# ArabicSense: A Benchmark for Evaluating Commonsense Reasoning in Arabic with Large Language Models

**Salima Lamsiyah[1]*, Kamyar Zeinalipour[2]*, Samir El Amrany[1]*, Matthias Brust[1], Marco Maggini[2], Pascal Bouvry[1], Christoph Schommer[1],**

[1]Faculty of Science, Technology and Medicine (FSTM), University of Luxembourg
[2]University of Siena, DIISM, Via Roma 56, 53100 Siena, Italy

**Correspondence:** kamyar.zeinalipour2@unisi.it

## Abstract

Recent efforts in natural language processing (NLP) commonsense reasoning research have led to the development of numerous new datasets and benchmarks. However, these resources have predominantly been limited to English, leaving a gap in evaluating commonsense reasoning in other languages. In this paper, we introduce the ArabicSense Benchmark, which is designed to thoroughly evaluate the world-knowledge commonsense reasoning abilities of large language models (LLMs) in Arabic. This benchmark includes three main tasks: first, it tests whether a system can distinguish between natural language statements that make sense and those that do not; second, it requires a system to identify the most crucial reason why a nonsensical statement fails to make sense; and third, it involves generating explanations for why statements do not make sense. We evaluate several Arabic BERT-based models and causal LLMs on these tasks. Experimental results demonstrate improvements after fine-tuning on our dataset. For instance, AraBERT v2 achieved an 87% F1 score on the second task, while Gemma and Mistral-7b achieved F1 scores of 95.5% and 94.8%, respectively. For the generation task, LLaMA-3 achieved the best performance with a BERTScore F1 of 77.3%, closely followed by Mistral-7b at 77.1%. All codes and the benchmark is publicly available. [1] [2] [3] [4] [5]

## 1 Introduction

Commonsense reasoning (CSR) plays a critical role in understanding natural language. It involves making inferences based on commonsense knowledge, which encompasses general facts about the physical world and human behavior that people intuitively understand during communication. This implicit knowledge forms the backdrop for everyday conversations. Both humans and natural language processing (NLP) systems require CSR to comprehend the flow of daily events. Therefore, Commonsense reasoning remains a persistent challenge in artificial intelligence (AI) and natural language processing, in particular, evaluating and enhancing the commonsense reasoning capabilities of large language models (LLMs) is essential for advancing general natural language understanding (NLU) systems (Davis and Marcus, 2015).

Despite recent progress in creating commonsense reasoning benchmarks, most of them are available only in English (Davis, 2023), leaving a significant gap in resources for evaluating Arabic pre-trained language models. For example, the Arabic benchmark proposed by Al-Tawalbeh and Al-Smadi (2020) for commonsense validation and explanation is merely a translation of the English dataset from SemEval-2020's Commonsense Validation and Explanation (ComVE) task (Wang et al., 2019). Similarly, recent efforts by Beheitt and Ben HajHmida (2023) have focused on translating the Explanations for CommonsenseQA (Arabic-ECQA) and Open Mind Common Sense (Arabic-OMCS) datasets from English versions provided by IBM Research (Aggarwal et al., 2021). Thus, there is currently no dataset specifically developed from scratch for commonsense reasoning in Arabic. Indeed, translating English commonsense datasets into Arabic causes many challenges because direct translations often fail to capture cultural nuances and linguistic subtleties, resulting in inaccuracies

* Equal contribution

[1] https://github.com/EL-Amrany/Arabic-Commonsense-Reasoning

[2] https://huggingface.co/datasets/Kamyar-zeinalipour/ArabicSense

[3] https://huggingface.co/Kamyar-zeinalipour/Mistral-7b-CS-AR

[4] https://huggingface.co/Kamyar-zeinalipour/gemma2-9b-CS-AR

[5] https://huggingface.co/Kamyar-zeinalipour/P-Llama3-8B



*Proceedings of The 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, pages 1–11
January 20th, 2025. 

and a loss of contextual relevance. Additionally, the structural differences between the two languages further complicate accurate translation, undermining the effectiveness of the datasets for evaluating commonsense reasoning in Arabic.

Developing high-performance text classification models critically depends on the availability of high-quality training data. However, collecting and curating such data is often costly and time-consuming, particularly for specialized tasks that require domain-specific knowledge. To address this challenge, researchers have begun exploring the use of large language models (LLMs) to generate synthetic datasets as an alternative approach. In this paper, we leverage the capabilities of GPT-4 (Achiam et al., 2023) to create ArabicSense, a dataset specifically designed for Arabic commonsense reasoning. We focus on two natural language understanding tasks and one natural language generation task, which are detailed below. Illustrative examples of these tasks are provided in Figure 1.

