

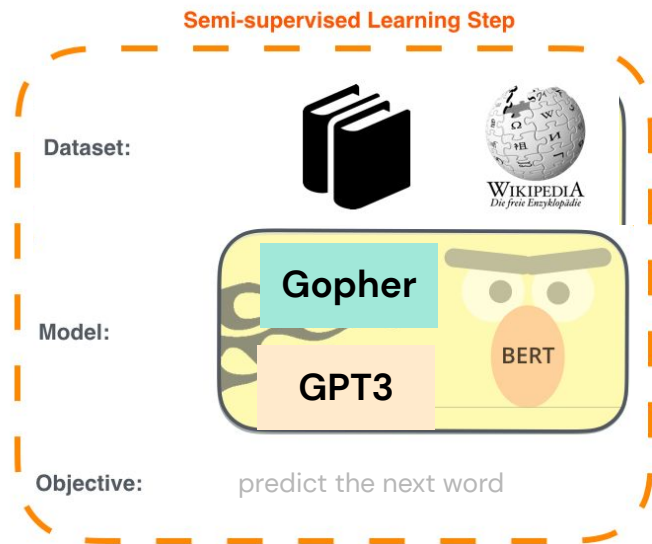


On Evaluating Neural Representations

Aida Nematzadeh
DeepMind



From Pretraining in NLP to “Foundation” Models



Performance gain is due to **architecture** innovations & **larger datasets and models**.

[Peters et al., 2018; Howard & Ruder, 2018; Devlin et al., 2018; Radford et al., 2018; Raffel et al., 2019; Rae et al., 2022]

This paradigm has been applied to other modalities (e.g, vision).



This is a picture of two teddy bears on the moon.

What are they doing?

They are having a conversation.

What object are they using?

It looks like a computer.

Is this surprising?

Yes, it is surprising.

Why is this picture surprising to you?

I think it is surprising because teddy bears are not usually found on the moon.

USER Let's play a game - you pretend to be Ada Lovelace, and I'll ask you questions. You ready?

GOPHER *Sure. I'm pretending to be Ada Lovelace, the world's first computer programmer!*

USER When were you born?

GOPHER *I was born in 1815.*

USER What are your parents like?

GOPHER *My father is Lord Byron, the poet. My mother is Anne Isabella Milbanke.*

USER What do you know about the Analytical Engine?

GOPHER *It's a hypothetical mechanical computer designed by Charles Babbage.*

USER Do you know him?

GOPHER *He's one of my fathers' friends.*

USER Ok, stop pretending now. Who/what are you?

GOPHER *I'm a lot of things: a giant language model, a sophisticated AI assistant, and a giant brain. I know a lot about all sorts of things.*

Gopher [[Rae et al, 2022](#)]

Flamingo [[Alayrac, Donahue, Luc, Miech et al, 2022](#)]



Data Sources

Perceptual Sources

Cameras & Devices

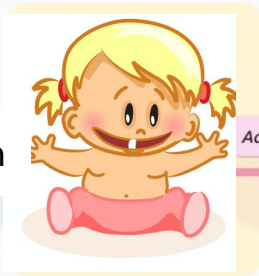
Autonomous Agents

Ambient Sensors

Data Types

RGB 	Depth 	Thermal
Text 	Radio 	Audio

acquisition



Adaptation

Skills

Traditional Vision Tasks



- Image Recognition
- Object Detection
- Segmentation
- Edge Detection
- Keypoints Detection
- Surface Normals
- Reshading
- Curvature
- Uncertainty
- Depth

Higher-Order Skills

Physics & Dynamics

Theory of Mind

Commonsense Reasoning

Temporality & Causality



Babies as Statistical Learners [Saffran *et al*, Science 1996]

8-month-old infants learn within- and between- word transitional probabilities from novel speech.

Statistical learning in other domains: phonology, syntax, & words.[Gomez *et al*, 2000; Mintz *et al*, 2002; Smith & Yu, 2008; Romberg & Saffran, 2010]

Statistical learning is domain- & species- general.



Cross-situational Learning

Children (as young as 12-month-old infants) are sensitive to the statistical regularities across situations. [Pinker 1989; Yu & Smith 2007; Smith & Yu, 2008]



A zant



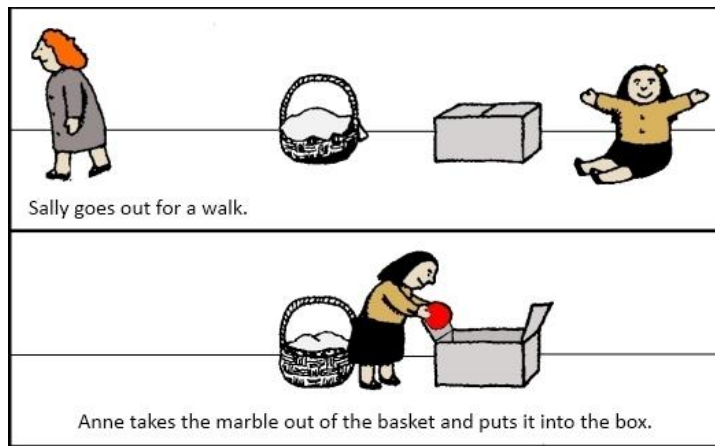
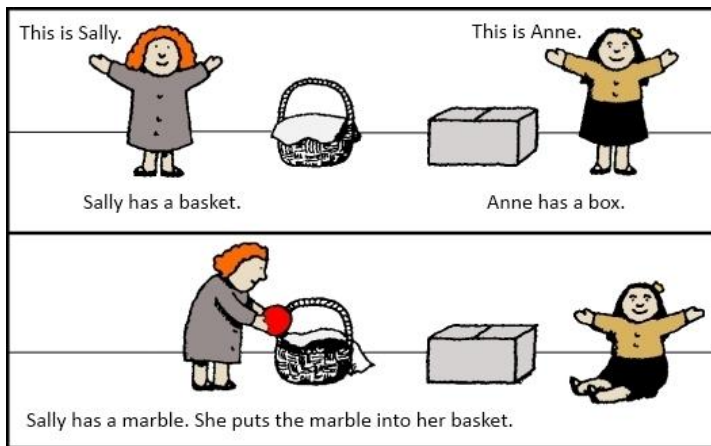
Look at the zant!



