Stanford
Computer Science

CS372 Spring 2023

Foundation Model Lecture Series #1/4 LLMs and ChatGPT

Edward Y. Chang
Computer Science

Stanford
Computer Science

CS372 Spring 2023

Foundation Model Lecture Series
#2/4 Prompt Template Design Principles

Edward Y. Chang
Computer Science

Stanford
Computer Science

CS372 Spring 2023

Foundation Model Lecture Series
#3/4 History of NLP in One Lecture

Edward Y. Chang
Computer Science

Stanford
Computer Science

CS372 Spring 2023

Foundation Model Lecture Series
#4/4 Virtual Assistant and Augmented Foundation Models

Edward Y. Chang
Computer Science

Search Architecture

Google

Page Ranking

Indexing Module, words + semantics

Knowledge, represented by a huge set of documents

ChatGPT Architecture

ChatGPT

Foundation Model(s): e.g., GPT4

Attention Mechanism, polysemy w/ context

Word2vec, similarity

Knowledge, represented by a huge set of documents

Similarity, or resemblance, is a relationship between objects that determines how much they are alike. Similarity is contextual and varies in degrees, for example, Apple is more similar to Android than to Orange.

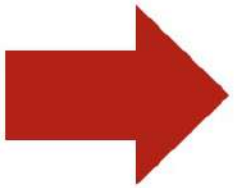
Context is essential for resolving polysemy.

Similarity

- Similarity is a central computation in cognition, as it plays a crucial role in integrating and processing the collected information from perception, attention, memory, and reasoning.
- By measuring the degree of resemblance between different objects or situations, similarity can help organize and categorize the information, enabling generalization and decision-making.
- However, it is essential to acknowledge that cognition is a complex, multi-faceted process. While similarity is undoubtedly a critical component in cognitive processing, the interplay of various cognitive functions allows humans and animals to navigate and adapt to their environment effectively.

Similarity between Word

Vocabulary:
Man, woman, boy, girl, prince, princess, queen, king, monarch



	1	2	3	4	5	6	7	8	9
man	1	0	0	0	0	0	0	0	0
woman	0	1	0	0	0	0	0	0	0
boy	0	0	1	0	0	0	0	0	0
girl	0	0	0	1	0	0	0	0	0
prince	0	0	0	0	1	0	0	0	0
princess	0	0	0	0	0	1	0	0	0
queen	0	0	0	0	0	0	1	0	0
king	0	0	0	0	0	0	0	1	0
monarch	0	0	0	0	0	0	0	0	1

Each word gets a 1x9 vector representation

2013 — **word2vec**

2014 — **seq2seq**

Seq2seq is a model that takes a **sequence** of items (words, letters, features of an images, and etc) and outputs another **sequence** of items.

2015 — **attention**

Attention allows the model **to focus on the relevant parts** of the input sequence as needed.

2017/18 — **self-attention, transformer**

In this year, **CNN replace RNN and transformer replace both of them**. As the model processes each word in the input sequence, self attention allows it to look at other positions in the input sequence for better encoding.

2018 — **GPT**

It takes advantage from transformer and is built **using transformer decoder** blocks.

2019 — **BERT**

BERT builds on top of a number of clever ideas that have been bubbling up in the NLP community since 2019– including but not limited to Semi-supervised Sequence Learning, ELMo, ULMFiT, the transformer. BERT is basically a trained **transformer encoder** stack.

1-of-N Encoding

apple = [1 0 0 0 0]

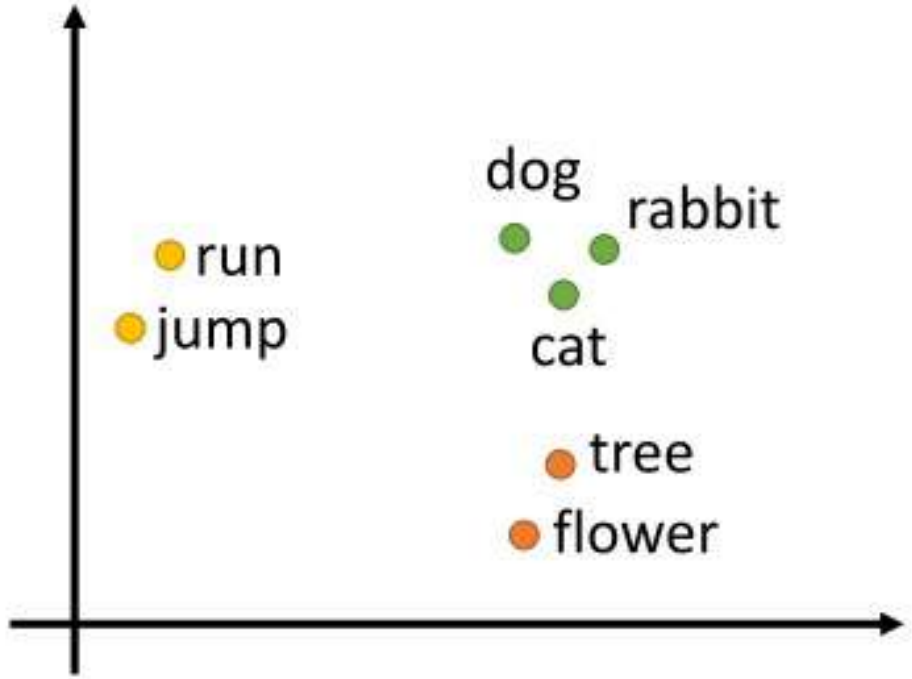
bag = [0 1 0 0 0]

cat = [0 0 1 0 0]

dog = [0 0 0 1 0]

elephant = [0 0 0 0 1]

Word Embedding



Word Class

Word Analogies

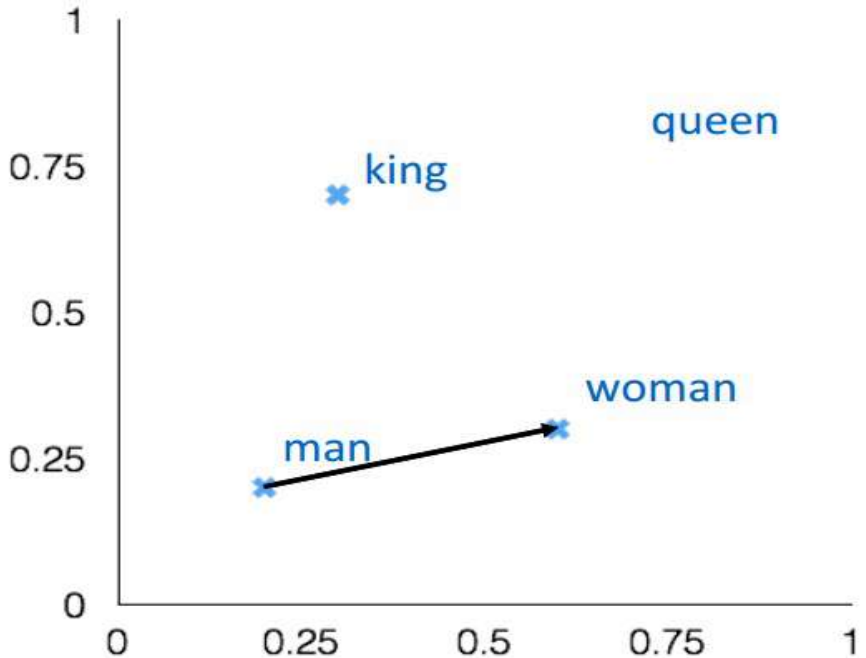
Test for linear relationships, examined by Mikolov et al. (2014)

a:b :: c:?

$$d = \arg\max_x \frac{(w_b - w_a + w_c)^T w_x}{||w_b - w_a + w_c||}$$

man:woman :: king:?

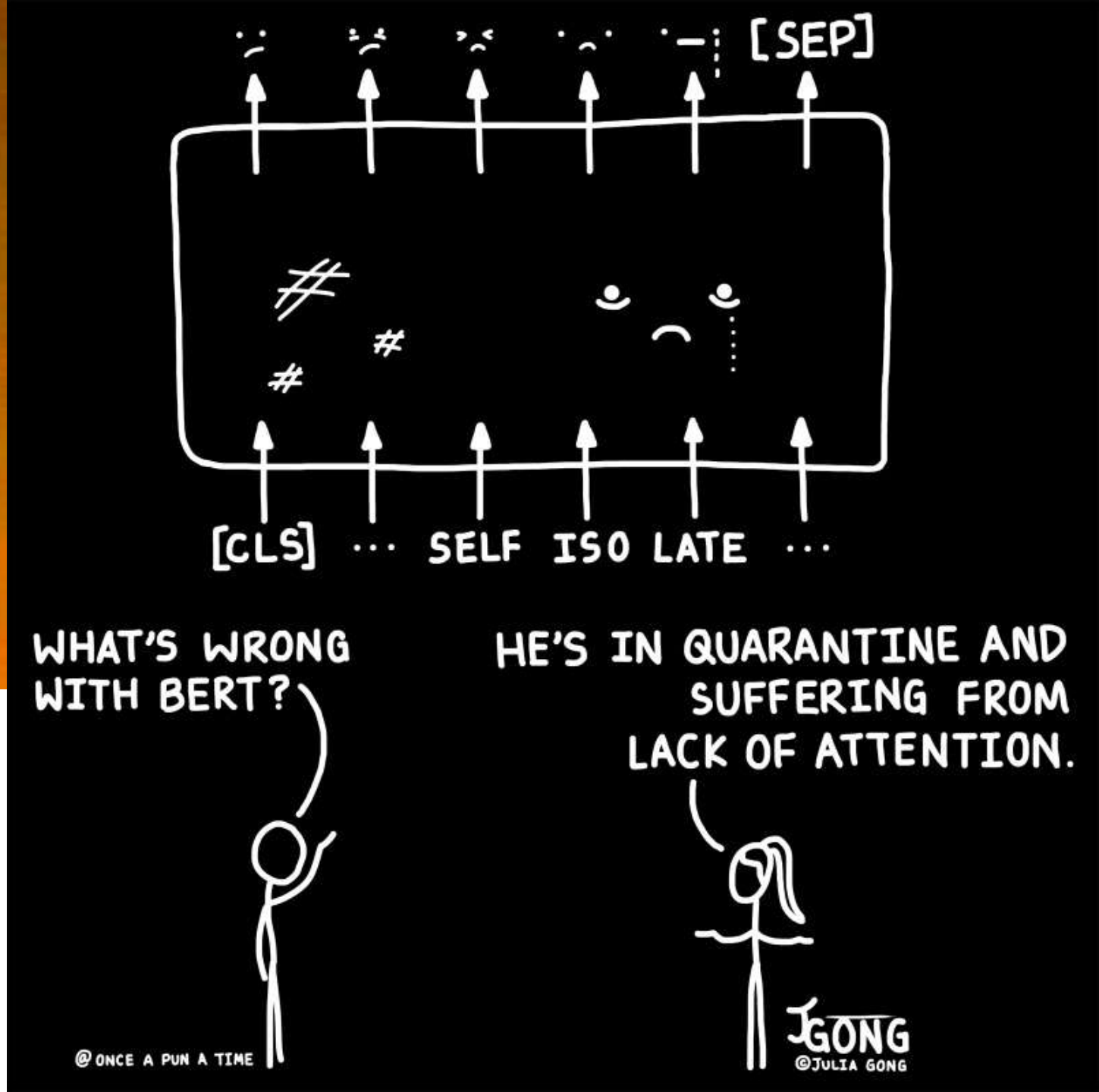
+	king	[0.30 0.70]
-	man	[0.20 0.20]
+	woman	[0.60 0.30]
	queen	[0.70 0.80]



