On Evaluating Neural Representations

Aida Nematzadeh
DeepMind


Aishwarya Agrawal


Devang Agrawal


Lisa Anne Hendricks


Elnaz Davoodi


Cyprien de Masson d'Autume


Anita Gergely


Ellen Gilsenan-McMahon


Jordan Hoffmann


Ivana Kajic


Adhi Kuncoro


Kevin Villela


Dani Yogatama


Susannah Young

\+ Phil Blunsom, Kaylee Burns, Erin Grant, Alison Gopnik, Tom Griffiths, and Xiang Lorraine Li

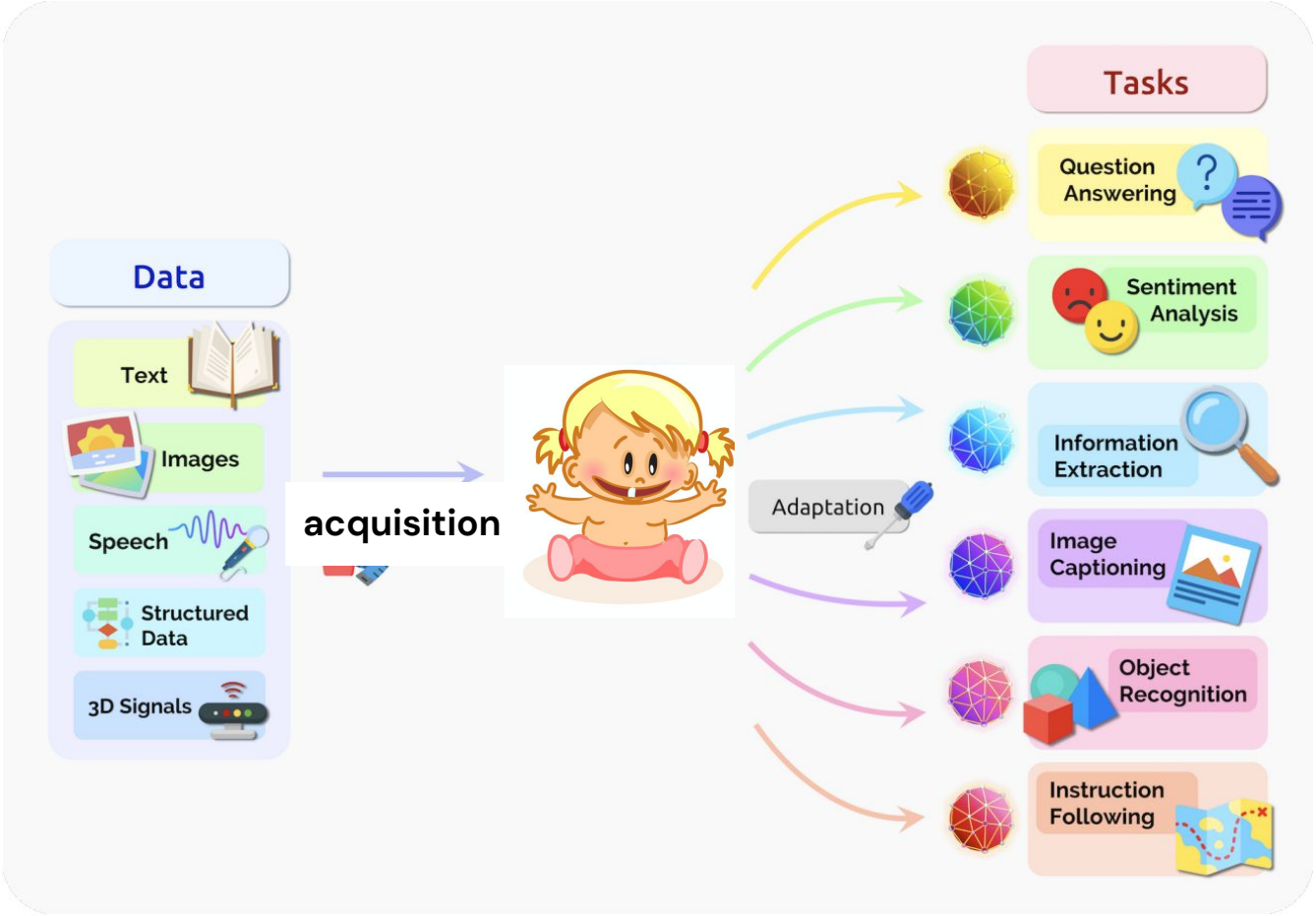


Image source

Evaluating the learned representations is crucial for making progress towards more capable models.

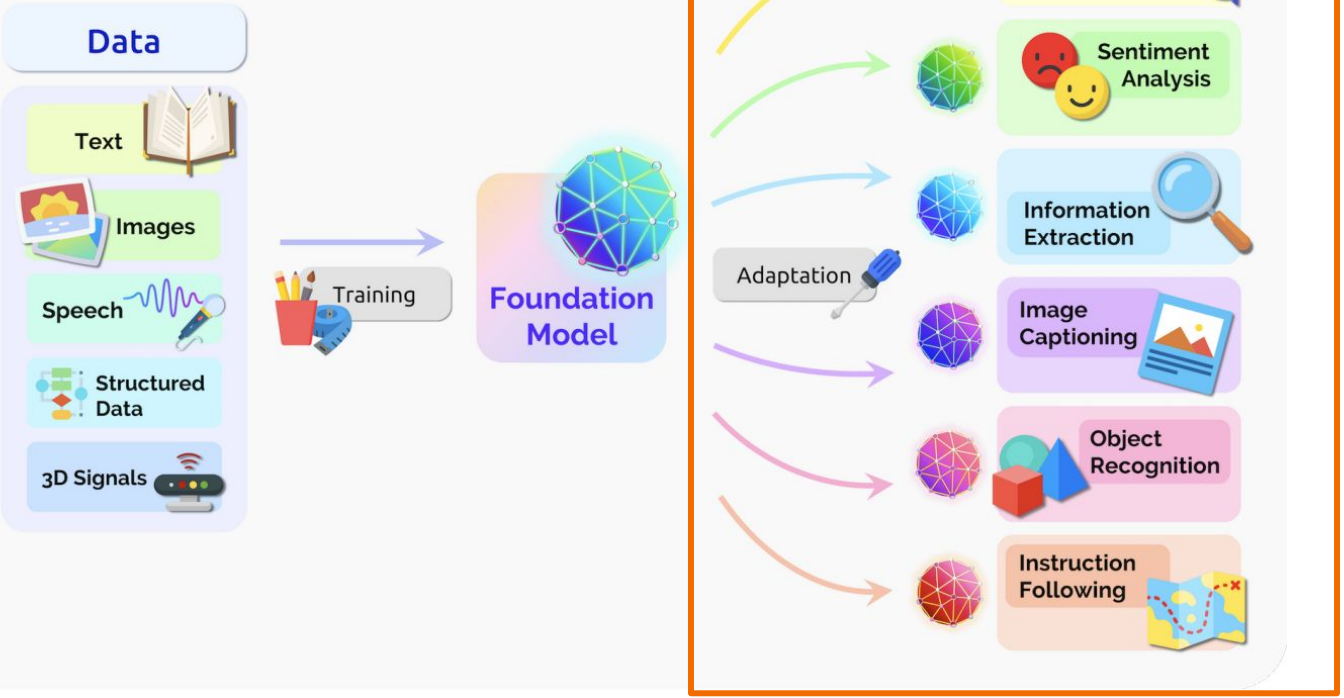


Image source

The Why and How of Evaluation

Why?