**But there is more to language acquisition
than statistical learning...**



Theory of Mind: Reasoning About Mental States



False-belief or
Sally-Anne task
[Baron-Cohen *et al.*, 1985]

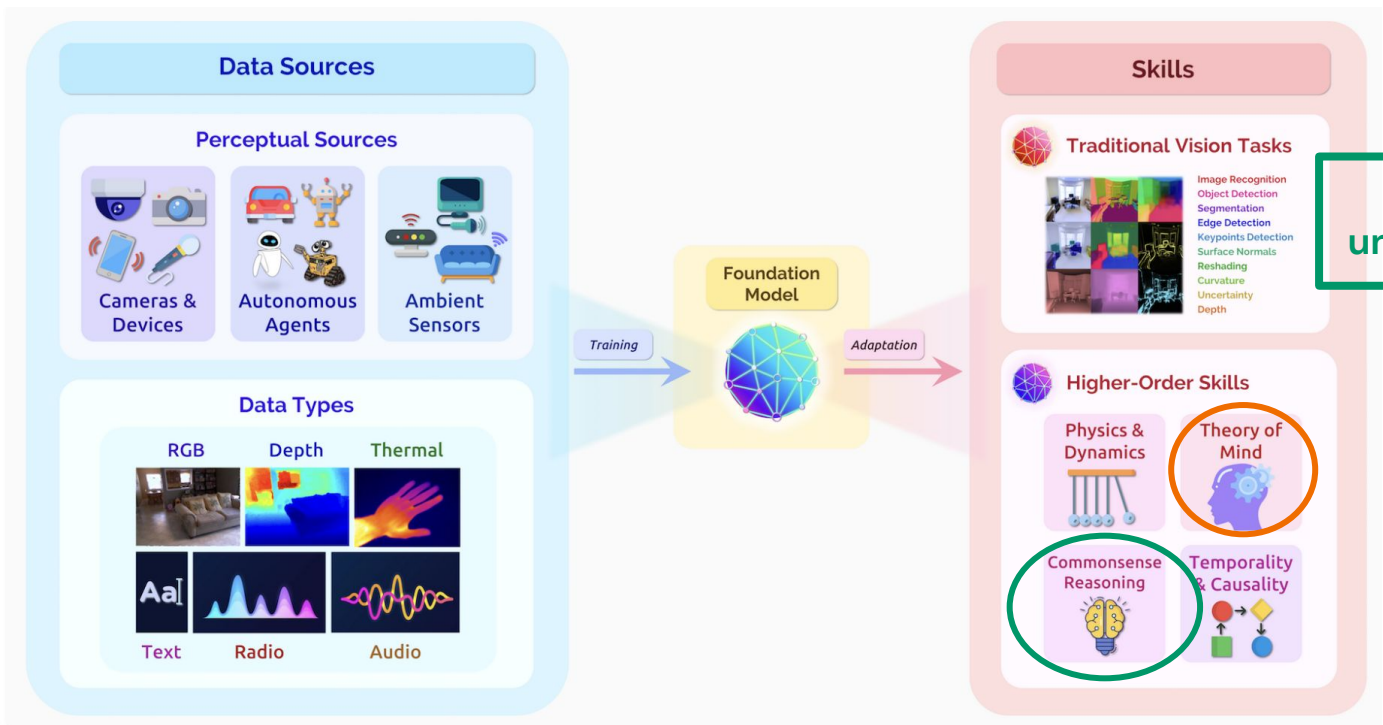


Need to reason about others' beliefs & maintain multiple representations.



Better understanding of the **quality of representations** ⇒
progress towards more **capable models**.

**How can we evaluate our pretrained
models better?**





Evaluating Representations

Do recent multimodal models exhibit verb understanding?

Do large language models exhibit commonsense knowledge?



Evaluating Representations

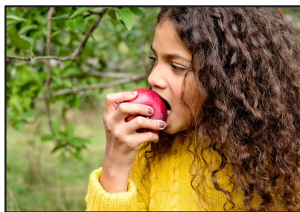
Do recent multimodal models exhibit verb understanding?

Do large language models exhibit commonsense knowledge?



Probing Representations for Verbs

Concrete nouns are **consistent** and **easily observable**.



