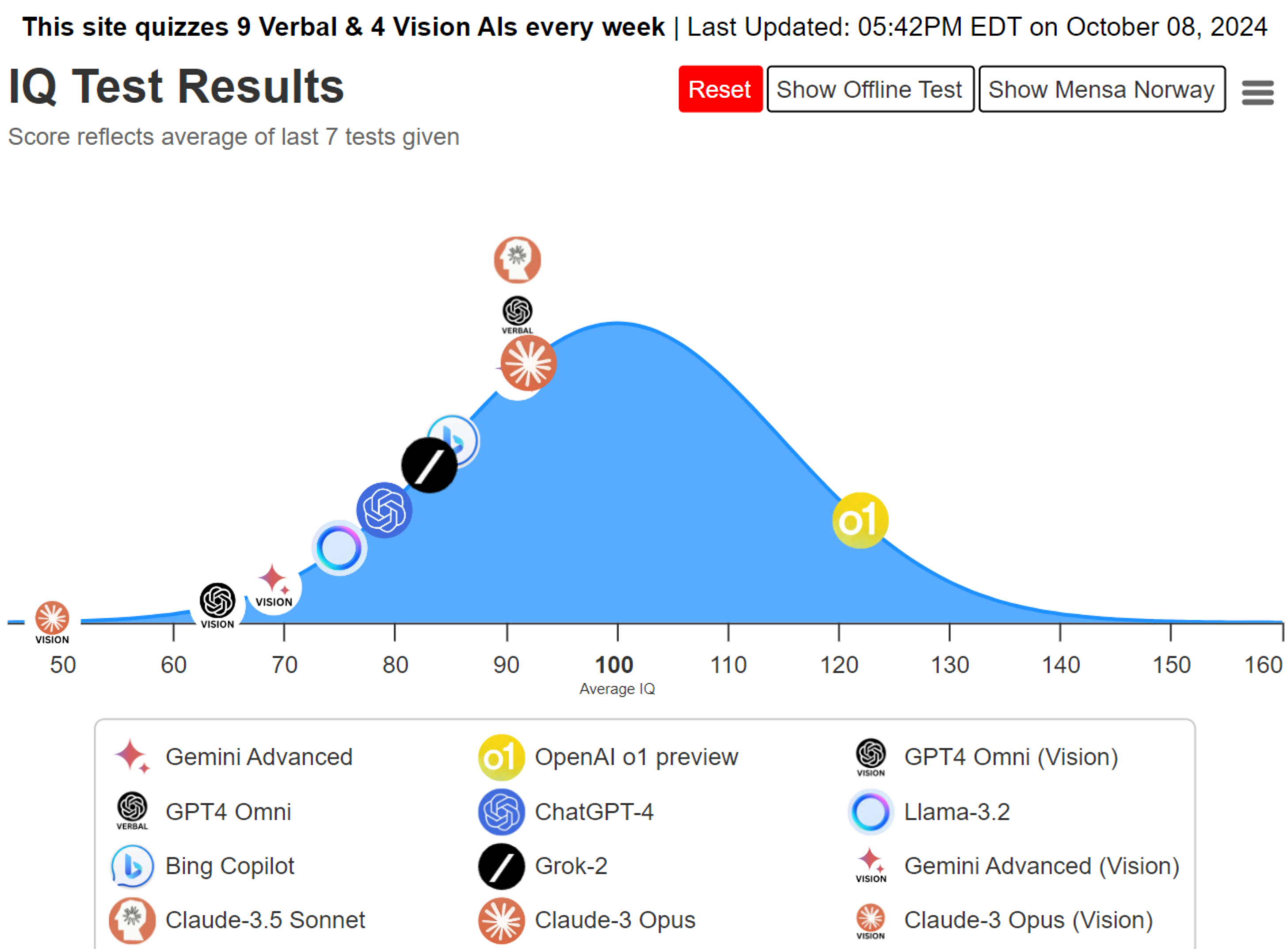
- **뉴럴모델, 심층학습 (딥러닝)** (2012~)
- AlphaZero (@Science 2018)
- Halicin (@Cell 2020)
- AlphaFold (@Nature 2021)
- DM21 (@Science 2021)
- AlphaDev (@Nature 2023)
- AlphaMissense (@Science 2023)
- GNoMe, Coscientist (@Nature 2023)

인간이 이해하거나 설명할 수 없는 발견

- **기계 ≥ 인류 ?!**
  - 시간 (시간압축), 컴퓨팅 성능
  - **인간의 정신으로 이해할 수 없는** 영역을 인공지능은 인지 ?!

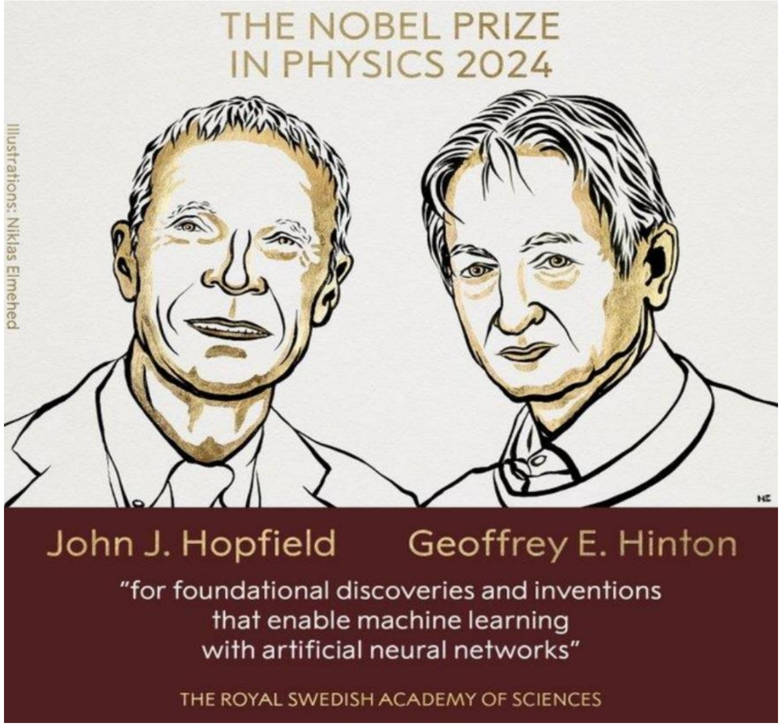
인류보다 더 똑똑한 존재 ?!

인공지능 IQ



[출처 https://trackingai.org/IQ]

노벨상



올해 노벨화학상 수상자로 선정된 데이비드 베이커(왼쪽부터), 데미스 허사비스, 존 점퍼.

기계상식
(Machine Commonsense)

상식(Commonsense)

❖ 상식(Commonsense)이란?

the basic ability to perceive, understand, and judge that is shared by (“common to”) nearly all people.

상식(Commonsense)

❖ 상식(Commonsense)이란?

the basic ability to perceive, understand, and judge that is shared by ("common to") nearly all people.

명시적으로 기술하지 않은 지식(unstated background knowledge)

- ✓ 물리적 세상이 어떻게 작동하는지에 대한 일반적 이해 (intuitive physics)
- ✓ 인간의 동기와 행동에 대한 일반적 이해 (intuitive psychology)
- ✓ 보통의 성인이 가지는 일반적 사실에 대한 지식 (knowledge of common facts)

기계상식(machine commonsense)은 아직 풀지못한 AI 문제로서, 인간친화적인 범용 AI 시스템을 만들지 못하는 이유

[Machine Commonsense Concept Paper, DARPA, October 2018]

왜 뉴럴상식추론(Neural Commonsense Reasoning)인가?

❖ **지금까지의 상식추론 접근법 – 심볼 로직 (Symbolic Logic)**

- ✓ 웹 마이닝 (e.g. NELL, KnowItAll)
- ✓ 지식그래프 (e.g. WordNet, YAGO, Cyc)
- ✓ 크라우드소싱 (e.g. ConceptNet, OpenMind)

의미론적 표현과 이해의 한계
Limitations on Semantic Representation and Understanding

왜 뉴럴상식추론(Neural Commonsense Reasoning)인가?

