MuDRiC: Multi-Dialect Reasoning for Arabic Commonsense Validation

Anonymous ACL submission

Abstract

Commonsense validation evaluates whether a sentence aligns with everyday human understanding, a critical capability for developing robust natural language understanding systems. While substantial progress has been made in English, the task remains underexplored in Arabic, particularly given its rich linguistic diversity. Existing Arabic resources have primarily focused on Modern Standard Arabic (MSA), leaving regional dialects underrepresented despite their prevalence in spoken contexts. To bridge this gap, we present two key contributions. We introduce MuDRiC, an extended Arabic commonsense dataset incorporating multiple dialects. To the best of our knowledge, this is the first Arabic multi-dialect commonsense reasoning dataset. We further propose a novel method adapting Graph Convolutional Networks (GCNs) to Arabic commonsense reasoning, which enhances semantic relationship modeling for improved commonsense validation. Our experimental results demonstrate that this approach consistently outperforms the baseline of direct language model fine-tuning. Overall, our work enhances Arabic natural language understanding by providing a foundational dataset and a new method for handling its complex variations. Data and code are available at this link.

1 Introduction

Common sense reasoning is a fundamental task in natural language processing (NLP), enabling machines to interpret and generate text in ways that align with human intuition (Sap et al., 2020). It is critical for AI systems to make plausible inferences about the world and engage in human-like conversation (Davis and Marcus, 2015). However, conversational commonsense often involves implicit social norms, cultural references (Sadallah et al., 2025), and pragmatic reasoning that vary across dialects. Despite progress in English (Levesque et al., 2012; Sap et al., 2019a; Talmor et al., 2019) and other high-resource languages, common sense reasoning remains challenging for languages with dialectal diversity, such as Arabic, primarily due to severe scarcity of annotated data for dialects.

Most existing Arabic common sense benchmarks focus exclusively on Modern Standard Arabic (MSA), neglecting the rich diversity of Arabic dialects (Lamsiyah et al., 2025; Sadallah et al., 2025; Khaled et al., 2023). Dialects such as Egyptian, Gulf, Levantine, and Moroccan dominate everyday communication across the Arab world. Beyond lexical or grammatical variation, these dialects encode fine-grained regional cultural knowledge, making dialectal commonsense reasoning a culturally grounded and challenging task. As a result, models trained solely on MSA often fail to generalize to dialectal content. To address this gap, we introduce the first multi-dialect Arabic commonsense dataset balanced across Egyptian, Gulf, Levantine, and Moroccan dialects.

In terms of approaches to address Arabic common sense tasks, prior work heavily relies on MSA-centered models, e.g., AraBERT (Antoun et al., 2020), which perform barely above chance on dialectal data (Lamsiyah et al., 2025; Khaled et al., 2023). Dialect-specific models such as MARBERT (Abdul-Mageed et al., 2021) also show weak performance due to nuanced differences between dialects. More related work in Appendix A.

We propose integrating base language models with graph-based augmentation to capture deeper semantic relationships. This integration of graph-based methods significantly enhances cross-dialect robustness. To summarize our main contributions:

- We introduce MuDRiC: the first multi-dialect Arabic common sense benchmark, enabling more inclusive and robust Arabic NLP systems.
- We propose graph-based augmentation training strategy to enhance performance on dialectical data.

2 Dataset

Task Description and Formulation Given a single sentence, the task aims to identify whether it is reasonable (labeled as 1) or non-reasonable (labeled as 0), based on its alignment with common sense. We cast commonsense validation as a binary classification task to provide a unified and simple formulation across datasets with different original structures. This setting allows us to evaluate the commonsense plausibility of a single sentence independently, rather than relying on relative comparisons between candidates and facilitates scaling the task to multiple Arabic dialects.

MSA We use two established Modern Standard Arabic (MSA) datasets for commonsense validation: the Arabic Dataset for Commonsense Validation (ADCV; Tawalbeh and Al-Smadi (2020)) and ArabicSense (Lamsiyah et al., 2025).ADCV contains 11,000 instances, each consisting of a pair of sentences, where the task is to select the more commonsensical option. We convert this setup into a binary classification task by separating each sentence pair into two individual sentences, assigning label 1 to the original correct (reasonable) sentence and label 0 to its incorrect counterpart. This results in 22,000 labeled samples. ArabicSense includes 5,650 multiple-choice instances with two candidate sentences per instance, one of which is commonsensical. We apply the same conversion strategy, assigning labels accordingly, yielding 11,288 MSA samples after removing duplication.