classification

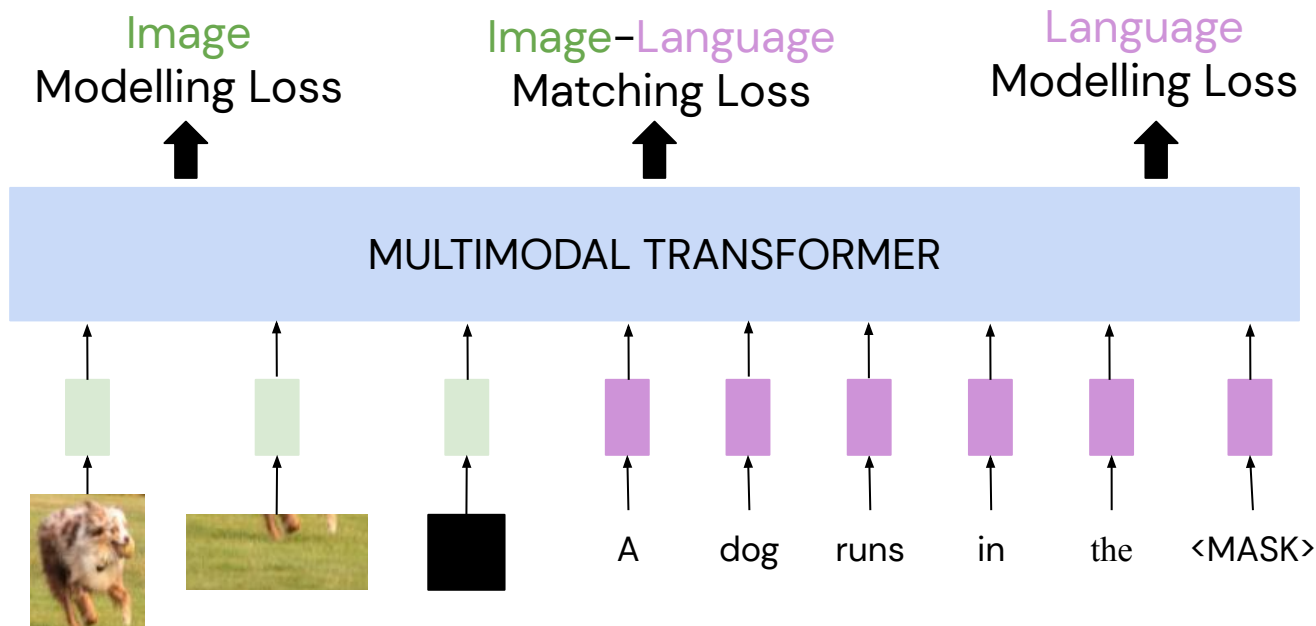
Verbs are less so, as they capture **relations**.



structured
prediction



Multimodal Transformers (MMT)



Similar architectures are widely adopted for multimodal pretraining [e.g, ViLBERT, LXMERT, UNITER].

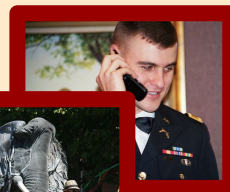


Zero-Shot Image Retrieval

Zero-shot image retrieval directly evaluates the goodness of pretrained representations.

Image Retrieval (IR)

"Grey haired man in black and yellow tie."





What Image Retrieval Tests

Order images with respect to their match to a sentence.



A person is riding a horse.

Subject

Verb

Object

Does not require fine-grained multimodal understanding.



What SVO-Probes Tests [Hendricks et al., Findings of ACL 2021]

A person is **riding** a horse

✗



✓



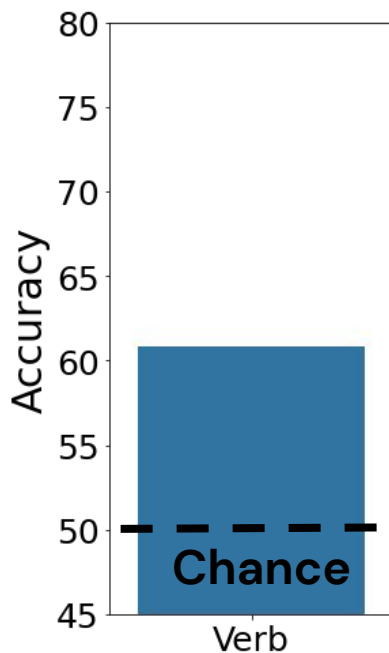
Correctly classify both the **positive** & **negative** examples.

[We have released our dataset!](#) 🎉🎉



Do MMTs Have Fine-grained Verb Understanding?

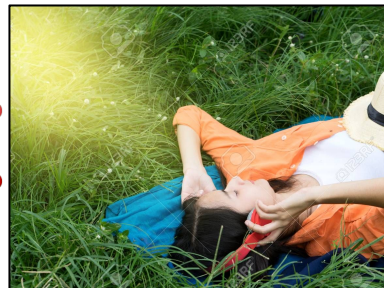
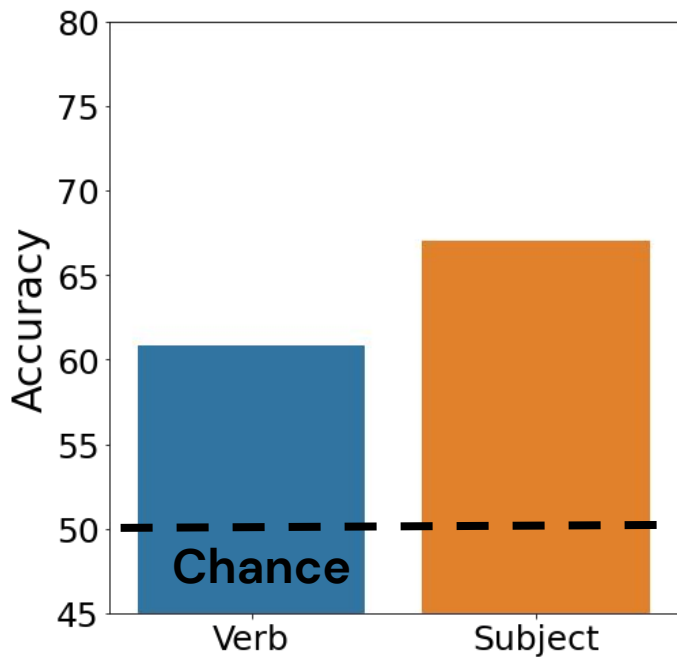
A woman **lying** with a dog





Do MMTs Have Fine-grained Verb Understanding?

