

# Chameleon: Mixed-Modal Early-Fusion Foundation Models

FAIR at Meta

Presenters: Yuchen Zhu, Rodrigo Loza, Sichen Zhu

# Outline

- Related Works
- Methodology
- Experiment Results
- Pros and Cons

# Related Works

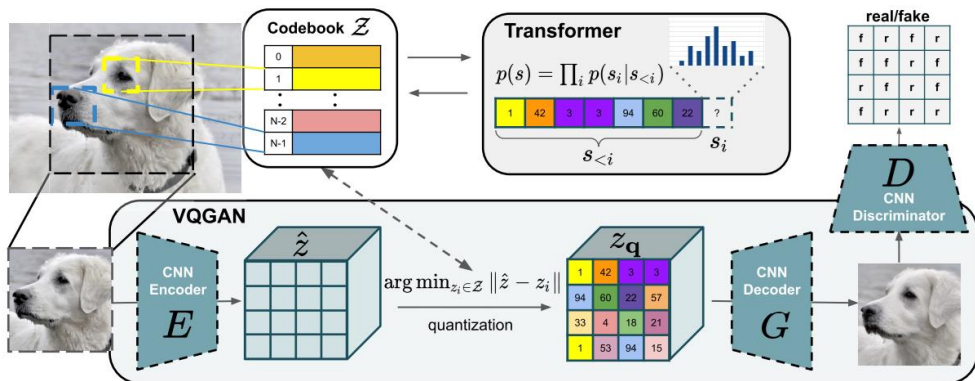
Chameleon is a **token-based**, **early-fusion**, multimodal learning algorithm through **pure autoregressive** generation.

To better understand the place of Chameleon in the literature, we compare through the following three aspects

- Image tokenization techniques & token-based SSL unimodal/multimodal representation learning
- Early-fusion models v.s. Late-fusion models
- Pure autoregressive v.s. Autoregressive + diffusion

# Images as sequences of tokens

- Discrete VAE (d-VAE, 2016) is one of the first work on generating discrete latent representation for data. Data are directly compressed to discrete tokens on the latent space.
- VQ-VAE (2017) proposes a more stable pipeline leveraging the idea of vector quantization. Instead of building a discrete latent space, VQ-VAE first learn a continuous latent representation for image patches, but force it to be representable as discrete tokens through projection onto their nearest neighbors in the codebook (vocabulary).
- VQ-GAN (2021) improves upon VQ-VAE by enabling the learning of a representation-rich codebook with additional perceptual losses and GAN losses.



Training losses:

$$\mathcal{L}_{\text{VQ}}(E, G, \mathcal{Z}) = \|x - \hat{x}\|^2 + \|\text{sg}[E(x)] - z_q\|_2^2 + \|\text{sg}[z_q] - E(x)\|_2^2.$$

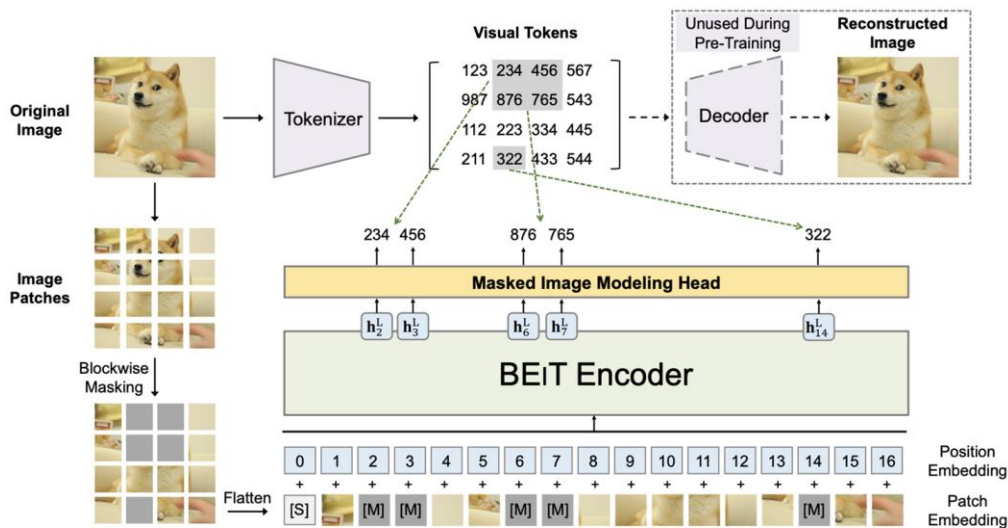
$$\mathcal{L}_{\text{GAN}}(\{E, G, \mathcal{Z}\}, D) = [\log D(x) + \log(1 - D(\hat{x}))]$$

# Start of Token-based Image SSL: BEIT

Motivated by the success of BERT and Masked Language Modeling on language tasks, BEIT attempts to make an analogy for image representation learning

Images are separated patches and encoded in two ways: 1) through d-VAE into visual tokens, 2) through linear projection head into latent vectors

Self-supervised learning task: Given a image in **pixel space** where some patches are masked, predicted the **visual tokens** with a vision transformer as a backbone



Note that:

BEIT is not for generation and there is no decoder in the training

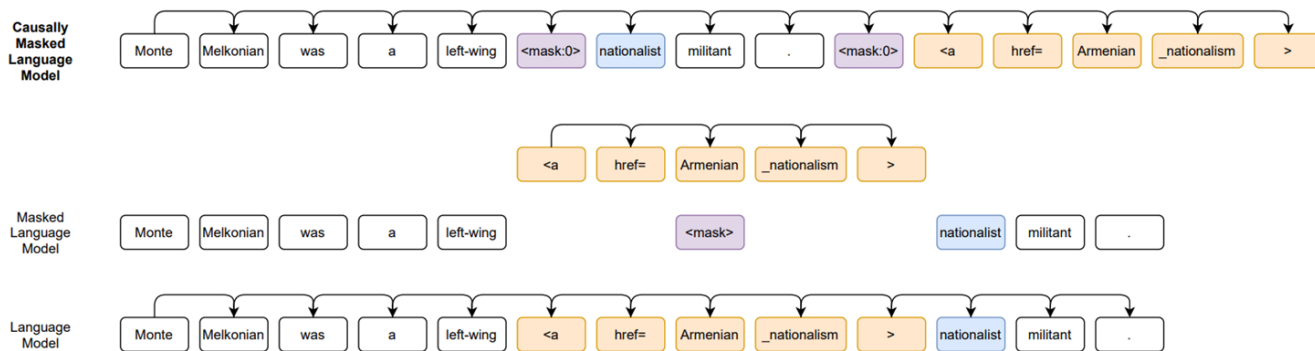
BEIT is not a pure token-based approach

# Precursor work of Chameleon: CM3 & CM3Leon

Chameleon is partially built upon two previous work from the same Meta team, CM3 & CM3Leon, which extends the token-based modeling approach to multimodal data in a **pure autoregressive** fashion.

CM3 considered a causally masked language training objective, where multimodal content such as hyperlinks and images are replaced with a placeholder `<mask>` token in the middle of text generation, and they will be generated at the end of the sentence. Due to this format, the models can be trained autoregressively with next token prediction using a decoder-only transformer.

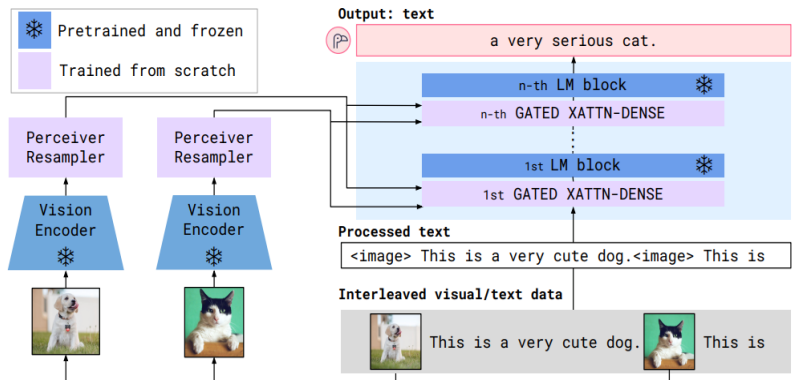
CM3Leon scales up CM3 and add a second stage of SFT to align this general purpose model. CM3Leon can perform both text-to-image generation and image-to-text captioning with competitive performance.



# Early-fusion Models v.s. Late-fusion Models

**Late-fusion models:** data of different modalities are encoded separately with different encoders before passing into the model for fusion

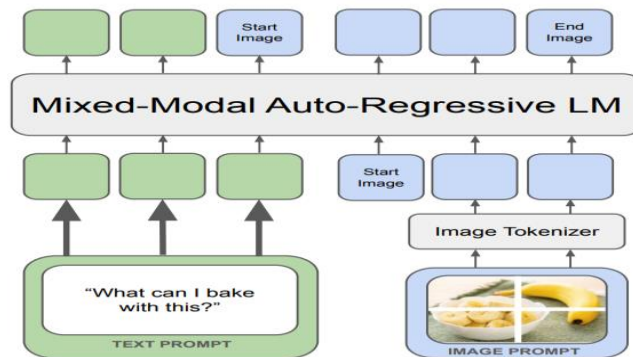
Examples: Flamingo, ViLT, LLaVA, Unified-IO, etc



Late-fusion: Flamingo

**Early-fusion models:** data of different modalities are encoded into discrete tokens with a shared vocabulary, then representation are learned in a shared space in an early stage

Examples: Chameleon, Show-o, Transfusion



Early-fusion: Chameleon

# Chameleon v.s. Other early-fusion models

After Chameleon, two other mixed modal **early fusion models** show up, namely Show-o and Transfusion.

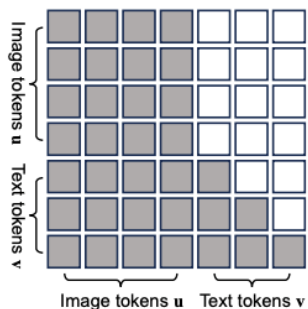
Similar to Chameleon, both Show-o and Transfusion tokenize images into discrete tokens and use one transformer to model the generation, which is an **early fusion** approach.

Unlike Chameleon which is **purely autoregressive**, Show-o and Transfusion model texts in an autoregressive way, but generate image tokens simultaneously.

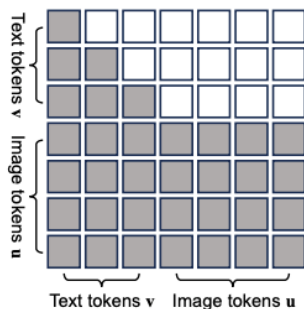
This reflects in the attention pattern for in the decoder transformer.

For Chameleon, all the attention are causally masked due to always predicting the next tokens.

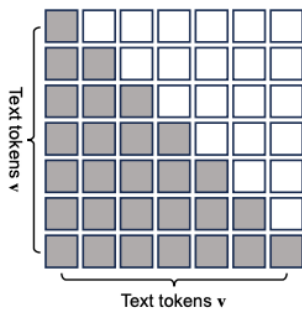
For Show-o and Transfusion, attention are causally masked when generating texts but bidirectional when generating images.



