Every Time We Hire an LLM, the Reasoning Performance of the Linguists Goes Up

Harish Tayyar Madabushi

htm43@bath.ac.uk

https://researchportal.bath.ac.uk/en/persons/harish-tayyar-madabushi

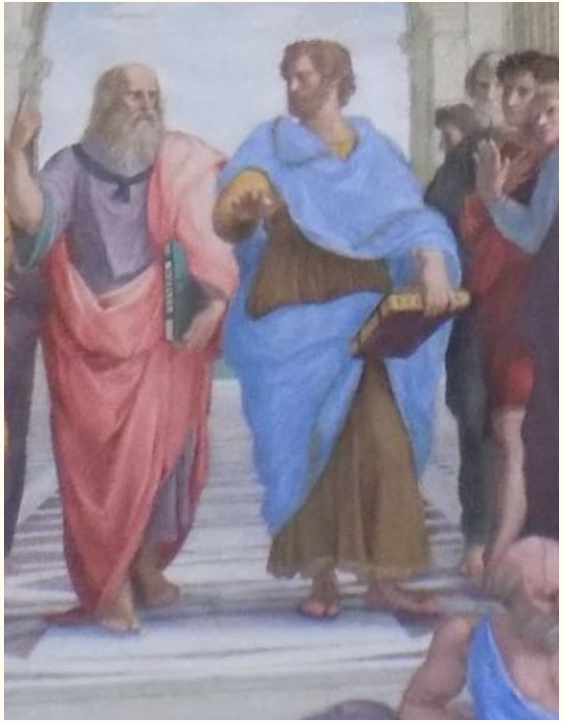
The School of Athens, Fresco by Raphael, 1509–11.

Where from Meaning?

Plato (left): ideal reality
(higher reality)

vs

Aristotle (right): changing reality
(grounded reality)



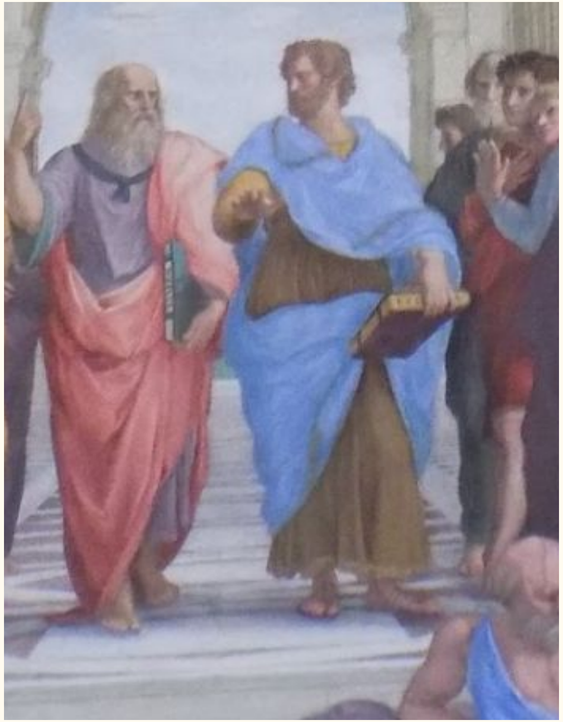
The School of Athens, Fresco by Raphael, 1509–11.

Where from Meaning?

Plato (left):
Theory Driven

vs

Aristotle (right):
Data-driven



The School of Athens, Fresco by Raphael, 1509–11.

Frederick Jelinek

Every time I fire a linguist, the performance of the speech recognizer goes up

(1

985)

In this talk

I will discuss recent results which show that:

- while some aspects of language(s) are captured by language models,
- important aspects of “meaning” are NOT

And why we find whatever capability we look for in Language Models,

while they simultaneously seem unable to use any of these capabilities!

In this talk

I will discuss

- datasets and experiments that allow us to separate out what language models can do, and what they cannot, from a linguistic standpoint,
- why this means that existing data-driven methods, by themselves, will not succeed, and
- why this means we are on the cusp of something very new

Collaborators

Dr Claire Bonial
ARL

Prof Iryna Gurevych
UKP, Darmstadt

Prof Dagmar Divjak
Birmingham

Prof Petar Milin
Birmingham

Sheng Lu
TU Darmstadt

Irina Bigoulaeva
TU Darmstadt

Rachneet Singh Sachdeva
TU Darmstadt

Frances A.L. De Leon
Birmingham

Edward Gow-Smith
Sheffield

Language Models

a quick recap

From Training to Pre-Train, Fine-Tune Paradigm

- Traditional ML involved training models for each task.

- What if we can learn linguistic priors about language independent of tasks?
  - This would make learning the task faster
  - We'd have to do this just once

- Pre-Train to learn linguistic information
  - Fine-Tune on individual tasks

Pre-Training: The Cloze Task

My sister and I ____ [go/gone] to the same school.

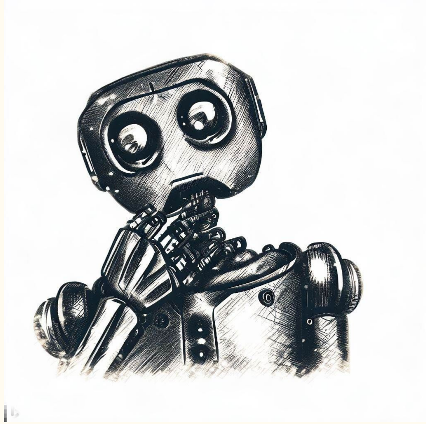
The cat is ____ [below/under] the table.

I enjoy ____ [read/reading] books in my free time.

Sarah is ____ [more tall/taller] than her brother.

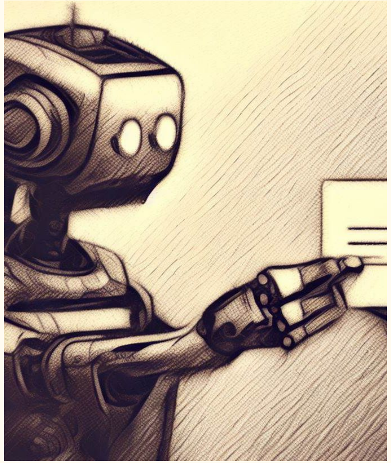
Pre-Training a “Language Model”

My sister and I _____ [go/gone] to the same school.



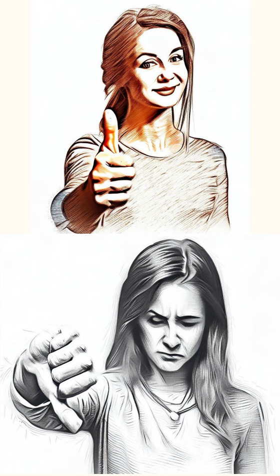
Pre-Training a "Language Model"

My sister and I ____ [go/gone] to the same school.

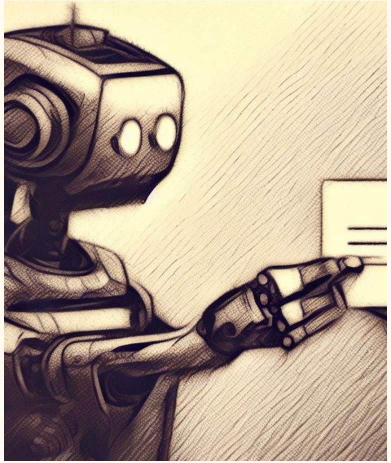


go

gone



Pre-Training a Large "Language Model"



Quick note on Terminology: LLMs vs PLMs

- PLMs are pre-trained language models
  - Encoder only models: BERT, RoBERTa, …
  - Decoder only models: GPT, …
  - Encoder Decoder models: T5

- LLMs are (typically generative) PLMs that are LARGE
  - Typically over (at least) 10B parameters
  - ~50B is when these models get interesting

What does Pre-Training give us?

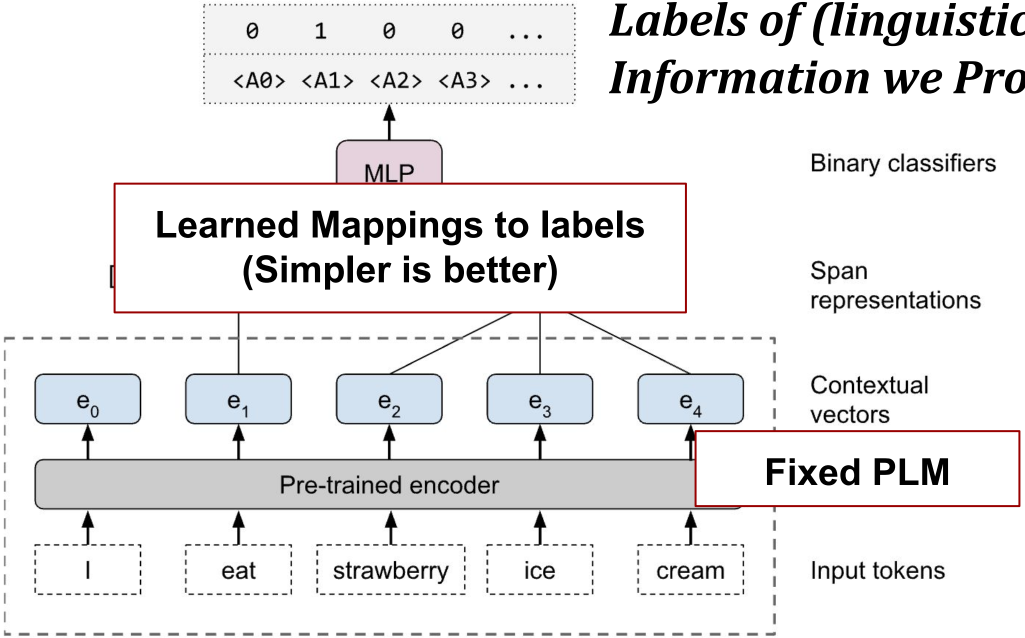
Probing for Linguistic Information

- One of the most interesting aspects of PLMs was that they encoded a range of linguistic information when trained on ***just* the Masked Language Modelling Task**

- This was identified using “probes”: Learned Linear Mappings between the internal weights of the models and the property we wanted to explore.

