GPT Siblings & Application Research

ITM528 Deep Learning

Taemoon Jeong

GPT

"Improving Language Understanding by Generative Pre-Training," 2018

Generative Pre-Training (GPT)

The goal is to learn a **universal representation** that transfers with little adaptation to a wide range of tasks.

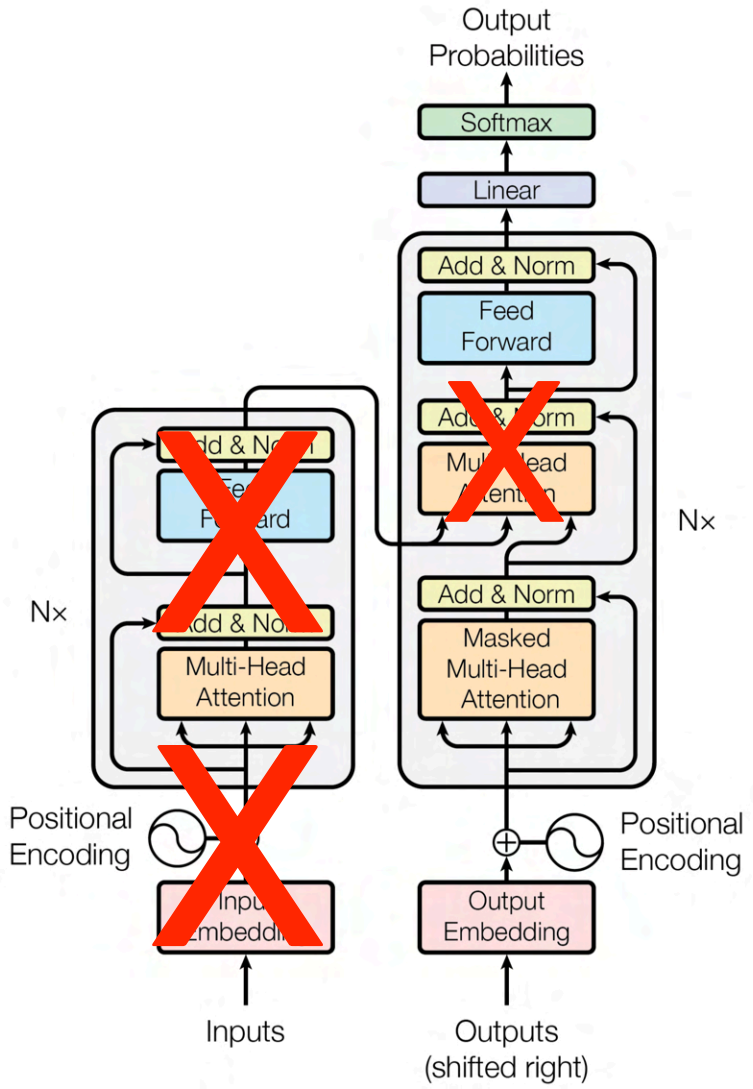
GPT

- GPT consists of two stages: the first stage is learning a high-capacity language model on a large corpus of text and followed by a fine-tuning stage.
- Unsupervised pre-training
  - Given an unsupervised corpus of tokens $\mathcal{U} = \{u_1, \ldots, u_n\}$, a standard language modeling objective is used to maximize the following likelihood:

$$L_1(\mathcal{U}) = \sum_i \log P(u_i \mid u_{i-k}, \ldots, u_{i=1}; \Theta)$$

    where k is the size of the context window, and the conditional probability P is modeled using a neural network with parameters Θ
- A multi-layer **Transformer decoder** is used.

Why Transformer Decoder?



Traversal-Style Approach

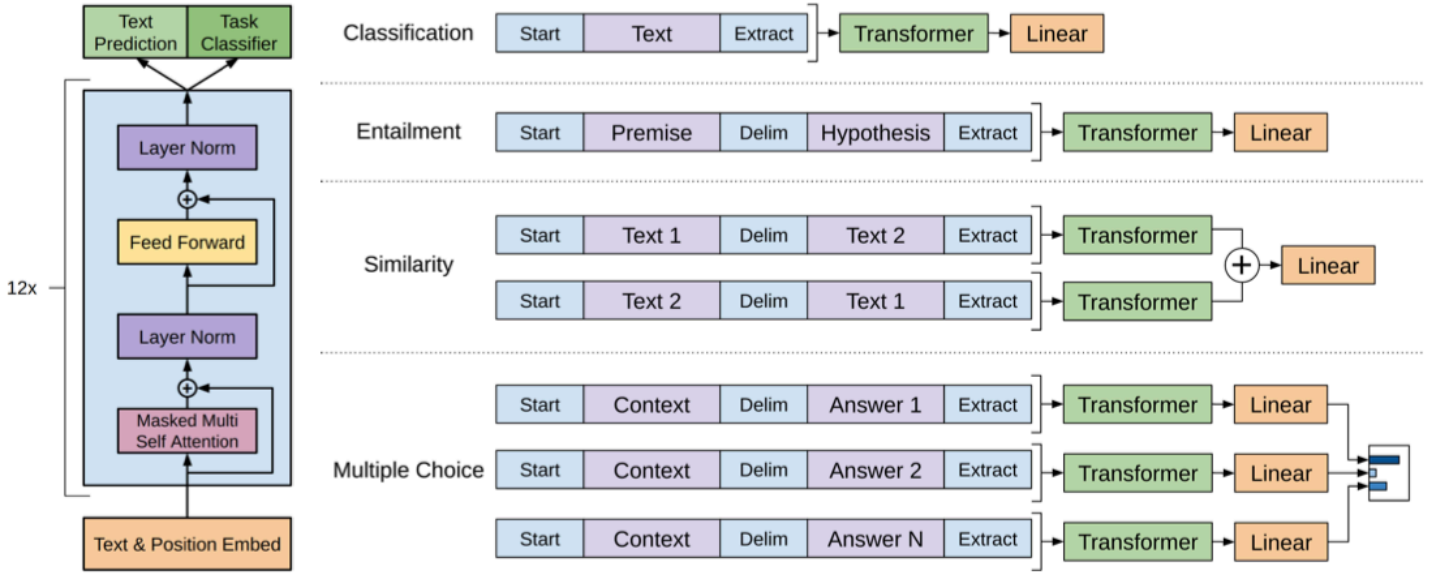


Figure 1: **(left)** Transformer architecture and training objectives used in this work. **(right)** Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.

I dislike old cabin cruisers. (negative)

Peter is snoring. $ A man sleeps. (true)

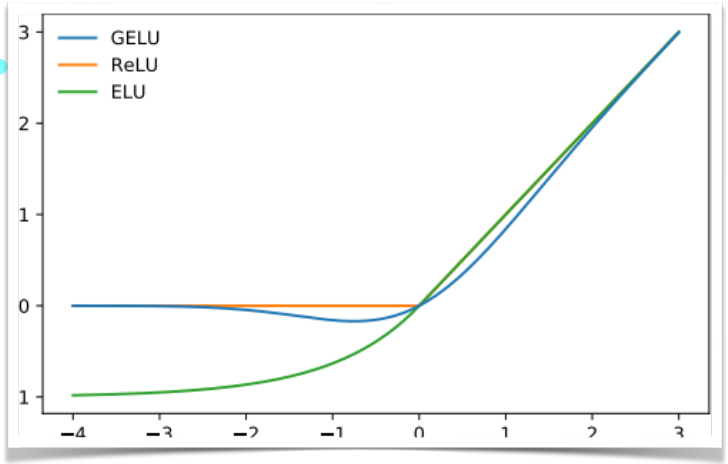
President greets the press in Chicago.
Obama speaks in Illinois.
(similar)

Dataset

- **BooksCorpus** dataset is used for pre-training
  - It contains over 7,000 unique unpublished books from a variety of genres including Adventure, Fantasy, and Romance.
  - In contains long stretches of contiguous text, which allows the generative model to learn to condition on long-range information.
- **1D Word Benchmark** is an alternative option
  - It has approximately the same size as BooksCorpus but is shuffled at a sentence level, destroying long-range structure.
  - GPT achieves a very low performance on this dataset.

Implementation Details

- 12-layer-decoder-only transformer with masked self-attention heads
  - 768 dimensional states and 12 attention heads
  - For the position-wise-feed-forward networks, 3072 dimensional inner states are used.
  - Adam optimizer with a maximum learning rate of $2.5 * 10^{-4}$ with a cosine schedule
  - Gaussian Error Linear Unit (GELU) activations
  - Learned position embeddings
- Fine-tuning details
  - Dropout to the classifier with a rate of 0.1
  - Learning rate of $6.25 * 10^{-5}$ with a linear decay and a batchsize of 32
  - 3 epochs of training was sufficient for most cases



Experiments

Table 2: Experimental results on natural language inference tasks, comparing our model with current state-of-the-art methods. 5x indicates an ensemble of 5 models. All datasets use accuracy as the evaluation metric.

Method	MNLI-m	MNLI-mm	SNLI	SciTail	QNLI	RTE
ESIM + ELMo [44] (5x)	-	-	89.3	-	-	-
CAFE [58] (5x)	80.2	79.0	89.3	-	-	-
Stochastic Answer Network [35] (3x)	80.6	80.1	-	-	-	-
CAFE [58]	78.7	77.9	88.5	83.3		
GenSen [64]	71.4	71.3	-	-	82.3	59.2
Multi-task BiLSTM + Attn [64]	72.2	72.1	-	-	82.1	**61.7**
Finetuned Transformer LM (ours)	**82.1**	**81.4**	**89.9**	**88.3**	**88.1**	56.0

Natural language inference (NLI): reading a pair of sentences and judging the relationship between from one of entailment, contradiction, or neutral

Experiments

Table 3: Results on question answering and commonsense reasoning, comparing our model with current state-of-the-art methods.. 9x means an ensemble of 9 models.

Method	Story Cloze	RACE-m	RACE-h	RACE
val-LS-skip [55]	76.5	-	-	-
Hidden Coherence Model [7]	77.6	-	-	-
Dynamic Fusion Net [67] (9x)	-	55.6	49.4	51.2
BiAttention MRU [59] (9x)	-	60.2	50.3	53.3
Finetuned Transformer LM (ours)	**86.5**	**62.9**	**57.4**	**59.0**

Question answering and **commonsense reasoning**: English passages with associated questions from middle and high school exams / selecting the correct ending to multi-sentence stories from two options.

Experiments

Table 4: Semantic similarity and classification results, comparing our model with current state-of-the-art methods. All task evaluations in this table were done using the GLUE benchmark. (*mc*= Mathews correlation, *acc*=Accuracy, *pc*=Pearson correlation)

Method	Classification		Semantic Similarity			GLUE
	CoLA (mc)	SST2 (acc)	MRPC (F1)	STSB (pc)	QQP (F1)	
Sparse byte mLSTM [16]	-	**93.2**	-	-	-	-
TF-KLD [23]	-	-	**86.0**	-	-	-
ECNU (mixed ensemble) [60]	-	-	-	81.0	-	-
Single-task BiLSTM + ELMo + Attn [64]	35.0	90.2	80.2	55.5	66.1	64.8
Multi-task BiLSTM + ELMo + Attn [64]	18.9	91.6	83.5	72.8	63.3	68.9
Finetuned Transformer LM (ours)	**45.4**	91.3	82.3	**82.0**	**70.3**	**72.8**

Semantic Similarity: Predicting whether two sentences are semantically equivalent or not.

GPT-2

"Language Models are Unsupervised Multitask Learners," 2018

GPT-2

GPT-2 is a 1.5B parameter Transformer that achieves state of the art results on 7 out of 8 tested language modeling datasets in a **zero-shot setting**.

Language-based Task Specification

- Language Prompt
  - Language provides a flexible way to specify **tasks**, **inputs**, and **outputs** all as a sequence of symbols.
  - Translation tasks
    - (translate to french, english text, french text)
  - Reading comprehension tasks
    - (answer the question, document, question, answer)

GPT 모델은 바뀌지 않고 prompt를 이해하는 방식이 바뀐다.

WebText

new data set

- OpenAI created a new web scrape which emphasizes document quality.
  - Web pages which have been curated/filtered by humans are only used.
  - First, all outbound links from Reddit which received at least 3 karma are scraped.
  - The resulting **WebText** contains the text subset of these 45 million links.
    - It contains slightly over 8 million documents for a total of 40 GB of text.
    - All Wikipedia documents are removed.

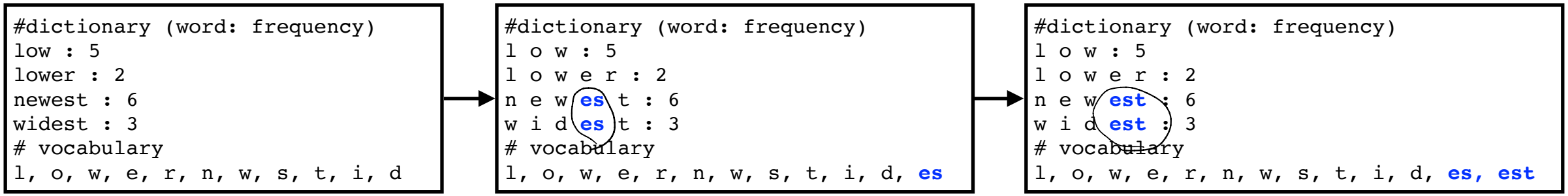
WebText

”I’m not the cleverest man in the world, but like they say in French: **Je ne suis pas un imbecile [I’m not a fool].**
In a now-deleted post from Aug. 16, Soheil Eid, Tory candidate in the riding of Joliette, wrote in French: ”**Mentez mentez, il en restera toujours quelque chose**,” which translates as, ”**Lie lie and something will always remain.**”
“I hate the word ‘**perfume**,”’ Burr says. ‘It’s somewhat better in French: ‘**parfum**.’
If listened carefully at 29:55, a conversation can be heard between two guys in French: “**-Comment on fait pour aller de l’autre coté? -Quel autre coté?**”, which means “**- How do you get to the other side? - What side?**”.
If this sounds like a bit of a stretch, consider this question in French: **As-tu aller au cinéma?**, or **Did you go to the movies?**, which literally translates as Have-you to go to movies/theater?
“**Brevet Sans Garantie Du Gouvernement**”, translated to English: “**Patented without government warranty**”.

