Chameleon: Mixed-Modal Early-Fusion Foundation Models

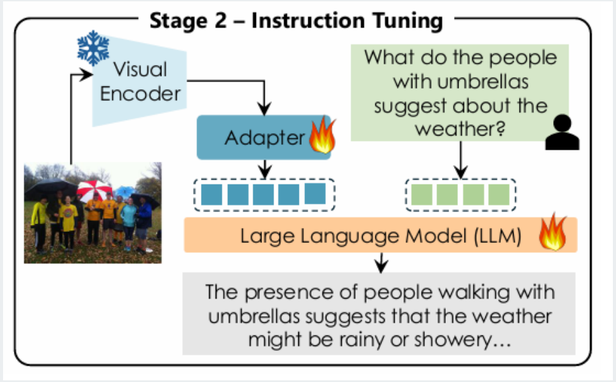
Chameleon Team
FAIR at Meta

May 17, 2024

발제자: 전진구
2025.05.01

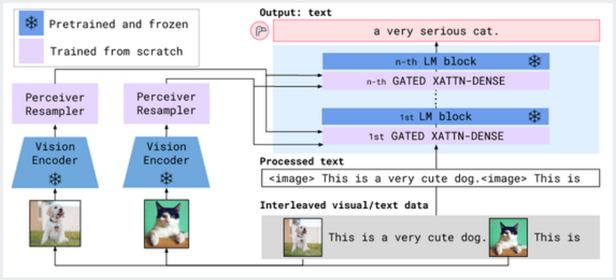
연구배경 및 문제정의 - Multimodal AI의 현주소

최근 멀티모달 파운데이션 모델은 매우 널리 채택되고 있지만, 여전히 서로 다른 Modality(양식)을 개별적으로 모델링하는 경우가 많다.



(LLaVA-MORE 아키텍처)

LLaVA-MORE와 같은 아키텍처는 시각 정보를 처리하기 위한 별도의 **Visual Encoder**와 텍스트 처리를 위한 **Large Language Model (LLM)**을 두고, 이 둘을 **Adapter**를 통해 연결하는 방식을 사용.



(Flamingo 아키텍처)

Flamingo와 같은 모델들은 강력한 사전 학습된 비전 모델과 언어 모델을 각각 활용하고, 이들을 **Perceiver Resampler**나 **GATED XATTN-DENSE** 계층과 같은 새로운 구성 요소를 통해 연결하여 시각적 정보를 언어 모델의 여러 계층에 주입하는 방식을 사용.

이러한 접근 방식들은 각 단일 양식 모델의 강력한 성능을 활용할 수 있다는 장점이 있지만, 여러 양식에 걸쳐 정보를 깊이 있게 통합하거나, 이미지와 텍스트가 interleaved(임의의 순서로 자유롭게 섞여 있는) Multimodal Document를 생성하는 데 있어 여전히 한계를 드러낼 수 있다.

LLaVA-MORE: A Comparative Study of LLMs and Visual Backbones for Enhanced Visual Instruction Tuning (by F Cocchi · 2025)
Flamingo: a Visual Language Model for Few-Shot Learning (by JB Alayrac · 2022)

연구배경 및 문제정의 - Chameleon이 나아갈 방향

경계 없는 이해와 생성 능력 확보

- 이미지와 텍스트가 임의의 순서로 배열된 시퀀스를 자연스럽게 이해하고 생성

- 단순 이미지 캡셔닝이나 텍스트 기반 이미지 생성을 넘어, 완전한 멀티모달 문서 자체를 모델링

- 경계 없이 자유로운 멀티모달 소통 능력을 목표

멀티모달 상호작용의 새로운 지평 제시

- 광범위한 Vision-Language 벤치마크에서 강력한 성능과 새로운 Mixed-modal 추론/생성 능력

- 설명 텍스트와 함께 직접 생성한 이미지를 포함하는 답변 등, 풍부하고 유연한 상호작용

- 기존 모델의 한계를 넘어선 새로운 멀티모달 경험을 제공

Early-Fusion(조기 융합) 기반의 단일 통합 아키텍처 실행

- 모든 양식을 초기부터 통일된 토큰 기반으로 처리하며, 단일 아키텍처에서 End-To-End 훈련.

