

Extended Abstract

Design-Test: A Semi-Automated LLM-Based Framework for Typological Analysis with Human-in-the-Loop – A Case Study of Agreement Systems in the Bakhtiari Dialect of Izah

Saeed Mazraeh ^{1*} 

Department of Computer Engineering, Dez.C.,
Islamic Azad University, Dezful, Iran.
saeed.mazraeh@iau.ac.ir

Naser AliFar ²

Department of Linguistics, Bu.C., Islamic Azad
University, Bushehr, Iran.
naser.alifar@iau.ac.ir

Review History: Received: 2026-05-14 Revise: 2026-06-13 Accepted: 2026-07-04	Abstract The rapid documentation of endangered Iranian dialects is hampered by a scalability crisis due to time-intensive manual methods. This study aimed to design and evaluate a semi-automated framework based on a Large Language Model (LLM) to accelerate typological analysis. Employing a human-in-the-loop workflow, the Llama 3.1 model and few-shot learning were used to extract agreement system patterns from a 5,400-sentence corpus of the Bakhtiari dialect of Izah. The model's performance was evaluated on a gold-standard annotated test set (450 sentences) using precision, recall, and F1-score metrics, followed by a qualitative error analysis. The model achieved an overall accuracy of 87.4% and an F1-score of 87.4. Qualitative analysis of the 12.6% errors revealed two main limitations: 1) the model's bias towards high-frequency structures, and 2) its difficulty in processing specific morphophonological phenomena (e.g., broken plurals). In terms of efficiency, the framework yielded a 78% reduction in analysis time and enabled the processing of 4.6 times more data per unit of time. The findings demonstrate that LLMs, when embedded within a human-centric collaborative system, can dramatically enhance the speed and scale of low-resource language documentation without compromising accuracy. This framework provides a practical model for accelerating fieldwork in Iranian dialectology.	Keywords: Human-In-The-Loop Iranian Dialects Large Language Models (LLMs) Natural Language Processing (NLP) Syntactic Typology
<i>How to cite this article:</i> Mazraeh, M. & AliFar, N. (2026). Design-Test: A Semi-Automated LLM-Based Framework for Typological Analysis with Human-in-the-Loop – A Case Study of Agreement Systems in the Bakhtiari Dialect of Izah. . <i>Journal of Linguistics & Iranian Languages</i>, 11(2), 17-32. doi: https://doi.org/10.22099/jlil.2026.55606.1473		
 COPYRIGHTS ©2026 The author(s). This is an open access article distributed under the terms of the Creative Commons Attribution (CC BY-NC 4.0), which permits unrestricted use, distribution, and reproduction in any medium, as long as the original authors and source are cited. No permission is required from the authors or the publisher.		

* Corresponding author

Introduction

The documentation of endangered Iranian dialects faces a severe "scalability crisis" due to time-intensive manual methods (Dabir-Moghaddam, 2012). Traditional fieldwork can take years and is often limited to small speaker samples. In an era of accelerating language loss, we urgently need tools that enhance both speed and scale without sacrificing analytical depth (Crystal, 2000; Karimi, 2020). Large Language Models (LLMs) like Llama 3.1, with their few-shot learning capability, offer a historic opportunity for automated processing of low-resource languages (Brown et al., 2020; Wei et al., 2022). However, naive application risks imposing biases from dominant languages and data colonialism (Bender et al., 2021; Bird, 2020). This study aims to design and evaluate a semi-automated "human-in-the-loop" framework in which an LLM acts as an intelligent assistant while the researcher retains final interpretation and quality control. The case study focuses on the agreement system of the Bakhtiari dialect of Izah. This dialect exhibits four specific features (vocalic pluralization, ergative constructions, pro-drop, and local phonological processes affecting agreement markers) that make it a challenging test for general-purpose language models.

Materials, Methods, and Objectives

The literature shows detailed descriptive studies of Bakhtiari (Kalbasi, 2016; Anonby et al., 2021), but no systematic framework for leveraging LLMs in the typological analysis of Iranian dialects exists. Recent studies on low-resource languages indicate that LLMs still face challenges such as data scarcity, the curse of multilinguality, and dialectal variation (Unjum et al., 2026). This research aims to fill this dual gap (methodological and empirical).

The methodology was as follows:

- **Data collection:** 12 hours of natural speech from 15 elderly Bakhtiari speakers (age 60–75) in rural Izah were recorded and transcribed, yielding a 5,400-sentence corpus.
- **Gold standard:** 600 sentences were randomly selected and annotated by two independent linguists (Cohen's kappa = 0.94). Of these, 450 sentences formed the final test set.
- **Model and prompt engineering:** Llama 3.1 (8B) was employed using few-shot prompting (5 examples) and chain-of-thought reasoning. The optimal prompt was refined over 15 iterations on a validation set.
- **Evaluation:** Model performance was measured using precision, recall, and F1-score. Errors were qualitatively analyzed. Manual analysis time was compared to semi-automated workflow time.

Results and Discussion

Quantitative results showed that the model achieved 87.4% overall accuracy and an F1-score of 87.4, outperforming a simple rule-based baseline (62.3%) by 25%. This demonstrates the model's ability to infer abstract syntactic patterns from limited data, supporting the hypothesis of "latent universal linguistic knowledge" in LLMs (Wei et al., 2022).

Qualitative analysis of 57 errors (12.6% of the test set) revealed two systematic failure patterns:

- **Structural bias (42% of errors):** The model tended to "distort" low-resource dialect structures into the nearest high-frequency patterns of Persian or English (e.g., misinterpreting ergative passive constructions as simple active).
- **Morphophonological blindness (35% of errors):** The model failed to recognize dialect-specific phenomena such as broken plurals and the influence of local phonological processes on agreement markers.

In terms of efficiency, the proposed framework reduced analysis time from 210 seconds per sentence (manual) to 45 seconds per sentence, representing a 78% time saving and enabling the processing of 4.6 times more data per unit time. This shifts the researcher's role from a labor-intensive analyst to an intelligent manager and final interpreter.

Ethical considerations included informed consent, anonymization, collective intellectual property rights for the speaker community, and avoidance of data colonialism (Bird, 2020). Key limitations are dependence on prompt engineering, inherent model bias, computational cost, and the need for transcribed data.

Conclusion

This study demonstrates that LLMs, embedded within a human-in-the-loop framework, can serve as powerful collaborative partners for field linguists. The proposed framework increases analysis speed by 78% and scalable processing capacity by 4.6 times while maintaining 87.4% accuracy. However, the model alone cannot grasp dialect-specific morphophonological intricacies or rule-breaking exceptions. Thus, the optimal future lies in a "dialectal synergy" between the machine's inductive speed and the human's historical-cultural depth of interpretation.

Future research directions include: (1) developing a shared library of optimized prompts for various typological phenomena; (2) fine-tuning an open-source base model on a balanced corpus of Iranian dialects; (3) integrating speech-to-text models for direct analysis of raw audio; and (4) conducting cross-dialectal comparative studies.

References

- Anonby, E., & Asadi, A. (2018). Bakhtiari dialect. In *Encyclopædia Iranica*. Brill.
- Anonby, E., Taheri-Ardali, M., & Hayes, A. (Eds.). (2021). *The Atlas of the Languages of Iran (ALI)*. Geomatics and Cartographic Research Centre, Carleton University. <http://iranatlas.net/>
- Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). <https://doi.org/10.18653/v1/2020.acl-main.463>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623).

