Exploring Graph Representations of Logical Forms for Language Modeling

Michael Sullivan
Saarland University
University at Buffalo

https://coli-saar.github.io/gfolds

What is a language model over logical forms (LFLM)?

- **Logical Form:** A sentence in a formal language *L*, such that *L* carries a predicate-argument structure that can be used to represent (aspects of) truth-conditional linguistic meaning.
  - e.g. FOL, (D)MRS (Copestake et al., 2005; Copestake, 2009), AMR (Banarescu et al., 2013), etc.

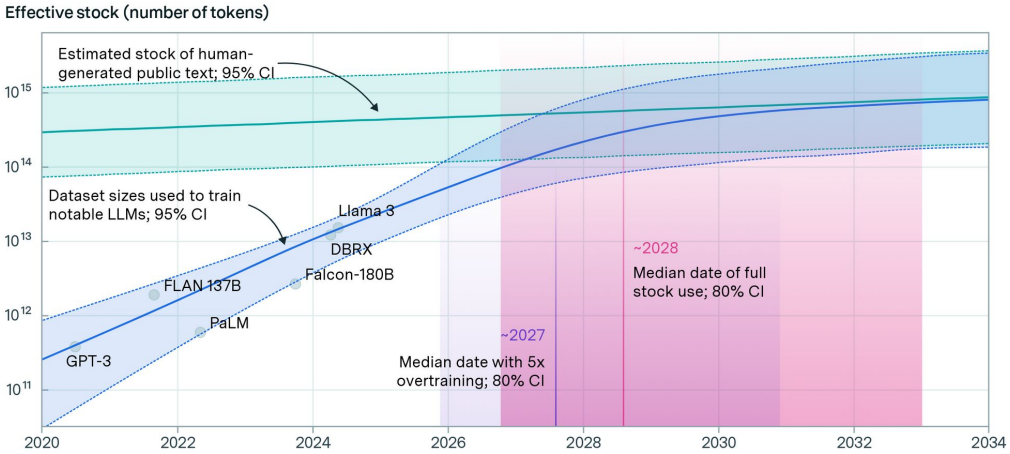
- **Language Model over Logical Forms (LFLM):** a language model that takes as input (sequences of) logical form representations of natural language.

Why LFLMs?

We are running out of training data

- Larger and larger LLMs require more and more training data:
  - GPT-2 (Radford et al., 2018): 1.5 billion parameters
  - GPT-3 (Brown et al., 2020): 175 billion parameters
  - GPT-4 (OpenAI, 2023): >1 trillion parameters (estimated)

- Villalobos et al. (2024): high-quality English training data will be exhausted sometime between 2026 and 2032



https://epoch.ai/blog/will-we-run-out-of-data-limits-of-llm-scaling-based-on-human-generated-data

Why logical forms?

1. **Linguistically-informed LMs can improve over textual models**, without using additional training data (e.g. Xu et al. 2021; Zhou et al., 2020; Zhang et al., 2020; etc.)

2. **Semantics is better than syntax:** semantically-informed LMs outperform syntactically-informed LMs (Wu et al., 2021; Prange et al. 2022)
    - I argue this is due to a syntactic/morphological de-noising effect:

"John saw Mary" ⇒ *see*(*j*, *m*)
"Mary was seen by John" ⇒ *see*(*j*, *m*)

"goose" ⇒ $goose_{\text{NUM:SG}}$
"geese" ⇒ $goose_{\text{NUM:PL}}$

The Linguistic Knowledge Catalysis Hypothesis (LKCH):

> *The (aspects of) linguistic knowledge incorporated into LFLMs greatly accelerates their learning of elementary linguistic phenomena, in turn accelerating the learning of more complex patterns.*

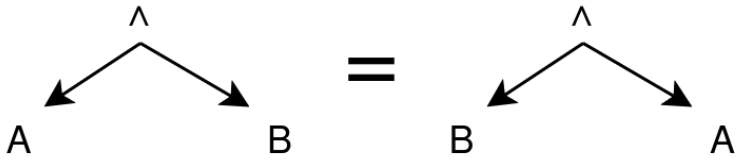
Contributions

- Experimental support for the LKCH
- Proof-of-concept of the downstream applicability of LFLMs
- Evidence supporting the scalability of LFLMs

Graph-based Formal-Logical Distributional Semantics (GFoLDS)

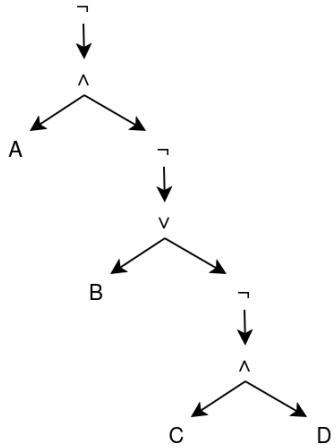
Why graphs?

- Graph representations permit permutation-invariance: the nodes in a graph are not ordered:



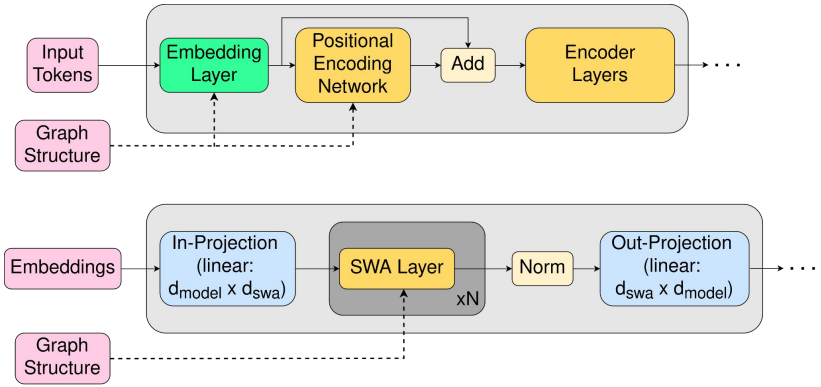
- Graphs facilitate the encoding of hierarchical structures:

$$\neg(A \wedge \neg(B \vee \neg(C \wedge D))) \quad \textit{vs.}$$



Instantiating an LFLM: GFoLDS

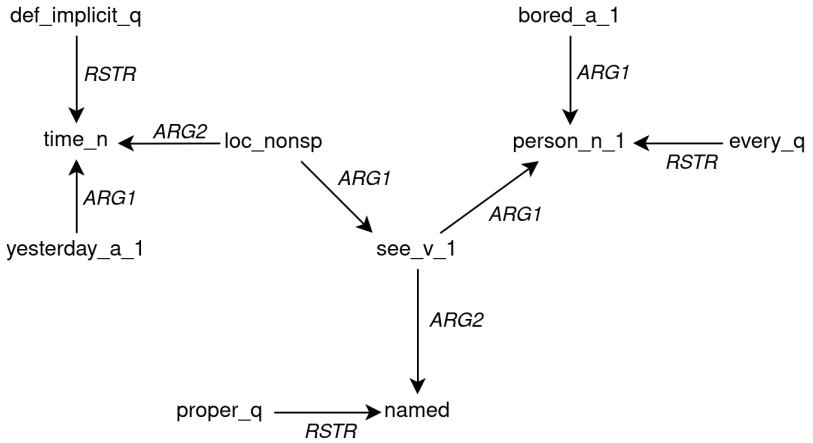
- I introduce the Graph-based Formal-Logical Distributional Semantics (GFoLDS) model, a pretrained graph transformer (Wu et al., 2021) over DMRS-derived representations



Top-level architecture of the GFoLDS model (top) and positional encoding network (bottom)

