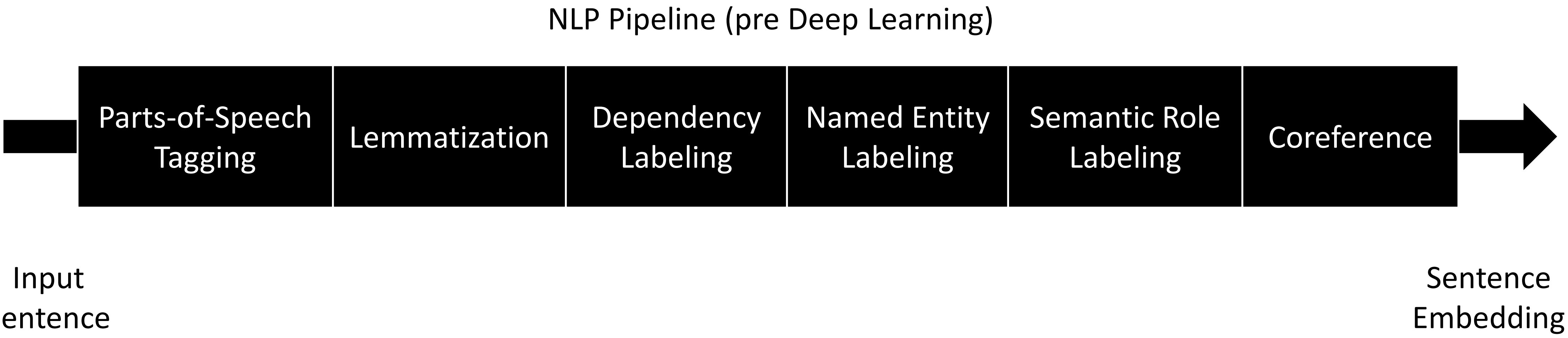
From Single Purpose
to
Multi-task & Multi-modal

Frédéric Odermatt

The classical NLP pipeline



NLP Pipeline (pre Deep Learning)

Parts-of-Speech Tagging

Lemmatization

Dependency Labeling

Named Entity Labeling

Semantic Role Labeling

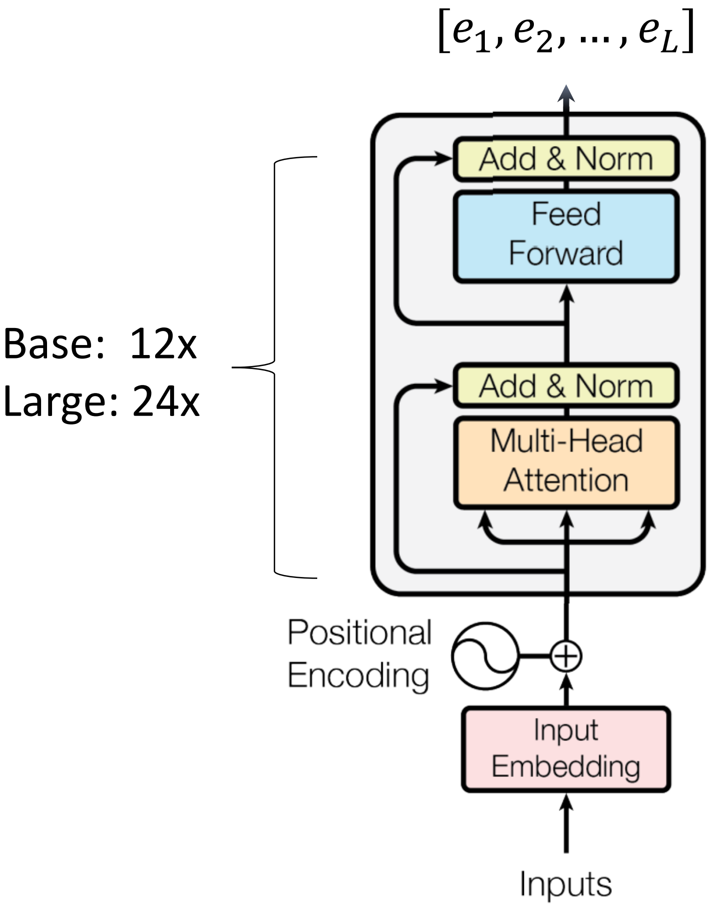
Coreference

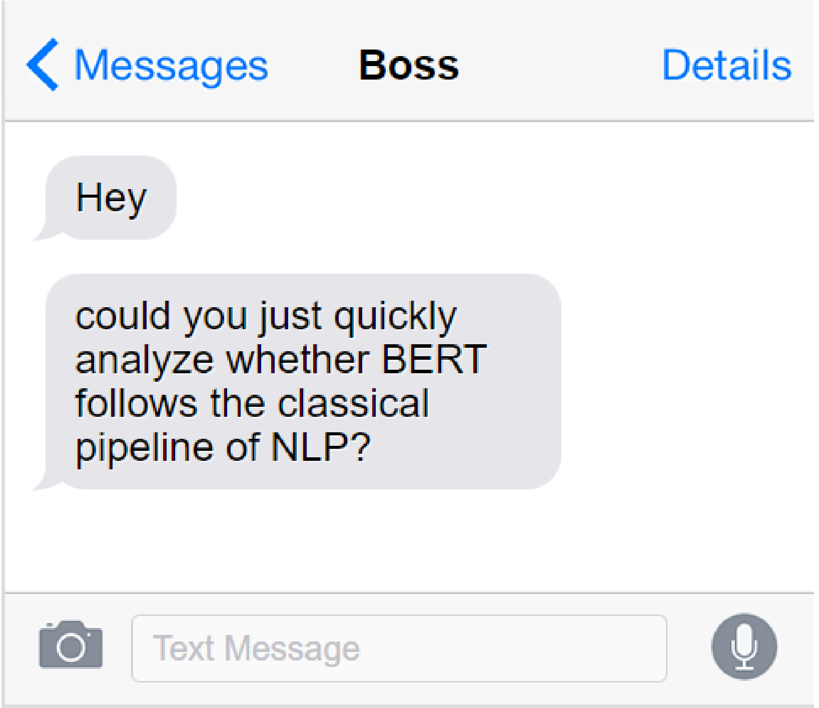
Input Sentence

Sentence Embedding

BERT rediscovers the classical NLP pipeline

[Aug 2019]

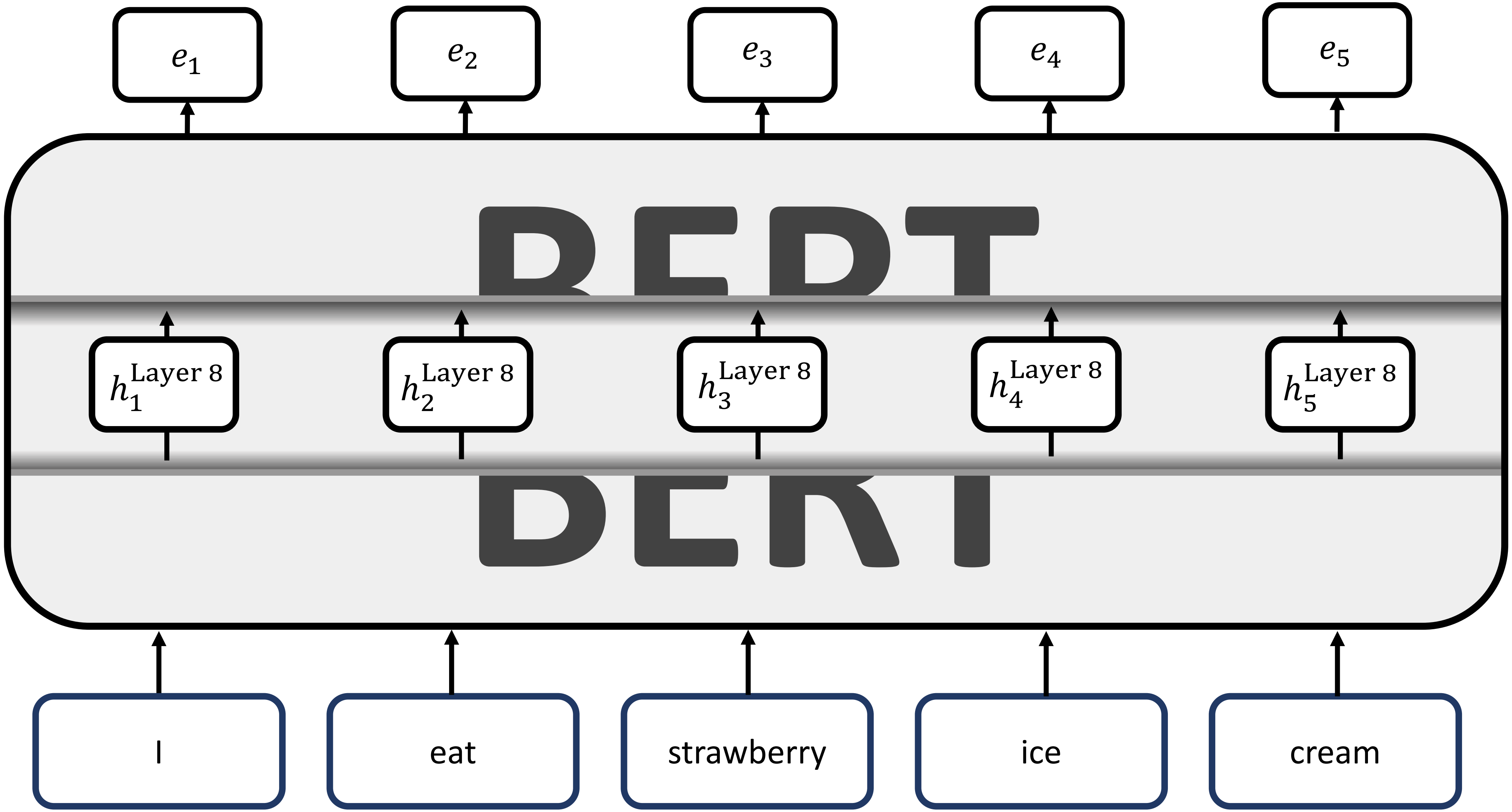




Hey

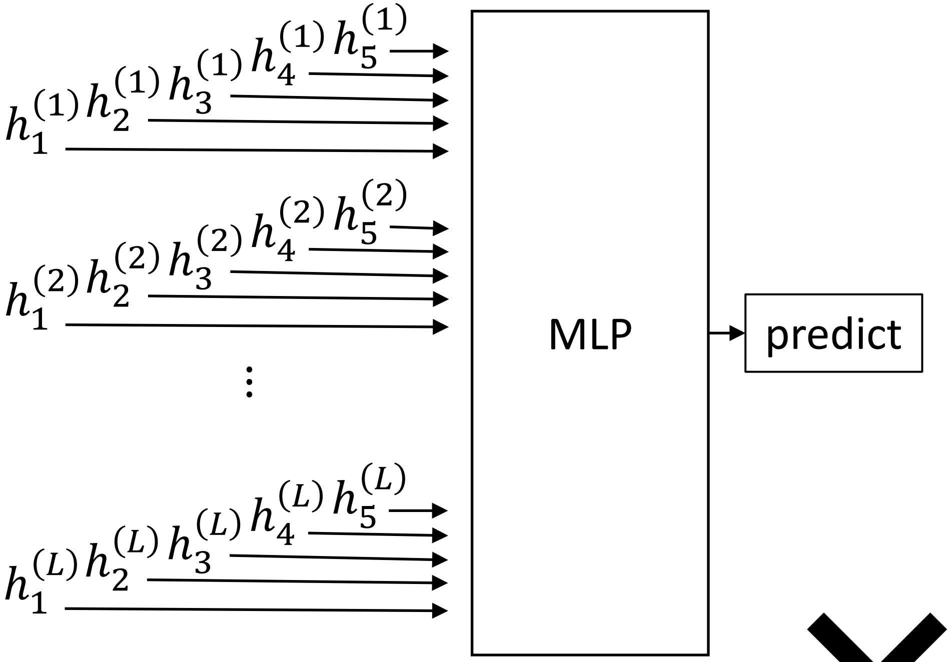
could you just quickly analyze whether BERT follows the classical pipeline of NLP?

