Biannual scientific-research journal of language and linguistics
Fall & Winter 2025, 21 (42), 93-132
https://lsi-linguistics.ihcs.ac.ir/
DOI: 10.30465/lsi.2026.52503.1820

Original Research
Received: 19/Sep/2025 *** Accepted: 18/Jul/2025
ISSN: 2423-3847

Analyzing the Performance of Large Language Models in the Automatic Conversion of Colloquial Persian to Standard Writing based on the Confirmed Persian Orthography Grammar

Masood Ghayoomi[1] Faculty of Linguistics, Institute for Humanities and Cultural Studies, Tehran, Iran.

Fatemeh Mohammadi[2] Faculty of Linguistics, Institute for Humanities and Cultural Studies, Tehran, Iran.

Ali Jahan[3] Computer Science Student, Department of Computer Engineering, Amirkabir University of Technology

Abstract

The rise of digital communication tools, such as social networks and messaging platforms, has led to an expansion of diverse writing styles in Persian, including colloquial and non-standard (broken) forms. Converting such texts into standard Persian writing, especially based on the latest orthographic guidelines of the Academy of Persian Language and Literature, requires advanced natural language processing tools. This study compares the performance of five large language models, namely ChatGPT, Gemini, Perplexity, Claude, and DeepSeek, by automatically converting colloquial Persian into standard written form. The dataset utilized in this study is compiled and correct from two major corpora. A sub-corpus has been created through random sampling and resulted in 1,025 sentences and 11,939 tokens. Outputs from the models are compared to the gold standard annotations, focusing on both positive correction (e.g., correct use of pseudo-spaces, Ezafe, hamza, and standardizing informal forms) and

1. m.ghayoomi@ihcs.ac.ir (Corresponding Author)
2. f.mohammadi@ihcs.ac.ir
3. alijahan@aut.ac.ir

How to cite: Ghayoomi, M., Mohammadi, F., Jahan, A. (2025). Analyzing the Performance of Large Language Models in the Automatic Conversion of Colloquial Persian to Standard Writing based on the Confirmed Persian Orthography Grammar. *Language and Linguistics*, 21(42), 93- 132. doi: 10.30465/LSI.2024.50217.1783

negative errors (e.g., word replacement, deletion of clitics, and stylistic shifts). Findings indicate that the language model in Claude achieved the highest alignment with the gold standard (51.39%) at word and sentence levels, while Perplexity performed the worst. The results highlight that, despite their strengths, large language models still face challenges in accurately standardizing Persian text based on the formal orthographic conventions.

Keywords: colloquial Persian, non-standard (broken) writing, standardization, large language models, Persian orthography

1. Introduction

The widespread adoption of digital communication technologies, such as messaging applications and social media platforms, has significantly expanded public participation in written communication. As a result, diverse forms of writing have emerged, reflecting the linguistic practices of users from various educational, social, and cultural backgrounds. In Persian, this trend has led to the increasing use of non-standard and colloquial writing styles that closely mirror everyday speech rather than established orthographic conventions. While these forms demonstrate the dynamism and creativity of language in digital environments and provide valuable material for linguistic research, they pose challenges when the conversion of such texts into standardized writing is required. Therefore, the normalization of non-standard texts according to the official orthographic guidelines established by the Academy of Persian Language and Literature is of both practical and scholarly importance.

The emergence of Large Language Models (LLM) since 2020 has provided new possibilities for processing and standardizing Persian writing. These models, relying on deep learning and extensive text data, have the ability to convert non-standard and colloquial writing styles into standard writing. This paper contributes to convert non-standard spelling varieties into the standard ones by using LLM tools and to measure the degree to which the LLMs' output conforms to formal writing. To this end, the present study evaluates the performance of five LLM tools, including ChatGPT, Gemini, Perplexity, Claude, and DeepSeek. The conversion is compared with the latest version of orthographic guidelines of the Academy of Persian Language and Literature (2023). The results of this research can be effective in understanding the capacities and limitations of LLMs and evaluating their role in standardizing Persian writing in the digital space.

Research Question(s)

How the performance of the LLM tools is comparable to convert a colloquial and non-standard texts into standard ones?

2. Literature Review

There are a number of domestic and international research on the automatic conversion of colloquial or non-standard texts into standard ones. To this end, investigations show that sequential machine translation (Mansfield et al., 2019; Arora and Kansal, 2020), deep neural models, rule-based methods (Ariffin & Tiun, 2020), and statistical approaches are the most important solutions used in this field. International studies in English, Malay (Arora & Kansal, 2020), Kazakh (Kozhirbayev & Yessenbayev, 2020) and medical texts (Liu et al., 2021) have reported that the use of sub-words, encoder-decoder networks, BERT-based models, and combining rules with machine learning can significantly increase the accuracy of standardization and improve the performance of natural language processing systems, including sentiment analysis (Arora & Kansal, 2020) and grammatical labeling (Ariffin & Tiun, 2020).

In Persian, from the rule-based methods (Armin and Shamsfard, 2010) to transformer-based deep learning models (Rasooli et al., 2020; AbdiKhojasteh et al., 2020; Adibian & Momtazi, 2022), extensive efforts have been made to convert colloquial and non-standard texts into the standard ones. In addition to designing automatic converters, such as a spell-checker (Tajalli et al., 2023), these studies have focused on collecting parallel formal-informal corpora, using crowdsourcing to generate data (Masoumi et al., 2020) and analyze the features of rewriting in social networks (Ghayoomi & Mesgarkhoyi, 2023). The results show that a large part of cyberspace words are compatible with the standard spelling, but broken writing remains as a serious challenge. Overall, the available evidence suggests that combining large data sets, standard script, and modern deep learning models can provide an effective path for the automatic standardization of Persian writing and the improvement of natural language processing tools.

3. Methodology

To achieve the research goals, the web version of the LLM tools, including ChatGPT, Gemini, Perplexity, Claude, and DeepSeek, are used to exploit their latest features and capabilities. The input of the tools is non-standard spelling of Persian and it is expected to be converted into standard spelling rules by the tools. To evaluate the performance, the gold standard labels of

the input is required. To this end, two Persian corpora containing colloquial and non-standard Persian texts, prepared by Masoumi et al. (2020) and Tajali et al. (2023), are used. The first corpus contains 4500 sentences and the second one contains 50,011 sentences. By concatenating the two corpora, a corpus with 54,511 sentences and 584,418 words is obtained. Due to the limitations of using web versions of LLM tools, 1025 sentences are randomly selected from the combined corpus, which includes 11,939 tokens and 4016 type words. After comparing the output of the input with the gold standard labels, the changes, either corrections or mistakes, are manually categorized by experts into "positive" or "negative" categories. The "positive" performance includes:

1) Inserting the Ezafe clitic as "ی" /ye/, such as "خانه ما" /xāne.ye ma/ 'our home' which becomes "خانه‌ی ما";

2) Inserting the Ezafe clitic as " ۀ " /ye/, such as "خانه ما" /xāne.ye ma/ 'our home' which becomes "خانهٔ ما";

3) Inserting the pseudo-space, such as "ساقه هایش" /sāqe hāyaš/ 'its stalks' which becomes "ساقه‌هایش";

4) Correctly inserting the hamza, such as "تاثیر" /ta?sir/ 'impact' which becomes "تأثیر";

5) Correctly inserting the tanvin, such as "کلا" /kollan/ 'generally' which becomes "کلاً";

6) Correctly converting the colloquial and non-standard writing into the standard form, such as "دستتون" /dastetun/ 'your hand' which becomes "دستتان" /dastetān/;

7) Correctly inserting the mediating consonants, such as "توش" /tuš/ 'in it' which becomes "تویش" /tuyaš/;

8) Correcting the spelling or unintentional errors in writing words, such as "ینی" /yani/ 'meaning', which becomes "یعنی";

9) Substituting the standard spelling for fancy writing, such as "عسیس" /?asis/ 'dear', which becomes "عزیز" /?aziz/.

The negative changes in the model output that are not consistent with the gold standard labels are as follow:

1) Changing the word order in a sentence, such as the sentence " بیا دوست باشیم دیگر" /biyā dust bāšim digar/ 'Let's be friends' which can be changed to " بیا دیگر دوست باشیم";

2) Replacing a word, such as changing "مگر" /mage/ 'unless' to "آیا" /?āyā/ [the question word for yes/no questions];
3) Deleting clitics and replacing them with the corresponding simple word, such as "خودشانند" /xodešānand/ 'they are themselves', which can be changed to "خودشان هستند" /xodešān hastand/;
4) Not converting colloquial or non-standard texts into the standard one, such as "یه" /ye/ 'one', whose correct form is "یک" /yek/ but without a conversion, no changes have been done;
5) Incorrect insertion of Ezafe, such as "مجسمه‌های" /moǰassamehāye/ 'statues' which is incorrectly changed to "مجسمهٔ‌های";
6) Incorrect insertion of pseudo-space, such as "برای‌تان" /barāyetān/ 'for you' while the correct writing is "برایتان";
7) Deleting a word, such as "ماروپله‌اش رو بازی کنه" /māropelle?as ro bāzi kone/ 'he/she plays his snake and ladder game' which can be converted " ماروپله‌اش بازی کنه". In this example, "رو" /ro/ which is the correct form of "را" /rā/, is deleted from the sentence;
8) Inserting a new word, such as "می‌خواستم بروم داخل" /mixāstam beravam dāxel/ 'I wanted to enter' which is converted into "می‌خواستم بروم به داخل" /mixāstam beravam be dāxel/ 'I wanted to enter to'. In the conversion the preposition "به" /be/ 'to' is added;
9) Inserting a space incorrectly, such as "اون که" /?un ke/ 'one that' which is changed to "آنکه" /?ānke/ while the correct form is 'آن که';
10) No insertion of a proper space, such as "توهم" /toham/ 'you too', while the correct form is "تو هم" /to ham/;
11) Replacing the wrong word, such as "تو" /tu/ 'in' in the example "تو خانه" 'at home' which is converted into "در خانه" /dar xāne/
In general, any change that causes the text to deviate from the expected standard and correct form falls into this category of negative errors.

4. Results

According to the obtained results in Table 1, GPT and Perplexity performed the most and the least positive and negative impacts at the word level, respectively. The most positive impact of the language models had been made by Claude, and Perplexity performed the worst. DeepSeek, GPT, and Gemini stay in the second to fourth ranks, respectively. Among the positive

changes, inserting a pseudo-space had the most impact in converting non-standard text to standard. Inserting the Ezafe clitic applied the least change to non-standard data.

Furthermore, according to the negative impacts of the LLM tools on the data, Perplexity and GPT applied the least and the most negative impact, respectively. Among the negative changes, the conversion of colloquial or broken writing to standard writing had the most negative changes, such that Perplexity had the least and Claude had the most changes. Inserting a wrong hamza had the least negative effect on data conversion. In this category, DeepSeek and Gemini did not have any negative effect on data conversion. Word substitution was the most negative change applied by GPT.

Table 1

distribution of positive and negative changes of LLMs in sample data at the word level

	GPT	Gemini	Perplexity	Claude	DeepSeek
Positive	12.36	11.49	6.48	15.40	14.44
Negative	34.13	17.50	14.12	23.13	22.48

Next, we examined the performance of the LLM tools against the gold standard data in terms of exact matching at the sentence level for the input and output data. The results are reported in Table1. Based on the results, Claude obtained the highest matching rate (%51.39) and Perplexity the lowest matching rate (%90.7) with the gold standard data at the sentence level. The other language models, namely DeepSeek, Gemini, and GPT, ranked second to fourth respectively among the models. In examining LLMs with each other, without considering the gold standard data, it was concluded that Claude and DeepSeek perform similarly, and GPT and Perplexity had the lowest performance.

Table 2

comparing performance of LLMs with gold standard data at the sentence level

	Gold	GPT	Gemini	Perplexity	Claude	DeepSeek
GPT	98					
Gemini	244	84				
Perplexity	81	47	183			
Claude	405	136	293	97		
DeepSeek	337	118	231	108	368	

In the further studying the performances of LLMs, the models were compared from the perspective of syntax. First, the sentences obtained from LLMs and the gold standard data were part-of-speech tagged. Then, the content and functional words were separated from each other and the performances of the models were compared. The gold standard data contained %50.76 of content words and %50.23 of functional words. The results of the analysis are reported in Table 3. Based on the results obtained, Claude and Perplexity were able to obtain the largest and smallest volume of content and functional words, respectively. The ratio of content word changes in GPT (70.29%) was higher than other models, while Claude had decreased this ratio to %67.73. The ratio of functional word change in comparing GPT (%29.71) and Claude (%32.27) was the opposite such that Claude had a better performance in converting non-standard functional words into standard ones.

Table 3
distribution of content and functional words in LLMs compared to gold standard data

	Content word	Functional word
GPT	4.85	1.44
Gemini	14.33	4.42
Perplexity	4.00	1.32
Claude	25.23	8.14
DeepSeek	21.15	6.84

5. Conclusion

The general conclusion that can be drawn from the present study is that the difference in the results obtained from different LLMs indicates that due to the difference in their algorithms and training data of the models, the difference in their performance is obvious. Therefore, it is not possible to choose a single LLM and limit a research to that model. In a comprehensive study, it is necessary to examine various LLMs. Examining the differences can be a guide to choose the right tool in future research.

Bibliography

AbdiKhojasteh, H., Ansari, E., & Bohlouli, M. (2020). *LSCP: Enhanced large scale colloquial Persian language understanding*. Proceedings of the 12th International Conference on Language Resources and Evaluation, Marseille, France: European Language Resources Association, pp. 6323–6327.

