EVALUATING THE EFFECTIVENESS OF UZBEK LANGUAGE PROCESSING MODELS BASED ON NEUROLINGUISTIC CRITERIA

Dushamova Nodirabegim – Independent Researcher

Abstract. This article examines the evaluation of the effectiveness of Uzbek language processing models based on neurolinguistic criteria. In conventional natural language processing systems, grammatical accuracy, lexical appropriateness, and semantic performance are considered the primary evaluation indicators. However, these indicators do not fully reflect the extent to which a model corresponds to human language comprehension processes. Therefore, this article explores the possibility of jointly applying such criteria as language comprehension time, contextual prediction, sensitivity to semantic incongruity, processing of grammatical complexity, and human evaluation. Taking into account the agglutinative structure, morphological richness, and context-dependent semantic features of the Uzbek language, a multi-criteria methodological approach to model evaluation is proposed. This approach may contribute to a more comprehensive investigation of the linguistic accuracy of artificial intelligence systems for Uzbek and their functional alignment with human language comprehension processes.

Keywords: artificial intelligence, neurolinguistics, natural language processing, Uzbek language, language comprehension, semantic processing, cognitive criteria, language model, EEG, eye-tracking analysis.

The development of artificial intelligence and natural language processing technologies is expanding the possibilities of automatically analyzing human language, identifying speech content, and generating context-based responses. Although modern language models have achieved high performance in text processing, the question of whether they possess mechanisms of meaning comprehension comparable to those of humans remains an important topic of scientific debate. Traditional evaluation methods measure model performance using accuracy, recall, the F1 score, perplexity, or the degree to which generated text corresponds to a reference answer. These indicators are useful for

evaluating performance in specific tasks; however, they do not directly represent the cognitive characteristics of human language processing.

Neurolinguistics investigates aspects of language perception and comprehension associated with brain activity, cognitive mechanisms, and temporal processes. Therefore, integrating neurolinguistic criteria with computational linguistics makes it possible to evaluate artificial intelligence models from a broader perspective.

This issue is particularly important for the Uzbek language. Its agglutinative nature, the expression of grammatical relationships through suffixes, flexible word order depending on communicative purpose, and the widespread use of polysemous units make automatic language analysis more complex. Consequently, Uzbek language models should be evaluated not only through general technical indicators but also through linguistic and cognitive criteria.

Natural language processing models perform tasks such as identifying linguistic units, analyzing grammatical relationships, representing textual meaning, and generating new text. Each of these tasks requires specific evaluation criteria. For example, the performance of a part-of-speech tagging model can be measured by the accuracy with which grammatical tags are assigned. In semantic analysis systems, correctly identifying meaning relationships between words and sentences is essential. Dialogue models must also be evaluated in terms of their ability to understand user intentions, maintain context, and generate semantically appropriate responses.

A neurolinguistic approach makes it possible to compare these results with human language processing mechanisms. In this context, two aspects should be distinguished:

Linguistic performance refers to the extent to which a model correctly performs grammatical, lexical, and semantic tasks.

Cognitive alignment refers to the degree to which model outputs and internal computational indicators correspond to certain observable characteristics of human language processing.

These two concepts may be related, but they do not represent the same phenomenon. A model's ability to produce a grammatically correct response does not prove that it thinks like a human or comprehends meaning in exactly the same way as a human.

Therefore, neurolinguistic evaluation should not aim to equate models with humans. Instead, it should focus on quantitatively and experimentally comparing specific characteristics of human language processing with the behavior of artificial intelligence systems.

Several complementary criteria can be used to evaluate Uzbek language processing models. During language comprehension, humans may use preceding words and sentence meaning to predict subsequent linguistic units. Language models also calculate the probability of the next word or token based on context.

For example:

“Talaba imtihonga puxta tayyorlandi, shuning uchun savollarga ...”

(“The student prepared thoroughly for the examination; therefore, they ...”)

In this context, continuations such as “javob berdi” (“answered”) are semantically expected. A model's ability to select a contextually appropriate continuation can help assess its predictive performance. However, high predictive accuracy does not, by itself, confirm that a model fully understands meaning. Therefore, this criterion should be applied alongside semantic, grammatical, and cognitive indicators. During language processing, humans respond differently to unexpected or semantically incongruous linguistic units. In neurolinguistic research, the N400 component recorded through electroencephalography (EEG) is widely used to investigate semantic expectancy and incongruity.

For example, the following sentences can be compared:

“U nonushtada issiq non yedi.” (“He/she ate warm bread for breakfast.”)

