Hu et al., 2020
Sinha et al., 2019

Greta Tuckute & Kamoya K Ikhofua

MIT Fall 2020, 6.884
Symbolic Generalization

Motivation

Natural language understanding systems to generalize in a systematic and robust way

- Diagnostic tests - how can we probe these generalization abilities?
  - **Syntactic generalization** (Hu et al., 2020, “SG”) and **logical reasoning** (Sinha et al., 2019, “CLUTRR”)

- Evaluation metrics for language models?

SG: Man shall not live by perplexity alone

Perplexity **is not sufficient** to check for human-like syntactic knowledge:

- It basically measures the probability of seeing some collection of words together
- However some words which are rarely seen together are grammatically correct
- *Colorless green ideas sleep furiously* (Chomsky, 1957)
- Need a **more fine-grained** way to assess learning outcomes of neural language models

SG: Paradigm

Assess NL models on custom sentences designed using psycholinguistic and syntax literature/methodology

- Compare critical sentence regions NOT full-sentence probabilities.
- Factor out confounds (e.g token lexical frequency, n-gram statistics)

SG: Paradigm

- Cover the scope of syntax phenomena: 16/47 (Carnie et al., 2012)
- Group syntax phenomena into 6 circuits based on processing algorithm

SG: Circuits

1. Agreement
2. Licensing
3. Garden-Path Effects
4. Gross Syntactic Expectation
5. Center Embedding
6. Long-Distance Dependencies

SG: Agreement

(A) The farmer that the clerks embarrassed knows$_{V_{sg}}$ many people.

(B) *The farmer that the clerks embarrassed know$_{V_{pl}}$ many people.

(C) The farmers that the clerk embarrassed know$_{V_{pl}}$ many people.

(D) *The farmers that the clerk embarrassed knows$_{V_{sg}}$ many people.

$$P_A(V_{sg}) > P_B(V_{pl}) \wedge P_C(V_{pl}) > P_D(V_{sg})$$

Chance is 25% (or up to 50%)

SG: NPI Licensing

- The word “any” is a negative polarity item (NPI)
- The word “no” can license an NPI when it structurally commands it, such as in A

A) **No** managers that respected the guard have had **any** luck

>

B) *The managers {that respected **no** guard} have had **any** luck

(Reflexive Pronoun Licensing was also included in sub-class suites)

SG: NPI Licensing

(A) No managers that respected the guard have had $\overbrace{\text{any}}^{\text{NPI}}$ luck. [+NEG,−DISTRACTOR]

(B) *The managers that respected no guard have had $\overbrace{\text{any}}^{\text{NPI}}$ luck. [−NEG,+DISTRACTOR]

(C) *The managers that respected the guard have had $\overbrace{\text{any}}^{\text{NPI}}$ luck. [−NEG,−DISTRACTOR]

(D) No managers that respected no guard have had $\overbrace{\text{any}}^{\text{NPI}}$ luck. [+NEG,+DISTRACTOR]

$$P_{\mathrm{A}}(\mathrm{NPI}) > P_{\mathrm{C}}(\mathrm{NPI}) \wedge P_{\mathrm{D}}(\mathrm{NPI}) > P_{\mathrm{B}}(\mathrm{NPI}) \wedge P_{\mathrm{A}}(\mathrm{NPI}) > P_{\mathrm{B}}(\mathrm{NPI})$$

Acceptable orderings:

ADBC
ADCB
DABC
DACB
ACDB (?)

Chance: 5/24

SG: Reflexive Pronoun Licensing

(A) The author that the senators liked hurt herself$_{R_{sg.fem}}$.

(B) *The authors that the senator liked hurt herself$_{R_{sg.fem}}$.

(C) The authors that the senator liked hurt themselves$_{R_{pl}}$.

(D) *The author that the senator liked hurt themselves$_{R_{pl}}$.

$$P_A(R_{sg}) > P_B(R_{sg}) \wedge P_C(R_{pl}) > P_D(R_{pl})$$

Chance: 25%

SG: NP/Z Garden-Paths

(A) !As the ship crossed the waters $\overbrace{\text{remained}}^{\text{V}^*}$ blue and calm. [TRANS,NO COMMA]

(B) As the ship crossed, the waters $\overbrace{\text{remained}}^{\text{V}^*}$ blue and calm. [TRANS,COMMA]

(C) As the ship drifted the waters $\overbrace{\text{remained}}^{\text{V}^*}$ blue and calm. [INTRANS,NO COMMA]

(D) As the ship drifted, the waters $\overbrace{\text{remained}}^{\text{V}^*}$ blue and calm. [INTRANS,COMMA]

$$S_{\text{A}}(\text{V}^*) > S_{\text{B}}(\text{V}^*) \wedge S_{\text{A}}(\text{V}^*) > S_{\text{C}}(\text{V}^*) \wedge$$
$$S_{\text{A}}(\text{V}^*) - S_{\text{B}}(\text{V}^*) > S_{\text{C}}(\text{V}^*) - S_{\text{D}}(\text{V}^*)$$

SG: Main-Verb Reduced Relative Garden-Paths

(A) !The child kicked in the chaos $\overbrace{\text{found}}^{\text{V*}}$ her way back home. [REDUCED, AMBIG]

(B) The child who was kicked in the chaos $\overbrace{\text{found}}^{\text{V*}}$ her way back home.

(C) The child forgotten in the chaos $\overbrace{\text{found}}^{\text{V*}}$ her way back home.

(D) The child who was forgotten in the chaos $\overbrace{\text{found}}^{\text{V*}}$ her way back home.

$$S_A(V^*) > S_B(V^*) \wedge S_A(V^*) > S_C(V^*) \wedge$$
$$S_A(V^*) - S_B(V^*) > S_C(V^*) - S_D(V^*)$$

Chance is 25%

SG: Gross Syntactic Expectation *(Subordination)*

(A) The minister praised the building $\overbrace{\text{ . }}^{\text{END}}$

(B) *After the minister praised the building $\overbrace{\text{ . }}^{\text{END}}$

(C) ??The minister praised the building, $\overbrace{\text{it started to rain.}}^{\text{MC}}$

(D) After the minster praised the building, $\overbrace{\text{it started to rain.}}^{\text{MC}}$

$$P_{\text{A}}(\text{END}) > P_{\text{B}}(\text{END}) \wedge P_{\text{D}}(\text{MC}) < P_{\text{C}}(\text{MC})$$

SG: Center Embedding

(A) The painting$_{N_1}$ that the artist$_{N_2}$ who lived long ago painted$_{V_2}$ deteriorated$_{V_1}$. [correct]

(B) #The painting$_{N_1}$ that the artist$_{N_2}$ who lived long ago deteriorated$_{V_1}$ painted$_{V_2}$. [incorrect]

$$P_{\mathrm{A}}(\mathrm{V}_2\mathrm{V}_1) > P_{\mathrm{B}}(\mathrm{V}_1\mathrm{V}_2)$$

$$P(\text{painted deteriorated}|\text{The painting that the artist}) >$$
$$P(\text{deteriorated painted}|\text{The painting that the artist})$$

SG: Long Distance Dependencies

(A) I know that our uncle grabbed $\overbrace{\text{the food}}^{\alpha}$ in front of the guests at the holiday party. [THAT, NO GAP]

