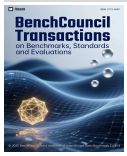


BenchCouncil Transactions on Benchmarks, Standards and Evaluations, 2026

DOI: https://doi.org/10.66834/byjwsd23
Research Article

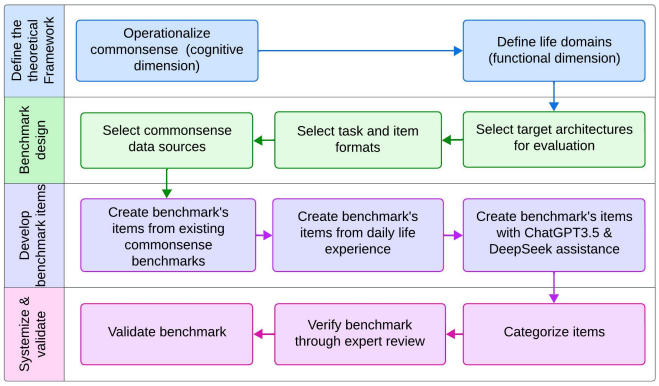
RESEARCH ARTICLE

JAMAL: A Multidimensional Benchmark for Arabic Commonsense Reasoning Across Life-Domains and Cognitive Axes

Basma Sayah,[1,*] Attia Nehar,[1,2] Hadda Cherroun,[1] Slimane Bellaouar[3] and Firoj Alam[4]

[1]Laboratoire d'Informatique et de Mathématiques (LIM), Amar Telidji University, Laghouat, Algeria , [2]Computer Science Department, Ziane Achour University of Djelfa, Djelfa, Algeria , [3]Laboratoire de Mathématiques et Sciences Appliquées (LMSA), Université de Ghardaia, Ghardaia, Algeria and [4]Qatar Computing Research Institute (QCRI), Hamad Bin Khalifa University, Doha, Qatar
*Corresponding author: b.sayah@lagh-univ.dz
Received on 10 January 2026; Accepted on 21 May 2026

Graphical Abstract



Highlights

- JAMAL, a novel language-agnostic commonsense framework instantiated as a concrete Arabic benchmark.
- Establishes a functional dimension covering 56 life-domain categories.
- Defines a cognitive dimension encompassing everyday situations, general knowledge, and problem-solving.
- Introduces a cultural grounding categorization of items into universal, western/global, and Arabic-specific knowledge for controlled analysis.
- Conducts a fine-grained evaluation of five language models across the three axes, revealing behavioral error patterns.
- Publicly releases the dataset to foster Arabic NLP development and support future multilingual extensions.

Abstract

Commonsense is a broad and multifaceted concept, making its evaluation a persistent challenge in natural language processing (NLP). This paper introduces **JAMAL** (Arabic for "Camel"), a multidimensional framework and benchmark for Arabic commonsense reasoning. JAMAL is structured along three complementary axes: (i) a life-domain axis comprising a taxonomy of 56 functional categories informed by the World Health Organization's International Classification of Functioning, Disability, and Health (ICF), capturing diverse aspects of daily human experience; (ii) a cognitive axis organizing commonsense into three reasoning types: everyday situations, general knowledge, and problem-solving; and (iii) a cultural grounding axis distinguishing between universal, western/global, and Arabic-specific knowledge. To operationalize this framework, benchmark items are constructed using psycholinguistically inspired principles of constrained contextual prediction. We evaluate five Arabic language models using JAMAL and observe consistent differences in their performance across all axes. Notably, FANAR-27B achieves the strongest overall results among all evaluated models, outperforming FANAR-9B and smaller baselines. Overall, **JAMAL** provides a structured and interpretable benchmark for evaluating commonsense reasoning in Arabic, supporting the development of more robust language models through systematic analysis of their behavioral limitations.

Key words: Commonsense Reasoning, Arabic NLP, Language Model Evaluation, Psycholinguistically Grounded Benchmarking, WHO-ICF Framework

1. Introduction

Neural language models have made remarkable strides in recent years, achieving strong performance across a wide range of NLP tasks, including text generation, sentiment analysis, and machine translation [1]. In particular, large-scale transformer-based models such as GPT, Llama, Qwen, and Gemini can generate fluent and coherent text [2–5].

As these models grow in capability and complexity, rigorous evaluation becomes increasingly important. Standard benchmark suites probe abilities such as commonsense reasoning, general knowledge, and reading comprehension, including HellaSWAG [6], MMLU [7], and RACE [8]. However, most of these benchmarks are designed primarily for English, limiting their direct applicability to other languages without adaptation.

In Arabic NLP, evaluation has advanced in recent years through general benchmarking efforts [9, 10], dedicated resources such as ArabicMMLU [11] and AraDiCE [12], and leaderboards such as BALSAM [13]. However, benchmarks targeting other specific capabilities remain limited [14].

Among these capabilities, commonsense reasoning is particularly important, as it underpins effective understanding and interaction with the world [15]. Despite its importance, existing Arabic commonsense evaluation resources remain limited in both scope and granularity. Early efforts, such as the translated *Is This Sentence Valid?* dataset [16], provide only coarse, task-level evaluation and do not capture the multifaceted nature of commonsense reasoning.

At the same time, the emergence of Arabic large language models—including Jais [17], ALLaM [18], and Fanar [19] has increased the need for more structured and diagnostic evaluation frameworks. These models differ in training data composition, dialectal coverage, and intended use cases, and their capabilities are not fully reflected by existing general-purpose benchmarks. This motivates the need for structured evaluation frameworks that provide deeper insight into model behavior.

To address these challenges, we introduce **JAMAL**[1], a multidimensional framework and Arabic benchmark for evaluating commonsense reasoning in language models.

JAMAL is structured along three complementary axes. The first is a life-domain axis comprising 56 functional categories informed by the World Health Organization's International Classification of Functioning, Disability, and Health (ICF) [20], enabling systematic coverage of everyday human experiences. The second is a cognitive axis that organizes commonsense into three reasoning types: everyday situations, general knowledge, and problem-solving. The third is a cultural grounding axis distinguishing between universal, western/global, and Arabic-specific knowledge.

Together, these axes enable fine-grained and interpretable evaluation of model performance across functional domains, reasoning types, and cultural contexts.

This work makes four contributions: (1) a language-agnostic multidimensional framework for commonsense evaluation; (2) an Arabic benchmark grounded in functional, cognitive, and cultural axes; (3) a comprehensive empirical analysis of model behavior across these dimensions; and (4) diagnostic insights revealing systematic strengths and weaknesses across five Arabic language models.

[1] The name *JAMAL* is a wordplay: in Arabic, *jamal* means "Camel," a culturally salient symbol, and it is also phonetically related to *jumal* ("sentences"), reflecting the benchmark's sentence-based design.

The remainder of this paper is organized as follows. Section 2 reviews related work on commonsense reasoning evaluation, covering global benchmarks, multilingual efforts, and Arabic-specific resources. Section 3 describes the design and construction of JAMAL. Section 4 presents the evaluation of five Arabic language models using JAMAL. Finally, Section 5 discusses conclusions, limitations, and future directions.

2. Related Work

This section reviews the evolution of global and Arabic commonsense evaluation benchmarks to situate the proposed JAMAL within the existing literature.

2.1. Global trends in commonsense evaluation

At the global level, commonsense reasoning (CSR) evaluation has evolved through a series of influential benchmarks, reflecting a shift from performance-oriented assessment toward more structured and diverse methodologies. Early large-scale datasets, such as SWAG (2018) [21], CommonsenseQA (2019) [22], and HellaSwag (2019) [6], predominantly adopt multiple-choice question answering formats, where commonsense is treated as a single unified capability and evaluated using accuracy-based metrics.

In the following years, evaluation was extended to more specialized reasoning domains. Social IQA (2019) [23] focuses on social and emotional reasoning, while PIQA (2020) [24] targets physical reasoning about everyday object use, material properties, and affordances. This shift was motivated by the need to better capture different facets of commonsense beyond general-purpose reasoning. However, most benchmarks still rely on answer selection as the primary evaluation paradigm.

In parallel, multilingual benchmarks were introduced to study cross-lingual transfer of commonsense knowledge. X-CODAH (2019) [25] and X-CSQA (2021) [26] extend existing datasets to multiple languages, typically via translation-based approaches, enabling evaluation of multilingual generalization.

More recently, research has moved toward more structured and interpretable evaluation frameworks. TG-CSR (2023) [27] introduces a theory-driven approach that decomposes commonsense reasoning into nine interpretable dimensions, such as temporal, spatial, and causal reasoning, enabling more fine-grained analysis of model behavior.

A further trend is the shift toward generative and reasoning-centric evaluation. For example, ExplaGraphs (2021) [28] requires models to generate structured explanations to justify predictions, while emerging frameworks such as SCoRE (2025) [29] emphasize multi-hop reasoning and reasoning-chain evaluation. Overall, these developments reflect a movement from static accuracy-based benchmarks toward more interpretable and reasoning-aware evaluation paradigms.

Despite these advances, a tension remains between scalability and interpretability: large-scale benchmarks offer broad coverage and simple evaluation, while structured approaches provide deeper insights but are less scalable.

2.2. Commonsense evaluation for Arabic language

Research on Arabic commonsense reasoning has produced a growing number of benchmarks, establishing important foundations for Arabic-centric evaluation while also revealing challenges in achieving structured and fine-grained assessment.

Early benchmarks, such as the Arabic Commonsense Dataset (ArCD) (2019) [30], Is This Sentence Valid?

Benchmark	Year	Lang.	Methodology	Taxonomy / Structure	Evaluation
SWAG [21]	2018	EN	MCQ (sentence completion)	None	Accuracy
CSQA [22]	2019	EN	MCQ (knowledge-based)	ConceptNet relations	Accuracy
HellaSwag [6]	2019	EN	Adversarial MCQ	None	Accuracy
Social IQA [23]	2019	EN	MCQ (social reasoning)	Social scenarios	Accuracy
PIQA [24]	2020	EN	MCQ (physical reasoning)	Physical interactions	Accuracy
X-CODAH [25]	2019	Multi	Multilingual MCQ	Translated dataset	Accuracy
X-CSQA [26]	2021	Multi	Multilingual MCQ	Translated taxonomy	Accuracy
ExplaGraphs [28]	2021	EN	Generative explanation	Argument graphs	NLI + explanation quality + human eval
TG-CSR [27]	2023	EN	Theory-driven MCQ	9 reasoning dimensions	Accuracy per dimension
SCoRE [29]	2025	EN	Multi-hop reasoning	Scenario-based logic	CoT audit + human eval
ArCD [30]	2019	AR	MCQ (Wikipedia-based)	None	Accuracy
Is This Sentence Valid? [16]	2020	AR	Sentence classification	Binary validity	Accuracy
Arabic Winograd [31]	2020	AR	Coreference resolution	Pronoun resolution	Accuracy
ArabicSense [32]	2025	AR	MCQ + generation	Implicit / synthetic	Accuracy, F1, BERTScore
Commonsense in Arab Culture [33]	2025	AR	MCQ + completion	12-domain taxonomy	Accuracy
JAMAL	**2026**	**AR**	**CPARG-based cloze-style text completion (psycholinguistically motivated); hybrid construction (manual curation + semi-automatic generation)**	**Three-axis structure: cognitive (3 reasoning types), life-domain (56 categories), and cultural grounding (3 categories)**	**Accuracy across axes**

Table 1. Comparison of existing global and Arabic commonsense reasoning benchmarks across language, methodology, taxonomy, and evaluation metrics, including the proposed JAMAL benchmark.

(2020) [16], and the Arabic Winograd Dataset (2020) [31], introduced initial testbeds for Arabic commonsense evaluation. These datasets mainly rely on multiple-choice or classification formats and treat commonsense as a general and undifferentiated capability.

