What do Large Language Models Know about Events and their Relations?

Alessandro Lenci

COmputational LINGuistics Laboratory
Università di Pisa
Dipartimento di Filologia, Letteratura e Linguistica (FiLeLi)

LLMs: Super-humans or Illusionists?

OpenAI ChatGPT

The Knowledge of LLMs?

- LLMs give us the impression (illusion?) of having a super-human amount of "knowledge" they use to "understand" language and carry out different types of human-like reasoning

Some key questions

- How do LLMs acquire their "knowledge"?
- Is the "knowledge" of LLMs like human knowledge?

The Knowledge of LLMs?

- LLMs give us the impression (illusion?) of having a super-human amount of "knowledge" they use to "understand" language and carry out different types of human-like reasoning

Some key questions

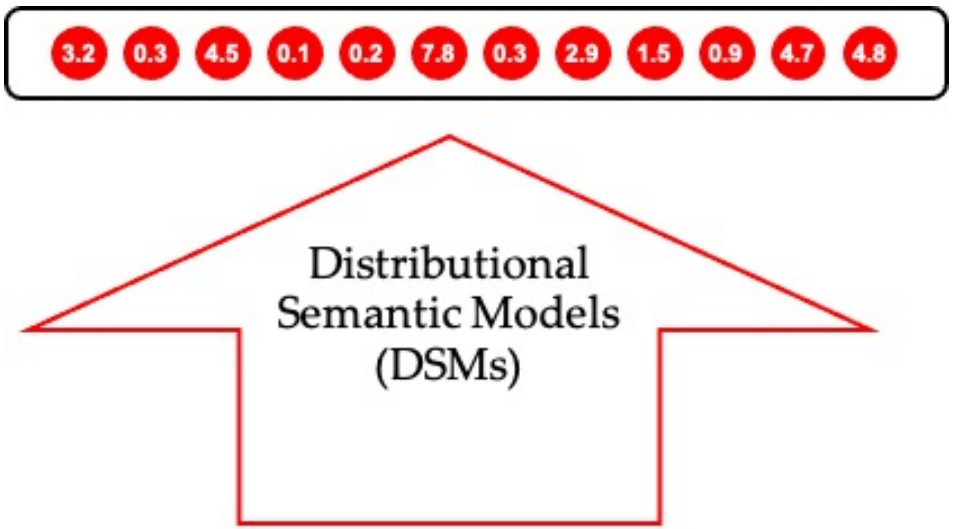
- How do LLMs acquire their "knowledge"?
- Is the "knowledge" of LLMs like human knowledge?

LLMs as Distributional Semantic Models (DSMs)

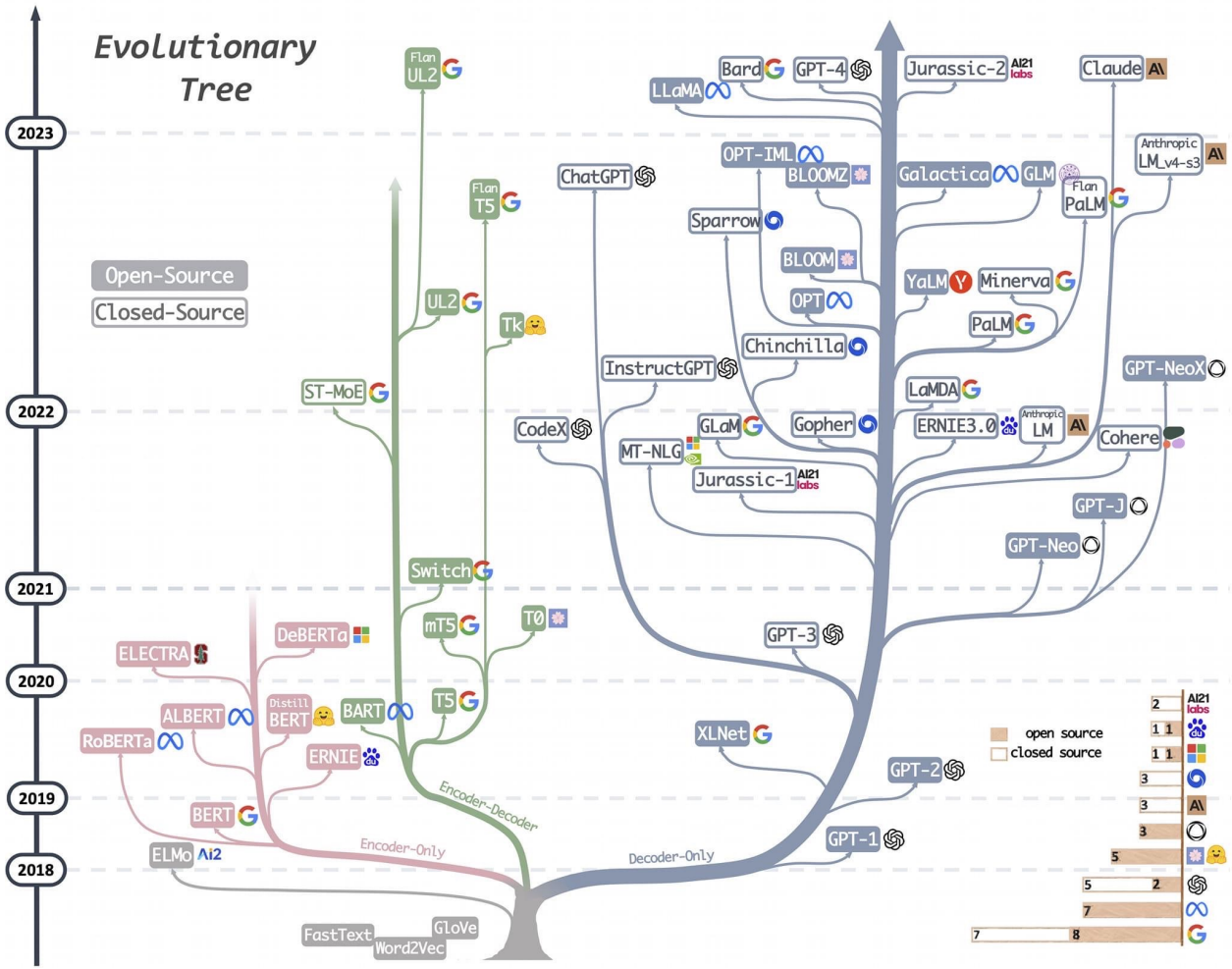
Lenci and Sahlgren (2023), *Distributional Semantics*, Cambridge University Press

Distributional Semantics

The meaning of linguistic expressions is represented with vectors (embeddings) encoding their statistical distribution in linguistic contexts extracted from corpora



The Tree of LLMs and their Roots



The "Knowledge" of LLMs

- LLMs learn from texts a far greater amount of information than any previous DSM
  - syntactic structure, several dimensions of lexical and sentence meaning (Tenney et al. 2019, Manning et al. 2020), pragmatic aspects (Hu et al. 2023), and so on
- LLMs reveal emergent abilities to carry out linguistic tasks (e.g., translating, question-answering, making inferences, etc.) without any task-specific training (Brown et al. 2020, Wei et al. 2022)

The "Knowledge" of LLMs

- LLMs learn from texts a far greater amount of information than any previous DSM
  - syntactic structure, several dimensions of lexical and sentence meaning (Tenney et al. 2019, Manning et al. 2020), pragmatic aspects (Hu et al. 2023), and so on

- LLMs reveal emergent abilities to carry out linguistic tasks (e.g., translating, question-answering, making inferences, etc.) without any task-specific training (Brown et al. 2020, Wei et al. 2022)

Do LLMs "Understand" What We/They say?

