On the Evaluation of LLM's Understanding of Logical Validity Through Minimal Pairs

Mauro Santelli

2025-05-13

(joint work with Ramiro Caso, Joaquín S. Toranzo Calderón, Ramiro Rodríguez Colmeiro and Nicolás Aguirre)

Introduction

- In different areas of AI research (NLP, NLU, ML, etc.) the capabilities of systems are tested in whatever ways the researchers of the time have found useful.
- While difficult to systematize, a number of approaches have been found useful (Chollet, 2019), of these, **benchmarks** are salient.

Introduction

- In different areas of AI research (NLP, NLU, ML, etc.) the capabilities of systems are tested in whatever ways the researchers of the time have found useful.
- While difficult to systematize, a number of approaches have been found useful (Chollet, 2019), of these, **benchmarks** are salient.
- `With benchmark we refer to a test that is:`

  `Reproducible: the test set is fixed`
  `Fair: the test set is the same for everyone`
  `Scalable: it is inexpensive to run the evaluation many times easy to set up`
  `Flexible enough to be applicable to a wide range of possible tasks (Chollet, 2019)`

`Benchmark is short for a test set of problems that has some or all of these properties. It claims to establish` *`evaluation benchmarks`* `towards some objective`

Reasoning Benchmarks

Summing up

A good benchmark presents a set of problems that if resolved correctly in enough cases by an agent/system, said benchmark tells us something about the possession by the agent of the capabilities the benchmark purportedly is an evaluation of.

Reasoning Benchmarks

Summing up

A good benchmark presents a set of problems that if resolved correctly in enough cases by an agent/system, said benchmark tells us something about the possession by the agent of the capabilities the benchmark purportedly is an evaluation of.

Benchmark

Benchmark of Capability X = Problem Set + Scoring Function

Reasoning Benchmarks

- Models are treated as black boxes
- Format: Question Answering
- Medium: Natural Language [1]
- What is evaluated: Inferential capacities
- Practical constraint: Saturation-resistant

[1] (Liang et al., 2023).

Reasoning Benchmarks

- Models are treated as black boxes
- Format: Question Answering
- Medium: Natural Language [1]
- What is evaluated: Inferential capacities
- Practical constraint: Saturation-resistant

[1] (Liang et al., 2023).

- LIME
- bAbI
- LSAT
- LegalReasoning
- GPQA

See Wu et al. (2021), Weston et al. (2016) and Liang et al. (2023).

Target capabilities are too broad

- It is difficult to link what these benchmarks measure with our theories of *good reasoning*, taking logic as some kind of normative standard.

Target capabilities are too broad

- It is difficult to link what these benchmarks measure with our theories of *good reasoning*, taking logic as some kind of normative standard.

Controversial or shallow theoretical assumptions

- Some of these benchmarks have a very superficial understanding of (good) reasoning
- They are rarely presented as trying compose larger measurements through relationships with other benchmarks / tasks / skills.

Target capabilities are too broad

- It is difficult to link what these benchmarks measure with our theories of *good reasoning*, taking logic as some kind of normative standard.

Controversial or shallow theoretical assumptions

- Some of these benchmarks have a very superficial understanding of (good) reasoning
- They are rarely presented as trying compose larger measurements through relationships with other benchmarks / tasks / skills.

Data quality is low

- Either low grade synthetic data
- Or just driven by availability of ground truth data (like legal standardized tests in the US)

- We can evaluate reasoning capabilities in isolation with benchmarks.

- We can evaluate reasoning capabilities in isolation with benchmarks...

  ...up to some point.

- We can evaluate reasoning capabilities in isolation with benchmarks...

  ...up to some point.
- Capabilities can be placed in hierarchies.
- Hierarchies can be enriched.
- Implementations should be seen as hypotheses liable for revision.

From Linguistic Minimal Pairs to Reasoning

We are trying to extrapolate a technique used successfully by linguists to test minimal syntactical competence through minimal pairs. A minimal pair involves two sentences (or smaller lexical units) in which one is derived from the other. The derived one involves just one change from the first, and that change renders it incorrect–from a grammatical point of view.