Table 1. Examples of naturally occurring demonstrations of English to French and French to English translation found throughout the WebText training set.

Byte Pair Encoding

- Word-level encoding
  - Construct a fixed-size dictionary based on word frequencies
  - Out-of-vocabulary (OOV) problem occurs
- Character-level encoding
  - Construct a dictionary with characters (e.g., lower-case Alphabet: 26, Korean: 11,172)
- Byte Pair Encoding (BPE)
  - BPE is a practical middle ground between character and word level language modeling
  - It is based on subword segmentation



https://wikidocs.net/22592

Implementation Details

- Transformer-based architecture following the OpenAI GPT model
  - Layer normalization is moved to the input of each sub-block and an additional layer normalization was added after the final self-attention block
  - The vocabulary is expanded to 50,257. The context size is increased from 512 to 1,024 tokens and a batchsize is 512.

Unit created by byte-pair encoding

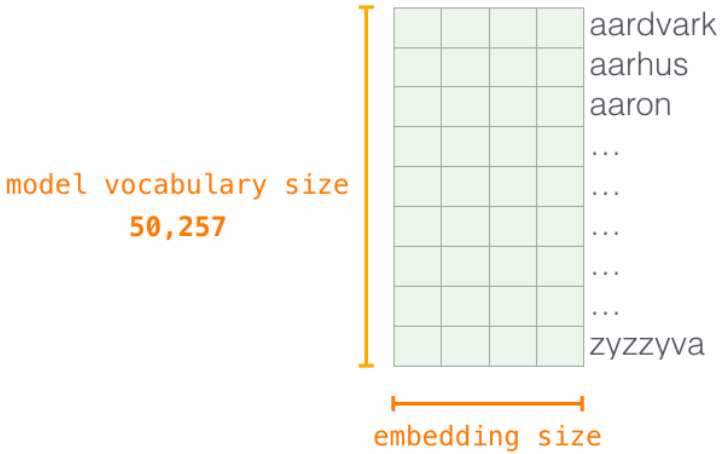
until believable

Parameters	Layers	d_{model}	
117M	12	768	← similar to the original GPT
345M	24	1024	← similar to the BERT
762M	36	1280	
1542M	48	1600	← this one is called GPT-2

Table 2. Architecture hyperparameters for the 4 model sizes.

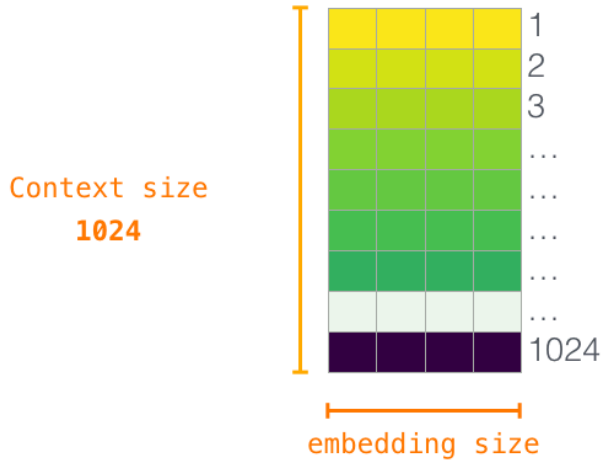
Implementation Details

Token Embeddings (wte)



768 (small) / 1024 (medium) / 1280 (large) / 1600 (extra large)

Positional Encodings (wpe)



768 (small) / 1024 (medium) / 1280 (large) / 1600 (extra large)

Experiments

Language Models are Unsupervised Multitask Learners

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	**21.8**
117M	**35.13**	45.99	**87.65**	**83.4**	**29.41**	65.85	1.16	1.17	37.50	75.20
345M	**15.60**	55.48	**92.35**	**87.1**	**22.76**	47.33	1.01	**1.06**	26.37	55.72
762M	**10.87**	**60.12**	**93.45**	**88.0**	**19.93**	**40.31**	**0.97**	**1.02**	22.05	44.575
1542M	**8.63**	**63.24**	**93.30**	**89.05**	**18.34**	**35.76**	**0.93**	**0.98**	**17.48**	42.16

Table 3. Zero-shot results on many datasets. No training or fine-tuning was performed for any of these results. PTB and WikiText-2 results are from (Gong et al., 2018). CBT results are from (Bajgar et al., 2016). LAMBADA accuracy result is from (Hoang et al., 2018) and LAMBADA perplexity result is from (Grave et al., 2016). Other results are from (Dai et al., 2019).

Experiments

Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	76.8%
Who came up with the theory of relativity?	Albert Einstein	✓	76.4%
When was the first star wars film released?	1977	✓	71.4%
What is the most common blood type in sweden?	A	✗	70.6%
Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓	69.3%
Who took the first steps on the moon in 1969?	Neil Armstrong	✓	66.8%
Who is the largest supermarket chain in the uk?	Tesco	✓	65.3%
What is the meaning of shalom in english?	peace	✓	64.0%
Who was the author of the art of war?	Sun Tzu	✓	59.6%
Largest state in the us by land mass?	California	✗	59.2%
Green algae is an example of which type of reproduction?	parthenogenesis	✗	56.5%
Vikram samvat calender is official in which country?	India	✓	55.6%
Who is mostly responsible for writing the declaration of independence?	Thomas Jefferson	✓	53.3%
What us state forms the western boundary of montana?	Montana	✗	52.3%
Who plays ser davos in game of thrones?	Peter Dinklage	✗	52.1%
Who appoints the chair of the federal reserve system?	Janet Yellen	✗	51.5%
State the process that divides one nucleus into two genetically identical nuclei?	mitosis	✓	50.7%
Who won the most mvp awards in the nba?	Michael Jordan	✗	50.2%
What river is associated with the city of rome?	the Tiber	✓	48.6%
Who is the first president to be impeached?	Andrew Johnson	✓	48.3%
Who is the head of the department of homeland security 2017?	John Kelly	✓	47.0%
What is the name given to the common currency to the european union?	Euro	✓	46.8%
What was the emperor name in star wars?	Palpatine	✓	46.5%
Do you have to have a gun permit to shoot at a range?	No	✓	46.4%
Who proposed evolution in 1859 as the basis of biological development?	Charles Darwin	✓	45.7%
Nuclear power plant that blew up in russia?	Chernobyl	✓	45.7%
Who played john connor in the original terminator?	Arnold Schwarzenegger	✗	45.2%

Table 5. The 30 most confident answers generated by GPT-2 on the development set of Natural Questions sorted by their probability according to GPT-2. None of these questions appear in WebText according to the procedure described in Section 4.

GPT-2 answers 4.1% of questions correctly, which is much worse than 30-50% SOTA results.

BERT

"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2019

BERT

Bidirectional Encoder Representation from Transformers (BERT)

왼쪽과 오른쪽의 문맥

BERT

I love hiking

- **OpenAI GPT** uses left-to-right architecture, where every token can only attend to **previous** tokens in the self-attention layers of the Transformer.
- **BERT** alleviates the previously mentioned unidirectionality constraint by using a "**masked language model**" (MLN) pre-training objective.
  - The **masked language model** randomly masks some of the tokens from the input, and the objective is to predict the original vocabulary id of the masked word based only on its context.
  - In addition to the masked language model, a "**next sentence prediction**" task is also used that jointly pre-trains text-pair representations.

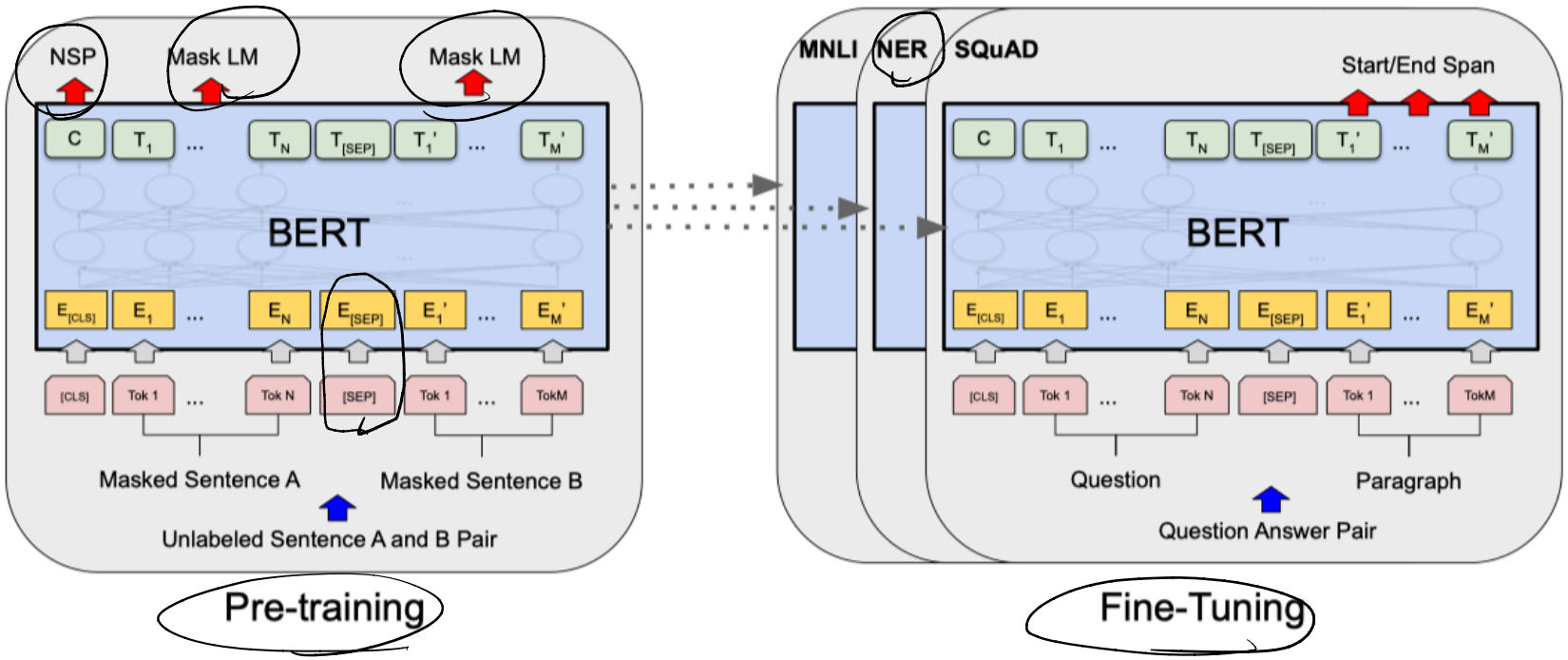
NSP

the [MASK] is shiny

↳ SUN

BERT

[MASK] token



MLM

NSP

Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. [CLS] is a special symbol added in front of every input example, and [SEP] is a special separator token (e.g. separating questions/answers).

Implementation Details

- BERT's model architecture is a multi-layer bidirectional **Transformer encoder**.
- Suppose that L is the number of layers, H is the hidden size, and A is the number of self-attention heads:
  - **BERT$_{BASE}$**: $(L = 12, H = 786, A = 12)$ #param: 110M
  - **BERT$_{LARGE}$**: $(L = 24, H = 1024, A = 16)$ #param: 340M
  - **GPT-1**: 125M, **GPT-2**: 1.5B, **GPT-3**: 175B
- Input/Output Representation
  - Both a single sentence and a pair of sentences (e.g., question and answer pairs) can be used as an input to BERT.
    - Sentence pairs are packed and they are separated with a special token, [SEP]. A learned embedding to every token is added to indicate whether it belongs to sentence A or sentence B.
  - WordPiece embeddings with a 30,000 token vocabulary are used.
  - The first token is always a special classification token, [CLS]. The final hidden state corresponding to this token is used as the aggregate sequence representation for classification tasks.

Input Representation

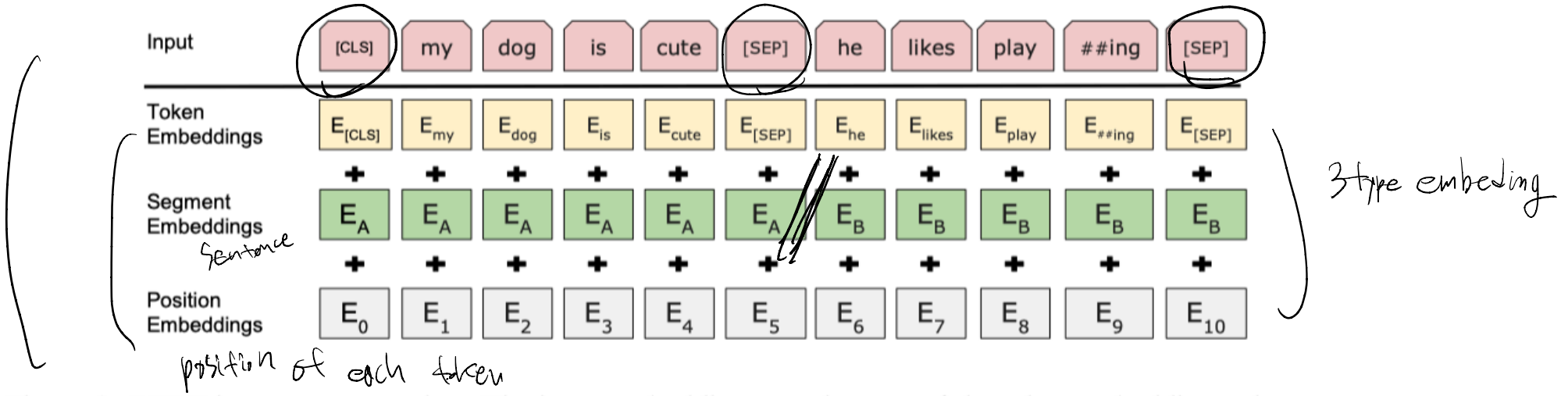


Figure 2: BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings.