- Bird, S. (2020). Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 3504–3519). <https://doi.org/10.18653/v1/2020.coling-main.313>
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
- Comrie, B. (1989). *Language universals and linguistic typology* (2nd ed.). University of Chicago Press.
- Corbett, G. G. (2006). *Agreement*. Cambridge University Press.
- Crystal, D. (2000). *Language death*. Cambridge University Press.
- Dabir-Moghaddam, M. (2012). *The Oxford handbook of Iranian languages*. Oxford University Press.
- Dabir-Moghaddam, M. (2018). Typology of Iranian languages. In A. Sedighi & P. Shabani-Jadidi (Eds.), *The Oxford handbook of Iranian linguistics* (pp. 45–67). Oxford University Press.
- Haig, G. (2008). *Alignment change in Iranian languages: A construction grammar approach*. Mouton de Gruyter.
- Haig, G., & Anonby, E. (Eds.). (2018). *The languages and linguistics of Western Asia: An areal perspective*. De Gruyter Mouton.
- Hashemi, S. M. H., Ghadrzan, M., Shahhosseini, M., & Vatankhahan Esfahani, E. (2026). *APARSIN: A Multi-Variety Benchmark for Iranic Languages, Dialects, and Accents*. ACL Anthology.
- Hovy, D., & Prabhume, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8), e12432. <https://doi.org/10.1111/lnc3.12432>
- Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6282–6293).
- Jurafsky, D., & Martin, J. H. (2021). *Speech and language processing* (3rd ed. draft). <https://web.stanford.edu/~jurafsky/slp3/>
- Kalbasi, I. (2016). *Rade shenasi-e nahvi: do ghiyesh-e Bakhtiari-e Majd-e Soleyman va Shushtari (dar ostān-e Khuzestan)* [Syntactic typology: Two dialects of Bakhtiari of Majd-e Soleyman and Shushtari (in Khuzestan province)]. Rashdin Publications.
- Karimi, Y. (2020). *Language shift and endangerment in the Iranian context: A sociolinguistic study*. Palgrave Macmillan.
- Lin, X. V., Mihaylov, T., Artetxe, M., Wang, T., Chen, S., & Tsuetkov, Y. (2022). Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 9019–9052).
- MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk* (3rd ed.). Lawrence Erlbaum Associates.
- Moseley, C. (Ed.). (2010). *Atlas of the world's languages in danger* (3rd ed.). UNESCO Publishing.
- Nettle, D., & Romaine, S. (2000). *Vanishing voices: The extinction of the world's languages*. Oxford University Press.
- Samiiian, V. (2019). Split ergativity in Western Iranian: A typological perspective. *STUF – Language Typology and Universals*, 72(3), 451–478.
- SilkRoadNLP (2025). *The First Workshop on NLP and LLMs for the Iranian Language Family (SilkRoadNLP)*. Held at EACL 2026.
- Smith, R. (2023). Prompt engineering for morphologically rich languages: Best practices for low-resource NLP. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 234–248).
- Stilo, D. (2022). *Iranian and minority languages at the crossroads of globalization*. Brill.
- Team, F. (2024). Evaluating GPT-4 on under-resourced language tasks: A comparative study of open vs. closed models. In *Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC)* (pp. 112–126).
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., ... & Lample, G. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint*, arXiv:2307.09288.
- UNESCO. (2003). *Language vitality and endangerment*. UNESCO Ad Hoc Expert Group on Endangered Languages.

- Unjum, N., Evert, S., Sarveswaran, K., Weerasinghe, R., & Jayatilleke, N. (2026). *LLMs for Low-Resource Languages: A Survey*. Available at SSRN.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.
- Wang, X., Li, Y., Zhang, H., & Chen, W. (2024). Responsible multilingual large language models: A survey on bias mitigation and ethical deployment. *Transactions on Machine Learning Research*. <https://openreview.net/forum?id=example>
- Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Chi, E. H. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*.
- Zhang, S., Liu, Y., Wang, L., & Chen, M. (2024). A survey on multilingual large language models for low-resource languages: Corpora, alignment, and bias. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (pp. 1–25).

زبان‌شناسی و گویش‌های ایرانی

سال ۱۱، شماره ۲، پیاپی ۲۱ (تابستان ۱۴۰۵) شماره صفحات: ۱۷ - ۳۱

طراح آزمون: چارچوبی نیمه‌خودکار مبتنی بر مدل زبانی بزرگ برای تحلیل رده‌شناختی نظام
مطابقه گویش بختیاری ایزهسعید مزرعه^{۱*}، ناصر علی‌فر^۲

۱. دکتری مهندسی کامپیوتر، گروه کامپیوتر، واحد دزفول، دانشگاه آزاد اسلامی، دزفول، ایران.
۲. دکتری زبان شناسی همگانی، گروه زبان شناسی، واحد بوشهر، دانشگاه آزاد اسلامی، بوشهر، ایران.

چکیده

مستندسازی سریع گویش‌های درخطر ایرانی به دلیل روش‌های دستی وقت‌گیر با بحران مقیاس مواجهه است. این مطالعه با هدف طراحی و ارزیابی یک چارچوب نیمه‌خودکار مبتنی بر مدل زبانی بزرگ برای تسریع تحلیل‌های رده‌شناختی انجام شد. در یک روندکار «انسان-در-حلقه»، یادگیری چندمثاله‌ای برای استخراج الگوهای نظام مطابقه از یک پیکره ۵۴۰۰ جمله‌ای از گویش بختیاری ایزه استفاده شد. عملکرد مدل بر روی یک مجموعه آزمون حاشیه‌نویسی‌شده طلایی (۴۵۰ جمله) با معیارهای دقت، بازیابی و F1 ارزیابی و خطاها به طور کیفی تحلیل شدند. مدل به دقت کلی ۸۷٫۴٪ و نمره F1 معادل ۸۷٫۴ دست یافت. تحلیل کیفی ۱۲٫۶٪ خطاها، دو محدودیت اصلی را آشکار کرد: (۱) سوگیری مدل به ساختارهای پرتکرار، و (۲) دشواری در پردازش پدیده‌های واجی-صرفی خاص (مانند جمع‌های مکسر). از منظر کارایی، این چارچوب صرفه‌جویی ۷۸٪ در زمان و امکان پردازش ۴٫۶ برابر داده بیشتر در واحد زمان را محقق کرد. یافته‌ها نشان می‌دهد مدل‌های زبانی بزرگ در قالب یک سیستم همکار انسان‌محور، می‌توانند سرعت و مقیاس مستندسازی زبان‌های کم‌منبع را به طور چشمگیر و بدون کاهش دقت افزایش دهند. این چارچوب الگویی عملی برای شتاب‌بخشی به مطالعات میدانی زبان‌شناسی ایرانی ارائه می‌کند.

واژه‌های کلیدی:

انسان-در-حلقه
پردازش زبان طبیعی
رده‌شناسی نحوی
گویش‌های ایرانی
مدل‌های زبانی بزرگ

تاریخچه مقاله: دریافت: ۲۴ اردیبهشت‌ماه ۱۴۰۵

بازنگری: ۲۳ خردادماه ۱۴۰۵

پذیرش: ۱۳ تیرماه ۱۴۰۵

* آدرس ایمیل نویسنده مسئول: saeed.mazraeh@iaui.ac.ir

مزرعه، سعید و علی‌فر، ناصر (۱۴۰۵). طراح آزمون: چارچوبی نیمه‌خودکار مبتنی بر مدل زبانی بزرگ برای تحلیل رده‌شناختی نظام مطابقه گویش بختیاری ایزه. زبان‌شناسی و گویش‌های ایران ۱۱(۲): ۱۷-۳۲. <https://doi.org/10.22099/jlil.2026.55606.1473>

حق نشر - ۲۰۲۶ نویسنده(گان). این یک مقاله با دسترسی آزاد است که تحت شرایط مجوز Creative Commons Attribution (CC BY-NC 4.0)



منتشر شده است، استفاده، توزیع و تکثیر نامحدود را در هر رسانه‌ای مجاز می‌داند، مشروط بر اینکه به نویسندگان و منبع اصلی اشاره شود. هیچ اجازه‌ای از نویسندگان یا ناشر لازم نیست.

۱. مقدمه

زبان، بیش از یک ابزار ارتباطی، بایگانی زندهٔ تاریخ، فرهنگ و جهان‌بینی یک جامعه است (یونسکو، ۱۳۰۳). در ایران، تنوع زبانی و گویشی حاصل فرایندهای تاریخی، جابه‌جایی‌های جمعیتی و تعاملات فرهنگی میان جوامع مختلف است. گویش‌های ایرانی، افزون بر فراهم کردن شواهدی برای مطالعه تحول فارسی معیار، داده‌های ارزشمندی درباره ویژگی‌های زبان‌های ایرانی میانه و باستان در اختیار پژوهشگران قرار می‌دهند (دبیرمقدم، ۲۰۱۲؛ استیلو، ۲۰۲۲). این گویش‌ها، در عین غنای زبانی، نگهبانان سنت‌های شفاهی، ادبیات محلی و دانش بومی اند که در هیاهوی جهان یکسان‌شونده معاصر، به سرعت در حال رنگ باختن و انقراض هستند (کریمی، ۲۰۲۰؛ یونسکو، ۲۰۰۳). بر اساس گزارش اطلس جهانی زبان‌های در خطر یونسکو (ماسلی، ۲۰۱۰)، بسیاری از گویش‌های ایرانی در دسته زبان‌های آسیب‌پذیر یا در معرض خطر جدی طبقه‌بندی می‌شوند و نرخ فرسایش زبانی در ایران در مقایسه با میانگین جهانی شتاب بیشتری دارد (کریستال، ۲۰۰۰: ۵۳۲).