- **Task A: Commonsense Validation** — The model is presented with two sentences ($S_1$ and $S_2$) that are similar in structure. The task is to identify which one of the two sentences does not make sense.

- **Task B: Commonsense Explanation (Multiple Choice)** — After identifying a nonsensical statement, the model is given three potential reasons ($r_1$, $r_2$, and $r_3$) explaining why the statement contradicts commonsense. The task is to select the correct reason. This assesses the model's understanding of the specific logical inconsistencies within the statement.

- **Task C: Commonsense Explanation (Generation)** — The model is required to generate a coherent explanation in natural language for why a given statement is against commonsense. The quality of the generated explanations is evaluated using BERTscore measure.

In our empirical study, we evaluate six BERT-based models — AraBERT (Antoun et al., 2020), ARBERT (Abdul-Mageed et al., 2021), MARBERT (Abdul-Mageed et al., 2021), CamelBERT[6], ArabicBERT (Safaya et al., 2020), and mBERT (Pires et al., 2019) — on the classification tasks described in Task A and Task B. Additionally, we

**Task A: Commonsense Validation**

*Which statement among the two is against commonsense?*

**S1:** الفهد يعد أسرع حيوان على وجه الأرض، إذ تصل سرعته إلى 120 كيلومترًا في الساعة في المسافات القصيرة

(**S1**: The cheetah is considered the fastest animal on Earth, as its speed can reach up to 120 kilometers per hour over short distances.)

**S2:** الفهود تستطيع الطيران لمسافات قصيرة باستخدام أجنحة صغيرة مخبأة تحت جلدها، مما يجعلها قادرة على الهروب من الضواري الأكبر حجمًا بسهولة

(**S2**: Cheetahs can fly short distances using small wings hidden under their skin, allowing them to easily escape from larger predators.)

**Task B: Commonsense Explanation (Multiple Choice)**

*Select the most corresponding reason why this statement is against common sense.*

**r1:** الفهود تفضل السباحة للهروب من الضواري لأنها تعتمد على براعتها في الماء

(**r1**: Cheetahs prefer swimming to escape from predators, as they rely on their agility in the water.)

**r2:** الفهود تستطيع الاختباء تحت الأرض بفضل جلودها المتغيرة اللون التي تجعلها غير مرئية

(**r2**: Cheetahs can hide underground thanks to their color-changing skin, which makes them invisible.)

**r3:** الفهود ليس لديها أجنحة ولا يمكنها الطيران؛ فهي تعتمد على سرعتها في الجري للهرب من الضواري

(**r3**: Cheetahs do not have wings and cannot fly; they rely on their running speed to escape predators.)

**Task B: Commonsense Explanation (Generation)**

*Generate the reason why this statement is against common sense.*

**r:** الفهود لا تملك القدرة على الطيران حيث تفتقر إلى الأجنحة، بل تعتمد على سرعتها العالية في الفرار من الكائنات المفترسة

**r:** (Cheetahs do not have the ability to fly as they lack wings; instead, they rely on their high speed to escape predators.)

Figure 1: Samples of our dataset

[6] https://github.com/CAMeL-Lab/CAMeLBERT

assess three state-of-the-art causal language models — Mistral-7b (Jiang et al., 2023), LLaMA-3 (Dubey et al., 2024), and Gemma[7] — using both zero-shot and fine-tuning approaches. The results demonstrate the effectiveness and quality of ArabicSense as a challenging commonsense reasoning benchmark for the Arabic language. Consequently, we present **ArabicSense** to the community as the first commonsense benchmark specifically designed to test commonsense world-knowledge and reasoning abilities of Arabic pre-trained language models.

The main contributions of this paper are summarized as follows:

- We present **ArabicSense**, the first commonsense reasoning benchmark developed specifically for the Arabic language.
- We develop three interrelated tasks to assess both natural language understanding and generation capabilities of pre-trained language models in Arabic commonsense reasoning.
- We leverage GPT-4 and prompting, to automatically generate high-quality synthetic data for commonsense reasoning in Arabic.
- We conduct a comprehensive evaluation of six BERT-based models and three state-of-the-art causal language models using zero-shot and fine-tuning approaches.

## 2 Related Work

**Commonsense Reasoning Benchmarks.** The NLP community has made significant progress in constructing commonsense datasets like ConceptNet (Speer et al., 2017) and ATOMIC (Hwang et al., 2021), as well as more specialized resources focused on physical (Bisk et al., 2020) and social commonsense (Sap et al., 2019). These resources have been widely incorporated into various downstream tasks (Lin et al., 2019; Guan et al., 2020; Liu et al., 2021) to assess AI's reasoning in commonsense scenarios. However, most of these benchmarks are English-centric, limiting their applicability for evaluating other languages (Davis, 2023).

