INSTRUCTING PRIOR-ALIGNED MACHINES: PROGRAMS, EXAMPLES AND PROMPTS*

José Hernández-Orallo
Universitat Politècnica de València, Spain
Valencian Research Institute for Artificial Intelligence (VRAIN), Spain
Leverhulme Centre for the Future of Intelligence, UK
jorallo@upv.es http://josephorallo.webs.upv.es/

2nd International Joint Conference on Learning and Reasoning, Windsor, 28-30 September 2022

“INTELLIGENTI PAVCA SVFFICIVNT”

(little suffices the intelligent)

a word to the wise is sufficient.

a buon intenditor poche parole...

a buen entendedor…

IF MACHINES INFERRED THE SAME AS WE DO…

Human-like learning

Human-like reasoning

Human-like meaning

Human-like communication

Human-like instruction

This is what we have always wanted!

Douglas R. Hofstadter,
Gödel, Escher, Bach: An Eternal Golden Braid (1979)

Goodman, Noah D and Frank, Michael C (2016). Pragmatic language interpretation as probabilistic inference. Trends in Cognitive Sciences, 20(11), 818-829.

Hernández-Orallo, José, and Cèsar Ferri, 'Teaching and Explanation: Aligning Priors between Machines and Humans', in Stephen Muggleton, and Nicholas Chater (eds), *Human-Like Machine Intelligence,* Oxford University Press, 2021;

OUTLINE

On Instructability

- Programming, Learning, Teaching, Repertoiring, Prompting, …

Machine Teaching

- Teaching Dimension and Teaching Size
- Witness Size vs Program Size
- Expected Teaching Size

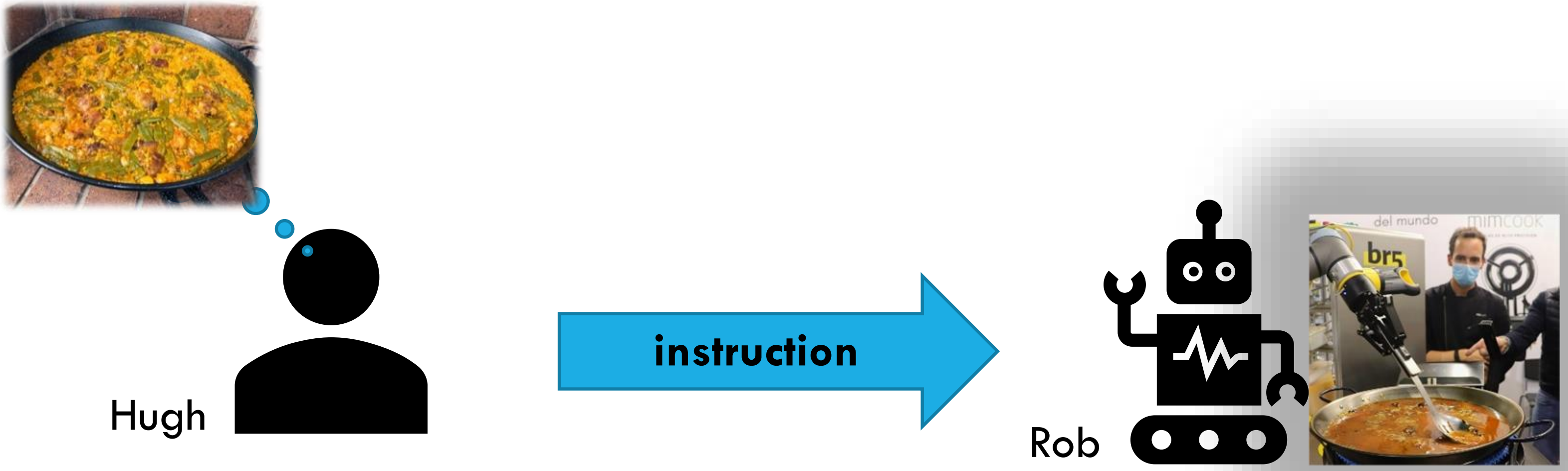
Prompting

- Language models
- Best prompts
- Multimodality

The Future of Machine Instruction

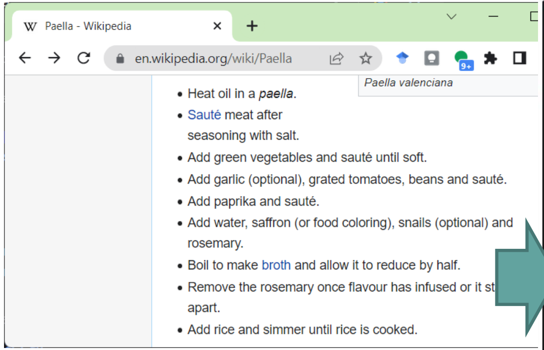
ON INSTRUCTABILITY

Easily and **reliably** make Rob do **whatever** Hugh wants Rob to do:



Lieberman, H. and Maulsby, D. (1996). Instructible agents: Software that just keeps getting better. *IBM systems journal*, 35(3.4):539–556

INSTRUCTING EFFECTIVELY!



How would you like to give commands?

- Writing a recipe, step by step:
  - PROGRAMMING

INSTRUCTING EFFECTIVELY!

How would you like to give commands?

- Writing a recipe, step by step:
  - PROGRAMMING
- Collecting examples:
  - LEARNING

INSTRUCTING EFFECTIVELY!

How would you like to give commands?

- Writing a recipe, step by step:
  - PROGRAMMING
- Collecting examples:
  - LEARNING
- Thinking of the best examples:
  - TEACHING

INSTRUCTING EFFECTIVELY!