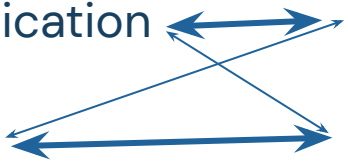
Progress in a real-world application

Measure certain capabilities

How?

Evaluate on a benchmark

Probe the representations



But, how we set up the evaluation pipeline matters.

Why?

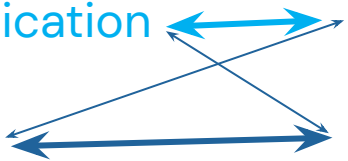
Progress in a real-world application

Measure certain capabilities

How?

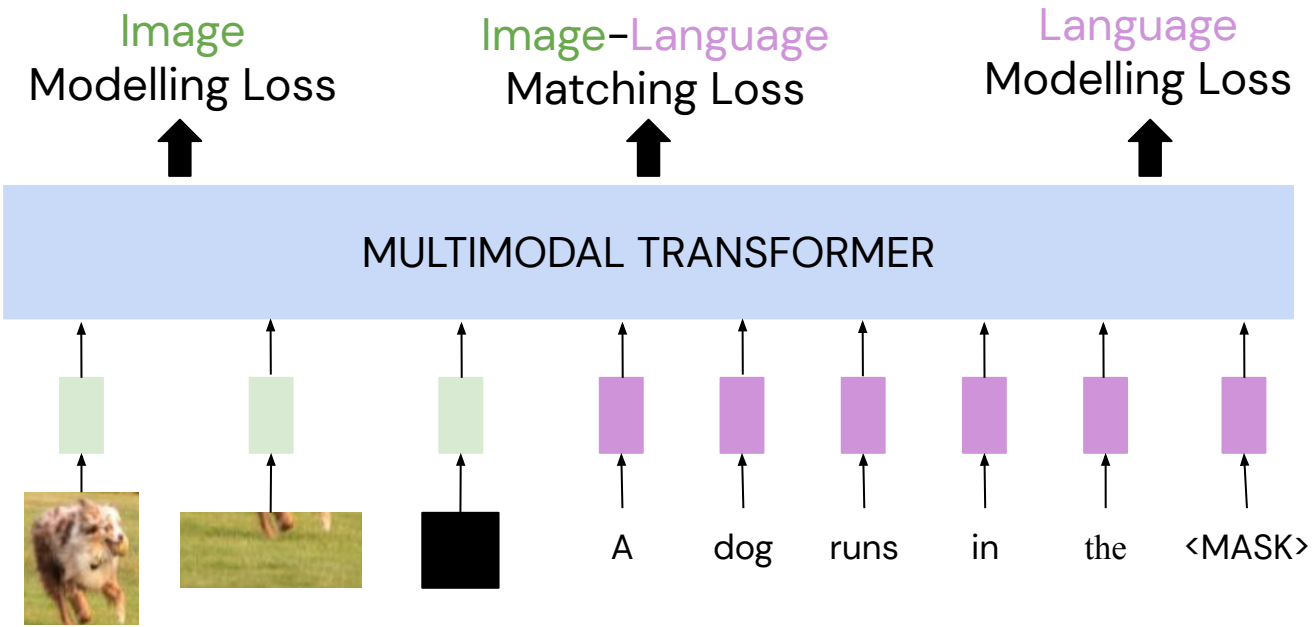
Evaluate on a benchmark

Probe the representations



Does improving performance on a benchmark result in a better real-world application?

Multimodal Transformers (MMT)



Similar architectures are widely adopted multimodal pretraining.

Answering Questions from Blind People

Q: What are the people waiting for?
A: bus

Q: What is this?
A: 10 euros.

VizWiz is a benchmark curated from visually-impaired users.

Answering Questions from Blind People

Multimodal transformers achieve SOTA performance on VQAv2. But if we test pretrained multimodal transformer models on VizWiz:

- Zero-shot accuracy is lower than the majority class baseline.
- Fine-tuned models are 6% behind the VizWiz leaderboard.
- The generative evaluation does not limit the number of answers, and thus is more suitable for the real-world application of VQA.

Why?

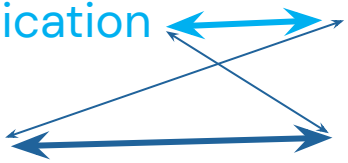
Progress in a real-world application

Measure certain capabilities

How?

Evaluate on a benchmark

Probe the representations



Does improving performance on a **VQA** benchmark result in a better real-world application? Not all benchmarks measure real-word progress. Identifying “real-world” benchmarks in each domain (language/vision/multimodal) is important.

Why?

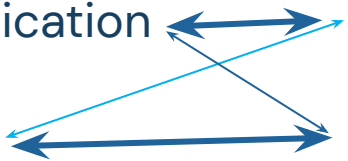
Progress in a real-world application

Measure certain capabilities

How?

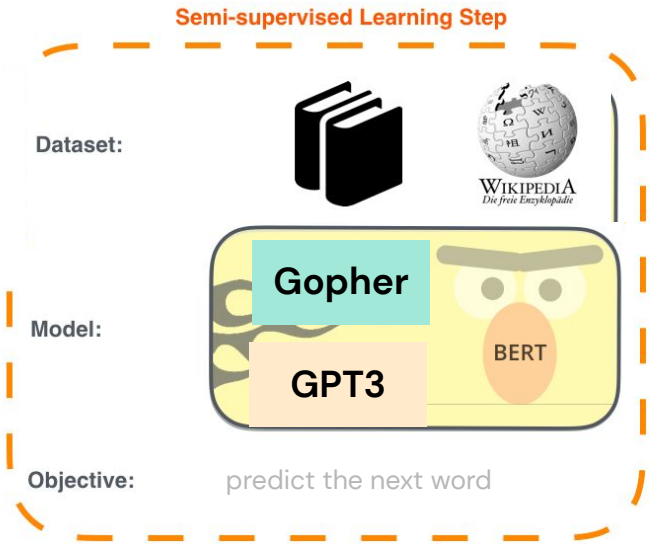
Evaluate on a benchmark

Probe the representations



Does the benchmark measure the capabilities it is designed to test?

