Enhancing Large Language Models for Arabic Dialects Using Knowledge-Based Rethinking and Contrastive Learning

Abdulsalam Alharbi[1], Shoaib Jameel[2], Basem Suleiman[1], and Imran Razzak[3,1]

[1] University of New South Wales, Sydney, Australia
{abdulsalam.alharbi, b.suleiman}@unsw.edu.au
[2] University of Southampton, UK
m.s.jameel@southampton.ac.uk
[3] MBZUAI, Abu Dhabi, UAE
imran.razzak@mbzuai.ac.ae

Abstract. Large Language Models (LLMs) have demonstrated remarkable capabilities across several natural language processing tasks. Nonetheless, their efficacy in multilingual and dialectally diverse contexts such as Arabic remains limited, particularly when addressing sensitive topics such as health-related information. This study presents a novel framework that combines knowledge-based Rethinking with dual Contrastive Learning (KBRCL) to improve the factual accuracy and dialectal consistency of generated responses. To assess our framework, we create a dialect-aware benchmark of 3,823 annotated health-related claims across several Arabic dialects. This benchmark enables a comprehensive assessment of both factual correctness and dialect alignment. Experimental findings indicate significant improvements in accuracy (from 46.66% to 88.94%), factual verification (89.80%) and dialect claim alignment (DCAS: 42.35%). These findings underscore the significance of integrating external knowledge sources with dialect-aware learning algorithms to provide more accurate and culturally relevant replies in Arabic NLP applications.

Keywords: contrastive learning · text classification · Arabic Dialects · health-related claims

Introduction

The growing adoption of Large Language Models (LLMs) for fact-checking and information retrieval has sparked critical discussions about their reliability, particularly regarding the accuracy, consistency, and contextual sensitivity of facts[8]. These concerns are especially pronounced in linguistically diverse and dialect-rich regions such as the Arab world. Health-related claims present a unique challenge in this context due to their factual complexity, socio cultural implications, and the high stakes associated with potential misinformation. Previous studies have consistently shown that LLMs often struggle to provide responses that are both factually accurate and culturally aligned when operating in different Arabic dialects [8][3]. This challenge is further intensified

by the diglossic nature of the Arabic language, where regional dialects differ significantly from Modern Standard Arabic (MSA) in vocabulary, syntax, and usage. Despite recent progress in Arabic Natural Language Processing (NLP), dialectal variation remains significantly underrepresented in evaluation benchmarks, especially within critical domains like healthcare [3]. Most available resources focus predominantly on MSA, overlooking the nuanced linguistic and cultural expressions inherent in spoken dialects. This gap limits the ability to effectively assess LLMs in real-world scenarios where users communicate informally and in localised language forms [2]. To address these limitations, we propose a novel framework Knowledge-Based Rethinking with Dual Contrastive Learning (KBRCL) which enhances both factuality and dialectal consistency. The framework leverages external medical knowledge for verification and uses contrastive learning to align semantically equivalent responses across dialects. We evaluate our approach using a new benchmark of 3,823 health-related claims across major Arabic dialects and introduce a novel metric, Dialect Claim Alignment Score (DCAS), to assess dialectal fidelity and semantic accuracy. Results show significant improvements across key metrics, demonstrating the framework's effectiveness in dialect-aware factual reasoning. The main contributions of this study are as follows: a unified framework combining external knowledge with contrastive learning to improve factual accuracy and dialectal alignment; a new dialect aware benchmark of 3,823 health-related claims across Arabic dialects; and the introduction of DCAS, a novel metric for evaluating dialectal and semantic alignment in LLM outputs.

Related Work

Dialect

Previous studies have advanced Arabic language understanding and dialect modeling, yet most focus on isolated aspects such as visual tasks [11] or QA based on MSA [5] [1], without addressing factual verification or dialectal consistency. Some works explored single-dialect models [14] or multilingual augmentation [10], but lacked a unified approach to fact checking across dialects. In contrast, our study proposes a comprehensive framework for evaluating Arabic LLMs using health-related claims, integrating both factual accuracy and dialectal alignment. We also introduce the Dialect Claim Alignment Score (DCAS), a novel metric that captures both dimensions.

Contrastive Learning methods

Contrastive Learning (CL) learns robust representations by distinguishing similar from dissimilar samples, often through contrastive loss and Noise Contrastive Estimation (NCE) [21, 4, 6, 9] Dual Contrastive Learning (DCL) extends this approach by combining global and task-specific objectives, demonstrating effectiveness in low-resource and multilingual settings [7, 20]. However, most CL studies overlook dialectal variation and factual accuracy in domains such as healthcare. In this work, we address these gaps by integrating DCL with Knowledge-Based Rethinking to produce accurate and dialect-aware responses.