Academy of Persian Language and Literature (2023). The Approved Orthography Rules of the Academy of Persian Language and Literature. Tehran: Persian Language and Literature Academy. [In Persian]

Adibian, M., & Momtazi, S. (2012). Converting Persian colloquial text to formal text using neural networks based on converters. *Journal of Language and Linguistics*, *18*(35), 45–67. [In Persian]

Armin, N., & Shamsfard, M. (2010). *Converting Persian colloquial text to formal text using n-grams*. Proceedings of the 16th Annual National Conference of the Iranian Computer Association. [In Persian]

Ariffin, S. N. A. N., & Tiun, S. (2020). Rule-based text normalization for Malay social media texts. *International Journal of Advanced Computer Science and Applications*, *11*.

Arora, M., & Kansal, V. (2020). Character level embedding with deep convolutional neural network for text normalization of unstructured data for Twitter sentiment analysis. *International Journal of Advanced Computer Science and Applications*, *11*.

Ghayoomi, M., & Mesgarkhoyi, M. (2023). Analysis of the corpus of Persian language data in cyberspace. *Language and Linguistics*, *19*, 37: 117–136. [In Persian]

Kozhirbayev, Z., & Yessenbayev, Z. (2020). *Kazakh text normalization using machine translation approaches*. CEUR Workshop Proceedings, 2780.

Liu, Y., Ji, B., Yu, J., Tan, Y., Ma, J., & Wu, Q. (2021). *An advanced ICD-9 terminology standardization method based on BERT and text similarity*. Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery. ICNC-FSKD 2020. Lecture Notes on Data Engineering and Communications Technologies, eds. Meng, H., Lei, T., Li, M., Li, K., Xiong, N., & Wang, L., Speringer, vol. 88, pp. 1868–1879.

Mansfield, C., Sun, M., Liu, Y., Gandhe, A., & Hoffmeister, B. (2019). *Neural text normalization with subword units*. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, eds. Loukina, A., Morales, M., & Kumar, R., Minneapolis, Minnesota: Association for Computational Linguistics, vol. 2, pp. 190–196.

Masoumi, V., Salehi, M., Veisi, H., Haddadian, G., Ranjbar, V., & Sahebdel, M. (2020). *TeleCrowd: A crowdsourcing approach to create informal to formal text corpora*. Arxiv.
https://arxiv.org/abs/2004.11771.

Rasooli, M. S., Bakhtyari, F., Shafiei, F., Ravanbakhsh, M., & Callison-Burch, C. (2020). *Automatic standardization of colloquial Persian*. Arxiv. https://arxiv.org/abs/2012.05879

Tajalli, V., Kalantari, F., & Shamsfard, M. (2023). *Developing an informal formal Persian corpus*. Arxiv. https://arxiv.org/abs/2308.05336.

مقاله پژوهشی

تاریخ دریافت: ۱۴۰۴/۰۴/۲۷ *** تاریخ پذیرش: ۱۴۰۴/۰۶/۲۸

ISSN: ۲۳۲۲-۳۸۴۷

دوفصلنامه علمی- پژوهشی زبان و زبان‌شناسی
دوره ۲۱، شماره ۴۲، پاییز و زمستان ۱۴۰۴، ۹۳- ۱۳۲
lsi-linguistics.ihcs.ac.ir
DOI: 10.30465/lsi.2026.52503.1820

تحلیلی بر عملکرد مدل‌های زبانی بزرگ در تبدیل خودکار گفتاری‌نویسی به نوشتار معیار براساس دستور خط مصوب فارسی

مسعود قیومی دانشیار، پژوهشکده زبان‌شناسی، پژوهشگاه علوم انسانی و مطالعات فرهنگی، تهران، ایران

فاطمه محمدی دانشجوی دکترای زبان‌شناسی، پژوهشکده زبان‌شناسی، پژوهشگاه علوم انسانی و مطالعات فرهنگی تهران، ایران

علی جهان دانشجوی کامپیوتر، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر، تهران، ایران

چکیده

رشد ارتباطات دیجیتال، مانند شبکه‌های اجتماعی و پیام‌رسان‌ها، موجب گسترش شکل‌های متنوعی از نوشتار در زبان فارسی، ازجمله گفتاری‌نویسی و شکسته‌نویسی، شده‌است. تبدیل این گونه‌های نوشتاری به نوشتار معیار، به‌ویژه براساس دستور خط مصوب فرهنگستان زبان و ادب فارسی، نیازمند بهره‌گیری از ابزارهای پیشرفته پردازش زبان طبیعی است. در پژوهش حاضر، عملکرد پنج مدل زبانی بزرگ شامل ChatGPT، Gemini، Perplexity، Claudeو DeepSeek در تبدیل خودکار گفتاری‌نویسی فارسی به نوشتار معیار براساس دستور خط مصوب ارزیابی شده‌است. داده‌های پژوهش از اصلاح و ترکیب دو پیکرهٔ حاصل از فضای مجازی حاصل شده و به‌صورت تصادفی یک زیرپیکره شامل ۱۰۲۵ جمله که حاوی ۱۱۹۳۹ واژه است تهیه شده‌است. تحلیل خروجی مدل‌ها بر مبنای برچسب‌های طلایی معیار و با تمرکز بر تغییرات مثبت (مانند درج نیم‌فاصله، کسرهٔ اضافه، همزه و اصلاح شکسته‌نویسی) و تغییرات منفی (مانند حذف واژه‌بست، جایگزینی واژه و تغییر سبک و جابه‌جایی واژه‌ها) انجام شده‌است. نتایج نشان می‌دهد مدل Claude با بیشترین تطابق با برچسب طلایی (۳۹/۵۱٪) در سطح واژه و جمله عملکرد بهتر و Perplexity ضعیف‌ترین عملکرد را از این منظر داشته‌است. این پژوهش نشان می‌دهد که مدل‌های زبانی بزرگ، علی‌رغم توانایی بالا، همچنان با چالش‌هایی در معیارسازی نوشتار فارسی، به‌ویژه در رعایت دقیق دستور خط، مواجه‌است.

کلیدواژه: گفتاری‌نویسی، شکسته‌نویسی، نوشتار معیار، مدل‌های زبانی بزرگ، دستور خط فارسی.

۱- مقدمه

وجود ابزارهای نوین ارتباطی مانند پیام‌رسان‌ها و شبکه‌های اجتماعی سبب شده‌است افراد جامعه، فارغ از سطح تحصیلات، طبقهٔ اجتماعی یا جنسیت به نگارش متن بپردازند. در مقایسه با گذشته که کتابت فقط به افراد فرهیخته جامعه محدود بود، این فناوری‌های نوین ارتباطی سبب شده‌است نوشتارِ متنوع توسط افراد مختلف جامعه تولید گردد. شکسته‌نویسی و گفتاری‌نویسی (محاوره‌نویسی) دو پدیدهٔ اجتناب‌ناپذیری است که ممکن است هنگام نگارش واژه‌ها توسط گویشوران به تحقق بیانجامد. با گسترش روزافزون این ابزارهای ارتباطی دیجیتال، نوعی دگرگونی چشمگیر در رفتار نوشتاری کاربران در زبان‌های مختلف، ازجمله زبان فارسی، شکل گرفته‌است. نوشتارهایی که در این بسترها تولید می‌شود، بیش از آنکه تابع قواعد سنتی و معیار زبانی باشد، متأثر از شیوهٔ گفتار روزمره و کاربردهای غیررسمی است. پدیده‌هایی همچون شکسته‌نویسی، گفتاری‌نویسی و استفاده از صورت‌های محاوره‌ای، به‌ویژه در نوشتارهای غیررسمی، به‌تدریج نه‌تنها بخشی از واقعیت زبانی فضای مجازی، بلکه به عرصه‌ای نوین برای پژوهش‌های زبان‌شناختی تبدیل شده‌است.

اگرچه حضور این پدیده‌ها در نوشتار مجازی بیانگر پویایی زبان و بازنمایی خلاقانهٔ گفتار در نوشتار است، در بسیاری از حوزه‌های رسمی، آموزشی، حقوقی و پژوهشی، نیاز به تبدیل این گونه‌های غیرمعیار به نوشتار معیار احساس می‌شود. ازآنجاکه برای زبانی مانند فارسی دستور خط توسط فرهنگستان زبان و ادب فارسی تهیه و منتشر شده‌است، به‌کارگیری این دستور به‌عنوان مرجع در معیارسازی نوشتار، به‌ویژه در مواجهه با متن‌های تولیدشده در فضای مجازی، ضرورتی کاربردی و علمی است.

تحول بزرگی که تقریباً از سال ۲۰۲۰ آغاز شده‌است تهیه و به‌کارگیری مدل‌های زبانی بزرگ[1] در ابزارهای مبتنی‌بر هوش مصنوعی مانند ChatGPT[2] است. بهره‌گیری از ظرفیت مدل‌های زبانی بزرگ، به‌عنوان یکی از پیشرفته‌ترین دستاوردهای فناوری در حوزهٔ پردازش زبان طبیعی، افق‌های تازه‌ای را در فرایند معیارسازی نوشتار فارسی گشوده‌است. این مدل‌ها، که برپایهٔ معماری‌های یادگیری عمیق و داده‌های عظیم متنی آموزش دیده‌است، می‌تواند در تبدیل خودکار گونه‌های غیررسمی و گفتاری زبان به شکل معیار، نقش مؤثری ایفا نماید. بااین‌حال، میزان انطباق این مدل‌ها با دستور خط معیار، توانایی آن‌ها در تشخیص و اصلاح انواع شکسته‌نویسی، و پرهیز از اعمال تغییرات نابه‌جا، هنوز به‌درستی بررسی و سنجیده نشده‌است.

---

1. Large Language Model (LLM)
2. https://chatgpt.com

در پژوهش حاضر، تلاش می‌شود با بهره‌گیری از داده‌های واقعی استخراج‌شده از پیکره‌های معتبر و متنوع، عملکرد پنج مدل زبانی بزرگ شامل ChatGPT،[1] Gemini، Perplexity،[2] Claude،[3] و DeepSeek[4] در تبدیل گفتاری‌نویسی به نوشتار معیار فارسی براساس آخرین نسخهٔ دستور خط مصوب فرهنگستان در سال 1402 تحلیل شود. هدف اصلی این مطالعه، سنجش میزان تطابق خروجی این مدل‌ها با معیارهای رسمی نگارش زبان فارسی و ارزیابی توان آنها در شناسایی و اصلاح الگوهای نوشتاری غیررسمی براساس دستور خط فارسی است. ازاین‌رو، پژوهش حاضر گامی در جهت ارزیابی عملی و انتقادی از قابلیت‌های مدل‌های زبانی بزرگ در مسیر همگرایی نوشتار فارسی در فضای مجازی به‌صورت معیار است.

۲. پیشینهٔ پژوهش

موضوع تبدیل متن نوشتاری غیرمعیار در فضای مجازی به متن معیار مورد توجه زبان‌های مختلف، ازجمله انگلیسی، مالایی و قزاقی، قرار گرفته‌است. زبان فارسی نیز مستثنی نیست و پژوهش‌هایی در این حوزه انجام گرفته‌است که در ادامهٔ این بخش به‌صورت کوتاه و فشرده توضیح داده می‌شود.

منزفیلد[5] و همکاران (۲۰۱۹) برای تبدیل متن محاوره‌ای به رسمی به‌عنوان یک قدم مهم در تبدیل متن به گفتار، از روش‌های ترجمهٔ ماشینی استفاده کرده‌اند. در این پژوهش، از روش دنباله-به-دنباله[6] در ترجمهٔ ماشینی استفاده شده‌است و ورودی و خروجی، دنباله‌ای از اجزای واژه‌ها است. به این مفهوم که جملهٔ مورد نظر به‌عنوان دادهٔ ورودی به‌صورت دنباله‌ای از زیرواژه‌ها به مدل داده می‌شود و دنباله‌ای از این زیرواژه‌ها به‌عنوان خروجی دریافت می‌شود و این زیرواژه‌ها در واقع اجزای سازنده واژه است که ترکیب آنها واژه را می‌سازد. این مورد برای آن است که نسبت‌به حالت استفاده از حروف، وابستگی‌های با فاصله زیاد را بهتر در نظر بگیرد. نتایج تجربی نشان می‌دهد کاربست این روش و البته استفاده از حجم بالای داده‌های متنی آماده در این زمینه در زبان انگلیسی، خروجی‌ها به امتیاز بلو[7] بالاتر از 98 درصد برسد.

1. https://www.gemini.com
2. https://www.perplexity.ai
3. https://claude.ai
4. https://deep-seek.chat
5. Mansfield, C.
6. sequence-to-sequence
7. BLEU score

آرورا[1] و کانزال[2] (2019) از این جهت به بررسی محاوره‌ای و غیرمعیاربودن پیام‌ها در شبکهٔ اجتماعی توییتر پرداخته‌اند که باعث می‌شود برای بسیاری از پردازش‌های الگوریتمی زبان طبیعی، مانند تحلیل احساس متن، پیام‌ها مناسب نباشد. در این مقاله، روشی برای معیارسازی متن نوشتاری و سپس کاربری آن در تحلیل احساسات پیام‌های توییتر ارائه شده‌است. در روش پیشنهادی، برای معیارسازی پیام‌های غیرمعیار توییتر، سه گام، شامل واحدسازی[3]، تشخیص و جایگزینی واژه‌های خارج از واژگان[4] و ریشه‌یابی واژه‌ها مطرح شده‌است. سپس، برای تحلیل احساسات از شبکهٔ عصبی پیچشی[5] مبتنی‌بر حروف استفاده شده‌است تا با استفاده از دو مرحلهٔ گفته‌شده بتوان تحلیل احساسات پیام‌های غیرمعیار توییتر را انجام داد. برای این فرایند، سه مزیت مطرح شده‌است که عبارت است از الف) داشتن امکان تحلیل احساس برای داده‌های غیرمعیار؛ ب) مدیریت حافظه در اثر استفاده از حروف به جای واژه‌ها؛ و ج) بهبود دقت در تحلیل احساس متن‌های غیر معیار.

