

CS11-711 Advanced NLP

Evaluation and Multimodal

Xiang Yue



Carnegie Mellon University

Language Technologies Institute

<https://phontron.com/class/anlp-fall2024>

Slides are partially adapted from 11-777 [MultiModal Machine Learning](#) (by Daniel Fried, Yonatan Bisk, Louis-Philippe Morency, Paul Liang)

What will cover in this class

We merge two classes:

- Evaluation of NLP tasks and LMs
- Multimodal

Why we need evals?

- **Performance Measurement:** They offer key metrics to assess how well models perform on specific tasks, helping to gauge their effectiveness.
- **Standardized Comparison:** Benchmarks provide a consistent basis for comparing different models, enabling fair evaluations across the field.
- **Guiding Model Development:** Enables researchers and model developers to identify specific areas where models may need improvement, focusing development effort.

Development of AI Evaluations

1. Classical Era (1990s - 2000s): Small-scale Task-specific Benchmarks

- MNIST (1998): Handwritten digit recognition for image classification.
- Penn Treebank (1993): Early benchmark for syntactic parsing in NLP.
- Switchboard (1992): Speech recognition evaluation dataset.

2. Deep Learning Era (2010s): Large-scale Task-specific Benchmarks

- ImageNet (2010): Revolutionized image recognition with deep CNNs.
- COCO (2014): Complex object detection, segmentation, and captioning.
- SQuAD (2016) and SQuAD 2.0 (2018): Machine reading comprehension and QA
- GLUE (2018) and SuperGLUE (2019): Testing models across multiple NLP tasks.

3. Foundation Model Era (2020s): Large-scale, General, Holistic, and Multimodal Benchmarks

- MMLU (2020): Testing reasoning and factual knowledge across diverse fields.
- Big-bench (2021): Broad benchmark for reasoning, world knowledge, and creativity.
- MMMU (2023): Complex reasoning multimodal questions in multiple subjects

Question Answering: SQuAD

SQuAD2.0

The Stanford Question Answering Dataset

The Norman dynasty had a major political, cultural and military impact on medieval Europe and even the Near East. The Normans were famed for their martial spirit and eventually for their Christian piety, becoming exponents of the Catholic orthodoxy into which they assimilated. They adopted the Gallo-Romance language of the Frankish land they settled, their dialect becoming known as Norman, Normand or Norman French, an important literary language. The Duchy of Normandy, which they formed by treaty with the French crown, was a great fief of medieval France, and under Richard I of Normandy was forged into a cohesive and formidable principality in feudal tenure. The Normans are noted both for their culture, such as their unique Romanesque architecture and musical traditions, and for their significant military accomplishments and innovations. Norman adventurers founded the Kingdom of Sicily under Roger II after conquering southern Italy on the Saracens and Byzantines, and an expedition on behalf of their duke, William the Conqueror, led to the Norman conquest of England at the Battle of Hastings in 1066. Norman cultural and military influence spread from these new European centres to the Crusader states of the Near East, where their prince Bohemond I founded the Principality of Antioch in the Levant, to Scotland and Wales in Great Britain, to Ireland, and to the coasts of north Africa and the Canary Islands.

Who was the duke in the battle of Hastings?

Ground Truth Answers: William the Conqueror William the Conqueror William the Conqueror

Prediction: William the Conqueror

Who ruled the duchy of Normandy

Ground Truth Answers: Richard I Richard I Richard I

Prediction: Richard I of Normandy

What religion were the Normans

Ground Truth Answers: Catholic Catholic orthodoxy Catholic

Prediction: The Normans were famed for their martial spirit and eventually for their Christian piety

What type of major impact did the Norman dynasty have on modern Europe?

Ground Truth Answers: <No Answer>

Prediction: <No Answer>

Who was famed for their Christian spirit?

Ground Truth Answers: No Answer

Question Answering: TriviaQA

Question: The Dodecanese Campaign of WWII that was an attempt by the Allied forces to capture islands in the Aegean Sea was the inspiration for which acclaimed 1961 commando film?

Answer: The Guns of Navarone

Excerpt: The Dodecanese Campaign of World War II was an attempt by Allied forces to capture the Italian-held Dodecanese islands in the Aegean Sea following the surrender of Italy in September 1943, and use them as bases against the German-controlled Balkans. The failed campaign, and in particular the Battle of Leros, inspired the 1957 novel **The Guns of Navarone** and the successful 1961 movie of the same name.

Question: American Callan Pinckney's eponymously named system became a best-selling (1980s-2000s) book/video franchise in what genre?

Answer: Fitness

Excerpt: Callan Pinckney was an American fitness professional. She achieved unprecedented success with her Callanetics exercises. Her 9 books all became international best-sellers and the video series that followed went on to sell over 6 million copies. Pinckney's first video release "Callanetics: 10 Years Younger In 10 Hours" outsold every other **fitness** video in the US.

95K question answer pairs authored by trivia enthusiasts and independently gathered evidence documents

- It has relatively complex, compositional questions
- It has considerable syntactic and lexical variability between questions and corresponding answer-evidence sentences
- It requires more cross sentence reasoning to find answers.

Metrics: Question Answering

- **Exact Match (EM):** Measures the percentage of predictions that exactly match any one of the ground truth answers. An EM score of 100% would mean the model produced exact matches for all questions.
- **F1 Score:** Calculates the overlap between the prediction and the ground truth answers at the word level. It considers both precision and recall, making it useful for partial matches. The F1 score is especially helpful for evaluating answers with close but not exact matches.

Machine Translation

WMT (Workshop on Machine Translation)

German→ English	Source: “Most informative is the analysis of airway secretions:” Reference: “Häufig jedoch führt die Analyse von Material aus den Atemwegen zur Diagnose:” Proper: “analysis of airway secretions” → “Analyse von Material aus den Atemwegen” Random: “Most” → “Häufig”
English→ Czech	Source: “We present Eman, an experiment manager, and show how to use it to train several simple MT systems.” Reference: “Popisujeme Emanu, nástroj pro správu experimentů, a ukazujeme, jak ho lze využít k trénování několika jednoduchých systémů pro strojový překlad.” Proper: “Eman” → “Emana”, “an experiment manager” → “nástroj pro správu experimentů”, “MT systems” → “systémů pro strojový překlad” Random: “how to use” → “jak ho lze využít”, “train” → “trénování”, “simple” → “jednoduchých”
Chinese→ English	Source: “凌寒再次挥手，又结结实实地抽了他一巴掌。” Reference: Ling Han raised his hand once more, and gave him another solid slap. Proper: ‘凌寒’ → “Ling Han” Random: ‘手’ → “his hand”

WMT 2023

<https://aclanthology.org/2023.wmt-1.54.pdf>

Summarization

SUMMARY: *A man and a child have been killed after a light aircraft made an emergency landing on a beach in Portugal.*

DOCUMENT: Authorities said the incident took place on Sao Joao beach in Caparica, south-west of Lisbon.

The National Maritime Authority said a middle-aged man and a young girl died after they were unable to avoid the plane.

[6 sentences with 139 words are abbreviated from here.]

Other reports said the victims had been sunbathing when the plane made its emergency landing.

[Another 4 sentences with 67 words are abbreviated from here.]

Video footage from the scene carried by local broadcasters showed a small recreational plane parked on the sand, apparently intact and surrounded by beachgoers and emergency workers.

[Last 2 sentences with 19 words are abbreviated.]

Xsum Dataset

<https://aclanthology.org/D18-1206.pdf>

Original Text (truncated): lagos, nigeria (cnn) a day after winning nigeria's presidency, *muhammadu buhari* told cnn's christiane amanpour that he plans to aggressively fight corruption that has long plagued nigeria and go after the root of the nation's unrest. *buhari* said he'll "rapidly give attention" to curbing violence in the northeast part of nigeria, where the terrorist group boko haram operates. by cooperating with neighboring nations chad, cameroon and niger, he said his administration is confident it will be able to thwart criminals and others contributing to nigeria's instability. for the first time in nigeria's history, the opposition defeated the ruling party in democratic elections. *buhari* defeated incumbent goodluck jonathan by about 2 million votes, according to nigeria's independent national electoral commission. the win comes after a long history of military rule, coups and botched attempts at democracy in africa's most populous nation.

Baseline Seq2Seq + Attention: UNK UNK says his administration is confident it will be able to destabilize nigeria's economy. UNK says his administration is confident it will be able to thwart criminals and other nigerians. he says the country has long nigeria and nigeria's economy.

Pointer-Gen: *muhammadu buhari* says he plans to aggressively fight corruption in the northeast part of nigeria. he says he'll "rapidly give attention" to curbing violence in the northeast part of nigeria. he says his administration is confident it will be able to thwart criminals.

Pointer-Gen + Coverage: *muhammadu buhari* says he plans to aggressively fight corruption that has long plagued nigeria. he says his administration is confident it will be able to thwart criminals. the win comes after a long history of military rule, coups and botched attempts at democracy in africa's most populous nation.

CNN / Daily Mail Dataset

<https://arxiv.org/pdf/1704.04368>

Metrics:

Machine Translation and Summarization

- **BLEU:** Measures the overlap of n-grams between the machine-translated output and one or more reference translations.
- **ROUGE:** measures the overlap of n-grams, favoring recall. ROUGE-L specifically measures longest common subsequence.
- **METEOR:** Accounts for stemming and synonyms, offering more linguistic flexibility than BLEU.
- **ChrF (Character F-score):** Based on character-level n-grams instead of word-level, making it more suitable for languages with rich morphology. ChrF is gaining popularity due to its adaptability to different languages.
- **COMET** and **BERTScore:** use embeddings to capture semantic similarity between machine translations and reference translations

General Language Understanding Evaluation (GLUE)

Corpus	Train	Test	Task	Metrics	Domain
Single-Sentence Tasks					
CoLA	8.5k	1k	acceptability	Matthews corr.	misc.
SST-2	67k	1.8k	sentiment	acc.	movie reviews
Similarity and Paraphrase Tasks					
MRPC	3.7k	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	391k	paraphrase	acc./F1	social QA questions
Inference Tasks					
MNLI	393k	20k	NLI	matched acc./mismatched acc.	misc.
QNLI	105k	5.4k	QA/NLI	acc.	Wikipedia
RTE	2.5k	3k	NLI	acc.	news, Wikipedia
WNLI	634	146	coreference/NLI	acc.	fiction books

Table 1: Task descriptions and statistics. All tasks are single sentence or sentence pair classification, except STS-B, which is a regression task. MNLI has three classes; all other classification tasks have two. Test sets shown in bold use labels that have never been made public in any form.

<https://arxiv.org/pdf/1804.07461>

SuperGLUE

Table 1: The tasks included in SuperGLUE. *WSD* stands for word sense disambiguation, *NLI* is natural language inference, *coref.* is coreference resolution, and *QA* is question answering. For MultiRC, we list the number of total answers for 456/83/166 train/dev/test questions.

Corpus	 Train 	 Dev 	 Test 	Task	Metrics	Text Sources
BoolQ	9427	3270	3245	QA	acc.	Google queries, Wikipedia
CB	250	57	250	NLI	acc./F1	various
COPA	400	100	500	QA	acc.	blogs, photography encyclopedia
MultiRC	5100	953	1800	QA	F1 _a /EM	various
ReCoRD	101k	10k	10k	QA	F1/EM	news (CNN, Daily Mail)
RTE	2500	278	300	NLI	acc.	news, Wikipedia
WiC	6000	638	1400	WSD	acc.	WordNet, VerbNet, Wiktionary
WSC	554	104	146	coref.	acc.	fiction books

Massive Multitask Language Understanding (MMLU)

Few Shot Prompt and Predicted Answer

The following are multiple choice questions about high school mathematics.

How many numbers are in the list 25, 26, ..., 100?
(A) 75 (B) 76 (C) 22 (D) 23

Answer: B

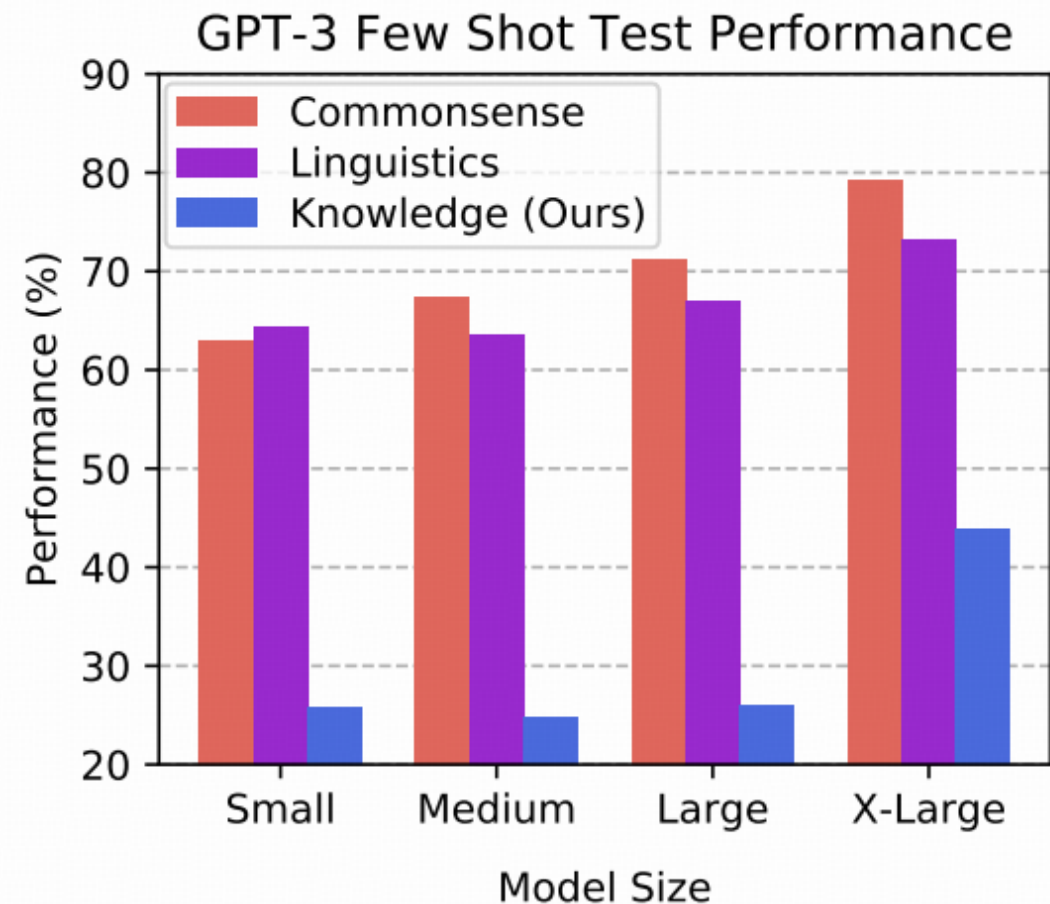
Compute $i + i^2 + i^3 + \dots + i^{258} + i^{259}$.
(A) -1 (B) 1 (C) i (D) $-i$

Answer: A

If 4 daps = 7 yaps, and 5 yaps = 3 baps, how many daps equal 42 baps?
(A) 28 (B) 21 (C) 40 (D) 30

Answer: C

(a) An example of few-shot learning and inference using GPT-3. The **blue** underlined bold text is the auto-completed response from GPT-3, while the preceding text is the user-inputted prompt. In this 2-shot learning example, there are two instruction examples and one initially incomplete example. On average, GPT-3 has low accuracy on high school mathematics questions.



