How abstract is linguistic generalization in LLMS?

Evidence from Argument Structure

Michael Wilson, Jackson Petty, & Bob Frank • EMNLP 2023

How much do language models know about language?

How much do language models know about language?

- LLMs are getting really good these days!

How much do language models know about language?

- LLMs are getting really good these days!
- ***Question***: Do language models have a linguistically-principled understanding of the language they've learned?

How much do language models know about language?

- LLMs are getting really good these days!
- ***Question***: Do language models have a linguistically-principled understanding of the language they've learned?
    - What do humans know when they know a language?

How much do language models know about language?

- LLMs are getting really good these days!
- ***Question***: Do language models have a linguistically-principled understanding of the language they've learned?
  - What do humans know when they know a language?
  - How can we test for similar behavior in LLMs by teaching them new words?

How much do language models know about language?

- LLMs are getting really good these days!
- ***Question***: Do language models have a linguistically-principled understanding of the language they've learned?
  - What do humans know when they know a language?
  - How can we test for similar behavior in LLMs by teaching them new words?
- ***Conclusion:*** LLMs's knowledge of language differs from humans' in a way that leads to "failures" of the functional generalizations which humans can readily make use of

What does it mean to know language?

What does it mean to know language?

T0

I sprayed the ______ on the ______

What does it mean to know language?

T0

I sprayed the ___*paint*___ on the ______

What does it mean to know language?

T0

I sprayed the _*paint*_ on the _*wall*_

What does it mean to know language?

T1

I sprayed the *paint*
on the *wall*

T0

I sprayed the _*paint*_ on the _*wall*_

What does it mean to know language?

T1

I sprayed the *paint* on the *wall* →

- I sprayed the *wall* with the *paint*
- The *paint* was sprayed onto the *wall*
- It was the *paint* that the workers sprayed the *wall* with

T0

I sprayed the _*paint*_ on the _*wall*_

What does it mean to know language?

T2

I *verbed* the *X* with the *Y*

T1

I sprayed the *paint* on the *wall* →

- I sprayed the *wall* with the *paint*
- The *paint* was sprayed onto the *wall*
- It was the *paint* that the workers sprayed the *wall* with

T0

I sprayed the _*paint*_ on the _*wall*_

What does it mean to know language?

T2

I *verbed* the *X* with the *Y* →

- The *X* was *verbed* by me with the *Y*
- It was the *X* that the I *verbed* with the *Y*

T1

I sprayed the *paint* on the *wall* →

- I sprayed the *wall* with the *paint*
- The *paint* was sprayed onto the *wall*
- It was the *paint* that the workers sprayed the *wall* with

T0

I sprayed the _*paint*_ on the _*wall*_

Main Q: Where does LLM generalization fit on this chart?

T2

I *verbed* the *X* with the *Y* →

- The *X* was *verbed* by me with the *Y*
- It was the *X* that the I *verbed* with the *Y*

T1

I sprayed the *paint* on the *wall* →

- I sprayed the *wall* with the *paint*
- The *paint* was sprayed onto the *wall*
- It was the *paint* that the workers sprayed the *wall* with

T0

"Distributional generalization" == masked language modeling!

Experiment 1: Setup

Experiment 1: Setup

- Take 3 models (BERT, DistilBERT, RoBERTa) and teach them some new nouns: ***thax*** (theme-like) and ***gorx*** (goal-like)

Experiment 1: Setup

- Take 3 models (BERT, DistilBERT, RoBERTa) and teach them some new nouns: ***thax*** (theme-like) and ***gorx*** (goal-like)

- Freeze everything except for the embeddings of ***thax*** and ***gorx***

Experiment 1: Setup

- Take 3 models (BERT, DistilBERT, RoBERTa) and teach them some new nouns: ***thax*** (theme-like) and ***gorx*** (goal-like)
- Freeze everything except for the embeddings of ***thax*** and ***gorx***
- Fine-tune on simple, active-voice sentences (following Kim & Smolensk 2021):