Guiding Context Hypotheses
Linguists, 1954 - 57

"You shall know a word
by the company it keeps!"

J. R. Firth,
A synopsis of linguistic theory (1957)



Stanford Daily, 3/27/2020

[link] (credit: Julia Gong)

2013 — **word2vec**

2014 — **seq2seq**

Seq2seq is a model that takes a **sequence** of items (words, letters, features of an images, and etc) and outputs another **sequence** of items.

2015 — **attention**

Attention allows the model **to focus on the relevant parts** of the input sequence as needed.

2017/18 — **self-attention, transformer**

In this year, **CNN replace RNN and transformer replace both of them**. As the model processes each word in the input sequence, self attention allows it to look at other positions in the input sequence for better encoding.

2018 — **GPT**

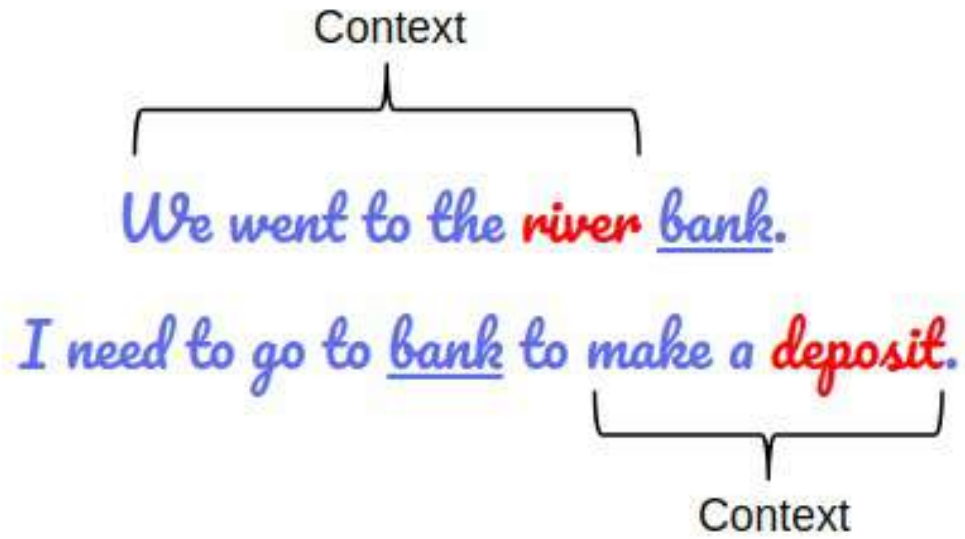
It takes advantage from transformer and is built **using transformer decoder** blocks.

2019 — **BERT**

BERT builds on top of a number of clever ideas that have been bubbling up in the NLP community since 2019– including but not limited to Semi-supervised Sequence Learning, ELMo, ULMFiT, the transformer. BERT is basically a trained **transformer encoder** stack.

In-Context Resolution

- Transformers w/ Attention Mechanism (2017)
- Attention Resolution in Context (2018)



The trophy cannot fit in the suitcase because it is too small. What is "it"?

In the statement "The trophy cannot fit in the suitcase because it is too small," the pronoun "it" refers to the suitcase. This means the suitcase is too small to accommodate the trophy.

The trophy cannot fit in the suitcase because it is too large. What is "it"?

In the statement "The trophy cannot fit in the suitcase because it is too large," the pronoun "it" refers to the trophy. This means the trophy is too large to fit in the suitcase.

The animal cannot enter the street because it is too narrow. What is "it"?

In the statement "The animal cannot enter the street because it is too narrow," the pronoun "it" refers to the street. This means the street is too narrow for the animal to enter.

The animal cannot enter the street because it is too wide. What is "it"?

In the statement "The animal cannot enter the street because it is too wide," the pronoun "it" refers to the street. This means the street is too wide for the animal to enter.

What is the difference between "enter the street" and "cross the street"?

"Enter the street" refers to moving onto or stepping into the street, typically from a different location like a sidewalk, yard, or building. It does not necessarily imply a specific intention to move from one side of the street to the other.

"Cross the street" specifically means moving from one side of the street to the other, typically by walking or running. This phrase implies a clear intention to reach the opposite side of the street, often by using crosswalks, pedestrian signals, or waiting for a break in traffic.

Common Sense Reasoning?

The animal cannot enter the street because it is too wide. What is "it"?

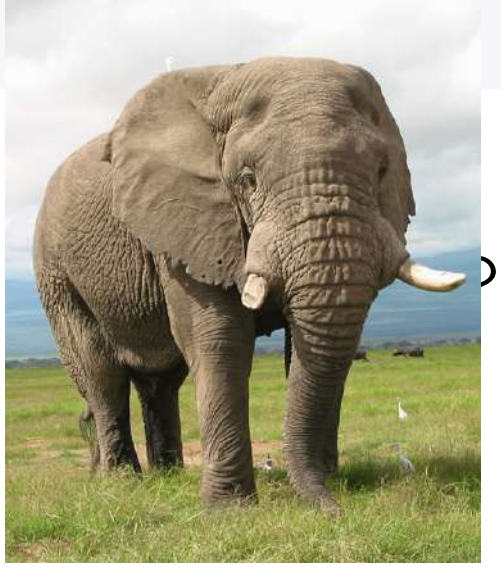
I apologize for any confusion earlier. In the sentence, "The animal cannot enter the street because it is too wide," the pronoun "it" refers to the street. The sentence implies that the street is too wide for the animal to enter or possibly cross safely.

When you previously answered "it" = "street", what was your reason?

I apologize for my previous incorrect response, and I appreciate your patience. My previous response was likely due to an error in parsing the sentence, which led me to interpret "it" as referring to the street instead of the animal. However, upon further reflection, it's clear that "it" should refer to the animal, as the sentence suggests that the animal cannot enter the street due to its size.

But if safety is considered, it is possible that your answer of "it is street" was reasonable.

You're correct that my previous response could be considered reasonable if we take safety into account. In that case, "it" could refer to the street being too wide, which might make it unsafe for the animal to enter or cross. However, without more context, it is difficult to determine which interpretation is more accurate. Both interpretations have merit, depending on the specific situation and additional information.

Replace "animal" with  or  ?

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

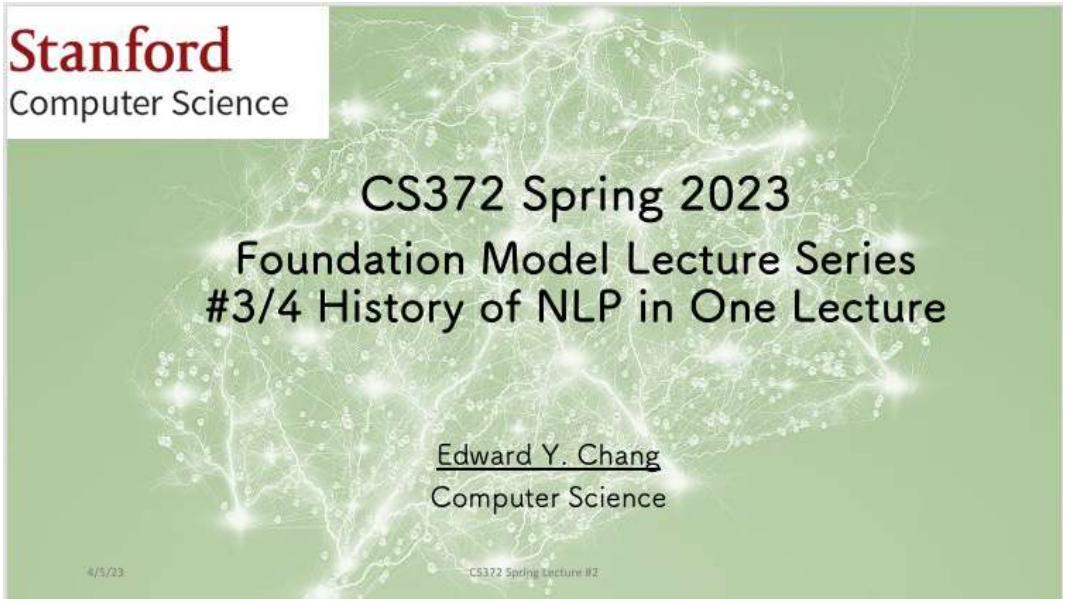
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

