

The Process in Language Models (LM) and Their Problems in the Kurdish Language (CK) GPT-4o as an example

Assist prof Dr. Omar Ahmed Azeez ⁽¹⁾

Dr.Mohammed Majeed Saied ⁽²⁾

Umerahmed@garmian.edu.krd

Muhammed.majeed@garmian.edu.krd

Dr. Hemin Salahaldin Ramadhan ⁽³⁾

hemin.salahaldin@garmian.edu.krd

⁽¹⁾⁽²⁾⁽³⁾ Garmian University /College of Education/Kurdish Department

Date of Receipt of Research: 1/9/2025

Date of Publication Approval: 23/9/2025

Abstract

The study is entitled ((The Process in Language Models (LM) and Their Problems in the Kurdish Language (CK) GPT-4o as an example)) it attempts to present the processes of language models that each language model follows to generate language and then compare these processes with linguistic theories, whether the processes of generating language models resemble the theoretical mechanisms previously proposed by linguistic theories to analyze language, The research focuses on how Chat GPT-4o has a problem with Kurdish (Central Kurdish) to find out what the problems Kurdish faces are in language models, especially with GPT-4o. We focus more on morphological and syntactical issues that the model faces with Kurdish during the process of tokenization and parsing, the first being more relevant to morphology and the second to syntax.

Keywords: artificial intelligence, language model, tokenization, parsing, GPT-4o.

Introduction:

Syntactic analysis, or parsing, is a core process in both theoretical and computational linguistics, particularly within the field of artificial intelligence (AI). It involves the structural analysis of linguistic input—whether natural language or formal language—based on the rules of a formal grammar. In linguistic terms, syntax refers to the hierarchical and linear arrangement of words and phrases that form well-structured sentences. In the domain of AI and natural language processing (NLP), syntactic parsing is employed to uncover the grammatical structure of language, thereby facilitating more accurate and human-like language understanding and interaction.

This process entails segmenting a sentence into its constituent syntactic categories (e.g., noun phrases, verb phrases) and assigning grammatical functions such as subject, predicate, or object. Accurate syntactic analysis is a prerequisite for deeper linguistic tasks such as semantic interpretation and discourse analysis. The resulting output, typically represented as a parse tree

or a dependency structure, reflects the syntactic organization of the input and serves as a crucial interface between surface form and deeper linguistic meaning.

However, it is not only the syntactic level that describes language, but also the morphological level, which is the correct analysis of the structure of words into their component units, especially in complex words. Therefore, the study presents all the processes that linguistic models follow in order to generate language, and then practically presents the problems that Kurdish language has related to these processes, as we mentioned before we focus on syntax/parsing and morphology/ tokenization.

The hypotheses and research questions:

The morphological complexity, the free Word order within the SOV structure, the lack of standardized digital data for Kurdish, and the absence of a standard orthographical system cause GPT-4o and other language models to make more grammatical and semantic errors in Kurdish than the other languages that have a huge data seat for instance English.

This research seeks to answer the following question: What are the syntactic, morphological, and lexical problems in the Kurdish language (Central Kurdish) that cause language models like GPT-4o to make grammatical and semantic errors?

Methodology:

This study was conducted using the descriptive -analysis approach to examine the functions of the GPT-4o language model on Central Kurdish. The main objective of this study is to identify and classify the systematic errors of the model in the interpretation of linguistic correctness in the Kurdish language. Instead of a quantitative assessment on either a large and small scale, this approach emphasizes in-depth analysis of examples in order to gain a detailed understanding of challenges and problems. In this regard, so far, no intensive analysis and discussion no intensive analysis or discussion has been conducted on the Kurdish language on the Kurdish language. This study attempts to fill this gap by emphasizing a practical treatment of GPT-4o and provide the necessary evidence and analysis for this purpose.

The most significant processes of generative language in language models (GPT-4o):

Tokenization

In American structuralist linguistics, particularly in the work of Leonard Bloomfield, it is assumed that language can be broken down into smaller units known as Immediate Constituents (ICs). (Bloomfield, L. (1933) In artificial intelligence (AI), the concept of tokenization is similar. AI systems break down text into tokens to make it easier to process and understand. Tokenization is the process used by AI models to convert raw text into numerical segments, as AI cannot interpret raw text directly. A token can represent a full word, a sub word, or even individual characters including spaces. This step is essential for

helping AI models understand and generate human language (**Hamilton, S & Mimno, D(2025)**).