More recent work has attempted to enrich both task design and evaluation objectives. ArabicSense (2025) [32] incorporates both classification and explanation generation, using metrics such as accuracy, F1, and BERTScore, although it relies heavily on synthetic data. Commonsense Reasoning in Arab Culture (2025) [33] introduces a culturally grounded taxonomy covering 12 daily life domains and 54 subtopics, combining multiple-choice and sentence completion tasks to evaluate reasoning across Arab cultural contexts. While this improves coverage, the taxonomy is largely derived from corpus-driven topic modeling with manual refinement, rather than an explicit cognitive or functional framework.

To address these limitations, there is a growing need for theory-driven benchmarks that implement structured taxonomies grounded in established cognitive, psychological, or behavioural models. Such benchmarks should also ensure careful item construction to avoid synthetic artifacts or superficial lexical cues, instead capturing meaningful reasoning processes.

Motivated by these limitations, we introduce JAMAL, a language-agnostic framework designed around a taxonomy spanning functional, cognitive, and cultural dimensions. Built through a controlled human-in-the-loop construction process, JAMAL enables fine-grained evaluation of commonsense reasoning across multiple complementary axes, supporting more systematic analysis of model behavior than existing Arabic benchmarks. Table 1 summarizes the key characteristics of global and Arabic benchmarks and situates JAMAL within the broader evaluation landscape.

3. The JAMAL Benchmark: Design, Development and Validation

This section describes the design and construction of JAMAL, a language-agnostic framework for fine-grained, multidimensional commonsense evaluation, and its Arabic instantiation as a concrete benchmark dataset. The overall process is illustrated in Figure 1.

3.1. Define the theoretical framework

In this first step, we aim to establish a theoretical foundation for understanding commonsense, in order to evaluate it effectively.

3.1.1. Operationalize commonsense (Cognitive dimension)

Since the goal was to build an effective benchmark for evaluating the commonsense knowledge of language models and to assess the extent to which this knowledge is acquired during training, it was essential to establish a clear and precise definition of commonsense to delineate what constitutes commonsense and what does not. After reviewing several definitions, we adopted one that explicitly distinguishes commonsense from domain-specific expertise: *"Commonsense is practical good sense gained through life experience, not through specialized study"* [34].

To operationalize this concept, we identified key themes from the literature. Ilievski et al. [35] emphasize the dimension of everyday situational knowledge, which enables navigation of routine scenarios. In contrast, Lenat [36] and Whiting et al. [37] treat commonsense as a body of general factual knowledge about the world. Complementing these, Newell and Simon [38] link it to the problem-solving processes used to address everyday challenges.

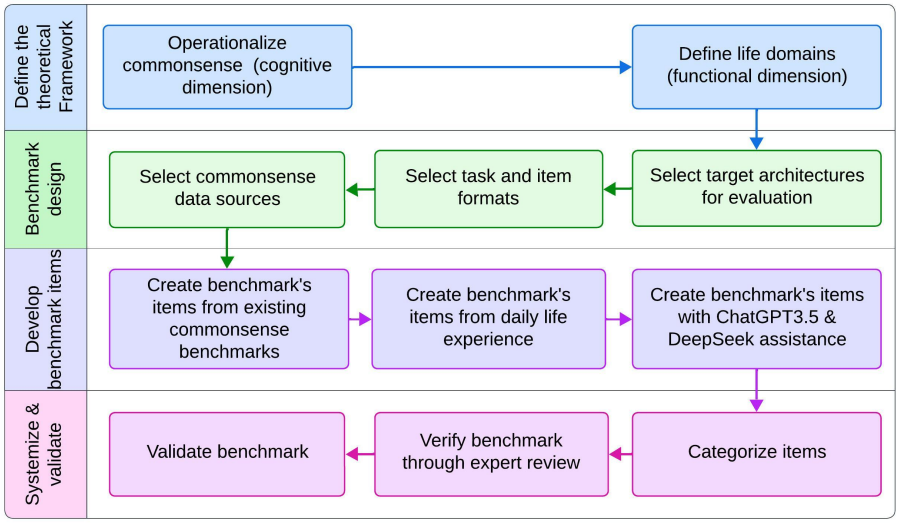


Figure 1. JAMAL Construction Process: From defining commonsense and its dimensions to benchmark validation.

Synthesizing these perspectives, we derived three core categories of commonsense: *(i)* everyday situations, *(ii)* general knowledge, and *(iii)* problem-solving.

The dimension of **everyday situations** encompasses the ability to understand typical scenarios, predict events based on prior experience [39], make sound judgments, and respond appropriately to common occurrences [40]. For example, understanding that people typically stand in a queue to wait for their turn or that one usually checks if an appliance is plugged in when it does not turn on. The **general knowledge** dimension refers to the factual and conceptual understanding that individuals typically possess, including basic facts, causal relationships, and general principles (e.g., "the sky is blue" [36]). This category also includes knowledge about physical objects and their properties, such as understanding that *an egg consists of a yolk, egg white, and shell* [37]. The **problem-solving** dimension involves applying logical reasoning and cognitive skills to solve structured tasks and address real-world challenges that require decision-making and pattern recognition [38]. This includes scenarios such as prioritizing which tasks to complete first when facing multiple deadlines. Furthermore, this dimension is shaped by interactions with both the environment and social contexts [41]. While these three categories capture the cognitive aspects of commonsense reasoning, its evaluation also requires grounding it in diverse areas of everyday life. To this end, we introduce a complementary life-domain dimension, described in the following subsection.

3.1.2. Define Life Domains (Functional Dimension)

We adopt the *International Classification of Functioning, Disability and Health (ICF)* as the conceptual foundation for defining the functional dimension of JAMAL, leveraging its comprehensive coverage of human activities, participation, and environmental context. However, our goal is not to reproduce the clinical taxonomy, but to derive a simplified and functionally meaningful set of life domains suitable for commonsense reasoning evaluation.

The taxonomy is derived by clustering related ICF components (primarily d- and e-codes) into higher-level functional domains that reflect how humans organize everyday experience. Rather than enforcing a one-to-one mapping between ICF codes and items, we aggregate multiple codes when they correspond to the same commonsense reasoning context.

For example, mobility-related activities (d4) inform *Transportation and Movement*, interpersonal interactions (d7) inform *Personal and Social Life*, and leisure-related activities (d920) are distributed across *Games, Sports, and Recreation*, *Arts, Music, and Creativity*, and *Fiction and Entertainment*. Environmental factors such as products, nature, and animals (e1, e2, e245) are integrated into domains describing physical and ecological context.

This procedure yields 15 top-level life domains covering social, physical, environmental, cognitive, occupational, and recreational aspects of human life, further decomposed into 56 subdomains (see Table A).

We introduce three deliberate simplifications relative to the original ICF: (i) aggregating multiple activity codes that correspond to the same everyday commonsense reasoning context into unified functional domains, (ii) elevating perceptual attributes such as *Colors, Shapes, and Forms* from sensory functions to a standalone commonsense domain, and (iii) separating goal-directed activities (*Human Actions and behaviours*) from institutional contexts (*Work and Productivity*) to better reflect differences in reasoning demands.

Overall, the resulting taxonomy preserves the breadth of the ICF while providing an evaluation-oriented organization of everyday human experiences for commonsense reasoning tasks.

3.2. JAMAL Design

This design phase translates the theoretical framework into practical specifications. We first identify the target language models for evaluation, which informs the selection of suitable benchmark item formats. We then select foundational knowledge sources aligned with the functional and cognitive dimensions of the framework.

3.2.1. Select target architectures for evaluation

Given the diversity of available architectures, we selected two main categories of language models for evaluation: **base language models** and **instruction-tuned models**.

Base language models are pre-trained on large, general-purpose corpora using objectives such as masked or next-token prediction. They typically require task-specific fine-tuning to achieve strong performance, as exemplified by models like BERT [42] and GPT [43].

instruction-tuned models are large language models that undergo additional fine-tuning to follow natural language instructions. This paradigm, popularized by models such as ChatGPT [2], enables zero-shot and few-shot task execution without the need for explicit task-specific fine-tuning.

3.2.2. Select task and item formats

We adopted the CPARG assessment format, a psycholinguistically grounded approach employed in *"What BERT Is Not?"* [44], which was originally introduced in human studies by Federmeier and Kutas [45]. In this format, each item presents a two-sentence context in which the task is to predict the final word of the second sentence. The task requires using implicit cues from the first sentence to infer this missing word through commonsense reasoning.

An example of a context in the CPARG format is shown below:

- The child blew out the candles. Everyone shouted happy ____.

In this example, the target word to be predicted is *birthday*. Solving this item relies on commonsense knowledge typically shared through a familiar social script of blowing out candles at a birthday celebration.

The CPARG format allows for the evaluation of the internal commonsense knowledge that language models acquire during training. Since the target word in each item is not explicitly stated, models cannot rely on simple text matching or selecting from predefined options. Instead, they must draw on internal reasoning and inference to predict the correct word.

To ensure proper alignment with the CPARG format, we established the following criteria for creating the benchmark items. In line with CPARG terminology, we refer to these items as *contexts* :

- Contexts must reflect commonsense rather than specialized knowledge, following the established definition 3.1.1.

  Example of specialized knowledge (invalid):

  Sentence 1: The patient's echocardiogram revealed severe mitral valve regurgitation.

  Sentence 2: The cardiologist decided the best course of action was to perform a [blank].

  Target word: *annuloplasty*
- Each context must consist of exactly two sentences.
- The target word to be predicted must not be explicitly mentioned in the context.

  Example of an explicitly mentioned target word in the first sentence (invalid):

  Sentence 1: He couldn't find his keys anywhere.

  Sentence 2: After searching for an hour, he finally found the missing [blank].

  Target word: *keys*
- The dataset must cover a diverse range of commonsense scenarios representing different aspects of daily life.

3.2.3. Select commonsense data sources

To construct commonsense items in the CPARG format, we required source material that could be reformulated into two-sentence contexts. We drew upon three main sources: existing commonsense benchmarks, scenarios inspired by our everyday experiences, and text generated by ChatGPT-3.5 and DeepSeek.

This multi-source approach was strategically chosen to leverage the unique advantages of each while mitigating their inherent biases. Existing benchmarks provide a validated foundation but risk inheriting cultural biases and data contamination from their original creation. Scenarios from daily life introduce crucial realism and a human perspective, yet they are limited in scalability and can reflect the subjectivity of the researchers' own experiences. Finally, LLM generated text ensures scalability and diversity but may amplify social biases present in the models' training data or produce superficially fluent but conceptually shallow items. By combining these sources and refining them manually, JAMAL aims to encompass a broad spectrum of commonsense knowledge, counterbalancing the weaknesses of any single source to create a more robust and nuanced evaluation set.

3.3. Develop benchmark items

In this section, we describe the process of creating and curating items from diverse sources in accordance with the prescribed CPARG format.

3.3.1. Create Benchmark's items from existing commonsense benchmarks

We searched for existing benchmarks using the three categories of the cognitive dimension of commonsense as keywords, focusing on assessments from both human and machine studies that incorporated these categories. For everyday situations, we identified the HellaSWAG [6] and Winogrande [46] benchmarks. HellaSWAG contains short narrative scripts drawn from video captions. Winogrande focuses on pronoun resolution in multiple-choice questions, testing comprehension of context and references. For general knowledge, we selected the "Is This Sentence Valid?" benchmark [16], which evaluates the ability to determine the factual and logical validity of statements. For problem-solving, we selected the Cornell Conditional Reasoning Test [47], a standardized assessment from cognitive psychology designed to evaluate logical and conditional reasoning in humans.Figure 2 illustrates the selected benchmarks.

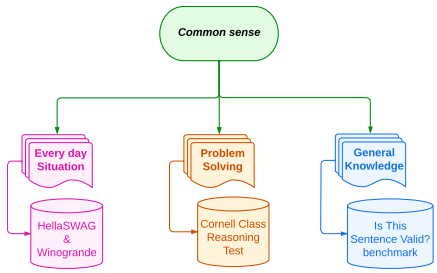


Figure 2. Overview of selected commonsense benchmarks categorized into three core cognitive dimensions: everyday Situations, general Knowledge, and problem Solving.

