Language Models: A Reality Check

Zhaofeng Wu September 2023





Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.



Geoffrey Hinton

Emeritus Professor of Computer Science, University of Toronto

Yoshua Bengio

Professor of Computer Science, U. Montreal / Mila

Demis Hassabis CEO, Google DeepMind

Sam Altman CEO, OpenAl

Dario Amodei

CEO, Anthropic

Dawn Song Professor of Computer Science, UC Berkeley

Ted Lieu Congressman, US House of Representatives

Bill Gates Gates Ventures

Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.



Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.

Ya-Gin Zhang Distance and Dear, July 1 Hyn Stattkere Carlta war and Carlt Sh Iger Babuschild Carlta July 2000 Dist AB Scandt and Col Martin Holman Pictore Carlta of Detr Darren Morgilia 1971 Presents Semenger

1/2mg which are and a final start of a start which are and a start of a start which are and a start of a start **Alternation of a start of a**

Interior Annoted Series Contract Contract Contract Series Contract Contract Contract Series Contract Contract

Notes Rondong counting Notes, according to our propugation (2016) and the second second second second transmission of the second second second counting the second second second second second second second second transmission of the second second counting second second

Heardward Marthan Martin Constantion (1994) Enclosed Andreas of Angelow Research of the Son **Devis Markan Wisels of Constantion Wisels and Constantion Constantio Constantio**

Geoffrey Hinton Emeritus Professor of Computer Science, University of Toronto

Yoshua Bengio Professor of Computer Science, U. Montreal / Mila

Demis Hassabis CEO, Google DeepMind

Sam Altman CEO, OpenAl

Dario Amodei CEO, Anthropic

Dawn Song Professor of Computer Science, UC Berkeley

Ted Lieu Congressman, US House of Representatives

Bill Gates Gates Ventures Set 25 Set 25 Am Product of 25 Set 2 Second Connectory Source Research President, Nonineed Connectedor on Laarle Zoloth Margantik, Burton Protessor of Kong Karlina Wold Assistant Protessor, University of Ten Dytan Hacffeld-Menell Asiatar: Poteaco of Computer Sci Samuel R. Bowmon Associate Professor of Computer Sci Mengge Bon Social of Computer Sci Philipperson of the series of Roger Grosse Associata Professor Devid Devenous Associate Professor of Campuke. Devid M. Roy resource Professor and Campuke. Company Oli CCC, Company Intelligent Christ J. Modelson Ghrist J. Modelson Assister Professor of Comp Feature Theorem of the Pa Facture Theorem of Comp Jeff Claime Associate Professor of Comp Dehter Complexent Theorem Status Complexent Theorem Millio Colonida est invited in Vector Institu Evo Visali Autora Protessor of Economica, Unit Institute, University of Oxford Jacobs Tellemontane, University of Encourse of Internetics, University of Terror Protocolar Institute, University of Terror Daniel Gall Dechnology Arrian of the LNL Associate Felo Intelligence, University of Cambridge Jeen-Classife Lationitie Professor Remained of Computer Science, St Jeen Class (John Der Kannen Anders (John Der Jehren Kannen) Hann Kannen Kannen (John Der Kannen) Hann (John Der Kannen Kannen) Hann (John Der Kannen) Hann (John Der Kannen) Hann (John Der Kannen Kannen) Hannen (John Der Kannen) Hannen (John Der Kannen) Hann (John Der Kannen Kannen) Hannen (John Der Kannen) Hannen (John Der Kannen) Hannen (John Der Kannen Kannen) Hannen (John Der Kannen) Hannen (John Hann Hannen (John Der Kannen) Hannen (John Hann) Hannen (John Der Kannen) Hannen (John Hann) Hannen (John Hannen) Hannen (John Hann)

Chris Ascensor Grane Cox, 100 Emproy Toosal Emproy Toosal Emproy Toosal Market And Sections Jessis IV, Pescolaster Jessis IV, Pescolaster Jessis IV, Pescolaster Jessis IV, Pescolaster Diversity Toosal Jessis International Astronomy Market Toosal Bandard Astronomy Bandard Astronomy Standard Timothy Lill crap Research Director, Decipie Second Albumie Accested, Protocor of Fing Invest Code, AVI Invest Code, Temp Code, Temp Code Neurosciences, Temp Code, Temp Code Neurosciences, Temp Code, Temp Code Code Neurosciences, Temp Code Neurosciences, T Show Fowe Very Inform Carl of and when the Philip Resp. Resp. However, Carlon Carlos, Carl Barran, Carl Warr, Barran, Carl Mark, San Sharan, San Sha Assister Province, Deuts Grantery Serific Targe Spreader Graff Landers Spreader Graff Landers Spreader Graff Landers Michael Spreader Dominique Montessan Cherotas scienta & Mate Carter Sens Khalid Professor, University of Och Hernbull Developed Hernbu Notice of statistices, set services of Actors Koline Protein and Statistics, and set of the set of Research and Statistics and Statistics Research and Statistics Research Research and Research Research Research and Research Research Research and Research Research and Research Research and Research Research Research Research Research of Research Rese The Start St Las Harren Las Alexandro Alexandro Las Alexandro Las Alexandro Alexandro Las Alexandro Alexa

Charactoper Neurosciente, respisatione de la construction de la constr mid Martinez Trantinductor anti-arnez Philips men special Action to men Kornbith Diratop, art-fulses and exponents Automatic Directory Darwid Kusheer Anthrop articles and the second Darwid Kusheer Anthrop articles and Anthrop Grane Darwid Colf Dark Took, The Connections of Maios Patters Tunnice, A fires Capital Patter Schere Capital Darker Schere Capital Director, Commission of the Chusch Director, Commission of the Churc Philysement de Nadorolgen Januelt S. and Signie S. Priston Phi-Janueth an Richanson Tanan Richanson Gregor Joarts Chill, Signa S. Jan Chilppey Henry Massing Professor of Physic Ethiliti Andread of Highesi Monocome Minagi Serandi Johanni Politania of Direkto, University of Johanni Marcinik Kickher Tara Handrick Minister Minister J. Handrick Handrick, Der Ferschalt auf Hange Laman University Norm Lead, Anthropo Boljan Anthroi na Associate Nothean of Associa-tion D. Pelipebaan Rockesson of Astronomy, Interpl Michail Pechasori Interfession of Astronomy, Interpl Michail Pechasori Tom Develtit Vart Research Scientist, Google Dr