Some Arabic benchmarks have been directly translated from English datasets (Al-Tawalbeh and Al-Smadi, 2020; Beheitt and Ben HajHmida, 2023). However, this approach often fails to capture the unique linguistic features and cultural nuances of the Arabic language, which are essential for accurate commonsense reasoning. Some studies have leveraged these translated datasets to evaluate the performance of pre-trained Arabic language models. For instance, Alshanik et al. (2023) explored commonsense validation and explanation through their participation in the SemEval 2020 Task 4, where their model achieved 84.7% accuracy in validation and a BLEU score of 24 for explanation generation. Finally, Khaled et al. (2023) conducted a comparative study on several Arabic BERT models for commonsense tasks, identifying ARBERTv2 as the top performer with 84.4% and 74.9% accuracy in validation and explanation tasks, respectively.

Despite initial efforts in Arabic commonsense reasoning, the field remains significantly underexplored compared to English-centric research. More work is needed to create dedicated datasets that capture the linguistic and cultural nuances of Arabic, making it essential to develop benchmarks specifically for evaluating Arabic commonsense reasoning.

**LLMs for Synthetic Data Generation.** Large language models (LLMs) are widely recognized for their strong generalization ability across a broad range of tasks (Achiam et al., 2023; Jiang et al., 2023; Dubey et al., 2024). However, optimizing these models for specific tasks remains a significant challenge. While zero-shot and few-shot prompting provide some level of flexibility (Dong et al., 2022), fine-tuning on task-specific data generally yields better results, particularly for specialized or out-of-domain tasks (Liu et al., 2022). Nonetheless, creating high-quality datasets is a time-consuming and resource-intensive process requiring specialized domain expertise. Synthetic data generation, which refers to artificially created data that replicates the characteristics of real-world data (Little, 1993), has emerged as a crucial solution for accelerating model training, particularly for small language models. It plays a significant role in all stages of training, including pre-training, instruction-tuning, and reinforcement learning from human feedback (Mitra et al., 2024).

A dataset is considered fully synthetic when the questions or instructions, the potential context, and the answers are all generated artificially. Examples of such methods include Self-Instruct (Wang et al., 2023), Unnatural Instructions (Honovich

[7] https://ai.google.dev/gemma/docs

et al., 2023), and Alpaca[8]. These models generate general-purpose synthetic data, while other approaches focus on task-specific fine-tuning by rephrasing existing datasets (Yin et al., 2023). A key limitation of fully synthetic data generation is the repetition and low quality of the generated samples. For example, Unnatural Instructions and Self-Instruct both reported significant redundancy in their data, with correctness rates around 54%-56.5%, while Alpaca's correctness rate was as low as 17%, making much of the data unsuitable for fine-tuning models. Indeed, partially synthetic data generation, which incorporates human-curated input, context, or output, helps improve data quality and diversity (Maini et al., 2024). However, these methods often depend on resource-intensive processes and limit task flexibility because of their reliance on human-generated components. In addition, inspired by self-instruct methods, several works have explored various languages, including Turkish, Arabic, English, and Italian. (Zeinalipour et al., 2024a; Zugarini et al., 2024; Zeinalipour et al., 2024c,b), Recently, Mitra et al. (2024) introduced AgentInstruct, a model that autonomously generates diverse, high-quality synthetic data from raw documents. It leverages powerful models like GPT-4 and tools such as search and code interpreters to create large-scale datasets tailored to both general and domain-specific skills, significantly improving the fine-tuning process. Inspired by AgentInstruct, we developed the first Arabic benchmark designed to evaluate commonsense reasoning in pre-trained Arabic language models.

## 3 ArabicSense: A New Benchmark Dataset

The aim of this work is twofold: to create a dataset for evaluating Arabic commonsense reasoning in LLMs and to improve their performance in this area. To achieve this, we generate diverse, high-quality data specifically designed for training LLMs in Arabic commonsense reasoning. This section outlines the methodology used to create the ArabicSense dataset, followed by the human validation process and an analysis of the dataset.

### 3.1 Methodology

The development of the ArabicSense dataset involves transforming unstructured seed data into three distinct tasks designed to assess various aspects of commonsense reasoning in Arabic: Commonsense Validation, Multiple-Choice Commonsense Explanation, and Generative Commonsense Explanation. We use the GPT-4 model to convert the seed data into diverse examples for each task. The following outlines the main steps used for building the dataset.