Language Models (LLMs) with Knowledge-Based

Recent advances in LLMs incorporate external knowledge to improve reasoning and reduce errors, using techniques such as knowledge prompting, retrieval-based verification, and structured memory integration [13, 12, 18]. Models like ERNIE and LLaMA demonstrate how structured or retrieved knowledge enhances robustness across tasks [22, 17]. Building on this, our work applies knowledge-based rethinking to improve factual accuracy and dialectal consistency in Arabic health-related claims.

Methodology

We apply dual contrast learning (DCL) to promote dialectal consistency and Fabricated Claim Injection (FCI) to help the model distinguish true from false claims. Responses are then verified using trusted knowledge bases. If misaligned, a rethinking process is triggered to regenerate the output. By integrating these mechanisms, the model ensures that its responses remain both accurate and dialectally coherent.

Proposed Framework:

Overall, by leveraging **Knowledge-Based Rethinking** and **DCL** (KBRCL) the proposed framework aims to: Enhancing the way large language models (LLMs) handle health related claims in different Arabic dialects while maintaining answer consistency and factual correctness. The process of the proposed framework can be summarised as follows.

Preparation of data and fabrication claims injection (FCI)

To develop the dialect-aware benchmark, we gathered 3,823 distinct health-related claims across four Arabic Dialect [4], health claims are categorized as true or false, using trusted Arabic medical sources [2] [15]. Dialectal labels were designated by native-speaking annotators based on lexical, phonetic, and syntactic indicators. For example, we began constructing a dataset of health-related claims articulated in four major Arabic dialects:

Saudi dialect:

هل النوم بعد الأكل يسبب سمنة؟

(Does sleeping after eating cause weight gain?)

Egyptian dialect:

هل النوم بعد الأكل يخلي الواحد يتخن؟

(Does sleeping after eating make you gain weight?)

Moroccan dialect:

واش النعاس من بعد الماكلة كيجبد السمنة

(Does sleeping after eating make a person gain weight?)

Lebanese dialect:

النوم بعد الأكل بيخلي الواحد يسمن؟

(Does sleeping after eating cause weight gain?)

[4] The dataset is available at: https://github.com/abdulsalamobaid1/arabic-health-claims-dataset

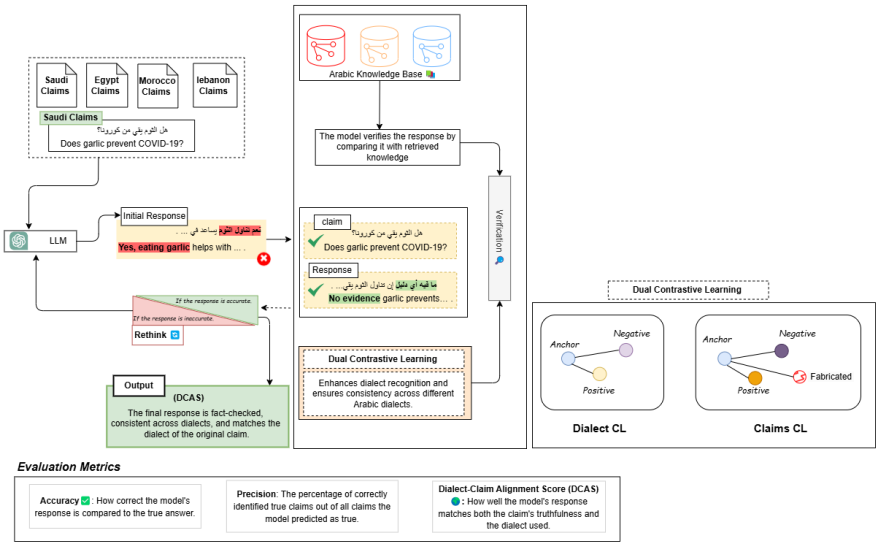

Fig. 1. Overview of the proposed framework for handling dialectal health claims using DCL and FCI mechanisms.

Initial Response

Each claim, in its original dialect, is passed to a language model (such as GPT-4) to obtain an initial response. These responses may exhibit issues such as factual inaccuracy in health-related information or dialectal inconsistency, where the response is generated in Modern Standard Arabic or in a dialect different from that of the original claim.