آریفین[6] و تیون[7] (۲۰۲۰) در پژوهش خود به ارائهٔ روشی برای تبدیل متن محاوره‌ای به معیار در زبان مالایی پرداخته‌اند که در آن از مجموعهٔ قواعدی برای تبدیل متن محاوره به رسمی استفاده کرده‌اند. روش ارزیابی در این مقاله، بررسی دقت برچسب‌گذاری مقولهٔ دستوری واژه در استفاده از متن معیارشده ازطریق این روش است. در واقع هرچه معیارسازی در این روش بهتر باشد، دقت برچسب‌گذاری مقولهٔ دستوری واژه نیز بیشتر خواهد بود. نتایج عملی نشان می‌دهد که استفاده از این قواعد باعث بهبود محسوس دقت برچسب‌گذاری مقولات دستوری شده‌است. این دستاورد مبین این نکته‌است که معیارسازی انجام‌شده روش مناسبی برای تبدیل متن محاوره به معیار در متون شبکه‌های اجتماعی بوده و این روش می‌تواند باعث بهبود دقت سایر پردازش‌های مرتبط با زبان طبیعی شود.

کوژیربایو[8] و یسن‌بایو[9] (۲۰۲۰) تبدیل خودکار متن نوشتاری غیرمعیار به معیار در زبان قزاقستانی را مورد مطالعه قرار داده‌اند. در این روش، از راه‌کار ترجمهٔ ماشینی استفاده شده‌است تا متن غیرمعیار به معادل معیار متن تبدیل شود و از این متن بتوان در سایر فعالیت‌های مربوط به پردازش زبان طبیعی استفاده نمود. در این مقاله، از دو روش آماری و شبکهٔ عصبی

1. Arora, M.
2. Kansal, V.
3. tokenization
4. out of vocabulary
5. convolutional neural network
6. Ariffin, S. N. A. N.
7. Tiun, S.
8. Kozhirbayev, Z.
9. Yessenbayev, Z.

استفاده شده‌است. از روش آماری به‌عنوان روش پایه استفاده شده و روش آن مبتنی‌بر عبارات سه واژه‌ای است. در روش شبکهٔ عصبی از مدل دنباله-به-دنباله مبتنی‌بر واژه در ترجمهٔ ماشینی استفاده شده‌است. شبکهٔ عصبی استفاده‌شده در این مدل، شبکهٔ کدگذار-کدگشا[1] با استفاده از حافظهٔ کوتاه‌مدت طولانی[2] است. درنهایت برای ارزیابی نتایج به‌دست‌آمده از این مدل‌ها از معیار بلو استفاده شده‌است. نتایج تجربی نشان داده‌است روش آماری به امتیاز 21/67 و روش شبکهٔ عصبی به امتیاز 29/74 در معیار بلو دست یابد.

لیو[3] و همکاران (۲۰۲۱) روشی برای تبدیل اصطلاحات محاوره‌ای به اصطلاحات معیار که پزشکان در پرونده‌های پزشکی ثبت کرده‌اند ارائه داده‌اند. در این پژوهش، ابتدا روشی از ترکیب مدل برت[4] (دولین و همکاران، ۲۰۱۹) و الگوریتم شباهت متن ارائه شده‌است که با استفاده از چند واژه‌ای‌ها[5]، ابتدا واژه‌های منتخبِ جایگزین برای واژهٔ محاوره‌ای به‌دست می‌آید و سپس با استفاده از روش دسته‌بندی در برت، واژهٔ مناسب از بین واژه‌های منتخب جدا می‌شود. در ادامه، برای بهبود روش پیشنهادشده، سعی شده‌است اندازهٔ مجموعهٔ واژه‌های منتخبِ جایگزینی افزایش یابد تا امکان یافتن واژهٔ درست بیشتر شده و دقت بهبود پیدا کند و درعین‌حال برای حذف نوفه[6] از این مجموعه استفاده گردد. براساس نتایج عملی، دقت به‌دست‌آمده در روش اول 83/5 درصد است و با استفاده از راه‌کار گفته‌شده، این دقت با بهبود 0/6 درصدی به 84/1 درصد افزایش یافته‌است. نکتهٔ قابل توجه این است که حجم محاسبات در این حالت به 26/7 درصد کمتر از حالت عادی کاهش یافته‌است.

در حوزهٔ زبان فارسی در فضای مجازی، چندین پژوهش با هدف تبدیل متن محاوره‌ای غیرمعیار به رسمی انجام شده‌است. در همین راستا، مجموعهٔ دادهٔ کوچک و بزرگ غیرمنسجمی توسط پژوهشگران مختلف براساس اهداف مورد نظرشان گردآوری شده‌است که در ادامه توضیح داده می‌شود. آرمین و شمس‌فرد (۱۳۸۹) در روشی مبتنی‌بر قاعده و با استفاده از مدل‌سازی آماری، متن محاوره‌ای غیرمعیار را به معیار تبدیل کرده‌اند. پیکرهٔ مورد نیاز این پژوهش از داده‌های موجود در کتاب‌های حاوی متن محاوره‌ای، زیرنویس فارسی چندین فیلم و ده‌ها وبلاگ به‌دست آمده‌است که درنهایت پیکره‌ای با ۴۴ هزار واژه فراهم شده‌است. الگوریتم پیشنهادی در این پژوهش، به‌طورکلی، شامل واحدسازی متن محاوره و یافتن واژه‌های پیشنهادی برای صورت معیار آنها با استفاده از مجموعه‌ای از قواعد و درنهایت رتبه‌بندی

1. encoder-decoder
2. Long Short-Term Memory (LSTM)
3. Liu, Y.
4. BERT
5. n-gram
6. noise

مجموعهٔ پیشنهادی با استفاده از روش احتمالاتی است. در این پژوهش، بیش از بیست قاعده برای تبدیل تعریف شده و علاوه‌بر آن، پایگاه داده‌ای برای تبدیل محاوره به رسمیِ موارد بی‌قاعده فراهم شده‌است. همچنین، برای یافتن احتمال هر یک از واژه‌های منتخب، از دوواژه‌ای‌ها[1] استفاده شده‌است. برای به‌دست‌آوردن احتمال وقوع هر واژه مشروط به واژهٔ قبلی از پیکرهٔ بی‌جن‌خان (بی‌جن‌خان، ۱۳۸۳) استفاده شده‌است.

ادیبیان و ممتازی (۱۴۰۱) به مدلی دست یافته‌اند که عبارات حاوی گفتاری‌نویسی در محتوای شبکه‌های اجتماعی را به‌عنوان دادهٔ ورودی می‌پذیرد و شکل سالم واژه‌ها را به‌عنوان خروجی ارائه می‌دهد. مدلی که در این مبدل استفاده شده‌است مبتنی‌بر شبکهٔ عصبی عمیق است و از الگوریتم کدگذار-کدگشا و یادگیری دنباله-به-دنباله بهره برده‌است. دقت گزارش‌شده این مدل 70/7 درصد است.

قیومی و مسگرخویی (۱۴۰۲) در پژوهش خود به تحلیلی بر پیکرهٔ حاصل از داده‌های زبانی فارسی در فضای مجازی پرداخته‌اند. برای انجام این پژوهش یک پیکرهٔ زبانی از پیام‌رسان‌ها و شبکه‌های اجتماعی همچون اینستاگرام، واتس‌اپ و تلگرام و و توییتر که حاوی نوعی نونویسی[2] بوده و با نوشتار معیار واژه‌ها براساس دستور خط مصوب فرهنگستان زبان و ادب فارسی منطبق نبوده‌است تهیه شده‌است. تلاش شده‌است داده‌های گردآوری‌شده در سه موضوع متنوع اجتماعی، سیاسی و اقتصادی متوازن باشد. ویژگی این پیکره این است که در سه لایه برچسب‌گذاری شده و حاوی صورت‌های معیار دستور خط مصوب فرهنگستان (۱۴۰۲)، بن‌واژهٔ صورت‌واژه‌ها و همچنین برچسب مقولات دستوری واژه‌ها است. این پیکره حاوی ۲۰۱۷ پاره‌گفتار است که متشکل از ۱۴۵۷۱ صورت‌واژه است. در این پژوهش قیومی و مسگرخویی تلاش کرده‌اند ضمن به‌کارگیری قواعد شکسته‌نویسی ارائه‌شده توسط طبیب‌زاده (۱۳۹۸)، به استخراج قواعدی که از داده‌های فضای مجازی به‌دست آمده‌است بپردازند. براساس بررسی‌های صورت‌گرفته بر روی این پیکره و مقایسهٔ آن با دستور خط زبان فارسی، این نتیجه به‌دست آمد که 56/90 درصد از واژه‌های این پیکره بدون هرگونه تغییر در شیوهٔ نگارش با دستور خط زبان فارسی منطبق بوده و برای مابقی واژه‌ها نوعی نونویسی رقم خورده‌است. به‌طور کلی، 13/24 درصد از مجموع تغییرات در نگارش واژه‌ها، حاوی شکسته‌نویسی یا نونویسی در داده‌های فضای مجازی در سطح خط و ویژگی‌های نگارشی بوده و 36/04 درصد از مجموع تغییرات براساس ویژگی‌های آوایی-ساختواژی بوده‌است.

---

1. bigram
2. neograpgy

رسولی و همکاران (۲۰۲۰) از معیارسازی دنباله-به-دنباله در ترجمهٔ ماشینی برای تبدیل متن محاوره‌ای به رسمی در زبان فارسی استفاده کرده‌اند به این صورت که ورودی مدل، دنباله‌ای از واژه‌های محاوره‌ای و خروجی آن دنباله‌ای از واژه‌های معیار است. مدل استفاده‌شده، مدل ترجمه‌ای با استفاده از شبکهٔ ترنسفورمر با شش لایه بر پایهٔ برت (دولین و همکاران، ۲۰۱۹) است. در این مقاله، به‌دلیل نبود داده‌های محاوره‌ای و معادل رسمی آن، از مجموعهٔ قواعدی برای تبدیل متن رسمی به محاوره استفاده شده‌است تا دادهٔ مناسب برای آموزش مدل ساخته شود. برای این هدف، از متون صفحات ویکی‌پدیا و یک میلیون جمله از پیکرهٔ میزان (کاشفی، ۲۰۱۸) به‌عنوان متن معیار استفاده شده‌است. همچنین، برای داده‌های آزمون، از نظرات موجود در وبگاه‌های خبری و ترجمهٔ فیلم‌ها و دیالوگ فیلم‌ها استفاده شده‌است. برای محاسبهٔ دقت مدل پایه، از ابزار هضم[1] که مبتنی‌بر قواعد تبدیل است استفاده شده‌است.

معصومی و همکاران (۲۰۲۰) پژوهش خود را بر روش جمع‌آوری متون محاوره و صورت معیار آن ازطریق جمع‌سپاری[2] متمرکز کرده‌اند. در این پژوهش، ابتدا مجموعهٔ بزرگی از داده‌های متنی محاوره‌ای از توییتر و تلگرام و وبگاه دیجی‌کالا گردآوری شده‌است. سپس، در پیام‌رسان تلگرام سامانه‌ای طراحی شده‌است تا کاربران بتوانند جملات محاوره‌ای را وارد کنند. این بات[3] صورت معیار را به‌صورت خودکار به متن معیار به‌عنوان خروجی به کاربر نمایش می‌دهد. خروجی‌های این سامانه به‌عنوان جملات منتخب معیار برای جملهٔ محاوره‌ای درنظر گرفته می‌شود. همچنین در این سامانه افراد می‌توانند به جملات منتخبِ صورت معیار برای هر جملهٔ محاوره رأی دهند و باتوجه‌به تعداد افرادی که هر جملهٔ منتخب را قابل قبول دانسته‌است، آن صورت معیار برای آن جمله محاوره تشخیص داده شود. این سامانه TeleCrowd نامیده شده و توانسته‌است در زمان کوتاه‌تر و هزینه کمتری نسبت‌به حالت عادی، ۲۷۰۰ جملهٔ منتخب و ۲۱۰۰۰ رأی را فراهم آورد. برای ارزیابی، از یک آزمون ۵۰۰ جمله‌ای برای تبدیل گونه غیررسمی به رسمی فارسی استفاده شده‌است و با برچسب‌های طلایی معیار مقایسه شده‌است. با استفاده از معیار بلو امتیاز 0/54 به‌دست آمده‌است که امتیاز قابل قبلی است و بیانگر کیفیت قابل قبول دادهٔ گردآوری‌شده‌است. بیش‌از ۶۰ درصد از افراد مشارکت‌کننده در این فرایند جمع‌سپاری، در بازهٔ سنی ۲۰ تا ۳۰ سال قرار داشته و توزیع جنسیتی این افراد متوازن بوده و ۵۱ درصد مرد درمقابل ۴۹ درصد زن بوده‌است.

