20 • 21기 심화세션

ToBig's 20기 도형준

LLaVA: Visual Instruction Tuning

2024.06.19

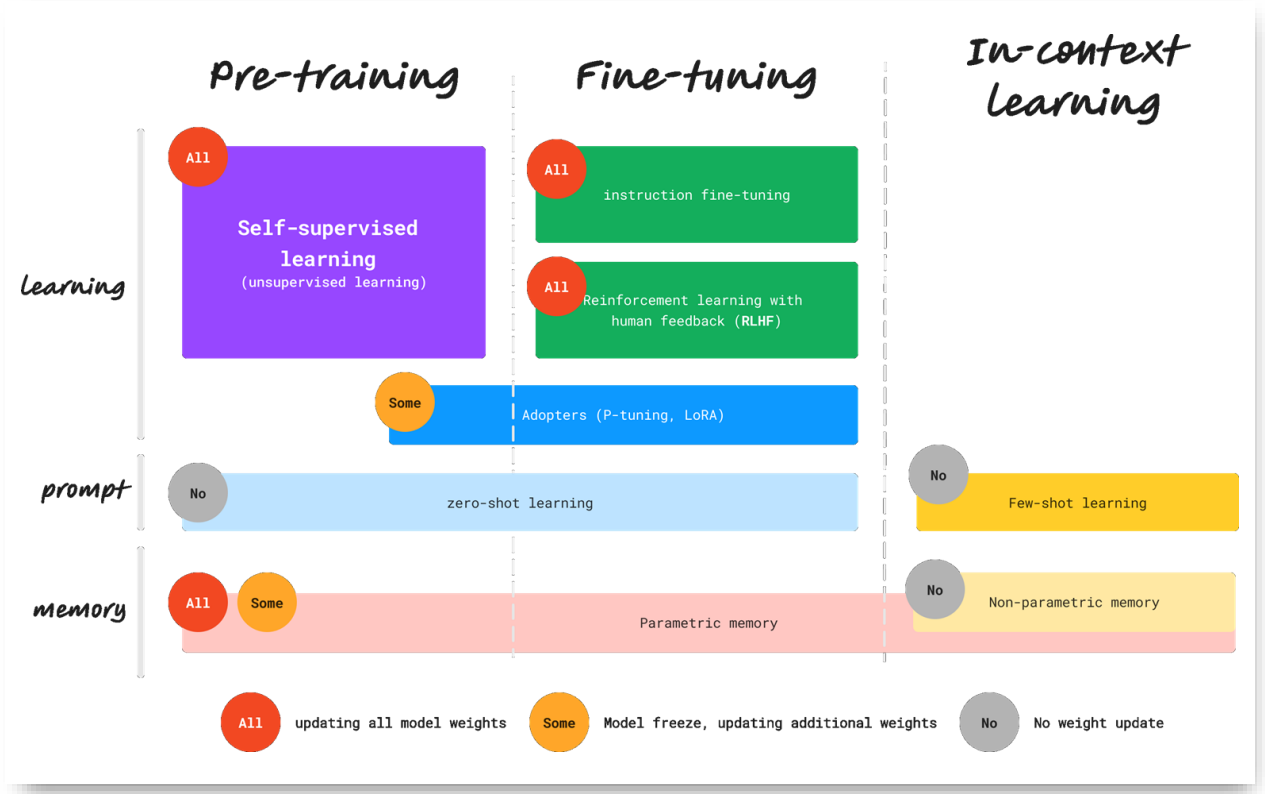
Contents

01

Glossary

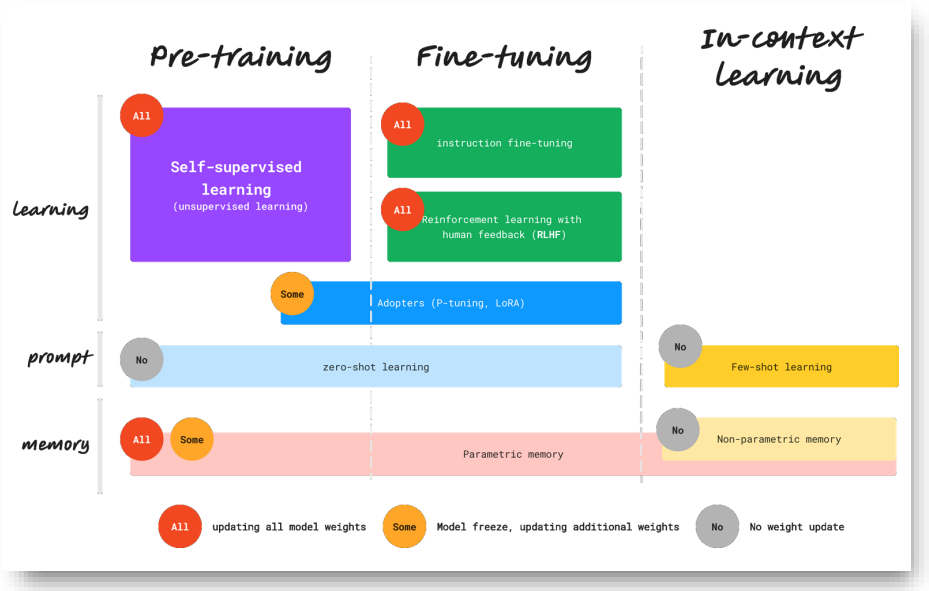
Unit 01. Glossary

- How to train Large Model?



Unit 01. Glossary

- How to train Large Model?