“U nonushtada issiq tosh yedi.” (“He/she ate a hot stone for breakfast.”)

In an ordinary situation, the second sentence is more semantically unexpected. However, it could convey meaning in a particular literary or unusual context. Therefore, identifying semantic incongruity requires consideration of the surrounding context rather than evaluating an isolated sentence alone.

It is possible to experimentally investigate how a language model evaluates semantic unexpectedness and to what extent its behavior correlates with the N400 responses observed in humans. This can be achieved by analyzing the relationship between model token probabilities or surprisal values and human EEG measurements.

When reading text or listening to spoken language, humans spend different amounts of time processing particular linguistic units. Rare words, complex grammatical structures, and contextually inappropriate expressions may affect processing time. Eye-tracking techniques can measure indicators such as first-pass fixation duration, regressions to previously encountered words, and total reading time. These data can be compared with the contextual predictions generated by language models.

For example, sentences containing morphologically complex Uzbek words can be compared with sentences containing grammatically simpler forms. If readers spend more time processing certain forms, the relationship between these differences and the model's probability estimates can be examined. It is important to recognize that reading time is not determined solely by language comprehension. Attention, word length, reader experience, and text complexity may also influence the results.

The agglutinative structure of Uzbek is characterized by a wide range of morphological forms. A single word may express several grammatical meanings. Therefore, the ability of a model to process grammatical relationships must be evaluated separately.

For example, the form “Talabalarimizga” (“to our students”) combines a stem with markers of plurality, possession, and the dative case. When analyzing such forms, it is

necessary to examine whether the model correctly identifies morphological tags and determines the word's grammatical function within the sentence.

To evaluate syntactic complexity, test samples can be designed to include subject-predicate agreement, case forms, subordinate clauses, and long-distance grammatical dependencies.

These indicators can be combined into a single overall score. However, the weights assigned to individual criteria must be justified in advance. Otherwise, the aggregate score may conceal important weaknesses in a model's performance. Accordingly, it is advisable to report results separately for each criterion before presenting an overall assessment.

In practical terms, this approach can be applied to evaluating machine translation systems, text editors, dialogue systems, educational platforms, and voice assistants for the Uzbek language. For example, in an educational system, evaluation may consider not only whether a model identifies a grammatical error but also whether it detects semantic inconsistencies in a student's answer and provides an understandable explanation.

The proposed methodology may help establish standardized test sets for comparing Uzbek language processing models, organize experiments involving human participants, and conduct more detailed analyses of model performance across different linguistic features.

Evaluating the effectiveness of Uzbek language processing models based on neurolinguistic criteria is a promising research direction for investigating the relationship between artificial intelligence and human language comprehension. Such evaluation supplements traditional technical indicators with criteria including contextual prediction, sensitivity to semantic incongruity, language processing time, and human judgments.

Taking into account the morphological richness and grammatical characteristics of Uzbek is an essential requirement for developing a reliable evaluation methodology. Therefore, it is advisable to combine high-quality annotated linguistic corpora,

standardized test samples, and, where necessary, experiments based on EEG or eye-tracking data.

At the same time, technical performance must be distinguished from cognitive alignment. A model's ability to predict certain characteristics of human language processing does not mean that it thinks like a human or comprehends meaning in precisely the same way. Careful scientific interpretation of the results and consideration of the limitations of each criterion will improve the reliability of the research.

Overall, integrating neurolinguistic and computational linguistic approaches provides a theoretical and practical foundation for more comprehensive evaluation of artificial intelligence systems for Uzbek, identifying their strengths and weaknesses, and improving future language technologies.

References

1. Abdurahmonova, N. *Kompyuter lingvistikasi* [Computational Linguistics]. Tashkent: Nodirabegim, 2021. 394 p.
2. Xolmanova, Z. *Kompyuter lingvistikasi* [Computational Linguistics]. Tashkent: Mehr-nuri, 2024. 234 p.
3. Hojiyev, A. *Tilshunoslik terminlarining izohli lug‘ati* [Explanatory Dictionary of Linguistic Terms]. Tashkent: O‘zbekiston Milliy Ensiklopediyasi, 2002. 168 p.
4. Ne’matov, H., and Bozorov, O. *Til va nutq* [Language and Speech]. Tashkent: O‘qituvchi, 1993.
5. Ne’matov, H., and Rasulov, R. *O‘zbek tili sistem leksikologiyasi asoslari* [Fundamentals of Systemic Lexicology of the Uzbek Language]. Tashkent: O‘qituvchi, 1995.