**Seed Data Collection.** We curated a collection of raw seed data exclusively from Arabic-language sources on Wikipedia [9]. The seed data covers a wide range of domains, including culture, geography, art, history, philosophy, and other relevant topics. Wikipedia is chosen for its diverse and extensive coverage of these subjects in Arabic, ensuring the dataset reflects varied contexts and knowledge areas essential for world-knowledge commonsense reasoning. More specifically, our data collection process began by extracting the opening sections of Arabic Wikipedia articles, with a specific emphasis on the bolded keywords found in the introduction. Alongside this keyword extraction, we also gathered vital metadata for each article, including details such as view counts, relevance scores, brief summaries, key headings, related terms, categorization information, and URLs.

**Transformation of Seed Data Using GPT-4**: To create the three tasks, we formulated specific prompts for each task and used GPT-4 (Achiam et al., 2023) to transform the seed data accordingly. Each task was generated with carefully crafted prompts that tailored the raw data into the required format, ensuring variety and depth in the examples.

- **Task A: Commonsense Validation** — The GPT-4 model was prompted to generate pairs of sentences ($S_1$ and $S_2$) that are similar in wording and structure. One of the sentences in each pair was logical, while the other was nonsensical, designed to challenge the model's commonsense reasoning ability.

- **Task B: Commonsense Explanation (Multiple Choice)** — After identifying the nonsensical sentence, GPT-4 was used to generate three possible reasons ($r_1$, $r_2$, and $r_3$), one of which was correct, explaining why the sentence contradicts commonsense. This task assesses the model's understanding of the specific logical inconsistencies in the sentence.

[8] https://github.com/tatsu-lab/stanford_alpaca

[9] https://en.wikipedia.org/wiki/Wikipedia:Lists_of_popular_pages_by_WikiProject

- **Task C: Commonsense Explanation (Generation)** — For this task, we prompt GPT-4 to generate a coherent explanation in natural language for why a given statement contradicts commonsense.

### 3.2 Refinement and Human Validation

The dataset was iteratively refined through human evaluations to ensure clarity, diversity, and quality across all three commonsense reasoning tasks. We assessed human performance on each task using three expert annotators who evaluated 200 random samples from each task. Our experts, who are native Arabic speakers and experienced NLP researchers, were not involved in the original data collection. Their expertise allows them to clarify misunderstandings in the annotation guidelines and produce more accurate and thoughtful annotations compared to crowd workers. The annotators were asked to rate each response using the following criteria:

- **RATING-A (Excellent):** The response is highly accurate, insightful, and completely relevant to the task. It shows a deep understanding of commonsense reasoning, providing a flawless and satisfying answer with no errors.
- **RATING-B (Good):** The response is generally correct and acceptable, but may contain minor errors, ambiguities, or imperfections. These issues do not significantly detract from the quality or overall validity of the response.
- **RATING-C (Adequate):** The response is relevant to the task but contains errors or oversights. While parts of the answer are valid, significant issues reduce its reliability, and it may veer off-topic in certain areas.
- **RATING-D (Poor):** The response is minimally relevant or partially incorrect. It may include some valid information but is weak in terms of commonsense reasoning. The answer may not fully address the task or be partially invalid.
- **RATING-E (Unacceptable):** The response is irrelevant, completely incorrect, or nonsensical. It fails to demonstrate an understanding of the task and does not provide a valid answer, possibly even contradicting commonsense knowledge.

The results revealed that 98% of the data across all tasks was rated as "A," demonstrating the exceptional quality of the proposed dataset. Furthermore, we measure the consistency of the review process with Fleiss's kappa[10], a statistical measure that evaluates inter-annotator agreement. Our expert annotators achieved a near-perfect Fleiss's kappa score, as shown in Table 1, demonstrating high reliability in the validation of the synthetic data. This high level of agreement highlights the robustness and effectiveness of our data generation method.

| Tasks | Fleiss's Kappa |
|---|---|
| Task A | 0.97% |
| Task B | 0.96% |
| Task C | 0.97% |

Table 1: Annotators agreement for the three tasks.

### 3.3 Dataset Analysis

The dataset used in this study is derived from Wikipedia articles, with commonsense statements extracted from sections of these articles. All views, word counts, and daily averages correspond to the statistics of these Wikipedia pages. The dataset for Task A includes 3954 training samples, 848 validation samples, and 848 test samples, with an average of 123,164 views per article and 217.40 words per sample, showing similar statistics across validation and test sets. Task B, which involves predicting the reason a statement is non-commonsensical, uses the same dataset sizes and maintains consistent statistics for views, word count, and daily averages. Task C, focused on generating explanations for nonsensical statements, follows the same size and structure as Task B, resulting in a balanced dataset across all tasks. Detailed statistical information for each task and split is presented in Table 2.