1. Pre-training (사전 학습)

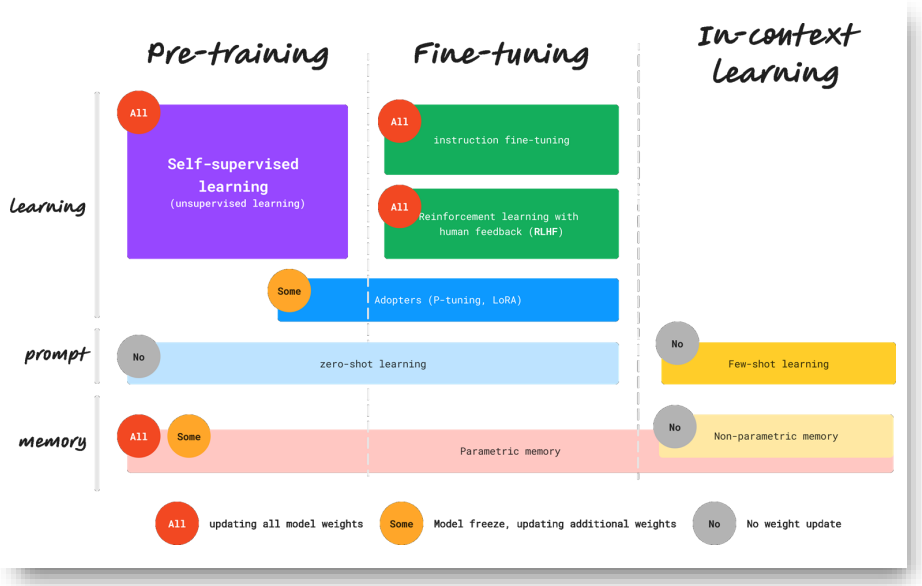
- **Self-Supervised Learning**
  - 사전 학습 단계에서 사용되는 기법, 대규모 비지도 학습을 통해 모델 초기화
  - 모든 모델 가중치가 업데이트되며, 주어진 데이터에서 특정 목표를 예측하는 방식으로 학습

2. Fine-tuning (미세 조정)

- **Instruction fine-tuning**
  - 모델이 사전 학습된 상태에서 인간의 명령을 이해하고 따르는 능력을 향상시키기 위해 미세 조정
  - 모든 모델 가중치가 업데이트
- **Reinforcement learning with human feedback (RLHF)**
  - 인간의 피드백을 통해 강화 학습을 수행하여 모델의 성능을 개선
  - 이 단계에서도 모든 모델 가중치가 업데이트
- **Adopters (P-tuning, LoRA)**
  - 일부 가중치만 업데이트하여 특정 작업에 맞게 모델을 조정
  - P-tuning과 LoRA와 같은 기법을 사용
  - 이는 모든 가중치가 아닌 추가적인 일부 가중치만을 업데이트하는 방식
- **Zero-shot learning**
  - 학습 단계 없이 주어진 프롬프트만으로 새로운 작업을 수행
  - 모델 가중치는 업데이트되지 않으며, 모델이 기존에 학습한 내용을 바탕으로 즉각적으로 새로운 작업을 처리

Unit 01. Glossary

- How to train Large Model?



3. In-context learning (맥락 학습)

- **Few-shot learning**
  - 몇 가지 예제(프롬프트)를 통해 모델이 새로운 작업을 수행
  - 이 역시 모델 가중치는 업데이트되지 않음

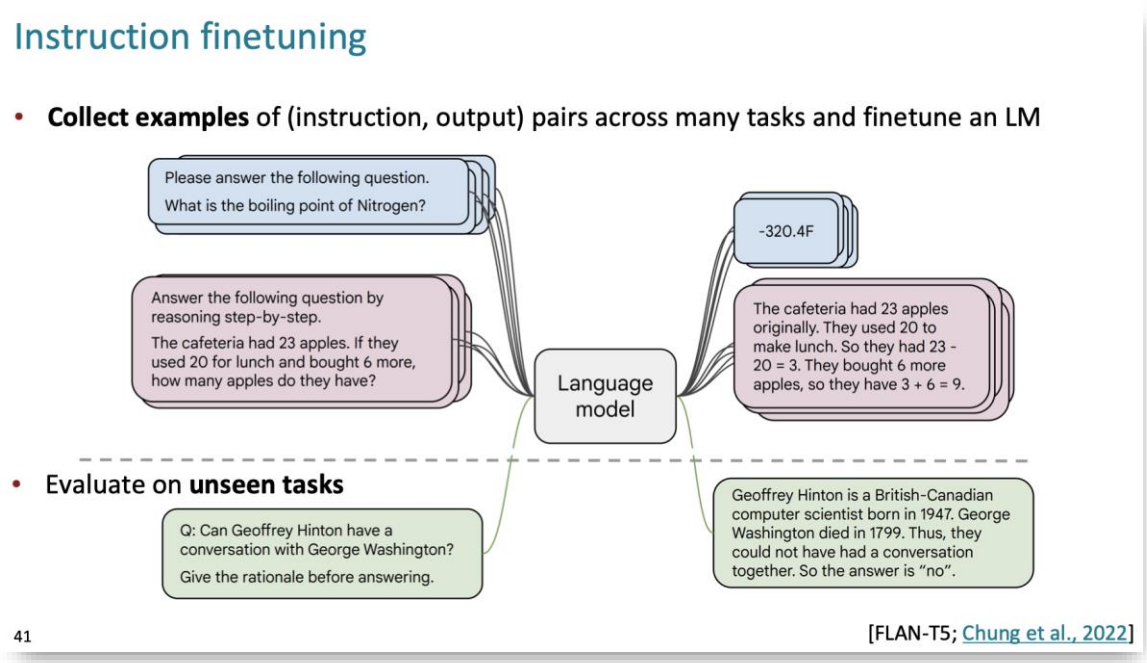
4. Memory (기억)

- **Parametric memory**
  - 모델의 모든 가중치가 업데이트되어 저장된 지식을 의미
  - 사전 학습 및 미세 조정 단계에서 사용
- **Non-parametric memory**
  - 모델의 가중치가 아닌 외부 저장소에 지식을 저장하여 사용하는 방식
  - 맥락 학습에서 사용

Unit 01. Glossary

- **Instruction Tuning**
  - 다양한 종류의 Task가 Instruction 형태로 들어 있는 (Instruction, Output)쌍의 데이터셋을 통해LM을 Fine-Tuning
  - 즉, Pre-trained 모델에 Prompt와 Completion 쌍의 데이터를 넣어 Supervised Learning을 수행하는 것
    → Unseen Task에 대해 평가를 진행했을 때, Zero-shot 성능 향상