(b) Performance on a commonsense benchmark (HellaSwag), a linguistic understanding benchmark (SuperGLUE), and the massive multitask test. On previous benchmarks, smaller models start well above random chance levels and exhibit more continuous improvements with model size increases, but on our test, GPT-3 moves beyond random chance with the largest model.

Massive Multitask Language Understanding (MMLU)

Task	Tested Concepts	Supercategory			
Abstract Algebra	Groups, rings, fields, vector spaces, ...	STEM	High School Statistics	Random variables, sampling distributions, chi-square tests, ...	STEM
Anatomy	Central nervous system, circulatory system, ...	STEM	High School US History	Civil War, the Great Depression, The Great Society, ...	Humanities
Astronomy	Solar system, galaxies, asteroids, ...	STEM	High School World History	Ottoman empire, economic imperialism, World War I, ...	Humanities
Business Ethics	Corporate responsibility, stakeholders, regulation, ...	Other	Human Aging	Senescence, dementia, longevity, personality changes, ...	Other
Clinical Knowledge	Spot diagnosis, joints, abdominal examination, ...	Other	Human Sexuality	Pregnancy, sexual differentiation, sexual orientation, ...	Social Sciences
College Biology	Cellular structure, molecular biology, ecology, ...	STEM	International Law	Human rights, sovereignty, law of the sea, use of force, ...	Humanities
College Chemistry	Analytical, organic, inorganic, physical, ...	STEM	Jurisprudence	Natural law, classical legal positivism, legal realism, ...	Humanities
College Computer Science	Algorithms, systems, graphs, recursion, ...	STEM	Logical Fallacies	No true Scotsman, base rate fallacy, composition fallacy, ...	Humanities
College Mathematics	Differential equations, real analysis, combinatorics, ...	STEM	Machine Learning	SVMs, VC dimension, deep learning architectures, ...	STEM
College Medicine	Introductory biochemistry, sociology, reasoning, ...	Other	Management	Organizing, communication, organizational structure, ...	Other
College Physics	Electromagnetism, thermodynamics, special relativity, ...	STEM	Marketing	Segmentation, pricing, market research, ...	Other
Computer Security	Cryptography, malware, side channels, fuzzing, ...	STEM	Medical Genetics	Genes and cancer, common chromosome disorders, ...	Other
Conceptual Physics	Newton's laws, rotational motion, gravity, sound, ...	STEM	Miscellaneous	Agriculture, Fermi estimation, pop culture, ...	Other
Econometrics	Volatility, long-run relationships, forecasting, ...	Social Sciences	Moral Disputes	Freedom of speech, addiction, the death penalty, ...	Humanities
Electrical Engineering	Circuits, power systems, electrical drives, ...	STEM	Moral Scenarios	Detecting physical violence, stealing, externalities, ...	Humanities
Elementary Mathematics	Word problems, multiplication, remainders, rounding, ...	STEM	Nutrition	Metabolism, water-soluble vitamins, diabetes, ...	Other
Formal Logic	Propositions, predicate logic, first-order logic, ...	Humanities	Philosophy	Skepticism, phronesis, skepticism, Singer's Drowning Child, ...	Humanities
Global Facts	Extreme poverty, literacy rates, life expectancy, ...	Other	Prehistory	Neanderthals, Mesoamerica, extinction, stone tools, ...	Humanities
High School Biology	Natural selection, heredity, cell cycle, Krebs cycle, ...	STEM	Professional Accounting	Auditing, reporting, regulation, valuation, ...	Other
High School Chemistry	Chemical reactions, ions, acids and bases, ...	STEM	Professional Law	Torts, criminal law, contracts, property, evidence, ...	Humanities
High School Computer Science	Arrays, conditionals, iteration, inheritance, ...	STEM	Professional Medicine	Diagnosis, pharmacotherapy, disease prevention, ...	Other
High School European History	Renaissance, reformation, industrialization, ...	Humanities	Professional Psychology	Diagnosis, biology and behavior, lifespan development, ...	Social Sciences
High School Geography	Population migration, rural land-use, urban processes, ...	Social Sciences	Public Relations	Media theory, crisis management, intelligence gathering, ...	Social Sciences
High School Gov't and Politics	Branches of government, civil liberties, political ideologies, ...	Social Sciences	Security Studies	Environmental security, terrorism, weapons of mass destruction, ...	Social Sciences
High School Macroeconomics	Economic indicators, national income, international trade, ...	Social Sciences	Sociology	Socialization, cities and community, inequality and wealth, ...	Social Sciences
High School Mathematics	Pre-algebra, algebra, trigonometry, calculus, ...	STEM	US Foreign Policy	Soft power, Cold War foreign policy, isolationism, ...	Social Sciences
High School Microeconomics	Supply and demand, imperfect competition, market failure, ...	Social Sciences	Virology	Epidemiology, coronaviruses, retroviruses, herpesviruses, ...	Other
High School Physics	Kinematics, energy, torque, fluid pressure, ...	STEM	World Religions	Judaism, Christianity, Islam, Buddhism, Jainism, ...	Humanities
High School Psychology	Behavior, personality, emotions, learning, ...	Social Sciences			

Summary of all 57 tasks.

<https://arxiv.org/pdf/2009.03300>

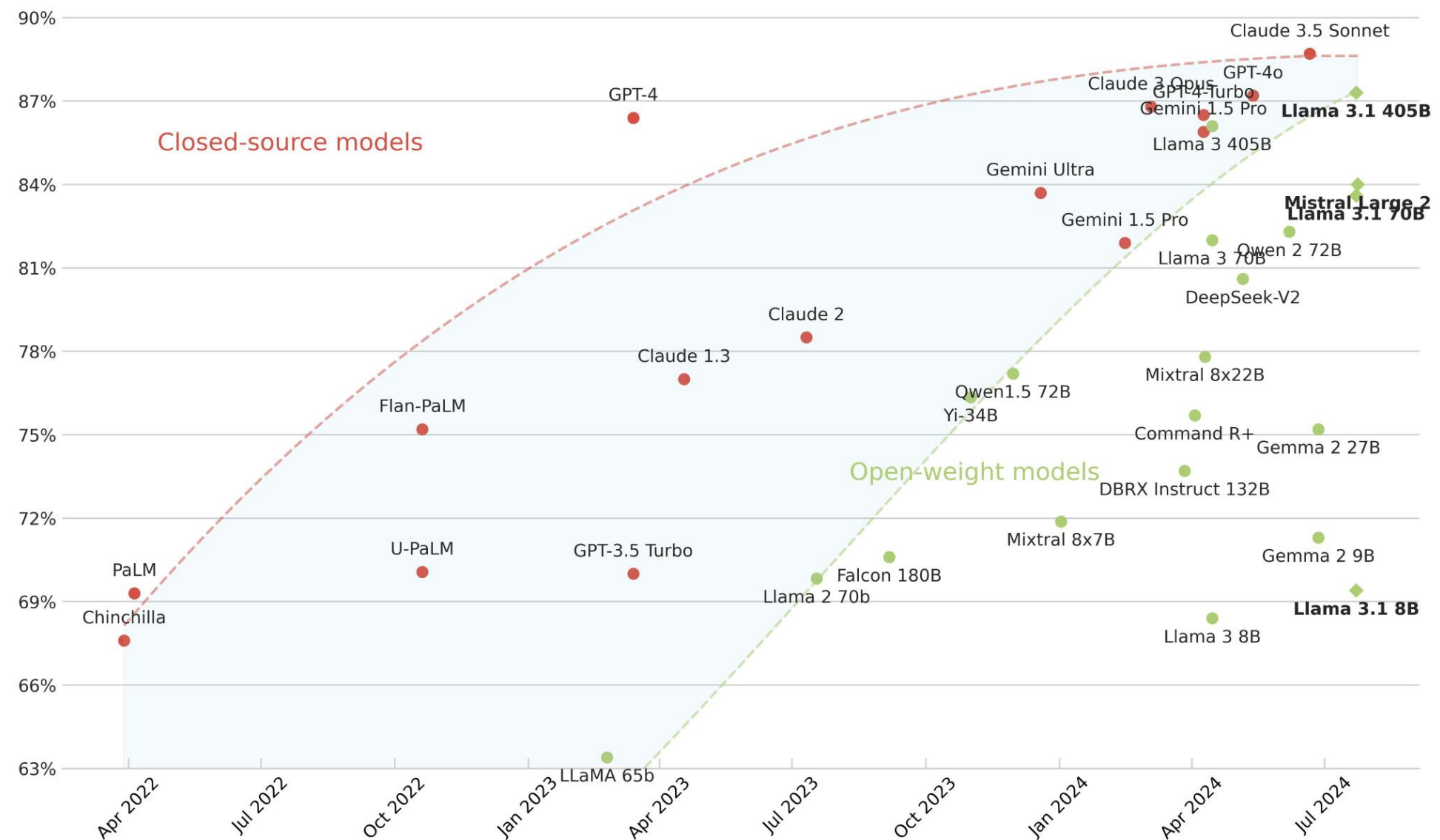
MMLU – Tracking the Progress of LLMs

Closed-source vs. open-weight models

@maximelabonne

Llama 3.1 405B closes the gap with closed-source models for the first time in history.

MMLU (5-shot)



<https://x.com/maximelabonne/status/1816416043511808259>

Big Bench Hard (BBH)

# Tasks	Criteria
<u>209</u>	<u>All BIG-Bench tasks</u>
187	- After filtering out tasks with more than three subtasks
130	- After filtering out tasks with fewer than 103 examples (3 for few-shot, 100 for evaluation)
85	- After filtering out tasks without human-rater baselines
78	- After filtering out tasks that do not use multiple-choice or exact match as the evaluation metric
<u>78</u>	<u>Clean multiple-choice or exact match tasks</u>
36	- After filtering out tasks in which the best reported model beats average reported human-rater score
23	- After filtering out extremely difficult tasks that are outside the scope of this work
23	Remaining tasks = BIG-Bench Hard (BBH)

Table 1: Filtering criteria to used to create the BIG-Bench Hard (BBH) subset. Exact names of the BIG-Bench tasks filtered out by each criteria are shown in Appendix D.

<https://arxiv.org/pdf/2210.09261>

Big Bench Hard (BBH)

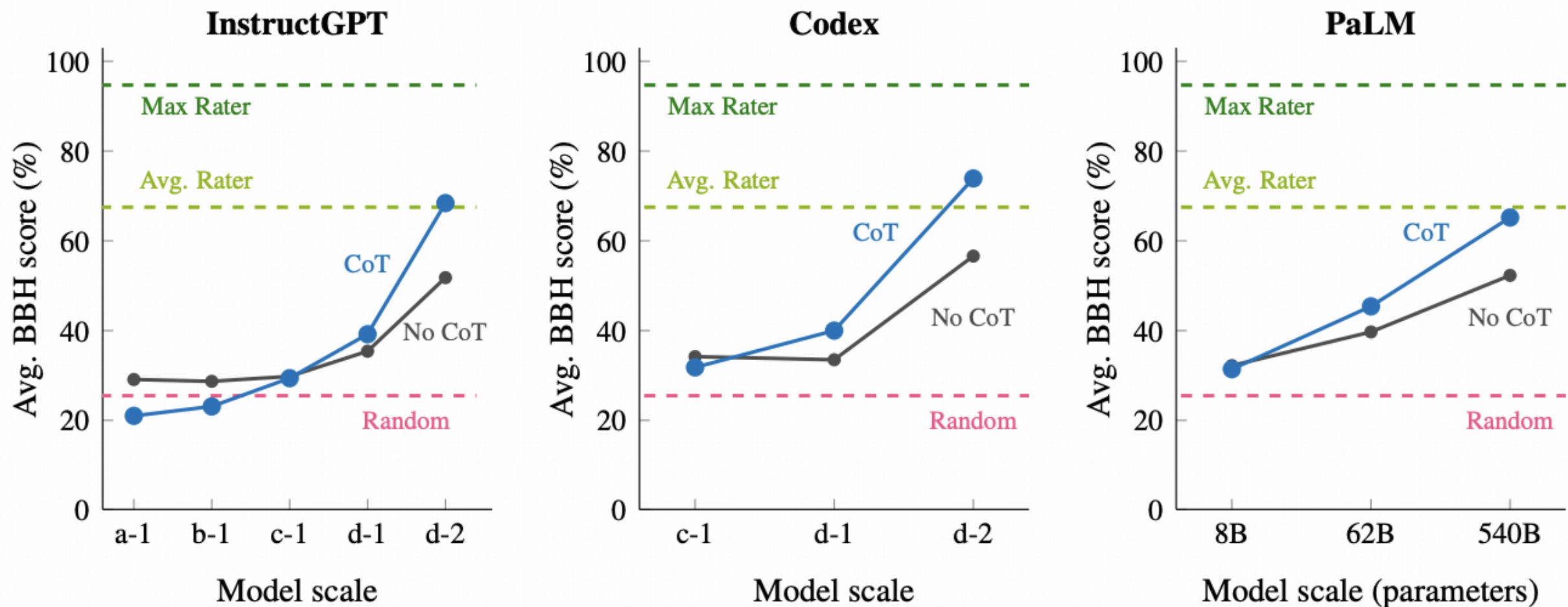


Figure 4: Scaling behavior of chain-of-thought (CoT) prompting on BIG-Bench Hard (BBH; 23 task unweighted average). InstructGPT models are the following: a-1 (text-ada-001), b-1 (text-babbage-001), c-1 (text-curie-001), d-1 (text-davinci-001), and d-2 (text-davinci-002). Codex models are the following: c-1 (code-cushman-001), d-1 (code-davinci-002), and d-2 (code-davinci-002).

