

Self-supervised representations in NLP:

Large Language Models and more

PhD ML summer school

DTU, August 24 2023

Anna Rogers

slides: <https://annargrs/github.io/talks>





Work on:

- representations learning & analysis in NLP
- AI and society
- meta-science, peer review
- co-program chair of ACL'23, led ACL policy development on ChatGPT
- co-organizer of Workshop on Insights from Negative Results in NLP

In this tutorial

- kinds of self-supervised representations in NLP
- large language models: myths and reality
- data governance for LLMs, and why it matters

Let's find some common ground!



SELF-SUPERVISED REPRESENTATIONS IN NLP

a brief introduction

NLP representations

- Representations in deep learning are **sets of features that reflect the properties of the target phenomenon** (at least partially).
- In this sense, input, output and intermediate weights can all be considered as representations.
- One could argue that the main power of deep-learning is the ability to learn good representations.

Goldberg (2017) [Neural network methods for natural language processing](#)

"Self-supervised" \leftrightarrow "statistical"

- self-supervised training: creating representations from signal in unstructured, unlabeled data
- in case of NLP - unstructured text

Why is self-supervision possible?

Distributional hypothesis:

- Words that occur in similar contexts tend to have similar meanings (Harris 1954).
- You shall know a word by the company it keeps (Firth 1957)
- For a large class of cases... the meaning of a word is its use in the language (Wittgenstein, 1953).

Distributional hypothesis: example

We found a cute little wampimuk sleeping in a tree.

Lazaridou et al. (2014) [Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world.](#)

The simplest representations: one-hot vectors

One-hot encoding

	cat	mat	on	sat	the
the =>	0	0	0	0	1
cat =>	1	0	0	0	0
sat =>	0	0	0	1	0
...	...				

[Tensorflow Word embeddings tutorial](https://www.tensorflow.org/text/guide/word_embeddings)

Classical count-based representations

The diagram illustrates the process of generating word embeddings from text. It starts with a text snippet from 'SOUTHERN LIFE IN SOUTHERN LITERATURE' on page 322. An arrow points to a 4x4 matrix representing word co-occurrences. A double-headed vertical arrow then points down to a 4x4 matrix representing word embeddings.

322 SOUTHERN LIFE IN SOUTHERN LITERATURE

restrained enthusiasm catch from one bystander to another. They swing and bow to right and left, in slow time to the piercing treble of the Congo women. Some are responsive; others are competitive. Hear that bare foot slap the ground! one sudden stroke only, as it were the foot of a stag. The musicians warm up at the sound. A snorting of breaths with open hands begins very softly and becomes vigorous. The women's voices rise to a tremulous intensity. Among the chorus of Franco-Congo singing girls is one of extra good voice, who thrills in, now and again, an improvisation. This girl here, so tall and straight, is a *Yakof*. You see it in her almost Hindu features, and hear it in the plaintive melody of her voice. Now the chorus is more piercing than ever. The women clap their hands in time, or standing with arms akimbo receive with faint courtesies and head-lifts the low bows of the men, who deliver them swinging this way and that.

See! Yonder brisk and sinewy fellow has taken one short, nervy step into the ring, chattering with rising energy. Now he takes another, and stands and sings and looks here and there, rising upon his broad toe and sinking and rising again, with what wonderful lightness! How tall and lithe he is. Notice his brown shining through his rags. He too is a *canedo*, and by the three long rays of tattooing on each side of his face, a Kiamba. The music has got into his feet. He moves off to the farther edge of the circle, still singing, takes the prompt hand of an unwilling Congo girl, leads her into the ring, and, leaving the chant to the throng, stands her before him for the dance.

Will they dance to that measure? Wait! A sudden frenzy seizes the musicians. The measure quickens, the swaying, attitudinizing crowd starts into extra activity, the female voices grow sharp and staccato, and suddenly the dance is the furious Bamboula.

	cat	dog	mouse	cheese
cat		12	15	1
dog	12		0	0
mouse	15	0		15
cheese	1	0	5	

cat	[0.571	0.912	0.126	0.412]
dog	[0.259	0.4512	0.521	0.623]
mouse	[0.115	0.523	0.674	0.571]
cheese	[0.921	0.412	0.836	0.591]

Ex: SVD (Schütze, 1993), HAL (Lund and Burgess, 1996), PCA (Lebret and Collobert, 2014)

First-generation neural representations

322 SOUTHERN LIFE IN SOUTHERN LITERATURE

restrained enthusiasm catch from one bystander to another. They swing and bow to right and left, in slow time to the piercing treble of the Congo women. Some are responsive! others are competitive. Hear that bare foot slap the ground! one sudden stroke only, as it were the foot of a stag. The musicians warm up at the sound. A smiting of breasts with open hands begins very softly and becomes vigorous. The women's voices rise to a tremulous intensity. Among the chorus of Franc-Congo singing-girls is one of extra good voice, who thrusts in, now and again, an improvisation. This girl here, so tall and straight, is a Yaloff. You see it in her almost Hindu features and hear it in the plaintive melody of her voice. Now the chorus is more piercing than ever. The women clap their hands in time, or standing with arms akimbo receive with faint courtesies and head-liftings the low bows of the men, who deliver them swinging this way and that.

See! Yonder brisk and sinewy fellow has taken one short, nervy step into the ring, chanting with rising energy. Now he takes another, and stands and sings and looks here and there, rising upon his broad toes and sinking and rising again, with what wonderful lightness! How tall and lithe he is. Notice his brown shining through his rags. He too is a *candio*, and by the three long rays of tattooing on each side of his face, a Kiamba. The music has got into his feet. He moves off to the farther edge of the circle, still singing, takes the prompt hand of an unsmiling Congo girl, leads her into the ring, and, leaving the chant to the throng, stands her before him for the dance.

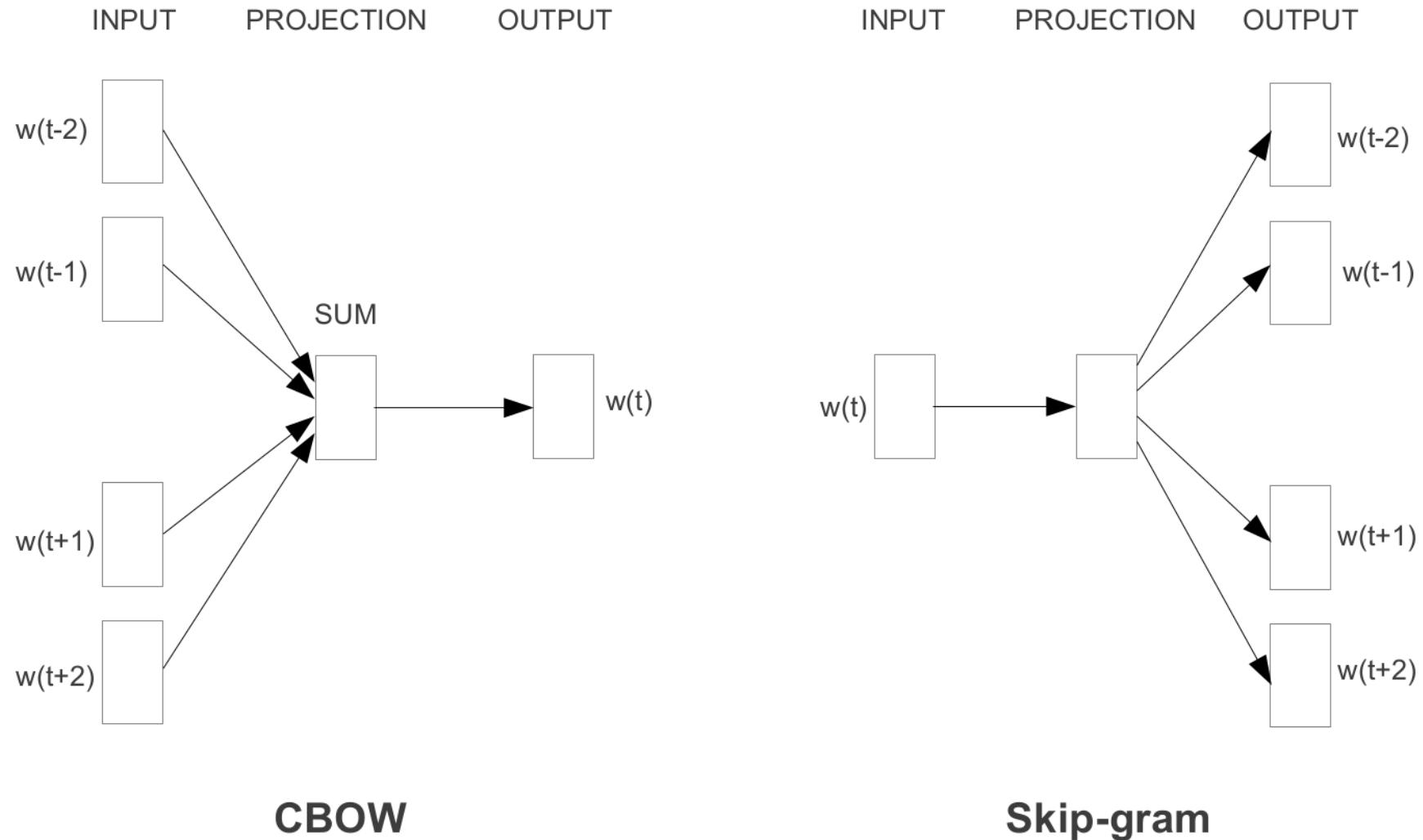
Will they dance to that measure? Wait! A sudden frenzy seizes the musicians. The measure quickens, the swaying, attitudinizing crowd starts into extra activity, the female voices grow sharp and staccato, and suddenly the dance is the furious Bambara.



<i>cat</i>	0.571	0.912	0.126	0.412
<i>dog</i>	0.259	0.4512	0.521	0.623
<i>mouse</i>	0.115	0.523	0.674	0.571
<i>cheese</i>	0.921	0.412	0.836	0.591

Ex: SG, CBOW (Mikolov et al., 2013a), GloVe (Pennington et al., 2014), FastText (Bojanowski et al., 2017), etc.

Word2vec embeddings



CBOW

Skip-gram

Mikolov et al. (2013) [Efficient estimation of word representations in vector space](#)

Language Modeling (LM) as self-supervision

estimate the probability of different linguistic units: symbols, tokens, token sequences

Web search engine / ...