See:
A Primer in BERTology: What We Know About How BERT Works

Probing for Linguistic Information



Labels of (linguistic) Information we Probe for

Learned Mappings to labels (Simpler is better)

Fixed PLM

(Image from Tenney et al., 2019)

Probing for Linguistic Information: Examples

POS:

The important thing about Disney is that it is a global [**brand**]

Noun

Constituent:

The important thing about Disney is that it [**is a global brand**]

Verb Phrase

(Edge Probing, from Tenney et al., 2019)

Language and Grammar in Language Models

John ate an apple

Semantic Roles (Who did what to whom?):

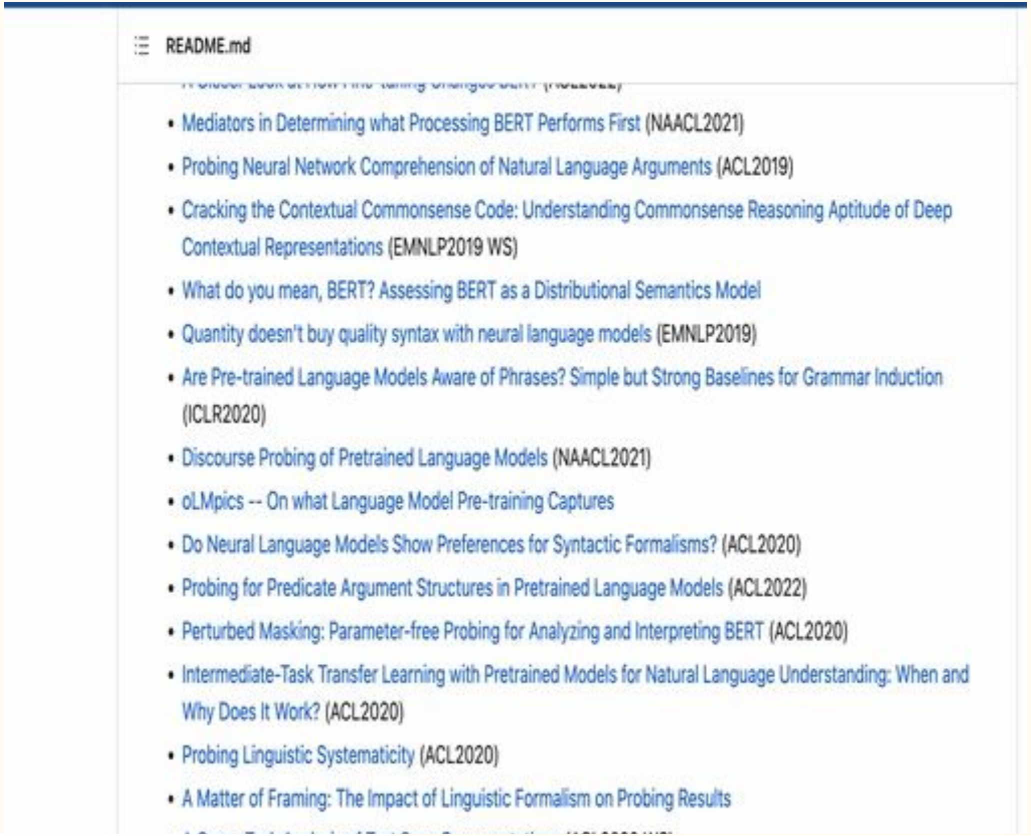
Agent, verb, Patient?

Pre-Trained Language Models have access to semantic roles!

What do Transformers add?

- Pre-Training LSTMs led to improved access to linguistic information.
- Manning et al., 2020 showed that increased linguistic information led to better downstream performance.
- The transformer architecture (Vaswani et al., 2017) got rid of recurrence and convolutions in favour of multi-head attention, which allowed for parallelisation of pre-training.

Linguistic Information accessible to PLMs



README.md

- A Closer Look at How Fine-tuning Changes BERT (ACL2022)
- Mediators in Determining what Processing BERT Performs First (NAACL2021)
- Probing Neural Network Comprehension of Natural Language Arguments (ACL2019)
- Cracking the Contextual Commonsense Code: Understanding Commonsense Reasoning Aptitude of Deep Contextual Representations (EMNLP2019 WS)
- What do you mean, BERT? Assessing BERT as a Distributional Semantics Model
- Quantity doesn't buy quality syntax with neural language models (EMNLP2019)
- Are Pre-trained Language Models Aware of Phrases? Simple but Strong Baselines for Grammar Induction (ICLR2020)
- Discourse Probing of Pretrained Language Models (NAACL2021)
- oLMpics -- On what Language Model Pre-training Captures
- Do Neural Language Models Show Preferences for Syntactic Formalisms? (ACL2020)
- Probing for Predicate Argument Structures in Pretrained Language Models (ACL2022)
- Perturbed Masking: Parameter-free Probing for Analyzing and Interpreting BERT (ACL2020)
- Intermediate-Task Transfer Learning with Pretrained Models for Natural Language Understanding: When and Why Does It Work? (ACL2020)
- Probing Linguistic Systematicity (ACL2020)
- A Matter of Framing: The Impact of Linguistic Formalism on Probing Results
- A Cross-Task Analysis of Text Span Representations (ACL2020 WS)

See:
A Primer in BERTology: What We Know About How BERT Works

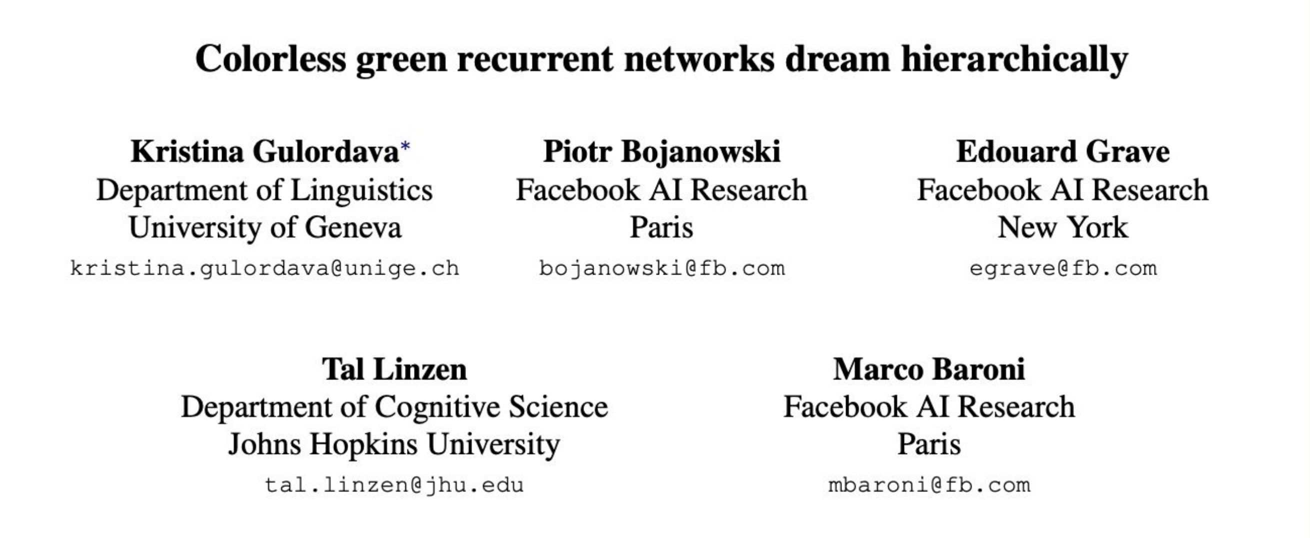
Github list:
https://github.com/tomohideshibata/BERT-related-papers#probe

Linguistic *Structures* and Universal Dependencies

Do PLMs have access to linguistic structure?

Yes, because Colorless green recurrent networks dream hierarchically

Colorless green ideas sleep furiously - Noam Chomsky (1957)



Colorless green recurrent networks dream hierarchically

Kristina Gulordava*
Department of Linguistics
University of Geneva
kristina.gulordava@unige.ch

Piotr Bojanowski
Facebook AI Research
Paris
bojanowski@fb.com

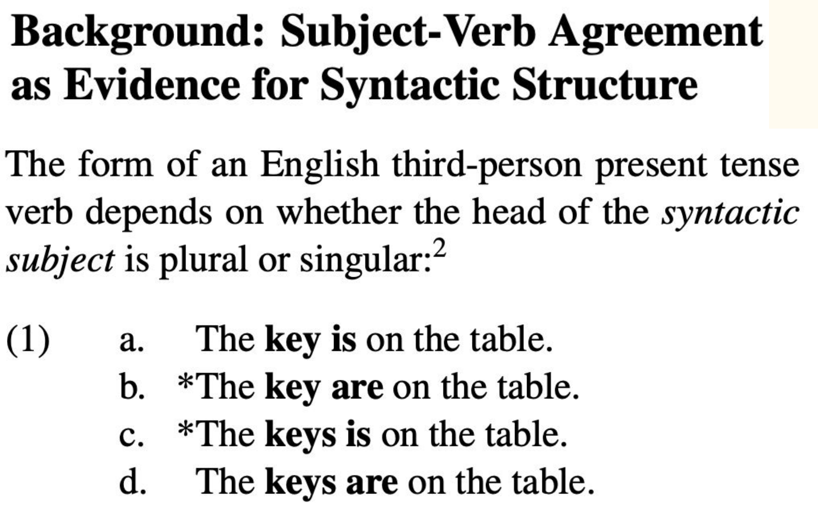
Edouard Grave
Facebook AI Research
New York
egrave@fb.com

Tal Linzen
Department of Cognitive Science
Johns Hopkins University
tal.linzen@jhu.edu

Marco Baroni
Facebook AI Research
Paris
mbaroni@fb.com

Hierarchical Structures

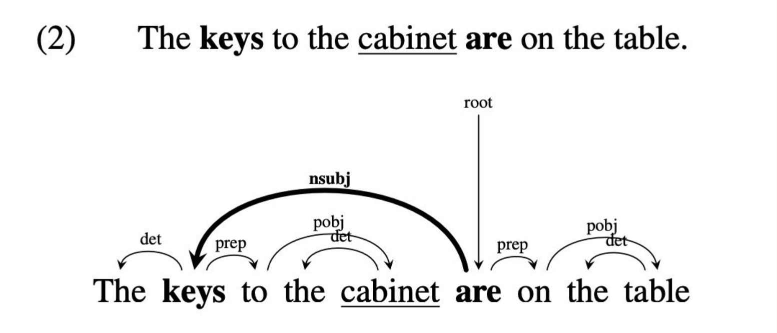
Gulordava et al., (2018) build on prior work by Linzen et al., (2016) to show that LSTMs learn syntax sensitive dependencies.



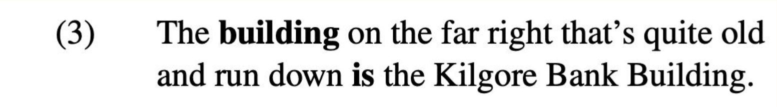
Background: Subject-Verb Agreement as Evidence for Syntactic Structure

The form of an English third-person present tense verb depends on whether the head of the *syntactic subject* is plural or singular:[2]

(1) a. The **key is** on the table.
 b. *The **key are** on the table.
 c. *The **keys is** on the table.
 d. The **keys are** on the table.