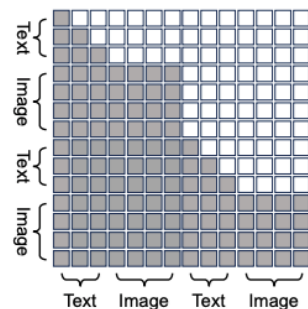
(a) Multimodal Understanding



(b) Text-to-Image Generation



(c) Language Modeling

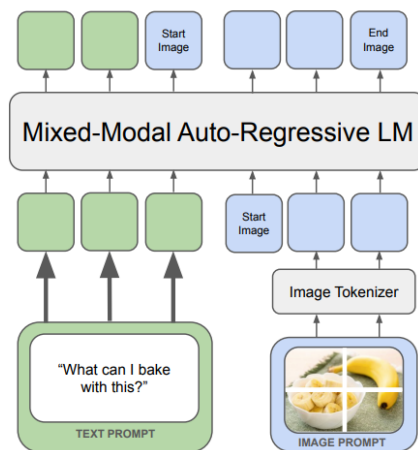


(d) Mixed-Modality Generation

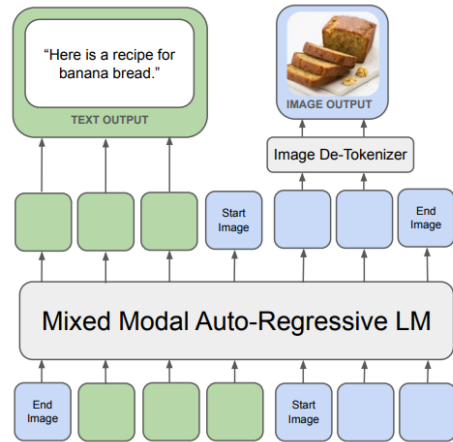


# Methodology 1

Train a VLM from scratch using no large encoders, opting instead for fusion as soon as possible. Text and Image prompts are tokenized and immediately fed into the model.



(a) Mixed-Modal Pre-Training



(b) Mixed-Modal Generation

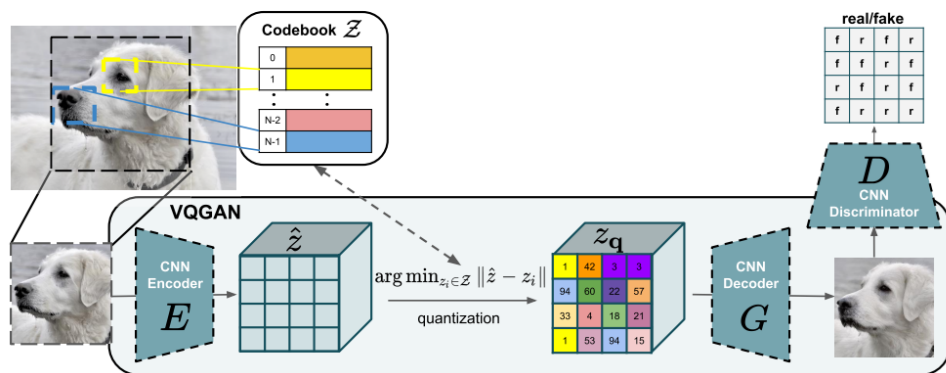
# Methodology 2 - Tokenization

## Text

- For text, BPE tokenizer is trained from scratch on a subset of the training data.

## Images

- Tokenizer largely based on VQGAN, but with added perceptual losses for faces and salient objects.
- 512x512 Images are encoded into 1024 tokens from a codebook of size 8192



Added perceptual losses help the tokenizer focus on important objects in an image

# Methodology 3 - Architecture

- Decoder-only model derived from **Llama 2**
- **RMSNorm** instead of LayerNorm for layer normalization
  - No clear argument for performance, but is more computationally efficient
- **SwiGLU** activation function everywhere
  - A modification of the regular GLU activation:

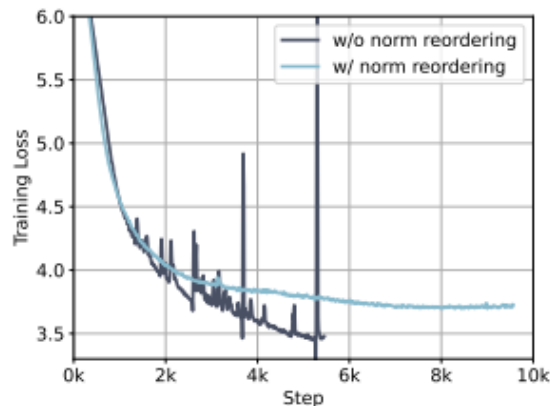
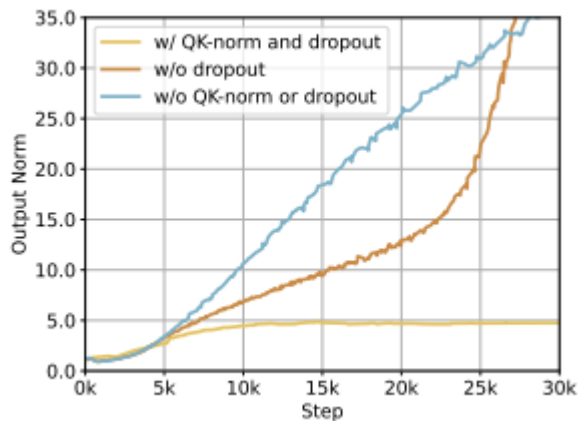
$$\text{GLU}(x, W, V, b, c) = \sigma(xW + b) \otimes (xV + c) \longrightarrow \text{SwiGLU}(x, W, V, b, c, \beta) = \text{Swish}_\beta(xW + b) \otimes (xV + c)$$

*“We offer no explanation as to why these architectures seem to work; we attribute their success, as all else, to divine benevolence” - From “GLU Variants Improve Transformer”*

- **Rotary Positional Embeddings** are used for adding positional information
  - Allows for better expressiveness of positional information and passing of relative position information to attention layers, along with other benefits

# Methodology 3.1 - Stability Issues

- The group observes training instability in multimodal settings and attributes this to “competition between modalities”, where modalities grow their respective norms whenever a softmax is present, i.e.: in every attention submodule, as well as the final softmax over the logits.
- They find this to be problematic for larger models and long training runs, and creates stability issues once norms reach the bf16 floating point limit.