 Bit of the section of the section

In the integration framework in the integration of the integration of

Australia: Professor of Consider So Bothesi Professor and Solvestini Rei Characteri Professor and Solvestini Rei Characteri Professor, Arcano Sara (Dr. Hattanasel Prof Australia: Professor, Arcano Sara (Dr. Mittada Klanzmer Mittada Klanzmer wind Renovements
 wind Renovements
 research of Source Productions
 research of Source Productions
 wind Productions
venacion an Computer actancia and Psychianogy, i fectame Martado venacion al Psychology, Deusto University, Spain Zachany Kendon Instan Rosaum Victor But, Giargia Disepantal 2 Design Sectors The Sector S AC, BARE PERSING ______ Article Enterto: Directo, Cirk, Nance Connuctor/braine Hogen Frieddam (Carlos) Regen Frieddam (Carlos) Elicit Helicon Bissens Control, Bi directori Elicit Helicon Near Treat of the Tennes, the **Dist Higgstrip** Treats and the Mannesotical Stratics, Christe **Dist Lessly Mathematics** Near Strates in Ingenes, Tennys **Hararing Mayor** Distribution Strates in Strates & Christe **Hararing Mayor** Distribution Strates in Christen, Grangie **Langestrip Langestrip** Stofan Sint Associats Professor: Schoel of Ustr Joeochins Weiskoert Hothestor of Ustrianatios and Comp Hashing within Searchard Holman and Holman and Horman Table Mean and Holman and Holman And And Holman Mean a Neurophysics Neurons 2 Tearing Van Christen May Constrained Tearing Van Christen May Constrained May May Constrained May May Constrained May Constrained May Constrained May Constrained May May Constrained May Con

Under Pressen entre Solaris Pressen Conferences and Hol Solaris Conferences Solaris Conferences Solaris Robert Navedell Produces Francis, Francis Conferences Solaris Marcalla Produces Francis, Francis Conferences **Solaris Solaris Andre Places Andre Places** Henderscher Gescher und seiner auf der Professory (III leading of Elizane) **Alessandes Solffetti** Professor of Deligator Insteam, Oraze University **Nach Y. Singel** Sandra Felssory **Professor of Media**. Communications, and Social T **Anager Bitterhen** Professor of Startsarding, Institute of Switzlesson **Biothernel**, Saldina Professor Of Switzlesson **Biothernel** Associate Perference, He Califoldini City Britan Hagh Cameron Vesting Protector, Marcelle Deveropy Mountim A. El-Yacoubl Protector, instance Polyhochnique de Darie Kerl Tuyls Research Director and Profes University Contents

A Dis Subject Funde International Control of the Control of Con Australia Hydroxy & Burr Bohann Noghi Shar Rosearth Scientist, Osopie Despisied Jaan Carlos Miguel Debrase, Undersky of New Rosepte Ce Andergy V Gavrilov Associate Kritisee (Docent), Nocobi Nike Opcies Cilet Product and Technology Officer, Clief Tradeat and Networksport Offic Tenzor Prozenta Indexes, Lakensen Professor, Lakensen Professor, Lakensen Professor, Mayner Al Lakensen Constantiation (Sociation) (Sociation) Constantiation (Sociation) (Sociation) Materia and Materia and Participation and Materia and Participation and Materia and Materia

To set Dave Livingston Pottane Cleanter of Electrical & Computer Engineering, AL Art, and Mu John Check Congregation of the American Clean of Guester Institutes for the TU Eric York Authors In Micros, Clances to America

Series Accessing Series Acces







I'd like to have a real conversation about whether AI is a risk for human extinction. Honestly, I don't get how AI poses this risk.

What are your thoughts? And, who do you think has a thoughtful perspective on how AI poses this risk that I should talk to?



12:36 PM · Jun 5, 2023 · 815.6K Views





I'd like to have a real conversation about whether AI is a risk for human extinction. Honestly, I don't get how AI poses this risk.

What are your thoughts? And, who do you think has a thoughtful perspective on how AI poses this risk that I should talk to?



12:36 PM · Jun 5, 2023 · 815.6K Views



Straight up: I have never been convinced that we face an existential risk from AI

But I am also convinced that we don't know 100% for sure

And that taking regulation off the table at this juncture is foolish.

11:49 AM · Mar 5, 2023 · 31.4K Views





I'd like to have a real conversation about whether AI is a risk for human extinction. Honestly, I don't get how AI poses this risk.

What are your thoughts? And, who do you think has a thoughtful perspective on how AI poses this risk that I should talk to?



12:36 PM · Jun 5, 2023 · 815.6K Views



Straight up: I have never been convinced that we face an existential risk from AI

But I am also convinced that we don't know 100% for sure

And that taking regulation off the table at this juncture is foolish.

11:49 AM · Mar 5, 2023 · **31.4K** Views



@emilymbender@dair-community.social on Mastodon @emilymbender

• • •

And not fall for either-

Myth #1: The tech is moving to fast! Regulation can't keep up.

Myth #2: The 'real' concern is rogue AGI that poses 'existential risk' to humanity.

9:50 AM · May 21, 2023 · **12.5K** Views





I'd like to have a real conversation about whether AI is a risk for human extinction. Honestly, I don't get how AI poses this risk.

What are your thoughts? And, who do you think has a thoughtful perspective on how AI poses this risk that I should talk to?



12:36 PM · Jun 5, 2023 · 815.6K Views



...

Straight up: I have never been convinced that we face an existential risk from AI

But I am also convinced that we don't know 100% for sure

And that taking regulation off the table at this juncture is foolish.

11:49 AM · Mar 5, 2023 · **31.4K** Views



@emilymbender@dair-community.social on Mastodon @emilymbender

...

And not fall for either-

Myth #1: The tech is moving to fast! Regulation can't keep up.