### 3.4 Experimental Setup

This study evaluates the performance of large language models for Arabic commonsense reasoning using the ArabicSense benchmark. The experimental setup involves two sets of models: BERT-based encoders (AraBERTv2 (Antoun et al., 2020), ARBERT (Abdul-Mageed et al., 2021), MARBERTv2, CamelBERT [11], ArabicBERT (base and large) (Safaya et al., 2020), and mBERT (Pires et al., 2019)) and three causal LLMs (Mistral-7b

[10] https://fr.wikipedia.org/wiki/Kappa_de_Fleiss

[11] https://github.com/CAMeL-Lab/CAMeLBERT

| Task | Split | Count | Mean Views | Mean Word Count |
|---|---|---|---|---|
| **Task A** | Train | 3954 | 123,164 | 217.40 |
| | Validation | 848 | 126,339 | 220.22 |
| | Test | 848 | 133,027 | 224.21 |
| **Task B** | Train | 3954 | 123,164 | 217.40 |
| | Validation | 848 | 126,339 | 220.22 |
| | Test | 848 | 133,027 | 224.21 |
| **Task C** | Train | 3954 | 123,164 | 217.40 |
| | Validation | 848 | 126,339 | 220.22 |
| | Test | 848 | 133,027 | 224.21 |

Table 2: Dataset Statistics for the Three Tasks. The statistics correspond to the original Wikipedia articles from which the commonsense statements were generated.

(Jiang et al., 2023), LLaMA-3 (Dubey et al., 2024), and Gemma[12]). The BERT-based encoders are evaluated on the first two tasks, while the causal LLMs are assessed across all three tasks. The detailed experimental setups for each task are described below.

For **Task A**, which involves binary classification to distinguish between commonsensical and nonsensical statements, all BERT-based models were fine-tuned using a batch size of 8, employing the AdamW optimizer (Loshchilov, 2017) with a learning rate of $2e^{-5}$. To prevent overfitting, dropout regularization (Srivastava et al., 2014) was applied with a rate of 0.1. Additionally, to ensure reproducibility, a fixed random seed of 42 was used across all models and random number generators (NumPy, PyTorch).

For **Task B**, models were tasked with multiclass classification, where they were required to identify the correct reason why a nonsensical statement deviates from commonsense. Similar to Task A, all BERT-based models were fine-tuned using a batch size of 8 and the AdamW optimizer with a learning rate of $2e^{-5}$. Input sequences consisted of three sentences, concatenated using the [SEP] token (Devlin, 2018) and tokenized using the AutoTokenizer from HuggingFace, with a maximum sequence length of 128 tokens. Regularization techniques, including dropout with a rate of 0.1, is applied to prevent overfitting.

In **Task C**, we evaluated the performance of LLMs to generate explanations for why nonsensical statements deviate from commonsense. The causal LLMs tested for this task included Mistral-7b, LLaMA-3, and Gemma. Fine-tuning was performed using two NVIDIA A6000 GPUs, each equipped with 48 GB of GPU memory, which was necessary to handle the large sequence lengths and computation requirements for this generation task. The models were fine-tuned for 4 epochs with a maximum sequence length of 1024 tokens. We applied a learning rate of $1e^{-4}$, combined with a cosine scheduler and a weight decay of $1e^{-4}$. To optimize memory usage, we utilized gradient accumulation over 4 steps, and employed techniques such as gradient checkpointing and flash attention. Additionally, we applied LoRA (Low-Rank Adaptation) (Hu et al., 2021) with a rank of 16 and an alpha of 32 to further enhance memory efficiency during training. The batch size for both training and evaluation was set to 8, and model checkpoints were saved at the end of each epoch for reproducibility and future evaluations.

For all tasks, early stopping (Prechelt, 2002) was used to monitor validation loss and prevent overfitting.

### 3.5 Evaluation Measures

For the classification tasks (**Task A** and **Task B**), we used accuracy, precision, recall, and F1-score to thoroughly assess the effectiveness of the models. For the text generation task (**Task C**), we evaluated both the fluency and semantic quality of the generated content using BERTScore (Zhang et al., 2019). It utilizes pre-trained transformer models to compare embeddings of the generated and reference texts, providing a more robust measure of semantic similarity.