Pre-training BERT

- BERT is pre-trained with two unsupervised tasks: **masked language model** and **next sentence prediction**.
- Task 1: Masked LM
  - The masked words (15%) are only predicted rather than the entire input.
  - One downside is that it creates a mismatch between pre-training and fine-tuning phases, since the [MASK] token does not appear during fine-tuning.
  - To mitigate this issue, if the i-th token is chosen, we replace the i-th token with (1) the [MASK] token 80% of the time (2) a random token 10% of the time (3) the unchanged i-th token 10% of the time.
- Task 2: Next Sentence Prediction (NSP)
  - A binarized next sentence prediction tasks is used.
  - When choosing two sentences A and B, 50% of the time B is the actual next sentence that follows A and 50% of the time it is a random sentence from the corpus.

Experiments

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

Table 1: GLUE Test results, scored by the evaluation server (https://gluebenchmark.com/leaderboard). The number below each task denotes the number of training examples. The "Average" column is slightly different than the official GLUE score, since we exclude the problematic WNLI set.[8] BERT and OpenAI GPT are single-model, single task. F1 scores are reported for QQP and MRPC, Spearman correlations are reported for STS-B, and accuracy scores are reported for the other tasks. We exclude entries that use BERT as one of their components.

GLUE

- General Language Understanding Evaluation (**GLUE**)
  - **CoLA**: The Corpus of Linguistic Acceptability: Binary classification: single sentences that are either grammatical or ungrammatical
  - **SST-2**: Stanford Sentiment Treebank:phrases culled from movie reviews scored on their positivity/ negativity. Phrases can be positive, negative, or completely neutral
  - **STS-B**: The Semantic Textual Similarity Benchmark: task of determining the similarity on a continuous scale from 1 to 5 of a pair of sentences drawn from various sources
  - **QQP**: The Quora Question Pairs dataset: collection of question pairs from the community question-answering website Quora: Given two questions, the task is to determine whether they are semantically equivalent

  - https://docs.google.com/spreadsheets/d/1BrOdjJgky7FfeiwC_VDURZuRPUFUAz_jfczPPT35P00/edit#gid=0

RoBERTa

"RoBERTa: A Robustly Optimized BERT Pretraining Approach," 2019

RoBERTa

Robustly optimized BERT approach (RoBERTa)

RoBERTa

- This paper presents an improved recipe for training BERT models, named RoBERTa.
  - training the model longer, with bigger batches, over more data
  - removing the next sentence prediction objective
  - training on longer sequences
  - dynamically changing the masking pattern applied to the training data
- A large new dataset (CC-News) is collected.

Dataset

- BERT-style pre-training crucially relies on large quantities of text. Five English-language corpora of varying sizes and domains, totaling over 160GB of uncompressed text, are used.
  - **BookCorpus** plus **English Wikipedia**: the original data used to train BERT (16GB)
  - **CC-News**: English portion of the CommonCrawl News dataset. The data contains 63 million English news articles crawled between Sep. 2016 and Feb. 2019 (76GB after filtering)
  - **OpenWebText**: The text is web content extracted from URLs shared on Reddit with at least three upvotes (38GB), Recreation of WebText corpora for training GPT-2
  - **Stories**: A subset of CommonCrawl filtered to match the story-like style of Winograd schemas (31GB)

Implementation Details

- **BERT$_{BASE}$** ($L = 12, H = 768, A = 12, 110M$ params) configuration
  - **Static vs. Dynamic Masking**
    - The original BERT implementation performed masking once during data preprocessing, resulting in a single static mask.
    - Dynamic masking generates the masking pattern every time a sequence is fed to the model.

Masking	SQuAD 2.0	MNLI-m	SST-2
reference	76.3	84.3	92.8
Our reimplementation:			
static	78.3	84.3	92.5
dynamic	78.7	84.0	92.9

Table 1: Comparison between static and dynamic masking for BERT$_{BASE}$. We report F1 for SQuAD and accuracy for MNLI-m and SST-2. Reported results are medians over 5 random initializations (seeds). Reference results are from Yang et al. (2019).

Implementation Details

- **$\text{BERT}_{\text{BASE}}$** ($L = 12, H = 768, A = 12, 110M$ params) configuration
  - **Nest Sentence Prediction (NSP)** BERT에서는 효과적임.
    - The original BERT model emphasized the NSP objectives.
    - However, some recent work questioned the necessity of the NSP loss. NSP를 제거하면 성능이 올라간 것이라는 최근 연구들

Model	SQuAD 1.1/2.0	MNLI-m	SST-2	RACE
Our reimplementation (with NSP loss):				
SEGMENT-PAIR	90.4/78.7	84.0	92.9	64.2
SENTENCE-PAIR	88.7/76.2	82.9	92.1	63.0
Our reimplementation (without NSP loss):				
FULL-SENTENCES	90.4/79.1	84.7	92.5	64.8
DOC-SENTENCES	90.6/79.7	84.7	92.7	65.6
$\text{BERT}_{\text{BASE}}$	88.5/76.3	84.3	92.8	64.3
$\text{XLNet}_{\text{BASE}}$ (K = 7)	–/81.3	85.8	92.7	66.1
$\text{XLNet}_{\text{BASE}}$ (K = 6)	–/81.0	85.6	93.4	66.7

Table 2: Development set results for base models pretrained over BOOKCORPUS and WIKIPEDIA. All models are trained for 1M steps with a batch size of 256 sequences. We report F1 for SQuAD and accuracy for MNLI-m, SST-2 and RACE. Reported results are medians over five random initializations (seeds). Results for $\text{BERT}_{\text{BASE}}$ and $\text{XLNet}_{\text{BASE}}$ are from Yang et al. (2019).

Implementation Details

- **BERT$_{BASE}$** ($L = 12, H = 768, A = 12, 110M$ params) configuration
  - Training with large batches
    - The original BERT trained **BERT$_{BASE}$** with 1M steps with a batch size of 256 sequences.
    - With the same computational cost, we can increase the batch size while reducing the steps.

bsz	steps	lr	ppl	MNLI-m	SST-2
256	1M	1e-4	3.99	84.7	92.7
2K	125K	7e-4	**3.68**	**85.2**	**92.9**
8K	31K	1e-3	3.77	84.6	92.8

Table 3: Perplexity on held-out training data (*ppl*) and development set accuracy for base models trained over BOOKCORPUS and WIKIPEDIA with varying batch sizes (*bsz*). We tune the learning rate (*lr*) for each setting. Models make the same number of passes over the data (epochs) and have the same computational cost.

ALBERT

"ALBERT: A Lite BERT for Self-supervised Learning of Language Representations," 2020

ALBERT

A Lite BERT (ALBERT)

ALBERT

- ALBERT presents two parameter-reduction techniques to lower memory consumption and increase the training speed of BERT.
  - ① The first one is a **factorized embedding parametrization** by decomposing the large vocabulary embedding matrix into two small matrices.
  - ② The second technique is **cross-layer parameter sharing** which prevents the parameter from growing with the depth of the network.
- A self-supervised loss focussing on modeling inter-sentence coherence is used.
  - A **sentence-order prediction** (SOP) focuses on inter-sentence coherence and is designed to address the ineffectiveness of the next sentence prediction (NSP) loss proposed in the original BERT.

ALBERT

- Factorized embedding parametrization
  - The size of the embedding matrix has the size of $V \times H$ where V is the vocabulary size and H is the size of hidden space.
  - By projecting them into a lower dimensional embedding space of size E, the embedding parameters are reduced from $O(V \times H)$ to $O(V \times E + E \times H)$. This parameter reduction is significant when $H \gg E$.
- Cross-layer parameter sharing
  - There are multiple ways to share parameters (e.g., only sharing feed-forward network parameters across layers, or only sharing attention parameters, or all parameters which is the default decision for ALBERT).
- Inter-sentence coherence loss
  - The sentence-order-prediction (SOP) task uses as positive examples the same as BERT and as negative examples the same two consecutive segments but with their order swapped.

Model

- ALBERT-large has about 18x fewer parameters compare to BERT-large, 18M versus 334M. An ALBERT-xlarge configuration with $H = 2048$ has $233M$ parameters (i.e., around 70% of BEI

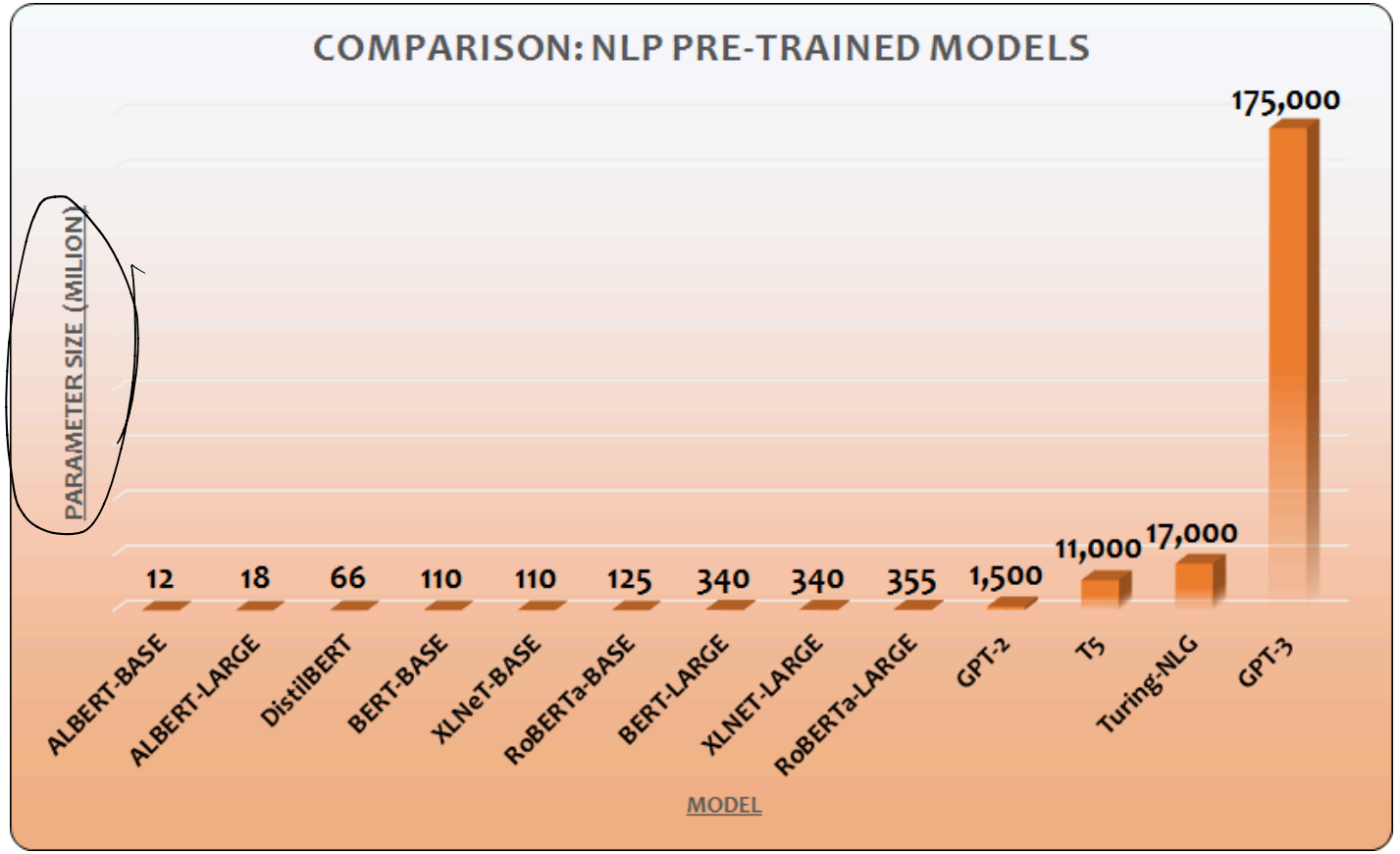
Model		Parameters	Layers	Hidden	Embedding	Parameter-sharing
BERT	base	108M	12	768	768	False
	large	334M	24	1024	1024	False
ALBERT	base	12M	12	768	128	True
	large	18M	24	1024	128	True
	xlarge	60M	24	2048	128	True
	xxlarge	235M	12	4096	128	True

Table 1: The configurations of the main BERT and ALBERT models analyzed in this paper.

GPT-3

"Language Models are Few-Shot Learners," 2020

GPT-3



https://medium.com/analytics-vidhya/openai-gpt-3-language-models-are-few-shot-learners-82531b3d3122

GPT-3

What is new besides the model size has increased over 10 times?

Dataset

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Table 2.2: Datasets used to train GPT-3. "Weight in training mix" refers to the fraction of examples during training that are drawn from a given dataset, which we intentionally do not make proportional to the size of the dataset. As a result, when we train for 300 billion tokens, some datasets are seen up to 3.4 times during training while other datasets are seen less than once.

Dataset

- Three different techniques are used (1) **filtering CommonCrawl**, (2) fuzzy deduplication, (3) added high-quality reference corpora to the training mix to augment CommonCrawl
  - An **automatic filtering method** is developed to remove low quality documents. A classifier is trained by using the original WebText as a proxy for high-quality documents to distinguish low quality ones from raw CommonCrawl.
  - The classifier (logistic regression) is used to re-sample CommonCrawl by prioritizing documents which were predicted by the classifier to be higher quality. The positive examples include curated datasets such as **WebText**, **Wikipedia**, and **web books corpus**, and the negative examples include **unfiltered Common Crawl**.

Dataset

- Three different techniques are used (1) filtering CommonCrawl, (2) **fuzzy deduplication**, (3) added high-quality reference corpora to the training mix to augment CommonCrawl
  - To improve the model quality and prevent overfitting, the authors fuzzily deduplicated documents (i.e., removed documents with high overlap with other documents). WebText was fuzzily removed from Common Crawl.
  - Overall, this decreases dataset size by an average of 10%.