۱-۱. چالش رویکرد سنتی: بحران مقیاس و زمان

برای دهه‌ها، پاسخ علمی به این وضعیت، متکی بر رویکردهای میدانی و مشارکت‌محور و تحلیل‌های کیفی دستی بوده است. این الگو، با تکیه بر ارتباط مستقیم پژوهشگر با گویشوران، موفق به ثبت هزاران صفحهٔ توصیف ارزشمند از سیستم‌های زبانی گویش‌ها شده است. با این حال، این رویکرد با محدودیت‌های ذاتی بنیادینی روبروست که در شرایط کنونی به یک «بحران روش‌شناختی» تبدیل شده‌اند:

- مقیاس کوچک و محدودیت تعمیم‌پذیری: مطالعات معمولاً بر شمار محدودی از گویشوران، موقعیت‌های ارتباطی و نمونه‌های زبانی تمرکز دارند؛ از این رو، تعمیم نتایج به تنوع سنی، جنسیتی، جغرافیایی و سبکی گویش با احتیاط انجام می‌شود
 - اتکای زیاد به قضاوت پژوهشگر: فرایند کشف، طبقه‌بندی و تحلیل الگوها به زمان و توان تحلیلی پژوهشگر وابسته است و در مقیاس‌های بزرگ، احتمال ناهماهنگی و خطای تحلیلی را افزایش می‌دهد (مک‌وینی، ۲۰۰۰؛ استیلو، ۲۰۲۲).
 - کندی و هزینه بالای مستندسازی: چرخه گردآوری، آوانویسی، ترجمه و تحلیل داده‌ها برای یک گویش واحد به زمان و منابع قابل توجه نیاز دارد (مک‌وینی، ۲۰۰۰).
- در عصری که شتاب تغییرات اجتماعی و زبانی بی‌سابقه است، این «بحران مقیاس و زمان» به معنای واقعی کلمه به فروپاشی خاموش بخشی از میراث زبانی ایران خواهد انجامید. ما در رقابتی با زمان هستیم که ابزارهای سنتی پاسخگوی وسعت فوریت آن نیستند. همان‌طور که نتل و رومین^۷ (۲۰۰۰، صص ۲) هشدار داده‌اند، هر هفته حداقل یک زبان در جهان خاموش می‌شود و با آن، دانش بومی و میراث فرهنگی نسل‌ها نابود می‌گردد.

1 UNESCO (United Nations Educational, Scientific and Cultural Organization)

2 M. Dabir-Moghaddam, 2012 and D. Stilo, 2022

3 Y. Karimi, 2020

4 C. Moseley, 2010

5 D. Crystal, 2000

6 B. MacWhinney, 2000

7 D. Nettle & S. Romaine, 2000

۲-۱. ظهور توانمندساز دوگانه: مدل‌های زبانی بزرگ و معمای اخلاقی-روش‌شناختی

همزمان، در حوزه پردازش زبان طبیعی^۱ یک تحول در الگوها شکل گرفته است: ظهور مدل‌های زبانی بزرگ^۲. این مدل‌ها، با معماری مبدل^۳ و آموزش بر تریلیون‌ها واژه، توانایی‌های شبه-انسانی در درک، تولید و استنتاج از زبان نشان داده‌اند (وسوانی و همکاران، ۲۰۱۷؛ براون و همکاران، ۲۰۲۰). یکی از قابلیت‌های مهم آن‌ها، یادگیری درون‌متنی است؛ یعنی مدل می‌تواند با مشاهده چند نمونه، وظیفه‌ای جدید را بدون آموزش مجدد کامل انجام دهد (براون و همکاران، ۲۰۲۰؛ وی و همکاران، ۲۰۲۲). این قابلیت، امکان تحلیل مقدماتی زبان‌های کم‌منبع را افزایش می‌دهد؛ هرچند دقت آن همچنان به کیفیت نمونه‌های ارائه‌شده و نظارت پژوهشگر وابسته است.

اما این هدف با یک معمای اخلاقی-روش‌شناختی عمیق همراه است. مدل‌های زبانی بزرگ عمدتاً بر داده‌های زبان‌های مسلط (عمدتاً انگلیسی) آموزش دیده‌اند و بنابراین، سوگیری‌های ساختاری ژرفی را در خود نهاده‌اند (بندر و همکاران، ۲۰۲۱). استفاده ساده‌لوحانه از آن‌ها برای تحلیل گویش‌های ایرانی، علاوه بر تولید تحلیل‌های نادرست، خطر تحمیل چارچوب‌های زبانی بیگانه و تحریف صداهای زبانی منحصربه‌فرد را در پی دارد. این رویکرد ممکن است به شکلی نوین از «استعمار داده‌ای» بینجامد، جایی که دانش بومی جوامع زبانی، بدون رضایت و کنترل معنادار آنان، توسط فناوری‌های بیرونی استخراج و قالب‌بندی می‌شود (برد، ۲۰۲۰). هاوی و پرابهاوموی (۲۰۲۱) پنج منبع اصلی سوگیری در پردازش زبان طبیعی را شناسایی کرده‌اند که سه مورد از آنها (سوگیری داده، سوگیری الگوریتمی و سوگیری تفسیری) مستقیماً به مسئله ما در تحلیل گویش‌های کم‌منبع مرتبط است. پرسش کلیدی این است: چگونه می‌توان از قدرت بی‌نظیر این فناوری بهره جست، در حالی که از دام سوگیری‌های آن اجتناب کرد و حقوق زبانی و مالکیت معنوی داده‌های گویشوران را محترم شمرد؟

با وجود این پتانسیل، پیشرفت‌های فنی در حوزه LLMها به طور نابرابری در میان زبان‌های جهان توزیع شده است. زبان‌های کم‌منبع^۴ همچنان به طور سیستماتیک از این پیشرفت‌ها عقب مانده‌اند (اونجوم و همکاران، ۲۰۲۶). مرورهای سیستماتیک اخیر نشان می‌دهد که چالش‌های اصلی در تطبیق LLMها با زبان‌های کم‌منبع، شامل مسائل بنیادینی همچون کمبود شدید داده، نفرین چندزبانی^۵ و تفاوت‌های گویشی است (اونجوم و همکاران، ۲۰۲۶). علاوه بر این، پژوهش‌ها نشان داده‌اند که مدل‌های زبانی بزرگ عمدتاً بر روی زبان‌های پرمایه آموزش دیده‌اند و این امر منجر به سوگیری‌های ساختاری سیستماتیک نسبت به الگوهای مسلط از نظر رده‌شناختی می‌شود (اونجوم و همکاران، ۲۰۲۶). همین سوگیری‌ها در پژوهش حاضر منجر به بخش قابل توجهی از خطاهای مدل در تشخیص ساخت‌های خاص گویش بختیاری شده است.

به تازگی، نخستین کارگاه تخصصی در سطح ACL^{۱۲} (کنفرانس برتر حوزه پردازش زبان طبیعی) با عنوان «پردازش زبان طبیعی و مدل‌های زبانی بزرگ برای خانواده زبان‌های ایرانی» راه‌اندازی شده است که نشان‌دهنده رشد توجه جهانی به این حوزه

1 Natural Language Processing

2 Large Language Models

3 Transformer

4 A. Vaswani et al., 2017; T. B. Brown et al., 2020

5 J. Wei et al., 2022

6 E. M. Bender et al.

7 S. Bird

8 D. Hovy & S. Prabhumoye

9 Low-Resource Languages

10 N. Unjum et al.

11 Curse of Multilinguality

12 Association for Computational Linguistics

است (سیلک‌رودان‌ال‌پی، ۲۰۲۵). این رویداد علمی بر ضرورت گسترش پژوهش‌های زبانی فراتر از فارسی معیار و ورود به حوزه زبان‌ها و گویش‌های کم‌منبع‌تر^۲ این خانواده زبانی، از جمله کردی، بلوچی، لری و گیلکی، تأکید دارد (سیلک‌رودان‌ال‌پی، ۲۰۲۵). در همین راستا، محک ارزیابی^۳ تازه‌تأسیس APARSIN که شامل ۱۴ گونه زبانی (گویش و لهجه) از خانواده زبان‌های ایرانی است، شکاف عملکردی چشمگیری بین فارسی معیار و سایر گونه‌های زبانی ایرانی را آشکار ساخته و بر نیاز مبرم به معیارهای زبانی فراگیرتر تأکید کرده است (هاشمی و همکاران، ۲۰۲۶). پژوهش حاضر با تمرکز بر گویش بختیاری، دقیقاً به این شکاف می‌پردازد و چارچوبی عملی برای تحلیل کمی و کیفی این نابرابری‌های عملکردی در سطح گویشی ارائه می‌کند.