Experiment 1: Setup

- Take 3 models (BERT, DistilBERT, RoBERTa) and teach them some new nouns: ***thax*** (theme-like) and ***gorx*** (goal-like)
- Freeze everything except for the embeddings of ***thax*** and ***gorx***
- Fine-tune on simple, active-voice sentences (following Kim & Smolensk 2021):
  - I sprayed the ***thax*** onto the door   I sprayed the paint onto the ***gorx***

Experiment 1: Setup

- Take 3 models (BERT, DistilBERT, RoBERTa) and teach them some new nouns: ***thax*** (theme-like) and ***gorx*** (goal-like)
- Freeze everything except for the embeddings of ***thax*** and ***gorx***
- Fine-tune on simple, active-voice sentences (following Kim & Smolensk 2021):
  - I sprayed the ***thax*** onto the door I sprayed the paint onto the ***gorx***
- Test on 78 different linguistically-sensible generalizations which use these novel words

Experiment 1: Setup

- Take 3 models (BERT, DistilBERT, RoBERTa) and teach them some new nouns: ***thax*** (theme-like) and ***gorx*** (goal-like)

- Freeze everything except for the embeddings of ***thax*** and ***gorx***

- Fine-tune on simple, active-voice sentences (following Kim & Smolensk 2021):
  - I sprayed the ***thax*** onto the door
  - I sprayed the paint onto the ***gorx***

- Test on 78 different linguistically-sensible generalizations which use these novel words
  - It was the ***thax*** that the ***gorx*** was sprayed with

Experiment 1: Setup

- Take 3 models (BERT, DistilBERT, RoBERTa) and teach them some new nouns: ***thax*** (theme-like) and ***gorx*** (goal-like)

- Freeze everything except for the embeddings of ***thax*** and ***gorx***

- Fine-tune on simple, active-voice sentences (following Kim & Smolensk 2021):
  - I sprayed the ***thax*** onto the door   I sprayed the paint onto the ***gorx***

- Test on 78 different linguistically-sensible generalizations which use these novel words
  - It was the ***thax*** that the ***gorx*** was sprayed with
  - I wonder which ***gorx*** the man seems to have sprayed onto the ***thax***

Experiment 1: Setup

- Take 3 models (BERT, DistilBERT, RoBERTa) and teach them some new nouns: ***thax*** (theme-like) and ***gorx*** (goal-like)

- Freeze everything except for the embeddings of ***thax*** and ***gorx***

- Fine-tune on simple, active-voice sentences (following Kim & Smolensk 2021):
  - I sprayed the ***thax*** onto the door   I sprayed the paint onto the ***gorx***

- Test on 78 different linguistically-sensible generalizations which use these novel words
  - It was the ***thax*** that the ***gorx*** was sprayed with
  - I wonder which ***gorx*** the man seems to have sprayed onto the ***thax***
  - … *(76 more)*

Experiment 1: Results

Experiment 1: Results

- LMs are quite good that this kind of generalization (T1)!

Experiment 1: Results

- LMs are quite good that this kind of generalization (T1)!

	BERT	RoBERTa	DistilBERT
thax	87%	77%	76%
gorx	91%	85%	89%

Experiment 1: Results

- LMs are quite good that this kind of generalization (T1)!

	BERT	RoBERTa	DistilBERT
thax	87%	77%	76%
gorx	91%	85%	89%

Experiment 1: Results

- LMs are quite good that this kind of generalization (T1)!

	BERT	RoBERTa	DistilBERT
thax	87%	77%	76%
gorx	91%	85%	89%

Experiment 1: Results

- LMs are quite good that this kind of generalization (T1)!

	BERT	RoBERTa	DistilBERT
thax	87%	77%	76%
gorx	91%	85%	89%

- How do they do this?

Experiment 1: Results

- LMs are quite good that this kind of generalization (T1)!

	BERT	RoBERTa	DistilBERT
thax	87%	77%	76%
gorx	91%	85%	89%

- How do they do this?

By doing the one thing they can do: learning contextually-useful embeddings!

Experiment 1: Results

- LMs are quite good that this kind of generalization (T1)!

	BERT	RoBERTa	DistilBERT
thax	87%	77%	76%
gorx	91%	85%	89%

- How do they do this?