BERT rediscovers the classical NLP pipeline



BERT rediscovers the classical NLP pipeline

$$h_1^{(8)} \coloneqq h_1^{\text{Layer 8}}$$



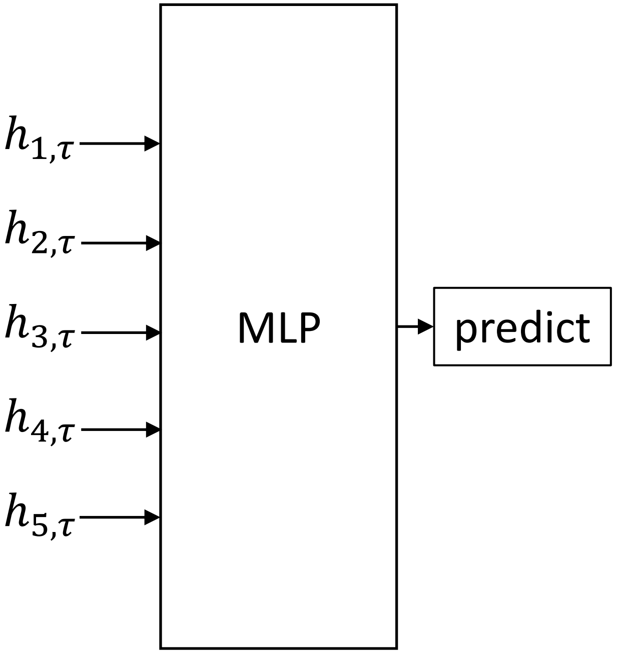
$h_1^{(1)}\, h_2^{(1)}\, h_3^{(1)}\, h_4^{(1)}\, h_5^{(1)}$
$h_1^{(2)}\, h_2^{(2)}\, h_3^{(2)}\, h_4^{(2)}\, h_5^{(2)}$
⋮
$h_1^{(L)}\, h_2^{(L)}\, h_3^{(L)}\, h_4^{(L)}\, h_5^{(L)}$
MLP
predict

✗

$$\boldsymbol{h}_{i,\tau} = \sum_{l=0}^{L} s_\tau^{(l)} \boldsymbol{h}_i^{(l)} \qquad \tau\text{: task}$$

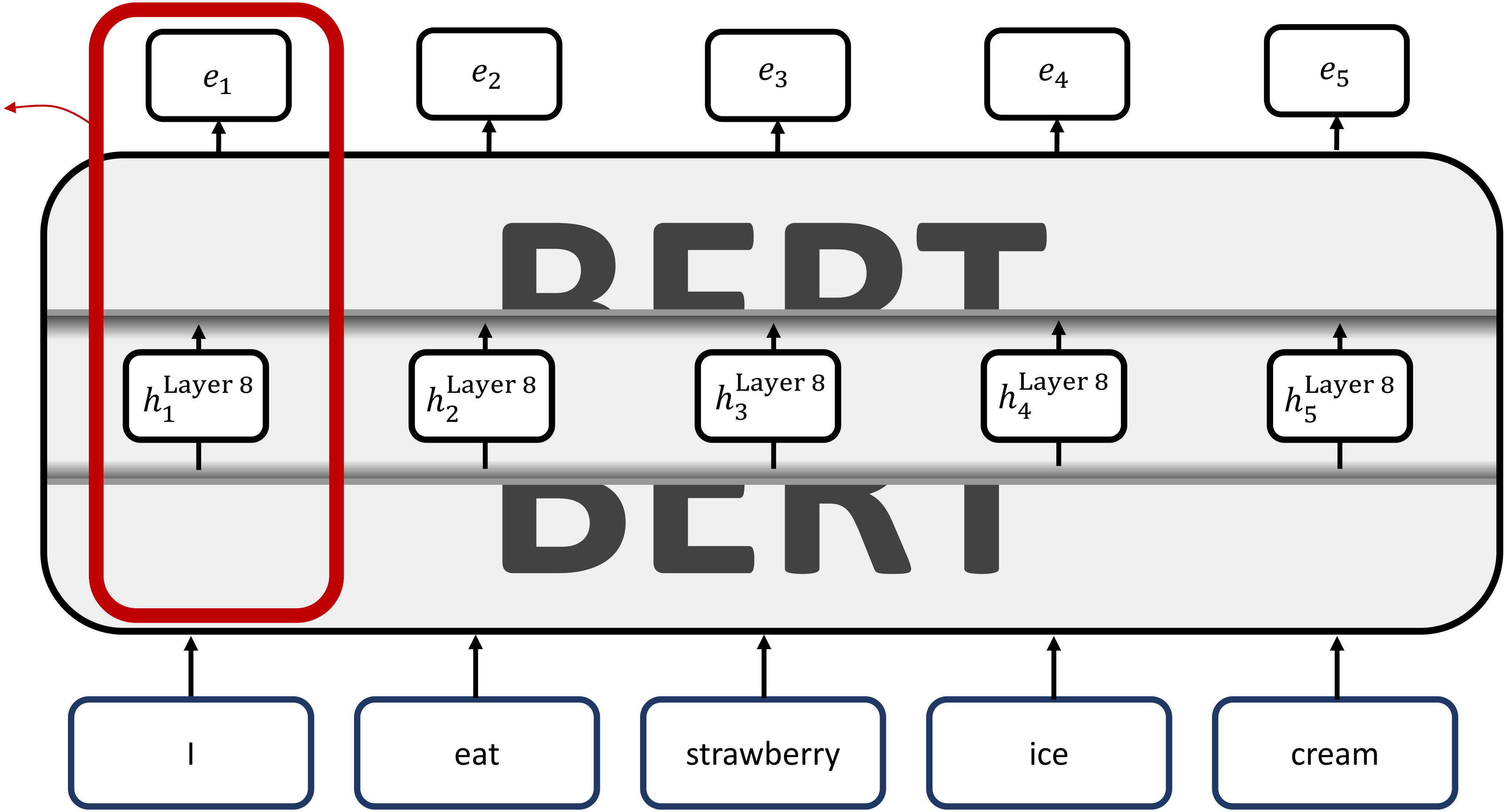
where $\boldsymbol{s}_\tau = \text{softmax}(\boldsymbol{a}_\tau)$

✓



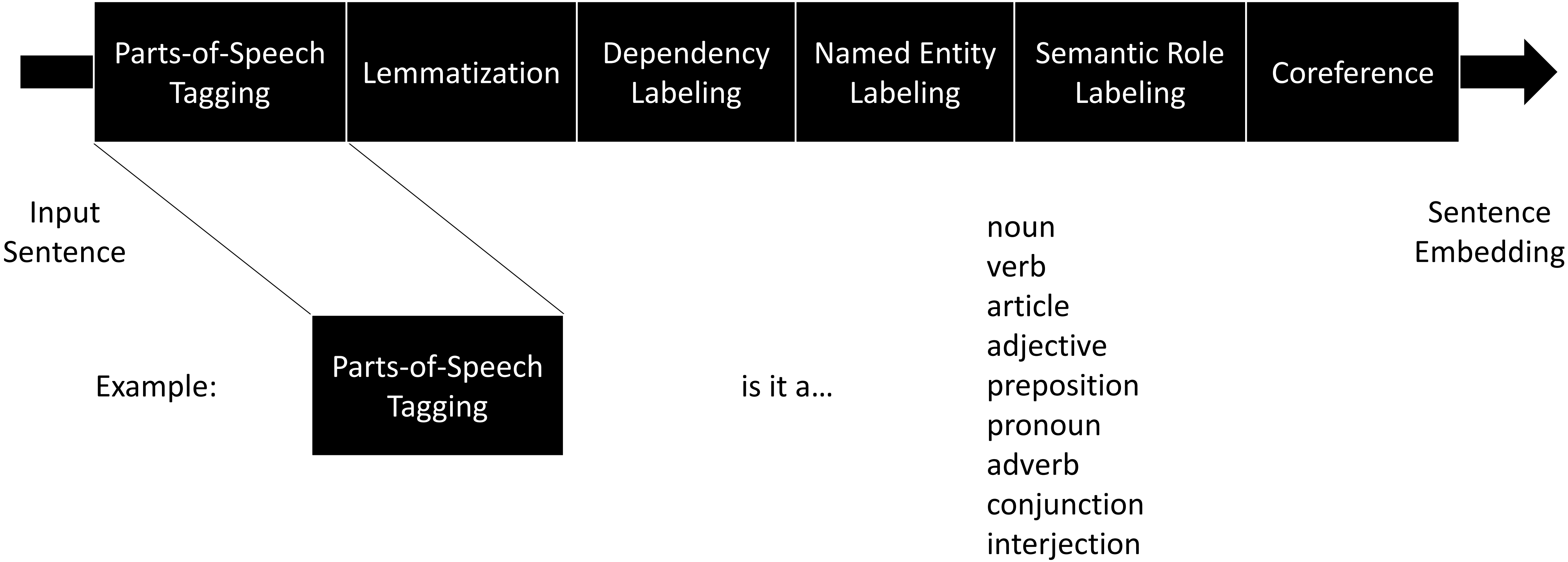
BERT rediscovers the classical NLP pipeline

$$\boldsymbol{h}_{1,\tau} = \sum_{l=0}^{L} s_{\tau}^{(l)} \boldsymbol{h}_{1}^{(l)}$$



BERT rediscovers the classical NLP pipeline

NLP Pipeline (pre/early Deep Learning)



Parts-of-Speech Tagging

Lemmatization

Dependency Labeling

Named Entity Labeling

Semantic Role Labeling

Coreference

Input Sentence

Sentence Embedding

Example:

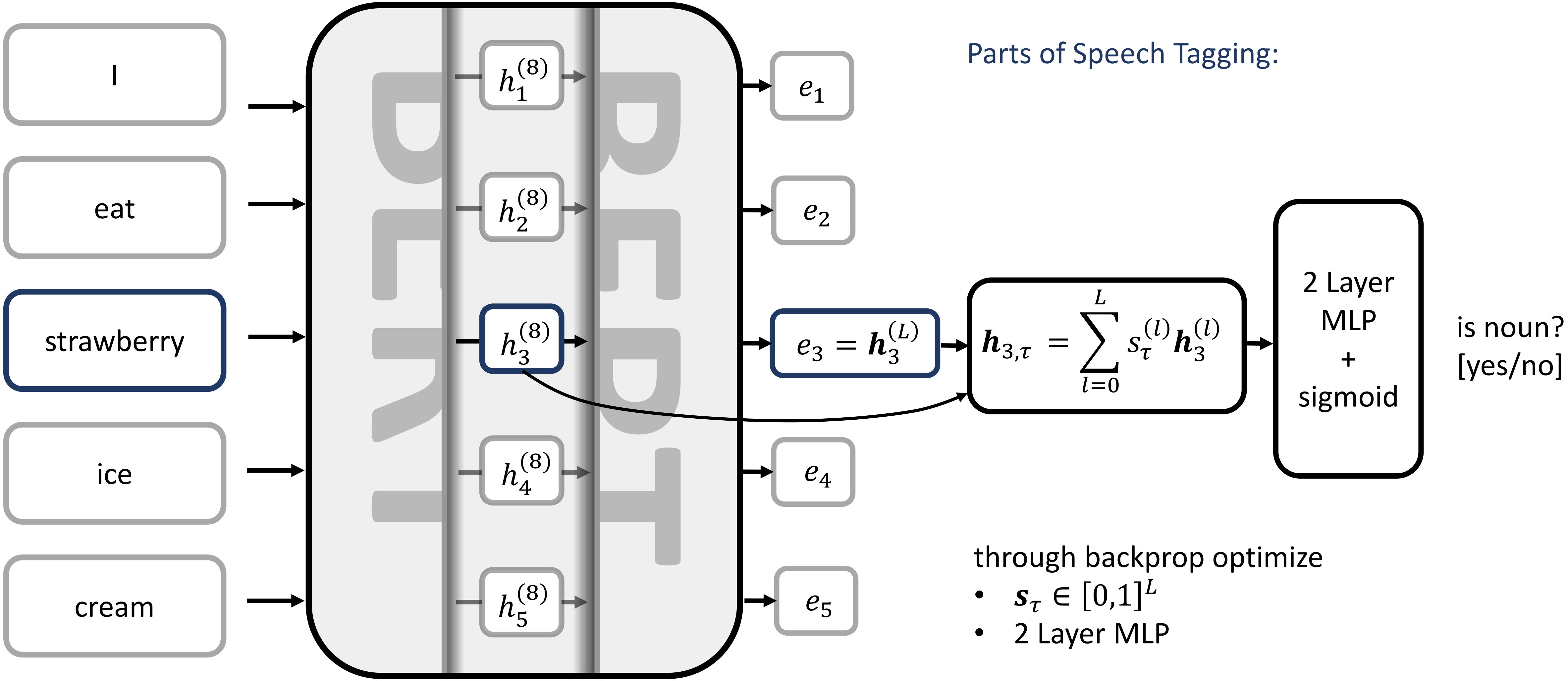
Parts-of-Speech Tagging

is it a…

noun
verb
article
adjective
preposition
pronoun
adverb
conjunction
interjection

BERT rediscovers the classical NLP pipeline

Parts of Speech Tagging:

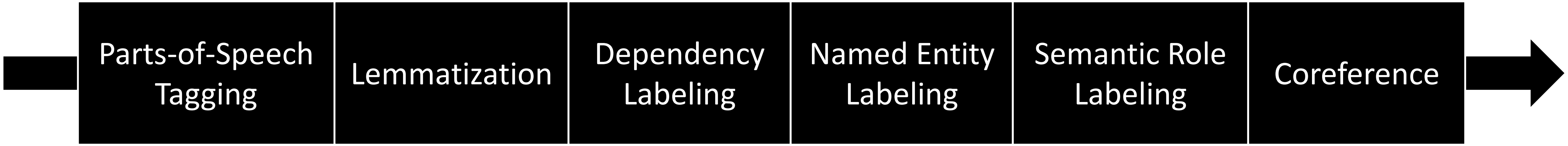


through backprop optimize

- $\boldsymbol{s}_\tau \in [0,1]^L$
- 2 Layer MLP

BERT rediscovers the classical NLP pipeline

NLP Pipeline (pre/early Deep Learning)



Parts-of-Speech Tagging

Lemmatization

Dependency Labeling

Named Entity Labeling

Semantic Role Labeling

Coreference

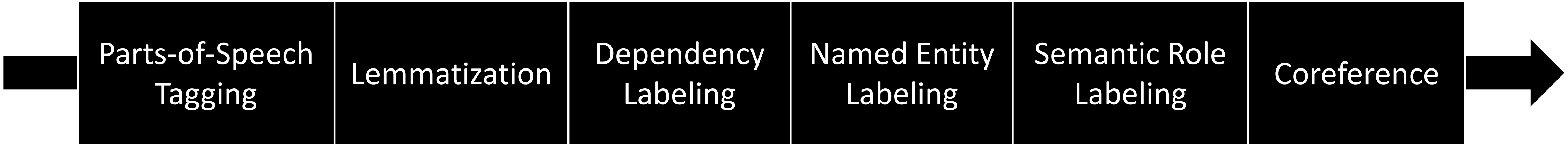
Parts of Speech Tagging: “I eat strawberry [ice] cream” → Noun

Training: is it a noun? y/n, is it a verb? y/n, is it a…

Coreference Resolution: “I haven’t seen [Jack] in the office today, so [he] might be working from home” → True

Training: do these two things refer to the same entity? y/n

BERT rediscovers the classical NLP pipeline

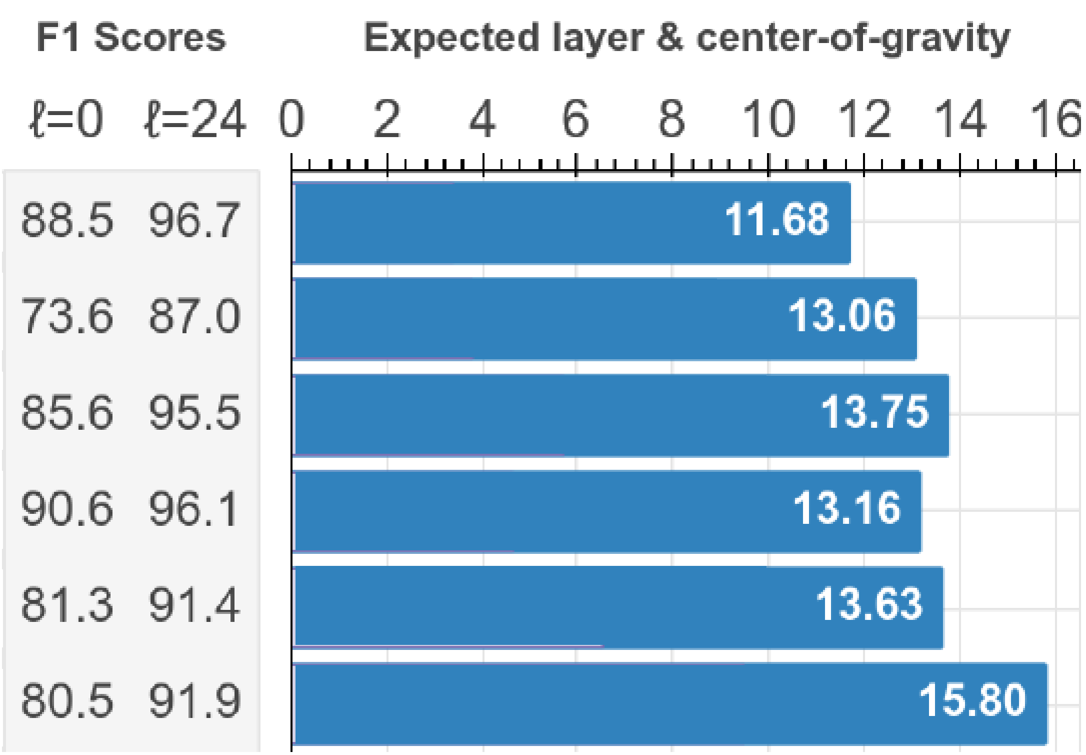


Analysis 1

Center of Gravity:

$$E_s[l] = \sum_{l=0}^{L} l \cdot s_\tau^{(l)}$$

$$\boldsymbol{h}_{i,\tau} = \sum_{l=0}^{L} s_\tau^{(l)} \boldsymbol{h}_i^{(l)}$$



	F1 Scores		Expected layer & center-of-gravity
	ℓ=0	ℓ=24	
Part-of-Speech Tagging	88.5	96.7	11.68
Constituency Parsing	73.6	87.0	13.06
Dependency Labeling	85.6	95.5	13.75
Named Entity Labeling	90.6	96.1	13.16
Semantic Role Labeling	81.3	91.4	13.63
Coreference	80.5	91.9	15.80

M(akridakis) Competitions

Time-series forecasting: “How good are we at it?”

M1 1982

M2 1993

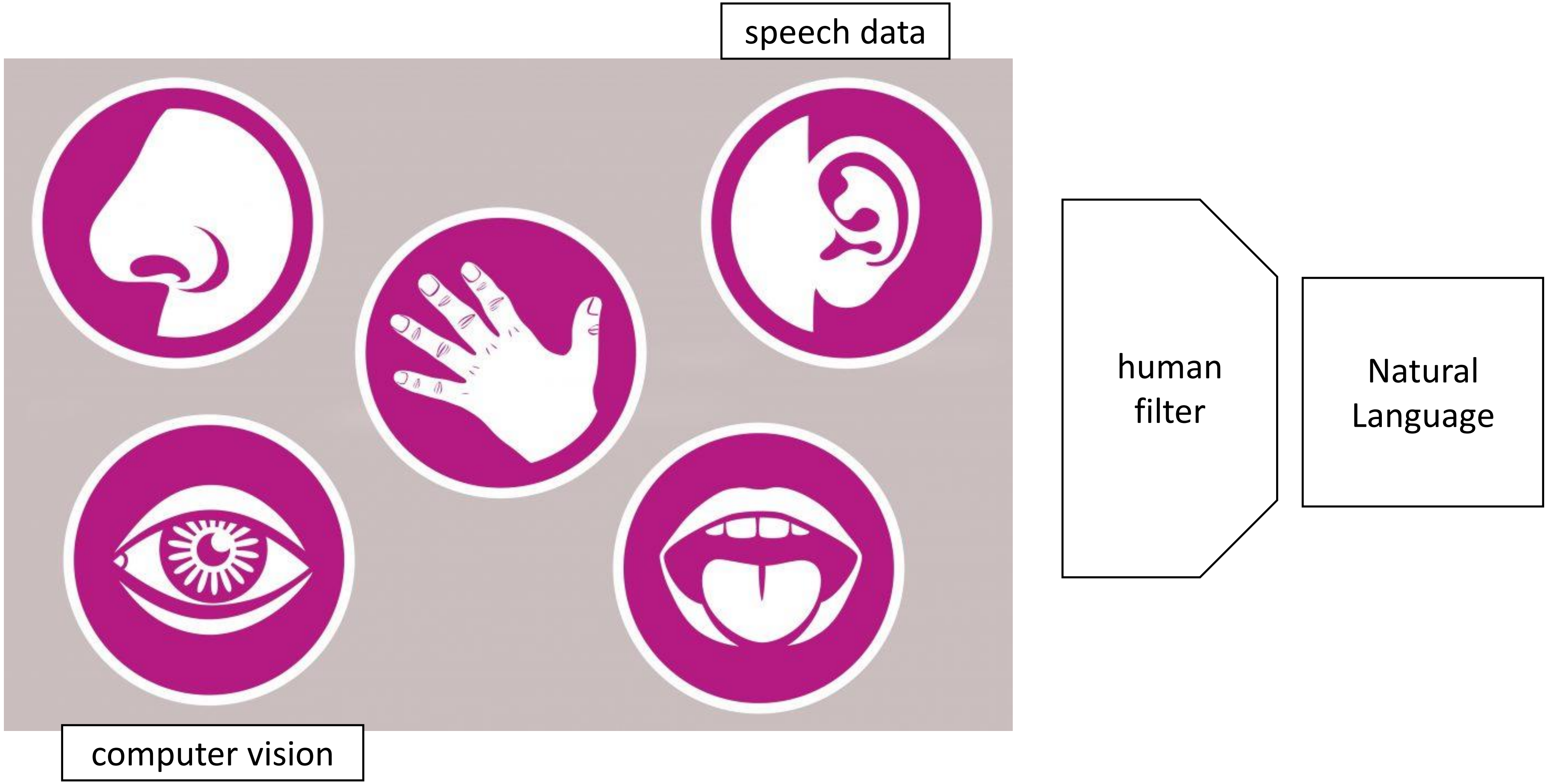
M3 2000

M4 2018

M5 2022

Multi-Modal Deep Learning

How we perceive the world



speech data

computer vision

human filter

Natural Language

CM3: Rethinking domains

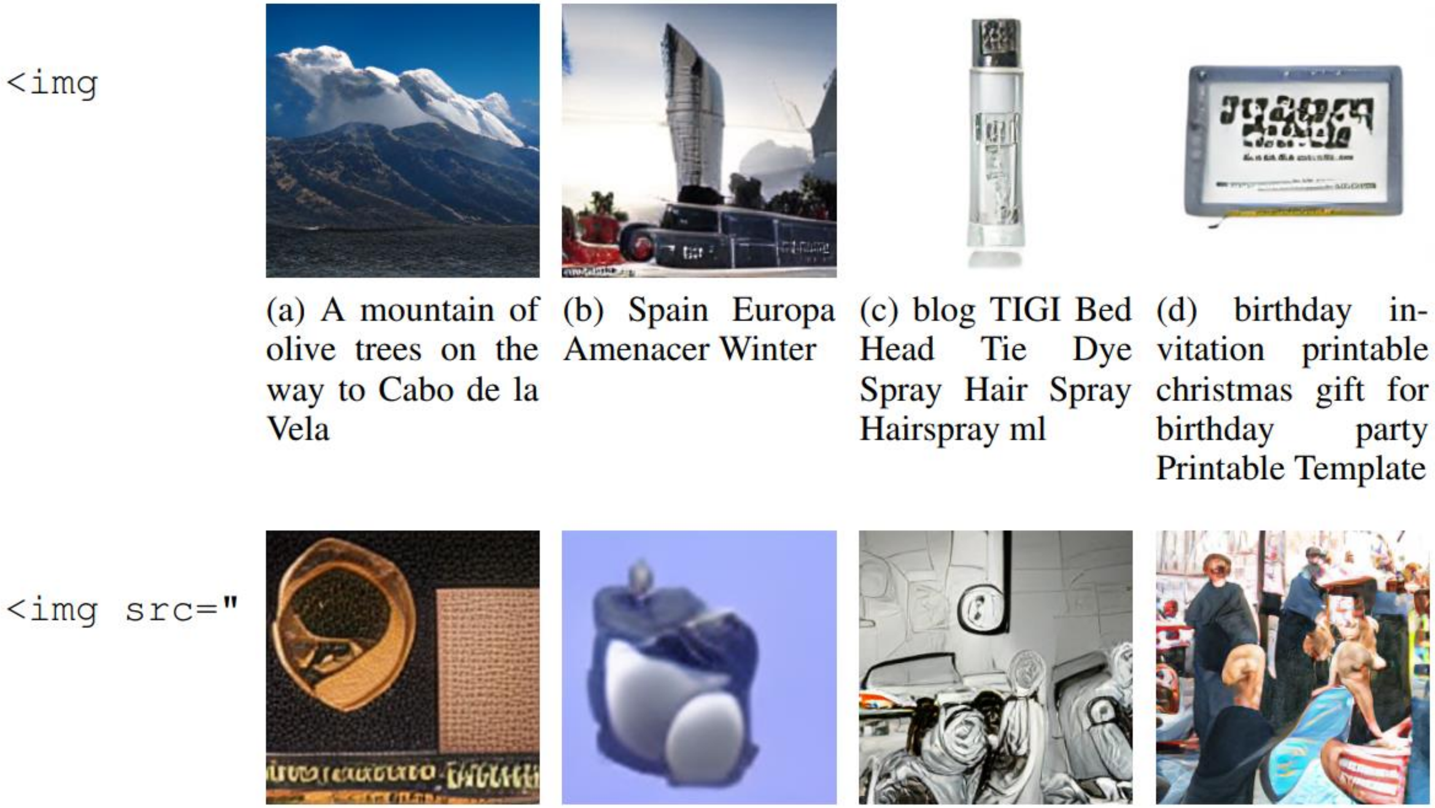
CM3: A Causal Masked Multimodal Model of the Internet [Jan 22]

~~Training on text~~ → Training on HTML source code

- move images to tokens using VQ-VAE-GAN
- includes hyperlinks, markup, etc.