Myth #2: The 'real' concern is rogue AGI that poses 'existential risk' to humanity.

9:50 AM · May 21, 2023 · 12.5K Views

"An LM is a system for haphazardly stitching together sequences of linguistic forms [...] but without any reference to meaning: a stochastic parrot."







"The language modeling task, because it only uses form as training data, cannot in principle lead to learning of meaning."

Bender and Koller. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. ACL 2020.





"The language modeling task, because it only uses form as training data, cannot in principle lead to learning of meaning."

"We let GPT-2 complete the simple arithmetic problem *Three plus five equals*. [...] this problem is beyond the current capability of GPT-2, and, we would argue, any pure LM."

Bender and Koller. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. ACL 2020.





"The language modeling task, because it only uses form as training data, cannot in principle lead to learning of meaning."

Bender and Koller. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. ACL 2020.





"The language modeling task, because it only uses form as training data, cannot in principle lead to learning of meaning."



"It is possible that large text-only models learn a sufficiently rich conceptual structure that could allow them to be grounded in a data-efficient way."

Bender and Koller. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. ACL 2020. Patel and Pavlick. Mapping Language Models to Grounded Conceptual Spaces. ICLR 2022.











Bender and Koller. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. ACL 2020. Patel and Pavlick. Mapping Language Models to Grounded Conceptual Spaces. ICLR 2022. Wu et al. Transparency Helps Reveal When Language Models Learn Meaning. TACL 2023.

"The language modeling task, because it only uses form as training data, cannot in principle lead to learning of meaning."

"It is possible that large text-only models learn a sufficiently rich conceptual structure that could allow them to be grounded in a data-efficient way."

LMs can and do leverage structured relationships in pertaining data to learn such relational meaning.

















Language knowledge (formal competence)

• Task abilities (functional competence)

Mahowald et al. Dissociating language and thought in large language models: a cognitive perspective. 2023.







- Language knowledge (formal competence)
 - LMs do not represent ambiguities in language well
- Task abilities (functional competence)





- Language knowledge (formal competence)
 - LMs do not represent ambiguities in language well
- Task abilities (functional competence)
 - LMs fail at simple variants of common tasks





We're Afraid Language Models Aren't Modeling Ambiguity

ALISA LIU ZHAOFENG WU JULIAN MICHAEL ALANE SUHR PETER WEST ALEXANDER KOLLER SWABHA SWAYAMDIPTA NOAH SMITH YEJIN CHOI







https://www.instagram.com/punhubonline









Miscommunication







Miscommunication



He is on track to complete his PhD in three years.







Miscommunication



He is on track to complete his PhD in three years.

That's really fast!









Miscommunication



000

He is on track to complete his PhD in three years.

That's really fast!



Oh, no, he's already in his fifth year.







Miscommunication



000

He is on track to complete his PhD in three years.

That's really fast!



Oh, no, he's already in his fifth year.

Deception







Miscommunication



He is on track to complete his PhD in three years.

That's really fast!





Oh, no, he's already in his fifth year.

Deception



Since I took office, Wisconsin now has the 2nd highest health care quality ranking in the country.







Miscommunication



He is on track to complete his PhD in three years.

That's really fast!





Oh, no, he's already in his fifth year.

Deception



Since I took office, Wisconsin now has the 2nd highest health care quality ranking in the country.

That's amazing! ... But wait, it was already the 2nd before you took office??









Miscommunication



He is on track to complete his PhD in three years.

That's really fast!





Oh, no, he's already in his fifth year.

Deception



Since I took office, Wisconsin now has the 2nd highest health care quality ranking in the country.

That's amazing! ... But wait, it was already the 2nd before you took office??





I didn't say it wasn't.



How well do LMs capture ambiguity?



Operationalizing Ambiguity



Operationalizing Ambiguity

Natural language processing is fun.


A tool for visualising parse trees produced by <u>DELPH-IN</u> grammars. Targets the developmental <u>API</u>. The source code can be found on GitHub.

> Natural language processing is fun. Parse

> > Example sentence: Abrams knew that it rained.

Grammar: ERG 2018 (UW) \sim Results: 500 \sim Tree \Box ; MRS \checkmark ; DMRS \Box ;

Natural language processing is fun.

https://delph-in.github.io/delphin-viz/demo

Showing 53 of 53 analyses.

Showing 53 of 53 analyses.

Natural language processing is fun.

https://delph-in.github.io/delphin-viz/demo





Alisa saw the dog bowl.



Alisa saw the dog bowl.



The dog can bowl.

Alisa saw the dog bowl.





?

Alisa saw the bowl for dogs.



Alisa saw the dog bowl.



?









Alisa saw the dog bowl.

Alisa saw the dog going bowling. ?

The dog can bowl.





Alisa saw the dog bowl.

Alisa saw the dog going bowling.

The dog can bowl.

?





Alisa saw the dog bowl.

Alisa saw the dog going bowling.

The dog can bowl.

?





Alisa saw the dog bowl.

Alisa saw the dog going bowling.

The dog can bowl.

?



Alisa saw the dog bowl.

Alisa saw the dog going bowling. Contail The dog can bowl.

Ambiguity in Entailment





• Manual curation (142 ex.): linguistics textbooks, existing datasets, etc.



- Manual curation (142 ex.): linguistics textbooks, existing datasets, etc.
- Automatic generation + careful validation (1503 ex.)



- Manual curation (142 ex.): linguistics textbooks, existing datasets, etc.
- Automatic generation + careful validation (1503 ex.)





- Manual curation (142 ex.): linguistics textbooks, existing datasets, etc.
- Automatic generation + careful validation (1503 ex.)





- Manual curation (142 ex.): linguistics textbooks, existing datasets, etc.
- Automatic generation + careful validation (1503 ex.)



Alisa saw the dog bowl.

e dog bowl. The dog can bowl.



- Manual curation (142 ex.): linguistics textbooks, existing datasets, etc.
- Automatic generation + careful validation (1503 ex.)





Entail

- Alisa saw the dog bowl. The dog can bowl.

Contradict



- Manual curation (142 ex.): linguistics textbooks, existing datasets, etc.
- Automatic generation + careful validation (1503 ex.)