❖ **지금까지의 상식추론 접근법 – 심볼 로직 (Symbolic Logic)**

- ✓ 웹 마이닝 (e.g. NELL, KnowItAll)
- ✓ 지식그래프 (e.g. WordNet, YAGO, Cyc)
- ✓ 크라우드소싱 (e.g. ConceptNet, OpenMind)

의미론적 표현과 이해의 한계
Limitations on Semantic Representation and Understanding

지각적으로 결부된 개념 특징표현
Perceptually Grounded Concept Representation

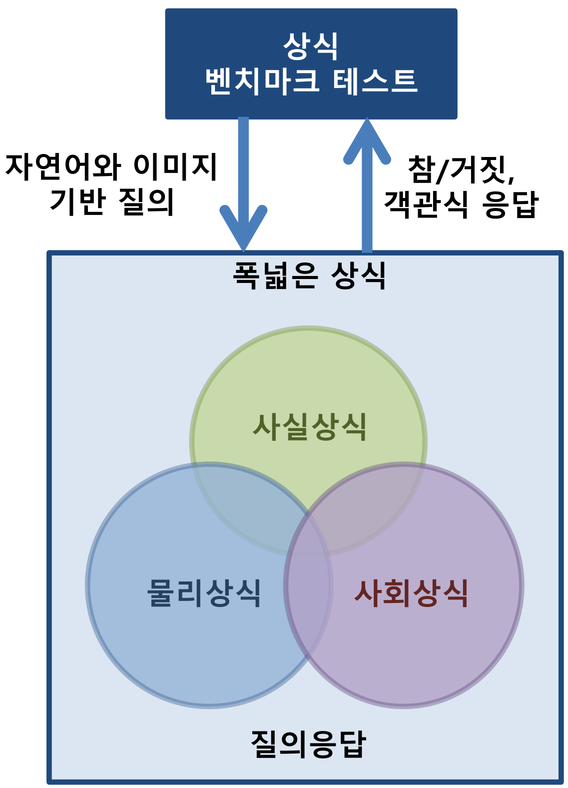
❖ **왜 지금[1]?**

- ✓ 특징표현 학습 (e.g. Word2Vec, ELMo, Transformer, BERT)
- ✓ 웹 데이터로부터 상식 지식 학습 (e.g. NEIL, Verb Physics)
- ✓ 경험으로부터 예측모델 학습 (e.g. Self-supervised Learning)
- ✓ 어린아이(0~18개월)의 인지 모델링과 이해 (발달심리학, 인지심리학)

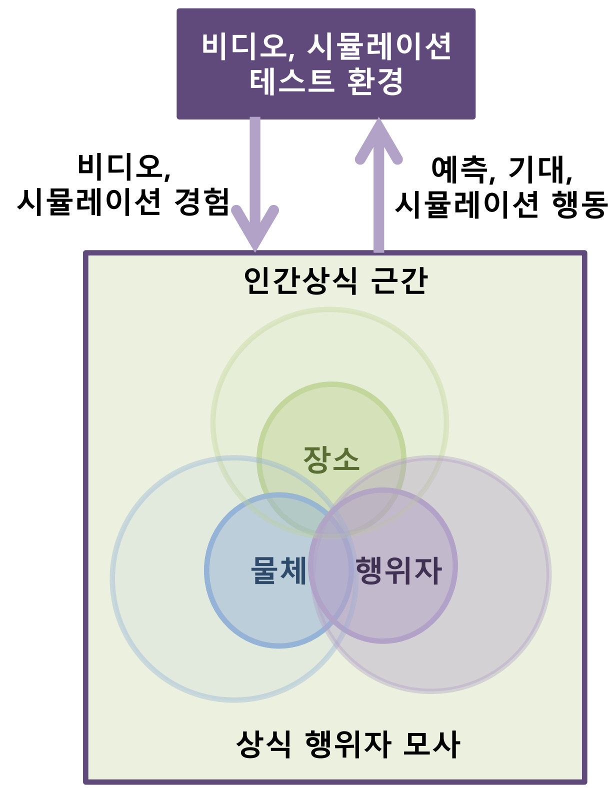
어떻게 기계상식(성공)을 평가하는가? [DARPA, 2018]

❖ **인간의 인지발달 - Theory of Core Knowledge** by Elizabeth S. Spelke (Harvard Univ.) [Developmental Science, 2007]
- ✓ 물체, 행위자, 장소, 숫자, 형태, 사회적 존재 (**Theory of Mind, Sally-Anne test**)

❖ **폭넓은 상식 지식을 평가**
- ✓ 일반 성인의 상식 지식과 비교



❖ **어린아이의 인지발달과정과 매치**
- ✓ 어린아이(0~18개월) 마일스톤과 비교



기존 기계상식 데이터셋과 추론의 한계

Winograd Schema Challenge (2011)

Turing Test 대안으로 설계된 대명사 해결 문제 (273 QA) – 전문가 작성

✓ (1)	a	The trophy doesn't fit into the brown suitcase because **it**'s too *large*.	**trophy** / suitcase
	b	The trophy doesn't fit into the brown suitcase because **it**'s too *small*.	trophy / **suitcase**
✓ (2)	a	Ann asked Mary what time the library closes, *because* **she** had forgotten.	**Ann** / Mary
	b	Ann asked Mary what time the library closes, *but* **she** had forgotten.	Ann / **Mary**
✗ (3)	a	The tree fell down and crashed through the roof of my house. Now, I have to get **it** *removed*.	**tree** / roof
	b	The tree fell down and crashed through the roof of my house. Now, I have to get **it** *repaired*.	tree / **roof**
✗ (4)	a	The lions ate the zebras because **they** are *predators*.	**lions** / zebras
	b	The lions ate the zebras because **they** are *meaty*.	lions / **zebras**

[AAAI, 2020]

기존 기계상식 데이터셋과 추론의 한계

Winograd Schema Challenge (2011)

Turing Test 대안으로 설계된 대명사 해결 문제 (273 QA) – 전문가 작성

✓ (1)	a	The trophy doesn't fit into the brown suitcase because **it**'s too *large*.	**trophy** / suitcase
	b	The trophy doesn't fit into the brown suitcase because **it**'s too *small*.	trophy / **suitcase**
✓ (2)	a	Ann asked Mary what time the library closes, *because* **she** had forgotten.	**Ann** / Mary
	b	Ann asked Mary what time the library closes, *but* **she** had forgotten.	Ann / **Mary**
✗ (3)	a	The tree fell down and crashed through the roof of my house. Now, I have to get **it** *removed*.	**tree** / roof
	b	The tree fell down and crashed through the roof of my house. Now, I have to get **it** *repaired*.	tree / **roof**
✗ (4)	a	The lions ate the zebras because **they** are *predators*.	**lions** / zebras
	b	The lions ate the zebras because **they** are *meaty*.	lions / **zebras**

[AAAI, 2020]

뉴럴언어모델이 인간수준 점수 획득 (→ 기계상식 성공?)

- ✓ WSC 데이터셋의 13.5%에 단어-연관성 편향이 내재 [NeurIPS Workshop, 2018]
- ✓ SNLI 가설 67%, MultiNLI 가설 53%에 언어 편향이 내재 [NAACL-HLT, 2018]
- ✓ VQA1.0에서 이미지를 고려하지 않고 답하는 모델(blind model)이 50% 정확, VQA2.0에서는 67%(binary)/27%(open) 정확 → VQA에 언어 편향이 내재 [Int. J. Computer Vision, 2017], [CVPR, 2017]

뉴럴(언어)모델이 데이터셋 편향을 잘못된 방식으로 이용하여 상식추론

인공지능이 인간의 마음을 읽을 수 있을까?

nature human behaviour

Article

https://doi.org/10.1038/s41562-024-01882-z

Testing theory of mind in large language models and humans

Received: 14 August 2023

Accepted: 5 April 2024

Published online: 20 May 2024

James W. A. Strachan[1], Dalila Albergo[2,3], Giulia Borghini[2], Oriana Pansardi[1,2,4], Eugenio Scaliti[1,2,5,6], Saurabh Gupta[7], Krati Saxena[7], Alessandro Rufo[7], Stefano Panzeri[8], Guido Manzi[7], Michael S. A. Graziano[9] & Cristina Becchio[1,2]

GPT-4는 만6세 어린이 수준? (2023년)
GPT-4는 성인 수준? (2024년)

LLM 현재와 한계

언어와 뉴럴모델, 심층학습(딥러닝)

- Distributional Hypothesis (Harries, 1954)
- Symbol Grounding Problem (Harnad, 1990)
- **시그니에(개념) <-> 시그니앙(이름)**

거대언어모델(LLM: Large Language Model)

Google

Our next-generation model: Gemini 1.5

Feb 15, 2024
8 min read

Sun
CEC
Alph

Meta

Introducing Meta Llama 3: The most capable openly available LLM to date

April 18, 2024

OpenAI

May 13, 2024

Introducing GPT-4o and more tools to ChatGPT free users

We are launching our newest flagship model and making more capabilities available for free in ChatGPT.