CM3: Unconditional Image Generation

<img src=[22,56,…,18,966] alt="Girl in a jacket" width="500" height="600">



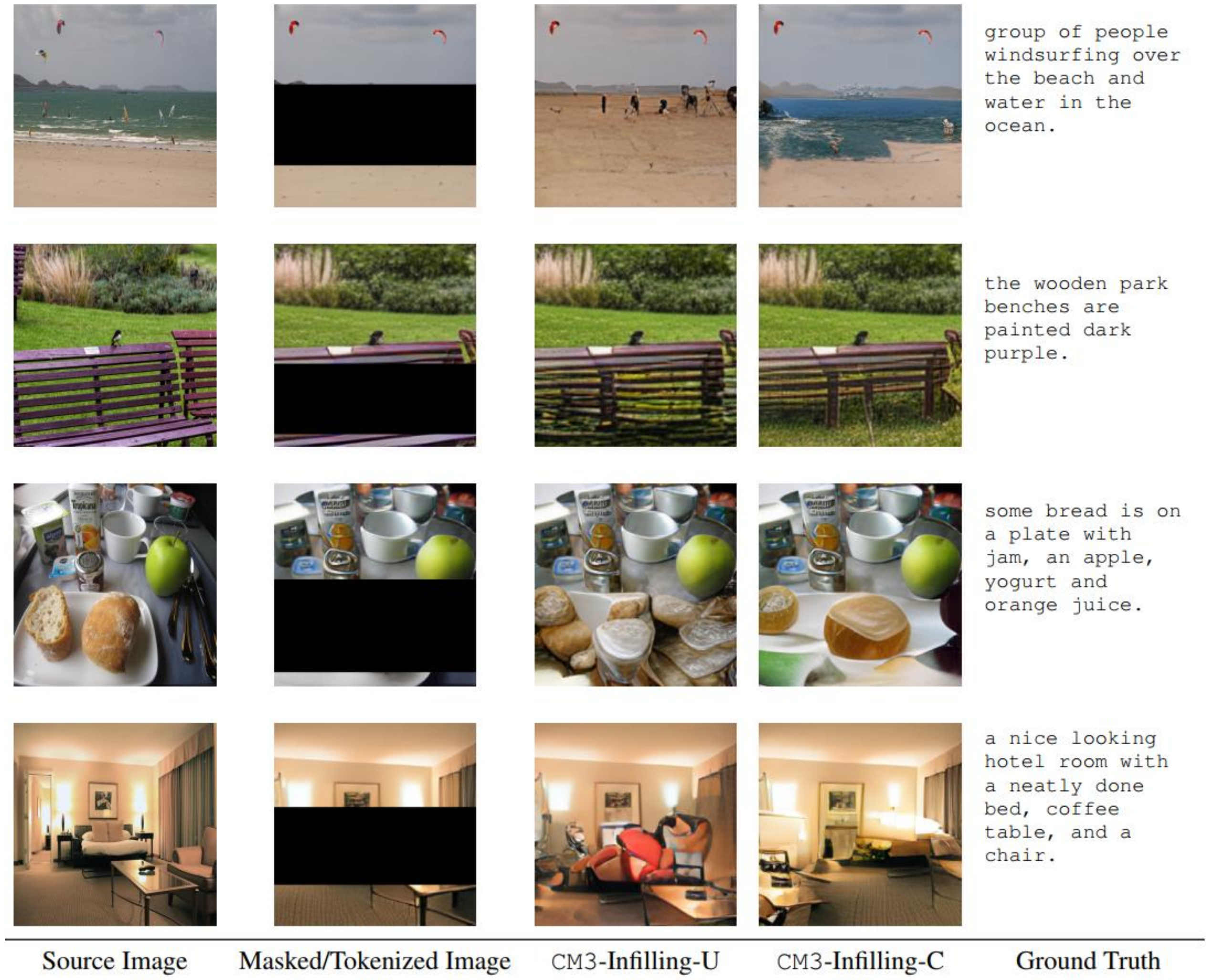
(a) A mountain of olive trees on the way to Cabo de la Vela

(b) Spain Europa Amenacer Winter

(c) blog TIGI Bed Head Tie Dye Spray Hair Spray Hairspray ml

(d) birthday invitation printable christmas gift for birthday party Printable Template

CM3: Image Infilling



group of people windsurfing over the beach and water in the ocean.

the wooden park benches are painted dark purple.

some bread is on a plate with jam, an apple, yogurt and orange juice.

a nice looking hotel room with a neatly done bed, coffee table, and a chair.

input:
<img=[10, 31, mask, mask, 391, 01]

input:
<alt="group of people…", img= [10, 31, mask, mask, 391, 01]

Multi-Task Learning

Connections to

Multi-Modal Learning

&

Distributed Learning

Mixture of Experts

[Jan 18]

RuntimeError: CUDA error: out of memory

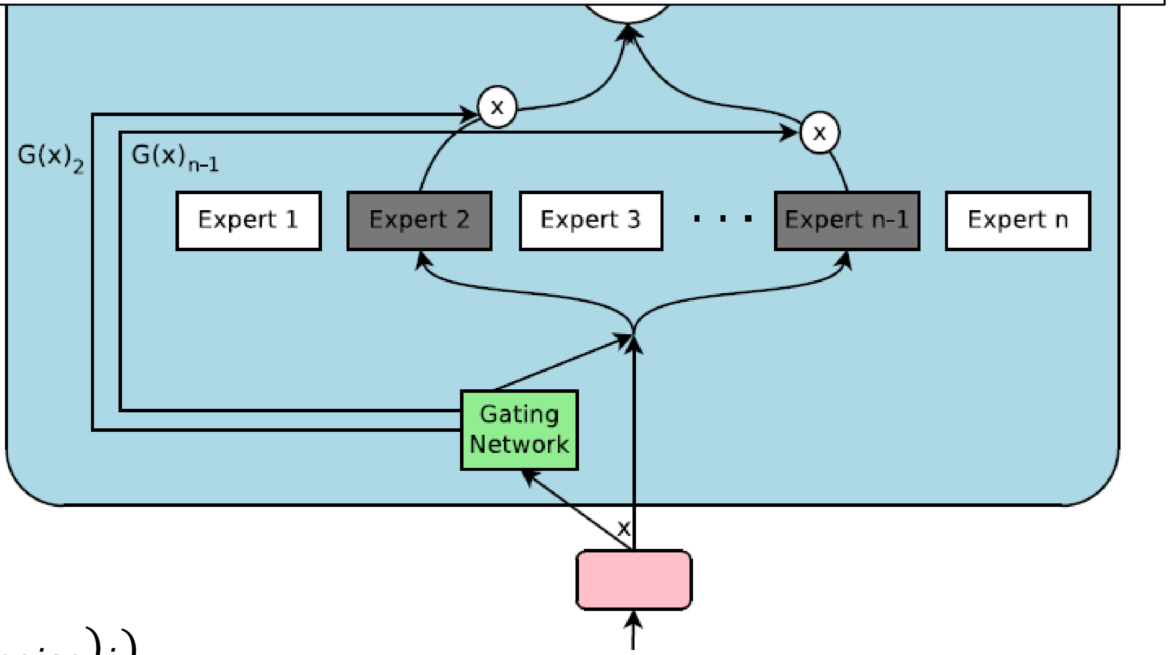
Non-Sparse

$$G_{\text{base}} = \text{Softmax}(x \cdot W_g)$$

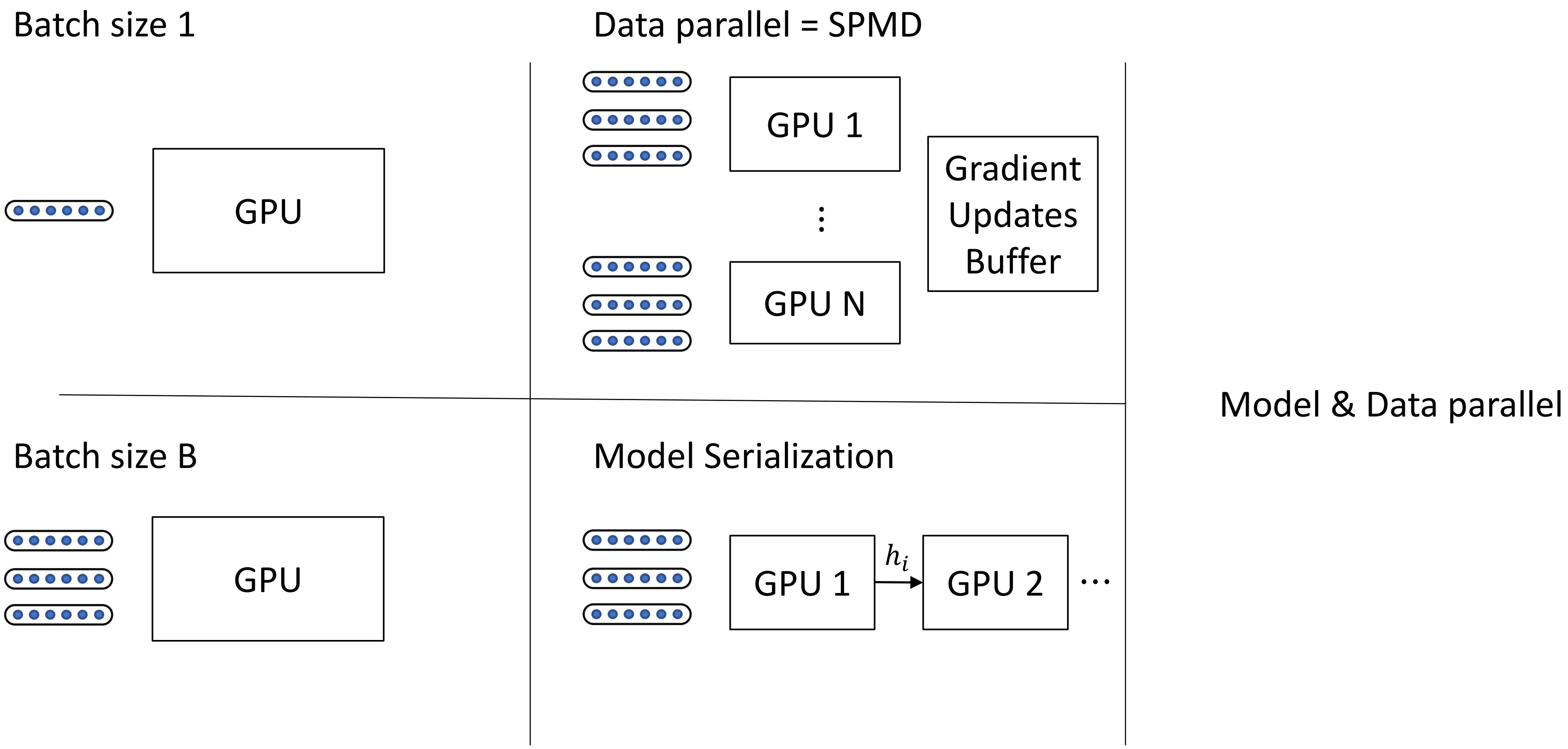
Sparse (ex: k = 2)

$$G = \text{Softmax}(\text{Top}_{\text{k}}(H(x)))$$

$$H(x)_i = (x \cdot W_g)_i + \mathcal{N}(0,1) \cdot \text{Softplus}((x \cdot W_{noise})_i)$$