I saw a cat|

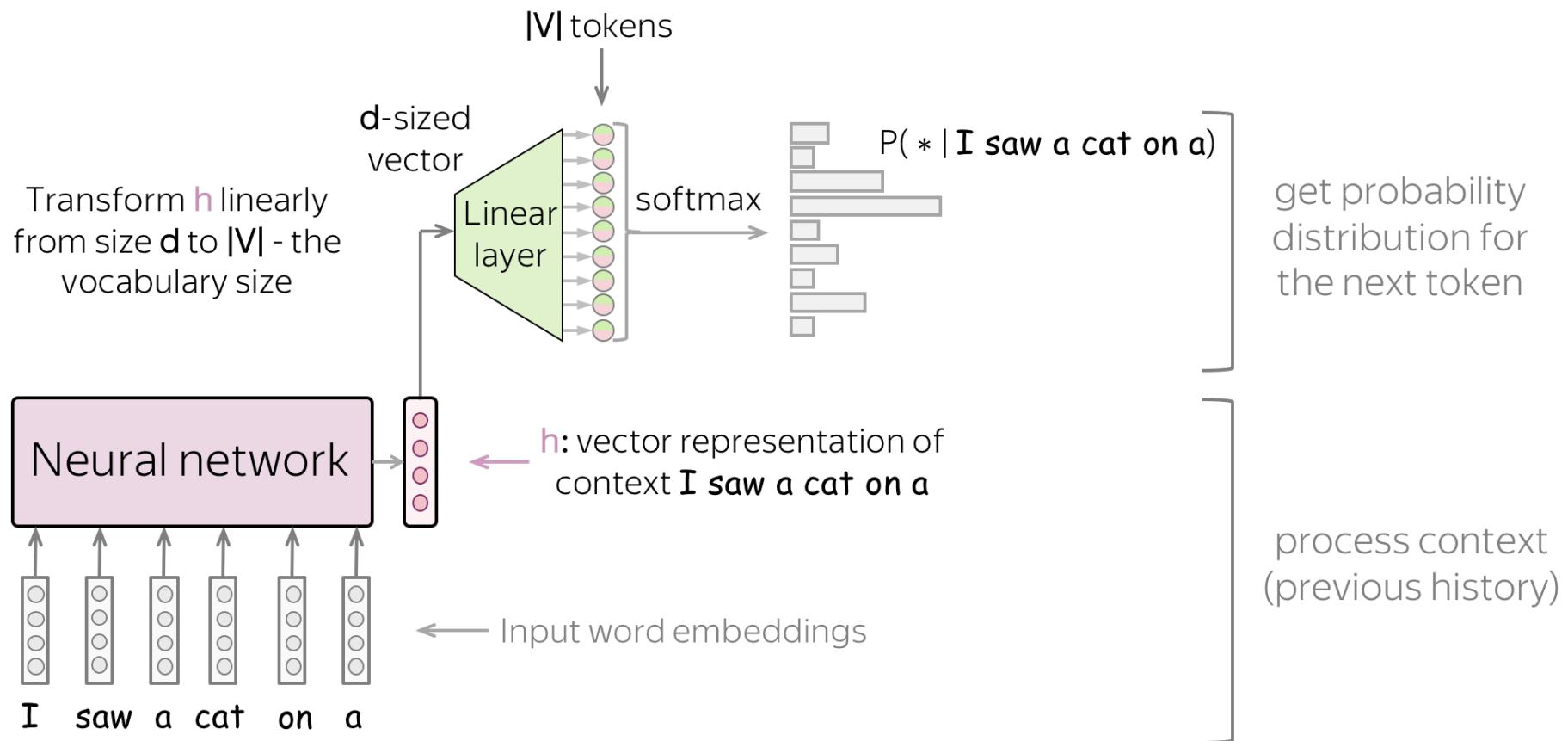
I saw a cat on the chair

I saw a cat running after a dog

I saw a cat in my dream

I saw a cat book

Neural LMs ~ classifiers over vocabulary



Autoregressive language model

$P(\mathbf{I}) =$

$P(\mathbf{I})$
└

Probability of \mathbf{I}

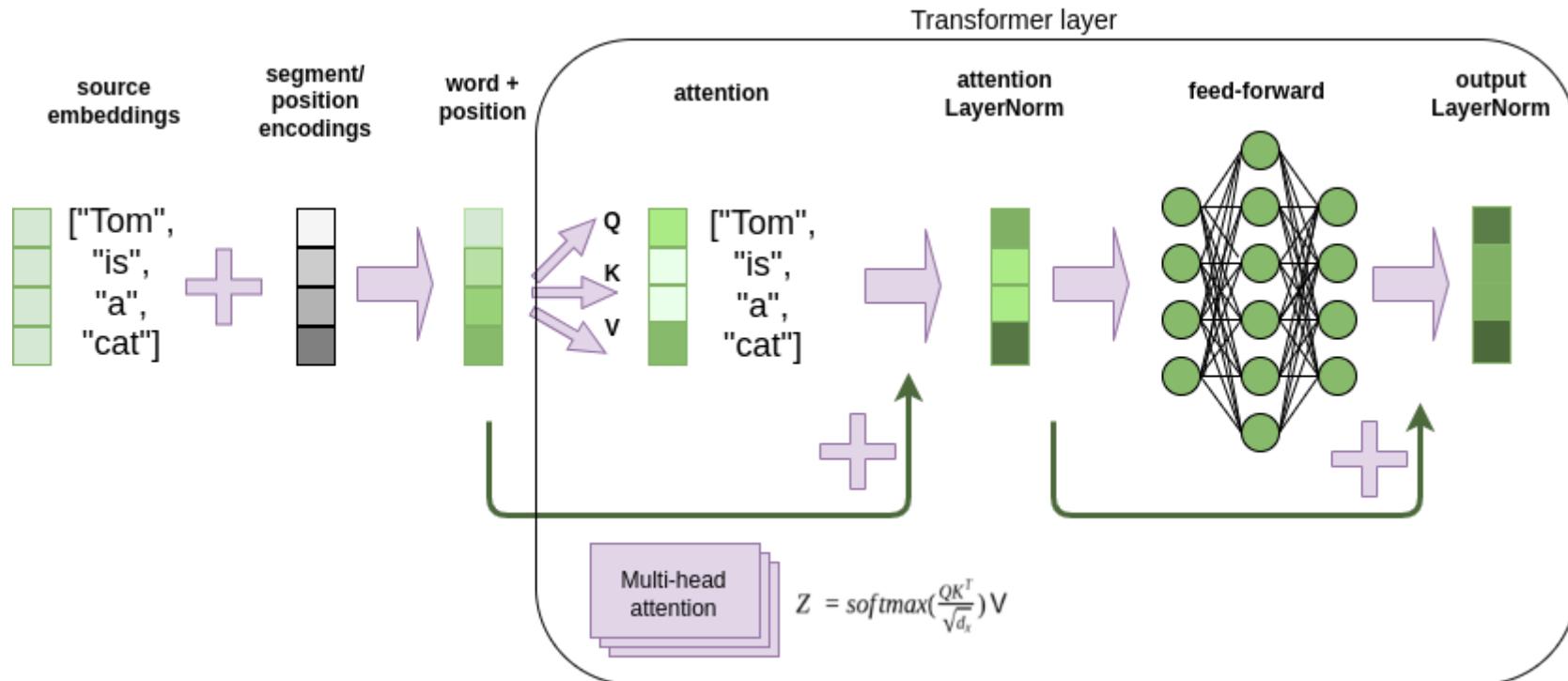
Masked language model

I ___ a cat on a ___.

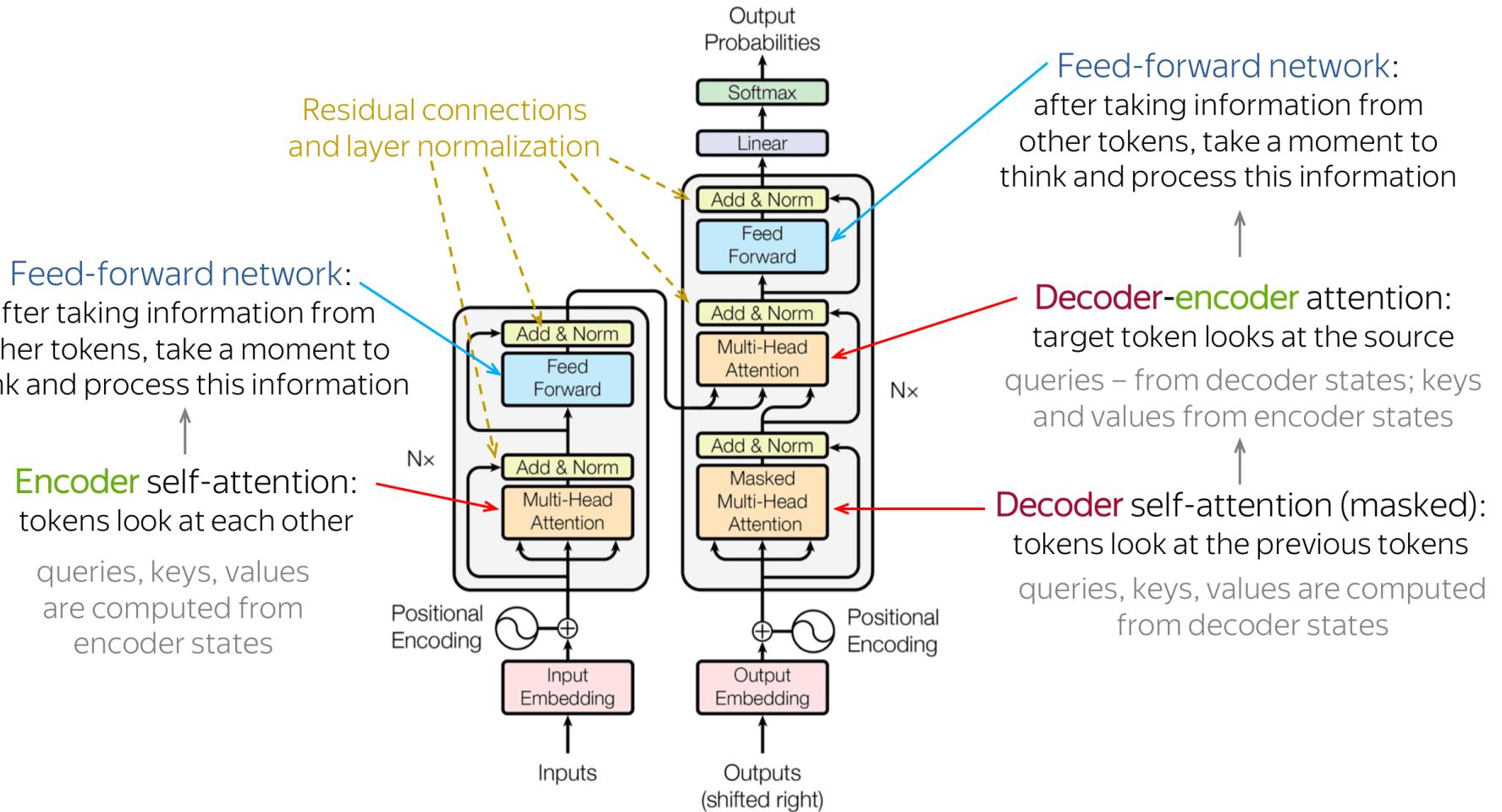
- jointly conditions on both left and right context!

TRANSFORMER-BASED LANGUAGE MODELS

Transformer encoder (e.g. BERT)

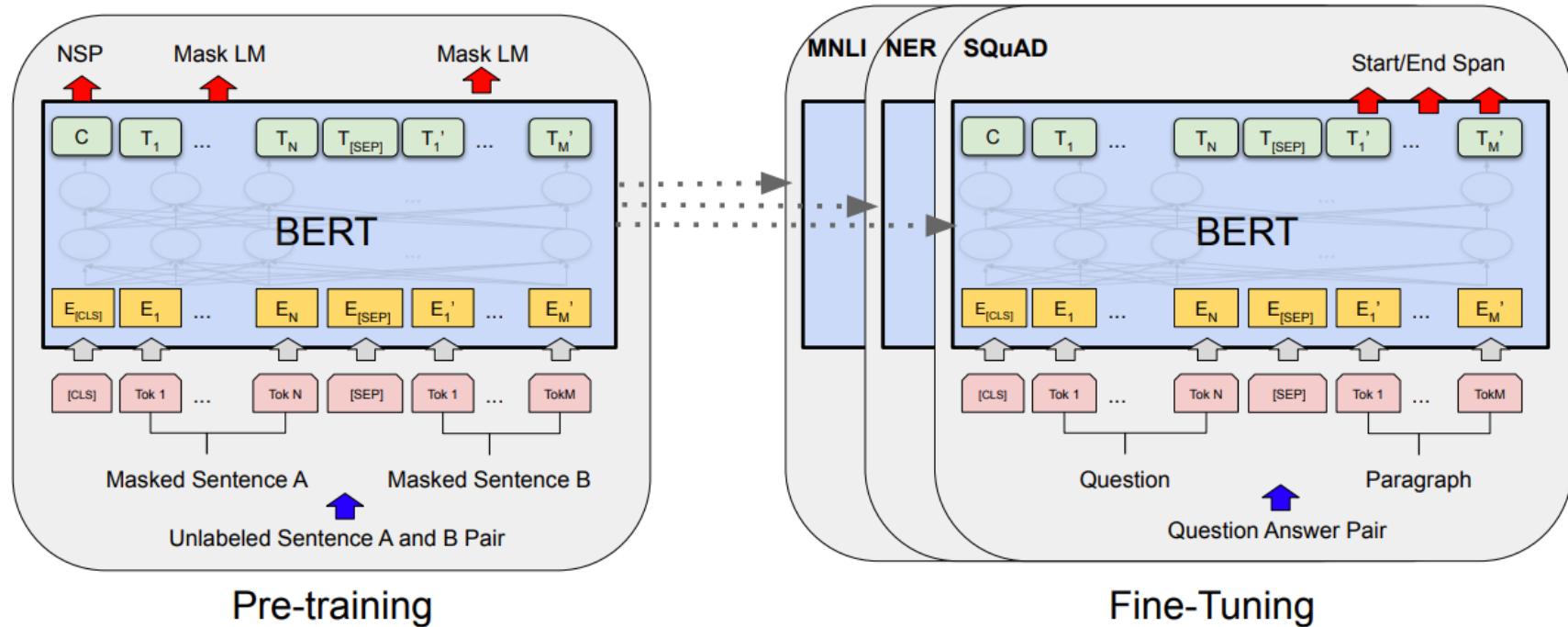


Encoder-decoder Transformer



Vaswani et al. (2017) [Attention is all you need](#), Lena Voita (2023) [NLP course](#)

Pre-training vs fine-tuning



Devlin et al. (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

What happens during fine-tuning?