Pretraining

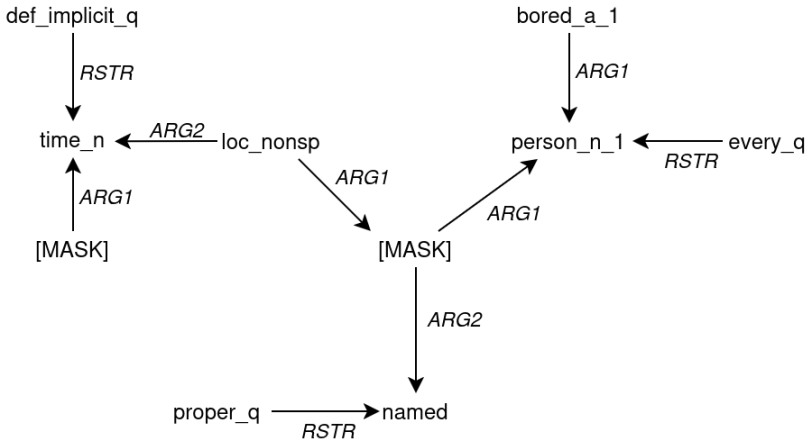
- I pretrained GFoLDS for four epochs on 17.5 million randomly-selected sentences (~84% parsable) from English Wikipedia[1]
  - ~6.5x smaller than BERT's (Devlin et al. 2019) pretraining dataset
- Pretraining objective: masked node modeling (MNM)



[1]https://huggingface.co/datasets/wikimedia/wikipedia/viewer/20231101.en

Pretraining

- I pretrained GFoLDS for four epochs on 17.5 million randomly-selected sentences (~84% parsable) from English Wikipedia[1]
  - ~6.5x smaller than BERT's (Devlin et al. 2019) pretraining dataset
- Pretraining objective: masked node modeling (MNM)



[1]https://huggingface.co/datasets/wikimedia/wikipedia/viewer/20231101.en

Experiments

Experiment 1: Evaluating the LKCH

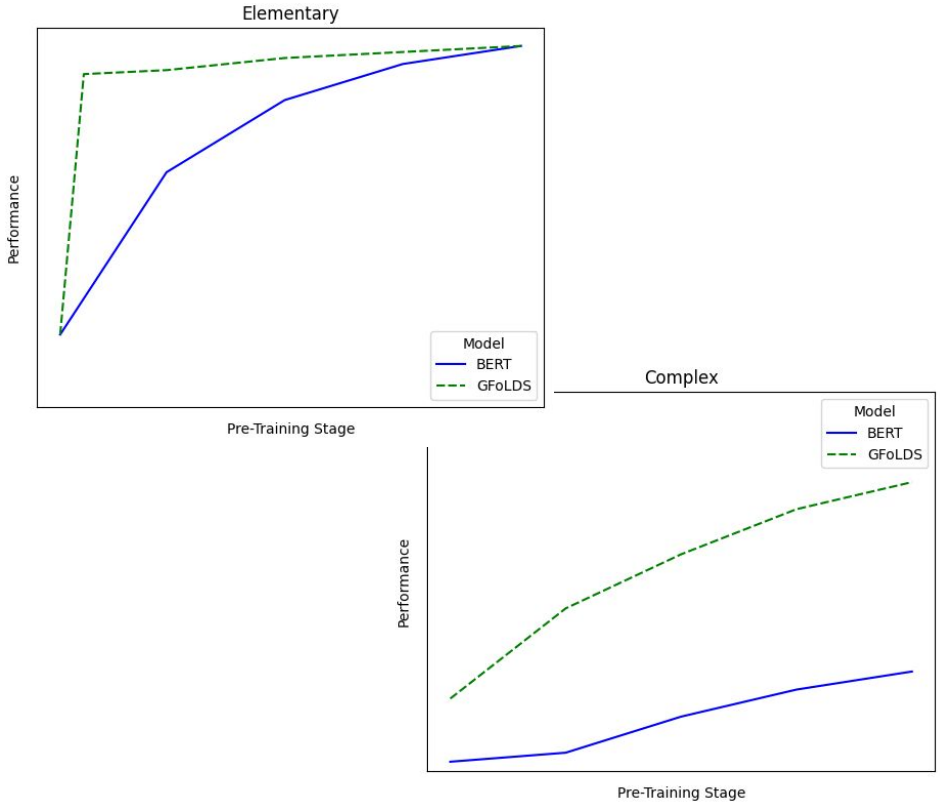
- Evaluated GFoLDS and BERT comparison models at 80 intervals throughout pretraining

- Elementary tasks:
  - POS prediction
  - Quantifier agreement prediction
- Complex task: RELPRON (Rimell et al. 2016)
  - *“a device that astronomers use is a ___”*
    (TARGET = *telescope*)

The Linguistic Knowledge Catalysis Hypothesis (LKCH):

> *The (aspects of) linguistic knowledge incorporated into LFLMs greatly accelerates their learning of elementary linguistic phenomena, in turn accelerating the learning of more complex patterns.*

Prediction

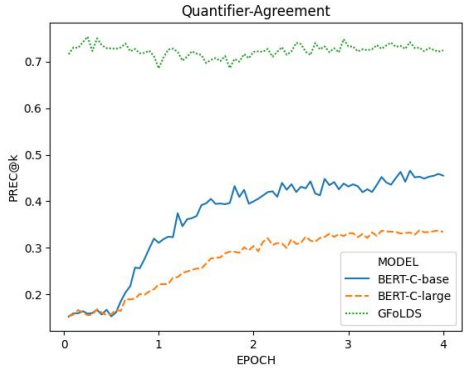


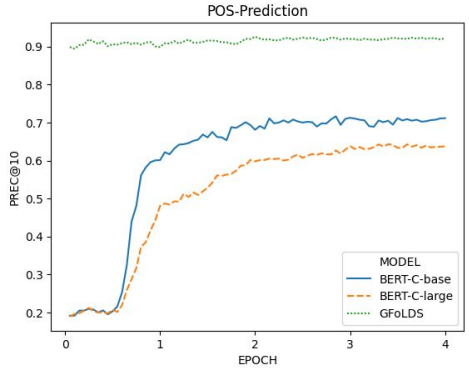
The Linguistic Knowledge Catalysis Hypothesis (LKCH):

The (aspects of) linguistic knowledge incorporated into LFLMs greatly accelerates their learning of elementary linguistic phenomena, in turn accelerating the learning of more complex patterns.

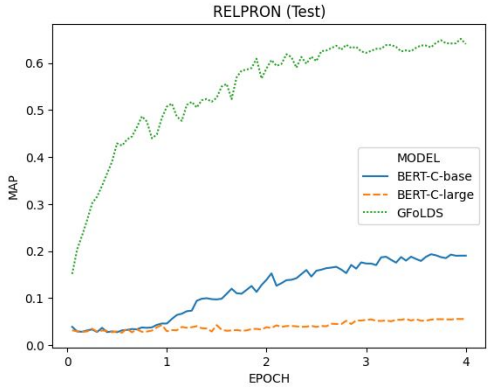
Results

Elementary:





Complex (RELPRON):



Contributions

✅ Experimental support for the LKCH

- Proof-of-concept of the downstream applicability of LFLMs

- Evidence supporting the scalability of LFLMs

Experiment 2: Downstream Tasks

1. RELPRON test set

Term/Hypernym	Properties	Corresponding Templates
telescope/ device	*astronomers use*	"A device that astronomers use is a [redacted]"
	detects planets	"A device that detects planets is a [redacted]"
assignment/ document	*student writes*	"A document that a student writes is an [redacted]"
	receives a grade	"A document that receives a grade is an [redacted]"
ruin/ building	*archaeologist discovers*	"A building that an archaeologist discovers is a [redacted]"
	jungle covers	"A building that the jungle covers is a [redacted]"

2. SNLI (Bowman et al., 2015)

Premise	Hypothesis	Label
It is raining	The ground is wet	*Entailment*
The man is lying down	The man is standing	*Contradiction*
It is December	It is five o'clock	*Neutral*

3. MegaVeridicality V2.1 (White et al., 2018)

Sentence	Label
A particular person didn't mean to do a particular thing	1
Someone didn't tell a particular person to do a particular thing	0
John wasn't upset that a particular thing happened	1
John didn't find that a particular thing happened	0
A particular person was thrilled to do a particular thing	1
A particular person yearned to have a particular thing	0

4. McRae et al. (2005)

Feature	Value
a-utensil	0.634 (19/30)
found-in-kitchens	0.600 (18/30)
used-with-forks	0.534 (16/30)
a-cutlery	0.500 (15/30)
is-dangerous	0.467 (14/30)
a-weapon	0.367 (11/30)