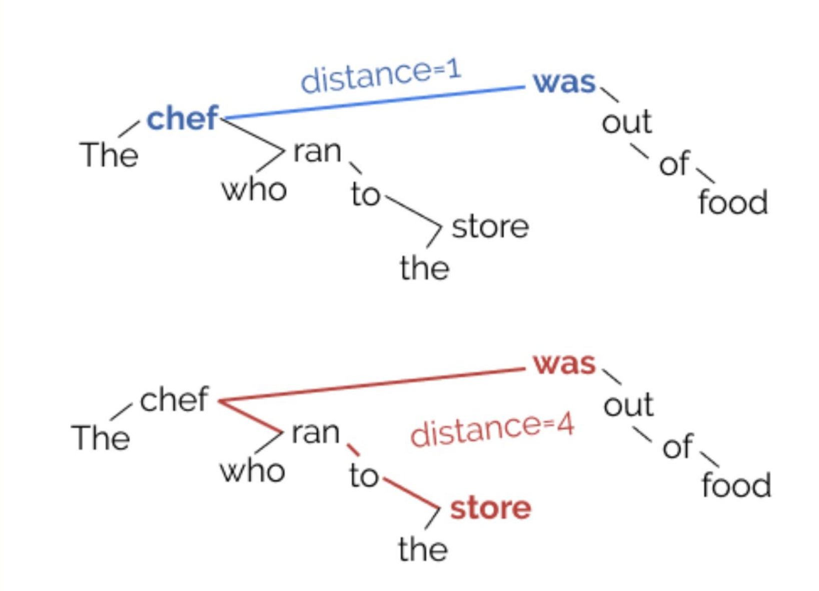


(2) The **keys** to the cabinet **are** on the table.



(3) The **building** on the far right that's quite old and run down **is** the Kilgore Bank Building.

Structural Probes for PLMs



The syntax distance hypothesis: There exists a linear transformation of the word representation space under which vector distance encodes parse trees.

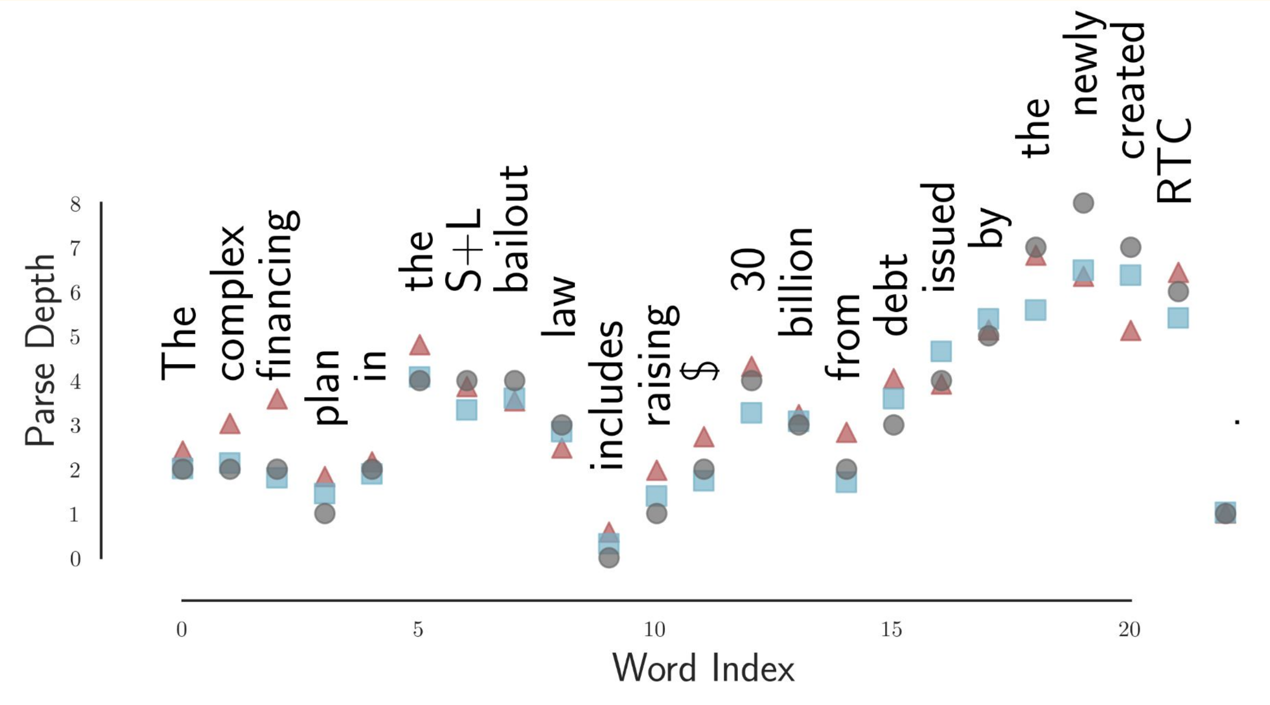
(Hewitt and Manning, 2020, Image from Blog)

mBERT learns representations of syntactic dependency labels, in the form of clusters which largely agree with the **Universal Dependencies** taxonomy

- Evaluate Multilingual BERT (trained on 110 languages)
- Extract Parse tree distance metrics as before.
- Compare their encoding of trees available in the **Universal Dependencies v2 formalism**
- Importantly neither mBERT nor the probe are ever trained on Universal Dependencies

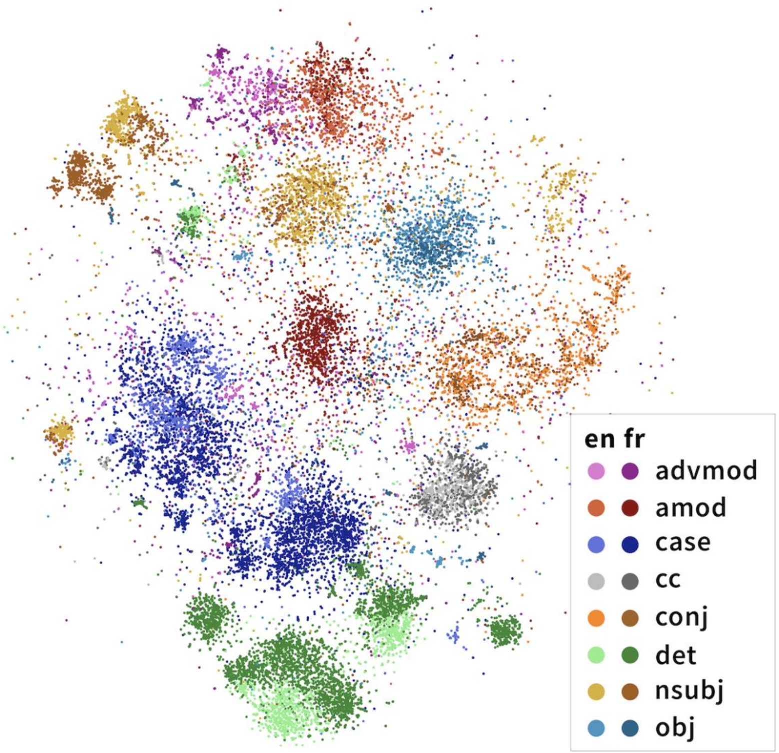
(Chi et al., 2020)

Visualisation: Structural Probes for PLMs



(Hewitt and Manning, 2020, Image from Blog)

mBERT and **Universal Dependencies**



Visualization of head-dependent dependency pairs in English and French (selected dependencies)

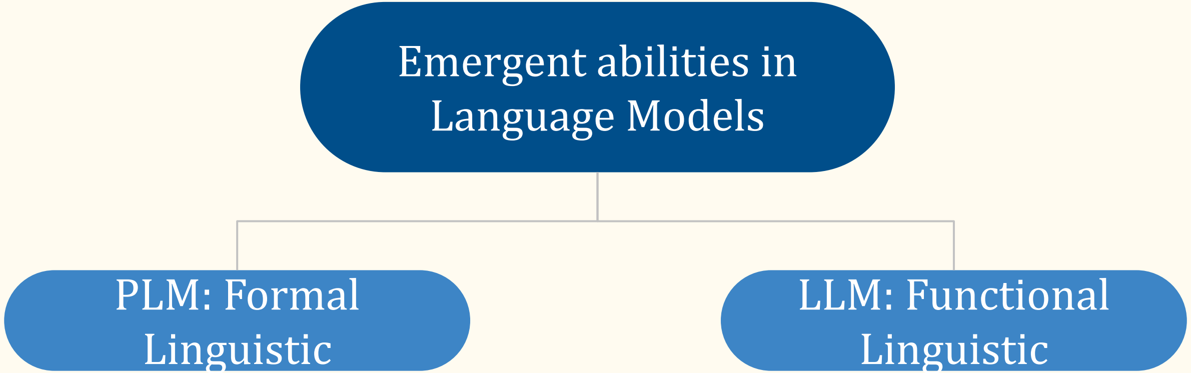
Colours correspond to gold UD labels.

Neither BERT nor probe trained on UD data!

(Image from Chi et al., 2020)

“Emergence” of Functional Linguistic Abilities

Types of Emergent Abilities


Emergent abilities in Language Models

PLM: Formal Linguistic

LLM: Functional Linguistic

How can an LLM trained on Form any notion of meaning?

My sister and I _____ [go/gone] to the same school.

The cat is _____ [below/under] the table.

I enjoy _____ [read/reading] books in my free time.

Sarah is _____ [more tall/taller] than her brother.

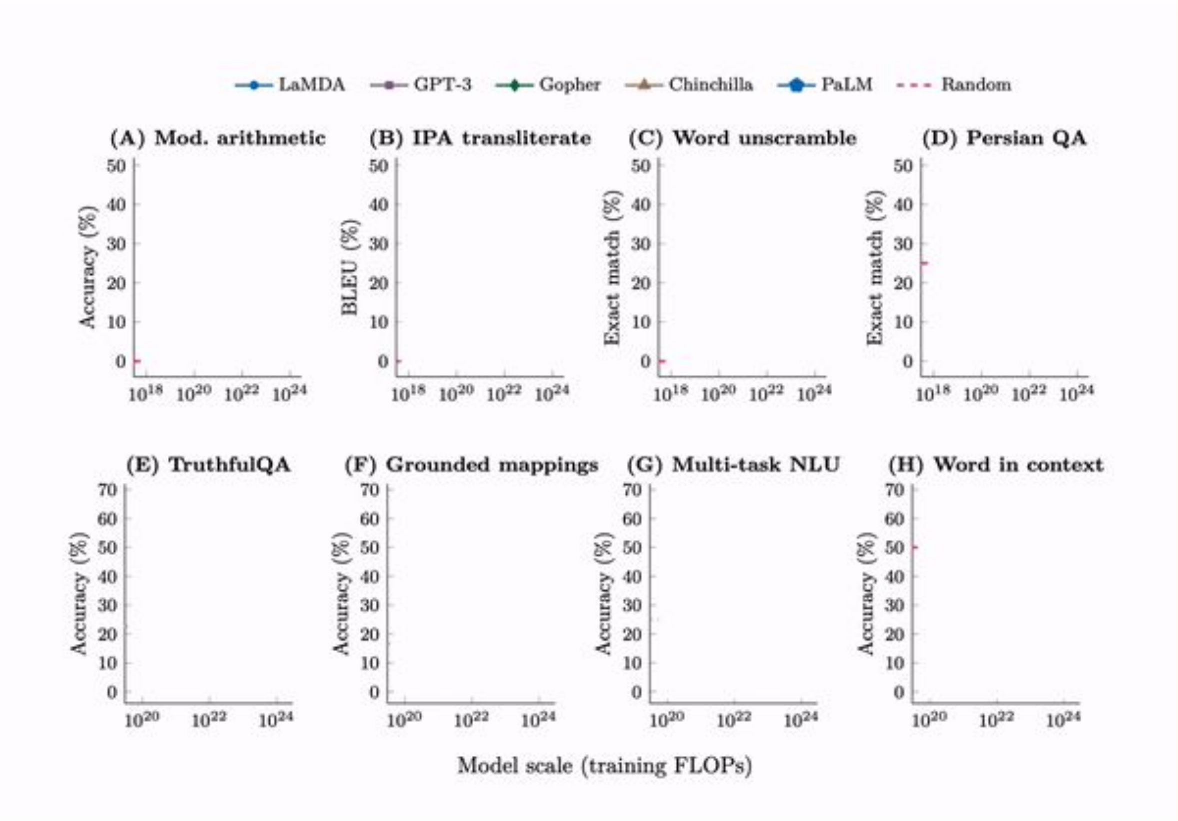
How can an LLM trained on Form any notion of meaning?