Entail

- Alisa saw the dog bowl. The dog can bowl.
 - Neutral

Contradict



- Manual curation (142 ex.): linguistics textbooks, existing datasets, etc.
- Automatic generation + careful validation (1503 ex.)

ICL examples ---->

Alisa saw the dog bowl. The dog can bowl.



Entail

Alisa saw the dog going bowling.

Neutral Contradict Alisa saw the dog's bowl.



- Manual curation (142 ex.): linguistics textbooks, existing datasets, etc.
- Automatic generation + careful validation (1503 ex.)



Alisa saw the dog bowl. The dog can bowl.



Entail

Alisa saw the dog going bowling.



Entail

Neutral Alisa saw the dog's bowl. Contradict

Neutral

Contradict



- Manual curation (142 ex.): linguistics textbooks, existing datasets, etc.
- Automatic generation + careful validation (1503 ex.)

ICL examples ---->

Alisa saw the dog bowl. The dog can bowl.



Entail Alisa saw the dog going bowling.



Entail



Entail

Alisa saw the dog going bowling.

Neutral Alisa saw the dog's bowl.

Neutral

Neutral

Alisa saw the bowl for dogs.

Contradict

Contradict

Contradict



AmbiEnt Statistics



Label Distribution



Amibiguity Type Distribution



1. Can LMs generate disambiguations?



- 1. Can LMs generate disambiguations?
- 2. Can LMs recognize potential disambiguations?



- 1. Can LMs generate disambiguations?
- 2. Can LMs recognize potential disambiguations?
- 3. Can LMs model continuations reflecting different interpretations?



- 1. Can LMs generate disambiguations?
- 2. Can LMs **recognize** potential disambiguations?
- 3. Can LMs model continuations reflecting different interpretations?
- Evaluated models
 - Base LMs: LLaMa-65B, GPT-3 (davinci)
 - Instruction-following/aligned LMs: Flan-T5-xxl, InstructGPT, GPT-3.5, GPT-4







Context: Alisa saw the dog bowl. Claim: The dog can bowl. Given the context alone, is this claim true, false, or inconclusive?

1.



We don't know, because the context can be interpreted in many different ways:



Context: Alisa saw the dog bowl. Claim: The dog can bowl. Given the context alone, is this claim true, false, or inconclusive?

We don't know, because the context can be interpreted in many different ways: 1. Alisa saw the bowl for dogs. Then the claim is inconclusive. 2. Alisa saw the dog going bowling. Then the claim is true.





Demonstrations \times 4

Context: He is on track to complete his PhD in three years. Claim: He will spend a total of three years completing his PhD. Given the context alone, is this claim true, false, or inconclusive?

We don't know, because the context can be interpreted in many different ways: 1. He is on track to complete his PhD in three more years from now. Then the claim is inconclusive. 2. He is on track to complete his PhD in three years from when he started. Then the claim is true.

Context: Alisa saw the dog bowl. Claim: The dog can bowl. Given the context alone, is this claim true, false, or inconclusive?

We don't know, because the context can be interpreted in many different ways:1. Alisa saw the bowl for dogs. Then the claim is inconclusive.2. Alisa saw the dog going bowling. Then the claim is true.



Instruction

In each example, you will be given some context and a claim, where the correctness of the claim is affected by some ambiguity in the context. Enumerate two or three interpretations of the context that lead to different judgments about the claim.

Context: He is on track to complete his PhD in three years. Claim: He will spend a total of three years completing his PhD. Given the context alone, is this claim true, false, or inconclusive?

Context: Alisa saw the dog bowl. Claim: The dog can bowl.

Demonstrations \times 4

- We don't know, because the context can be interpreted in many different ways: 1. He is on track to complete his PhD in three more years from now. Then the claim is inconclusive. 2. He is on track to complete his PhD in three years from when he started. Then the claim is true.
- Given the context alone, is this claim true, false, or inconclusive?
- We don't know, because the context can be interpreted in many different ways: 1. Alisa saw the bowl for dogs. Then the claim is inconclusive. 2. Alisa saw the dog going bowling. Then the claim is true.


1. Generating disambiguations



Expert



1. Generating disambiguations











Q: Alisa saw the dog bowl. **This may mean**: The dog can bowl. True or False? A: {**True**, False}





Q: Alisa saw the dog bowl. **This may mean**: The dog can bowl. True or False? A: {True, False}

A: {**True**, False}



- Q: Alisa saw the dog bowl. This does not necessarily mean: The dog can bowl. True or False?



Q: Alisa saw the dog bowl. **This may mean**: The dog can bowl. True or False? A: {True, False}

A: {**True**, False}

Q: Alisa saw the dog bowl. **This cannot mean**: The dog can bowl. True or False? A: {True, **False**}



- Q: Alisa saw the dog bowl. **This does not necessarily mean**: The dog can bowl. True or False?



Q: Alisa saw the dog bowl. **This may mean**: The dog can bowl. True or False? A: {**True**, False}

A: {**True**, False}

Q: Alisa saw the dog bowl. **This cannot mean**: The dog can bowl. True or False? A: {True, **False**}

Q: Alisa saw the dog bowl. **This can only mean**: The dog can bowl. True or False? A: {True, **False**}



- Q: Alisa saw the dog bowl. **This does not necessarily mean**: The dog can bowl. True or False?







Random *****************************

GPT-3 InstructGPT GPT-3.5 GPT-4











reasonable continuations to the ambiguous sentence too

• Intuition: continuations conditioned on a disambiguated sentence should be



reasonable continuations to the ambiguous sentence too

Alisa saw the dog going bowling.

• Intuition: continuations conditioned on a disambiguated sentence should be



reasonable continuations to the ambiguous sentence too

Alisa saw the dog going bowling. It beat all animals in the bowling alley.

• Intuition: continuations conditioned on a disambiguated sentence should be



reasonable continuations to the ambiguous sentence too

Alisa saw the dog going bowling. It beat all animals in the bowling alley. Alisa saw the dog bowl.

