# 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
Department 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

**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).

**How can we compare the quality of different predictors?** We want to design the most accurate predictor. Thus, we need to be abkle 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? 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 that the predictor with the larger combined probability is better. Which combination operation $f_n(p_1, \ldots, p_n)$ should we choose?

**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?

**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 – and confirms that no other yet untried function will be better.

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

**Second natural requirement.** If all the probabilities $p_i$ are the same, i.e., if $p_1 = \ldots = p_n = p$, then it is reasonable to use this common probability $p$ as the measure of the predictor's quality.