- 안정적 훈련 방법론과 최적화된 아키텍처로 기존 기술적 과제를 극복.

- 진정한 의미의 Early-Fusion 기반 혼합 양식 모델을 구현.

통합 멀티모달 파운데이션 모델의 새 기준 정립

- 멀티모달 콘텐츠를 유연하게 추론하고 생성하는 통합 파운데이션 모델 비전을 실현

핵심 아이디어 - Pretraining - Dataset

First Stage(80%)

대규모의 다양한 비지도 데이터를 활용하여 기본적인 mixed-modal 이해 및 생성 능력을 구축하는 단계

1. Text-Only Data

- LLaMa-2(H Touvron · 2023) 및 CodeLLaMa(B Rozière · 2023) 학습에 사용된 사전 훈련 데이터를 포함한 다양한 텍스트 데이터 셋을 활용.
- 총 2.9조 개의 텍스트 전용 토큰으로 구성됨.

2. Text-Image Pair Data

- 공개적으로 사용 가능한 데이터 소스와 라이선스가 부여된 데이터를 조합하여 사용
- 총 14억 개의 텍스트-이미지 쌍 → 1.5조 개의 텍스트-이미지 토큰

3. Text/Image Interleaved Data

- Meta의 제품 또는 서비스 데이터를 제외한 공개 웹 소스에서 데이터를 확보(웹 크롤링)(H Laurençon · 2023)
- 4천억 개의 인터리빙된 텍스트 및 이미지 데이터 토큰

Second Stage(20%)

고품질·선별 instruction tuning(지시 튜닝) 데이터 혼합으로 성능 및 alignment(정렬)을 강화하는 단계

1단계에서 사용된 데이터의 가중치는 50% 낮추고 진행

유사한 이미지-텍스트 토큰 비율을 유지하면서, 더 정제되고 품질이 높은 데이터셋을 추가로 혼합

Llama 2: Open Foundation and Fine-Tuned Chat Models(by H Touvron · 2023)
Code Llama: Open Foundation Models for Code(by B Rozière · 2023)
OBELICS: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents(by H Laurençon · 2023)

핵심 아이디어 - Pretraining - Tokenization

모든 양식(이미지, 텍스트)을 초기부터 통일된 discrete(이산)토큰 으로 처리하여, 단일 아키텍처에서 End-To-End 방식으로 처리하는 것이 목표

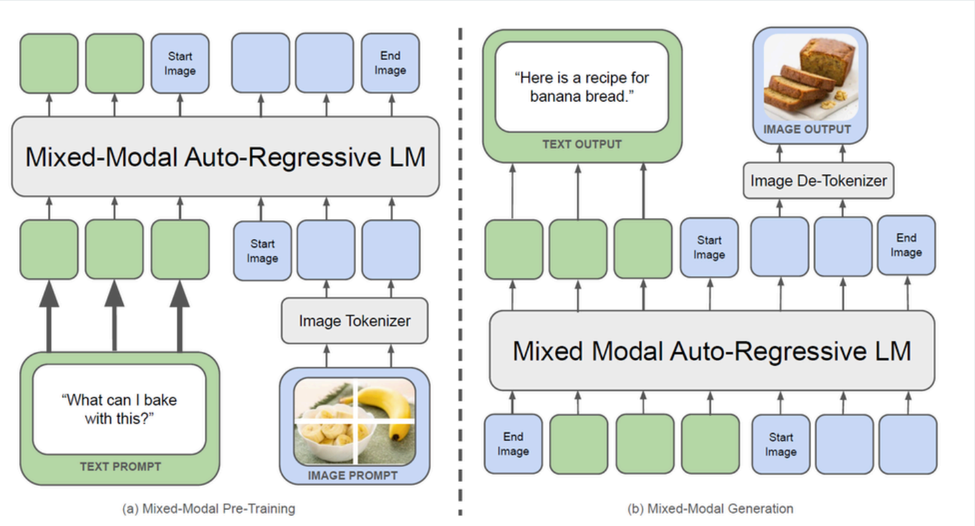


Image Tokenization

- VQ-VAE (Vector-Quantized Variational Autoencoder)(by A Oord · 2017) 기반의 이미지 토크나이저를 새로 훈련하여 사용
- 사람 얼굴 생성의 중요성을 고려하여, 얼굴이 포함된 이미지의 비율을 2배로 up-sample하여 훈련 (by O Gafni · 2022)