From Linguistic Minimal Pairs to Reasoning

We are trying to extrapolate a technique used successfully by linguists to test minimal syntactical competence through minimal pairs. A minimal pair involves two sentences (or smaller lexical units) in which one is derived from the other. The derived one involves just one change from the first, and that change renders it incorrect–from a grammatical point of view.

- Take this example from BLiMP (Warstadt et al., 2020):

```
Correct: Many girls insulted themselves.
Incorrect: Many girls insulted herself.
```

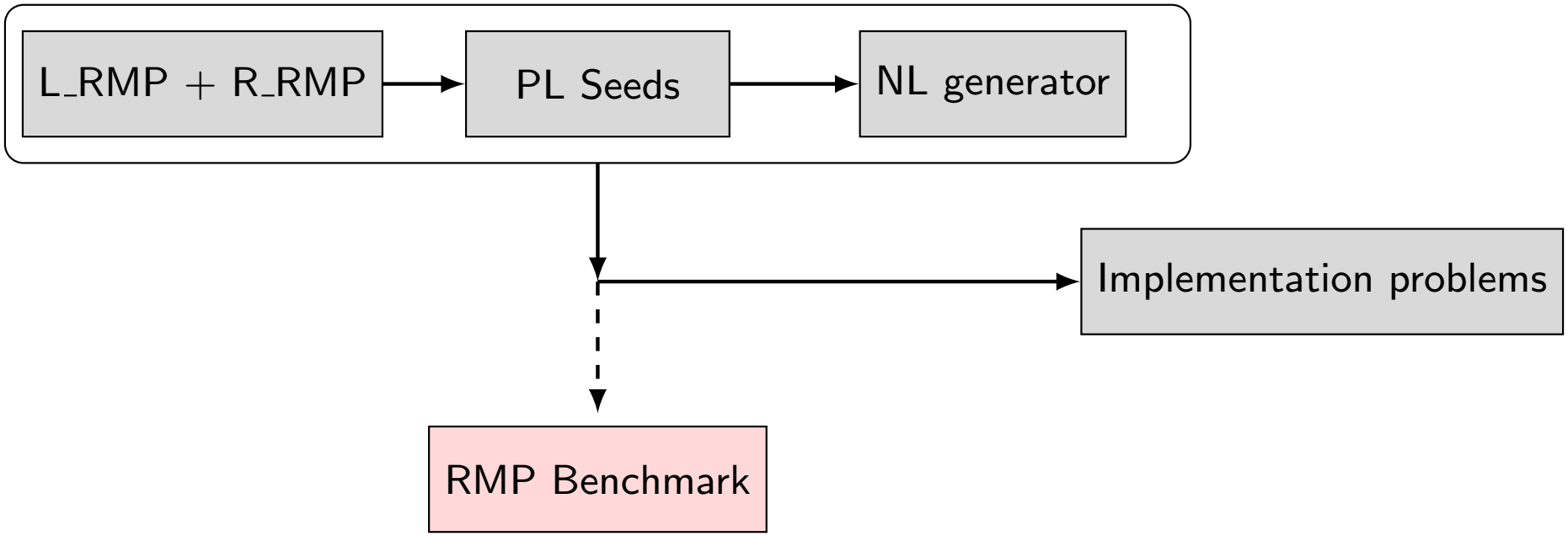
- This minimal pair is meant to test the following phenomenon: "the requirement that reflexive pronouns like himself (a.k.a. anaphora) agree with their antecedents in person, number, gender and animacy" (Warstadt et al., 2020).

Reasoning Minimal Pairs

- We propose an extension of this methodology to apply it to valid forms of reasoning.
  - This is partially motivated by the role in teaching the opposition between valid forms and invalid arguments.
  - For instance, the deceptive similarity between Modus Ponens and the affirming the consequent fallacy (or Modus Tollens and denying the antecedent)

.