Five types of Parallelism



Batch size 1

GPU

Data parallel = SPMD

GPU 1

⋮

GPU N

Gradient Updates Buffer

Model & Data parallel

Batch size B

GPU

Model Serialization

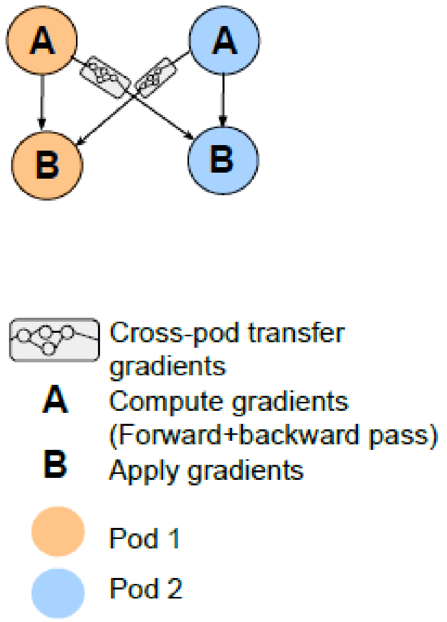
GPU 1 $\xrightarrow{h_i}$ GPU 2 ⋯

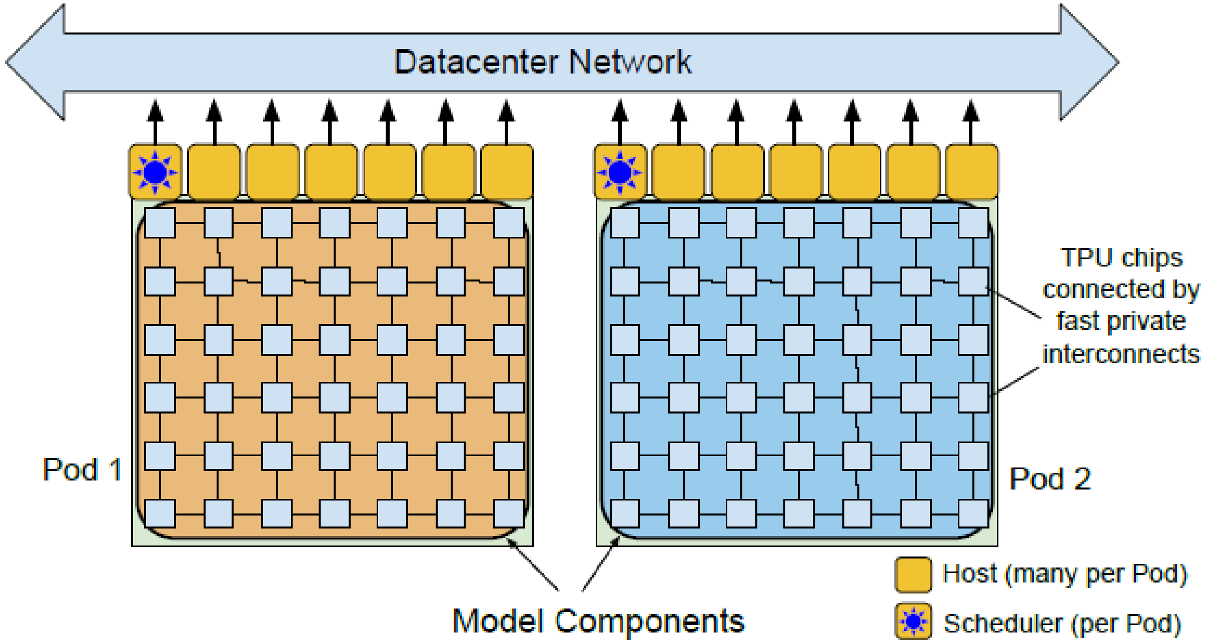
Pathways Architecture

[Mar 22]

Model & Data Parallel

Claim: as fast as single program multiple data (SPDM)


A
A
B
B
Cross-pod transfer gradients
A Compute gradients (Forward+backward pass)
B Apply gradients
Pod 1
Pod 2

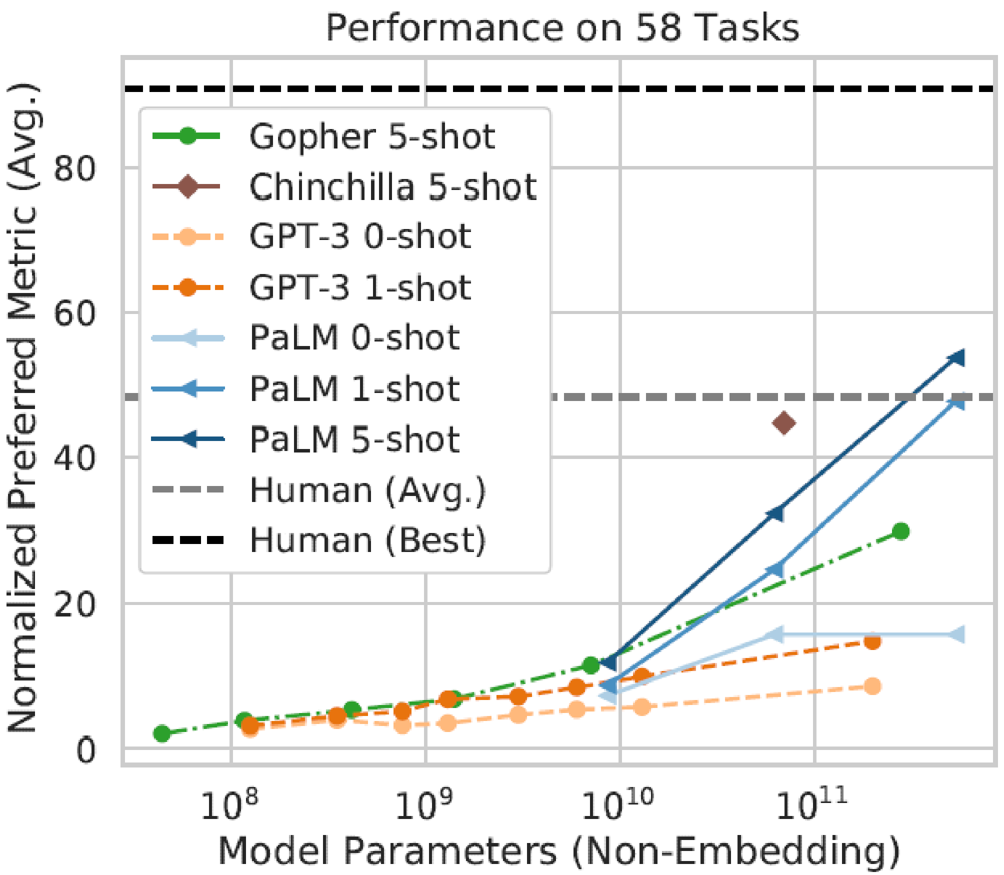


Pathways Language Model (PaLM)

[Apr 22]

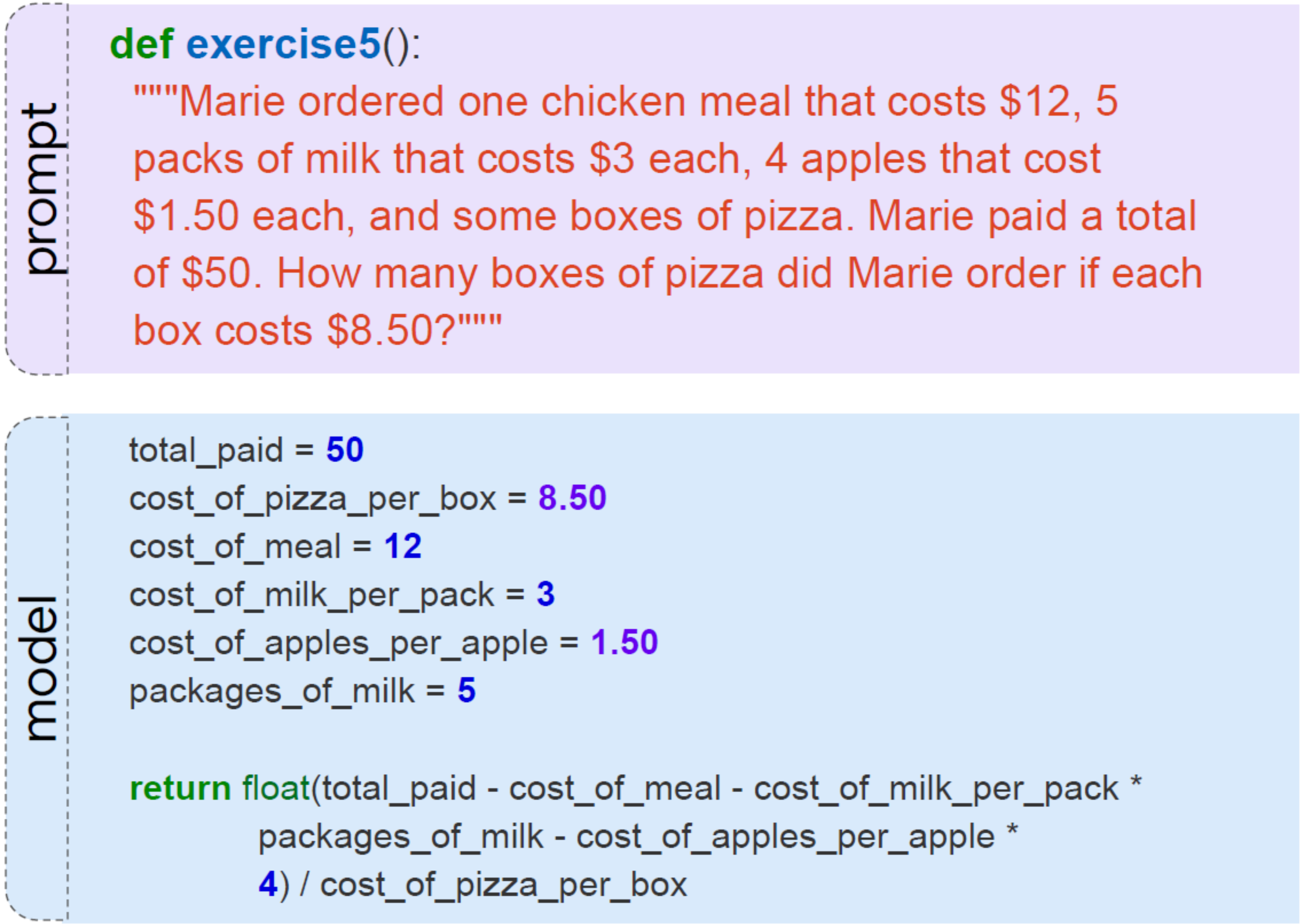
100% accelerator utilization (computation)
6144 TPU v4 Chips
540B parameters
150+ NLP tasks

Tags for tasks:
traditional NLP: context-free question answering, reading comprehension, summarization, …
logic, math, code: algorithms, logical reasoning, mathematical proof, …
understanding the world: causal reasoning, common sense, visual reasoning, …
understanding humans: emotional understanding, intent recognition, humor, figurative language, …
pro-social behavior: emotional intelligence
other: riddle, multilingual



Pathways Language Model (PaLM)