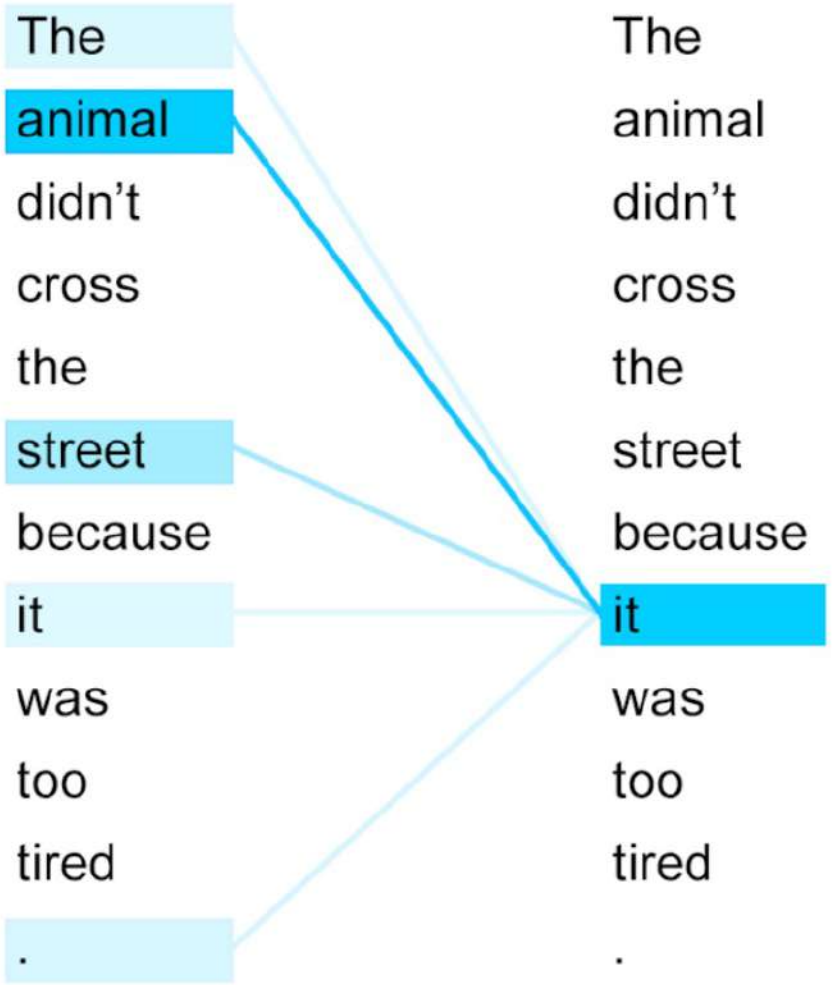
Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to

Key Steps of Transformer (4/17)

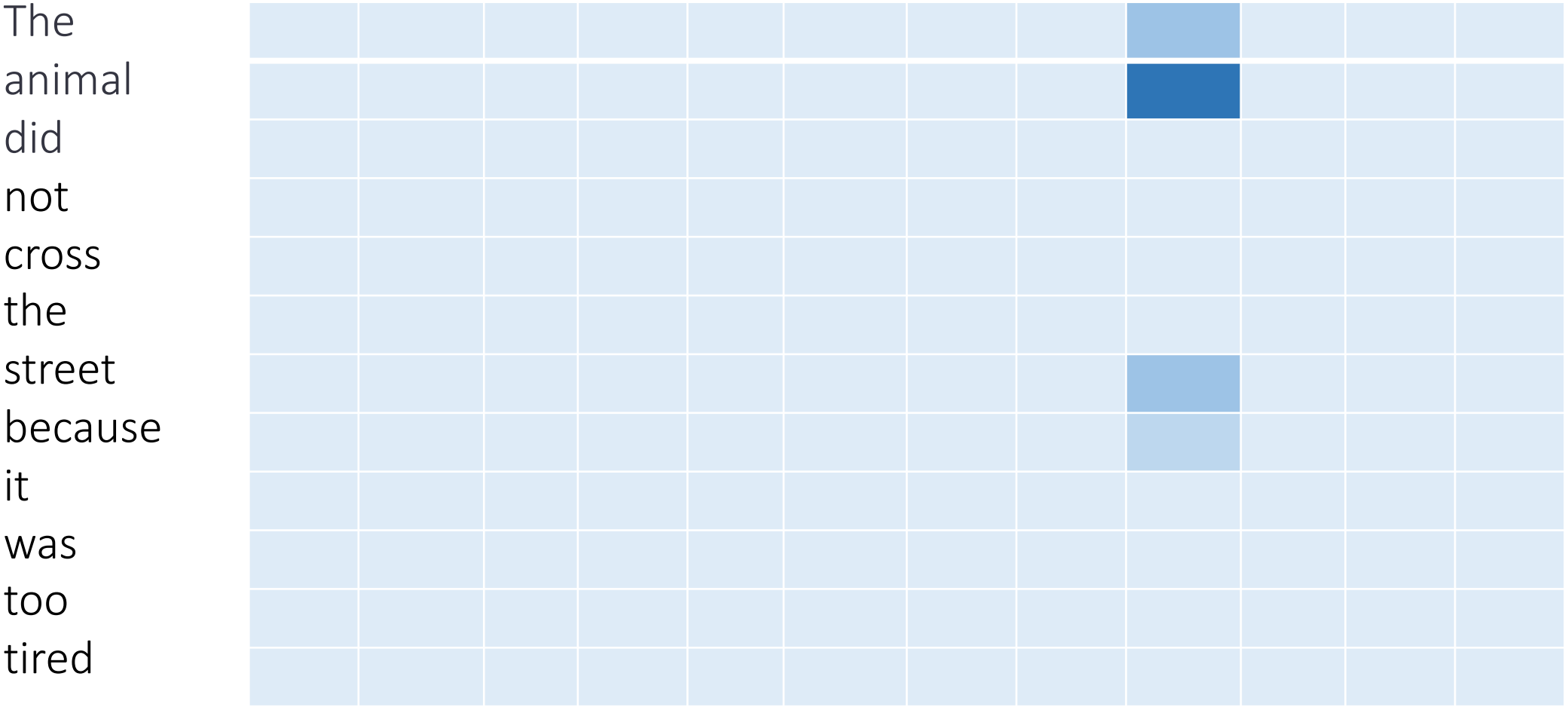
- Input: words
  - The trophy cannot fit in the suitcase because it is too large. What is "it"?
- Words embedding: word2vec
- Position embedding (order matters)
- Self-attention
  - Query (Q)
  - Key (K)
  - Value (V)
- $$\text{similarity}(Q, K) = \frac{Q \cdot K^T}{scaling}$$




The
animal
didn't
cross
the
street
because
it
was
too
tired
.
The
animal
didn't
cross
the
street
because
it
was
too
tired
.

Attention Window
= 12 words

The animal did not cross the street because it was too tired.