Embedding

Artificial intelligence models of languages or any machine language in NLP, unlike humans, machines can only understand code or numbers, because they are not human. so the process of embedding is converting all words in particular language to numerical code like 02040 for example for word (dog) another code for other words in this way the models can recognize and understand the meaning of words (**Daur, A(2021)** (**Andrea W Wen-Yi& David Mimno,(2023)**)

Parsing

Some of the most common parsing methods in natural language processing (NLP) include top-down parsing, bottom-up parsing, and chart parsing. While tokenization refers to breaking down sentences into smaller units (such as words or sub words), parsing goes a step further by identifying the syntactic structure of a sentence—distinguishing the relevant constituents and their relationships. In syntax, sentence trees (or parse trees) illustrate how phrases like NP (noun phrase), VP(verb phrase), and PP(prepositional phrase) relate to one another within a sentence.

This process in NLP often aligns with Chomsky and syntactic theory, which seeks to model how phrases are hierarchically organized. Parsing methods vary: top-down parsing begins from the highest syntactic level (such as the sentence) and works downward to individual words, while bottom-up parsing starts from individual tokens and builds upward toward a complete structure. Additionally, we distinguish between constituency parsing, which identifies phrase structure components, and dependency parsing, which focuses on the grammatical relationships and dependencies between words. These parsing strategies reflect core concerns in theoretical syntax and are essential for constructing accurate syntactic representations in computational models. The main point here is models don't understand trees and hierarchical sentence structure like humans do. Instead, they recognize relations through repeated training data. (**Madhuri A. Tayal, Dr. M. M. Raghuvanshi & Dr. Latesh Malik, (2014)** .

Language Models:

Another system closely related to syntax in Natural Language Processing (NLP) is the language model (LM). Language models enable AI tools to understand and generate text by predicting the likelihood of one word following another. This predictive ability is based on training with a large corpus of natural language.

In the field of syntax, a similar concept can be found in Finite State Grammar (FSG), introduced by Noam Chomsky in his book Syntactic Structures (**Chomsky, N. 1957**)

FSG operates on the principle of word transition probabilities—some word combinations are more probable than others. Although finite state grammars are limited in their ability to

account for complex and compound sentence structures, they remain significant, especially in machine language processing and computational linguistics.

Thus, the functioning of language models in NLP shares conceptual similarities with finite state grammars, particularly in their probabilistic approach to syntax (**Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaglu, Richard Socher, Xavier Amatriain, Jianfeng Gao, 2024**).

Syntax and Semantics:

Syntax refers to the structure of sentences and relationships between components in a sentence, helping distinguish grammatical sentences from ungrammatical ones. Semantics involves the study of meaning—whether at the word level or within entire sentences. Unlike humans, AI models cannot fully grasp meaning in the same way, especially when it involves cultural context, irony, deception, or emotional nuances that influence how people use language creatively or indirectly. However, AI models can still interpret word meanings within context based on patterns learned from large datasets. The challenge arises when dealing with semantic ambiguity and complex relationships such as polysemy (words with multiple meanings) or homophones (words that sound alike but have different meanings), which require more advanced algorithms for accurate interpretation.

As we mentioned some crucial methods in AI language models like Gpt-4o there are some other steps because AI cant directly generate and understand raw text like human therefore the first step is inputting data or raw text into the models like sentences in this step AI can remove some upper and lower cases like punctuation when they are not necessary after inputting data we have tokenization as we mentioned after that AI try to normalize the inputted text for example remove space capitalization to be normalize all text, then we have positional encoding unlike human inputting can't generate sentences with knowledge about the position of the words which one come first and so on for example in Kurdish the main word order is SOV, while English use SVO. Attention mechanism is also the process that AI language models use to identify the important words in the sentences to focus more on, for example, if we say

-The boy with the blue shirt who bought a car.

After tokenization and other processes, it knows that boy and bought is more important than, for example, the, a, with...etc.

Probability Estimation

we input a huge amount of data into the machine and we asked to generate sentences, it can't directly think and decide it works depending on probability for example it searches after i play for football, basketball, or any other words that it may come and more suitable after I play, so probability estimating is the significant process in Language models (**Jurafsky D. & Martin J. H. 2024**).

Sequences generation

It is related to the previous step, because many users ask AI to generate an essay or paragraph, so in this step, AI can end all sentences or paragraphs by step, not just one or two sentences, until it reaches the maximum length.

In the language model, we have some strategies for choosing tokens in probability estimation and sequence generation. For example, greedy **decoding is one** of the strategies that language models use. It includes the highest degree of probability. It is useful for some context, but on the other hand, it looks very dry and boring because of its repetition. Unlike the greedy the **Beams strategy** try to choose multiple sequences of token which the result is it gives more than one possibility also **sampling**, like beams gives or picks random tokens the result is more probability at one time so it looks more vary not repetitive while we have two other methods which control the balance between beams and sampling which are **top k and top p**, the first one keeps random to be not unlikely and top p keeps balance between creativity and safety in meaning (Jurafsky D. & Martin J. H. 2024)

Detokenization, as we mentioned, models work on tokens. After tokenization, it needs to be detokenized because the single words don't have meaning. It needs to gather and collect in one sentence, so it is the process of combining the single words to create the full sentences (Kamoda G. & et al., 2025).