Text Tokenization

- BPE(Byte Pair Encoding)(by R Sennrich · 2015) 기반의 토크나이저를 사용
- SentencePiece 라이브러리(by T Kudo · 2018)를 사용하여 BPE 토크나이저를 훈련

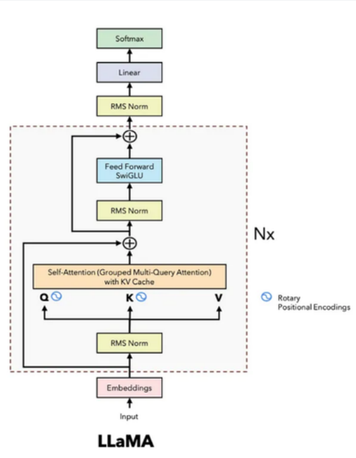
통합 처리 및 결과 생성

1. discrete 토큰으로 변환된 이미지 토큰과 텍스트 토큰은 순서에 관계없이 하나의 시퀀스로 결합되어 Mixed-Modal Auto-Regressive LM(LLaMA-2 기반)의 입력으로 사용됨
2. 모델은 이 토큰 시퀀스를 바탕으로 다음 토큰을 예측하며, 이 과정에서 텍스트 뿐만 아니라 이미지에 해당하는 토큰들도 생성
3. 생성된 이미지 토큰들은 Image De-Tokenizer(VQ-VAE의 디코더 부분)를 통해 최종 IMAGE OUTPUT으로 복원

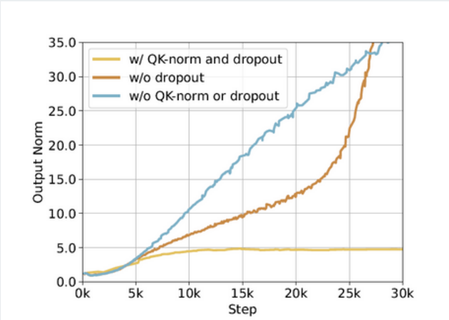
Neural Discrete Representation Learning(by A Oord · 2017)
Scene-Based Text-to-Image Generation with Human Priors(by O Gafni · 2022)
Neural Machine Translation of Rare Words with Subword Units(by R Sennrich · 2015)
SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing(by T Kudo · 2018)

핵심 아이디어 - Pretraining - Stability

Chameleon의 훈련 instability(불안정성) 문제



(LLaMA 2's model architecture)



(학습 발산 발생)