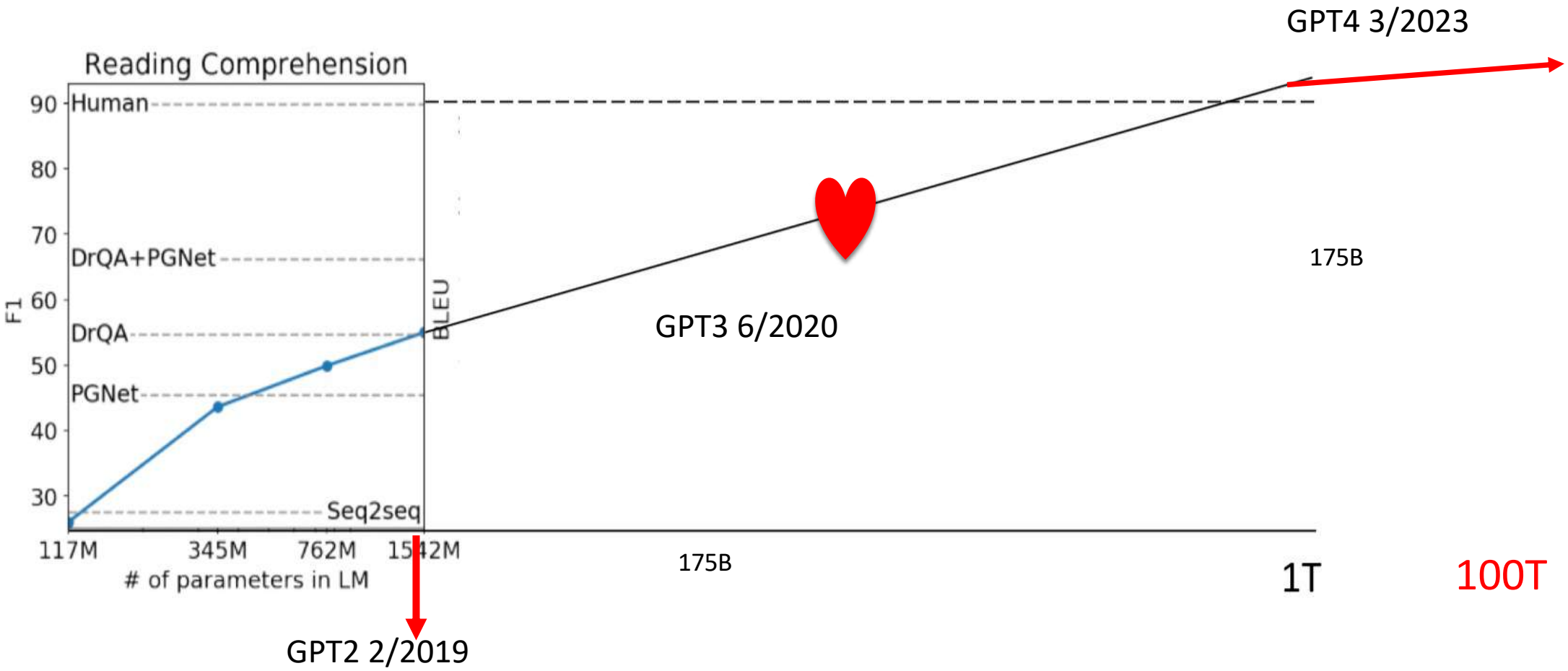


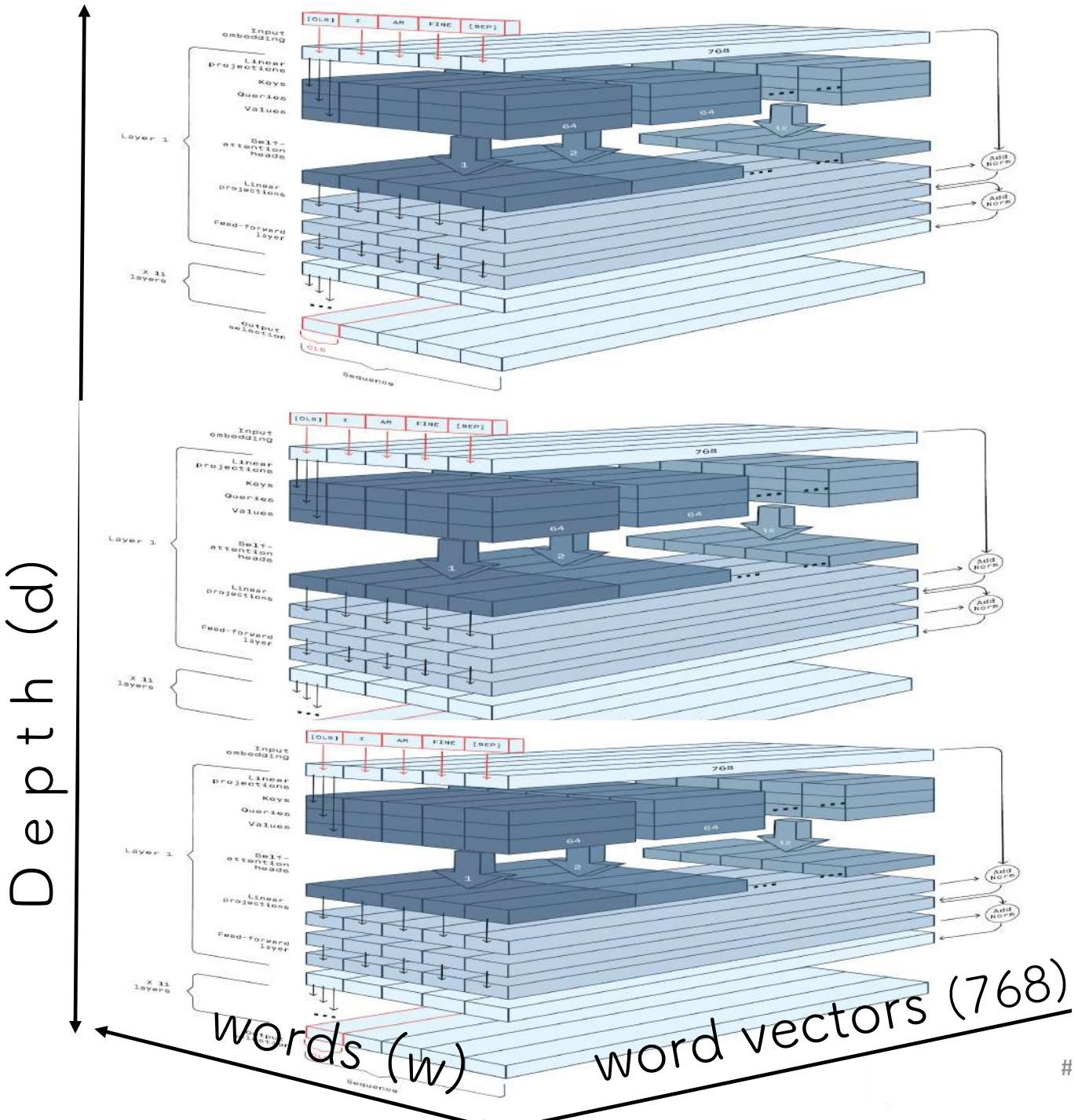
Attention Window, Context Width

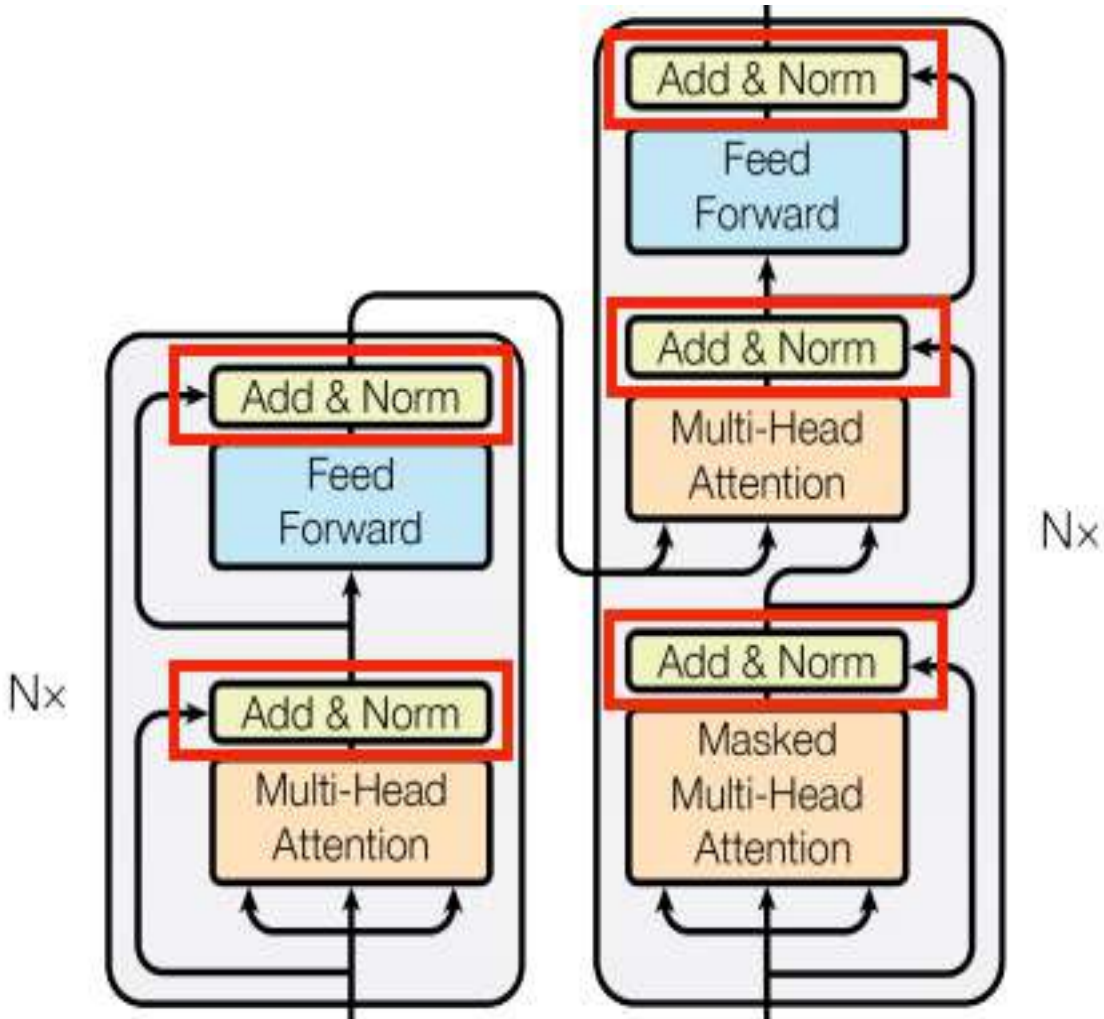
- After the dog chased its tail, it ran into the house and jumped on the couch, where it fell asleep for hours.

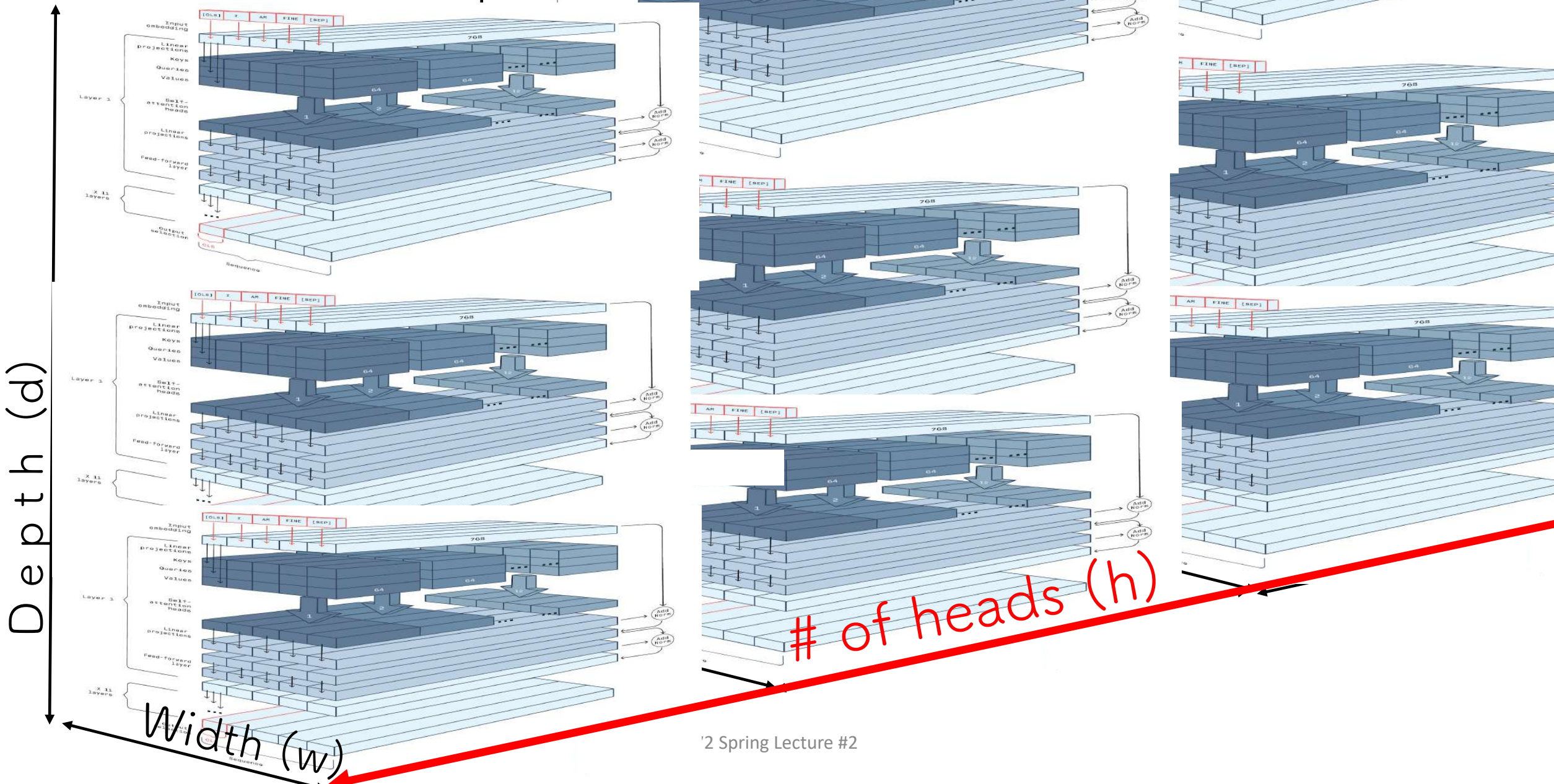
- After the dog chased its tail, it ran into the house and jumped on the couch, where it fell asleep for hours.