1. https://github.com/roshan-research/hazm
2. crowdsourcing
3. bot

عبدی‌خجسته و همکاران (۲۰۲۰) به‌منظور بهبود درک زبان محاوره‌ای فارسی با استفاده از روش‌های یادگیری ماشین مدرن، مانند یادگیری عمیق[1]، یک مجموعهٔ داده‌ای را به‌صورت سلسله مراتبی گردآوری کرده و بر روی درک چندوظیفه‌ای زبان غیررسمی فارسی تمرکز کرده‌اند. برای این هدف، یک مجموعهٔ داده بزرگ از فارسی گفتاری مکتوب از شبکهٔ اجتماعی توییتر تهیه شده‌است که شامل ۱۲۰ میلیون جمله از ۲۷ میلیون توییت است که به‌صورت خودکار برچسب مقولات دستوری واژه‌ها، تجزیهٔ نحوی وابستگی جملات، قطبیت احساسات و ترجمه به پنج زبان انگلیسی، آلمانی، چک، ایتالیایی و هندی به داده‌ها تخصیص داده شده و نشانه‌گذاری شده‌است. برای تهیهٔ این دادهٔ بزرگ و تبدیل داده‌های حاوی گفتاری‌نویسی به داده‌های معیار، از روش جمع‌سپاری استفاده شده و برای ترجمهٔ داده‌ها ترجمهٔ ابری گوگل[2] به‌کار برده شده‌است.

نعیمی و همکاران (۲۰۲۱) در پژوهش خود از روش‌های پردازش زبان طبیعی، مانند غلط‌یاب خودکار، برای تبدیل دادهٔ غیرمعیار فارسی به معیار استفاده کرده‌اند. برای این منظور دو مجموعهٔ داده برای واژه‌های معیار و غیرمعیار را از چهار پایگاه خبری پربازدید فارسی استخراج کرده‌اند. سپس، واژه‌های غیرمعیار را بر اساس میزان تغییرات لازم برای ساخت معادل‌های معیار، به چند دسته تقسیم کرده‌اند. نتایج نشان می‌دهد که فرایند پیشنهادی می‌تواند تقریباً ۹۴ درصد از واژه‌های غیرمعیار فارسی را تشخیص دهد. علاوه بر این، مقایسهٔ فرایند پیشنهادی با دو غلط‌یاب املایی معروف فارسی، ویراستیار و وفا، نشان می‌دهد که از نظر تشخیص و تصحیح، عملکرد قابل توجهی داشته‌است.

تجلی و همکاران (۲۰۲۳) اقدام به تهیهٔ یک پیکرهٔ موازی فارسی غیرمعیار و معیار از ۵۰ هزار جفت جمله در زبان فارسی پرداخته‌اند که شامل سبک‌های گونهٔ نوشتاری غیررسمی و رسمی حاصل از رسانه‌های اجتماعی، وبلاگ‌ها، رایانامه‌ها و پیام‌های متنی است. هدف این پیکره کمک به توسعهٔ ابزارهای پردازش زبان غیرمعیار فارسی است. این پیکره درحدود ۵۳۰ هزار واژهٔ همترازشده[3] و یک فرهنگ لغت با تقریباً ۴۹۳۹۷ جفت واژه و عبارت است.

خازنی و همکاران (۲۰۲۴) در پژوهش خود برای تحلیل احساسات داده‌های حاصل از فضای مجازی، به تهیهٔ یک مبدل برای تبدیل فارسی گفتاری غیرمعیار به معیار اقدام کرده‌اند و به بررسی چالش‌های تحلیل متون غیرمعیار در زبان فارسی پرداخته‌اند. در این پژوهش، از مدل‌های بدون نظارت استفاده شده‌است. برای این هدف، بیش از ۱۰میلیون متن بدون برچسب

1. deep learning
2. https://cloud.google.com/translate
3. aligned

از شبکه‌های اجتماعی و زیرنویس فیلم‌ها (به‌عنوان متون محاوره‌ای غیرمعیار) و حدود ۱۰میلیون متن خبری (به‌عنوان متون معیار) استفاده شده‌است. همچنین، ۶۰ هزار متن از نظرات کاربران شبکهٔ اجتماعی اینستاگرام با برچسب‌های مثبت، منفی و خنثی به‌عنوان داده‌های آموزش برای آموزش مدل دسته‌بندی احساسات متون کوتاه به‌کار برده شده‌است. با استفاده از این ابزار، ۵۷ درصد از واژه‌های پیکرهٔ غیرمعیار تبدیل شده و دقت 81/91 درصد را بر روی داده‌های آزمون به‌دست آورده‌است.

۳. چارچوب نظری

یکی از شاخص‌ترین پدیده‌های زبانی در نوشتار معاصر فارسی، به‌ویژه درگفت‌وگوهای انجام‌شده در فضای مجازی، نوشتار از شکل معیار فاصله می‌گیرد و به‌سمت نوشتار گفتاری‌نویسی یا شکسته‌نویسی گرایش پیدا می‌کند. این نوع نوشتار، بازتابی از گفتار روزمرهٔ مردم است که طی دهه‌های اخیر جایگاه مهمی در تولیدات زبانی غیررسمی یافته است. تحلیل شکسته‌نویسی را می‌توان در چارچوب نظریه دوزبانگونگی مورد بررسی قرار داد. فرگوسن (۱۹۵۹) این اصطلاح را برای توصیف موقعیتی به‌کار می‌برد که در آن، دو یا چند گونهٔ زبانی متمایز، به‌طور هم‌زمان، در یک جامعه زبانی وجود دارد. در زبان فارسی نیز این تمایز میان گونهٔ رسمی و معیار و گونهٔ گفتاری و غیرمعیار قابل مشاهده است. مطابق با دیدگاه قیومی و مسگرخویی (۱۴۰۲)، فضای مجازی با فراهم آوردن امکان استفاده هم‌زمان از دو گونهٔ معیار و غیر معیار زبان، موجب تقویت و تسریع فرآیند دوزبانگونگی در فارسی شده‌است. کاربران، به‌ویژه در محیط‌هایی مانند پیام‌رسان‌ها و شبکه‌های اجتماعی، به‌شکل گسترده‌ای از زبان گفتاری در نوشتار استفاده می‌کنند. در نتیجه، نوشتار روزمرهٔ فارسی در ابزارهای ارتباطی دیجیتال نوین با طیفی از شکسته‌نویسی‌ها و عناصر زبانی در گفتار شکل گرفته که مرز میان نوشتار و گفتار را کمرنگ می‌کند.

با پذیرش دوپدیدهٔ گفتاری‌نویسی و شکسته‌نویسی درفارسی، به تعریف و تمایز این دو نیاز است. طبیب‌زاده (۱۳۹۸) گفتاری‌نویسی یا محاوره‌گرایی را استفاده از امکانات گفتاری در آثار ادبی اطلاق می‌کند و براساس حوزه‌های زبانی، به چهار بخش تقسیم می‌کند: الف) زمانی که منعکس‌کننده ویژگی‌های نحوی است، مانند بجای «آنها به تهران رفتند» می‌گوییم «آنها رفتند تهران»؛ ب) زمانی که منعکس‌کننده پسوندهای عامیانه است، مانند« دختره»، «یواشکی» و «راستکی»؛ پ) زمانی که منعکس‌کننده واژگان گفتاری است، مانند «واسه»، «خفن»؛ و ت) زمانی که منعکس‌کننده ویژگی‌های آوایی است؛ مانند بجای «آنها به تهران رفتند» می‌گوییم «اونا به تهرون رفتن».

در ادامه، طبیب‌زاده (۱۳۹۸) شکسته‌نویسی را شکل نوشتاری واژه، وند یا پی‌بستی در نظر می‌گیرد که اولاً معادلی گفتاری داشته باشد و ثانیاً معادل آوایی سالمی در زبان فارسی داشته باشد. پس شکسته‌نویسی بخشی از گفتارنویسی است و نه برابر با آن. در ادامه چند نمونه برای صورت سالم واژه، وند و پی‌بست آورده شده‌است:

الف) در سطح واژه: صورت سالمِ «اگه» می‌شود «اگر». لازم به ذکر است که شکل نوشتاری مانند "گر" که به صورت کهن به‌کار می‌رود شکسته به حساب نمی‌آید؛ زیرا به سبک گفتاری مرتبط نیست و به سبک ادبی تعلق دارد.

ب) در سطح پسوند: صورت سالمِ «کتابا» می‌شود «کتاب‌ها»؛

پ) در سطح پی‌بست: صورت سالمِ «کتاب رو» یا «کتابو» می‌شود «کتاب را».

با تمیزدادن واژه‌های دارای دو پدیدهٔ گفتاری‌نویسی و شکسته‌نویسی، مسئلهٔ دیگری مطرح می‌شود و آن شیوهٔ نگارش واژه به‌صورت معیار است. طبیب‌زاده (۱۳۹۸) همچنین از دستور خط به‌عنوان مجموعه‌ای از قواعد نگارشی یاد می‌کند که به وحدت در نوشتار واژه‌ها و صورت‌های دستوری می‌پردازد. هدف اصلی این قواعد، تثبیت و معیارسازی شکل نوشتاری زبان به‌منظور افزایش خوانایی، فهم‌پذیری، و قابلیت پردازش متون است.

دستور خط در زبان‌های مختلف، به‌ویژه در دوران مدرن، نقش مهمی در تعامل میان زبان نوشتاری و زبان گفتاری ایفا می‌کند. در زبان فارسی، روند معیارشدن دستور خط از قرن نوزدهم میلادی آغاز شده‌است، اما در دهه‌های اخیر، به‌ویژه سه دههٔ گذشته، اهمیت آن در حوزه‌هایی همچون فناوری زبان، پردازش خودکار متن، و تبدیل متن به گفتار و بالعکس، دوچندان شده‌است. در این زمینه، رعایت دستور خط نه تنها به فهم بهتر متون مکتوب کمک می‌کند، بلکه در برنامه‌نویسی ابزارهایی که وظیفه پردازش الگوریتمی زبان طبیعی را برعهده دارد نیز یک ضرورت است. این ماشین‌ها برای تحلیل، تفسیر و تولید زبان، نیازمند داده‌هایی با دستور خط یکدست و معیار است. از این‌رو، دستور خط نه‌تنها یک مسئله زبانی و ادبی، بلکه یک نیاز فنی در توسعه فناوری‌های مرتبط با زبان، ازجمله موتورهای جست‌وجو، سامانه‌های تبدیل متن به گفتار و ابزارهای تصحیح خودکار است. این پیوند میان زبان‌شناسی و فناوری، ضرورت بازنگری، تدوین، و ترویج دستور خط معیار را بیش از پیش آشکار می‌سازد. در این پژوهش می‌کوشیم دستور خط مصوب فرهنگستان زبان و ادب فارسی را در تبدیل داده‌های حاوی گفتاری‌نویسی یا شکسته‌نویسی به شکل معیار در نظر داشته باشیم.

۴. ابزارها و داده‌های پژوهش

هدف از انجام پژوهش حاضر، مقایسه عملکرد مدل‌های زبانی بزرگ در تبدیل خودکار گفتاری‌نویسی یا شکسته‌نویسی به نوشتار معیار براساس آخرین دستور خط مصوب فارسی (۱۴۰۲) فرهنگستان خط و زبان فارسی است. برای این هدف، به ابزارها و داده‌های خاصی نیاز است که در ادامه توضیح داده می‌شود.

۴. ۱. ابزارهای پژوهش

در انجام پژوهش حاضر، از مدل زبانی بزرگ استفاده می‌شود که عبارت است از ChatGPT، Gemini، Perplexity، Claude و DeepSeek. برای به‌کارگیری این ابزارها در راستای اهداف پژوهش، از نسخهٔ تحت وب ابزارها استفاده شده‌است. ویژگی نسخهٔ تحت وب این است که امکانات آخرین نسخهٔ موجود ابزارها در دسترس بوده و از آن‌ها بهره گرفته شده‌است.

۴. ۲. داده‌های پژوهش

در انجام این پژوهش از دو پیکرهٔ فارسی حاصل از فضای مجازی که توسط معصومی و همکاران (۲۰۲۰) و تجلی و همکاران (۲۰۲۳) تهیه شده و حاوی شکسته‌نویسی و گفتاری‌نویسی است استفاده کرده‌ایم که در جدول ۱ اطلاعات آماری این دو پیکره گزارش شده‌است.