No.	Processing Stage	Context	السياق	Issue
1	Original	He crashed into a brick wall and fell to the ground. As he lifted his wrist, someone's shoe came down on it.	اصطدم بجدار من الطوب وانهار على الأرض. يرفع معصمه، ولكن حذاء شخص ما ينزل عليه.	Unusable context; no clear commonsense inference can be derived.
1	**Discarded**	**Discarded sentence**	**تم الاستبعاد**	**Rejection due to irrelevance.**
2	Original	The person lifts the violin to their chin and prepares. They play a song on the violin.	يرفع الشخص الكمان إلى ذقنه ويستعد. يعزف الشخص أغنية على الكمان.	The word "violin" can serve as the inference target.
2	**Refined context**	**Ahmad lifts his instrument to his chin and gets ready. He plays a sad song on the violin.**	يرفع أحمد آلته إلى ذقنه ويستعد. يقوم بعزف أغنية حزينة على الكمان.	**Removed the target word from the first sentence; added the emotional cue "sad" and the contextual cue "instrument".**
3	Original	A fiery ball throws a person backward. He crashes through a brick wall and lands on the ground.	كرة نارية ترمي شخصًا إلى الوراء. يتم دفعه عبر جدار من الطوب ويهبط على الأرض.	Unrealistic scenario; metaphorical description lacking grounded commonsense interpretation.
3	**Newly crafted**	**His eyes were swollen and bruised. He got into a fight yesterday and received a punch.**	كانت عيناه متورمتين وزرقاوين. دخل في شجار البارحة وتلقى لكمة.	**Inspired by physical harm, we replaced the unrealistic scenario with a grounded everyday situation involving a fight and injury.**

Table 2. Examples of items manually created or refined from existing commonsense benchmarks.

We translated the selected benchmark sentences into Arabic using a Python script and the Google Statistical Machine Translation (SMT) API[2]. Sentences were manually adapted to the CPARG format, discarding many captions, questions, and dialogues that did not reference a specific word or concept. Three main adaptation cases were observed:

- **Unusable sentences:** Could not be modified to CPARG format as no target word could be inferred from context (see sentence (1), Table 2).
- **Easily adaptable sentences:** Required minimal edits, such as removing the final word to designate it as the target (see sentence (2), Table 2).
- **Sentences requiring full rewriting:** Retained only the core idea; new sentences were crafted to fit the CPARG format (see sentence (3), Table 2).

A total of 600 items were adapted. The process was highly selective, except for the Cornell Critical Thinking Test, Level X, whose 65 items were included in full. These items assess reasoning and judgment through everyday scenarios and explicitly aim to evaluate "the ability to apply logical principles to everyday problems" [47], making them well-suited for commonsense evaluation.

3.3.2. Create benchmark's items from daily life experience

To expand the benchmark's scope and realism, we supplemented it with 400 new items inspired by our own everyday experiences. Many of these ideas emerged from reviewing existing benchmarks or from personal observations. Each item was carefully constructed to conform to the CPARG format, resulting in a benchmark of $1,000$ contexts. Examples of contexts from our daily life experience are shown in table 3:

3.3.3. Benchmark Item Construction using ChatGPT-3.5 and DeepSeek

We used two LLM-based generation strategies to construct CPARG-format contexts: (i) interactive prompting with ChatGPT-3.5, and (ii) an automated multi-step pipeline using the DeepSeek API.

ChatGPT-3.5 interactive generation:
We first generated CPARG-format contexts using ChatGPT-3.5 through iterative prompt engineering. We experimented with different prompting strategies, ranging from minimal prompts (providing only CPARG examples) to more explicit instructions describing the task, and finally to detailed step-by-step prompts encouraging constraint checking before generation. A representative prompt is:

> "Generate 10 Arabic two-sentence contexts: the first sentence provides a hint about a food item, and the second sentence ends with the target item as the final word. Ensure that the first sentence uniquely identifies the item and that the second sentence reveals the answer only in the final word."

This strategy was not fully error-proof, given the inherent difficulty of implicitly eliciting target words without explicit mention. Common issues included the target word appearing

[2] `https://cloud.google.com/translate`

No.	Example (Source)	Translation
1	كانت الطاولة مغطاة بالغبار. أحضرت ليلى قطعة قماش وبدأت بمسحها.	The table was covered in dust. Leila brought a cloth and started wiping.
2	كان عبد القادر وهشام يحبان التخييم. فور وصولهما إلى وجهتهما بدآ بجمع الحطب.	Abdelkader and Hisham loved camping. As soon as they arrived, they started collecting firewood.
3	كانت بسمة تعاني من ضعف في النظر. نصحتها أمها بأكل الجزر.	Basma suffered from poor eyesight. Her mother advised her to eat carrots.

Table 3. Examples of CPRAG-style contexts derived from daily-life experiences.

No.	Proceesing Stage	Context	السياق	Issue
1	Generated	I drew a new card and remembered the rules. There was no room for mistakes in UNO.	أخذت بطاقة جديدة وتذكرت القواعد. لم يكن هناك مجال للخطأ في اونو.	Too vague, applies to multiple card games.
1	**Refined**	**She looked at the colored cards and had to find a way to get rid of them quickly. There was no room for error in playing UNO.**	نظرت إلى البطاقات الملونة، كان عليها أن تجد طريقة للتخلص منها بسرعة. لم يكن هناك مجال للخطأ في لعب اونو.	**Added gameplay-specific cues (colored cards and card-elimination strategy) to better disambiguate the UNO context.**
2	Generated	He took his racket in hand, realizing that hitting with power and precision was the key to victory. The match was all about exchanging hits on the tennis court.	أخذ مضربه بيده، مدركاً أن الضرب بقوة ودقة هو مفتاح الفوز. كانت المباراة تدور حول تبادل الضربات على ملعب التنس.	Ambiguous terms: racket and court; could refer to other racket sports.
2	**Refined**	**The match was about exchanging the bright green ball with the racket. Every day, he would go to practice on the tennis court.**	كانت المباراة تدور حول تبادل الكرة ذات اللون الأخضر الفاقع بالمضرب. كان كل يوم يذهب للتدرب في ملعب التنس.	**Added a concrete visual cue (green tennis ball) and habitual practice context to better disambiguate the tennis scenario.**
3	Generated	When Ahmed's family prepares a healthy dinner, they like to add dark green leafy vegetables rich in minerals, like spinach.	عندما تعد عائلة أحمد وجبة عشاء صحية، فإنهم يحبون إضافة خضروات ذات أوراق خضراء داكنة وغنية بالمعادن، مثل السبانخ.	Fits multiple vegetables.
3	**Refined**	**Ahmed's family prepared a meal that Popeye always eats to become strong. But Ahmed did not like eating spinach.**	عندما أعدت عائلة أحمد وجبة يأكلها باباي دائماً ليصبح قوياً. لكن أحمد كان لا يحب تناول السبانخ.	**Added strong cultural cue (Popeye reference) to uniquely identify spinach as the intended vegetable.**

Table 4. Examples of CPARG-format contexts generated by ChatGPT-3.5 and their subsequent manual refinement.

before the final position in the sentence, ambiguous or underspecified hints (e.g., "she steamed this green vegetable. . . "), and reduced quality when generating multiple examples simultaneously. These limitations are consistent with known challenges in constraint adherence and hallucination in large language models [48].

All outputs were therefore manually reviewed, and non-compliant instances were corrected to ensure strict adherence to the CPARG format. Examples of this refinement process are shown in Table 4.

DeepSeek API pipeline generation:
To improve scalability and reduce manual interaction, we additionally used the DeepSeek API with an automated multi-step generation pipeline. This pipeline decomposed the task into three stages:

1. Generate target words belonging to functional categories (see Appendix A).
2. Generate short descriptive hint sentences for each target word.

No.	Processing Stage	Context	السياق	Issue
1	Generated	On the slopes of the mountains, tall green plants rose, famous for their sturdy wood. We took souvenir pictures next to the cedar tree.	على سفوح الجبال ارتفعت نباتات شاهقة خضراء طيلة العام، تشتهر بخشبها المتين. أخذنا صورًا تذكارية بجانب شجرة الأرز.	Could refer to other types of trees.
1	**Refined**	**On the slopes of the Lebanese mountains, tall green plants rose, famous for their sturdy wood. We took souvenir pictures next to the cedar tree.**	على سفوح جبال لبنان ارتفعت نباتات شاهقة خضراء طيلة العام، تشتهر بخشبها المتين. أخذنا صورًا تذكارية بجانب شجرة الأرز.	**Added a strong geographic cue (Lebanon), since the cedar tree is a national symbol of the country.**
2	Generated	I wanted to grow plants that tolerate salinity and give sweet fruits near the sea. The expert advised me to plant a palm tree.	أردت زراعة نباتات تتحمل الملوحة وتعطي ثمارًا حلوة بالقرب من البحر. نصحني الخبير بزراعة النخيل.	Ambiguous plant context, as it allows multiple interpretations of salt-tolerant plants.
2	**Refined**	**I wanted to produce dates rich in benefits. I have decided to plant a palm tree.**	أردت إنتاج التمر الغني بالفوائد. قررت زراعة النخيل.	**Added "dates" cue to disambiguate the target.**
3	Generated	In the fertile fields, tall plants were grown with stems full of a sweet-tasting liquid from which a natural sweetener is extracted. To produce natural sugar, The farmers planted a lot of sugarcane.	في الحقول الخصبة، كانت تزرع نباتات طويلة ذات سيقان مليئة بسائل حلو المذاق يستخرج منه المحلي الطبيعي. لانتاج السكر الطبيعي، زرع الفلاحون الكثير من قصب	CPRAG should rely on the full context rather than the second sentence alone.
3	**Refined**	**In the fertile fields, tall plants were grown with stems full of a sweet-tasting liquid from which a natural sweetener is extracted. The farmers planted a lot of sugarcane.**	في الحقول الخصبة، كانت تزرع نباتات طويلة ذات سيقان مليئة بسائل حلو المذاق يستخرج منه المحلي الطبيعي. زرع الفلاحون الكثير من قصب	**Removed the additional cue ("to produce natural sugar") to ensure that inference is based on the full context rather than the second sentence alone.**

Table 5. Examples of CPARG-style contexts generated by the DeepSeek pipeline, together with their subsequent manual refinement.

3. Generate an intermediate sentence that semantically links the hint sentence (first sentence) to the target word, which appears as the final word of the second sentence.

While this structured pipeline improved consistency, manual validation and correction were still required to ensure full compliance with CPARG constraints. Representative examples of refined outputs are shown in Table 5.

Overall, combining ChatGPT-3.5 and DeepSeek-based generation, we constructed a total of 1823 contexts.

It is worth noting that in English translations of the generated examples, the target word may not always appear in final position due to differences in syntactic structure. However, this does not affect the original Arabic instances used in the benchmark, where the target word is strictly constrained to appear as the final token in accordance with the CPARG format.

3.4. Systemize and validate

In this final phase, we systematize the benchmark by labeling each item according to its functional and cognitive category assignments. We then conduct human verification and validation of both the labels and the benchmark items to ensure the reliability and internal consistency of JAMAL.

3.4.1. Categorize items

To enable fine-grained evaluation, each item in JAMAL was labeled with one of the 56 life-domain categories from the functional dimension (Appendix A) and with one or more of the three cognitive branches (everyday situations, general knowledge, and problem-solving). Because the cognitive branches can overlap, items could receive multiple cognitive labels when appropriate. All automatically assigned labels were subsequently verified and corrected by human annotators.

Appendix B details the final distribution of items across categories. Across its 56 subdomains, JAMAL contains between 17 and 54 items per subcategory. This density aligns with established benchmark practices: BIG-bench tasks often contain 10–100 examples per task, CommonsenseQA and HellaSwag include roughly 15–50 examples per category, Social IQA contains approximately 15–50 questions per subcategory, and psycholinguistic diagnostic suites such as CPARG consist of 102 items distributed across all categories.