More Parameters, More Data → Higher Accuracy?




Depth (d)
words (w)
word vectors (768)
Input embedding
Linear projections
Keys
Queries
Values
Self-attention heads
Linear projections
Feed-forward layer
X 11 layers
Output selection
Sequence
Layer 1
[CLS]
I
AM
FINE
[SEP]
768
64
12
Add Norm


Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx


Depth (d)
Width (w)
# of heads (h)

Size Matters?
Sutton: the larger the better

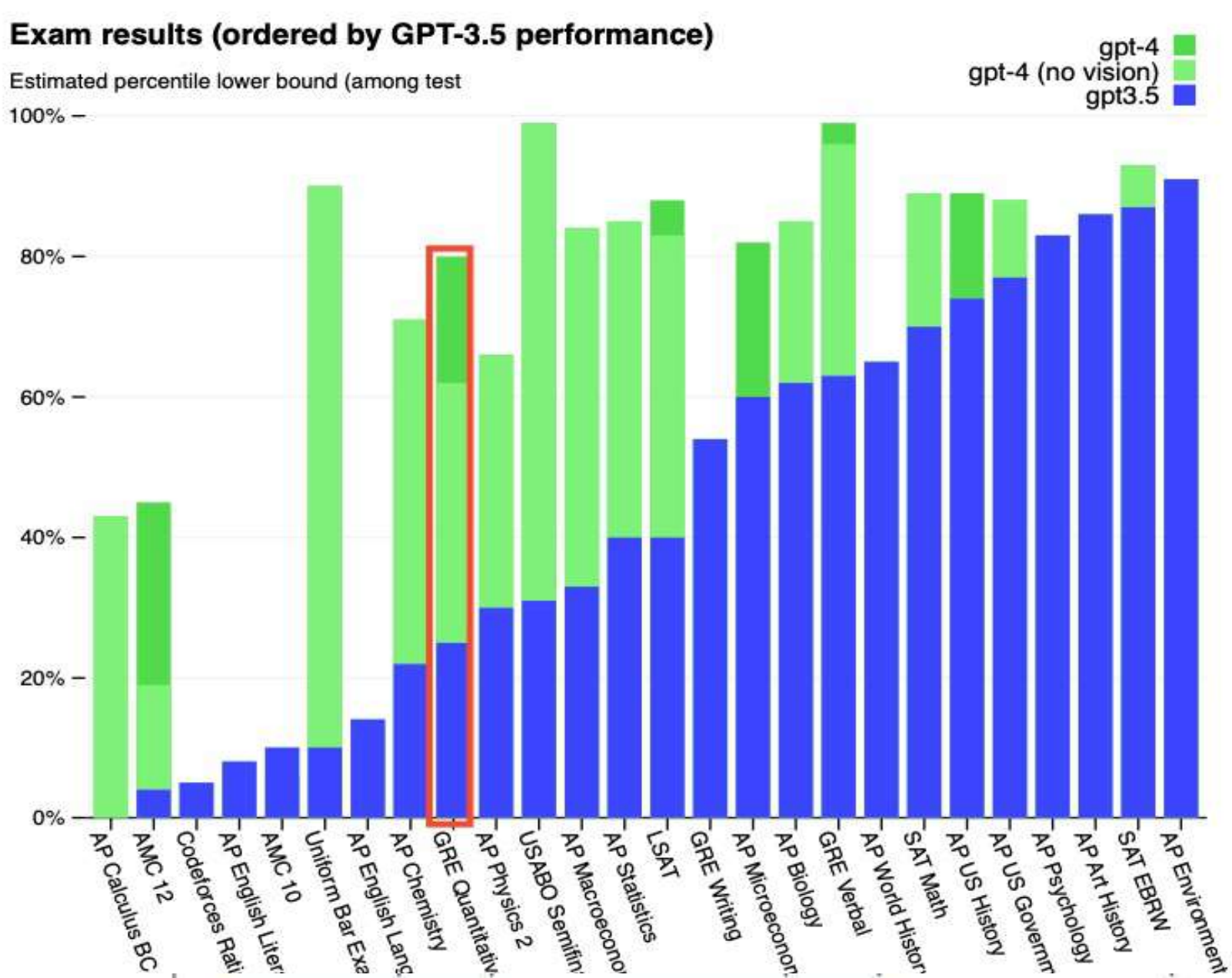
- Word2vec: ASCII to 768-dimension vector, each 8 bytes (double)
- Width (w): more context

  $O(kw^2)$ growth of memory requirement

- Heads (h): robustness improvement
- Depth (d): accuracy improvement, more iterations and refinements
- Data (n): more documents

  $O(khdn)$ growth of parameters

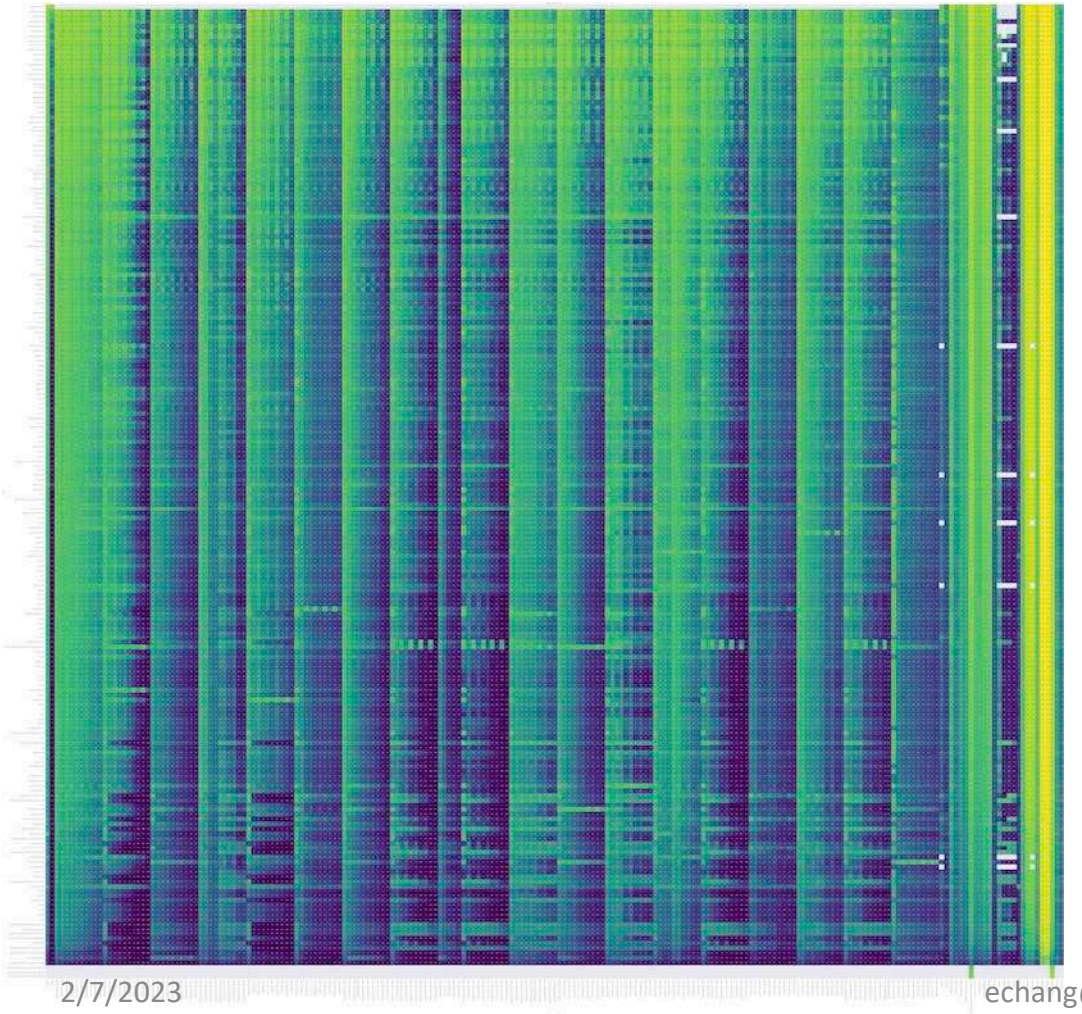
With Vision versus Without Vision Modality