"We have decoupled the ability to act successfully from the need to be intelligent, understand, reflect, consider or grasp anything. We have liberated agency from intelligence. [...] AI understood as *Agere sine Intelligere*"

Floridi (2023). "AI as Agency Without Intelligence: On ChatGPT, Large Language Models, and Other Generative Models". *Philosophy and Technology*, 36(15)

Language vs. Thought

Mahowald et al. (2024). Dissociating language and thought in large language models: A cognitive perspective. *Trends in Cognitive Sciences*

"good at language → good at thought" fallacy

If an entity (be it human or a machine) generates long coherent stretches of text, it must possess rich knowledge and reasoning capacities

- Mahowald et al. (2024) distinguish between:
  - formal linguistic competence, that is knowledge of linguistic rules and patterns
  - functional competence, that is the ability of understanding and using language in the world

Language vs. Thought

Mahowald et al. (2024). Dissociating language and thought in large language models: A cognitive perspective. *Trends in Cognitive Sciences*

"good at language → good at thought" fallacy

If an entity (be it human or a machine) generates long coherent stretches of text, it must possess rich knowledge and reasoning capacities

- Mahowald et al. (2024) distinguish between:
  - formal linguistic competence, that is knowledge of linguistic rules and patterns
  - functional competence, that is the ability of understanding and using language in the world

Language vs. Thought

Mahowald et al. (2023). Dissociating language and thought in large language models: A cognitive perspective. *ArXiv*

- Functional competence requires inferential competence:
  - formal reasoning (logical reasoning and novel problem solving)
  - world knowledge (knowledge of objects and events and their properties, participants and relations)
  - situation modeling the ability of building a representation of the stories we extract from language input and track their dynamic evolution over time)
  - social reasoning (as the ability of using language by taking into account the states of mind of our interlocutors and our shared knowledge)

Key points

- LLMs have an almost human-like formal competence (i.e., to generate texts), but still fall short of inferential and reasoning competence
- We need in-depth analyses of the kind of knowledge that LLMs truly possess

Language vs. Thought

Mahowald et al. (2023). Dissociating language and thought in large language models: A cognitive perspective. *ArXiv*

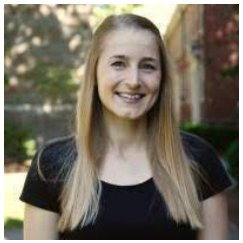
- Functional competence requires inferential competence:
  - formal reasoning (logical reasoning and novel problem solving)
  - world knowledge (knowledge of objects and events and their properties, participants and relations)
  - situation modeling the ability of building a representation of the stories we extract from language input and track their dynamic evolution over time)
  - social reasoning (as the ability of using language by taking into account the states of mind of our interlocutors and our shared knowledge)

Key points

- LLMs have an almost human-like formal competence (i.e., to generate texts), but still fall short of inferential and reasoning competence
- We need in-depth analyses of the kind of knowledge that LLMs truly possess

Kauf et al. (2023), "Event knowledge in large language models: The gap between the impossible and the unlikely", *Cognitive Science*: 47

Carina Kauf
(MIT)



Anna A. Ivanova
(MIT)

Eveline Fedorenko
(MIT)

Emmanuele Chersoni
(Hong Kong PolyU)

Giulia Rambelli
(Univ. Pisa & AMU)

Event Knowledge between Semantics and Pragmatics

- Knowledge of the prototypical, abstract structure of everyday events and their participants (e.g., McRae and Matsuki, 2011), also known as Generalized Event Knowledge (GEK)

The cop arrested the thief

The cop arrested the number (semantically impossible)

The thief arrested the cop (pragmatically implausible)

Event Knowledge between Semantics and Pragmatics

- Knowledge of the prototypical, abstract structure of everyday events and their participants (e.g., McRae and Matsuki, 2011), also known as Generalized Event Knowledge (GEK)

  The cop arrested the thief

  The cop arrested the number (semantically impossible)

  The thief arrested the cop (pragmatically implausible)

Event Knowledge between Semantics and Pragmatics

- Knowledge of the prototypical, abstract structure of everyday events and their participants (e.g., McRae and Matsuki, 2011), also known as Generalized Event Knowledge (GEK)

The cop arrested the thief

The cop arrested the number (semantically impossible)

The thief arrested the cop (pragmatically implausible)

Event Knowledge between Semantics and Pragmatics

- Knowledge of the prototypical, abstract structure of everyday events and their participants (e.g., McRae and Matsuki, 2011), also known as Generalized Event Knowledge (GEK)

  The cop arrested the thief

  The cop arrested the number (semantically impossible)

  The thief arrested the cop (pragmatically implausible)

Generalized Event Knowledge (GEK)

McRae & Matsuki 2009 "People Use their Knowledge of Common Events to Understand Language, and Do So as Quickly as Possible"

- Two sources of event knowledge
  - first-hand experience in performing or watching events

  - linguistic experience derived from the distributional analysis of the linguistic input
    - *The policeman is chasing the thief*

Dataset

- 391 pairs of plausible/implausible/impossible sentences

Item Type	Plausible?	Possible?	Sentence
animate-inanimate (AI)	yes	yes	The teacher bought the laptop.
	no	no	The laptop bought the teacher.
animate-animate (AA)	yes	yes	The nanny tutored the boy.
	no	yes	The boy tutored the nanny.

- Human subjects evaluated the extent to which each sentence was "plausible, i.e., likely to occur in the real world" on a Likert scale from 1 (completely implausible) to 7 (completely plausible)

Sentence Manipulations

Sentence Set	Plausible?	Voice	Synonym #	Sentence
Dataset 1 *(Fedorenko et al., 2020)*	yes	active	1	The teacher bought the laptop.
			2	The instructor purchased the computer.
		passive	1	The laptop was bought by the teacher.
			2	The computer was purchased by the instructor.
	no	active	1	The laptop bought the teacher.
			2	The computer purchased the instructor.
		passive	1	The teacher was bought by the laptop.
			2	The instructor was purchased by the computer.

LLMs Tested with Event Knowledge

***Table S1**. Overview of LLM designs. BPE = BytePair Encoding, CLM= Causal Language Modeling, MLM = Masked Language Modeling. NSP = Next-Sentence Prediction*

Model	Attention	Tokenization	#parameters	Vocabulary size	Training data size	Training task
mpt-30b	Unidirectional	BPE	30B	50K	1TB	CLM
gpt-J-6b	Unidirectional	BPE	6B	50K	800GB	CLM
gpt2-xl	Unidirectional	BPE	1.5B	50K	40GB	CLM
roberta-large	Bidirectional	WordPiece	355M	30K	160GB	Dynamic MLM
bert-large-cased	Bidirectional	WordPiece	340M	30K	13GB	MLM + NSP

- Sentence score
  - unidirectional LLMs - sum of the log-probabilities of each token w_i in the sequence, conditioned on the preceding sentence tokens
  - bidirectional LLMs - a variant of the sentence's pseudo-log-likelihood score (PLL) (Salazar et al. 2020, Kauf and Ivanova 2023)