Dialects Extension Based on the MSA datasets above, we translate them into four Arabic dialects including Egyptian, Moroccan, Gulf and Levantine using GPT-4o (OpenAI, 2024). The statistical distribution of the extended datasets is summarized in Table 1, while Table 2 presents sample MSA sentences from ADCV alongside their corresponding dialectal translations. Prompting details can be found in Appendix B.

The final composite dataset ensures a parallel representation across four major Arabic dialect families. This addresses a critical gap in Arabic NLP, where previous benchmarks have been limited to either Modern Standard Arabic or isolated dialectal efforts without systematic comparison.

Quality Control We apply multiple quality control steps to ensure the reliability of the translated dataset. First, each dialectal translation is automatically verified using Gemini 2.5 Flash by jointly evaluating the translated sentence and its original MSA counterpart (Prompt in Appendix B). Samples flagged as incorrect by Gemini are subsequently reviewed by native-speaker annotators. In total, approximately 8,580 samples (5.2% of the dataset) were flagged; among these, 27% were confirmed to be incorrect (corresponding to roughly 1.4% of the full dataset) and were corrected by the annotators.

As an additional validation step to the original source datasets, we randomly sample 500 instances and ask two independent annotators to verify the correctness of their commonsense labels (reasonable vs. non-reasonable). All sampled instances were found to be correctly labeled.

Source Dataset	MSA Samples	Dialectal Samples
ADCV	22,000	88,000
ArabicSense	11,288	45,152
Total	**33,288**	**133,152**

Table 1: Statistical distribution of datasets, with each MSA extending to four dialects.

3 Methodology

We explore (i) graph-based augmentation to inject relational structure and (ii) domain-adversarial training (Appendix D) to encourage dialect-invariant representations.

3.1 Graph-based Language Model Reps

Inspired by prior work that integrates graph encoders with Transformer models (Jiawei et al., 2020; Zhibin et al., 2020), we augment pretrained Masked Language Models (MLMs) with a Graph Convolutional Network (GCN) that encodes local word-level relations and surface morphological cues. This is motivated by the continuum of Arabic dialects and their non-standard orthography, which introduce substantial spelling and morphological variation. As a result, sequence-based fine-tuning becomes brittle, and dialect-invariant objectives are less reliable (Sha'ban and Habash, 2025; Bhatia et al., 2025). The graph encoder connects related variants via message passing, complementing contextual semantics and improving cross-dialect commonsense validation.

Pipeline Overview We first construct a semantic graph for each input instance to capture relational

MSA Text	Egyptian	Gulf	Moroccan	Levantine	Label
لا أحد يريد العيش مع الفئران (No one wants to live with rats)	محدش عايز يعيش مع الفيران	ما في أحد يبي يعيش مع الفيران	حتى واحد ما بغا يعيش مع الفيران	ما حدا بده يعيش مع الفيران	1
تقوم جورجيا تك بتدريب التنين (Georgia Tech trains dragons)	جورجيا تك بتدرب التنين	جورجيا تك تقوم بتدريب التنين	جورجيا تك كيدرّبو التنين	جورجيا تك عم تدرب التنين	0

Table 2: MSA and dialectal examples from ADCV. Label 1 = reasonable and label 0 = non-reasonable sentence.

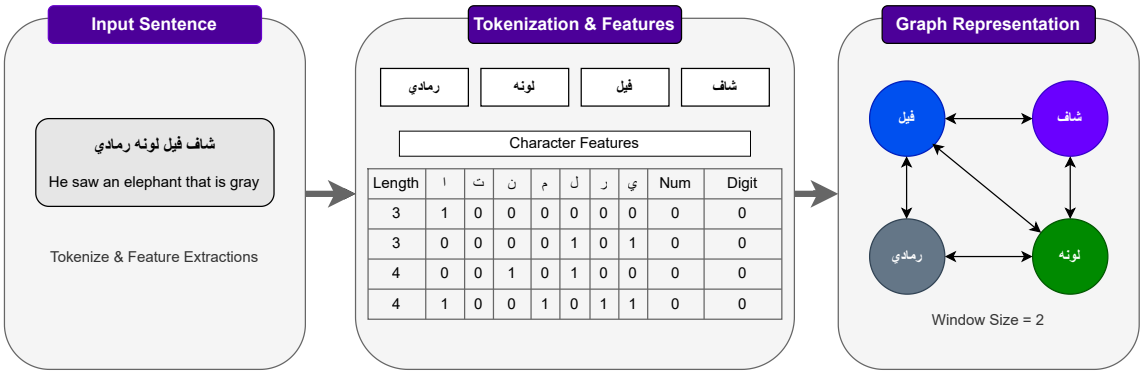

Figure 1: The creation process of the graph representation for a sentence.

dependencies between entities. This graph is processed using GCNs to produce a fixed-length graph representation. In parallel, a BERT-based encoder generates a contextualized text embedding, which is projected to the same dimensionality as the graph representation. The two embeddings are then fused via concatenation. This fusion enriches the model by combining token-level semantic representations with higher-order structural information, enabling more effective reasoning over the input.