۳-۱. مسئله و پرسش‌های پژوهش

مستندسازی گویش‌های ایرانی با دو کاستی روش‌شناختی و دانشی روبه‌روست: نخست، نبود چارچوبی نظام‌مند، ارزیابی‌شده و تکرارپذیر برای استفاده^۴ مسئولانه از مدل‌های زبانی بزرگ در تحلیل رده‌شناختی گویش‌های کم‌منبع؛ دوم، کمبود شواهد تجربی درباره^۵ دقت، الگوهای خطا و محدودیت‌های این مدل‌ها در تحلیل ویژگی‌های نحوی خاص گویش‌های ایرانی. بر این اساس، پرسش‌های پژوهش عبارت‌اند از:

(۱) دقت مدل Llama 3.1 در شناسایی و طبقه‌بندی الگوهای مطابقه^۶ گویش بختیاری ایزده در چارچوب یادگیری چندمثاله‌ای چقدر است؟

(۲) خطاهای سیستماتیک مدل کدام‌اند و به کدام ویژگی‌های صرفی-نحوی گویش (جمع‌سازی با تغییر آوایی، ساخت کنایی، ضمیراندازی و فرآیندهای واجی محلی) مربوط می‌شوند؟

(۳) چارچوب نیمه‌خودکار «انسان-در-حلقه» چه تأثیری بر زمان تحلیل و مقیاس پردازش داده در مقایسه با تحلیل کاملاً دستی دارد؟

۴-۱. مطالعه موردی: نظام مطابقه در گویش بختیاری ایزده

برای آزمون عملی و عینی این چارچوب، پژوهش بر تحلیل نظام مطابقه در گویش بختیاری شهرستان ایزده متمرکز شده است. این انتخاب از دلایل روش‌شناختی مستحکمی برخوردار است:

۱. اهمیت رده‌شناختی و پیچیدگی ذاتی: مطابقه از هسته‌ای‌ترین و پیچیده‌ترین مؤلفه‌های نحوی است که تنوع بالایی در گویش‌های ایرانی نشان می‌دهد و آزمون‌ی دشوار برای هر مدل تحلیلی محسوب می‌شود (کلباسی، ۲۰۱۶)؛ همچنین رجوع کنید به هیگ، ۲۰۰۸^۶ و سامیان، ۲۰۱۹^۷.

۲. نمایندگی گویش کم‌منبع: با وجود جمعیت زیاد گویشور، گویش بختیاری از نظر منابع دیجیتال و پیکره‌های حاشیه‌نویسی‌شده، زبانی کم‌منبع به شمار می‌رود.

1 SilkRoadNLP (The First Workshop on NLP and LLMs for the Iranian Language Family)

2 lower-resource

3 Benchmark

4 S. M. H. Hashemi et al.

5 I. Kalbasi

6 G. Haig

7 V. Samiian

۳. وجود تنوع درونی و ویژگی‌های خاص: این گویش دارای گونه‌های محلی و پدیده‌های واجی-صرفی خاص (مانند جمع‌سازی با تغییر آوایی و بازمانده‌های ساخت کنایی) است که امکان بررسی حساسیت و محدودیت‌های مدل را فراهم می‌آورد.

۵-۱. ویژگی‌های خاص گویش بختیاری ایزده در نظام مطابقه

پیش از پرداختن به روش‌شناسی و تحلیل، لازم است چهار ویژگی صرفی-نحوی گویش بختیاری ایزده که این گویش را برای آزمون مدل‌های زبانی عمومی دشوار می‌سازد، معرفی گردد (کلباسی، ۲۰۱۶؛ هیگ، ۲۰۰۸؛ سامیان، ۲۰۱۹؛ آنونبی و همکاران، ۲۰۲۱):

- (۱) جمع‌سازی با تغییر آوایی درون‌واژه (مانند /koræ/ به معنای «بچه‌ها» بدون پسوند جمع)؛
 - (۲) بازمانده‌های ساخت کنایی در زمان گذشته (مطابقه فعل با مفعول منطقی به جای نهاد دستوری)؛
 - (۳) ضمیراندازی^۲ (تشخیص نهاد تنها از طریق شناسه فعل)؛
 - (۴) تأثیر فرآیندهای واجی محلی بر شناسه‌های مطابقه (همگونی و حذف واکه در گفتار سریع).
- شناسایی خودکار این ویژگی‌ها برای مدل‌های پیش‌آمخته‌ای که عمدتاً بر داده‌های فارسی معیار و انگلیسی آموزش دیده‌اند چالش‌برانگیز است.

۲. پیشینه پژوهش

۱-۲. سنت توصیفی در مطالعات گویش‌های ایرانی

مطالعات گویش‌شناسی ایران ریشه در سنت زبان‌شناسی توصیفی-ساختاری دارد و میراثی غنی از تحلیل‌های عمیق تک‌گویشی بر جای گذاشته است (دبیرمقدم، ۲۰۱۲). در مورد گویش بختیاری، پژوهش‌های پایه‌ای با دقت نظری بالا به توصیف نظام‌مند دستگاه واجی، ساخت صرفی و ویژگی‌های نحوی از جمله نظام پیچیده مطابقه پرداخته‌اند (کلباسی، ۲۰۱۶؛ آنونبی و همکاران، ۲۰۲۱)؛ برای تحلیل نظام کنایی به هیگ، ۲۰۰۸ و برای مطابقه شکافته به سامیان، ۲۰۱۹ رجوع کنید. از منظر رده‌شناختی، نظام مطابقه در زبان‌های ایرانی با چارچوب‌های نظری کلاسیک مانند کوربت (۲۰۰۶)^۳ و کمری (۱۹۸۹)^۴ قابل تحلیل است. همچنین، تغییر نظام مطابقه از کنایی به فاعلی-مفعولی در زبان‌های ایرانی به تفصیل در هیگ (۲۰۰۸) بررسی شده است و نظام مطابقه شکافته^۵ در گویش‌های غربی ایران موضوع پژوهش‌هایی مانند سامیان (۲۰۱۹) بوده است. برای توصیف نظام‌مند گویش بختیاری، علاوه بر منابع پیشین، مدخل دائرةالمعارف ایرانیکا (آنونبی و اسدی، ۲۰۱۸)^۶ نیز قابل مراجعه است.

اگرچه این مطالعات از حیث دقت تحلیلی بی‌بدیل هستند، اما عمدتاً در چارچوب «مطالعات موردی عمیق با مقیاس محدود» قرار می‌گیرند. محدودیت‌های ذاتی این الگو—شامل زمان‌بری فرآیندها، دشواری مطالعات پویاشناسی زبانی در مقیاس جمعیتی، و چالش مقایسه سیستماتیک میان گویش‌ها—امکان پاسخگویی به پرسش‌های کلان رده‌شناختی و جامعه‌زبان‌شناختی را به

1 E. Anonby et al.

2 Pro-Drop

3 G. G. Corbett

4 B. Comrie

5 Split Agreement

6 E. Anonby & A. Asadi

شدت محدود ساخته است (هیگ و آنونبی، ۲۰۱۸). در عصر انقراض شتابان زبان‌ها، این «بحران مقیاس» مستلزم تحول در روش‌شناسی است.

۲-۲. کاربرد مدل‌های زبانی بزرگ در زبان‌شناسی میدانی

مدل‌های زبانی بزرگ با معماری ترنسفورمر و قابلیت «یادگیری با چند مثال» امکان تحلیل خودکار زبان‌های کم‌منبع را بدون نیاز به پیکره‌های بزرگ فراهم کرده‌اند (براون و همکاران، ۲۰۲۰؛ زوی و همکاران، ۲۰۲۲). پژوهش‌های اولیه نشان داده‌اند که این مدل‌ها می‌توانند در ترجمه و تولید متن برای زبان‌های کم‌منبع عملکرد قابل قبولی داشته باشند (لین و همکاران، ۲۰۲۲). با این حال، در حوزه زبان‌شناسی میدانی و به ویژه برای گویش‌های ایرانی، هنوز چارچوب نظام‌مندی برای ارزیابی دقت و محدودیت‌های این مدل‌ها در تحلیل ویژگی‌های نحوی خاص ارائه نشده است. همچنین بررسی سوگیری‌های ساختاری و خطر تحمیل الگوهای زبان‌های مسلط (مانند انگلیسی یا فارسی معیار) بر گویش‌های کم‌منبع، موضوعی است که در پژوهش‌های جهانی مطرح شده (بندر و همکاران، ۲۰۲۱) اما در مورد گویش بختیاری بررسی تجربی نشده است.

۲-۳. جایگاه و هدف متمایز این پژوهش

این پژوهش با اتخاذ رویکرد انسان-در-حلقه و مطالعه موردی نظام مطابقه در گویش بختیاری ایزده، در پی پر کردن این شکاف دوگانه است.

۳. چارچوب نظری پژوهش

در این پژوهش، تحلیل نظام مطابقه در گویش بختیاری ایزده بر پایه چهار رکن نظری اصلی استوار است.

۳-۱. رده‌شناسی مطابقه در زبان‌های ایرانی

مطابقه رابطه ساختاری میان نهاد و فعل است که در ویژگی‌های شخص، شمار و گاه جنس بازنمایی می‌شود (کوربت، ۲۰۰۶). در گویش بختیاری، نظام مطابقه دارای بازمانده‌هایی از ساخت کنایی در زمان گذشته و جمع‌سازی با تغییر آوایی است (هیگ، ۲۰۰۸؛ سامیان، ۲۰۱۹).

۳-۲. یادگیری چندمثاله‌ای و دانش زبانی نهفته

مدل‌های زبانی بزرگ از طریق پیش‌آموزش بر روی داده‌های چندزبانه، دانش زبانی جهان‌شمول را در پارامترهای خود ذخیره کرده‌اند (براون و همکاران، ۲۰۲۰). قابلیت یادگیری با چند مثال امکان تطبیق سریع با الگوهای یک گویش جدید را فراهم می‌کند (وی و همکاران، ۲۰۲۲).