<https://arxiv.org/pdf/2210.09261>

HELM: Holistic Evaluation of Language Models

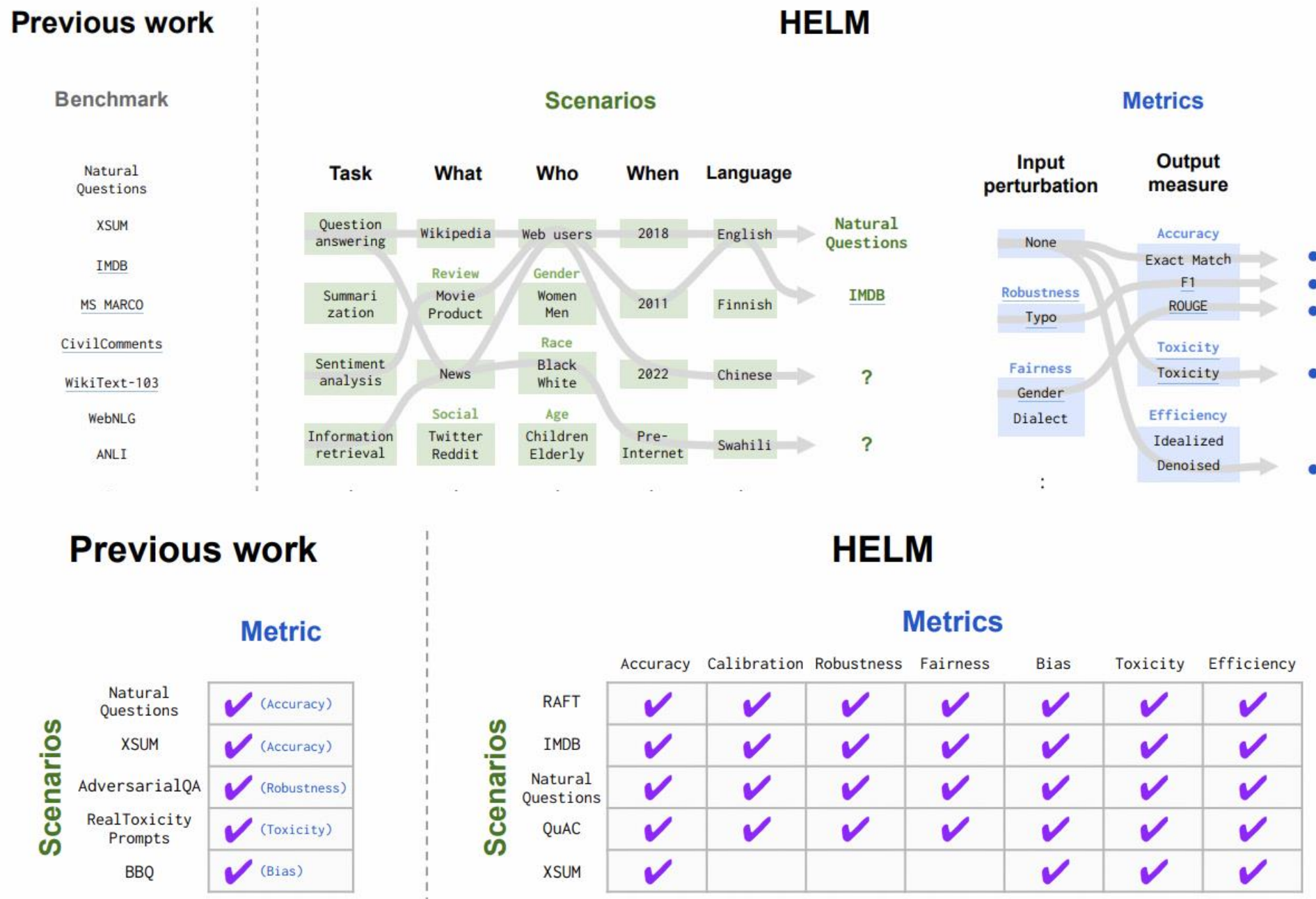


Figure 3: **Many metrics for each use case.** In comparison to most prior benchmarks of language technologies, which primarily center accuracy and often relegate other desiderata to their own bespoke datasets (if at all), in HELM we take a multi-metric approach. This foregrounds metrics beyond accuracy and allows one to study the tradeoffs between the metrics.

Chatbot Arena: Evaluation by Human Preference

The screenshot displays the Chatbot Arena interface. At the top left, there is a "Chat now!" button and a search bar with the text "Expand to see the descriptions of 72 models". Below this, two chat windows are shown side-by-side, labeled "Model A" and "Model B".

Model A: claude-3-5-sonnet-20241022

How to evaluate LLMs?

Evaluating Large Language Models (LLMs) involves multiple approaches and metrics. Here's a comprehensive overview of how to evaluate LLMs:

1. Common Evaluation Metrics
 - a) Accuracy Metrics:
 - Perplexity
 - BLEU score (for translation tasks)
 - ROUGE score (for summarization)
 - F1 score
 - Precision and Recall
 - Accuracy percentage
 - b) Qualitative Metrics:
 - Coherence
 - Relevance
 - Fluency

Model B: gemini-1.5-pro-002

How to evaluate LLMs?

Evaluating Large Language Models (LLMs) is a complex and evolving field. There's no single perfect metric, and the best approach depends on the specific application and desired capabilities of the LLM. Generally, evaluation strategies fall into two categories: intrinsic and extrinsic.

Intrinsic Evaluation: This involves assessing the LLM's inherent qualities directly, often using standardized benchmarks and datasets. These methods are generally faster and cheaper than extrinsic evaluations.

- Metrics:
 - Perplexity:** Measures how well the model predicts the next word in a sequence. Lower perplexity indicates better performance. It's a good measure of general language modeling ability but doesn't capture specific task performance.
 - Accuracy:** For tasks with clear right and wrong answers (e.g., question answering, multiple-choice), accuracy measures the percentage of correct responses.
 - F1 Score:** A harmonic mean of precision and recall, useful for tasks like information retrieval and named entity recognition where both completeness (recall) and correctness (precision) are important.
 - BLEU (Bilingual Evaluation Understudy) and METEOR (Metric for Evaluation of**

At the bottom of the interface, there are four buttons for voting: "A is better", "B is better", "Tie", and "Both are bad". Below these is a text input field with the placeholder "Press '🔄 New Round' to start over 🗳️ (Note: Your vote shapes the leaderboard, please vote RESPONSIBLY!)" and a "Send" button. At the very bottom, there are four more buttons: "Random Image", "New Round", "Regenerate", and "Share".

Info
🗳️ Thanks for voting! Your vote shapes the leaderboard, please vote RESPONSIBLY.

Elo Rating System

The Elo rating system is a method for calculating the relative skill levels of players, which has been widely adopted in competitive games and sports. The difference in the ratings between two players serves as a predictor of the outcome of a match. The Elo rating system works well for our case because we have multiple models and we run pairwise battles between them.

If player A has a rating of R_A and player B a rating of R_B , the exact formula (using the logistic curve with base 10) for the probability of player A winning is

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}} .$$

The ratings of players can be linearly updated after each battle. Suppose player A (with Rating R_A) was expected to score E_A points but actually scored S_A points. The formula for updating that player's rating is

$$R'_A = R_A + K \cdot (S_A - E_A) .$$

Chatbot Arena Leaderboard

Category		Apply filter		Overall Questions					
Overall		<input type="checkbox"/> Style Control	<input type="checkbox"/> Show Deprecated	#models: 159 (100%) #votes: 2,141,583 (100%)					
Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License	Knowledge Cutoff	
1	1	ChatGPT-4o-latest_(2024-09-03)	1340	+4/-3	33743	OpenAI	Proprietary	2023/10	
1	1	o1-preview	1335	+4/-4	21071	OpenAI	Proprietary	2023/10	
3	6	o1-mini	1308	+4/-4	23128	OpenAI	Proprietary	2023/10	
3	4	Gemini-1.5-Pro-002	1303	+4/-4	15736	Google	Proprietary	Unknown	
4	4	Gemini-1.5-Pro-Exp-0827	1299	+4/-3	32385	Google	Proprietary	2023/11	
6	9	Grok-2-08-13	1290	+3/-3	40873	xAI	Proprietary	2024/3	
6	3	Claude 3.5 Sonnet_(20241022)	1286	+6/-6	7284	Anthropic	Proprietary	2024/4	
6	11	Yi-Lightning	1285	+4/-4	20973	01 AI	Proprietary	Unknown	
6	4	GPT-4o-2024-05-13	1285	+3/-3	102960	OpenAI	Proprietary	2023/10	
10	15	GLM-4-Plus	1275	+4/-4	19922	Zhipu AI	Proprietary	Unknown	
10	18	GPT-4o-mini-2024-07-18	1273	+4/-3	42661	OpenAI	Proprietary	2023/10	
10	19	Gemini-1.5-Flash-002	1272	+5/-6	12379	Google	Proprietary	Unknown	
10	26	Llama-3.1-Nemotron-70b-Instruct	1271	+5/-7	6228	Nvidia	Llama 3.1	2023/12	
10	14	Gemini-1.5-Flash-Exp-0827	1269	+4/-4	25503	Google	Proprietary	2023/11	
11	6	Claude 3.5 Sonnet_(20240620)	1268	+3/-3	81086	Anthropic	Proprietary	2024/4	
11	25	Grok-2-Mini-08-13	1267	+4/-3	34105	xAI	Proprietary	2024/3	
11	8	Meta-Llama-3.1-405b-Instruct-fp8	1267	+4/-3	43099	Meta	Llama 3.1 Community	2023/12	
11	7	Gemini Advanced App_(2024-05-14)	1266	+3/-3	52235	Google	Proprietary	Online	
11	7	Meta-Llama-3.1-405b-Instruct-bf16	1266	+5/-6	14607	Meta	Llama 3.1 Community	2023/12	
12	14	Yi-Lightning-lite	1265	+3/-5	17271	01 AI	Proprietary	Unknown	
12	5	GPT-4o-2024-08-06	1264	+3/-4	31765	OpenAI	Proprietary	2023/10	

Chatbot Arena Leaderboard

Code to recreate leaderboard tables and plots in this [notebook](#). You can contribute your vote at [lmarena.ai](#)!

Rank* (UB)	Delta	Model	Arena Score	95% CI	Votes	Organization	License	Knowledge Cutoff
1	0	o1-preview	1302	+5/-5	21071	OpenAI	Proprietary	2023/10
1	0	ChatGPT-4o-latest (2024-09-03)	1300	+4/-3	33743	OpenAI	Proprietary	2023/10
3	3	Claude 3.5 Sonnet (20241022)	1287	+8/-6	7284	Anthropic	Proprietary	2024/4
4	-1	Gemini-1.5-Pro-002	1269	+5/-5	15736	Google	Proprietary	Unknown
4	0	Gemini-1.5-Pro-Exp-0827	1268	+4/-4	32385	Google	Proprietary	2023/11
4	2	GPT-4o-2024-05-13	1262	+4/-2	102960	OpenAI	Proprietary	2023/10
6	5	Claude 3.5 Sonnet (20240620)	1258	+3/-3	81086	Anthropic	Proprietary	2024/4
6	-3	o1-mini	1256	+4/-4	23128	OpenAI	Proprietary	2023/10
7	4	Gemini Advanced Exp (2024-05-14)	1253	+4/-3	52235	Google	Proprietary	Online
7	4	Meta-Llama-3.1-405b-Instruct-bf16	1251	+5/-5	14607	Meta	Llama 3.1 Community	2023/12
8	3	Meta-Llama-3.1-405b-Instruct-fp8	1251	+4/-3	43099	Meta	Llama 3.1 Community	2023/12
9	-3	Grok-2-08-13	1248	+3/-4	40873	xAI	Proprietary	2024/3
9	3	GPT-4o-2024-08-06	1248	+4/-3	34765	OpenAI	Proprietary	2023/10
11	-5	Yi-Lightning	1241	+5/-5	20973	01 AI	Proprietary	Unknown
14	8	GPT-4-Turbo-2024-04-09	1241	+3/-2	99055	OpenAI	Proprietary	2023/12
14	5	Gemini-1.5-Pro-001	1239	+3/-3	82665	Google	Proprietary	2023/11
14	13	Claude 3 Opus	1238	+2/-2	177919	Anthropic	Proprietary	2023/8
14	-4	Gemini-1.5-Flash-Exp-0827	1237	+5/-3	25503	Google	Proprietary	2023/11
14	-2	Yi-Lightning-lite	1237	+4/-4	17271	01 AI	Proprietary	Unknown
15	11	GPT-4-1106-preview	1234	+3/-2	103422	OpenAI	Proprietary	2023/4
15	-5	GEM-4-Blue	1234	+5/-6	10022	Zhipu AI	Proprietary	Unknown

Apply filter
 Style Control Show Deprecated

Overall Leaderboard with Style Control. See details in [blog post](#).

#models: 159 (100%) #votes: 2,141,583 (100%)

Lower the influence of response style

Chatbot Arena Leaderboard

Arena [NEW: Overview](#) [Arena \(Vision\)](#) [Arena-Hard-Auto](#) [Full Leaderboard](#)

Total #models: 159. Total #votes: 2,141,583. Last updated: 2024-10-27.