McRae et al. (2005) feature norms for the concept *knife*

Experiment 2: Results

	GFoLDS	Comparison BERT Models		Actual BERT Models	
		Large	Base	Large	Base
RELPRON (MAP)	0.651	0.056 (+0.595)	0.193 (+0.458)	0.769 (-0.118)	0.690 (-0.039)
SNLI (Acc)	81.0%	62.0% (+19.0%)	79.9% (+1.1%)	91.1% (-10.1%)	90.7% (-9.7%)
MegaVeridicality (Acc)	81.3%	76.2% (+5.1%)	78.1% (+3.2%)	85.6% (-4.3%)	84.2% (-2.9%)
McRae et al. (ρ)	0.205	0.134 (+0.071)	0.167 (+0.038)	0.241 (-0.036)	0.247 (-0.042)

	GFoLDS	Comparison BERT Models		Actual BERT Models	
		Large	Base	Large	Base
Parameters (Millions)	174	335 (-161)	110 (+64)	335 (-161)	110 (+64)
Pretraining Data: Base/Actual (Millions of Words)	508/472	508/508 (-0/-81)		3300/3300 (-2792/-2828)	
Pretraining Epochs	4	4 (-0)		∼40 (-36)	

Contributions

✅ Experimental support for the LKCH

✅ Proof-of-concept of the downstream applicability of LFLMs

- Evidence supporting the scalability of LFLMs

Experiment 3: Scalability

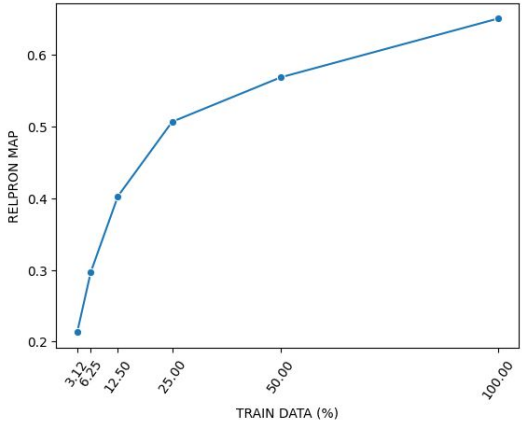
- Applied the Chinchilla Scaling Laws (CSLs; Hoffmann et al., 2022, Muennighoff et al., 2024) to GFoLDS to establish its scalability

- Pretrained five additional GFoLDS models on 50%, 25%, 12.5%, 6.25%, and 3.125% of the original pretraining data

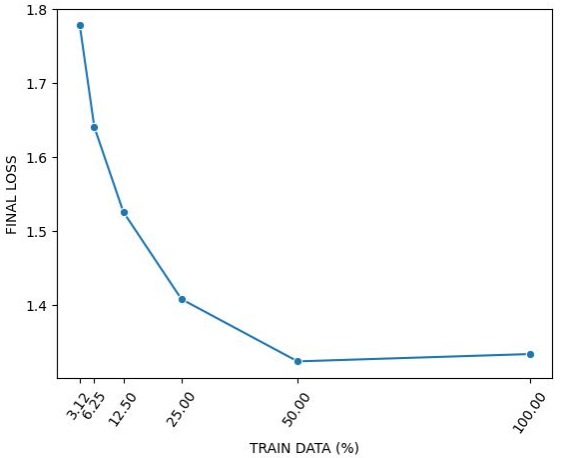
- Measured final pretraining loss and performance on a downstream task (RELPRON)

Experiment 3: Results

- Continued performance increases on the RELPRON dataset: GFoLDS is likely to scale with dataset size



- Final loss plateau: GFoLDS is *underparameterized* (too small) even for half of its pretraining dataset
  - For comparison: CSLs predict that LMs of the same size become underparameterized with a dataset **20x** larger



Contributions

✅ Experimental support for the LKCH

✅ Proof-of-concept of the downstream applicability of LFLMs

✅ Evidence supporting the scalability of LFLMs

Appendix

Preprocessing

- Due to difficulties with tokenization, I removed CARGs and out-of-vocabulary items from the DMRS input graphs:

"*John went to the park in the spring of 2017*"

⇓

"[NAMED] *went to the park in* [SEASON] *of* [YEAR]"

Preprocessing

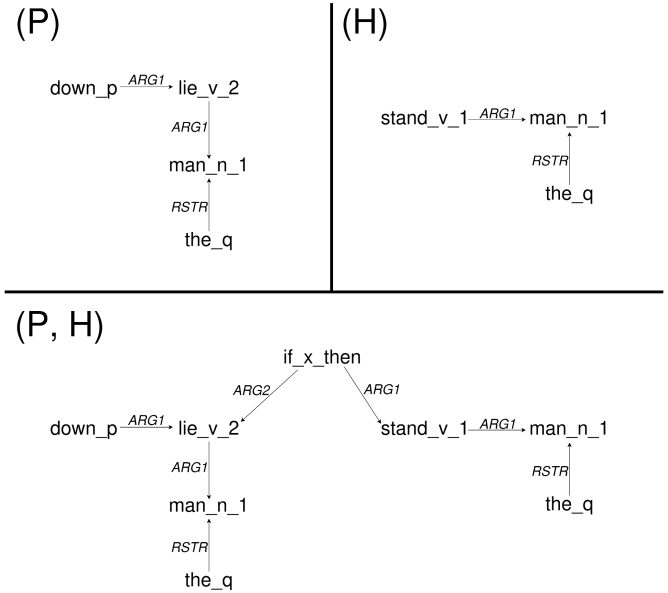
- Removed *focus_d* and *parg_d*

- Replaced edge labels:

Original Label	Interpretation/Role	Replacement
ARG1	First-place argument	—
ARG2	Second-place argument	—
ARG3	Third-place argument	—
ARG4	Fourth-place argument	—
MOD	Indicates a shared handle between two predicates	—
RSTR	Restriction of a quantifier	—
ARG	Argument of the "*unknown*" predicate	*MOD*
L-INDEX	Left-hand conjunct of two coordinated variables	*INDEX*
R-INDEX	Right-hand conjunct of two coordinated variables	*INDEX*
L-HNDL	Left-hand conjunct of two coordinated handles	*HNDL*
R-HNDL	Right-hand conjunct of two coordinated handles	*HNDL*

SNLI preprocessing

- Problem: GFoLDS can only take one input sentence at a time
  - Solution: turn each (premise, hypothesis) pair into a single sentence