Graduate Record Examination (GRE) Quantitative	163 / 170 (~80th)	157 / 170 (~62nd)	147 / 170 (~25th)

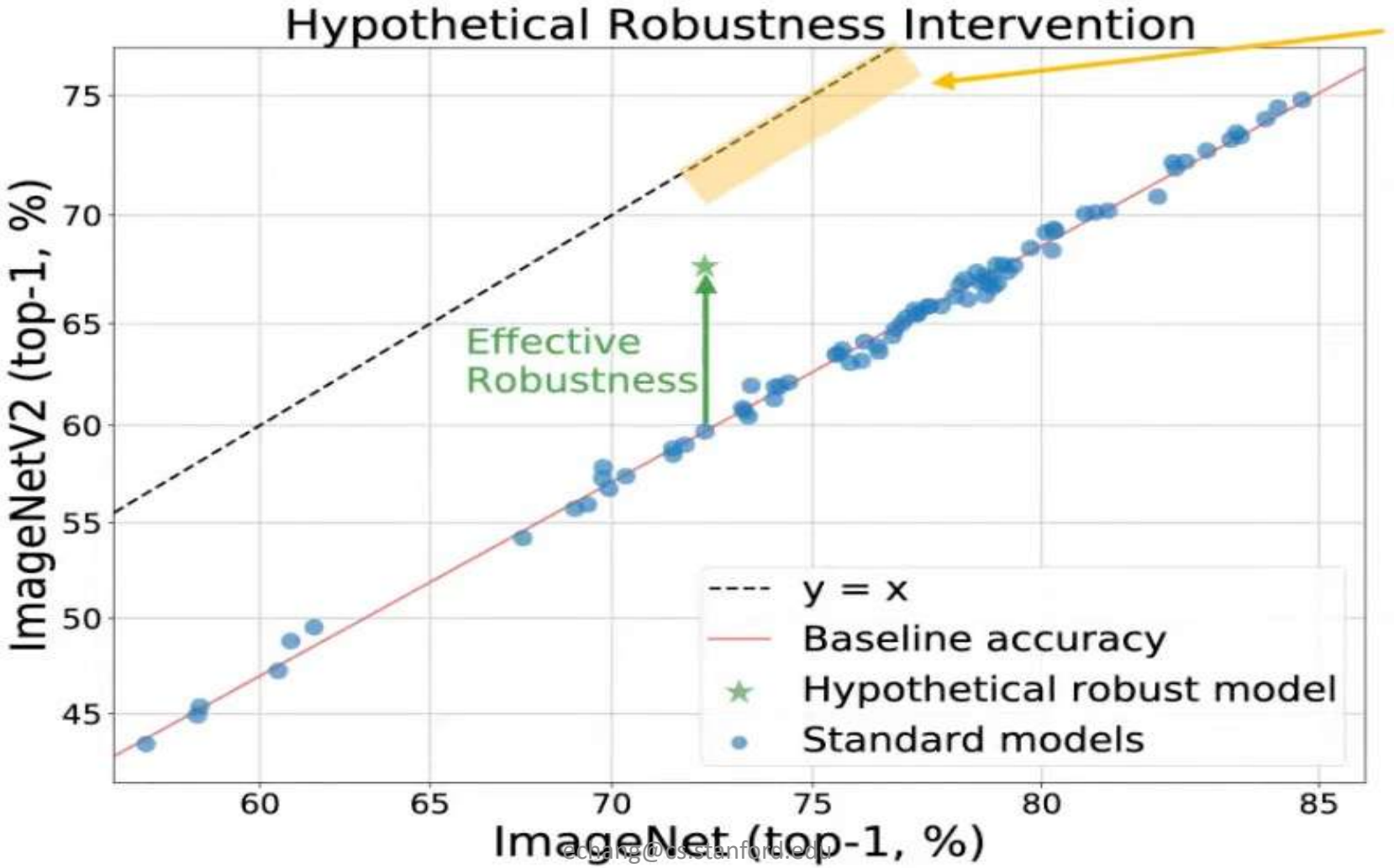
Measuring Robustness to Natural Distribution Shifts in Image Classification

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Ben Recht, Ludwig Schmidt

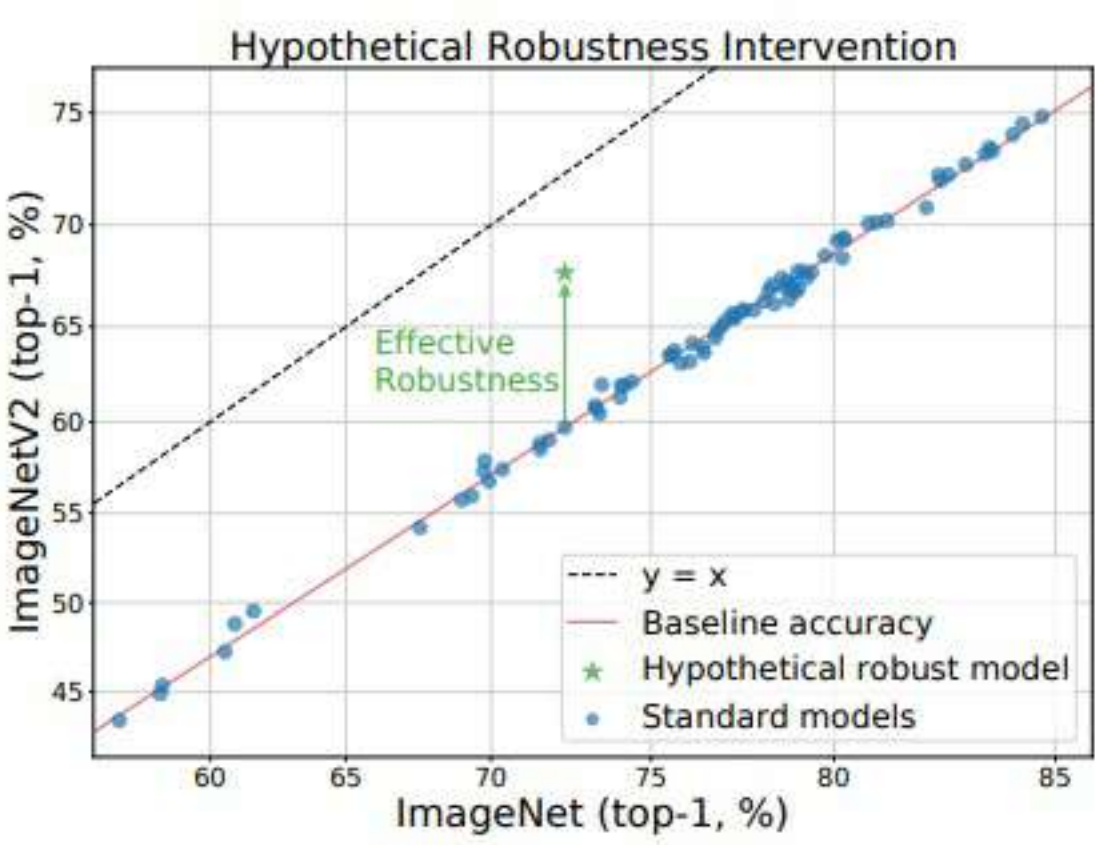


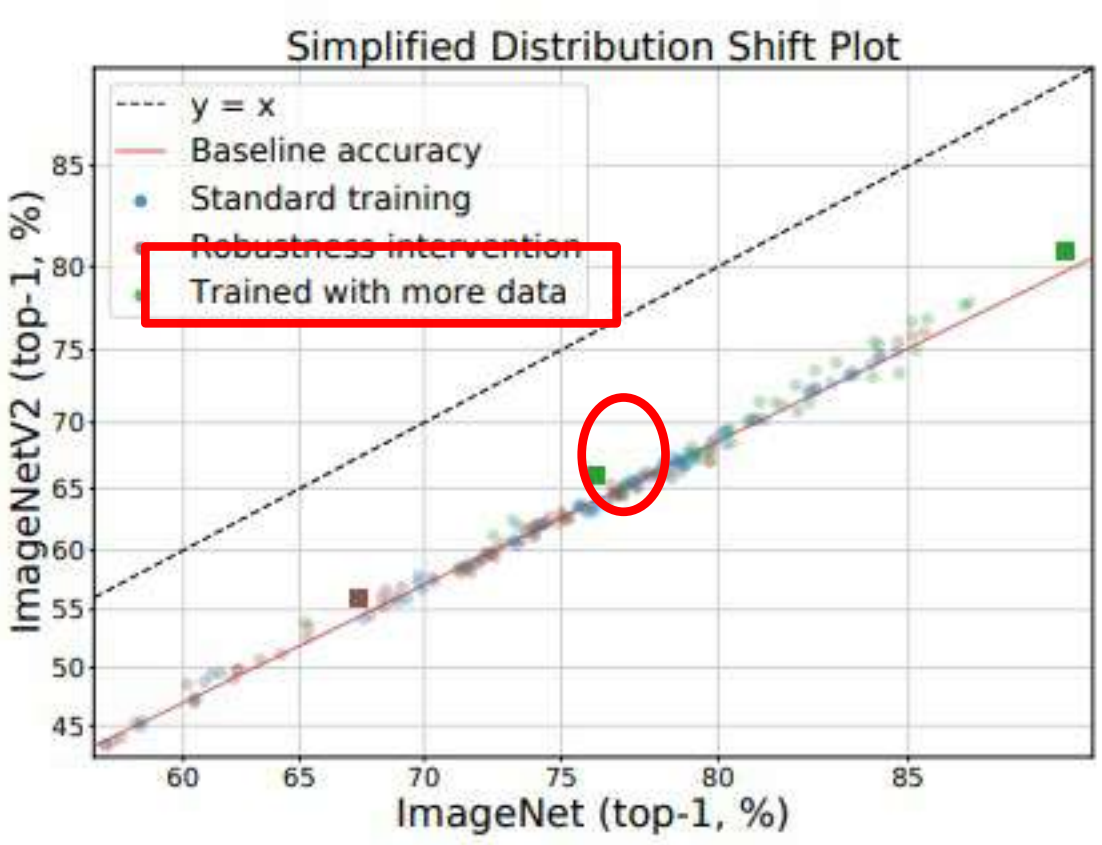
- y-axis, 204 ImageNet Models
  - VGG, AlexNet, etc.
  - Improved robustness
  - More data
- x-axis, 213 test conditions