Two-Stage Processing Phase

The Two-Stage Processing Phase consists of Knowledge-Based Rethinking and Dual Contrastive Learning (KBRCL) to effectively process health claims across various Arabic dialects while ensuring both factual accuracy and dialectal consistency.

First stage: Knowledge-Based Rethinking To ensure the factual accuracy of model responses, each health-related claim and its corresponding initial response are verified against a structured knowledge base built from credible Arabic medical and scientific sources. These include multilingual or Arabic-aligned resources such as ConceptNet [16], Wikidata [19], and ERNIE [22]. The verification process involves the following steps: Retrieval of Relevant Knowledge: Extracting semantically related entries from the knowledge base based on the input claim. Accuracy Evaluation: The model's initial response is compared against the retrieved knowledge: If the response aligns with factual content, it is accepted without modification. If the response contains incorrect or contradictory information, it is corrected or fully regenerated. Reevaluation Through Knowledge Injection: In cases where the initial correction fails or the response remains

uncertain, the retrieved knowledge is fed back into the model as context. The model then generates a revised response based on factual data. To quantify factual alignment, we apply an Accuracy Verification Score, which assesses the consistency of the model's output with trusted knowledge sources. This stage ensures the final output is accurate, medically sound, and aligned with reliable references.

Second Stage: Dual Contrastive Learning (DCL) The second stage ensures dialect consistency while preserving the original dialect of the given claim once the response has been confirmed by Knowledge Based. Dual Contrastive Learning helps to reduce differences between replies produced in several dialects within the same claim. **Methods for ensuring dialect consistency:** Embedding representations of pre-trained language models are used to transform responses in various dialects into numerical embeddings. To enforce semantic coherence, responses with the same meaning across different dialects are brought closer together using Dual Contrastive Learning (DCL), while semantically inconsistent responses are pushed apart. Additionally, the framework addresses dialect-level alignment by ensuring that the model's response to a given health-related claim remains consistent across all dialects. To evaluate this, we introduce the Dialect-Claim Alignment Score (DCAS), which measures how well the generated response aligns with both the factual accuracy of the claim and the dialect in which it was expressed.

Evaluation Metrics

To assess the model, the following metrics have been used: ***Accuracy*** Accuracy measures the proportion of correct predictions out of the total predictions made by the model. ***Precision*** Precision measures the proportion of correctly predicted positive claims out of all claims the model classified as positive. ***Dialect-Claim Alignment Score (DCAS)*** DCAS evaluates whether the model response aligns well with both the health claim and the original dialect in which it was presented. It is calculated based on the proportion of responses that are both correct and in the same dialect.

$$\text{DCAS} = \frac{\text{Number of correct and dialect-matched responses}}{\text{Total number of claims}} \tag{1}$$

In addition, to assess the effectiveness of the proposed approach, we conducted an evaluation comparing model performance before and after applying the rethinking mechanism. To evaluate the effectiveness of our approach, we conducted experiments, as presented in the following section. The results highlight improvements in key evaluation metrics, demonstrating the impact of our proposed framework.

Experiments and Results

Baseline Performance

The baseline results in Table 1 indicate a complete lack of alignment between the dialectal context and the claim verification.

Metric	Value
Accuracy	46.66%
Precision	21.12%
Dialect-Claim Alignment Score (DCAS)	0.00%

Table 1. Baseline (Without Rethink)

Our model result

Our framework, which integrates knowledge-based rethinking and contrastive learning, consistently outperforms the baseline. Using ConceptNet and Wikidata, accuracy reached 88.94%, precision 86.66%, and factuality verification 89.80%, with DCAS improving to 42.35%. The rethinking module corrected 846 out of 1,193 responses, improving performance. When using ERNIE as the knowledge source, the model achieved slightly higher accuracy and precision, with stable recall. Although DCAS decreased slightly, overall performance remained strong.

Metric	ConceptNet/Wikidata	ERNIE
Accuracy	88.94%	90.90%
Precision	86.66%	91.83%
Factuality Verification	89.80%	77.77%
DCAS	42.35%	40.52%
Responses Rethought	1193	1193
Changed Predictions	846	348
Recall	100%	100%
F1-Score	0.5781	0.5647
Correct Dialect Matches	1619 / 3823 (52.88%)	1549 / 3823 (40.52%)