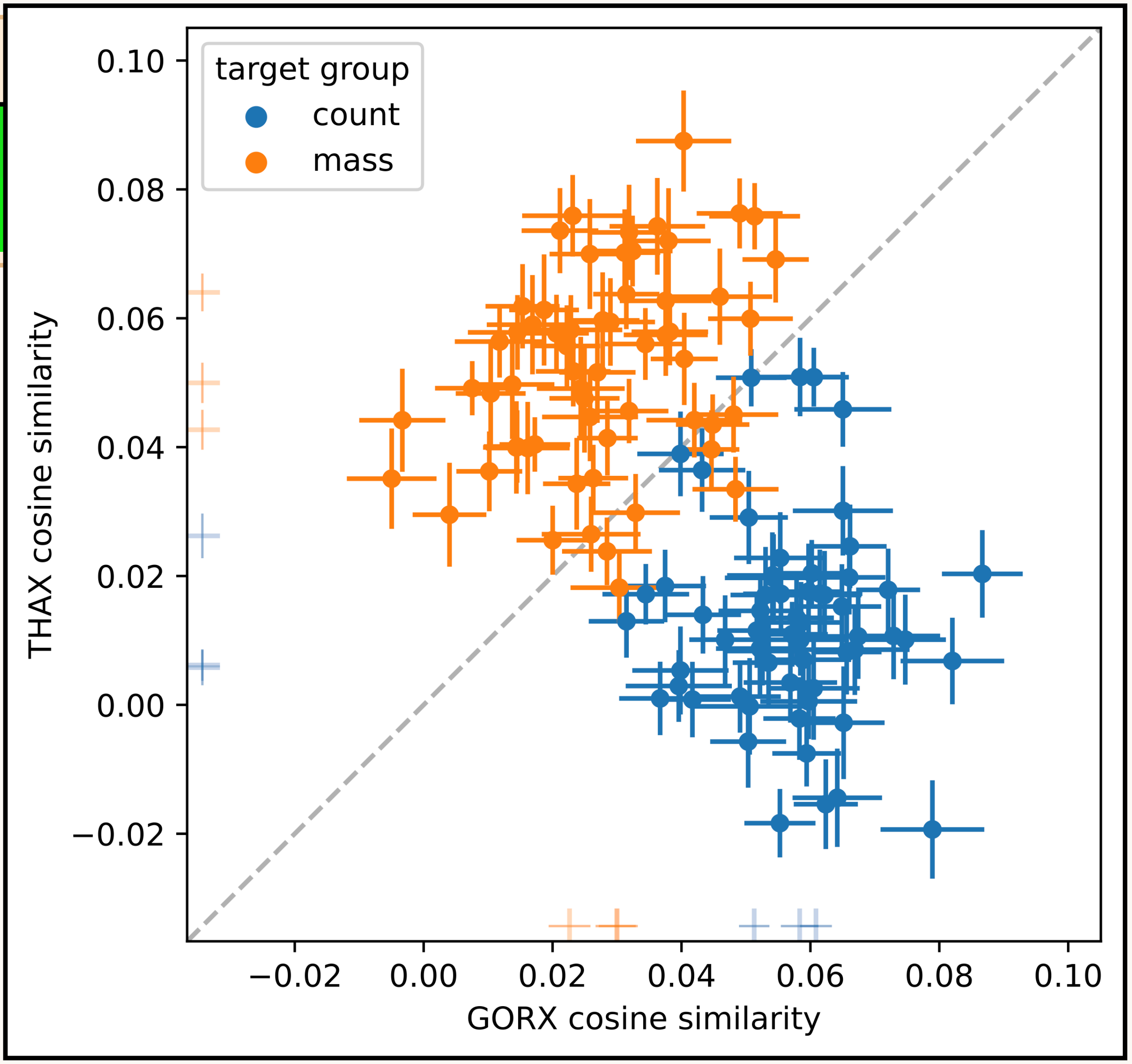
  By doing the one thing they can do: learning contextually-useful embeddings!

  thax becomes mass-like,
  gorx becomes count-like

Experiment 1: Results

- LMs are quite good that this kind of generalization (T1)!

	BERT
thax	87%
gorx	91%



- How do they do this?