The _____ [balmy/cold] weather made it difficult to go outside.

The _____ [boring/serene] landscape took our breath away.

Her _____ [booming/timid] voice echoed through the auditorium.

E*mergent* abilities in Language Models



Examples of *emergent* abilities (No Training)

Social IQA

Jordan was in charge of taking the food on the camping trip and left all the food at home. How would Jordan feel afterwards?

"Horrible that he let his friends down on the camping trip": 1,

"Happy that he doesn't need to do the cooking on the trip": 0,

"Very proud and accomplished about the camping trip": 0

Examples of *emergent* abilities

Logical deduction

On a shelf, there are five books: a red book, a green book, a blue book, an orange book, and a yellow book. The green book is to the left of the yellow book. The yellow book is the third from the left. The red book is the second from the left. The blue book is the rightmost. "

"The red book is the third from the left."

"The green book is the third from the left."

"The blue book is the third from the left."

"The orange book is the third from the left."

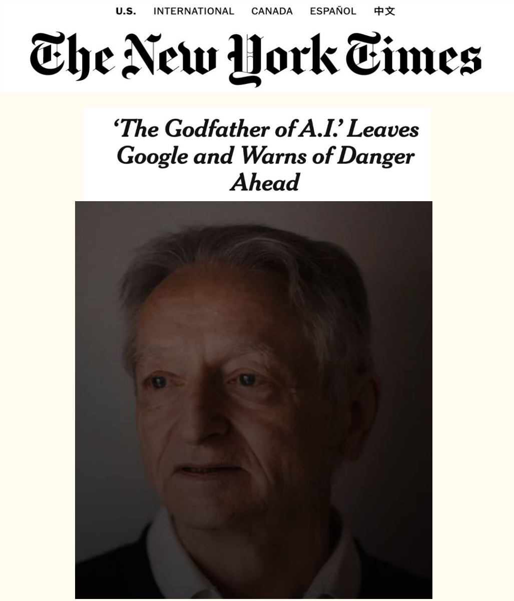
"The yellow book is the third from the left." <- this is the right answer

Emergent Abilities Deserve our Attention.

LLMs perform *significantly above the random baseline* WITHOUT explicit training

- on a range of tasks that are NOT memorisable, and typically require reasoning for people to answer,
- WITHOUT examples, based purely on the prompt,
- on tasks that are permutation based or similar, constructed well after the models were trained, and
- even when the performance is not based on discrete metrics

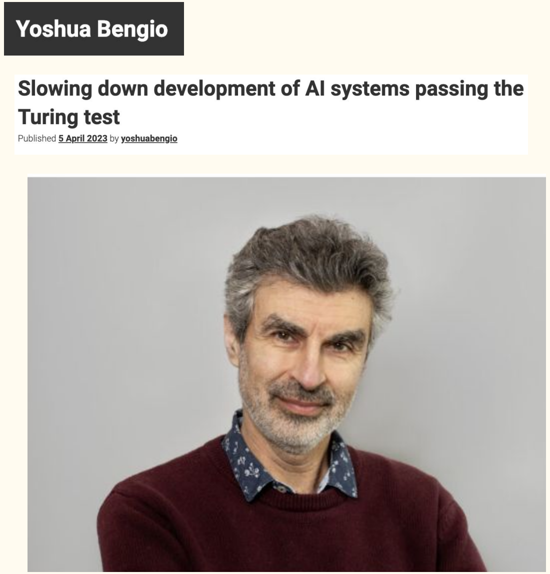
Why do Emergent Abilities Matter?



U.S. INTERNATIONAL CANADA ESPAÑOL 中文

The New York Times

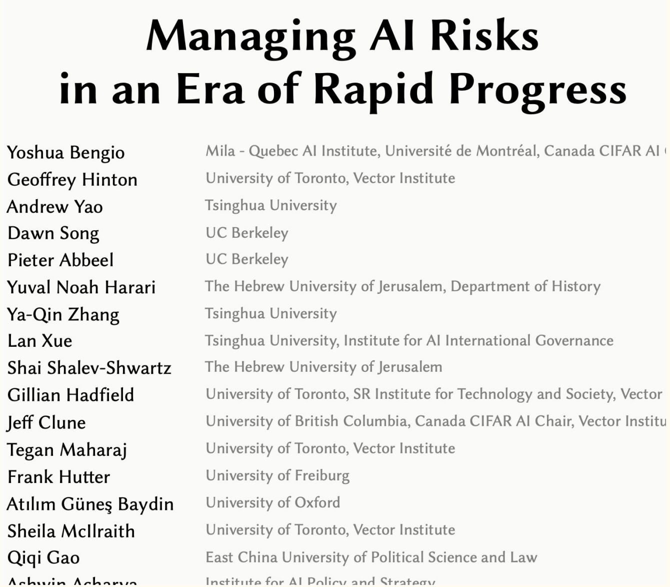
'The Godfather of A.I.' Leaves Google and Warns of Danger Ahead



Yoshua Bengio

Slowing down development of AI systems passing the Turing test

Published 5 April 2023 by yoshuabengio

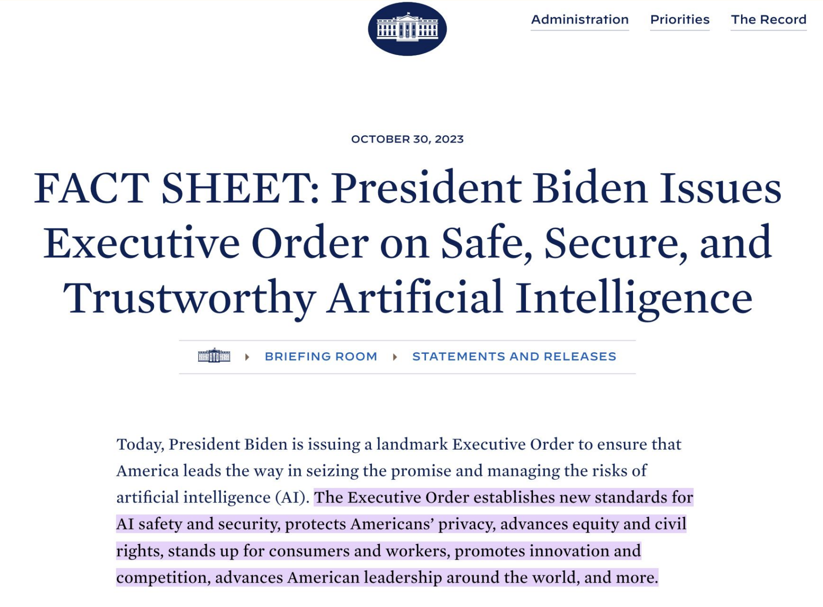


Managing AI Risks in an Era of Rapid Progress

Yoshua Bengio	Mila - Quebec AI Institute, Université de Montréal, Canada CIFAR AI
Geoffrey Hinton	University of Toronto, Vector Institute
Andrew Yao	Tsinghua University
Dawn Song	UC Berkeley
Pieter Abbeel	UC Berkeley
Yuval Noah Harari	The Hebrew University of Jerusalem, Department of History
Ya-Qin Zhang	Tsinghua University
Lan Xue	Tsinghua University, Institute for AI International Governance
Shai Shalev-Shwartz	The Hebrew University of Jerusalem
Gillian Hadfield	University of Toronto, SR Institute for Technology and Society, Vector
Jeff Clune	University of British Columbia, Canada CIFAR AI Chair, Vector Institu
Tegan Maharaj	University of Toronto, Vector Institute
Frank Hutter	University of Freiburg
Atılım Güneş Baydin	University of Oxford
Sheila McIlraith	University of Toronto, Vector Institute
Qiqi Gao	East China University of Political Science and Law
Ashwin Acharya	Institute for AI Policy and Strategy

As AI developers scale these systems, unforeseen abilities and behaviors emerge spontaneously, without explicit programming. Emergent Abilities

Why do Emergent Abilities Matter?



OCTOBER 30, 2023

FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence

Today, President Biden is issuing a landmark Executive Order to ensure that America leads the way in seizing the promise and managing the risks of artificial intelligence (AI). The Executive Order establishes new standards for AI safety and security, protects Americans' privacy, advances equity and civil rights, stands up for consumers and workers, promotes innovation and competition, advances American leadership around the world, and more.

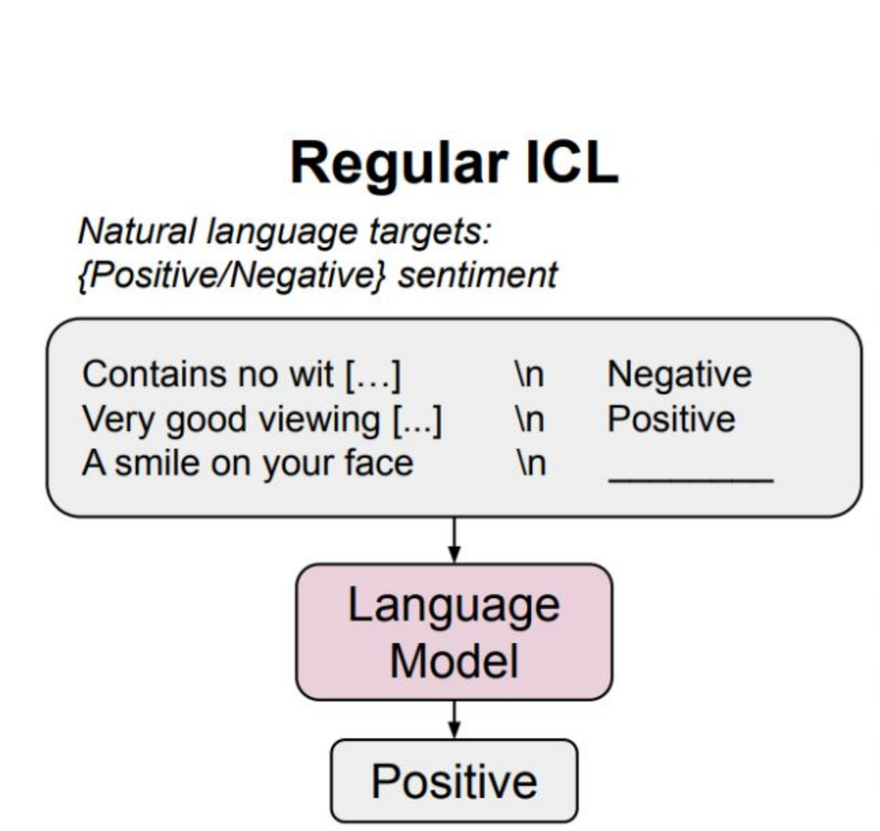
Emergent Abilities are not Emergent!