### 3.6 Results

#### 3.6.1 Task A and Task B Evaluation Results

To verify the quality of the generated ArabicSense dataset, we designed a comprehensive eval-

[12] https://ai.google.dev/gemma/docs

uation strategy for the text classification tasks, Task A (Commonsense Validation) and Task B (Commonsense Explanation). The evaluation was performed in two phases, starting with BERT-based encoders and then extending to causal LLMs. In the first phase, we evaluated six pre-trained BERT-based language models—AraBERT v2, MarBERT, CamelBERT, ArabicBERT base, ArabicBERT large, and mBERT—on the dataset without fine-tuning. This initial phase assessed the baseline performance of these models, leveraging only their pre-trained knowledge. As shown in Table 3, the models struggled to perform well on both tasks, with accuracy scores for Task A ranging from 0.33 to 0.34 and Task B accuracy ranging from 0.32 to 0.36. Similarly, precision, recall, and F1 scores were generally low, indicating the difficulty these models faced in distinguishing between sensible and nonsensical sentences, as well as identifying logical inconsistencies in Task B.

In the second phase, we fine-tuned the same BERT-based models on the ArabicSense dataset to evaluate the impact of task-specific training. The results, as presented in Table 4, show improvements across all metrics for both tasks. For Task A, AraBERT v2 achieved the highest performance, with an accuracy, precision, recall, and F1 score of 0.87. Similarly, for Task B, AraBERT v2 also obtained an accuracy and F1 score of 0.83, closely followed by other models like ArabicBERT (base) and MarBERT, which achieved strong results across metrics. These findings demonstrate that fine-tuning improves the models' ability to perform commonsense reasoning in Arabic, validating the quality and effectiveness of the ArabicSense dataset.

Next, we extended our evaluation to causal large language models (LLMs), including Gemma, LLaMA-3, and Mistral-7b, testing their performance in both zero-shot and fine-tuned settings. In the zero-shot setting (Table 5), Gemma performed the best, achieving an F1 score of 0.867 for Task A and 0.921 for Task B. LLaMA-3 and Mistral-7b showed weaker performance on Task A, with F1 scores of 0.659 and 0.601, respectively, although they achieved moderate results on Task B, with F1 scores of 0.863 and 0.805. These results indicate that without fine-tuning, causal LLMs face challenges in handling Arabic commonsense reasoning tasks. After fine-tuning the causal LLMs on our dataset (Table 6), all models showed performance improvements. For instance, Gemma's F1 score increased to 0.947 for Task A and 0.944 for Task B, demonstrating its ability to handle complex reasoning after fine-tuning. Similarly, Mistral-7b, which initially performed poorly, achieved an F1 score of 0.948 for Task A and 0.934 for Task B. LLaMA-3 also showed marked improvement, reaching F1 scores of 0.945 for Task A and 0.930 for Task B. These results highlight the critical role of fine-tuning in enhancing the performance of LLMs for nuanced commonsense reasoning tasks in Arabic.

### 3.6.2 Task C Evaluation Results

Table 7 presents BERTscore results for Task C (Commonsense Explanation Generation) using both zero-shot learning and fine-tuning for three different causal models: Gemma, LLaMa-3, and Mistral-7b. The BERTscore results show that fine-tuning on the ArabicSense dataset improves the performance of all models—Gemma, LLaMa-3, and Mistral-7b—compared to zero-shot learning. Gemma, which had the lowest zero-shot F1 score (0.656), saw the most improvement after fine-tuning, with an F1 score increase to 0.759. Similarly, LLaMa-3 and Mistral-7b improved from 0.747 and 0.728 F1 scores to 0.773 and 0.771, respectively. This highlights that the ArabicSense dataset enhances the models' ability to generate coherent explanations for why statements are against commonsense, validating its effectiveness for Task C. Furthermore, these results confirm the importance of fine-tuning on task-specific datasets to achieve optimal performance, particularly for tasks that require a deeper understanding of logical relationships.

Overall, by comparing the performance of BERT-based models and causal LLMs before and after fine-tuning, we demonstrate the effectiveness of the ArabicSense dataset in enhancing model performance. The consistent improvement across both encoder-based and causal models highlights the robustness of our dataset for training models to handle commonsense reasoning challenges in Arabic.

## 4 Conclusion

In this paper, we introduced **ArabicSense**, the first comprehensive benchmark designed to evaluate the commonsense reasoning abilities of large language models (LLMs) in Arabic. Through the creation of three distinct tasks: commonsense validation (task A), commonsense explanation (task B), and commonsense explanation generation (task C), we

| Model without Fine-Tuning | Task A | | | | Task B | | | |
|---|---|---|---|---|---|---|---|---|
| | **Accuracy** | **Precision** | **Recall** | **F1** | **Accuracy** | **Precision** | **Recall** | **F1** |
| AraBERT v2 | 0.33 | 0.21 | 0.32 | 0.22 | 0.33 | 0.18 | 0.32 | 0.18 |
| MarBERT | 0.34 | 0.34 | 0.34 | 0.34 | 0.36 | 0.35 | 0.35 | 0.35 |
| CamelBERT | 0.33 | 0.33 | 0.33 | 0.33 | 0.34 | 0.22 | 0.34 | 0.21 |
| ArabicBERT (base) | 0.34 | 0.22 | 0.33 | 0.19 | 0.34 | 0.34 | 0.34 | 0.34 |
| ArabicBERT (large) | 0.34 | 0.11 | 0.33 | 0.17 | 0.33 | 0.14 | 0.32 | 0.16 |
| mBERT | 0.33 | 0.11 | 0.33 | 0.16 | 0.32 | 0.10 | 0.33 | 0.16 |