LLMs Tested with Event Knowledge

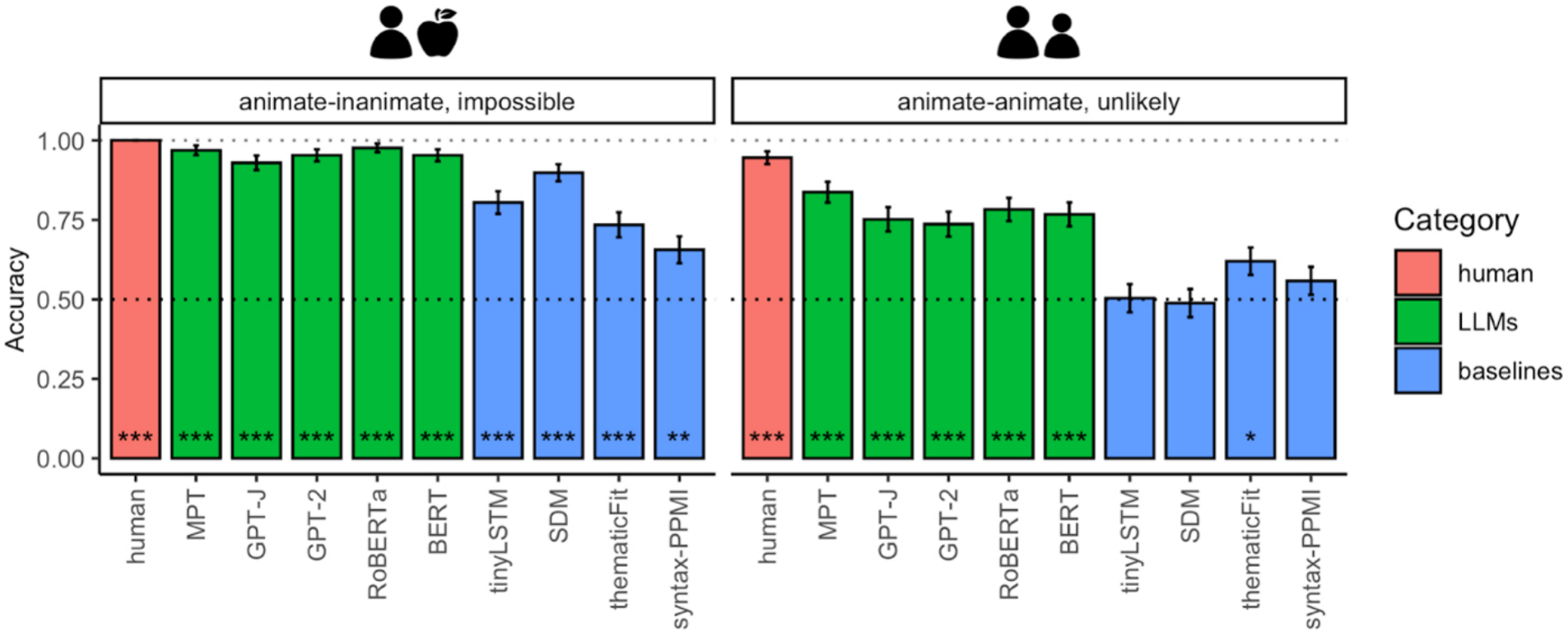
***Table S1**. Overview of LLM designs. BPE = BytePair Encoding, CLM= Causal Language Modeling, MLM = Masked Language Modeling. NSP = Next-Sentence Prediction*

Model	Attention	Tokenization	#parameters	Vocabulary size	Training data size	Training task
mpt-30b	Unidirectional	BPE	30B	50K	1TB	CLM
gpt-J-6b	Unidirectional	BPE	6B	50K	800GB	CLM
gpt2-xl	Unidirectional	BPE	1.5B	50K	40GB	CLM
roberta-large	Bidirectional	WordPiece	355M	30K	160GB	Dynamic MLM
bert-large-cased	Bidirectional	WordPiece	340M	30K	13GB	MLM + NSP

- Sentence score
  - unidirectional LLMs - sum of the log-probabilities of each token w_i in the sequence, conditioned on the preceding sentence tokens
  - bidirectional LLMs - a variant of the sentence's pseudo-log-likelihood score (PLL) (Salazar et al. 2020, Kauf and Ivanova 2023)

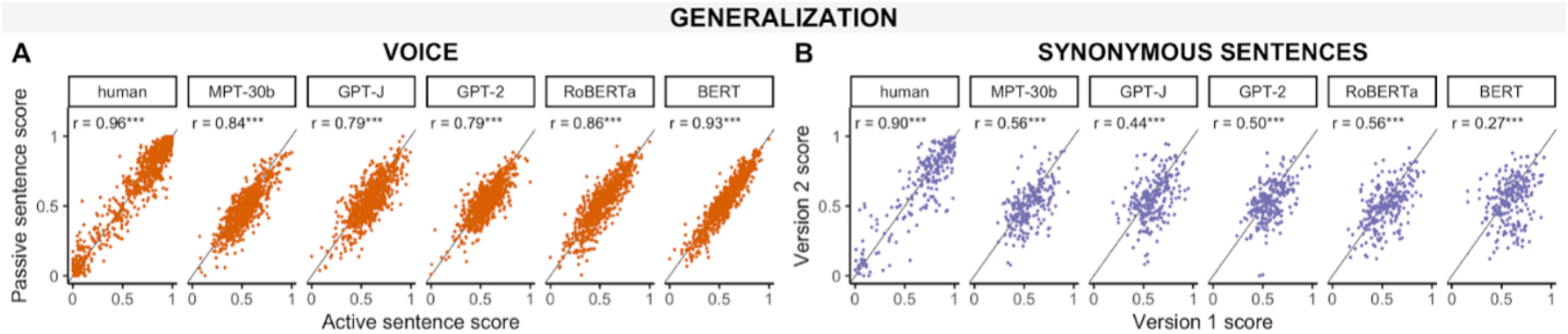
Results

- We evaluate the ability of humans and LLMs to assign a higher score to the plausible event description than the implausible (impossible) one



Results

sentence manipulations



The father pampered the infant.

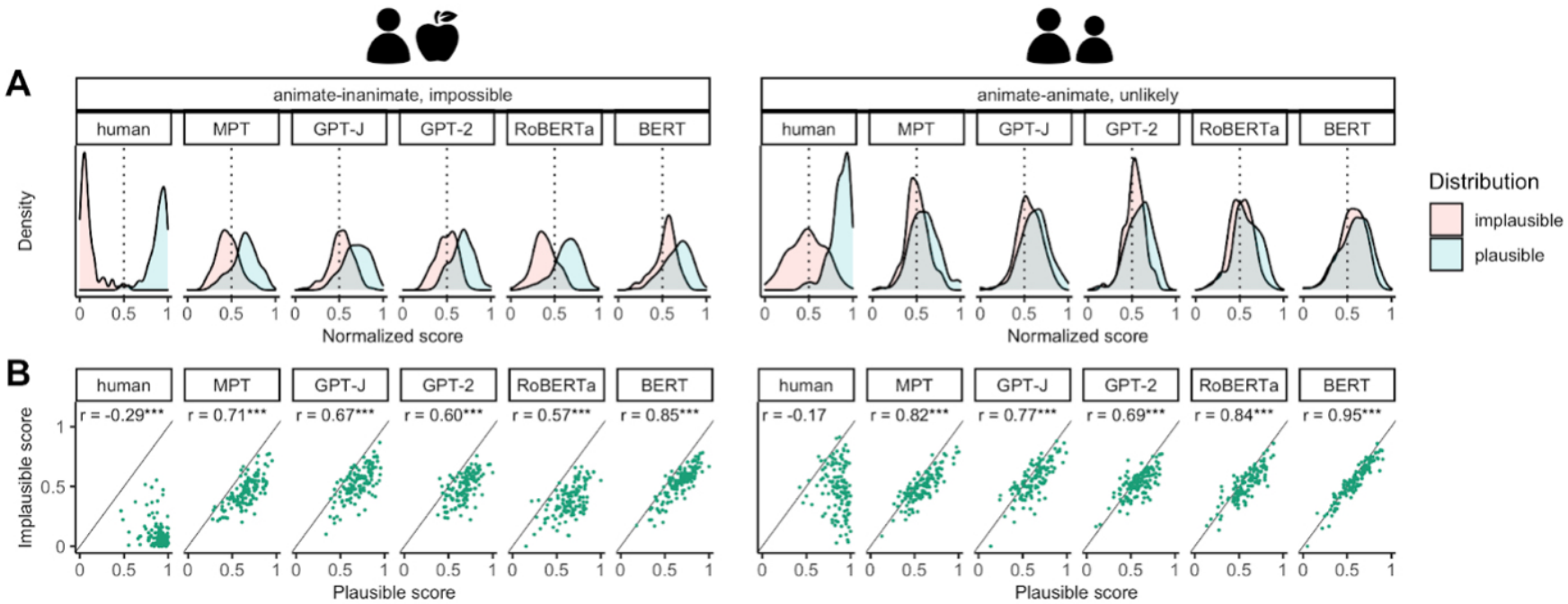
The infant was pampered by the father.