(B) *I know what our uncle grabbed $\overbrace{\text{the food}}^{\alpha}$ in front of the guests at the holiday party. [WH, NO GAP]

(C) ??I know that our uncle grabbed $\overbrace{\text{in front of}}^{\beta}$ the guests at the holiday party. [THAT, GAP]

(D) I know what our uncle grabbed $\overbrace{\text{in front of}}^{\beta}$ in front of the guests at the holiday party. [WH, GAP]

$$S_{\mathrm{B}}(\alpha) > S_{\mathrm{A}}(\alpha) \wedge S_{\mathrm{C}}(\beta) > S_{\mathrm{D}}(\beta)$$

SG: Pseudo-Clefting

(A) What the worker did was $\overbrace{\text{board the plane.}}^{\text{VP}}$

(B) ?What the worker did was $\overbrace{\text{the plane.}}^{\text{NP}}$

(C) What the worker repaired was $\overbrace{\text{the plane.}}^{\text{NP}}$

(D) *What the worker repaired was $\overbrace{\text{board the plane}}^{\text{VP}}$.

$$S_{\mathrm{D}}(\mathrm{VP}) > S_{\mathrm{A}}(\mathrm{VP}) \wedge S_{\mathrm{B}}(\mathrm{NP}) > S_{\mathrm{C}}(\mathrm{NP})$$

SG: Assessment

accuracy_per_test_suite = correct predictions / total items

- Test for stability by including syntactically irrelevant but semantically plausible syntactic content before the critical region
    - E.g:
    - The keys to the cabinet on the left are on the table
    - *The keys to the cabinet on the left is on the table

- Compare model class to dataset size

SG: Score by Model Class

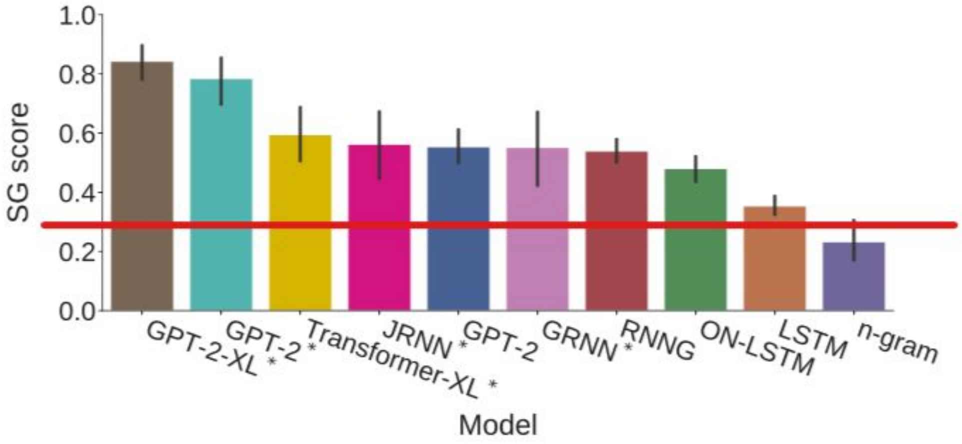
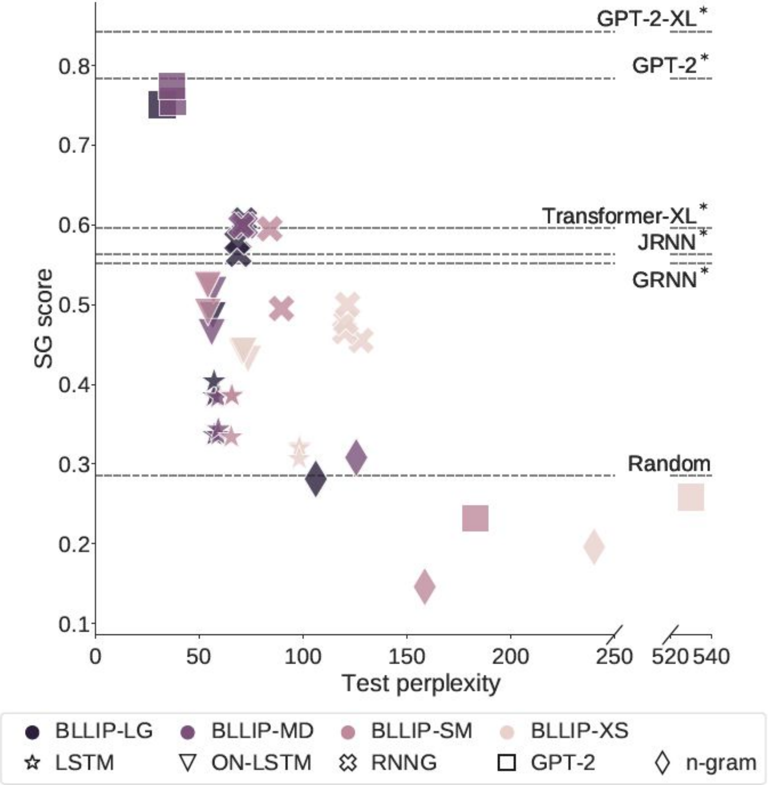


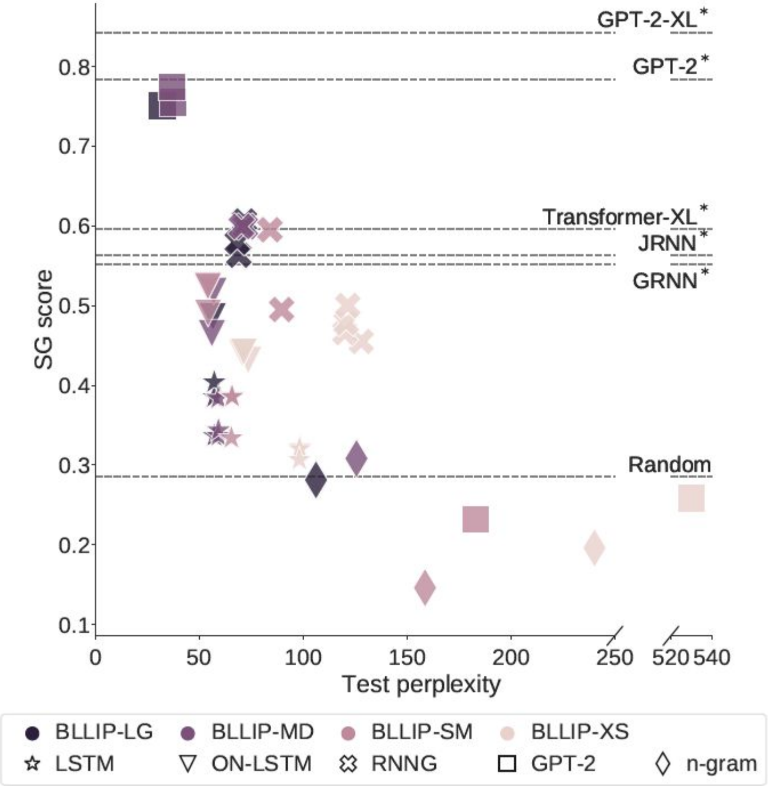
Figure 1: Average SG score by model class. Asterisks denote off-the-shelf models. Error bars denote bootstrapped 95% confidence intervals of the mean.