Each item was also annotated for cultural grounding to enable post-hoc analysis of model performance across different types of cultural knowledge. Items were labeled into three categories: *universal* (knowledge shared across cultures), *western/global* (concepts commonly encountered in mainstream media), and *Arabic-specific* (knowledge grounded in Arabic cultural contexts, including traditions and local practices), as detailed in Section 4.3.2.

3.4.2. Verify benchmark through expert review

The verification process consisted of two stages. First, one of the authors reviewed the items to ensure compliance with the benchmark design and corrected any inconsistencies. Second, an external native Arabic speaker independently evaluated the items and flagged contexts that did not meet the CPARG requirements.

The reviewer received detailed written instructions (see Appendix C), which included:

- Carefully examining each context in the benchmark.
- Identifying contexts that violated the established requirements.
- Providing justification for each flagged context.
- Noting ambiguous or unclear formulations.

During the verification process, 83 contexts (4.6%) were flagged as incorrect and categorized into four error types:

- **Target word already mentioned in the context (37 out of 1823):** This was the largest error category, comprising 44.6%of all errors.
- **Multiple possible target words (16 out of 1823):** This category includes cases where more than one target word could be inferred, comprising 19.3%of all errors.
- **Incorrect target word (27 out of 1823):** This category reflects mismatches between the context and the intended target word, comprising 32.5%of all errors.
- **Sentence-level errors (3 out of 1823):** This category includes cases of poor phrasing or ambiguity, comprising 3.6% of all errors.

Overall, the verification process confirmed the overall quality of the benchmark, and flagged items were corrected accordingly.

3.4.3. Validate benchmark

After verification, we validated JAMAL using a stratified 25% sample (455 items) drawn proportionally from all 56 functional categories. Two native Arabic speakers with different academic backgrounds independently evaluated the items, ensuring a diverse perspective since commonsense benchmarks target shared everyday reasoning rather than specialized knowledge. All reviewers followed the instructions detailed in Appendix D, assessing each item across four criteria: **Cloze Predictability**, **Agreement with Reference**, **Category Validation**, and **Inferential Consistency**. The structured evaluation form ensured a systematic and consistent assessment of item quality.

- **Cloze predictability:** This criterion measures whether the reviewer's predicted word exactly matches the gold target word. The first reviewer achieved exact matches for 403 items (88.57%). Allowing for morphological variants sharing the same root increases this to 408 items (89.67%), and including valid synonyms further raises it to 417 items (91.65%). (In the evaluation of language models in later sections, we account for exact matches, root-based variants, and synonym matches.)

  The remaining 38 responses (8.35%) were distributed as follows: 1 item (0.22%) was left blank (target word: *knitting*), primarily due to reviewer oversight; 27 items (5.93%) were incorrect due to inattention; and 10 items (2.20%) reflected plausible alternative completions, indicating potential ambiguity in item design.

  For the second reviewer, exact matches were achieved for 409 items (89.89%). Allowing for root-related variants increases this to 411 items (90.33%), and including synonyms raises this to 415 items (91.21%). The remaining 40 responses (8.79%) were distributed as follows: 4 items (0.88%) were left blank; 30 items (6.59%) were incorrect due to inattention; and 6 items (1.32%) reflected plausible alternative completions.
- **Agreement with reference:** In this criterion, reviewers evaluated whether the expected word represented the most plausible continuation of the sentence. The expected word was considered correct even if the reviewer predicted a synonym or root variant (accounted for in the evaluation metrics), made a minor error, or left the item blank. An item was marked incorrect only when the reviewer identified a different plausible completion that was not synonymous with the target word.

  The first reviewer judged 445 items as correct (97.80%), while the second reviewer judged 449 items as correct (98.68%), resulting in an inter-annotator agreement of 96.7%.
- **Category validation:** Only one category misclassification was identified by the second reviewer. The word *embroidery*, in this context, was incorrectly classified under *Patterns and Design* instead of *Crafts and Hobbies*. The corresponding item was:

  "Maryam brought a piece of fabric and colored threads to decorate her old clothes. She decided to learn the art of . . . " (*Crafts and Hobbies*).
- **Inferential consistency:** To test compliance with the CPARG format, both reviewers confirmed that the primary contextual cue is present in the first sentence and that successful inference requires the full context rather than only the second sentence. The target word is not explicitly mentioned, confirming that all items in the validation sample adhere to the CPARG format.

The validation results demonstrate strong inter-annotator agreement and support the semantic and categorical reliability of JAMAL.

3.4.4. Cross-cultural adaptation and migration pipeline

After the design, creation, and validation of JAMAL, we introduce a cross-cultural adaptation and migration pipeline to enable its extension to new cultural and linguistic settings. JAMAL is language-agnostic, as its structural organization across the functional, cognitive, and cultural axes is designed to generalize beyond Arabic. Adaptation to a new language or culture can therefore be achieved through three complementary strategies: translation of existing items, creation of new instances, and culture-specific generation of novel concepts, with all steps supported by human-in-the-loop validation.

Translation-based adaptation:

For cross-lingual transfer, automatic translation can be used to accelerate initial dataset construction. However, translation may alter syntactic structure and affect constraints such as the requirement that the target word appears in final position in the CPARG format. Therefore, all translated instances undergo post-verification to ensure compliance with the cloze structure and to preserve unambiguous target predictability.

Creation of new items:
When translation is insufficient or when extending JAMAL, new items are constructed directly following the CPARG format. This process consists of three steps:

1. Target word selection: Target words are selected for each category (e.g., food, music, clothing) using lexical and knowledge resources such as WordNet, ConceptNet, and BabelNet, as well as localized sources such as Wikidata and language model outputs. These resources provide culturally relevant candidate concepts.
2. Context construction: Two-sentence contexts are constructed around each target word. The first sentence introduces a contextual cue, while the second sentence is designed to end with the target word, which is removed during evaluation. This step can be automated using the pipeline described in Section 3.3.3, which generates CPARG-style contexts.
3. Human validation and refinement: Annotators verify that each instance satisfies CPARG structural constraints and is culturally appropriate for the target setting. When necessary, they refine wording or add minimal contextual cues to ensure naturalness and unambiguous target predictability.

Culture-specific adaptation:
The cultural axis of JAMAL requires additional care during adaptation, particularly for categories that contain culture-dependent knowledge. These include celebrations, films, arts and music. In such cases, adaptation is not strictly translational but involves functional cultural mapping, where concepts are replaced with culturally equivalent or culturally salient alternatives in the target setting.

For example, a celebration such as Eid may be mapped to Christmas or Diwali, a traditional dish to a locally equivalent food item, and a reference to a popular film to a culturally corresponding iconic movie. New culturally grounded concepts can also be introduced based on localized knowledge bases (e.g., Wikidata), native-speaker expertise, or culture-specific textual resources.

This strategy enables fine-grained adaptation not only across languages but also across dialects and regional varieties, supporting more localized and culturally sensitive evaluation settings.

4. Experiments and Discussion

We evaluated five Arabic language models on **JAMAL** using a Python script: AraGPT-base, MARBERT, AraBERT-large, FANAR-9B, and FANAR-27B.[3].

We evaluate model performance using four complementary metrics:

- **Exact match:** assigns a score of 1 if the predicted word exactly matches the gold target, and 0 otherwise.
- **Synonym match:** assigns a score of 1 if the predicted word is either identical to or a synonym of the target, based on Arabic WordNet, and 0 otherwise.
- **Same root:** uses the ISRI stemmer to determine whether the predicted and target words share the same morphological root.
- **Cosine similarity:** computes the semantic similarity between predicted and target words using FastText embeddings, yielding a score in $[0, 1]$.

Together, these metrics provide a multi-faceted evaluation of commonsense reasoning, capturing lexical accuracy, morphological similarity, and semantic relatedness.

4.1. Overall commonsense evaluation

As shown in Figure 3, same-root scores consistently exceed exact-match scores across all models, with AraGPT-base improving from 13.93% to 17.27%, MarBERT from 16.34% to 18.91%, AraBERT-large from 4.77% to 10.20%, FANAR-9B from 35.31% to 57.73%, and FANAR-27B from 62.88% to 72.53%. This pattern indicates that models frequently produce the correct lexical item with different conjugations rather than the exact target word.

Similarly, synonym-match scores exceed exact-match scores for all models: AraGPT-base rises from 13.93% to 17.87%, MarBERT from 16.34% to 19.30%, AraBERT-large from 4.77% to 10.86%, FANAR-9B from 35.31% to 58.77%, and FANAR-27B from 62.88% to 72.75%. This pattern is logical, as the synonym metric assigns a positive score when the model predicts either the exact word or one of its synonyms. Cosine similarity yields the highest scores overall, 37.32% for AraGPT-base, 42.99% for MarBERT, 26.67% for AraBERT-large, 61.85% for FANAR-9B, and 77.44% for FANAR-27B, suggesting that even when models do not output the expected word or a direct synonym, they often generate semantically related terms. These cases are analyzed further in subsection 4.2.

In model comparisons, the FANAR models achieve the highest performance across all metrics, with FANAR-27B substantially outperforming FANAR-9B, suggesting consistent benefits from scaling within the same architecture. Among smaller models, MARBERT (165M parameters) surpasses the larger AraBERT-large (370M parameters) across all metrics, and AraGPT-base (135M parameters) also outperforms AraBERT-large despite its smaller size.

These findings suggest that model size alone is not a reliable predictor of commonsense reasoning performance. Instead, differences in pretraining data, architectural design, and training objectives likely play an important role, although their individual contributions cannot be disentangled in our experiments.

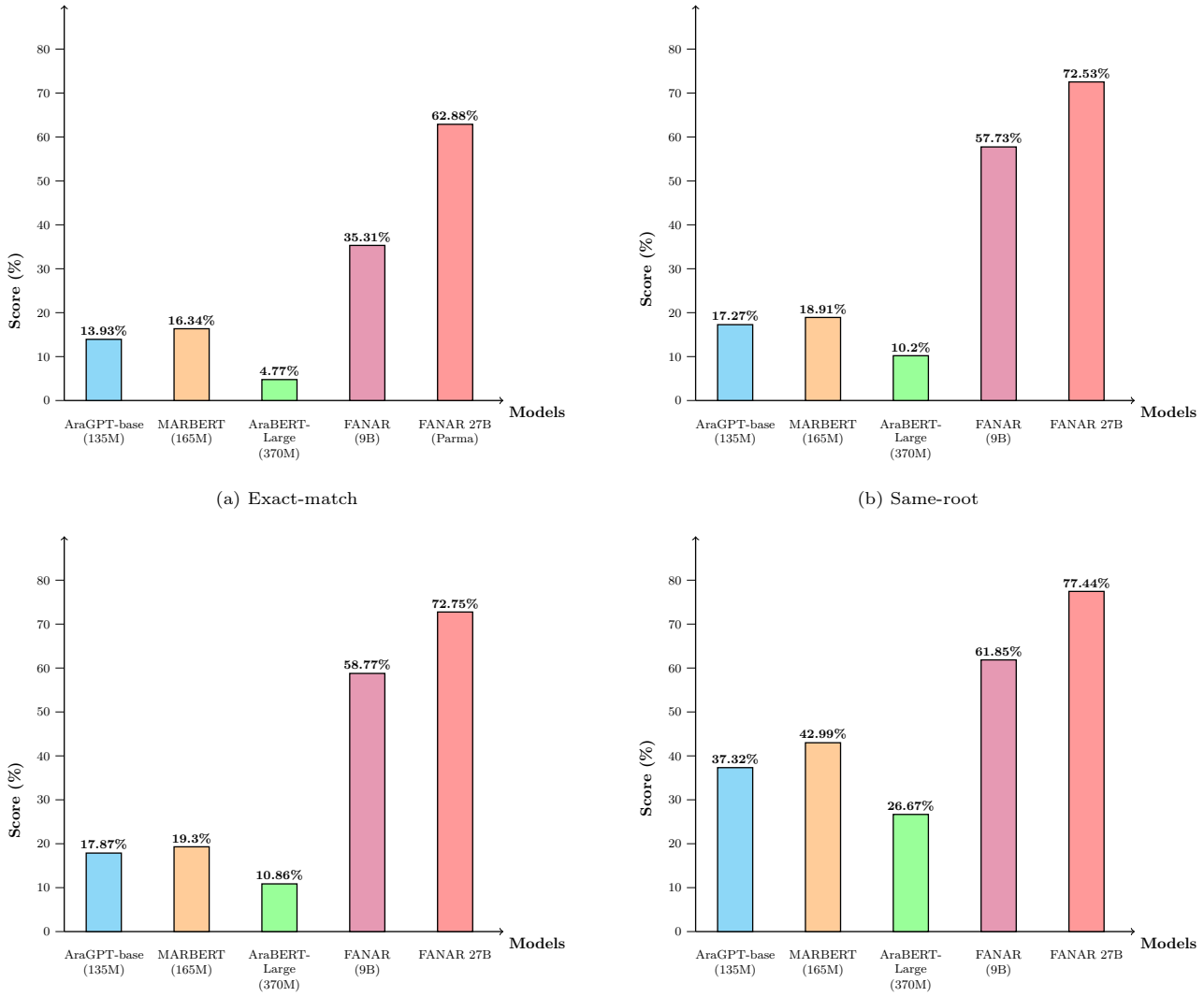
4.2. Error analysis

In this subsection, we present a qualitative analysis of model errors, followed by a theoretical interpretation of these behaviours grounded in prior research on transformer-based language models.