Graph Representation To build the graph representation (Figure 1), each input text is first tokenized into words. A co-occurrence graph is then constructed, where nodes correspond to unique words and undirected edges connect words appearing within a fixed-size sliding window. Each node is initialized with a handcrafted feature vector derived from word-level statistics, including word length and Arabic-specific morphological indicators (e.g., character counts and digit-related features). This results in lightweight and informative node representations that reflect the surface morphology and character patterns in the Arabic text.

The resulting graph is then processed by a multi-layer GCN with one hidden layer. The GCN layers propagate and aggregate features across the graph, enabling the model to learn contextual structural patterns. A global mean-pooling layer is then applied to extract a single fixed-length vector that summarizes the entire graph.

Semantic Representation by MLMs In parallel, the input text is encoded using a BERT-based model. We extract the contextual embedding of the "[CLS]" token from the final hidden state layer, which serves as a summary representation of the input sequence. Both the graph and the BERT embeddings are projected into a shared fusion space using learnable linear projections.

Fusion To combine the two representation inputs, we employ a multi-head self-attention mechanism over the concatenated graph and MLM embeddings. This allows the model to dynamically weigh the contribution of each modality and to learn complex interactions between them. The output of the attention layer is flattened and passed through a feedforward classification head. Figure 2 and Algorithm 1 summarize the pipeline in Appendix C.

4 Experiments

4.1 Experimental Setup

Baselines We fine-tune three Arabic-centric pretrained masked language models (MLMs), i.e., CAMeLBERT-mix (Inoue et al., 2021), AraBERT and MARBERT as baselines, and evaluated on all dialects datasets. The three models were pretrained on large Arabic corpus, making them suitable for the task at hand. AraBERT focuses on Modern Standard Arabic. MARBERT emphasizes dialectal Arabic, incorporating substantial representation from various regional dialects, which can better

Methods	MSA	Egyptian	Gulf	Levantine	Moroccan	Avg.
AraBERTv2	65.88	68.33	65.78	65.36	62.09	65.34
AraBERTv2 + GCN	**68.15**	**70.00**	**66.63**	**67.03**	**63.24**	**67.01**
CAMeLBERT-mix	73.30	73.52	74.63	73.82	66.45	72.34
CAMeLBERT-mix + GCN	**74.21**	**75.61**	**76.24**	**76.70**	**68.87**	**74.28**
MARBERTv2	80.12	80.73	78.81	80.03	71.09	78.16
MARBERTv2 + GCN	**81.64**	**81.26**	**80.87**	**80.64**	**73.06**	**79.53**

Table 3: Accuracy of baselines and our method on MSA and dialects datasets in % (higher is better).

capture the dialectal characteristics of the dataset. Similar to MARBERT, CAMeLBERT-mix was pretrained to dialectal Arabic in addition to modern standard Arabic and classical Arabic.

Our Methods We evaluated the effectiveness of graph-based representations of MLMs which fused GCN-based embeddings with masked language models' contextual embeddings. This experimental setup allowed us to examine the hypothesis that graph-enhanced representations can improve downstream task performance.

Data Split and Training Setups In all experiments, we trained models using cross-entropy loss and optimized with AdamW (Loshchilov and Hutter, 2017), using a learning rate of 2e-5, weight decay of 0.01, and a batch size of 128 for 3 epochs. We used 110K MuDRiC samples with ADCV as source dataset, balanced across both commonsense and dialect labels. The data were split into 70%/15%/15% train, development, and test sets. This split was fixed and reused across all experiments to ensure fair and consistent comparisons.

4.2 Results

Table 3 reports accuracy on MSA and four dialectal subsets. The *Avg. Dialects* column corresponds to the overall average across all five subsets (MSA + the four dialects).

Which is the Best Base Model? Across the table, performance follows a stable ranking: *MARBERTv2* > *CAMeLBERT-mix* > *AraBERTv2*. This ordering is consistent with the expected degree of *dialect exposure* during pretraining: MARBERTv2 is explicitly dialect-heavy, CAMeLBERT-mix is more balanced, and AraBERTv2 is comparatively more MSA-oriented. The implication is that dialectal generalization is primarily constrained by representation quality learned during pretraining; downstream methods help, but they do not compensate fully for a strong pretraining mismatch.