SG: Perplexity and SG Score

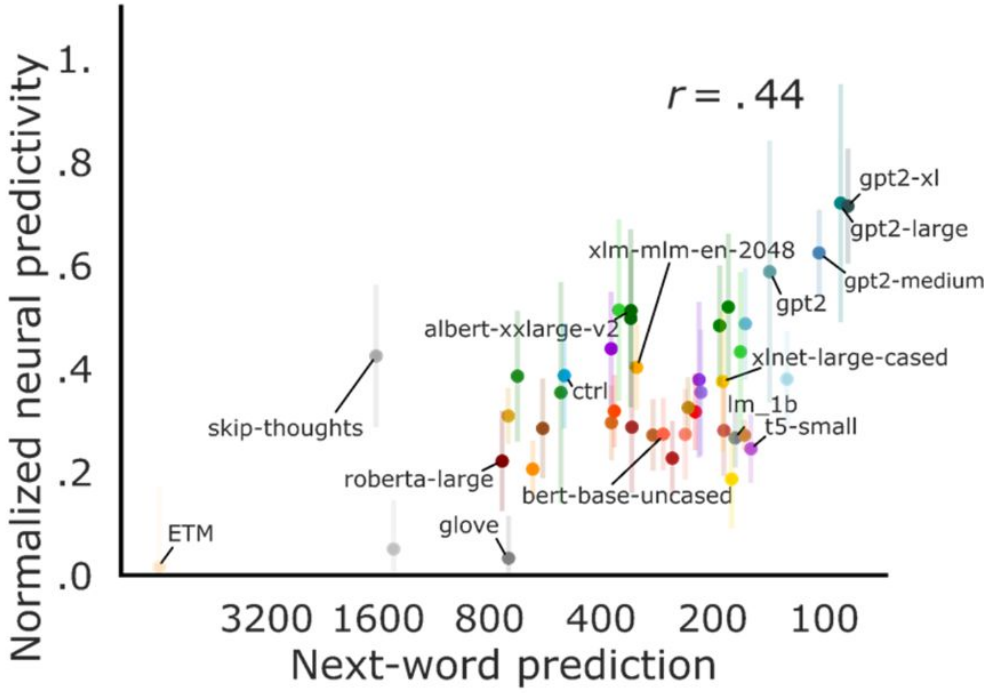


- BLLIP-XS: 1M tokens
- BLLIP-S: 5M tokens
- BLLIP-M: 14M tokens
- BLLIP-LG: 42M tokens