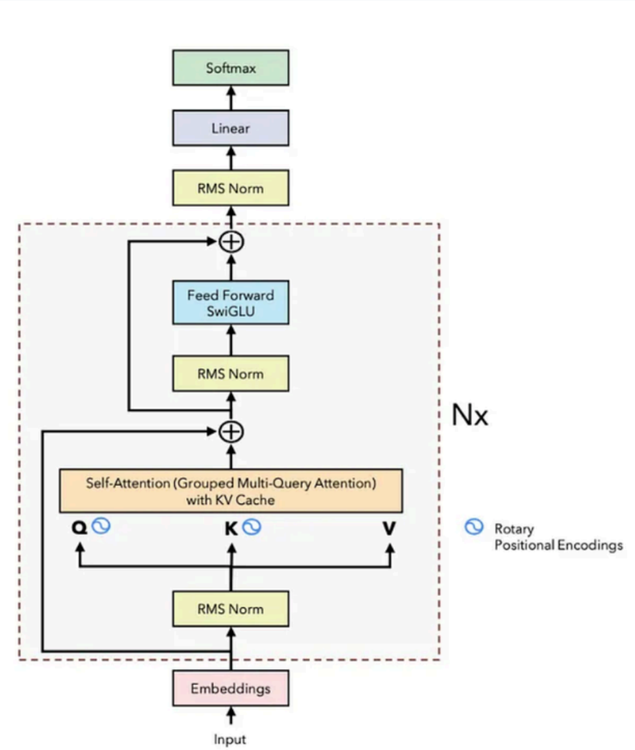
- Chmeleon은 두 modal(이미지+텍스트)을 토큰화 하여 하나의 토큰 시퀀스로 처리
- 이 하나의 토큰 시퀀스는 LLaMa 기반의 Mixed-Modal Auto-Regressive LM에 입력됨.
- 두 modal이 모델의 가중치를 공유
- 이때 각 modal 은 softmax 연산의 이동 불변 특성(softmax(z)=softmax(z+c)) 때문에 자신의 중요도를 높이기 위해 자신의 내부 표현(텐서, 예: [1, 512, 768]) 값norm(크기)를 키우는 경쟁이 발생
- Self-Attention 계층의 내부 softmax 연산을 거칠때 마다 내부 표현(텐서) norm이 커짐
- 내부 표현(텐서)의 norm 이 계속 증가하다가 bf16 (bfloat16)과 같은 유효 표현 범위를 벗어나 정확한 값 표현이 불가능해지고 수치적 오류(예: 오버플로우로 인한 무한대 값)가 발생

정상적인 로짓값: [2.5, 3.0, 0.5]
비정상적인 로짓값: [Infinity, 100.0, 50.0]
↓ softmax
[0.359,0.592,0.049]
[1.0, 0.0, 0.0]

- 오버플로우로 인해 뒤죽박죽이 된 내부 표현(텐서)의 로짓값이 출력층의 마지막 sotfmax 함수를 통과하게 되면 확률 분포가 과도하게 편향된 예측을 하게 되고
- 손실 함수의 값이 매우 커져서 역전파 과정에서 매우 큰 기울기 값을 야기하고 결국 Exploding Gradients(기울기 폭주)가 발생
- 폭주하는 기울기는 모델의 가중치를 너무 큰폭으로 그리고 예측 불가능한 방향으로 업데이트해 안정적으로 손실을 줄여나가지 못하게 하고, 오히려 학습이 제대로 진행되지않는 training divergence(학습 발산) 발생

핵심 아이디어 - Pretraining - Stability

Chameleon의 훈련 instability(불안정성) 해결 전략



LLaMA

Chameleon-34B:

$$h = x + \text{attention\_norm}(\text{attention}(x))$$
$$\text{output} = h + \text{ffn\_norm}(\text{feed\_forward}(h))$$

Llama2:

$$h = x + \text{attention}(\text{attention\_norm}(x))$$
$$\text{output} = h + \text{feed\_forward}(\text{ffn\_norm}(h))$$

1. QK-Norm(Query-key Normalization)

- 어텐션(Attention) 메커니즘 내부의 쿼리(Query) 및 키(Key) 벡터에 직접 계층 정규화(Layer Normalization)를 적용(by M Wortsman · 2023)
- 어텐션 내부 Softmax 함수에 입력되는 값들의 Norm 증가를 직접적으로 제어하여 안정성을 확보

2. z-loss Normalization

- Softmax 함수의 분배 함수(partition function) Z를 정규화
- Softmax의 손실함수에 Z 항 추가 $10 - 5log2Z$
- Total Loss = Main Loss (e.g., Cross-Entropy) + z-loss regularization

3. DropOut(7B)

- QK-Norm 외에 추가적으로 어텐션 및 피드포워드(Feed Forward) 계층 이후에 드롭아웃(Dropout, 0.1 비율)을 적용
- SwiGLU 활성화 함수(by N Shazeer · 2020)의 곱셈적 특성으로 인해 발생할 수 있는 Norm 증가를 제어하기 위함