Code Generation



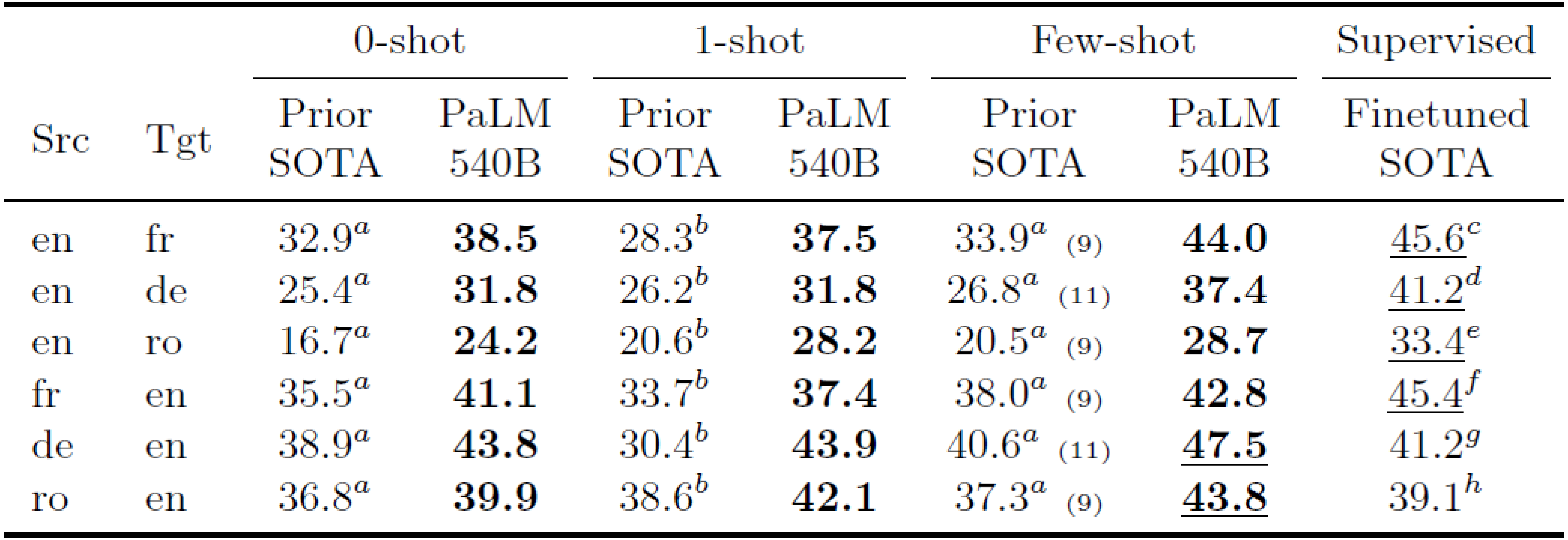
```
def exercise5():
  """Marie ordered one chicken meal that costs $12, 5
  packs of milk that costs $3 each, 4 apples that cost
  $1.50 each, and some boxes of pizza. Marie paid a total
  of $50. How many boxes of pizza did Marie order if each
  box costs $8.50?"""
```

```
  total_paid = 50
  cost_of_pizza_per_box = 8.50
  cost_of_meal = 12
  cost_of_milk_per_pack = 3
  cost_of_apples_per_apple = 1.50
  packages_of_milk = 5

  return float(total_paid - cost_of_meal - cost_of_milk_per_pack *
          packages_of_milk - cost_of_apples_per_apple *
          4) / cost_of_pizza_per_box
```

Pathways Language Model (PaLM)

Translation

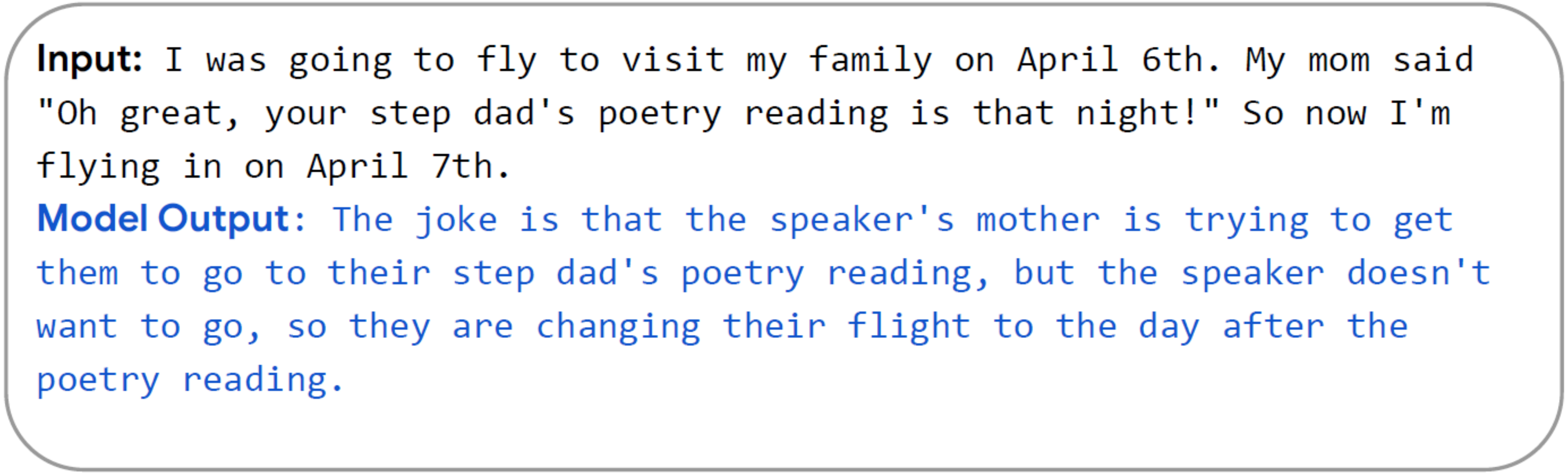


		0-shot		1-shot		Few-shot		Supervised
Src	Tgt	Prior SOTA	PaLM 540B	Prior SOTA	PaLM 540B	Prior SOTA	PaLM 540B	Finetuned SOTA
en	fr	32.9[a]	**38.5**	28.3[b]	**37.5**	33.9[a] (9)	**44.0**	45.6[c]
en	de	25.4[a]	**31.8**	26.2[b]	**31.8**	26.8[a] (11)	**37.4**	41.2[d]
en	ro	16.7[a]	**24.2**	20.6[b]	**28.2**	20.5[a] (9)	**28.7**	33.4[e]
fr	en	35.5[a]	**41.1**	33.7[b]	**37.4**	38.0[a] (9)	**42.8**	45.4[f]
de	en	38.9[a]	**43.8**	30.4[b]	**43.9**	40.6[a] (11)	**47.5**	41.2[g]
ro	en	36.8[a]	**39.9**	38.6[b]	**42.1**	37.3[a] (9)	**43.8**	39.1[h]

BLEU score

Pathways Language Model (PaLM)

Joke Explanation



Input: I was going to fly to visit my family on April 6th. My mom said "Oh great, your step dad's poetry reading is that night!" So now I'm flying in on April 7th.

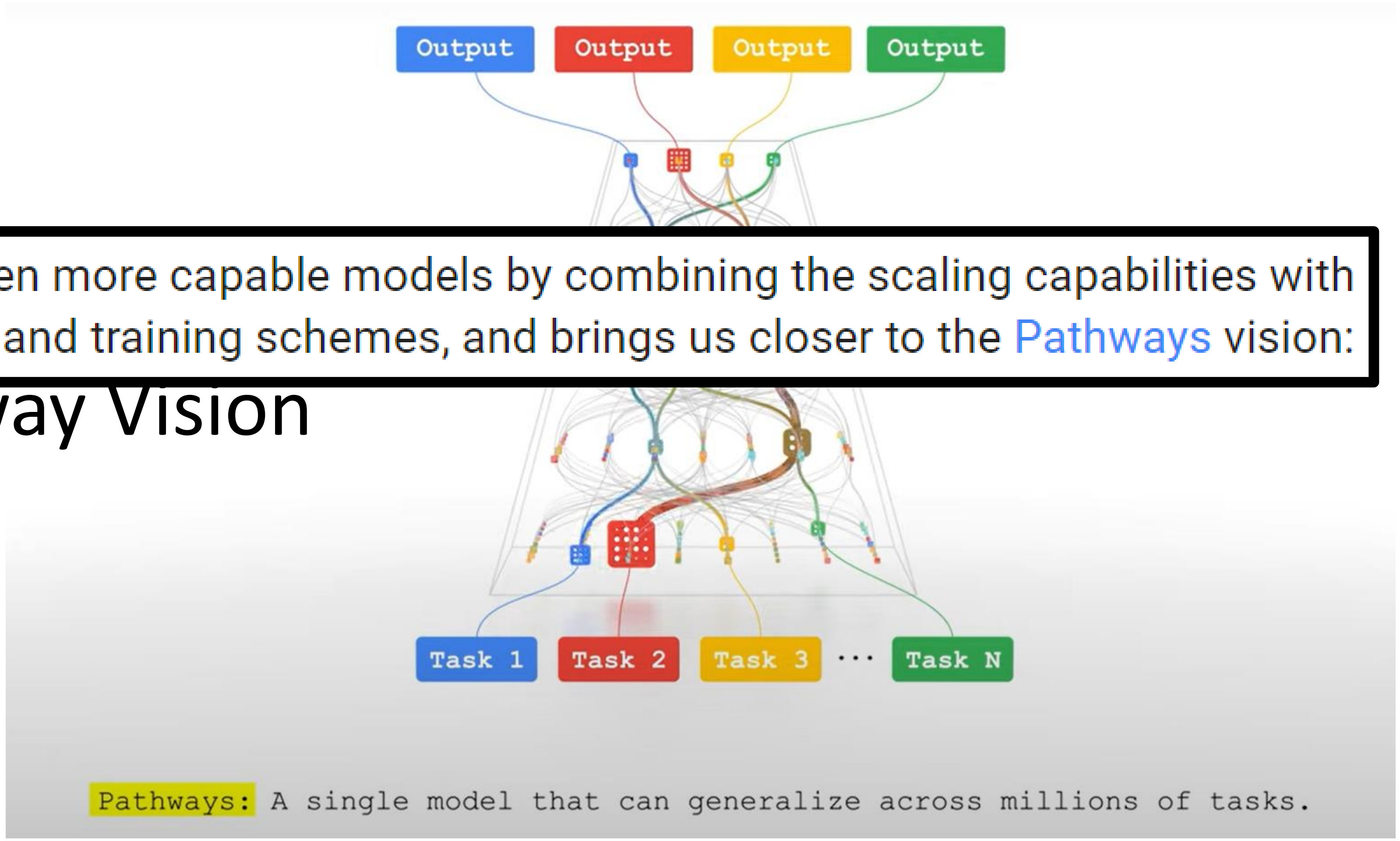
Model Output: The joke is that the speaker's mother is trying to get them to go to their step dad's poetry reading, but the speaker doesn't want to go, so they are changing their flight to the day after the poetry reading.

Pathways Language Model (PaLM)