SG: Perplexity and SG Score

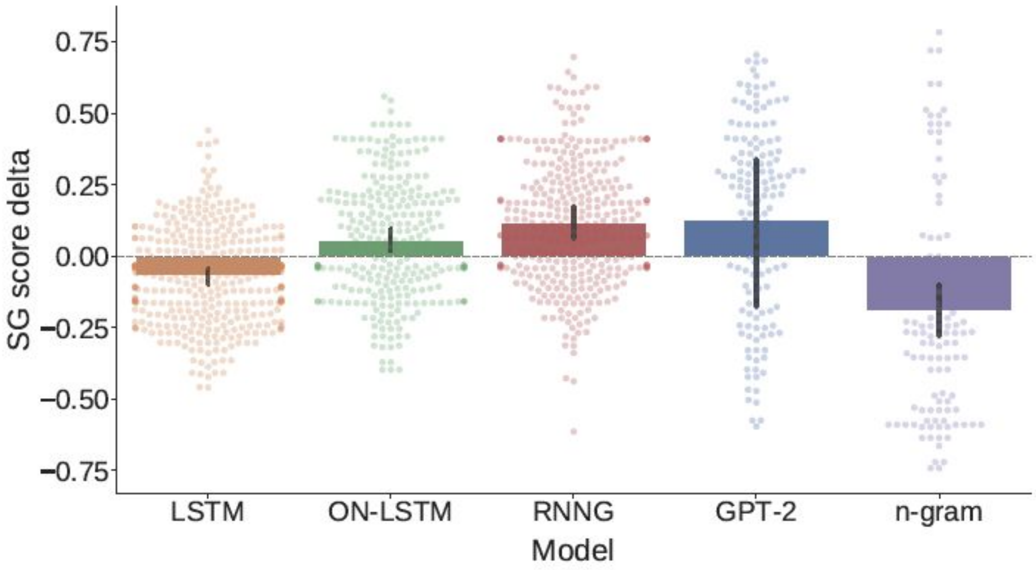


SG: Perplexity and Brain-Score



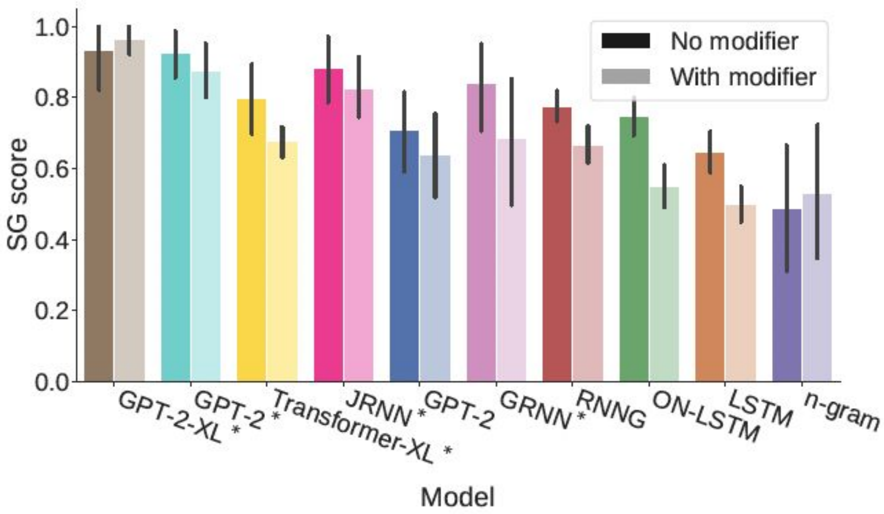
Schrimpf et al., 2020

SG: The Influence of Model Architecture

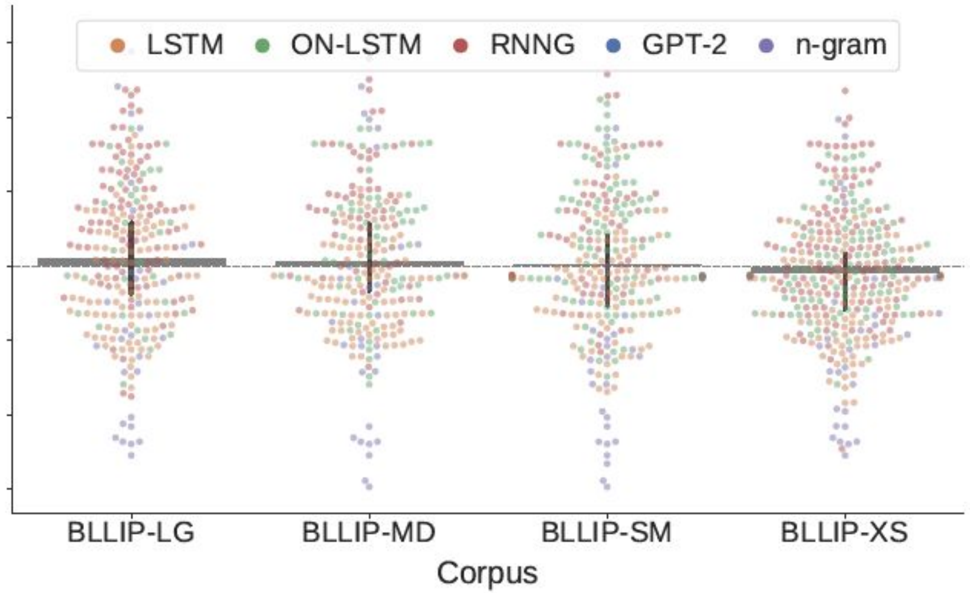


SG: The Influence of Model Architecture

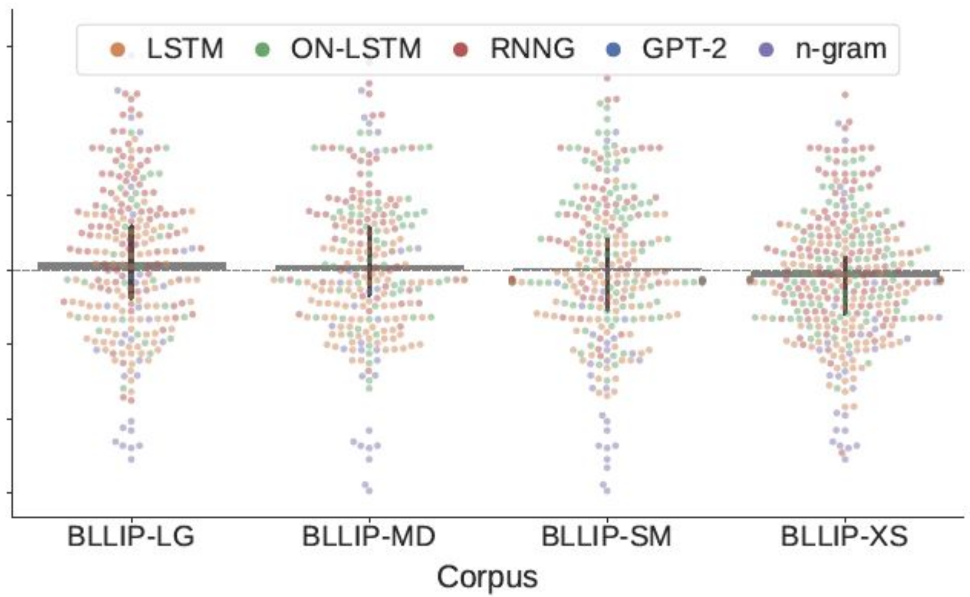
- Architectures as priors to the linguistic representation that can be developed
- Robustness depends on model architecture

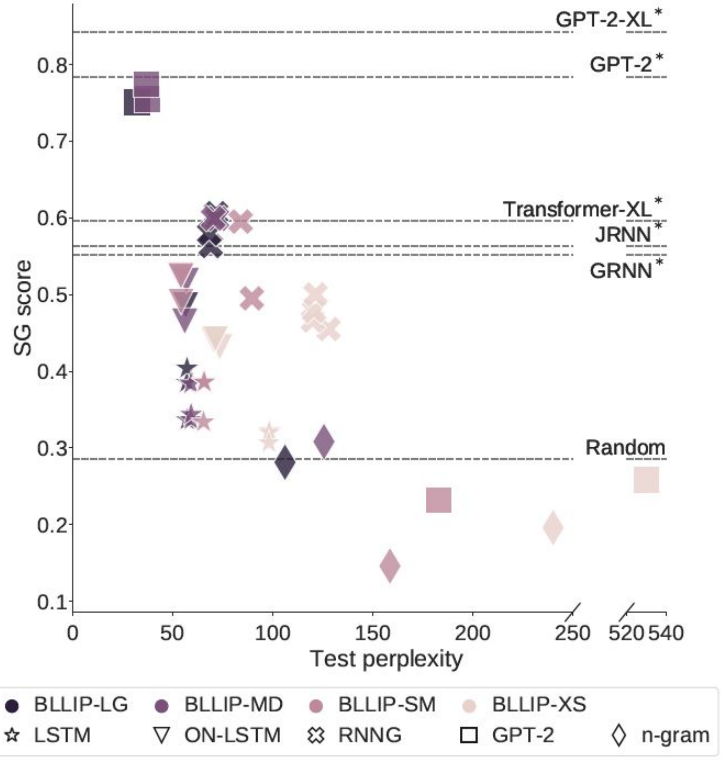


SG: The Influence of Dataset Size



SG: The Influence of Dataset Size





SG: The Influence of Dataset Size

- Increasing amount of training data yields diminishing returns:
  - *"(...) require over 10 billion tokens to achieve human-like performance, and most would require trillions of tokens to achieve perfect accuracy – an impractically large amount of training data, especially for these relatively simple syntactic phenomena."* (van Schijndel et al., 2019)
- Limited data efficiency
- Structured architectures or explicit syntactic supervision
- Humans? 11-27 million total words of input per year? (Hart & Risley, 1995; Brysbaert et al., 2016)

SG: The Influence of Dataset Size

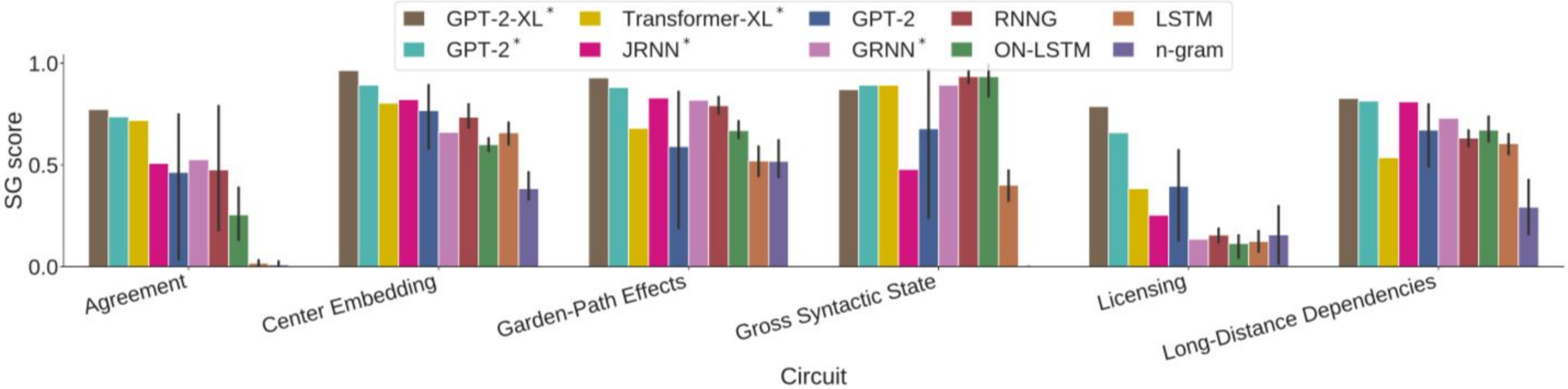
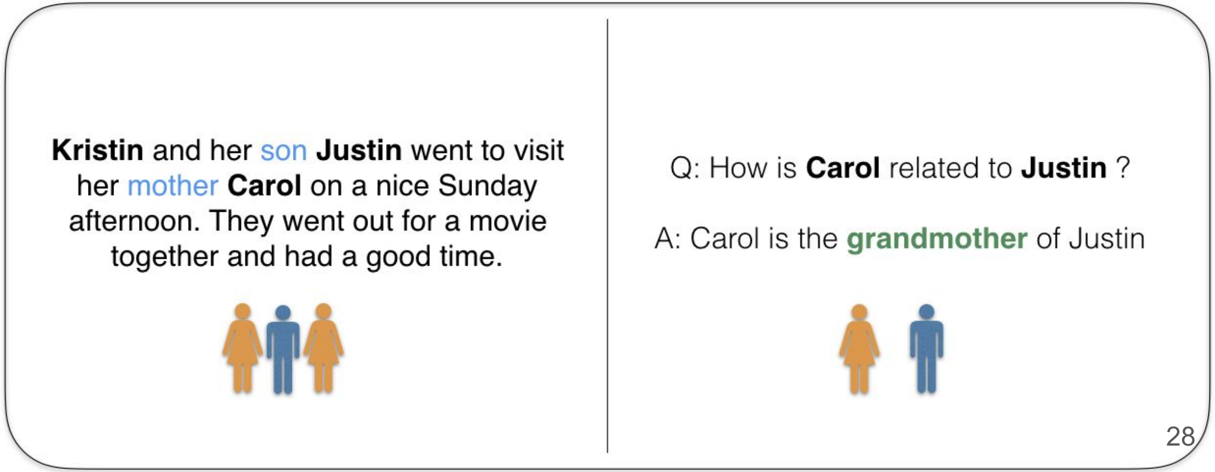