How it would work



- Reasoning Minimal Pairs would be composed of a Left (valid form) Pair and a Right (invalid) Pair. L_RMP and R_RMP
- We would use inference rules in Classical Propositional Logic as L_RMPs, and minimal invalid variations of them as R_RMPs.
- These would become the set of reasoning seeds.
- We would use this with a set of natural language sentences and a set of logical connectives to instantiate the seeds in Natural Language

- Reasoning Minimal Pairs would be composed of a Left (valid form) Pair and a Right (invalid) Pair. L_RMP and R_RMP
- We would use inference rules in Classical Propositional Logic as L_RMPs, and minimal invalid variations of them as R_RMPs.
- These would become the set of reasoning seeds.
- We would use this with a set of natural language sentences and a set of logical connectives to instantiate the seeds in Natural Language
- This would be enough for a rather easy benchmark for LLMs to saturate

- Reasoning Minimal Pairs would be composed of a Left (valid form) Pair and a Right (invalid) Pair. L_RMP and R_RMP
- We would use inference rules in Classical Propositional Logic as L_RMPs, and minimal invalid variations of them as R_RMPs.
- These would become the set of reasoning seeds.
- We would use this with a set of natural language sentences and a set of logical connectives to instantiate the seeds in Natural Language
- This would be enough for a rather easy benchmark for LLMs to saturate
- The idea is to complicate the story at the level of the NL generator.
- A further desiderata is to expand it to FOL and FO Modal Logic (and other systems from there)

Some considerations

About the PL seeds

- Pure inference rules are easier to implement
- Mixed inference rules pose challenges (for they cannot really be tested in isolation, they become part of a system of connectives)
- Inference rules that require the use of assumptions and Falsum are another challenge.

Some considerations

About the PL seeds

- Pure inference rules are easier to implement
- Mixed inference rules pose challenges (for they cannot really be tested in isolation, they become part of a system of connectives)
- Inference rules that require the use of assumptions and Falsum are another challenge.
  - Assumptions towards a desired derivation impose a pressure to generate versions in which a lot of different connectives appear in the sentences bracketed in the assumption.
  - A freestanding Falsum is, in NL, any contradiction from which Falsum is derivable.

Some considerations

- These difficulties in implementation pose an opportunity to increase the challenge of the target problem set.
- The NL generator would ideally take full use of premise and conclusion markers in natural language (given that, due to, ..., etc.; in conclusion, therefore, as a result, ...)
- it should also take full advantage of antecedent and consequent markers for conditionals, necessary and sufficient conditions equivalents, and the same goes for the rest of the connectives.
- A further complication could involve normative biletarist considerations regarding pragmatic force of assertions and denials in their natural language counterparts, including assertive dampeners like I suppose, I believe (as a commitment debilitator), etc.
- The NL generator should mix the order of premises and conclusions in the resulting paragraphs, and mix the order of the premises treating different renderings of MP as just MP accordingly.

pure and mixed rules

$$\frac{\begin{array}{c}A\\B\end{array}}{A \wedge B}\ I\wedge$$

$$\frac{\begin{array}{c}[A]\\\vdots\\B\end{array}}{A \rightarrow B}\ I\rightarrow$$

$$\frac{A \wedge B}{\begin{array}{ll}A & [B]\end{array}}\ E\wedge$$

$$\frac{\begin{array}{c}A\\A \rightarrow B\end{array}}{B}\ E\rightarrow$$

$$\frac{A \quad [B]}{A \vee B} \, I\vee$$

$$\frac{\begin{array}{c}[A] \\ \vdots \\ \neg A\end{array}}{\neg A} \, I\neg$$

$$\frac{\begin{array}{ccc}[A] & & [B] \\ \vdots & & \vdots \\ C & C & C\end{array}}{C} \, E\vee$$

$$\frac{\begin{array}{c}A \\ \neg A\end{array}}{B} \, E\neg$$

- A very simple RMP for Modus Ponens:

  `L_RMP: If libraries contain books, then astronauts explore space. Libraries contain books. Therefore, astronauts explore space.`

  `R_RMP: If libraries contain books, then astronauts explore space. Astronauts explore space. Therefore, libraries contain books.`

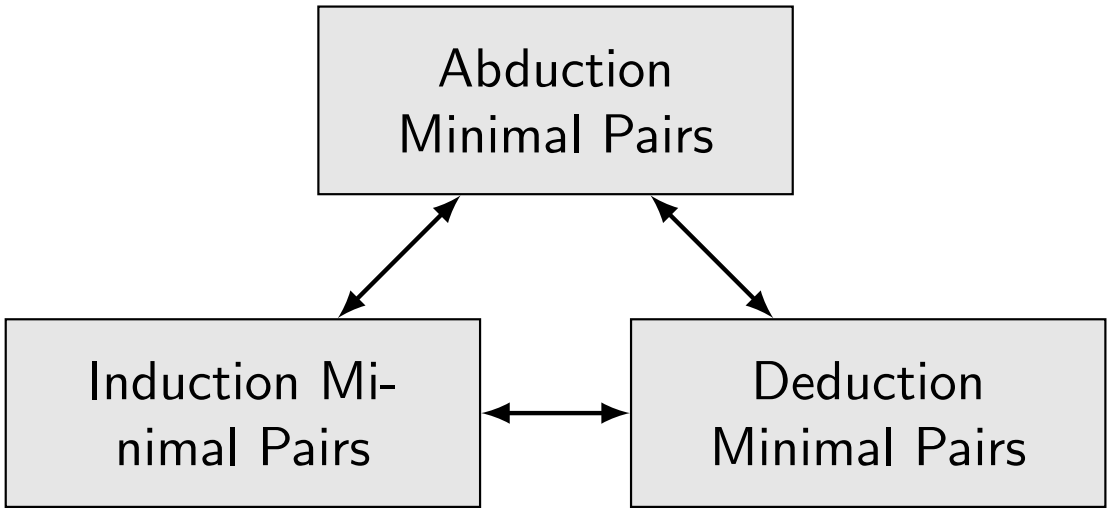
- A very simple RMP for Modus Ponens:

```
L_RMP: If libraries contain books, then astronauts explore
space. Libraries contain books. Therefore, astronauts explore
space.
```

```
R_RMP: If libraries contain books, then astronauts explore
space. Astronauts explore space. Therefore, libraries contain
books.
```

- This RMP is easily tackled by current frontier models. The Benchmark would need to do a good job of augmenting its difficulty through the NL generator in order to test the capabilities we want to test.

Logical Reasoning Triangle



Minimal Peirce

Obrigado! Thank you! ¡Gracias!

Chollet, F. (2019). On the measure of intelligence. *arXiv*.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C. A., Manning, C. D., Re, C., Acosta-Navas, D., Hudson, D. A., Zelikman, E., Durmus, E., Ladhak, F., Rong, F., Ren, H., Yao, H., Wang, J., Santhanam, K., Orr, L., Zheng, L., Yuksekgonul, M., Suzgun, M., Kim, N., Guha, N., Chatterji, N. S., Khattab, O., Henderson, P., Huang, Q., Chi, R. A., Xie, S. M., Santurkar, S., Ganguli, S., Hashimoto, T., Icard, T., Zhang, T., Chaudhary, V., Wang, W., Li, X., Mai, Y., Zhang, Y., and Koreeda, Y. (2023). Holistic Evaluation of Language Models. *Transactions on Machine Learning Research*.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., and Bowman, S. R. (2020). BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Weston, J., Bordes, A., Chopra, S., Rush, A. M., van Merriënboer, B., Joulin, A., and Mikolov, T. (2016). Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. In *4th International Conference on Learning Representations*.

Wu, Y., Rabe, M., Li, W., Ba, J., Grosse, R., and Szegedy, C. (2021). Lime: Learning inductive bias for primitives of mathematical reasoning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, page 11251–11262.