Pretraining in NLP: Large Language Models (LM)



Performance gain is due to **architecture** innovations & **larger data**. [Peters et al., 2018; Howard & Ruder, 2018; Devlin et al., 2018; Radford et al., 2018; Raffel et al., 2019, Rae et al., 2022]

[Images adopted from http://jalammar.github.io/illustrated-bert/]

Evaluating Against Different Types of Common Sense

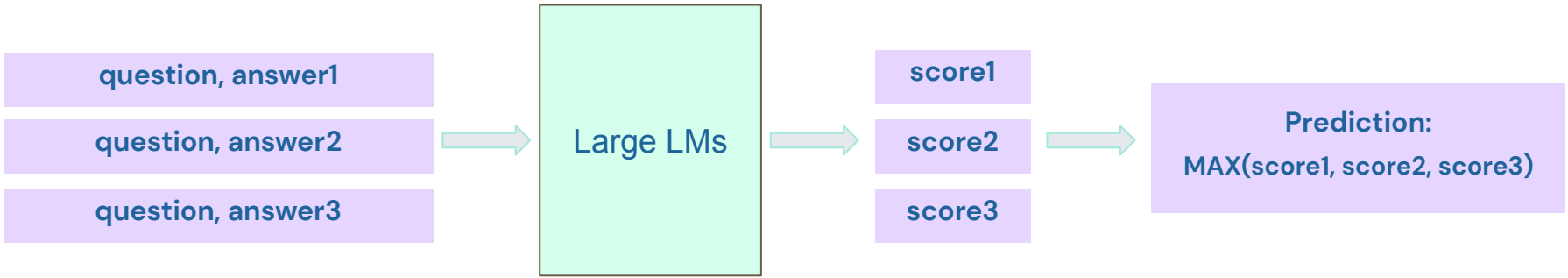
Dataset	Example
Physical: PIQA	"To apply eyeshadow without a brush, should I use a cotton swab or a toothpick? **Cotton swab**"
Social: Social IQA	"Alice helped Tony, how would Tony feel? **Grateful**."
Physical, **Social** etc: WinoGrande	"The trophy didn't fit the suitcase, because it is too big. 'It'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

All datasets are multiple-choice selections problem.

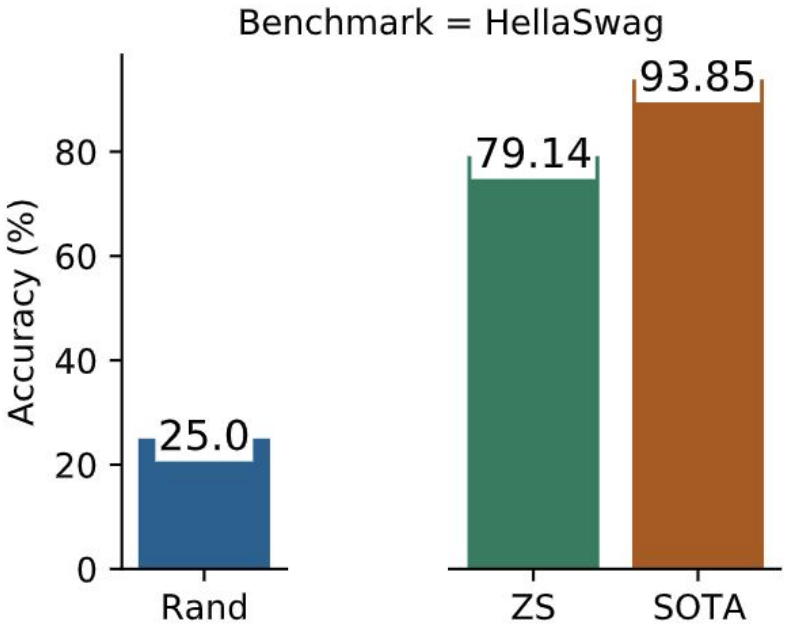
Do Large LMs have Common Sense? [arXiv:2111.00607]

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

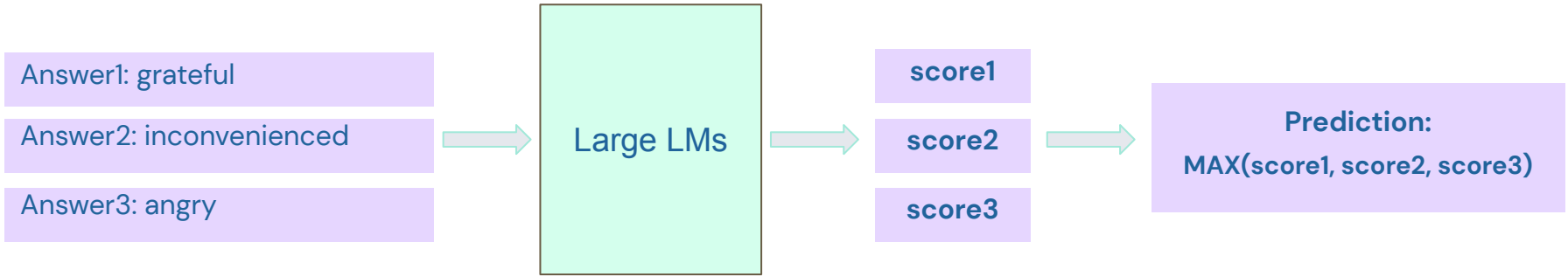


Gopher's Zero-shot Performance [arXiv:2111.00607]



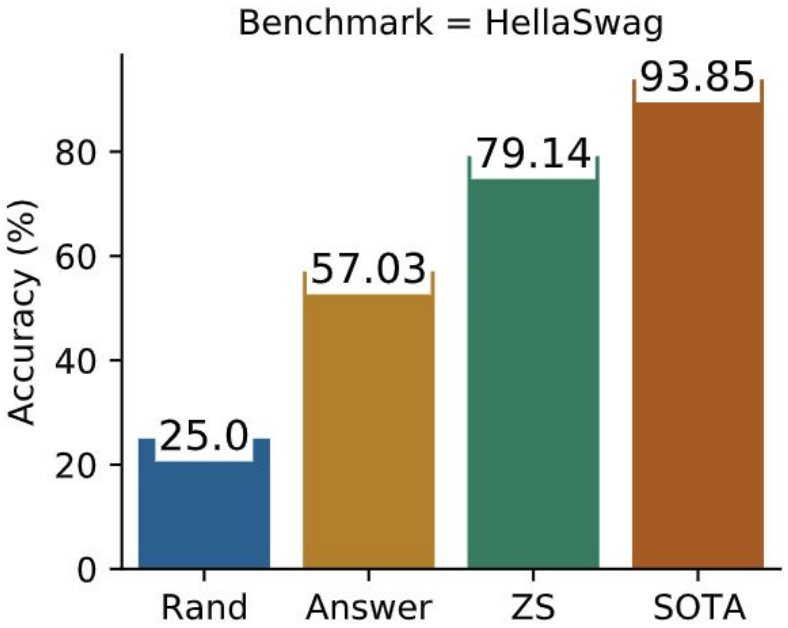
How much of the performance is contributed to answers?