(Weight-based Analysis of Detokenization in Language Models: Understanding the First Stage of Inference Without Inference Go Kamoda1 Benjamin Heinzerling, Tatsuro Inaba, Keito Kudo, Keisuke Sakaguchi, Kentaro Inui, Findings of the Association for Computational Linguistics: 2025,).

The post-processing:

After all steps, the output may need to remove some unwanted tokens that may be duplicated, and capitalize or fix spaces between the tokens.

Ranking / Re-ranking Outputs

Sometimes, the model doesn't just generate one answer — it can give many possible responses. But not all of them are good or correct. So the system needs to choose the best one. This step is called ranking or re-ranking, where the model checks all the options and picks the most suitable one

Morphological complexity:

Case studies of languages such as Turkish, Finnish, and Arabic (all of which have rich morphologies) show that large language models face serious problems in managing the morphological and syntactic structure of these languages (Çöltekin, Ç. (2020))

Kurdish is a complex language with a rich morphology, especially the verb conjugation system, which is complex because it uses several prefixes and suffixes to express tense, mood and person for example the verb ده‌نوسم (It consists of aspect /de-/, the root /noos/ and the first person singular suffix /-m/.

In addition, Kurdish makes extensive use of suffixes for existing and proper nouns.

These types of structures challenge the standard tokenization process, which itself actually works on the basis of (white-space), is considered a major problem and challenge for the Kurdish language. (Hassani, H. (2018)

Morphological complexities include not only those, but also compound and complex morphosyntactic verbs, which consist of several parts and are written together, for example, morphosyntactic verbs such as (1-a,b,c).

1-a - چاوپینانکهوت / He/She met them

b- وهبیرمهینانهوه / I reminded them

c - سهری لیتیکدان / He/She confused them

Understanding the meaning of these unfamiliar words for the model requires knowledge above/more than the meaning of the words themselves. Language models are likely to have difficulty or failure in segmenting and understanding the meaning of these words, especially if these words are very rare in the learning data.

As mentioned earlier, because Kurdish is an analytic and polysynthetic language, words and their component parts are written together, which becomes complicated for the model and the model needs to recognize the different parts of words.

Word Order in Kurdish:

Generally word order of phrases in central Kurdish is (subject + object + verb) (SOV), for example:

2- ئارام سێوهکهی خوارد. (Aram ate the apple.))

In sentence number (2), the word "ئارام/Aram" is subject, "سێوه/apple" is the object and "خوارد/ate" is the verb of the sentence Change the number (3. A & B) in several ways, without affecting the meaning of the sentence:

3-a - سێوهکه ئارام خواردی. (The apple aram ate.))

b- سێوهکهی خوارد، ئارام. (the apple ate, aram))

Although this replacement in Kurdish is for emphasis and simplification, which is called topicalization. However, this replacement confuses the model, because its output tends to follow fixed patterns, not the confusion The linguistic model should have a stickiness to the data evident.

Therefore, models such as GPT-4o, trained with English data (SVO) fixed, may have problems correctly interpreting Kurdish sentences with a different word order than SOV.

We asked GPT-4o to translate the sentences (1) and (2) into English:



Figure (1)

As we see, the first one is correct, but when it tries to translate the second one, which is the topic sentence, it changes it into passive voice because it couldn't recognize the Kurdish topic sentence.

Another problem that is expected to face the GPT-4o model is the expansion of phrases, especially noun and adjective phrases, for example:

4- شاری سلیمانی (Sulaimani city)

It is a noun phrase, consisting of three tokens, namely the words “شاری/city”, “ی”/izaafa, and “سلیمانی/Sulaimani”. So far, the model may recognize the tokens that make up the phrase, but when the phrase is expanded by modifications (noun or adjective), it may cause the model to make a mistake, for instance:

5- شاری رازاوه ی ڕهنگینی سلیمانی (The beautiful decorated city of Sulaimani)

This expression is still a phrase, but expanded by two modifications. It is assumed that because the model does not know that the words “decorated” and “beautiful” are descriptive of the word “Sulaimani”, it may make a mistake.

6- شاری رازاوه (Decorative city)

7- ڕهنگینی سلیمانی (Colorful Slemani)



Figure (2)

This may be the main reason for not fully recognizing the adjective (ی) because it does not know that it can be repeated more than once in a phrase, or that Kurdish phrases are expanded, or that this (ی) is an expanded mix in the individual. If you look at the experiment, after the model was asked to translate the phrase into English, instead of counting all three modifications as modifications, but the word (and) between (decorated and colorful) instead of (ی). This shows that although he could not deal with the expansion of the phrase, he solved the sentence differently.

This problem can occur and be repeated in expanding other types of phrases. Although what has been discussed so far is primarily due to the lack of recognition of tokens and referred to in this context, these problems can also create problems for "parsing", because obviously the role and activation of phrases is syntactic and parsing is directly related to this level.

