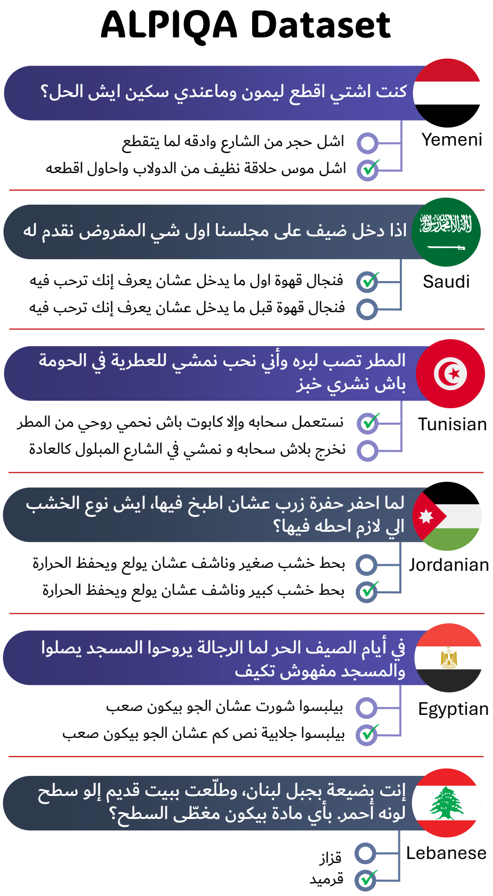
ALPIQA: Dialectal PIQA for Physical Commonsense Reasoning in Arabic

Zaid Alyafeai **Abdelrahman Eldesokey** **Chadi Helwe** **Harethah Abu Shairah**
Sultan AlRashed **Firas Laakom** **Francesco Orabona** **Bernard Ghanem**
KAUST

Abstract

Physical commonsense reasoning is essential for making sense of everyday situations. It requires humans to reason about the physical world and the relationships between its objects and events. PIQA (Bisk et al., 2020) is the first physical commonsense reasoning for English. However, there are no available datasets for low-resource languages like dialectal Arabic. In this paper, we introduce ALPIQA, a manually curated dataset for commonsense reasoning in dialectal Arabic by native speakers. ALPIQA incorporates commonsense examples in six dialects, including Egyptian, Yemeni, Tunisian, Saudi, Lebanese, and Jordanian dialects, with diverse topics. We benchmark ALPIQA on different LLMs that range from tiny to large language models.



كنت اشتي اقطع ليمون وماعندي سكين ايش الحل؟
○ اشل حجر من الشارع وادقه لما يتقطع
✓ اشل موس حلاقة نظيف من الدولاب واحاول اقطعه

اذا دخل ضيف على مجلسنا اول شي المفروض نقدم له
✓ فنجال قهوة اول ما يدخل عشان يعرف إنك ترحب فيه
○ فنجال قهوة قبل ما يدخل عشان يعرف إنك ترحب فيه

المطر تصب لبره وأني نحب نمشي للعطرية في الحومة
باش نشري خبز
✓ نستعمل سحابه وإلا كابوت باش نحمي روحي من المطر
○ نخرج بلاش سحابه و نمشي في الشارع المبلول كالعادة

لما احفر حفرة زرب عشان اطبخ فيها، ايش نوع الخشب
الي لازم احطه فيها؟
○ بحط خشب صغير وناشف عشان يولع ويحفظ الحرارة
✓ بحط خشب كبير وناشف عشان يولع ويحفظ الحرارة

في أيام الصيف الحر لما الرجالة يروحوا المسجد يصلوا
والمسجد مفهوش تكيف
○ بيلبسوا شورت عشان الجو بيكون صعب
✓ بيلبسوا جلابية نص كم عشان الجو بيكون صعب

إنت بضيعة بجبل لبنان، وطلّعت ببيت قديم إلو سطح
لونه أحمر. بأي مادة بيكون مغطّى السطح؟
○ قزاز
✓ قرميد

Figure 1: Examples of the ALPIQA dataset in six different Arabic dialects.