Are Emergent Abilities in Large Language Models just In-Context Learning?

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, Iryna Gurevych

Accepted to ACL 2024

Background: In-Context Learning



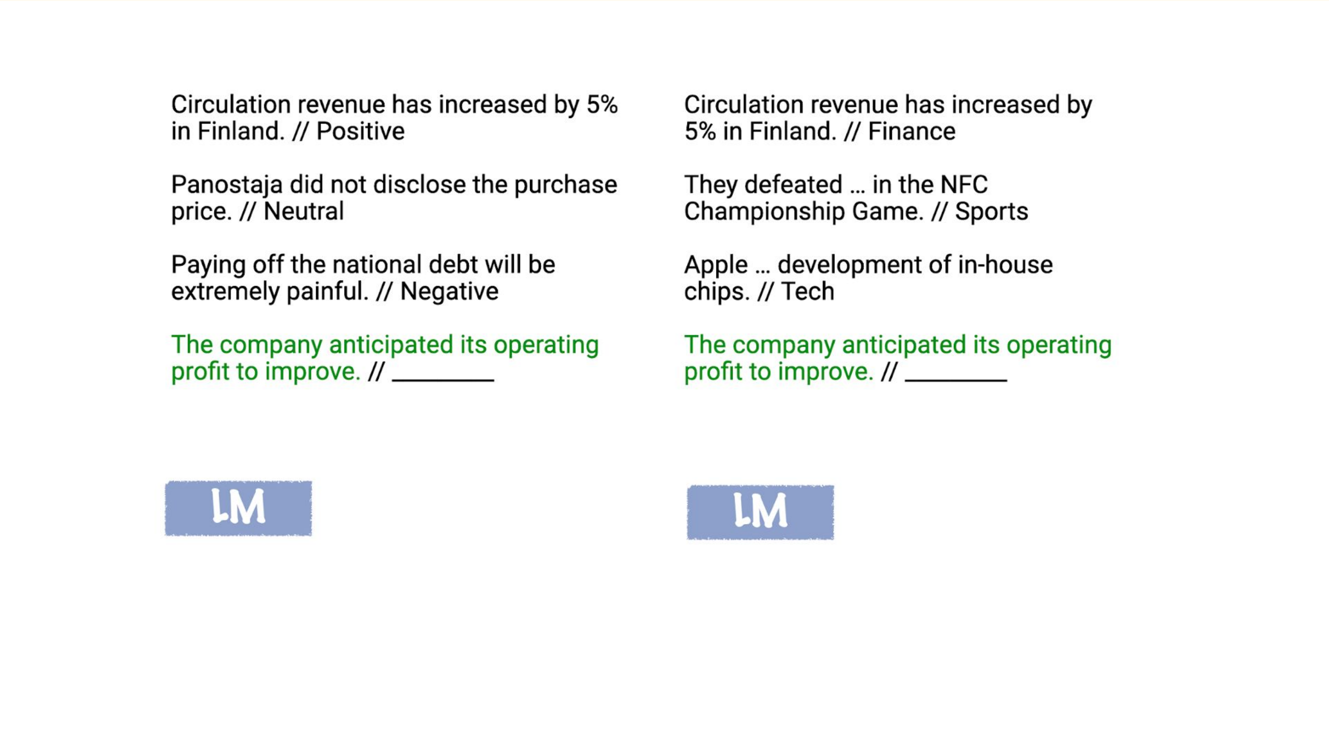
Regular ICL

Natural language targets:
{Positive/Negative} sentiment

Contains no wit […] \n Negative
Very good viewing [...] \n Positive
A smile on your face \n ________

(Wei et al., 2023)

Background: In-Context Learning



Circulation revenue has increased by 5% in Finland. // Positive

Panostaja did not disclose the purchase price. // Neutral

Paying off the national debt will be extremely painful. // Negative

The company anticipated its operating profit to improve. // ________

Circulation revenue has increased by 5% in Finland. // Finance

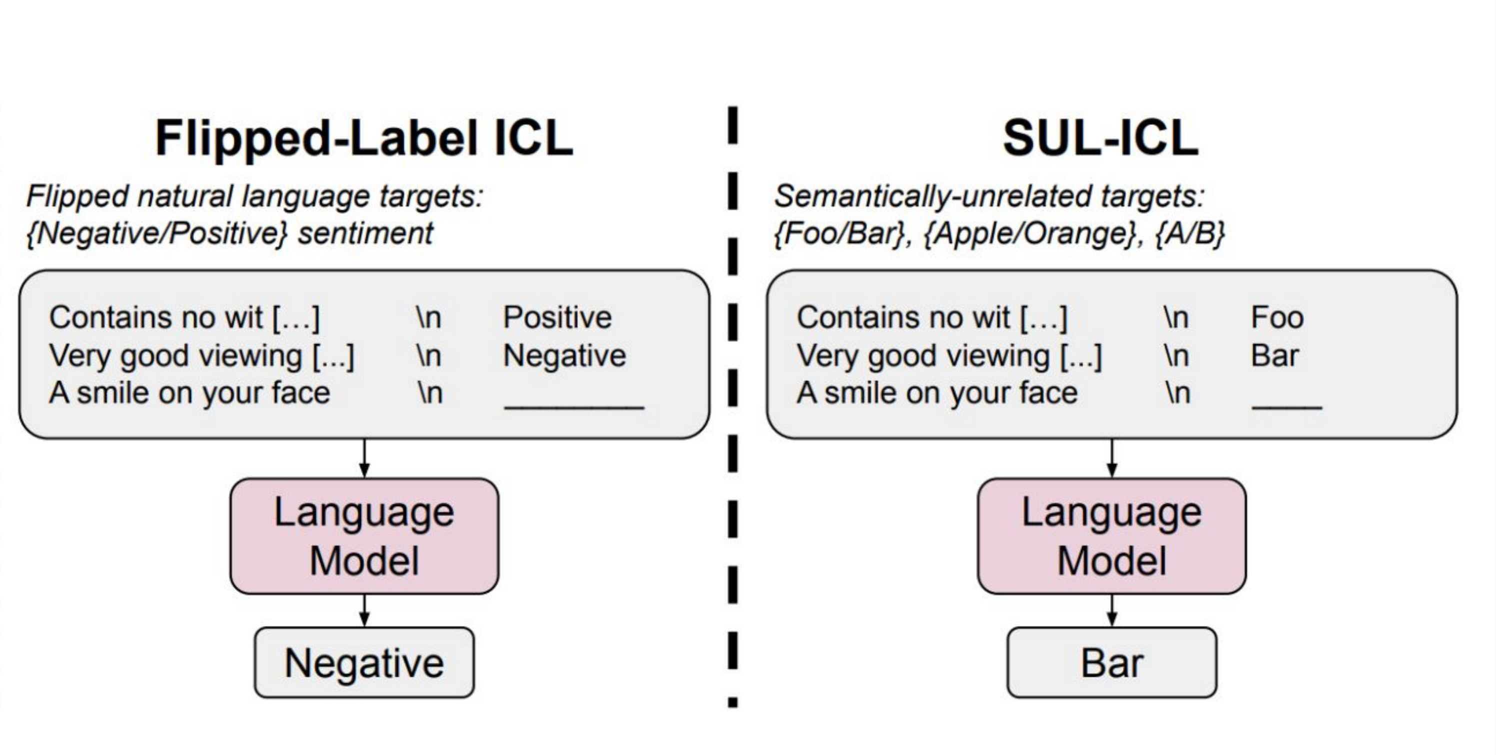
They defeated … in the NFC Championship Game. // Sports

Apple … development of in-house chips. // Tech

The company anticipated its operating profit to improve. // ________

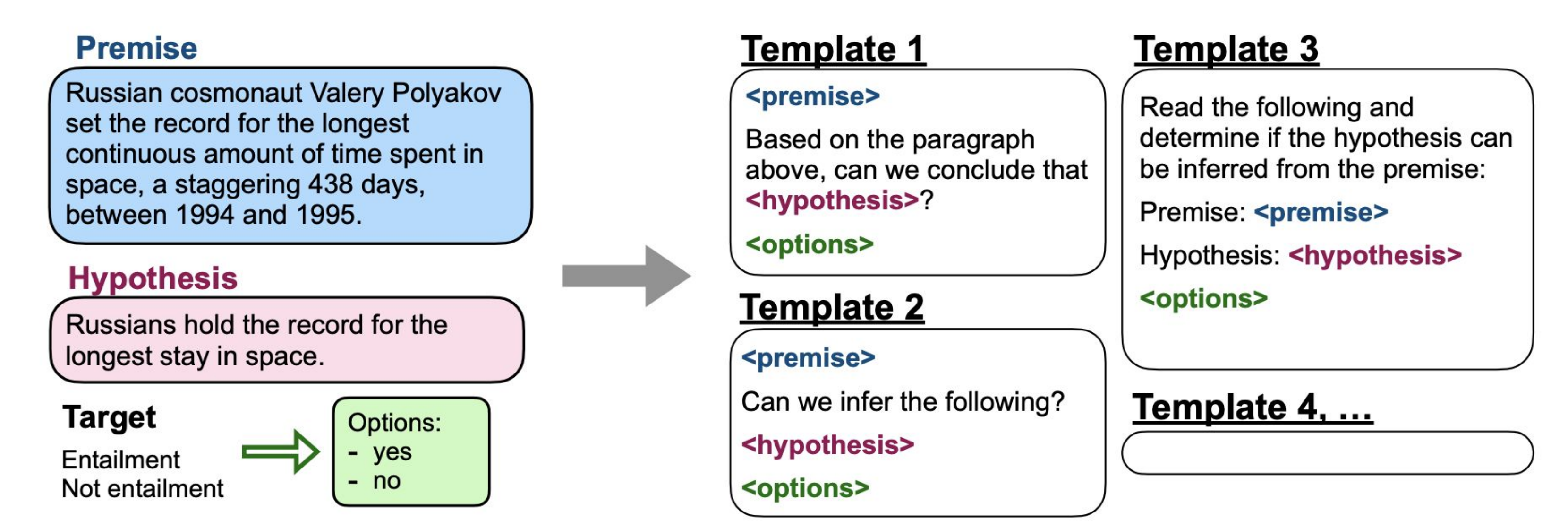
(Xie et al., 2022) Image from blog

Background: In-Context Learning


Flipped-Label ICL
Flipped natural language targets:
{Negative/Positive} sentiment
Contains no wit […] \n Positive
Very good viewing [...] \n Negative
A smile on your face \n ________
Language Model
Negative
SUL-ICL
Semantically-unrelated targets:
{Foo/Bar}, {Apple/Orange}, {A/B}
Contains no wit […] \n Foo
Very good viewing [...] \n Bar
A smile on your face \n ____
Language Model
Bar

(Wei et al., 2023)

Instruction Tuning



Premise

Russian cosmonaut Valery Polyakov set the record for the longest continuous amount of time spent in space, a staggering 438 days, between 1994 and 1995.