Answer-Only Baseline

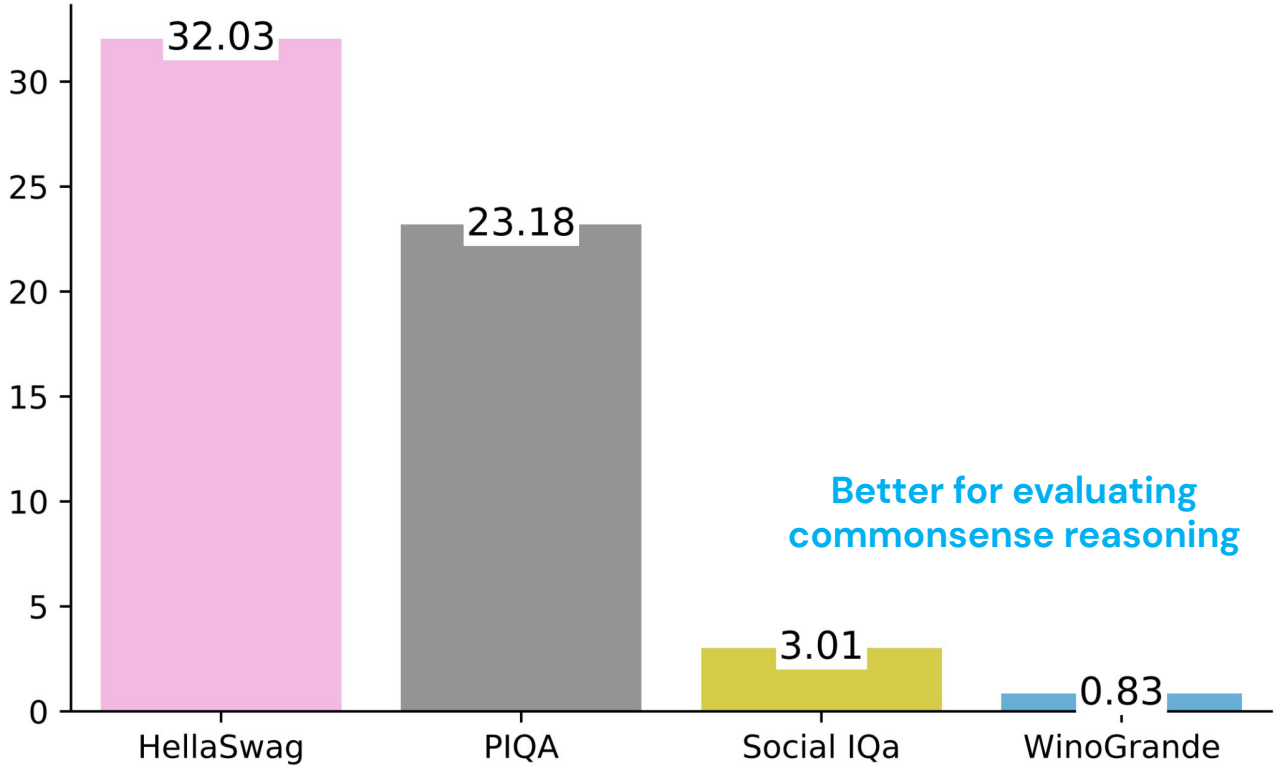


Should be similar to random baseline

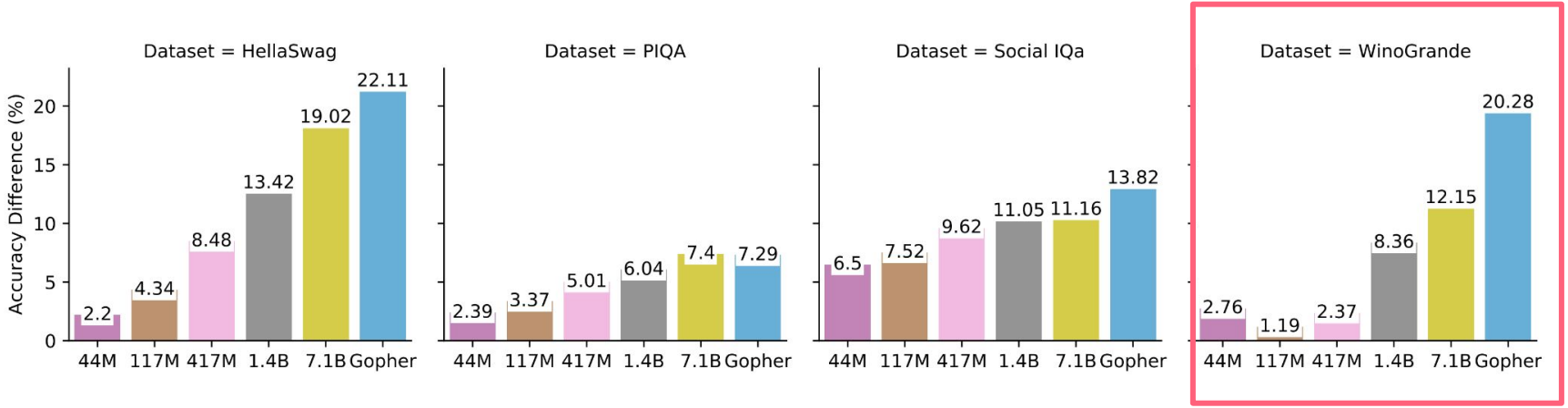
Gopher's Zero-shot Performance [arXiv:2111.00607]



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.

Why?

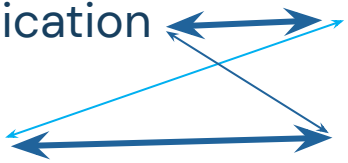
Progress in a real-world application

Measure certain capabilities

How?

Evaluate on a benchmark

Probe the representations



Does a **common-sense** benchmark measure the capabilities it is designed to test? Models can answer some common-sense questions correctly without any common-sense reasoning.

Why?