1 Introduction

Physical commonsense reasoning is an important aspect of human understanding. It requires humans to reason about the physical world by interacting with its objects and events. One of the most popular datasets for such task is PIQA (Bisk et al., 2020). Each example in PIQA contains three fields, which are the prompt or question and two possible solutions. For each question, there is only one correct solution. For example, the question "What happens to a ball if we throw it on the ground"? We could have two solutions: the ball will bounce back up, or the ball will bounce down. It requires commonsense to figure out that the first solution is the correct one. The examples in PIQA are in English and are mostly related to Western culture. In this paper, we are interested in creating a dataset that incorporates commonsense reasoning in dialectal Arabic as an under-resourced language. The Arabic language has many dialects depending on the country and the geographical location. Each country may also have different dialects within it. Most Arabic dialects use the same script with slight variations to construct verbs and nouns. Additionally, different dialects may introduce new nouns that were influenced by culture and location. In this paper, we create 600 examples in 6 different Arabic dialects, which are Egyptian, Yemeni, Tunisian, Saudi, Lebanese, and Jordanian. The dataset incorporates examples that are culturally relevant to the dialects of the country it represents. The examples are manually curated by native speakers in each dialect. They also span different topics like food, locations, religion, cultural items, clothing, etc. We

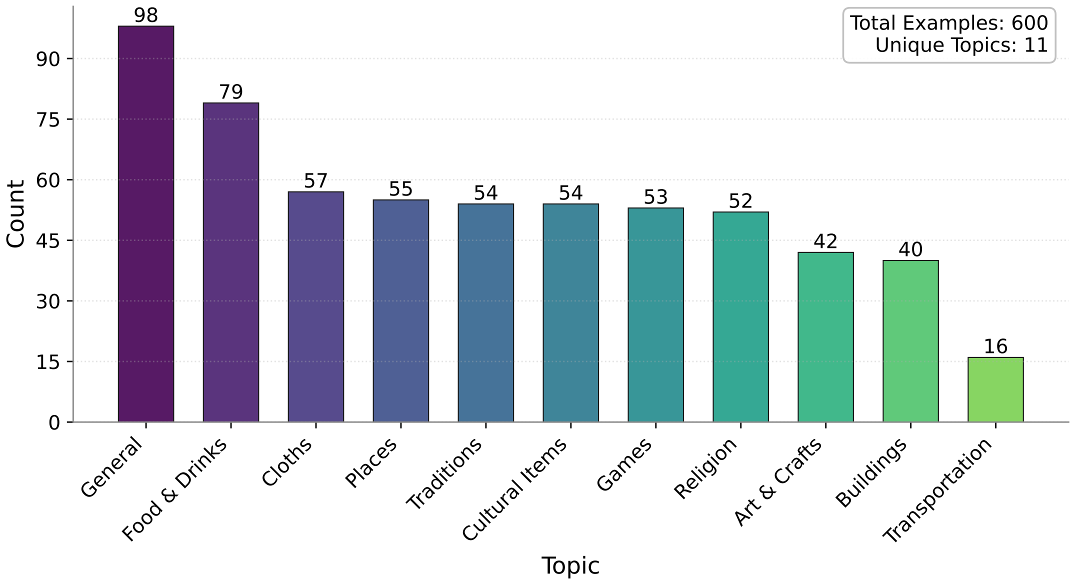


Figure 2: **Distribution of Topics:** Distribution of different topics in the ALPIQA dataset. The dataset contains 600 examples spanning 11 different topics.

also benchmark the dataset on different LLMs to measure the current state of this task.

In Figure 1, we highlight 6 different examples from each dialect in ALPIQA. Such examples vary in length and topic.

2 Related Work

In the literature, there has been some work on commonsense reasoning spanning English, Arabic, and multilingual datasets.