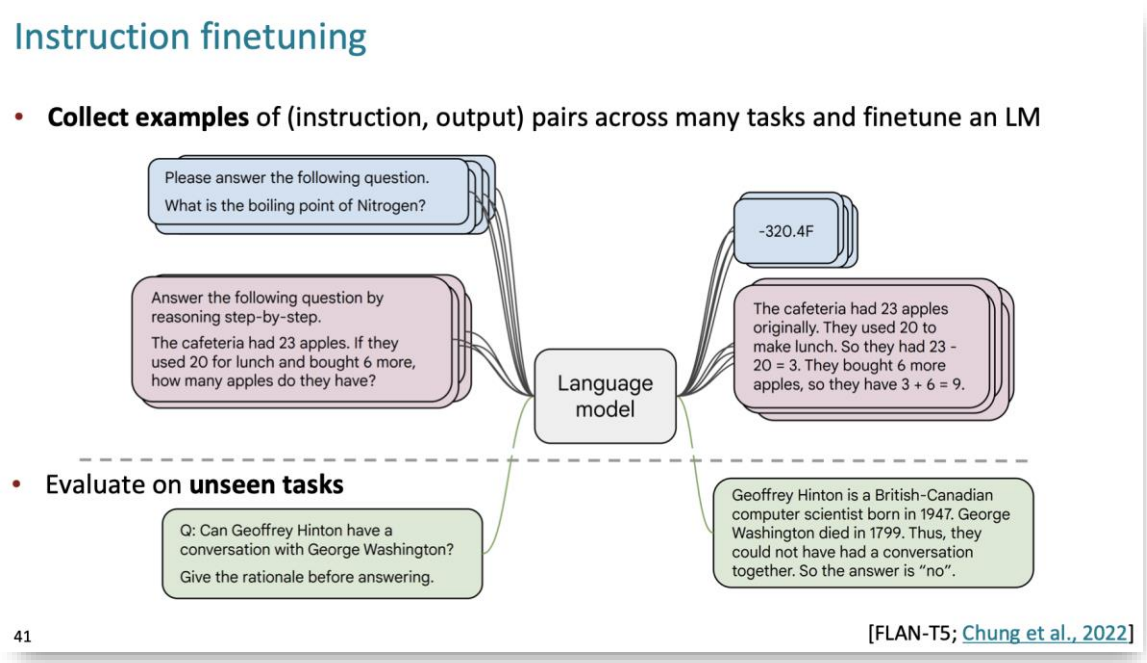
Instruction



Output

Unit 01. Glossary

- **Instruction Tuning**
  - 이를 통해, LM이 조금 더 사람이 원하는 대답의 형태로 말할 수 있도록 전체 Weight를 업데이트
  - 이는 Domain-Specific하게 Zero-Shot Learning 가능
    즉, Prompt를 주지 않아도 Domain Response를 만들 수 있음



Unit 01. Glossary

- **Instruction Tuning**

- **Instruction Tuning 장점**
  - 간단(Simple)하고 직관적(Straightforward)인 방법을 통해 높은 성능 향상을 냄
  - Unseen Task까지 Generalize할 수 있음

- **Instruction Tuning 단점**
  - 너무 많은 Task에 대한 Demonstration 수집 비용이 큼
  - LM의 Objective와 함께 사람의 Preference 사이의 Mismatch가 있음
  - 그럴 듯한 거짓말을 하는 등의 Hallucination 문제 발생
    → Hallucination 문제를 보정해주기 위해 RLHF 방법론 등장

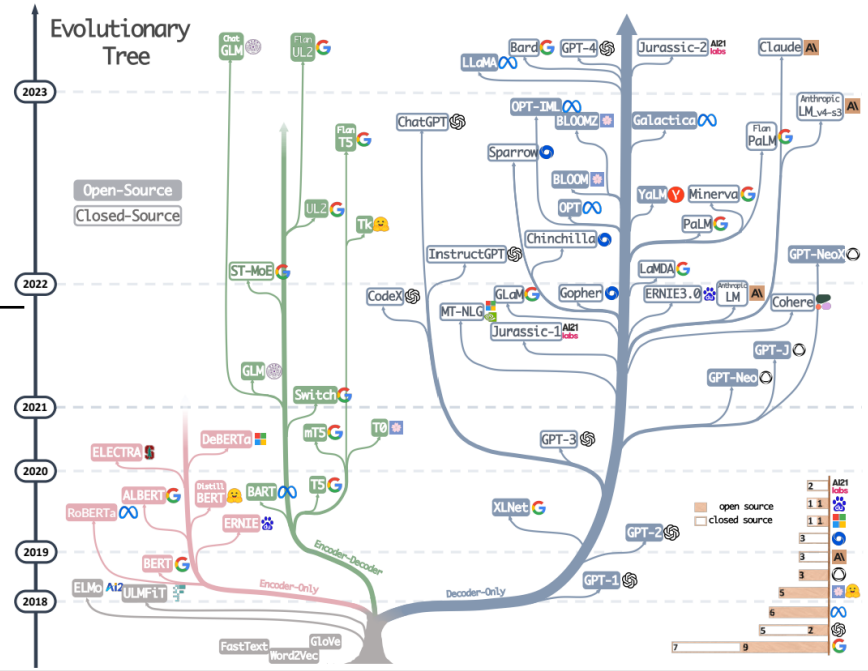
02

Background

Unit 02. Background

- LLM Progression

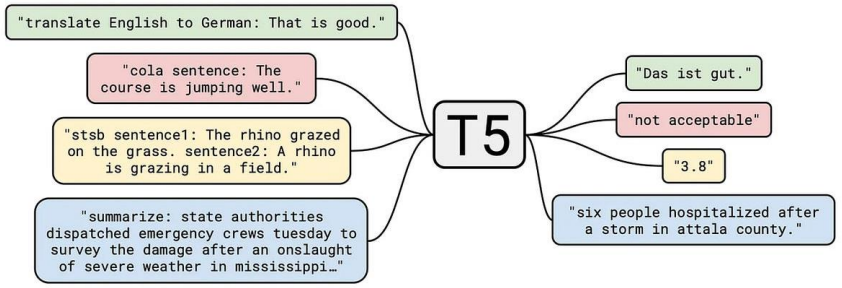
FLAN 등장 ←
Instruction Tuning의 시작
2021.09

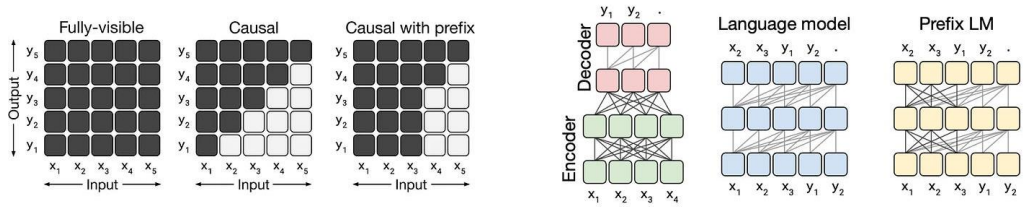


	Characteristic	LLMs
Encoder-Decoder or Encoder-only (BERT-style)	Training: Masked Language Models Model type: Discriminative Pretrain task: Predict masked words	ELMo [80], BERT [28], RoBERTa [65], DistilBERT [90], BioBERT [57], XLM [54], Xlnet [119], ALBERT [55], ELECTRA [24], T5 [84], GLM [123], XLM-E [20], ST-MoE [133], AlexaTM [95]
Decoder-only (GPT-style)	Training Autoregressive Language Models Model type: Generative Pretrain task: Predict next word	GPT-3 [16], OPT [126]. PaLM [22], BLOOM [92], MT-NLG [93], GLaM [32],Gopher [83], chinchilla [41], LaMDA [102], GPT-J [107], LLaMA [103], GPT-4 [76], BloombergGPT [117]