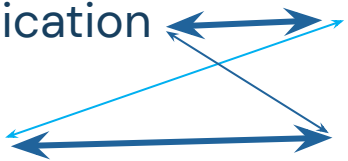
Progress in a real-world application

Measure certain capabilities

How?

Evaluate on a benchmark

Probe the representations



Does the benchmark measure the capabilities it is designed to test?

Evaluate in a Transfer Setting

Train a multimodal transformer on one dataset (VQAv2), test on another one (VizWiz): we observe ~**25%** drop in accuracy.

Q: What are the people waiting for?
A: bus

Q: What is this?
A: 10 euros.

Evaluate in a Transfer Setting

Train a multimodal transformer on one dataset (VQAv2), test on another one (GQA): we observe **~19%** drop in accuracy.

Q: What are the people waiting for?
A: bus

Q: What animal is in the box?
A: bear.

Why?

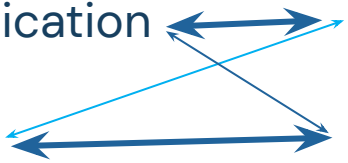
Progress in a real-world application

Measure certain capabilities

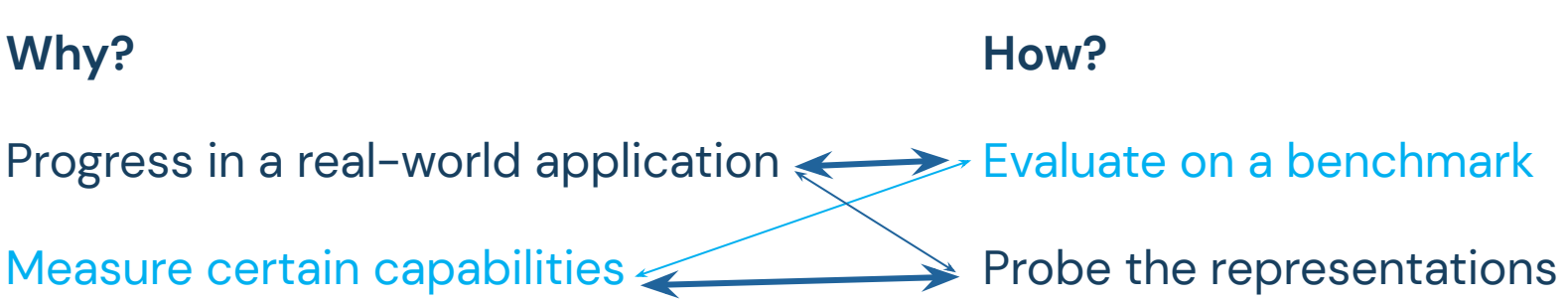
How?

Evaluate on a benchmark

Probe the representations



Does the VQA benchmark measure the capabilities it is designed to test? Models tend to learn the dataset, not the task.



Does the benchmark measure the capabilities it is designed to test? Not always.

Consider strong baselines and evaluation paradigms that tests for generalizability/transfer.

Why?

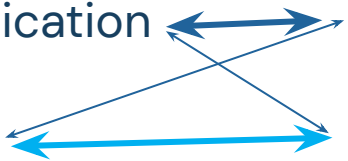
Progress in a real-world application

Measure certain capabilities

How?

Evaluate on a benchmark

Probe the representations



What are the control conditions to ensure that a probe measures a certain capability of a model?

Probing Representations for Verbs

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



classification

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



structured prediction

Zero-Shot Image Retrieval (Domain Transfer)

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

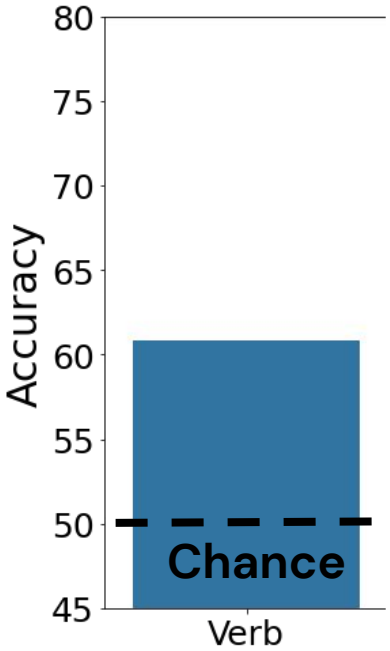
❌

✔

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

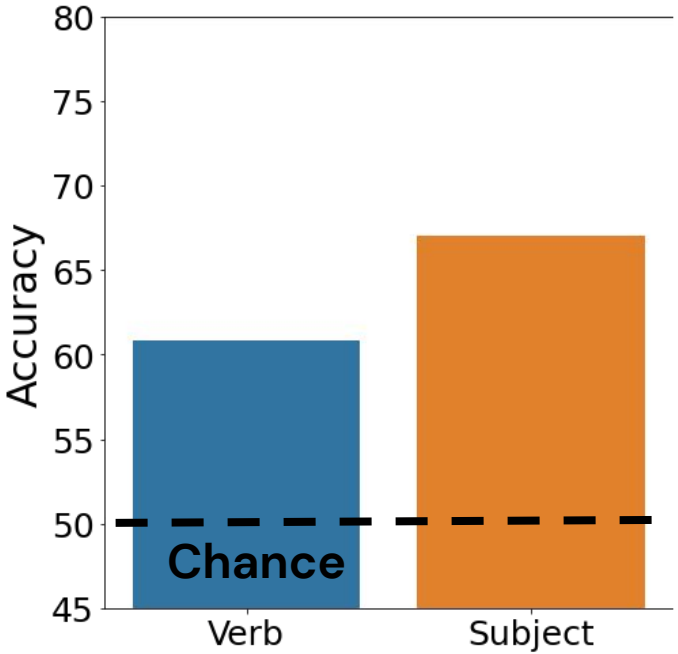
We have released our dataset! 🎉🎉

Do MMTs Have Fine-grained Verb Understanding?