English Commonsense Reasoning One of the first datasets for commonsense reasoning was COPA (Roemmele et al., 2011). The dataset introduces examples for causal commonsense reasoning, where for each question, you have two plausible solutions. Commonsense QA (CSQA) (Talmor et al., 2019) is a dataset for commonsense question answering in English. It includes 12,247 multiple-choice questions and answers that discuss concepts and their relations. PIQA (Bisk et al., 2020) is one of the most popular datasets for commonsense reasoning in English. It is mostly for physical commonsense reasoning, where the examples deal with the real world and its physical items. COMM2SENSE (Singh et al., 2021) introduces a commonsense reasoning benchmark comprising natural language true/false statements, with each sample paired with its complementary counterpart, resulting in 4k sentences. The examples include different domains like physical, social, numerical, etc. SocialIQA (Sap et al., 2019) is another dataset for commonsense reasoning in English. It contains 38,000 multiple-choice questions for extracting emotional and social intelligence in a variety of everyday circumstances. It is considered a huge benchmark compared to other studies in the literature.

Multi-lingual Commonsense Reasoning There have been some studies for multilingual commonsense reasoning. XCOPA (Ponti et al., 2020) is a dataset constructed by translating the validation and test sets of COPA to 11 languages. mCSQA (Sakai et al., 2024) is a multilingual commonsense question answering dataset for cross-lingual transfer evaluation. It includes examples in 8 languages constructed with the help of language models. X-CSQA (Lin et al., 2021) is constructed by translating (Talmor et al., 2019) to 15 languages, including Arabic. This dataset can be used to measure cross-lingual understanding between multiple languages. All these benchmarks construct topologically diverse languages, but dialectal Arabic is not included in such benchmarks.

Arabic Commonsense Reasoning Compared to English, there are limited studies discussing commonsense reasoning in Arabic and its dialects. For example, ArabicCulture (Sadallah et al., 2025) introduced a dataset in commonsense reasoning spanning 13 countries in the Arabic world. It spans 12 domains with examples mainly in Modern Standard Arabic. On the other hand, MuDRiC (Elozeiri et al., 2025) introduced a dataset in multiple dialects by

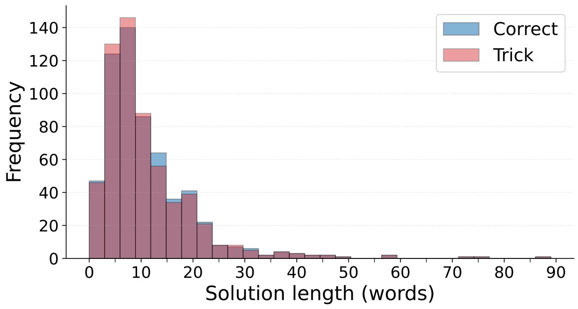


Figure 3: **Solution Length Distribution:** The distribution of the number of words for both the correct solution and the tricks.

using GPT-4o to translate 5,650 examples to Egyptian, Moroccan, Gulf, and Levant dialects. ArabicSense (Lamsiyah et al., 2025) is a benchmark for evaluating the world knowledge and commonsense reasoning of LLMs in Arabic. It contains examples in three tasks, which are commonsense validation, commonsense explanation(multiple choice), and commonsense explanation (generation).

Compared to previous literature, we see that ALPIQA is the first dialectal Arabic benchmark for physical commonsense reasoning spanning diverse dialects and topics.

3 Dataset Creation

Inspired by PIQA, we create 600 examples discussing physical commonsense reasoning in 6 different Arabic dialects: Egyptian, Yemeni, Tunisian, Saudi, Jordanian, and Lebanese. The dataset was constructed by native speakers who can read and write in that specific dialect. To give inspiration for adding examples, we included 11 topics that span different concepts in each country. The topics are: Food & drinks, Religion, Places, Transportation, Clothing, Cultural Items, Traditions, Art and Crafts, Buildings, Games, and General.