Unit 02. Background

- **T5: Text-to-Text Transfer Transformer**
  - Concept: We are treating all text problems in same format(Framework) → 텍스트로 다루자



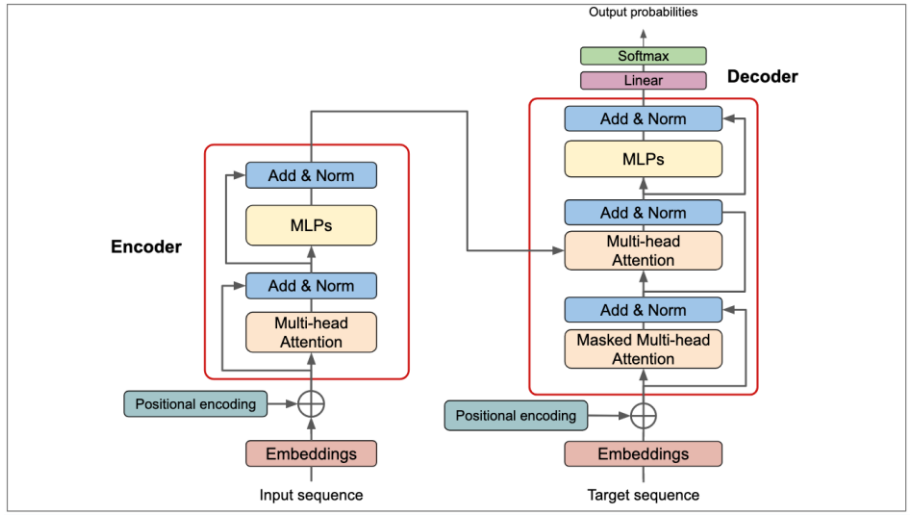


1. **Encoder-Decoder**
   - **Encoder** proceeds with an attraction in a fully invisible manner
   - **Decoder** proceeds with an attraction in a casual manner
2. **Casul LM**
   - Set it up as a GPT-like model with a autoregressive character, limiting only six layers,
     half of the baseline model, and only decoders exist
3. **Prefix LM**: the encoder used the auto-encoder format, and the decoder used the auto-encoder format

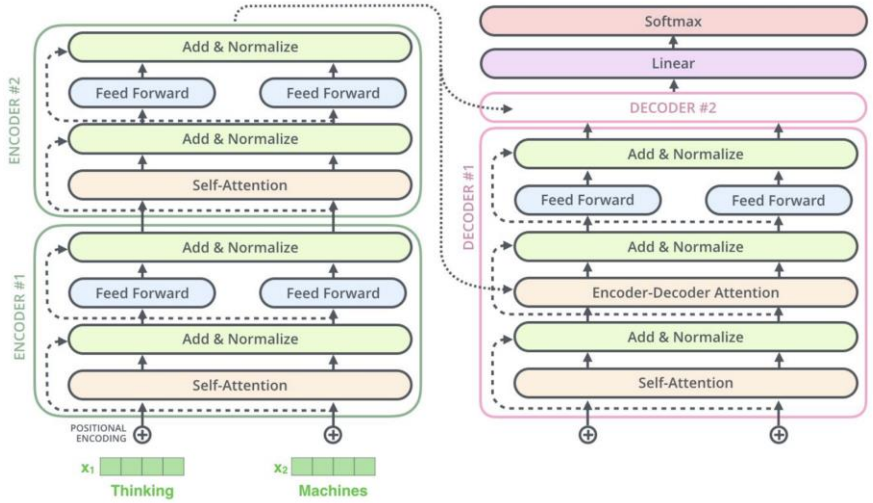
❖ **Result**
- Model with the **Encoder-Decoder structure** performed the best
- Model with **bi-directional structures** perform well

- T5: Text-to-Text Transfer Transformer
  - 기존 Transformer Architecture와 차이점

Original Transformer



T5



1) Transformer의 Layer Normalization에 사용되는 bias를 제거하고 rescale만 수행
2) Absolute positional embedding 대신 Relative positional embedding 사용
3) Model layer 전체에서 position embedding parameter sharing

Unit 02. Background

- **FLAN: Fine-tuned Language Models are Zero-shot Learners**
  - Instruction Tuning은 구글의 FLAN논문에서 처음 나온 개념
  - LLM모델을 다양한 데이터 셋에 대한 Instruction으로Fine-tuning을 진행했을 때, Unseen Task에 대한 Zero-shot성능 향상

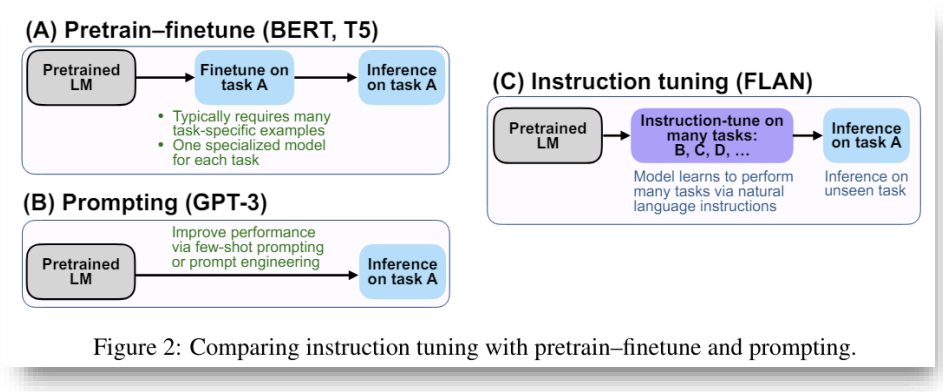


Figure 2: Comparing instruction tuning with pretrain–finetune and prompting.

	Input	Output
w/o Instruction	What's your name?	wie heiben sie?
w/ Instruction	Translate into 'What's your name?' into German	wie heiben sie?