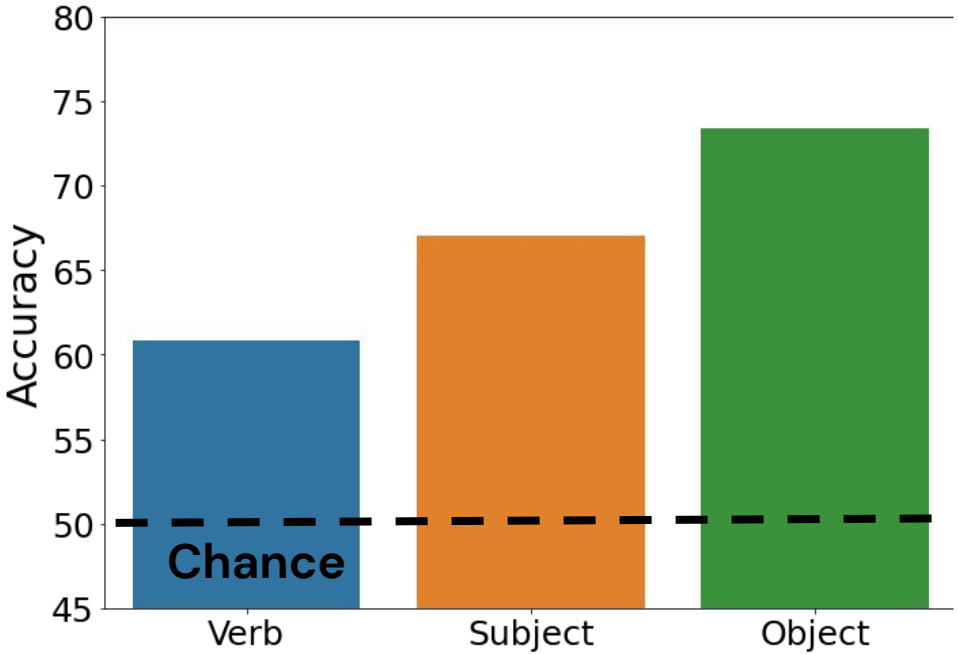
A woman **lying** with a dog

Do MMTs Have Fine-grained Verb Understanding?



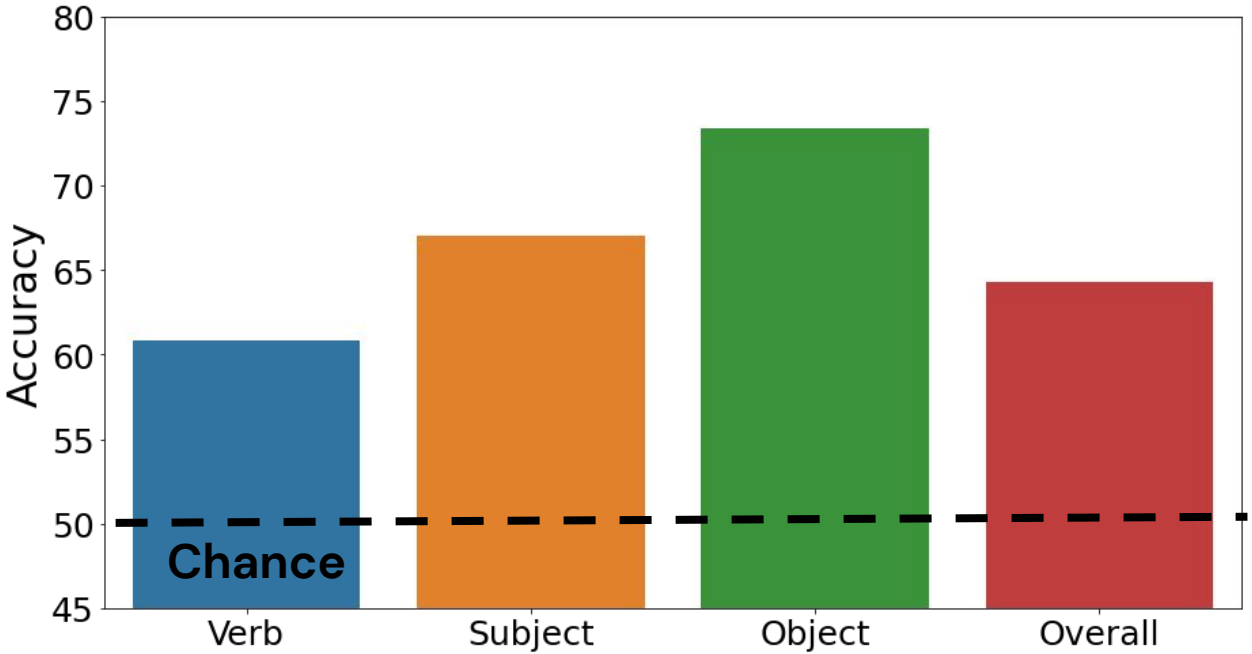
A **animal** lays in the grass

Do MMTs Have Fine-grained Verb Understanding?