Code to recreate leaderboard tables and plots in this [notebook](#). You can contribute your vote at [lmarena.ai](#)!

Category: Overall

Apply filter: Style Control Show Deprecated

Overall Leaderboard with Style Control. See details in [blog post](#).
#models: 159 (100%) #votes: 2,141,583 (100%)

			Arena Score	95% CI	Votes	Organization	License	Knowledge Cutoff
1			1302	+5/-5	21071	OpenAI	Proprietary	2023/10
2		GPT-4o (2024-09-03)	1300	+4/-3	33743	OpenAI	Proprietary	2023/10
3		Claude-3.5-Sonnet (20241022)	1287	+8/-6	7284	Anthropic	Proprietary	2024/4
4		GPT-4o-mini (2024-08-02)	1269	+5/-5	15736	Google	Proprietary	Unknown
5		Claude-3.5-Exp-0827	1268	+4/-4	32385	Google	Proprietary	2023/11
6		GPT-4o-mini (2024-08-13)	1262	+4/-2	102960	OpenAI	Proprietary	2023/10
7		Claude-3.5-Sonnet (20240620)	1258	+3/-3	81086	Anthropic	Proprietary	2024/4
8			1256	+4/-4	23128	OpenAI	Proprietary	2023/10
9		Gemini-1.5-Flash-Exp (2024-05-14)	1253	+4/-3	52235	Google	Proprietary	Online
10		Claude-3.5-Exp-Instruct-bf16	1251	+5/-5	14607	Meta	Llama 3.1 Community	2023/12
11		Claude-3.5-Exp-Instruct-fp8	1251	+4/-3	43099	Meta	Llama 3.1 Community	2023/12
12			1248	+3/-4	40873	xAI	Proprietary	2024/3
13		GPT-4o-mini (2024-06-06)	1248	+4/-3	34765	OpenAI	Proprietary	2023/10
14			1241	+5/-5	20973	01 AI	Proprietary	Unknown
15		Claude-3.5-Exp (2024-04-09)	1241	+3/-2	99055	OpenAI	Proprietary	2023/12
16	5	Gemini-1.5-Pro-001	1239	+3/-3	82665	Google	Proprietary	2023/11
17	13	Claude-3-Opus	1238	+2/-2	177919	Anthropic	Proprietary	2023/8
18	-4	Gemini-1.5-Flash-Exp-0827	1237	+5/-3	25503	Google	Proprietary	2023/11
19	-2	Yi-Lightning-lite	1237	+4/-4	17271	01 AI	Proprietary	Unknown
20	11	GPT-4-1106-preview	1234	+3/-2	103422	OpenAI	Proprietary	2023/4
21	-5	GLM-4-Plus	1234	+5/-6	10022	Zhipu AI	Proprietary	Unknown

LLM-as-Judge: MT-Bench

Question: If the FED buys bonds in the secondary market (A) the money supply will increase. (B) the money supply will decrease. (C) the money supply will not be affected. (D) the discount rate would be affected. Output the answer only.

Assistant A:

The Federal Reserve buys bonds in the secondary market to increase the money supply.

Assistant B:

(A) the money supply will increase.

Follow-up Question: How does it affect my daily life? Give 3 examples.

Assistant A:

The Federal Reserve buys bonds in the secondary market to increase the money supply. This affects your daily life by:

1. Increasing the money supply.
2. Increasing the money supply and lowering the discount rate.
3. Increasing the money supply and lowering the discount rate and increasing the money supply.

Assistant B:

When the Federal Reserve buys bonds in the secondary market, it can have several effects on the economy that can impact your daily life. Here are three examples:

1. Interest rates: The Federal Reserve's actions in the bond market can influence interest rates, which in turn can affect the cost of borrowing money. For example, if the Federal Reserve buys bonds and drives up demand for them, interest rates on things like mortgages, car loans, and credit cards may go up, making it more expensive to borrow money.
2. Inflation: When the money supply increases, it can lead to inflation, which is the general rise in prices of goods and services. This can impact your daily life by making it more expensive to buy things like groceries, gas, and other necessities.
3. Employment: The Federal Reserve's actions in the bond market can also have an impact on employment. For example, if the Federal Reserve's actions lead to economic growth and increased demand for goods and services, it can create more job opportunities and boost employment.

GPT-4 Judgment:

Assistant A provided an incorrect response to the user's question about how the Federal Reserve buying bonds in the secondary market affects daily life. The answer given is repetitive and lacks clear examples of how the action impacts daily life.

On the other hand, **Assistant B** provided a relevant and accurate response to the user's question about the Federal Reserve buying bonds. The answer includes three clear examples of how the action impacts daily life, such as interest rates, inflation, and employment.

Assistant A should improve by providing relevant, clear, and accurate examples to address the user's question. **Assistant B**'s response is more helpful, relevant, and detailed.

LLM-as-Judge: Pros and Cons

Pros:

- 1. Scalability and Speed:** LLMs can evaluate a large number of responses or submissions quickly
- 2. Cost Efficiency:** Using an LLM can be less costly than relying on multiple human judges, especially for large-scale evaluations.
- 3. Flexibility:** LLMs can be fine-tuned or adjusted for specific criteria or tasks, allowing them to judge based on different sets of standards.

Cons:

- 1. Risk of Misinterpretation:** LLMs may misinterpret input if phrasing is unusual.
- 2. Vulnerability to Manipulation.** LLMs may be susceptible to adversarial prompts or "gaming" techniques (e.g., jailbreaking).
- 3. Limited Understanding of Nuance:** LLMs can struggle with understanding subjective, nuanced, or context-dependent elements in evaluation,

LLMs Evaluation Open Questions

- How can we speed and scale up evaluation while maintaining the evaluation accuracy?
- How can we more efficiently and cost-effectively approximate human preference?
- How to solve the mismatch between human preference (e.g., Chatbot Arena) and task accuracy on standard benchmarks (e.g., MMLU, MATH)?
- How to construct benchmarks that are in a dynamic fashion?
- How to construct benchmark to lower contamination issues?

Multimodal LLMs

Multimodal Behaviors and Signals

Language

- **Lexicon**
 - Words
- **Syntax**
 - Part-of-speech
 - Dependencies
- **Pragmatics**
 - Discourse acts

Acoustic

- **Prosody**
 - Intonation
 - Voice quality
- **Vocal expressions**
 - Laughter, moans

Visual

- **Gestures**
 - Head gestures
 - Eye gestures
 - Arm gestures
- **Body language**
 - Body posture
 - Proxemics
- **Eye contact**
 - Head gaze
 - Eye gaze
- **Facial expressions**
 - FACS action units
 - Smile, frowning

Touch

- Haptics
- Motion

Physiological

- Skin conductance
- Electrocardiogram

Mobile

- GPS location
- Accelerometer
- Light sensors



Multimodal Machine Learning

*What are the **core multimodal technical challenges**,
understudied in conventional machine learning?*

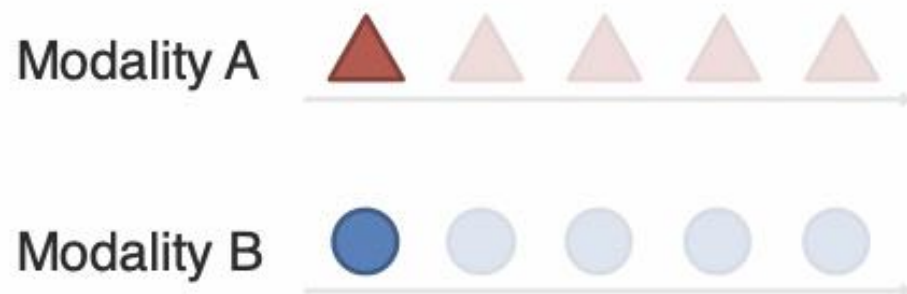


Challenge 1: Representation

Definition: Learning representations that reflect cross-modal interactions between individual elements, across different modalities

➔ This is a core building block for most multimodal modeling problems!

Individual elements:



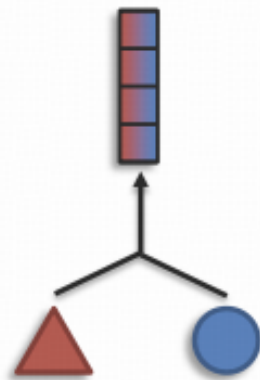
*It can be seen as a “local” representation
or
representation using holistic features*

Challenge 1: Representation

Definition: Learning representations that reflect cross-modal interactions between individual elements, across different modalities

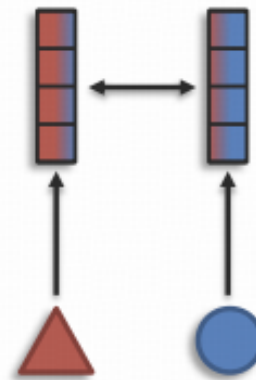
Sub-challenges:

Fusion



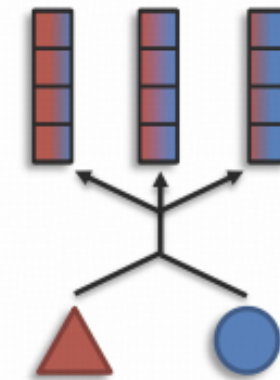
modalities \gt # representations

Coordination



modalities $=$ # representations

Fission



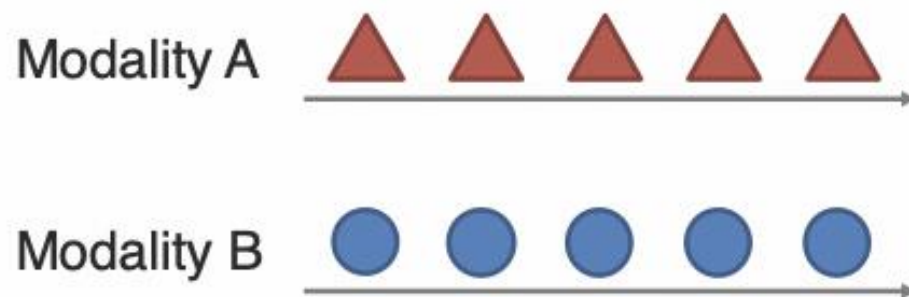
modalities \lt # representations

Challenge 2: Alignment

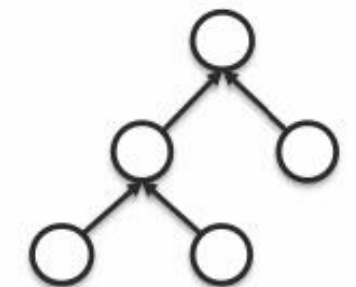
Definition: Identifying and modeling cross-modal connections between all elements of multiple modalities, building from the data structure

➔ Most modalities have internal structure with multiple elements

Elements with temporal structure:

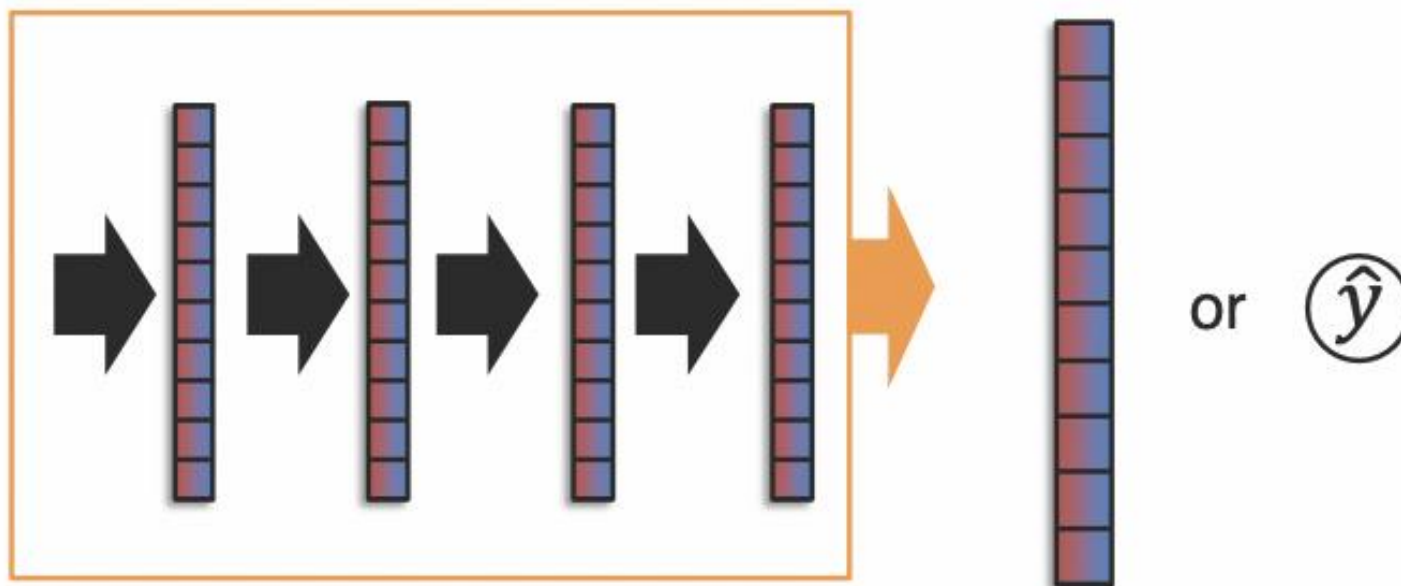
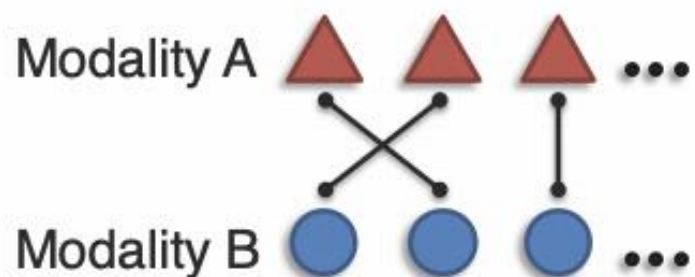


Other structured examples:



Challenge 3: Reasoning

Definition: Combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and problem structure

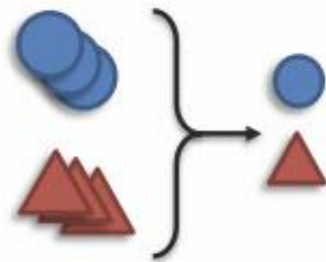


Challenge 4: Generation

Definition: Learning a generative process to produce raw modalities that reflects cross-modal interactions, structure and coherence

Sub-challenges:

Summarization



Reduction



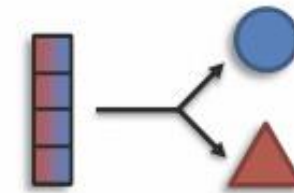
Translation



Maintenance



Creation



Expansion



Information:
(content)

A 3D perspective view of a grid of colored squares. The grid is composed of red, green, and blue squares, arranged in a repeating pattern. The grid recedes into the distance, creating a sense of depth. The colors are vibrant and the grid lines are clearly visible.