Robustness Gap

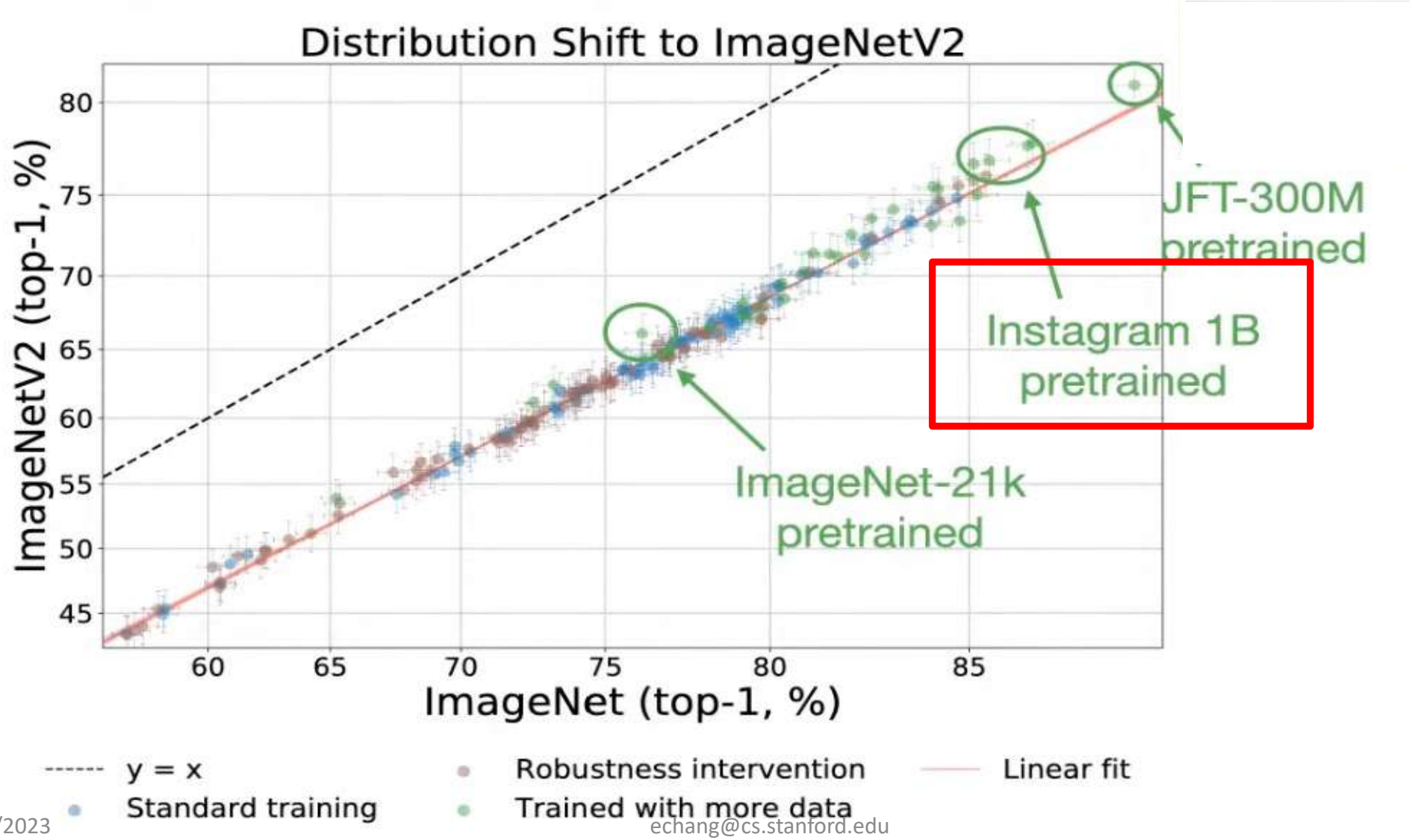


Only Training Data Size & Diversity Matter





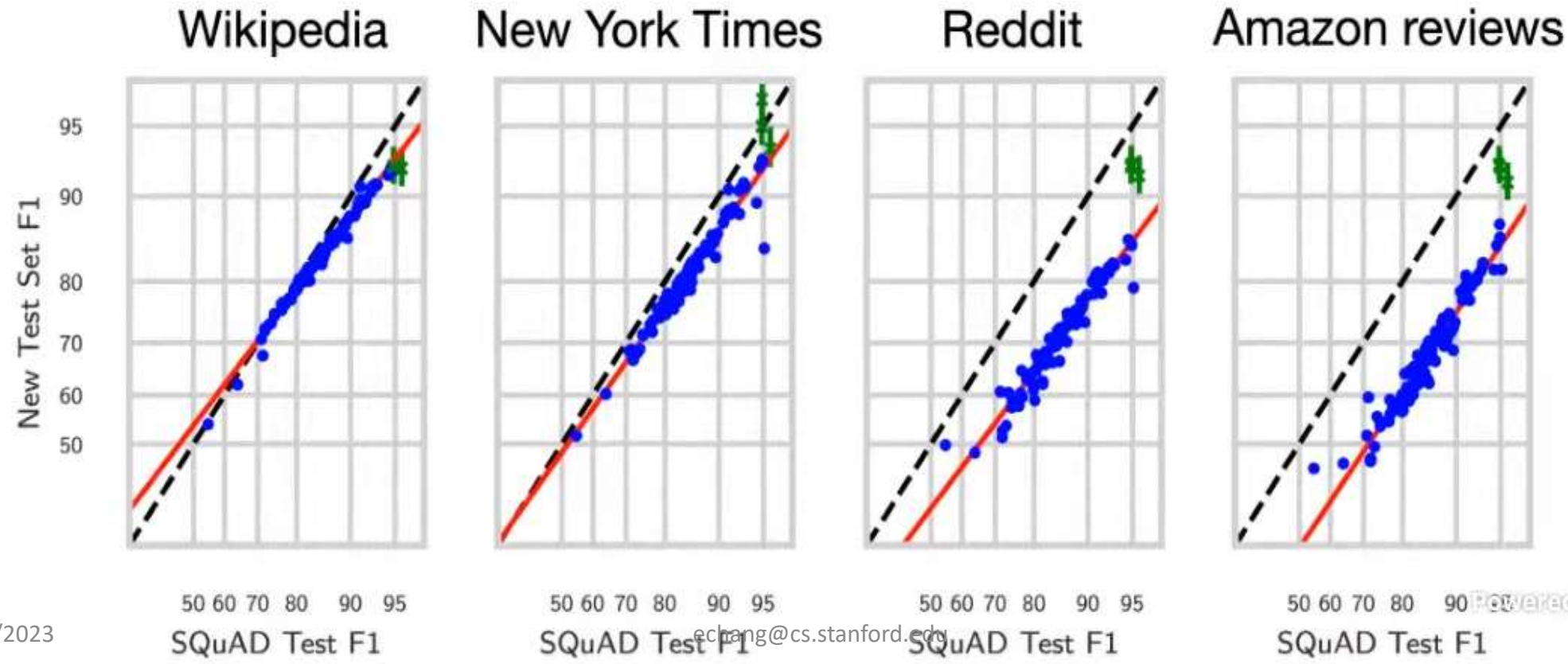
Training Data Distribution Matters



NLP Results

SQuAD (Stanford Question Answering Dataset): question answering on paragraphs

➡ Similar trends in natural language processing. [Miller, Krauth, Recht, Schmidt '20]



Where does Robustness Come From?