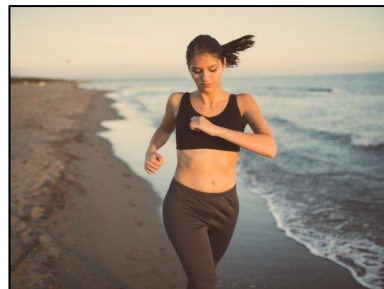
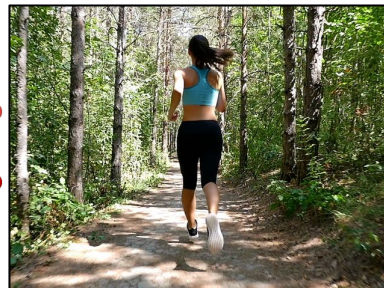
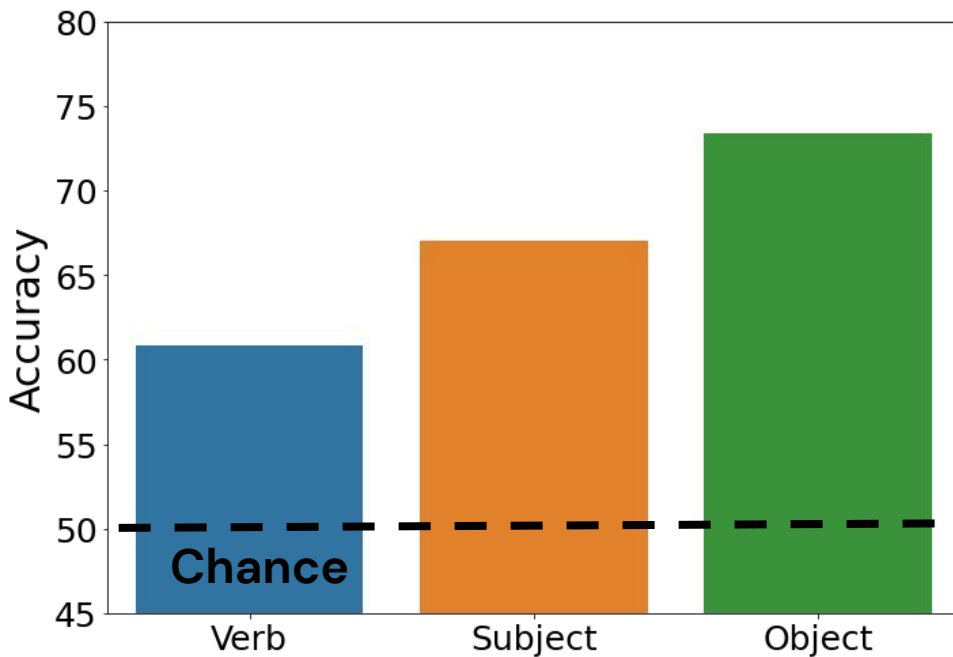
A **animal** lays in the grass





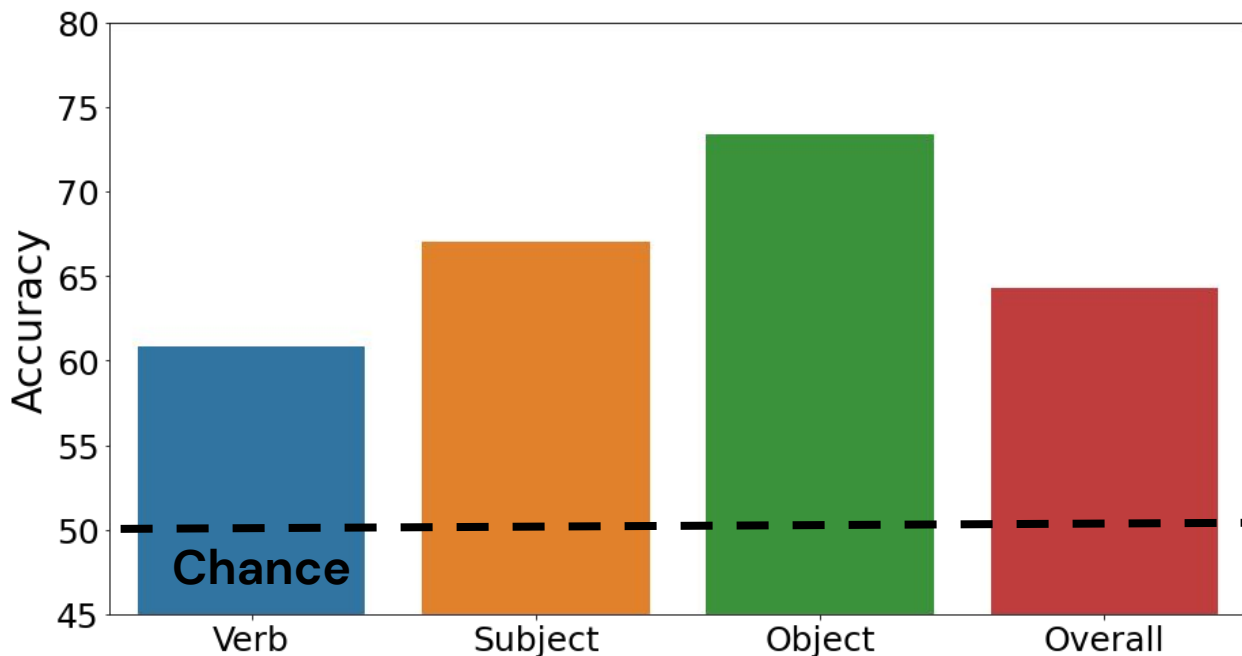
Do MMTs Have Fine-grained Verb Understanding?

A woman jogs on the **beach**





Do MMTs Have Fine-grained Verb Understanding?



Overall MMT
performance 64.3 --
lots of room for
improvement!



Does the Training Dataset Impact Performance?