۳-۳. رویکرد انسان-در-حلقه

1 G. Haig & E. Anonby

2 X. V. Lin et al.

در رویکرد حلقه در انسان، مدل خروجی اولیه را تولید می کند و پژوهشگر متخصص آن را اعتبارسنجی، اصلاح و تفسیر می کند (برد، ۲۰۲۰). این فرایند ضمن افزایش دقت، از تحمیل چارچوب های بیگانه بر داده های گویش جلوگیری می کند.

۳-۴. سوگیری و عدالت داده ای

مدل های زبانی بزرگ^۱ (LLMs) عمدتاً بر داده های زبان های مسلط آموزش دیده اند و دچار سوگیری ساختاری هستند (بندر و همکاران، ۲۰۲۱؛ هاوی و پرابهاوموی، ۲۰۲۱). این پژوهش با شناسایی و طبقه بندی سیستماتیک خطاها، مصادیق این سوگیری را در گویش بختیاری مستند می کند.

۴. روش شناسی

۴-۱. طرح پژوهش: یک رهیافت ترکیبی اکتشافی-کاربردی

این پژوهش با هدف تدوین و اعتبارسنجی یک چارچوب روش شناختی نوین و مسئولانه برای تحلیل ساختارهای رده شناختی در گویش های کم منبع ایرانی طراحی شد. هسته مرکزی این چارچوب، یک روندکار نیمه خودکار انسان محور است که در آن یک مدل زبانی بزرگ پیش آموخته به عنوان یک دستیار هوشمند اولیه عمل می کند و تفسیر نهایی، اعتبارسنجی و کنترل کیفیت بر عهده پژوهشگر متخصص باقی می ماند. مطالعه حاضر به صورت یک پژوهش اکتشافی-کاربردی با تمرکز بر مطالعه موردی استخراج و طبقه بندی الگوهای نظام مطابقه در گویش بختیاری شهرستان ایذه انجام گرفت.

۴-۲. جامعه آماری، نمونه گیری و گردآوری داده ها

جامعه آماری، گویشوران بومی مسن (بازه سنی ۶۰-۷۵ سال) بختیاری ساکن در مناطق روستایی شهرستان ایذه بودند. معیارهای ورود شامل: (۱) زادگاه و اقامت دائم در منطقه، (۲) عدم تسلط به فارسی معیار به عنوان زبان اول، و (۳) تمایل به مشارکت بود. با استفاده از روش نمونه گیری هدفمند و نمونه گیری زنجیره ای (گلوله برفی^۲)، ۱۵ گویشور (۸ مرد و ۷ زن) انتخاب شدند. این حجم نمونه با توجه به ماهیت کیفی-اکتشافی مطالعه و تمرکز بر تنوع درونی گفتار طبیعی (نه نمایندگی آماری کل جامعه) کافی تشخیص داده شد.

داده های زبانی از طریق مصاحبه نیمه ساختاریافته و ضبط گفتار طبیعی در بافت های ارتباطی خودجوش (گفتگوهای دونفره، شرح خاطرات و آیین ها) گردآوری شد. برای حداکثر کیفیت صوتی، از ضبط کننده حرفه ای Zoom H5 و میکروفون یقه ای Rode Lavalier استفاده شد. در مجموع، حدود ۱۲ ساعت گفتار طبیعی ضبط گردید.

فایل های صوتی بر اساس پروتکل پیکره بنیاد گفتار^۳ و با رعایت اصول اخلاقی (گمنام سازی) به صورت لایه ای آوانویسی شدند. این فرآیند شامل: الف) رونویسی آوایی (با الفبای IPA)، ب) ارائه ترجمه تحت اللفظی به فارسی، و ج) یادداشت گذاری اولیه ناقض های نحوی بود. پیکره متنی نهایی حاوی ۵۴۰۰ جمله بود که در یک پایگاه داده ساختاریافته ذخیره گردید.

۱ Large Language Models (LLMs)

۲ Snowball Sampling

۳ TalkBank

۳-۴. ایجاد داده استاندارد طلایی و تقسیم‌بندی پیکره

برای آموزش و ارزیابی عینی مدل، نیاز به یک زیرمجموعه با حاشیه‌نویسی دستی دقیق بود.

- حاشیه‌نویسی دستی: ۶۰۰ جمله از پیکره ($\approx 1\%$) به صورت تصادفی ساده انتخاب شدند. این جملات توسط دو زبان‌شناس مستقل با تخصص در زبان‌های ایرانی، از نظر الگوهای مطابقه (شامل مطابقه فعل با نهاد بر اساس شخص، شمار، و در مواردی جنس) حاشیه‌نویسی شدند. میزان توافق بین ارزیاب‌ها با استفاده از ضریب کاپای کوهن ۰٫۹۴ محاسبه شد که نشان‌دهنده قابلیت اعتماد بسیار بالا است. اختلاف‌های ناچیز در یک جلسه مشترک رفع گردید. این مجموعه به عنوان «داده استاندارد طلایی» در نظر گرفته شد.

- تقسیم‌بندی: داده طلایی به سه بخش تقسیم شد: (۱) مجموعه تمرین-آموزش ($n=100$): برای مهندسی دستور و ارائه مثال به مدل، (۲) مجموعه اعتبارسنجی ($n=50$): برای تنظیم دقیق و بهینه‌سازی دستورها، و (۳) مجموعه آزمون ($n=450$): برای ارزیابی نهایی و تحلیل خطا. مابقی پیکره (۴۸۰۰ جمله) به عنوان مجموعه اکتشافی برای اجرای نهایی مدل و سنجش کارایی در مقیاس بزرگ استفاده گردید.

۴-۴. معماری چارچوب رایانشی: انتخاب مدل، مهندسی دستور و پروتکل اجرا

پس از بررسی مقدماتی چند مدل (شامل GPT-3.5 Turbo و Claude Instant) از نظر هزینه، دسترسی و عملکرد اولیه بر روی نمونه‌های کوچک، مدل Llama 3.1 (8B) پارامتر) به دلایل دسترسی باز، شفافیت نسبی، توانایی زبانی قوی در چندزبانه و ملاحظات محاسباتی انتخاب گردید. این مدل بر روی یک سرور مجهز به دو کارت گرافیک NVIDIA RTX 4090 با استفاده از چارچوب استنتاج تولید متن^۱ مستقر شد.

استراتژی اصلی مبتنی بر یادگیری چندمثاله‌ای و یادگیری زنجیره‌ای تفکر^۲ بود. فرآیند مهندسی دستور تکرارپذیر و نظام‌مند بود و طی ۱۵ دور تکرار بر روی مجموعه اعتبارسنجی انجام شد. در هر دور، معیارهای دقت، بازیابی و خوانایی خروجی بهبود یافت. دستور بهینه نهایی چهاربخشی بود: (۱) تعیین نقش و وظیفه، (۲) ارائه قالب ساختاریافته خروجی، (۳) بیان محدودیت‌ها (مثلاً پاسخ «نامشخص» در صورت ابهام)، و (۴) ارائه ۵ مثال آموزشی متنوع که چالش‌های کلیدی (جمع مکرر، ضمیر پنهان، ساخت‌های منفی و مجهول) را پوشش می‌دادند. نمونه کامل دستور نهایی و مثال‌ها در پیوست ۱ ارائه شده است.

تعامل با مدل از طریق یک اسکریپت پایتون خودکارسازی شد. هر جمله از مجموعه‌های آزمون و اکتشافی، در چارچوب دستور نهایی به مدل تزریق و پاسخ آن ثبت می‌گردید. برای اطمینان از پایداری، هر جمله دو بار با سکشن تصادفی متفاوت اجرا و در صورت اختلاف پاسخ، توسط پژوهشگر بررسی می‌شد.

۴-۵. روش‌های تحلیل داده‌ها و ارزیابی

عملکرد مدل بر روی مجموعه آزمون با معیارهای استاندارد دقت، بازیابی، نمره F1 و دقت کلی ارزیابی شد. برای بررسی معناداری آماری تفاوت عملکرد در زیررده‌های مختلف، از آزمون کای-دو استفاده گردید. علاوه بر این، تمامی خطاهای مدل در مجموعه آزمون، بر اساس یک چارچوب رده‌بندی علی قیاسی مورد تحلیل کیفی قرار گرفتند. این رده‌بندی از داده‌ها استخراج شد و شامل دسته‌های اصلی زیر بود: الف) سوگیری به ساختار فارسی معیار/انگلیسی، ب) عدم تشخیص نشانه‌های جمع نامنظم یا پدیده‌های

1 Text Generation Inference
2 Chain-of-Thought

واجبی خاص، ج) بدفهمی نقش نحوی کلمات در بافت گویش، د) توهم محض (تولید تحلیل بدون شواهد)، و ه) خطای ناشی از پیچیدگی ساختاری جمله.