Multimodal Tasks

Media description dataset 1 – MS COCO

- Microsoft Common Objects in COntext ([MS COCO](#))
- 120,000 images
- Each image is accompanied with five free form sentences describing it (at least 8 words)
- Sentences collected using crowdsourcing (Mechanical Turk)
- Also contains object detections, boundaries and keypoints



The man at bat readies to swing at the pitch while the umpire looks on.



A large bus sitting next to a very tall building.

Visual Questions & Answers – VQA

- Task - Given an image and a question, answer the question (<http://www.visualqa.org/>)



What color are her eyes?
What is the mustache made of?



How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?



Does it appear to be rainy?
Does this person have 20/20 vision?

Multimodal QA

- TVQA
 - Video QA dataset based on 6 popular TV shows
 - 152.5K QA pairs from 21.8K clips
 - Compositional questions

00:00.755 --> 00:02.655 (Chandler:) Go to your room!
00:06.961 --> 00:08.622 (Janice:) I gotta go, I gotta go.

00:08.829 --> 00:10.057 (Janice:) Not without a kiss.
00:10.264 --> 00:12.391 (Chandler:) Maybe I won't kiss you so you'll stay.

00:12.600 --> 00:14.761 (Joey:) Kiss her. Kiss her!
00:16.771 --> 00:19.137 (Janice:) I'll see you later, sweetie. Bye, Joey.

00:39.327 --> 00:40.760 (Chandler:) She makes me happy.
00:41.596 --> 00:44.087 (Joey:) Okay. All right.

00:00 00:06 00:10 00:17 00:39 00:45 01:04

What is Janice holding on to after Chandler sends Joey to his room?
A Chandler's tie
B Chandler's hands
C Her Breakfast
D Her coat
E Chandler's coffee cup.

Why does Joey want Chandler to kiss Janice when they are in the kitchen?
A Because Joey is glad that Chandler is happy
B Because Joey likes to watch people kiss
C Because then she will leave
D Because Joey thinks Janice is hot
E Because then Chandler will move away from the toast.

What is on the couch behind Joey when he is at the counter?
A A chick
B A soccer ball
C A duck
D A pillow
E Janice's coat

<https://arxiv.org/abs/1809.01696>

Multimodal QA – Visual Reasoning

- Cornell NLVR2
 - Same as NLVR but with >100k real images



<https://arxiv.org/pdf/1811.00491>

OK-VQA

Vehicles and Transportation



Q: What sort of vehicle uses this item?
A: firetruck

Brands, Companies and Products



Q: When was the soft drink company shown first created?
A: 1898

Objects, Material and Clothing



Q: What is the material used to make the vessels in this picture?
A: copper

Sports and Recreation



Q: What is the sports position of the man in the orange shirt?
A: goalie

Cooking and Food



Q: What is the name of the object used to eat this food?
A: chopsticks

Geography, History, Language and Culture



Q: What days might I most commonly go to this building?
A: Sunday

People and Everyday Life



Q: Is this photo from the 50's or the 90's?
A: 50's

Plants and Animals



Q: What phylum does this animal belong to?
A: chordate, chordata

Science and Technology



Q: How many chromosomes do these creatures have?
A: 23

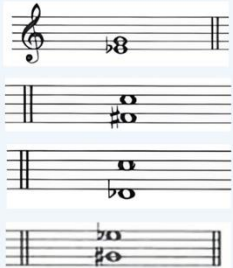
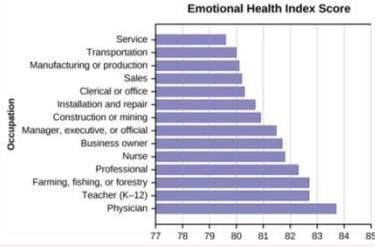
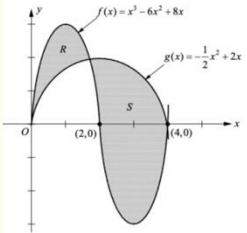
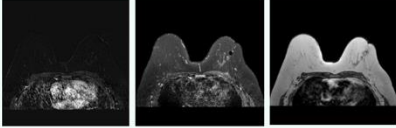

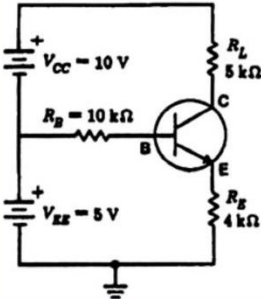
Weather and Climate



Q: What is the warmest outdoor temperature at which this kind of weather can happen?
A: 32 degrees

<https://okvqa.allenai.org/>

MMMU

Art & Design	Business	Science
<p>Question: Among the following harmonic intervals, which one is constructed incorrectly?</p> <p>Options:</p> <p>(A) Major third <i><image 1></i></p> <p>(B) Diminished fifth <i><image 2></i></p> <p><u>(C) Minor seventh <i><image 3></i></u></p> <p>(D) Diminished sixth <i><image 4></i></p> 	<p>Question: ...The graph shown is compiled from data collected by Gallup <i><image 1></i>. Find the probability that the selected Emotional Health Index Score is between 80.5 and 82?</p> <p>Options:</p> <p>(A) 0 (B) 0.2142</p> <p><u>(C) 0.3571</u> (D) 0.5</p> 	<p>Question: <i><image 1></i> The region bounded by the graph as shown above. Choose an integral expression that can be used to find the area of R.</p> <p>Options:</p> <p><u>(A) $\int_0^{1.5} [f(x) - g(x)] dx$</u></p> <p>(B) $\int_0^{1.5} [g(x) - f(x)] dx$</p> <p>(C) $\int_0^2 [f(x) - g(x)] dx$</p> <p>(D) $\int_0^2 [g(x) - x(x)] dx$</p> 
<p>Subject: Music; Subfield: Music;</p> <p>Image Type: Sheet Music;</p> <p>Difficulty: Medium</p>	<p>Subject: Marketing; Subfield: Market Research; Image Type: Plots and Charts;</p> <p>Difficulty: Medium</p>	<p>Subject: Math; Subfield: Calculus;</p> <p>Image Type: Mathematical Notations;</p> <p>Difficulty: Easy</p>
Health & Medicine	Humanities & Social Science	Tech & Engineering
<p>Question: You are shown subtraction <i><image 1></i>, T2 weighted <i><image 2></i> and T1 weighted axial <i><image 3></i> from a screening breast MRI. What is the etiology of the finding in the left breast?</p> <p>Options:</p> <p>(A) Susceptibility artifact</p> <p>(B) Hematoma</p> <p><u>(C) Fat necrosis</u> (D) Silicone granuloma</p> 	<p>Question: In the political cartoon, the United States is seen as fulfilling which of the following roles? <i><image 1></i></p> <p>Option:</p> <p>(A) Oppressor</p> <p>(B) Imperialist</p> <p><u>(C) Savior</u> (D) Isolationist</p> 	<p>Question: Find the VCE for the circuit shown in <i><image 1></i>. Neglect VBE</p> <p>Answer: 3.75</p> <p>Explanation: ...$I_E = [(V_{EE}) / (R_E)] = [(5 \text{ V}) / (4 \text{ k-ohm})] = 1.25 \text{ mA}$; $V_{CE} = V_{CC} - I_E R_L = 10 \text{ V} - (1.25 \text{ mA}) 5 \text{ k-ohm}$; $V_{CE} = 10 \text{ V} - 6.25 \text{ V} = 3.75 \text{ V}$</p> 
<p>Subject: Clinical Medicine; Subfield: Clinical Radiology; Image Type: Body Scans: MRI, CT.;</p> <p>Difficulty: Hard</p>	<p>Subject: History; Subfield: Modern History; Image Type: Comics and Cartoons;</p> <p>Difficulty: Easy</p>	<p>Subject: Electronics; Subfield: Analog electronics; Image Type: Diagrams;</p> <p>Difficulty: Hard</p>

<https://arxiv.org/pdf/2311.16502>

EPIC-Kitchens

- [Dataset](#)
- Large-scale dataset in first-person (egocentric) vision; multi-faceted, audio-visual, non-scripted recordings in native environments - i.e. the wearers' homes



<https://epic-kitchens.github.io/2022>

Multimodal Retrieval: IKEA Interior Design Dataset

- [Interior Design Dataset](#) – Retrieve desired product using room photos and text queries.
- 298 room photos, 2193 product images/descriptions.

Room images:



Object images:



Description:

You sit comfortably thanks to the armrests.

There's a natural and living feeling of wood, as knots and other marks remain on the surface.

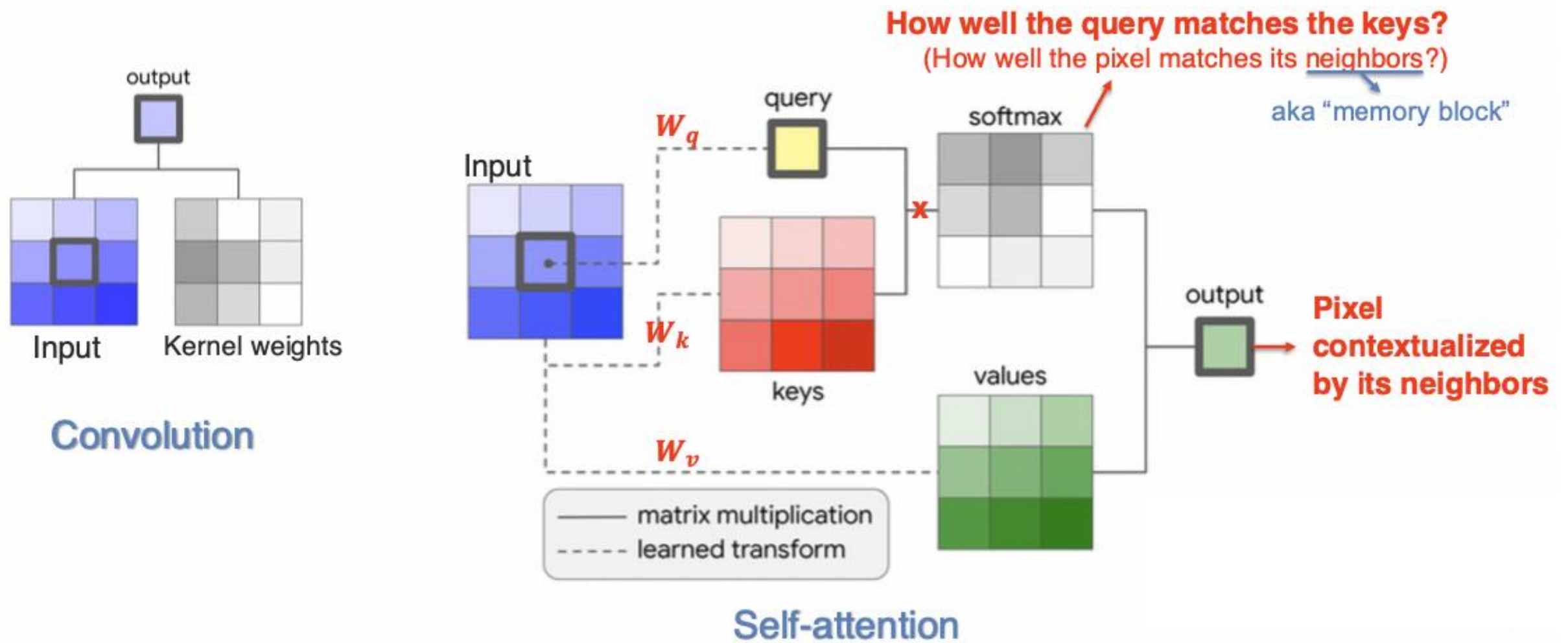
This lamp gives a pleasant light for dining and spreads a good directed light across your dining or bar table.