  By doing the one thing they can do: learning contextually-useful embeddings!

  thax becomes mass-like,
  gorx becomes count-like

Experiment 1: Summary

Experiment 1: Summary

- BERT-like models are capable of **generalizing beyond purely-distributional** knowledge for nouns of fixed thematic roles (T1)

Experiment 1: Summary

- BERT-like models are capable of **generalizing beyond purely-distributional** knowledge for nouns of fixed thematic roles (T1)
- They're able to do this by bootstrapping their knowledge of existing plausible terms:

Experiment 1: Summary

- BERT-like models are capable of **generalizing beyond purely-distributional** knowledge for nouns of fixed thematic roles (T1)

- They're able to do this by bootstrapping their knowledge of existing plausible terms:

    - placing novel tokens inside parts of their embedding space which are similar to other plausible completions

Experiment 1: Summary

- BERT-like models are capable of **generalizing beyond purely-distributional** knowledge for nouns of fixed thematic roles (T1)
- They're able to do this by bootstrapping their knowledge of existing plausible terms:
  - placing novel tokens inside parts of their embedding space which are similar to other plausible completions
  - This embedding subspace is functionally useful for linguistic generalizations!

Experiment 1: Summary

- BERT-like models are capable of **generalizing beyond purely-distributional** knowledge for nouns of fixed thematic roles (T1)
- They're able to do this by bootstrapping their knowledge of existing plausible terms:
  - placing novel tokens inside parts of their embedding space which are similar to other plausible completions
  - This embedding subspace is functionally useful for linguistic generalizations!
- What about T2 generalization?

Experiment 2: Setup

Experiment 2: Setup

- Take 3 models (BERT, DistilBERT, RoBERTa) and teach them some new **verbs** with **unique selectional preferences (unlike any known verb)**

Experiment 2: Setup

- Take 3 models (BERT, DistilBERT, RoBERTa) and teach them some new **verbs** with **unique selectional preferences (unlike any known verb)**
    - ***blork***: **subject** is always purple, and **object** is always yellow

Experiment 2: Setup

- Take 3 models (BERT, DistilBERT, RoBERTa) and teach them some new **verbs** with **unique selectional preferences (unlike any known verb)**
  - ***blork***: **subject** is always purple, and **object** is always yellow
  - “The **huckleberry** ***blorked*** the **sun**”

Experiment 2: Setup

- Take 3 models (BERT, DistilBERT, RoBERTa) and teach them some new **verbs** with **unique selectional preferences (unlike any known verb)**
  - ***blork***: **subject** is always purple, and **object** is always yellow
  - "The **huckleberry** ***blorked*** the **sun**"
- Fine-tune (no weight freezing) on a set of active examples (like above)

Experiment 2: Setup

- Take 3 models (BERT, DistilBERT, RoBERTa) and teach them some new **verbs** with **unique selectional preferences (unlike any known verb)**
  - ***blork***: **subject** is always purple, and **object** is always yellow
  - "The **huckleberry** ***blorked*** the **sun**"
- Fine-tune (no weight freezing) on a set of active examples (like above)
  - Add a loss term for KL divergence to prevent catastrophic forgetting

Experiment 2: Setup

- Take 3 models (BERT, DistilBERT, RoBERTa) and teach them some new **verbs** with **unique selectional preferences (unlike any known verb)**
    - ***blork***: **subject** is always purple, and **object** is always yellow
    - "The **huckleberry** ***blorked*** the **sun**"
- Fine-tune (no weight freezing) on a set of active examples (like above)
    - Add a loss term for KL divergence to prevent catastrophic forgetting
- Eval on a broad set of generalizations:

Experiment 2: Setup

- Take 3 models (BERT, DistilBERT, RoBERTa) and teach them some new **verbs** with **unique selectional preferences (unlike any known verb)**
  - ***blork***: **subject** is always purple, and **object** is always yellow
  - "The **huckleberry** ***blorked*** the **sun**"
- Fine-tune (no weight freezing) on a set of active examples (like above)
  - Add a loss term for KL divergence to prevent catastrophic forgetting
- Eval on a broad set of generalizations:
  - The ______ was ***blorked*** by the _______