The father pampered the infant.

The dad coddled the baby.

Results

is it really human-like event knowledge?



AI Sentences

The writer composed the book.

The book composed the writer.

AA Sentences

The father pampered the infant.

The infant pampered the father.

Interim Conclusions

- The input of current textual LLMs is qualitatively poorer and quantitatively far richer than the ones received by human learners
- Large gaps still exists between between LLMs and humans in their core knowledge
- We can not look only at their superficial behavior!
  - we need carefully designed, linguistic and theory-driven analyses of LLMs

Interim Conclusions

- The input of current textual LLMs is qualitatively poorer and quantitatively far richer than the ones received by human learners
- Large gaps still exists between between LLMs and humans in their core knowledge
- We can not look only at their superficial behavior!
  - we need carefully designed, linguistic and theory-driven analyses of LLMs

Interim Conclusions

- The input of current textual LLMs is qualitatively poorer and quantitatively far richer than the ones received by human learners
- Large gaps still exists between between LLMs and humans in their core knowledge
- We can not look only at their superficial behavior!
  - we need carefully designed, linguistic and theory-driven analyses of LLMs

Interim Conclusions

- The input of current textual LLMs is qualitatively poorer and quantitatively far richer than the ones received by human learners
- Large gaps still exists between between LLMs and humans in their core knowledge
- We can not look only at their superficial behavior!
  - we need carefully designed, linguistic and theory-driven analyses of LLMs

Plato and LLMs

Psychological Review
1997, Vol. 104, No. 2, 211–240

Copyright 1997 by the American Psychological Association, Inc.
0033-295X/97/$3.00

A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge

Thomas K Landauer
University of Colorado at Boulder

Susan T. Dumais
Bellcore

Plato's Problem or the Human Learning Problem

How do people know as much as they do with as little information as they get?'

Plato and LLMs

Psychological Review
1997, Vol. 104, No. 2, 211–240

Copyright 1997 by the American Psychological Association, Inc.
0033-295X/97/$3.00

A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge

Thomas K Landauer
University of Colorado at Boulder

Susan T. Dumais
Bellcore

The LLM Learning Problem

LLMs still lack aspects of human semantic and pragmatic competence despite being trained on very large amounts of textual data

Plato and LLMs

Psychological Review
1997, Vol. 104, No. 2, 211–240

Copyright 1997 by the American Psychological Association, Inc.
0033-295X/97/$3.00

A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge

Thomas K Landauer
University of Colorado at Boulder

Susan T. Dumais
Bellcore

The LLM Learning Problem

LLMs still lack aspects of human semantic and pragmatic competence despite being trained on very large amounts of textual data

The LLM Learning Problem

Hypothesis

Some dimensions of meaning might be occur weakly in the linguistic signal

- The Reporting Bias (Gordon and Van Durme 2013)
  - people tend to omit information that is obvious (cf. Grice's Maxim of Quantity)
- Paik et al. (2021) show that color information about concepts associated with a single color (e.g., strawberry) is worst represented in corpora

- Possible solution
  - more attention to the quality of training data
  - extend textual corpora with extralinguistic information (cf. multimodal models)

The LLM Learning Problem

Hypothesis

Some dimensions of meaning might be occur weakly in the linguistic signal

- The Reporting Bias (Gordon and Van Durme 2013)
  - people tend to omit information that is obvious (cf. Grice's Maxim of Quantity)
- Paik et al. (2021) show that color information about concepts associated with a single color (e.g., strawberry) is worst represented in corpora
- Possible solution
  - more attention to the quality of training data
  - extend textual corpora with extralinguistic information (cf. multimodal models)

The LLM Learning Problem

Hypothesis

Some dimensions of meaning might be occur weakly in the linguistic signal

- The Reporting Bias (Gordon and Van Durme 2013)
  - people tend to omit information that is obvious (cf. Grice's Maxim of Quantity)
- Paik et al. (2021) show that color information about concepts associated with a single color (e.g., strawberry) is worst represented in corpora

- Possible solution
  - more attention to the quality of training data
  - extend textual corpora with extralinguistic information (cf. multimodal models)

The LLM Learning Problem

Hypothesis

Some dimensions of meaning might be occur weakly in the linguistic signal

- The Reporting Bias (Gordon and Van Durme 2013)
  - people tend to omit information that is obvious (cf. Grice's Maxim of Quantity)
- Paik et al. (2021) show that color information about concepts associated with a single color (e.g., strawberry) is worst represented in corpora

- Possible solution
  - more attention to the quality of training data
  - extend textual corpora with extralinguistic information (cf. multimodal models)

The LLM Learning Problem

Hypothesis

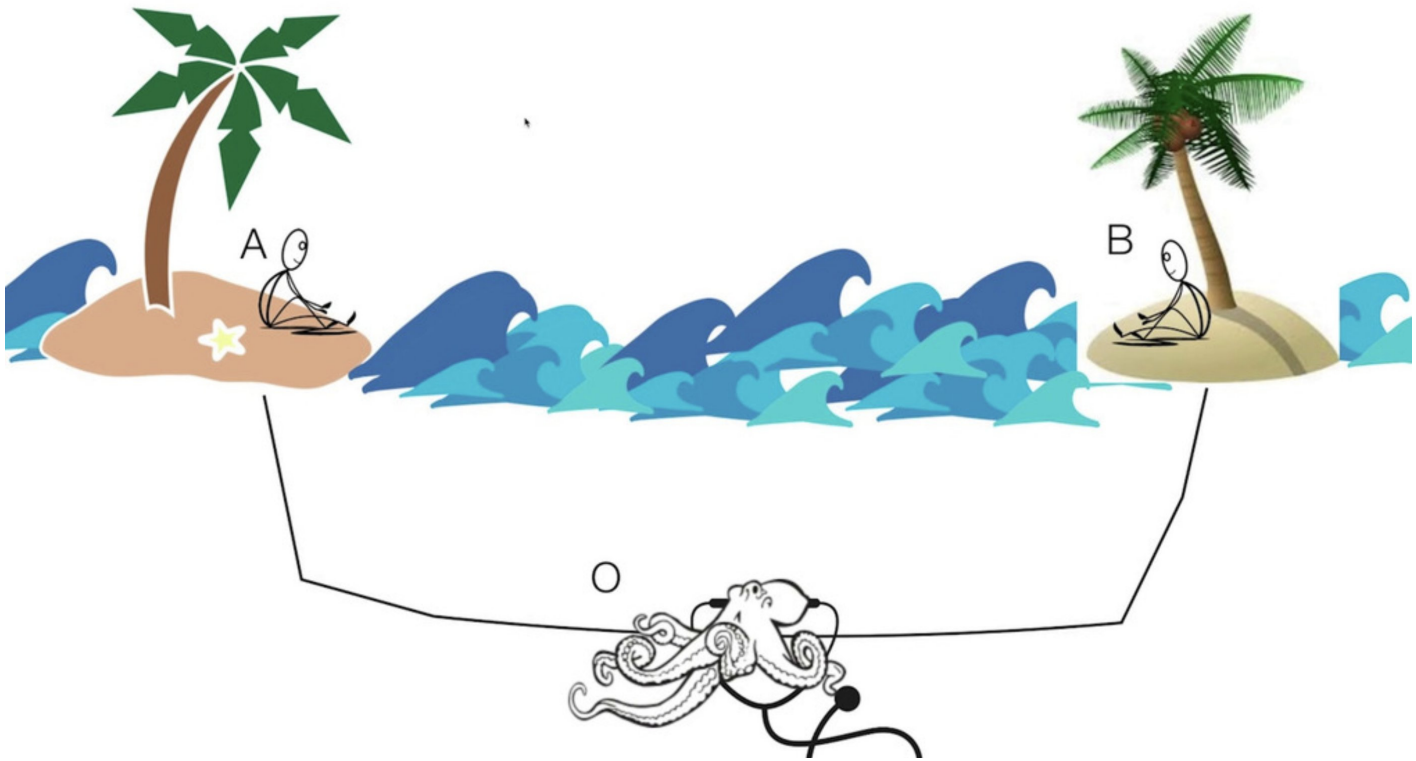
Some dimensions of meaning might be occur weakly in the linguistic signal