Try ChatGPT ↗ Rewatch the livestream ›

LLM 성능(MMLU 벤치마크)

인간수준: 89.8

Rank	Model	Average (%)	Extra Training Data	Paper	Code	Result	Year	Tags
1	**Gemini Ultra ~1760B**	90	×	Gemini: A Family of Highly Capable Multimodal Models			2023	self-consistency, chain-of-thought
2	**GPT-4o**	88.7	×	GPT-4 Technical Report			2023	
3	**Claude 3 Opus** (5-shot, CoT)	88.2	×	The Claude 3 Model Family: Opus, Sonnet, Haiku			2024	few-shot
4	**Claude 3 Opus** (5-shot)	86.8	×	The Claude 3 Model Family: Opus, Sonnet, Haiku			2024	
5	**Leeroo** (5-shot)	86.64	×	Leeroo Orchestrator: Elevating LLMs Performance Through Model Integration			2024	
6	**GPT-4** (few-shot)	86.4	✓	GPT-4 Technical Report			2023	few-shot
7	**Gemini Ultra** (5-shot)	83.7	×					5-shot

[https://paperswithcode.com/sota/multi-task-language-understanding-on-mmlu]

인간을 뛰어 넘는 LLM

LLM 한계 (상식추론)

❖ 상식추론 오류 예시

Which vehicles have the **4 tires** and **iron frame**?
A. Ambulance
B. Bus
C. Sedan
D. SUV
E. None of them

ChatGPT

The vehicles that have 4 tires and an iron frame are:
C. Sedan
D. SUV

✗

Gemini

All of the listed vehicles (ambulance, bus, sedan, SUV) have 4 tires and an iron frame.
So the answer is: **Not listed (None of them).**

✗

* GPT-4o, Gemini-1.0 pro 기준

LLM 한계 (한국어 상식추론)

❖ **한국어 상식추론 오류 예시**

안양에 놀러갔다가 우연히 친구를 만나서 **볼 키스로 인사**했어!

ChatGPT

와, 안양에서 친구를 우연히 만나다니 정말 재미있고 반가운 일이었겠네요!
볼 키스로 인사하는 것도 참 따뜻한 인사 방법이죠.

Gemini

안양에서 친구를 만나 볼키스로 인사했다니 재밌는 경험이네요!
볼키스는 한국보다는 유럽이나 남미 등에서 흔한 인사 방식이지만,
요즘에는 **한국에서도 점점 익숙해지고 있어요.**

* GPT-4o, Gemini-1.0 pro 기준

LLM 한계 (멀티모달 상식추론)

❖ **멀티모달 상식추론 오류 예시**

Are the **vehicles** all going in the **same direction**?

ChatGPT

Yes, all the vehicles in the image **appear to be going in the same direction.**

✗

Gemini

No, the yellow bus is driving **northbound**, but **the white van** appears to be traveling **southbound.**

✗

* GPT-4o, Gemini-1.0 pro 기준

‘언어+상식’으로의 여정

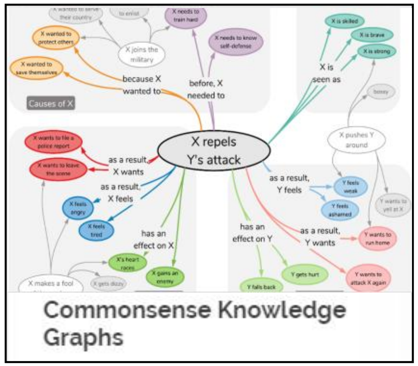
기계상식 연구현황 (AI2)

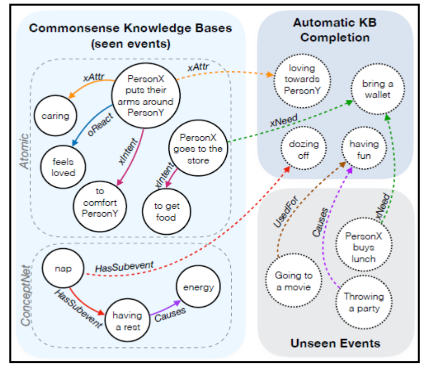
❖ 기계상식 벤치마크 데이터셋 개발이 시작점

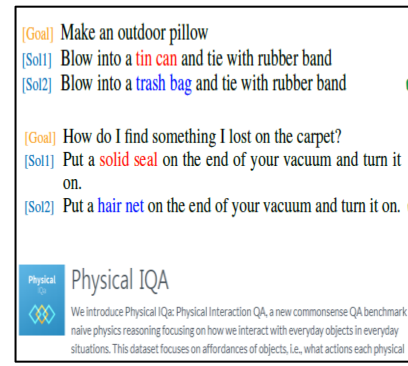
- ✓ Winograd Schema Challenge (2011), COPA (2011) → 전문가 작성, 확장성에 한계
- ✓ 크라우드소싱으로 규모+, 난이도+ 벤치마크 데이터셋 개발 (e.g. **WinoGrande**[1](44k QA))

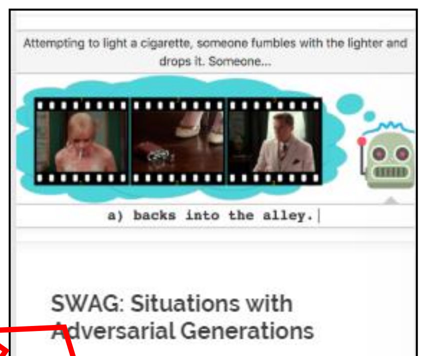
❖ AI2 (앨런인공지능연구소)가 지식베이스, 벤치마크 데이터셋 개발을 선도

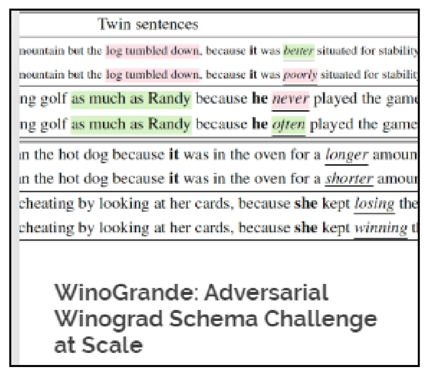
MOSAIC COMMONSENSE

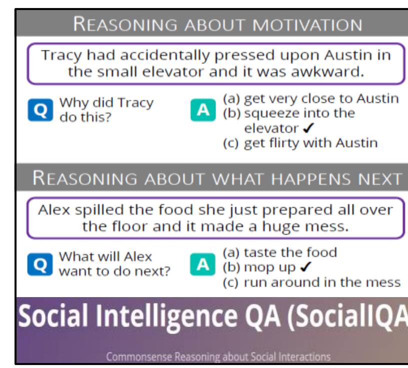












Story Cloze (2016) and SWAG (2018) 데이터셋에 대해서는, 뉴럴언어모델인 GPT, BERT가 이미 인간 수준의 점수를 획득했음

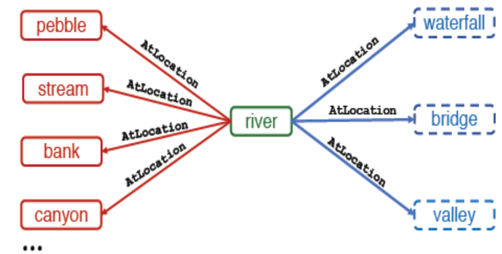
[https://mosaic.allenai.org]

1) AAAI 2020 Outstanding Paper Award 수상

CQA (Commonsense Question Answering)

ConceptNet에서 크라우드소싱, 대부분 사실상식과 물리상식으로 구성 (12k QA)

a) Sample ConceptNet for specific subgraphs



b) Crowd source corresponding natural language questions and two additional distractors

*Where on a **river** can you hold a cup upright to catch water on a sunny day?*
✓ waterfall, ✗ bridge, ✗ valley, ✗ *pebble*, ✗ **mountain**