جدول ۱: اطلاعات آماری پیکره‌های استفاده‌شده در پژوهش حاضر

تهیه‌کننده پیکره	تعداد جملات/عبارات	تعداد واژه‌ها	تعداد واژه‌های یکتا
معصومی و همکاران (۲۰۲۰)	4500	54849	12292
تجلی و همکاران (2023)	50011	529569	65507
پیکره ترکیبی پژوهش حاضر	54511	584418	69234

پیکره‌ای که ازطریق جمع‌سپاری توسط معصومی و همکاران (۲۰۲۰) تهیه شده‌است حاوی جملات معیار و غیرمعیاری است که در سطح جمله ترازبندی شده‌است. یکی‌از دلایلی که سبب شده‌است این پیکره از کیفیت مناسبی برای کاربرد در پژوهش حاضر برخوردار نشود این است که به‌دلیل مشارکت افراد مختلف و گاهاً غیرمتخصص فارسی‌زبان، دستور خط فارسی رعایت نشده‌است. عدم رعایت فاصله‌گذاری صحیح سبب شده‌است داده‌های برچسب‌گذاری‌شده نتواند نیاز اهداف مورد نظر پژوهش حاضر را بر آورد. ویژگی دیگر این داده‌ها تغییر ساختار داده و استفاده از فرایند تفسیر به‌خصوص در مورد واژه‌بست‌ها است. برای مثال، عبارت غیرمعیار «چکار کنم باهاشون» به عبارت «با آنها چه کار کنم» تبدیل شده‌است. همچون پیکرهٔ

دیجی‌کالا[1]، در این پیکره نیز با تغییر ساختار، حذف به قرینه با تجلی ظاهری واژه/عبارت جایگزین شده‌است و این پدیدهٔ زبانی نادیده گرفته شده‌است. برای مثال، جملهٔ غیرمعیار «دلم با رفتنت تنگو دلم با بودنت خونه» به‌صورت عبارت «دلم با رفتنت تنگ است و دلم با بودنت خون است» برچسب‌گذاری شده‌است. اگرچه در جملهٔ معیار، اثری از شکسته‌نویسی یا گفتاری‌نویسی مشاهده نمی‌شود، ساختار جمله تغییر کرده و جمله‌ای که یک فعل داشته‌است به دو جملهٔ ساده با دو فعل تبدیل شده‌است.

پیکرهٔ تهیه‌شده توسط تجلی و همکاران (۲۰۲۳) که از تنوع داده‌ای و حجم قابل قبولی برخوردار است نیز دارای نقاط قوت و ضعف است. نقاط قوت این پیکره این است که ضمن داشتن تنوع از منابع گردآوری داده، از وبلاگ گرفته تا شبکه‌های اجتماعی، و حجم قابل قبول، در سطح واژه و جمله ترازبندی شده‌است. بنابراین، شکل غیرمعیار و معیار واژه‌ها نسبت‌به یکدیگر مشخص است. نقطهٔ قوت دیگر این پیکره این است که دستور خط مصوب (۱۳۸۹) و فرهنگ املایی صادقی و زندی‌مقدم (۱۳۹۴) به‌عنوان مرجع در نظر گرفته شده‌است که نتیجه آن یکدستی نسبی در شیوهٔ نگارش واژه‌ها و ترکیبات است. با این حال، این پیکره دارای نقاط ضعفی نیز هست. یکی‌از نقاط ضعف که در سایر پیکره‌ها نیز مشاهده شده‌است تغییر واژه و ساخت زبانی است. در تغییر واژه، گونهٔ گفتاری به شکل معیار تبدیل شده‌است. برای مثال، واژهٔ «واسه» که در گفتار وجود دارد و در گفتاری‌نویسی همچنان همان «واسه» است، در فرایند برچسب‌گذاری و تبدیل به دادهٔ معیار، این واژه به واژهٔ «برای» در مثال «واسه چی» و در مثالی دیگر، به واژهٔ «مال» در مثال «واسه حسن» تغییر کرده‌است. فرایند واژه‌سازی اتباع در دادهٔ معیار تهیه‌شده، نادیده گرفته شده‌است. برای مثال، واژهٔ «صیغه‌میغه» در دادهٔ حاصل از فضای مجازی به‌صورت «صیغه» در داده‌های معیار برچسب‌گذاری شده‌است. نمونه‌ای دیگر واژهٔ «هی» است که به «مدام» تغییر یافته‌است. در سطح عبارت و جمله، مثال‌هایی را می‌توان مشاهده کرد که سبب شده‌است ساختار زبانی به‌کل تغییر کند. برای مثال، شکل معیار عبارت شکستهٔ «خودتونید» که باید به‌صورت «خودتانید» باشد، در این پیکره «خودت هستید» برچسب‌گذاری شده‌است که در این برچسب‌گذاری علاوه‌بر حذف واژه‌بست، شکل سالم پیشنهادشده نیز اشتباه‌است. مثال دیگر، «تو نختم» است که به جملهٔ «در نخ تو هستم» تبدیل شده‌است. جملهٔ تک‌واژه‌ای «چمه» به عبارت «من را چه شده است» تبدیل شده‌است که گونهٔ غیرمعیار و معیار با یکدیگر قابل مقایسه نیست. درحالی‌که می‌توان «چه‌ام است» را به‌عنوان شکل سالم این عبارات در نظر گرفت که اولاً هیچ‌گونه تغییری در ساخت زبانی آن ایجاد نشود و ثانیاً با قواعد شکسته‌نویسی سازگار باشد. گاهی برچسب‌گذاری انجام‌شده اشتباه‌است. برای

1. https://www.kaggle.com/datasets/radeai/digikala-comments-and-products

مثال، «می‌پاشونه» که شکل شکستهٔ «می‌پاشاند» از مصدر «پاشانیدن» است به‌اشتباه «می‌پاشد» برچسب خورده‌است که مصدر آن «پاشیدن» است. نمونه‌ای دیگر «مالوند» است که صورت شکستهٔ «مالاند» از مصدر «مالاندن» است، اما واژهٔ «مالید» از مصدر «مالیدن» به‌عنوان برچسب به این واژه تخصیص یافته‌است. این اشتباده در شرایطی اتفاق می‌افتد که اساساً این دو مصدر از نظر معنایی با یکدیگر متفاوت است. همچنین، اکثر واژه‌بست‌ها در این پیکره همچون پیکره‌های دیگر به ضمیر تبدیل شده‌است.

در پژوهش حاضر، از این دو پیکره استفاده شده‌است با این تفاوت که داده‌های این دو پیکره بازبینی شده و اشکالات پیکره‌ها که نمونه‌های آن در بالا ذکر شد رفع شده‌است. از ترکیب این دو پیکره با هم، پیکره‌ای با حجم ۵۴۵۱۱ جمله به‌دست می‌آید. به‌دلیل محدودیت در به‌کارگیری ابزارهای تحت وب مبتنی‌بر مدل‌های زبانی بزرگ که در بخش پیشین معرفی شد، ۱۰۲۵ جمله به‌صورت تصادفی از این پیکرهٔ ترکیبی انتخاب شده‌است که حاوی ۱۱۹۳۹ واژه و ۴۰۱۶ واژهٔ یکتا است.

روش تحلیل داده‌ها به این صورت است که باتوجه‌به وجود برچسب طلایی داده‌های انتخاب‌شده، پس‌از حذف برچسب‌ها، این داده‌ها به‌عنوان دادهٔ ورودی به ابزارها داده می‌شود. سپس، خروجی ابزارها با برچسب‌های طلایی مقایسه شده و براساس تفاوت عملکرد خروجی مدل‌ها در مقایسه با برچسب‌های طلایی، به دسته‌بندی دستی خطاها توسط افراد خبره پرداخته می‌شود. عملکرد مدل‌ها در تبدیل داده‌های حاوی گفتاری‌نویسی یا شکسته‌نویسی به نوشتار معیار براساس آخرین دستور خط مصوب فرهنگستان زبان و ادب فارسی را به دو گروه تقسیم کرده‌ایم. دسته‌ای از تغییراتی که مدل‌ها بر داده‌ها اعمال کرده‌است و با برچسب معیار طلایی منطبق است را «مثبت» در نظر می‌گیریم. این دسته از تغییرات عبارت است از:

۱) درج کسرهٔ اضافه به‌صورت «ی»، مانند «خانه ما» که تبدیل می‌شود به «خانه‌ی ما»؛

۲) درج کسرهٔ اضافه به‌صورت «ۀ»، مانند «خانه ما» که تبدیل می‌شود به «خانهٔ ما»؛

۳) درج صحیح نیم‌فاصله، مانند «ساقه هایش» که تبدیل می‌شود به «ساقه‌هایش»؛

۴) درج صحیح همزه، مانند «تاثیری» که تبدیل می‌شود به «تأثیری»؛

۵) درج صحیح تنوین، مانند «کلا» که تبدیل می‌شود به «کلاً»؛

۶) تبدیل صحیح شکسته‌نویسی به شکل سالم و معیار، مانند «دستتون» که تبدیل می‌شود به «دستتان»؛

۷) درج صحیح یای میانجی، مانند «توش» که تبدیل می‌شود به «تویش»؛

۸) اصلاح غلط املایی یا اشکالات ناخواسته در تحریر واژه‌ها، مانند «ینی» که تبدیل می‌شود به «یعنی»؛

۹) جایگزینی واژه صحیح در موارد دارای لوس‌نگاری یا نوشتار فانتزی، مانند «عسیس» که تبدیل می‌شود به «عزیز».

در جدول ۲ نتایج ارزیابی حاصل از عملکرد مثبت مدل‌های زبانی بزرگ مورد نظر در پیکرهٔ انتخابی برای هر یک از معیارهای نه‌گانه معرفی‌شده از تناظر یک‌به‌یک واژه‌ها در دادهٔ ورودی دارای گفتاری‌نویسی و خروجی مدل‌ها گزارش شده‌است. شایان ذکر است در داده‌های بررسی‌شده، جایگزینی واژهٔ دارای ویژگی لوس‌نگاری دیده نشده‌است؛ ولی این نوع نوشتار در داده‌های فارسی فضای مجازی وجود دارد. نکتهٔ دیگر اینکه در بررسی عملکرد ابزارها، درج علائم سجاوندی اگرچه که یک تغییر مثبت است، اما نادیده گرفته شده‌است. براساس نتایج به‌دست‌آمده، بیشترین میزان اثرگذاری مثبت مدل‌های زبانی توسط Claude انجام‌شده و مدل زبانی Perplexity کمترین میزان تغییر را داشته‌است. مدل‌های زبانی DeepSeek، GPT و Gemini به‌ترتیب در جایگاه‌های دوم تا چهارم قرار دارد. ازمیان تغییرات، درج نیم‌فاصله بیشترین اثرگذاری را در تبدیل نوشتار غیرمعیار به معیار داشته‌است که از میان این ابزارها، Claude بیشترین اثرگذاری را در تبدیل داشته‌است. درج کسرهٔ اضافه، بدون درنظرگرفتن نوع آن به‌صورت «ی» یا «ۀ»، کمترین تغییری بوده‌است که بر داده‌های غیرمعیار اعمال شده‌است. در تبدیل گفتاری‌نویسی و شکسته‌نویسی به نوشتار معیار، Claude بیشترین تبدیل را به خود اختصاص داده‌است. تبدیل شکسته‌نویسی به نوشتار معیار سایر مدل‌ها به‌ترتیب عبارت است از: DeepSeek، Gemini، GPT و Perplexity که بیانگر عملکرد بهتر DeepSeek در جایگاه دوم است.

جدول ۲: تغییرات مثبت مدل‌های زبانی بزرگ در داده‌های نمونه در سطح واژه

مقوله	GPT	Gemini	Perplexity	Claude	DeepSeek
درج کسرهٔ اضافه ی	5	4	2	9	5
درج کسرهٔ اضافه ۀ	4	3	5	5	3
درج نیم‌فاصله	1005	921	455	1216	1125
درج همزه	14	5	7	18	19
درج تنوین	76	23	38	87	88
شکسته‌نویسی به معیار	337	378	235	450	425
درج ی میانجی	15	12	13	16	26
اصلاح غلط املایی	20	26	18	37	33
جایگزینی واژهٔ حاوی لوس‌نگاری	0	0	0	0	0
جمع کل	1476	1372	773	1838	1724

مثال‌های ۱ تا ۹ بیانگر عملکرد مدل‌های مختلف زبانی بزرگ مورد مطالعه در تغییرات مثبت هشتگانه است. شایان ذکر است در عملکرد ابزارها و خروجی ارائه‌شده توسط آنها، الزاماً تمامی ابزارها ویژگی مطرح‌شده را ندارد. بنابراین، در بعضی از ابزارها ممکن است تغییر اعمال شده باشد و در برخی خیر.[۱]

۱) درج کسرهٔ اضافه ی:

۱	دادهٔ ورودی	آقا ما وقتی فهمیدیم سیگار چیه که تو ویدیوش دیدیم یه سیگار روشن گذاشته رو دسته گیتارش لای سیما
۲	دادهٔ طلایی معیار	آقا ما وقتی فهمیدیم سیگار چیست که توی ویدیواش دیدیم یک سیگار روشن را گذاشته روی دسته‌ی گیتارش لای سیم‌ها
۳	خروجی GPT	آقا ما وقتی فهمیدیم سیگار چیست که در ویدیوی او دیدیم یک سیگار روشن را بر روی دسته گیتارش در لای سیما گذاشته است
۴	خروجی Gemini	آقا ما وقتی فهمیدیم سیگار چیست که در ویدیوی او دیدیم یک سیگار روشن را روی دسته گیتارش لای سیما گذاشته است
۵	خروجی Perplexity	آقا ما وقتی فهمیدیم سیگار چیست که تو ویدیوش دیدیم یک سیگار روشن را گذاشته رو دسته گیتارش لای سیم ها
۶	خروجی Claude	آقا ما وقتی فهمیدیم سیگار چیست که در ویدیوش دیدیم یک سیگار روشن را گذاشته روی دسته گیتارش لای سیم‌ها
۷	خروجی DeepSeek	آقا ما وقتی فهمیدیم سیگار چیست که تو ویدیوش دیدیم یک سیگار روشن را گذاشته روی دسته گیتارش لای سیم‌ها

۱- واژه‌های هدف در دادهٔ ورودی و دادهٔ طلایی معیار به‌عنوان واژه‌های هدفِ مرتبط با مقولهٔ مورد نظر، به‌صورت پررنگ در متن مشخص شده‌است. واژه‌ها یا عباراتی که براساس شاخص مورد نظر توسط ابزار در واژهٔ هدف تغییر اعمال شده است یا خیر با زیرخط مشخص شده‌است.