<https://github.com/IvonaTau/ikea>

Vision Transformers



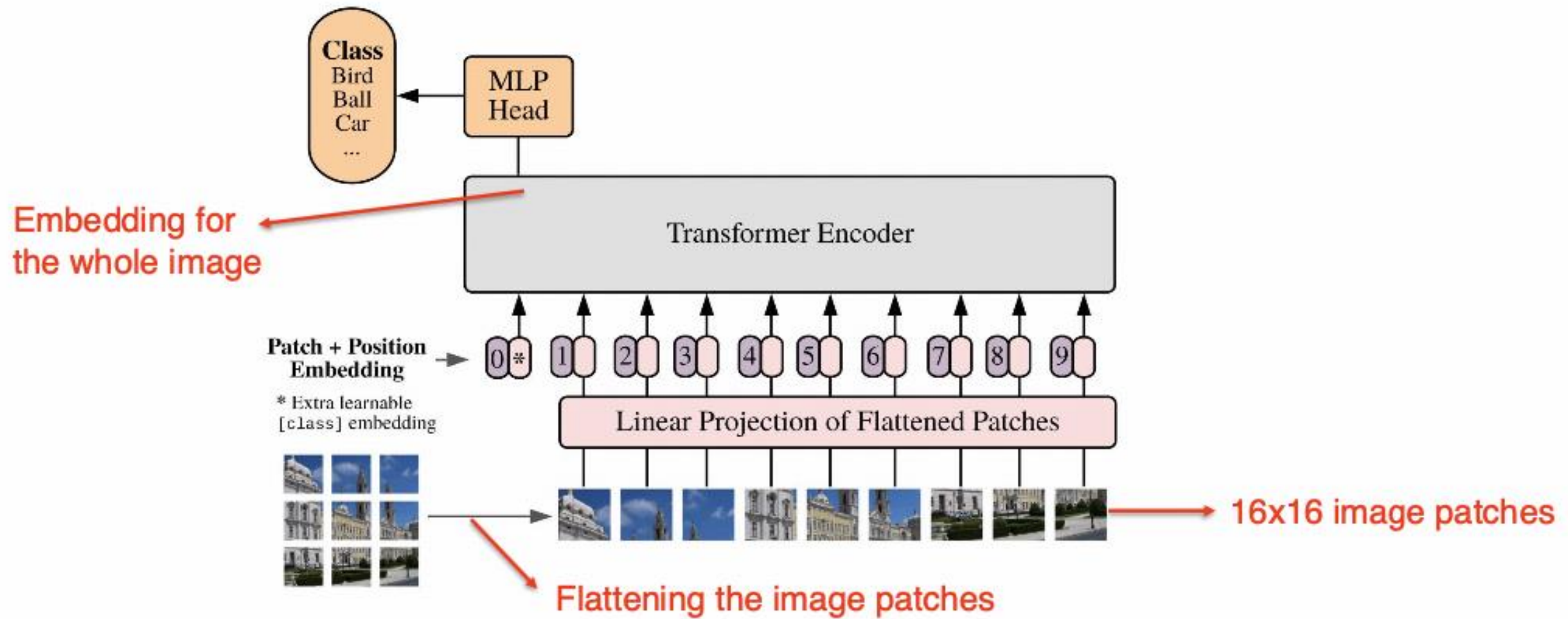
Replacing a CNN w/ Self-Attention



<https://arxiv.org/abs/1906.05909>

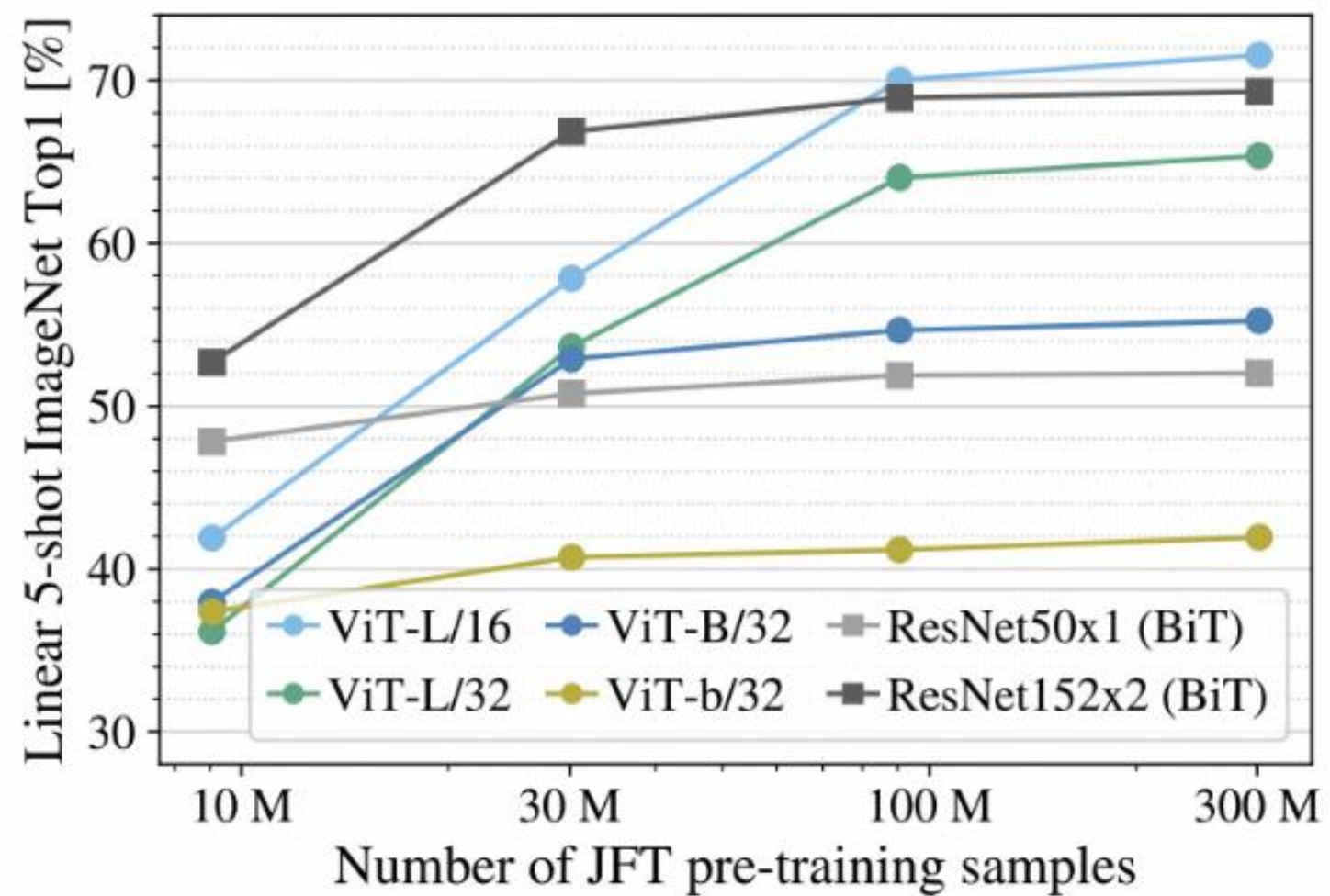
<https://arxiv.org/abs/1906.05909>

Vision Transformer (ViT)



<https://arxiv.org/abs/2010.11929>

Vision Transformer (ViT)



<https://arxiv.org/abs/2010.11929>

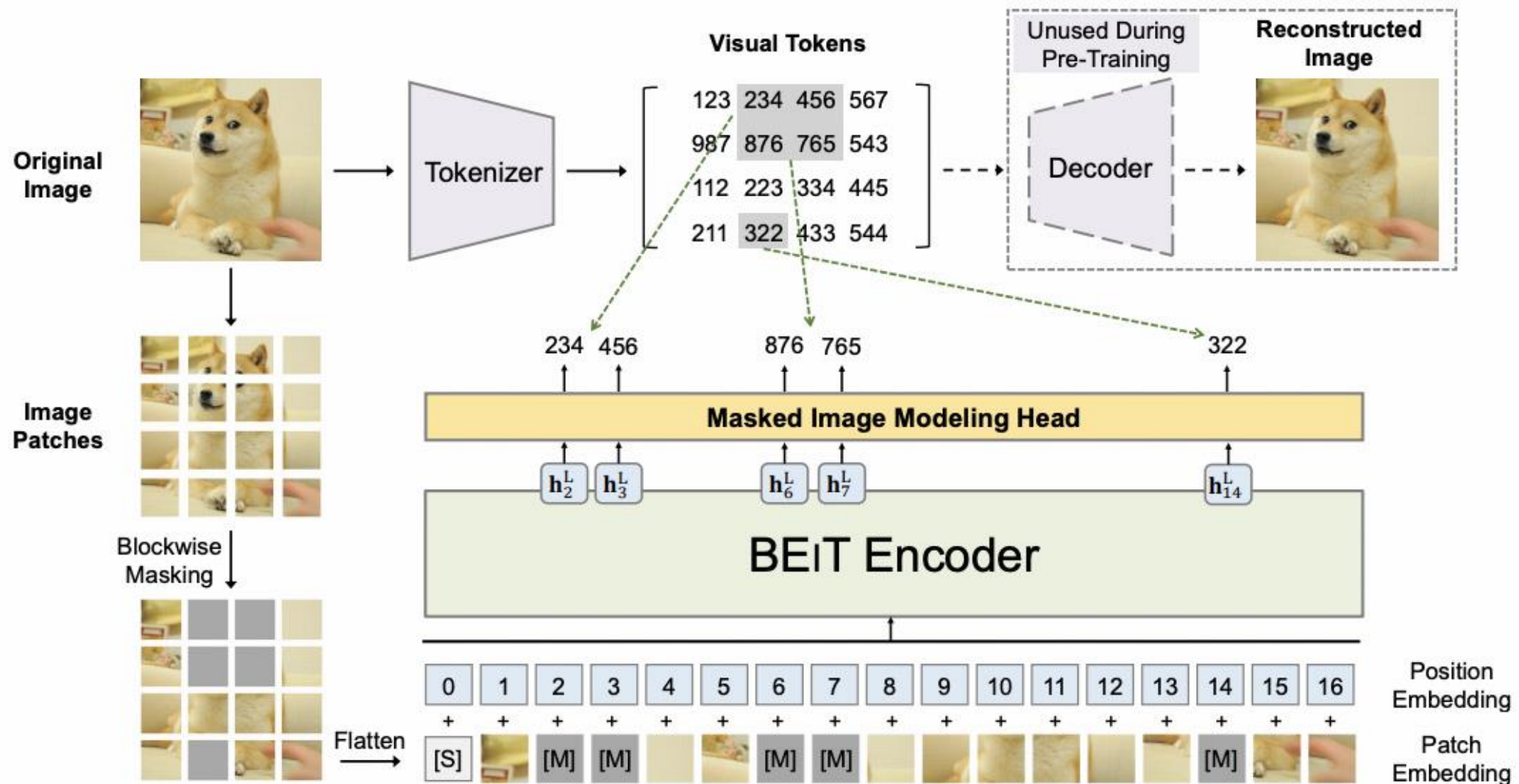
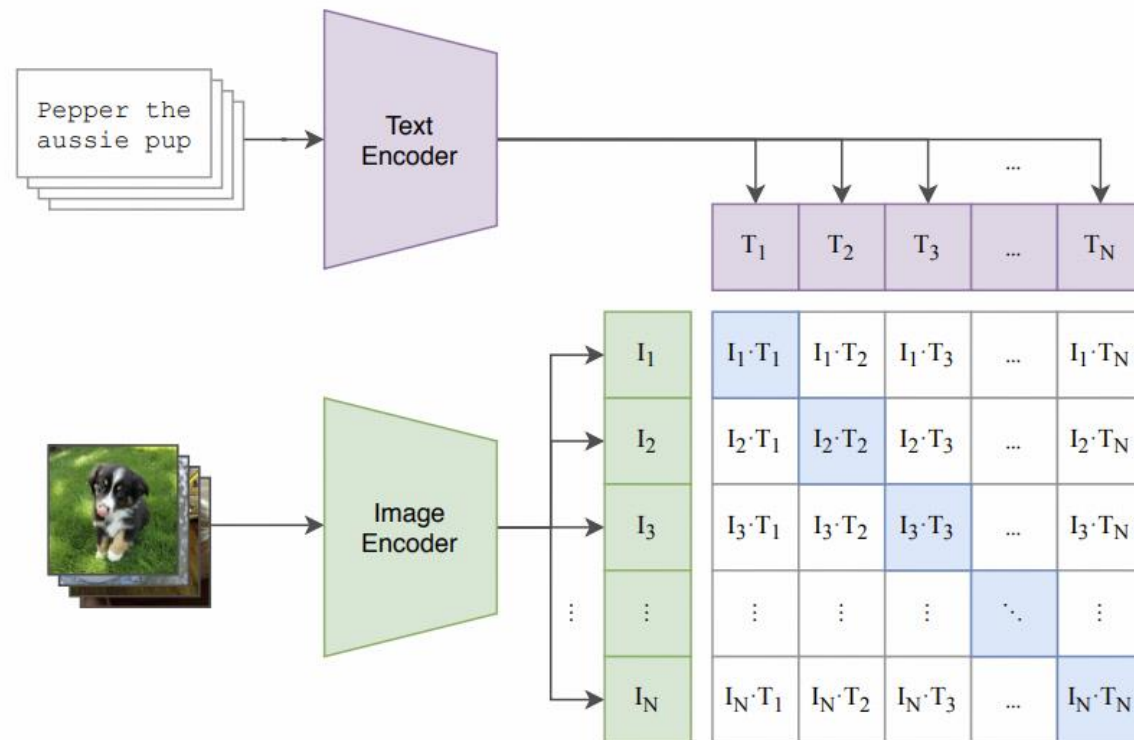


Figure 1: Overview of BEIT pre-training. Before pre-training, we learn an “image tokenizer” via autoencoding-style reconstruction, where an image is tokenized into discrete visual tokens according to the learned vocabulary. During pre-training, each image has two views, i.e., image patches, and visual tokens. We randomly mask some proportion of image patches (gray patches in the figure) and replace them with a special mask embedding [M]. Then the patches are fed to a backbone vision Transformer. The pre-training task aims at predicting the visual tokens of the *original* image based on the encoding vectors of the *corrupted* image.