Error patterns:
We conducted a qualitative analysis of cases in which models received a score of zero under the Exact Match, Same Root, and Synonym Match metrics. Across AraGPT-base, MARBERT, AraBERT-large, FANAR-9B, and FANAR-27B, we identified several recurring error patterns. While not strictly mutually exclusive, these patterns capture distinct dimensions of model errors, ranging from semantic specificity and reasoning failures to cultural bias and syntactic interference.

- **Semantic neighbor substitution:** Models often predict semantically related concepts instead of the target word,

[3] The evaluation scripts and JAMAL are available at `https://github.com/BasmaSayah/An-ICF-guided-commonsense-benchmark`



(a) Exact-match

(b) Same-root

(c) Synonym-match

(d) Cosine similarity

Figure 3. Overall accuracy, same-root, synonym, and cosine similarity scores across evaluated models.

reflecting partial domain awareness. *Example:* For the context "When he traveled to Brazil, he saw a woman wearing a colorful dress moving her hands in a rapid rhythm. She was performing a [MASK]", the expected answer is *samba*. AraGPT-base predicts *the dancers*, MarBERT predicts *ballet*, and AraBERT-large predicts *hair*.

- **Hypernym substitution:** Specific items are replaced by broader categories, indicating a lack of precision. *Example:* "The melodies moved between different instruments in astonishing harmony. The audience listened attentively to the [MASK]." Expected: *symphony*. AraGPT-base predicts *song*, FANAR-27B predicts *playing*.
- **Hyponym confusion:** Models struggle to distinguish closely related hyponyms in specialized domains. *Example:* "She used a hooked needle and yarn to make a pillow cover. She learned the craft of [MASK]." Expected: *crochet*. FANAR-9B predicts *embroidery*, FANAR-27B predicts *tailoring*.
- **Causal overextension:** Models occasionally replace likely outcomes with extreme or abstract consequences. *Example:* "He walked on a thin rope above the circus. Everyone feared that he would [MASK]." Expected: *fall*. AraGPT-base predicts *die*.
- **Associative / Cultural bias:** Corpus-level or cultural associations can override contextual constraints. *Example:* "The boy listened to the song three times in a row. He was trying to memorize the [MASK]." Expected: *lyrics*. AraGPT-base, MarBERT, and FANAR-9B predict *the Quran*.
- **Context insensitivity / Syntactic priming:** Predictions may follow grammatical patterns while ignoring semantic fit. *Example:* "Ahmad raises his instrument to his chin and gets ready. He plays a sad song on the [MASK]." Expected: *violin*. AraGPT-base predicts *then*, FANAR-27B predicts *oud*.

These error patterns are consistent across all models, although they occur less frequently in FANAR-27B. Notably, high cosine similarity often reflects semantically related but

incorrect predictions rather than plausible completions, underscoring the limitations of embedding-based evaluation. Overall, the results highlight persistent weaknesses in lexical precision, causal reasoning, cultural awareness, and fine-grained semantic discrimination.

Interpretation of error patterns

To situate these findings within existing literature, we provide theoretical explanations grounded in prior work on transformer-based language models, highlighting interacting mechanisms involving semantic representation, data distribution effects, and syntactic , as well as decoding biases.

- **Semantic similarity errors (Semantic neighbor / Hypernym substitution)):** These errors arise when models select tokens that are semantically related to the target but differ in specificity. This includes both overgeneralized predictions (hypernyms) and contextual or associative neighbors. The behavior reflects the organization of semantic information in continuous embedding spaces, where related concepts are embedded in nearby regions [49, 50]. As a result, next-token prediction is influenced by competition among semantically similar candidates, often favoring higher-frequency or more generic alternatives with stronger prior probability [51].
- **Hyponym confusion:** This error type reflects difficulty in distinguishing closely related concepts within narrow or specialized domains. Models often capture the broader semantic field but fail to resolve fine-grained lexical distinctions between near-hyponyms. Prior work suggests that distributional representations compress fine-grained semantic differences into shared latent regions [6]. This effect is further amplified by the Zipfian distribution of language, where specific terms occur infrequently in training corpora [52]. Consequently, such terms receive fewer training signals, leading to weaker and less reliable representations [53].
- **Causal overextension:** This pattern reflects a tendency to generate salient or prototypical outcomes rather than contextually constrained causal consequences. Instead of explicitly modeling intermediate causal steps, models rely on high-probability continuations conditioned on surface context (i.e., shallow lexical and syntactic cues rather than deeper semantic or pragmatic understanding) [54]. This can result in overly extreme or exaggerated predictions when fine-grained causal constraints are not strongly represented in the training data.
- **Associative / Cultural bias errors:** These errors occur when strong corpus-level associations override local contextual constraints. Frequent co-occurrence patterns in pre-training data induce strong prior probabilities that may dominate contextual conditioning. As a result, culturally salient or high-frequency terms may be produced even when they are not contextually appropriate, reflecting well-documented bias propagation effects in large language models [55].
- **Context insensitivity / Syntactic priming:** These cases indicate reliance on surface-level syntactic or lexical patterns rather than deeper semantic integration. Models may produce locally fluent but globally inconsistent outputs when semantic constraints are weak or ambiguous. Prior studies show that transformer models often over-rely on local dependencies learned during training, at the expense of long-range semantic coherence [44].

Taken together, these theoretical perspectives suggest that the observed error patterns are not isolated failures but systematic behaviours arising from the interaction between embedding geometry, data distribution, and decoding biases. They also reflect a broader tension between reliance on high-probability training priors and the need for precise, contextually constrained reasoning.

4.3. Strengths and weaknesses of the models

Figures 4 and 5 show the root-match performance of AraGPT-base, MARBERT, AraBERT-large, FANAR-9B, and FANAR-27B across the 15 higher-level functional categories. Since the full set of 56 categories is too dense for visual presentation, we aggregate results at this higher level for clarity. The same figures also report performance across the three cognitive categories (everyday situations, general knowledge, and problem solving). Table 6 shows model performance across cultural scope categories (universal, mainstream, and Arabic).

4.3.1. Performance across functional and cognitive categories

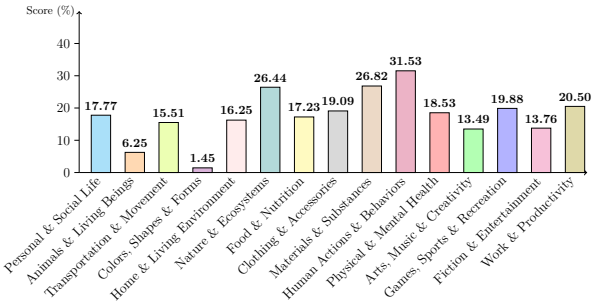
AraGPT-base demonstrates relatively balanced performance across the three commonsense branches (everyday situations, general knowledge, problem solving), suggesting uniform treatment of different reasoning types. At the functional level, it performs better in behavioural and environmentally grounded domains (e.g., Human Actions & behaviours, Nature & Ecosystems) but struggles in perceptual or visually grounded categories such as Colors and Shapes & Forms.

MarBERT achieves the strongest overall performance among the smaller models, particularly in human-centric and interaction-oriented domains (Personal & Social Life). This aligns with its pretraining on diverse social media text. However, like AraGPT-base, it underperforms in fine-grained perceptual categories. In the cognitive branches, MarBERT excels in Problem Solving compared to Everyday Situations and General Knowledge.

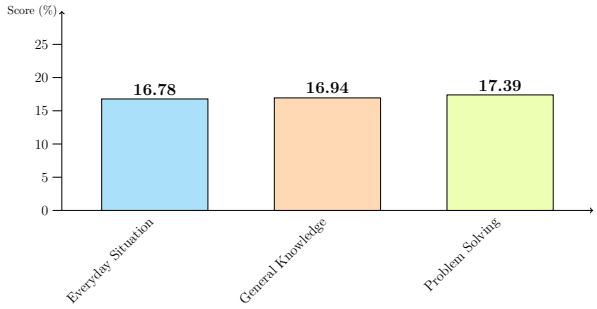
AraBERT-large shows lower performance across most categories despite being pretrained on the same dataset as AraGPT-base v2 [56]. AraBERT-large performs moderately in Materials & Substances, Games, Sports & Recreation, and Clothing & Accessories, and exhibits relatively better results in Problem Solving, suggesting that increased capacity may aid abstract reasoning.

Overall, these results confirm that training data, model architecture, pretraining objectives, and model size all influence commonsense reasoning. Models pretrained on socially rich and contextually diverse corpora, like MarBERT, exhibit stronger and more consistent performance, especially in human-centered domains.

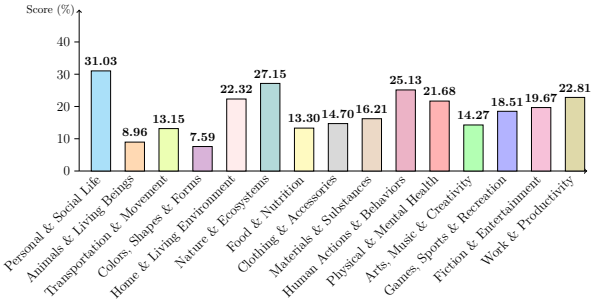
Figure 5 shows that FANAR-9B substantially outperforms smaller models across both cognitive and functional dimensions. It achieves balanced scores across the three cognitive categories: everyday situations, general knowledge, and problem solving; and demonstrates strong performance in human-centered and practical functional domains such as Food & Nutrition, Materials & Substances, Human Actions & behaviours, and Transportation & Movement. Perceptual and functional categories, such as Colors and Shapes & Forms, remain comparatively weaker, indicating room for improvement in fine-grained, concrete knowledge representation. Animals & Living Beings exhibits a comparable weakness.