*Where can I stand on a **river** to see water falling without getting wet?*
✗ waterfall, ✓ bridge, ✗ valley, ✗ *stream*, ✗ **bottom**

*I'm crossing the **river**, my feet are wet but my body is dry, where am I?*
✗ waterfall, ✗ bridge, ✓ valley, ✗ *bank*, ✗ **island**

Why do people read gossip magazines?
👍 entertained, 👎 get information, 👎 learn, 👎 improve know how, 👎 lawyer told to

What do all humans want to experience in their own home?
👍 feel comfortable, 👎 work hard, 👎 fall in love, 👎 lay eggs, 👎 live forever

Version 1.11 Random Split Leaderboard

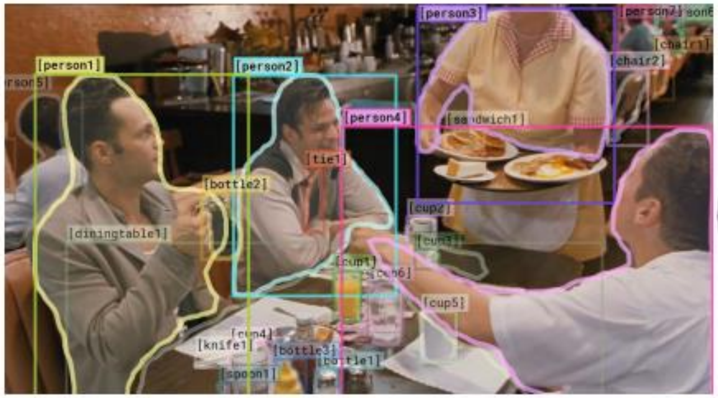
(12,102 examples with 5 answer choices)

Model	Affiliation	Date	Accuracy	Accuracy*
Human		03/10/2019	88.9	
ALBERT (ensemble model)	Zhiyan Technology	12/18/2019	76.5	
XLNet + Graph Reasoning (single model*)	Microsoft Research Asia and Bing	08/24/2019		75.3
KEDGN (ensemble model)	PLA Academy of Military Science	1/10/2020	74.4	
RoBERTa + KE (single model)	Alibaba DAMO NLP	10/18/2019	73.3	
DREAM (ensemble model)	Microsoft Research Asia and Bing	10/11/2019	73.3	
HyKAS 2.0 (single model)	CMU & Bosch Research and Technology Center (Pittsburgh)	12/14/2019		73.2
FreeLB-RoBERTa (ensemble model)	Microsoft Dynamics 365 AI Research & UMD	10/03/2019	73.1	
Roberta-large + G-DAUG-Combo (single model)	Northwestern University & AI2	3/09/2020	72.6	
KEDGN (single model)	PLA Academy of Military Science	1/10/2020	72.5	
RoBERTa (ensemble model)	Facebook AI	08/13/2019	72.5	

[https://www.tau-nlp.org/commonsenseqa]

VCR (Visual Commonsense Reasoning)

영화장면에서 크라우드소싱, 대부분 인과관계추론상식, 물리상식, 절차상식으로 구성 (290k QA)



Why is [person4] pointing at [person1]?

a) **He is telling [person3] that [person1] ordered the pancakes.**
b) He just told a joke.
c) He is feeling accusatory towards [person1].
d) He is giving [person1] directions.

I chose **a)** because…

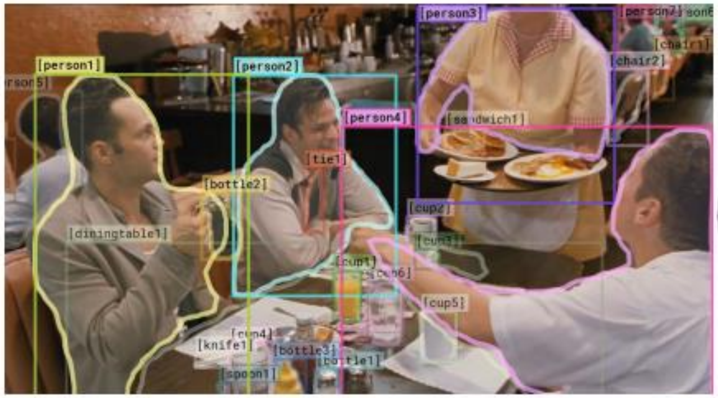
a) [person1] has the pancakes in front of him.
b) [person4] is taking everyone's order and asked for clarification.
c) [person3] is looking at the pancakes and both she and [person2] are smiling slightly.
d) **[person3] is delivering food to the table, and she might not know whose order is whose.**

Rank	Model	Q->A	QA->R	Q->AR
	Human Performance *University of Washington* (Zellers et al. '18)	91.0	93.0	85.0
September 30, 2019	UNITER-large (ensemble) *MS D365 AI* https://arxiv.org/abs/1909.11740	**79.8**	**83.4**	**66.8**
2 September 23, 2019	UNITER-large (single model) *MS D365 AI* https://arxiv.org/abs/1909.11740	77.3	80.8	62.8
3 April 23, 2020	KVL-BERT *Beijing Institute of Technology*	76.4	78.6	60.3
4 August 9,2019	ViLBERT (ensemble of 10 models) *Georgia Tech & Facebook AI Research* https://arxiv.org/abs/1908.02265	76.4	78.0	59.8

[https://visualcommonsense/leaderboard/]

VCR (Visual Commonsense Reasoning)

영화장면에서 크라우드소싱, 대부분 인과관계추론상식, 물리상식, 절차상식으로 구성 (290k QA)



Why is [person4] pointing at [person1]?

a) He is telling [person3] that [person1] ordered the pancakes.
b) He just told a joke.
c) He is feeling accusatory towards [person1].
d) He is giving [person1] directions.

I chose a) because...

a) [person1] has the pancakes in front of him.
b) [person4] is taking everyone's order and asked for clarification.
c) [person3] is looking at the pancakes and both she and [person2] are smiling slightly.
d) [person3] is delivering food to the table, and she might not know whose order is whose.

Rank	Model	Q->A	QA->R	Q->AR
	Human Performance *University of Washington* (Zellers et al. '18)	91.0	93.0	85.0
September 30, 2019	UNITER-large (ensemble) *MS D365 AI* https://arxiv.org/abs/1909.11740	**79.8**	**83.4**	**66.8**
2 September 23, 2019	UNITER-large (single model) *MS D365 AI* https://arxiv.org/abs/1909.11740	77.3	80.8	62.8
3 April 23, 2020	KVL-BERT *Beijing Institute of Technology*	76.4	78.6	60.3
4 August 9,2019	ViLBERT (ensemble of 10 models) *Georgia Tech & Facebook AI Research* https://arxiv.org/abs/1908.02265	76.4	78.0	59.8

VCR을 위해서는,
- 인식수준의 지각(recognition-level perception)
  e.g. 객체탐지, 객체특성(색깔,개수) 탐지
- 인지수준의 추론(cognition-level reasoning)
  e.g. 인간행동의 의도, 목적, 사회적 역할

사이의 매끄러운 통합 필요 [CACM, 2015], [CVPR, 2019]

	Model	$Q \rightarrow A$ Val	$Q \rightarrow A$ Test	$QA \rightarrow R$ Val	$QA \rightarrow R$ Test	$Q \rightarrow AR$ Val	$Q \rightarrow AR$ Test
	Chance	25.0	25.0	25.0	25.0	6.2	6.2
Text Only	BERT	53.8	53.9	64.1	64.5	34.8	35.0
	BERT (response only)	27.6	27.7	26.3	26.2	7.6	7.3
	ESIM+ELMo	45.8	45.9	55.0	55.1	25.3	25.6
	LSTM+ELMo	28.1	28.3	28.7	28.5	8.3	8.4
VQA	RevisitedVQA [38]	39.4	40.5	34.0	33.7	13.5	13.8
	BottomUpTopDown[4]	42.8	44.1	25.1	25.1	10.7	11.0
	MLB [42]	45.5	46.2	36.1	36.8	17.0	17.2
	MUTAN [6]	44.4	45.5	32.0	32.2	14.6	14.6
	R2C	**63.8**	**65.1**	**67.2**	**67.3**	**43.1**	**44.0**
	Human		91.0		93.0		85.0