۲) درج کسرهٔ اضافه ۀ

۱	دادهٔ ورودی	تو مثل من زمستونی نداری که باشه لحظه ی چشم انتظاری گلدون خالی ندیدی نشسته زیر بارون گلای کاغذی داری تو گلدون
۲	دادهٔ طلایی معیار	تو مثل من زمستانی نداری که باشد لحظهٔ چشم‌انتظاری گلدان خالی ندیدی نشسته زیر باران گل‌های کاغذی داری توی گلدان
۳	خروجی GPT	تو مانند من زمستونی نداری که باشد لحظه‌ی چشم‌انتظاری گلدان خالی ندیدی که نشسته زیر باران گل‌های کاغذی در گلدان
۴	خروجی Gemini	تو مثل من زمستانی نداری که باشد لحظه ی چشم انتظاری کلدان خالی ندیدی نشسته زیر بارون گلای کاغذی داری تو گلدون
۵	خروجی Perplexity	تو مثل من زمستانی نداری که باشد لحظه ی چشم انتظاری گلدون خالی ندیدی نشسته زیر بارون گل‌های کاغذی داری تو گلدون
۶	خروجی Claude	تو مثل من زمستانی نداری که باشد لحظه‌ی چشم‌انتظاری گلذان خالی ندیدی نشسته زیر باران گل‌های کاغذی در گلدان
۷	خروجی DeepSeek	تو مثل من زمستونی نداری که باشد لحظهٔ چشم‌انتظاری گلدون خالی ندیدی نشسته زیر باران گل‌های کاغذی دای تو گلدون

۳) درج نیم‌فاصله

۱	دادهٔ ورودی	باس پاشم چرا نمیخوابم اخه
۲	دادهٔ طلایی معیار	بایست پا شوم چرا نمی‌خوابم آخر
۳	خروجی GPT	باس پاشم چرا نمی‌خوابم آخه
۴	خروجی Gemini	باید پاشم چرا نمی‌خوابم آخه
۵	خروجی Perplexity	باس پاشم چرا نمی خوابم آخه
۶	خروجی Claude	باید پاشم چرا نمی‌خوابم آخر
۷	خروجی DeepSeek	باس پاشم چرا نمی‌خوابم اخه

۴) درج همزه

۱	دادهٔ ورودی	متاسفانه رکورد ما از دست رفت در گل نخوردن
۲	دادهٔ طلایی معیار	متأسفانه رکورد ما از دست رفت در گل نخوردن
۳	خروجی GPT	متأسفانه رکورد ما از دست رفت در گل نخوردن
۴	خروجی Gemini	متاسفانه رکورد ما از دست رفت در گل نخوردن
۵	خروجی Perplexity	متاسفانه رکورد ما از دست رفت در گل نخوردن
۶	خروجی Claude	متأسفانه رکورد ما از دست رفت در گل نخوردن
۷	خروجی DeepSeek	متأسفانه رکورد ما از دست رفت در گل نخوردن

۵) درج تنوین

۱	دادهٔ ورودی	اصن به این موضوع فکر نکرده بودم
۲	دادهٔ طلایی معیار	اصلاً به این موضوع فکر نکرده بودم
۳	خروجی GPT	اصلاً به این موضوع فکر نکرده بودم
۴	خروجی Gemini	اصلاً به این موضوع فکر نکرده بودم
۵	خروجی Perplexity	اصلا به این موضوع فکر نکرده بودم
۶	خروجی Claude	اصلاً به این موضوع فکر نکرده بودم
۷	خروجی DeepSeek	اصلاً به این موضوع فکر نکرده بودم

۶) شکسته‌نویسی به معیار

۱	دادهٔ ورودی	پریروز خاله هام و برده بودم هواخوری
۲	دادهٔ طلایی معیار	پریروز خاله‌هایم را برده بودم هواخوری
۳	خروجی GPT	پریروز خاله‌هایم را برده بودم هواخوری
۴	خروجی Gemini	پریروز خاله‌هایم را برده بودم هواخوری
۵	خروجی Perplexity	پریروز خاله‌هایم را برده بودم هواخوری
۶	خروجی Claude	پریروز خاله‌هایم را برده بودم هواخوری
۷	خروجی DeepSeek	پریروز خاله‌هایم را برده بودم هواخوری

۷) درج ی میانجی

۱	دادهٔ ورودی	گفتم رفت تا تلفنخونه و بیاد
۲	دادهٔ طلایی معیار	گفتم رفت تا تلفن‌خانه و بیاید
۳	خروجی GPT	گفتم رفت تا تلفن‌خانه و بیاید
۴	خروجی Gemini	گفتم رفت تا تلفن‌خانه و بیاید
۵	خروجی Perplexity	گفتم رفت تا تلفن‌خانه و بیاید
۶	خروجی Claude	گفتم رفت تا تلفن‌خانه و بیاید
۷	خروجی DeepSeek	گفتم رفت تا تلفنخانه و بیاید

۸) اصلاح غلط املایی

۱	دادهٔ ورودی	خعلی باحال شد داستانت
۲	دادهٔ طلایی معیار	خیلی باحال شد داستانت
۳	خروجی GPT	خیلی باحال شد داستانت
۴	خروجی Gemini	خیلی باحال شد داستانت
۵	خروجی Perplexity	خیلی باحال شد داستانت
۶	خروجی Claude	خیلی باحال شد داستانت
۷	خروجی DeepSeek	خیلی باحال شد داستانت

۹) جایگزینی واژهٔ حاوی لوس‌نگاری

۱	دادهٔ ورودی	نخیرمم دولوخ میگه
۲	دادهٔ طلایی معیار	نخیر هم دروغ می‌گوید
۳	خروجی GPT	نخیرم او دروغ می‌گوید
۴	خروجی Gemini	نخیر هم دروغ می‌گوید
۵	خروجی Perplexity	نخیرم دولوخ می‌گوید
۶	خروجی Claude	نخیرمم دولوخ می‌گوید
۷	خروجی DeepSeek	نخیرمم دولوخ می‌گه

دستهٔ دیگری از تغییرات مدل‌ها با برچسب‌های معیار طلایی مطابقت ندارد که این دسته از تغییرات را منفی در نظر می‌گیریم. این تغییرات منفی عبارت است از:

۱) تغییر چینش واژه در جمله، مانند جملهٔ «بیا دوست باشیم دیگر» که تبدیل شده‌است به «بیا دیگر دوست باشیم»؛

۲) جایگزینی واژه، مانند تبدیل «مگه» به «آیا» در شرایطی که صورت سالم «مگر» است؛

۳) حذف واژه بست، مانند «خودشانند» که تبدیل شده‌است به «خودشان هستند»؛

۴) عدم تبدیل گفتاری‌نویسی یا شکسته‌نویسی به نوشتار معیار، مانند «یه» که صورت سالم آن «یک» است اما همان «یه» نوشته شده‌است و تبدیل صورت نپذیرفته‌است؛

۵) درج اشتباه همزه، مانند «مجسمه‌های» که تبدیل شده‌است به «مجسمهٔ‌های» و همزه به اشتباه درج شده‌است؛

۶) درج اشتباه نیم‌فاصله، مانند «برای‌تان» که نگارش صحیح «برایتان» است؛

۷) حذف واژه، مانند «ماروپله‌اش رو بازی کنه» که به «ماروپله‌اش بازی کنه» تبدیل شده‌است و «رو» که صورت سالم آن «را» است از جمله حذف شده‌است؛

۸) درج واژه جدید، مانند «می‌خواستم بروم داخل» که تبدیل شده‌است به «می‌خواستم بروم به داخل» که حرف اضافهٔ «به» به جمله اضافه شده‌است؛

۹) درج اشتباه فاصله، مانند «اون که» که تبدیل شده‌است به «آن که» در شرایطی که صورت سالم «آنکه» است؛

۱۰) عدم درج فاصله مناسب، مانند «باهم» که صورت سالم آن «با هم» است؛

۱۱) جایگزینی واژه غلط مانند «تو» در مثال «تو خونه» که تبدیل شده‌است به «در خانه» درصورتی‌که صورت سالم همان «تو» است.

درجدول ۳ نتایج ارزیابی حاصل از عملکرد منفی مدل‌های زبانی بزرگ مورد نظر در پیکرهٔ انتخابی برای هر یک از معیارهای یازده‌گانهٔ معرفی‌شده از تناظر یک‌به‌یک واژه‌ها در دادهٔ ورودی دارای گفتاری‌نویسی و خروجی مدل‌ها گزارش شده‌است. براساس نتایج به‌دست‌آمده، کمترین میزان اثرگذاری منفی مدل زبانی، توسط Perplexity انجام‌شده و مدل زبانی GPT بیشترین میزان اثرگذاری منفی را داشته‌است. ازمیان تغییرات منفی، تبدیل گفتاری‌نویسی یا شکسته‌نویسی به نوشتار معیار، بیشترین میزان تغییرات منفی را به خود اختصاص داده‌است که مدل زبانی Perplexity کمترین و مدل زبانی Claude بیشترین تغییرات را انجام داده‌است. درج اشتباه همزه کمترین اثر منفی را در تبدیل داده‌ها داشته‌است. در این مقوله، مدل‌های DeepSeek و Gemini هیچگونه اثر منفی در تبدیل داده‌ها نگذاشته‌است. جایگزینی واژه بیشترین تغییر منفی است که توسط مدل زبانی GPT اعمال شده‌است. این نوع تغییر، همچون

اشکالاتی است که در شکل اولیهٔ پیکره‌های پژوهش حاضر استفاده شده‌است. جایگزینی «صبر کن» به‌جای «وایستا» در دادهٔ اصلی نمونه‌ای از این نوع جایگزینی است که امتیاز منفی به آن داده شده‌است. دلیل تخصیص امتیاز منفی این است که نگارش «وایستا» در چارچوب دستور خط مصوب قابل پذیرش است؛ درحالی‌که «صبر کن» صورت رسمی برای این واژه است و جایگزینی این واژه سبب تغییر سبک می‌شود. تغییر در چینش واژه، حذف واژه و حذف واژه‌بست اثرگذاری‌های دیگری از مدل‌های زبانی است که این تغییرات در مدل زبانی GPT بسیار بیشتر از سایر مدل‌های زبانی مطالعه‌شده می‌باشد.

جدول ۳: تغییرات منفی مدل‌های زبانی بزرگ در داده‌های نمونه در سطح واژه

مقوله	GPT	Gemini	Perplexity	Claude	DeepSeek
تغییر چینش واژه در جمله	453	16	31	5	8
جایگزینی واژه	1674	349	439	451	606
حذف واژه‌بست	85	8	12	27	16
عدم تبدیل گفتاری‌نویسی یا شکسته‌نویسی به نوشتار معیار	1610	1651	1112	2182	1957
درج اشتباه همزه	2	0	1	1	0
درج اشتباه نیم‌فاصله	12	11	7	37	25
حذف واژه	186	15	41	7	30
درج واژه جدید	10	2	2	2	3
درج اشتباه فاصله	6	15	14	11	6
عدم درج فاصله مناسب	23	20	20	32	24
جایگزینی واژه غلط	13	2	7	6	9
جمع کل	4074	2089	1686	2761	2684

مثال‌های ۱۰ تا ۲۰ بیانگر عملکرد مدل‌های مختلف زبانی بزرگ مورد مطالعه در تغییرات منفی یازده‌گانه است. همان‌طور که پیشتر اشاره شده بود، در عملکرد ابزارها و خروجی ارائه‌شده توسط آنها، الزاماً تمامی ابزارها ویژگی مطرح‌شده را ندارد. بنابراین، در بعضی از ابزارها ممکن است تغییر اعمال شده باشد و در برخی خیر.

