How to compare prediction abilities of different predictors – such as Large Language Models: a theoretical explanation of an empirical formula

Fernando Desantiago, Nathan Diamond, Nestor Escobedo, Nicole Favela, Nicholas Jara, Dayanna Ontiveros Alejandro Pedregon, Nayeli Ramirez, Franz Reyes, David Velez, and Vladik Kreinovich

Departments of Computer Science, University of Texas at El Paso
500 W. University, El Paso, TX 79968, USA
fdesantiago@miners.utep.edu, nkdiamond@miners.utep.edu, naescobedo@miners.utep.edu, nefavela@miners.utep.edu, najara1@miners.utep.edu, daontiveros3@miners.utep.edu, arpedregon2@miners.utep.edu, nramirez18@miners.utep.edu, fareyes4@miners.utep.edu, djvelez@miners.utep.edu, vladik@utep.edu

1. LLMs are one of the tools for predicting future

- One of the main goals of science is to predict future events:
  - we know the situations $s_1, \ldots, s_t$ at several previous moments of time, and
  - we want to predict the situation s_{t+1} that will happen in the next moment of time.
- In particular, recently, one of the tools that is currently used for such prediction is Large Language Models (LLMs).

2. How can we compare the quality of different predictors?

- We want to design the most accurate predictor.
- Thus, we need to be able to compare the quality of different predictors.
- Several natural characteristics can be used to describe this quality.
- For example, for each i, we can compute the probability $p_i \stackrel{\text{def}}{=} p(s_i|s_1, \ldots, s_{i-1})$ that:
  - the predictor correctly predicts the state s_i at moment i
  - based on the previous states $s_1, \ldots, s_{i-1}$.
- For each i, the larger the probability p_i, the better.
- But what if, for two predictors, we have $p_1 > p'_1$ but $p_2 < p'_2$?
- Which predictor should we select?

3. How can we compare the quality of different predictors?

- To be able to always compare the quality of different predictors:
  - we need to combine all these values $p_1, \ldots, p_n$ into a single number
    $$p = f_n(p_1, \ldots, p_n),$$
  - so the predictor with the larger combined probability is better.
- Which combination operation $f_n(p_1, \ldots, p_n)$ should we choose?

4. Empirical fact and the corresponding challenge

- It has been shown that:
  - among all proposed functions,
  - the most adequate comparison occurs when we select

$$f_n(p_1, \ldots, p_n) = \sqrt[n]{p_1 \cdot \ldots \cdot p_n}.$$

- But is this function indeed the best?
- Or is it simply the best of all the functions that have been tried, and a yet untried function will work better?

5. What we do in this talk

- We show that the empirically successful function is the only one that satisfies natural requirements.
- This explains why this function is empirically successful.
- It also confirms that no other yet untried function will be better.

6. First natural requirement

- Time is continuous.
- Our selection of the moments of time is arbitrary.
- Instead of the original time units – e.g., days, we could use weeks or months, etc.
- A natural requirement is that our measure of quality should not change if we simply choose a different unit.
- Let us describe this in precise terms.
- Suppose that we use a new unit which is k times larger; then:
  - predicting the state of the system in the new unit means
  - predicting the state which is k moments ahead if we use the original unit.
- For example, prediction for the next week means predicting 7 days ahead.

7. First natural requirement (cont-d)

- To correctly predict k old moments ahead, we need to:
  - correctly predict the next state s_{t+1} (for which the probability of successis p_{t+1}),
  - then correctly predict s_{t+2} based on s_t and s_{t+1} (for which the probability of success is p_{t+2}), etc.
- It is reasonable to assume that the predictions at different moments of time are statistically independent.
- So in the new units, the probability of the correct prediction is the product $p_{i+1} \cdot \ldots \cdot p_{i+k}$. In these terms:
  - the requirement that the measure of quality should not change if we select a different unit of time
  - takes the following form:

$$f_n(p_1, \ldots, p_n) = f_{n/k}(p_1 \cdot \ldots p_k, p_{k+1} \cdot \ldots \cdot p_{2k}, \ldots).$$

8. First natural requirement (cont-d)

- This must be true for all k, in particular, for $k = n$.
- Thus, we have $f_n(p_1, \ldots, p_n) = f_1(p_1 \cdot \ldots \cdot p_n)$.
- So, we reduced our problem to finding an appropriate function $f_1(x)$ of one variable.

9. Second natural requirement

- Suppose that all the probabilities p_i are the same, i.e.,

$$p_1 = \ldots = p_n = p.$$

- Then it is reasonable to use this common probability p as the measure of the predictor's quality.
- In view of the above formula, this means that we should have

$$f_n(p^n) = p.$$

- In particular, for each x, if we take $p = \sqrt[n]{x}$, then we get $f_1(x) = \sqrt[n]{x}$.
- For this function $f_1(x)$, the above equality leads to the desired formula

$$f_n(p_1, \ldots, p_n) = \sqrt[n]{p_1 \cdot \ldots \cdot p_n}.$$

10. Acknowledgments

This work was supported in part:

- by the US National Science Foundation grants:
  - 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science),
  - HRD-1834620 and HRD-2034030 (CAHSI Includes),
  - EAR-2225395 (Center for Collective Impact in Earthquake Science C-CIES),
- by the AT&T Fellowship in Information Technology, and
- by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Focus Program SPP 100+ 2388, Grant Nr. 501624329,