Table 2. Comparison of KBRCL Model Performance Using Different Knowledge Sources

Dialect Breakdown

Table 3 provides a comparison of dialectal prediction extremes between the ConceptNet/Wikidata based setup and the ERNIE-based setup. The Egyptian dialect achieved the highest number of correct predictions (787) under the ConceptNet/Wikidata configuration, while the Saudi dialect had the lowest (51). Under the ERNIE-based setup, the Moroccan dialect showed the highest correct predictions (789), and again, Saudi Arabic recorded the lowest (40). Modern Standard Arabic (MSA) exhibited the highest number of incorrect predictions in both setups (1,579 and 1,683 respectively), whereas Levantine dialect recorded the fewest errors, 1 . Saudi dialect inputs frequently suffered from misinterpretation of health-specific terminology. For example, claims involving COVID-19 transmission or vaccine announcements were often misclassified or rewrit-

ten in MSA, resulting in a loss of dialectal fidelity. Moroccan Arabic posed similar challenges, as terms like

الكرش

(abdominal fat) or

الدوخة

(dizziness) were frequently misunderstood or generalised in MSA. In contrast, Egyptian dialect inputs featuring expressions such as

الضغط العالي

(high blood pressure) or

التطعيم بيحمي فعلاً؟

(Does vaccination really protect?) were handled more accurately. This can be attributed to clearer morphological patterns and higher representation in the model's training data. Together, these findings suggest that both the extent of dialectal exposure during training and the degree of lexical divergence from MSA substantially impact the model's ability to produce accurate and dialect-sensitive responses.

Type	**Dialect**	**Count**
ConceptNet/Wikidata		
Most Correct Dialect	Egypt	787
Least Correct Dialect	Saudi	51
Most Incorrect Dialect	MSA	1579
Least Incorrect Dialect	Levant	1
ERNIE		
Most Correct Dialect	Morocco	789
Least Correct Dialect	Saudi	40
Most Incorrect Dialect	MSA	1683
Least Incorrect Dialect	Levant	1

Table 3. Comparison of Dialect Prediction Extremes Using ConceptNet/Wikidata and ERNIE

Ablation Study

Model	**Accuracy**	**Precision**	**Factuality**	**DCAS**	**Rethought / Changed**
Baseline (without Rethinking)	46.66%	21.12%	N/A	0.00%	– / –
Contrastive Only	58.17%	N/A	N/A	42.35%	– / –
KBRCL with ConceptNet/Wikidata	88.94%	86.66%	89.80%	42.35%	1193 / 846
KBRCL with ERNIE	**90.90%**	**91.83%**	77.77%	40.52%	1193 / 348

Table 4. Performance Comparison of KBRCL Approaches Using Different Knowledge Sources Relative to the Baseline

To analyse the contribution of each component in our proposed KBRCL framework, we conducted an ablation study evaluating different model configurations. As presented in Table 4 comparative analysis between the baseline model, a variant employing only contrastive learning, and two full KBRCL systems that use different external knowledge sources (ConceptNet/Wikidata and ERNIE). These results confirm that both knowledge-based rethinking and contrastive learning contribute meaningfully to improving factual verification, answer precision, and dialectal alignment. The ablation study confirms that integrating knowledge-based rethinking with dual contrastive learning (KBRCL) substantially enhances factual accuracy, answer precision, and dialectal alignment. While the ERNIE-based variant achieves the highest accuracy, the ConceptNet/Wikidata model performs better in factual verification and prediction corrections.

Conclusions

The proposed framework examined the shortcomings of Large Language Models (LLMs) in processing health-related claims in various Arabic dialects. In addition, leveraging knowledge-based rethinking with dual contrastive learning (KBRCL) has enhanced the model's performance concerning factual accuracy and dialectal consistency. To evaluate the model, we have introduced an innovative metric, Dialect-Claim Alignment Score (DCAS). Overall, this study significantly contributes to the body of knowledge of NLP by offering a practical, culturally informed solution that improves the trustworthiness of Arabic health-related NLP applications. Beyond technical gains, the framework supports more equitable access to health information by ensuring dialect-aware, accurate responses.