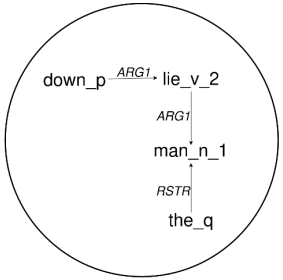


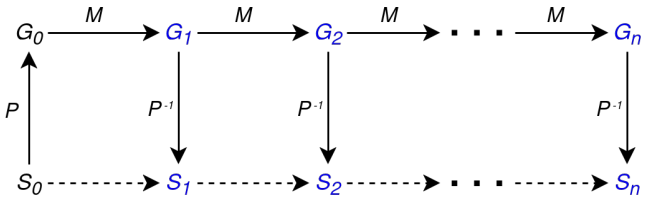
Future directions: multi-sentence model

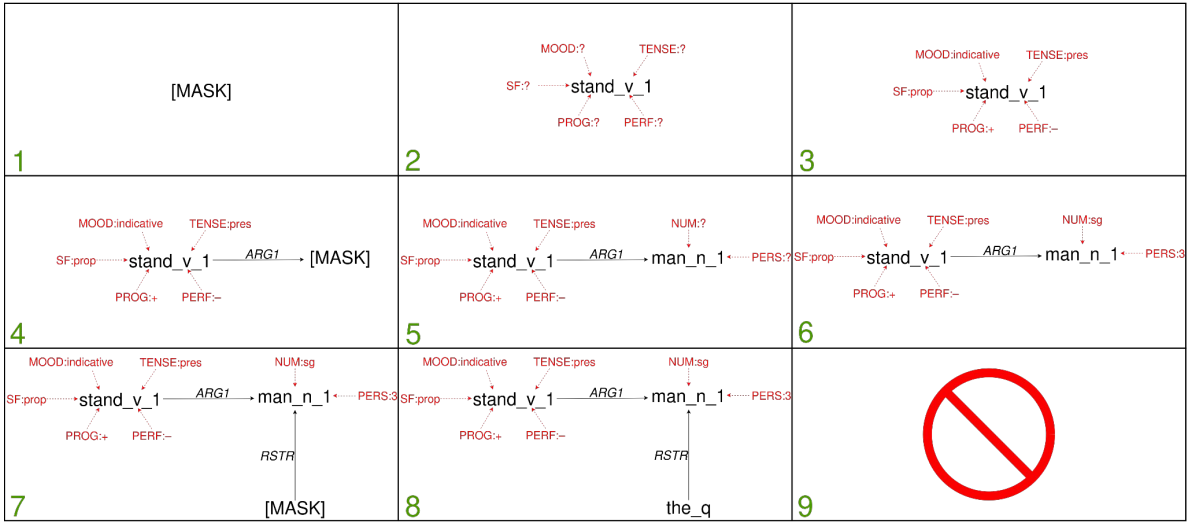
$$\vec{e_i} = E(n_i, G) = \mathcal{E}_T(n_i) + Norm\left(\sum_{\phi \in F(n_i, G)} \mathcal{E}_F(\phi)\right) \tag{7.7a}$$

$$\vec{e_{k,i}} = E(n_{k,i}, G_k) = \mathcal{E}_T(n_{k,i}) + Norm\left(\sum_{\phi \in F(n_{k,i}, G_k)} \mathcal{E}_F(\phi)\right) + \mathcal{E}_S(k) \tag{7.7b}$$

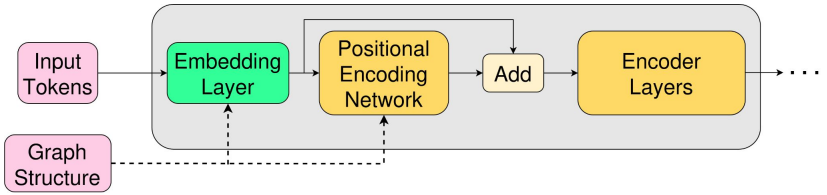
Future directions: graph-to-graph generative model

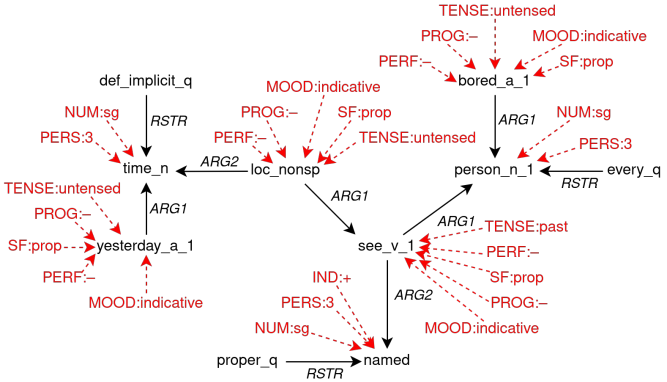






GFoLDS Architecture

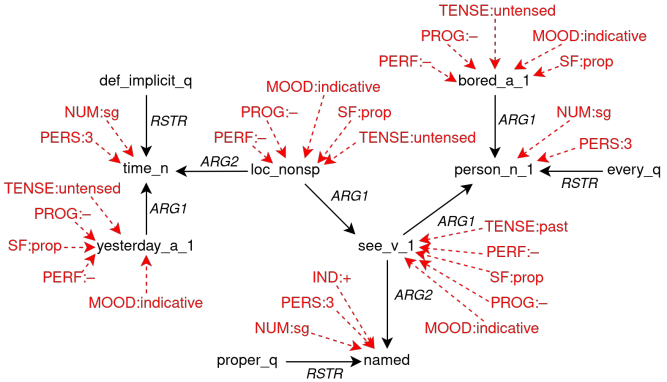


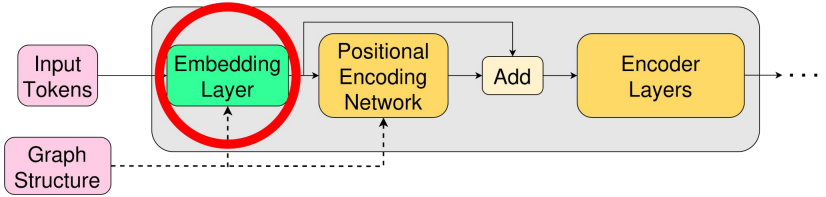


"*Every bored person saw Mary yesterday*"

- GFoLDS is a **Graph Transformer** (Wu et al., 2021): a graph neural network (GNN) coupled with a transformer encoder

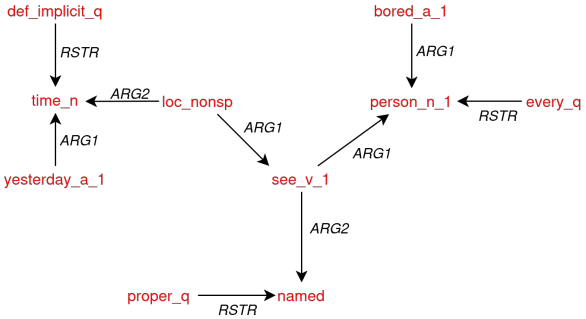
GFoLDS Architecture

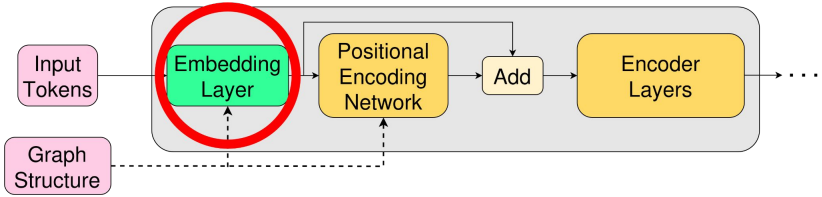




$$\vec{e_i} = \mathcal{E}_T(n_i) + Norm\left(\sum_{\phi \in F(n_i)} \mathcal{E}_F(\phi)\right)$$

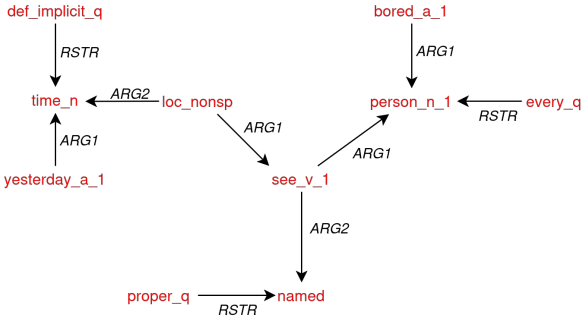
GFoLDS Architecture

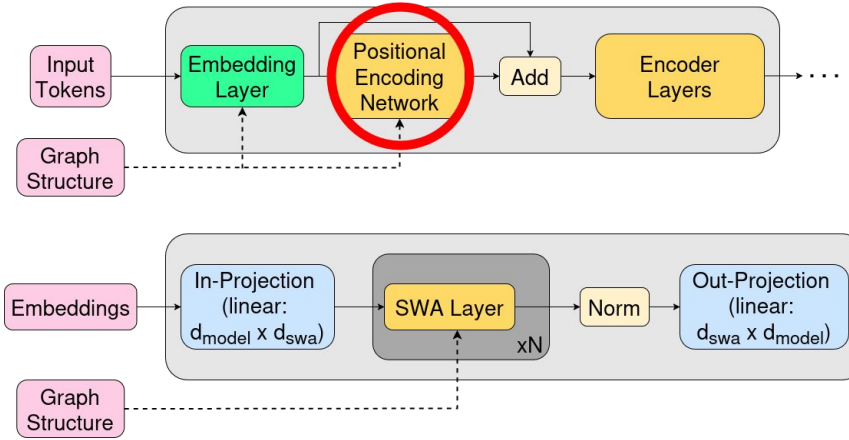




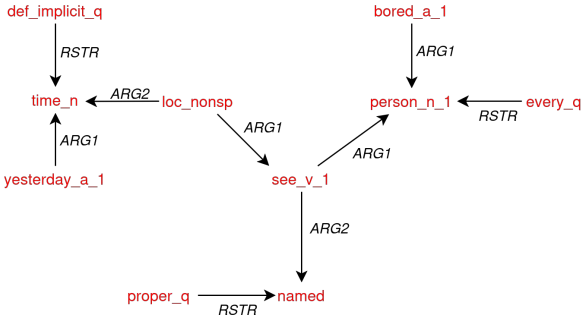
$$\vec{e_i} = \mathcal{E}_T(n_i) + Norm\left(\sum_{\phi \in F(n_i)} \mathcal{E}_F(\phi)\right)$$