سپس زمان صرفه‌جویی شده با مقایسه دو روش محاسبه شد: ۱) زمان متوسط تحلیل کاملاً دستی هر جمله (برآوردشده از زمان حاشیه‌نویسی داده طلایی)، و ۲) زمان متوسط تحلیل نیمه خودکار (شامل زمان اجرای خودکار مدل + زمان بررسی و اصلاح نهایی توسط پژوهشگر).

۴-۶. ملاحظات اخلاقی و محدودیت‌های روش‌شناختی

کلیه پروتکل‌های اخلاقی رعایت شد: اخذ رضایت آگاهانه کتبی، تضمین گمنام ماندن کامل داده‌ها و اطلاعات هویتی، و تصریح مالکیت معنوی جمعی داده‌ها برای جامعه گویشوران. داده‌های خام و پیکره نهایی در یک مخزن بایگانی امن با دسترسی کنترل‌شده نگهداری می‌شوند.

محدودیت‌های ذاتی روش حاضر عبارتند از: ۱) وابستگی عملکرد به مهندسی دستور که فرآیندی تجربی و نیازمند تخصص دوگانه (زبان‌شناسی و NLP) است، ۲) سوگیری بنیادین مدل پایه به سوی زبان‌های پرمنبع، ۳) هزینه محاسباتی نسبتاً بالا برای استقرار مدل‌های بزرگ، و ۴) مشکل تحلیل داده‌های کاملاً شفاهی و فاقد متن. این محدودیت‌ها به صراحت در بخش بحث مورد تحلیل قرار خواهند گرفت.

۵. نتایج

تحلیل داده‌ها بر اساس چارچوب روش‌شناختی تشریح‌شده یافته‌هایی را در سه سطح کلان ارائه می‌دهد: ۱) عملکرد کمی مدل، ۲) تحلیل کیفی عمیق خطاها و الگوهای موفق، و ۳) سنجش کارایی روندکار. همچنین، یک تحلیل مقایسه‌ای اولیه با یک روش قاعده‌بنیاد ساده انجام شد تا جایگاه واقعی عملکرد مدل مشخص گردد.

چهار ویژگی صرفی-نحوی گویش بختیاری ایزده که به عنوان چالش‌های اصلی تشخیص مطابقت در نظر گرفته شده‌اند، پیشتر در بخش مقدمه معرفی گردیدند. در ادامه، عملکرد مدل بر پایه همین ویژگی‌ها ارزیابی و خطاها دسته‌بندی می‌شوند.

۵-۱. عملکرد کمی مدل

ارزیابی مدل بر روی مجموعه آزمون استاندارد طلایی (۴۵۰ جمله) نشان داد که مدل در شناسایی و طبقه‌بندی الگوهای مطابقت از دقت قابل‌توجهی برخوردار است. جدول ۱ خلاصه‌ای از معیارهای عملکرد را نشان می‌دهد. نتایج نشان می‌دهد موازنه خوبی بین دقت^۱ و بازیابی^۲ برقرار است که در نمره F1 متعادل ۸۷٫۴٪ منعکس شده است.

جدول (۱)

عملکرد مدل Llama 3.1 در استخراج الگوهای مطابقت (مجموعه آزمون، n=450)

معیار	نمره	تفسیر
دقت کلی ^۲	۸۷٫۴٪	درصد کل جملاتی که الگوی مطابقت در آنها به درستی شناسایی شد.

1 Precision

2 Recall

3 Accuracy

معیار	نمره	تفسیر
دقت ^۱	۸۹,۲٪	از میان تمامی الگوهای که مدل به عنوان «مطابقه» برچسب زد، درصد موارد صحیح.
بازیابی	۸۵,۷٪	از میان تمامی الگوهای مطابقه موجود در داده، درصد مواردی که مدل موفق به شناسایی آنها شد.
نمره F1	۸۷,۴	میانگین توازن یافته هارمونیک دقت و بازیابی.

پایداری در مقیاس بزرگ: عملکرد مدل در مجموعه اکتشافی (۴۸۰۰ جمله) با نمونه‌گیری تصادفی و ارزیابی دستی ۵۰۰ جمله از آن، نرخ دقت مشابهی در حدود ۸۶,۱٪ را نشان داد. این نتیجه، حاکی از پایداری نسبی مدل در تعمیم به داده جدید و حجم بزرگتر است.

مقایسه با یک روش قاعده‌بنیاد^۲: برای درک بهتر عملکرد LLM، یک روش قاعده‌بنیاد ساده طراحی شد که مبتنی بر جستجوی پسوندهای فعل پرتکرار فارسی معیار و منطبق کردن آن با نشانه‌های آشکار نهاد بود. این روش پایه بر روی همان مجموعه آزمون تنها به دقت کلی ۶۲,۳٪ دست یافت. این شکاف عملکرد (~۲۵٪)، قدرت استنتاج الگوی مدل را در مواجهه با تنوع و پیچیدگی گویش به وضوح نشان می‌دهد.

۲-۵. تحلیل کیفی خطاها و الگوهای بهینه

تحلیل خطاها بر اساس چهار ویژگی معرفی شده در بخش مقدمه انجام شده است. علیرغم عملکرد کلی خوب، تحلیل کیفی ۵۷ خطا (معادل ۱۲,۶٪ از مجموعه آزمون) الگوهای نظام‌مندی را آشکار کرد که برای درک محدودیت‌های مدل حیاتی هستند. خطاها در چهار دسته اصلی طبقه‌بندی علی شدند (جدول ۲). علاوه بر این، تحلیل خروجی‌های صحیح مدل نیز الگوهای موفقیت آن را مشخص کرد.

جدول (۲)

توزیع انواع خطاهای سیستماتیک مدل (بر اساس تحلیل ۵۷ خطا در مجموعه آزمون)

نوع خطا	درصد	ویژگی گویش مرتبط
سوگیری ساختاری	۴۲٪	ویژگی‌های ۲ و ۳ (ساخت کنایی، ضمیراندازی)
نابینایی واجی-صرفی	۳۵٪	ویژگی‌های ۱ و ۴ (جمع آوایی، فرآیندهای واجی محلی)
بدفهمی نقش نحوی	۱۵٪	ویژگی ۲ (ساخت کنایی)
توهم محض	۸٪	تلفیقی از ویژگی‌های ۱+۳ یا ۲+۴

- **سوگیری ساختاری (۴۲٪ خطاها).** این خطاها عمدتاً در مواجهه با ویژگی‌های شماره‌های ۲ و ۳ (بازمانده کنایی و ضمیراندازی) رخ دادند، زیرا مدل انتظار داشت نهاد همیشه به صورت آشکار و با مطابقه فاعلی-مفعولی ظاهر شود.
- **نابینایی واجی-صرفی (۳۵٪ خطاها).** بیشترین آسیب در این دسته مربوط به ویژگی‌های شماره‌های ۱ و ۴ (جمع‌سازی با تغییر آوایی و تأثیر فرآیندهای واجی محلی) بود، زیرا مدل الگوهای قاعده‌مند پسوندی را جستجو می‌کند و تغییرات درون‌واژه یا تضعیف شناسه را تشخیص نمی‌دهد.

1 Precision
2 Baseline

- بدفهمی نقش نحوی (۱۵٪ خطاها). این دسته عمدتاً متأثر از ویژگی شماره ۲ (ساخت کنایی) بود، که در آن مدل نهاد منطقی را به اشتباه به عنوان فاعل دستوری برچسب زد.
- توهم محض (۸٪ خطاها). این خطاها عمدتاً در جملاتی با دو ویژگی همزمان (مثلاً هم جمع آوایی و هم ضمیراندازی) رخ داد که مدل قواعد متقاطع را تشخیص نداد.

۱-۲-۵. یافته‌های کلان از تحلیل خطا

- الگوی موفقیت: مدل در شناسایی الگوهای قاعده‌مند و پرتکرار (مانند مطابقت شخص و شمار در جملات ساده خبری) بسیار موفق است (دقت > ۹۵٪). این نشان می‌دهد مدل به سرعت می‌تواند نگاشت‌های پایه بین فرم و عملکرد در گویش هدف را یاد بگیرد.
- الگوی شکست سیستماتیک (الف): سوگیری به فارسی معیار. مدل تمایل داشت ساختارهای کم‌منبع یا نامتعارف گویش را به نزدیک‌ترین ساختار پرتکرار در فارسی معیار «تحریف» کند (مانند تفسیر ساخت مجهول ارگتیو به عنوان معلوم ساده). این خطاها به طور خاص متأثر از ویژگی شماره ۲ (ساخت کنایی) و ویژگی شماره ۳ (ضمیراندازی) بودند، زیرا مدل در داده‌های آموزش خود الگوی مطابقت فاعلی-مفعولی صریح دیده است و ساخت‌های کنایی یا نهادهای پنهان را به درستی پردازش نمی‌کند.
- الگوی شکست سیستماتیک (ب): نابینایی واجی-صرفی. مدل در تشخیص نشانه‌های جمع نامتعارف و تأثیر فرآیندهای واجی محلی (مانند ادغام یا حذف واکه) بر نشانه مطابقت، مستعد خطا بود. این دسته عمدتاً ناشی از ویژگی شماره ۱ (جمع‌سازی با تغییر آوایی) و ویژگی شماره ۴ (تأثیر فرآیندهای واجی محلی بر شناسه‌ها) بود. مدل به دنبال نشانه‌های جمع پسوندی (مانند «-ها» در فارسی) می‌گردد و تغییر واکه را نادیده می‌گیرد. همچنین تضعیف شناسه در گفتار سریع مانع بازشناسی مطابقت می‌شود.

