In the Mood for Inference: Logic-Based Natural Language Inference with Large Language Models

Bill Noble, Rasmus Blanck, and Gijs Wijnholds

Natural Logic Meets Machine Learning

4 August 2025

Natural Language Inference (NLI)

- Task: Given a **premise** and **hypothesis** sentence, decide whether the relation between them is an `entailment`, a `contradiction`, or `neutral`.
- Inference has a deep theoretical importance in linguistic semantics — NLI is understood as testing understanding and reasoning in natural language
  - Improvements in neural language modeling saw renewed focus on NLI (rebranded from RTE) in the 2010's: SICK (Marelli et al., 2014), SNLIBowman et al. (2015)
  - Subsequent datasets seek to probe for understanding of specific linguistic phenomena: MED (Yanaka et al., 2019a; Richardson et al., 2020), CURRICULUM benchmark (Chen and Gao, 2022)

Generative LLMs for Neurosymbolic NLI

General strategy:

1. Prompt an LLM to generate a logical representation of the premise and hypothesis
2. Use a theorem prover to determine the entailment relation.
   - Try to prove the hypothesis from the premise $\rightarrow$ `entailment`
   - Try to prove the negation of the hypothesis from the premise $\rightarrow$ `contradiction`
   - Otherwise $\rightarrow$ `neutral`

Generative LLMs for Neurosymbolic NLI

General strategy:

1. Prompt an LLM to generate a logical representation of the premise and hypothesis
2. Use a theorem prover to determine the entailment relation.
    - Try to prove the hypothesis from the premise $\rightarrow$ `entailment`
    - Try to prove the negation of the hypothesis from the premise $\rightarrow$ `contradiction`
    - Otherwise $\rightarrow$ `neutral`

Generative LLMs for Neurosymbolic NLI

Examples:

- Olausson et al. (2023): LINC – majority voting for multiple FoL formula samples
- Pan et al. (2023): Logic-LM – Problem formulator chooses between Logic Programming, FoL prover, Constraint Optimizatio, or SMT Solver
-

Cf.:

- Chen et al. (2021): *NeuralLog* – inference via incremental monotonic rephrasing
- Tomihari and Yanaka (2023): Supplement word-to-word knowledge with visual LLM phrase representations

An abbreviated list of challenges ...

- There may not be much of [your favorite formalism] in the training data
- The LLM may generate syntactically incorrect translations
- There may be translation inconsistencies between preemies and hypothesis
- World knowledge may not be encoded (lexical inference, common knowledge, topoi)

... and things one can do to mitigate them

- have premise formalisation in context window when generating for the hypothesis
- explicitly prompt for world knowledge
- few-shot prompting
- grammar-constrained decoding (Geng et al., 2024)
- re-prompt the model when the representation is incorrect (Pan et al., 2024)

Our approach

1. We prompt the LLM with the premise and hypothesis and ask for FoL translations; we *also* ask directly for a NLI label.
   - As a baseline strategy, we also *only* ask for the direct answer
   - In another condition, we ask the model to produce FoL formulas for any relevant lexical world knowledge
2. The prompt includes three few-shot examples from the same dataset (formatted in the same way as the current item)
3. The FoL formulas are provided to the Prover9 FoL theorem prover, with the hypothesis as a goal
   - In the *world knowledge* condition, both premise and world knowledge formulas are used

Our approach

Prompts

- *label-only* – we only ask the LLM for the NLI label
- *forms* – we ask for FoL formulas representing the premise and hypothesis *then* ask for the NLI label
- *forms+wk* – we additionally ask for formulas representing any relevant lexical knowledge

Inference schemas

- DA – the *direct answer* label provided by the LLM
- P9 – the label inferred from whether *Prover9* can find a proof from the LLM-written premise formula (and any lexical knowledge) to that of the hypothesis or its negation

Our approach

Model Zephyr 7b – freely available training code & data, reasonably good *label-only* baseline, small enough to run on our GPUs

FoL Prover Prover9 – well-established FoL theorem prover, permissive and intuitive syntax

Datasets[1]

comparative – **P**: *John is taller than Gordon and Erik..., and Mitchell is as tall as John*; **H**: *Erik is taller than Gordon*; `neutral`

conditional – **P**: *Francisco has visited Potsdam and if Francisco has visited Potsdam then Tyrone has visited Pampa*; **H**: *Tyrone has visited Pampa*; `entailment`

negation – **P**: *Laurie has only visited Nephi, Marion has only visited Calistoga*; **H**: *Laurie didn't visit Calistoga*; `contradiction`

quantifier – **P**: *Everyone has visited every place*; **H**: *Virgil didn't visit Barry*; `neutral`

lexical entailment – **P**: *Sadat beat Jimmy Carter*; **H**: *Jimmy Carter secluded Sadat*; `not-entailed`

monotonicity – **P**: *Not all new workers joined for a dinner* **H**: *Not all workers joined for a dinner*; `entailment`

[1] (Chen and Gao, 2022; Richardson et al., 2020; Schmitt and Schütze, 2021; Glockner et al., 2018; Yanaka et al., 2019b)