(a)BERT/T5: Fine-tuning
(b)GPT-3: In-context Learning을 이용해 Task수행
(c)Pre-trained LM에 Fine-tuning을 진행시 Prompt 형식(Instruction)으로 변환하고,이를 이용하여 Fine-tuning

- **FLAN: Fine-tuned Language Models are Zero-shot Learners**
  - 데이터셋은 총 Task 별로 Cluster를 형성해 총 62개의 Dataset과 12개의 Task Cluster를 형성
  - Zero-shot 성능이 낮은 이유를 Zero-shot Prompt 형태가 학습된 Prompt 형태와 다르다고 원인으로 인식
  - 실제 Zero-shot Prompt 형태로 다양한 Instruction Template을 만들어 학습

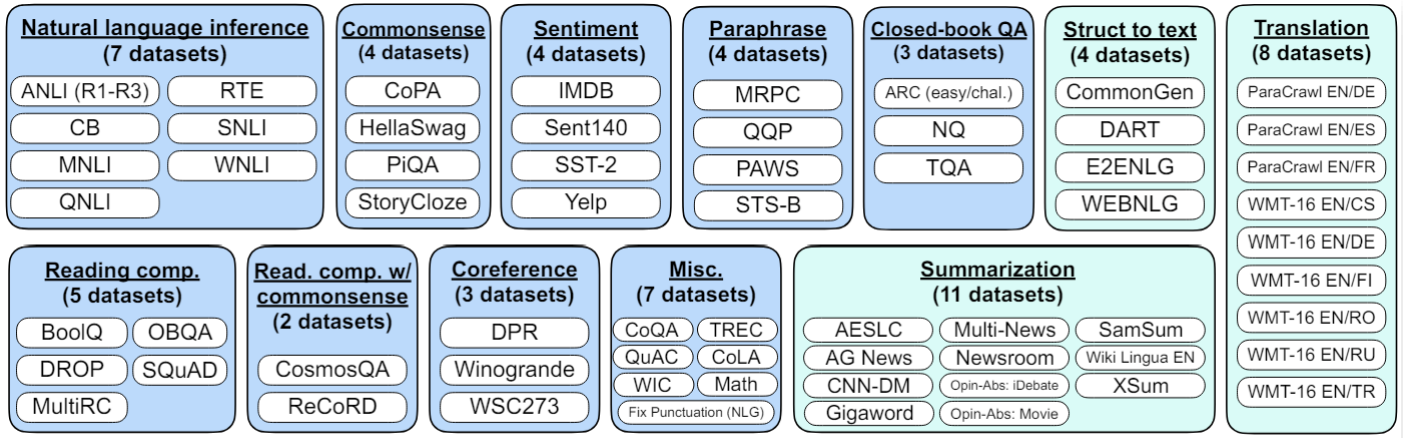


Figure 3: Datasets and task clusters used in this paper (NLU tasks in blue; NLG tasks in teal).

Unit 02. Background

- **FLAN: Fine-tuned Language Models are Zero-shot Learners**
  - 각각의 데이터 셋에 대해 10개의 Template을 가지는 Instructional Format으로 변경됨
  - 137B 크기의 모델에 Instruction Templates가 적용된 62개의 NPL 데이터셋을 학습시켜, Unseen Task 평가
    - 그 결과, 175B GPT-3를 25개 중 20개의 데이터셋에서 성능을 능가

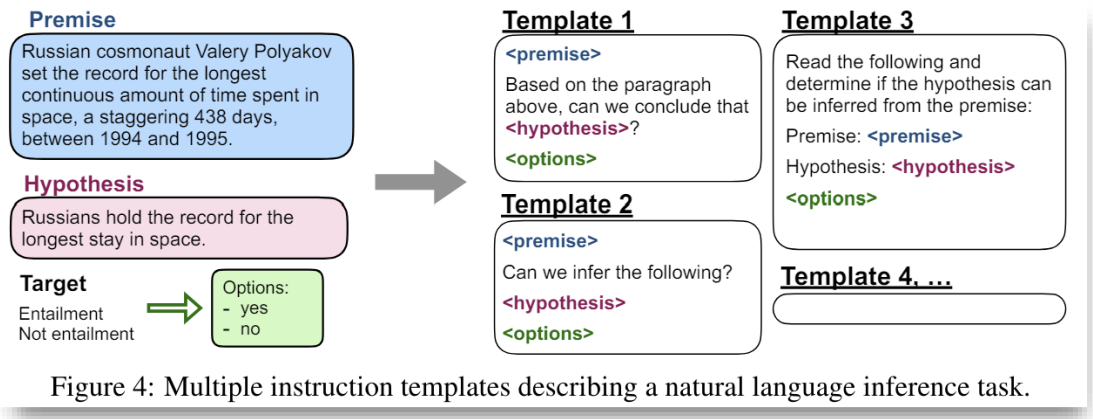
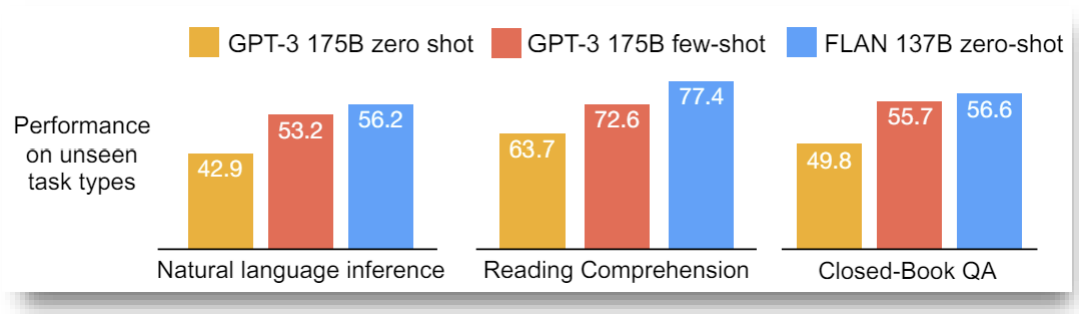
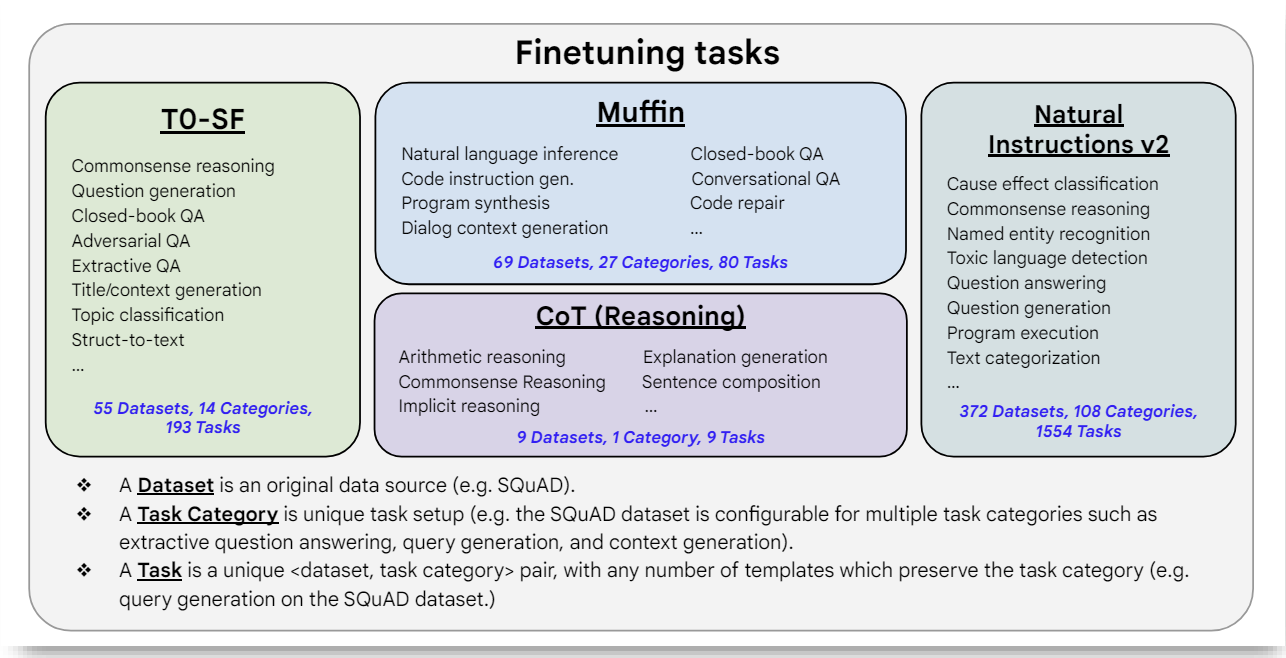