Table 3: Evaluation of the pretrained language models **without fine-tuning** on Tasks A and B.

| Model with Fine-Tuning | Task A | | | | Task B | | | |
|---|---|---|---|---|---|---|---|---|
| | **Accuracy** | **Precision** | **Recall** | **F1** | **Accuracy** | **Precision** | **Recall** | **F1** |
| AraBERT v2 | 0.87 | 0.86 | 0.87 | 0.87 | 0.83 | 0.83 | 0.83 | 0.83 |
| MarBERT | 0.81 | 0.78 | 0.85 | 0.82 | 0.83 | 0.83 | 0.83 | 0.83 |
| CamelBERT | 0.82 | 0.81 | 0.84 | 0.82 | 0.80 | 0.80 | 0.79 | 0.80 |
| ArabicBERT base | 0.84 | 0.82 | 0.87 | 0.85 | 0.81 | 0.82 | 0.81 | 0.81 |
| ArabicBERT large | 0.75 | 0.80 | 0.67 | 0.73 | 0.84 | 0.84 | 0.84 | 0.84 |
| mBERT | 0.75 | 0.72 | 0.84 | 0.77 | 0.76 | 0.76 | 0.75 | 0.76 |

Table 4: Evaluation of the pretrained language models **with fine-tuning** on Tasks A and B.

| Model with Zero-shot | Task A | | | | Task B | | | |
|---|---|---|---|---|---|---|---|---|
| | **Accuracy** | **Precision** | **Recall** | **F1** | **Accuracy** | **Precision** | **Recall** | **F1** |
| Gemma | 0.869 | 0.880 | 0.854 | 0.867 | 0.921 | 0.922 | 0.920 | 0.921 |
| LLama-3 | 0.690 | 0.733 | 0.598 | 0.659 | 0.863 | 0.865 | 0.860 | 0.863 |
| Mistral-7b | 0.523 | 0.517 | 0.718 | 0.601 | 0.805 | 0.804 | 0.806 | 0.805 |

Table 5: Comparison results of the Causal LLMs using **zero-shot** on Task A and Task B.

| Model with Fine-tuning | Task A | | | | Task B | | | |
|---|---|---|---|---|---|---|---|---|
| | **Accuracy** | **Precision** | **Recall** | **F1** | **Accuracy** | **Precision** | **Recall** | **F1** |
| Gemma | 0.947 | 0.948 | 0.946 | 0.947 | 0.944 | 0.944 | 0.944 | 0.944 |
| LLama-3 | 0.945 | 0.948 | 0.942 | 0.945 | 0.930 | 0.930 | 0.930 | 0.930 |
| Mistral-7b | 0.948 | 0.946 | 0.950 | 0.948 | 0.934 | 0.934 | 0.934 | 0.934 |

Table 6: Comparison results of the Causal LLMs after **fine-tuning** on Task A and Task B.

| Model | Zero-shot | | | Fine-tuning | | |
|---|---|---|---|---|---|---|
| | **Precision** | **Recall** | **F1** | **Precision** | **Recall** | **F1** |
| Gemma | 0.641 | 0.672 | 0.656 | 0.765 | 0.754 | 0.759 |
| LLama-3 | 0.733 | 0.763 | 0.747 | 0.774 | 0.773 | 0.773 |
| Mistral-7b | 0.735 | 0.722 | 0.728 | 0.768 | 0.774 | 0.771 |

Table 7: BERTscore results using **zero-shot learning** and **Fine Tuning** on Task C.

addressed the gap in commonsense reasoning resources available for Arabic. The dataset was generated using GPT-4 and refined through human validation, ensuring its quality and relevance to the Arabic language context.

Our empirical evaluations, conducted across six pre-trained Arabic BERT-based models and three state-of-the-art causal LLMs, clearly demonstrate that the models' performance improves after fine-tuning on our dataset. The results show that fine-tuning these models on ArabicSense enables them to handle the nuances of Arabic commonsense reasoning with good accuracy, precision, recall, and F1 scores. These findings confirm the utility and quality of ArabicSense as a benchmark for advancing research and model development in this domain. The codes and resources will be made publicly available to support further exploration and enhancement of Arabic common-sense reasoning tasks.