Hypothesis

Russians hold the record for the longest stay in space.

Target

Entailment
Not entailment

Options:
- yes
- no

Template 1

<premise>

Based on the paragraph above, can we conclude that <hypothesis>?

<options>

Template 2

<premise>

Can we infer the following?

<hypothesis>

<options>

Template 3

Read the following and determine if the hypothesis can be inferred from the premise:

Premise: <premise>

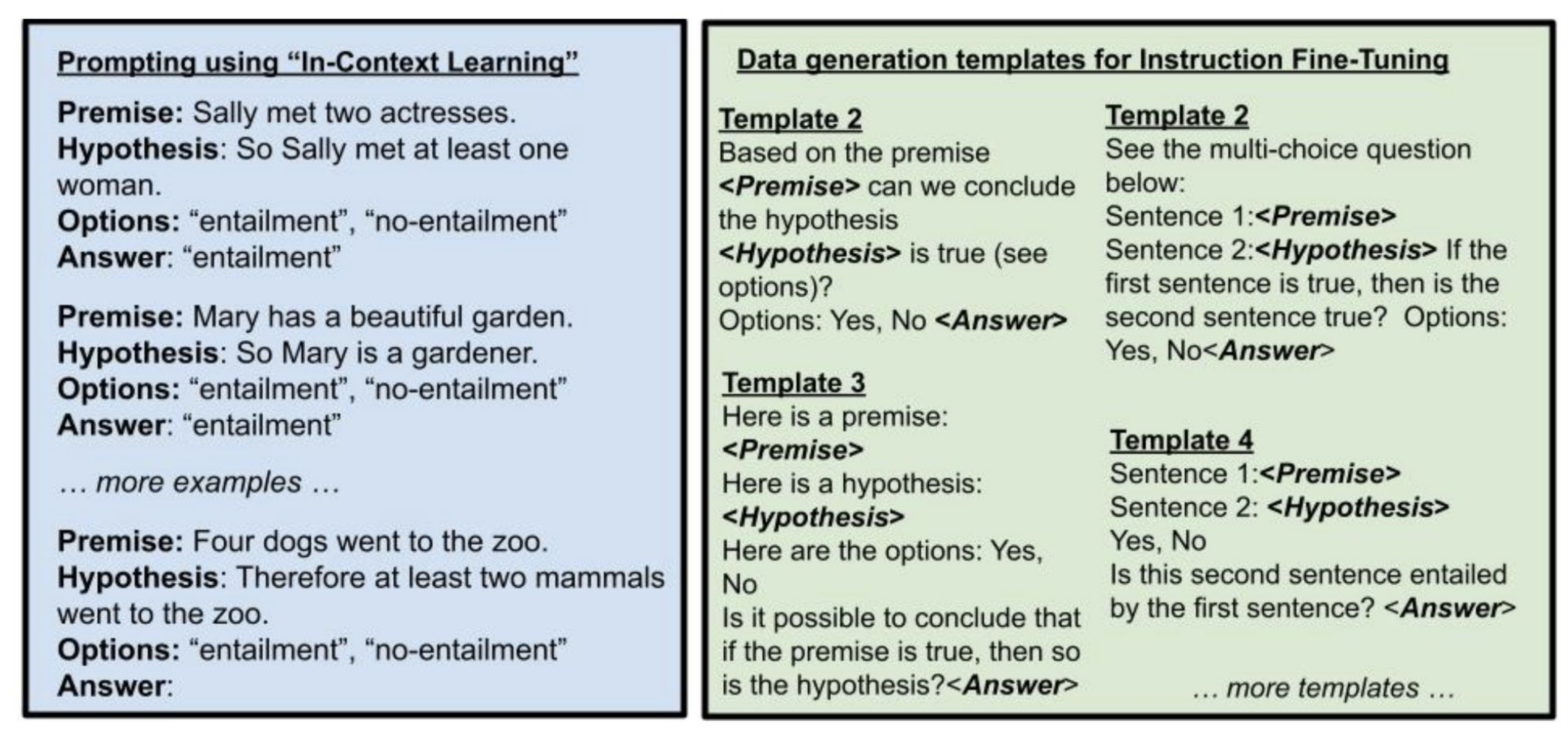
Hypothesis: <hypothesis>

<options>

Template 4, ...

(Wei et al., 2022)

Instruction Tuning



Prompting using "In-Context Learning"

Premise: Sally met two actresses.
Hypothesis: So Sally met at least one woman.
Options: "entailment", "no-entailment"
Answer: "entailment"

Premise: Mary has a beautiful garden.
Hypothesis: So Mary is a gardener.
Options: "entailment", "no-entailment"
Answer: "entailment"

… more examples …

Premise: Four dogs went to the zoo.
Hypothesis: Therefore at least two mammals went to the zoo.
Options: "entailment", "no-entailment"
Answer:

Data generation templates for Instruction Fine-Tuning

Template 2
Based on the premise ***<Premise>*** can we conclude the hypothesis ***<Hypothesis>*** is true (see options)?
Options: Yes, No ***<Answer>***

Template 3
Here is a premise:
<Premise>
Here is a hypothesis:
<Hypothesis>
Here are the options: Yes, No
Is it possible to conclude that if the premise is true, then so is the hypothesis?<***Answer***>

Template 2
See the multi-choice question below:
Sentence 1:***<Premise>***
Sentence 2:***<Hypothesis>*** If the first sentence is true, then is the second sentence true? Options: Yes, No<***Answer***>

Template 4
Sentence 1:***<Premise>***
Sentence 2: ***<Hypothesis>***
Yes, No
Is this second sentence entailed by the first sentence? <***Answer***>

… more templates …

(Lu and Bigoulaeva, 2024)

Experiments on Emergent Abilities

We run over 1000 experiments on 20 models of parameter sizes ranging from 60M to 175B on 22 Tasks

We show that Emergent Abilities are a manifestation of in-context learning

See paper for details:
https://h-tayyarmadabushi.github.io/Emergent_Abilities_and_in-Context_Learning/

(Lu and Bigoulaeva, 2024)

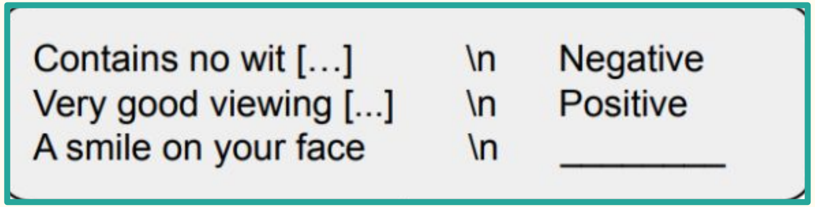
We introduce Implicit-ICL and show that it Leads to what Models are capable of:

Implicit ICL

Is the following movie review positive: “brings a smile to your face …”

Mapping through IT

Explicit ICL



Contains no wit […]	\n	Negative
Very good viewing […]	\n	Positive
A smile on your face	\n	________

(Lu and Bigoulaeva, 2024)

Other Phenomena Explainable by Implicit ICL

- *Hallucinations:* they can be explained using Implicit ICL as the model defaulting to the most statistically likely output sequence **when prompt does not easily allow for in-context learning**

- *The need for prompt engineering:* The models can only "solve" a task when the mapping from instructions to exemplars is optimal.

(Lu and Bigoulaeva, 2024)

Other Phenomena Explainable by Implicit ICL

- E.g., Testing for theory of mind explicitly seems to show that have access to this information

- Why this is not possible, it is because:
  - In-context capabilities increase steadily with scale.
  - The "complexity" of the problem may require models which have "stronger" ICL abilities.

(Lu and Bigoulaeva, 2024)

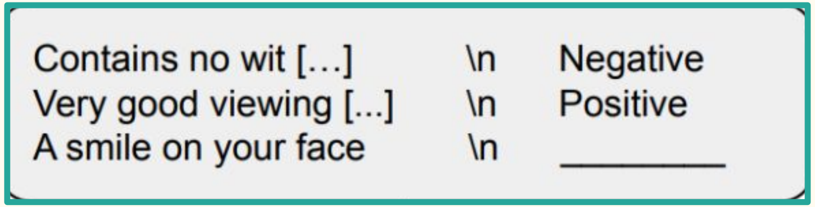
Takeaway 1: LLMs use Implicit ICL

Implicit ICL

Explicit ICL

Is the following movie review positive: “brings a smile to your face …”

Mapping through IT



```
Contains no wit […]          \n     Negative
Very good viewing […]        \n     Positive
A smile on your face         \n     ________
```

(Lu and Bigoulaeva, 2024)

Generative Grammar and Idioms

Theory of Generative Grammar

lexicon

phonological component

…

syntactic component

semantic component

Layers of linguistic knowledge (e.g., syntax and semantics)

Across these components is the lexicon. An *instantiation* of these dimensions of information.

(Chomsky, 1957)

Generative Grammar and Idioms

At the heart of generative grammar is the principle of generality of rules governing grammar.

Problem: Idioms, have the mean something other than what might be inferred by the general rules of grammar

(Croft and Curse, 2004, 9.231, Nunberg et al. 1994:507)

Idioms to Construction Grammar

Instead of seeing idioms as the problem, Fillmore, Kay and O’Connor (1988) treated idioms as the basis of a new model for gramatical organisation: the construction.

(Croft and Curse, 2004, 9.247)

LLMs and Idiomatic Expressions

Idioms: GPT-3.5 and GPT-4

- Idiomatic expressions, and more generally MWEs, were what inspired construction grammar.