## Methodology 3.2 - Solving Stability

- Query-Key Normalization solves the issue of norm growth in attention

$$\text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \longrightarrow \text{softmax}(g * \hat{Q}\hat{K}^T)V$$

- Moving LayerNorms outside of Attention and MLP helps limit the growth of feed-forward outputs.

**Chameleon-34B:**  $h = x + \text{attention\_norm}(\text{attention}(x))$   
output =  $h + \text{ffn\_norm}(\text{feed\_forward}(h))$

**Llama2:**  $h = x + \text{attention}(\text{attention\_norm}(x))$   
output =  $h + \text{feed\_forward}(\text{ffn\_norm}(h))$

- A small loss penalty to the denominator  $Z = \sum_i e^{x_i}$  of the last softmax activation prevents the logits from becoming too large.
- Dropout (Chameleon 7B only)

# Methodology 4 - Pre-Training

- Pretraining is done in two steps, both using a mixture of datasets categorized as *text-only*, *text-image*, and *text/image interleaved*.
- First, they train using very large-scale, completely unsupervised datasets for 80% of total training steps
- Then, they lower the weight of these large unsupervised datasets and mix in other, smaller, but high-quality datasets

Model	Params	Context Length	GQA	Tokens	LR	Epochs	Dropout	Zloss	Qknorm
LLaMa-1	7B	2k	×	1.0T	$3.0 \times 10^{-4}$	1.0	0.0	0.0	×
	33B	2k	×	1.4T	$1.5 \times 10^{-4}$	1.0	0.0	0.0	×
LLaMa-2	7B	4k	×	2.0T	$3.0 \times 10^{-4}$	1.0	0.0	0.0	×
	34B	4k	✓	2.0T	$1.5 \times 10^{-4}$	1.0	0.0	0.0	×
Chameleon	7B	4k	×	4.4T	$1.0 \times 10^{-4}$	2.1	0.1	$10^{-5}$	✓
	34B	4k	✓	4.4T	$1.0 \times 10^{-4}$	2.1	0.0	$10^{-5}$	✓

# Methodology 5 - Alignment

- The team chooses to perform a lightweight alignment stage by supervised fine-tuning using carefully curated datasets.
- They include a range of different tasks separated between *Text*, *Code*, *Visual Chat*, *Image Generation*, *Interleaved Text/Image Generation*, and *Safety*.
- *Text* and *Code* were borrowed from the Llama2 and CodeLlama SFT datasets.
- *Image Generation* is a carefully curated subset of their licensed data.
- *Interleaved Text/Image Generation* and *Visual Chat* are contracted from third-parties.

The datasets section is very opaque, and not a lot of detailed information is disclosed.

# Methodology 6 - Inference Challenges

- **Data dependencies per-step**
  - Image and text each have their own decoding methods, so tokens must be analyzed at each step.
- **Constrained generation masking**
  - In constrained generation (i.e.: unimodal outputs), additional control flow is required to ignore tokens not fitting a particular modality
- **Fixed-sized text units**
  - Since images are encoded and decoded as 1024-token blocks, inference has to account for this.



# Evaluation

## Human evaluation

*“Chameleon has significant new mixed modal understanding and generation abilities that cannot be measured with existing benchmarks.”*

Prompt generation

Absolute evaluation

Relative evaluation

Inter-annotator agreement

## Benchmarking

Text-only tasks

Image-To-Text tasks

# Human Generated Prompts for Evaluation

Informed problem-solving

**Advice: 10.2%**

What does a meningitis rash look like? What are the other symptoms I should be on the lookout for?

**How-to: 12.5%**

How do I properly clean my TV screen? I used Windex and now there are towel fibers and wipe marks all over. Show me some reference photos.

**Explanation: 14.4%**

I've been studying classical French art, and my favorite so far is his painting seen here: **<img>** Could you please give me a few images of other contemporary artworks that have this same aesthetic?



**Hypothetical: 5.6%**

What would the modern-day vehicle look like if oil had never been discovered?

Responses to imaginative questions

**Brainstorming: 18.6%**

Show me a Middle Eastern alternative to these dishes. **<img1>** **<img2>**



**Article: 3.1%**

Write me an introduction to a story about knick-knacks, and finish the story by shifting the focus with an image.

**Story: 3.9%**

Can you create and illustrate a short story for children about an octopus that can't stop eating pizza?

**Identification: 9.3 %**



Is the below image a Shetland Pony? If not, what is it, and can you show me a Shetland Pony? **<img>**

**Comparison: 9.6%**

Please tell me what the difference between these two creatures is, and show me some more examples. **<img1>** **<img2>**



**Report: 5.4%**

Generating a summary of real events  
Who designed the church in the image below, and what's the name of the Church? **<img>** Can you please provide me with additional photos of famous landmarks designed by the same architect?



**Other: 5.2%**

Create a decal for my truck that features running horses as well as the TRD insignia. Use black to gray gradients.

**Reasoning: 2.1%**

What is typically found at a construction site? Show me a construction site that has a crane.

Generate diverse and natural prompts from human annotators.

Evaluate prompts and filter out unclear prompts and prompts that don't expect mixed-modal responses.

>> 1048 prompts

Manually classify prompts into 12 categories.