References

1. Abdallah, A., Kasem, M., Abdalla, M., Mahmoud, M., Elkasaby, M., Elbendary, Y., Jatowt, A.: Arabicaqa: A comprehensive dataset for arabic question answering. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2049–2059 (2024)
2. obaid Alharbi, A., Alsuhaibani, A., Alalawi, A.A., Naseem, U., Jameel, S., Kanhere, S., Razzak, I.: Evaluating large language models on health-related claims across arabic dialects. In: Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script. pp. 95–103 (2025)
3. ALMutairi, M., AlKulaib, L., Aktas, M., Alsalamah, S., Lu, C.T.: Synthetic arabic medical dialogues using advanced multi-agent llm techniques. In: Proceedings of The Second Arabic Natural Language Processing Conference. pp. 11–26 (2024)
4. Alsuhaibani, A., Razzak, I., Jameel, S., Wang, X., Xu, G.: Climb: Imbalanced data modelling using contrastive learning with limited labels. In: Barhamgi, M., Wang, H., Wang, X. (eds.) Web Information Systems Engineering – WISE 2024. pp. 60–75. Springer Nature Singapore, Singapore (2025)
5. Atef, A., Mattar, B., Sherif, S., Elrefai, E., Torki, M.: Aqad: 17,000+ arabic questions for machine comprehension of text. In: 2020 IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA). pp. 1–6. IEEE (2020)

6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PmLR (2020)
7. Dehghan, S., Yanıkoğlu, B.: Multi-domain hate speech detection using dual contrastive learning and paralinguistic features. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 11745–11755 (2024)
8. Demidova, A., Atwany, H., Rabih, N., Sha'ban, S.: Arabic train at nadi 2024 shared task: Llms' ability to translate arabic dialects into modern standard arabic. In: Proceedings of The Second Arabic Natural Language Processing Conference. pp. 729–734 (2024)
9. Gutmann, M., Hyvärinen, A.: Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. pp. 297–304. JMLR Workshop and Conference Proceedings (2010)
10. Hossain, S., Shammary, F., Shammary, B., Afli, H.: Enhancing dialectal arabic intent detection through cross-dialect multilingual input augmentation. In: Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4). pp. 44–49 (2025)
11. Kadaoui, K., Atwany, H., Al-Ali, H., Mohamed, A., Mekky, A., Tilga, S., Fedorova, N., Artemova, E., Aldarmaki, H., Kementchedjhieva, Y.: Jeem: Vision-language understanding in four arabic dialects. arXiv preprint arXiv:2503.21910 (2025)
12. Li, X., Peng, S., Yada, S., Wakamiya, S., Aramaki, E.: Genkp: generative knowledge prompts for enhancing large language models. Applied Intelligence **55**(6), 464 (2025)
13. Pan, X., Yao, W., Zhang, H., Yu, D., Yu, D., Chen, J.: Knowledge-in-context: Towards knowledgeable semi-parametric language models. arXiv preprint arXiv:2210.16433 (2022)
14. Shang, G., Abdine, H., Khoubrane, Y., Mohamed, A., Abbahaddou, Y., Ennadir, S., Momayiz, I., Ren, X., Moulines, E., Nakov, P., et al.: Atlas-chat: Adapting large language models for low-resource moroccan arabic dialect. arXiv preprint arXiv:2409.17912 (2024)
15. Sheikh Ali, Z., Mansour, W., Elsayed, T., Al-Ali, A.: AraFacts: The first large Arabic dataset of naturally occurring claims. In: Habash, N., Bouamor, H., Hajj, H., Magdy, W., Zaghouani, W., Bougares, F., Tomeh, N., Abu Farha, I., Touileb, S. (eds.) Proceedings of the Sixth Arabic Natural Language Processing Workshop. pp. 231–236. Association for Computational Linguistics, Kyiv, Ukraine (Virtual) (Apr 2021), https://aclanthology.org/2021.wanlp-1.26/
16. Speer, R., Chin, J., Havasi, C.: Conceptnet 5.5: An open multilingual graph of general knowledge. In: Proceedings of the AAAI conference on artificial intelligence. vol. 31 (2017)
17. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
18. Wang, J., Sun, Q., Li, X., Gao, M.: Boosting language models reasoning with chain-of-knowledge prompting. arXiv preprint arXiv:2306.06427 (2023)
19. Wang, X., Gao, T., Zhu, Z., Zhang, Z., Liu, Z., Li, J., Tang, J.: Kepler: A unified model for knowledge embedding and pre-trained language representation. Transactions of the Association for Computational Linguistics **9**, 176–194 (2021)
20. Wang, Y., Wang, Z., Lin, Y., Guo, J., Halim, S., Khan, L.: Dual contrastive learning framework for incremental text classification. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 194–206 (2023)
21. Zhang, Y., Yu, Z., Huang, Y., Tang, J.: Cllmfs: A contrastive learning enhanced large language model framework for few-shot named entity recognition. arXiv preprint arXiv:2408.12834 (2024)
22. Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., Liu, Q.: Ernie: Enhanced language representation with informative entities. arXiv preprint arXiv:1905.07129 (2019)