مثال ۱ (خطای ناشی از جمع‌سازی با تغییر آوایی - ویژگی شماره ۱):

گویش (خط فارسی محلی): گُرا هُواغش شکست

/ˈkoræ hævɒ:ˈyɛʃ ʃikʰæst/

تجزیه واژه‌به‌واژه: /ˈkoræ/ (بچه) /hævɒ:ˈyɛʃ/ (اسباب‌بازی-ش) /ʃikʰæst/ (شکست. PST.3SG)

ترجمه آزاد: بچه‌ها اسباب‌بازی‌اش را شکستند.

تحلیل خطا: مدل نهاد /ˈkoræ/ را مفرد تشخیص داد و فعل /ʃikʰæst/ را سوم شخص مفرد گرفت در حالی که در این بافت /ˈkoræ/ معنای جمع دارد و فعل صورت سوم شخص جمع است. این خطا مصداق ویژگی شماره ۱ (جمع‌سازی با تغییر آوایی) است که مدل قادر به تشخیص آن نبود.

مثال ۲ (خطای ناشی از ساخت کنایی - ویژگی شماره ۲):

گویش (خط فارسی محلی): وِگِه مداره شو

/væˈgʲeʰ moˈdɒ:ræ ˈʃo/

تجزیه واژه‌به‌واژه: /væˈgʲeʰ/ (گویا) /moˈdɒ:ræ/ (مدیر) /ʃo/ (شد. PST.3SG)

ترجمه آزاد: گویا مدیر شد / منصوب شد.

تحلیل خطا: مدل نهاد /mo'dɒ:ræ/ را فاعل دستوری و فعل /ʃo'/ را مطابق با آن تشخیص داد. اما این ساخت مجهول کنایی با نشانه /væ'gje^h/ است و نهاد نحوی در نقش مفعول منطقی قرار دارد و فعل به صورت بی‌شخص به کار رفته است. این خطا مصداق ویژگی شماره ۲ (ساخت کنایی) است که مدل آن را به ساختار فاعلی-مفعولی فارسی معیار تحریف کرد.

مثال ۳ (خطای ناشی از ضمیراندازی - ویژگی شماره ۳):

گویش (خط فارسی محلی): دینام بازی خاھن

/dɪnɒ:m bɒ:zi: xɒ:hæn/

تجزیه واژه‌به‌واژه: /dɪnɒ:m/ (فردا) /bɒ:zi:/ (بازی) /xɒ:hæn/ (خواهیم کرد، FUT.1PL)

ترجمه آزاد: فردا بازی خواهیم کرد.

تحلیل خطا: در این جمله نهاد به صورت جداگانه ذکر نشده است و شخص و شمار تنها از طریق شناسه فعل (xɒ:hæn = اول شخص جمع) قابل تشخیص است. مدل نهاد را «نامشخص» برگرداند (در حالی که باید اول شخص جمع تشخیص می‌داد). این خطا مصداق ویژگی شماره ۳ (ضمیراندازی) است که مدل در شناسایی نهاد از طریق مطابقت فعل ناتوان است.

مثال ۴ (خطای ناشی از تلفیق دو ویژگی - هم جمع آوایی و هم ضمیراندازی):

گویش (خط فارسی محلی): رفته‌ن د شو

/ræftæn de ʃo/

تجزیه واژه‌به‌واژه: /ræftæn/ (رفته‌اند، PERF.3PL) de (در) ʃo (شب)

ترجمه آزاد: شب رفته‌اند.

تحلیل خطا: در این جمله نهاد آشکار وجود ندارد و فعل /ræftæn/ (رفته‌اند) خود حاوی شناسه سوم شخص جمع است. مدل به دلیل نبود نهاد ظاهری، «مطابق نامشخص» برگرداند. این خطا مصداق تلفیق ویژگی شماره ۳ (ضمیراندازی) و تأثیر فرآیند واجی (حذف شناسه در گفتار سریع) است.

جدول (۲) توزیع خطاها را بر اساس ویژگی‌های گویش نشان می‌دهد. همان‌طور که مشاهده می‌شود، ویژگی‌های ۱ و ۴ بیشترین سهم را در خطاهای نابینایی واجی-صرفی و ویژگی‌های ۲ و ۳ بیشترین سهم را در خطاهای سوگیری ساختاری دارند.

۳-۵. سنجش کارایی روندکار نیمه خودکار

مهم‌ترین یافته کاربردی، صرفه‌جویی چشمگیر در زمان تحلیل بود. برآورد زمان بر اساس زمان‌سنجی دقیق نشان داد:

- تحلیل کاملاً دستی: به طور متوسط ۲۱۰ ثانیه برای هر جمله (شامل گوش دادن دقیق، آوانگاری، تحلیل نحوی و ثبت).
- روندکار نیمه خودکار: به طور متوسط ۴۵ ثانیه برای هر جمله (شامل زمان پردازش خودکار مدل + زمان بررسی، اعتبارسنجی و تصحیح نهایی خروجی توسط پژوهشگر).
- این ارقام به معنای کاهش ۷۸٪ در زمان تحلیل هر جمله و امکان پردازش حجمی ۴٫۶ برابری داده در واحد زمان مشابه است. کیفیت نهایی خروجی پس از مرحله نظارت پژوهشگر، معادل با تحلیل دستی کامل ارزیابی شد.

۶. بحث

در این بخش، یافته‌های پژوهش در پرتو چارچوب نظری تفسیر و پیامدهای روش‌شناختی و اخلاقی آن استخراج می‌شوند. یافته‌های این پژوهش نشان می‌دهد که مدل Llama 3.1 با استفاده از یادگیری چندمثاله‌ای می‌تواند الگوهای مطابقه را در گویش بختیاری با دقت ۸۷,۴٪ شناسایی کند که ۲۵٪ بهتر از یک روش قاعده‌بنیاد ساده است. با این حال، دو محدودیت عمده مشاهده شد: (۱) سوگیری ساختاری (۴۲٪ خطاها) ناشی از تحمیل الگوهای فارسی معیار بر ساخت‌های کنایی و ضمیرانداز، و (۲) نابینایی واجی-صرفی (۳۵٪ خطاها) ناشی از ناتوانی مدل در تشخیص جمع‌سازی با تغییر آوایی و تأثیر فرآیندهای واجی محلی. چارچوب «انسان-در-حلقه» پیشنهادی زمان تحلیل را ۷۸٪ کاهش می‌دهد و امکان پردازش ۴,۶ برابری داده را فراهم می‌کند، اما وابستگی به مهندسی دستور و سوگیری ذاتی مدل پایه از محدودیت‌های اصلی آن است.

۱-۶. تفسیر عملکرد مدل: دانش زبانی جهان‌شول، تعمیم‌پذیری محدود

عملکرد موفق مدل با نظریه یادگیری انتقالی^۱ قابل توضیح است: Llama 3.1 از طریق پیش‌آموزش بر داده‌های چندزبانه، مفاهیم نحوی جهان‌شول (مانند نهاد، فعل، شخص، شمار) را در پارامترهای خود ذخیره کرده است (براون و همکاران، ۲۰۲۰؛ وی و همکاران، ۲۰۲۲). یادگیری چندمثاله‌ای به مدل امکان می‌دهد این دانش را به گویش بختیاری نگاشت کند. اما تحلیل خطاها مرزهای این تعمیم‌پذیری را نشان می‌دهد. سوگیری ساختاری (۴۲٪ خطاها) ناشی از تحمیل الگوهای زبان‌های مسلط است (بندر و همکاران، ۲۰۲۱) و نابینایی واجی-صرفی (۳۵٪ خطاها) نشان می‌دهد که مدل در سطح نحو صوری متوقف می‌ماند و تعامل واج‌شناسی و صرف را درک نمی‌کند (بندر و کولر، ۲۰۲۰).

۲-۶. پیامدهای روش‌شناختی

صرفه‌جویی ۷۸٪ در زمان، سه پیامد اصلی دارد: (۱) تغییر نقش پژوهشگر از تحلیلگر دستی به مدیر هوشمند و مفسر نهایی. (۲) امکان مطالعات رده‌شناختی در مقیاس کلان (۴,۶ برابر داده بیشتر). (۳) دموکراتیک‌سازی فناوری با ارائه الگوهای دستوری از پیش بهینه‌شده (پیوست ۱) که بار فنی استفاده از LLM را برای زبان‌شناسان میدانی کاهش می‌دهد.