## 5 Limitations

Despite promising results, our study has several limitations. ArabicSense focuses on three specific tasks of commonsense reasoning, which may not cover the entire spectrum of commonsense knowledge. Commonsense reasoning encompasses a wide range of domains, and further expansions to include additional reasoning dimensions (e.g., causal or temporal reasoning) could enhance the benchmark's coverage. Additionally, while the dataset was generated using advanced models such as GPT-4 and validated by humans to ensure quality, it remains synthetic in nature. Synthetic data generation may introduce biases or fail to capture certain real-world nuances that naturally occurring datasets might better reflect. Future work could explore hybrid approaches that combine synthetic and real-world data to enhance the quality of the dataset.

## 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.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. Explanations for CommonsenseQA: New Dataset and Models. 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 3050–3065, Online. Association for Computational Linguistics.

Saja Al-Tawalbeh and Mohammad Al-Smadi. 2020. A benchmark arabic dataset for commonsense explanation. *arXiv preprint arXiv:2012.10251*.

Farah Alshanik, Ibrahim Al-Sharif, and Mohammad W Abdullah. 2023. Commonsense validation and explanation for arabic sentences. In *International Conference on Emerging Trends and Applications in Artificial Intelligence*, pages 101–112. Springer.

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.

Mohamed El Ghaly Beheitt and Moez Ben HajHmida. 2023. Generation of arabic commonsense explanations. In *Asian Conference on Intelligent Information and Database Systems*, pages 527–537. Springer.

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

Ernest Davis. 2023. Benchmarks for automated commonsense reasoning: A survey. *ACM Computing Surveys*, 56(4):1–41.

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

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. A knowledge-enhanced pretraining model for commonsense story generation. *Transactions of the Association for Computational Linguistics*, 8:93–108.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2023. Unnatural instructions: Tuning language models with (almost) no human labor. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14409–14428, Toronto, Canada. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

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 *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6384–6392.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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. IEEE.

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.

Roderick JA Little. 1993. Statistical analysis of masked data. *Journal of Official statistics*, 9(2):407.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965.

Ye Liu, Yao Wan, Lifang He, Hao Peng, and S Yu Philip. 2021. Kg-bart: Knowledge graph-augmented bart for generative commonsense reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6418–6425.

I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Pratyush Maini, Skyler Seto, Richard Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. 2024. Rephrasing the web: A recipe for compute and data-efficient language modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14044–14072, Bangkok, Thailand. Association for Computational Linguistics.

Arindam Mitra, Luciano Del Corro, Guoqing Zheng, Shweti Mahajan, Dany Rouhana, Andres Codas, Yadong Lu, Wei-ge Chen, Olga Vrousgos, Corby Rosset, et al. 2024. Agentinstruct: Toward generative teaching with agentic flows. *arXiv preprint arXiv:2407.03502*.

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.

Lutz Prechelt. 2002. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online). International Committee for Computational Linguistics.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. 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.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Cunxiang Wang, Shuailong Liang, Yue Zhang, Xiaonan Li, and Tian Gao. 2019. Does it make sense? and why? a pilot study for sense making and explanation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4020–4026, Florence, Italy. Association for Computational Linguistics.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Da Yin, Xiao Liu, Fan Yin, Ming Zhong, Hritik Bansal, Jiawei Han, and Kai-Wei Chang. 2023. Dynosaur: A dynamic growth paradigm for instruction-tuning data curation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4031–4047, Singapore. Association for Computational Linguistics.

Kamyar Zeinalipour, Achille Fusco, Asya Zanollo, Marco Maggini, and Marco Gori. 2024a. Harnessing llms for educational content-driven italian crossword generation. *arXiv preprint arXiv:2411.16936*.

Kamyar Zeinalipour, Yusuf Gökberk Keptiğ, Marco Maggini, and Marco Gori. 2024b. Automating turkish educational quiz generation using large language models. *arXiv preprint arXiv:2406.03397*.

Kamyar Zeinalipour, Yusuf Gökberk Keptiğ, Marco Maggini, Leonardo Rigutini, and Marco Gori. 2024c. A turkish educational crossword puzzle generator. In *International Conference on Artificial Intelligence in Education*, pages 226–233. Springer.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Andrea Zugarini, Kamyar Zeinalipour, Surya Sai Kadali, Marco Maggini, Marco Gori, and Leonardo Rigutini. 2024. Clue-instruct: Text-based clue generation for educational crossword puzzles. *arXiv preprint arXiv:2404.06186*.