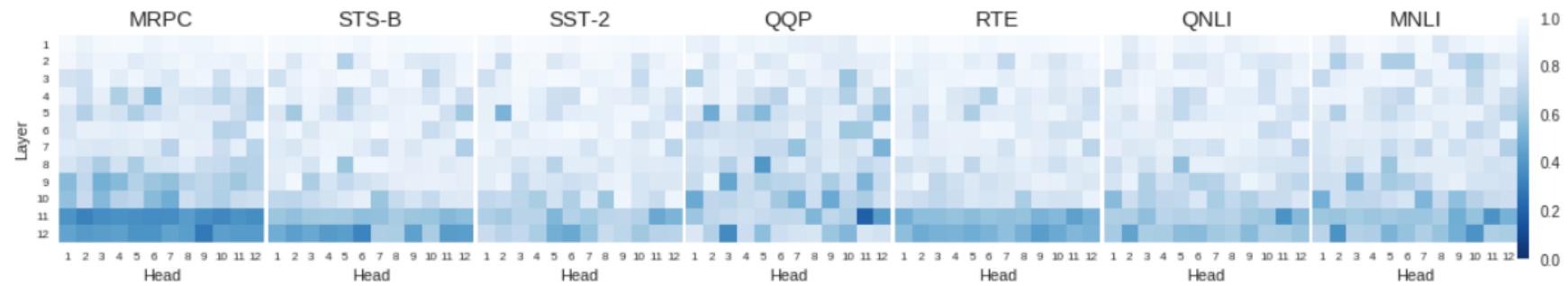


Figure 5: Per-head cosine similarity between pre-trained BERT's and fine-tuned BERT's self-attention maps for each of the selected GLUE tasks, averaged over validation dataset examples. Darker colors correspond to greater differences.

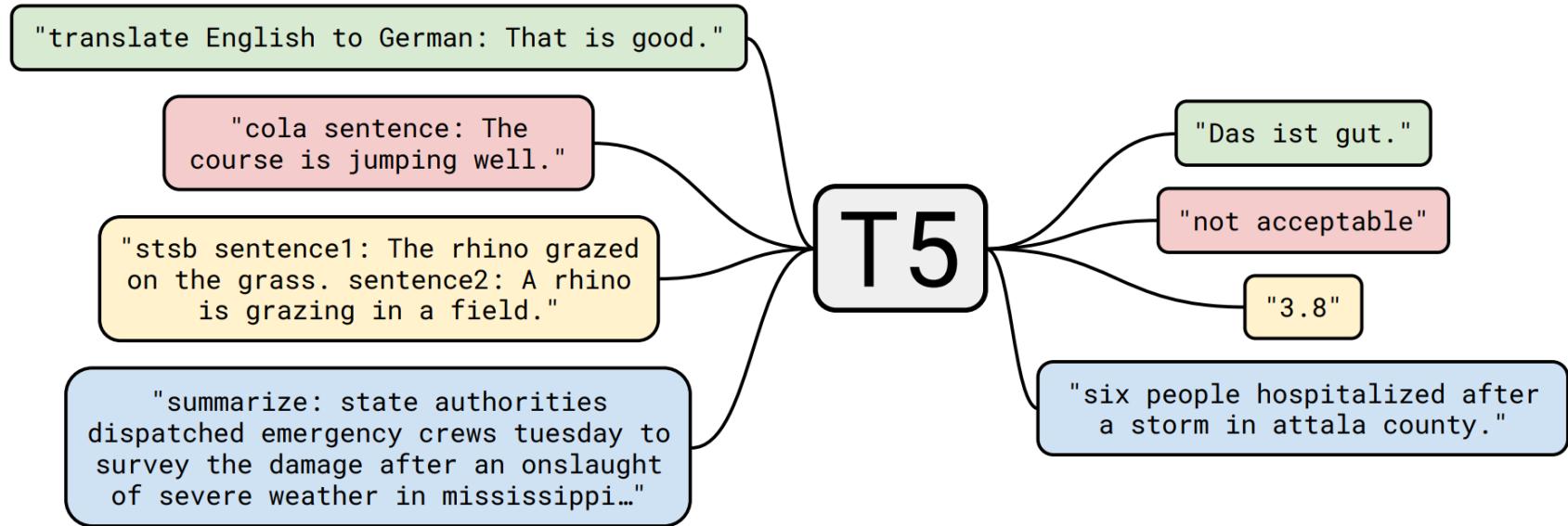
Kovaleva et al. (2019) [Revealing the Dark Secrets of BERT](#)

Can we fine-tune on many tasks at once?

- "pre-fine-tuning": an additional step between regular pre-training and fine-tuning
- a massive multi-task learning step (4.8 million total training examples) performed on around 50 classification, summarization, question answering, and common sense reasoning tasks
- pre-fine-tuned models consistently better than pre-training+fine-tuning!
- using different heads for different objectives

Aghajanyan et al. (2021) [Muppet: Massive Multi-task Representations with Pre-Finetuning](#)

Can we fine-tune and pre-train with the same objective?



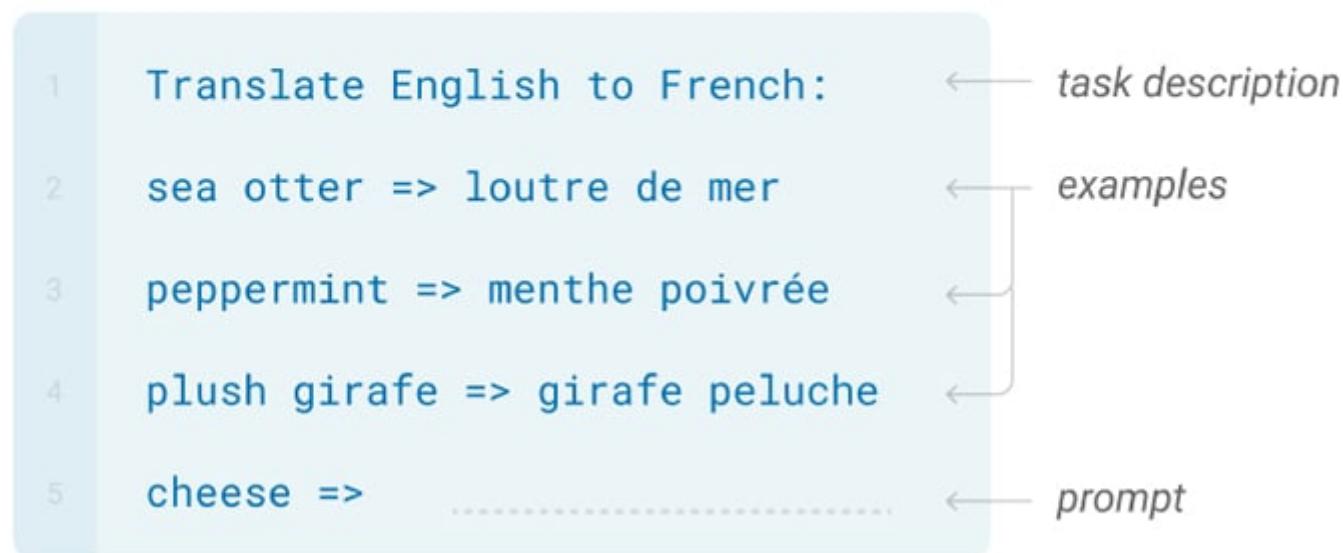
! adding multi-task learning to larger models does not improve upon the standard pre-training / finetuning

Raffel et al. (2020) [Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer](#)

"In-context/few-shot learning"

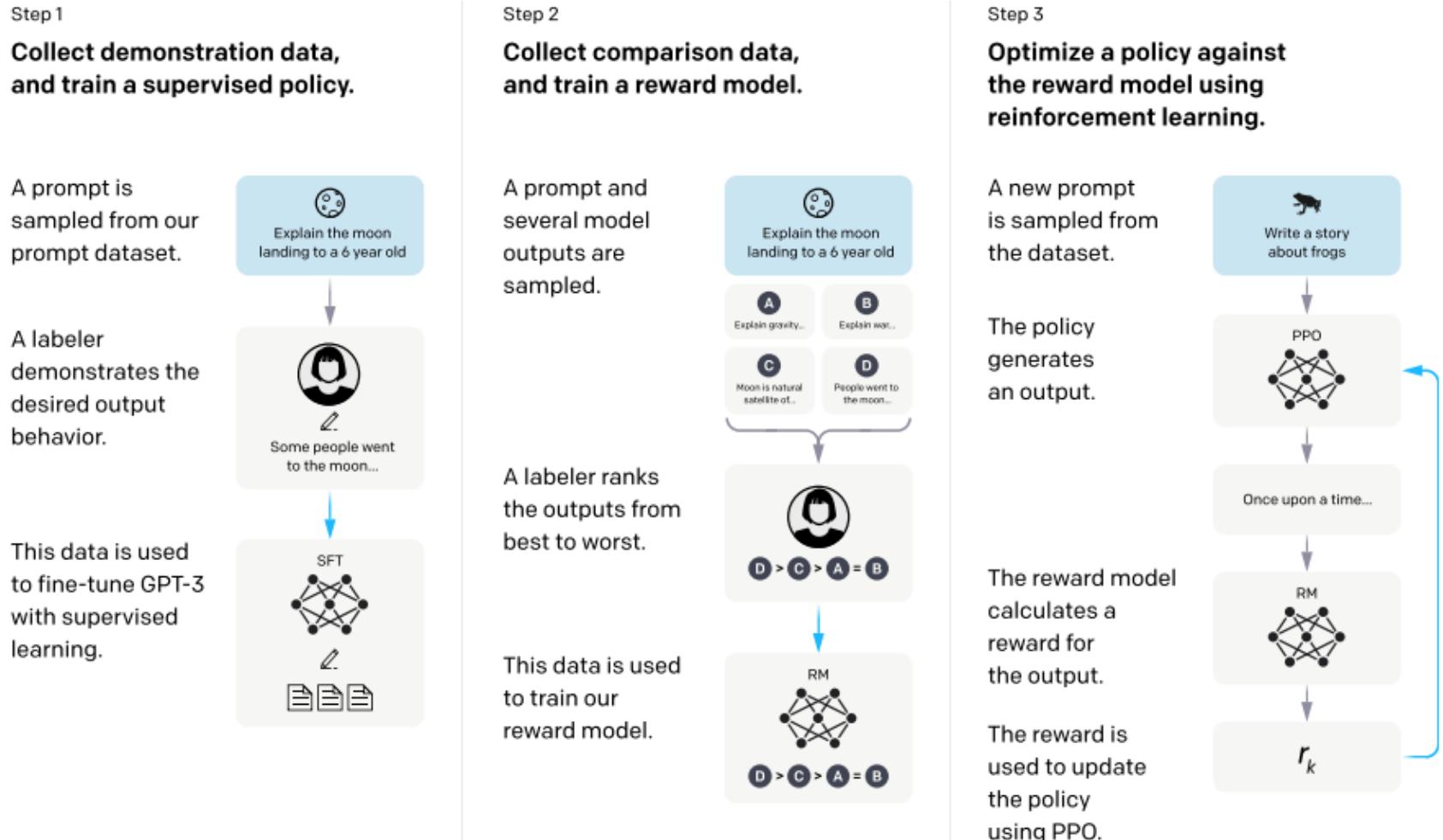
Few-shot

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



Brown et al. (2020) [Language Models are Few-Shot Learners](#), illustration by [Anna Popovych](#)

Instruction tuning and RLHF



Ouyang et al. (2022) Training language models to follow instructions with human feedback

- dialogue version of InstructGPT
- new OpenAI in-house data (humans both writing and rating model responses)
- new ranking data for RLHF
- keeps changing under the hood
- that's all we know!

WHAT ARE LANGUAGE MODELS GOOD FOR?

Input assistance

- predictive keyboards
- speech transcription
- spellchecking
- grammar correction

See also [Emily M. Bender's talk on uses of generative AI](#)

(Well-defined) text classification based on LLM encoders

- Sorting user input for further handling, e.g. intent detection
- hate speech, toxicity detection
- fact-checking assistance
- ... many others!

(Well-tested) information extraction pipelines

- identifying target information in unstructured text
- summarizing (if texts are relatively standard)

... but marketing of LLM-based solutions got a bit out of hand

What is the advertised role of this tool?