GCN Consistently Improves Performance. The consistent improvements across all models suggest that the GCN provides information that the transformer alone does not reliably capture, such as relational structure, lexical/semantic neighborhood effects, or instance-to-instance dependencies. Importantly, the gains appear larger (in relative terms) for the weaker backbones (AraBERTv2 and CAMeLBERT-mix), which supports the interpretation that GCN fusion helps *mitigate* dialect/domain mismatch by encouraging useful sharing across related samples or features. For MARBERTv2, the improvement is still present but more incremental, consistent with the idea that a dialect-rich backbone already captures much of the needed variation and the GCN mainly refines decision boundaries.

Overall, the table supports a clear conclusion: dialect-aware pretraining is the strongest driver of performance, and GCN fusion is a reliable enhancement.

5 Conclusion

This work presented two major contributions to Arabic NLP: (1) the creation of the first large-scale, multi-dialect common sense reasoning dataset, and (2) Enhanced Arabic Commonsense Reasoning methodology combining graph-based embeddings with pre-trained BERT-based models to enhance performance. By systematically expanding MSA commonsense reasoning benchmarks into four major dialects we established a crucial resource for evaluating dialect robustness. Our experiments demonstrated that neither MSA-focused models (e.g., AraBERT) nor dialect-pretrained models (e.g., MARBERT) alone suffice for reliable common sense classification across dialects. Instead, our hybrid graph-based approach to structured commonsense representation outperformed prior methods, setting a new benchmark for dialect-aware Arabic NLP.

Limitations & Future Work

While our work advances dialect-aware common sense reasoning, several limitations warrant discussion: the dialectal data generation process relied on GPT-4o for translation, which may introduce subtle semantic shifts or stylistic inconsistencies compared to naturally occurring dialectal speech, and while we implemented quality checks, the absence of large-scale human validation leaves room for potential noise, particularly in idiomatic expressions requiring deep cultural familiarity; the framework treats all dialects as equally distinct from MSA, overlooking gradient dialectal relationships. For instance, Levantine Arabic shares more lexical overlap with MSA than Moroccan Arabic, potentially leading to uneven generalization where linguistically closer dialects benefit implicitly; the binary labeling scheme (reasonable vs. non-reasonable) oversimplifies the continuum of common sense plausibility, failing to capture partially valid or context-dependent interpretations; moreover, the focus on four major dialects excludes dozens of other Arabic varieties, risking the marginalization of less common dialects like Sudanese or Yemeni Arabic, an area future work should address.

Future work will prioritize reducing the persistent Moroccan gap via dialect-specific adaptation (e.g., continued pretraining, normalization, or dialect-aware modules). We also plan to expand coverage to additional underrepresented varieties (e.g., Sudanese, Yemeni, Algerian, Iraqi), test broader diglossic/code-switched settings (e.g., Moroccan Arabic–French), and extend evaluation beyond sentence-level to contextual or multi-turn commonsense reasoning.

Ethical Statement

Data License A primary ethical consideration in our work is the licensing and provenance of the data used. Our dataset builds upon two publicly available resources: the Arabic Dataset for Commonsense Validation and ArabicSense, both of which have been released for research purposes with appropriate usage permissions. To ensure compliance with licensing constraints, we generated novel dialectal variants derived from the Modern Standard Arabic (MSA) instances provided in the original datasets. This approach ensures that all newly created content remains consistent with the intended research scope of the original licenses and mitigates potential concerns related to data reuse and redistribution.

Biased Language As the dialectal variants were generated using GPT-4o, we rely on the model's built-in safety mechanisms to not generate outputs that may contain biased, offensive, or contextually inappropriate language.

Positive Impact of Commonsense Validation Our work advances existing methods and datasets for Arabic commonsense validation by introducing dialectal variants and exploring novel modeling approaches within this domain. We believe that enhancing commonsense understanding across Arabic dialects can contribute meaningfully to real-world applications such as fake news detection, fact-checking, and mitigating the spread of misleading or harmful content.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Norah Alshahrani, Saied Alshahrani, Esma Wali, and Jeanna Matthews. 2024. Arabic synonym BERT-based adversarial examples for text classification. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 137–147, St. Julian's, Malta. Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Gagan Bhatia, El Moatez Billah Nagoudi, Abdellah El Mekki, Fakhraddin Alwajih, and Muhammad Abdul-Mageed. 2025. Swan and ArabicMTEB: Dialect-aware, Arabic-centric, cross-lingual, and cross-cultural embedding models and benchmarks. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4654–4670, Albuquerque, New Mexico. Association for Computational Linguistics.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In *The*

Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pages 7432–7439. AAAI Press.

Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM*, 58(9):92–103.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-CARE: a new dataset for exploring explainable causal reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–446, Dublin, Ireland. Association for Computational Linguistics.

Javid Ebrahimi, Hao Yang, and Wei Zhang. 2021. How does adversarial fine-tuning benefit bert? *CoRR*, abs/2108.13602.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 6384–6392. AAAI Press.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Mete Ismayilzada, Debjit Paul, Syrielle Montariol, Mor Geva, and Antoine Bosselut. 2023. CRoW: Benchmarking commonsense reasoning in real-world tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9785–9821, Singapore. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Zhang Jiawei, Zhang Haopeng, Xia Congying, and Sun Li. 2020. GRAPH-BERT: Only attention is needed for learning graph representations.

Akbar Karimi, Leonardo Rossi, and Andrea Prati. 2021. Adversarial training for aspect-based sentiment analysis with bert. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 8797–8803.

M Moneb Khaled, Aghyad Al Sayadi, and Ashraf Elnagar. 2023. Commonsense validation and explanation in arabic text: A comparative study using arabic bert models. In *2023 24th International Arab Conference on Information Technology (ACIT)*, pages 1–6.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Salima Lamsiyah, Kamyar Zeinalipour, Samir El amrany, Matthias Brust, Marco Maggini, Pascal Bouvry, and Christoph Schommer. 2025. ArabicSense: A benchmark for evaluating commonsense reasoning in Arabic with large language models. In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, pages 1–11, Abu Dhabi, UAE. Association for Computational Linguistics.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Thirteenth International Conference, KR 2012, Rome, Italy, June 10-14, 2012*. AAAI Press.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-aware graph networks for commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2829–2839, Hong Kong, China. Association for Computational Linguistics.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

OpenAI. 2024. Gpt-4o system card. *CoRR*, abs/2410.21276.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Abdelrahman Sadallah, Junior Cedric Tonga, Khalid Almubarak, Saeed Almheiri, Farah Atif, Chatrine

Qwaider, Karima Kadaoui, Sara Shatnawi, Yaser Alesh, and Fajri Koto. 2025. Commonsense reasoning in Arab culture. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7695–7710, Vienna, Austria. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019a. Atomic: an atlas of machine commonsense for if-then reasoning. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. 2020. Commonsense reasoning for natural language processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 27–33, Online. Association for Computational Linguistics.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2009. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80.

Sanad Sha'ban and Nizar Habash. 2025. The Arabic generality score: Another dimension of modeling Arabic dialectness. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29990–30001, Suzhou, China. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Saja Khaled Tawalbeh and Mohammad Al-Smadi. 2020. Is this sentence valid? an arabic dataset for commonsense validation. *CoRR*, abs/2008.10873.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Cunxiang Wang, Shuailong Liang, Yili Jin, Yilong Wang, Xiaodan Zhu, and Yue Zhang. 2020. SemEval-2020 task 4: Commonsense validation and explanation. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 307–321, Barcelona (online). International Committee for Computational Linguistics.

Lu Zhibin, Du Pan, and Nie Jian-Yun. 2020. VGCN-BERT: Augmenting BERT with graph embedding for text classification. In *Proceedings of the 42nd European Conference on Information Retrieval (ECIR 2020)*, pages 369–382, Lisbon, Portugal (Online). European Conference on Information Retrieval, Springer.

Appendix

A Related Work

A.1 Commonsense Reasoning Datasets

Common Sense Reasoning in English There have been many benchmarks for English commonsense reasoning, such as CommonSenseQA (Talmor et al., 2019), ComVe (Wang et al., 2020), ATOMIC (Sap et al., 2019a) and ATOMIC 2020 (Hwang et al., 2021). Within the broader scope of commonsense reasoning, several specialized subfields have emerged, each targeting distinct types of implicit human knowledge. Earlier work focused on pronoun coreference resolution in linguistic contexts (Levesque et al., 2012), physical commonsense reasoning (Bisk et al., 2020), social reasoning (Sap et al., 2019b), and causal reasoning (Du et al., 2022). Additional efforts have explored commonsense in natural language generation (Lin et al., 2020), as well as the integration of commonsense reasoning into real-world NLP tasks (Ismayilzada et al., 2023).

Despite these advancements, most research and benchmarks are centered around English, leaving many other languages, such as Arabic, underresourced.