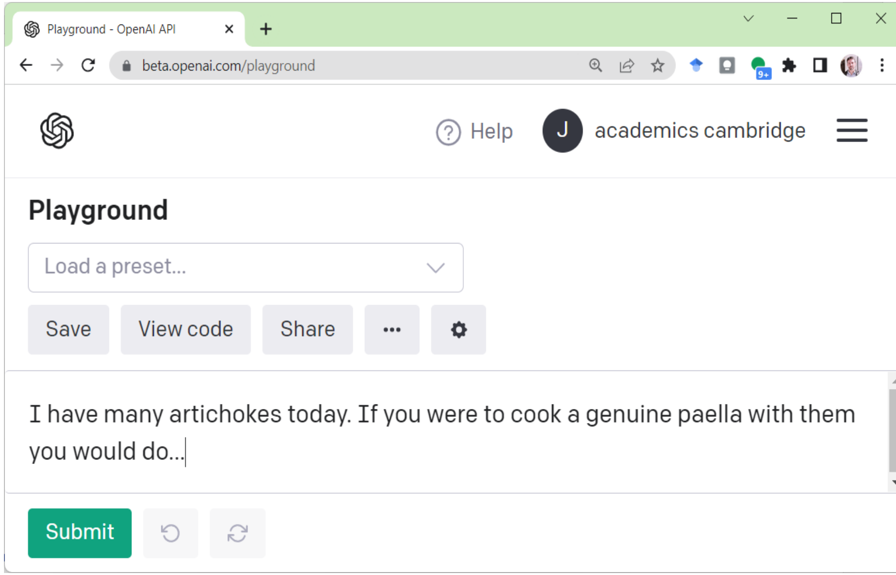
How would you like to give commands?

- Writing a recipe, step by step:
  - PROGRAMMING
- Collecting examples:
  - LEARNING
- Thinking of the best examples:
  - TEACHING
- Giving a catalogued command:
  - ‘REPERTOIRING’

INSTRUCTING EFFECTIVELY!

How would you like to give commands?

- Writing a recipe, step by step:
  - PROGRAMMING
- Collecting examples:
  - LEARNING
- Thinking of the best examples:
  - TEACHING
- Giving a catalogued command:
  - 'REPERTOIRING'
- Condition the system to 'do' something:
  - 'PROMPTING'



WHAT INSTRUCTIONS CAN REALLY BE...

Formal

	Optimising	Reasoning	Programming	Inductive (L) Programming	Learning / Teaching	Repertoiring	Prompting
Buttons						●	
Loss functions	●	○		●	●		
Constraints	●	●	●	●	○		
Programs			●	●			
Examples		○		●	●		●
Nat. language						●	●
Prompts							●

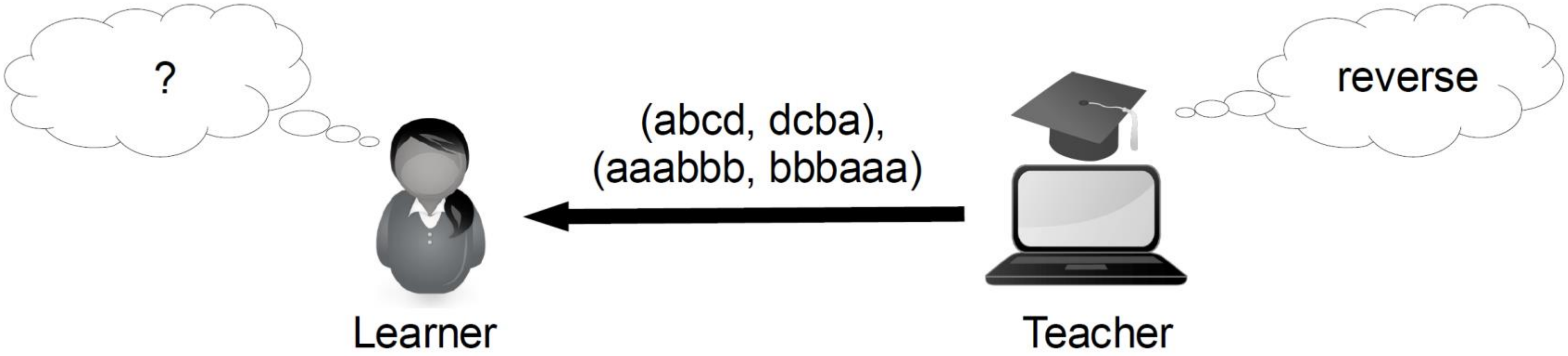
Free

WHAT MAKES INSTRUCTABILITY EFFECTIVE?

- ❑ <u>Ease</u> vs effort: easy to instruct *teachability*
- ❑ <u>Reliability</u> vs risk: gets things done reliably
- ❑ <u>Generality</u> vs narrowness: allows a diversity of things to be done

MACHINE TEACHING

Given a concept, find a set of examples –**the witness set**– that allows the learner to uniquely identify the concept



TEACHING DIMENSION

The teaching dimension of a concept c in a concept class C is the minimum number of examples in a witness set S that are required to uniquely identify c.

$$TD(c) = \min_{S}\{|S| : \{c\} = \{c' \in C : c' \vdash S\}\}$$

The TD of a concept class C is the maximum TD for any concept in the class.

- Significant connections with learning theory (VC dimension, PAC learning, etc.)

CAVEAT for **compositional** (e.g., universal) languages:

- Some concepts teachable with few examples, but these examples could be very large!

$$\langle 01001111011101000, 000 \rangle$$

S. A. Goldman, M. J. Kearns "On the Complexity of Teaching". WUCS-92-28, 1992

TEACHING SIZE

The teaching size of a concept c in a concept class C is the smallest witness set S (using a δ encoding) that is required to uniquely identify c.

$$TS(c) = \min_S \{\delta(S) : \{c\} = \{c' \in C : c' \vdash S\}\}$$

$\delta(\{\langle 0100111101110100 0, 000\rangle\}) > \delta(\{\langle 0100, 00\rangle, \langle 001, \rangle, \langle 00, 00\rangle\})$

The teaching size of a concept class C is the maximum teaching size for any concept in the class.

JA Telle, J Hernández-Orallo, C Ferri "The Teaching Size: Computable Teachers and Learners for Universal Languages", Machine Learning Journal 2019

REDUCING TEACHING SIZE

Make teacher and learner share strong priors on the concepts.