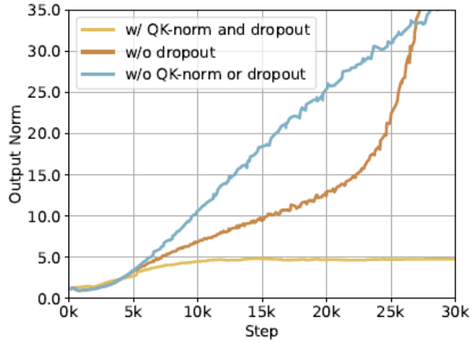
4. Swim-Transformer Normalization(34B)(by Z Liu · 2021))

- 7B 모델의 안정화 방법(QK-Norm + Dropout)만으로는 34B 모델 안정화에 충분하지 않아, 추가적인 정규화가 필요
- Swin Transformer의 정규화 전략을 차용하여, 피드포워드 계층 뒤에 RMSNorm(by B Zhang · 2019)를 위치.

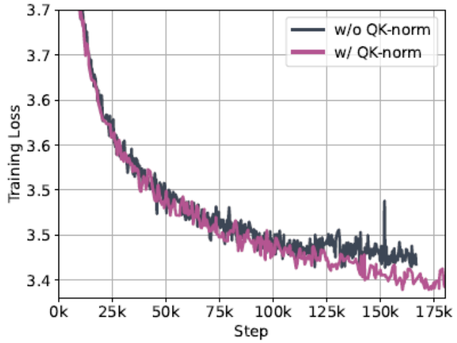
Swin Transformer: Hierarchical Vision Transformer using Shifted Windows(by Z Liu · 2021)
Small-scale proxies for large-scale Transformer training instabilities(by M Wortsman · 2023)
GLU Variants Improve Transformer(by N Shazeer · 2020)
Root Mean Square Layer Normalization(by B Zhang · 2019)

핵심 아이디어 - Pretraining - Stability

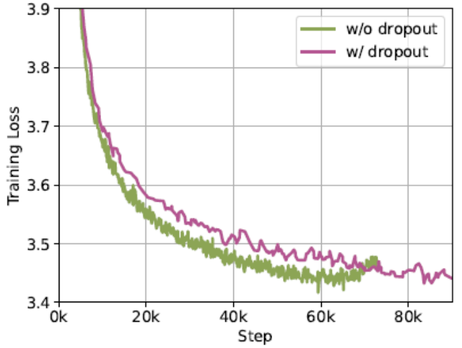
Chameleon 모델에 대한 output norm 및 훈련 손실 곡선



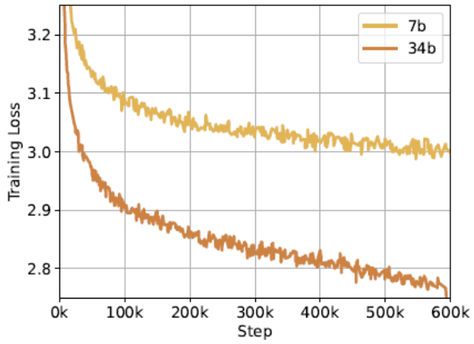
(a)



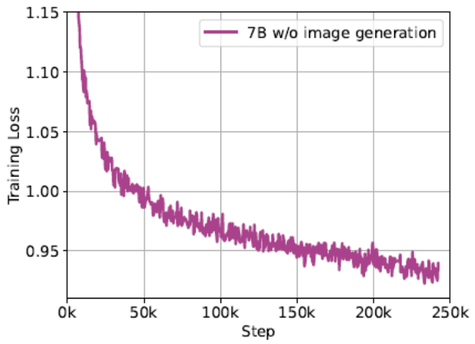
(b)



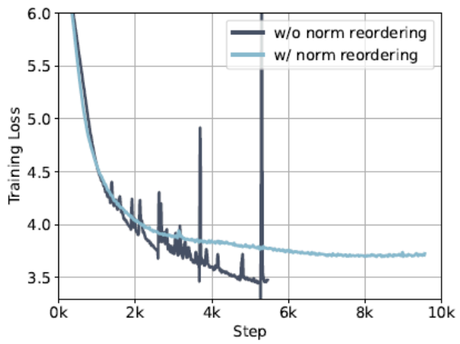
(c)