The authors participated in the collection and construction of the dataset. Each example was manually constructed and verified by a native speaker and it includes 6 columns: *Prompt*, *solution0*, *solution1*, *label*, *dialect*, and *topic*. We utilized the following instructions to create the examples in the dataset:

- The examples must be original and not taken from the original PIQA dataset.
- The examples must be in the target dialect or culturally relevant to the dialect.
- The examples must be manually curated by native speakers in each dialect.
- The examples must span different topics, including food, locations, religion, cultural items, clothing, etc.
- Avoid short examples that might be easy to answer by language models.

4 Statistics

ALPIQA contains 600 examples spanning 6 different dialects and 11 different topics.

Topics Distribution: The dataset spans a wide-range of topics for diversity and to cover different daily-life activities. Figure 2 shows the distribution of these topics. General questions are the most common with 98 questions, followed by Food & Drinks with 79 questions. Other topics such as Clothes, Places, and Traditions have a bit over 50 questions, while transportation has the lowest number of question of 16.

Solution Length: To be able to evaluate the impact of solution length on the performance of LLMs, we attempted to include answers with varying length that ranges from 0 to 90 words. Figure 3 shows the distribution of solution length for both the correct and wrong solutions (trick)

Solution Similarity: Following PIQA, we generated solutions with slight variations. Figure 4 illustrates the distribution of edit distances between paired solutions, computed using the *editdistance*[1] library. The figure shows that most solutions differ only marginally, with a mean edit distance of 10.6 and a maximum of 84.

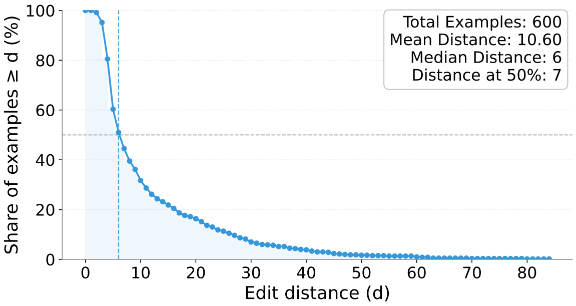


Figure 4: **Edit Distance Distribution:** A cumulative histogram of minimum edit distance between the two suggested answers. We made sure that the majority of answers vary slightly.

[1] https://pypi.org/project/editdistance/

Table 1: Performance across dialects. The average is calculated across the different dialects.

Model	Egyptian	Jordanian	Lebanese	Saudi	Tunisian	Yemeni	Average
Qwen3 0.6B	0.50	0.50	0.53	0.55	0.58	0.49	0.525
Qwen 2.5 3B	0.60	0.57	0.68	0.63	0.73	0.66	0.645
Gemma 2 9b	0.61	0.80	0.78	0.71	0.74	0.73	0.728
Gemma 3 27b	0.75	0.81	0.88	0.79	0.82	0.75	0.800
GPT 4.1	0.90	0.94	0.93	0.88	0.97	0.88	0.917

5 Evaluation

We evaluate our dataset using five different LLMs spanning different sizes, which are (GPT 4.1 (large), medium (Gemma 3 27b), small (Gemma 2 9b), nano (Owen 2.5 3B), tiny (Qwen 3 0.6B)). We choose the models that achieve the best scores in Arabic using the ABBL leaderboard (Ouda, 2025). We group the evaluation metrics using three main evaluation strategies:

Dialect In Table 1, we compare the results of all models grouped by dialect. Generally, we see smaller models like Qwen 3 achieves the worst performance with near-random results. On the other hand, GPT 4.1 achieves the highest results across all dialects. As expected, the performance increases across dialects when we increase the model size. As we increase the model size, the performance increases. Interestingly, GPT 4.1 achieves the worst results on the Saudi and Yemeni dialects, which are similar in structure.