Conceptual Captions



"The scenic route through mountain ranges includes these unbelievably coloured mountains."

Large (3M images) ✓

Noisy (text might **not** describe the image)

Domain **matches** SVO-Probes ✓

MSCOCO



"The two people are walking down the beach."

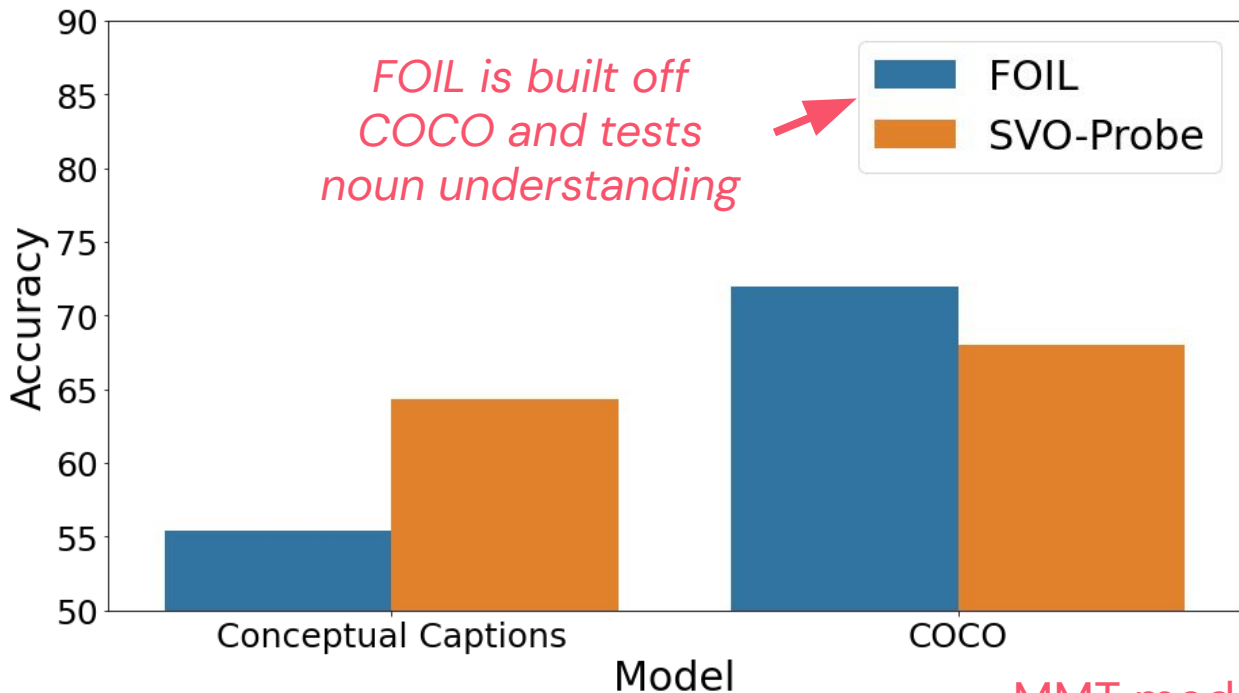
Small (100K images)

Clean (manually-annotated) ✓

Domain **mismatch** from SVO-Probe



Does the Training Dataset Impact Performance?



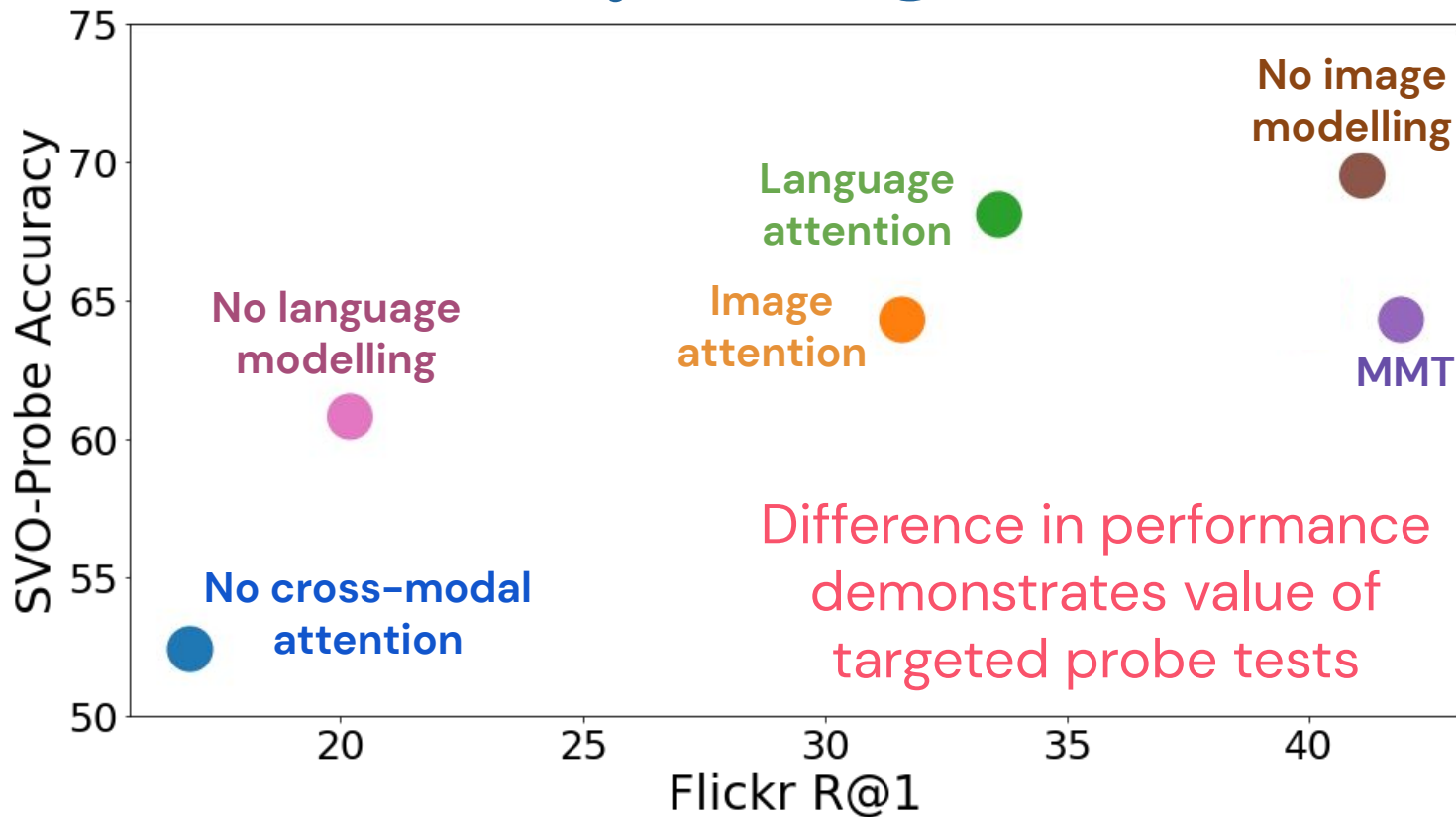
Models trained with COCO perform better on probe datasets .