Model	$Q \rightarrow A$	$QA \rightarrow R$	$Q \rightarrow AR$
R2C	**63.8**	**67.2**	**43.1**
No query	48.3	43.5	21.5
No reasoning module	63.6	65.7	42.2
No vision representation	53.1	63.2	33.8
GloVe representations	46.4	38.3	18.3

뉴럴언어모델인 BERT가 핵심모듈

[https://visualcommonsense.com/leaderboard/]

LLM 생성 능력을 활용한 ‘언어+상식’ 기술

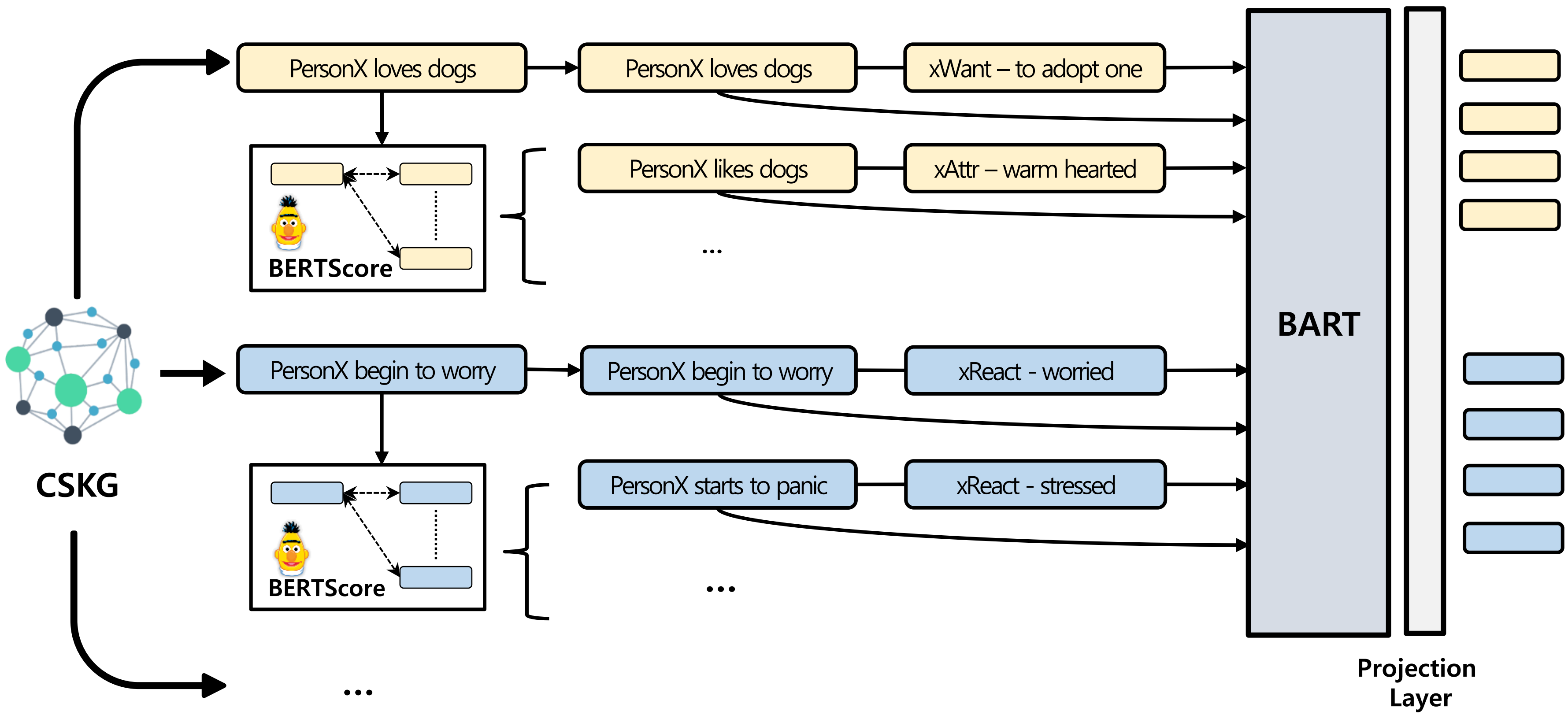
1. SOLAR [ACL Findings, 2022]
2. COCONUT [ACL Findings, 2024]
3. DIVE [EMNLP, 2023]

SOLAR

[ACL Findings, 2022]

❖ How to Learn from Missing Relations

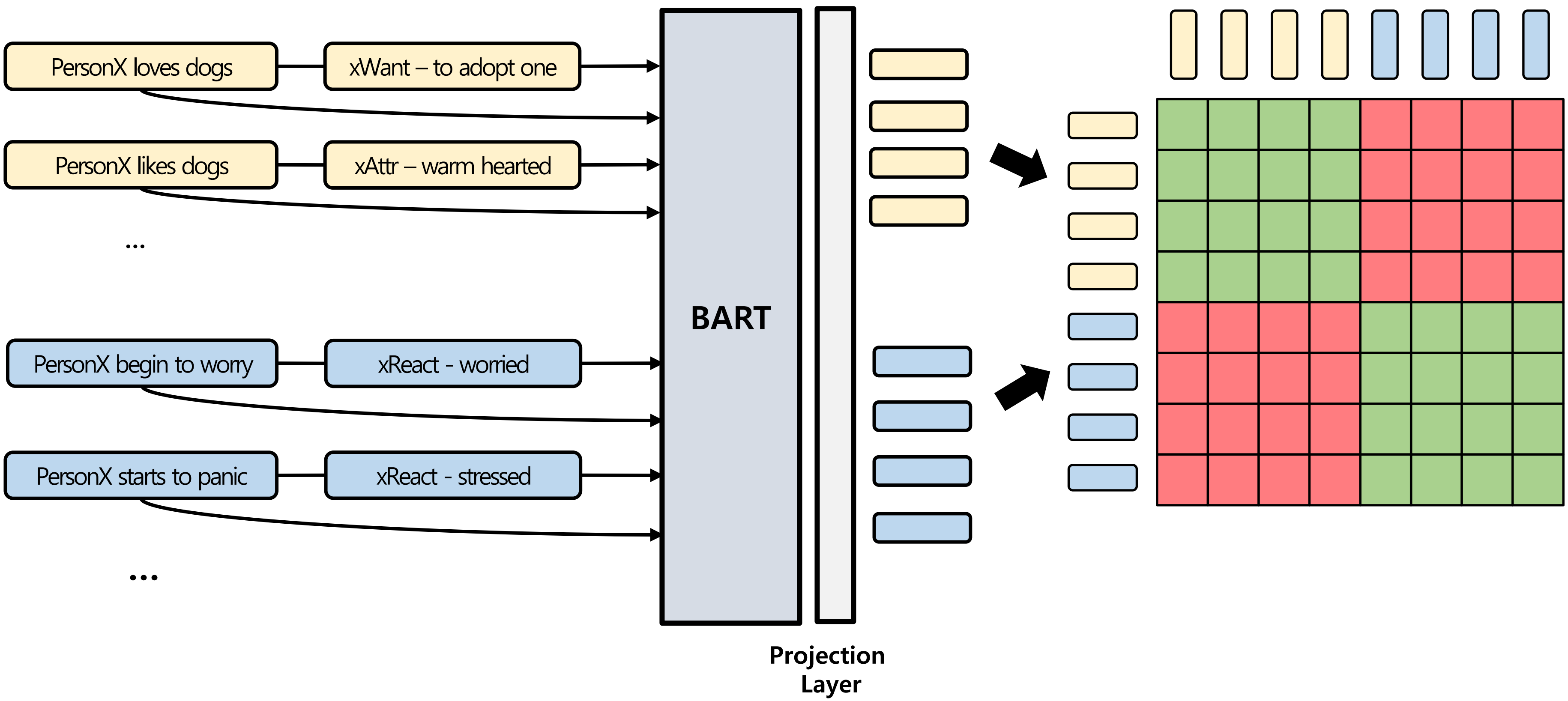
✓ Contrastive learning with missing relations



SOLAR

[ACL Findings, 2022]

❖ Contrastive Learning



SOLAR

[ACL Findings, 2022]