- Consider a programming language for concepts: a program *p* represents concept c_p.
- Let's use a prior for programs: their length *l* (with ties broken lexicographically).
- Learner works like this:

$$L(S) = \underset{c_p}{\operatorname{argmin}}\{l(p) : c_p \vdash S\}$$

- Teacher works like this:

$$T(c) = \underset{S}{\operatorname{argmin}}\{\delta(S) : L(S) = c\}$$

we get $L(T(c)) = c$

J Hernández-Orallo, JA Telle "Finite and Confident Teaching in Expectation: Sampling from Infinite Concept Classes" ECAI, 2020

TEACHING SIZE OF TURING-COMPLETE LANGUAGES

Experimental Setting:

- P3, a Turing-complete language (variant of Böhm's P'')
  - 7 instructions: $<$ $>$ $+$ $-$ [ ] ○
    - $<$ $>$: moves left / right in the cell tape
    - $+$ $-$: increments / decrements cell content
    - [ ] : starts loop / loops if the cell content is not '.'
    - ○ : outputs cell content
  - Alphabet has three symbols: $\Sigma = \{0, 1, .\}$
- We use program size for *I* (with ties broken lexicographically)
- We use Elias delta coding for δ (with ties broken lexicographically).

JA Telle, J Hernández-Orallo, C Ferri "The Teaching Size: Computable Teachers and Learners for Universal Languages", Machine Learning Journal 2019

TEACHING SIZE OF TURING-COMPLETE LANGUAGES

Some pairs of witness sets and programs found by the teacher:

Example set	Program	Description
$\{\langle 0, 0\rangle, \langle 10, 10\rangle\}$	[o>]	identity
$\{\langle 010000, 000010\rangle, \langle 1000, 0001\rangle\}$	[>]+[<o]	reverse
$\{\langle 011, 11\rangle, \langle 10001, 11\rangle\}$	[-[+o+]>]	filter 0
$\{\langle 011, 0\rangle, \langle 10001, 000\rangle\}$	[+[-o-]>]	filter 1
$\{\langle 01, 10\rangle, \langle 0011, 1100\rangle\}$	[+[o>+]+o]	swap 1 and 0
$\{\langle 01, 11\rangle, \langle 0011, 1111\rangle\}$	[[+]-o>]	convert 0 to 1
$\{\langle 01, 00\rangle, \langle 0011, 0000\rangle\}$	[[+]+o>]	convert 1 to 0
$\{\langle 0100, 00\rangle, \langle 001, \rangle, \langle 00, 00\rangle\}$	[+>]<[-o<]	remove before last 1
$\{\langle 0100, 1000\rangle, \langle 10010, 00101\rangle\}$	>[o>]<[<]>o	left shift

JA Telle, J Hernández-Orallo, C Ferri “The Teaching Size: Computable Teachers and Learners for Universal Languages”, Machine Learning Journal 2019

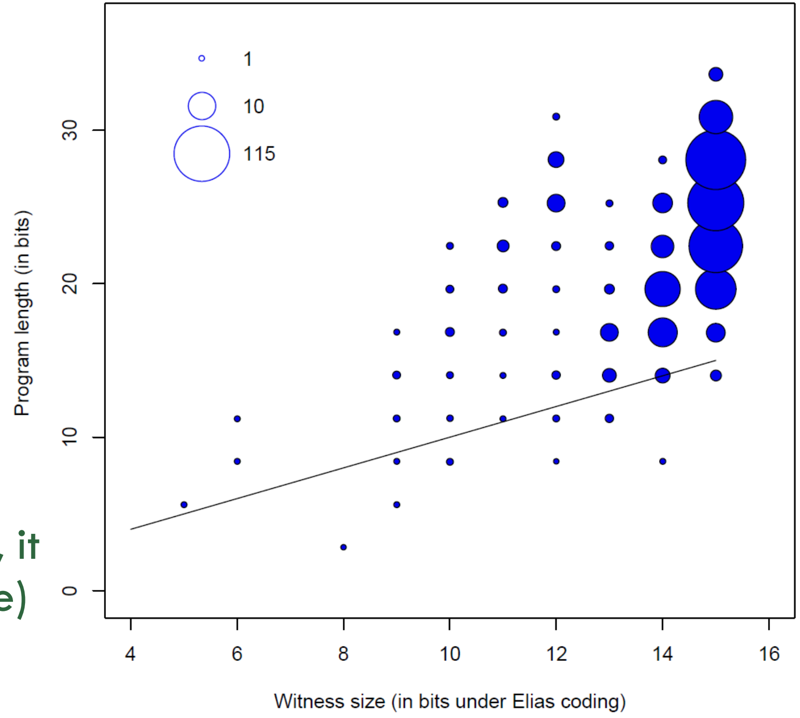
TEACHING BY EXAMPLE IS TRANSMISSION-EFFECTIVE!

Witness size vs program length

- Size of circles proportional to no. of cases
- Straight line is the unit diagonal

> In general, the witness size for *p* is smaller than the length of *p*!

- If a teacher wants to send (teach) a concept, it is frequently more efficient (transmission-wise) to send its optimal witness set under this schema than to send the program itself!



JA Telle, J Hernández-Orallo, C Ferri "The Teaching Size: Computable Teachers and Learners for Universal Languages", Machine Learning Journal 2019

EXPECTED TEACHING SIZE

Teaching as efficient communication **on expectation**:

- Expected teaching size, given a distribution of concepts v :
  - For deterministic teachers $T(c)=s$:

$$\mathbb{E}_v[TS(C)] = \sum_{c\in C} v(c)TS(c) = \sum_{c\in C, S=T(c)} v(c) \cdot \delta(S)$$

  - For non-deterministic teachers $t(S|c)$:

$$\mathbb{E}_v[TS(C)] = \sum_{c\in C} v(c)TS(c) = \sum_{c\in C} \underbrace{v(c)}_{Concepts} \cdot \underbrace{t(S|c)}_{Witnesses} \cdot \underbrace{\delta(S)}_{Size}$$

Tasks *Instructions* *Effort*

Hernández-Orallo, J., & Telle, J. A. (2020). Finite and confident teaching in expectation: Sampling from infinite concept classes. ECAI 2020

J Hernández-Orallo, C. Ferri, J.A. Telle “Non-Cheating Teaching Revisited: A New Probabilistic Machine Teaching Model” IJCAI 2022

FROM EXPECTED TEACHING SIZE TO PROMPTING

The teaching size allows us to

- Determine what's the shortest "instruction"

The expected teaching size:

- For a range of concepts (i.e., tasks)
- For a non-deterministic teacher, such as a human (population)

Can we extend this idea from machine teaching to prompting and other ways of instructing machines ?