Experiment 2: Setup

- Take 3 models (BERT, DistilBERT, RoBERTa) and teach them some new **verbs** with **unique selectional preferences (unlike any known verb)**
  - ***blork***: **subject** is always purple, and **object** is always yellow
  - "The **huckleberry** ***blorked*** the **sun**"
- Fine-tune (no weight freezing) on a set of active examples (like above)
  - Add a loss term for KL divergence to prevent catastrophic forgetting
- Eval on a broad set of generalizations:
  - The [obj] was ***blorked*** by the [subj]

Experiment 2: Setup

- Take 3 models (BERT, DistilBERT, RoBERTa) and teach them some new **verbs** with **unique selectional preferences (unlike any known verb)**
  - ***blork***: **subject** is always purple, and **object** is always yellow
  - "The **huckleberry** ***blorked*** the **sun**"
- Fine-tune (no weight freezing) on a set of active examples (like above)
  - Add a loss term for KL divergence to prevent catastrophic forgetting
- Eval on a broad set of generalizations:
  - The __[obj]__ was ***blorked*** by the __[subj]__
  - It was the _______ that the ________ ***blorked***

Experiment 2: Setup

- Take 3 models (BERT, DistilBERT, RoBERTa) and teach them some new **verbs** with **unique selectional preferences (unlike any known verb)**
  - ***blork***: **subject** is always purple, and **object** is always yellow
  - "The **huckleberry** ***blorked*** the **sun**"
- Fine-tune (no weight freezing) on a set of active examples (like above)
  - Add a loss term for KL divergence to prevent catastrophic forgetting
- Eval on a broad set of generalizations:
  - The __[obj]__ was ***blorked*** by the __[subj]__
  - It was the __[obj]__ that the __[subj]__ ***blorked***

Experiment 2: Setup

- Take 3 models (BERT, DistilBERT, RoBERTa) and teach them some new **verbs** with **unique selectional preferences (unlike any known verb)**
  - ***blork***: **subject** is always purple, and **object** is always yellow
  - "The **huckleberry** ***blorked*** the **sun**"
- Fine-tune (no weight freezing) on a set of active examples (like above)
  - Add a loss term for KL divergence to prevent catastrophic forgetting
- Eval on a broad set of generalizations:
  - The __[obj]__ was ***blorked*** by the __[subj]__
  - It was the __[obj]__ that the __[subj]__ ***blorked***
  - The ______ always ***blorked*** the ______

Experiment 2: Setup

- Take 3 models (BERT, DistilBERT, RoBERTa) and teach them some new **verbs** with **unique selectional preferences (unlike any known verb)**
  - ***blork***: **subject** is always purple, and **object** is always yellow
  - "The **huckleberry** ***blorked*** the **sun**"

- Fine-tune (no weight freezing) on a set of active examples (like above)

  - Add a loss term for KL divergence to prevent catastrophic forgetting

- Eval on a broad set of generalizations:
  - The __[obj]__ was ***blorked*** by the __[subj]__
  - It was the __[obj]__ that the __[subj]__ ***blorked***
  - The __[subj]__ always ***blorked*** the __[obj]__

Experiment 2: Results

Experiment 2: Results

- Models are **unable** to robustly generalize the argument position of novel verbs

Experiment 2: Results

- Models are **unable** to robustly generalize the argument position of novel verbs

- They succeed **only when the linear ordering** of arguments in the generalization set matches what was learned during training

Experiment 2: Results

- Models are **unable** to robustly generalize the argument position of novel verbs
- They succeed **only when the linear ordering** of arguments in the generalization set matches what was learned during training

	Fine-tune acc.	Active SO	Active OS	Passive SO	Passive OS
RoBERTa	79.2	66.7	53.6	44.8	39.5
BERT	86.1	75.5	54.2	54.4	55.7
DistilBERT	88.9	74.8	49.4	51.9	39.9
MultiBERT 00	80.6	63.9	50.8	48.7	40.8
MultiBERT 05	84.7	80.4	49.8	51.0	21.4
MultiBERT 10	82.6	70.1	55.2	44.1	42.7
MultiBERT 15	76.4	68.8	49.7	53.3	43.6
MultiBERT 20	79.2	66.4	65.2	41.0	43.0