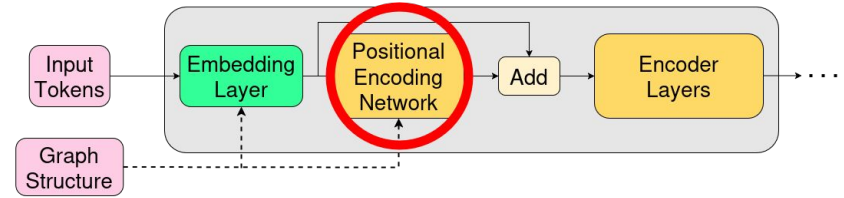
GFoLDS Architecture

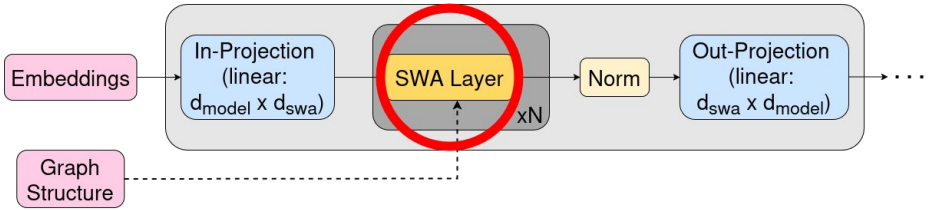


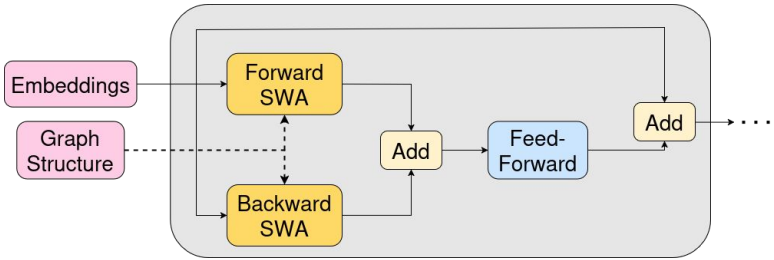


GFoLDS Architecture

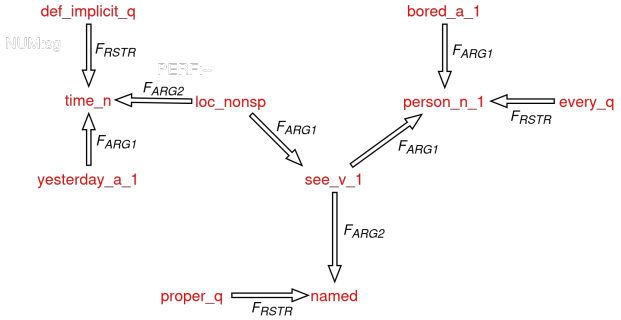


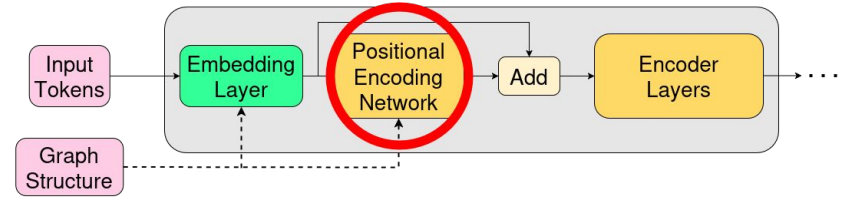


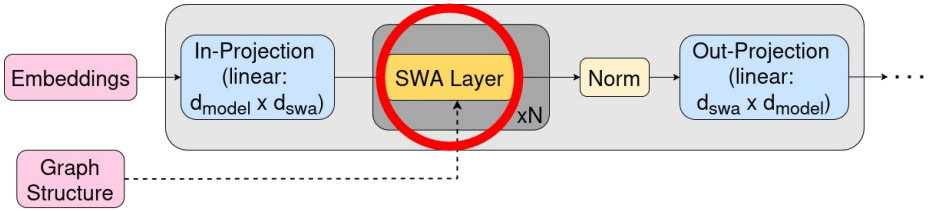


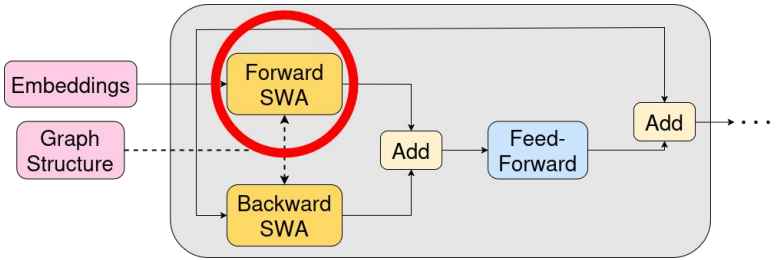


GFoLDS Architecture



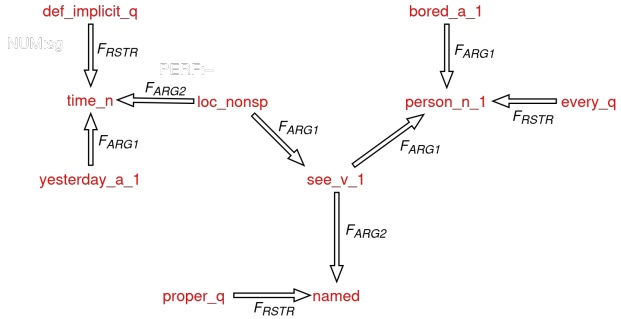


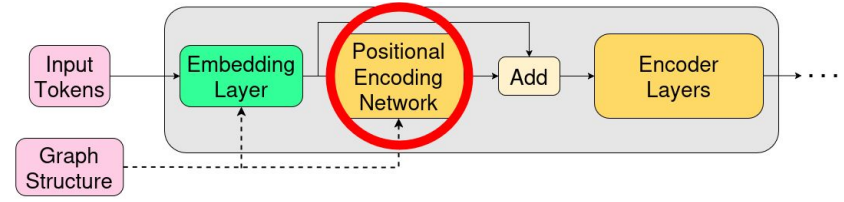


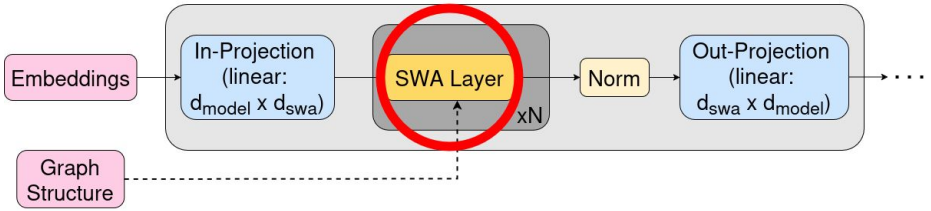


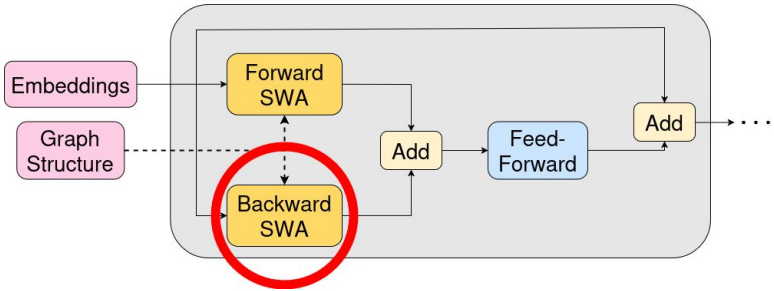
$$\vec{f}_i = Norm\left(\sum_{n_k \xrightarrow{\ell} n_i \in G} W_\ell^{(f)} \vec{h}_k\right)$$