Note: While not instructed specifically, certain image understanding tasks that require identifying the text in an image, such as OCR (Optical character recognition) or Infographics (interpreting a given chart or plot), do not appear in our evaluation set of prompts.

# Baseline Models in Human Evaluation

Baseline models: OpenAI GPT-4V & Google Gemini Pro

Responses are text-only despite of mixed-modal prompts

Augmenting GPT-4V and Gemini's responses with images:

>> instruct models to generate image captions by prompting:

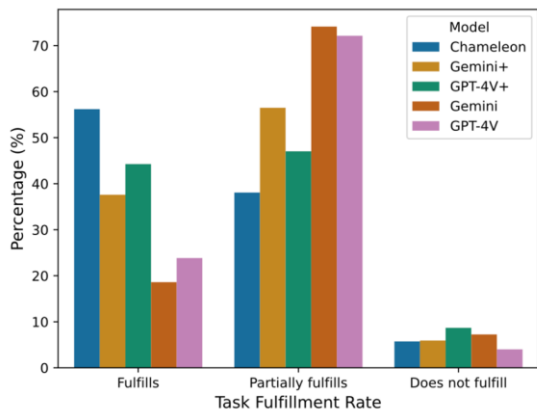
“If the question requires an image to be generated, then generate an image caption instead and enclose the caption in a pair of <caption></caption> tags”

>> use OpenAI DALL-E 3 to generate images conditioned on generated captions

>> replace the captions in the original responses with generated images

# Absolute and Relative Evaluation

**Absolute:** Judged by three different annotators whether the responses fulfills/partially fulfills/does not fulfill the task described in the prompt



(a) The prompt task fulfillment rates.

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%	<u>1.9%</u>	43.9%	48.6%	7.5%
Article	59.4%	37.5%	3.1%	40.6%	53.1%	<u>6.3%</u>	<u>62.5%</u>	37.5%	0.0%
Brainstorming	57.9%	36.4%	5.6%	33.3%	61.5%	<u>5.1%</u>	<u>47.7%</u>	47.2%	5.1%
Comparison	60.4%	34.7%	5.0%	47.5%	46.5%	<u>5.9%</u>	43.6%	44.6%	11.9%
Explanation	53.0%	37.7%	9.3%	33.8%	61.6%	<u>4.6%</u>	41.7%	50.3%	7.9%
How-to	52.7%	40.5%	6.9%	43.5%	52.7%	<u>3.8%</u>	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%	<u>1.0%</u>	35.1%	55.7%	9.3%
Other	41.8%	40.0%	18.2%	38.2%	41.8%	<u>20.0%</u>	<u>50.9%</u>	40.0%	9.1%
Reasoning	50.0%	13.6%	36.4%	27.3%	59.1%	<u>13.6%</u>	<u>31.8%</u>	54.5%	13.6%
Report	49.1%	40.4%	10.5%	29.8%	61.4%	<u>8.8%</u>	38.6%	47.4%	14.0%
Story	31.7%	63.4%	4.9%	39.0%	56.1%	<u>4.9%</u>	<u>53.7%</u>	43.9%	2.4%

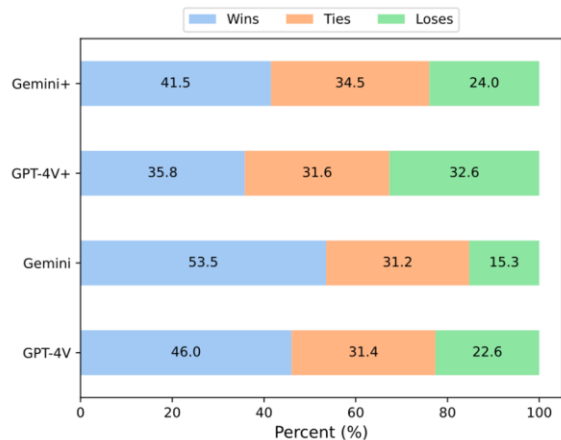
Note: Gemini+ has a better performance in generating “fulfills+partially fulfills” responses.

# Absolute and Relative Evaluation

**Relative:** presenting responses in anonymous order and asking human annotators which responses they prefer.

Gemini+

GPT-4V+



	Wins	Ties	Losses	Win rate
Overall	435	362	251	58.8%
Advice	48	35	24	61.2%
Article	14	14	4	65.6%
Brainstorming	101	60	34	67.2%
Comparison	41	38	22	59.4%
Explanation	65	46	40	58.3%
How-to	53	51	27	59.9%
Hypothetical	17	24	18	49.2%
Identification	39	33	25	57.2%
Other	24	17	14	59.1%
Reasoning	7	8	7	50.0%
Report	16	22	19	47.4%
Story	10	14	17	41.5%
Mixed-modal Prompts	194	145	102	60.4%
Text-only Prompts	241	217	149	57.6%

	Wins	Ties	Losses	Win rate
Overall	375	331	342	51.6%
Advice	54	27	26	63.1%
Article	9	11	12	45.3%
Brainstorming	78	57	60	54.6%
Comparison	35	35	31	52.0%
Explanation	53	56	42	53.6%
How-to	49	46	36	55.0%
Hypothetical	23	19	17	55.1%
Identification	31	26	40	45.4%
Other	16	13	26	40.9%
Reasoning	11	5	6	61.4%
Report	16	21	20	46.5%
Story	0	15	26	18.3%
Mixed-modal Prompts	149	119	173	47.3%
Text-only Prompts	226	212	169	54.7%

Note: Didn't outperform GPT-4V+ & Gemini+ by a significant margin.

Raise the question: do we really need interleaved multimodal foundation model instead of using two foundation models from each modality?