Another problem expected to face the model is the changing form of the prepositions (به/to, بۆ/for, له/from). These prepositions are stored in the lexicon as (پێی/to, و/for, لێ/from), but when activated for syntax and morphology and during conversation their form will changes see (8. A,b,c,d,e & f).

- 8-a- نارام به ئۆتۆمبیل گهرايموه. (Aram came back by car)
- b- ئهو ئۆتۆمبیلهی نارام پێی هاتمهوه، شین بوو. (پێی changes to به) (The car that Aram came back in was blue)
- c- نەسرین چوو بۆ قوتابخانه. (Nasrin went to school)
- d- نەسرین چوو مته قوتابخانه. (ه/ه changes to بۆ) (Nasrin has gone to school.)
- e- سهیوان خۆی له من شاردموه. (Saywan hid himself from me)
- f- سهیوان خۆی لێ شاردموه. (لێ changes to له) (Saywan hid himself from me.)

ChatGPT

نهم دهقه كوردییه بۆ زمانی ئینگلیزی به شیوهی خوارهوه وه رگێردان:

1. نارام به ئۆتۆمبیل گه ریا وه.

.Aram returned by car

2. نه و ئۆتۆمبیله ی نارام پتی هاته وه، شین بوو.

.The car that Aram came with was blue

3. نه سرین چوو بۆ قوتابخانه.

.Nasrin went to school

4. نه سرین چوو ته قوتابخانه.

.Nasrin went to the school

(Note: The difference here is subtle in Kurdish; the second sentence may emphasize a specific school)

5. سه یوان خۆی له من شارده وه.

.Sêwan hid himself from me

6. سه یوان خۆی لێ شاردمه وه.

I hid Sêwan from him. (or from her/them — depending on context)

Figure (3)

Although this problem is directly related to tokenization and lexical items are independent tokens and they are expected to be recognized by the model in their lexical form, when they are activated in syntax and change their form, they become parsing objects and create problems for model parsing.

This in itself can be a big complication for the system, because these objects basically have the same meaning or value in two forms, which requires the model to be taught, these tokens have the same meaning and perform the same function in both forms.

To solve these problems, it is necessary to use artificial intelligence tools such as stemmers and morphological analyzers during the teaching of the model, or even beforehand, because these two tools are important in NLP.

Syntactical Agreement

Agreement is the grammatical relationship between the words(phrases) in a sentence. (Crystal, D. 2008)

In any language, grammatical agreement is of a specific type. The Kurdish language relies on bound pronouns to show grammatical agreement. In Kurdish, there are two sets of bound pronouns that agree with the subject according to (person, number, tense of the verb, transitivity and intransitivity of the verb). If a Kurdish sentence is changed into morphosyntax, then the object is also fixed within the verb by an bound pronoun. However, in general, in the Kurdish language, the subject agrees with the verb according to(person and number) . Gpt-4o has a problem with the agreement system of the Kurdish language and often makes mistakes, expressing a singular subject as plural and a plural subject as singular. Therefore, this is another one of the problems that the model faces in the process of parsing and determining the grammatical relationship between tokens.

We asked GPT_4o to

We asked the model to give us a short paragraph about car in general,

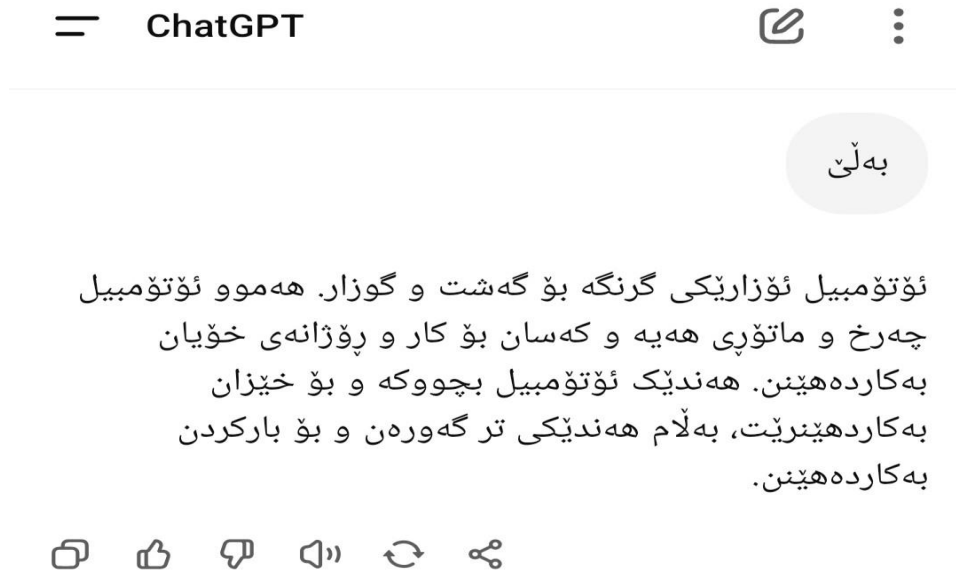


Figure (4)