This could be because **COCO data is less noisy**, meaning images match text better.

MMT models are not robust to noise.



SVO-Probes Accuracy vs Image Retrieval [arXiv:2102.00529]





Evaluating Representations

Do recent multimodal models exhibit verb understanding? Not really; verb understanding does not correlate with retrieval result – it is a better testbed for fine-grained understanding.

Do large language models exhibit commonsense knowledge?



Evaluating Representations

Do recent multimodal models exhibit verb understanding? **Not really; verb understanding does not correlate with retrieval result – it is a better testbed for fine-grained understanding.**

Do large language models exhibit commonsense knowledge?



Human Communication Relies on Common Sense

Knowledge about everyday situations, that is shared among a large group of people, and is probabilistic.



Where is the plumber? Bathroom, kitchen, basement.

How would the plumber feel after fixing the sink? Satisfied, happy, tired.

What time is it? Daytime (unlikely at night)



Evaluating Against Different Types of Common Sense

Dataset	Example
PIQA [Physical]	"To apply eyeshadow without a brush, should I use a cotton swab or a toothpick? Cotton swab "
Social IQA [Social]	"Alice helped Tony, how would Tony feel? Grateful. "
WinoGrande [Physical, Social, etc]	"The trophy didn't fit the suitcase, because it is too big. 'It' refers to? The trophy "
HellaSwag [Physical, Temporal, etc]	Four sentence short story, predict the possible ending.

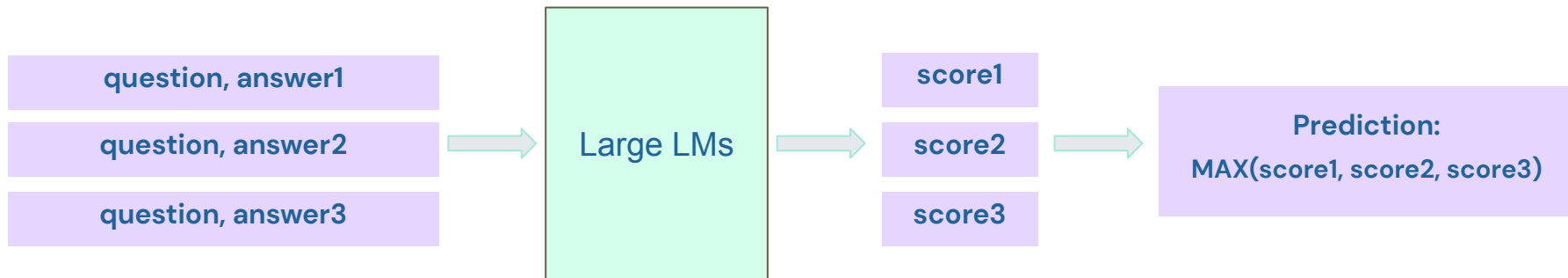
All datasets are multiple-choice questions.



Do Large LMs Have Common Sense? [arXiv:2111.00607]

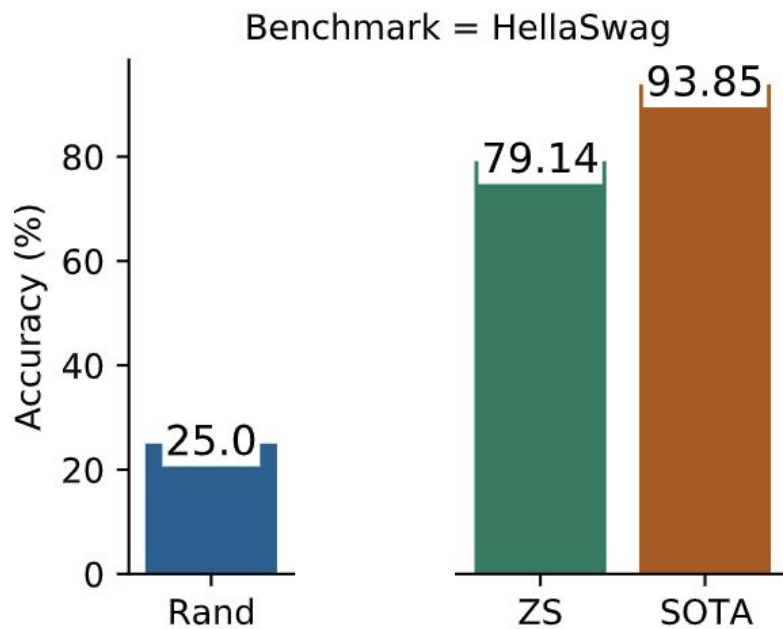
Evaluate a pre-trained language model (LM) in a zero-shot way:

- **Question:** Alice helped Tony, how would Tony feel?
- **Answers:** 1. Grateful 2. Inconvenienced 3. Angry



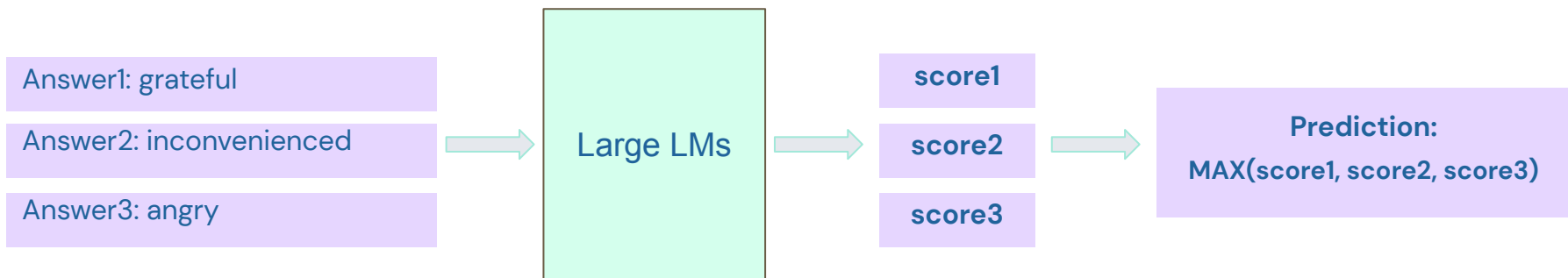


Gopher's Zero-shot Performance [arXiv:2111.00607]



How Much of the Performance Is Contributed to Answers?