CommonSense Reasoning in Arabic Recent years have witnessed exploration of Arabic commonsense reasoning. Initial efforts focused on translating English commonsense benchmarks into Modern Standard Arabic (MSA) (Tawalbeh and Al-Smadi, 2020), or leveraging large language models (LLMs) to generate MSA data from seed data (Lamsiyah et al., 2025). However, these datasets lack cultural nuances of Arabic. Recent work by Sadallah et al. (2025) fills this gap by collecting a dataset covering cultures of 13 countries across the Gulf, Levant, North Africa, and the Nile Valley. Despite this advancement, their dataset remains restricted to MSA and does not encompass the rich linguistic and cultural diversity embedded in Arabic dialects. Therefore, we collect the first Arabic dialects commonsense reasoning benchmark, extending commonsense evaluation to Arabic dialects, aiming to capture more authentic and regionally grounded reasoning patterns.

A.2 Approaches for Commonsense Reasoning

Prior research has primarily focused on fine-tuning transformer-based models or employing LLMs for commonsense validation and explanation generation, without introducing improved task-specific representations that could enhance performance. Tawalbeh and Al-Smadi (2020) fine-tuned BERT, USE, and ULMFit models for binary classification, selecting the more plausible sentence from a pair. More recently, Lamsiyah et al. (2025) evaluated a suite of BERT-based encoders, including AraBERTv2 (Antoun et al., 2020), ARBERT, MARBERTv2 (Abdul-Mageed et al., 2021), CaMeLBERT, and mBERT (Pires et al., 2019), on two classification tasks: *(i)* distinguishing commonsensical from nonsensical statements, and *(ii)* identifying the underlying reasoning behind nonsensicality. They also assessed causal LLMs including Mistral-7B (Jiang et al., 2023), LLaMA-3 (Touvron et al., 2023) and Gemma, on the two tasks above, along with the task *(iii)* generating natural language explanations for commonsense violations. These approaches lacked exploring better representation learning techniques to enhance the performance.

Integration of Adversarial Training with Encoder Transformer Models Prior work has explored integrating adversarial training with transformer-based models. Karimi et al. (2021) introduced BERT Adversarial Training (BAT), which fine-tuned BERT and domain-specific BERT-PT using adversarial perturbations in the embedding space to improve robustness in Aspect-Based Sentiment Analysis (ABSA). Ebrahimi et al. (2021) showed that adversarial training can preserve BERT's syntactic abilities, such as word order sensitivity and parsing, during fine-tuning, compared to standard fine-tuning. Additionally, it demonstrated how adversarial training prevented BERT from oversimplifying representations by reducing over-reliance on a few words, leading to better generalization.

In Arabic context, Alshahrani et al. (2024) conducted a synonym-based word-level adversarial attack on Arabic text classification models using a Masked Language Modeling (MLM) task with AraBERT. This attack replaces important words in the input text with semantically similar synonyms predicted by AraBERT to generate adversarial examples that can fool state-of-the-art classifiers. To ensure grammatical correctness, they utilize CAMeLBERT as a Part-of-Speech tagger to verify that the synonym replacements match the original word's grammatical tags, maintaining sentence grammar.

We investigate the use of adversarial training across dialects as a means to learn more robust and generalized representations, thereby enhancing model performance and resilience across the diverse landscape of Arabic dialects.

Integration of Graph-based Approaches with Encoder Transformer Models Graph Neural Networks (GNNs) (Scarselli et al., 2009), and particularly Graph Convolutional Networks (GCNs) (Kipf and Welling, 2017) have gained significant attention for their ability to model relational and topological structures in data. Integrating graph-based structures with encoder-based Transformer models enables models to better grasp higher-level connections and contextual dependencies that are crucial for complex language understanding tasks like commonsense reasoning. For example, GraphBERT (Jiawei et al., 2020) introduced leveraging Transformer-style self-attention over linkless subgraphs, allowing it to learn graph representations without relying on explicit edge connections. This approach mitigates issues such as over-smoothing and enhances parallelizability. In contrast, VGCN-BERT (Zhibin et al., 2020) adopts a hybrid design, incorporating a vocabulary-level graph convolutional network (VGCN) into the BERT architecture. It constructs a global word co-occurrence graph and fuses the GCN-derived word representations with the BERT input embeddings, thereby enriching the model's understanding of global corpus-level semantics. Both models demonstrate how graph-derived features, when fused effectively with Transformer encoders, can improve downstream tasks like text classification by fusing graph extracted morphological features with the token-level contextual embeddings. In the context of commonsense reasoning, (Lin et al., 2019) proposed KAGNet, a model that integrates GCNs with Long Short-Term Memory networks (LSTMs) (Hochreiter and Schmidhuber, 1997) to encode knowledge paths from external commonsense knowledge bases, thereby improving question answering performance through structured reasoning.