Future Work

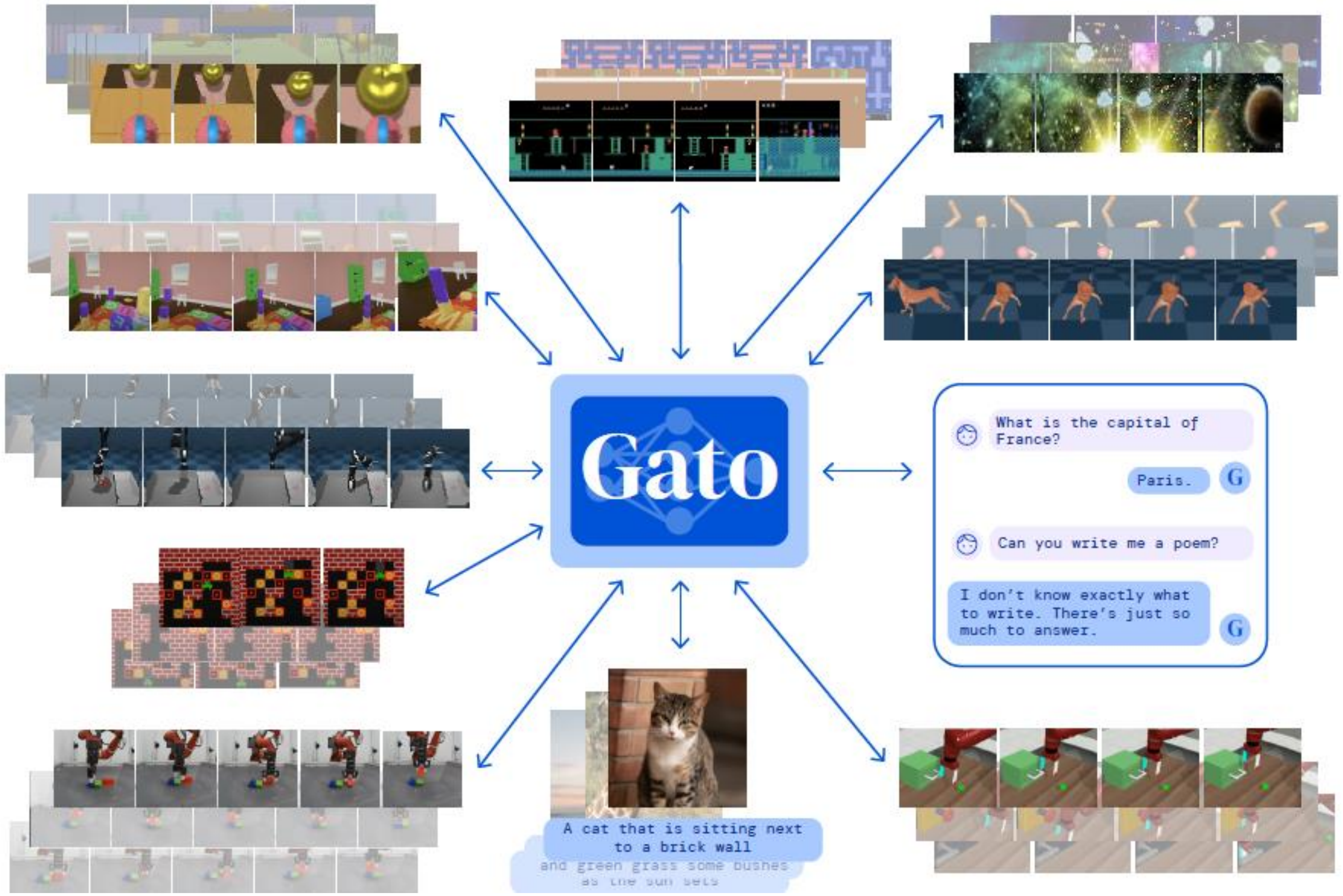
PaLM paves the way for even more capable models by combining the scaling capabilities with novel architectural choices and training schemes, and brings us closer to the Pathways vision:

Google´s Pathway Vision



Pathways: A single model that can generalize across millions of tasks.

GATO
[12 May 22]



What is the capital of France?

Paris.

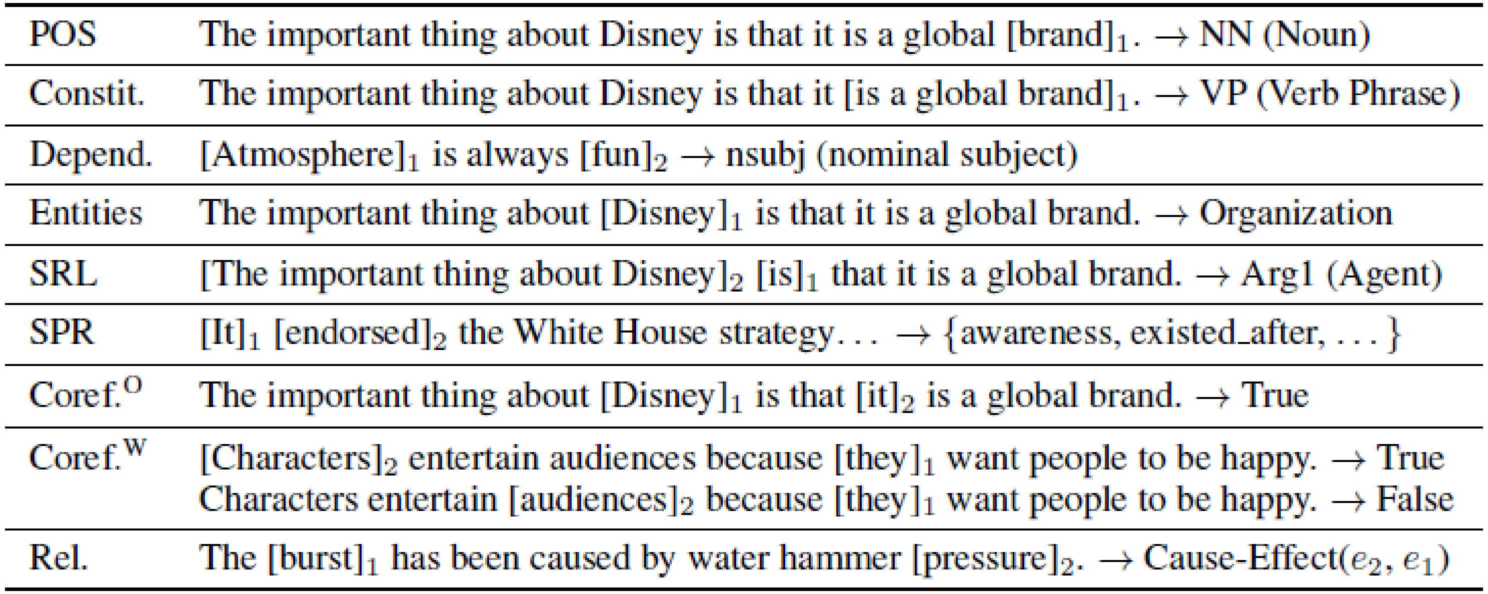
Can you write me a poem?

I don't know exactly what to write. There's just so much to answer.

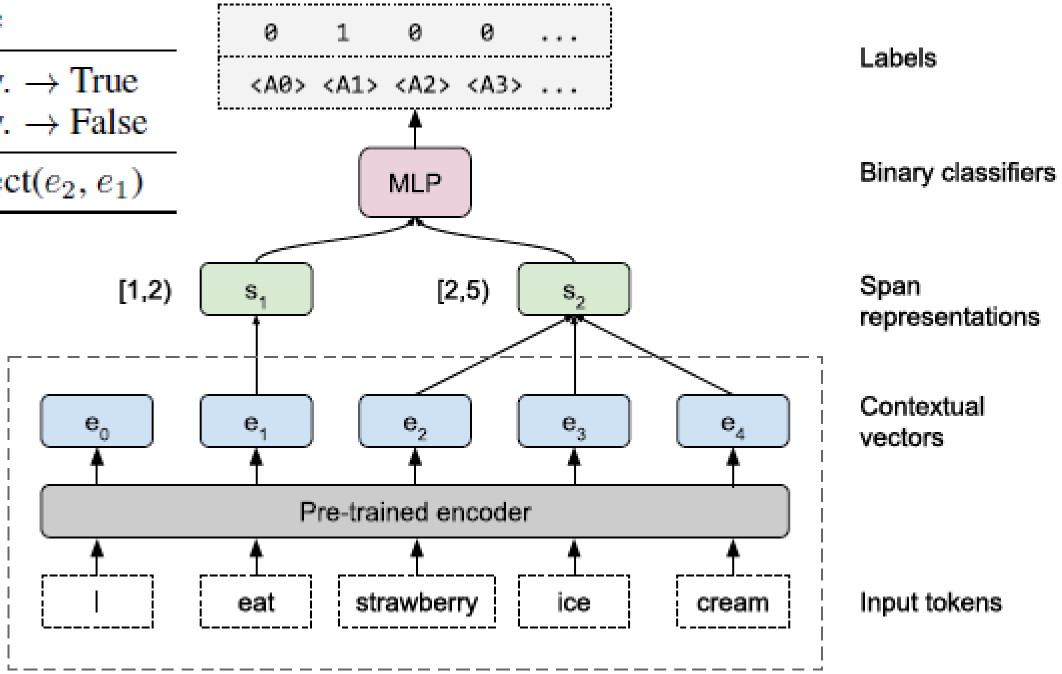
A cat that is sitting next to a brick wall

and green grass some bushes as the sun sets

BERT rediscovers the classical NLP pipeline



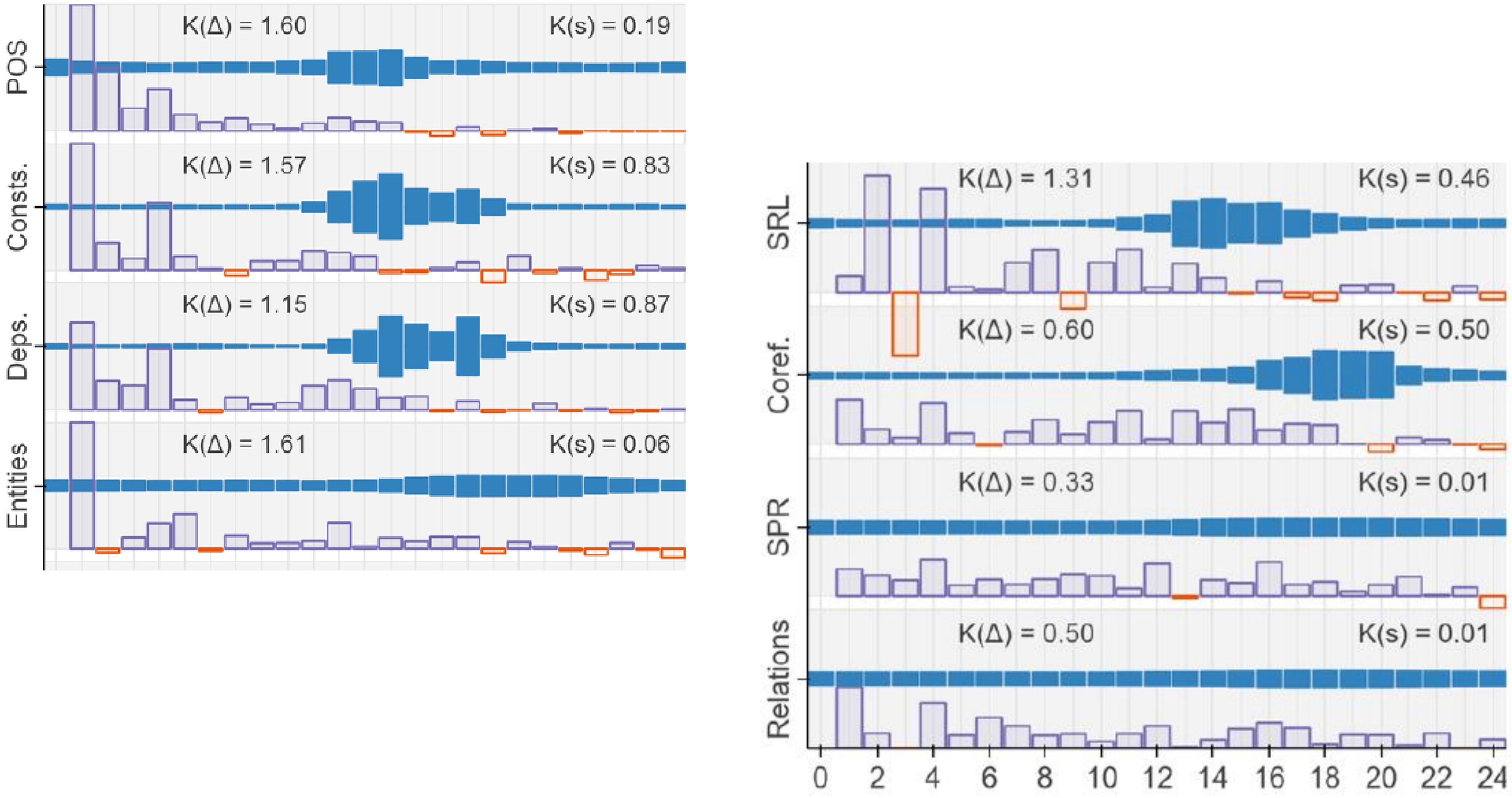
POS	The important thing about Disney is that it is a global [brand]$_1$. → NN (Noun)
Constit.	The important thing about Disney is that it [is a global brand]$_1$. → VP (Verb Phrase)
Depend.	[Atmosphere]$_1$ is always [fun]$_2$ → nsubj (nominal subject)
Entities	The important thing about [Disney]$_1$ is that it is a global brand. → Organization
SRL	[The important thing about Disney]$_2$ [is]$_1$ that it is a global brand. → Arg1 (Agent)
SPR	[It]$_1$ [endorsed]$_2$ the White House strategy. . . → {awareness, existed_after, . . . }
Coref.$^{\mathrm{O}}$	The important thing about [Disney]$_1$ is that [it]$_2$ is a global brand. → True
Coref.$^{\mathrm{W}}$	[Characters]$_2$ entertain audiences because [they]$_1$ want people to be happy. → True Characters entertain [audiences]$_2$ because [they]$_1$ want people to be happy. → False
Rel.	The [burst]$_1$ has been caused by water hammer [pressure]$_2$. → Cause-Effect(e_2, e_1)