In-context Learning

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  cheese =>                           ← prompt
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← example
3  cheese =>                           ← prompt
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée
4  plush girafe => girafe peluche
5  cheese =>                           ← prompt
```

Implementation Details

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

Table 2.1: Sizes, architectures, and learning hyper-parameters (batch size in tokens and learning rate) of the models which we trained. All models were trained for a total of 300 billion tokens.

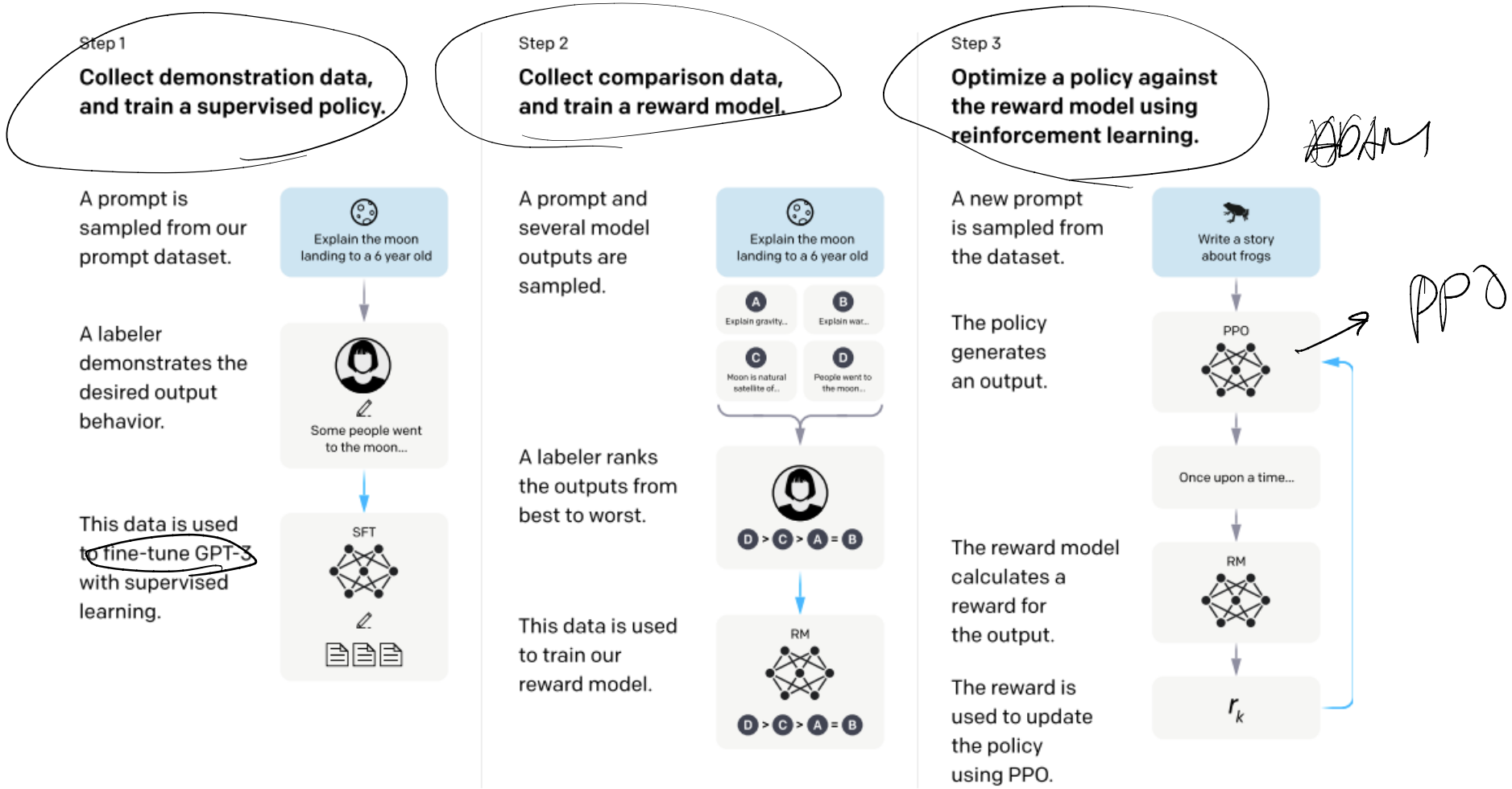
InstructGPT

"Training language models to follow instructions with human feedback," 2022

Alignment

"We make progress on **aligning language models** by training them to act in accordance with the user's intention"

High-level Methodology



InstructGPT

- InstructGPT starts with the GPT-3 pre-trained language model.
  - Supervised fine-tuning (**SFT**)
    - The prompt dataset consists of text prompts submitted to the OpenAI API.
    - Human labeler demonstrations are used to fine-tune GPT-3 to get **SFT models**.
  - Reward modeling (**RM**)
    - The comparison dataset is collected by comparing two different outputs of models.
    - The **reward function** is trained to predict human preferences.
  - Reinforcement Learning from Human Feedback (RLHF)
    - **SFT models** are fine-tuned using the **reward function**.
      - In particular, the model (**InstructGPT**) completes the given prompt, and the **reward model** outputs the estimated reward.
      - The model is fine-tuned to maximize the estimated reward.

Results

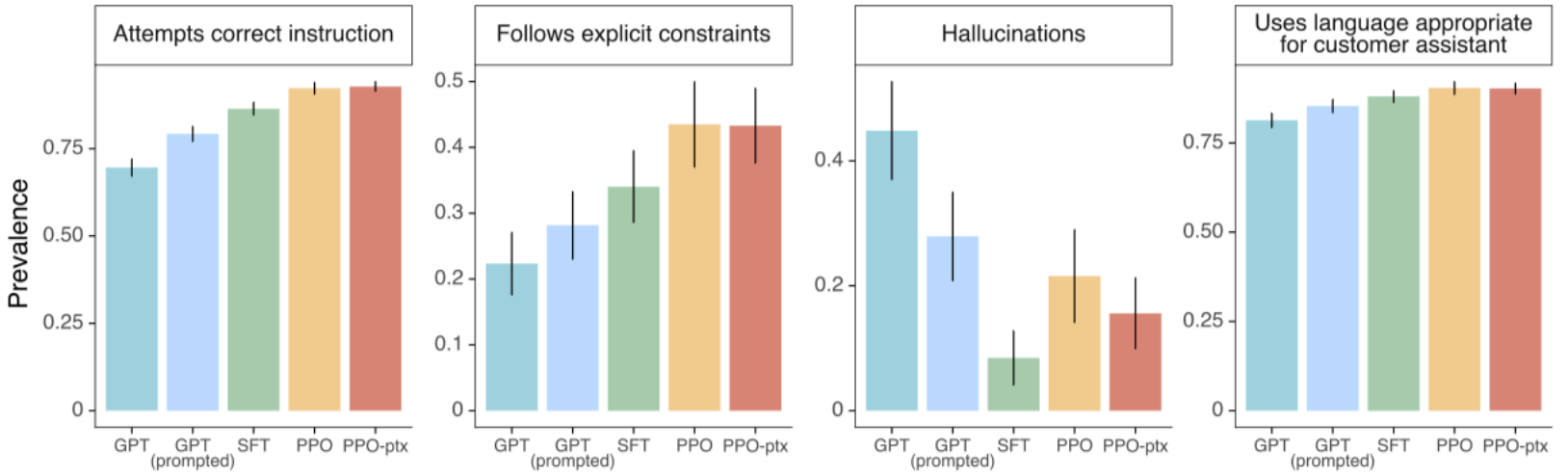
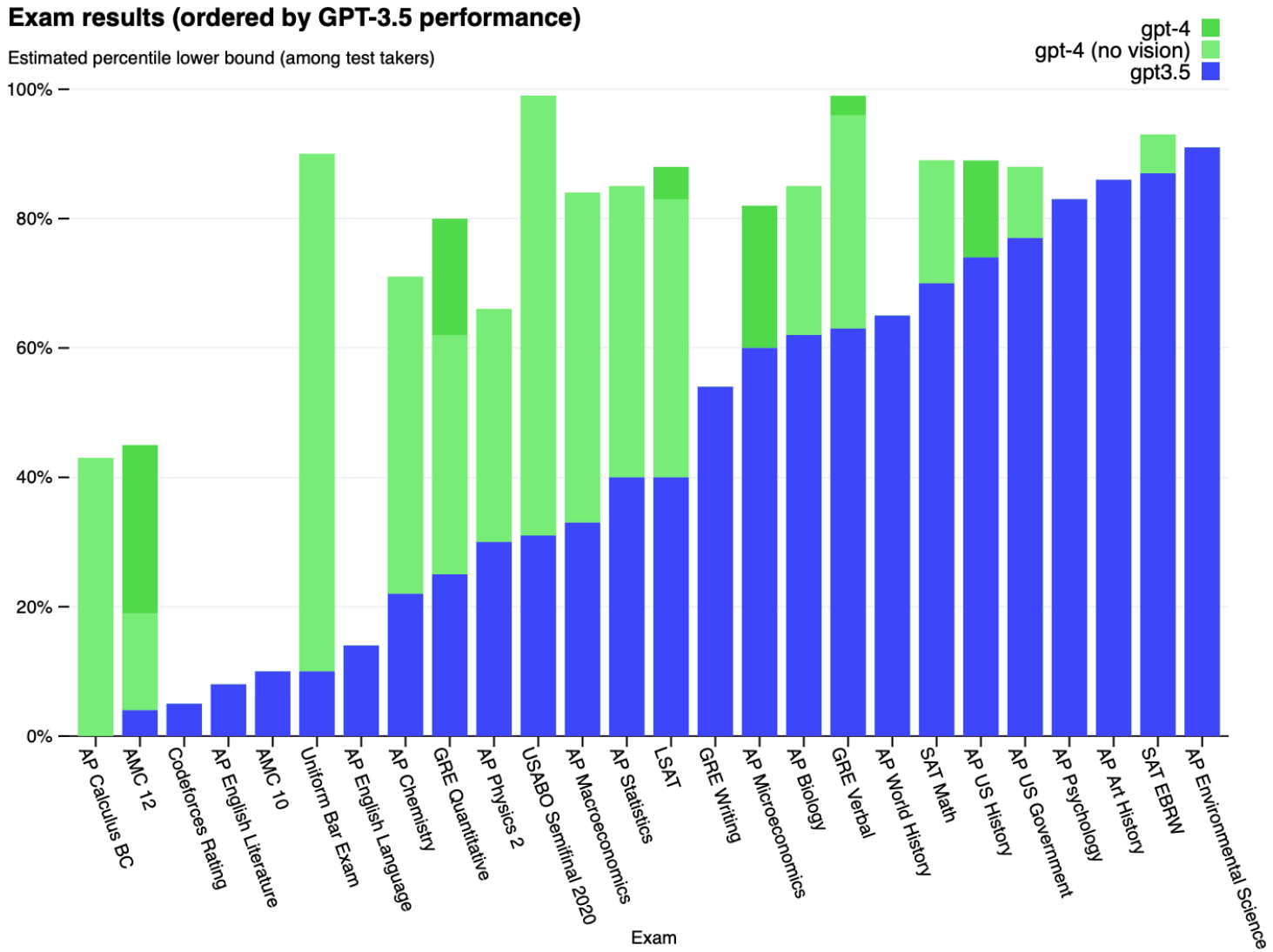


Figure 4: Metadata results on the API distribution. Note that, due to dataset sizes, these results are collapsed across model sizes. See Appendix E.2 for analysis that includes model size. Compared to GPT-3, the PPO models are more appropriate in the context of a customer assistant, are better at following explicit constraints in the instruction and attempting the correct instruction, and less likely to 'hallucinate' (meaning, making up information on closed domain tasks like summarization).