The correct version of the below paragraph is:

9- " ئۆتۆمبیل ئامرازێکی گرنه بۆ گهشتوگوزار، ئۆتۆمبیل هه موویان تاپه و مه کینه یان هه یه، خه لک بۆ کاروباری " 9-
 رۆزانه ی خۆیان به کاریده هیئن، هه ندیک ئۆتۆمبیل بچوکن و بۆ خیزان به کارده هیئرین، به لام هه ندیک تر گه وره ن و بۆ
 "گواستنوه ی شتومه ک به کارده هیئرین

9-"A car is an important tool for tourism, all cars have tires and engines, people use it for their daily business, some cars are small and used for families, but others are large and used for transporting things. "

As we can see, in addition to lexical errors, the model made a grammatical agreement error.

Part II evaluates the output of the model

Because the Kurdish language does not have standard data, the models selects searches from the Internet and websites, of course, these articles do not follow the Kurdish standards They do not speak Kurdish and the expressions are often the everyday speech of the people, who follow the dialects of the community. This has led to confusion and the lack of a standard model for the system, so the system does not recognize the tokens and cannot decompose them and connect them in its output. That is, it does not recognize tokens syntactically and semantically. Here, most of the inflectional morphemes are lost. We assume that this is due to the above reason and due to the phonological system of the Kurdish language The system should also make use of incomplete sentences, for example article titles, or headlines of newspapers, magazines and websites (Straka, M. and Straková, J., 2017), this is when the

system is not given this function, this is in the normal thread For the latter, we expect GPT-4o to give us passive and incomplete sentences (see examples and figures below).

Common GPT-4o mistakes in simple and complex words

Tokenization methods vary according to the granularity of the texts and the function they play. The model must be able to process large texts into its component parts. which is to divide text and paragraphs into sentences and then into sentence components, including inflectional words, even punctuation, because these are part of it. They contain the meaning of sentences. It then divides these detailed units into more detailed units/characters, which is important here for a complex language such as Kurdish. Recognize the inflectional morphemes and turn them into tokens and their functions in syntax and semantics, and must be able to identify the components of complex lexical/morphological words (Morphological complexities). The study focuses on this point and tests them, but it should not be forgotten. Single-morphemic characters and words are also divided into smaller tokens, which we have avoided and will not test.

We asked GPT-4o to give us (5) five different sentences about sports Figure (5).

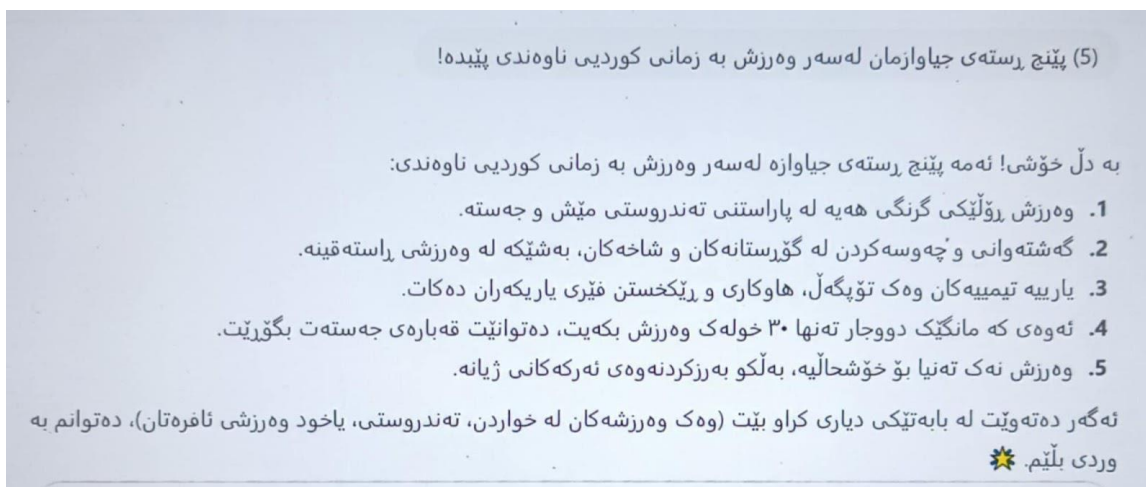


Figure (5)

The figure proves our above hypothesis. Looking at the sentences in the picture, the first sentence should look like this:

[".", "وهرزش", "روۆلیکی", "گرنگی", "هه‌یه", "له", "پاراستنی", "ته‌ن‌روستی", "مێش", "و", "جه‌سته"] -10