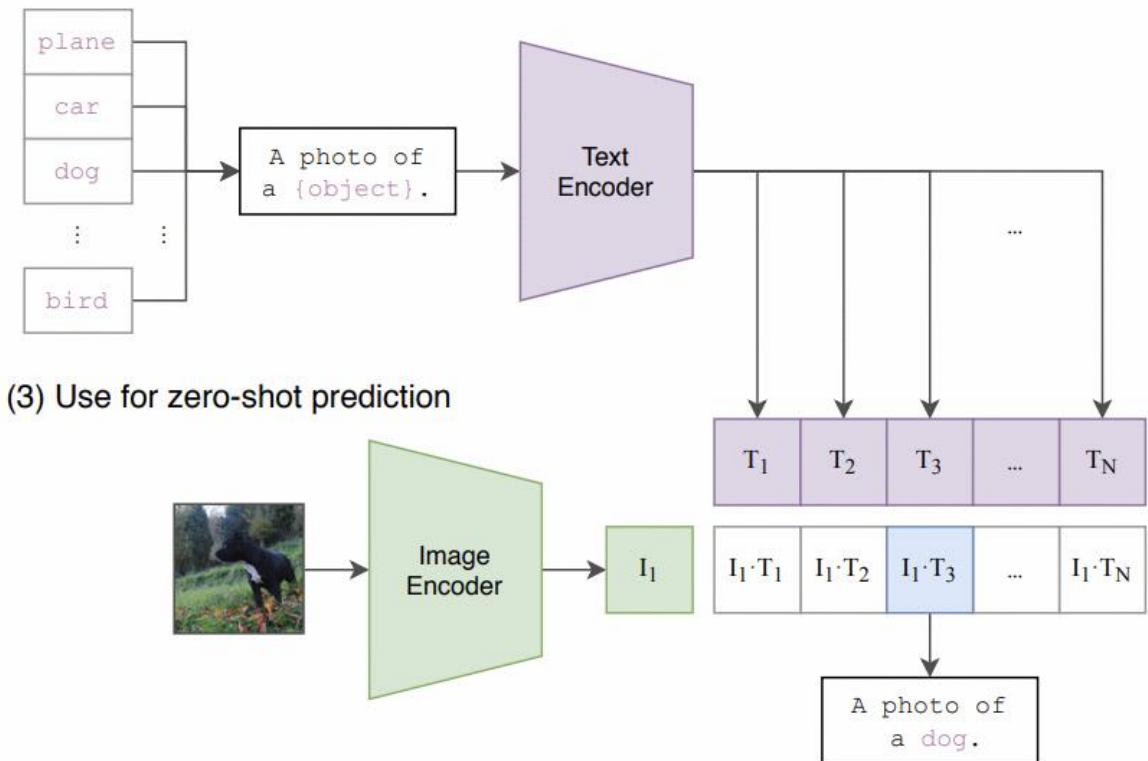
<https://arxiv.org/pdf/2106.08254>

CLIP

(1) Contrastive pre-training



(2) Create dataset classifier from label text



(3) Use for zero-shot prediction

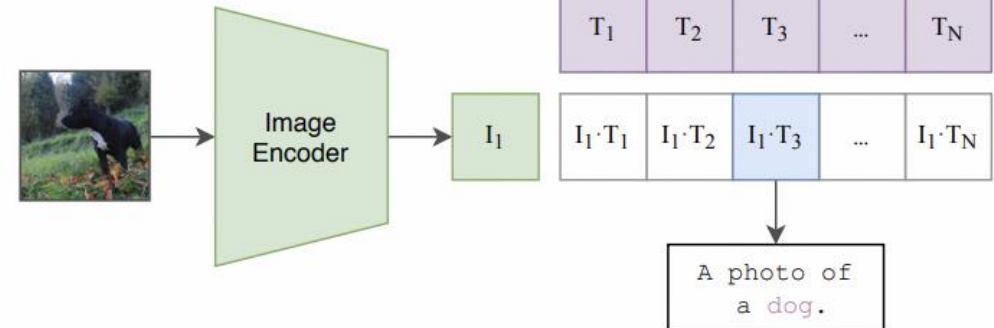


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

<https://arxiv.org/pdf/2103.00020>


```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

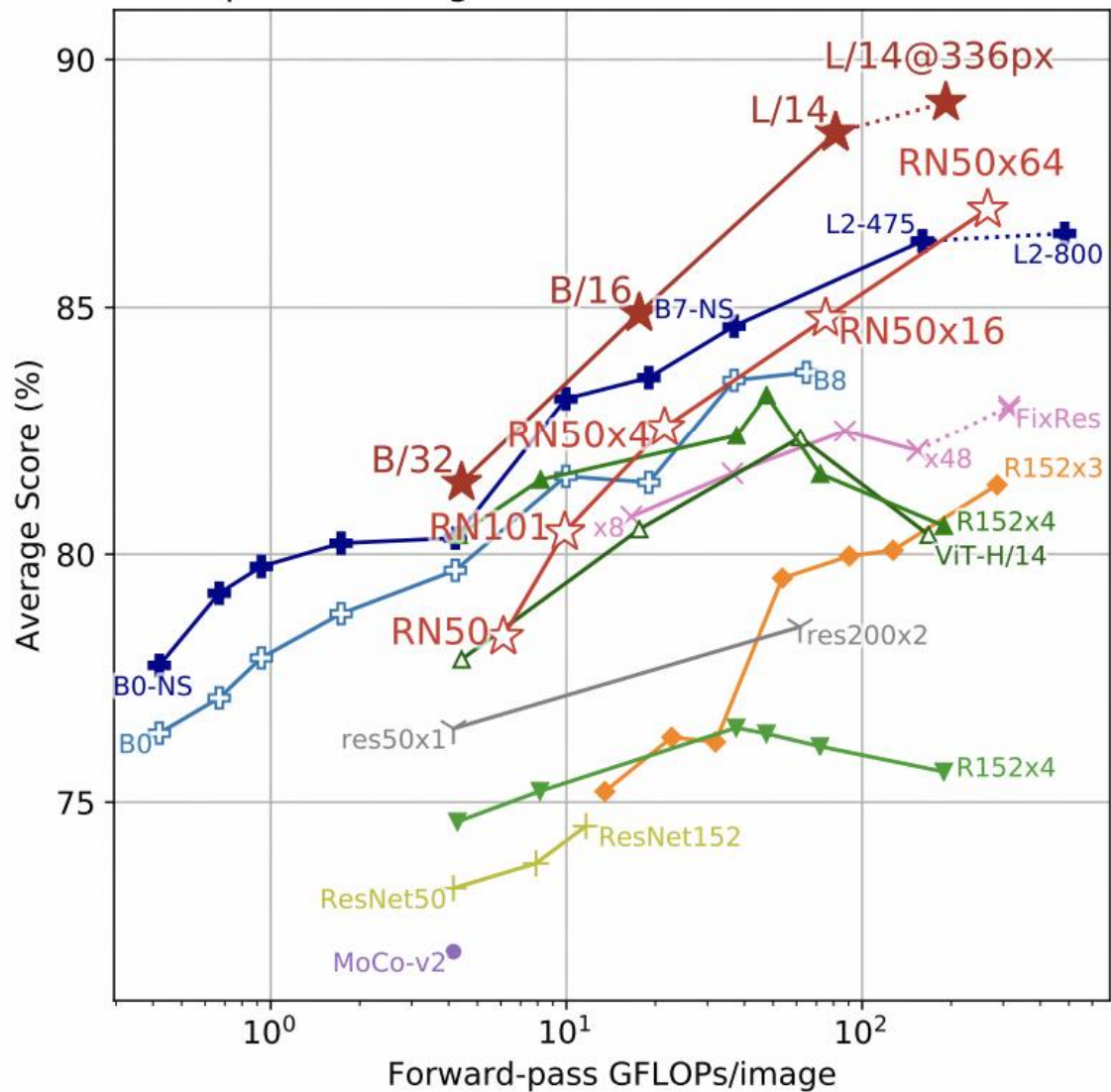
# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

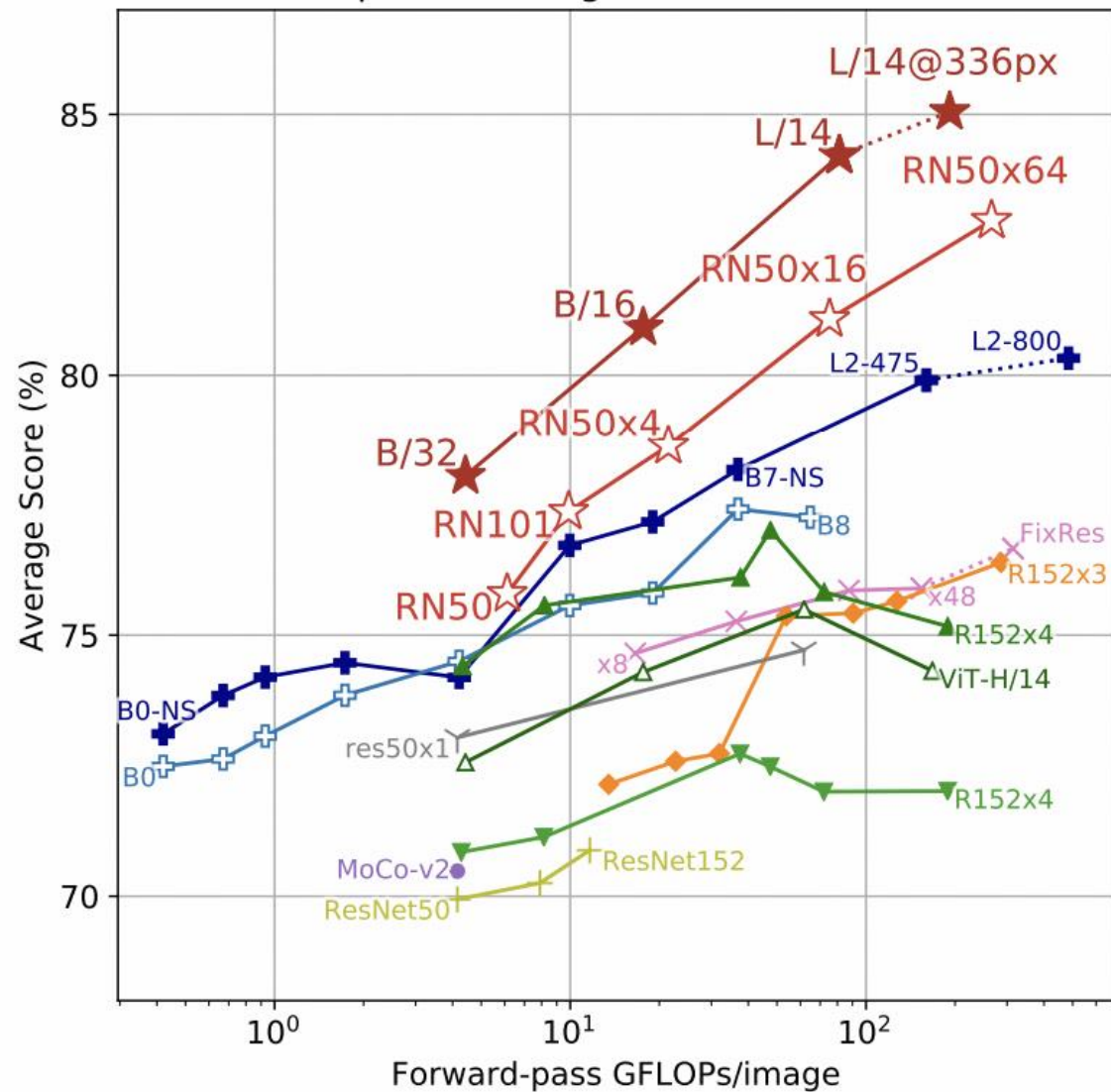
# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

<https://arxiv.org/pdf/2103.00020>

Linear probe average over Kornblith et al.'s 12 datasets



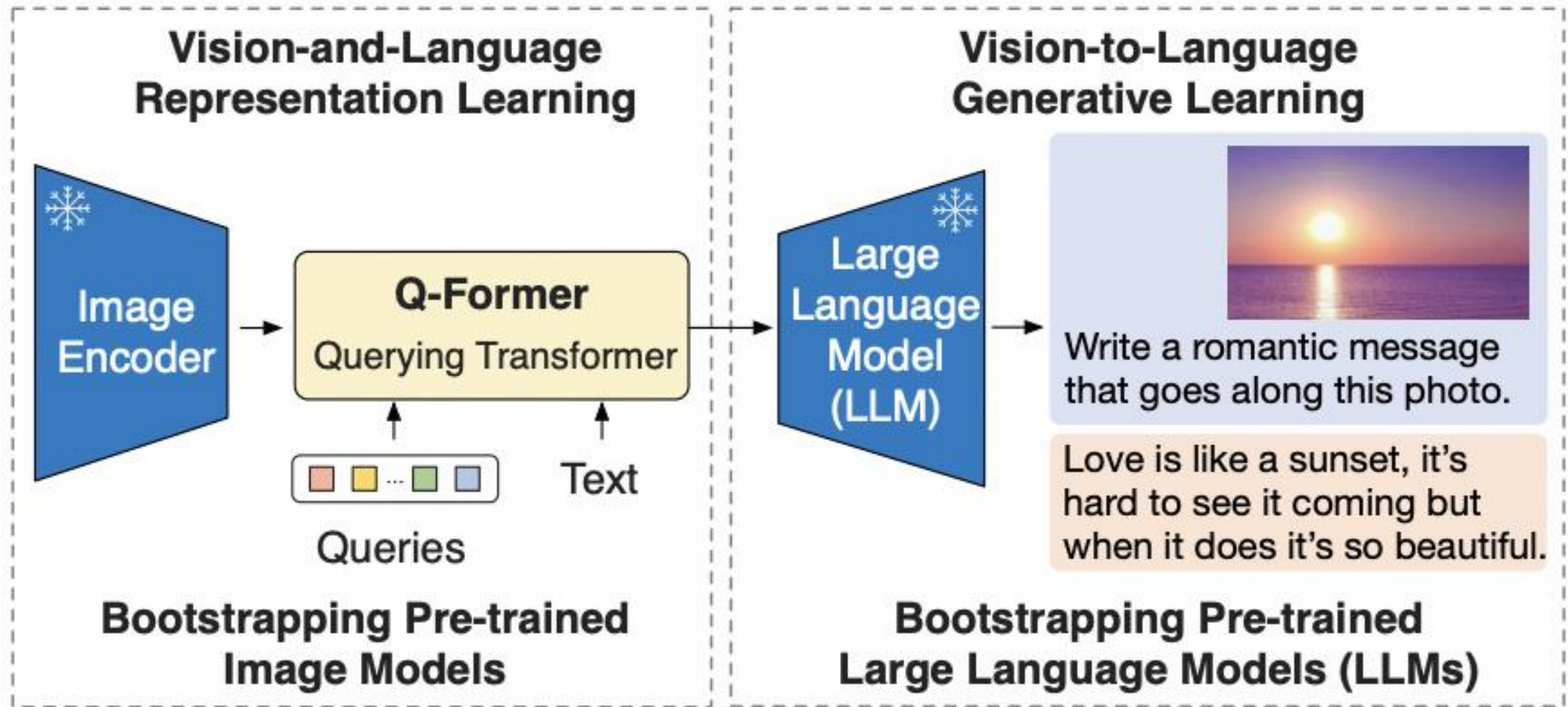
Linear probe average over all 27 datasets



- ★ CLIP-ViT
- ✕ Instagram-pretrained
- △ ViT (ImageNet-21k)
- ☆ CLIP-ResNet
- ◆ SimCLRv2
- ▲ BiT-M
- ◆ EfficientNet-NoisyStudent
- BYOL
- ▼ BiT-S
- + EfficientNet
- MoCo
- + ResNet

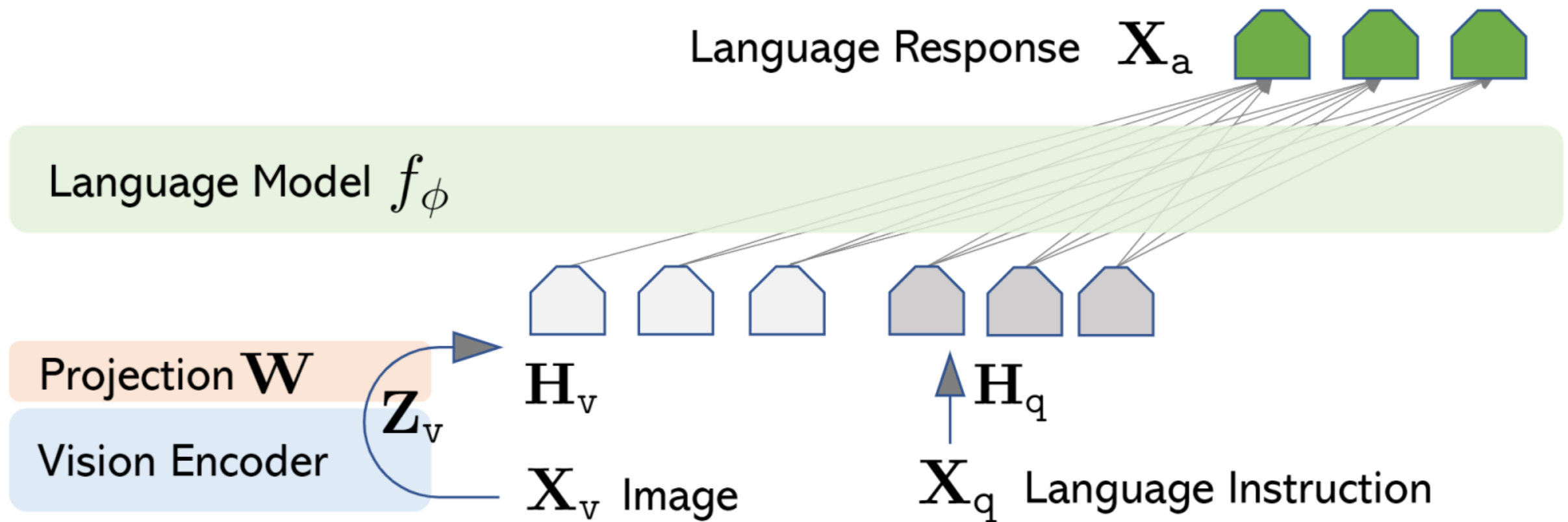
<https://arxiv.org/pdf/2103.00020>

BLIP2



<https://arxiv.org/pdf/2301.12597>

LLaVA



<https://llava-vl.github.io/>

LLaVA Instruction

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], 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: One example to illustrate the instruction-following data. The top block shows the contexts such as captions and boxes used to prompt GPT, and the bottom block shows the three types of responses. Note that the visual image is not used to prompt GPT, we only show it here as a reference.