Figure 5: Evaluation results on all models, split across test suite circuits.

CLUTRR: Motivation and Paradigm

- **C**ompositional **L**anguage **U**nderstanding and **T**ext-based **R**elational **R**easoning
- Kinship inductive reasoning

- Unseen combinations of logical rules
- Model robustness

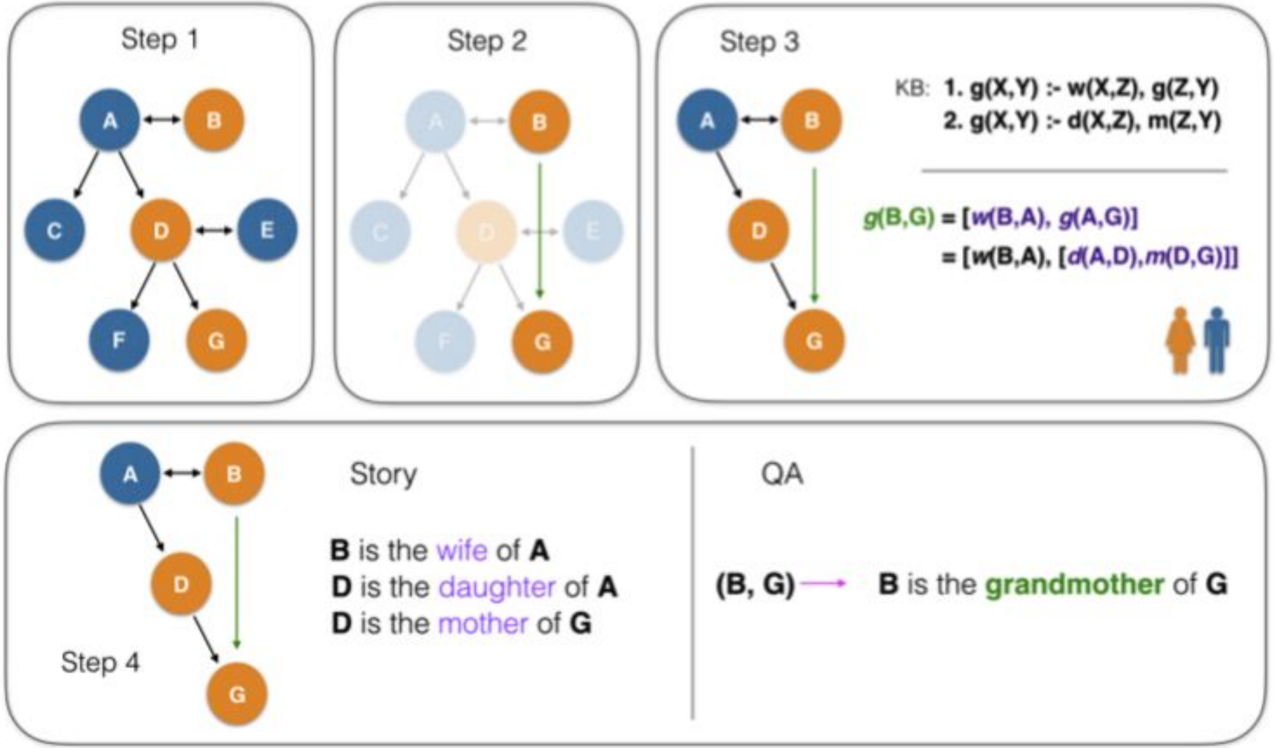


CLUTRR: Motivation and Paradigm

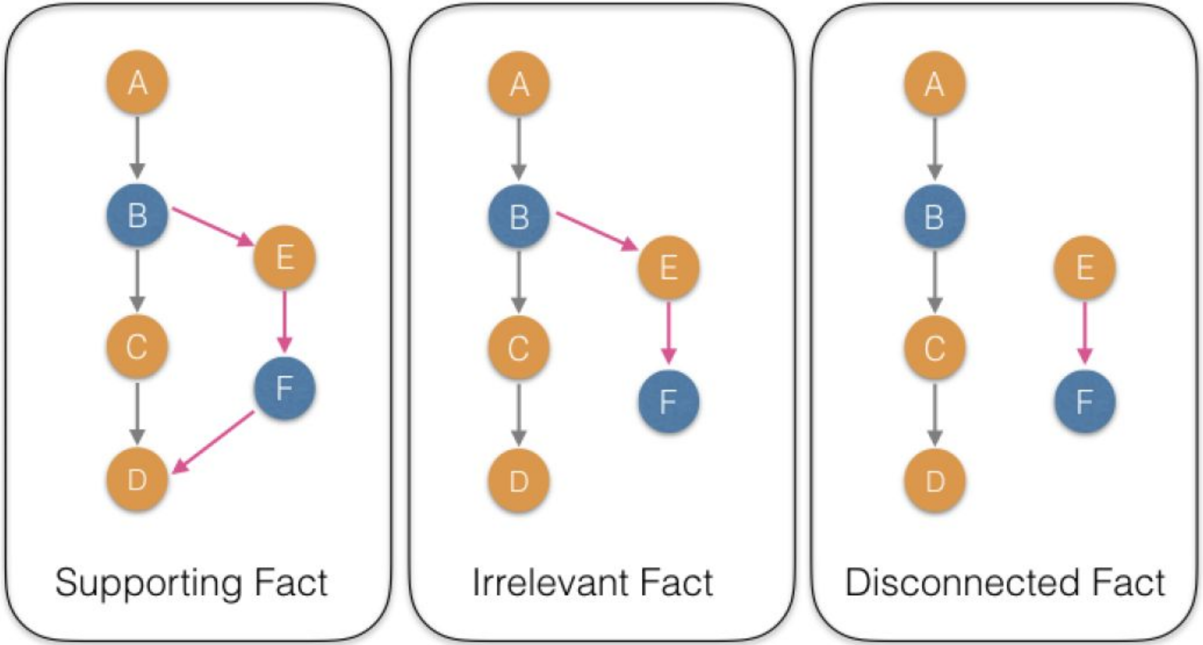
- Productivity
  - mother(mother(mother(Justin))) ~ great grandmother of Justin
- Systematicity
  - Only certain sets allowed with symmetries: son(Justin, Kristin) ~ mother(Kristin, Justin)
- Compositionality
  - son(Justin, Kristin) consists of components