- The Reporting Bias (Gordon and Van Durme 2013)
  - people tend to omit information that is obvious (cf. Grice's Maxim of Quantity)
- Paik et al. (2021) show that color information about concepts associated with a single color (e.g., strawberry) is worst represented in corpora

- Possible solution
  - more attention to the quality of training data
  - extend textual corpora with extralinguistic information (cf. multimodal models)

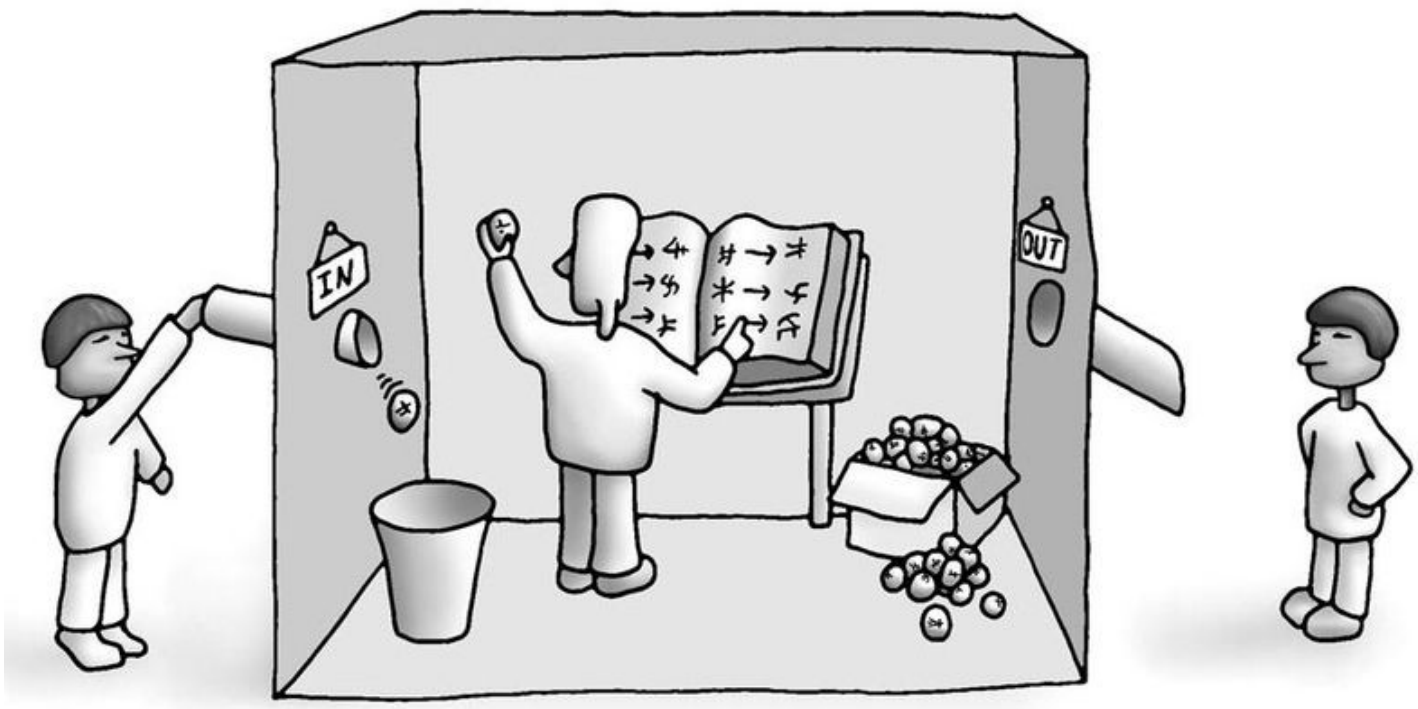
The Octopus Argument

Bender, E. M. and Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data.. *Proc. ACL*: 5185–5198



The Chinese Room Argument

Searle, J.R. (1980). Minds, brains, and programs. *The Behavioral and Brain Sciences*, 3, pp. 417-424



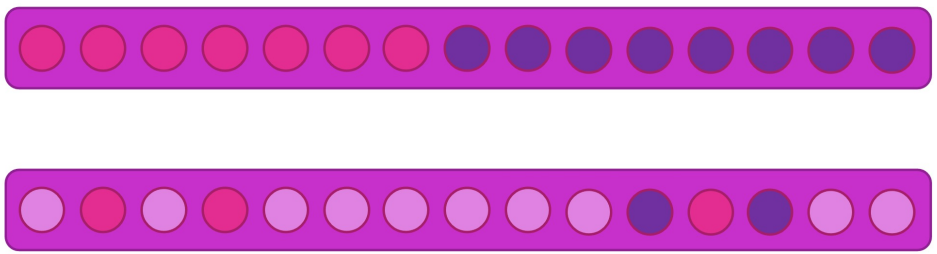
The Symbol Grounding Problem (Harnad 1990)

- Distributional vectors encode associations between symbols (i.e., the orthographic words observed in corpora), but meaning can not spring from symbol-symbol relations only
- The distributional vectors produced by DSMs are ungrounded, but this does not entail that they can not be grounded
- Vectors represent a shared format between linguistic and extralinguistic information

Visual Embeddings

- Multi-modal embeddings encode both visual and textual information
- Textual information can be at both the word (e.g., object) and sentence (e.g., caption) levels

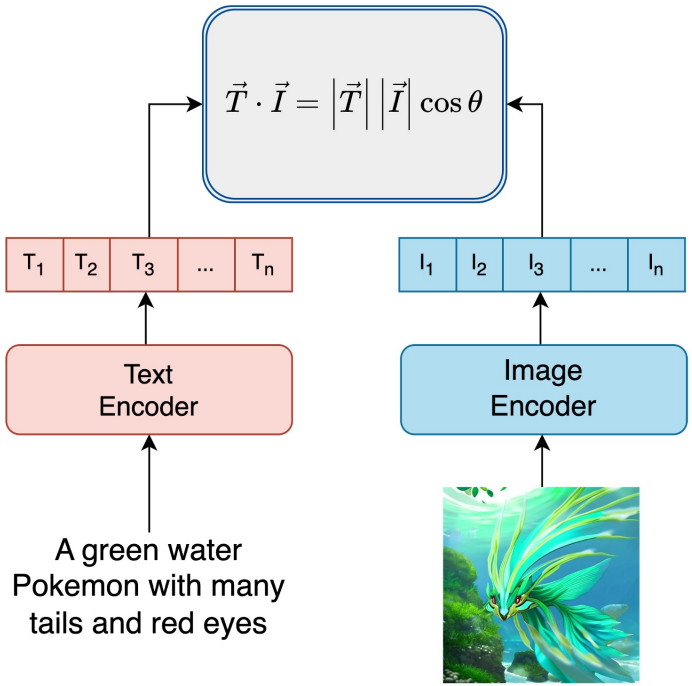
…
a young tennis player preparing to serve a tennis ***ball*** *to her opponent*
a tennis player leaps up to return the ***ball***
a tennis player is trying to hit a ***ball*** *with a tennis racket*
…