	Label	E		C		N	
		DA	P9	DA	P9	DA	P9
Dataset	*Prompt*						
comparative	label-only	77.6	–	73.7	–	6.8	–
	forms	71.4	8.2	**94.9**	0.0	**17.5**	<u>99.0</u>
	forms+wk	**98.0**	<u>100.0</u>	56.6	0.0	9.7	0.0
conditional	label-only	48.4	–	0.0	–	50.9	–
	forms	**74.7**	59.8	**54.7**	50.0	**43.6**	<u>100.0</u>
	forms+wk	50.5	37.0	51.6	48.4	0.9	<u>100.0</u>
negation	label-only	0.0	–	15.6	–	36.5	–
	forms	0.0	0.0	60.0	<u>98.9</u>	**90.4**	<u>100.0</u>
	forms+wk	**2.2**	0.0	**62.2**	<u>98.8</u>	9.6	<u>98.2</u>
quantifier	label-only	70.0	–	3.5	–	**96.9**	–
	forms	72.2	59.8	**82.5**	27.2	29.2	<u>100.0</u>
	forms+wk	**98.9**	64.0	1.8	<u>23.5</u>	5.2	<u>100.0</u>

	Label	**E**		**N/C**	
		DA	P9	DA	P9
Dataset	*Prompt*				
lexical	label-only	6.7	–	94.6	–
	forms	**25.0**	1.5	**97.3**	100.0
	forms+wk	**25.0**	65.3	96.6	34.8
monotonicity	label-only	**58.3**	–	34.8	–
	forms	46.8	16.7	**47.8**	90.9
	forms+wk	47.5	50.0	47.2	54.6

Discussion

- P9 recall is often very good for `neutral`; in conjunction with poor performance on other labels, this indicates missing assumptions and/or inadequate FoL representations
- DA is *generally* higher when the model is asked to generate formulas first, even when the P9 score lower than DA
  - exception: in the **quantifier** dataset the *label-only* prompt performs better than *forms*
- DA performance has an unpredictable response to prompting for world knowledge
  - e.g., improves recall for **comparative** and **quantifier** `entailments` but at the expense of `contradictions`

Conclusion

- Performance given different prompts and inference schemata is highly idiosyncratic WRT dataset
- That said, combining LLM-generated formulas with theorem proving is promising for pushing the limits of challenging NLI datasets
- We only experimented with one LLM (Zephyr)
- We may get better performance with a formalisation that balances ubiquitousness (in LLM training data) with linguistic expressivity

Conclusion

- Performance given different prompts and inference schemata is highly idiosyncratic WRT dataset
- That said, combining LLM-generated formulas with theorem proving is promising for pushing the limits of challenging NLI datasets
- We only experimented with one LLM (Zephyr)
- We may get better performance with a formalisation that balances ubiquitousness (in LLM training data) with linguistic expressivity

Thank you!

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Zeming Chen and Qiyue Gao. 2022. Curriculum: A broad-coverage benchmark for linguistic phenomena in natural language understanding. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3204–3219, Seattle, United States. Association for Computational Linguistics.

Zeming Chen, Qiyue Gao, and Lawrence S. Moss. 2021. NeuralLog: Natural language inference with joint neural and logical reasoning. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 78–88, Online. Association for Computational Linguistics.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych.

2024. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595, Mexico City, Mexico. Association for Computational Linguistics.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI Systems with Sentences that Require Simple Lexical Inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Theo Olausson, Alex Gu, Ben Lipkin, Cedegao Zhang, Armando Solar-Lezama, Joshua

Tenenbaum, and Roger Levy. 2023. LINC: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5153–5176, Singapore. Association for Computational Linguistics.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Wang. 2023. Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3806–3824, Singapore. Association for Computational Linguistics.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2024. Automatically correcting large language models: Surveying the landscape of diverse automated correction strategies. *Transactions of the Association for Computational Linguistics*, 12:484–506.

Kyle Richardson, Hai Hu, Lawrence Moss, and Ashish Sabharwal. 2020. Probing natural language inference models through semantic fragments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8713–8721.

Martin Schmitt and Hinrich Schütze. 2021. Language Models for Lexical Inference in Context. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1267–1280, Online. Association for Computational Linguistics.

Akiyoshi Tomihari and Hitomi Yanaka. 2023. Logic-based inference with phrase abduction using vision-and-language models. *IEEE Access*, 11:45645–45656.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019a. Can neural networks understand monotonicity reasoning? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 31–40, Florence, Italy. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha

Abzianidze, and Johan Bos. 2019b. Can Neural Networks Understand Monotonicity Reasoning? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 31–40, Florence, Italy. Association for Computational Linguistics.

Filtered and unfiltered Prover9 Recall

	Label	**E**		**C**		**N**	
Dataset	*Prompt*						
comparative	forms	8.2	8.2	0.0	0.0	99.0	99.0
	forms+wk	100.0	100.0	0.0	0.0	0.0	0.0
conditional	forms	57.9	59.8	48.4	50.0	100.0	100.0
	forms+wk	35.8	37.0	47.4	48.4	100.0	100.0
negation	forms	0.0	0.0	97.8	98.9	96.5	100.0
	forms+wk	0.0	0.0	93.3	98.8	95.7	98.2
quantifier	forms	57.8	59.8	24.6	27.2	97.9	100.0
	forms+wk	61.1	64.0	21.1	23.5	96.9	100.0

Filtered and unfiltered Prover9 Recall

	Label	**E**		**N/C**	
Dataset	*Prompt*				
lexical	forms	1.3	1.5	89.9	100.0
	forms+wk	52.0	65.3	27.2	34.8
monotonicity	forms	13.7	16.7	74.5	90.9
	forms+wk	36.0	50.0	36.6	54.6