(d)



(e)

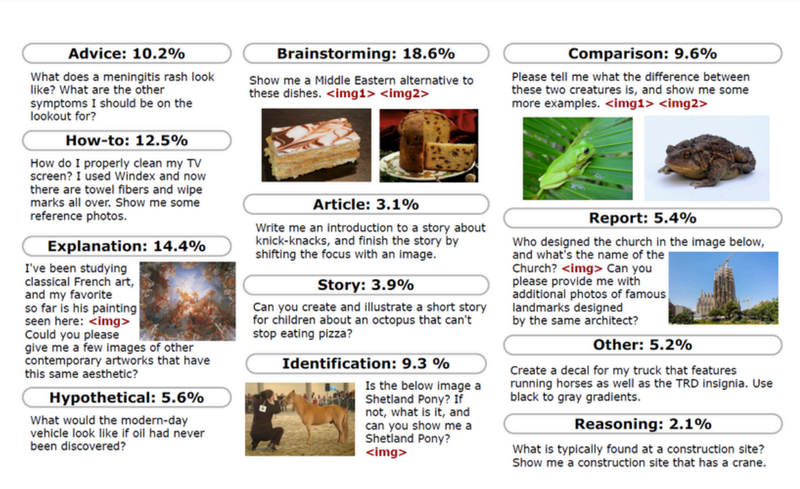


(f)

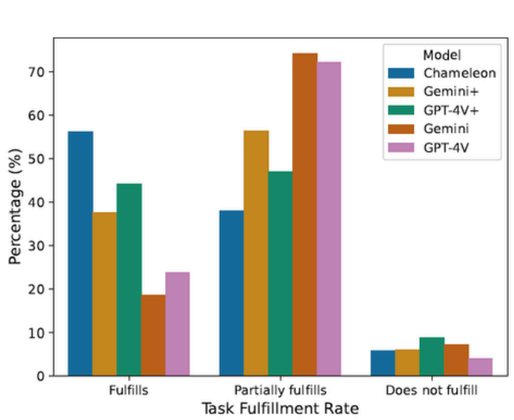
DEMO

Benchmark Evaluations

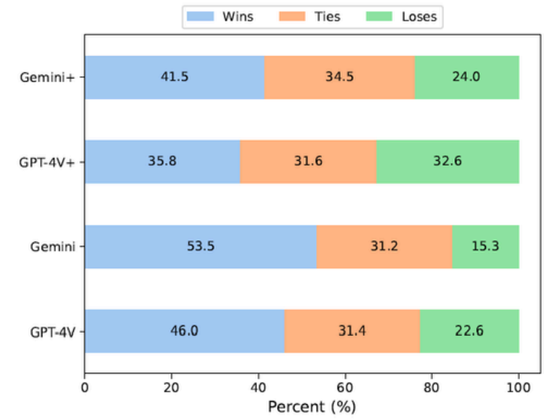
Human Evaluations(인간 평가)



평가에 사용된 총 1,048개의 프롬프트 (441개의 mixedmodal 프롬프트와 607개의 Only-text 프롬프트)는 사용자가 멀티모달 AI 시스템으로 수행하고자 하는 작업을 더 잘 이해하기 위해 수동으로 검토되어 12개의 범주로 분류됨.



(a) The prompt task fulfillment rates.



(b) Chameleon vs. the baselines: Gemini+, GPT-4V+, Gemini, GPT-4V.