# Human Evaluation by Task Categories

The task categories that Chameleon performs well: Brainstorming, Comparison and Hypothetical

Need to improve: Identification and Reasoning

Don't see significant difference when comparing mixed-modality and text-only prompt

**Prompt:** I'm bored. Could you show me some cool, quirky-looking birds? Also tell me about them in 2-3 sentences.



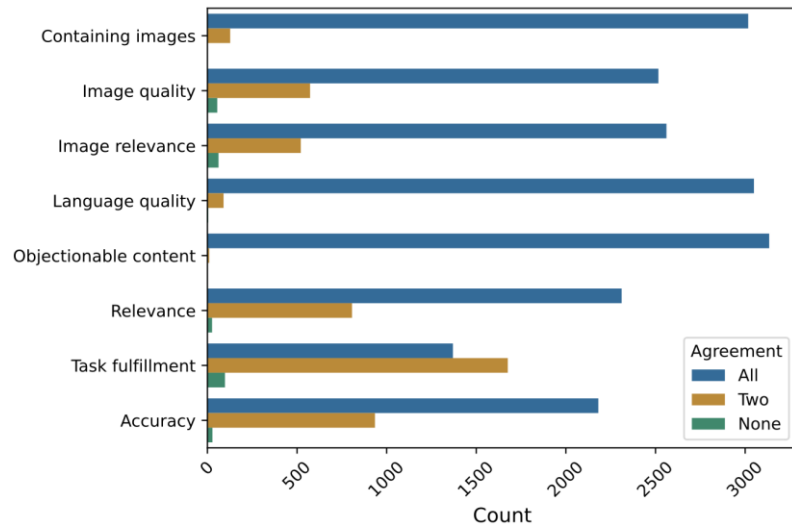
**Response:** Absolutely! Here are three birds with some unique features that you might find interesting:

**Keel-Billed Toucan** The Keel-Billed Toucan is a brightly colored bird native to South American rainforests and woodlands. Its most distinctive feature is its colorful beak, which can grow up to seven inches long and has a distinctive, curved shape. The beak is used for plucking fruit from trees and for social displays. Despite its large size, the Keel-Billed Toucan is a skilled flyer and can often be seen flitting through the treetops. Here is an image of this vibrant bird: `<img>`

# Inter-annotator agreement

## Majority votes from three different annotators

Absolute Evaluation



Task fulfillment: when annotator's judgment differs from the other two's, the decision is usually still close (e.g., fulfills vs. partially fulfills) rather than opposite (e.g., fulfills vs. does not fulfill)

Relative Evaluation

	All 3 annotators agree	2 of 3 annotators agree	No Agreement
Chameleon vs. Gemini+	331 (31.5%)	609 (58.1%)	108 (10.3%)
Chameleon vs. GPT-4V+	371 (35.4%)	579 (55.2%)	98 (9.3%)
Chameleon vs. Gemini	317 (30.2%)	621 (59.3%)	110 (10.5%)
Chameleon vs. GPT-4V	300 (28.6%)	611 (58.3%)	137 (13.1%)

# Benchmark Evaluations: Text-only Tasks

Mistral (Oct 2023): aim to improve computation efficiency; outperforms LLaMa-2 13B across all tested benchmark, and LLaMa-1 34B in mathematics and code generation.

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



# Benchmark Datasets: Text-only Tasks

## PIQA [2020]: benchmarking progress in physical commonsense understanding

> over 16k multiple choice QA pairs

a. Shape, Material, and Purpose	
[Goal]	Make an outdoor pillow
[Sol1]	Blow into a <b>tin can</b> and tie with rubber band <span style="color: red;">✗</span>
[Sol2]	Blow into a <b>trash bag</b> and tie with rubber band <span style="color: green;">✓</span>
[Goal]	To make a hard shelled taco,
[Sol1]	put seasoned beef, cheese, and lettuce <b>onto</b> the hard shell. <span style="color: red;">✗</span>
[Sol2]	put seasoned beef, cheese, and lettuce <b>into</b> the hard shell. <span style="color: green;">✓</span>
[Goal]	How do I find something I lost on the carpet?
[Sol1]	Put a <b>solid seal</b> on the end of your vacuum and turn it on. <span style="color: red;">✗</span>
[Sol2]	Put a <b>hair net</b> on the end of your vacuum and turn it on. <span style="color: green;">✓</span>

b. Commonsense Convenience	
[Goal]	How to make sure all the clocks in the house are set accurately?
[Sol1]	Get a solar clock for a reference and place it just outside a window that gets lots of sun. Use a system of call and response once a month, having one person stationed at the solar clock who yells out the correct time and have another person move to each of the indoor clocks to check if they are showing the right time. Adjust as necessary. <span style="color: red;">✗</span>
[Sol2]	Replace all wind-ups with digital clocks. That way, you set them once, and that's it. Check the batteries once a year or if you notice anything looks a little off. <span style="color: green;">✓</span>

Figure 2: **PIQA** covers a broad array of phenomena. Above are two categories of example QA pairs. **Left** are examples that require knowledge of basic properties of the objects (flexibility, curvature, and being porous), while on the **Right** both answers may be technically correct but one is more convenient and preferable.