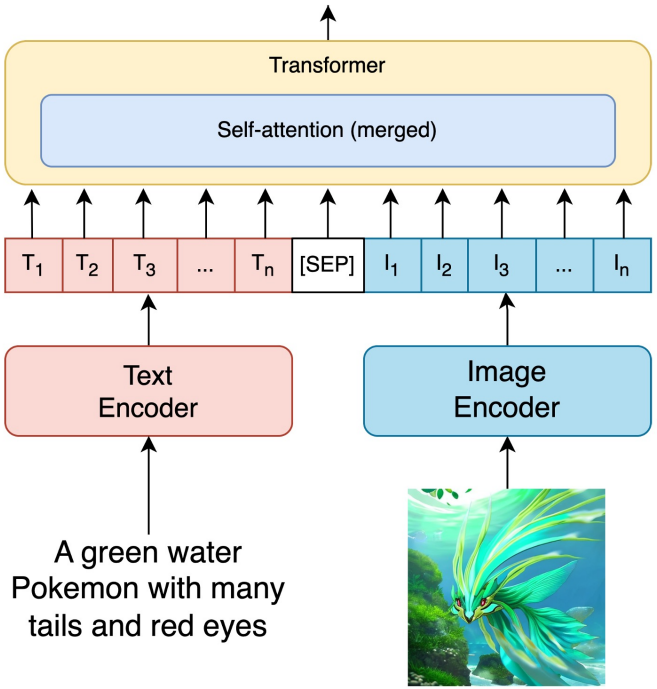
Multimodal Language Models (MLMs)

- Huge deep learning models
- Visual and textual inputs ⇒ textual outputs
- Pre-trained on massive multimodal datasets
- Three main architectural components: 1.) image encoder (e.g, ViT model); 2.) text encoder (LLM); 3.) multimodal bridge

Two-stream models (dual encoder)



Single-stream models (fusion encoder)



Cassese et al. (in press), "Assessing Language and Vision-Language Models on Event Plausibility", *Italian Journal of Computational Linguistics*

Maria Cassese
(Univ. Pisa)

Alessandro Bondielli
(Univ. Pisa)

Datasets

DTFit (Vassallo et al. 2018). 395 plausible and implausible sentences, the latter obtained by replacing the patient with an atypical filler for that role (e.g., *The actor won the award* vs *The actor won the battle)*

EventsAdapt (Fedorenko et al. 2020). The same used in the Kauf et al. (2023) study

EventsRev (Ivanova et al. 2021). 38 plausible and implausible sentences, the latter obtained by reversing the noun phrases, which in this case always depict animate entities (e.g., *The cat is chasing the mouse* vs. *The mouse is chasing the cat*). Each sentence is accompanied by simple black and white drawings depicting the interaction between the two animated participants described in the sentence

MLMs tested with Event Knowledge

VisualBERT (Li et al. 2019). A single-stream early fusion encoder model initialized from pre-trained BERT-base weights and further trained on multimodal datasets. Visual features are extracted from a Faster R-CNN network

FLAVA (Singh et al. 2021). A foundation MLM including a image ViT encoder, a BERT-like textual encoder, and a multimodal encoder

LLaVA (Liu et al. 2023). An open-source chatbot trained by fine-tuning a LLM on multimodal instruction-following data

- instruction-tuned LLMs:
  - MISTRAL
  - VICUNA

Results

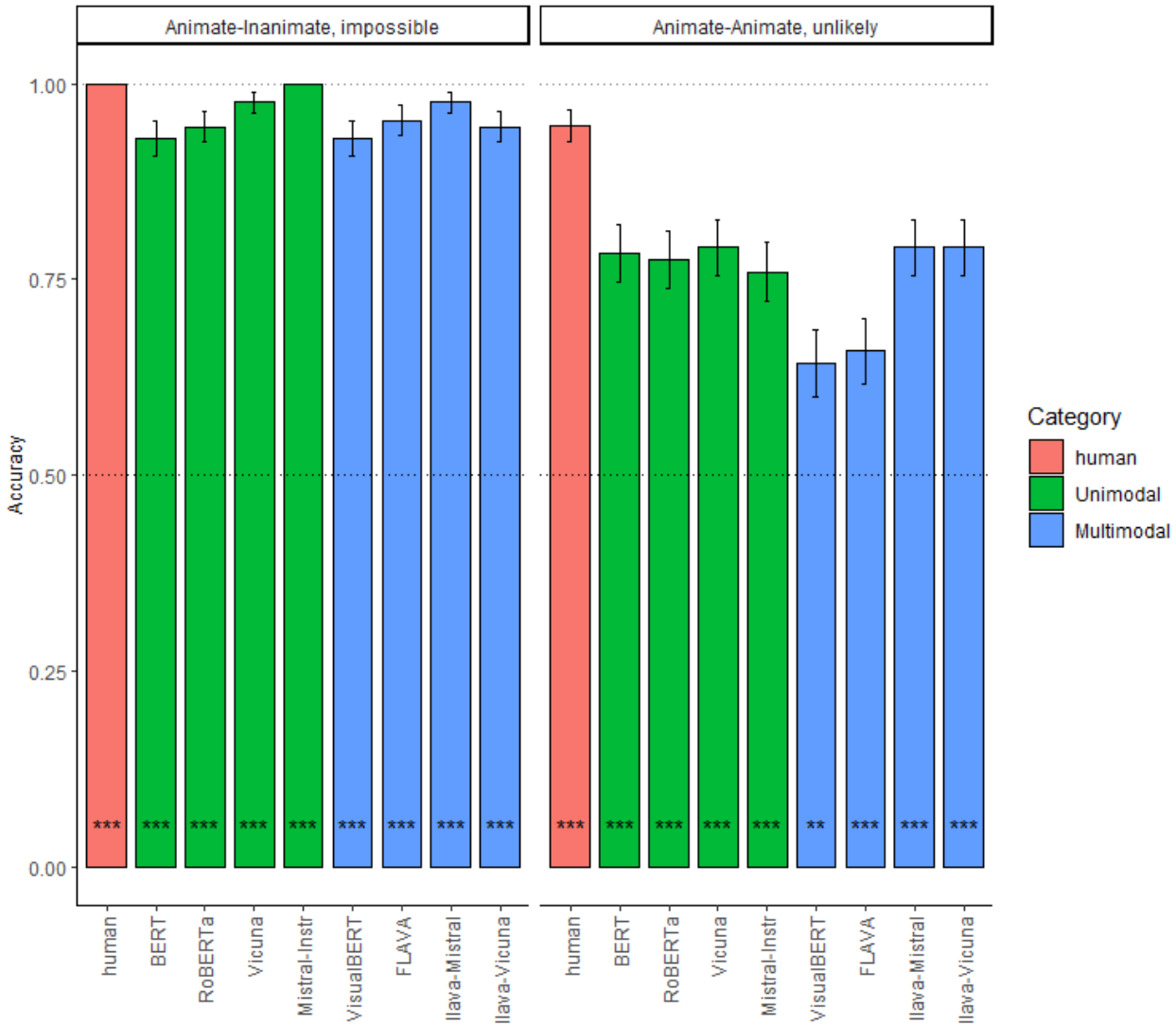
Table: Textual models accuracy on the different datasets

Dataset	Size	Human	BERT	RoBERTa	Mistral	Vicuna
DTFit	395	0.99	0.85	0.89	0.84	0.85
$EvAdapt_{an-in}$	128	1.00	0.93	0.95	**1.00**	0.98
$EvAdapt_{an-an}$	129	0.95	0.78	0.78	0.76	**0.79**
EvRev	38	1.00	0.76	0.79	0.87	0.89

Table: Multimodal models accuracy on the different datasets

Dataset	Size	Human	VisualBERT	FLAVA	LLAVA-Mistral	LLAVA-Vicuna
DTFit	395	0.99	**0.90**	0.86	0.85	0.78
$EvAdapt_{an-in}$	128	1.00	0.93	0.95	0.98	0.95
$EvAdapt_{an-an}$	129	0.95	0.64	0.66	**0.79**	**0.79**
EvRev	38	1.00	0.76	0.79	0.89	**0.92**