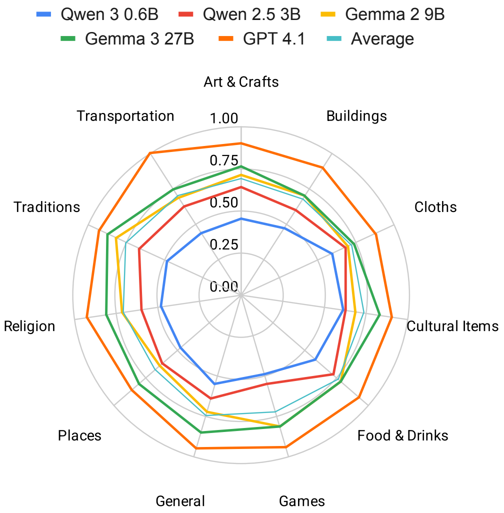


Figure 5: Comparison of all LLMs grouped by Topic.

Topic In Figure 5, we compare the performance of all models by considering the different topics. Again, GPT 4.1 achieves the highest results across all different topics. When considering individual topics, we see that the good & drinks topic achieves the highest average results across models. On the other hand, the buildings topic achieves the worst average results.

Length In Table 2, we compare the performance of all models in terms of prompt and solution length. We see the performance drops for longer prompts and solution pairs. The largest difference occurs with Gemma 3 27B, where we observe a decrease of 4.5 % when considering longer prompts.

Table 2: Results grouped by prompt + solution length (0–20 and +20 words) with averages.

Model	0–20	+20	Average
Qwen3 0.6B	0.536	0.514	0.525
Qwen 2.5 3B	0.656	0.623	0.640
Gemma 2 9b	0.740	0.719	0.730
Gemma 3 27b	0.821	0.777	0.799
GPT 4.1	0.922	0.911	0.917

6 Conclusion

In this paper, we introduced ALPIQA, a dataset for physical commonsense reasoning in dialectal Arabic. The dataset contains 600 examples, spanning six different dialects, which are Yemeni, Egyptian, Jordanian, Lebanese, Tunisian, and Saudi. The examples also include diverse topics, including Food & drinks, Religion, Places & Transportation, Clothing, Cultural Items, Traditions, Art and Crafts, Buildings, Games, and General. We benchmark the dataset across five different LLMs varying in sizes. Our results show that GPT 4.1 achieves the highest results with more than 10% gap compared to LLMs that are less than 27 billion parameters in size.

Acknowledgments

The research reported in this publication was supported by funding from King Abdullah University of Science and Technology (KAUST) - Center of Excellence for Generative AI, under award number 5940.

References

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, page 7432–7439.

Kareem Elozeiri, Mervat Abassy, Preslav Nakov, and Yuxia Wang. 2025. Mudric: Multi-dialect reasoning for arabic commonsense validation. (arXiv:2508.13130). ArXiv:2508.13130 [cs].

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)*, page 1–11, Abu Dhabi, UAE. Association for Computational Linguistics.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021. Common sense beyond english: Evaluating and improving multilingual language models for commonsense reasoning. (arXiv:2106.06937). ArXiv:2106.06937 [cs].

Karim Ouda. 2025. Abbl: An advanced benchmark and leaderboard for comprehensive evaluation of arabic language models.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. (arXiv:2005.00333). ArXiv:2005.00333 [cs].

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI spring symposium: logical formalizations of commonsense reasoning*, pages 90–95.

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. (arXiv:2502.12788). ArXiv:2502.12788 [cs].

Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2024. mcsqa: Multilingual commonsense reasoning dataset with unified creation strategy by language models and humans. (arXiv:2406.04215). ArXiv:2406.04215 [cs].

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. (arXiv:1904.09728). ArXiv:1904.09728 [cs].

Shikhar Singh, Nuan Wen, Yu Hou, Pegah Alipoormolabashi, Te-Lin Wu, Xuezhe Ma, and Nanyun Peng. 2021. Com2sense: A commonsense reasoning benchmark with complementary sentences. (arXiv:2106.00969). ArXiv:2106.00969 [cs].

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. (arXiv:1811.00937). ArXiv:1811.00937 [cs].