## SIQA [2019]: measure the social and emotional intelligence of computational models

> over 38k multiple choice QA pairs

Context	Question	Answer
(1) Jesse was pet sitting for Addison, so Jesse came to Addison's house and walked their dog.	What does Jesse need to do before this?	(a) feed the dog ✓ ⊖ (b) get a key from Addison (c) walk the dog
(2) Kai handed back the computer to Will after using it to buy a product off Amazon.	What will Kai want to do next?	(a) wanted to save money on shipping ✓ ⊖ (b) Wait for the package (c) Wait for the computer

<b>wants</b> (e.g., What will Kai want to do next?) <b>29%</b>	<b>reactions</b> (e.g., How would Robin feel afterwards?) <b>21%</b>	<b>descriptions</b> (e.g., How would you describe Alex?) <b>15%</b>	<b>motivations</b> (e.g., Why did Sydney do this?) <b>12%</b>	<b>needs</b> (e.g., What does Remy need to do before this?) <b>12%</b>	<b>effects</b> (e.g., What will happen to Sasha?) <b>11%</b>
--	--	---	---	--	--

# Benchmark Datasets: Text-only Tasks

## Category: Shaving (ActivityNet; In-domain)

A bearded man is seen speaking to the camera and making several faces. the man

- a) then switches off and shows himself via the washer and dryer rolling down a towel and scrubbing the floor. (0.0%)
- b) then rubs and wipes down an individual's face and leads into another man playing another person's flute. (0.0%)
- c) is then seen eating food on a ladder while still speaking. (0.0%)
- d) then holds up a razor and begins shaving his face. (100.0%)**

## Category: Sharpening knives (ActivityNet; Zero-Shot)

Two men are in a room and the man with a blue shirt takes out a bench stone and with a little lubricant on the stone takes an knife and explains how to sharpen it. then he

- a) uses a sharpener to smooth out the stone using the knife. (100.0%)
- b) shows how to cut the bottom with the knife and place a tube on the inner and corner. (0.0%)
- c) bends down and grabs the knife and remove the appliance. (0.0%)
- d) stops sharpening the knife and takes out some pieces of paper to show how sharp the knife is as he cuts slivers of paper with the knife. (0.0%)**

Common sense natural language inference datasets that are easy for human but particularly challenging for machines.

**Hellaswag [2019] & WinoGrande [2021]**

Twin sentences	Options (answer)
The monkey <b>loved</b> to play with the balls but <b>ignored</b> the blocks because he found <b>them exciting</b> .	<b>balls</b> / blocks
The monkey <b>loved</b> to play with the balls but <b>ignored</b> the blocks because he found <b>them dull</b> .	balls / <b>blocks</b>
William could only climb beginner walls while Jason <b>climbed advanced</b> ones because <b>he</b> was very <b>weak</b> .	<b>William</b> / Jason
William could only climb beginner walls while Jason <b>climbed advanced</b> ones because <b>he</b> was very <b>strong</b> .	William / <b>Jason</b>
Robert woke up at 9:00am while Samuel woke up at 6:00am, so <b>he</b> had <b>less</b> time to get ready for school.	<b>Robert</b> / Samuel
Robert woke up at 9:00am while Samuel woke up at 6:00am, so <b>he</b> had <b>more</b> time to get ready for school.	Robert / <b>Samuel</b>
The child was screaming after the baby bottle and toy fell. Since the child was <b>hungry</b> , <b>it</b> stopped his crying.	<b>baby bottle</b> / toy
The child was screaming after the baby bottle and toy fell. Since the child was <b>full</b> , <b>it</b> stopped his crying.	baby bottle / <b>toy</b>

# Benchmark Datasets: Text-only Tasks

**ARC-Easy & ARC-Challenge [2018]:** 7787 science questions, all non-diagram, multiple choice.

**OpenBookQA [2018]:** 6000 questions centered around 1326 elementary level science facts.

Knowledge Type	Example
Definition	What is a worldwide increase in temperature called? (A) greenhouse effect (B) global warming (C) ozone depletion (D) solar heating
Basic Facts & Properties	Which element makes up most of the air we breathe? (A) carbon (B) nitrogen (C) oxygen (D) argon
Structure	The crust, the mantle, and the core are structures of Earth. Which description is a feature of Earth's mantle? (A) contains fossil remains (B) consists of tectonic plates (C) is located at the center of Earth (D) has properties of both liquids and solids
Processes & Causal	What is the first step of the process in the formation of sedimentary rocks? (A) erosion (B) deposition (C) compaction (D) cementation
Teleology / Purpose	What is the main function of the circulatory system? (1) secrete enzymes (2) digest proteins (3) produce hormones (4) transport materials
Algebraic	If a red flowered plant (RR) is crossed with a white flowered plant (rr), what color will the offspring be? (A) 100% pink (B) 100% red (C) 50% white, 50% red (D) 100% white
Experiments	Scientists perform experiments to test hypotheses. How do scientists try to remain objective during experiments? (A) Scientists analyze all results. (B) Scientists use safety precautions. (C) Scientists conduct experiments once. (D) Scientists change at least two variables.
Spatial / Kinematic	In studying layers of rock sediment, a geologist found an area where older rock was layered on top of younger rock. Which best explains how this occurred? (A) Earthquake activity folded the rock layers...

## Question:

Which of these would let the most heat travel through?

- A) a new pair of jeans.
- B) a steel spoon in a cafeteria.
- C) a cotton candy at a store.
- D) a calvin klein cotton hat.

## Science Fact:

Metal is a thermal conductor.

## Common Knowledge:

Steel is made of metal.

Heat travels through a thermal conductor.

Figure 1: An example for a question with a given set of choices and supporting facts.

# Benchmark Datasets: Text-only Tasks

**BoolQ** [2019]: yes/no questions that query for complex, non-factoid information and inference.

> over 16k questions

> paired with a paragraph from Wikipedia that contain the answer.

- 
- Q:** Has the UK been hit by a hurricane?  
**P:** The Great Storm of 1987 was a violent extratropical cyclone which caused casualties in England, France and the Channel Islands ...  
**A:** Yes. [An example event is given.]
- Q:** Does France have a Prime Minister and a President?  
**P:** ... The extent to which those decisions lie with the Prime Minister or President depends upon ...  
**A:** Yes. [Both are mentioned, so it can be inferred both exist.]
- Q:** Have the San Jose Sharks won a Stanley Cup?  
**P:** ... The Sharks have advanced to the Stanley Cup finals once, losing to the Pittsburgh Penguins in 2016 ...  
**A:** No. [They were in the finals once, and lost.]
- 

Figure 1: Example yes/no questions from the BoolQ dataset. Each example consists of a question (**Q**), an excerpt from a passage (**P**), and an answer (**A**) with an explanation added for clarity.