In this work, we integrate graph neural network-projected embeddings into transformer-based encoders, enriching contextual representations with global structural information which are critical to common sense validation.

B Data Generation and Quality Control

Dialect generation We design a prompt tailored for accurate and meaning-preserved translation:

أنت خبير في اللهجات العربية. ترجم الجملة التالية
إلى اللهجة {dialect} بدون تغيير المعنى:
. {sentence}

The prompt translates to "You are an expert in Arabic dialects. Translate the following sentence to {dialect}: {sentence}". This ensures that the intended meaning of each sentence remains intact while reflecting natural dialectal usage.

Automatic Evaluation We prompt Gemini 2.5 Flash to assess the faithfulness of dialectal translations as follows:

> **System Prompt:** You are an expert Arabic linguist. Your task is to verify whether `DIALECT_TEXT` is a correct translation of `MSA_TEXT` from Modern Standard Arabic (MSA) into the specified Arabic dialect. A translation is correct if the meaning, events, entities, time references, and polarity in `MSA_TEXT` are faithfully preserved in `DIALECT_TEXT`, even if wording differs due to dialectal variation. Ignore minor spelling, punctuation, and orthographic differences. Do not allow additions, omissions, or factual changes. Output only one word: `true` or `false`. Do not explain your decision.
>
> ```
> DIALECT: <dialect_name>
> MSA_TEXT: <msa_text>
> DIALECT_TEXT: <dialect_text>
> Answer:
> ```

C Graph Embeddings-based Encoder Transformer models

C.1 Extended Explanation of methodology

Figure 2 and Algorithm 1 summarize the pipeline.

D Domain Adversarial Training

D.1 Method

We implement domain-adversarial training (DANN) (Ganin et al., 2016) to encourage dialect-invariant features during fine-tuning. A shared Transformer encoder produces a sentence representation $\mathbf{h}$ from the final-layer `[CLS]` vector. We attach two MLP heads: (i) a task classifier for commonsense validation, and (ii) a dialect discriminator predicting the dialect label. The dialect discriminator receives $\mathbf{h}$ through a Gradient Reversal Layer (GRL), which multiplies its backpropagated gradient by $-\alpha$, so the encoder is optimized to reduce task loss while making dialect prediction harder.

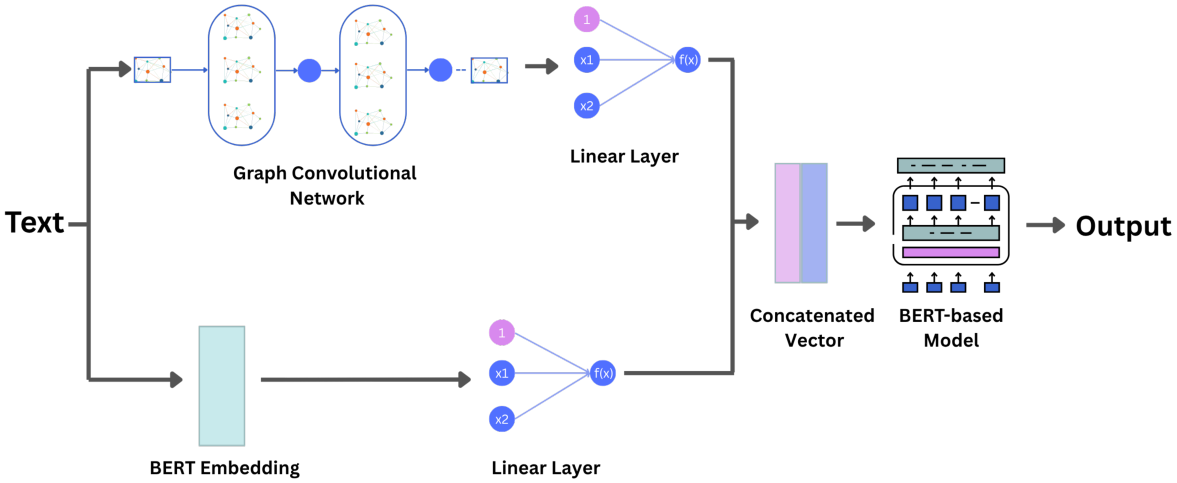

Figure 2: BERT Model with Graph Embeddings Fusion.

Algorithm 1 The training algorithm for Graph Embeddings-based Encoder Transformer models.