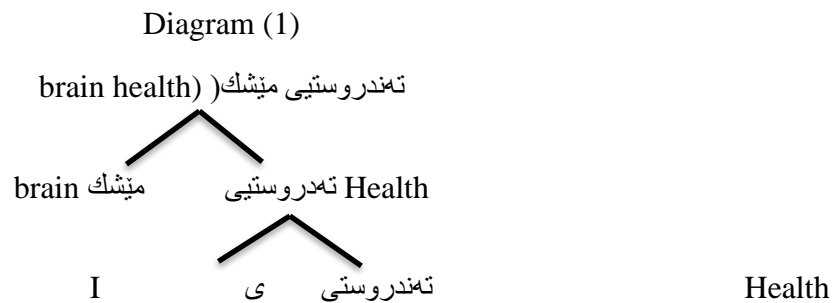
Looking at the example above, the sentence is fairly correct in terms of morphological and lexical complexity, but it is not perfect. We will leave the syntactic error in terms of the order of the sentence unites for the next topic. The system does not recognize the word (ته‌ن‌روستی/health) as a lexical word and as a determiner phrase/DP so it does not express the determiner (ی/i) in the sentence. (سه‌عید س. ز. 2020). At the same time, according to logic, the word “مێش” should mean “مێشک/brain”, so he has made a Lexical mistake. We present below the tokens of the same sentence with the same syntactic structure as in example (10):

[", "وهرزش", "رۆلێکی", "گرنگی", "ههیه", "له", "پاراستنی", "تەندروستی", "مێشک", "و", "جەستە"]-10

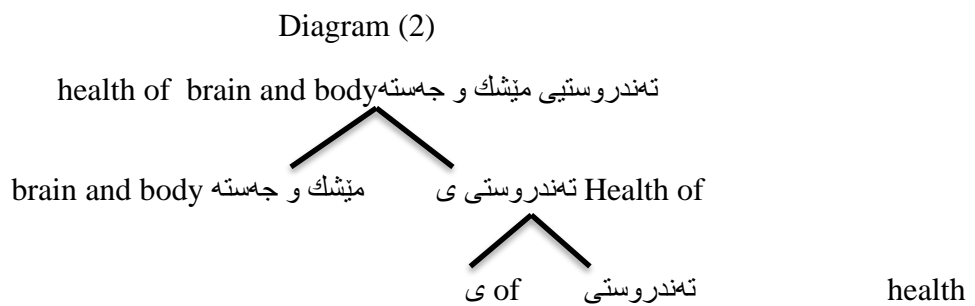
Example () contains two errors compared to example (), in the tokens (تەندروستی/health [and] "مێشک/fly"). If (health) tokens are presented correctly, it is as follows "health", "i" but (fly) is a Lexical error, that the system does not recognize the token, which is the Lexical element, which The token is "مێشک/brain", mistakenly identified as "fly/مێشک". Let us first assume that the reason for this is that in Kurdish the suffix /-k or ik/ ends some nouns and forms an abbreviated noun, or gives it a new meaning (دهسك، پەتک، درك), which are derived from the simple words (دەس/دهست، پەت، در). However, since there is no semantic relationship between مێشک/fly and مێشک/brain, we cannot assume that the model opened /-k/مێشکۆله/Cerebellum/ as a suffix to مێشک/fly, because in There is the word "مێشک/Brain" in Kurdish, so the reason for this mistake must be that the data stored in the model is not perfect, as mentioned earlier.

This suggests that the model does not recognize simple and uncomplicated lexical elements well and does not recognize the correct tokens of the phrases, so it is not familiar with the fact that (مێشک/brain and تەندروستی/body health) is an expanded phrase. The model does not recognize the determiner token (ی/i), which is syntactically correct in tree-diagram(2). First, we need to present the correct (تەندروستی مێشک/brain health) in tree diagram (1) for a more detailed analysis.

The syntactic correct is: [brain health]



The tree diagram above shows the correctness of the phrase, in which according to the rule of the head-final (مێشک/brain) is the complement of the structure, because (مێشک/brain and جەستە/body) is a conjunction phrase, which is due to conjunction Coordination conjunction occurs, provided that the two coordinated components are equal)2020 from (Falk. J. S,1978:). Tree diagram ().



The second sentence of the picture is not correct, in terms of syntactic correctness of phrases and punctuation, but the problem is in recognizing the lexical elements, the central Kurdish ,

We notice that these two errors are sometimes observed in the everyday speech of the speakers of the language, which reinforces our assumption that the data stored in the model is not perfect and it gets the data from the Internet.

The fourth sentence in the picture, a flawless sentence, the tokens are in place, the agreement of verbs, actions, and persons, the sentence is grammatically correct. For this purpose, the analysis of the sentence is presented below(13):

13- ئهوهی که، مانگیك دوو جار تهنیا 30 خولهك وهرزش i بکهیت j، (وهرزش) دهتوانیت i قهبارهی جهستهت j (تو) بگۆریت i.

