Hey there!

Thank you for sharing your interest in the *Indo-European* languages.

If you also agree that studying their common roots and semantic similarities by applying state of the art **NLP** techniques on their respective ancient texts can answer a lot of pending questions we have about the **PIE** cultural origins – and also help in the reconstruction of the **PIE** language, please feel free to share some technical/ historical/ philosophical thoughts here along with some reference links..

The Indo-European languages are a family of languages spoken in most of Europe and areas of European settlement and in much of Southwest and South Asia¹. They are derived from a hypothetical *Proto-Indo-European* language (***PIE***) that is reconstructed by linguists based on similarities and differences among the modern languages. The oldest attested Indo-European languages are Anatolian, Indo-Iranian, Greek, Italic, Celtic, Germanic, Armenian, Tocharian, Balto-Slavic and Albanian¹. The speakers of PIE are called *Proto-Indo-Europeans*, and they are “believed” to have lived during the late Neolithic, or roughly the 4th millennium BC, in the Pontic–Caspian steppe zone in Eurasia (present-day Ukraine and southern Russia) ⁵. They had a society that was based on pastoralism, semi-nomadism, and a hierarchical kinship system⁷. They also had a religion that involved worshipping a sky father, a thunder god, a fire god, and other natural forces⁷.

(1) Indo-European languages - Wikipedia. https://en.wikipedia.org/wiki/Indo-European\_languages Accessed 16/4/2023.

(2) Proto-Indo-Europeans - Wikipedia. https://en.wikipedia.org/wiki/Proto-Indo-Europeans Accessed 16/4/2023.

(3) Indo-European languages | Definition, Map, Characteristics, & Facts. https://www.britannica.com/topic/Indo-European-languages Accessed 16/4/2023.

(5) Indo-European Languages - World History Encyclopedia. https://www.worldhistory.org/Indo-European\_Languages/ Accessed 16/4/2023.

(7) Proto-Indo-European society - Wikipedia. [https://en.wikipedia.org/wiki/Proto-Indo-European\_society Accessed 16/4/2023](https://en.wikipedia.org/wiki/Proto-Indo-European_society%20Accessed%2016/4/2023).

There are many words that are common in Indo-European languages, which show their shared origin from ***PIE***. Some of these words are related to basic concepts such as numbers, body parts, animals, and kinship terms. For example, the word for 'mother' is similar in many Indo-European languages: Sanskrit mātār, Greek mētēr, Latin mater, English mother, Armenian mayr, Old Irish mathair, Lithuanian motina¹. The word for 'two' is also similar: Sanskrit dva, Greek duo, Latin duo, English two, Armenian erku, Old Irish do, Lithuanian du¹. The word for 'cow' is another example: Sanskrit gauh, Greek bous, Latin bos, English cow, Armenian kov².

These words are called cognates, and they can be traced back to their reconstructed PIE forms by applying sound changes and comparing different branches of Indo-European languages. Linguists use these cognates to reconstruct PIE vocabulary and grammar. So, I am planning to do the same, but want to experiment by handing off this task to AI using **NLP** to get closest to **PIE** as possible using an appropriate Language Model.

(1) Indo-European vocabulary - Wikipedia. https://en.wikipedia.org/wiki/Indo-European\_vocabulary Accessed 16/4/2023.

(2) Common words of Indo-European languages - Hindu Website. https://www.hinduwebsite.com/general/indoeuro.asp Accessed 16/4/2023.

(3) indo european - Why do English, Italian, German, Spanish, French and .... [https://linguistics.stackexchange.com/questions/13284/why-do-english-italian-german-spanish-french-and-latin-share-a-common-alphab Accessed 16/4/2023](https://linguistics.stackexchange.com/questions/13284/why-do-english-italian-german-spanish-french-and-latin-share-a-common-alphab%20Accessed%2016/4/2023).

A language model is a type of artificial intelligence that can understand and generate human language. It works by analyzing large amounts of text data and using statistical methods to identify patterns and relationships between words and phrases. Once trained on this data, a language model can generate new text that is similar in style and content to the original data.

There are many different types of language models, but some of the most common include neural networks, Markov models, and n-gram models. Neural networks are a type of machine learning algorithm that can learn to recognize patterns in data by adjusting the weights of connections between neurons. Markov models use probability theory to predict what word or phrase is most likely to come next in a sentence based on the words that have come before it. N-gram models are similar to Markov models, but they use a sliding window of n words instead of just one or two. Even though each of these models uses a different approach to analyzing text data, but they all rely on statistical methods to identify patterns and relationships between words.

Language models are used in a wide variety of applications, including speech recognition, machine translation, chatbots, and more. They are also used in natural language processing (NLP) tasks such as sentiment analysis, named entity recognition, and text classification.

They have become increasingly important as more and more people rely on digital assistants like Siri, Alexa and ChatGPT to help them with everyday tasks.

(1) Large Language Models Explained: Purpose and Applications of LLMs. https://www.profolus.com/topics/large-language-models-explained-purpose-applications-llms/ Accessed 20/4/2023.

(2) A Beginner's Guide to Language Models | Built In - Medium. https://builtin.com/data-science/beginners-guide-language-models Accessed 20/4/2023.

(3) How Do Large Language Models Work? - Speak Ai. https://speakai.co/how-do-large-language-models-work/ Accessed 20/4/2023.

(4) What Are Large Language Models (LLMs) and How Do They Work? - MUO. https://www.makeuseof.com/what-are-large-langauge-models-how-do-they-work/ Accessed 20/4/2023.