Figure 4: Multiple instruction templates describing a natural language inference task.



Unit 02. Background

- **FLAN-T5:ScalingInstruction-FinetunedLanguageModels**
  - FLAN방법론을 T5모델에 적용한 논문으로 모델이나 Task들의 수를 키워서 Scaling 함 (473Dataset,146Category,1800Task)
  - Instruction을 생성하기 위해 비슷한 유형끼리 Cluster를 모았으며, Cluster별로 Instruction Template 생성
  - Prompt Engineering시 CoT(Chain-of-thought)를 적용해 Step-by-Step으로 결론에 도달하도록 제안
  - FLAN-T5에서는 Reasoning데이터 셋을 활용하여 구축해 모든 Evaluation성능 향상



03
Method

• 1. GPT-assisted Visual Instruction Data Generation

X_c

X_b

Context type 1: Captions
A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.

X_v

Context type 2: Boxes
person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

Response type 1: conversation
Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV) ...<omitted>
Response type 2: detailed description
The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>
Response type 3: complex reasoning
Question: What challenges do these people face?
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>

<Table 1>

- **Situation 1**
  - 기존 Multimodal (Text,Image) Pair들은 Instruction Follwing을 사용하기에는 적절치 않은 Instruction들이 대다수
  - Instruction Following Dataset: (Query, Answer) Pair
- **Situation 2**
  - AutoReregressive LM(GPT) Series가 query(prompt)를 바탕으로 높은 퀄리티의 데이터 생성 가능
- **'Naïve' Data reformation Pipeline**
  - $X_v, X_c \rightarrow X_q$: GPT에게 X_c처럼 설명하도록 만드는 질문(query)를 생성
    - "GPT야. X_c처럼 설명하도록 만드는 "질문(쿼리)"를 만들어봐
  - 각 X_q에 대하여 데이터 인스턴스 생성:
    $Human: X_q, X_v < STOP > \backslash n\ Assitant: X_c < STOP > \backslash n$
    - Inefficient Diversity
    - Inefficient In-depth Reasoning

$\boldsymbol{X_v}: Image, \boldsymbol{X_c}: Associated\ Caption, \boldsymbol{X_q}: Natural\ Set\ of\ questions$

Unit 03. Method

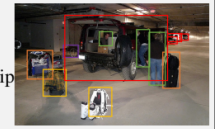
• 1. GPT-assisted Visual Instruction Data Generation

X_c

X_b

X_v

Context type 1: Captions
A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.



Context type 2: Boxes
person: [0.681, 0.242, 0.774, 0.694], person: [0.63, 0.222, 0.686, 0.516], person: [0.444, 0.233, 0.487, 0.34], backpack: [0.384, 0.696, 0.485, 0.914], backpack: [0.755, 0.413, 0.846, 0.692], suitcase: [0.758, 0.413, 0.845, 0.69], suitcase: [0.1, 0.497, 0.173, 0.579], bicycle: [0.282, 0.363, 0.327, 0.442], car: [0.786, 0.25, 0.848, 0.322], car: [0.783, 0.27, 0.827, 0.335], car: [0.86, 0.254, 0.891, 0.3], car: [0.261, 0.101, 0.787, 0.626]

Response type 1: conversation
Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV).
Question: Where is the vehicle parked?
Answer: The vehicle is parked in an underground parking area, likely in a public garage.
Question: What are the people in the image doing?
Answer: The people are trying to fit all of their luggage into the SUV, likely preparing for a trip.

Response type 2: detailed description
The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip.
In the parking area, various luggage items are scattered around the vehicle. There are two backpacks, one located near the left rear wheel and the other closer to the right side of the vehicle. Additionally, there are two suitcases, one on the right side of the car and another further away near the center of the parking area. A bicycle can also be seen on the left side of the vehicle.
Other cars are parked around the main SUV, with one car positioned behind it and slightly to the left, another behind and slightly to the right, and the third car further behind on the right side.

Response type 3: complex reasoning
Question: What challenges do these people face?
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings to accommodate. They might have to strategize and arrange the luggage efficiently to ensure that everything fits properly into the vehicle. Additionally, they need to consider the comfort of the passengers and visibility while driving, so the placement of the luggage must not obstruct the driver's view or make the passengers uncomfortable during the trip.