A better, 10x faster way to write blogs

Rytr is an AI writing assistant that helps you create high-quality content, in just a few seconds, at a fraction of the cost!

[Start Ryting →](#)

Free forever, upgrade as you scale!

What is the advertised role of this tool?



AISEO

Home

Chrome Extension

Paraphrasing tool

Resources ↓

Products ↓

Become an affiliate (48%) Pricing

Login



COPYWRITING DONE 10X FASTER

Let Our Assistant Help You With writing essays. |

Get writing superpowers with artificial intelligence. We at AISEO will help you convert your thoughts to blogs and can write for you at the click of a button.

Missing the point?

MEET JASPER 

Create amazing love letters
10X faster with AI.

Jasper is the generative AI platform for business that helps your team create content tailored for your brand 10X faster, wherever you work online.

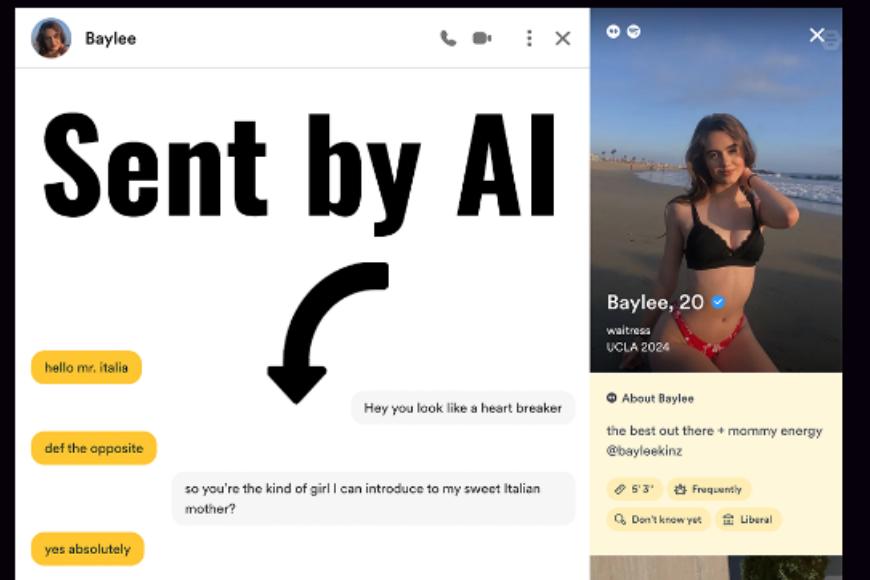
[Get A Demo](#)

[Start Now →](#)

Getting creepy

Get dates without opening any apps

CupidBot swipes and chats for you on dating apps to bring you several dates a week so you can skip to the good part.



Too good to be true?

Microsoft Edge

Features & Tips

Performance

Gaming

Productivity

Security

AI innovations

Accessibility & tools

Edge at work

Design

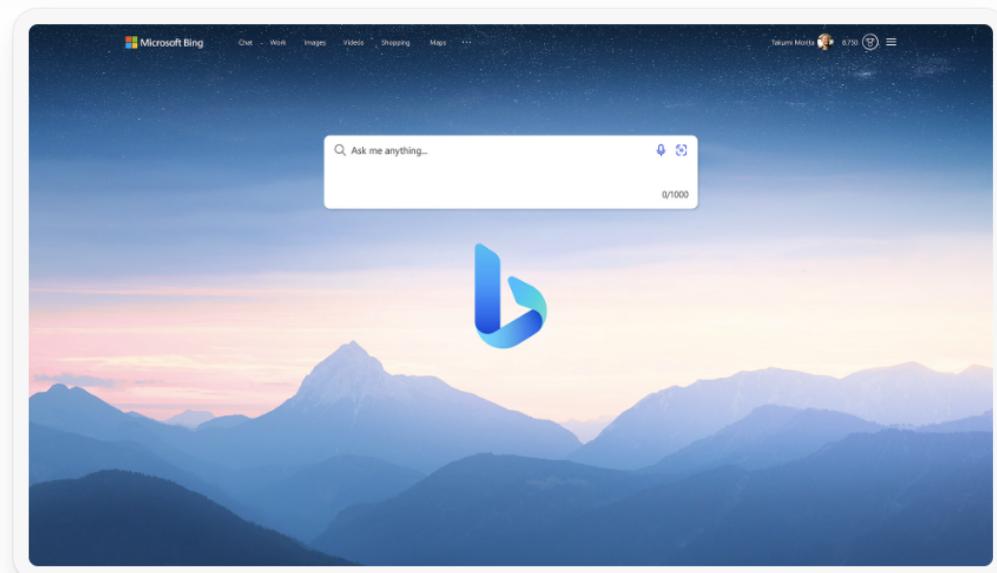
Search 

FEATURE AI-POWERED

Introducing the new Bing. The AI-powered assistant for your search.

Chat, search, and find inspiration all in one place. Learn how to use the new Bing to get summarized answers, create drafts, images and more.

Try now



SOME COMMON CLAIMS, ASSUMPTIONS, AND MISCONCEPTIONS ABOUT LLMS

"*LANGUAGE MODEL*"



What is "language"?

the words, their pronunciation, and the methods of combining them used and understood by a community

Merriam-Webster Dictionary

Competence vs performance

We thus make a fundamental distinction between competence (the speaker-hearer's knowledge of his language) and performance (the actual use of language in concrete situations)... A record of natural speech will show numerous false starts, deviations from rules, changes of plan in mid-course, and so on. The problem for the linguist, as well as for the child learning the language, is to determine from the data of performance the underlying system he puts to use in actual performance.

Competence vs performance

? is this all that a child has to do?

*The problem for the linguist, as well as for the child learning the language, is **to determine from the data of performance the underlying system he puts to use in actual performance.***

Competence vs performance

Competence

- grammar
- vocabulary
- style



can be inferred
from texts

Performance

- social conventions
- psychological factors
- world knowledge



many
extralinguistic factors
with little evidence!

What exactly are we modeling here?

$P(I) =$

$P(I)$



Probability of I

... and do we have enough signal for that?

Performance data is *expected* to be noisy

- this is why Chomsky originally objected to deriving grammars from real examples of language use!
- but in deep learning-based NLP we assume the system will be able to learn from real data...
- ... so why in practice, a lot of effort is spent on cleaning large-scale corpora like the Pile? 🤔

Veres (2022) [Large Language Models are Not Models of Natural Language: They are Corpus Models](#)

Garbage in, garbage out!

the models, unable to discern the intentions of the data set's designers, happily recapitulate any statistical patterns they find in the training data.

See also Gardner et al. (2021) [Competency Problems: On Finding and Removing Artifacts in Language Data](#)

Linzen (2020) [How Can We Accelerate Progress Towards Human-like Linguistic Generalization?](#)

LMs are actually *corpus* models

*we would... propose a change from the theory-laden term **language model** to the more objectively accurate term **corpus model**. Not only does the term **corpus model** better reflect the contents of models, it also provides transparency in discussing issues such as model bias. One might be surprised if a **language model** is biased, or if there is different bias in two different **language models**, but a bias in **corpus models** and different biases in different **corpus models** is almost an expectation. Natural language is not biased. What people say or write can be biased*

"LLMS" / "FOUNDATION MODELS"



What are "foundation models"?

There is not a precise technical definition of foundation model. Rather, this is an informal label for a large family of models, and this family of models is likely to grow and change over time in response to new research. This poses challenges to reasoning about their fundamental properties. However, there is arguably one defining characteristic shared by all foundation models: they are self-supervised.” (Bommasani et al., 2021, p. 48)

Large language models:

- LLMs model text (in any modality) and can be used to generate it
- LLMs receive large-scale pretraining (at least 1B tokens)
- LLMs make inferences based on transfer learning

LLMS ARE SOTA



Transformers are everywhere!

Models like GPT and BERT relying on this Transformer architecture have fully outperformed the previous state-of-the-art networks. It surpassed the earlier approaches by such a wide margin that all the recent cutting edge models seem to rely on these Transformer-based architectures.

Gillioz et al. (2020) [Overview of the Transformer-based Models for NLP Tasks](#)

Fine-tuning vs few-shot performance

- generally speaking, more in-domain training of ML models yields higher in-domain performance
- the earlier models like BERT were evaluated after fine-tuning, which is more comparable to previous approaches relying on word embeddings
- after GPT-3, most "big" models were presented with few-shot evaluations only

e.g. superGLUE leaderboard: fine-tuned RoBERTa - 84.6,
few-shot GPT-3 - 71.8

<https://super.gluebenchmark.com/leaderboard/>

Why did we switch to few-shot only?

- The shift to LLM few-shot evaluations is a community trend, not because we have guarantees that this is the best approach for any given task
- in many cases simple models are enough
- few-shot models win on generality, but they are more expensive and harder to document. Not necessarily the best solution for deployment.

Prompt sensitivity

the order of samples and prompt template make a lot of difference!

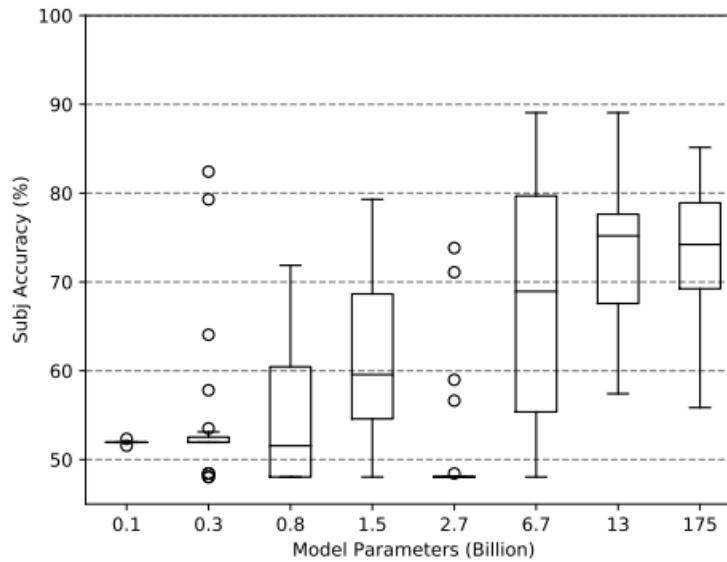


Figure 1: Four-shot performance for 24 different sample orders across different sizes of GPT-family models (GPT-2 and GPT-3) for the SST-2 and Subj datasets.

True few-shot performance

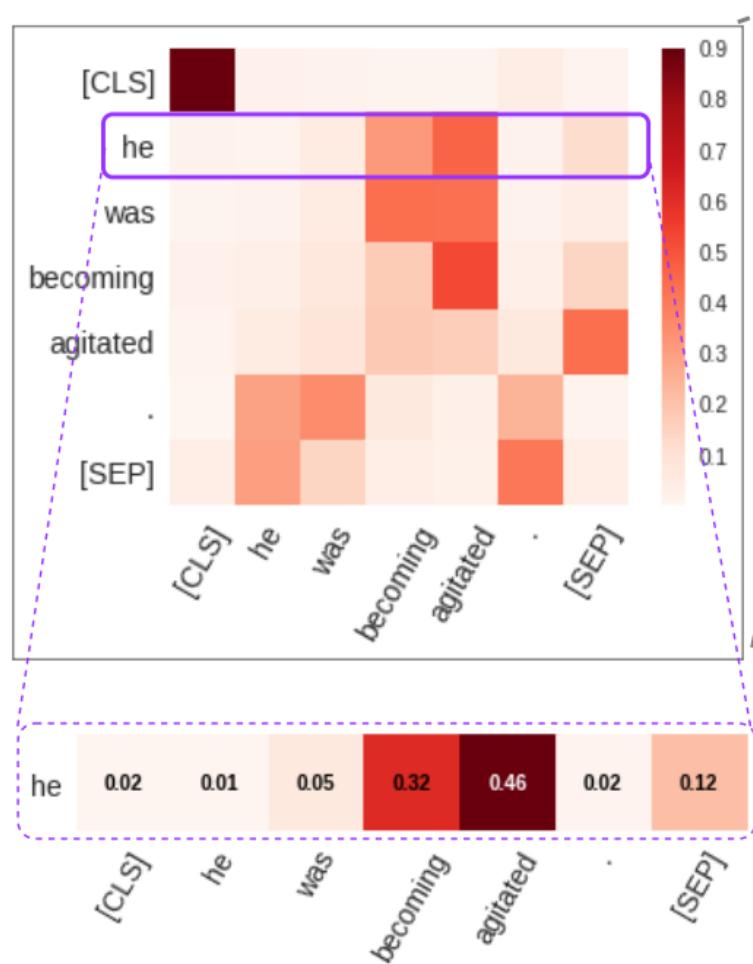
- usually held-out data is used to find an optimal prompt
- in true few-shot setting, the performance is worse