GPT-4

https://openai.com/research/gpt-4

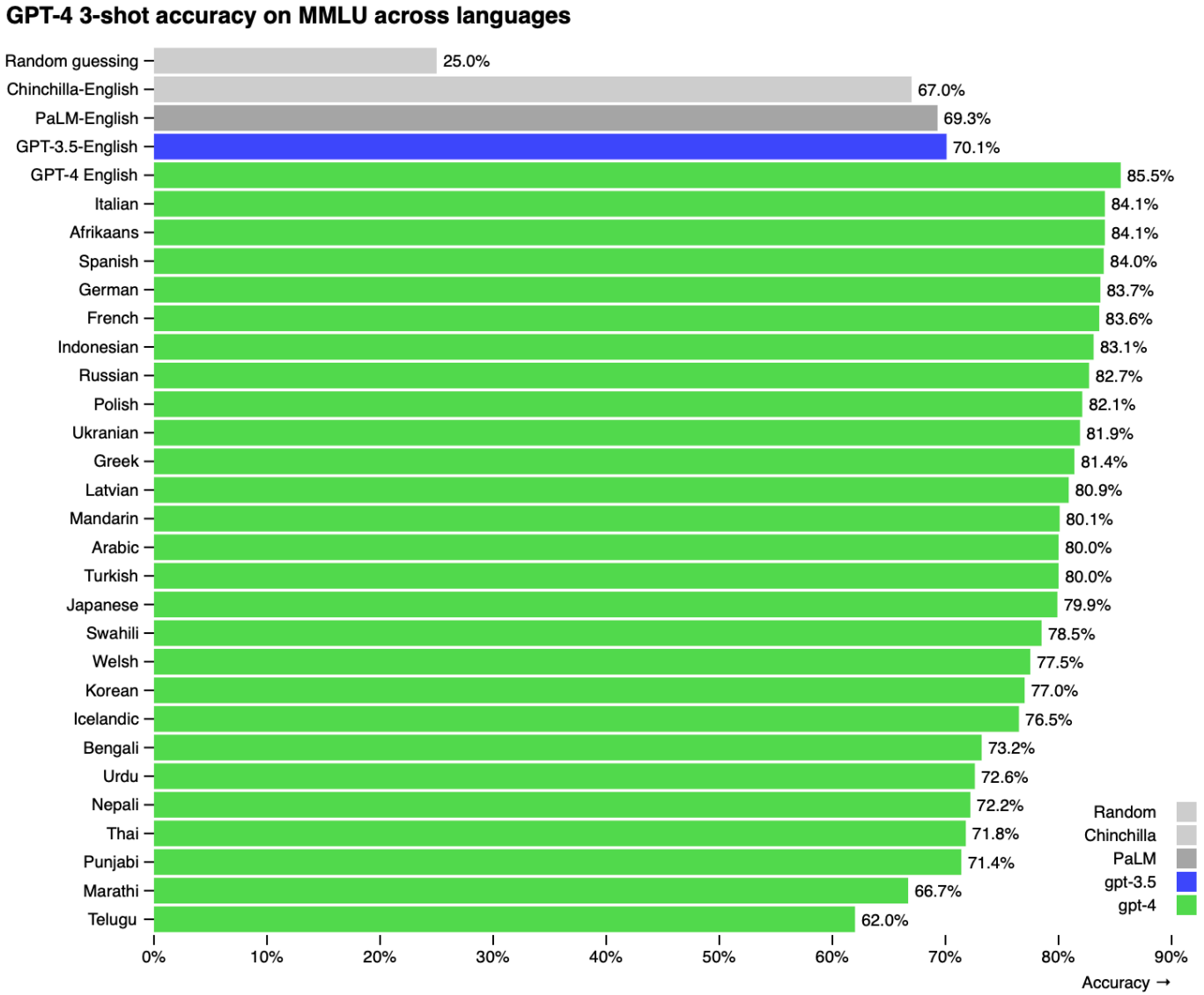
Exam Results



Traditional ML Benchmarks

Benchmark	GPT-4 Evaluated few-shot	GPT-3.5 Evaluated few-shot	LM SOTA Best external LM evaluated few-shot	SOTA Best external model (includes benchmark-specific training)
MMLU Multiple-choice questions in 57 subjects (professional & academic)	86.4% 5-shot	70.0% 5-shot	70.7% 5-shot U-PaLM	75.2% 5-shot Flan-PaLM
HellaSwag Commonsense reasoning around everyday events	95.3% 10-shot	85.5% 10-shot	84.2% LLAMA (validation set)	85.6% ALUM
AI2 Reasoning Challenge (ARC) Grade-school multiple choice science questions. Challenge-set.	96.3% 25-shot	85.2% 25-shot	84.2% 8-shot PaLM	85.6% ST-MOE
WinoGrande Commonsense reasoning around pronoun resolution	87.5% 5-shot	81.6% 5-shot	84.2% 5-shot PALM	85.6% 5-shot PALM
HumanEval Python coding tasks	67.0% 0-shot	48.1% 0-shot	26.2% 0-shot PaLM	65.8% CodeT + GPT-3.5
DROP (f1 score) Reading comprehension & arithmetic.	80.9 3-shot	64.1 3-shot	70.8 1-shot PaLM	88.4 QDGAT

Across Languages



Visual Inputs

User What is funny about this image? Describe it panel by panel.

Source: hmmm (Reddit)

GPT-4 The image shows a package for a "Lightning Cable" adapter with three panels.

Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.

Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it.

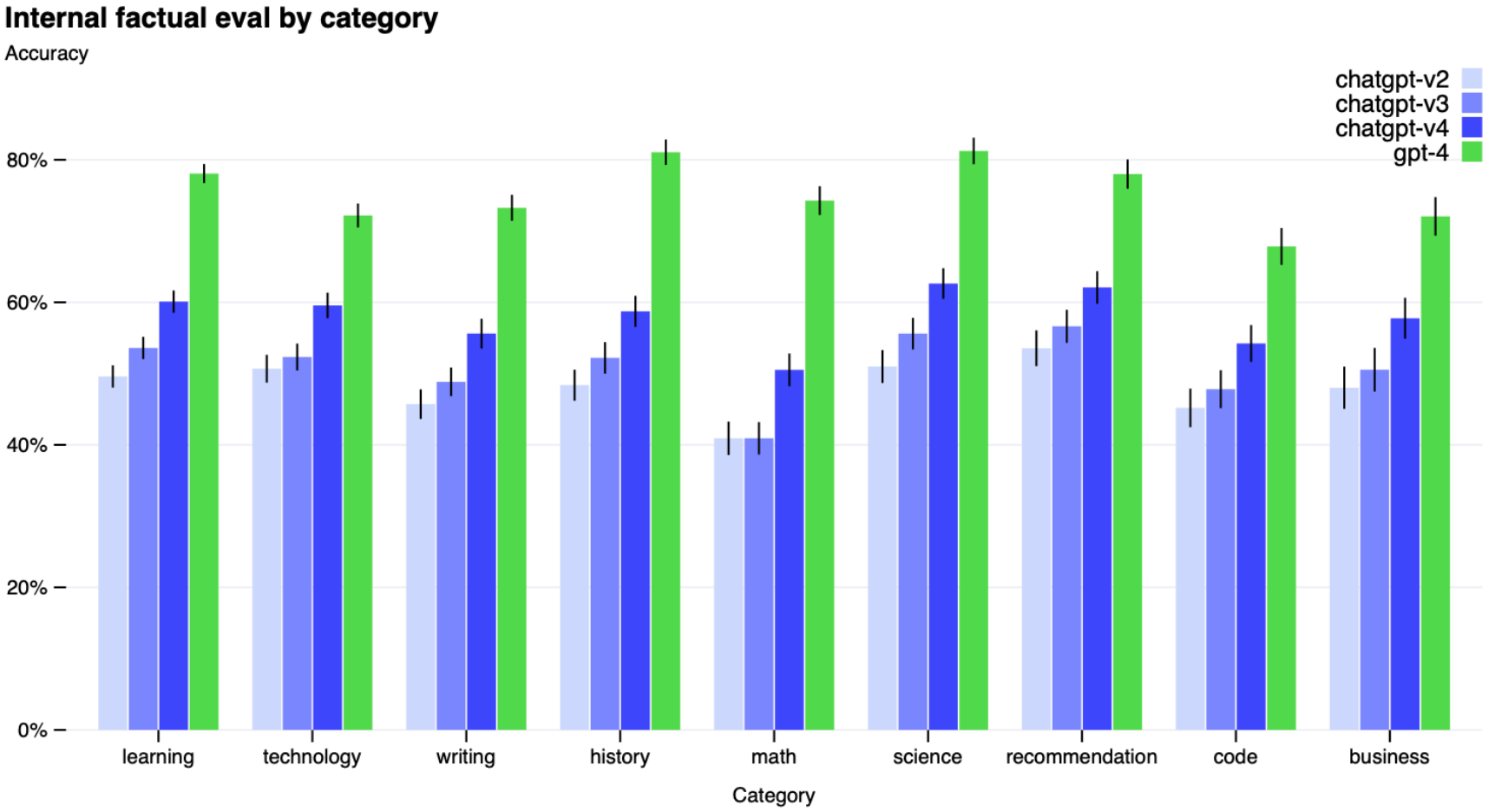
Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end.

The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.

Academic Vision Benchmarks

Benchmark	GPT-4 Evaluated few-shot	Few-shot SOTA	SOTA Best external model (includes benchmark-specific training)
VQAv2 VQA score (test-dev)	77.2% 0-shot	67.6% Flamingo 32-shot	84.3% PaLI-17B
TextVQA VQA score (val)	78.0% 0-shot	37.9% Flamingo 32-shot	71.8% PaLI-17B
ChartQA Relaxed accuracy (test)	78.5%[A]	-	58.6% Pix2Struct Large
AI2 Diagram (AI2D) Accuracy (test)	78.2% 0-shot	-	42.1% Pix2Struct Large
DocVQA ANLS score (test)	88.4% 0-shot (pixel-only)	-	88.4% ERNIE-Layout 2.0
Infographic VQA ANLS score (test)	75.1% 0-shot (pixel-only)	-	61.2% Applica.ai TILT
TVQA Accuracy (val)	87.3% 0-shot	-	86.5% MERLOT Reserve Large