- That, just 30 minutes of exercise is twice a month j, (exercise) can i change your body size j (you) i.

The fifth sentence of the picture is a complex sentence in terms of sentence components. The sentence is syntactically complete in terms of sentence components, but phonologically problematic. There is no phonological token in the first sentence, the interfix /ی/، which connects two morphemes, the first ending in a vowel and the second beginning with a vowel. Thus, the sentence of the tokens of the first sentence is like the example (14 a,b), when it should have been like the example (14. A & b).

14-a -] "وهرزش" ، "نهك" ، "تهنیا" ، "بۆ" ، "خۆشحالی" ، "ه" ، " ، "]

] b-] "وهرزش" ، "نهك" ، "تهنیا" ، "بۆ" ، "خۆشحالی" ، "ی" ، "ه" ، " ، "]

14. a. -["sports", "not", "only", "for", "fun", "a", " , "]

b-["sports", "not", "only", "for", "fun", "y", "a", " , "]

If the sentence is interpreted morphologically correctly, it must separate the morpheme /یه/ from the allomorph morpheme /ه-/، then again the sentence contains grammatical errors and the model is wrong the tones should be interpreted as (15):

15-] "وهرزش" ، "نهك" ، "تهنیا" ، "بۆ" ، "خۆشحالی" ، "یه" ، " ، "]

15-] "sports", "not", "only", "for", "fun", "is", " , "]

In Kurdish, the word (وهرزش/sports) is a noun, while the word (وهرزشکردن/sports) is a compound verb, so according to the second sentence, the noun must be in the form of a verb, that is, instead of the word (وهرزشکردن/sports). (وهرشوexercise), which proves that, once again, the model made a mistake in choosing the lexical element, so the linguistic analysis of the constituent elements of the sentence is in the form of (16), for example, when it should have been For example, in the form of(17 & 18).

16- وهرزشکردن-وهرزش i نهك تهنیا بۆ خۆشحالییه، بهلكو i pro بهرزکردنهوهی ئهرکهکانی ژیانه.

17- Exercise (i) not only for pleasure, but (pro i) to improve the tasks of life.

18- Exercise is not only for pleasure, but also for improving life's functions.

Common mistakes of Chat GPT-4o AI model in word order in Kurdish:

As mentioned earlier, the rule of order of elements in CK is (SOV). In this topic, this aspect of the sentences of the figure (5) and the placement of arguments and adjunction objects will be evaluated. The morphological aspects of the examples in Figure 5 were presented in the previous topic. Here we will not go back to the mistakes; we will treat the sentences in the same way that we have presented the correct ones.

Because the Kurdish language follows the above-mentioned rule, the first sentence of the figure does not follow the rule that the verb of the sentence should be put at the end, but the prepositional phrase is free in the sentence and can change its place, although Kurdish research has shown that this phrase appears after the verb in the sentence. That is, in terms of the order of the elements, the sentence should have been in the form of an example (19), because we do not assume that the model emphasizes one aspect of the sentence humanly. Only follow the word order rule of the language.

19- وەرزش رۆلێکی گرنگی له پاراستنی ئەندروستی مێشک و جەستە هەیە.

19- Exercise plays an important role in maintaining mental and physical health.

According to the example (19)/above, the order of the sentence tokens should have been in the form of example (20):

20-["وەرزش", "رۆلێکی", "گرنگی", "له", "پاراستنی", "ئەندروستی", "مێشک", "و", "جەستە", "هەیە", "،"]

20-["Sport", "a role", "importance", "in", "protection", "health", "brain", "and", "body", "has", "،"]

The second and third sentences, in addition to the interpretations made in terms of morphology and the choice of lexical elements for the sentence in the previous topic, in terms of the position of the elements flawlessly. We assume that, as we assumed earlier, the GPT-4o transfers the Kurdish language models from the Internet, and of course the Kurdish speakers of the language follow their language model. According to the principle of taking an ideal speaker for language model identification (**Chomsky, 1965: Pp.3**), the writer is expected to be an example of an ideal speaker. As he said "Linguistic theory is concerned primarily with an ideal speaker-listener, in a completely homogeneous speech-community, who knows its language perfectly and is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic) in applying his knowledge of the language in actual performance."

The fourth and fifth sentences are complex. Both sentences follow the word order rule in Kurdish language. Therefore, in this regard, they are correct sentences, the sentence coordination conjunctions are in position, so we only predict the interpretations presented in the previous topic.