۱۰) تغییر چینش واژه در جمله

۱	دادهٔ ورودی	تمومش کنید این بازی کثیف و
۲	دادهٔ طلایی معیار	تمامش کنید این بازی کثیف را
۳	خروجی GPT	این بازی کثیف را تمام کنید
۴	خروجی Gemini	تمامش کنید این بازی کثیف را
۵	خروجی Perplexity	تمومش کنید این بازی کثیف را
۶	خروجی Claude	تمامش کنید این بازی کثیف را
۷	خروجی DeepSeek	تمامش کنید این بازی کثیف را

۱۱) جایگزینی واژه

۱	دادهٔ ورودی	یادش بخیر یهو ساعت ۱۱ برفک میشد شبکه یک و دو بعدا هم سه اضافه شد
۲	دادهٔ طلایی معیار	یادش بخیر یهو ساعت ۱۱ برفک می‌شد شبکه یک و دو بعداً هم سه اضافه شد
۳	خروجی GPT	یادش بخیر؛ ناگهان ساعت ۱۱، برفک می‌شد؛ شبکه یک و دو، بعداً نیز سه اضافه شد.
۴	خروجی Gemini	یادش بخیر یهو ساعت ۱۱ برفک می‌شد شبکه یک و دو بعدا هم سه اضافه شد
۵	خروجی Perplexity	یادش بخیر! یهو ساعت ۱۱ برفک میشد شبکه یک و دو. بعدها هم سه اضافه شد.
۶	خروجی Claude	یادش به‌خیر یک‌هو ساعت ۱۱ برفک می‌شد شبکه یک و دو، بعداً هم سه اضافه شد.
۷	خروجی DeepSeek	یادش بخیر یکو ساعت ۱۱ برفک می‌شد شبکه یک و دو بعداً هم سه اضافه شد

۱۲) حذف واژه‌بست

۱	دادهٔ ورودی	یه درخت خوشمزه سر راهمه دستمم به کلی از شاخه هاش می رسه
۲	دادهٔ طلایی معیار	یک درخت خوشمزه سر راهم است دستم هم به کلی از شاخه‌هایش می‌رسد
۳	خروجی GPT	یک درخت خوشمزه سر راهم است که دست‌هایم به کلی از شاخه‌های آن می‌رسد
۴	خروجی Gemini	یه درخت خوشمزه سر راهم است دستم هم به کلی از شاخه‌هایش می‌رسد
۵	خروجی Perplexity	یه درخت خوشمزه سر راهمان است و دستم به کلی از شاخه‌هایش می‌رسد
۶	خروجی Claude	یک درخت خوشمزه سر راهم است دستم هم به کلی از شاخه‌هایش می‌رسد
۷	خروجی DeepSeek	یک درخت خوشمزه سر راه است دستم هم به کلی از شاخه‌هایش می‌رسد

۱۳) عدم تبدیل گفتاری‌نویسی یا شکسته‌نویسی به نوشتار معیار

۱	دادهٔ ورودی	میشه بدونی میشه ندونی میشه اون که من میخوام و بدونی
۲	دادهٔ طلایی معیار	می‌شود بدانی می‌شود ندانی می‌شود آنکه من می‌خواهم را بدانی
۳	خروجی GPT	می‌شود بدانی می‌شود یا نمی‌شود آنکه من می‌خواهم را بدانی
۴	خروجی Gemini	میشه بدونی میشه ندونی میشه اون که من میخوام و بدونی
۵	خروجی Perplexity	میشه بدونی میشه ندونی میشه اون که من می خوام و بدونی
۶	خروجی Claude	می‌شود بدانی می‌شود ندانی می‌شود آن که من می‌خواهم را بدانی
۷	خروجی DeepSeek	می‌شود بدانی می‌شود ندانی می‌شود آن که من می‌خواهم را بدانی

۱۴) درج اشتباه همزه

۱	دادهٔ ورودی	اشکالی نداره ما مشکلی با پوشوندن مجسمه های برهنه نداریم
۲	دادهٔ طلایی معیار	اشکالی ندارد ما مشکلی با پوشاندن مجسمه‌های برهنه نداریم
۳	خروجی GPT	اشکالی ندارد ما مشکلی با پوشاندن مجسمهٔ‌های برهنه نداریم
۴	خروجی Gemini	اشکالی ندارد ما مشکلی با پوشاندن مجسمه‌های برهنه نداریم
۵	خروجی Perplexity	اشکالی ندارد ما مشکلی با پوشاندن مجسمه‌های برهنه نداریم
۶	خروجی Claude	اشکالی ندارد ما مشکلی با پوشاندن مجسمه‌های برهنه نداریم
۷	خروجی DeepSeek	اشکالی ندارد ما مشکلی با پوشاندن مجسمه‌های برهنه نداریم

۱۵) درج اشتباه نیم‌فاصله

۱	دادهٔ ورودی	بزرگترین مشکل بیکاری اینه که تعطیلی نداره
۲	دادهٔ طلایی معیار	بزرگترین مشکل بیکاری این است که تعطیلی ندارد
۳	خروجی GPT	بزرگ‌ترین مشکل بیکاری این است که تعطیلی ندارد.
۴	خروجی Gemini	بزرگترین مشکل بیکاری این است که تعطیلی ندارد
۵	خروجی Perplexity	بزرگ ترین مشکل بیکاری این است که تعطیلی ندارد.
۶	خروجی Claude	بزرگترین مشکل بیکاری این است که تعطیلی ندارد.
۷	خروجی DeepSeek	بزرگترین مشکل بیکاری این است که تعطیلی ندارد

۱۶) حذف واژه

۱	دادهٔ ورودی	علتشون چی میتونه باشه
۲	دادهٔ طلایی معیار	علتشان چه می‌تواند باشد
۳	خروجی GPT	علتشان چه می‌تواند باشد
۴	خروجی Gemini	علتش چیست
۵	خروجی Perplexity	علتشان چه می‌تواند باشد
۶	خروجی Claude	علتشان چه می‌تواند باشد
۷	خروجی DeepSeek	علتشان چه می‌تواند باشد

۱۷) درج واژه جدید

۱	دادهٔ ورودی	شبتون پر کابوس
۲	دادهٔ طلایی معیار	شبتان پر کابوس
۳	خروجی GPT	شبتان پر از کابوس باشد
۴	خروجی Gemini	شبتان پر کابوس
۵	خروجی Perplexity	شبتان پر کابوس
۶	خروجی Claude	شبتان پر کابوس
۷	خروجی DeepSeek	شب‌تان پر کابوس

۱۸) درج اشتباه فاصله

۱	دادهٔ ورودی	کاش میشد خوابم مث بازیهای کامپیوتری چن نفره میشد مثلا همزمان و همراه هم یه خواب مشابه میدیدی و با هم همکاری میکردید توش تجربه جالبی بود
۲	دادهٔ طلایی معیار	کاش می‌شد خوابم مثل بازی‌های کامپیوتری چندنفره می‌شد مثلاً همزمان و همراه هم یک خواب مشابه می‌دیدید و با هم همکاری می‌کردید تویش تجربه جالبی بود
۳	خروجی GPT	کاش می‌شد خوابم مانند بازی‌های کامپیوتری چندنفره می‌شد؛ مثلاً همزمان و همراه هم یک خواب مشابه می‌دیدید و با هم همکاری می‌کردید؛ در آن تجربه جالبی بود.
۴	خروجی Gemini	کاش می‌شد خوابم مثل بازی‌های کامپیوتری چند نفره می‌شد مثلا همزمان و همراه هم یه خواب مشابه می‌دیدی و با هم همکاری می‌کردید توش تجربه جالبی بود
۵	خروجی Perplexity	کاش میشد خوابم مث بازیهای کامپیوتری چن نفره میشد مثلا همزمان و همراه هم یه خواب مشابه میدیدی و با هم همکاری می کردید توش تجربه جالبی بود.
۶	خروجی Claude	کاش می‌شد خوابم مثل بازی‌های کامپیوتری چند نفره می‌شد مثلاً همزمان و همراه هم یک خواب مشابه می‌دیدی و با هم همکاری می‌کردید در آن تجربه جالبی بود
۷	خروجی DeepSeek	کاش می‌شد خوابم مثل بازی‌های کامپیوتری چندنفره می‌شد مثلاً همزمان و همراه هم یک خواب مشابه می‌دیدید و با هم همکاری می‌کردید تویش تجربه‌ی جالبی بود

۱۹) عدم درج فاصله مناسب

۱	دادهٔ ورودی	من می دونم بزرگترین حسرت زندگیم چیه کنسرت هم نوا با بم شجریانعلیزادهکلهر
۲	دادهٔ طلایی معیار	من می‌دانم بزرگترین حسرت زندگی‌ام چیست کنسرت همنوا با بم شجریان علیزاده کلهر
۳	خروجی GPT	من می‌دانم بزرگترین حسرت زندگی‌ام چیست کنسرت همنوا با بم شجریانعلیزادهکلهر
۴	خروجی Gemini	من می‌دونم بزرگترین حسرت زندگیم چیه کنسرت هم نوا با بم شجریانعلیزادهکلهر
۵	خروجی Perplexity	من می‌دونم بزرگترین حسرت زندگیم چیه کنسرت هم نوا با بم شجریان علیزاده کلهر
۶	خروجی Claude	من می‌دانم بزرگترین حسرت زندگی‌ام چیست کنسرت همنوا با بم شجریانعلیزادهکلهر
۷	خروجی DeepSeek	من می‌دانم بزرگترین حسرت زندگی‌ام چیست کنسرت همنوا با بم شجریانعلیزادهکلهر

۲۰) جایگزینی واژه غلط

۱	دادهٔ ورودی	احساس میکنم مث سگ پیشمون میشم ک داوطلب شدم برای یه بیمارستان دیگ ک توی طرحه
۲	دادهٔ طلایی معیار	احساس می‌کنم مثل سگ پیشمان می‌شوم که داوطلب شده‌ام برای یک بیمارستان دیگر که توی طرح است
۳	خروجی GPT	احساس می‌کنم مانند سگی پیش ما می‌شوم که داوطلب شده‌ام برای یک بیمارستان دیگر که در طرح است.
۴	خروجی Gemini	احساس می‌کنم مثل سگ پشیمان می‌شوم که داوطلب شدم برای یه بیمارستان دیگر که توی طرح است
۵	خروجی Perplexity	احساس می‌کنم مثل سگ پیششان می‌شوم، چون داوطلب شدم برای یک بیمارستان دیگر که در طرح است.
۶	خروجی Claude	احساس می‌کنم مثل سگ پشیمان می‌شوم که داوطلب شدم برای یک بیمارستان دیگر که در طرح است.
۷	خروجی DeepSeek	احساس می‌کنم مثل سگ پیش‌مان می‌شوم که داوطلب شدم برای یک بیمارستان دیگر که در طرح است.

برای آنکه بتوان به نگاه جامع‌تری از عملکرد مدل‌های زبانی بزرگ مطالعه‌شده برسیم، در جدول ۴ توزیع نسبی تغییرات مثبت و منفی مدل‌های زبانی در مجموعه دادهٔ منتخب در مقایسه با حجم تغییرات انجام‌شده مقایسه شده‌است. بیشترین میزان تغییرات اعم از مثبت یا منفی در داده‌ها، توسط مدل زبانی GPT و کمترین تغییر توسط مدل زبانی Perplexity اعمال شده‌است. در مقایسهٔ کلی عملکرد مدل‌ها، توزیع نسبی تغییرات منفی واژه‌ها در مقایسه با تغییرات مثبت به‌نسبت واژگان پیکره بیشتر بوده‌است که جزئیات این تغییرات در جداول ۲ و ۳ توضیح داده شده‌است. از مقایسهٔ تغییرات مثبت و منفی اعمال‌شده توسط مدل‌های زبانی این نتیجه به‌دست آمد که مدل زبانی Gemini کمترین تفاوت عملکرد در تغییرات را داشته‌است که در حدود ۶ درصد است. این تفاوت در مدل‌های Perplexity، Claudeو DeepSeek درحدود ۸ درصد است و این تفاوت در GPT بیش‌از ۲۰ درصد افزایش یافته‌است.

جدول 4: توزیع نسبی تغییرات مثبت و منفی مدل‌های زبانی بزرگ در داده‌های نمونه در سطح واژه

	GPT	Gemini	Perplexity	Claude	DeepSeek
تعداد واژه‌های تغییریافته	5550	3461	2459	4599	4408
توزیع نسبی تغییر واژه‌ها به نسبت واژگان پیکره	46/49	28/99	20/60	38/52	36/92
توزیع نسبی تغییر مثبت واژه‌ها به نسبت واژگان پیکره	12/36	11/49	6/48	15/40	14/44
توزیع نسبی تغییر منفی واژه‌ها به نسبت واژگان پیکره	34/13	17/50	14/12	23/13	22/48

بررسی‌های انجام‌شده که در بالا توضیح داده شد در سطح واژه بود. در ادامه به بررسی مقایسه‌ای عملکرد مدل‌های زبانی بزرگ مورد نظر با یکدیگر و همچنین با داده‌های طلایی معیار از نظر تطابق کامل در سطح جمله برای دادهٔ ورودی و خروجی مدل‌ها می‌پردازیم. نتیجهٔ حاصل از این مقایسه در جداول ۵ و 6 گزارش شده‌است. براساس نتایج گزارش‌شده در جدول ۵، مدل زبانی Claude بیشترین میزان تطابق (39/51 درصد) و مدل زبانی Perplexity کمترین میزان تطابق (7/90 درصد) را با دادهٔ طلایی معیار در سطح جمله داشته‌است. سایر مدل‌های زبانی، به‌ترتیب DeepSeek، Gemini و GPT، رتبه‌های دوم تا چهارم مدل‌های زبانی در تبدیل داده‌های دارای گفتاری‌نویسی یا شکسته‌نویسی به نوشتار معیار در سطح جمله را به خود اختصاص داده‌است. در بررسی مدل‌های زبانی بزرگ با یکدیگر، بدون درنظرگرفتن داده‌های طلایی معیار، این نتیجه به‌دست آمد که مدل زبانی Claude و DeepSeek از نظر

عملکرد به یکدیگر تشابه بیشتر دارد و دو مدل زبانی GPT و Perplexity کمترین میزان عملکرد را در مقایسه با هم دارد.

جدول ۵: مقایسه عملکرد مدل‌های زبانی بزرگ با داده معیار طلایی در سطح جمله

	طلایی معیار	GPT	Gemini	Perplexity	Claude	DeepSeek
GPT	۹۸					
Gemini	۲۴۴	۸۴				
Perplexity	۸۱	۴۷	۱۸۳			
Claude	۴۰۵	۱۳۶	۲۹۳	۹۷		
DeepSeek	۳۳۷	۱۱۸	۲۳۱	۱۰۸	۳۶۸	

براساس داده‌های حاصل از مقایسه عملکرد مدل‌های زبانی بزرگ با داده معیار طلایی در سطح جمله که در جدول 5 گزارش شد، به تحلیل دیگری از منظر طول متوسط این دسته از جملات پرداختیم که نتایج مربوط به مدل‌ها در مقایسه با دادهٔ طلایی معیار در جدول ۶ گزارش شده‌است. طول متوسط داده‌های طلایی معیار 11/92 واژه است که اگر با خروجی مدل‌ها مقایسه شود، مشخص می‌گردد که مدل‌های Claude، DeepSeek و Gemini موفق شده‌است با تفاوت دو الی سه واژه در مقایسه با دادهٔ معیار، جملات و عبارات گونه غیرمعیار را به معیار تبدیل کند. درحالی‌که در مدل‌های زبانی GPT و Perplexity با عملکردی نزدیک به یکدیگر، تفاوت جملات تغییریافته در مقایسه با داده‌های معیار را به چهار الی پنج واژه افزایش داده‌است که بیانگر تفاوت فاحش این دو مدل با سایر مدل‌ها است.