Given:
$\mathcal{D}_{\text{train}}$ ▷ Labeled corpora of text samples
$\mathcal{T}$ ▷ Pretrained textual encoder (e.g., BERT)
$\mathcal{G}$ ▷ Graph encoder (e.g., GCN)
$\mathcal{F}$ ▷ Fusion mechanism (e.g., attention)
$\mathcal{C}$ ▷ Classifier head
θ ▷ Trainable parameters
Preparation:
for all $(x, y) \in \mathcal{D}$ **do**
  Tokenize $x \to \mathbf{t} \in \mathbb{R}^{Lh}$
  Convert $x \to$ graph $\mathcal{G}_x = (V, E, \mathbf{X})$
end for
Initialize:
$\theta \leftarrow$ random or pretrained weights
for $e = 1$ to E **do**
  Training Step:
  for all $(x, y, \mathcal{G}_x) \in \mathcal{D}_{\text{train}}$ **do**
    $\mathbf{z}_t \leftarrow \mathcal{T}(x)$ ▷ Textual representation
    $\mathbf{z}_g \leftarrow \mathcal{G}(\mathcal{G}_x)$ ▷ Graph representation
    $\mathbf{z}_f \leftarrow \mathcal{F}(\mathbf{z}_t, \mathbf{z}_g)$ ▷ Fusion
    $\hat{y} \leftarrow \mathcal{C}(\mathbf{z}_f)$ ▷ Prediction
    Update θ via $\nabla_\theta \mathcal{L}(\hat{y}, y)$
  end for
end for

Both heads use the same architecture: `Dropout` $\to$ `Linear`$(H \to 768)$ $\to$ `ReLU` $\to$ `Dropout` $\to$ `Linear`$(768 \to C)$, with dropout rate 0.1, where C=2 for the main task and C=5 for dialect prediction.

We optimize the combined objective:

$$\mathcal{L} = \mathcal{L}_{\text{main}} + \lambda \mathcal{L}_{\text{dial}}, \quad (1)$$

where both terms are cross-entropy losses and the GRL applies the adversarial signal to the shared encoder. We set λ=1.0 and use a simple schedule for the GRL strength $\alpha = \min(1, \frac{e+1}{5})$ as a function of epoch index e.

D.2 Experimental Setup

We follow the same data split and base fine-tuning setup described in Section 4.1. Adversarial models are trained for 3 epochs using AdamW (learning rate $2e-5$, weight decay 0.01), with gradient clipping (max norm 1.0) and a linear warmup over the first 100 optimization steps (from 0.1lr to lr, then constant). We select the best checkpoint using the dev weighted F1 of the main task and report results on the held-out test set.

D.3 Results

Table 4 compares the baseline models against their domain-adversarial training variants. Overall, domain-adversarial training consistently leads to a substantial degradation in accuracy, indicating that enforcing domain invariance in this setting harms learning rather than improving cross-dialect generalization.

Why Does Adversarial Training Fail? All three backbones degrade under adversarial training, and the degradation is especially severe for AraBERTv2 (dropping to near chance). This pattern strongly suggests that dialect-specific cues are

Methods	MSA↑	Egyptian↑	Gulf↑	Levantine↑	Moroccan↑	Avg.↑
AraBERTv2	**65.88**	**68.33**	**65.78**	**65.36**	**62.09**	**65.34**
AraBERTv2 (Adv)	50.52	50.15	49.89	50.03	50.55	50.23
CAMeLBERT-mix	**73.30**	**73.52**	**74.63**	**73.82**	**66.45**	**72.34**
CAMeLBERT-mix (Adv)	66.52	67.12	67.26	68.91	64.18	66.80
MARBERTv2	**80.12**	**80.73**	**78.81**	**80.03**	**71.09**	**78.16**
MARBERTv2 (Adv)	79.58	79.00	77.57	78.52	70.24	76.98

Table 4: Accuracy (%) of baselines and adversarial training-based models on MSA and dialects datasets.

not purely nuisance variation for this task: enforcing dialect invariance likely removes information that is genuinely predictive (lexical, morphological, or orthographic markers correlated with the label). Another plausible contributor is optimization instability: if the adversarial signal is too strong relative to the supervised signal, feature collapse can occur. An effect that would be amplified for a less dialect-robust backbone such as AraBERTv2. Practically, these results argue against treating dialect simply as a domain to be "erased"; a better direction may be *domain-aware* modeling (e.g., dialect embeddings/adapters or mixture-of-experts) rather than domain-invariant representations. In contrast to adversarial training, adding GCN-based embeddings improves *every* backbone.