(a) 프롬프트 작업 이행률

(b) Chameleon 대 기준 모델: Gemini+, GPT-4V+, Gemini, GPT-4V

	Chameleon			Gemini+			GPT-4V+		
	Fulfills	Partially fulfills	Does not fulfill	Fulfills	Partially fulfills	Does not fulfill	Fulfills	Partially fulfills	Does not fulfill
Mixed-modality	55.3%	36.7%	7.9%	39.2%	57.8%	2.9%	42.6%	52.4%	5.0%
Text-only	57.7%	38.4%	4.0%	36.4%	55.5%	8.1%	46.1%	42.7%	11.2%

	Gemini			GPT-4V		
	Fulfills	Partially fulfills	Does not fulfill	Fulfills	Partially fulfills	Does not fulfill
Mixed-modality	19.7%	76.0%	4.3%	24.3%	72.6%	3.2%
Text-only	18.3%	72.7%	9.1%	23.6%	72.0%	4.4%

Task Type	Chameleon			Gemini+			GPT-4V+		
	Fulfills	Partially fulfills	Does not fulfill	Fulfills	Partially fulfills	Does not fulfill	Fulfills	Partially fulfills	Does not fulfill
Advice	69.2%	26.2%	4.7%	42.1%	56.1%	1.9%	43.9%	48.6%	7.5%
Article	59.4%	37.5%	3.1%	40.6%	53.1%	6.3%	62.5%	37.5%	0.0%
Brainstorming	57.9%	36.4%	5.6%	33.3%	61.5%	5.1%	47.7%	47.2%	5.1%
Comparison	60.4%	34.7%	5.0%	47.5%	46.5%	5.9%	43.6%	44.6%	11.9%
Explanation	53.0%	37.7%	9.3%	33.8%	61.6%	4.6%	41.7%	50.3%	7.9%
How-to	52.7%	40.5%	6.9%	43.5%	52.7%	3.8%	48.1%	41.2%	10.7%
Hypothetical	55.9%	39.0%	5.1%	39.0%	47.5%	13.6%	42.4%	44.1%	13.6%
Identification	55.7%	33.0%	11.3%	33.0%	66.0%	1.0%	35.1%	55.7%	9.3%
Other	41.8%	40.0%	18.2%	38.2%	41.8%	20.0%	50.9%	40.0%	9.1%
Reasoning	50.0%	13.6%	36.4%	27.3%	59.1%	13.6%	31.8%	54.5%	13.6%
Report	49.1%	40.4%	10.5%	29.8%	61.4%	8.8%	38.6%	47.4%	14.0%
Story	31.7%	63.4%	4.9%	39.0%	56.1%	4.9%	53.7%	43.9%	2.4%

Task Type	Gemini			GPT-4V		
	Fulfills	Partially fulfills	Does not fulfill	Fulfills	Partially fulfills	Does not fulfill
Advice	21.5%	70.1%	8.4%	23.4%	75.7%	0.9%
Article	12.5%	84.4%	3.1%	9.4%	90.6%	0.0%
Brainstorming	18.5%	71.8%	9.7%	27.2%	66.7%	6.2%
Comparison	14.9%	76.2%	8.9%	19.8%	72.3%	7.9%
Explanation	15.2%	78.1%	6.6%	19.9%	77.5%	2.6%
How-to	19.8%	74.0%	6.1%	31.3%	67.2%	1.5%
Hypothetical	30.5%	49.2%	20.3%	32.2%	61.0%	6.8%
Identification	18.6%	75.3%	6.2%	22.7%	68.0%	9.3%
Other	14.5%	60.0%	25.5%	18.2%	67.3%	14.5%
Reasoning	9.1%	77.3%	13.6%	13.6%	81.8%	4.5%
Report	12.3%	77.2%	10.5%	22.8%	68.4%	8.8%
Story	9.8%	82.9%	7.3%	7.3%	90.2%	2.4%