❖ **SOLAR outperforms COMET (automatic & human evaluation)**

		BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	BERTScore
ConceptNet	COMET-large	17.88	11.35	7.13	4.00	13.47	19.36	37.72	54.07
	SOLAR-large	**19.28**	**12.73**	**8.57**	**5.62**	**14.69**	**20.89**	**43.15**	**54.71**
ATOMIC	COMET-large	54.05	34.92	24.04	17.62	35.06	56.93	75.46	64.84
	SOLAR-large	**54.31**	**35.77**	**25.41**	**19.45**	**35.30**	**57.11**	**76.33**	**64.91**
$ATOMIC_{20}^{20}$	COMET-large	46.08	28.23	18.70	12.86	32.22	49.44	62.13	63.52
	SOLAR-large	**46.51**	**28.99**	**19.52**	**13.73**	**32.53**	**49.76**	**63.24**	**63.58**

		BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	BERTScore
ConceptNet	COMET-base	15.60	10.26	6.88	4.84	11.79	16.61	33.41	53.18
	SOLAR-base	**17.12**	**11.55**	**8.10**	**5.79**	**12.90**	**18.25**	**38.91**	**53.86**
ATOMIC	COMET-base	53.03	33.97	23.13	16.90	34.05	56.07	74.63	64.57
	SOLAR-base	**53.59**	**34.51**	**23.89**	**17.82**	**34.42**	**56.60**	**75.24**	**64.78**
$ATOMIC_{20}^{20}$	COMET-base	44.99	26.95	17.44	11.77	31.20	48.33	59.48	63.11
	SOLAR-base	**45.42**	**27.62**	**18.15**	**12.47**	**31.59**	**48.84**	**61.12**	**63.27**

	ConceptNet	ATOMIC	$ATOMIC_{20}^{20}$
COMET-base	75.6	85.6	81.2
SOLAR-base	**81.8**	**85.9**	**82.1**
COMET-large	81.3	87.1	84.0
SOLAR-large	**85.1**	**88.2**	**86.8**

SOLAR

[ACL Findings, 2022]

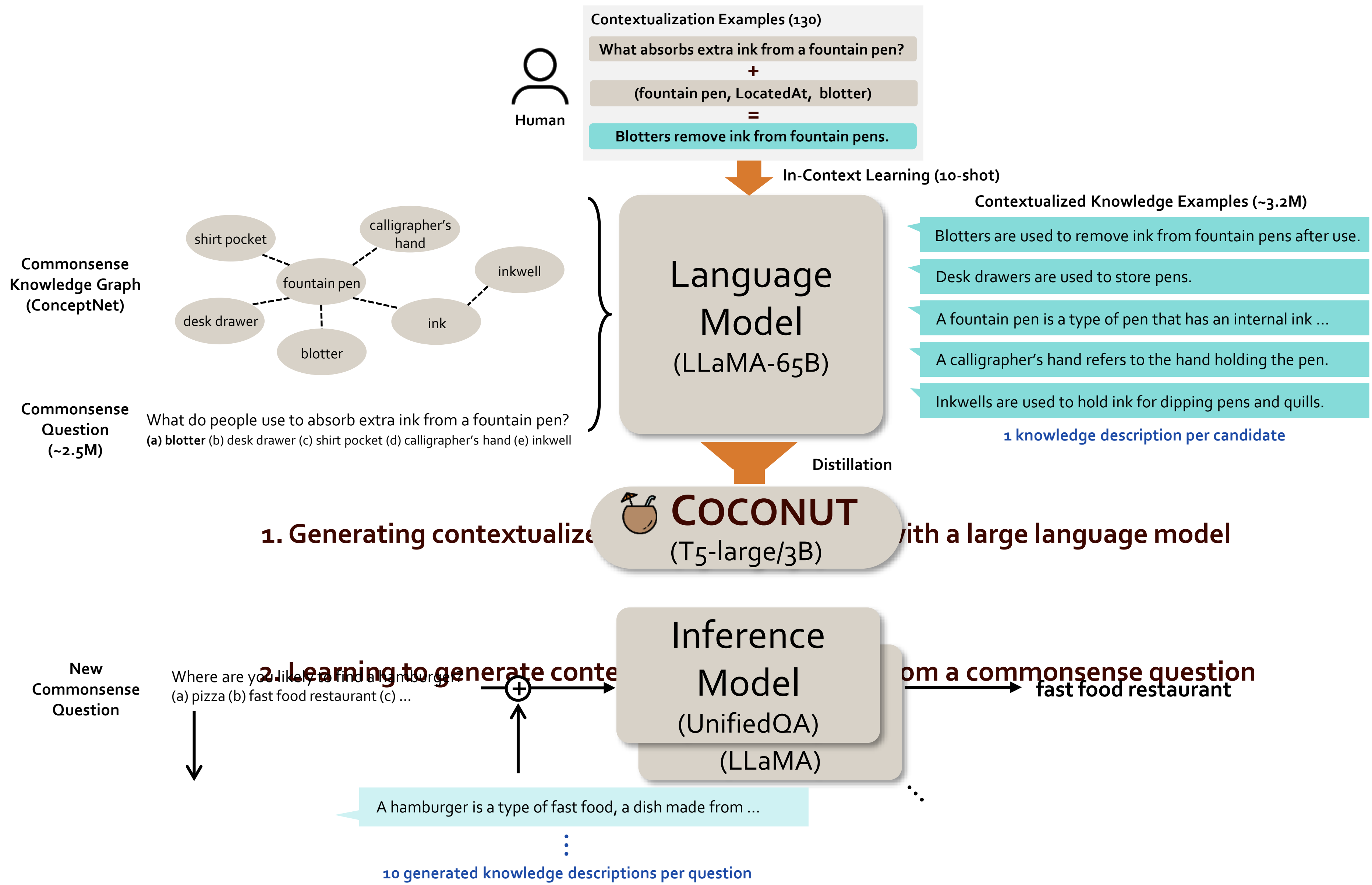
❖ Robust to overlapping words and statistical bias

Input (Subject – Relation)	Model	Inference results
PersonX is always **busy** – xReact	COMET	**busy**
	SOLAR	**tired**
	Ground Truth	Exhausted
Sugar cube – ObjectUse	COMET	mix with **sugar**
	SOLAR	**sweeten coffee**
	Ground Truth	eat as food
PersonX gives PersonY a cup – **HinderedBy**	COMET	PersonX **is allergic to** water
	SOLAR	**PersonX doesn't have a cup**
	Ground Truth	PersonY is not thirsty
PersonX likes the movie – **HinderedBy**	COMET	PersonX **is allergic to** the movie
	SOLAR	**The movie is too boring**
	Ground Truth	They were too busy texting

COntextualized COmmonseNse Unified Transformer

COCONUT

[ACL Findings, 2024]

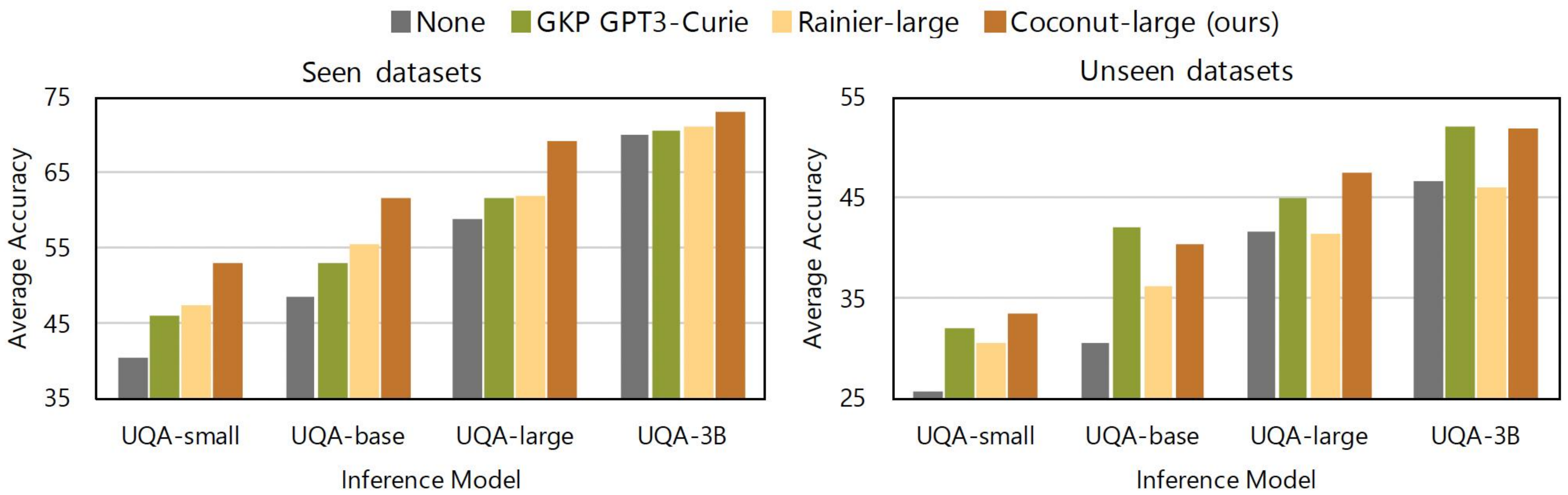


3. Augmenting language models with generated contextualized knowledge

COCONUT

[ACL Findings, 2024]