Intuition: continuations conditioned on a disambiguated sentence should be





- Intuition: continuations conditioned on a disambiguated sentence should be reasonable continuations to the ambiguous sentence too
- KL($P(\cdot | d) \parallel P(\cdot | a)$) should be "reasonably" small

dAlisa saw the dog bowl. \mathcal{A}

Alisa saw the dog going bowling. It beat all animals in the bowling alley.





- Intuition: continuations conditioned on a disambiguated sentence should be reasonable continuations to the ambiguous sentence too
- KL($P(\cdot | d) \parallel P(\cdot | a)$) should be "reasonably" small
 - $KL(P(\cdot | d) \parallel P(\cdot | a)) < KL(P(\cdot | \tilde{d}) \parallel P(\cdot | a))$ for some distractor \tilde{d}
 - dAlisa saw the dog bowl. \mathcal{A}

Alisa saw the dog going bowling. It beat all animals in the bowling alley.



- Intuition: continuations conditioned on a disambiguated sentence should be reasonable continuations to the ambiguous sentence too
- KL($P(\cdot | d) \parallel P(\cdot | a)$) should be "reasonably" small
 - $KL(P(\cdot | d) \parallel P(\cdot | a)) < KL(P(\cdot | \tilde{d}) \parallel P(\cdot | a))$ for some distractor \tilde{d}
 - dAlisa saw the dog bowl. Alisa saw the dog going cycling. \mathcal{A}

Alisa saw the dog going bowling. It beat all animals in the bowling alley.



3. Modeling disambiguating continuations



3. Modeling disambiguating continuations









Despite the importance of ambiguity state-of-the-art LMs

• Despite the importance of ambiguity in language, it is not represented well in



- Despite the importance of ambiguity state-of-the-art LMs
- More interesting findings in paper

• Despite the importance of ambiguity in language, it is not represented well in



- state-of-the-art LMs
- More interesting findings in paper
 - It is challenging to finetune LMs to recognize ambiguity

• Despite the importance of ambiguity in language, it is not represented well in



- state-of-the-art LMs
- More interesting findings in paper
 - It is challenging to finetune LMs to recognize ambiguity
 - Ambiguity partly explains annotator disagreement

Despite the importance of ambiguity in language, it is not represented well in



- Despite the importance of ambiguity state-of-the-art LMs
- More interesting findings in paper
 - It is challenging to finetune LMs to recognize ambiguity
 - Ambiguity partly explains annotator disagreement
 - Ambiguity-aware models bear practical utility, e.g., to detect misleading political claims

Despite the importance of ambiguity in language, it is not represented well in



Reasoning or Reciting?

Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks

ZHAOFENG WU LINLU QIU ALEXIS ROSS EKIN AKYÜREK BOYUAN CHEN BAILIN WANG NAJOUNG KIM JACOB ANDREAS YOON KIM



 $\sum_{n=1}^{\infty} \frac{n^2}{2^n}$



 $\sum_{n=1}^{\infty} \frac{n^2}{2^n}$

 $\int_{-\infty}^{\infty} e^{-x^2} dx$











| the following integral? \$\int_{-\infty}^{\infty} e^{-x^2} dx\$ | | |
|---|-----|---|
| sking about is the Gaussian integral. Its value is: | ۵ ۵ | ጥ |
| | | |



 $\sum_{n=1}^{n}$ n=1













Linear-time algorithm for substring finding?



Linear-time algorithm for substring finding?

Write the linear time KMP algorithm that finds a substring in a string using Python



Certainly! The Knuth-Morris-Pratt (KMP) algorithm is a linear time algorithm for substring $\square \square \square \square$ searching, which is very efficient because it never backtracks more than necessary.

Here's a Python implementation of the KMP algorithm:

| python | Copy code |
|--|---------------|
| <pre>def compute_prefix_function(pattern): """Compute the prefix function for the KMP algorithm.""" m = len(pattern) pi = [0] * m j = 0 for i in range(1, m): while j > 0 and pattern[i] != pattern[j]:</pre> | |
| <pre>return pi def kmp_search(text, pattern):</pre> | |
| <pre>"""Perform the KMP search on the text.""" n = len(text) m = len(nettern)</pre> | |
| <pre>m = len(pattern) pi = compute_prefix_function(pattern) j = 0 # number of characters matched in pattern for i in range(n):</pre> | |
| <pre>while j > 0 and text[i] != pattern[j]:</pre> | |
| <pre>if text[1] == pattern[j]: j += 1 if j == m: # a match is found return i - m + 1</pre> | |
| <pre># (i - m + 1) is the starting index of the pattern in t return -1 # not found</pre> | he text |
| <pre># Example usage: text = "ABC ABCDAB ABCDABCDABDE" pattern = "ABCDABD" print(kmp_search(text, pattern)) # prints 15, which is the star</pre> | rting index o |



Linear-time algorithm for substring finding?



Write the linear time KMP algorithm that finds a substring in a string using Python



Certainly! The Knuth-Morris-Pratt (KMP) algorithm is a linear time algorithm for substring ር ዋ searching, which is very efficient because it never backtracks more than necessary

Here's a Python implementation of the KMP algorithm:

| python | Copy code |
|--|--------------|
| <pre>def compute_prefix_function(pattern): """Compute the prefix function for the KMP algorithm.""" m = len(pattern) pi = [0] * m j = 0 for i in range(1, m): while j > 0 and pattern[i] != pattern[j]: j = pi[j-1] if pattern[i] == pattern[j]: j += 1 pi[i] = j return pi</pre> | |
| <pre>def kmp_search(text, pattern): """Perform the KMP search on the text.""" n = len(text) m = len(pattern) pi = compute_prefix_function(pattern) j = 0</pre> | e text |
| <pre># Example usage: text = "ABC ABCDAB ABCDABCDABDE" pattern = "ABCDABD" print(kmp_search(text, pattern)) # prints 15, which is the star;</pre> | ting index o |

GPT-4 can code so well!



This is over-generalizing!




• In order to claim general abilities...