Experiment 2: Results

- Models are **unable** to robustly generalize the argument position of novel verbs
- They succeed **only when the linear ordering** of arguments in the generalization set matches what was learned during training

The [subj] *blorked* the [obj]

Which [obj] has the [subj] *blorked*

Which [subj] was the [obj] *blorked* by

The [obj] was *blorked* by the [subj]

	Fine-tune acc.	Active SO	Active OS	Passive SO	Passive OS
RoBERTa	79.2	66.7	53.6	44.8	39.5
BERT	86.1	75.5	54.2	54.4	55.7
DistilBERT	88.9	74.8	49.4	51.9	39.9
MultiBERT 00	80.6	63.9	50.8	48.7	40.8
MultiBERT 05	84.7	80.4	49.8	51.0	21.4
MultiBERT 10	82.6	70.1	55.2	44.1	42.7
MultiBERT 15	76.4	68.8	49.7	53.3	43.6
MultiBERT 20	79.2	66.4	65.2	41.0	43.0

Echoes of linearity

Echoes of linearity

- We actually observe a similar sensitivity to linear ordering in Experiment 1, although the general success of models on this task hides the effect

Echoes of linearity

- We actually observe a similar sensitivity to linear ordering in Experiment 1, although the general success of models on this task hides the effect

	BERT	RoBERTa	DistilBERT
same order	93%	83%	88%
reversed order	86%	79%	78%
Δ	7 pts	4 pts	10 pts

Experiment 2: Summary

Experiment 2: Summary

- Models don't seem to display T2 generalization of argument structure to arbitrary verbs regardless of thematic role

Experiment 2: Summary

- Models don’t seem to display T2 generalization of argument structure to arbitrary verbs regardless of thematic role

- One possible reason: unlike with T1 generalization, there isn’t anything to bootstrap knowledge onto

Experiment 2: Summary

- Models don't seem to display T2 generalization of argument structure to arbitrary verbs regardless of thematic role
- One possible reason: unlike with T1 generalization, there isn't anything to bootstrap knowledge onto
  - By design, our novel verbs have unique selection preferences

Experiment 2: Summary

- Models don’t seem to display T2 generalization of argument structure to arbitrary verbs regardless of thematic role
- One possible reason: unlike with T1 generalization, there isn’t anything to bootstrap knowledge onto
  - By design, our novel verbs have unique selection preferences
  - No existing embedding which matches these values

Experiment 2: Summary

- Models don’t seem to display T2 generalization of argument structure to arbitrary verbs regardless of thematic role
- One possible reason: unlike with T1 generalization, there isn’t anything to bootstrap knowledge onto
  - By design, our novel verbs have unique selection preferences
  - No existing embedding which matches these values
  - Unable to make generalizations on the basis of structure alone, independent from a known similar context

Main Q: Where does LLM generalization fit on this chart?

T2

I *verbed* the *X* with the *Y* →

- The *X* was *verbed* by me with the *Y*
- It was the *X* that the I *verbed* with the *Y*

T1

I sprayed the *paint* on the *wall* →

- I sprayed the *wall* with the *paint*
- The *paint* was sprayed onto the *wall*
- It was the *paint* that the workers sprayed the *wall* with

T0

"Distributional generalization" == masked language modeling!

Main Q: Where does LLM generalization fit on this chart?

T2

I *verbed* the *X* with the *Y* →

- The *X* was *verbed* by me with the *Y*
- It was the *X* that the I *verbed* with the *Y*

T1

I sprayed the *paint* on the *wall* →

- I sprayed the *wall* with the *paint*
- The *paint* was sprayed onto the *wall*
- It was the *paint* that the workers sprayed the *wall* with

T0

"Distributional generalization" == masked language modeling!

Main Q: Where does LLM generalization fit on this chart?

T2

I *verbed* the *X* with the *Y* →

- The *X* was *verbed* by me with the *Y*
- It was the *X* that the I *verbed* with the *Y*

T1

I sprayed the *paint* on the *wall* →

- I sprayed the *wall* with the *paint*
- The *paint* was sprayed onto the *wall*
- It was the *paint* that the workers sprayed the *wall* with

T0

"Distributional generalization" == masked language modeling!

Questions?

Thank you to for coming to our talk!