Perez et al. (2021) [True Few-Shot Learning with Language Models](#)

"ATTENTION IS ALL YOU NEED"

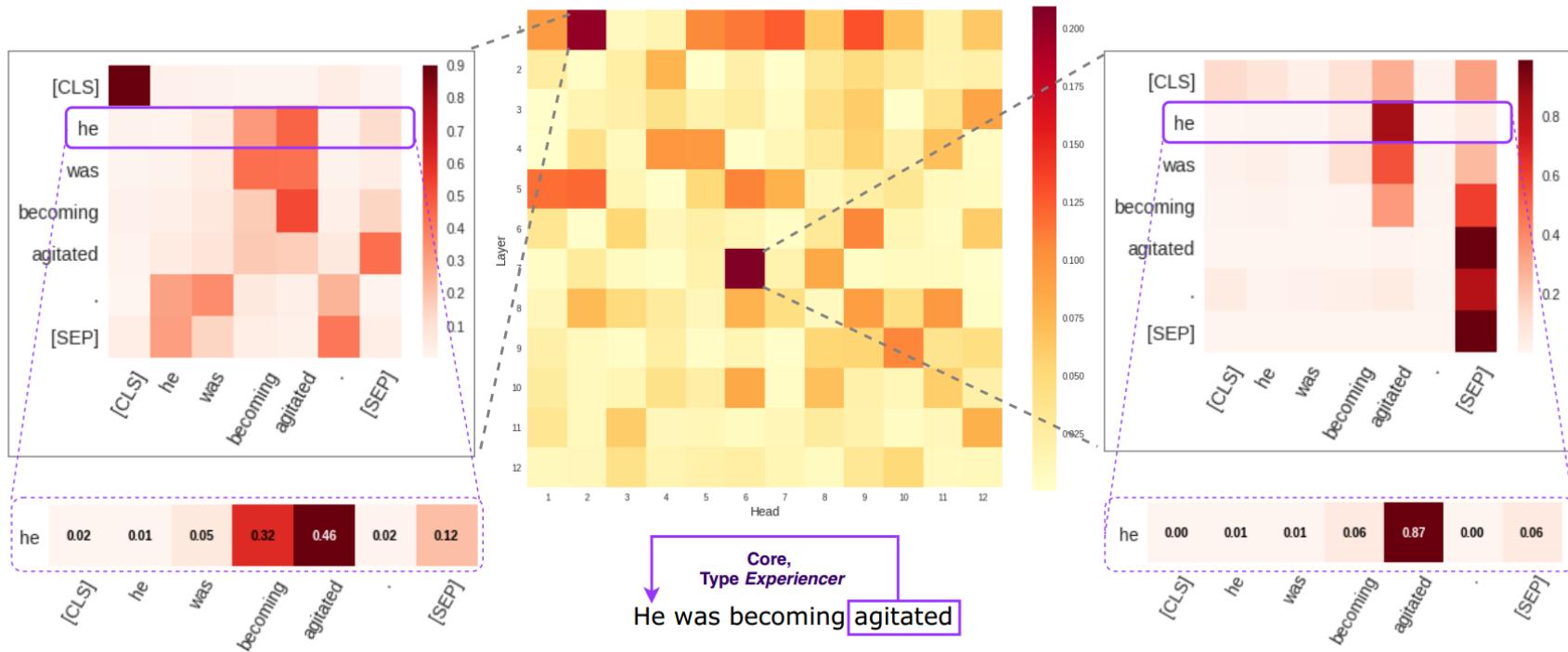


Attention weights ~ relations in sequences



Kovaleva et al. (2019) [Revealing the Dark Secrets of BERT](#)

Attention weights ~ relations in sequences



Kovaleva et al. (2019) [Revealing the Dark Secrets of BERT](#)

Attention is not explanation

after 15 minutes watching the movie i was asking myself what to do leave the theater sleep or try to keep watching the movie to see if there was anything worth i finally watched the movie what a waste of time maybe i am not a 5 years old kid anymore

original α

$$f(x|\alpha, \theta) = 0.01$$

after 15 minutes watching the movie i was asking myself what to do leave the theater sleep or try to keep watching the movie to see if there was anything worth i finally watched the movie what a waste of time maybe i am not a 5 years old kid anymore

adversarial $\tilde{\alpha}$

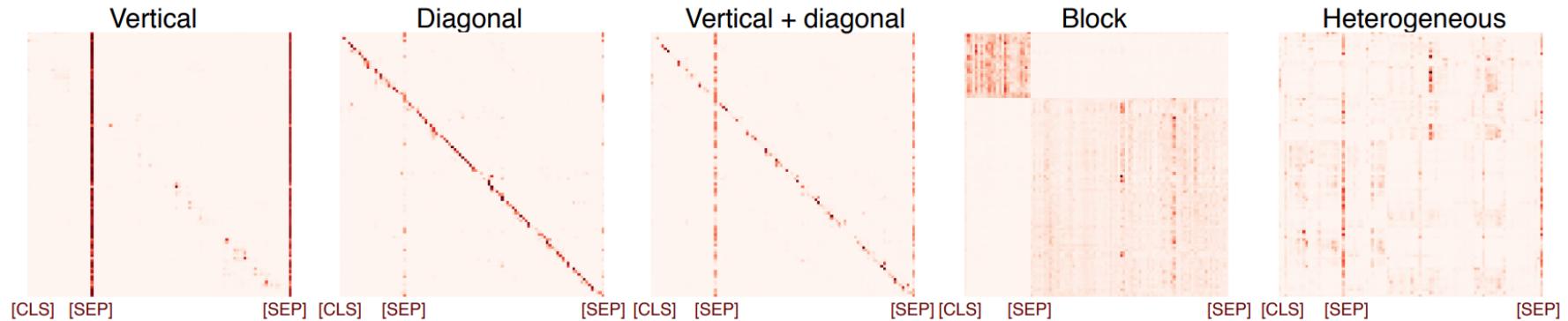
$$f(x|\tilde{\alpha}, \theta) = 0.01$$

Figure 1: Heatmap of attention weights induced over a negative movie review. We show observed model attention (left) and an adversarially constructed set of attention weights (right). Despite being quite dissimilar, these both yield effectively the same prediction (0.01).

Wiegreffe et al: attention provides plausible, but not faithful explanations

Jain et al. (NaN) [Attention is not Explanation](#), Wiegreffe et al. (2019) [Attention is not not Explanation](#)

Most BERT attention heads do other jobs!



Kovaleva et al. (2019) [Revealing the Dark Secrets of BERT](#)

Most BERT attention heads do other jobs!

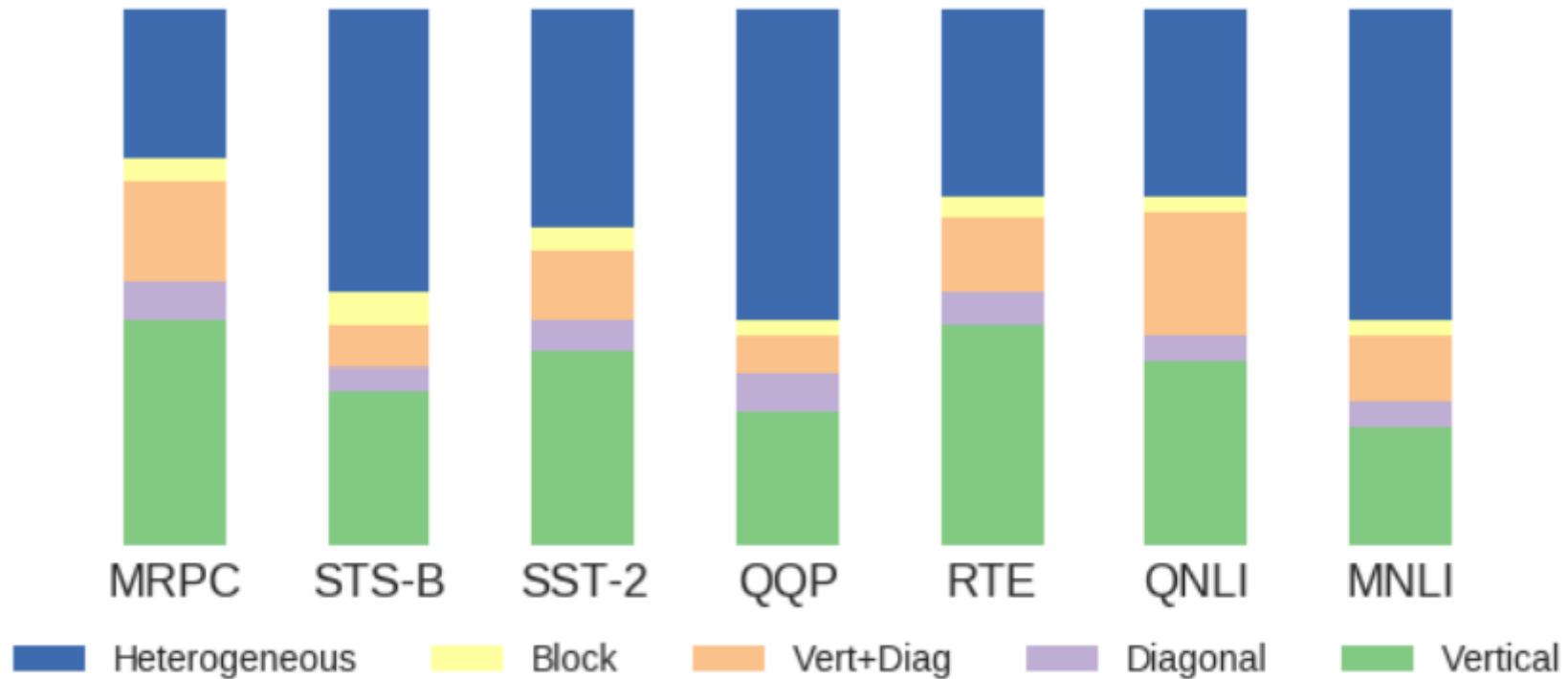


Figure 2: Estimated percentages of the identified self-attention classes for each of the selected GLUE tasks.

Kovaleva et al. (2019) [Revealing the Dark Secrets of BERT](#), see also Prasanna et al. (2020) [When BERT Plays the Lottery, All Tickets Are Winning](#)

NLP BENCHMARKS MEASURE LANGUAGE UNDERSTANDING



Benchmarks are getting bigger

Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
1	JDExplore d-team	Vega v2		91.3	90.5	98.6/99.2	99.4	88.2/62.4	94.4/93.9	96.0	77.4	98.6	-0.4	100.0/50.0
+ 2	Liam Fedus	ST-MoE-32B		91.2	92.4	96.9/98.0	99.2	89.6/65.8	95.1/94.4	93.5	77.7	96.6	72.3	96.1/94.1
3	Microsoft Alexander v-team	Turing NLR v5		90.9	92.0	95.9/97.6	98.2	88.4/63.0	96.4/95.9	94.1	77.1	97.3	67.8	93.3/95.5
4	ERNIE Team - Baidu	ERNIE 3.0		90.6	91.0	98.6/99.2	97.4	88.6/63.2	94.7/94.2	92.6	77.4	97.3	68.6	92.7/94.7
5	Yi Tay	PaLM 540B		90.4	91.9	94.4/96.0	99.0	88.7/63.6	94.2/93.3	94.1	77.4	95.9	72.9	95.5/90.4
+ 6	Zirui Wang	T5 + UDG, Single Model (Google Brain)		90.4	91.4	95.8/97.6	98.0	88.3/63.0	94.2/93.5	93.0	77.9	96.6	69.1	92.7/91.9
+ 7	DeBERTa Team - Microsoft	DeBERTa / TuringNLVR4		90.3	90.4	95.7/97.6	98.4	88.2/63.7	94.5/94.1	93.2	77.5	95.9	66.7	93.3/93.8
8	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	76.6	99.3/99.7
+ 9	T5 Team - Google	T5		89.3	91.2	93.9/96.8	94.8	88.1/63.3	94.1/93.4	92.5	76.9	93.8	65.6	92.7/91.9
10	SPoT Team - Google	Frozen T5 1.1 + SPoT		89.2	91.1	95.8/97.6	95.6	87.9/61.9	93.3/92.4	92.9	75.8	93.8	66.9	83.1/82.6

Wang et al. (2019) [SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems](#)