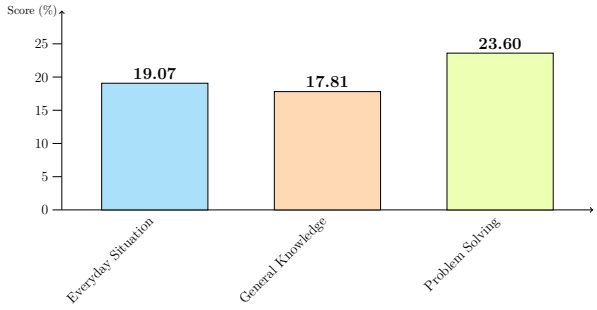
(A1) AraGPT-base root-match scores across life domains



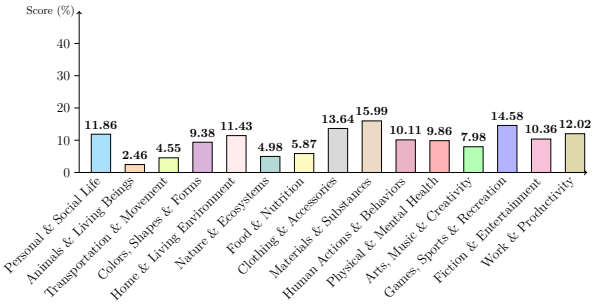
(A2) AraGPT-base root-match scores across cognitive dimension



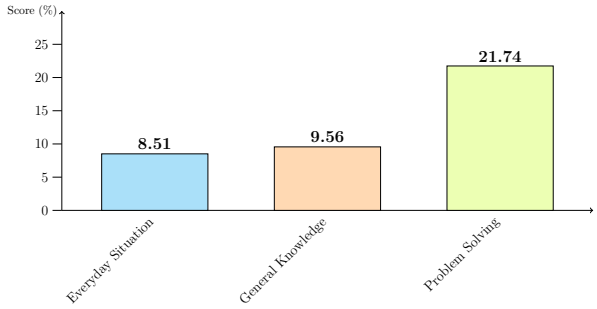
(B1) MarBERT root-match scores across life domains



(B2) MarBERT root-match scores across cognitive dimension



(C1) AraBERT-large root-match scores across life domains



(C2) AraBERT-large root-match scores across cognitive dimension

Figure 4. Strengths and weaknesses of AraGPT-base, MarBERT, and AraBERT-large.

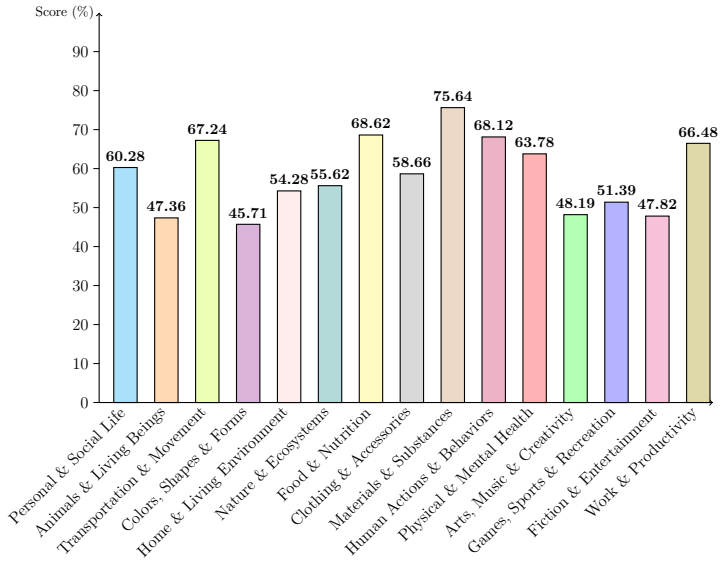
FANAR-27B consolidates the gains of FANAR-9B, outperforming it across nearly all functional and cognitive categories. Notable improvements are seen in visually grounded domains such as Colors, Shapes & Forms (+35%) , as well as in Animals & Living Beings (+21%), and Home & Living Environment (+23%). Cognitive categories also benefit, with Everyday Situations (+14%), General Knowledge (+15%), and Problem Solving (+11%) showing clear gains.

FANAR-27B achieves strong performance across functional categories and shows improved contextual integration compared to smaller models, with fewer extreme or irrelevant predictions. Its performance is closer to human cloze predictability in several visually grounded and everyday domains, suggesting improved semantic precision with increased model capacity. Despite these advances, the model still exhibits a tendency toward overgeneralization in some cases, and persistent gaps in problem solving indicate that challenges in higher-order reasoning remain.

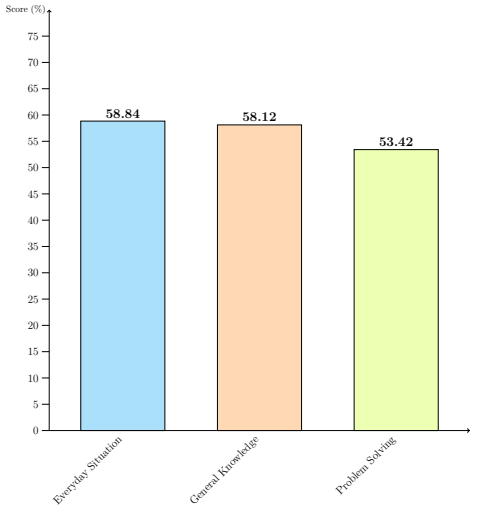
4.3.2. Performance analysis by cultural grounding

Building on the previous analysis, we examine model performance across the three cultural grounding categories: *Universal* (N = 1,615 items; knowledge shared across cultures, such as physical laws and daily routines), *Western/Global* (N = 78 items; mainstream concepts common in global media, such as karaoke or Monopoly), and *Arabic-specific* (N = 131 items; knowledge grounded in Arabic cultural contexts). This distribution reflects the natural skew in commonsense knowledge, where universal concepts are inherently more frequent than localized ones, thereby preserving realistic real-world frequencies rather than enforcing an artificially uniform design. The results, reported in Table 6 show that:

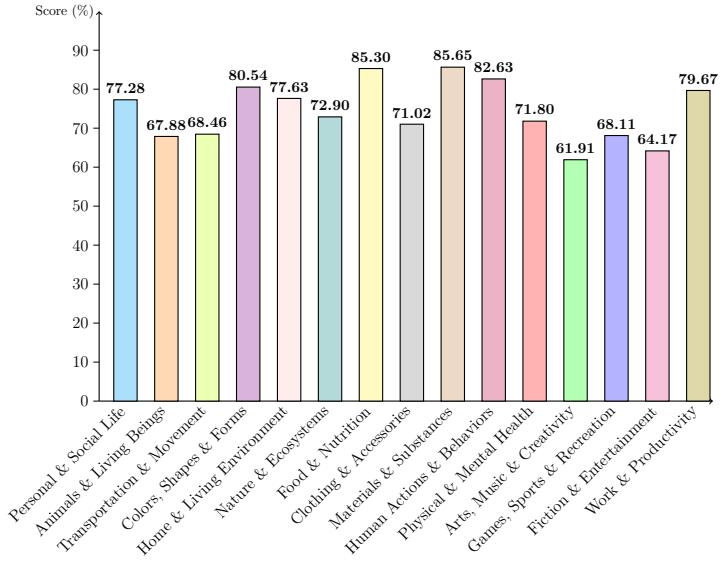
- Across all models, universal items consistently achieve higher performance than Arabic-specific items. For example, AraGPT-base achieves 14.49% on universal vs. 10.69% on Arabic, MarBERT 17.28% vs. 12.21%, AraBERT-large 4.95% vs. 2.29%, FANAR-9B 35.60% vs. 30.53%, and



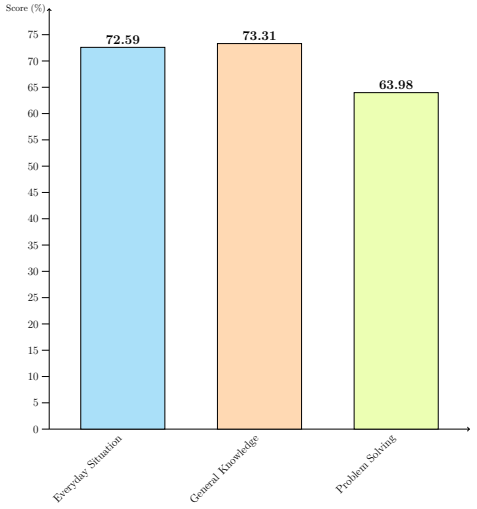
(A1) Fanar-9B root-match scores across life domains



(A2) Fanar-9B cognitive dimension scores



(B1) Fanar-27B root-match scores across life domains



(B2) Fanar-27B cognitive dimension scores

Figure 5. Strengths and weaknesses of Fanar-9B and Fanar-27B.

FANAR-27B 64.00% vs. 52.56%. This indicates a persistent performance gap between universal and culturally specific knowledge across all models.

- Performance on western/global items varies across models. AraGPT-base (7.69%) and MarBERT (3.85%) score lower on western items than on either Universal or Arabic-specific categories. In contrast, AraBERT-large (5.13%) and FANAR-9B (37.17%) achieve slightly higher scores on western items than on Universal items. FANAR-27B (54.96%) shows intermediate performance, with western scores lower than universal (64.00%) but higher than Arabic-specific items (52.56%). This variation may reflect differences in model capacity and exposure to culturally diverse content during training.
- Scaling from FANAR-9B to FANAR-27B yields higher absolute performance across all categories, with FANAR-27B reaching 64.00% (universal), 54.96% (western), and 52.56% (Arabic). However, the Universal–Arabic gap remains substantial, suggesting that while increased model capacity improves overall performance, disparities across cultural categories persist.

Overall, all models consistently show lower performance on Arabic-specific items than on universal items. This gap may reflect the comparatively limited representation of culturally specific knowledge in large-scale pretraining data rather than differences in item quality or ambiguity, particularly given the high human validation performance on these items.

Model	Overall	Universal (U)	Western (W)	Arabic (A)	Gap (U-A)	Gap (U-W)
AraGPT-base	13.93%	14.49%	7.69%	10.69%	3.80%	6.80%
MarBERT	16.34%	17.28%	3.85%	12.21%	5.06%	13.43%
AraBERT-large	4.77%	4.95%	5.13%	2.29%	2.66%	-0.17%
FANAR-9B	35.31%	35.60%	37.17%	30.53%	5.07%	-1.58%
FANAR-27B	62.88%	64.00%	54.96%	52.56%	9.06%	11.46%

Table 6. Performance comparison across cultural grounding categories. Overall accuracy is reported in the *Overall* column.

5. Conclusion

This study introduces **JAMAL**, a language-agnostic framework for evaluating commonsense reasoning in language models, alongside its Arabic instantiation as a benchmark. JAMAL adopts a three-axis taxonomy spanning a functional dimension (56 life-domain categories), a cognitive dimension (everyday situations, general knowledge, and problem-solving), and a cultural grounding axis (universal, western/global, and Arabic-specific knowledge). Together, these axes enable fine-grained, multi-dimensional evaluation that diagnoses model commonsense reasoning capabilities.

JAMAL is constructed through a multi-stage pipeline combining manual curation and LLM-assisted generation with human refinement. This hybrid design leverages the scalability of LLMs while maintaining quality through human verification. Its items follow the CPARG format [44], which imposes strict structural constraints and relies on subtle contextual cues to elicit commonsense inference.

The empirical evaluation of five Arabic language models reveals consistent performance gaps across functional, cognitive, and cultural dimensions, with FANAR-27B achieving the strongest overall results. Overall, JAMAL provides a structured and extensible benchmark for interpretable evaluation of commonsense reasoning, addressing a key gap in Arabic NLP and supporting future cross-lingual and culturally aware model assessment.

6. Limitations and Future Perspectives

While JAMAL provides a structured framework for evaluating contextual commonsense reasoning in Arabic language models, it is not exhaustive. Commonsense knowledge is broad, dynamic, and culturally situated. Future work could expand the benchmark with additional examples across functional life-domain categories and the problem-solving dimension, as well as include a wider range of culturally specific cases reflecting the diversity of Arabic-speaking communities.

The current benchmark primarily uses Modern Standard Arabic (MSA). Extending this work to Arabic dialects and more diverse linguistic settings is a natural direction for future research and would allow for a broader evaluation of model robustness across different forms of Arabic usage.