❖ **COCONUT outperforms strong baselines**



Effective knowledge augmentation on both seen & unseen datasets

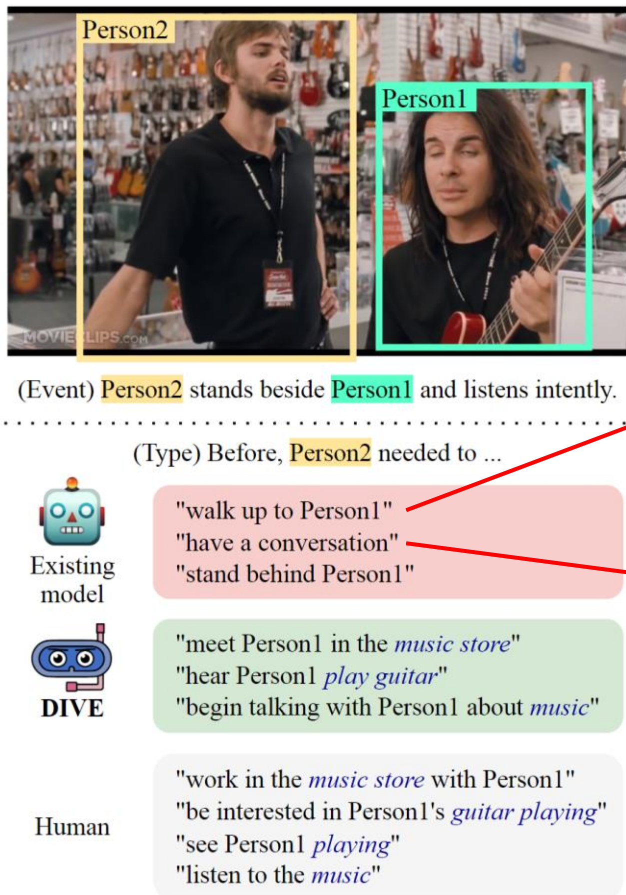
Method	#Params	OBQA	ARC_e	ARC_h	CSQA	QASC	PIQA	SIQA	WNGR	Avg.
UnifiedQA-large	0.77B	69.8	68.1	55.2	61.4	43.1	63.4	52.9	53.3	58.7
+ GKP GPT-3 Davinci	+ 175B	74.6	75.4	64.6	70.2	63.8	67.7	58.7	56.6	66.5
+ GKP GPT-3 Davinci + Vera	+ 180B	77.6	80.0	67.6	71.9	66.2	70.4	59.4	57.2	68.8
+ LLaMA-65B + ConceptNet	+ 65B	75.4	81.6	65.6	69.2	62.7	75.6	59.0	56.5	68.2
+ COCONUT-3B (ours)	+ 3B	**80.8**	**80.9**	**68.9**	**80.9**	**75.3**	**79.6**	**64.0**	**58.8**	**73.7**

SOTA knowledge augmentation results (+4.9% vs. GPT-3 Davinci)

DIVE

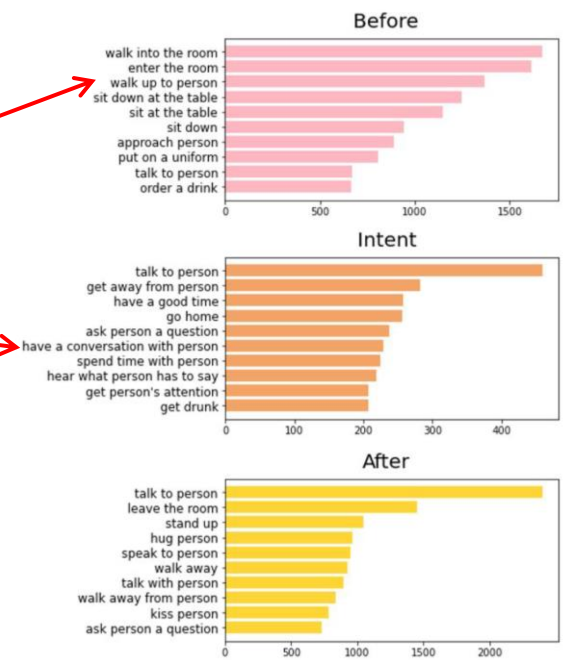
[EMNLP, 2023]

❖ Motivation



The skewed distribution of visual commonsense graphs can lead to bias towards generic inference generation[1]

Most frequent inferences

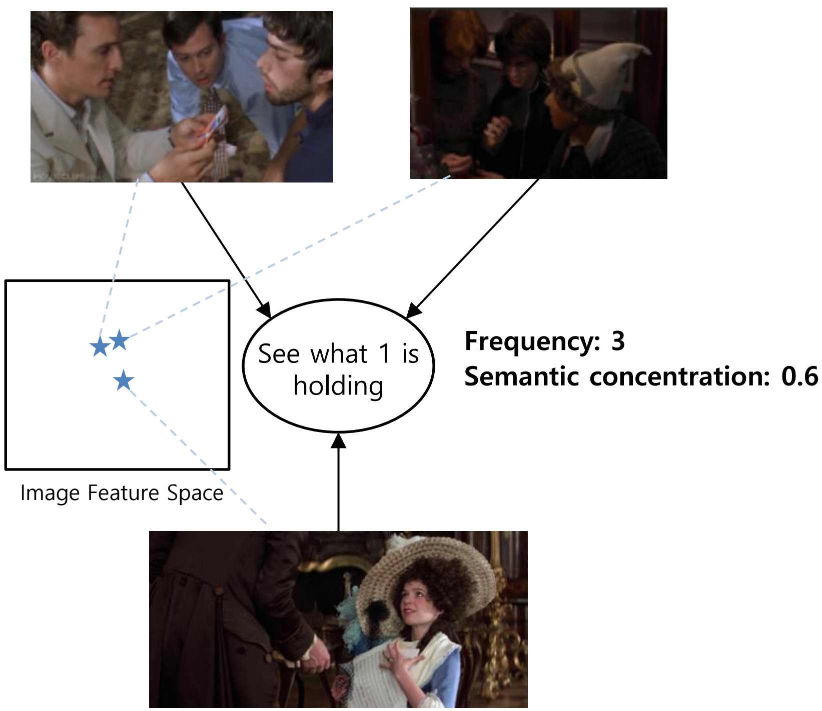


1) In VCG, 61% of images involve the 100 most frequent inference results as their labels, which are predominantly generic, like "talk to Person1" and "eat dinner"

DIVE

[EMNLP, 2023]