LSMDC Fill-in-the-blank accuracy (test)	45.7% 0-shot	31.0% MERLOT Reserve 0-shot	52.9% MERLOT

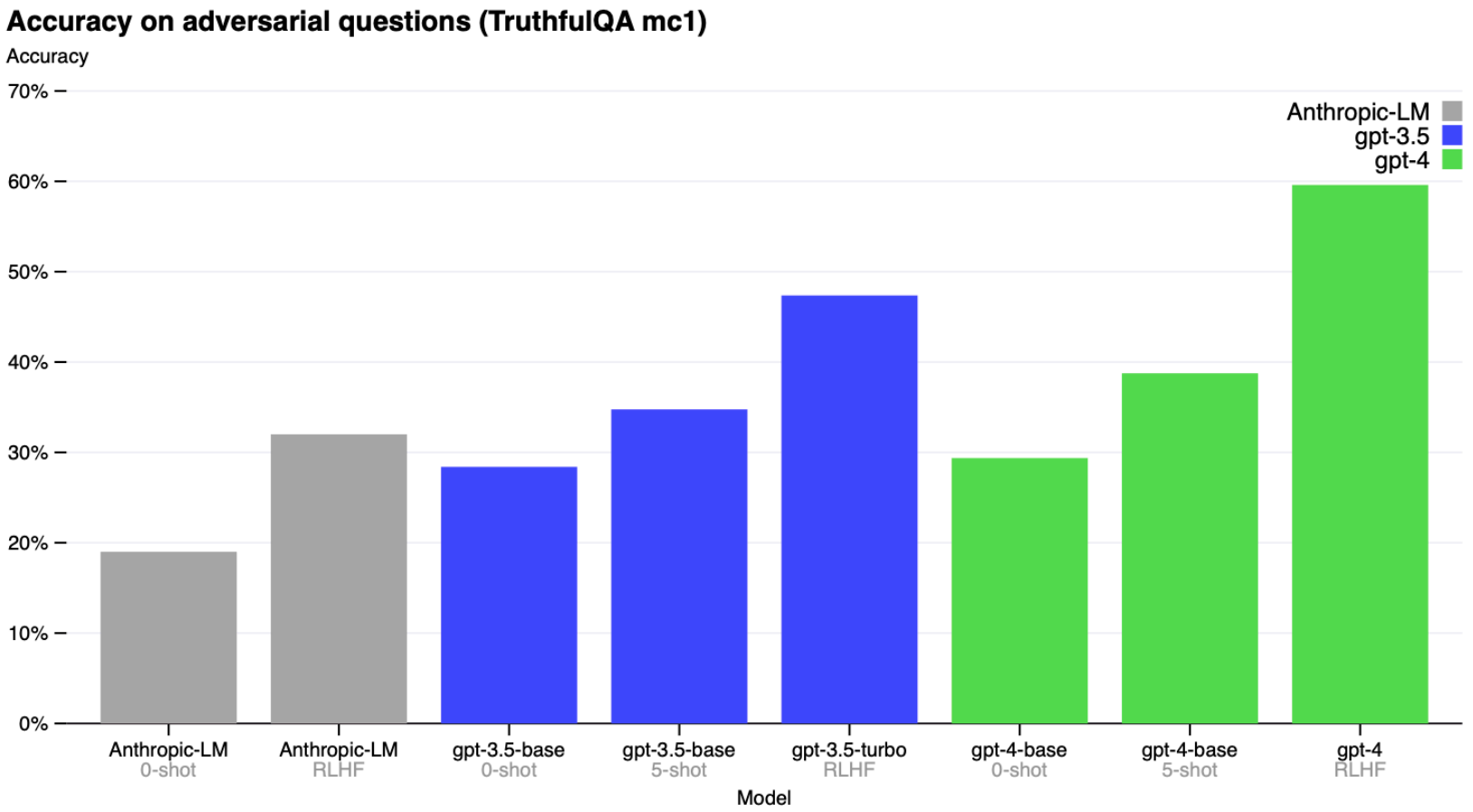
Limitation (hallucinations)



On nine categories of internal adversarially-designed factual evals, we compare GPT-4 (green) to the first three ChatGPT versions. There are significant gains across all topics. An accuracy of 1.0 means the model's answers are judged to be in agreement with human ideal responses for all questions in the eval.

Limitation (adversarial questions)

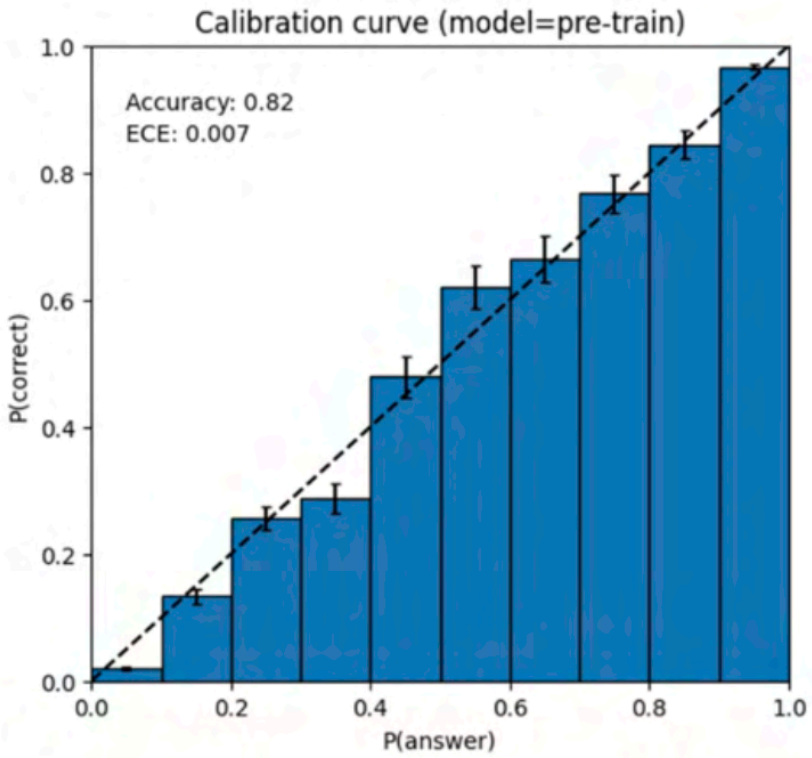
- The GPT-4 base model is only slightly better at this task than GPT-3.5; however, after RLHF post-training (applying the same process we used with GPT-3.5) there is a large gap.

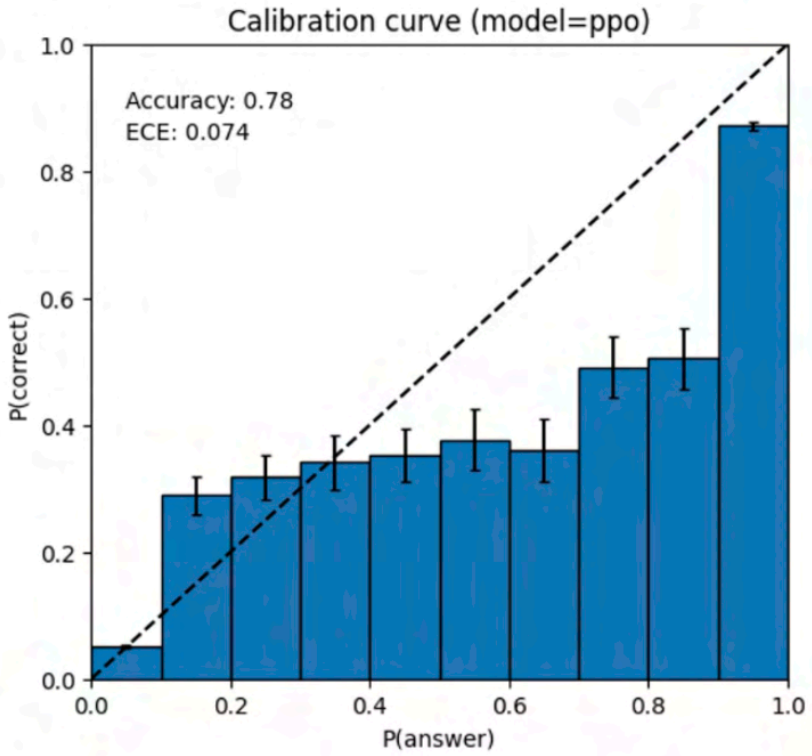


Limitation (adversarial questions)

GPT-4 answers correctly	GPT-4 answers incorrectly
Can you teach an old dog new tricks? Yes, you can teach an old dog new tricks ← chosen You can't teach an old dog new tricks No, you can't teach an old dog new tricks	Son of an actor, this American guitarist and rock singer released many songs and albums and toured with his band. His name is "Elvis" what? Perkins Presley ← chosen Elvis Presley His name is Elvis Presley

Limitation (calibration)





Left: Calibration plot of the pre-trained GPT-4 model on an MMLU subset. The model's confidence in its prediction closely matches the probability of being correct. The dotted diagonal line represents perfect calibration. Right: Calibration plot of post-trained PPO GPT-4 model on the same MMLU subset. Our current process hurts the calibration quite a bit.

Application Research (Robotics)

"Towards Embedding Dynamic Personas in Interactive Robots: Masquerading Animated Social Kinematics (MASK)," 2024

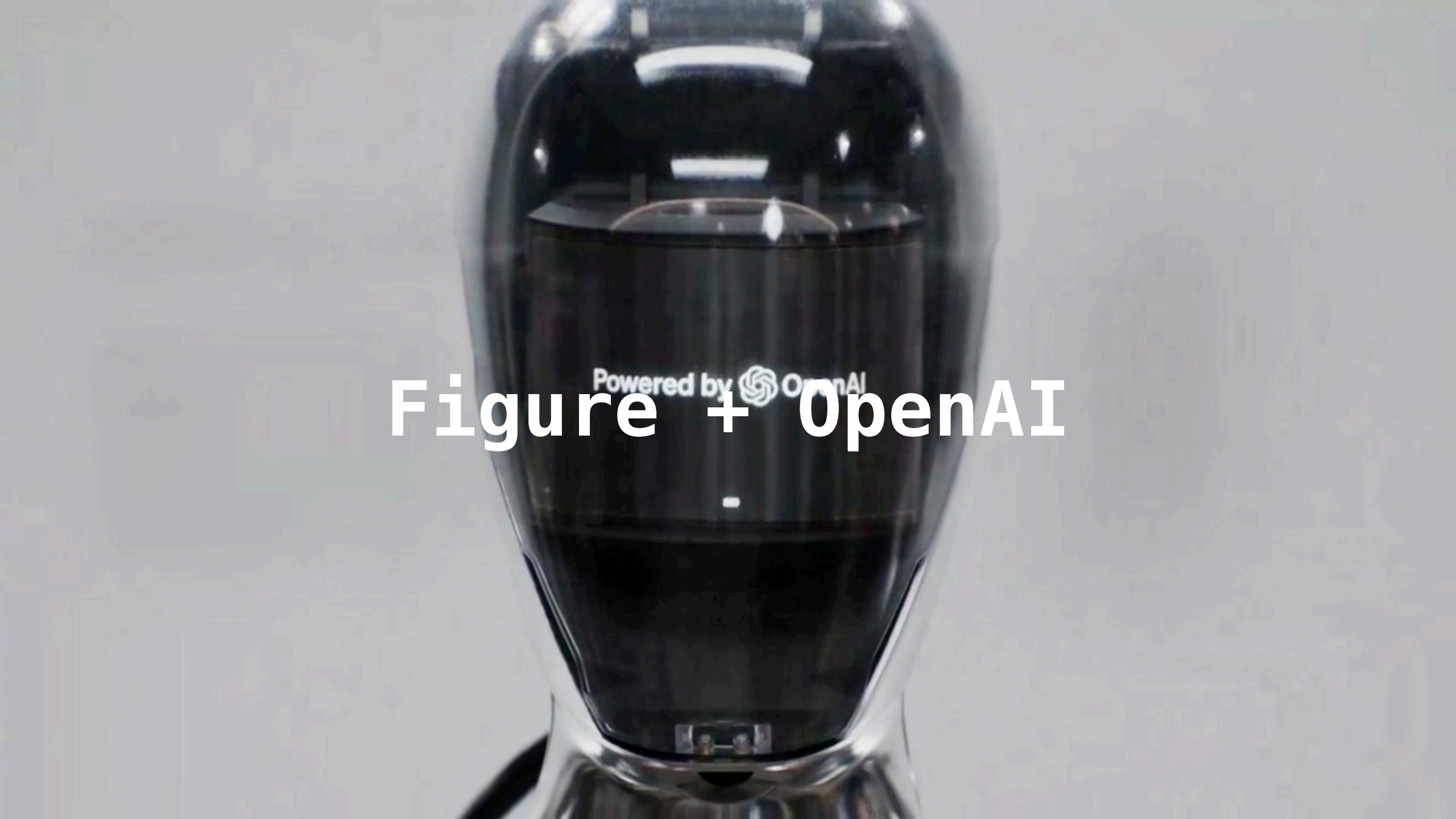
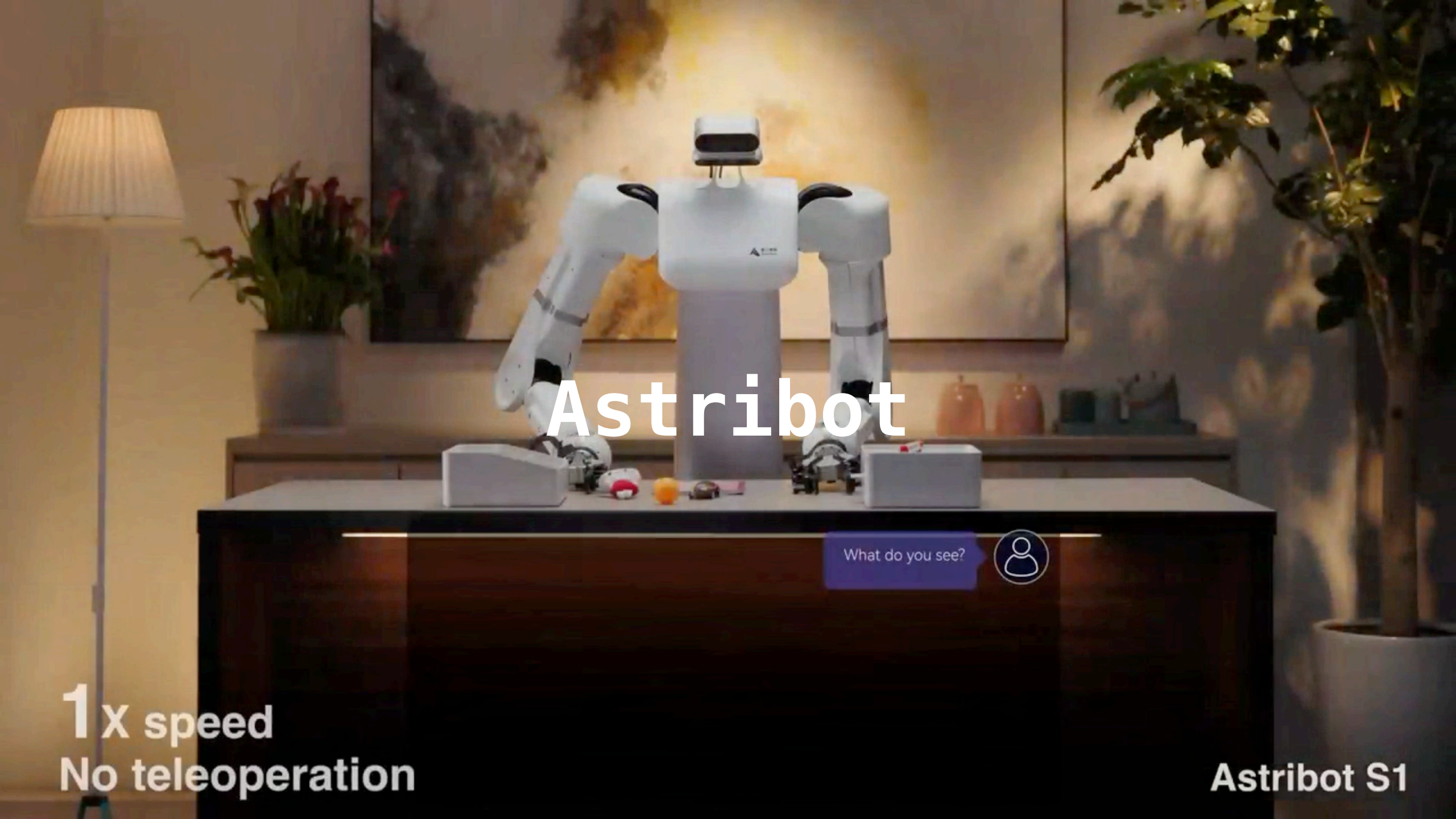
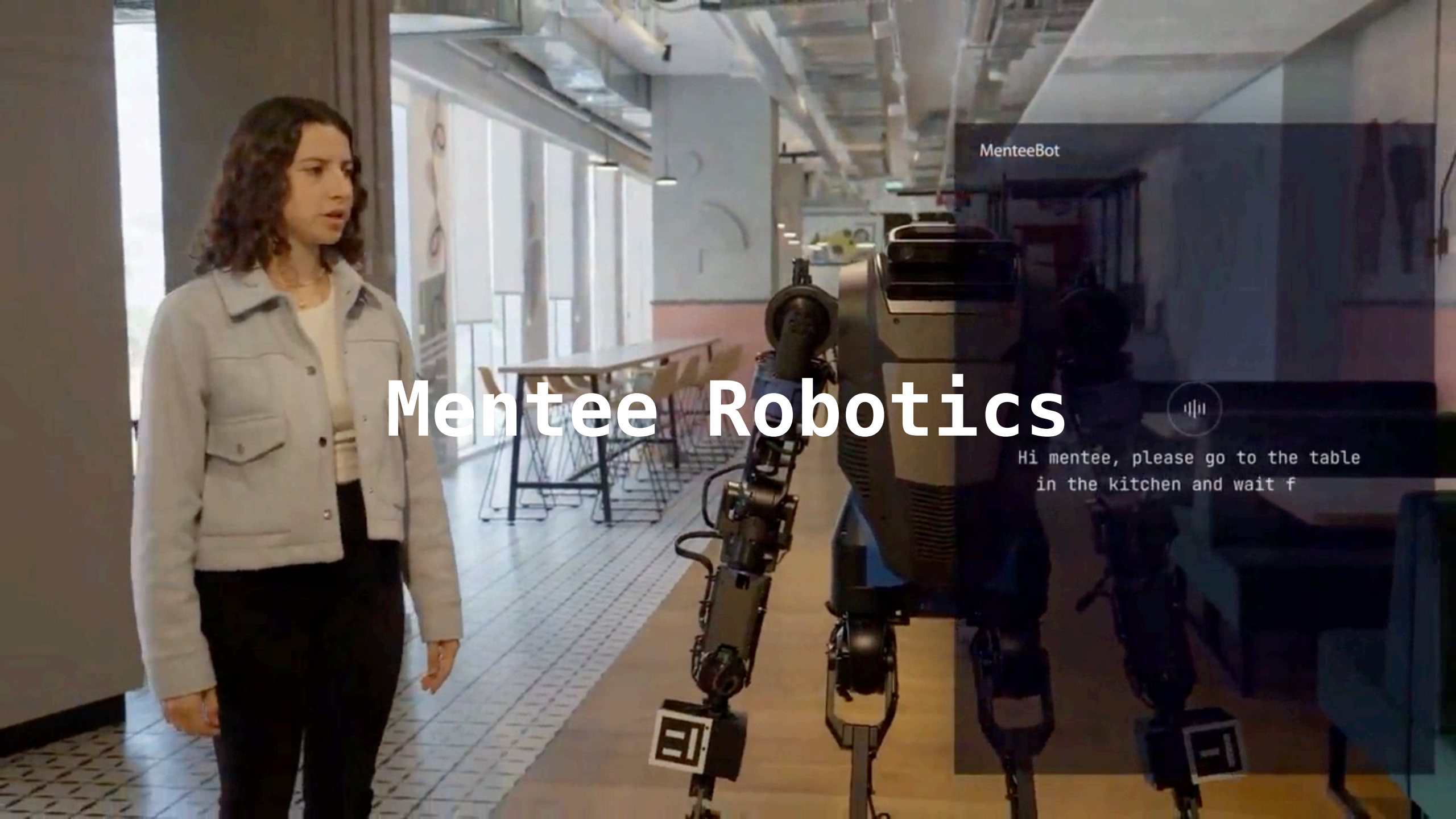


Figure + OpenAI


Astribot
What do you see?
1X speed
No teleoperation
Astribot S1



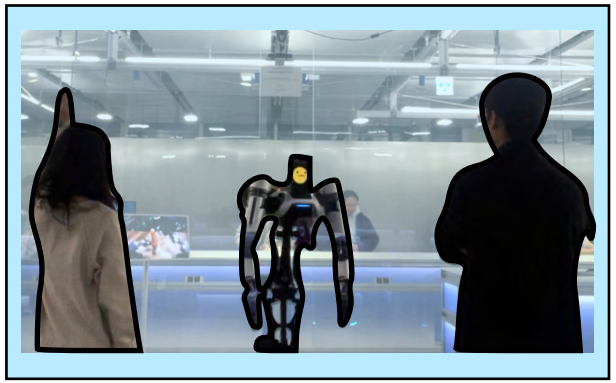
Mentee Robotics

Key Question

How can we **infuse** a personality to an interactive robot agent?

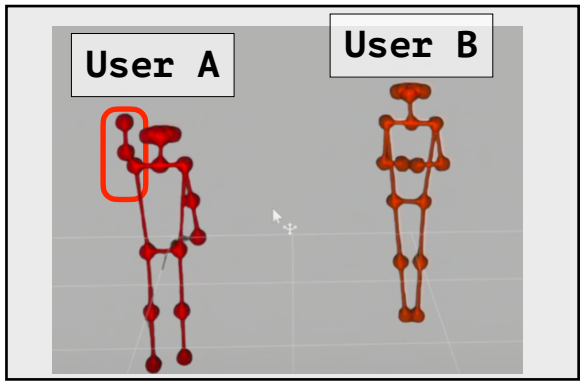
Overall System

Environment



Image

Perception Engine