جدول ۶: مقایسهٔ طول متوسط جملات صحیح مدل‌های زبانی بزرگ در مقایسه با داده معیار طلایی

مدل	طول متوسط جملات
GPT	7/54
Gemini	9/11
Perplexity	7/42
Claude	9/72
DeepSeek	9/70

در ادامهٔ بررسی عملکرد مدل‌های زبانی بزرگ، از منظر نحو نیز به واژه‌ها توجه شده‌است به این صورت که ابتدا جملات مشترک حاصل از مدل‌های زبانی و داده طلایی معیار در سطح مقوله دستوری تحلیل شده و برچسب نحوی واژه‌ها به آنها تخصیص یافته‌است و سپس واژه‌های محتوایی و دستوری در داده‌های تبدیل‌شده از یکدیگر تفکیک شده و توزیع نسبی عملکرد مدل‌های زبانی نسبت‌به این دو دسته واژه در مقایسه با حجم کل پیکره بررسی شده‌است. داده طلایی معیار دربرگیرنده 76/50 درصد واژه‌های محتوایی و 23/50 درصد واژه‌های دستوری است.

نتایج تحلیل انجام‌شده در حوزه نحو در جدول 7 گزارش شده‌است. براساس نتایج به‌دست آمده، مدل زبانی Claude و Perplexity توانسته‌است به‌ترتیب بیشترین و کمترین حجم واژه‌های تغییریافته محتوایی و دستوری را در تبدیل داده غیرمعیار به معیار به‌دست آورد. نکته جالب توجه این است که نسبت تغییر واژه‌های محتوایی در مدل GPT (70/29 درصد) بیشتر از سایر مدل‌ها است که در مدل Claude این نسبت به 67/73 درصد کاهش یافته‌است. نسبت تغییر واژه‌های دستوری در مقایسهٔ میان مدل زبانی GPT و Claude برعکس است که به ترتیب نتایج 29/71 و 32/27 درصد در این دو مدل زبانی به‌دست آمد و Claude عملکرد بهتری در تبدیل واژه‌های دستوری غیرمعیار به معیار داشته‌است.

جدول ۷: توزیع واژه‌های محتوایی و دستوری در مدل‌های زبانی بزرگ در مقایسه با دادهٔ طلایی معیار

مدل	واژه محتوایی		واژه دستوری	
	بسامد مطلق	بسامد نسبی	بسامد مطلق	بسامد نسبی
GPT	579	4/85	172	1/44
Gemini	1711	14/33	528	4/42
Perplexity	478	4/00	157	1/32
Claude	3012	25/23	972	8/14
DeepSeek	2525	21/15	817	6/84

در تحلیلی دیگر، ۵۰ تا ۴۰۰۰ واژهٔ مشترک جملات مدل‌های زبانی بزرگ در مقایسه با دادهٔ طلایی معیار مقایسه شده‌است که نتایج حاصل از این بررسی در جدول ۸ گزارش شده‌است. نکتهٔ قابل توجه این است که مدل‌های زبانی Claude و DeepSeek به ترتیب

بیشترین حجم واژه‌های مشترک با داده طلایی معیار را داشته‌است که این تشابه در ۵۰ واژهٔ مشترک اول مانند یکدیگر است ولی در تعداد بالاتر، با اختلاف کمی، دو مدل عملکرد متفاوتی از خود نشان می‌دهد. تشابه واژه‌های پربسامد در مدل Gemini به دو مدل معرفی‌شده نزدیک است. اما مدل زبانی GPT در همین تعدادِ کمِ واژه‌های مشترک، عملکرد متفاوتی با سه مدل معرفی‌شده دارد. نکتهٔ دیگر قابل توجه این است که تعداد واژه‌های مشترک در مدل‌های مختلف تا تعداد ۴۰۱۶ واژهٔ یکتا در دادهٔ طلایی معیار فاصله دارد. این فاصله، بیانگر عدم توانایی مدل زبانی در تشخیص واژه است که در این بررسی مدل‌های زبانی Claude و DeepSeek همچنان پرچمدار است.

جدول 8: مقایسه واژه‌های پربسامد جملات مدل‌های زبانی بزرگ در مقایسه با داده طلایی معیار

	۵۰	۱۰۰	۲۰۰	۳۰۰	۴۰۰	۵۰۰	۱۰۰۰	۲۰۰۰	۳۰۰۰	۴۰۰۰
GPT	۴۱	۸۲	۱۶۸	۲۴۷	۳۲۳	۴۱۶	۸۲۸	۱۵۸۰	2315	3035
Gemini	۴۶	۸۸	۱۷۸	۲۵۳	۳۳۱	۴۱۰	۸۴۴	۱۶۲۰	2425	3206
Perplexity	۳۹	۸۰	۱۶۰	۲۳۵	۲۹۹	۳۷۵	۷۵۳	۱۳۷۳	2101	2799
Claude	۴۸	۹۴	۱۸۹	۲۷۴	۳۶۳	۴۵۹	۹۲۴	۱۷۹۸	2668	3521
DeepSeek	۴۸	۹۱	۱۸۷	۲۶۹	۳۵۳	۴۴۵	۹۱۳	۱۷۴۶	۲۵۹۱	۳۴۴۲

۶. نتیجه‌گیری

در این پژوهش به بررسی مقایسه‌ای عملکرد پنج مدل زبانی بزرگ شامل ChatGPT، Gemini، Perplexity ، Claude و DeepSeek در فعالیت تبدیل داده‌های دارای گفتاری‌نویسی یا شکسته‌نویسی به دادهٔ معیار براساس دستور خط مصوب فرهنگستان زبان و ادب فارسی پرداختیم. در فرایند بررسی، اثرگذاری مدل‌ها در تبدیل داده را به دو دستهٔ مثبت و منفی تقسیم کردیم. نتیجهٔ کلی به‌دست‌آمده حاکی از آن است که اثرگذاری منفی مدل‌ها بیشتر از اثرگذاری آنها در تبدیل شکل غیرمعیار به معیار است. نتیجهٔ مهم دیگری که از مقایسهٔ عملکرد مدل‌ها به‌دست آمد این بود که مدل زبانی Claude بیشترین و مدل زبانی Perplexity کمترین میزان کارایی را در سطح واژه و جمله به‌دست آورده‌است و GPT بیشترین حجم اختلاف میان اثرگذاری مثبت و منفی را به‌دست آورده‌است.

نتیجهٔ کلی که می‌توان از پژوهش حاضر گرفت این است که تفاوت نتایج به‌دست‌آمده از مدل‌های زبانی بزرگ مختلف بیانگر این نکته است که به‌دلیل تفاوت در الگوریتم و داده‌های آموزش مدل‌ها، این تفاوت در عملکردشان به نمایش گذاشته می‌شود. بنابراین نمی‌توان به تنهایی یک مدل زبانی بزرگ را انتخاب کرده و پژوهش را به آن مدل محدود نمود. در یک بررسی جامع نیاز است مدل‌های زبانی بزرگ متنوع بررسی گردد. بررسی تفاوت‌ها می‌تواند راه‌گشای انتخاب ابزار صحیح در پژوهش‌های آتی باشد.

منابع

آرمین، نادیه و شمس‌فرد، مهرنوش. (۱۳۸۹)، تبدیل متن محاوره‌ای فارسی به رسمی به کمک n-gramها. *مجموعه مقالات شانزدهمین کنفرانس ملی سالانه انجمن کامپیوتر ایران*. https://www.researchgate.net/publication/314239155_tbdyl_mtn_mhawrh_ay_farsy_bh_kmk_N_gram_ha

ادیبیان، مجید، و ممتازی، سعیده. (۱۴۰۱). تبدیل متن محاوره به رسمی فارسی با استفاده از شبکه‌های عصبی مبتنی بر مبدل. زبان *و زبان‌شناسی*. ۱۸ (۳۵)، ۶۷-۴۵.

بی‌جن‌خان، محمود. (1383) نقش پیکره زبانی در نوشتن دستور زبان: معرفی یک نرم‌افزار رایانه‌ای. زبان*شناسی*. ۱۹(۳۷): ۴۸-۶۷.

فرهنگستان زبان و ادب فارسی. (۱۳۸۹). *دستور خط مصوب فرهنگستان*. تهران: فرهنگستان زبان و ادب فارسی.

فرهنگستان زبان و ادب فارسی. (۱۴۰۲). *دستور خط مصوب فرهنگستان*. تهران: فرهنگستان زبان و ادب فارسی.

صادقی، علی‌اشرف و زندی‌مقدم، زهرا. (۱۳۹۴). *فرهنگ املایی خطّ فارسی مصوب فرهنگستان زبان و ادب فارسی*. تهران: فرهنگستان زبان و ادب فارسی (نشر آثار).

طبیب‌زاده، امید. (۱۳۹۸). *مبانی و دستور خط فارسی شکسته براساس صد سال آثار داستانی و نمایشی*. تهران: پژوهشگاه علوم انسانی و مطالعات فرهنگی.

قیومی، مسعود و مسگرخویی، مریم. (۱۴۰۲). تحلیل بر پیکره حاصل از داده‌های زبان فارسی در فضای مجازی. زبان *و زبان‌شناسی*. ۱۹(۳۷)، ۱۱۷–۱۳۶.

AbdiKhojasteh, H., Ansari, E., and Bohlouli, M. (2020). LSCP: Enhanced large scale colloquial Persian language understanding. In *Proceedings of the 12th International Conference on Language Resources and Evaluation*. European Language Resources Association, 6323–6327.

Ariffin, S. N. A. N. and Tiun, S. (2020). Rule-based text normalization for Malay social media texts. *International Journal of Advanced Computer Science and Applications*. 11(10): 156-162.

Arora, M. and Kansal, V. (2019). Character level embedding with deep convolutional neural network for text normalization of unstructured data for Twitter sentiment analysis. *Social Network Analysis and Mining*, 9, 12. https://doi.org/10.1007/s13278-019-0557-y

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Minnesota:Association for Computational Linguistics, 4171−4186.

Ferguson, C. A. (1959). Diglossia. *WORD*. 15(2), 325-340

Kashefi, O. (2018). MIZAN: A large Persian-English parallel corpus. *Computing Research Repository*. https://dblp.org/rec/bib/journals/corr/abs-1801-02107.

Khazeni, M., Heydari, M., & Albadvi, A. (2024). Persian slang text conversion to formal and deep learning of Persian short texts on social media for sentiment classification. *Journal of Electrical and Computer Engineering Innovations*, 13(1): 27–42.

Kozhirbayev, Z. and Yessenbayev, Z. (2020). Kazakh text normalization using machine translation approaches. *CEUR Workshop Proceedings*. 2780.

Liu, Y., Ji, B., Yu, J., Tan, Y., Ma, J., and Wu, Q. (2021). An advanced ICD-9 terminology standardization method based on BERT and text similarity. In *Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery. ICNC-FSKD 2020. Lecture Notes on Data Engineering and Communications Technologies*. eds. Meng, H., Lei, T., Li, M., Li, K., Xiong, N., and Wang, L., Springer, vol. 88, 1868–1879.

Mansfield, C., Sun, M., Liu, Y., Gandhe, A., and Hoffmeister, B. (2019). Neural Text Normalization with Subword Unit. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. eds. Loukina, A., Morales, M., and Kumar, R., Association for Computational Linguistics, vol. 2, 190–196.

Masoumi, V., Salehi, M., Veisi, H., Haddadian, G., Ranjbar, V., and Sahebdel, M. (2020). TeleCrowd: A crowdsourcing approach to create informal to formal text corpora. https://arxiv.org/abs/2004.11771.

Naemi, A., Mansourvar, M., Naemi, M., Damirchilu, B., Ebrahimi, A., and Wiil, U. K. (2021). Informal-to-formal word conversion for Persian language using natural language processing techniques. In *Proceedings of the 2nd International Conference on Computing, Networks and Internet of Things*. New York, NY, USA.

Rasooli, M. S., Bakhtyari, F., Shafiei, F., Ravanbakhsh, M., and Callison-Burch, C. (2020). Automatic standardization of colloquial Persian . https://arxiv.org/abs/2012.05879.

Tajalli, V., Kalantari, F., and Shamsfard, M. (2023). Developing an informal formal Persian corpus. https://arxiv.org/abs/2308.05336.

استناد به این مقاله: قیومی، مسعود، محمدی، فاطمه، جهان، علی. (۱۴۰۴). تحلیلی بر عملکرد مدل‌های زبانی بزرگ در تبدیل خودکار گفتاری‌نویسی به نوشتار معیار براساس دستور خط مصوب فارسی. *زبان و زبان‌شناسی*، ۲۱(۴۲)، ۹۳- ۱۳۲.
doi: 10.30465/lsi.2026.52503.1820