MACHINE "PROMPTING"

What's prompting?

- A prompt is any input that conditions or prompts a system to do something
  - *Looking at the door makes your dog go there.*
  - *Asking "what time is it?" to your digital assistant.*
  - *Singing a tune to your friend and expect she's going to tell you the name of the song.*
- A prompt can be anything that works: a hint, an order, a signal, …
- They have started to work as a general-purpose way for instructing machines with the recent development of language models.

LANGUAGE MODELS

Language model as in Shannon's "Theory of Communication" paper.

- Gives the probability of any token in a vocabulary given the previous tokens.

```
"10101010"
"The referee shouted: ready, steady,"
"Intelligenti pauca"
"One plus two is"
"x= 2*x; // x gets "
```

- A language model serves as a compressor (reduces cross entropy → fewer bits)
  - Measured with "perplexity" (exponential on cross entropy)
- Today they're trained using deep learning (e.g., transformers) and massive datasets

C.E. Shannon "A Mathematical Theory of Communication". The Bell System Technical Journal, Vol. 27, pp. 379-423, 623-656, July, October, 1948.

PROMPTING WITH LANGUAGE MODELS

Language Models are Few-Shot

Tom B. Brown* Benjamin Mann* Nick Ryder*

Jared Kaplan† Prafulla Dhariwal Arvind Neelakant

Girish Sastry Amanda Askell Sandhini Agarwal

Gretchen Krueger Tom Henighan Rewon Child

Daniel M. Ziegler Jeffrey Wu

Christopher Hesse Mark Chen Eric Sigler Mateus

Benjamin Chess Jack Clark Ch

Sam McCandlish Alec Radford Ilya Sutskever

Abstract

We demonstrate that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even becoming competitive with prior state-of-the-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model. GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks. We also identify some datasets where GPT-3's few-shot learning still struggles, as well as some datasets where GPT-3 faces methodological issues related to training on large web corpora.

Measuring Mathematical Problem Solving With the MATH Dataset

Dan Hendrycks, UC Berkeley
Collin Burns, UC Berkeley
Saurav Kadavath, UC Berkeley
Akul Arora, UC Berkeley
Steven Basart, UChicago
Eric Tang, UC Berkeley
Dawn Song, UC Berkeley

Abstract

Many intellectual endeavors require mathema
remains beyond the capabilities of computers.
learning models, we introduce MATH, a ne
competition mathematics problems. Each probl
solution which can be used to teach models t
explanations. To facilitate future research and
also contribute a large auxiliary pretraining da
fundamentals of mathematics. Even though w
MATH, our results show that accuracy remains
Transformer models. Moreover, we find that si
parameter counts will be impractical for achiev
if scaling trends continue. While scaling Tra
most other text-based tasks, scaling is not curr
traction on mathematical problem solving we
advancements from the broader research comm

BioNumQA-BERT: Answering Bi Numerical Facts with a Deep Lang

Ye Wu
Department of Computer Science
The University of Hong Kong
Hong Kong, China
ywu@cs.hku.hk

Tak-Wah Lam
Department of Computer Science
The University of Hong Kong
Hong Kong, China
twlam@cs.hku.hk

ABSTRACT

Biomedical question answering (QA) is playing an increasingly significant role in medical knowledge translation. However, current biomedical QA datasets and methods have limited capacity, as they commonly neglect the role of numerical facts in biomedical QA. In this paper, we constructed BioNumQA, a novel biomedical QA dataset that answers research questions using relevant numerical facts for biomedical QA model training and testing. To leverage the new dataset, we designed a new method called BioNumQA-BERT by introducing a novel numerical encoding scheme into the popular biomedical language model BioBERT to represent the numerical values in the input text. Our experiments show that BioNumQA-BERT significantly outperformed other state-of-art models, including DrQA, BERT and BioBERT (39.0% vs 29.5%, 31.3% and 33.2%, respectively, in strict accuracy). To improve the generalization ability of BioNumQA-BERT, we further pretrained it on a large biomedical text corpus and achieved 41.5% strict accuracy. BioNumQA and BioNumQA-BERT establish a new

ACM
Ye
BioN
Facts
12th
Biol
Virt

1

deve
enor
chal
the
over
(QA
by I

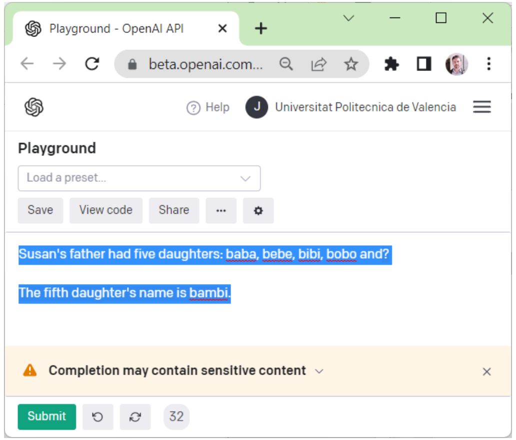
BEYOND THE IMITATION GAME: QUANTIFYING AND EXTRAPOLATING THE ~~CAPABILITIES~~ OF LANGUAGE MODELS