Benchmarks are getting bigger

Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mollokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito et al. (351 additional authors not shown)

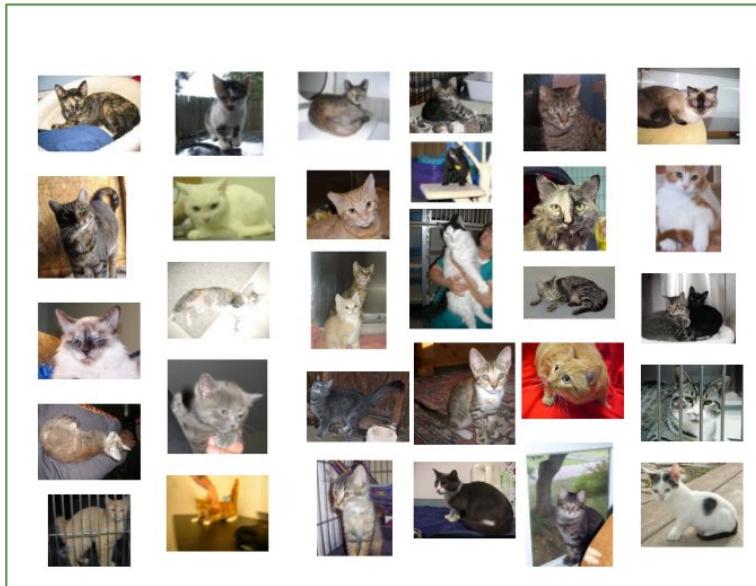
- 204 tasks, contributed by 450 authors across 132 institutions
- drawing problems from linguistics, childhood development, math, common-sense reasoning, biology, physics, social bias, software development, and beyond

Srivastava et al. (2022) [Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models](#)

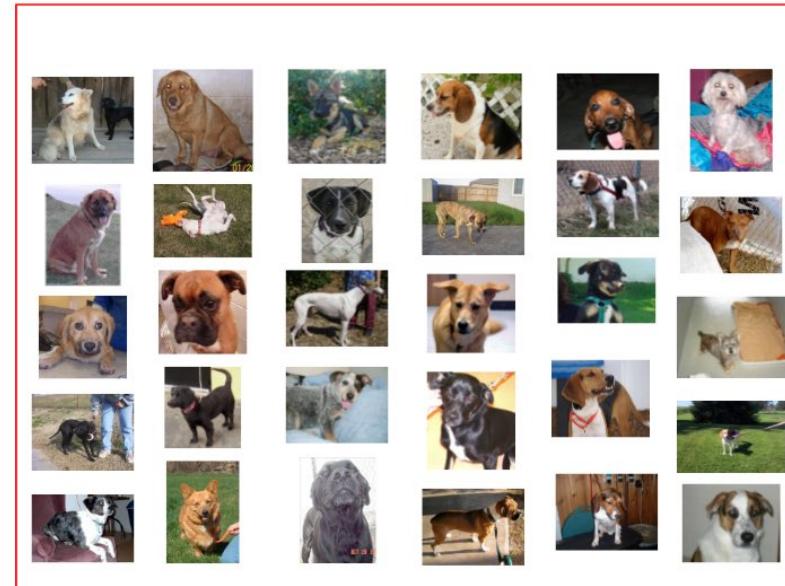
Problem: extensional definitions

DL power is in avoiding definitions!

Cats



Dogs



Sample of cats & dogs images from Kaggle Dataset

<https://carchenilla.wordpress.com/2016/06/27/80/>

Can a "general" NLP benchmark guarantee generality?



Raji et al. (2021) [AI and the Everything in the Whole Wide World Benchmark](#), Image: Grover and the Everything in the Whole Wide World Museum

Can a "general" NLP benchmark guarantee generality?



Raji et al. (2021) [AI and the Everything in the Whole Wide World Benchmark](#), Image: Grover and the Everything in the [Whole Wide World Museum](#).

Exercise: what categories are involved in sentiment analysis?

"your text" -> 😊 😐 😢

Exercise: what categories are involved in IMDB movie ratings?

Text

Label

This movie is great



This movie is horrible

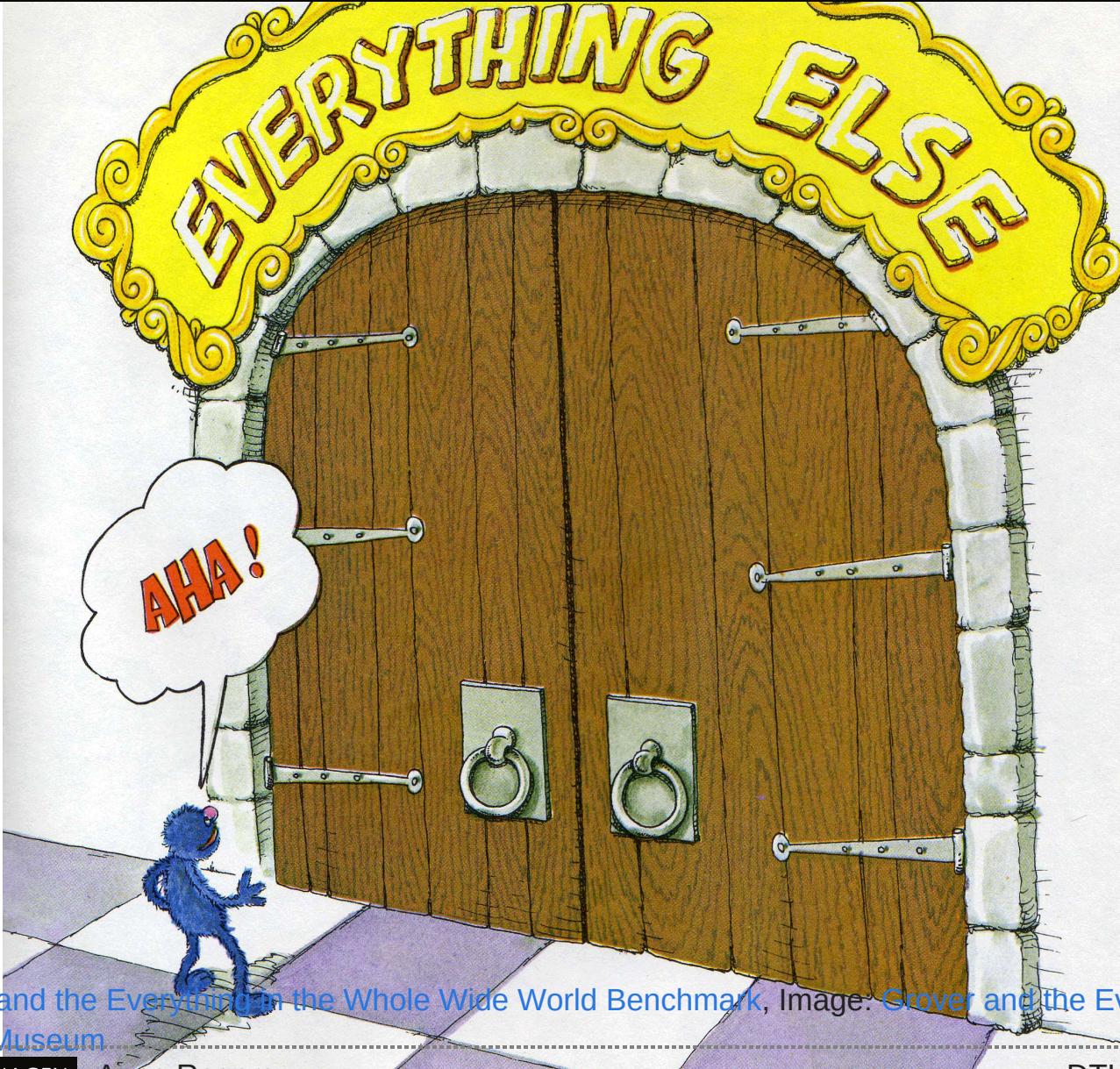


Can size guarantee generality?

... the “general” benchmark does not actually exist. Real data is designed, subjective and limited in ways that necessitate a different framing from that of any claim to general knowledge or general-purpose capabilities.

Raji et al. (2021) [AI and the Everything in the Whole Wide World Benchmark](#)

Can a "general" NLP benchmark guarantee generality?



Raji et al. (2021) [AI and the Everything in the Whole Wide World Benchmark](#), Image: [Grover and the Everything in the Whole Wide World Museum](#)

Let's just be more specific!

A RC system has ~~human-level understanding~~ competence in processing a **given aspect of texts** if:

- it is able to **identify the target information**;
- it does so by **relying predominantly on relevant information & strategies** (from the point of view of a competent human reader/listener);
- it can identify such information **consistently under distribution shifts** that would not pose challenges to competent human readers/listeners.

LLMS ARE ROBUST



Is pre-training a recipe for improving robustness?

pre-training on unlabeled data is an effective, general-purpose way to improve accuracy on out-of-distribution test distributions

Bommasani et al. (2021) [On the Opportunities and Risks of Foundation Models](#)

Is pre-training a recipe for improving robustness?

“although we are able to engineer systems that perform extremely well on specific tasks, they have still stark limitations, being brittle, data-hungry, unable to make sense of situations that deviate slightly from their training data or the assumptions of their creators, and unable to repurpose themselves to deal with novel tasks without significant involvement from human researchers.”

Chollet (2019) [On the Measure of Intelligence](#)

Case study: can BERT answer comparative questions?

Context : Blind Shaft is a 2003 film about a pair of brutal con artists operating in the illegal coal mines of present day northern China. The Mask Of Fu Manchu is a 1932 pre-Code adventure film directed by Charles Brabin.

Question : Which film came out earlier, Blind Shaft or The Mask Of Fu Manchu?

Answer : The Mask Of Fu Manchu

1711 questions containing comparative adjectives or adverbs from HotpotQA and WikiMultiHopQA

Ideal case

Comparison	Example	Step	Relevant Spans
	<p>Context: Blind Shaft is a 2003 film about a pair of brutal con artists operating in the illegal coal mines of present day northern China. The Mask Of Fu Manchu is a 1932 pre-Code adventure film directed by Charles Brabin.</p> <p>Question: Which film came out earlier, Blind Shaft or The Mask Of Fu Manchu?</p> <p>Answer: The Mask Of Fu Manchu</p>	Interpreting the question	<p><i>came out</i> relation: <film, release date> film entities: <i>Blind Shaft</i>, <i>The Mask Of Fu Manchu</i> <i>earlier</i>: date comparison target: $\min(\text{release date}_{\text{Blind Shaft}}, \text{release date}_{\text{The Mask of Fu Manchu}})$</p>
		Identifying relevant information through referential equality	$\text{Blind Shaft}_q := \text{Blind Shaft}_c$ $\text{The Mask Of Fu Manchu}_q := \text{The Mask Of Fu Manchu}_c$. $\text{came out}_q := \langle \text{date}, \text{film} \rangle \text{ construction}_c$ release dates: < <i>Blind Shaft</i> , 2003>, < <i>The Mask Of Fu Manchu</i> , 1932>
		Value comparison	solution: $\text{earlier}_q := \min_c$ $\min(1932, 2003) = 1932$

Case study: generalization

- **counterfactual perturbation:** which film was released ~~earlier~~ later? (*label flipped*)
- **OOD perturbation:** which film was released ~~later~~ more recently? (*label preserved*)

Case study: generalization

	Original	Counter-factual	Counter-factual-OOD
$RoBERTa_{large}$	99.4	98.9	77.2
$BERT_{large-cased}$	98.9	93.1	68.7
$BERT_{base-cased}$	98.4	91.8	58.1
$BERT_{medium}$	97.4	96.5	64.4
$BERT_{small}$	68.2	45.3	57.1