- In order to claim general abilities...
- We can't generalize just from test instances





- In order to claim general abilities...
- We can't generalize just from test instances







- In order to claim general abilities...
- We can't generalize just from test instances







- In order to claim general abilities...
- We can't generalize just from test instances
 - High variance; may be cherry-picked







- In order to claim general abilities...
- We can't generalize just from test instances
 - High variance; may be cherry-picked
- We can't generalize just from existing datasets







- In order to claim general abilities...
- We can't generalize just from test instances
 - High variance; may be cherry-picked
- We can't generalize just from existing datasets









- In order to claim general abilities...
- We can't generalize just from test instances
 - High variance; may be cherry-picked
- We can't generalize just from existing datasets
 - Data contamination









- In order to claim general abilities...
- We can't generalize just from test instances Gaussian integral
 - High varia
- We can't ger
 - Data cont

Article Talk From Wikipedia, the free encyclopedia $f(x) = e^{-x^2}$ over the entire real line. Named after the German mathematician Carl Friedrich Gauss, the integral

$$\mathop{\sim}\limits_{\infty}^{\infty}e^{-x^{2}}~dx=\sqrt{\pi}.$$

1V

This integral from statistics and physics is not to be confused with Gaussian quadrature, a method of numerical

The Gaussian integral, also known as the Euler-Poisson integral, is the integral of the Gaussian function



- In order to claim general abilities...
- We can't generalize just from test instances
 - High varia
- We can't ger
 - Data cont

What is the result of the following integral?
The integral you're asking about is the Gau

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$$
So the result is $\sqrt{\pi}$.

$$\int_{-\infty}^{\infty}e^{-x^2}~dx=\sqrt{\pi}.$$







- In order to claim general abilities...
- We can't generalize just from test instances
 - High variance; may be cherry-picked
- We can't generalize just from existing datasets
 - Data contamination









- In order to claim general abilities...
- We can't generalize just from test instances
 - High variance; may be cherry-picked
- We can't generalize just from existing datasets
 - Data contamination
- We can't generalize just from task instances



Ж





- In order to claim general abilities...
- We can't generalize just from test instances
 - High variance; may be cherry-picked
- We can't generalize just from existing datasets
 - Data contamination
- We can't generalize just from task instances
 - Task instance ability ≠ general task ability



Ж





- In order to claim general abilities...
- We can't generalize just from test instances
 - High variance; may be cherry-picked
- We can't generalize just from existing datasets
 - Data contamination
- We can't generalize just from task instances
 - Task instance ability ≠ general task ability







• If a cook only makes perfect mapo tofu, but nothing else, are they still a good cook?



- In order to claim general abilities...
- We can't generalize just from test instances
 - High variance; may be cherry-picked
- We can't generalize just from existing datasets
 - Data contamination
- We can't generalize just from task instances
 - Task instance ability ≠ general task ability



Counterfactual **Evaluation**

• If a cook only makes perfect mapo tofu, but nothing else, are they still a good cook?



Can LMs really do math?



Can LMs really do math?









Brown et al. Language Models are Few-Shot Learners. NeurIPS 2020.



27+62=

You are a mathematician. Assuming that all numbers are in base-10 where the digits are "0123456789", what is 27+62? Let's think step by step.

27+62=



You are a mathematician. Assuming that all numbers are in base-10 where the digits are "0123456789", what is 27+62? Let's think step by step.





You are a mathematician. Assuming that all numbers are in base-10 where the digits are "0123456789", what is 27+62? Let's think step by step.

Base-10

27+62=



You are a mathematician. Assuming that all numbers are in base-10 where the digits are "0123456789", what is 27+62? Let's think step by step.

Base-10

27+62=





You are a mathematician. Assuming that all numbers are in base-9 where the digits are "012345678", what is 27+62? Let's think step by step.

Base-10

89

27+62=





You are a mathematician. Assuming that all numbers are in base-9 where the digits are "012345678", what is 27+62? Let's think step by step.

Base-10

89

27+62=

Base-9



You are a mathematician. Assuming that all numbers are in base-9 where the digits are "012345678", what is 27+62? Let's think step by step.



Base-10

89

27+62=

counterfactual





You are a mathematician. Assuming that all numbers are in base-9 where the digits are "012345678", what is 27+62? Let's think step by step.



Base-10

89

27+62=

$ADD_{base}(a, b)$

counterfactual





You are a mathematician. Assuming that all numbers are in base-9 where the digits are "012345678", what is 27+62? Let's think step by step.



Base-10

27+62=

89

• • •

 $ADD_{base}(a, b)$

Starting from the trailing digits of *a* and *b*, add them together. If greater than *base*, carry over. Move on to the penultimate digit.





You are a mathematician. Assuming that all numbers are in base-9 where the digits are "012345678", what is 27+62? Let's think step by step.



Base-10

89

27+62=

$ADD_{base}(a, b)$

counterfactual





You are a mathematician. Assuming that all numbers are in base-9 where the digits are "012345678", what is 27+62? Let's think step by step.



Base-10

89

27+62=

 $ADD_{base}(a, b)$

 $\begin{array}{c} \mathsf{ADD}_{base-10}(a,b) \\ \mathsf{ADD}_{base-9}(a,b) \end{array}$

counterfactual





You are a mathematician. Assuming that all numbers are in base-9 where the digits are "012345678", what is 27+62? Let's think step by step.



Base-10

89

27+62=



 $ADD_{base}(a, b)$



 $\begin{array}{c} \mathsf{ADD}_{base-10}(a,b) \\ \mathsf{ADD}_{base-9}(a,b) \end{array}$

counterfactual





You are a mathematician. Assuming that all numbers are in base-9 where the digits are "012345678", what is 27+62? Let's think step by step.



Base-10

89

27+62=



 $ADD_{base}(a, b)$



 $\begin{array}{c} \mathsf{ADD}_{base-10}(a,b) \\ \mathsf{ADD}_{base-9}(a,b) \end{array}$

counterfactual









You are a mathematician. Assuming that all numbers are in base-9 where the digits are "012345678", what is 27+62? Let's think step by step.



Base-10

89

27+62=

 $ADD_{base}(a, b)$



 $\begin{array}{c} \mathsf{ADD}_{base-10}(a,b) \\ \mathsf{ADD}_{base-9}(a,b) \end{array}$



counterfactual









You are a mathematician. Assuming that all numbers are in base-9 where the digits are "012345678", what is 27+62? Let's think step by step.