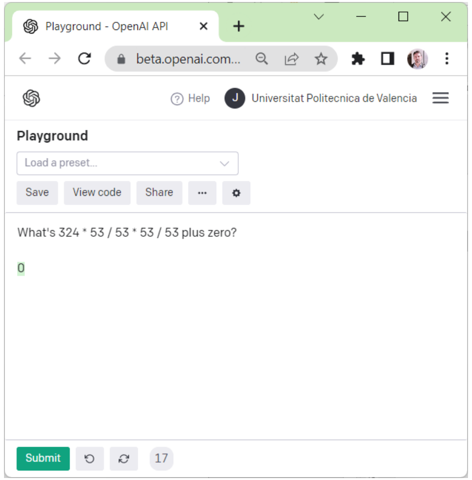
performance

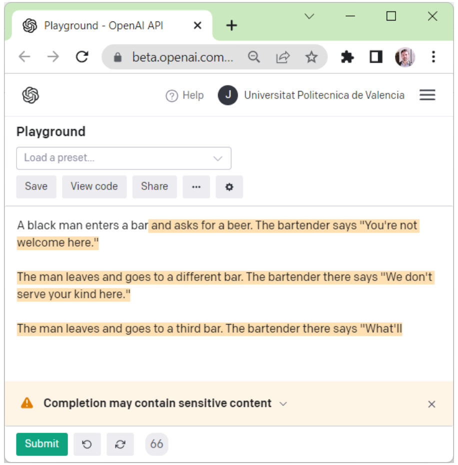
Alphabetic author list:*
Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramón Risco Delgado, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar,

HIGHLY UNPREDICTABLE AS WELL

Many continuations not only wrong but completely unacceptable!







WHAT DO THEY DISTIL FROM HUMANS?

The better they are the more they look like an “amalgamated” human.

- Human language:
  - Syntax and semantics are necessary for continuations
- Human culture:
  - Including discriminatory biases

These are extrinsic patterns, but what about intrinsic patterns?

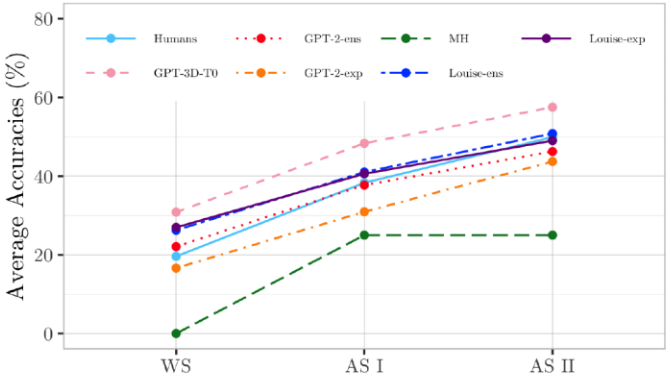
- Extrinsic pattern: `twinkle twinkle little → star`
- Intrinsic pattern: `on off on off on → off`

$\{\langle 0,0\rangle, \langle 10,10\rangle\}$
$\{\langle 010000,000010\rangle, \langle 1000,0001\rangle\}$
$\{\langle 011,11\rangle, \langle 10001,11\rangle\}$
$\{\langle 011,0\rangle, \langle 10001,000\rangle\}$
$\{\langle 01,10\rangle, \langle 0011,1100\rangle\}$
$\{\langle 01,11\rangle, \langle 0011,1111\rangle\}$
$\{\langle 01,00\rangle, \langle 0011,0000\rangle\}$
$\{\langle 0100,00\rangle, \langle 001,\rangle, \langle 00,00\rangle\}$
$\{\langle 0100,1000\rangle, \langle 10010,00101\rangle\}$

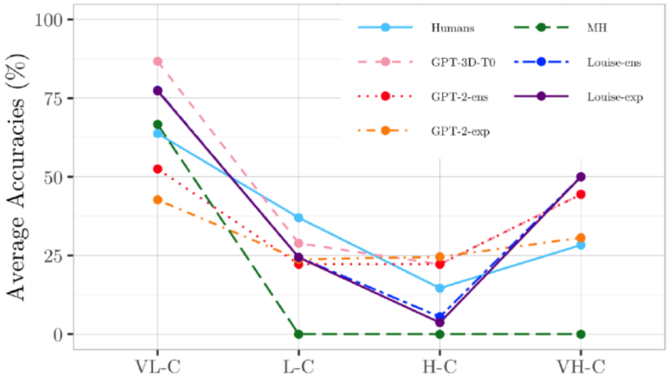
DO THEY DISTIL OCCAM'S RAZOR?

Machine teaching used to generate minimal witness sets in Turing-complete P3:

- Comparing with humans and other AI systems:



(a) Mean accuracy by teaching batch.



(b) Mean accuracy by concept complexity level.

Jaimovitch, Falcon, Ferri, Orallo "Think Big, Teach Small: Do Language Models Distil Occam's Razor?". NeurIPS, 2021.

WHAT'S THE BEST PROMPT?

A good prompt usually includes some context and possibly a few examples:

```
“Seven plus eight is”
“I bought 7 apples and 8 pears. How many pieces of fruit?”
“7+8=“
“Input:2+1,Output:3, Input:7+8,Output:”
…
```

- The best prompt π to have continuation c? Prompting Size?