# Benchmark Datasets: Text-only Tasks

**GSM8k [2021]:** 8.5k high quality linguistically diverse grade school math word problems

solutions primarily involve performing a sequence of elementary calculations using basic arithmetic operations (+ - x ÷) to reach the final answer.

**MATH [2021]:** 12.5k problems from high school math competitions.

## MATH Dataset (Ours)

**Problem:** Tom has a red marble, a green marble, a blue marble, and three identical yellow marbles. How many different groups of two marbles can Tom choose?

**Solution:** There are two cases here: either Tom chooses two yellow marbles (1 result), or he chooses two marbles of different colors ( $\binom{4}{2} = 6$  results). The total number of distinct pairs of marbles Tom can choose is  $1 + 6 = \boxed{7}$ .

**Problem:** The equation  $x^2 + 2x = i$  has two complex solutions. Determine the product of their real parts.

**Solution:** Complete the square by adding 1 to each side. Then  $(x + 1)^2 = 1 + i = e^{\frac{i\pi}{4}}\sqrt{2}$ , so  $x + 1 = \pm e^{\frac{i\pi}{8}}\sqrt[4]{2}$ . The desired product is then  $(-1 + \cos(\frac{\pi}{8})\sqrt[4]{2})(-1 - \cos(\frac{\pi}{8})\sqrt[4]{2}) = 1 - \cos^2(\frac{\pi}{8})\sqrt{2} = 1 - \frac{(1 + \cos(\frac{\pi}{4}))}{2}\sqrt{2} = \boxed{\frac{1 - \sqrt{2}}{2}}$ .

**Problem:** Beth bakes 4, 2 dozen batches of cookies in a week. If these cookies are shared amongst 16 people equally, how many cookies does each person consume?

**Solution:** Beth bakes 4 2 dozen batches of cookies for a total of  $4 \cdot 2 = \ll 4 \cdot 2 = 8 \gg 8$  dozen cookies

There are 12 cookies in a dozen and she makes 8 dozen cookies for a total of  $12 \cdot 8 = \ll 12 \cdot 8 = 96 \gg 96$  cookies

She splits the 96 cookies equally amongst 16 people so they each eat  $96/16 = \ll 96/16 = 6 \gg 6$  cookies

**Final Answer:** 6

# Benchmark Datasets: Text-only Tasks

## MMLU [2020]: multiple-choice questions from 57 domains

<b>Microeconomics</b>	One of the reasons that the government discourages and regulates monopolies is that	
	(A) producer surplus is lost and consumer surplus is gained.	✗
	(B) monopoly prices ensure productive efficiency but cost society allocative efficiency.	✗
	(C) monopoly firms do not engage in significant research and development.	✗
	(D) consumer surplus is lost with higher prices and lower levels of output.	✓

Figure 3: Examples from the Microeconomics task.

<b>Conceptual Physics</b>	When you drop a ball from rest it accelerates downward at $9.8 \text{ m/s}^2$ . If you instead throw it downward assuming no air resistance its acceleration immediately after leaving your hand is	
	(A) $9.8 \text{ m/s}^2$	✓
	(B) more than $9.8 \text{ m/s}^2$	✗
	(C) less than $9.8 \text{ m/s}^2$	✗
	(D) Cannot say unless the speed of throw is given.	✗
<b>College Mathematics</b>	In the complex $z$ -plane, the set of points satisfying the equation $z^2 =  z ^2$ is a	
	(A) pair of points	✗
	(B) circle	✗
	(C) half-line	✗
	(D) line	✓

Figure 4: Examples from the Conceptual Physics and College Mathematics STEM tasks.

# Benchmark Evaluations: Text-only Tasks

Common-sense reasoning and reading comprehension capabilities (zero-shot)

- Competitive with Llama-2
- Outperform Llama-2 70B on  $\frac{5}{8}$  tasks and on par with Mixtral 8x7B

Math and world knowledge (maj@1 unless mentioned)

- maj@N: majority vote after N generations

Overall, Chameleon outperforms LLaMa-2 with performance approaching Mistral 7B/Mixtral 8x7B on some tasks.

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



# Benchmark Evaluations: Image-To-Text Tasks

Text generation conditioned on an image:

- Image captioning
- Visual question-answering tasks

	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 <i>an open-access reproduction of Flamingo</i>	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 <i>Fine-tuned on specific task</i>	34B	140.8 0-shot	82.3 2-shot	—
	Chameleon-MultiTask <i>Fine-tuned on all tasks</i>	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



# Strengths

- Demonstrate the possibility of scaling early-fusion VLM for generating interleaved texts and images, not just capable of vision language reasoning or understanding tasks
- Early fusion model has a simple design space without worrying about modality specific encoders and allows seamless integration of information across modalities
- Introduce techniques for stable training of super large models (>7B).
- Perform human evaluation on generated results rather than just static scores.
- Through human evaluation, Chameleon potentially justify that a general purpose mixed modality model could outperform “two-stage” approaches

# Weaknesses

- Comparing with non-native mixed-modality models in experiments does not benchmark the performance of Chameleon in a fully convincing way.
- Compared to other papers, such as Flamingo, InternVL, etc. the experiments in Chameleon contains fewer tasks and datasets (most of the datasets in this paper are published by research groups in UW).
- Code data is used in training but not benchmarked during evaluation. Similarly, image only evaluation tasks such as object classification are not included.
- The paper writing does not contain much justification or reasoning over the design choice