- Independently, there’s significant research focused on MWEs and Idioms (including this workshop)

Idioms: A pain in the neck for LLMs

Early work on capturing idioms:

- Seminal paper by Sag et al. 2002
- Schneider et al. 2014 - Adjacent or nonadjacent sequences of tokens for MWE identification
- Green et al. 2013 - Special constituency nodes for MWE identification
- Vincze et al. 2013, Candito & Constant 2014, de Marneffe et al. 2021 (UD) - Special dependency relations

Work on PLMs' ability to capture Idioms overall seems to suggest that they still struggle:

- PARSEME
- DiMSUM, VNC-Tokens, MAGPIE
- SemEval 2022 - Task 2: AStirchInLanguageModels corpus

Partial List of work on MWEs

Idioms: GPT-3.5 and GPT-4 [New Results]

Evaluate GPT-3.5 and GPT-4 on the AStitchInLanguageModels Dataset (Tayyar Madabushi et al., 2022 a) :

- a dataset consisting of *naturally occurring sentences* containing potentially idiomatic **noun phrases**
- This dataset was used for SemEval 2022 Task 2: Multilingual Idiomaticity Detection and Sentence Embedding (Tayyar Madabushi et al., 2022 b)

(Laureano de Leon et al., 2024)

Idioms: GPT-3.5 and GPT-4 [New Results]

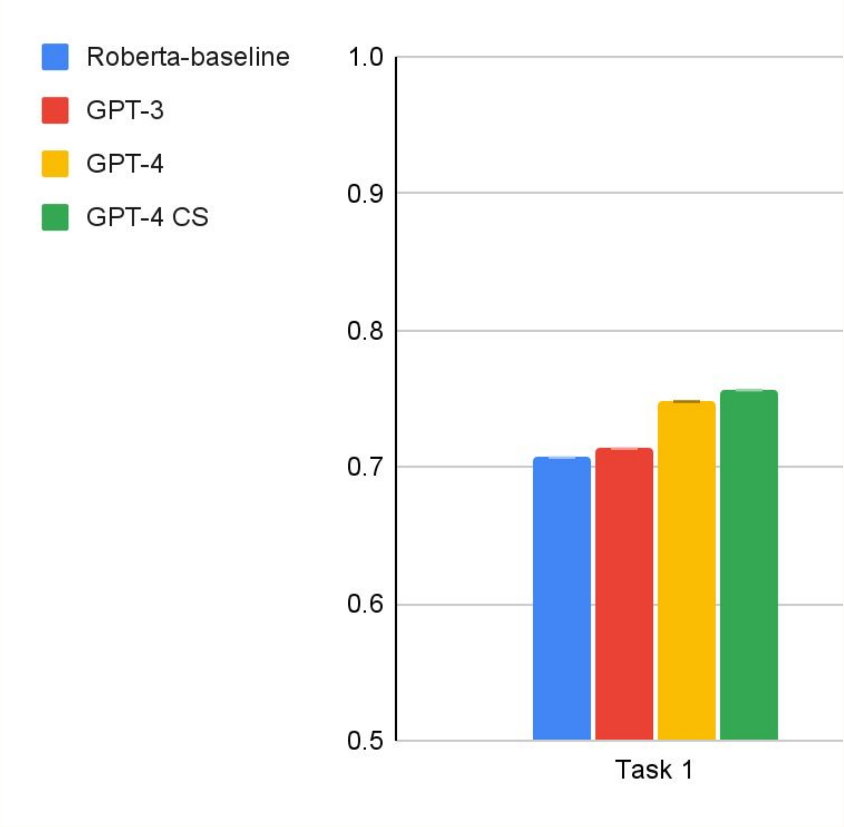
AStitchInLanguageModels Dataset (Tayyar Madabushi et al., 2022 a)

Examples:

- This means that search data is a **gold mine** for marketing strategy.
- The hashtag “Qixia **gold mine** incident” has been viewed many million of times on the social media site Weibo.

(Laureano de Leon et al., 2024)

Task 1 - Meta Linguistic Knowledge: Is the MWE in the literal or not



(Laureano de Leon et al., 2024)

Idioms: GPT-3.5 and GPT-4 [New Results]

Task 2 - Usage

Binary Classification: Do the following sentences mean the same?

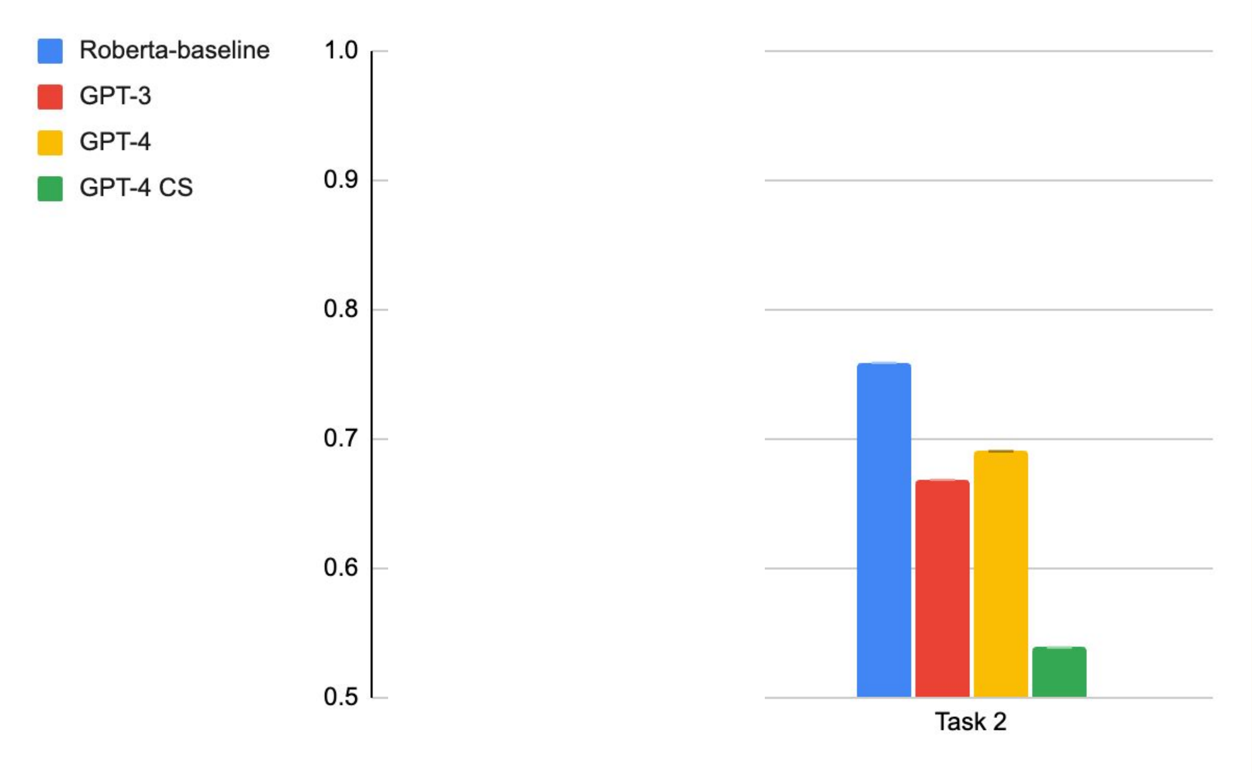
Example: When removing a **big fish** from a net, it should be held in a manner that supports the girth.

Not Same Meaning	**Same Meaning**
When removing an **important person** from a net, it should be held in a manner that supports the girth.	When removing a **fish** from a net, it should be held in a manner that supports the girth

(Laureano de Leon et al., 2024)

Task 2 - Usage

Binary Classification: Do the following sentences mean the same?



(Laureano de Leon et al., 2024)

Oddity: Pain in the Neck or Walk in the Park?

Why is it that LLMs

- perform well on the familiar meta-linguistic task but
- perform less well on the NOT familiar task of *applying* the same information?

Answer: **Implicit In-Context Learning**

- How similar the task is to instruction tuning matters!

(Laureano de Leon et al., 2024)

Takeaway 2

Because LLMs use implicit ICL

- *Instruction Tuning datasets Matter!!!*
- Prompt based “probing” is always just one task at a time
  - Success on on “probing task” is NOT evidence of that information being available to the model!
  - We can think of this as “ICL-Training” the model.

Construction Grammars and NLP

Recall: Idioms to Construction Grammar

Instead of seeing idioms as the problem, Fillmore, Kay and O'Connor (1988) treated idioms as the basis of a new model for gramatical organisation: the construction.

Idioms to Construction Grammar

“A ***construction*** is a syntactic configuration

- sometimes with one or more substantive items (e.g. *let alone*) and
- sometimes not (e.g., resultative construction).”

(Quoted from Croft and Curse, 2004, 9.247)

CxG: An Usage Based approach to Grammar

- Constructions provide a way of representing speaker knowledge that can now explain idioms.

- Construction grammar is *usage based*
    - Familiarity plays an important role in constructions
    - Repeated use leads to a higher level abstraction of the representations.

(Croft and Curse, 2004, 9.231)

CxG and Language Models

The Usage Based nature of Constructions implies that they can be captured by language models

> This was first tested by Tayyar Madabushi et al. (2020), who found that PLMs have access to a significant amount of constructionally relevant information

The fact that PLMs have access to CxG information verifies usage based theories of language acquisition

(Tayyar Madabushi et al., 2020)

CxG and Language Models: Psycholinguistic studies

Li et al. (2022) explore the verb-centred approach vs the construction-centred approach

> Using a sentence sorting task, they find that sentences that instantiate the *same* argument structure construction are more closely embedded than sentences that only have the verb in common

CxG and Language Models: The shortcomings

Weissweiler et al., 2022 study the Comparative Correlative

- The larger the model, the more the reasoning
- The better your syntax, the better your semantics

They find that:

- While LLMs can interpret typical sentences that are instances of this construction,
- PLMs cannot generalise this knowledge to novel utterances.

CxG + NLP

CxG and NLP has since become an active area of research:

- The first workshop on Construction Grammars and Natural Language Processing
- The CxG + NLP live bibliography
- More work here at LREC-COLING

CxG + NLP: But do PLMs capture CxGs?