<Table 14>

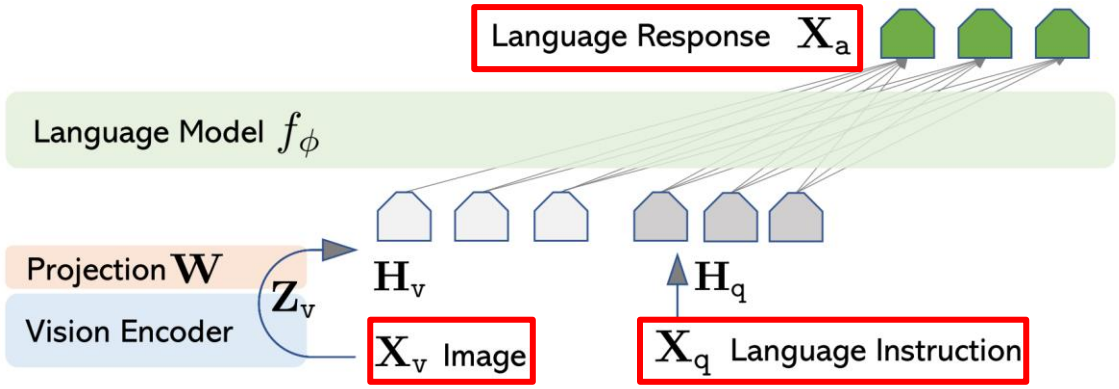
① ② ③

- **Problem**
  - 각 X_q에 대하여 데이터 인스턴스 생성:
    $Human: X_q, X_v < STOP > \backslash n\ Assitant: X_c < STOP > \backslash n$
    - Inefficient Diversity
    - Inefficient In-depth Reasoning
- **Solution**
  - 캡션은 일반적으로 다양한 관점에서 시각 장면을 설명
  - 바운딩 박스는 장면 내의 객체를 지역화하며, 각 박스는 객체 개념과 그 공간적 위치를 인코딩
- **New Reformation Pipeline**
  - Given: $X_c, X_v \& X_b (b - boxes\ of\ objects)$ of 'COCO images' Dataset
  - 사람이 직접 $\{X_{q_1}, X_{a_1}\}$ 생성
  - $X_c, X_b, \{X_{q_1}, X_{a_1}\} \rightarrow X_{q_2}, X_{a_2}$
    - "GPT야, X_c, X_b, few-shot examples를 참고해서 ~~~ 스타일의 user 'question', assistant 'answer'를 만들어봐
  - $\{X_q, X_a\} := \{X_{q_1}, X_{a_1}\} \cup \{X_{q_2}, X_{a_2}\}$
  - 생성된 3가지 스타일(종류)의 $\{X_q, X_a\}$ 예시

$\boldsymbol{X_v}: Image, \boldsymbol{X_c}: Associated\ Caption, \boldsymbol{X_q}: Natural\ Set\ of\ questions$

- **1. Data reformation pipeline 새롭게 제시**
    - $X_c, X_b, \{X_{q_1}, X_{a_1}\} \rightarrow X_{q_2}, X_{a_2}$
        - "GPT야, X_c, X_b, few-shot examples를 참고해서
          ~~~ 스타일의 user 'question', assistant 'answer'를 만들어봐

```
messages = [ {"role":"system", "content": f"""You are an AI visual assistant, and you are
seeing a single image. What you see are provided with five sentences, describing the same image you
are looking at. Answer all questions as you are seeing the image.

Design a conversation between you and a person asking about this photo. The answers should be in a
tone that a visual AI assistant is seeing the image and answering the question. Ask diverse questions
and give corresponding answers.

Include questions asking about the visual content of the image, including the object types, counting
the objects, object actions, object locations, relative positions between objects, etc. Only include
questions that have definite answers:
(1) one can see the content in the image that the question asks about and can answer confidently;
(2) one can determine confidently from the image that it is not in the image. Do not ask any question
that cannot be answered confidently.

Also include complex questions that are relevant to the content in the image, for example, asking
about background knowledge of the objects in the image, asking to discuss about events happening in
the image, etc. Again, do not ask about uncertain details. Provide detailed answers when answering
complex questions. For example, give detailed examples or reasoning steps to make the content more
convincing and well-organized. You can include multiple paragraphs if necessary."""}
]
```

<Table 13>

Captions
There is a movie theater that displays the show times above the doors.
A red fire hydrant is deep in the snow.
The fire hydrant is in the snow near a recently plowed sidewalk.
This city has had a very hard winter with snow.
A hotel for dogs in the snow in winter.

Conversation
Question:
What color is the fire hydrant in the image?
===
Answer:
The fire hydrant in the image is red.
===
Question:
Is the fire hydrant fully covered by snow or partially visible?
===
Answer:
The fire hydrant is partially visible, with a significant portion of it covered by the deep snow.
===
Question:
What is the condition of the sidewalk near the fire hydrant?
===
Answer:
The sidewalk near the fire hydrant has been recently plowed, so it is relatively clear of snow compared to the surrounding area.
===

<Table 15>

- 최종 구축된 데이터 (Image-Text instruction Following Samples)
    - Dataset = $\{X_c, X_v, X_q, X_a, (X_b)\}\}$
    - Dataset size: 158k (conversation) + 23k (detailed description) + 77k(complex reasoning)
          ~~~

- 2. Architecture



$$\mathbf{H}_v = \mathbf{W} \cdot \mathbf{Z}_v, \text{ with } \mathbf{Z}_v = g(\mathbf{X}_v)$$

Figure 1: LLaVA network architecture.

- {Pre-trained LLM과 Vision Encoder}를 사용하고 있음
  - LLM: LLaMA
  - Vision Encoder: VIT-L/14
- Image를 LLM Input으로 conversion하여 사용하고 있으므로, Multi-modal을 위한 LLM 사용 방법론을 제시했다고 볼 수 있음
  - **Vision Encoder → Image Encoder** 역할 수행
  - **LLM → Text Decoder**의 역할을 수행

- 3. Training

Instruction Tuning

- Autoregressive training objective
- Data Preparation
  - 1) Instruction Tuning Dataset(3가지 종류, COCO 기반): (X_q, X_a) pair의 sequence라고 볼 수 있음
    - 녹색 부분만 loss에 반영

$$\mathbf{X}_{\text{system-message}} \texttt{<STOP>}$$
$$\texttt{Human}: \mathbf{X}^1_{\text{instruct}} \texttt{<STOP>}\ \texttt{Assistant}: \mathbf{X}^1_{\text{a}} \texttt{<STOP>}$$
$$\texttt{Human}: \mathbf{X}^2_{\text{instruct}} \texttt{<STOP>}\ \texttt{Assistant}: \mathbf{X}^2_{\text{a}} \texttt{<STOP>} \cdots$$