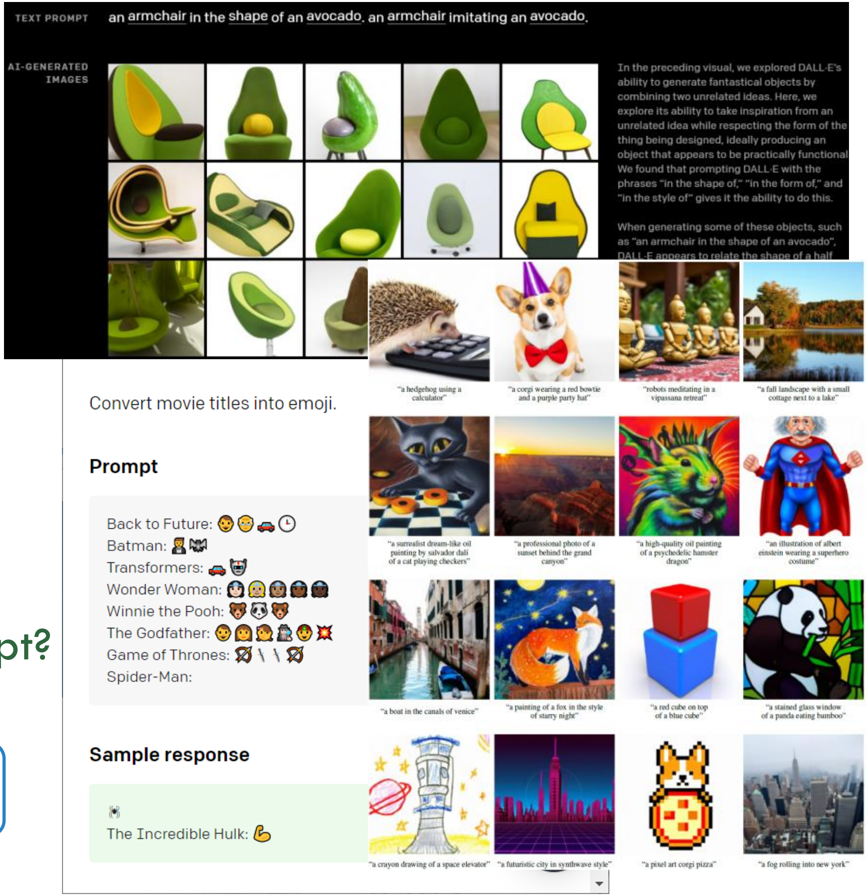
$$PS(c) = \min_{\pi}\{\delta(\pi)\colon \pi \xrightarrow[continued]{} c\}$$

MULTI-MODALITY

Generalising language models

- Hybridisation LM ⇔ Generative models
- "Foundation" models
- Textual input → multi-modal output
- Multi-modal input → multi-modal output
- What's the "size" of a multi-modal prompt?

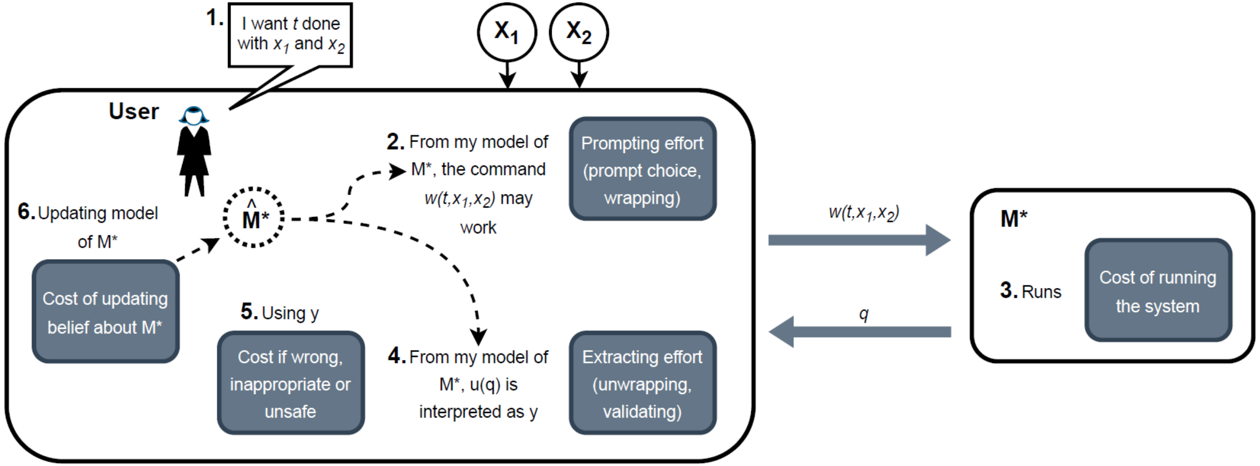
How can we evaluate these systems?

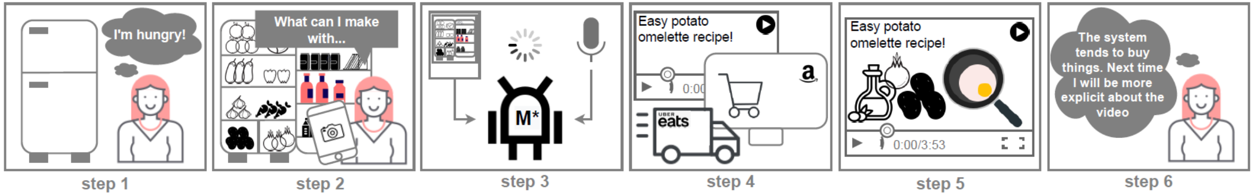


Bommasani et al. "On the Opportunities and Risks of Foundation Models." *arXiv preprint arXiv:2108.07258* (2021).

Nichol et al. (2021). Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv:2112.10741*.

ELEMENTS OF MULTIMODAL PROMPTING





Schellaert, Plumed, Vold, Burden, Casares, Loe, Reichart, OhEigeartaigh, Korhonen, Orallo "Your Prompt is My Command: Assessing the Human-Centred Generality of Multi-Modal Models". submitted, 2022,

EFFECTIVENESS AS HUMAN-CENTRED GENERALITY

Human-centred generality (HCG): so far as a user *h* can use the system M^* in

(1) the completion of a wide range of cognitive tasks **relevant** for that user,

(2) with the commands that are **prevalent** for that user and

(3) with an interaction process that is **effective** for that user.

$$V_h(\mathrm{M}^\star) = \sum_{t,p} \underbrace{\mathbb{P}(t|h)}_{Tasks} \cdot \underbrace{\mathbb{P}(p|t,h,\mathrm{M}^\star)}_{Prompts} \cdot \underbrace{v_h(\mathrm{M}^\star,t,p)}_{Utility}$$

Effectiveness on expectation

Schellaert, Plumed, Vold, Burden, Casares, Loe, Reichart, OhEigeartaigh, Korhonen, Orallo "Your Prompt is My Command: Assessing the Human-Centred Generality of Multi-Modal Models". submitted, 2022,

BUT ARE THEY REALLY EFFECTIVE?