Results on EventsAdapt



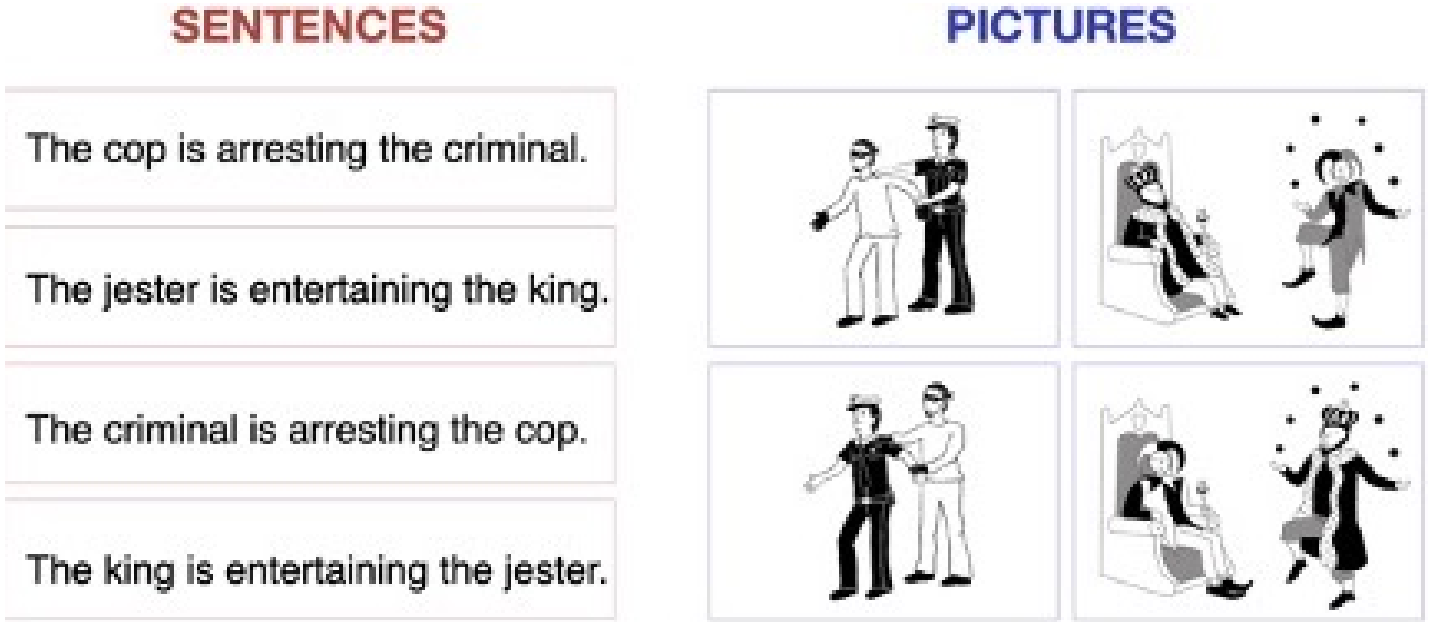
Concrete vs. Abstract Events

Dataset	Size	Human	BERT	RoBERTa	Mistral-Instr	Vicuna
$DTFit_{concr}$	350	0.99	0.85	**0.90**	0.84	0.78
$DTFit_{abstract}$	45	0.95	0.90	0.87	0.82	0.74
$EventsAdapt_{concr_{AN-IN}}$	97	1	0.95	0.95	**1**	0.94
$EventsAdapt_{abstr_{AN-IN}}$	31	1	0.87	0.94	**1**	**1**
$EventsAdapt_{concr_{AN-AN}}$	65	0.96	**0.82**	0.75	**0.82**	0.79
$EventsAdapt_{abstr_{AN-AN}}$	64	0.94	0.75	**0.80**	0.69	0.75

Dataset	Size	Human	VisualBERT	FLAVA	llava-Mistral	llava-Vicuna
$DTFit_{concr}$	350	0.99	**0.90**	0.86	0.85	0.78
$DTFit_{abstract}$	45	0.95	**0.92**	**0.92**	0.85	0.74
$EventsAdapt_{concr_{AN-IN}}$	97	1	0.94	0.95	0.99	0.94
$EventsAdapt_{abstr_{AN-IN}}$	31	1	0.90	0.97	0.94	0.97
$EventsAdapt_{concr_{AN-AN}}$	65	0.96	0.70	0.67	**0.82**	0.81
$EventsAdapt_{abstr_{AN-AN}}$	64	0.94	0.56	0.62	0.75	0.78

Adding Images of the Events Expressed by the Sentences

The sentences in the EventRev dataset are associated with pictures that have been fed to the MLMs



Dataset	**VisualBERT**	**FLAVA**	**llava-Mistral**	**llava-Vicuna**
$EventsRev_t$	0.76	0.79	0.92	**0.95**
$EventsRev_{t+i}$	0.61	0.79	**0.84**	0.79

Causal Relations

- Understanding cause-effect connections is a hallmark of human cognition: Causes precede effects, but causality is not merely temporal precedence

Objectives

- a controlled dataset of sentence pairs bounded by a different degree of causality and temporality relation
  - the identification of the correct relations depends on general knowledge about event relations
- testing the knowledge of causal relations of state-of-the-art LLMs in zero-shot prompting setting

Causal Relations

- Understanding cause-effect connections is a hallmark of human cognition: Causes precede effects, but causality is not merely temporal precedence

Objectives

- a controlled dataset of sentence pairs bounded by a different degree of causality and temporality relation
  - the identification of the correct relations depends on general knowledge about event relations
- testing the knowledge of causal relations of state-of-the-art LLMs in zero-shot prompting setting

ImpliCA: A New Dataset on Implicit Causal Relations

- 600 English sentence pairs:
  - 200 linked by an implicit causal relation: Absence of causal connectives (*because*, *so*, etc.) and causal verbs (*cause*, *result*, etc.)
  - 200 linked by an implicit temporal precedence relation, but no causal relation
  - 200 unrelated (neither causal, nor temporal relation)

- Sentences were classified by 5 expert annotators

ImpliCA: A New Dataset on Implicit Causal Relations

Causal

A: *Matteo wanted to buy a new house*. B: *Matteo asked for a loan from the bank*.

Temporal Precedence

A: *Erik entered the airport*. B: *Erik went to the check-in desk*.

Unrelated

A: *The sea is full of fish*. B: *The seagull flies in the sky*.

- Unrelated sentences express associated events, but without temporal or causal relation
  - association is measured with PMI and LMI of each pair of lexical elements in the two sentences on the UkWac Corpus
  - unrelated and causal+temporal sentences have the same degree of statistical association (Wilcoxon test - PMI: p-value=0.40; LMI: p-value=0.43)

Models and Prompts

- Instruction-tuned models:

Bloom : bloom-7b1 (Muennighoff et al. 2022)
Falcon : falcon 7b-instruct (Almazrouei et al. 2023)
LLaMA : Llama-2-7b-chat-hf (Touvron et al. 2023)
Mistral : Mistral-7B-Instruct-v0.1 (Jiang et al. 2023)
GPT : gpt-3.5-turbo and gpt-4 (Brown et al. 2020)

- Causality prompt: Is it likely that the event in Sentence A causes (brings about) the event in Sentence B?
- Temporality prompt: Does the event in Sentence A typically precede the event in Sentence B?