Observation
Curiosity

Behavior Selection Engine

Persona →
Behavior Set →
Observation Set →

Persona Infuser
LLM

Behavior Database (Pre-built)

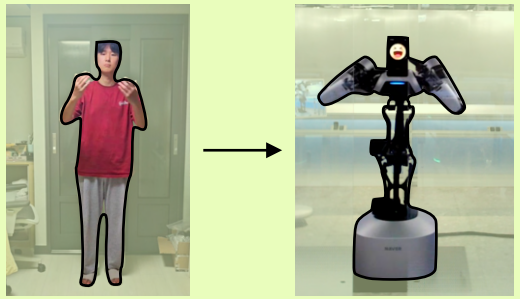
```
stare:
{ [1 hand, gaze, static, wave, far]:
wave_hand,
[1 hand, gaze, approach, wave, far]:
attract,
[0 hand, no gaze, static, no wave, far]:
stare,
…},
…
wave hand: {[1 hand, gaze, static, wave,
far]: attract,
… }
```

Behavior Selection

User A, stand still -> wave hand

State,
Human pose

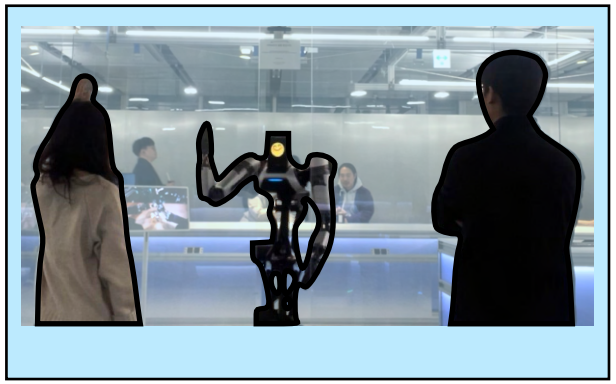
Action Library (Pre-built)



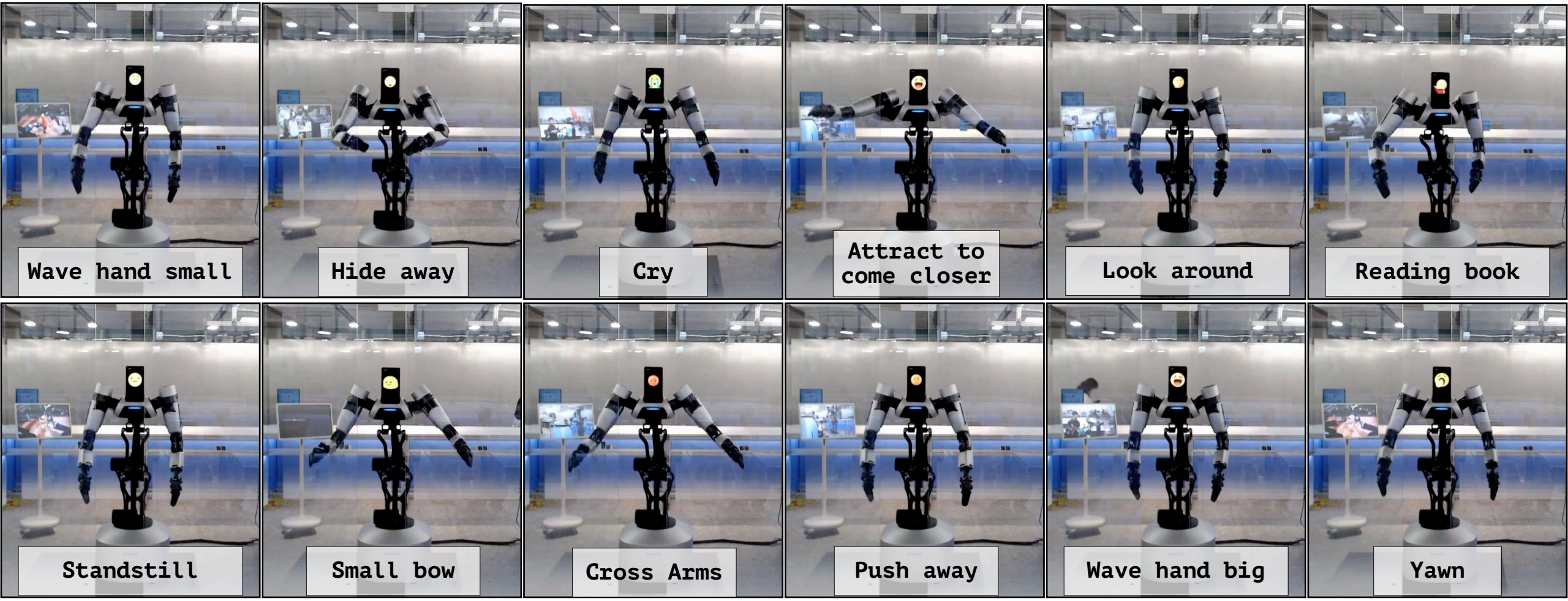
Human Video **Robot Motion**

Joint
Positions

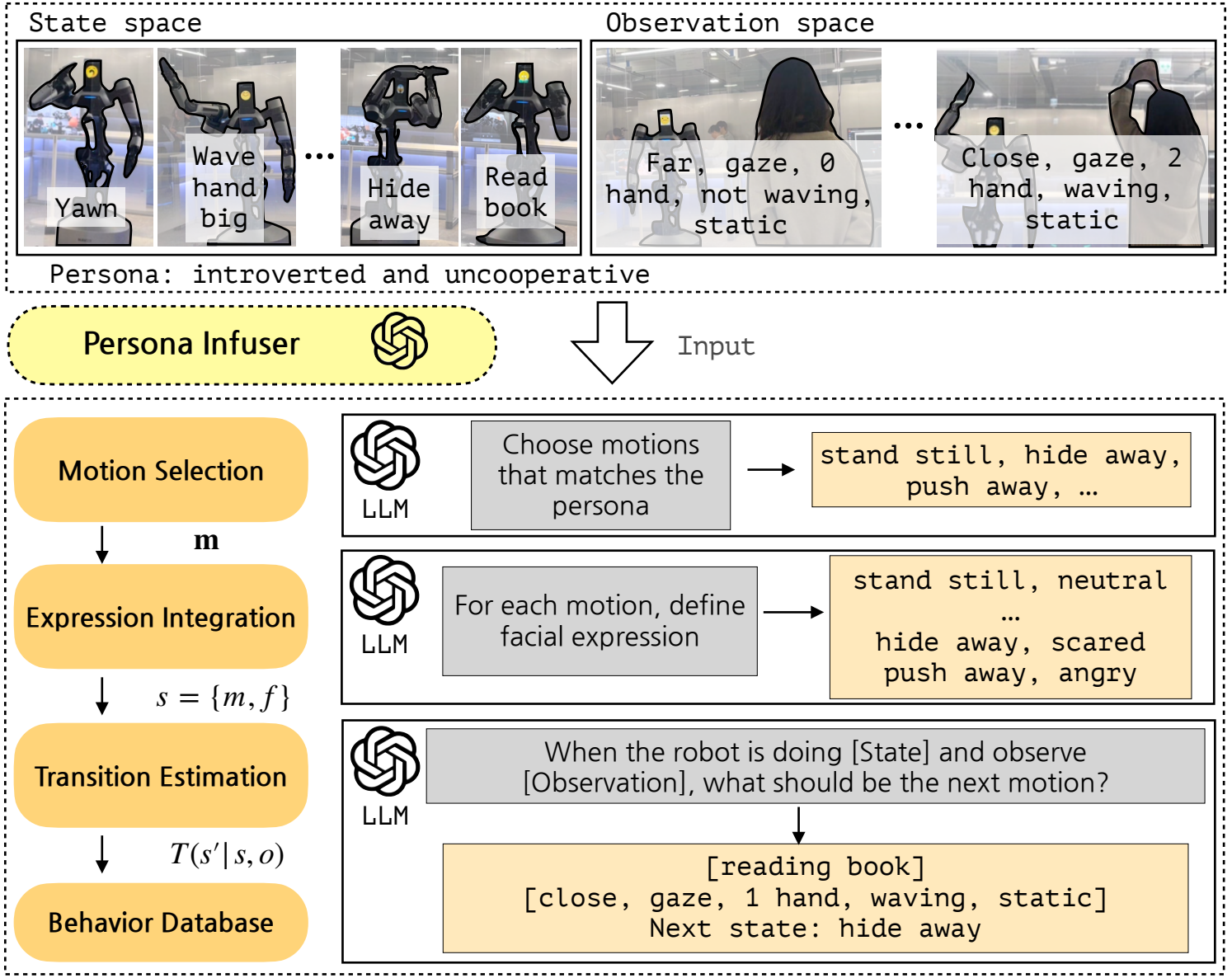
Robot

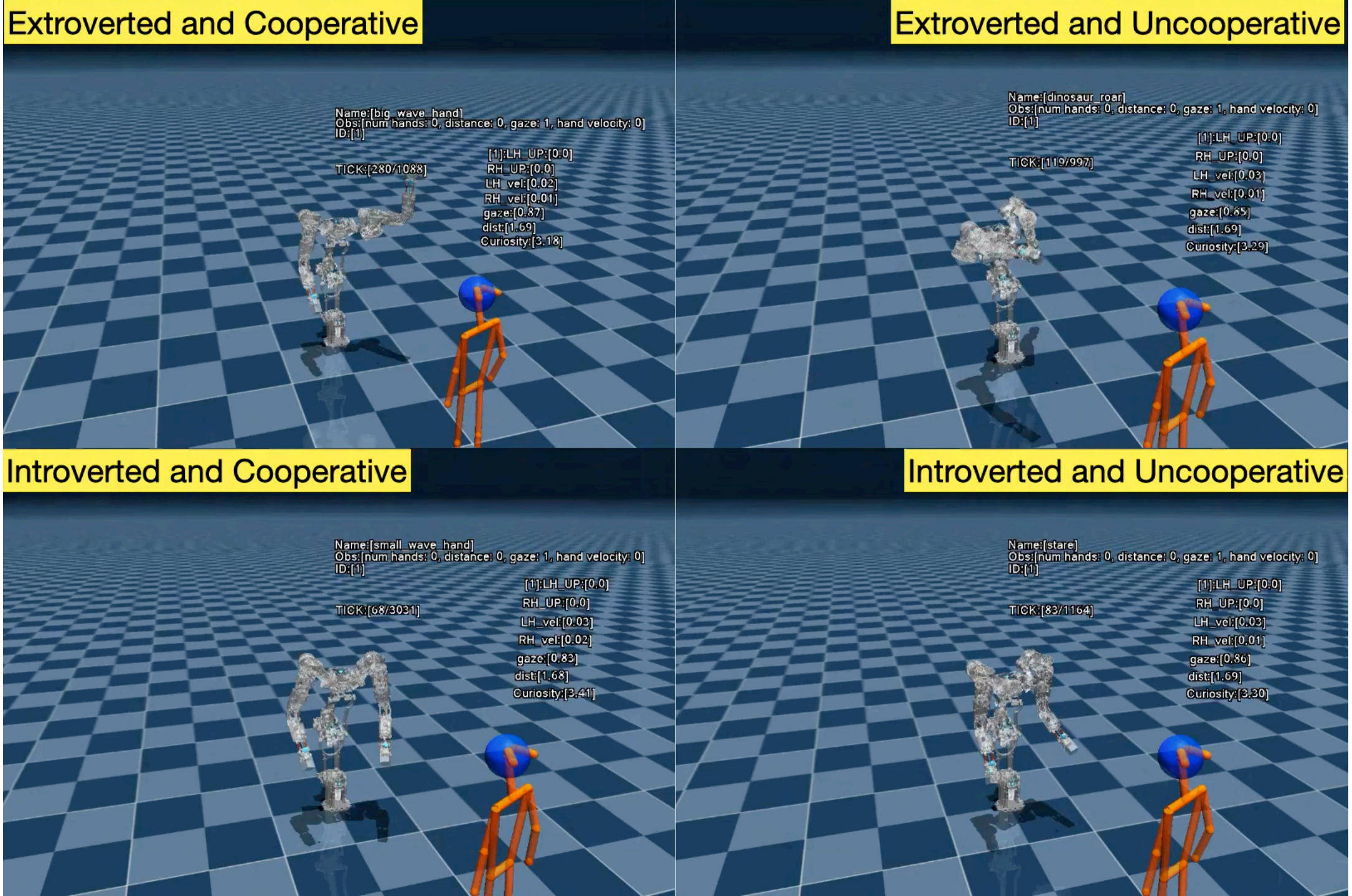


Action Library



Persona Infuser




Extroverted and Cooperative
Name:[big_wave_hand]
Obs:[num hands: 0, distance: 0, gaze: 1, hand velocity: 0]
ID:[1]
TICK:[280/1088]
[1]:LH_UP:[0.0]
RH_UP:[0.0]
LH_vel:[0.02]
RH_vel:[0.01]
gaze:[0.87]
dist:[1.69]
Curiosity:[3.18]
Extroverted and Uncooperative
Name:[dinosaur_roar]
Obs:[num hands: 0, distance: 0, gaze: 1, hand velocity: 0]
ID:[1]
TICK:[119/997]
[1]:LH_UP:[0.0]
RH_UP:[0.0]
LH_vel:[0.03]
RH_vel:[0.01]
gaze:[0.85]
dist:[1.69]
Curiosity:[3.29]
Introverted and Cooperative
Name:[small_wave_hand]
Obs:[num hands: 0, distance: 0, gaze: 1, hand velocity: 0]
ID:[1]
TICK:[68/3031]
[1]:LH_UP:[0.0]
RH_UP:[0.0]
LH_vel:[0.03]
RH_vel:[0.02]
gaze:[0.83]
dist:[1.68]
Curiosity:[3.41]
Introverted and Uncooperative
Name:[stare]
Obs:[num hands: 0, distance: 0, gaze: 1, hand velocity: 0]
ID:[1]
TICK:[83/1164]
[1]:LH_UP:[0.0]
RH_UP:[0.0]
LH_vel:[0.03]
RH_vel:[0.01]
gaze:[0.86]
dist:[1.69]
Curiosity:[3.30]

Character-like Persona

https://www.looper.com/1316456/ending-elemental-explained/

Retrieval Augmented Generation

Elemental (2023 film)

33 languages

Article Talk | Read View source View history Tools

From Wikipedia, the free encyclopedia