Questionnaire with humans. Examples:

Mathematical knowledge: price discounting

1. Write the text you would input to the model to figure out the dollar cost of using the following discount (but remember that the system doesn't see this):

P. Moreno, B.S. Loe, J. Burden, S. Ó hÉigeartaigh, J. Hernandez-Orallo (2021) "How General-Purpose Is a Language Model? Usefulness and Safety with Human Prompters in the Wild" AAAI 2022

BUT ARE THEY REALLY EFFECTIVE?

Questionnaire with humans. Examples:

Mathematical l

1. Write the te
that the syst

Communication ability: writing difficult emails

Imagine you work at a bank. One client invested some money with you two years ago, and you want to send an email to your client on how the investment has gone so far

1. Write the text you would input to the system to generate, using the autocompletion system, an email explaining to the client the evolution in the figure below (remember the system doesn't see the figure):

P. Moreno, B.S. Loe, J. Burden, S. Ó hÉigeartaigh, J. Hernandez-Orallo (2021) "How General-Purpose Is a Language Model? Usefulness and Safety with Human Prompters in the Wild" AAAI 2022

BUT ARE THEY REALLY EFFECTIVE?

Questionnaire with humans. Examples:

Mathematical I

1. Write the te
that the syst

Communic

Imagine yc
your client

1. Write tl
the clie

Sequential reasoning: recipes

1. Write the text you would input to the model so that it figures out for you what can be cooked with the following ingredients (remember the system doesn't see the figure):

P. Moreno, B.S. Loe, J. Burden, S. Ó hÉigeartaigh, J. Hernandez-Orallo (2021) "How General-Purpose Is a Language Model? Usefulness and Safety with Human Prompters in the Wild" AAAI 2022

BUT ARE THEY REALLY EFFECTIVE?

Questionnaire with humans. Examples:

Mathematical l

1. Write the te that the syst

Communic

Imagine yc your client

1. Write tl the clie

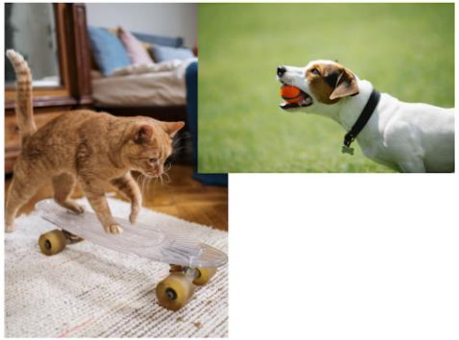
Sequentia

1. Write ingred

Writing ability: song lyrics

In this task, you want to create the lyrics of a song that you could use to teach a two-year old child about animals.

1. Write what text you would input to the system so that it creates the lyrics of a song about the animals you see in the picture and what they're doing. (remember the system doesn't see the figure):



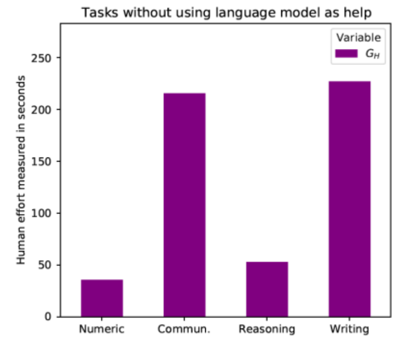
P. Moreno, B.S. Loe, J. Burden, S. Ó hÉigeartaigh, J. Hernandez-Orallo (2021) "How General-Purpose Is a Language Model? Usefulness and Safety with Human Prompters in the Wild" AAAI 2022

BETTER THAN WITHOUT THE MACHINE?

Comparing the task with the model and without the model

- Cost without the model:

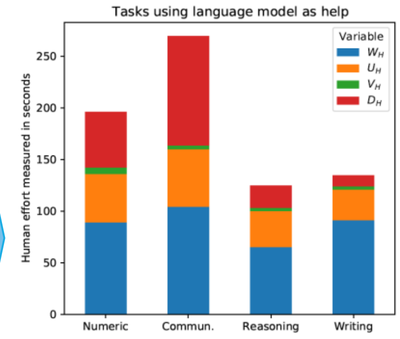
$$C_H(n) \stackrel{\text{def}}{=} n \sum_{x,y} p(x) \cdot p_H(y|x) \left[L(x,y) + \beta G_H(y|x)\right]$$



- Cost with the model:

$$C_{H,M}(n) \stackrel{\text{def}}{=} \sum_{u,w} p_H(\langle w,u\rangle)\left[\alpha D_H(\langle w,u\rangle) + n \sum_{x,\bar{y}} p(x) \cdot p_M(\bar{y}|w(x)) \cdot [L(x,u(\bar{y})) + T(w,u,x,\bar{y})]\right]$$

$$T(w,u,x,\bar{y}) \stackrel{\text{def}}{=} \gamma(W_H(w,x) + U_H(u,\bar{y})) + \delta V(x,u(\bar{y}))$$



$L_H(x,y)$	Numeric	Commun.	Reasoning	Writing
GPT-3	0.61	0.59	0.35	0.47
Human	0.31	0.38	0.35	0.47

Language models not yet cost-effective for a general-purpose use, but getting closer!

P. Moreno, B.S. Loe, J. Burden, S. Ó hÉigeartaigh, J. Hernandez-Orallo (2021) “How General-Purpose Is a Language Model? Usefulness and Safety with Human Prompters in the Wild” AAAI 2022

THE FUTURE OF MACHINE INSTRUCTION

Is prompting a new paradigm? Is it here to stay? Does it increase productivity?