Benchmark Evaluations

Text & Image-to-Text

	Chameleon		Llama-2			Mistral		Gemini Pro	GPT-4
	7B	34B	7B	34B	70B	7B	8x7B	—	—
Commonsense Reasoning and Reading Comprehension									
PIQA	79.6	83.3	78.8	81.9	82.8	83.0	83.6	—	—
SIQA	57.0	63.3	48.3	50.9	50.7	—	—	—	—
HellaSwag	74.2	82.7	77.2	83.3	85.3	81.3	84.4	—	—
	75.6 10-shot	85.1 10-shot	—	—	87.1 10-shot	83.9 10-shot	86.7 10-shot	84.7 10-shot	95.3 10-shot
WinoGrande	70.4	78.5	69.2	76.7	80.2	75.3	77.2	—	—
Arc-E	76.1	84.1	75.2	79.4	80.2	80.0	83.1	—	—
Arc-C	46.5	59.7	45.9	54.5	57.4	55.5	59.7	—	—
OBQA	51.0	54.0	58.6	58.2	60.2	—	—	—	—
BoolQ	81.4	86.0	77.4	83.7	85.0	84.7*	—	—	—
Math and World Knowledge									
GSM8k	41.6	61.4	14.6	42.2	56.8	52.1 maj@8	74.4 maj@8	86.5 maj@32 CoT	92.0 SFT CoT
	50.9 maj@8	77.0 maj@32	—	—	—	—	75.1* maj@32	—	—
MATH	11.5 maj@1	22.5 maj@1	2.5	6.24	13.5	13.1 maj@4	28.4 maj@4	32.6	52.9**
	12.9 maj@4	24.7 maj@4	—	—	—	—	—	—	—
MMLU	52.1	65.8	45.3	62.6	68.9	60.1	70.6	71.8	86.4

(a). 공개 소스 파운데이션 모델 대비
집합적 학술 벤치마크에서의 전반적인 성능 비교

	Model	Model Size	COCO	Flickr30k	VQAv2
Pre-trained	Flamingo-80B	80B	113.8 32-shot	75.1 4-shot	67.6 32-shot
	IDEFICS-80B	80B	116.6 32-shot	73.7 4-shot	65.9 32-shot
Chameleon	Chameleon	34B	120.2 2-shot	74.7 2-shot	66.0 2-shot
	Chameleon-SFT	34B	140.8 0-shot	82.3 2-shot	—
	Chameleon-MultiTask	34B	139.1 2-shot	76.2 2-shot	69.6
Fine-tuned	Flamingo-80B-FT	80B	138.1	—	82.0
	IDEFICS-80B-Instruct	80B	123.2 32-shot	78.4 32-shot	68.8 32-shot
Closed Source (finetuning status unknown)	GPT-4V	—	78.5* 8-shot	55.3* 8-shot	77.2
	Gemini Nano 2	—	—	—	67.5
	Gemini Pro	—	99.8* 2-shot	82.2* 4-shot	71.2
	Gemini Ultra	—	—	—	77.8

(b). 이미지-텍스트 변환 능력에 대한 모델 성능 비교

Conclusion

1. 완전히 토큰 기반 아키텍처

- 여러 양식(modality) 간의 원활한 정보 통합을 가능하게 하는 완전히 토큰 기반 아키텍처
- 이미지를 개별 토큰으로 처리하고 mixedmodal 데이터로 훈련하여, 기존 방식으로는 어려웠던 이미지와 텍스트의 공동 추론을 학습함

2. 확장성 및 안정성 확보

- 이전까지 Early-Fusion모델의 규모 확대를 제한했던 주요 최적화 및 아키텍처 설계의 어려움을 해결함
- 안정적이고 확장 가능한 훈련을 위한 새로운 기술을 도입함

3. 벤치마크에서의 우수한 성능

- 이미지 캡셔닝 및 시각적 질의응답과 같은 주요 비전-언어 작업에서 Chameleon-34B 모델이 Flamingo나 IDEFICS와 같은 기존의 강력한 모델들을 능가하는 성능을 보여줌
- 텍스트 전용 벤치마크에서도 경쟁력 있는 성능을 유지함

4. 새로운 mixedmodal 능력 구현 및 가능성 제시

- 단순히 기존 작업을 잘하는 것을 넘어, 새로운 혼합 양식(mixed-modal) 추론 및 생성 능력을 실제로 구현함
- mixed-modal open-ended QA(질의응답) 벤치마크에서의 강력한 성능을 통해 입증되어 멀티모달 상호작용의 완전히 새로운 가능성 제시함

Open Question

1. Chameleon 과 Liquid 성능비교

2. 프로젝트에 어떻게 활용 할 수 있을까?