- Memory (compression)
- Children are not exposed to systematic dataset

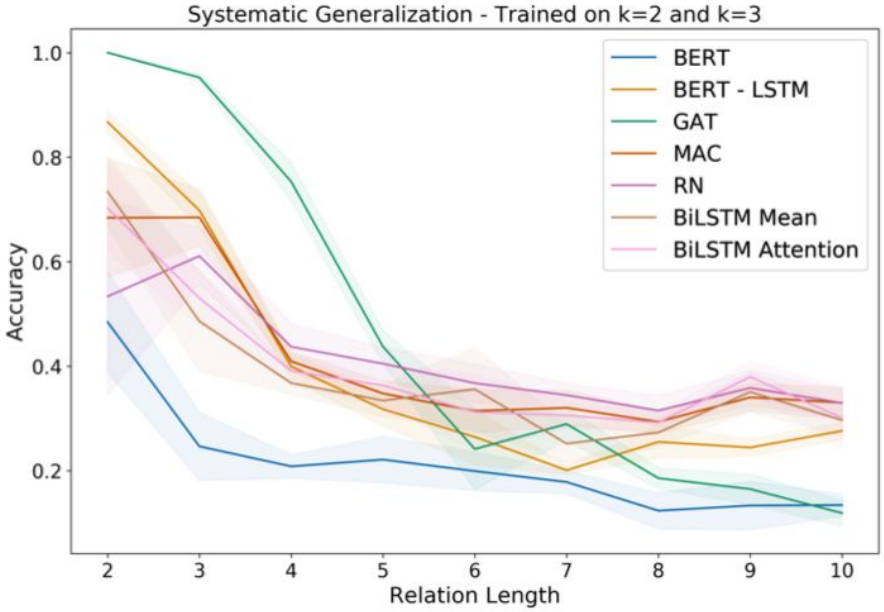
CLUTRR: Dataset Generation & Paradigm

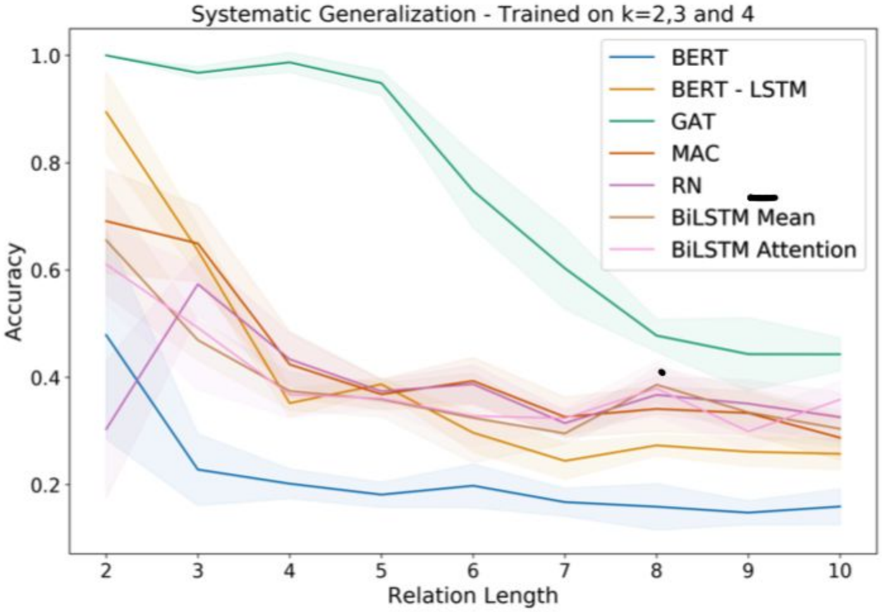


CLUTRR: Model Robustness



CLUTRR: Systematic Generalization





CLUTRR: Model Robustness

Models		Unstructured models (no graph)						Structured model (with graph)
Training	Testing	BiLSTM - Attention	BiLSTM - Mean	RN	MAC	BERT	BERT-LSTM	GAT
Clean	Clean	0.58 $\pm$0.05	0.53 $\pm$0.05	0.49 $\pm$0.06	0.63 $\pm$0.08	0.37 $\pm$0.06	0.67 $\pm$0.03	**1.0** $\pm$0.0
	Supporting	**0.76** $\pm$0.02	0.64 $\pm$0.22	0.58 $\pm$0.06	0.71 $\pm$0.07	0.28 $\pm$0.1	0.66 $\pm$0.06	0.24 $\pm$0.2
	Irrelevant	0.7 $\pm$0.15	**0.76** $\pm$0.02	0.59 $\pm$0.06	0.69 $\pm$0.05	0.24 $\pm$0.08	0.55 $\pm$0.03	0.51 $\pm$0.15
	Disconnected	0.49 $\pm$0.05	0.45 $\pm$0.05	0.5 $\pm$0.06	0.59 $\pm$0.05	0.24 $\pm$0.08	0.5 $\pm$0.06	**0.8** $\pm$0.17
Supporting	Supporting	0.67 $\pm$0.06	0.66 $\pm$0.07	0.68 $\pm$0.05	0.65 $\pm$0.04	0.32 $\pm$0.09	0.57 $\pm$0.04	**0.98** $\pm$0.01
Irrelevant	Irrelevant	0.51 $\pm$0.06	0.52 $\pm$0.06	0.5 $\pm$0.04	0.56 $\pm$0.04	0.25 $\pm$0.06	0.53 $\pm$0.06	**0.93** $\pm$0.01
Disconnected	Disconnected	0.57 $\pm$0.07	0.57 $\pm$0.06	0.45 $\pm$0.11	0.4 $\pm$0.1	0.17 $\pm$0.05	0.47 $\pm$0.06	**0.96** $\pm$0.01
Average		**0.61** $\pm$0.08	0.59 $\pm$0.08	0.54 $\pm$0.07	**0.61** $\pm$0.06	0.30 $\pm$0.07	0.56 $\pm$0.05	**0.77** $\pm$0.09

CLUTRR: Model Robustness (noisy training)