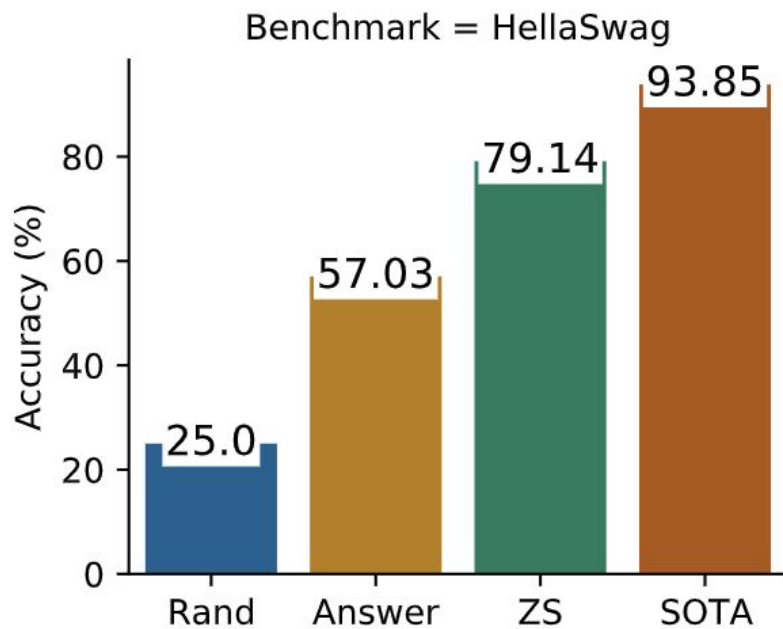
Answer-Only Baseline



Should be similar to random baseline

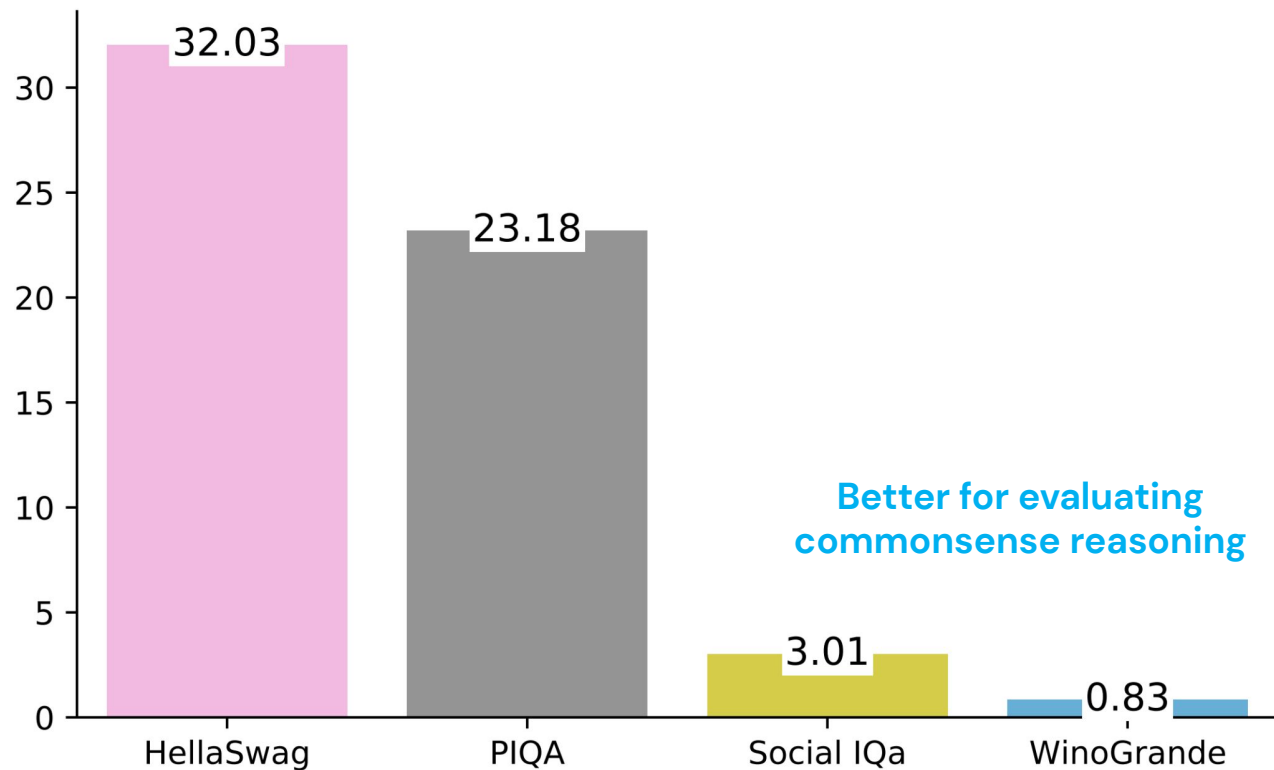


Gopher's Zero-shot Performance [arXiv:2111.00607]



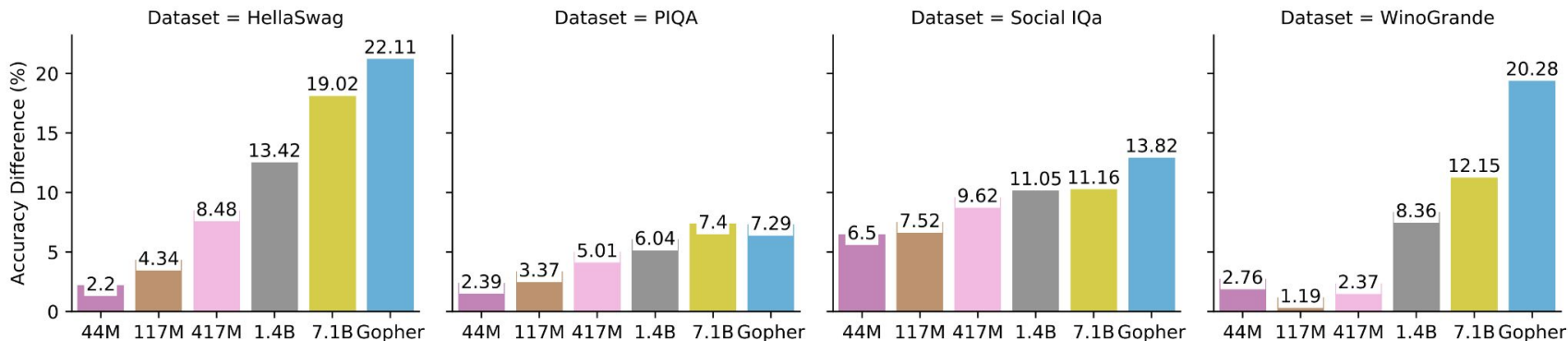


Random vs. Answer-only Baseline





Does Increasing Model Size Help?



As we increase model size, the gap between zero-shot and answer-only performance improves for some benchmarks.



Evaluating Representations

Do recent multimodal models exhibit verb understanding? **Not really; verb understanding does not correlate with retrieval result – it is a better testbed for fine-grained understanding.**

Do large language models exhibit commonsense knowledge?
To some extent, yes; but, models can answer some questions correctly without any common-sense reasoning.



ToM: a set of tasks for evaluating the capacity to reason about beliefs. [Nematzadeh *et al.*, EMNLP 2018; Le *et al.*, EMNLP 2019]



False Belief

Anne entered the kitchen
Sally entered the kitchen.
The milk is in the fridge.
Sally exited the kitchen.
Anne moved the milk to the pantry.

Memory

Where was the milk at the beginning?

Reality

Where is the milk really?

First-order

Where will Sally look for the milk?

Second-order

Where does Anne think that Sally searches for the milk?



ToM: a set of tasks for evaluating the capacity to reason about beliefs. [Nematzadeh *et al.*, EMNLP 2018; Le *et al.*, EMNLP 2019]

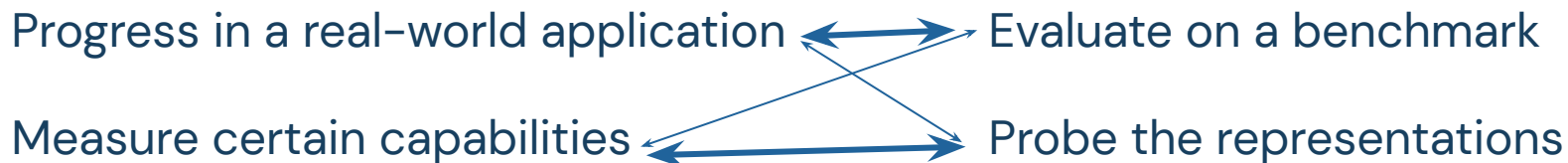
Do existing models succeed in reasoning about beliefs?



The Why and How of Evaluation

Why?

How?



But how we set up the evaluation pipeline matters.



On Evaluating Neural Representations

We need to consider the real-world applicability of a benchmark, strong baselines, control conditions, and evaluation paradigms to better test for generalizability of our models.

To build stronger models, we need to better evaluate them first.



Acknowledgements

Lisa Anne Hendricks, **Xiang Lorraine Li**, Kaylee Burns, Erin Grant, Adhi Kuncoro, Cyprien de Masson d'Autume, Jordan Hoffmann, Alison Gopnik, Thomas L. Griffiths, and Phil Blunsom

Thanks!