GFoLDS Architecture



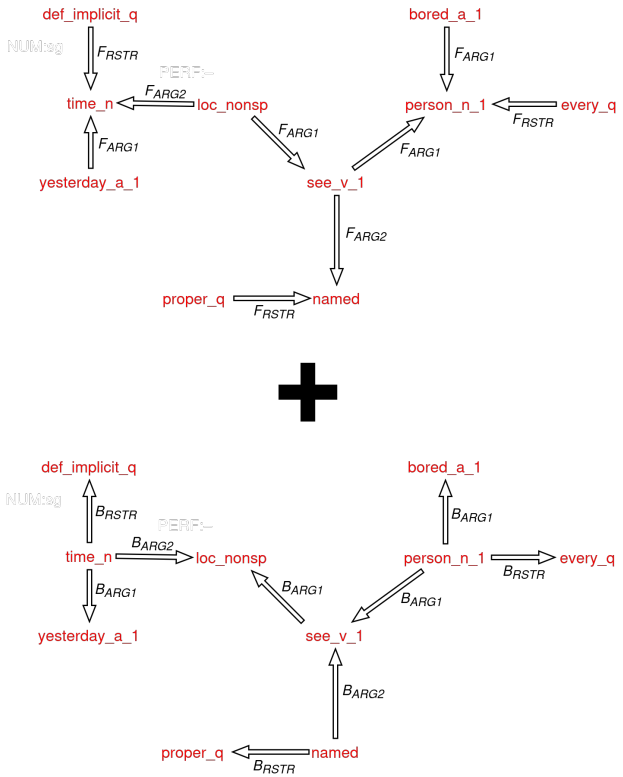


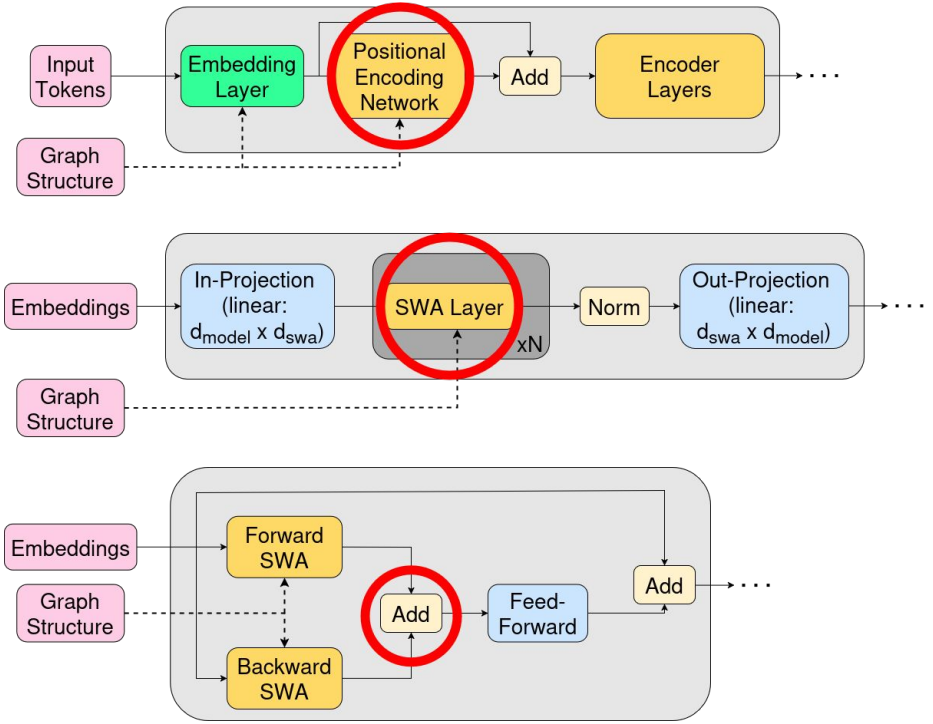




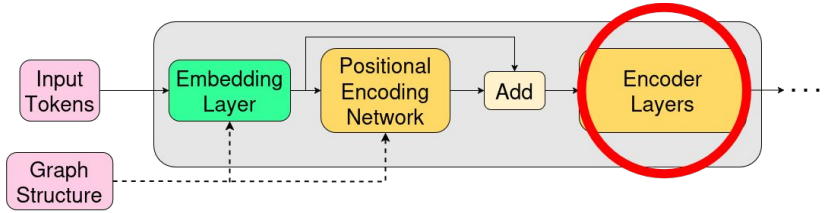
$$\vec{b_i} = Norm\left(\sum_{n_i \xrightarrow{\ell} n_k \in E} W_\ell^{(b)} \vec{x_k}\right)$$

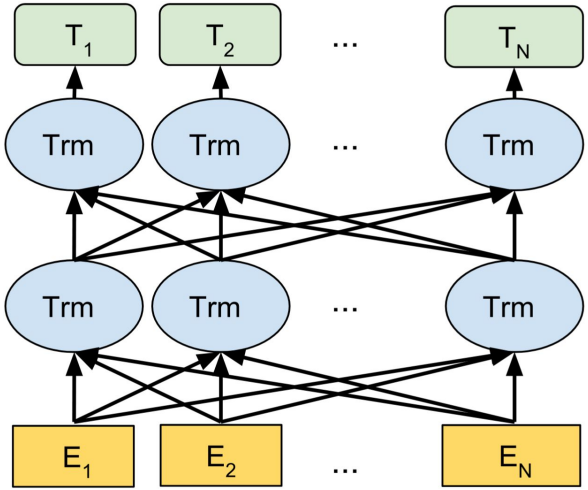
GFoLDS Architecture



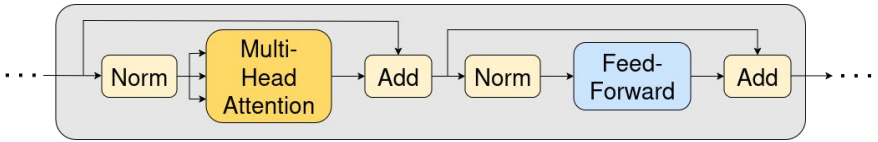


GFoLDS Architecture

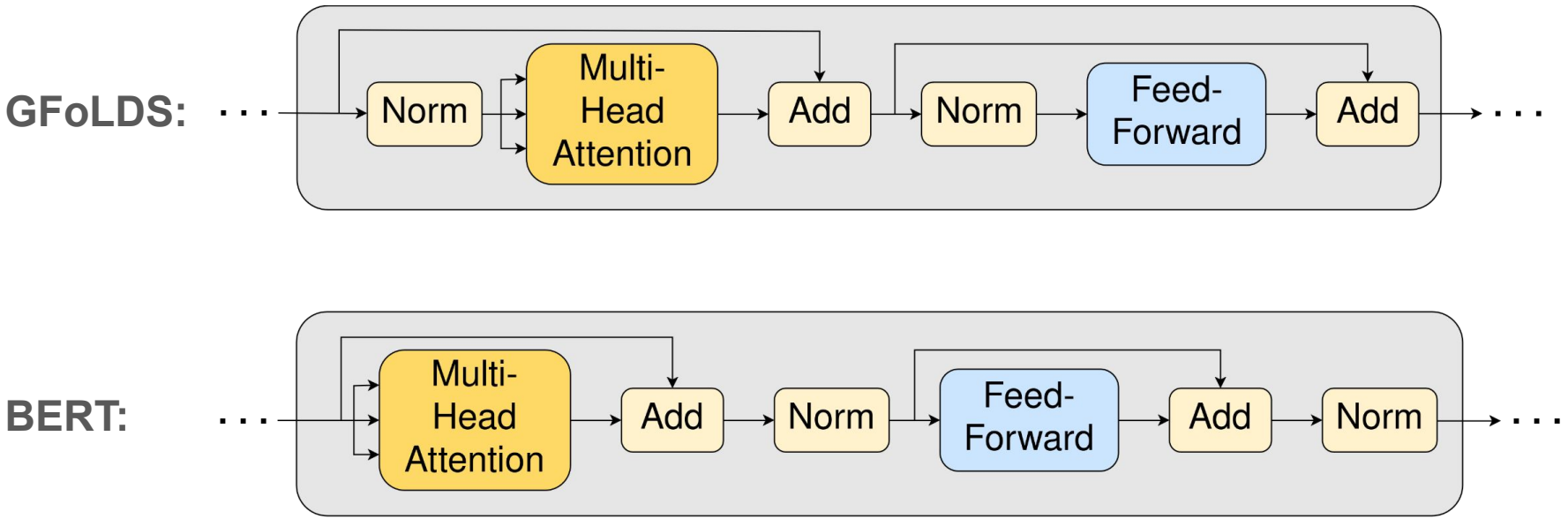




(Devlin et al., 2019)