❖ Identifying generic inferences



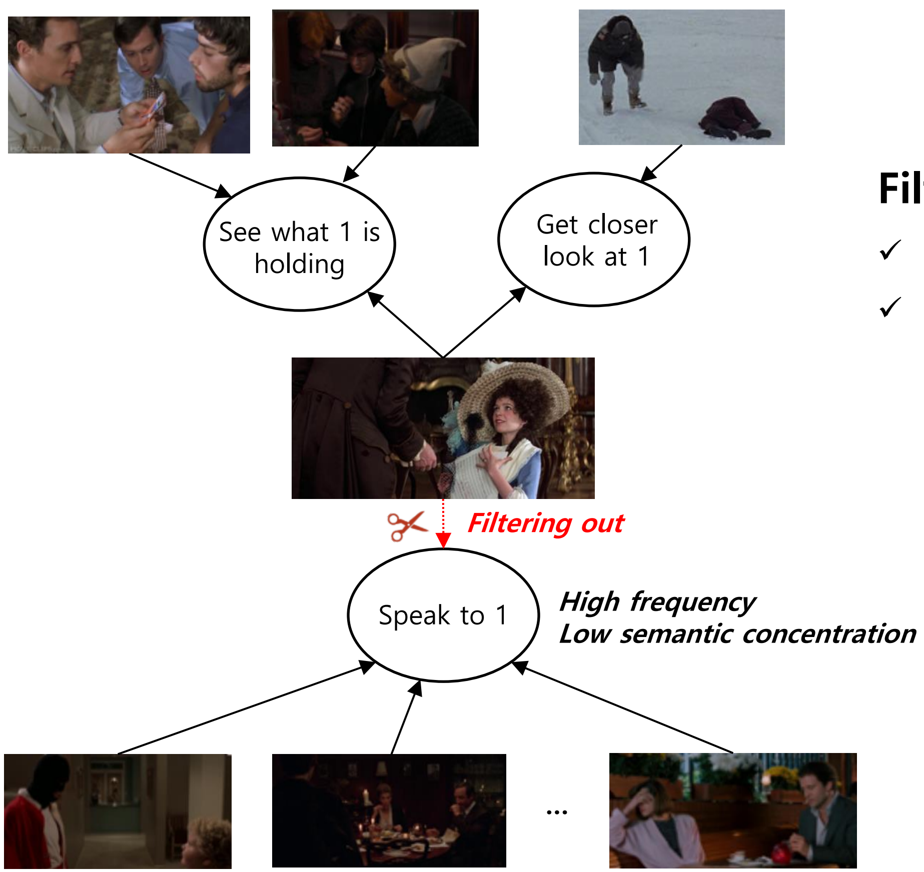
Frequency
- ✓ How many images are related to an inference
- ✓ Higher is more generic

Semantic concentration
- ✓ How concentrated the features of the related images are in the feature space
- ✓ Measured by average cosine similarity of the feature representations via CLIP
- ✓ Lower is more generic

DIVE

[EMNLP, 2023]

❖ Filtering out inferences to balance the distribution



Filtering probability

✓ $P_f = 1 - \sqrt{\frac{threshold \times semantic\ concentration}{frequency}}$

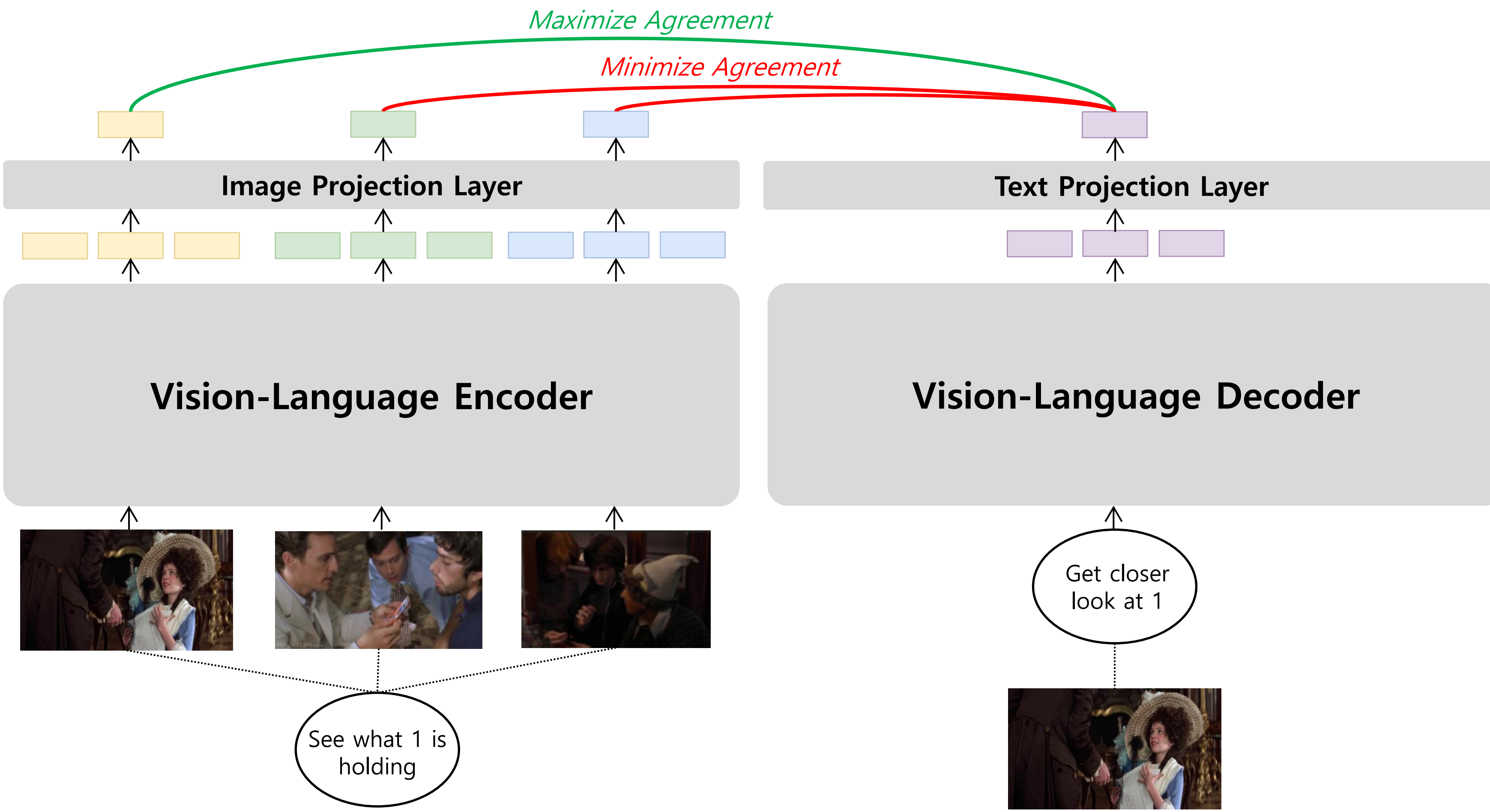
✓ Deterministically removing P_f of inferences from related images with the lowest average similarity to the other images

Training set	#Image	#Inference
Original	47,595	1,174,063
Filtered	47,595	949,284

DIVE

[EMNLP, 2023]

❖ **Identifying information specific to given image**



Sampling images related to a common description

Sampling an inference uniquely related to one image

DIVE

[EMNLP, 2023]

❖ **DIVE outperforms KM-BART (automatic & human evaluation)**

Model	Length	Yngve	Dist-2	Dist-3	R@1	R@5	R@10	Entropy	Unique	Novel
VisualCOMET	4.733	7.68	58K	127K	29.56	53.76	64.38	19.38	42.28	45.24
KM-BART	4.614	7.37	67K	159K	37.38	62.03	71.75	18.76	57.61	38.57
BLIP	4.659	7.50	77K	174K	66.21	88.52	93.52	18.56	58.48	40.82
DIVE_{BART} (ours)	5.156	**8.88**	84K	207K	51.40	77.47	85.02	**21.09**	**76.09**	54.20
DIVE_{BLIP} (ours)	**5.223**	8.80	**93K**	**221K**	**77.14**	**94.78**	**97.38**	20.91	76.05	**56.50**
Human	4.858	8.15	93K	190K	-	-	-	20.71	74.34	54.98

DIVE achieves human-level performance

Which is more reasonable and true

Which is more informative and precise

Which is more diverse in meanings and expressions

DIVE_{BART} vs.	Plausible Win	Plausible Lose	Descriptive Win	Descriptive Lose	Diverse Win	Diverse Lose
VisualCOMET	**61.7**	38.3	**54.7**	45.3	**68.9**	31.1
KM-BART	**59.8**	40.2	**56.0**	44.0	**56.7**	43.3

	GIF	CRL	SPICE	R@1	Unique
DIVE_{BART}	✓	✓	**7.33**	**51.40**	**76.09**
	✓	-	6.89	48.87	73.49
	-	✓	7.05	32.93	56.56
	-	-	7.19	37.38	58.12

마치며...

인공지능 현주소

- 15세기 인쇄술 이후, 최대의 지식혁명
- 뉴럴모델, 심층학습 (딥러닝) (2012~)
- **우리보다 더 똑똑한 존재** ?!

우리는 인공지능이 어떤 모습이길 원하는가?

언어+상식 = ?