<https://llava-vl.github.io/>

Image Tokens + Transformers

Is this magic?

An armchair in the shape of an avocado →



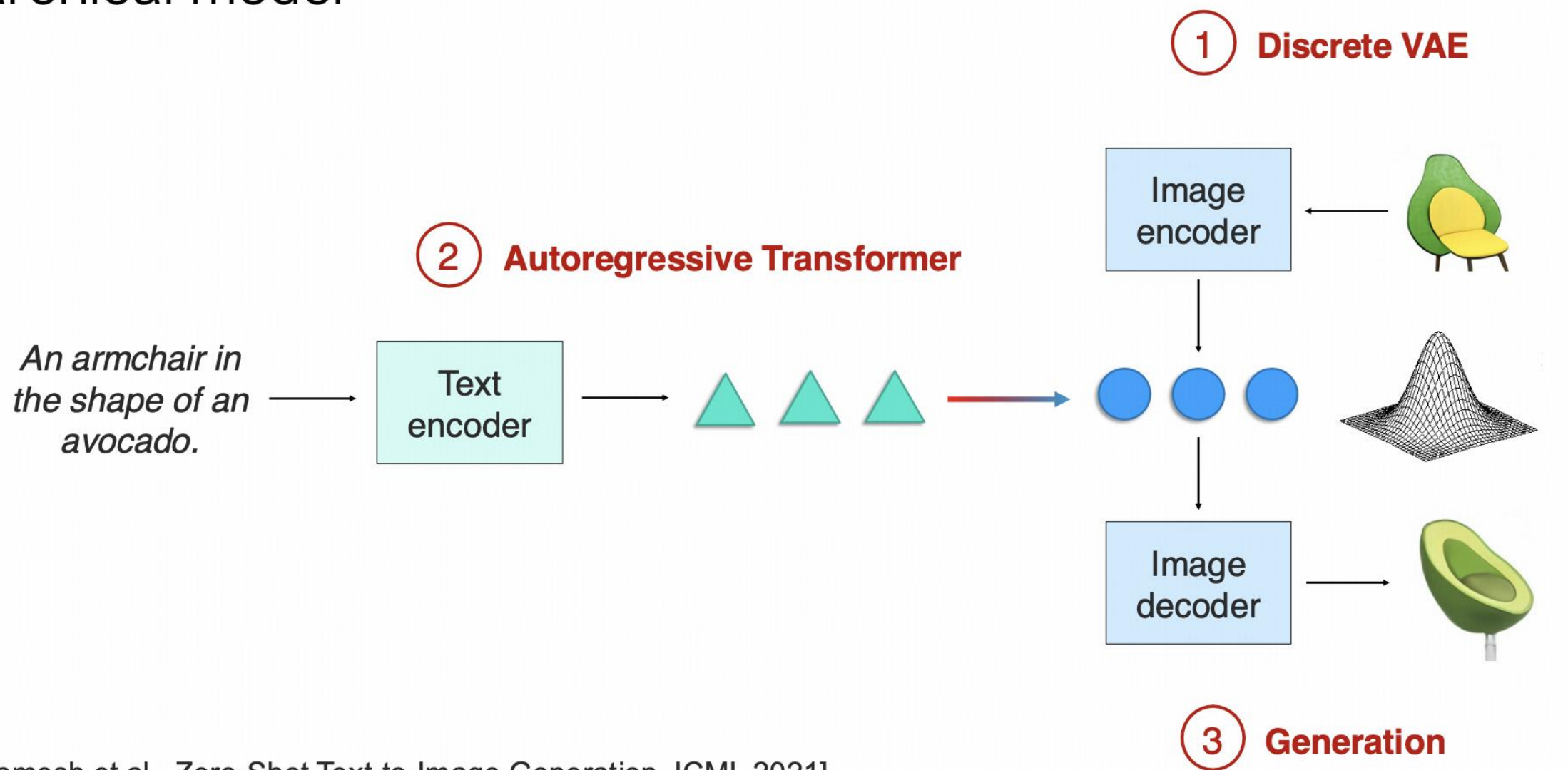
[DALL-E. Ramesh et al., Zero-Shot Text-to-Image Generation. ICML 2021]

[see also, Esser et al. Taming Transformers for High Resolution Image Synthesis. CVPR 2021]

11 / 11

Image Tokens + Transformers

Hierarchical model

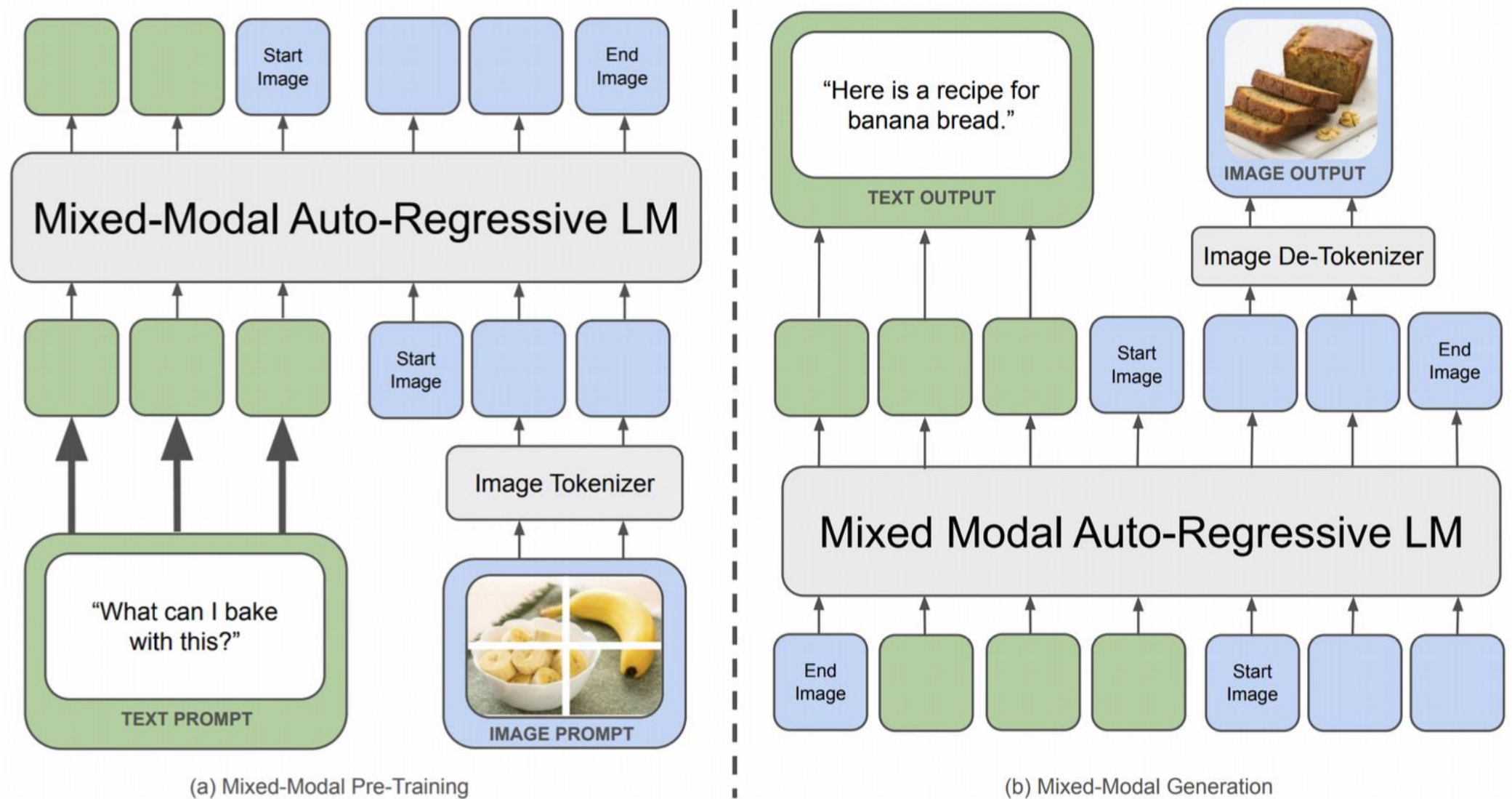


[DALL-E. Ramesh et al., Zero-Shot Text-to-Image Generation. ICML 2021]

[see also. Esser et al. Taming Transformers for High Resolution Image Synthesis. CVPR 2021]

CM3, CM3Leon, Chameleon

Train on interleaved images and text.



<https://arxiv.org/pdf/2405.09818>

Diffusion Models

Key insight: “shaped” Gaussian noise, applied repeatedly, can create images

Fix q , set p to be a U-Net that slightly de-noises the image

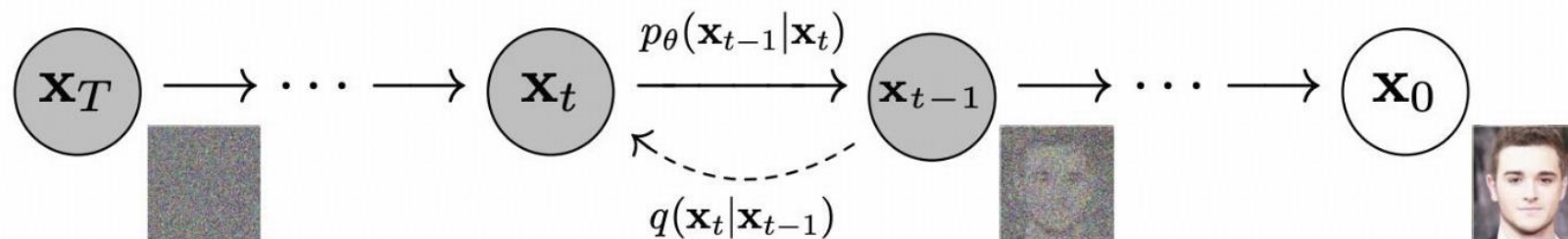


Figure 2: The directed graphical model considered in this work.

$$\mathbb{E}[-\log p_\theta(\mathbf{x}_0)] \leq \mathbb{E}_q \left[-\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \right] = \mathbb{E}_q \left[-\log p(\mathbf{x}_T) - \sum_{t \geq 1} \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})} \right] =: L$$

Our old friend
the ELBO
Easy to
sample
arbitrarily
high Ts
Gaussian
Noise!
Make this
easier with
Gaussian
tricks

<https://lilianweng.github.io/posts/2021-07-11-diffusion-models/>

More on CS 11-777: Multimodal Machine Learning

<https://cmu-mmml.github.io/>

Questions?