GFoLDS vs. BERT Encoder Layers



https://github.com/google-research/tuning_playbook?tab=readme-ov-file#how-can-optimization-failures-be-debugged-and-mitigated

Limitations and Weaknesses

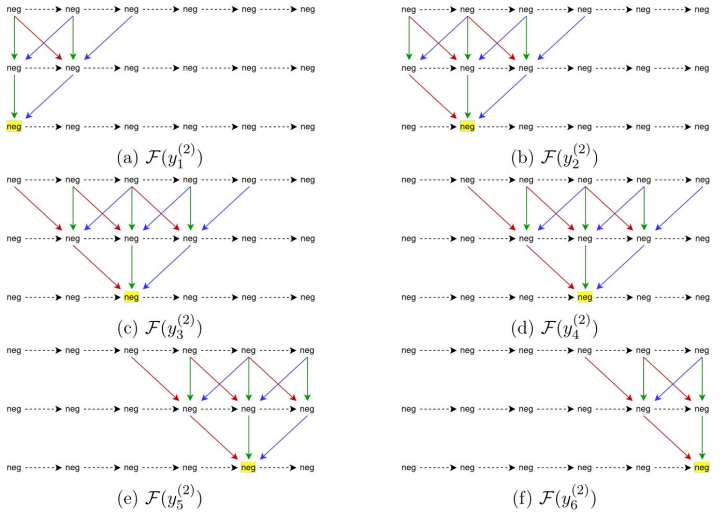
- A serious limitation in GFoLDS' architecture: it can't count repeated sequences of the same token (node label)
  - Caused by local node aggregation of its GNN
  - Prevents learning double-negation cancellation
  - Limits ability to detect node sentence membership in SNLI

Theorem 2. *Given a GFoLDS model (as defined in Chapter 4) M with n SWA layers and an input graph G, let $M(G)_i$ denote the embedding that M assigns to the i^{th} node x_i of G. Suppose that G contains a path $p = x_1 \xrightarrow{\ell} \ldots \xrightarrow{\ell} x_k$ of length k such that all nodes (and all edges) in p have the same label (respectively), and that for all $1 < i < k$, x_i has no incoming or outgoing edges not in p. Then:*

i. for all $1 \leq i \leq n$, and all $1 \leq j \leq k$ such that $i \neq j$: $M(G)_i \neq M(G)_j$

ii. for all $k - n \leq i \leq k$, and all $1 \leq j \leq k$ such that $i \neq j$: $M(G)_i \neq M(G)_j$

iii. for all $n < i, j < k - n$: $M(G)_i = M(G)_j$



Limitations and Weaknesses

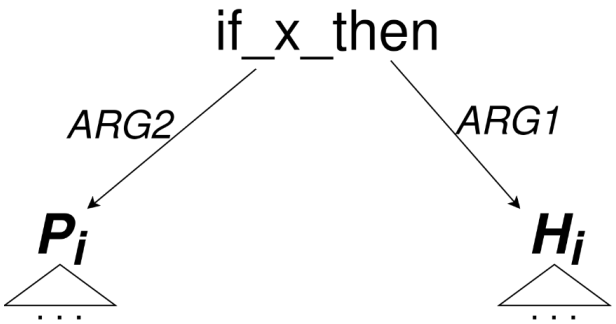
Premise	Template	Hypothesis	Label
A young boy dressed in plaid about to take a picture.	(P, H)	A young boy is about to take a picture.	Entailment
	$(P, (T_{NT})^1 H)$	It is not true that a young boy is about to take a picture.	Contradiction
	$(P, (T_{NT})^2 H)$	It is not true that it is not true that a young boy is about to take a picture.	Entailment
	$(P, (T_{NT})^3 H)$	It is not true that it is not true that it is not true that a young boy is about to take a picture.	Contradiction
A race between friends at the park.	(P, H)	The park is deserted.	Contradiction
	$(P, (T_{NT})^1 H)$	It is not true that the park is deserted.	Entailment
	$(P, (T_{NT})^2 H)$	It is not true that it is not true that the park is deserted.	Contradiction
	$(P, (T_{NT})^3 H)$	It is not true that it is not true that it is not true that the park is deserted.	Entailment
People kneeling on the ground.	(P, H)	People are praying.	Neutral
	$(P, (T_{NT})^1 H)$	It is not true that people are praying.	Neutral

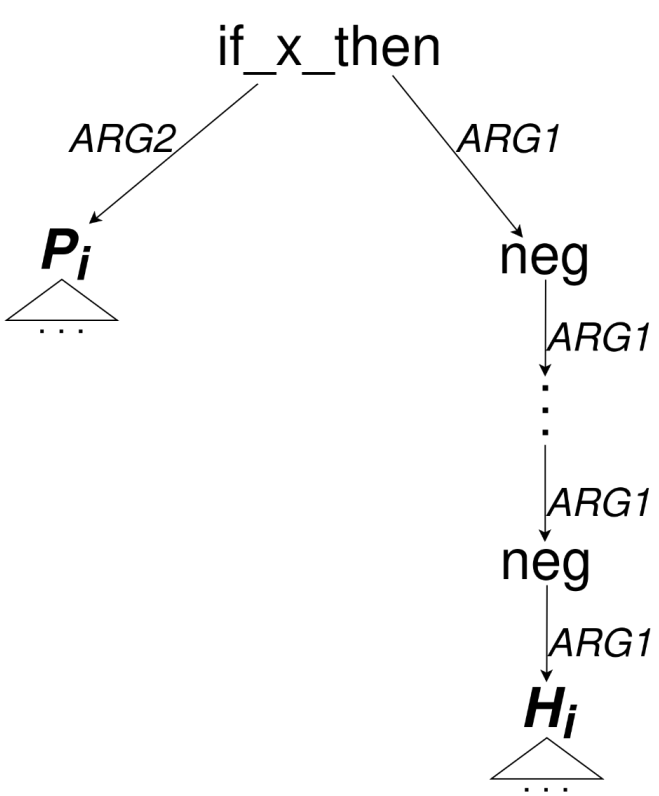
Table 2.1: Examples of depth-n negated challenge data points $(P, (T_{NT})^n H)$ generated from SNLI (some examples have been slightly modified for presentability).