۳-۶. ملاحظات اخلاقی و مسئولیت اجتماعی

رعایت رضایت آگاهانه، گمنام‌سازی و مالکیت جمعی داده‌ها برای جامعه گویشوران از اصول این پژوهش بوده است. چارچوب «انسان-در-حلقه» با قراردادن پژوهشگر به عنوان دروازه‌بان نهایی، از «استعمار داده‌ای» جلوگیری می‌کند (برد، ۲۰۲۰). محصولات نهایی باید به جامعه بازگردند (کتاب‌های آموزشی، فرهنگ‌های دیجیتال).

۴-۶. محدودیت‌ها و نقشه راه پژوهشی آینده

مهم‌ترین محدودیت‌ها: وابستگی به مهندسی دستور (فرآیندی تجربی)، سوگیری ذاتی مدل پایه، هزینه محاسباتی و نیاز به داده آوانویسی‌شده. پژوهش‌های آینده باید به (۱) توسعه کتابخانه استاندارد از دستورهای بهینه‌شده برای پدیده‌های رده‌شناختی

1 Transfer Learning

2 E. M. Bender & A. Koller

مختلف، (۲) تنظیم دقیق^۱ یک مدل متن‌باز بر روی پیکره متوازن از گویش‌های ایرانی، و (۳) ادغام با مدل‌های صوتی-زبانی برای تحلیل مستقیم گفتار خام پردازند.

۷. نتیجه‌گیری

این پژوهش با هدف طراحی و اعتبارسنجی یک چارچوب نیمه خودکار «انسان-در-حلقه» مبتنی بر مدل زبانی بزرگ برای تحلیل نظام مطابقه در گویش بختیاری ایده انجام شد. یافته‌های کمی نشان داد که مدل Llama 3.1 با استفاده از یادگیری چندمثاله‌ای به دقت کلی ۸۷٫۴٪ و نمره F1 معادل ۸۷٫۴ دست می‌یابد که ۲۵٪ برتر از یک روش قاعده‌بنیاد ساده است. تحلیل کیفی خطاها، دو محدودیت اصلی مدل را آشکار ساخت: (۱) سوگیری به ساختارهای پرتکرار فارسی معیار و انگلیسی (۴۲٪ خطاها) که منجر به تحریف ساخت‌های خاص گویش می‌شود، و (۲) نابینایی واجی-صرفی (۳۵٪ خطاها) که در مواجهه با پدیده‌هایی مانند جمع مکسر ظاهر می‌گردد. از منظر کاربردی، چارچوب پیشنهادی زمان تحلیل را ۷۸٪ کاهش داده و پردازش ۴٫۶ برابری داده را در واحد زمان ممکن ساخته است.

چارچوب ارائه‌شده سه پیامد مهم دارد: (۱) تغییر نقش پژوهشگر از تحلیلگر دستی به مدیر هوشمند و مفسر نهایی، (۲) امکان پذیر شدن مطالعات رده‌شناختی در مقیاس کلان، و (۳) دموکراتیک‌سازی فناوری از طریق ارائه الگوهای دستوری از پیش‌بهینه‌شده (پیوست ۱). با این حال، پژوهش با محدودیت‌هایی مانند وابستگی به مهندسی دستور، سوگیری ذاتی مدل پایه، هزینه محاسباتی و نیاز به داده‌های آوانویسی‌شده روبرو است. رعایت اصول اخلاقی (رضایت آگاهانه، گمنام‌سازی، مالکیت جمعی داده‌ها) و اجتناب از استعمار داده‌ای از ارکان اصلی این چارچوب بوده است.

در نهایت، این پژوهش اثبات می‌کند که مدل‌های زبانی بزرگ می‌توانند شریکی توانمند برای زبان‌شناسان میدانی باشند، مشروط بر اینکه در قالب یک همکاری دیالکتیکی میان سرعت استقرار ماشینی و عمق تفسیر تاریخی-فرهنگی انسان به کار گرفته شوند. پیشنهاد می‌شود پژوهش‌های آتی بر (۱) توسعه کتابخانه‌ای از دستورهای بهینه‌شده برای پدیده‌های رده‌شناختی مختلف، (۲) تنظیم دقیق یک مدل پایه بر روی پیکره متوازن از گویش‌های ایرانی، و (۳) ادغام با مدل‌های صوتی-زبانی برای تحلیل مستقیم گفتار خام متمرکز شوند.

منابع

- Anonby, E., & Asadi, A. (2018). Bakhtiari dialect. In *Encyclopædia Iranica*. Brill.
- Anonby, E., Taheri-Ardali, M., & Hayes, A. (Eds.). (2021). *The Atlas of the Languages of Iran (ALI)*. Geomatics and Cartographic Research Centre, Carleton University. <http://iranatlas.net/>
- Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). <https://doi.org/10.18653/v1/2020.acl-main.463>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623).
- Bird, S. (2020). Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 3504–3519). <https://doi.org/10.18653/v1/2020.coling-main.313>

1 Fine-Tuning

- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
- Comrie, B. (1989). *Language universals and linguistic typology* (2nd ed.). University of Chicago Press.
- Corbett, G. G. (2006). *Agreement*. Cambridge University Press.
- Crystal, D. (2000). *Language death*. Cambridge University Press.
- Dabir-Moghaddam, M. (2012). *The Oxford handbook of Iranian languages*. Oxford University Press.
- Dabir-Moghaddam, M. (2018). Typology of Iranian languages. In A. Sedighi & P. Shabani-Jadidi (Eds.), *The Oxford handbook of Iranian linguistics* (pp. 45–67). Oxford University Press.
- Haig, G. (2008). *Alignment change in Iranian languages: A construction grammar approach*. Mouton de Gruyter.
- Haig, G., & Anonby, E. (Eds.). (2018). *The languages and linguistics of Western Asia: An areal perspective*. De Gruyter Mouton.
- Hashemi, S. M. H., Ghadrdan, M., Shahhosseini, M., & Vatankhahan Esfahani, E. (2026). *APARSIN: A Multi-Variety Benchmark for Iranic Languages, Dialects, and Accents*. ACL Anthology.
- Hovy, D., & Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8), e12432. <https://doi.org/10.1111/lnc3.12432>
- Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6282–6293).
- Jurafsky, D., & Martin, J. H. (2021). *Speech and language processing* (3rd ed. draft). <https://web.stanford.edu/~jurafsky/slp3/>
- Kalbasi, I. (2016). *Rade shenasi-e nahvi: do ghiyesh-e Bakhtiari-e Majd-e Soleyman va Shushtari (dar ostān-e Khuzestan)* [Syntactic typology: Two dialects of Bakhtiari of Majd-e Soleyman and Shushtari (in Khuzestan province)]. Rashdin Publications.
- Karimi, Y. (2020). *Language shift and endangerment in the Iranian context: A sociolinguistic study*. Palgrave Macmillan.
- Lin, X. V., Mihaylov, T., Artetxe, M., Wang, T., Chen, S., & Tsvetkov, Y. (2022). Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 9019–9052).
- MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk* (3rd ed.). Lawrence Erlbaum Associates.
- Moseley, C. (Ed.). (2010). *Atlas of the world's languages in danger* (3rd ed.). UNESCO Publishing.
- Nettle, D., & Romaine, S. (2000). *Vanishing voices: The extinction of the world's languages*. Oxford University Press.
- Samiiian, V. (2019). Split ergativity in Western Iranian: A typological perspective. *STUF – Language Typology and Universals*, 72(3), 451–478.
- SilkRoadNLP (2025). *The First Workshop on NLP and LLMs for the Iranian Language Family (SilkRoadNLP)*. Held at EACL 2026.
- Smith, R. (2023). Prompt engineering for morphologically rich languages: Best practices for low-resource NLP. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 234–248).
- Stilo, D. (2022). *Iranian and minority languages at the crossroads of globalization*. Brill.
- Team, F. (2024). Evaluating GPT-4 on under-resourced language tasks: A comparative study of open vs. closed models. In *Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC)* (pp. 112–126).
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., ... & Lample, G. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint*, arXiv:2307.09288.
- UNESCO. (2003). *Language vitality and endangerment*. UNESCO Ad Hoc Expert Group on Endangered Languages.
- Unjum, N., Evert, S., Sarveswaran, K., Weerasinghe, R., & Jayatilleke, N. (2026). *LLMs for Low-Resource Languages: A Survey*. Available at SSRN.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.

- Wang, X., Li, Y., Zhang, H., & Chen, W. (2024). Responsible multilingual large language models: A survey on bias mitigation and ethical deployment. *Transactions on Machine Learning Research*. <https://openreview.net/forum?id=example>
- Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Chi, E. H. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*.
- Zhang, S., Liu, Y., Wang, L., & Chen, M. (2024). A survey on multilingual large language models for low-resource languages: Corpora, alignment, and bias. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (pp. 1–25).