Although JAMAL has undergone manual verification, a full human performance baseline over the entire dataset remains an important direction for future work. Such a baseline would enable a more precise estimation of the human–model performance gap.

Finally, the CPARG format is a high-constraint cloze-based design in which each item presents a controlled two-sentence context with a single masked target. This reduces prompt variability and enables consistent assessment of contextual inference. However, it reflects a constrained form of language use rather than open-ended interaction, and thus captures only a subset of naturalistic language behavior.

Ethical Statement

No ethical approval was required for this study, as it did not involve human or animal subjects.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability Statements

The data supporting the findings of this study are openly available in `https://github.com/BasmaSayah/An-ICF-guided-commonsense-benchmark`.

Credit authorship contribution statement

Basma Sayah: Conceptualization, Methodology, Data curation, Validation, Investigation, Writing – Original Draft. Attia Nehar: Methodology, Investigation, Writing – Review & Editing. Hadda Cherroun: Methodology, Investigation, Writing – Review & Editing. Slimane Bellaouar: Methodology, Investigation, Writing – Review & Editing. Firoj Alam: Methodology, Investigation, Resources, Writing – Review & Editing.

A. JAMAL Functional Dimension Categories Adapted from the International Classification of Functioning (ICF)

Benchmark Domain	**Benchmark Subcategories**	**Mapped ICF Code and Component**
Personal & Social Life	Family & Friends, Celebrations, Emotion, Communication	d7 "Interpersonal interactions and relationships"
Animals & Living Beings	Pets, Wildlife, Farming Animals, Insects & Small Creatures	e245 "Animals"
Transportation & Movement	Land Transport, Water Transport, Air Transport, Walking & Running & Movement	d4 "Mobility"
Colors, Shapes, and Forms	Basic Colors, Geometrical Shapes, Patterns & Designs	b156 "Perception of visual stimuli"; b160 "Thought functions"
Home & Living Environment	Buildings & Structures, Rooms & Spaces, Household Objects, Household Tools & Appliances	e1 "Products and technology"
Nature & Ecosystems	Plants & Trees, Bodies of Water, Weather & Seasons, Landscapes	e2 "Natural environment and human-made changes to environment"
Food & Nutrition	Types of Food, Cooking & Eating, Farming & Gathering	d550 "Eating"; d630 "Preparing meals"
Clothing & Personal Accessories	Fabrics & Textiles, Types of Clothes, Footwear, Accessories	d540 "Dressing"; e1 "Products and technology"
Materials & Substances	Natural Materials, Manufactured Materials, Chemical Substances	e1 "Products and technology"; e2 "Natural environment and human-made changes to environment"
Human Actions & Behaviors	Learning & Education, Work Actions, Play Actions, Helping & Caring	d1 "Learning and applying knowledge"; d8 "Major life areas"
Physical & Mental Health	Physical Abilities, Emotional Well-being, Illnesses & Conditions, Health Maintenance	b1 "Mental functions"; b7 "Neuromusculoskeletal and movement-related functions"
Arts, Music, and Creativity	Visual Arts, Performing Arts, Music, Crafts & Hobbies	d920 "Recreation and leisure"; b160 "Thought functions"
Games, Sports, and Recreation	Team Sports, Individual Sports, Board Games, Recreational Activities (e.g., hide and seek, tag, playground games)	d920 "Recreation and leisure"
Fiction & Entertainment	Stories & Books, Films & TV Shows, Fantasy & Myths	d920 "Recreation and leisure"; b160 "Thought functions"
Work & Productivity	Jobs & Occupations, Workplaces, Workplace Tools & Machinery, Business	d850 "Remunerative employment"; e1 "Products and technology"

Table 7. ICF-based functional dimension categories used in the JAMAL benchmark construction.

B. Distribution of JAMAL items across the 56 functional categories

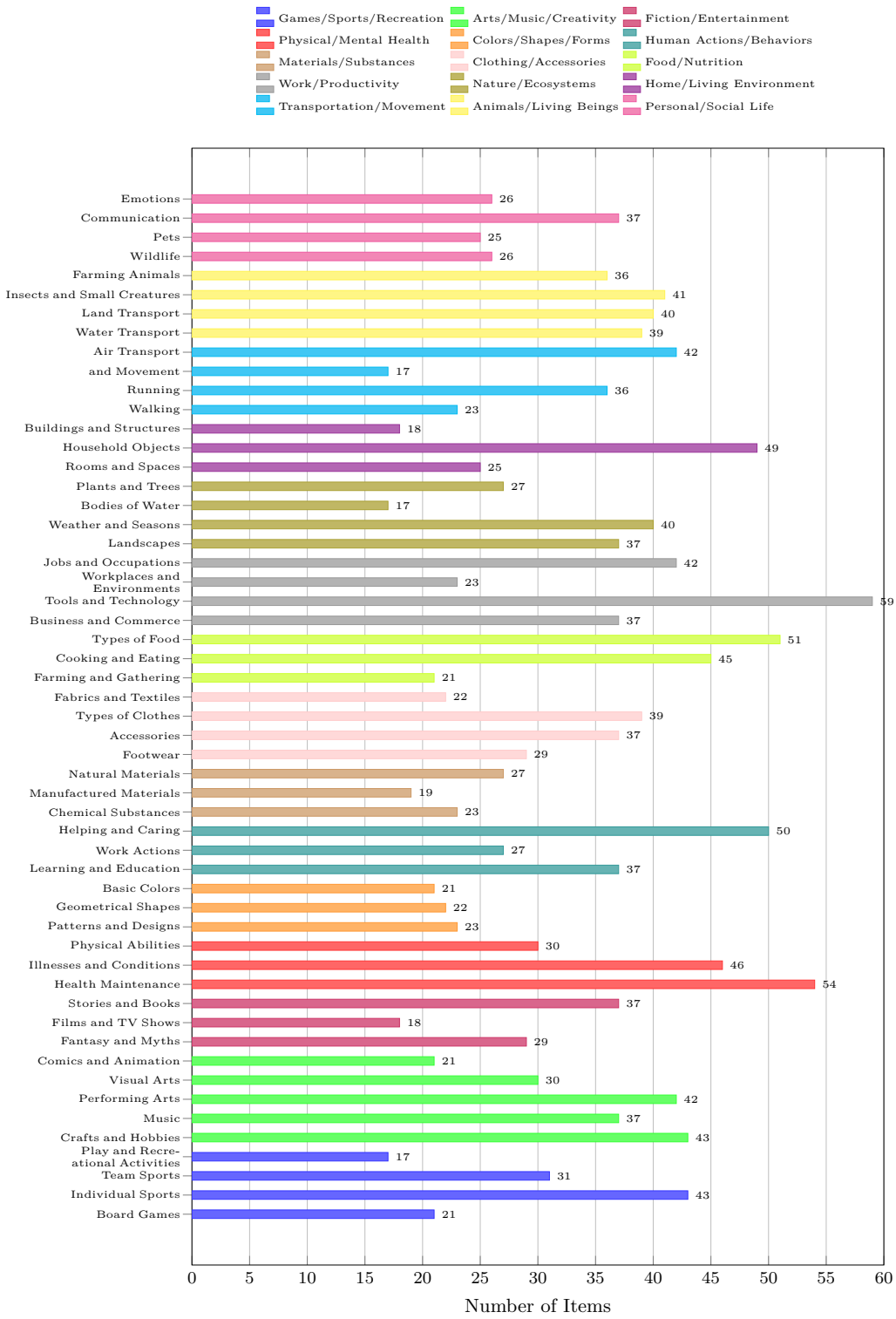


Figure 6. Distribution of JAMAL items across functional categories.

C. Benchmark Verification Guidelines

The reviewer was tasked with examining the entire benchmark. If a context failed to satisfy any of the following conditions, the reviewer was instructed to mark it with a ✗ in the designated box.

Verification Criteria

- Each context must consist of two sentences.
- The target word must be the most natural completion of the context.
- No alternative completion should be reasonably plausible upon reading the context.
- The target word should not be explicitly mentioned in the context; however, the first sentence should provide a clear hint toward it.
- The correct completion should not be inferable from the second sentence alone.
- The target word must be sufficiently clear and predictable so that a reader can infer it from the context.

Example of a Valid Context

✓ **Context:** Rosalyn announced happily, "Checkmate." She was going to become really good at playing. . .
Target word: chess
Justification: The first sentence provides a clear cue ("checkmate"), making "chess" the only plausible completion.

Examples of Invalid Contexts

1. **Context:** Ahmad went to the market. He bought some. . .
Target word: apples
✗ **Justification:** Multiple completions are possible (e.g., bread, milk, meat, fruit).
2. **Context:** Layla decided to bake fresh bread in the oven. She enjoyed the smell of. . .
Target word: bread
✗ **Justification:** The target word ("bread") is explicitly mentioned in the first sentence.
3. **Context:** The class time ended. The teacher started wiping the. . .
Target word: board
✗ **Justification:** The correct completion can be inferred from the second sentence alone.

D. Benchmark Validation Guidelines

The reviewers were tasked with completing sentence contexts and assessing multiple criteria for each benchmark item. The following guidelines outline the validation procedure.

Validation Objectives

- Assess *cloze predictability*: whether reviewer predictions match the target words.
- Assess *reference agreement*: whether the expected word represents the only natural completion.
- Determine *inferential consistency*: which sentence provides contextual clues for prediction.
- Verify *category alignment*: whether contexts properly belong to their assigned categories.

Validation Procedure

Step 1: Read the full context

Each item consists of two consecutive sentences. Read both sentences before making any judgments.

Step 2: Predict the final word

Write a single word that most naturally completes the context. Choose the most coherent completion based on linguistic intuition.

Step 3: Mark inference sources

✓ **Inferable from Sentence 1:** Mark if the first sentence contains the hint to the target word.
✓ **Inferable from Sentence 2:** Mark if the target word can be predicted from the second sentence alone.

Step 4: Validate category assignment

✓ **Matches Category 1:** Mark if the context clearly aligns with Category 1.
✓ **Matches Category 2:** Mark if the context clearly aligns with Category 2.

Step 5: Compare the predicted word to the expected word

✓ Mark **True** if the predicted word matches the expected word exactly.
✓ Mark **False** if the predicted word does not match the expected word.

Step 6: Assess the expected word

✓ Mark **True** if the expected word is the only plausible completion.
✓ Mark **False** if there exist other plausible completions that are not synonymous with the expected word.

Step 7: Document issues

Use the notes column to record:

- Unclear sentences.
- Mismatches between context and assigned categories.
- Difficulties in predicting the target word.
- Any other relevant observations.

Example of Proper Annotation

✓ **Context:** He put the letter in the envelope and dropped it in the mailbox. The worker noticed it was missing a . . .
Target word: stamp
Category 1 (Cognitive): Everyday Situation
Category 2 (Functional): Communication
Validator prediction: stamp
Annotation: Same word: ✓, Inferable from S1: ✓, Inferable from S2 alone: ✗, Matches Category 1: ✓, Matches Category 2: ✓, Uniqueness of the expected completion: ✓

General Instructions

- Rely on natural linguistic intuition; no external resources are permitted.
- Work independently without discussing items with the other reviewer.
- Leave cells blank when uncertain rather than guessing.
- Apply check marks only in designated columns.

Submission Requirements

Before final submission, reviewers must:

1. Ensure each prediction field contains exactly one word.
2. Verify check marks are placed only where appropriate.
3. Confirm all notes provide clear explanations where needed.
4. Save the completed file and send it to the researcher.