Model	Depth-m test	No inoc.	Depth-1 inoc.	Depth-$\leq$2 inoc.	Depth-$\leq$3 inoc.
$BART_M$	2	0.71	0.32	—	—
$BART_M$	3	0.36	0.93	0.31	—
$BART_M$	4	0.82	0.36	0.94	0.31
$BART_M$	5	0.33	0.88	0.31	0.94
$BART_M$	6	0.86	0.41	0.94	0.31
$RoBERTa_M$	2	0.77	0.36	—	—
$RoBERTa_M$	3	0.34	0.89	0.33	—
$RoBERTa_M$	4	0.85	0.33	0.97	0.32
$RoBERTa_M$	5	0.32	0.88	0.33	0.95
$RoBERTa_M$	6	0.89	0.34	0.97	0.33
$DeBERTa_S$	2	0.56	0.62	—	—
$DeBERTa_S$	3	0.4	0.61	0.32	—
$DeBERTa_S$	4	0.84	0.64	0.96	0.5
$DeBERTa_S$	5	0.3	0.51	0.32	0.96
$DeBERTa_S$	6	0.88	0.77	0.96	0.36
$RoBERTa_S$	2	0.74	0.32	—	—
$RoBERTa_S$	3	0.4	0.89	0.3	—
$RoBERTa_S$	4	0.84	0.35	0.94	0.39
$RoBERTa_S$	5	0.34	0.88	0.3	0.74
$RoBERTa_S$	6	0.83	0.33	0.93	0.53
$BART_{SMFA}$	2	0.77	0.37	—	—
$BART_{SMFA}$	3	0.33	0.91	0.31	—
$BART_{SMFA}$	4	0.84	0.34	0.94	0.29
$BART_{SMFA}$	5	0.3	0.85	0.31	0.92
$BART_{SMFA}$	6	0.86	0.41	0.94	0.28
$RoBERTa_{SMFA}$	2	0.79	0.35	—	—
$RoBERTa_{SMFA}$	3	0.32	0.93	0.32	—
$RoBERTa_{SMFA}$	4	0.83	0.31	0.95	0.32
$RoBERTa_{SMFA}$	5	0.32	0.94	0.32	0.94
$RoBERTa_{SMFA}$	6	0.84	0.32	0.95	0.32
Mean	2	0.72	0.39	—	—
Mean	3	0.36	0.86	0.32	—
Mean	4	0.84	0.39	0.95	0.35
Mean	5	0.32	0.82	0.32	0.91
Mean	6	0.86	0.43	0.95	0.35

Table 2.3: Accuracy for all models on depth-(m>n) external negation (D^m_{NT}) after depth-$\leq n$ inoculation ($n \in \{1, 2, 3\}$) on $D^{\leq n}$. For the sake of convenience, mean accuracy across the models is reported at the bottom of the table; most individual model accuracies do not substantially deviate from these mean values.

Limitations and Weaknesses





Limitations and Weaknesses

Sequence Length	1-SWA	2-SWA	3-SWA	4-SWA
1	✓	✓	✓	✓
2	✓	✓	✓	✓
3	✓	✓	✓	✓
4	X	✓	✓	✓
5	X	✓	✓	✓
6	X	X	✓	✓
7	X	X	✓	✓
8	X	X	X	✓
9	X	X	X	✓
10	X	X	X	✓
11	X	X	X	✓
12	X	X	X	X

Table 6.5: Results of the mod-2 counting experiment for four GFoLDS models with n SWA layers ($1 \leq n \leq 4$). A ✓ symbol at row r indicates that the model was able to correctly classify each $p_{neg}^{(k)}$ sequence as odd or even within 100 training epochs, for all $1 \leq k \leq r$.

Depth	S_1 Accuracy	S_2 Accuracy	Total Accuracy
Overall	87.0%	72.3%	82.1%
1	80.3%	82.4%	81.3%
2	86.3%	89.0%	87.6%
3	83.9%	66.8%	76.5%
4	88.5%	72.5%	83.6%
5	89.3%	61.9%	82.8%
6	89.6%	50.5%	82.6%
7	89.2%	45.3%	83.3%
8	88.5%	46.3%	84.3%
9	87.4%	45.4%	84.1%
10	88.2%	42.1%	85.3%
11	88.5%	43.4%	86.5%
12	87.1%	26.9%	84.6%
13	88.0%	0.0%	85.9%
14	87.4%	0.0%	85.4%
15	77.1%	—	77.1%
16	62.5%	—	62.5%
17	58.8%	—	58.8%
18	66.7%	—	66.7%
19	33.3%	—	33.3%

Table 6.6: GFoLDS' test set accuracy on the S_1/S_2 classification task, by node depth (undirected distance from the *if_x_then* node). *Total accuracy* denotes the accuracy for all S_1 and S_2 nodes at the depth in question. The — symbol in the S_2 column for depths 15-19 indicates that there were no S_2 nodes in the test set at those depths.

References

Copestake, A.; Flickinger, D.; Pollard, C.; and Sag, I. A. 2005. Minimal Recursion Semantics: An Introduction. *Research on Language and Computation*, 3: 281–332.

Ann Copestake. 2009. Invited Talk: Slacker Semantics: Why Superficiality, Dependency and Avoidance of Commitment Can be the Right Way to Go. In Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), 1–9.

Banarescu, L.; Bonial, C.; Cai, S.; Georgescu, M.; Griffitt, K.; Hermjakob, U.; Knight, K.; Koehn, P.; Palmer, M.; and Schneider, N. 2013. Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, 178–186.

Xu, Z.; Guo, D.; Tang, D.; Su, Q.; Shou, L.; Gong, M.; Zhong, W.; Quan, X.; Jiang, D.; and Duan, N. 2021. Syntax-Enhanced Pre-trained Model. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers), 5412–5422.

Zhou, J.; Zhang, Z.; Zhao, H.; and Zhang, S. 2020. LIMIT-BERT: Linguistics Informed Multi-Task BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 4450–4461.

Zhang, Z.; Wu, Y.; Zhao, H.; Li, Z.; Zhang, S.; Zhou, X.; and Zhou, X. 2020c. Semantics-Aware BERT for Language Understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 9628–9635.

Wu, Z.; Peng, H.; and Smith, N. A. 2021. Infusing Finetuning with Semantic Dependencies. *Transactions of the Association for Computational Linguistics*, 9: 226–242.

Prange, J.; Schneider, N.; and Kong, L. 2022. Linguistic Frameworks Go Toe-to-Toe at Neuro-Symbolic Language Modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4375–4391.

Wu, Z.; Jain, P.; Wright, M.; Mirhoseini, A.; Gonzalez, J. E.; and Stoica, I. 2021. Representing Long-Range Context for Graph Neural Networks with Global Attention. *Advances in Neural Information Processing Systems*, 34: 13266–13279.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

Rimell, L.; Maillard, J.; Polajnar, T.; and Clark, S. 2016. RELPRON: A Relative Clause Evaluation Data Set for Compositional Distributional Semantics. *Computational Linguistics*, 42(4): 661–701.

Bowman, S. R.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A Large Annotated Corpus for Learning Natural Language Inference. *arXiv preprint arXiv:1508.05326*.

White, A. S.; Rudinger, R.; Rawlins, K.; and Van Durme, B. 2018. Lexicosyntactic Inference in Neural Models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4717–4724.

McRae, K.; Cree, G. S.; Seidenberg, M. S.; and McNorgan, C. 2005. Semantic Feature Production Norms for a Large Set of Living and Nonliving Things. *Behavior Research Methods*, 37(4): 547–559.

Hoffmann, J.; Borgeaud, S.; Mensch, A.; Buchatskaya, E.; Cai, T.; Rutherford, E.; Casas, D. d. L.; Hendricks, L. A.; Welbl, J.; Clark, A.; Hennigan, T.; Noland, E.; Millican, K.; van den Driessche, G.; Damoc, B.; Guy, A.; Osindero, S.; Simonyan, K.; Elsen, E.; Rae, J. W.; Vinyals, O.; and Sifre, L. 2022. Training Compute-Optimal Large Language Models. *arXiv preprint arXiv:2203.15556*.

Muennighoff, N.; Rush, A.; Barak, B.; Le Scao, T.; Tazi, N.; Piktus, A.; Pyysalo, S.; Wolf, T.; and Raffel, C. A. 2024. Scaling Data-Constrained Language Models. *Advances in Neural Information Processing Systems*, 36.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2018. Language Models are Unsupervised Multitask Learners.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, 1877–1901.

OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774.

Villalobos, P.; Sevilla, J.; Heim, L.; Besiroglu, T.; Hobbhahn, M.; and Ho, A. 2022. Will We Run out of Data? An Analysis of the Limits of Scaling Datasets in Machine Learning. *arXiv preprint arXiv:2211.04325*.