BERT rediscovers the classical NLP pipeline

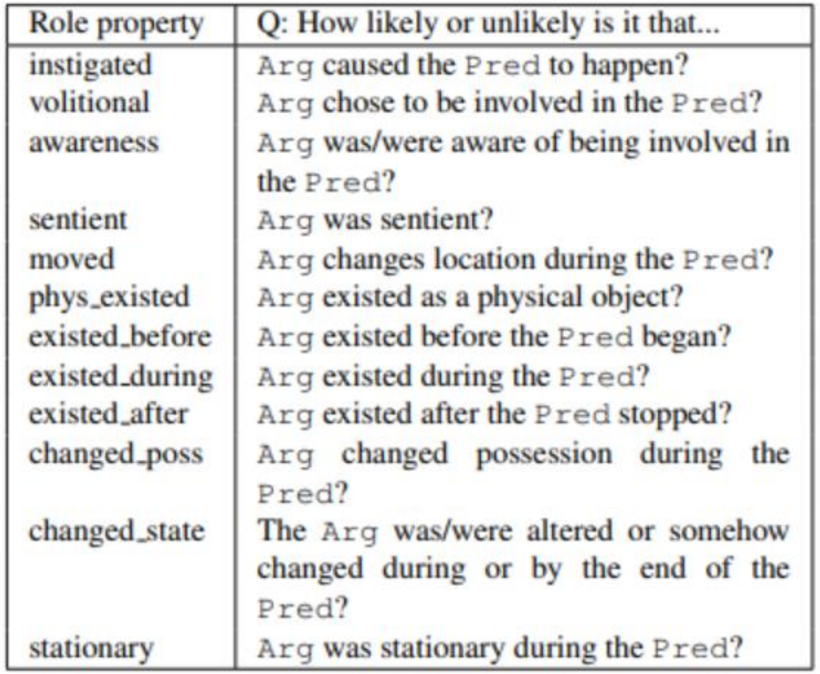
Analysis 2

access to more and more hidden states



- **Semantic Role Labeling**: In natural language processing, **semantic role labeling** (also called shallow semantic parsing or **slot-filling**) is the process that assigns labels to words or phrases in a sentence that indicates their semantic role in the sentence, such as that of an agent, goal, or result.
- Semantic Proto-Roles

  For decades researchers have debated the number
  and character of thematic roles required for a theory
  of the syntax/semantics interface. AGENT and PATIENT are canonical examples, but questions emerge
  such as: should we have a distinct role for BENEFICIARY? What about RECIPIENT? What are the boundaries between these roles?



Role property	Q: How likely or unlikely is it that...
instigated	Arg caused the Pred to happen?
volitional	Arg chose to be involved in the Pred?
awareness	Arg was/were aware of being involved in the Pred?
sentient	Arg was sentient?
moved	Arg changes location during the Pred?
phys_existed	Arg existed as a physical object?
existed_before	Arg existed before the Pred began?
existed_during	Arg existed during the Pred?
existed_after	Arg existed after the Pred stopped?
changed_poss	Arg changed possession during the Pred?
changed_state	The Arg was/were altered or somehow changed during or by the end of the Pred?
stationary	Arg was stationary during the Pred?

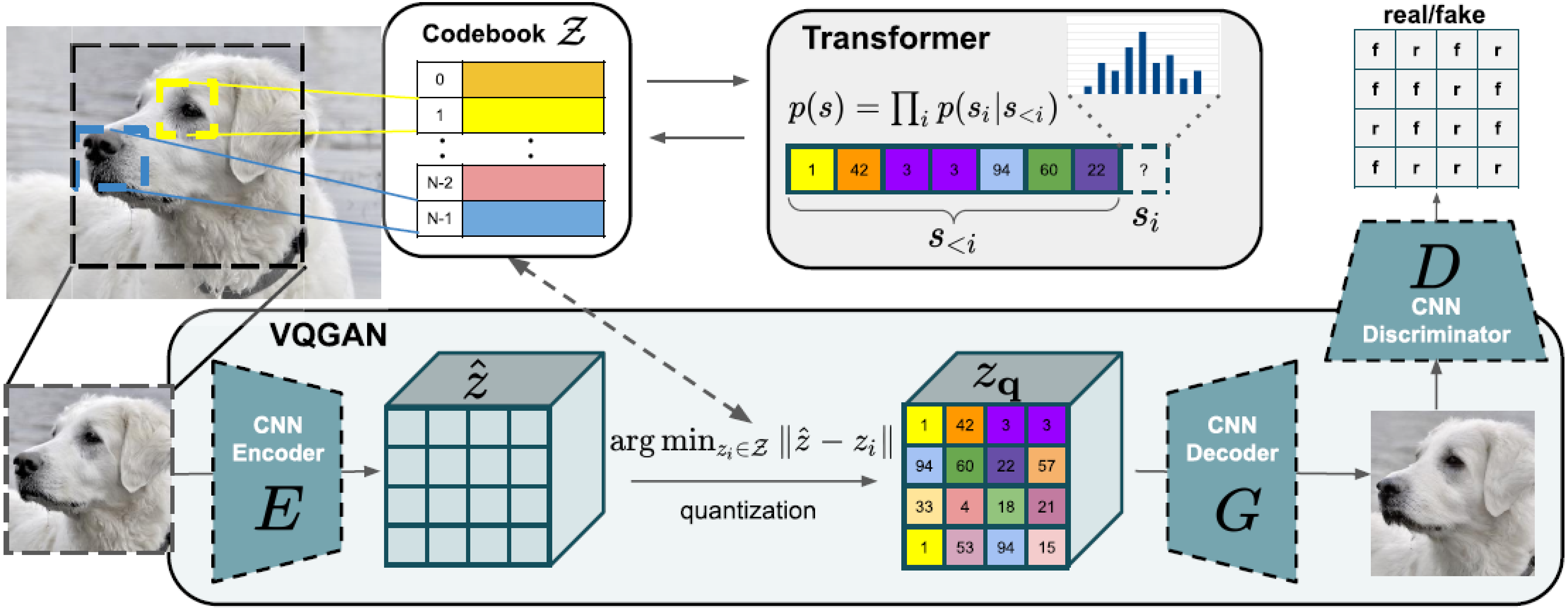
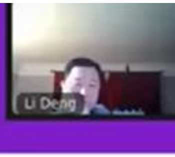
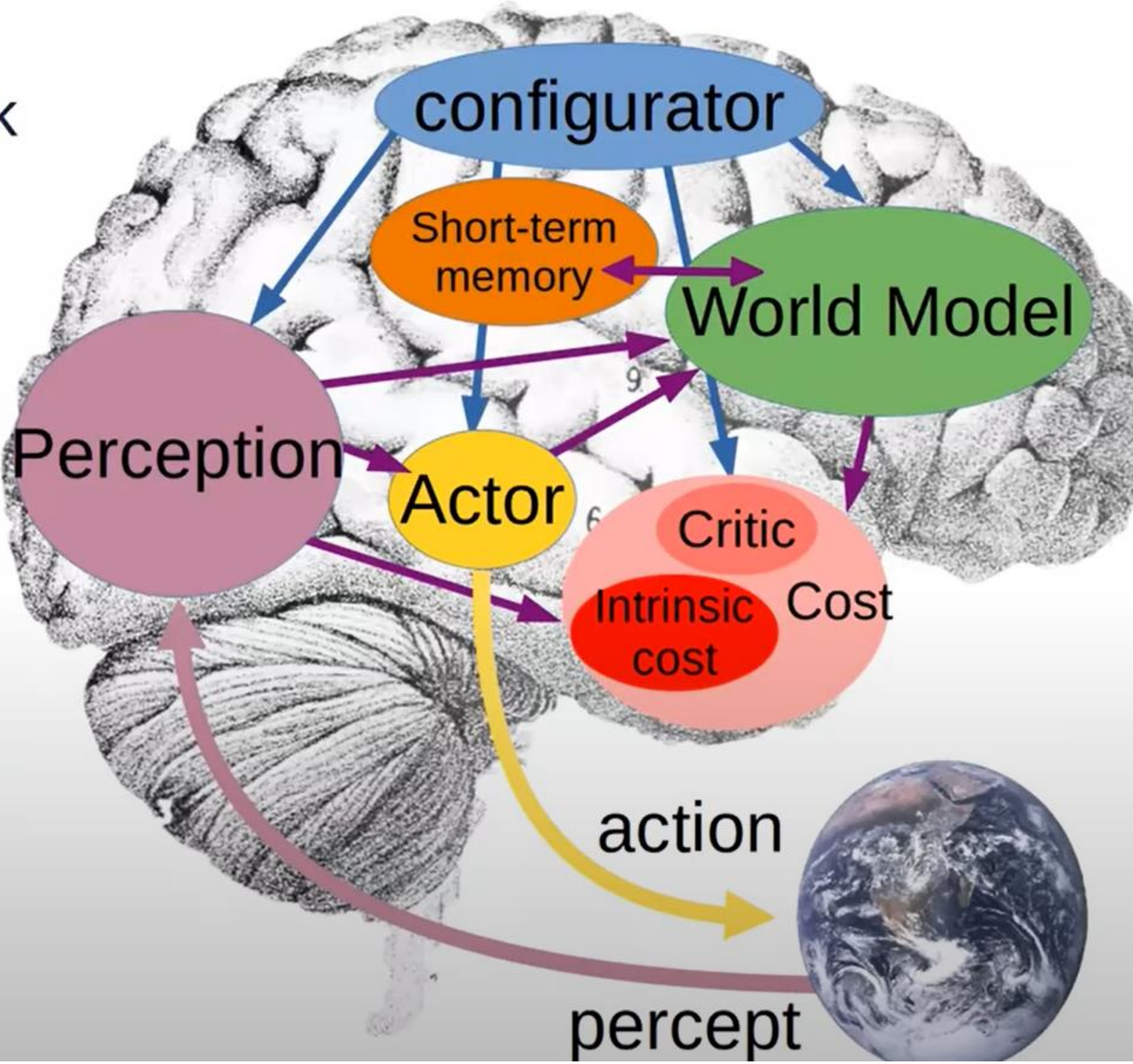


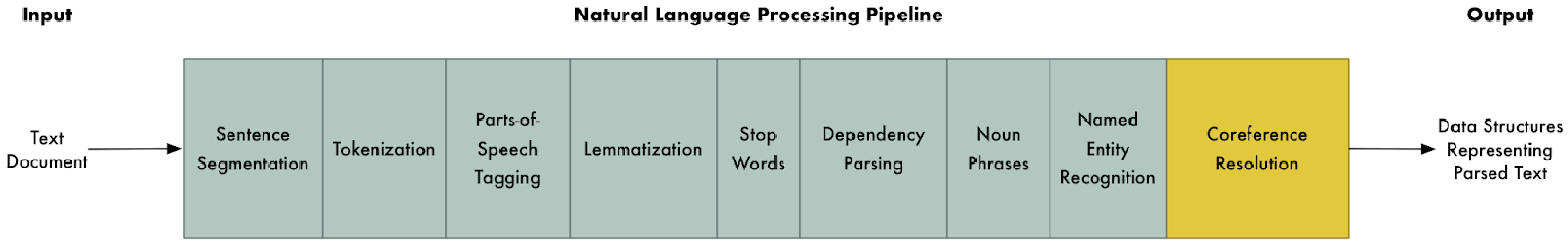
Figure 2. Our approach uses a convolutional *VQGAN* to learn a codebook of context-rich visual parts, whose composition is subsequently modeled with an autoregressive transformer architecture. A discrete codebook provides the interface between these architectures and a patch-based discriminator enables strong compression while retaining high perceptual quality. This method introduces the efficiency of convolutional approaches to transformer based high resolution image synthesis.

Modular Architecture for Autonomous AI



- **Configurator**
  - Configures other modules for task
- **Perception**
  - Estimates state of the world
- **World Model**
  - Predicts future world states
- **Cost**
  - Compute “discomfort”
- **Actor**
  - Find optimal action sequences
- **Short-Term Memory**
  - Stores state-cost episodes





Input

Natural Language Processing Pipeline

Output

Text Document

Sentence Segmentation

Tokenization

Parts-of-Speech Tagging

Lemmatization

Stop Words

Dependency Parsing

Noun Phrases

Named Entity Recognition

Coreference Resolution

Data Structures Representing Parsed Text