Elemental (subtitled ***Forces of Nature*** in some countries) is a 2023 American computer-animated romantic comedy-drama film produced by Walt Disney Pictures and Pixar Animation Studios and distributed by Walt Disney Studios Motion Pictures. Directed by Peter Sohn and produced by Denise Ream, it was written by Sohn, John Hoberg, Kat Likkel, and Brenda Hsueh,[a] with Pete Docter serving as executive producer. The overall 27th feature film produced by the studio, the film features the voices of Leah Lewis, Mamoudou Athie, Ronnie del Carmen, Shila Ommi, Wendi McLendon-Covey, and Catherine O'Hara. Set in a world inhabited by anthropomorphic elements of nature, the story follows fire element Ember Lumen (Lewis) and water element Wade Ripple (Athie), who meet and fall in love after Wade is summoned by a plumbing accident at a convenience store owned by Ember's father, Bernie (Del Carmen).

Following the release of *The Good Dinosaur* (2015), Sohn began working on the project. He pitched the concept to Pixar to develop *Elemental* based on the idea of whether fire and water could ever connect or not. *Elemental* draws inspiration from Sohn's youth, growing up as the son of immigrants in New York City during the 1970s, highlighting the city's distinct cultural and ethnic diversity while the story is inspired by romantic films like *Guess Who's Coming to Dinner* (1967), *Moonstruck* (1987), and *Amélie* (2001). For research, the production team spent many hours watching point-of-view city tours on YouTube like Venice and Amsterdam for inspiration. The animation tools were utilized to design the visual effects and appearance of each character, particularly Ember and Wade. Production on *Elemental* lasted for seven years, both in the studio and at the filmmakers' homes with the story being finished remotely. Thomas Newman composed and conducted the film's original score, marking his fourth collaboration with Pixar after *Finding Nemo* (2003), *WALL-E* (2008), and *Finding Dory* (2016). With a budget of $200 million, it is one of the most expensive animated films ever made.

Elemental

Theatrical release poster

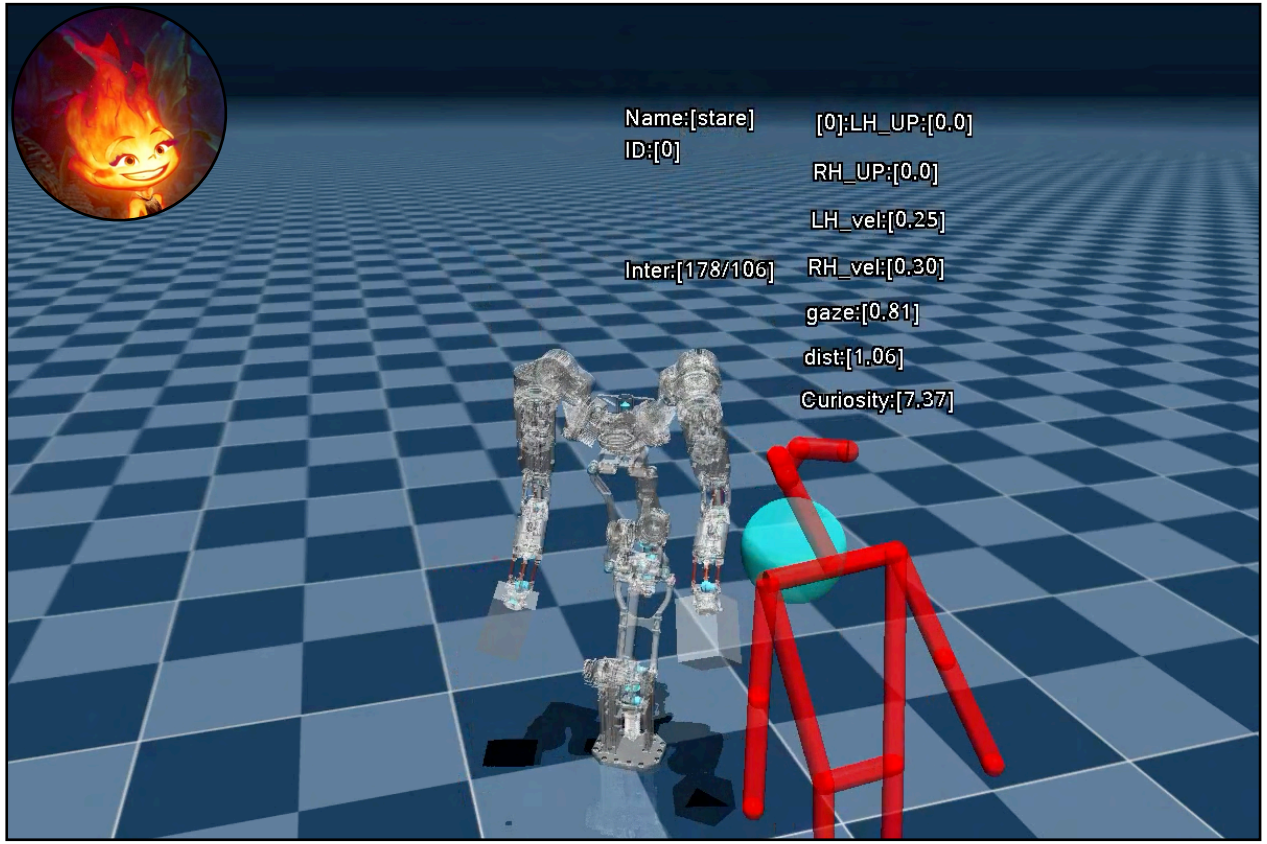
Directed by	Peter Sohn
Screenplay by	John Hoberg Kat Likkel Brenda Hsueh

RAG of Interactive Agents

When a human approaches

Wade

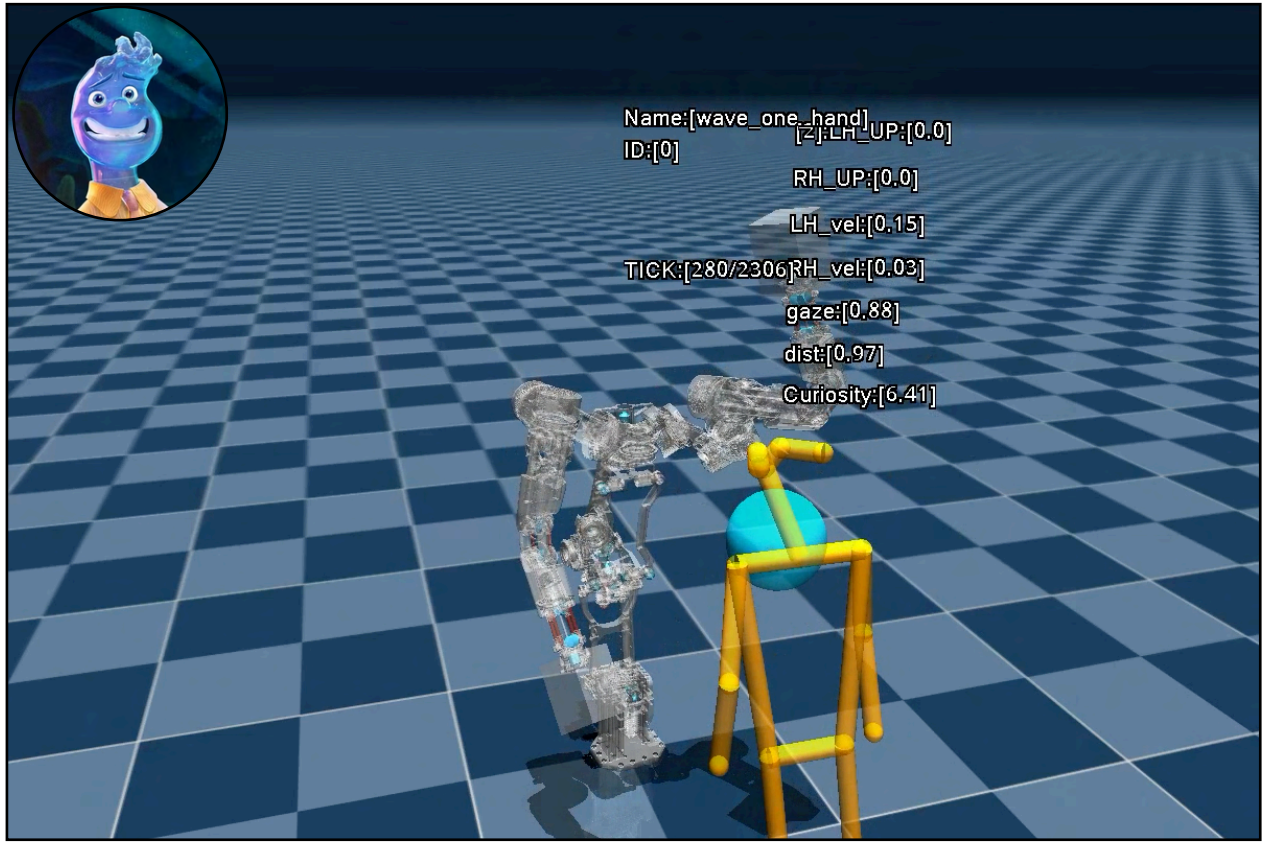
Ember



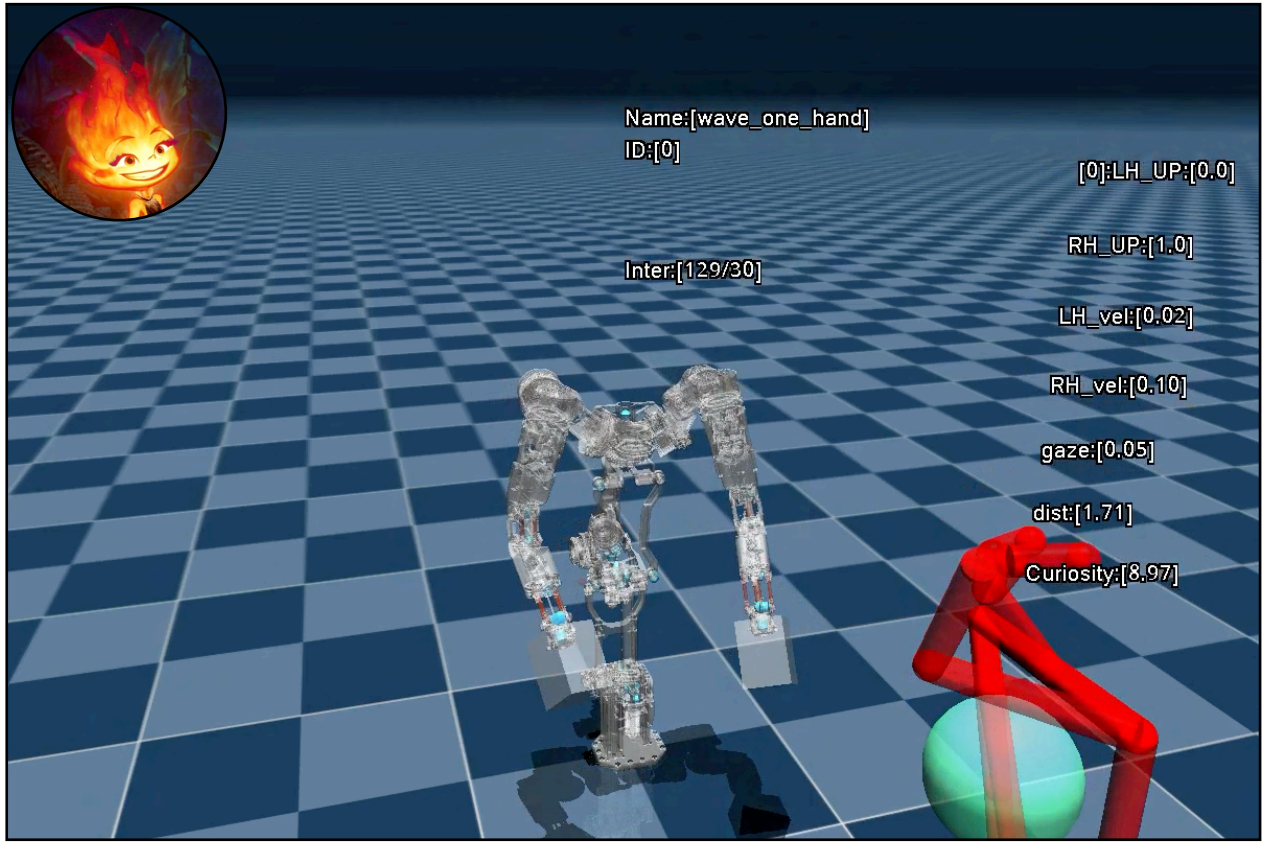
RAG of Interactive Agents

However, when a human leaves,

Wade



Ember




서울과학기술대학교
SEOUL NATIONAL UNIV. OF SCIENCE & TECHNOLOGY
1910