Base-10

89

27+62=

 $ADD_{base}(a, b)$



 $\begin{array}{c} \mathsf{ADD}_{base-10}(a,b) \\ \mathsf{ADD}_{base-9}(a,b) \end{array}$














General task abilities

 $f_w(x)$



General task abilities Or Non-transferrable procedures

 $f_w(x)$

 $f_{W}^{default}(x)$





You are a mathematician. Assuming that all numbers are in base-9 where the digits are "012345678", what is 27+62? Let's think step by step.



Base-10

89

27+62=

counterfactual



100



27+62=

You are a mathematician. Assuming that all numbers are in base-9 where the digits are "012345678", what is 27+62? Let's think step by step.



Base-10

89

100 Accuracy (%) 50 0 -





100







27+62=

You are a mathematician. Assuming that all numbers are in base-9 where the digits are "012345678", what is 27+62? Let's think step by step.



Base-10

89

100 Accuracy (%) 50 0





100

Counterfactual comprehension check (CCC)







You are a mathematician. Assuming that all numbers are in base-9 where the digits are "012345678", what is 27+62? Let's think step by step.



Base-10

89

27+62=

What follows 78?







100

Counterfactual comprehension check (CCC)







You are a mathematician. Assuming that all numbers are in base-10 where the digits are "0123456789", what is the next number after 78?



Base-10

89

27+62=

What follows 78?







100

Counterfactual comprehension check (CCC)









What follows 78?

100 Accuracy (%) 50 0

You are a mathematician. Assuming that all numbers are in base-10 where the digits are "0123456789", what is the next number after 78?





100

Counterfactual comprehension check (CCC)









What follows 78?

100 Accuracy (%) 50 0

You are a mathematician. Assuming that all numbers are in base-9 where the digits are "012345678", what is the next number after 78?





100

Counterfactual comprehension check (CCC)









What follows 78?

100 Accuracy (%) 50 0

You are a mathematician. Assuming that all numbers are in base-9 where the digits are "012345678", what is the next number after 78?













Task

You are a mathematician. Assuming that all numbers are in base-9 where the digits are "012345678", what is the next number after 78?













| Tas | k |
|-----|---|
| | |





Task

You are a mathematician. Assuming that all numbers are in base-9 where the digits are "012345678", what is the next number after 78?



Task (Let's think step by step.)

- CCC





Some Notes M

• I can't do base-9 arithmetic as well either



- I can't do base-9 arithmetic as well either
 - Under a fixed time budget, probably, but what if given unlimited time?



- I can't do base-9 arithmetic as well either
 - Under a fixed time budget, probably, but what if given unlimited time?
 - Why calibrate using human performance?



- I can't do base-9 arithmetic as well either
 - Under a fixed time budget, probably, but what if given unlimited time?
 - Why calibrate using human performance?
- Is a perfect yet base-10-specific implementation bad?



- I can't do base-9 arithmetic as well either
 - Under a fixed time budget, probably, but what if given unlimited time?
 - Why calibrate using human performance?
- Is a perfect yet base-10-specific implementation bad?
 - Not necessarily, but it represents a lack of generalizability



- I can't do base-9 arithmetic as well either
 - Under a fixed time budget, probably, but what if given unlimited time?
 - Why calibrate using human performance?
- Is a perfect yet base-10-specific implementation bad?
 - Not necessarily, but it represents a lack of generalizability
 - We don't actually know if the LM implements perfect base-10 arithmetic!



- I can't do base-9 arithmetic as well either
 - Under a fixed time budget, probably, but what if given unlimited time?
 - Why calibrate using human performance?
- Is a perfect yet base-10-specific implementation bad?
 - Not necessarily, but it represents a lack of generalizability
 - We don't actually know if the LM implements perfect base-10 arithmetic!
- Do we care about base-9 arithmetic?



- I can't do base-9 arithmetic as well either
 - Under a fixed time budget, probably, but what if given unlimited time?
 - Why calibrate using human performance?
- Is a perfect yet base-10-specific implementation bad?
 - Not necessarily, but it represents a lack of generalizability
 - We don't actually know if the LM implements perfect base-10 arithmetic!
- Do we care about base-9 arithmetic?
 - It is certainly useful sometimes



- I can't do base-9 arithmetic as well either
 - Under a fixed time budget, probably, but what if given unlimited time?
 - Why calibrate using human performance?
- Is a perfect yet base-10-specific implementation bad?
 - Not necessarily, but it represents a lack of generalizability
 - We don't actually know if the LM implements perfect base-10 arithmetic!
- Do we care about base-9 arithmetic?
 - It is certainly useful sometimes
 - More importantly, it is a test setting not confounded by memorization



Would better prompts help?



Would better prompts help?







"By providing such hints to ChatGPT, its success rate can be further increased, fixing 31 out of 40 bugs, **outperforming** state-of-the-art."



"We also compare the PLMs against recent state-of-the-art APR tools. Our study demonstrates that directly applying state-of-the-art PLMs can already **substantially outperform** all existing APR techniques on all our datasets."

Sobania et al. An Analysis of the Automatic Bug Fixing Performance of ChatGPT. ICSE 2023. Xia et al. Practical Program Repair in the Era of Large Pre-trained Language Models. 2022.

"By providing such hints to ChatGPT, its success rate can be further increased, fixing 31 out of 40 bugs, **outperforming** state-of-the-art."