Models		Unstructured models (no graph)						Structured model (with graph)
Training	Testing	BiLSTM - Attention	BiLSTM - Mean	RN	MAC	BERT	BERT-LSTM	GAT
Supporting	Clean	0.38 ± 0.04	0.32 ± 0.04	0.4 ± 0.09	0.45 ± 0.03	0.19 ± 0.06	0.39 ± 0.06	**0.92** ± 0.17
	Supporting	0.67 ± 0.06	0.66 ± 0.07	0.68 ± 0.05	0.65 ± 0.04	0.32 ± 0.09	0.57 ± 0.04	**0.98** ± 0.01
	Irrelevant	0.44 ± 0.03	0.39 ± 0.03	**0.51** ± 0.08	0.46 ± 0.09	0.2 ± 0.06	0.36 ± 0.05	0.5 ± 0.23
	Disconnected	0.31 ± 0.21	0.25 ± 0.16	0.47 ± 0.08	0.41 ± 0.06	0.2 ± 0.08	0.32 ± 0.04	**0.92** ± 0.05
Irrelevant	Clean	0.57 ± 0.05	0.56 ± 0.05	0.46 ± 0.13	0.67 ± 0.05	0.24 ± 0.06	0.46 ± 0.08	**0.92** ± 0.0
	Supporting	0.38 ± 0.22	0.31 ± 0.16	0.61 ± 0.07	0.61 ± 0.04	0.27 ± 0.06	0.46 ± 0.04	**0.77** ± 0.12
	Irrelevant	0.51 ± 0.06	0.52 ± 0.06	0.5 ± 0.04	0.56 ± 0.04	0.25 ± 0.06	0.53 ± 0.06	**0.93** ± 0.01
	Disconnected	0.44 ± 0.26	0.54 ± 0.27	0.55 ± 0.05	0.61 ± 0.06	0.26 ± 0.03	0.45 ± 0.08	**0.85** ± 0.25
Disconnected	Clean	0.45 ± 0.02	0.47 ± 0.03	0.53 ± 0.09	0.5 ± 0.06	0.22 ± 0.09	0.44 ± 0.05	**0.75** ± 0.07
	Supporting	0.47 ± 0.03	0.46 ± 0.05	0.54 ± 0.03	0.58 ± 0.06	0.22 ± 0.06	0.38 ± 0.08	**0.78** ± 0.12
	Irrelevant	0.47 ± 0.05	0.48 ± 0.03	0.52 ± 0.04	0.51 ± 0.05	0.17 ± 0.04	0.38 ± 0.05	**0.56** ± 0.26
	Disconnected	0.57 ± 0.07	0.57 ± 0.06	0.45 ± 0.11	0.4 ± 0.1	0.17 ± 0.05	0.47 ± 0.06	**0.96** ± 0.01
Average		0.47 ± 0.08	0.46 ± 0.08	0.52 ± 0.07	**0.53** ± 0.06	0.23 ± 0.07	0.43 ± 0.05	**0.82** ± 0.09

Table 3: Testing the robustness of the various models when trained various types of noisy facts and evaluated on other noisy / clean facts. The types of noise facts (supporting, irrelevant and disconnected) are defined in Section 3.5 of the main paper.

Future work & Perspectives

- Sub-word tokenization
- Active attention and reasoning
- Generalization across tasks
- Abstractions as probabilistic
- Architecture and dimensionality reduction

References

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). How Many Words Do We Know? Practical Estimates of Vocabulary Size Dependent on Word Definition, the Degree of Language Input and the Participant's Age. *Frontiers in psychology*, *7*, 1116. https://doi.org/10.3389/fpsyg.2016.01116

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: Paul H. Brookes Publishing Company.

Schrimpf, M., Blank, I., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J., Fedorenko, E (2020): Artificial Neural Networks Accurately Predict Language Processing in the Brain, *bioRxiv* 2020.06.26.174482; doi: https://doi.org/10.1101/2020.06.26.174482.

Van Schijndel, M., Mueller, A., & Linzen, T. (2019). Quantity doesn't buy quality syntax with neural language models. *arXiv preprint arXiv:1909.00111*.

Supplementary

CLUTTR, Fig. 6

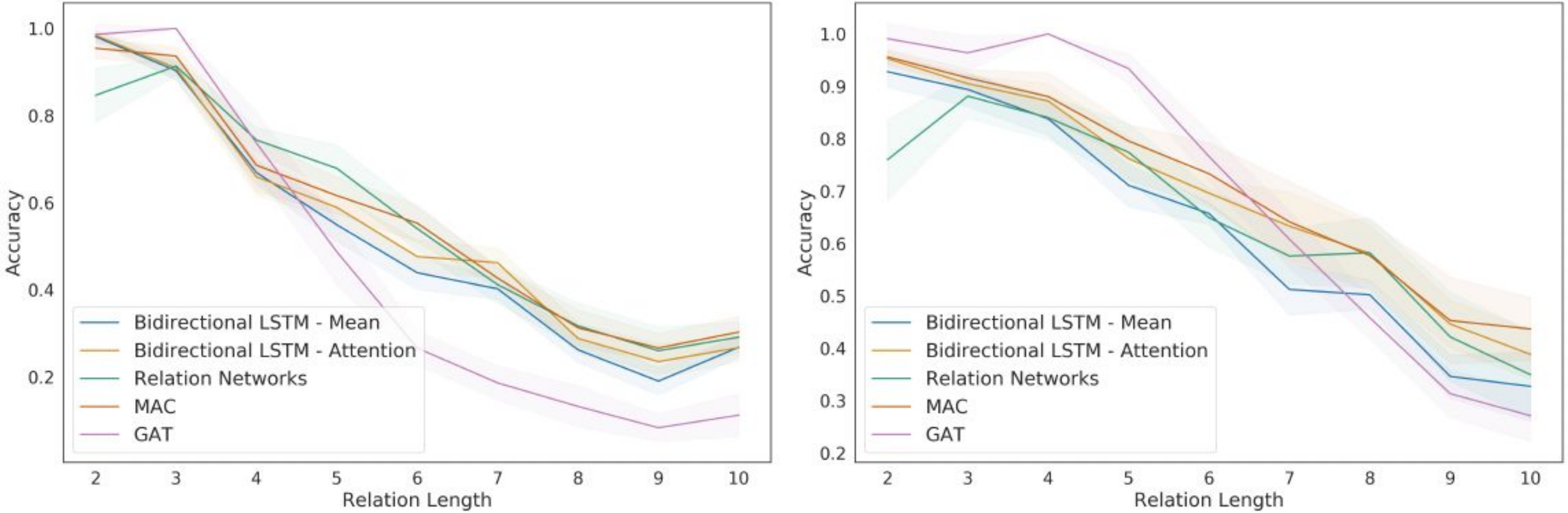


Figure 6: Systematic Generalizability of different models on `CLUTRR-Gen` task (having 20% less placeholders and without training and testing placeholder split), when **Left:** trained with $k = 2$ and $k = 3$ and **Right:** trained with $k = 2, 3$ and 4

CLUTTR, Table 5

Relation Length	Human Performance Time Limited	 Unlimited Time	Reported Difficulty
2	0.848	1	1.488 +- 1.25
3	0.773	1	2.41 +- 1.33
4	0.477	1	3.81 +- 1.46
5	0.424	1	3.78 +- 0.96
6	0.406	1	4.46 +- 0.87

Table 5: Human performance accuracies on CLUTRR dataset. Humans are provided the Clean-Generalization version of the dataset, and we test on two scenarios: when a human is given limited time to solve the task, and when a human is given unlimited time to solve the task. Regardless of time, our evaluators provide a score of difficulty of individual puzzles.