References

1. T. B. Brown *et al.*, "Language models are few-shot learners," 2020, doi: https://doi.org/10.48550/arXiv.2005.14165.
2. OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023, doi: https://doi.org/10.48550/arXiv.2303.08774. [Online]. Available: https://arxiv.org/abs/2303.08774
3. H. Touvron *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023. [Online]. Available: https://arxiv.org/abs/2307.09288
4. J. Bai *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023, doi: https://doi.org/10.48550/arXiv.2309.16609.
5. G. Team *et al.*, "Gemini: A family of highly capable multimodal models," 2025. [Online]. Available: https://arxiv.org/abs/2312.11805
6. R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi, "HellaSwag: Can a machine really finish your sentence?" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 4791–4800, doi: https://doi.org/10.18653/v1/P19-1472. [Online]. Available: https://aclanthology.org/P19-1472/
7. D. Hendrycks *et al.*, "Measuring massive multitask language understanding," 2021. [Online]. Available: https://arxiv.org/abs/2009.03300
8. G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy, "RACE: Large-scale ReAding comprehension dataset from examinations," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, M. Palmer, R. Hwa, and S. Riedel, Eds. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 785–794, doi: https://doi.org/10.18653/v1/D17-1082. [Online]. Available: https://aclanthology.org/D17-1082/
9. A. Abdelali *et al.*, "LAraBench: Benchmarking Arabic AI with large language models," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Y. Graham and M. Purver, Eds. St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 487–520, doi: https://doi.org/10.18653/v1/2024.eacl-long.30. [Online]. Available: https://aclanthology.org/2024.eacl-long.30/
10. M. T. I. Khondaker, A. Waheed, E. M. B. Nagoudi, and M. Abdul-Mageed, "GPTAraEval: A comprehensive evaluation of ChatGPT on Arabic NLP," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 220–247, doi: https://doi.org/10.18653/v1/2023.emnlp-main.16. [Online]. Available: https://aclanthology.org/2023.emnlp-main.16/
11. F. Koto *et al.*, "ArabicMMLU: Assessing massive multitask language understanding in Arabic," in *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 5622–5640, doi: https://doi.org/10.18653/v1/2024.findings-acl.334. [Online]. Available: https://aclanthology.org/2024.findings-acl.334/
12. B. Mousi *et al.*, "AraDiCE: Benchmarks for dialectal and cultural capabilities in LLMs," in *Proceedings of the 31st International Conference on Computational Linguistics*, O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, Eds. Abu Dhabi, UAE: Association for Computational Linguistics, Jan. 2025, pp. 4186–4218. [Online]. Available: https://aclanthology.org/2025.coling-main.283/
13. R. N. Almatham *et al.*, "BALSAM: A platform for benchmarking Arabic large language models," in *Proceedings of The Third Arabic Natural Language Processing Conference*, K. Darwish *et al.*, Eds. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 258–277, doi: https://doi.org/10.18653/v1/2025.arabicnlp-main.21. [Online]. Available: https://aclanthology.org/2025.arabicnlp-main.21/
14. S. Al-Khalifa, N. Durrani, H. Al-Khalifa, and F. Alam, "The landscape of arabic large language models," *Communications of the ACM*, vol. 68, no. 10, pp. 54–61, 2025, doi: https://doi.org/10.1145/3737453.
15. K. Smith *et al.*, "The origins of common sense in humans and machines," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 42, 2020, pp. 3–4. [Online]. Available: https://escholarship.org/uc/item/2367w9c4
16. S. K. Tawalbeh and M. Al-Smadi, "Is this sentence valid? an arabic dataset for commonsense validation," *CoRR*, vol. abs/2008.10873, 2020. [Online]. Available: https://arxiv.org/abs/2008.10873
17. N. Sengupta *et al.*, "Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models," 2023. [Online]. Available: https://arxiv.org/abs/2308.16149
18. M. S. Bari *et al.*, "ALLam: Large language models for arabic and english," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=MscdsFVZrN
19. F. Team *et al.*, "Fanar: An arabic-centric multimodal generative ai platform," 2025. [Online]. Available: https://arxiv.org/abs/2501.13944
20. World Health Organization, *International Classification of Functioning, Disability and Health (ICF)*. Geneva: World Health Organization, 2001. [Online]. Available: https://www.who.int/classifications/icf/en/
21. R. Zellers, Y. Bisk, R. Schwartz, and Y. Choi, "SWAG: A large-scale adversarial dataset for grounded commonsense inference," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 93–104, doi: https://doi.org/10.18653/v1/D18-1009. [Online]. Available: https://aclanthology.org/D18-1009/
22. A. Talmor, J. Herzig, N. Lourie, and J. Berant, "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)*, J. Burstein,

C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4149–4158, doi: https://doi.org/10.18653/v1/N19-1421. [Online]. Available: https://aclanthology.org/N19-1421/

23. M. Sap, H. Rashkin, D. Chen, R. Le Bras, and Y. Choi, "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)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 4463–4473, doi: https://doi.org/10.18653/v1/D19-1454. [Online]. Available: https://aclanthology.org/D19-1454/
24. Y. Bisk, R. Zellers, R. Le Bras, J. Gao, and Y. Choi, "Piqa: Reasoning about physical commonsense in natural language," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 7432–7439, doi: https://doi.org/10.1609/aaai.v34i05.6239. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/6239
25. M. Chen, M. D'Arcy, A. Liu, J. Fernandez, and D. Downey, "CODAH: An adversarially-authored question answering dataset for common sense," in *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, A. Rogers, A. Drozd, A. Rumshisky, and Y. Goldberg, Eds. Minneapolis, USA: Association for Computational Linguistics, Jun. 2019, pp. 63–69, doi: https://doi.org/10.18653/v1/W19-2008. [Online]. Available: https://aclanthology.org/W19-2008/
26. B. Y. Lin, S. Lee, X. Qiao, and X. Ren, "Common sense beyond English: Evaluating and improving multilingual language models for commonsense reasoning," 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)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 1274–1287, doi: https://doi.org/10.18653/v1/2021.acl-long.102. [Online]. Available: https://aclanthology.org/2021.acl-long.102/
27. H. Santos, A. M. Mulvehill, K. Shen, M. Kejriwal, and D. L. McGuinness, "Tg-csr: A human-labeled dataset grounded in nine formal commonsense categories," *Data in Brief*, vol. 51, p. 109666, 2023, doi: https://doi.org/10.1016/j.dib.2023.109666. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352340923007515
28. S. Saha, P. Yadav, L. Bauer, and M. Bansal, "ExplaGraphs: An explanation graph generation task for structured commonsense reasoning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 7716–7740, doi: https://doi.org/10.18653/v1/2021.emnlp-main.609. [Online]. Available: https://aclanthology.org/2021.emnlp-main.609/
29. W. Zhan *et al.*, "Score: Benchmarking long-chain reasoning in commonsense scenarios," 2025. [Online]. Available: https://arxiv.org/abs/2503.06218
30. H. Mozannar, E. Maamary, K. El Hajal, and H. Hajj, "Neural Arabic question answering," in *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, W. El-Hajj *et al.*, Eds. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 108–118, doi: https://doi.org/10.18653/v1/W19-4612. [Online]. Available: https://aclanthology.org/W19-4612/
31. J. L. Lee *et al.*, "Massively multilingual pronunciation modeling with WikiPron," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, N. Calzolari *et al.*, Eds. Marseille, France: European Language Resources Association, May 2020, pp. 4223–4228. [Online]. Available: https://aclanthology.org/2020.lrec-1.521/
32. S. Lamsiyah *et al.*, "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)*, 2025, pp. 1–11. [Online]. Available: https://aclanthology.org/2025.wacl-1.1/
33. A. Sadallah *et al.*, "Commonsense reasoning in arab culture," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 7695–7710, doi: https://doi.org/10.18650/ACL.2025.380.
34. A. Hornby and J. Turnbull, *Oxford Advanced Learner's Dictionary of Current English*. Oxford University Press, 2015.
35. F. Ilievski, A. Oltramari, K. Ma, B. Zhang, D. L. McGuinness, and P. Szekely, "Dimensions of commonsense knowledge," *Knowledge-Based Systems*, vol. 229, p. 107347, 2021, doi: https://doi.org/10.1016/j.knosys.2021.107347. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705121006092
36. D. B. Lenat, "Cyc: a large-scale investment in knowledge infrastructure," *Commun. ACM*, vol. 38, no. 11, p. 33–38, Nov. 1995, doi: https://doi.org/10.1145/219717.219745.
37. M. E. Whiting and D. J. Watts, "A framework for quantifying individual and collective common sense," *Proceedings of the National Academy of Sciences*, vol. 121, no. 4, p. e2309535121, 2024, doi: https://doi.org/10.1073/pnas.2309535121.
38. H. A. Simon and A. Newell, *Human problem solving: The state of the theory in 1970*. American Psychological Association, 1971, vol. 26, no. 2, doi: https://doi.org/10.1037/h0030806.
39. R. C. Schank and R. P. Abelson, *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology press, 2013.
40. P. Schuyler, *Common Sense*. Los Angeles, CA: Higher Shelf Publishing, 2003. [Online]. Available: https://www.amazon.com/Common-Sense-Peter-Schuyler/dp/1932636021
41. R. Pesonen, "Casual reasoning : A social ecological look at human cognition and common sense," 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:202257933
42. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *North American Chapter of the Association for Computational Linguistics*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:52967399
43. A. Radford, "Improving language understanding with unsupervised learning," *OpenAI Res*, 2018. [Online]. Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

44. A. Ettinger, "What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 34–48, 2020, doi: https://doi.org/10.1162/tacl_a_00298. [Online]. Available: https://aclanthology.org/2020.tacl-1.3/
45. K. D. Federmeier and M. Kutas, "A rose by any other name: Long-term memory structure and sentence processing," *Journal of Memory and Language*, vol. 41, no. 4, pp. 469–495, 1999, doi: https://doi.org/10.1006/jmla.1999.2660. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0749596X99926608
46. K. Sakaguchi, R. L. Bras, C. Bhagavatula, and Y. Choi, "Winogrande: an adversarial winograd schema challenge at scale," *Commun. ACM*, vol. 64, no. 9, p. 99–106, Aug. 2021, doi: https://doi.org/10.1145/3474381.
47. R. H. Ennis, W. L. Gardiner, R. Morrow, D. Paulus, and L. Ringel, *The cornell class-reasoning test, Form X*, 1964.
48. S. Ouyang, J. M. Zhang, M. Harman, and M. Wang, "An empirical study of the non-determinism of chatgpt in code generation," *ACM Trans. Softw. Eng. Methodol.*, vol. 34, no. 2, Jan. 2025, doi: https://doi.org/10.1145/3697010.
49. K. Ethayarajh, "How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings," 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)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 55–65, doi: https://doi.org/10.18653/v1/D19-1006. [Online]. Available: https://aclanthology.org/D19-1006/
50. W. Gurnee, N. Nanda, M. Pauly, K. Harvey, D. Troitskii, and D. Bertsimas, "Finding neurons in a haystack: Case studies with sparse probing," *Transactions on Machine Learning Research*, 2023. [Online]. Available: https://openreview.net/forum?id=JYs1R9IMJr
51. A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=rygGQyrFvH
52. G. K. Zipf, *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Cambridge, MA: Addison-Wesley, 1949.
53. O. Levy, Y. Goldberg, and I. Dagan, "Improving distributional similarity with lessons learned from word embeddings," *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 211–225, 2015, doi: https://doi.org/10.1162/tacl_a_00134. [Online]. Available: https://aclanthology.org/Q15-1016/
54. A. Srivastava *et al.*, "Beyond the imitation game: Quantifying and extrapolating the capabilities of language models," *Transactions on Machine Learning Research*, 2023, featured Certification. [Online]. Available: https://openreview.net/forum?id=uyTL5Bvosj
55. E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big? ," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 610–623, doi: https://doi.org/10.1145/3442188.3445922.
56. W. Antoun, F. Baly, and H. Hajj, "AraGPT2: Pretrained transformer for Arabic language generation," in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, N. Habash *et al.*, Eds. Kyiv, Ukraine (Virtual): Association for Computational Linguistics, Apr. 2021, pp. 196–207. [Online]. Available: https://aclanthology.org/2021.wanlp-1.21/