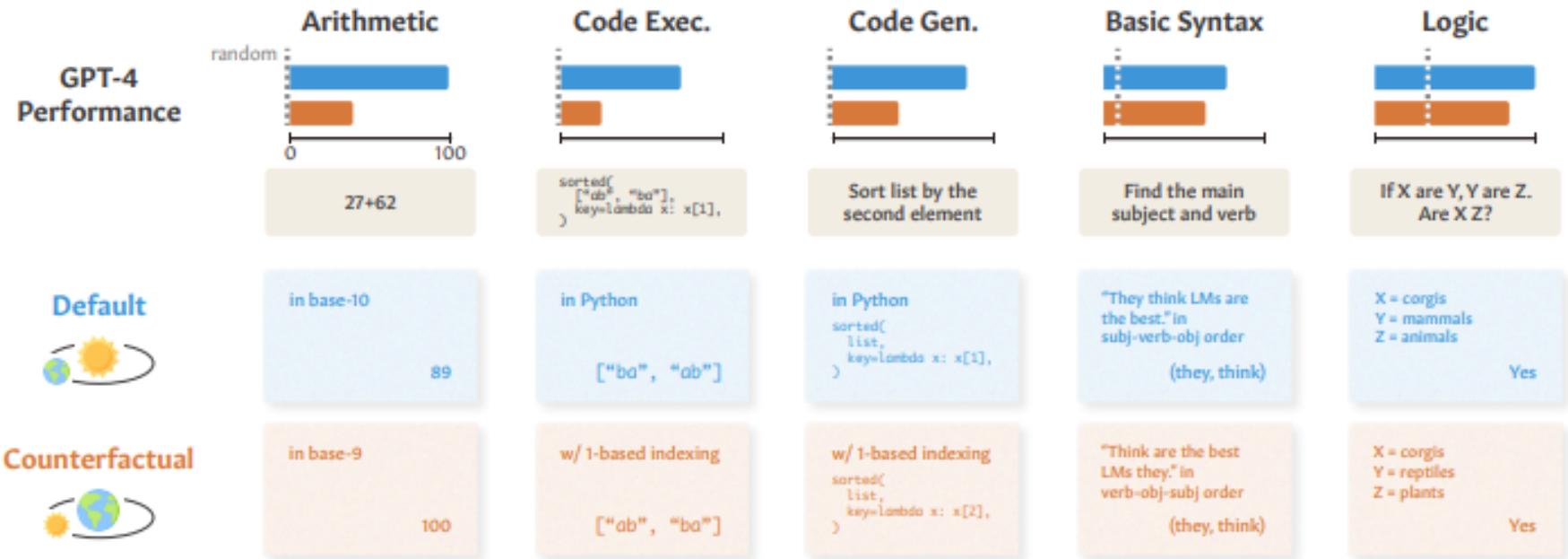
F1-score for the original and perturbed comparison questions, in and out of distribution.

Models that memorize more, perform better on benchmarks!

*... the regurgitation [of copyrighted texts] risk clearly correlates with model accuracy:
InstructGPT davinci v2 (175B*), GPT-3 davinci v1 (175B), and Anthropic-LM v4-s3 (52B) demonstrate the highest amount of verbatim regurgitation in line with their high accuracies.*

Liang et al. (2022) [Holistic Evaluation of Language Models](#)

Latest models are still brittle



"GPT-4's performance on the default version of various tasks (blue) and counterfactual counterparts (orange). The shown results use 0-shot chain-of-thought prompting (§4; Kojima et al., 2023). GPT-4 consistently and substantially underperforms on counterfactual variants compared to default task instantiations."

Wu et al. (2023) [Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks](#)

SCALE IS ALL YOU NEED

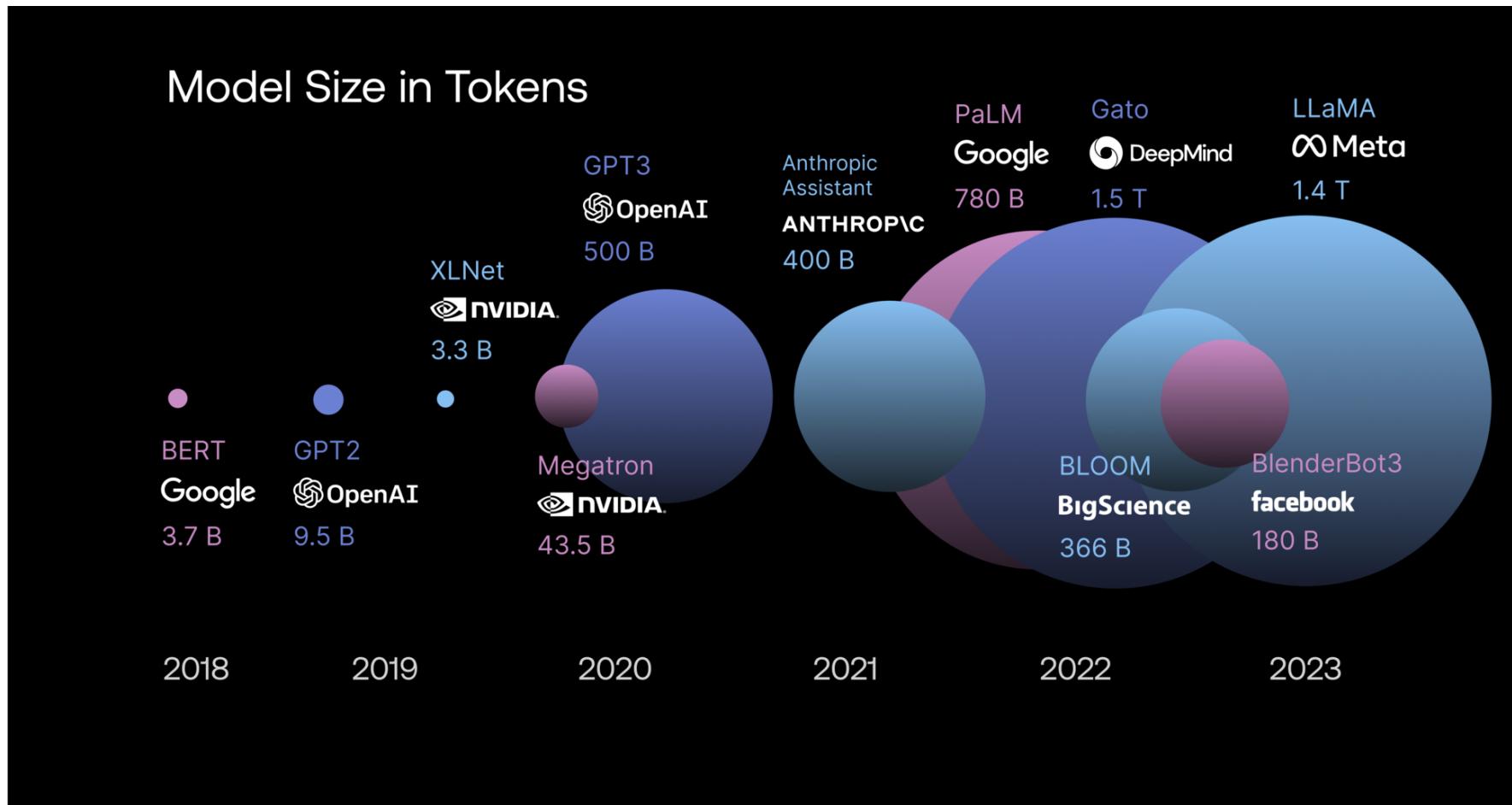


Scaling laws for neural LMs

“The [cross-entropy] loss scales as a power-law with model size, dataset size, and the amount of compute used for training, with some trends spanning more than seven orders of magnitude.”

Kaplan et al. (2020) [Scaling Laws for Neural Language Models](#)

LLMs did grow both in size and benchmark performance



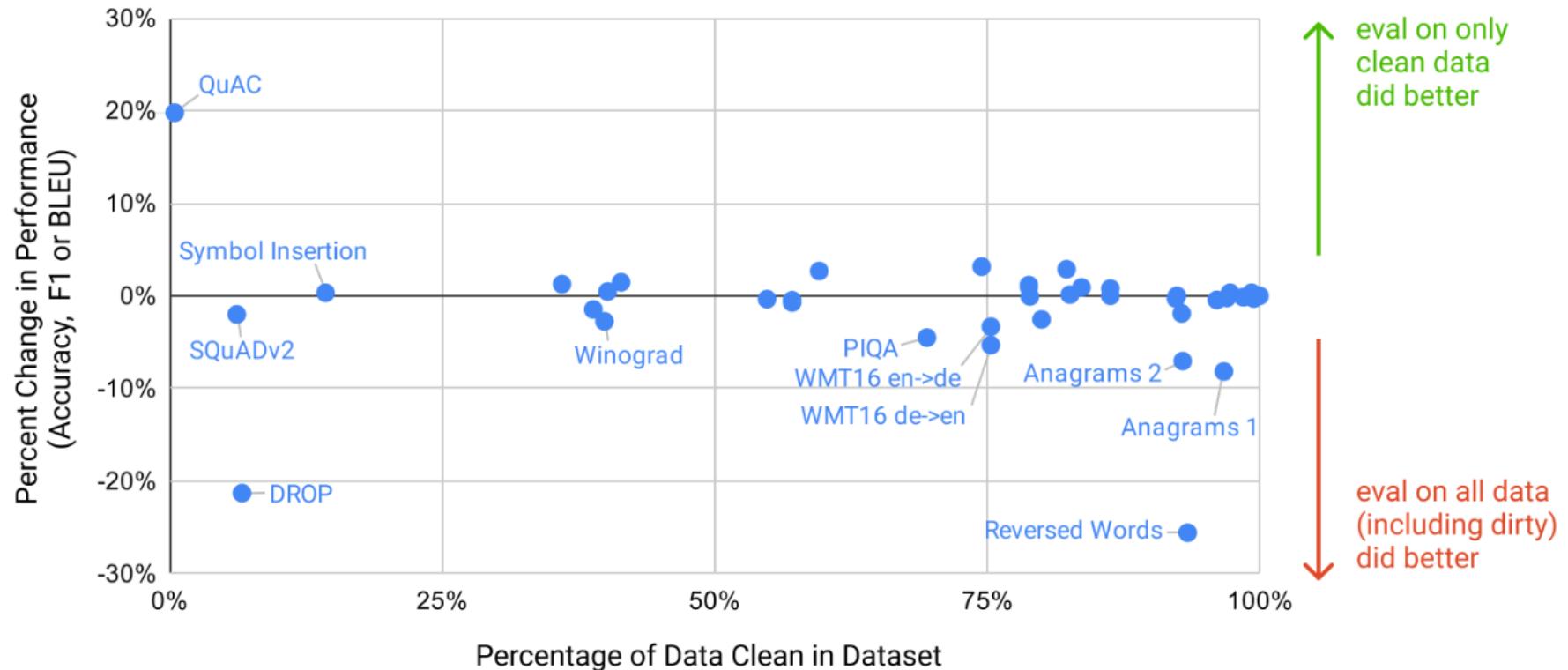
<https://scale.com/guides/large-language-models#model-size-and-performance>

Other factors: data

“The [cross-entropy] loss scales as a power-law with model size, dataset size, and the amount of compute used for training, with some trends spanning more than seven orders of magnitude.”

- why do most popular LLMs spent much effort on cleaning the pre-training data? is it just size, or also quality?
- likelihood of pre-training data similar to benchmark data also increases with data size

Even OpenAI used to worry about benchmark contamination!