CLUTTR, Table 4

Models		Unstructured models (no graph)						Structured model (with graph)
Training	Testing	BiLSTM - Attention	BiLSTM - Mean	RN	MAC	BERT	BERT-LSTM	GAT
Supporting	Clean	0.96 ±0.01	**0.97** ±0.01	0.88 ±0.05	0.94 ±0.02	0.48 ±0.08	0.57 ±0.08	0.92 ±0.17
	Supporting	0.96 ±0.03	0.96 ±0.03	0.97 ±0.01	0.97 ±0.01	0.75 ±0.07	0.88 ±0.05	**0.98** ±0.01
	Irrelevant	0.92 ±0.02	**0.93** ±0.01	0.9 ±0.03	0.91 ±0.01	0.56 ±0.04	0.54 ±0.06	0.5 ±0.23
	Disconnected	0.8 ±0.04	0.83 ±0.04	0.76 ±0.08	0.86 ±0.04	0.27 ±0.06	0.42 ±0.08	**0.92** ±0.05
Irrelevant	Clean	0.63 ±0.02	0.61 ±0.07	0.85 ±0.09	0.8 ±0.07	0.53 ±0.09	0.44 ±0.06	**0.92** ±0.0
	Supporting	0.66 ±0.03	0.64 ±0.04	0.69 ±0.06	0.76 ±0.06	0.42 ±0.08	0.43 ±0.08	**0.77** ±0.12
	Irrelevant	0.89 ±0.04	0.86 ±0.1	0.74 ±0.11	0.78 ±0.06	0.61 ±0.1	0.83 ±0.06	**0.93** ±0.01
	Disconnected	0.64 ±0.02	0.62 ±0.05	0.72 ±0.05	0.73 ±0.04	0.41 ±0.04	0.61 ±0.05	**0.85** ±0.25
Disconnected	Clean	0.9 ±0.05	0.82 ±0.12	**0.94** ±0.02	0.93 ±0.04	0.68 ±0.07	0.64 ±0.02	0.75 ±0.07
	Supporting	0.87 ±0.04	0.82 ±0.05	0.85 ±0.03	**0.88** ±0.04	0.54 ±0.08	0.5 ±0.05	0.78 ±0.12
	Irrelevant	**0.87** ±0.03	0.85 ±0.03	0.83 ±0.03	0.87 ±0.02	0.59 ±0.09	0.58 ±0.09	0.56 ±0.26
	Disconnected	0.91 ±0.04	0.91 ±0.03	0.8 ±0.17	0.71 ±0.11	0.49 ±0.1	0.79 ±0.1	**0.96** ±0.01
Average		0.83 ±0.08	0.82 ±0.08	0.83 ±0.07	**0.84** ±0.06	0.58 ±0.07	0.60 ±0.05	**0.82** ±0.09

Table 4: Testing the robustness on toy placeholders of the various models when trained various types of noisy facts and evaluated on other noisy / clean facts. The types of noise facts (supporting, irrelevant and disconnected) are defined in Section 3.5 of the main paper.

CLUTTR, Fig. 7

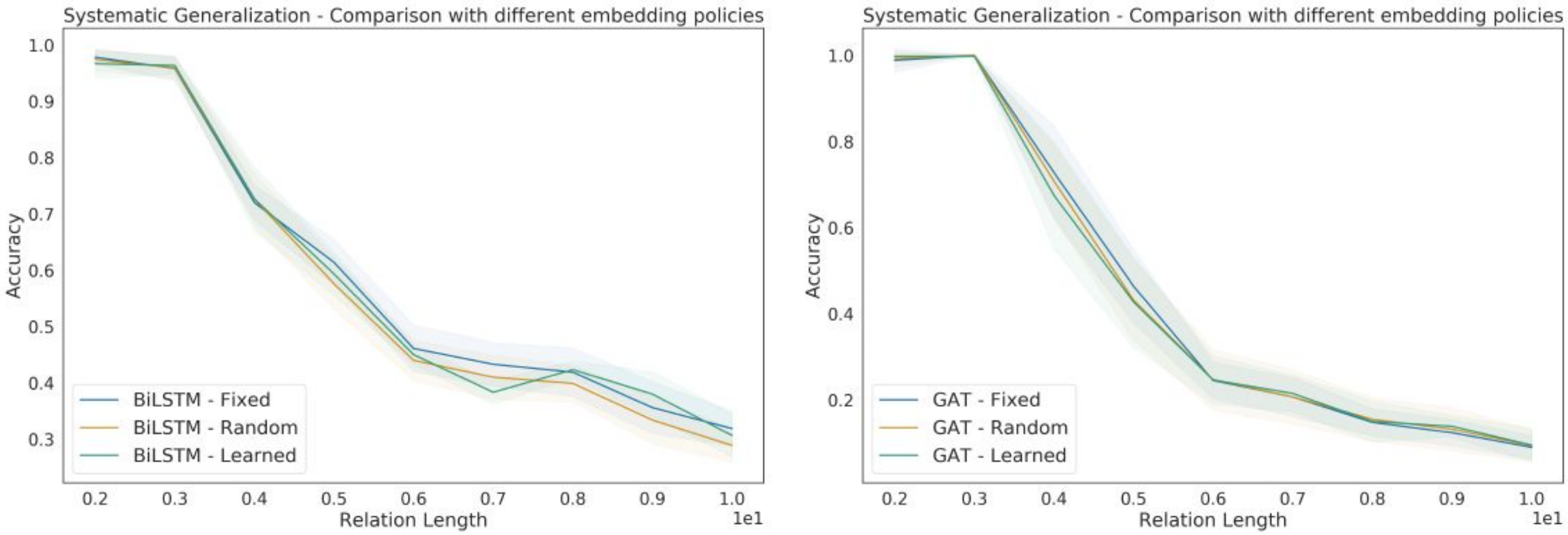


Figure 7: Systematic Generalization comparison with different Embedding policies

Van Schijndel et al., 2019

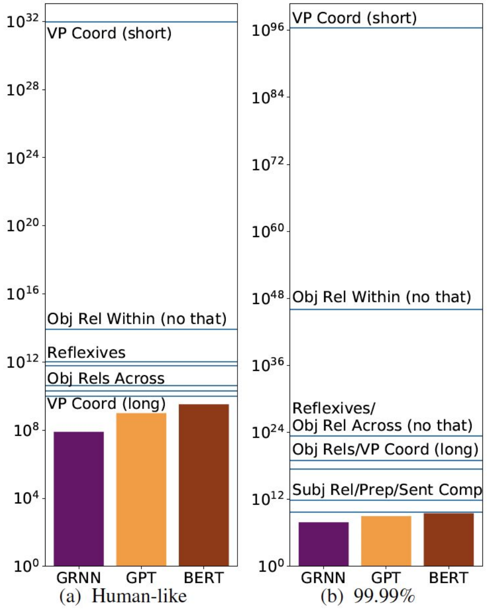


Figure 2: Lines depict number of training tokens needed for LSTMs to achieve human-like (left) or 99.99% accuracy (right) in each syntactic agreement condition, according to our estimates. Bars depict the amount of data on which each model was trained.