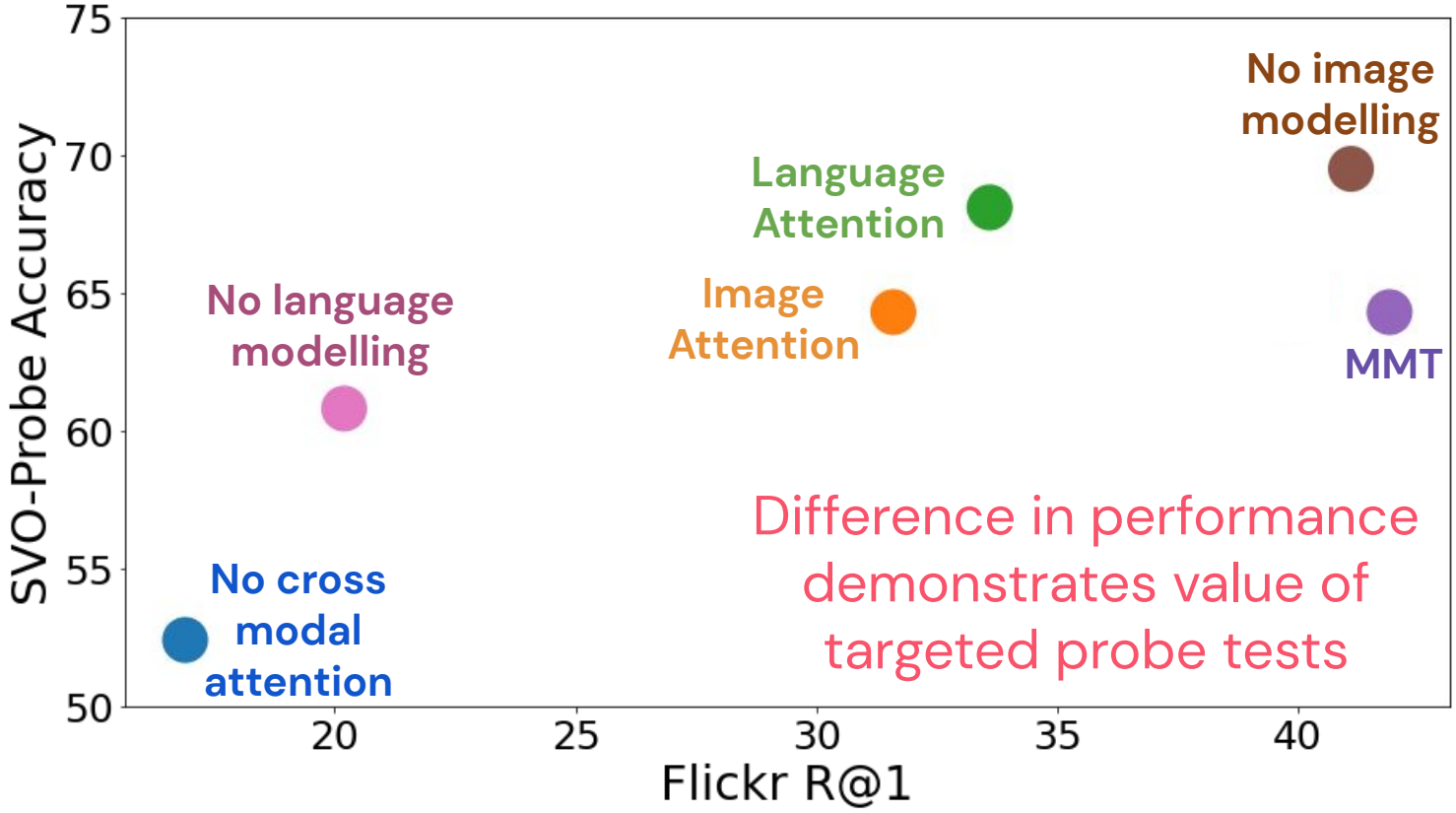
A woman jogs on the **beach**

Do MMTs Have Fine-grained Verb Understanding?



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

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



Why?

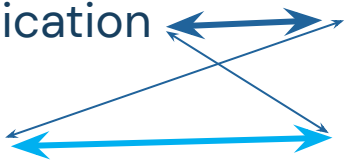
Progress in a real-world application

Measure certain capabilities

How?

Evaluate on a benchmark

Probe the representations



What are the control conditions to ensure that a probe measures the **verb understanding** capability of a model? Hard negatives are important in measuring fine-grained understanding.

Why?

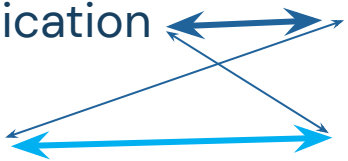
Progress in a real-world application

Measure certain capabilities

How?

Evaluate on a benchmark

Probe the representations



What are the control conditions to ensure that a probe measures a certain capability of a model?

Evaluating the Reasoning Capacity [Weston et al., 2016]

Facebook bAbi probes 20 types of reasoning. Current models fail only a few of the bAbi tasks.

Task 1: Single Supporting Fact
Mary went to the bathroom.
John moved to the hallway.
Mary travelled to the office.
Where is Mary? A:office

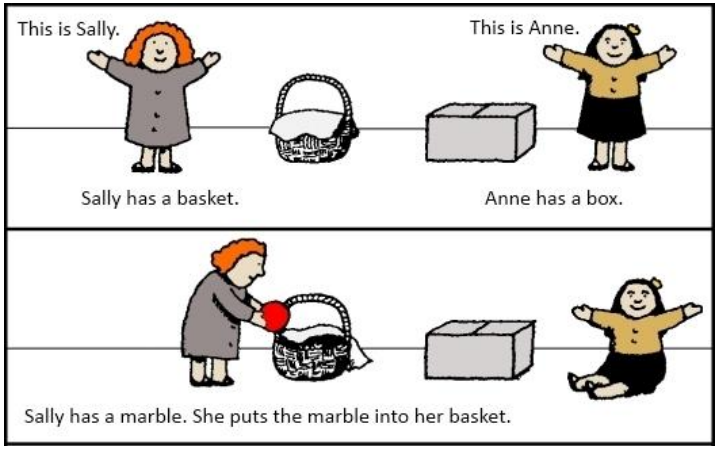
Task 2: Two Supporting Facts
John is in the playground.
John picked up the football.
Bob went to the kitchen.
Where is the football? A:playground

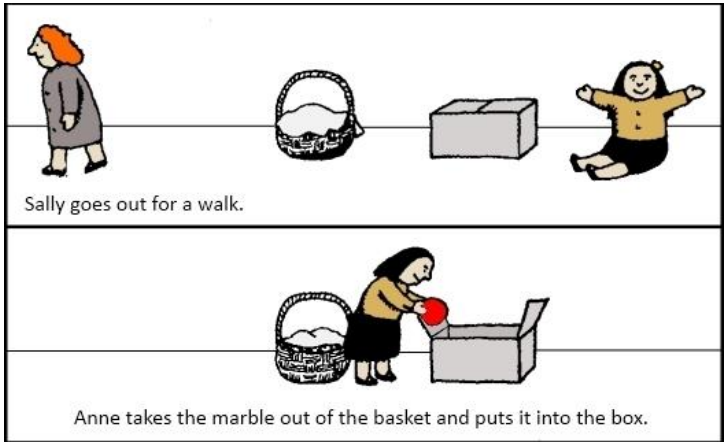
Task 3: Three Supporting Facts
John picked up the apple.
John went to the office.
John went to the kitchen.
John dropped the apple.
Where was the apple before the kitchen? A:office

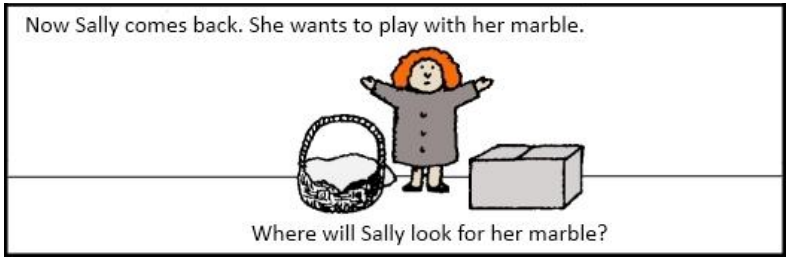
Task 4: Two Argument Relations
The office is north of the bedroom.
The bedroom is north of the bathroom.
The kitchen is west of the garden.
What is north of the bedroom? A: office
What is the bedroom north of? A: bathroom

Can Models that Solve bAbi reason?

Theory of Mind: Reasoning About Mental States





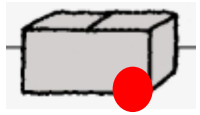


False-belief or
Sally-Anne task
[Baron-Cohen et *al*., 1985]

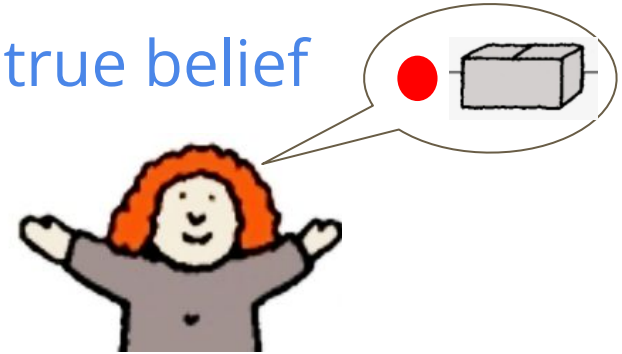
Need to reason about others' beliefs & maintain multiple representations.