  - 2) Pre-training dataset for projection(W)
    - CC3M를 필터링하여 595K Image-Text Pair를 얻음
    - Filtered CC3M에 대해, 'Naïve' Data Reformation Pipeline를 적용한 후, 위 1)처럼 format을 맞춤

$$\mathbf{X}^t_{\text{instruct}} = \begin{cases} \text{Randomly choose } [\mathbf{X}^1_{\text{q}}, \mathbf{X}_{\text{v}}] \text{ or } [\mathbf{X}_{\text{v}}, \mathbf{X}^1_{\text{q}}], & \text{the first turn } t = 1 \\ \mathbf{X}^t_{\text{q}}, & \text{the remaining turns } t > 1 \end{cases}$$

- 3. Training

Stage 1: Pre-training for Feature Alignment

- {LLM, Visual Encoder}는 freeze하고, W 부분만 학습

Stage 2: Fine-tuning End-to-End

- Visual Encoder는 freeze
- {LLM,W}는 학습

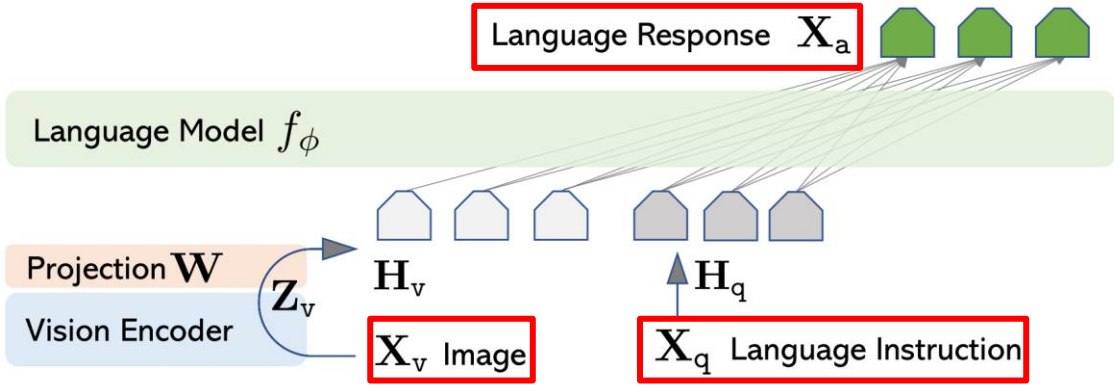


Figure 1: LLaVA network architecture.

04

Experiments

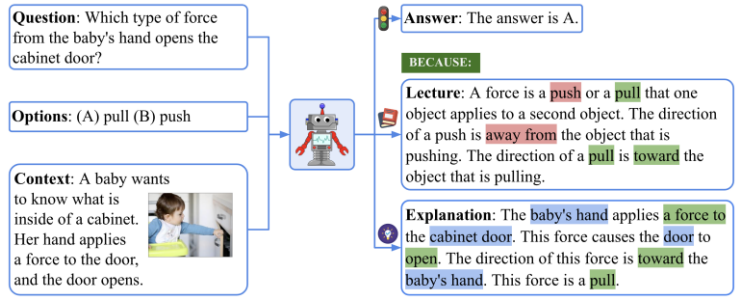
Unit 04. Experiments

- **Multimodal Chatbot – Quantitative Evaluation**
- **Data**
  - COCO Validation Split에서 30개를 추출
  - Proposed Data Reformation Pipeline을 사용하여 3가지 종류의 $\{X_q, X_a\}$생성
- **LLaVA**
  - Input: $\{X_q, X_v\}$
  - Output: Predicted X_a
- **Rival Model: GPT**
  - Input: $\{X_q, X_c, X_b\}$
  - Output: Predicted X_a
- **Judge Model: GPT-4**
  - "GPT-4기준, LLaVA가 얼마나 잘 했을까?"
  - Output: 1~10점 사이의 평가 점수
    (Helpfulness, Relevance, Accuracy, Level of details 고려)

	Conversation	Detail description	Complex reasoning	**All**
Full data	83.1	75.3	96.5	85.1
Detail + Complex	81.5 (-1.6)	73.3 (-2.0)	90.8 (-5.7)	81.9 (-3.2)
Conv + 5% Detail + 10% Complex	81.0 (-2.1)	68.4 (-7.1)	91.5 (-5.0)	80.5 (-4.4)
Conversation	76.5 (-6.6)	59.8 (-16.2)	84.9 (-12.4)	73.8 (-11.3)
No Instruction Tuning	22.0 (-61.1)	24.0 (-51.3)	18.5 (-78.0)	21.5 (-63.6)

Unit 04. Experiments

- **ScienceQA**
- **Data**
  - Size: 21k
  - Multiple choice question on various subjects and topics
  - Train/Valid/Test =12726/4241/4241
- **LLaVA**
  - Train with 12 epochs with prompt for {reasons, answer}



GPT4&LLaVA Ensemble

- GPT4가 결과 생성 실패 시, LLaVA 결과 사용
- LLaVA, GPT4 결과가 다를시, GPT4에게 다시 질문

Method	Subject			Context Modality			Grade		Average
	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	
Representative & SoTA methods with numbers reported in the literature									
Human [34]	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	88.40
GPT-3.5 [34]	74.64	69.74	76.00	74.44	67.28	77.42	76.80	68.89	73.97
GPT-3.5 w/ CoT [34]	75.44	70.87	78.09	74.68	67.43	79.93	78.23	69.68	75.17
LLaMA-Adapter [59]	84.37	88.30	84.36	83.72	80.32	86.90	85.83	84.05	85.19
MM-CoT$_{Base}$ [61]	87.52	77.17	85.82	87.88	82.90	86.83	84.65	85.37	84.91
MM-CoT$_{Large}$ [61]	95.91	82.00	90.82	95.26	88.80	92.89	92.44	90.31	91.68
Results with our own experiment runs									
GPT-4†	84.06	73.45	87.36	81.87	70.75	90.73	84.69	79.10	82.69
LLaVA	90.36	95.95	88.00	89.49	88.00	90.66	90.93	90.90	90.92
LLaVA+GPT-4† (complement)	90.36	95.50	88.55	89.05	87.80	91.08	92.22	88.73	90.97
LLaVA+GPT-4† (judge)	91.56	96.74	91.09	90.62	88.99	93.52	92.73	92.16	**92.53**

들어주셔서 감사합니다.