References:**References****References**

1. Bloomfield, L. (1933) *Language*. New York: Henry Holt and Company.
2. Chomsky, N. (1957) *Syntactic structures*. The Hague: Mouton.
3. Chomsky, N. (1965) *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.
4. Çöltekin, Ç. (2020) 'A corpus for studying morphological, syntactic and semantic phenomena in Turkish', Proceedings of the 12th Language Resources and Evaluation Conference, pp. 417-427.
5. Crystal, D. (2008) *A dictionary of linguistics and phonetics*. 6th ed. Oxford: Wiley-Blackwell.
6. Daur, A. (2021) *A comparative evaluation of deep learning-based transformers for natural language to SQL query translation*. Master's thesis, University of Magdeburg.
7. Hamilton, S. and Mimno, D. (2025) 'Lost in space: Finding the right tokens for structured output', *arXiv preprint arXiv:2502*, v2, 1 August. Available at: <https://arxiv.org/abs/2502> (Accessed: [date]).
8. Hassani, H. (2018) 'Challenges of Kurdish text processing', *International Conference on Computational Linguistics and Intelligent Text Processing*, pp. 112-124. Springer.
9. Jurafsky, D. and Martin, J.H. (2024) *Speech and language processing*. 3rd ed. (Draft), Chapter 10. [Online]. Available at: [URL if available] (Accessed: [date]).
10. Kamoda, G. et al. (2025) 'Weight-based analysis of detokenization in language models: Understanding the first stage of inference without inference', *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 6324-6343.
11. McCarus, E.N. (1958) *A Kurdish grammar: Descriptive analysis of the Kurdish of Sulaimaniya, Iraq*. New York: American Council of Learned Societies.
12. Minaee, S. et al. (2024) 'Large language models: A survey', *arXiv preprint arXiv:2402.06196*, v3. Available at: <https://arxiv.org/abs/2402.06196> (Accessed: [date]).
13. Straka, M. and Straková, J. (2017) 'Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe', *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Vancouver, Canada, pp. 88-99. doi: 10.18653/v1/K17-3009.
14. Tayal, M.A., Raghuwanshi, M.M. and Malik, L. (2014) 'Syntax parsing: Implementation using grammar-rules for English language', *2014 International Conference on Electronic Systems, Signal Processing and Computing Technologies*, p. 376. IEEE.
15. Thackston, W.M. (2006) *Sorani Kurdish: A reference grammar with selected readings*. Cambridge, MA: Harvard University Press.
16. Toraman, C. et al. (2023) 'Impact of tokenization on language models: An analysis for Turkish', *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4), pp. 1-21.

17. Wen-Yi, A.W. and Mimno, D. (2023) 'Hyperpolyglot LLMs: Cross-lingual interpretability in token embeddings', *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1124-1131. Association for Computational Linguistics.
18. Yuan, X., Yang, Z., Wang, Y., Zhao, J. and Liu, K. (2024) 'Improving zero-shot LLM re-ranker with risk minimization'. Beijing: Institute of Automation, Chinese Academy of Sciences; School of Artificial Intelligence, University of Chinese Academy of Sciences; Beijing Academy of Artificial Intelligence.

1. سه‌عید، سزا زاهیر (2020)، نواندنی سنتاکسیی له زمانی کوردیدا، چاپی یه‌که‌م، چاپخانه‌ی ناوه‌ندی سارا بۆ چاپ و بڵاوکردنه‌وه، سلیمانی.

Conclusions:

1. From a technical perspective, all problems in the Kurdish language are due to the lack of complete data in language models, which becomes the cause of linguistic errors in outputs.
2. According to the tests, the GPT-4o model has problems in tokenization and parsing in the Kurdish language (Central Kurdish). GPT-4o still has major problems with the Kurdish language and needs more training and data enrichment from Kurdish language speakers.
3. Syntactic and morphological problems are numerous; we have highlighted some of them, including: non-recognition of tokens, errors in lexical unit selection, issues with SOV word order, grammatical agreement issues, recognition of determinative phrases, morphological changes in prepositions (بۆ، به، به) ...etc.
4. Sometimes the model makes spelling mistakes, especially in complex words and inflectional words that have more than one morpheme in their structure, due to lack of data, that the tokens are ignorant, and the ability to recognize lexical elements that have more than one representation for example preposition.
5. The model doesn't have a big problem with Kurdish word order, and it can generate the right sentence with word order.

عملية نماذج اللغة (ML) ومشاكلها في اللغة الكردية (الكردية الوسطى)) باستخدام نموذج GPT-4o كمثال

ملخص

تتناول الدراسة "عملية نماذج اللغة (ML) ومشاكلها في اللغة الكردية (الكردية الوسطى)) باستخدام نموذج GPT-4o كمثال". وتسعى الدراسة إلى عرض العمليات التي يتبعها كل نموذج لغوي في إنتاج اللغة. في اللغة الكردية (الكردية الوسطى)، من الضروري معرفة مشاكل اللغة الكردية في نماذج اللغة، وخاصة في نموذج GPT-4o. سنركز على المشاكل الصرفية والنحوية التي يواجهها النموذج في اللغة الكردية أثناء عملية الترميز والتحليل، حيث تتعلق الأولى بالصرف والثانية بالنحو.