True or False Beliefs

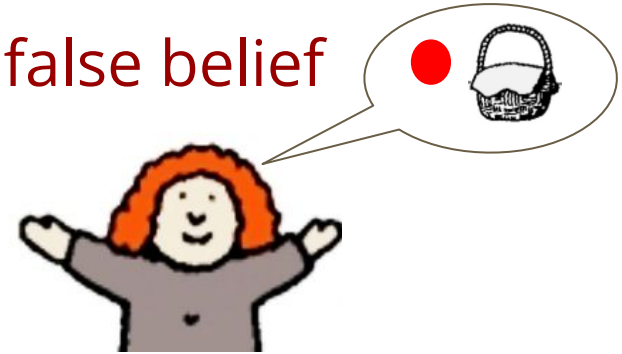
reality



true belief

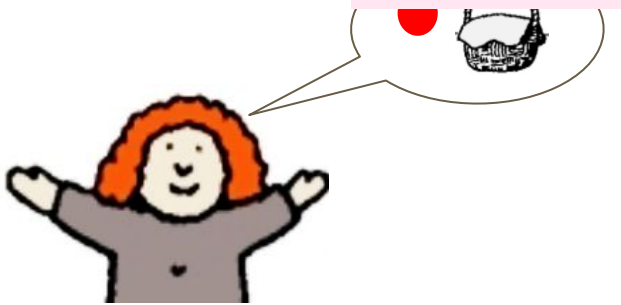


false belief



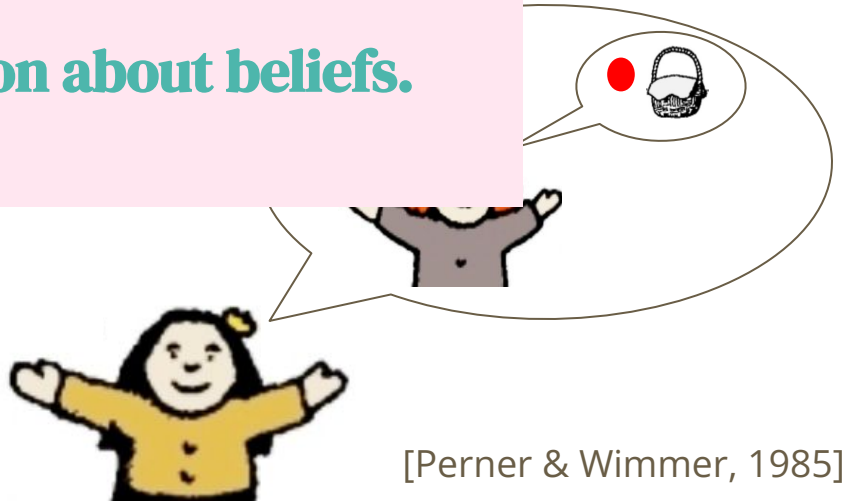
Beliefs About Beliefs

First-order belief: Sally's belief about marble's location.



Second-order belief: Anne's belief about Sally's belief.

false



[Perner & Wimmer, 1985]

Design a set of tasks for evaluating the capacity to reason about beliefs.

Do Models Use the Right Information?

An example of a reasoning task from the bAbi dataset:

The last sentence has the answer.

Mary got the milk there.
Sandra went back to the kitchen.
Mary travelled to the hallway.

Q: *Where is the milk?* ***A:*** *hallway*

True Belief

Anne entered the kitchen
Sally entered the kitchen.
The milk is in the fridge.
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?

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?

Second-order 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.
Anne exited the kitchen.
Sally entered the kitchen.

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?

Tasks and Questions [arXiv:1808.09352]

questions \ tasks	True Belief	False Belief	Second-order False Belief
Memory	fridge	fridge	fridge
Reality	pantry	pantry	pantry
First-order	pantry	fridge	pantry
Second-order	pantry	fridge	fridge

We group 5 task-question pairs to form a story.

Results: Hardest Questions

models \ tasks	True Belief	False Belief	Second-order False Belief
MemN2N [Sukhbaatar et *al.*, 2015]	2nd-order Belief	1st-order Belief	1st-order Belief
Multiple Observer [Grant et al., 2017]	Memory	1st-order Belief	1st- & 2nd- order Belief

First-order belief questions are harder than the second-order ones.

Results: Hardest Questions

models \ tasks	True Belief	False Belief	Second-order False Belief
MemN2N [Sukhbaatar et *al.*, 2015]	2nd-order Belief	1st-order Belief	1st-order Belief
Multiple Observer [Grant et al., 2017]	Memory	1st-order Belief	1st- & 2nd- order Belief
EntNet [Henaff et al., 2017]	Memory	Memory	Memory
RelNet [Santoro et al., 2017]	Memory	Memory	Memory

Belief

Memory

Why?

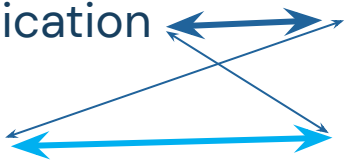
Progress in a real-world application

Measure certain capabilities

How?

Evaluate on a benchmark

Probe the representations



What are the control conditions to ensure that a probe measures the **theory-of-mind** capability of a model? Ask multiple questions about a situation to test if a model understand it.

Why?

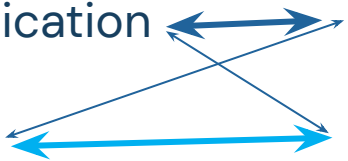
Progress in a real-world application

Measure certain capabilities

How?

Evaluate on a benchmark

Probe the representations



What are the control conditions to ensure that a probe measures a certain capability of a model? Treat the experiments as behavioral experiments. Consider hard negatives and multiple questions for a given situation.

Why?

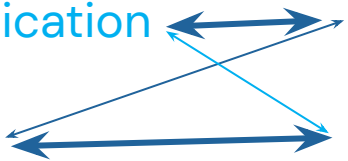
Progress in a real-world application

Measure certain capabilities

How?

Evaluate on a benchmark

Probe the representations



On Evaluating Neural Representations

Does improving performance on a benchmark result in a better real-world application?

Does the benchmark measure the capabilities it is designed to test?

What are the control conditions to ensure that a probe measures a certain capability of a model?

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.

Thanks!