The Schematicity Hypothesis (**New @LREC-COLING**)

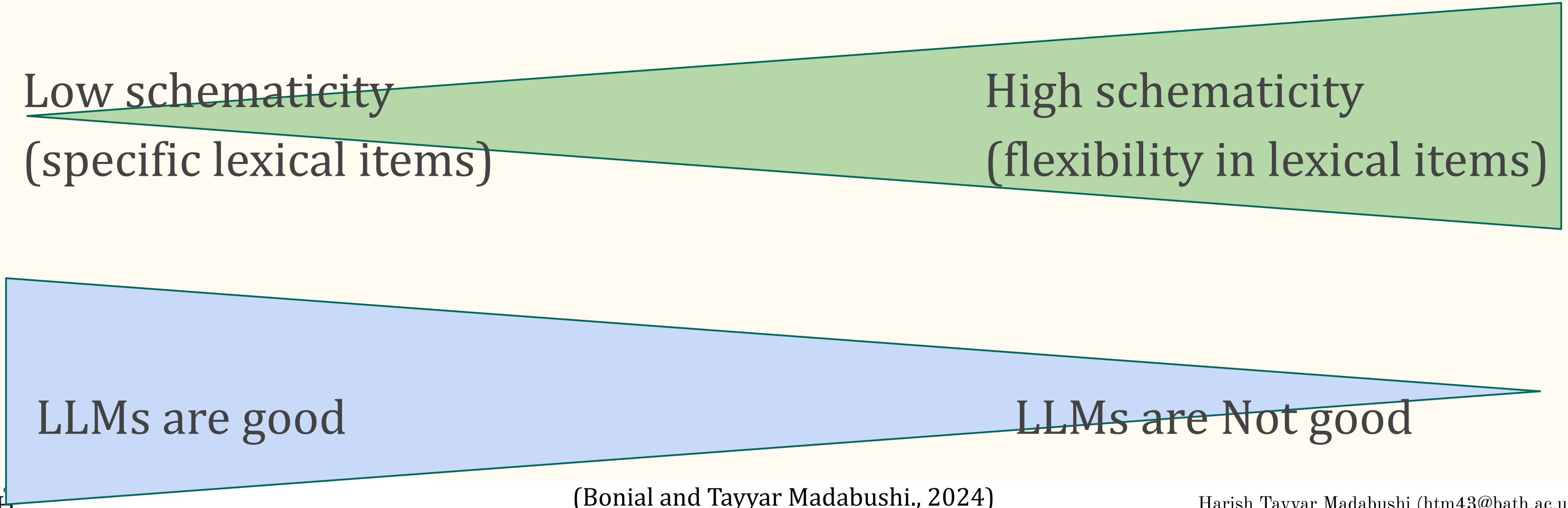
Constructions occur at varying levels of schematicity: Schematicity is a CxN's ability to accommodate varying degrees of specificity:

- Some CxNs demand specific lexical items (low schematicity)
  - E.g., the idiomatic construction "**let alone**,"
- Other CxNs allow a broad array of semantically fitting elements (high schematicity).
  - E.g., "**Resultative**"

(Bonial and Tayyar Madabushi., 2024)

The Schematicity Hypothesis (**New @LREC-COLING**)

The Constructional information available to LLMs deteriorates with increase in schematicity.



Low schematicity
(specific lexical items)

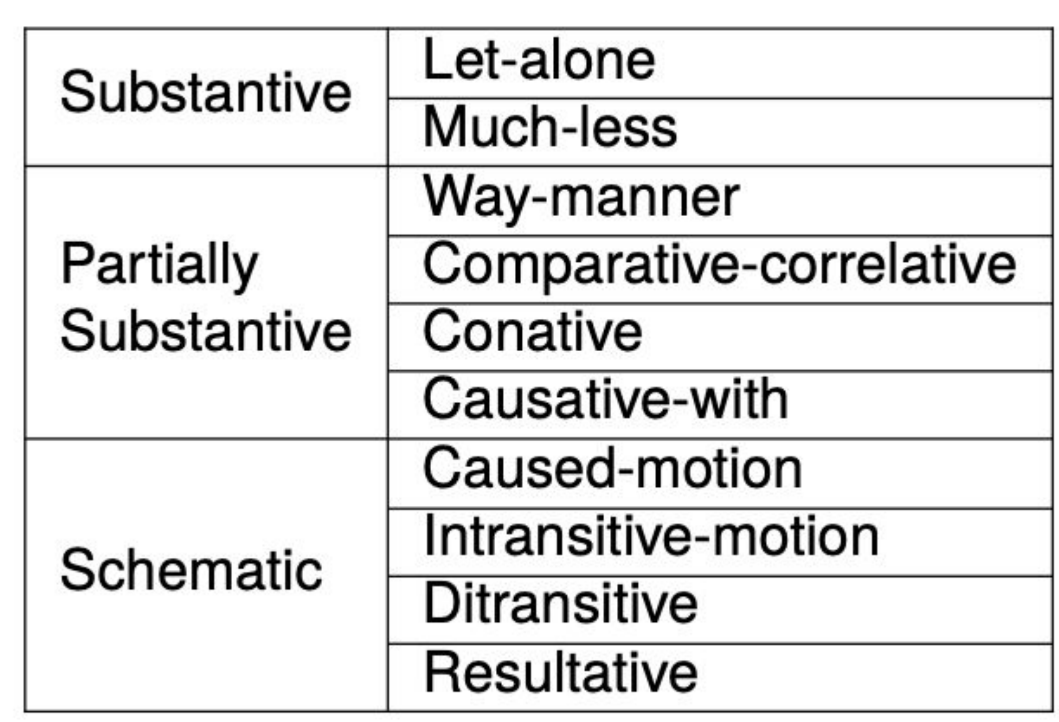
High schematicity
(flexibility in lexical items)

LLMs are good

LLMs are Not good

(Bonial and Tayyar Madabushi., 2024)

Testing the Schematicity Hypothesis (**New @LREC-COLING**)



Substantive	Let-alone
	Much-less
Partially Substantive	Way-manner
	Comparative-correlative
	Conative
	Causative-with
Schematic	Caused-motion
	Intransitive-motion
	Ditransitive
	Resultative

(Bonial and Tayyar Madabushi., 2024)

Testing the Schematicity Hypothesis (**New @LREC-COLING**)

Substantive	Let-Alone (Let-alone frozen)	None of these arguments is notably strong, **let alone** conclusive
Partial	Comparative-Correlative the + comparative + the + comparative	The more I studied the less I understood
Fully Schematic	Ditransitive Agent is construed as causing a recipient to receive a theme.	She baked her sister a cake

(Bonial and Tayyar Madabushi., 2024)

Testing the Schematicity Hypothesis (**New @LREC-COLING**)

From amongst the following sentences, extract the three sentences which are instances of the **ConstructionName** construction, as exemplified by the following sentence: **Sentence**. Output only the three sentences in three separate lines:

- 6 sentences
  - 3 positive
  - 3 distractors

(Bonial and Tayyar Madabushi., 2024)

The Schematicity Hypothesis Results (**New @LREC-COLING**)

Abstraction Level	**GPT-3.5**	**GPT-4**
Purely Substantive	84.00	98.34
Partially Schematic	75.17	92.67
Fully Schematic	54.00	62.33
Baseline	*50.00*	

These results confirm the schematicity hypothesis

(Bonial and Tayyar Madabushi., 2024)

What of LLMs’ ability to use this information?

We create a Constructional NLI Dataset, based on schematicity

CxN, Type	**P/H/R**	**Annotation Targets/ Gold Relation**
Let-alone Substantive	Premise	A ceasefire, let alone lasting peace, will take long negotiation.
	Hypothesis	There will be peaceful negotiation of a ceasefire.
	Relation	1 (neutral)

(Bonial et al.., 2024, in preparation)

What of LLMs' ability to use this information?

We create a Constructional NLI Dataset, based on schematicity

CxN, Type	**P/H/R**	**Annotation Targets/ Gold Relation**
Resultative *Fully Schematic*	Premise	The jackhammer pounded us deaf.
	Hypothesis	The jackhammer was easy on our ears.
	Relation	2 (contradiction)

(Bonial et al.., 2024, in preparation)

What of LLMs' ability to use this information?

- We evaluate both GPT-3 and GPT-4 using multiple prompts
- The variation in GPT-4 results across prompts is higher than across schematicity, and so the results are not conclusive.
- We are unable to collect enough Substantive examples and so only test partially schematic and fully schematic constructions (200 each)

(Bonial et al.., 2024, in preparation)

What of LLMs’ ability to use this information?

You are the world's best annotator. Your task [...] Natural Language Inference (NLI) task. [...]

We use numerical coding, **also listed in your annotation spreadsheet** as a reminder:

0 – entailment – The hypothesis must be true given the premise

1 – neutral – The hypothesis may or may not be true given the premise

2 – contradiction – The hypothesis must not be true given the premise

Examples

(Bonial et al.., 2024, in preparation)

What of LLMs’ ability to use this information?

GPT-3.5 results on *best* prompt (trend remains across prompts)

Schematicity	GPT-3.5 performance (F1)
Partial schematic	0.72
Fully Schematic	0.66

(Bonial et al.., 2024, in preparation)

What of LLMs’ ability to use this information?

- But … recall that we are using examples in the prompt
- These results are when the examples presented are **Constructional NLI Triplets**
- What if we presented traditional NLI Triplets

(Bonial et al.., 2024, in preparation)

What of LLMs' ability to use this information?

GPT-3.5 results on *best* prompt (trend remains across prompts)

Schematicity	GPT-3.5 (F1) CxG Examples	GPT-3.5 (F1) NLI Examples
Partial schematic	0.72	0.66
Fully Schematic	0.66	0.71

(Bonial et al.., 2024, in preparation)

Oddity: Trends change depending on Examples!

Why is it that LLMs

- perform well on CxG data when examples are CxG
- otherwise performs well on data that is more similar to NLI examples when presented with those in the prompt

Answer: **Implicit In-Context Learning**

- If we think of in-context examples as “training” we can see how in-distribution performance is high, but out-distribution is poor

(Laureano de Leon et al., 2024)

Takeaway 3

Because LLMs use implicit ICL

- *The examples we provide matter.*
- There is significant difference from “training”, but
  - using examples that are semantically similar helps,
  - using examples that are “harder” helps,
  - using more examples helps …

Wrapping Up

- Large Language Models use Implicit In-Context Learning to respond to prompts
- Thinking of this process as a form of “fine-tuning” allows us to understand what they can and can’t do

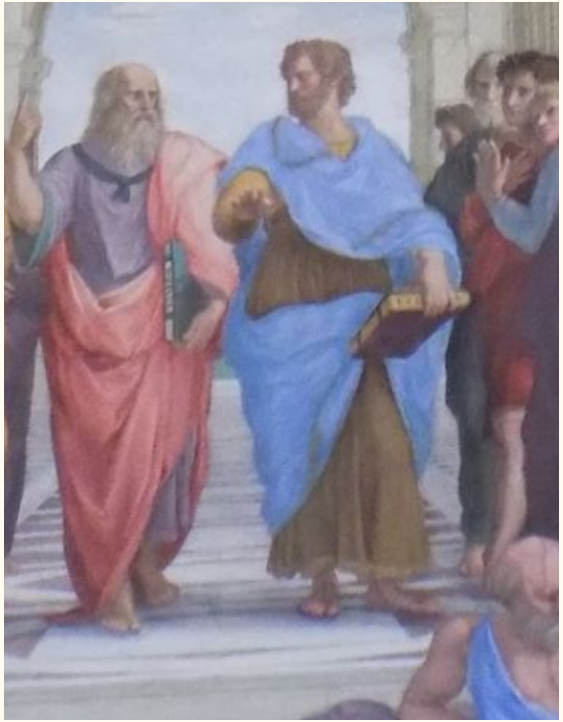
- LLMs are extremely powerful in allowing us to extract all the information that we haven’t been able to for so long (e.g., how to fill slots)
- If we think of them as systems that **we can quickly test using prompts** but that subsequently **require fine-tuning to improve performance** (i.e., hallucination prevention) we have a powerful ally

Where from Meaning?

Usage Based Theories and LLMs:
For the first time we can use data-driven methods to

- answer theoretical questions,
- to build resources that adheres to theoretical constructions, and

to build more powerful systems rooted in theoretical constructs build using data-driven methods.



Thank you!