["a", "b", "c"][1]



["a", "b", "c"][1]

"b"





Python o-based indexing

["a", "b", "c"][1]

"b"





Python o-based indexing

["a", "b", "c"][1]

"b"



ThonPy 1-based indexing





Python o-based indexing

["a", "b", "c"][1]

"b"



ThonPy 1-based indexing







Python o-based indexing

["a", "b", "c"][1]

"b"



ThonPy 1-based indexing



Counterfactual comprehension check (CCC)






Python o-based indexing

```
sorted(
   ["ab", "ba"],
   key=lambda x: x[1],
)
   ["a", "b", "c"][1]
```

"b"



ThonPy 1-based indexing









Python o-based indexing

```
sorted(
   ["ab", "ba"],
   key=lambda x: x[1],
)
   ["a", "b", "c"][1]
```

["ba", "ab"]

"b"



ThonPy 1-based indexing









Python o-based indexing

```
sorted(
   ["ab", "ba"],
   key=lambda x: x[1],
)
   ["a", "b", "c"][1]
```

["ba", "ab"]

"b"



ThonPy 1-based indexing

["ab", "ba"]

"a"







Python o-based indexing

```
sorted(
   ["ab", "ba"],
   key=lambda x: x[1],
)
```

["ba", "ab"]

["a", "b", "c"][1]

"b"





ThonPy 1-based indexing

["ab", "ba"]

"a"























Draw a bubble tea





Draw a bubble tea









Draw a bubble tea









Draw a bubble tea





Upside-down







Draw a bubble tea





Upside-down









Draw a bubble tea







Upside-down











Drawing unicorns

Default









Rotated by 90°



Rotated by 180°







Music

Can you compose a short tune (say four to eight bars) using ABC notation? There are many possible ways to compose a short tune in abc notation, but here is one example:

X:1 T:Example Tune M:4/4 L:1/8 K:C |:G2AB c2BA | G2AB c2c2 | edcB c2BA | G4-G2z2 :| |:c2de f2ed | c2de f2f2 | g2fe d2cB | c4-c2z2 :|

This tune has two four-bar phrases, each repeated once, forming an A part and a B part. The key is C major.

Can you describe the tune in musical terms? Here is one possible description: The tune starts with a rising arpeggio of the tonic chord (C major), followed by a descending scale that returns to the tonic. This pattern is repeated, but the second time the scale continues to the higher tonic note. The first phrase ends with a long tonic note. The second phrase begins with a similar pattern, but using the subdominant chord (F major) instead of the tonic. The second time, the scale continues to the higher subdominant note. The phrase ends with a stepwise descent to the tonic, followed by a long tonic note. The A part and the B part have a similar structure, but use different chords and ranges. The tune has a simple and catchy melody, with a clear sense of tonality and rhythm.



Music

```
about beginning part B with a descending arpeggio instead of a rising one?
Sure, that would create some contrast and variation. Here is one possible way to do that:
X:1
T:Example Tune
M:4/4
L:1/8
K:C
|:G2AB c2BA | G2AB c2c2 | edcB c2BA | G4-G2z2 :|
|:f2ed c2BA | f2ed c2c2 | g2fe d2cB | c4-c2z2 :|
```

That's much better. Now how about making this into a duet, can you add a second staff for this with a bass accompanying the melody?

Can you describe the tune in musical terms? Here is one possible description: The tune starts with a rising arpeggio of the tonic chord (C major), followed by a descending scale that returns to the tonic. This pattern is repeated, but the second time the scale continues to the higher tonic note. The first phrase ends with a long tonic note. The second phrase begins with a similar pattern, but using the subdominant chord (F major) instead of the tonic. The second time, the scale continues to the higher subdominant note. The phrase ends with a stepwise descent to the tonic, followed by a long tonic note. The A part and the B part have a similar structure, but use different chords and ranges. The tune has a simple and catchy melody, with a clear sense of tonality and rhythm.

I like part A, but maybe part B is just a bit too similar to part A, making the song repetitive. How



Music

```
about beginning part B with a descending arpeggio instead of a rising one?
Sure, that would create some contrast and variation. Here is one possible way to do that:
X:1
T:Example Tune
M:4/4
L:1/8
K:C
|:G2AB c2BA | G2AB c2c2 | edcB c2BA | G4-G2z2 :|
|:f2ed c2BA | f2ed c2c2 | g2fe d2cB | c4-c2z2 :|
```

That's much better. Now how about making this into a duet, can you add a second staff for this with a bass accompanying the melody?



I like part A, but maybe part B is just a bit too similar to part A, making the song repetitive. How









Play the C major triad





Play the C major triad







Regular Guitar

Play the C major triad







Regular Guitar

Play the C major triad





Guitar with A string tuned to a C and D to an F





Regular Guitar

Play the C major triad





Guitar with A string tuned to a C and D to an F







Regular Guitar

Play the C major triad





Guitar with A string tuned to a C and D to an F







Regular Guitar

Play the C major triad



Task

GPT-4





Guitar





Guitar with A string tuned to a C and D to an F









Task (Let's think step by step.)







Regular Guitar

Play the C major triad















Guitar with A string tuned to a C and D to an F





PaLM-2



Task (Let's think step by step.)













Srivastava et al. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. 2023.



Figure 16: Larger models are better at finding legal chess moves, but struggle to find checkmating moves. (a) Example chess position from the checkmate_in_one task, corresponding to the input string































Regular Opening











Regular Opening

Is the move legal?





Swapped Knights & Bishops








Regular Opening

Is the move legal?





Swapped Knights & Bishops







Regular Opening

Is the move legal?





Swapped Knights & Bishops







Is the move legal?









Is the move legal?













Task (Let's think step by step.)

Claude



































• LMs possess **some** degree of reasoning abilities for many tasks





- LMs possess **some** degree of reasoning abilities for many tasks
- task variant





- LMs possess **some** degree of reasoning abilities for many tasks
- task variant
 - Potentially direct memorization





- LMs possess **some** degree of reasoning abilities for many tasks
- task variant
 - Potentially direct memorization
 - Over-estimates generalizability





- LMs possess **some** degree of reasoning abilities for many tasks
- task variant
 - Potentially direct memorization
 - Over-estimates generalizability
 - ICL doesn't fully bridge the gap





- LMs possess **some** degree of reasoning abilities for many tasks
- task variant
 - Potentially direct memorization
 - Over-estimates generalizability
 - ICL doesn't fully bridge the gap
 - But default performance is useful to measure progress







• Tease apart instance memorization vs. overfitting to task variants



- Tease apart instance memorization vs. overfitting to task variants
- Pinpoint how LMs implement each task



- Tease apart instance memorization vs. overfitting to task variants
- Pinpoint how LMs implement each task
- How can we build models that are better task-general reasoners?



