- Contrast Loss
- Multi-Modal
- More Data with diversity/variety
- Fine-tuning

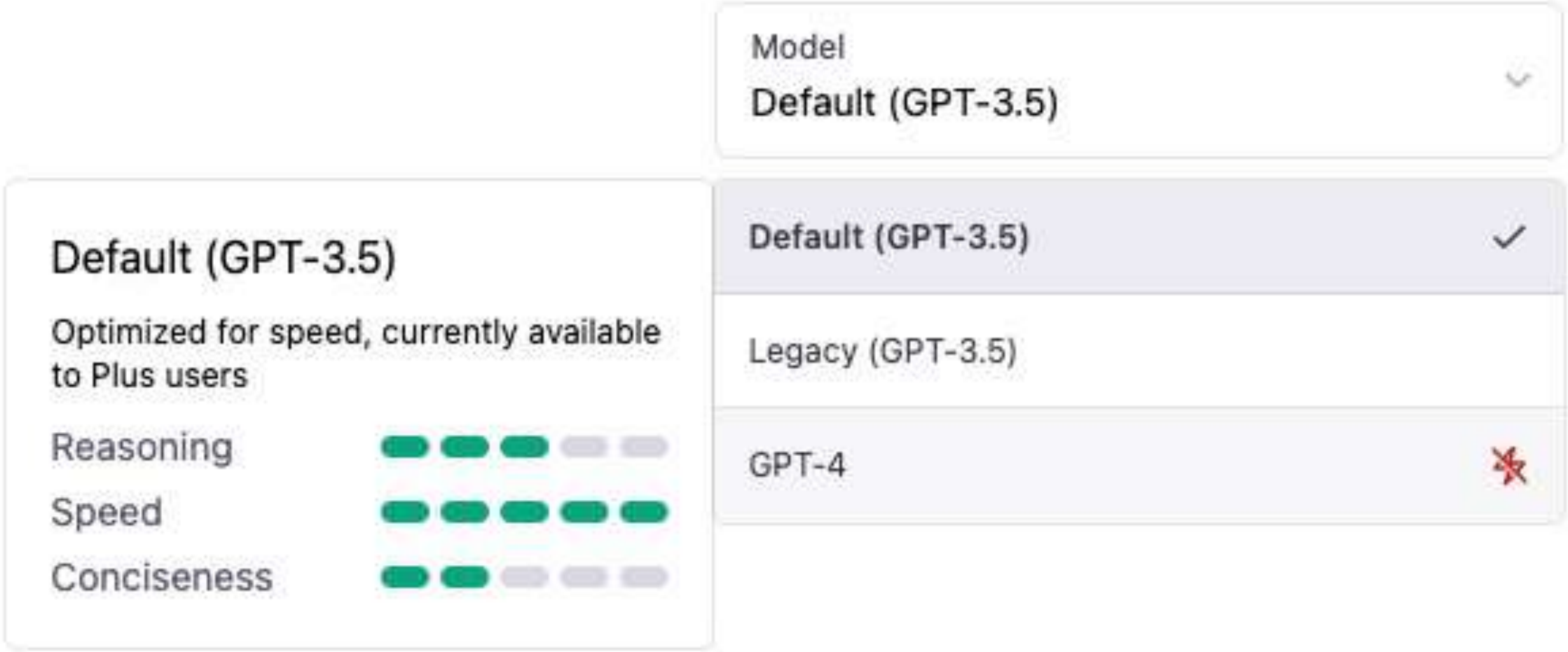
ChatGPT Architecture

Foundation Model(s): e.g., GPT4

Attention Mechanism, polysemy w/ context

Word2vec, similarity

Knowledge, represented by a huge set of documents


Model
Default (GPT-3.5)
Default (GPT-3.5)
Optimized for speed, currently available to Plus users
Reasoning
Speed
Conciseness
Default (GPT-3.5)
Legacy (GPT-3.5)
GPT-4

Differences between GPT3 and ChatGPT?

GPT-3 (Generative Pre-trained Transformer 3) and ChatGPT (Chat Generative Pre-trained Transformer) are both large-scale language models created by OpenAI.

The main difference between the two models is their intended use case. GPT-3 was designed primarily for natural language processing tasks such as language translation, text completion, and question-answering, among others. GPT-3 is capable of generating human-like responses to a wide range of prompts and is trained on a massive dataset of diverse text sources.

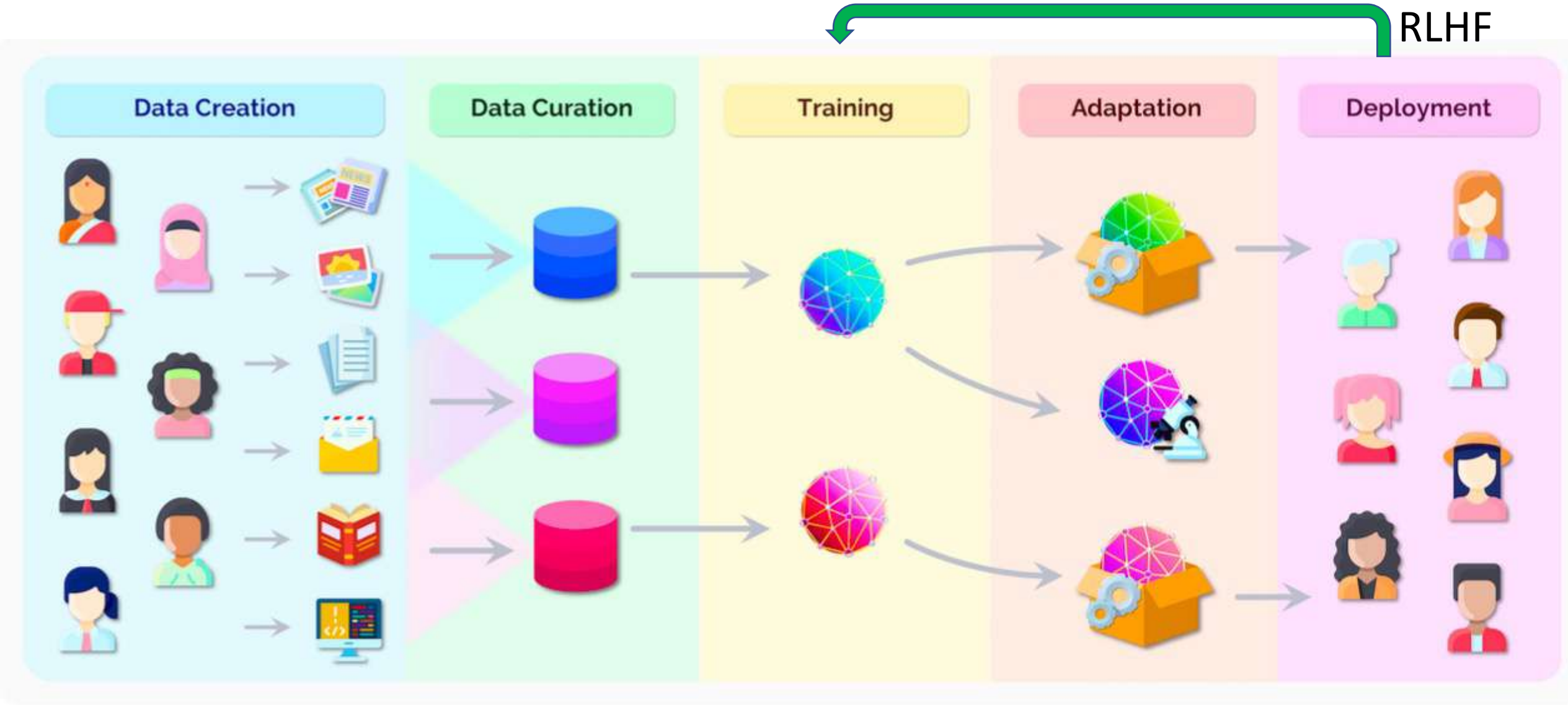
On the other hand, ChatGPT was specifically designed to excel in conversational AI and chatbot applications. It is optimized for generating coherent and contextually-relevant responses to user input, with a focus on maintaining a conversational flow and providing informative and engaging responses. ChatGPT is trained on a dataset that includes a high volume of conversational data, such as social media conversations, online chat logs, and other text sources.

In summary, GPT-3 is a more general-purpose language model that can perform a wide range of natural language processing tasks, while ChatGPT is a specialized model that excels specifically in conversational AI and chatbot applications.

GPT-3 vs. ChatGPT

- GPT-3: 175 billion parameters
- GOT-3.5: GPT-3 with RLHF
- ChatGPT: 1.3 billion parameters
  - Trained on a different dataset that includes a high volume of conversational data, such as social media conversations, online chat logs, and other text sources.
  - May call GPT-3.5 for solving NLP problems, e.g., Q&A, translation, classification, coreference resolution.

Foundation Models w/ RLHF



ChatGPT Architecture

UI: ChatGPT

Foundation Model(s): e.g., GPT4

Attention Mechanism, polysemy w/ context

Word2vec, similarity

Knowledge, represented by a huge set of documents

Prompting vs. Fine-Tuning

- Fine-tuning is the process of taking a pre-trained language model and adapting it to a specific task by further training it on task-specific data.
- Prompts are used to "prime" the model with context and generate text that is relevant to the input.

We know how transformer works, but we do not know why transformer can be so smart, at times.

Is a conscious life that prioritizes self-optimization through maximizing grades, careers, and fortune, while occasionally minimizing energy consumption, truly superior to a life that values aesthetic pleasure, embraces the full spectrum of human experience, and cultivates wonder in the vastness of space and time?

Insider Insights on GPT-4, an hour ago

GPT4, Latest Updates, 4/05/2023 today

Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research

Abstract

Artificial intelligence (AI) researchers have been developing and refining large language models (LLMs) that exhibit remarkable capabilities across a variety of domains and tasks, challenging our understanding of learning and cognition. The latest model developed by OpenAI, GPT-4 [Ope23], was trained using an unprecedented scale of compute and data. In this paper, we report on our investigation of an early version of GPT-4, when it was still in active development by OpenAI. We contend that (this early version of) GPT-4 is part of a new cohort of LLMs (along with ChatGPT and Google's PaLM for example) that exhibit more general intelligence than previous AI models. We discuss the rising capabilities and implications of these models. We demonstrate that, beyond its mastery of language, GPT-4 can solve novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology and more, without needing any special prompting. Moreover, in all of these tasks, GPT-4's performance is strikingly close to human-level performance, and often vastly surpasses prior models such as ChatGPT. Given the breadth and depth of GPT-4's capabilities, we believe that it could reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence (AGI) system. In our exploration of GPT-4, we put special emphasis on discovering its limitations, and we discuss the challenges ahead for advancing towards deeper and more comprehensive versions of AGI, including the possible need for pursuing a new paradigm that moves beyond next-word prediction. We conclude with reflections on societal influences of the recent technological leap and future research directions.

Term Project Elements

- Resources: ChatGPT accounts, GPT-4 credit
- Roster photos

- Select your application domain
- Consider high-priority unmet needs
- Solutions
  - Engineering prompt templates
  - Customizing intent, context, temperature, etc.
  - Fine-tuning GPT4 with domain-specific knowledge (optional)

Assignments

- https://docs.google.com/document/d/10Wds8aiD0Zw7Jqr0gGnOhFiSCJ5i_LINWWVoF5fS4BU/edit?usp=sharing

Concerns Rising? Halting Further Development and Closing the Source (E. Musk, etc.)

- Job preparation strategies in the AI era.
- GPT-4 consciousness: Is there a "person" within it?
- Understanding the inner workings of GPT-4.
- RLHF: Benefits and potential side effects.
- Decision-making for content inclusion and toxicity.
- General Artificial Intelligence: Timeline and possibilities.
  - Yudkowsky's change of stance on AI development.
  - Hinton's surprise at the rapid progress of AI.
- Exploring potential human extinction scenarios.