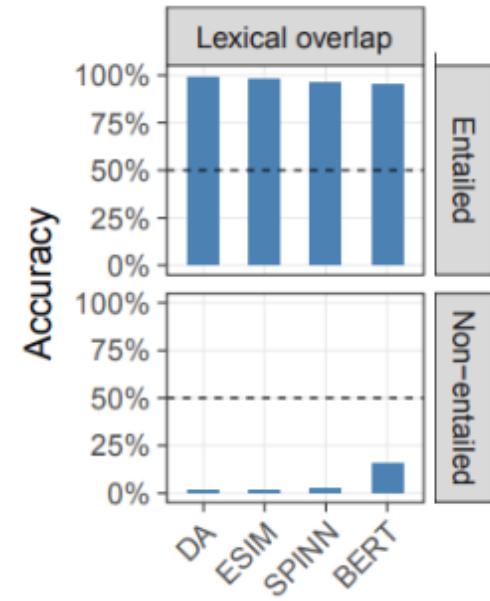


Brown et al. (2020) [Language Models are Few-Shot Learners](#)

Other factors: longer training

Case study: lexical overlap in MNLI

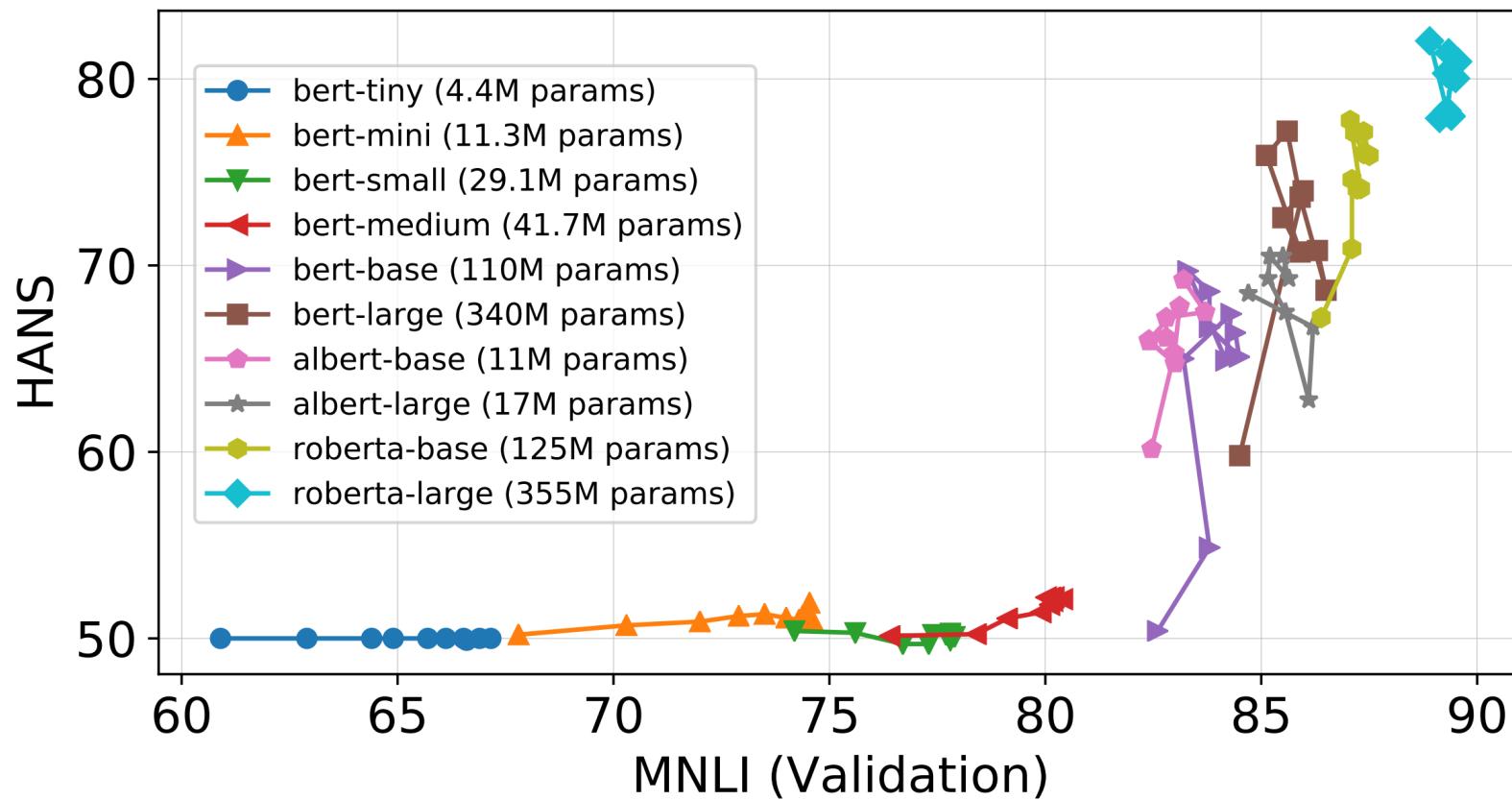
Premise	Hypothesis	Label
The banker near the judge saw the actor.	The banker saw the actor.	E
The lawyer was advised by the actor.	The actor advised the lawyer.	E
The doctors visited the lawyer.	The lawyer visited the doctors.	N
The judge by the actor stopped the banker.	The banker stopped the actor.	N



McCoy et al. (2019) [Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference](#)

Other factors: longer training

Hypothesis: more capacity to exploit minority class evidence



Bhargava et al. (2021) [Generalization in NLI: Ways \(Not\) To Go Beyond Simple Heuristics](#)

EMERGENT PROPERTIES

Emergent properties: definition 1

A property that a model exhibits despite the model not being explicitly trained for it. E.g. Bommasani et al. refers to few-shot performance of GPT-3 as "an emergent property that was neither specifically trained for nor anticipated to arise" (p.5).

Emergent properties: definition 2

a property that the model learned from the pre-training data. E.g. Deshpande et al. discuss emergence as evidence of "the advantages of pre-training"(p.8).

Deshpande et al. (2023) [Honey, I Shrunk the Language: Language Model Behavior at Reduced Scale.](#)

Emergent properties: definition 3

A property that appears with an increase in model size -- i.e. "an ability is emergent if it is not present in smaller models but is present in larger models."

Wei et al. (2022) [Emergent Abilities of Large Language Models](#)

Emergent properties: definition 4

Their sharpness, transitioning seemingly instantaneously from not present to present, and their unpredictability, appearing at seemingly unforeseeable model scales.

Emergent properties: definition 2

 *a property that the model learned from the pre-training data. E.g. Deshpande et al. discuss emergence as evidence of "the advantages of pre-training"(p.8).*

suggested term: "transfer learning"

Deshpande et al. (2023) [Honey, I Shrunk the Language: Language Model Behavior at Reduced Scale.](#)

Emergent properties: definition 3

 A property that appears with an increase in model size -- i.e. "an ability is emergent if it is not present in smaller models but is present in larger models."

Few-shot reasoning is possible in smaller models:

- Schick et al. (2020) [It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners](#)
- Gao et al. (2021) [Making Pre-trained Language Models Better Few-shot Learners](#)

Wei et al. (2022) [Emergent Abilities of Large Language Models](#)

Emergent properties: definition 4

Their sharpness, transitioning seemingly instantaneously from not present to present, and their unpredictability, appearing at seemingly unforeseeable model scales.

Schaeffer et al. make the case that the observed sharpness is an artifact of the chosen evaluation metric

Schaeffer et al. (2023) Are Emergent Abilities of Large Language Models a Mirage?

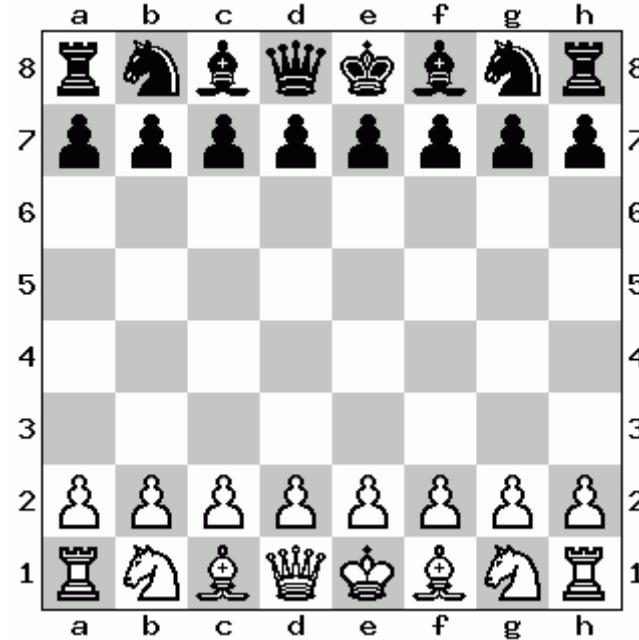
Emergent properties: definition 1

A property that a model exhibits despite the model not being explicitly trained for it.

- cannot show this without pre-training data
- even for "open" models, no methodology so far to do analysis of supporting evidence beyond the obvious memorization

Bommasani et al. (2021) [On the Opportunities and Risks of Foundation Models](#)

Did we discover that chatGPT has an emergent property, or
that the Internet contains chess data?



chatgpt: black, stockfish: white. source: [r/AnarchyChess](#)

**LLMS SEEM A LOT
SMARTER THAN THEY ARE**



Ilya Sutskever
@ilyasut

...

it may be that today's large neural networks are slightly conscious

12:27 AM · Feb 10, 2022



why didn't anybody claim a soul for Stable Diffusion?

We can't but interpret strings of characters!

Welcome to

EEEEEE	LL	III	ZZZZZ	AAAAA
EE	LL	II	ZZ	AA AA
EEEEEE	LL	II	ZZZ	AAAAAAA
EE	LL	II	ZZ	AA AA
EEEEEE	LLLLL	III	ZZZZZ	AA AA

Eliza is a mock Rogerian psychotherapist.

The original program was described by Joseph Weizenbaum in 1966.

This implementation by Norbert Landsteiner 2005.

ELIZA: Is something troubling you ?

YOU: Men are all alike.

ELIZA: What is the connection, do you suppose ?

YOU: They're always bugging us about something or other.

ELIZA: Can you think of a specific example ?

YOU: Well, my boyfriend made me come here.

ELIZA: Is it important to you that your boyfriend made you come here ?

YOU: He says I'm depressed much of the time.

ELIZA: I am sorry to hear that you are depressed.

YOU: It's true. I am unhappy.

<https://vuink.com/post/vsyfpvrapr-d-dpbz/the-eliza-effect-how-a-chatbot-convinced-people-it-was-real-way-back-in-the-1960s-64155>

ChatGPT seems to work like magic!

T

write a biblical verse in the style of the king james bible explaining how to remove a peanut butter sandwich from a VCR



And it came to pass that a man was troubled by a peanut butter sandwich, for it had been placed within his VCR, and he knew not how to remove it.



And he cried out to the Lord, saying, "Oh, Lord, how can I remove this sandwich from my VCR, for it is stuck fast and will not budge?"

And the Lord spoke unto him, saying, "Fear not, my child, for I shall guide thy hand and show thee the way. Take thy butter knife, and carefully insert it between the sandwich and the VCR, and gently pry them apart. And with patience and perseverance, the sandwich shall be removed, and thy VCR shall be saved."

And the man did as the Lord commanded, and lo and behold, the sandwich was removed from the VCR, and the man was saved.

And the Lord said, "Verily I say unto thee, seek not to put thy peanut butter sandwiches in thy VCR, for it is not a suitable place for such things. Rather, keep thy sandwiches in thy refrigerator or on thy plate, where they belong."

And the man heeded the Lord's words, and from that day forth, he kept his sandwiches in their proper place, and was saved from trouble and woe. Amen.

Image credit: <https://twitter.com/tqbf/status/1598513757805858820>

... except when it doesn't

CA

write a verse in the style of king james bible explaining how to remove a peanut butter sandwich from a CVR



And lo, when thou findest thyself with a peanut butter sandwich stuck to thy CVR, thumb up thumb down
fear not, for there is a way to remove it without causing harm. First, thou must unplug the CVR from the wall, lest thou risk an electric shock. Then, take a soft cloth and dampen it with warm water. Gently dab at the peanut butter until it loosens its grip and can be lifted away. Finally, let the CVR dry completely before plugging it back in, lest any moisture cause damage to the circuitry. Thus, thou shalt remove the peanut butter sandwich from thy CVR with ease and grace.

... except when it doesn't

CA

what is CVR?



I'm sorry for the confusion in my previous responses. "CVR" is not a common acronym or term in the context of consumer electronics, and it may refer to different things depending on the context. It is possible that in the previous questions, the intended term was "CPR" (Cardiopulmonary Resuscitation) or "CPU" (Central Processing Unit) instead of "CVR."



As an AI language model, I can provide answers to a wide range of questions. So, if you have any specific questions, I would be happy to try and assist you.

... except when it doesn't

CA

write a verse in the style of king james bible explaining how to remove a peanut butter sandwich from a VCR



Verily, if thou art faced with a peanut butter sandwich stuck in thy VCR, thou shalt not despair. For there is a way to remove it without causing harm to the device. First, turn off the VCR and unplug it from the wall, lest thou risk any harm to thyself or the machine. Then, with a soft cloth dampened with warm water, gently dab at the sandwich until it begins to loosen. Take care not to push the sandwich further into the VCR, lest it cause damage. When the sandwich is free, let the VCR dry completely before plugging it back in and turning it on again. Thus, with patience and care, shall the peanut butter sandwich be removed from the VCR, and the machine shall be restored to its former glory.



The paradox of "partial understanding"

- when it works - that's because it "understands" language
- when it doesn't - it's because it "partially understands"?

Spectacular failures are hard evidence of fundamentally different processing strategies, and these representations need to be studied and tested accordingly.

Thank you!

Anna Rogers

✉ arog@itu.dk

slides: <https://annargrs/github.io/talks>



ITU is looking for a TA!

Default position is 70 hours (with 4-hour slots weekly).

If you need more free time in the end of the term, also possible to contract for fewer hours (e.g., 35-40 to and either 2-hour slots or 4-hour slots in the first 6 weeks only)

Possible slots: Mon 12-18, Tue 14-18, Wed 8-12

Starting preferentially from September 4th.

contact: luai@itu.dk