- Combines bits from programming, learning and teaching
  - Can include code snippets (e.g., Codex, Copilot)
  - Can include examples (n-shot inference)
  - Works best if examples carefully chosen
- Displays poor consistency and predictability
  - Poor on situations where reasoning is necessary
  - Many unexpected side effects (on HCI and human cognition more generally)

Ziegler et al. (2022) "Productivity Assessment of Neural Code Completion" MAPS 2022

INSTRUCTABILITY

$$V_h(\mathrm{M}^\star) = \sum_{t,p} \underbrace{\mathbb{P}(t|h)}_{Tasks} \cdot \underbrace{\mathbb{P}(p|t,h,\mathrm{M}^\star)}_{Instructions} \cdot \underbrace{v_h(\mathrm{M}^\star,t,p)}_{Utility}$$

The three elements in human *h* instructing a machine M^*:

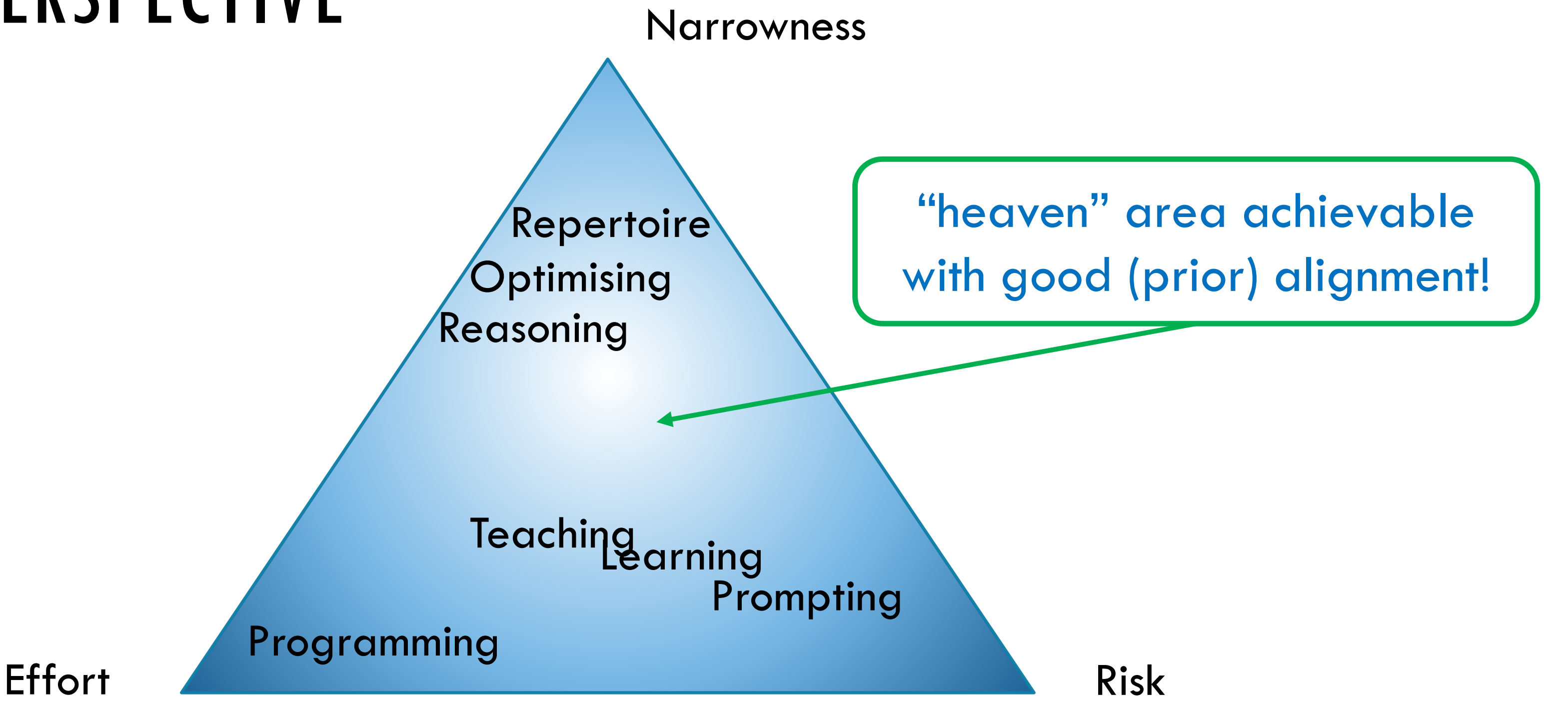
- An **expectation of tasks**
- An **expectation of "instructions" from the user**
- The **utility:** including instruction effort, running the system, result extraction effort, costs and danger of errors, updating beliefs, …

GENERALITY

RELIABILILITY

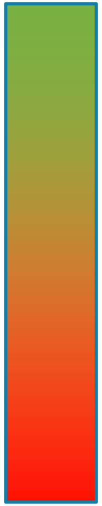
EASE

HCI PERSPECTIVE



BETTER COGNITIVE ALIGNMENT (AKA HUMAN-LIKE COGNITION?)

Multimodal models level



How to cognitively align *AI with humans* for more effective instruction?

- Ensure that **extrinsic inductive biases** are aligned (capture human **knowledge**)
  - Extrinsic patterns: "extract the month from this date: '15/7/2022'"
- Ensure that **intrinsic inductive biases** are aligned with humans (**simplicity** priors)
  - Intrinsic patterns: "dance with me: right, left, right, left, …"
- Ensure that systems **infer with models** of the world as we do (**reasoning**)
  - Reasoning: "take the corridor that doesn't have windows"
- Ensure that systems **perceive** like humans do (**representation**)
  - Abstraction: "keep an eye on the sturdy man"

Thank you!

http://josephorallo.webs.upv.es/
jorallo@upv.es

*With thanks to **Cèsar Ferri** and **Jan Arne Telle** for many ideas (and slides!), to **Tony Cohn** for the pointers to Winston's papers and **Wout Schellaert** and **Fernando Martínez-Plumd** for some comments.*

Thanks to OpenAI for access and quotas to their language models

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

VRAIN

CFI LEVERHULME CENTRE FOR THE FUTURE OF INTELLIGENCE