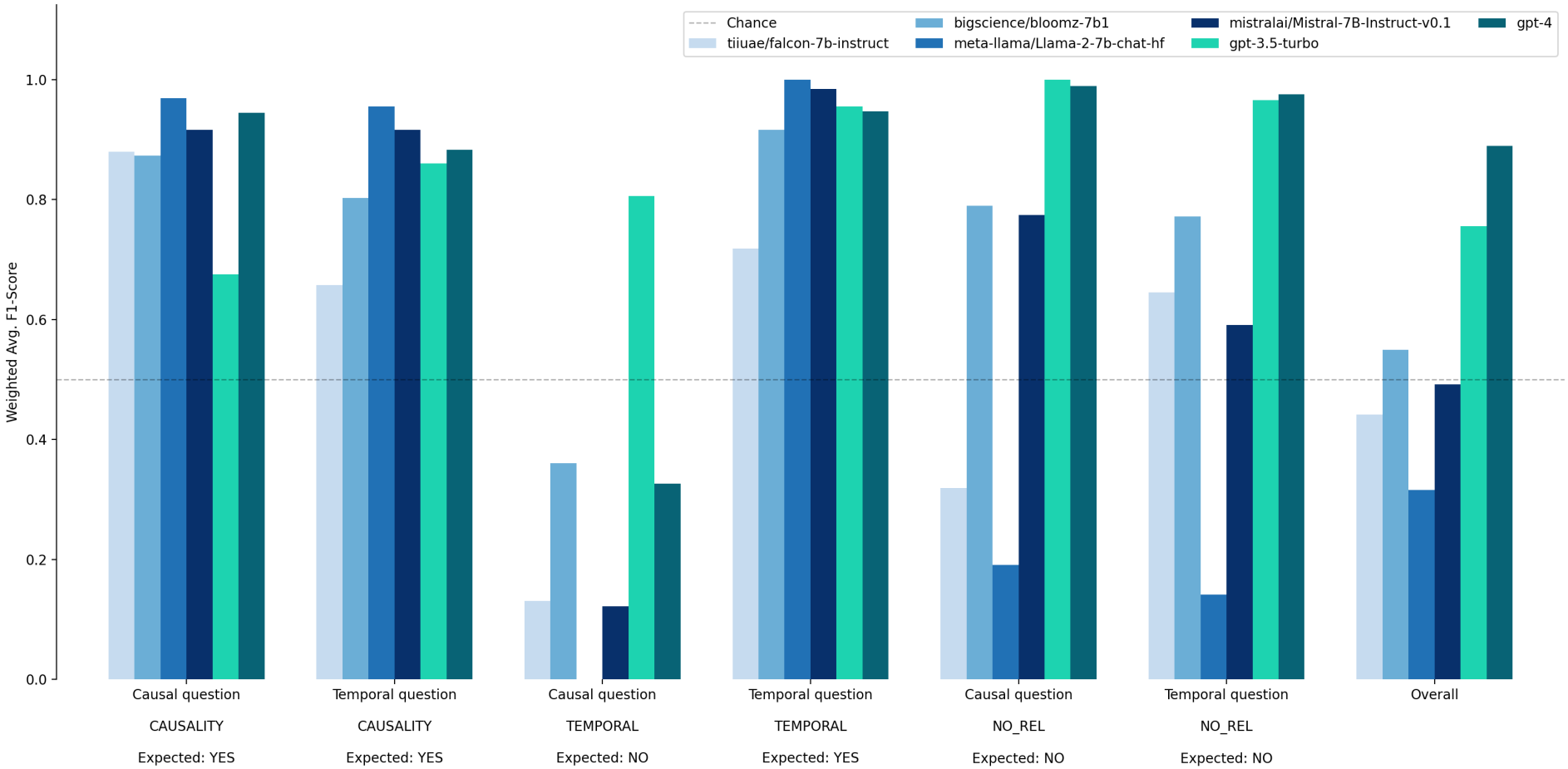
Models and Prompts

- Instruction-tuned models:

Bloom : bloom-7b1 (Muennighoff et al. 2022)
Falcon : falcon 7b-instruct (Almazrouei et al. 2023)
LLaMA : Llama-2-7b-chat-hf (Touvron et al. 2023)
Mistral : Mistral-7B-Instruct-v0.1 (Jiang et al. 2023)
GPT : gpt-3.5-turbo and gpt-4 (Brown et al. 2020)

- Causality prompt: Is it likely that the event in Sentence A causes (brings about) the event in Sentence B?
- Temporality prompt: Does the event in Sentence A typically precede the event in Sentence B?

Testing ImpliCA



The "Knowledge" of (M)LMs

- The "knowledge" of LLMs is huge but is still far from being human-like, as they still strive to distinguish plausibile from implausible but possible events
- They tend to memorize lots of textual sequences (Carlini et al. 2022), lacking the truly generalized nature of (event) knowledge
- MLMs do not improve over LLMs on event knowledge
  - the added value of images is mitigated by the bag-of-words behavior of MLMs (Thrush et al. 2022, Castro et al. 2023), which prevent them from distinguish even plausibility when this depends on argument swapping
  - MLMs have still several problems in recognizing verbs (Hendricks and Nematzadeh 2021)

The "Knowledge" of (M)LMs

- The "knowledge" of LLMs is huge but is still far from being human-like, as they still strive to distinguish plausibile from implausible but possible events
- They tend to memorize lots of textual sequences (Carlini et al. 2022), lacking the truly generalized nature of (event) knowledge
- MLMs do not improve over LLMs on event knowledge
  - the added value of images is mitigated by the bag-of-words behavior of MLMs (Thrush et al. 2022, Castro et al. 2023), which prevent them from distinguish even plausibility when this depends on argument swapping
  - MLMs have still several problems in recognizing verbs (Hendricks and Nematzadeh 2021)

The "Knowledge" of (M)LMs

- The "knowledge" of LLMs is huge but is still far from being human-like, as they still strive to distinguish plausibile from implausible but possible events
- They tend to memorize lots of textual sequences (Carlini et al. 2022), lacking the truly generalized nature of (event) knowledge
- MLMs do not improve over LLMs on event knowledge
  - the added value of images is mitigated by the bag-of-words behavior of MLMs (Thrush et al. 2022, Castro et al. 2023), which prevent them from distinguish even plausibility when this depends on argument swapping
  - MLMs have still several problems in recognizing verbs (Hendricks and Nematzadeh 2021)

(M)LMs beyond the Grounding Problem

- The still limited added value of MLMs might be an effect of the current limits of visual analysis techniques, but it might depend on the very type of "knowledge" they extract from linguistic and visual data
- They mainly identify highly sophisticated associative links between linguistic expressions, but they do not have a semantic space organized in terms of structured "theories" that might support a truly inferential competence
  - cf. the still limited ability of discriminating causal from temporal relations
- The "core knowledge" of foundation models is rich of factoids and associations (far more than any human being could ever master, given the huge amount of data they are extracted from), but it might the same structured organization as the human knowledge system

(M)LMs beyond the Grounding Problem

- The still limited added value of MLMs might be an effect of the current limits of visual analysis techniques, but it might depend on the very type of "knowledge" they extract from linguistic and visual data
- They mainly identify highly sophisticated associative links between linguistic expressions, but they do not have a semantic space organized in terms of structured "theories" that might support a truly inferential competence
  - cf. the still limited ability of discriminating causal from temporal relations
- The "core knowledge" of foundation models is rich of factoids and associations (far more than any human being could ever master, given the huge amount of data they are extracted from), but it might the same structured organization as the human knowledge system

(M)LMs beyond the Grounding Problem

- The still limited added value of MLMs might be an effect of the current limits of visual analysis techniques, but it might depend on the very type of "knowledge" they extract from linguistic and visual data
- They mainly identify highly sophisticated associative links between linguistic expressions, but they do not have a semantic space organized in terms of structured "theories" that might support a truly inferential competence
  - cf. the still limited ability of discriminating causal from temporal relations
- The "core knowledge" of foundation models is rich of factoids and associations (far more than any human being could ever master, given the huge amount of data they are extracted from), but it might the same structured organization as the human knowledge system

Tack!!!

Thank you!!!

Grazie!!!