

# BALSAM: A Platform for Benchmarking Arabic Large Language Models

Rawan Al-Matham,<sup>1</sup> Kareem Darwish,<sup>2</sup> Raghad Al-Rasheed,<sup>1</sup> Waad Alshammari,<sup>1</sup>  
Muneera Alhoshan,<sup>1</sup> Amal Almazrua,<sup>1</sup> Asma Al Wazrah,<sup>1</sup> Mais Alheraki,<sup>1</sup> Firoj Alam,<sup>2</sup>  
Preslav Nakov,<sup>3</sup> Norah Alzahrani,<sup>4</sup> Eman AlBilali,<sup>5</sup> Nizar Habash,<sup>3,6</sup>  
Abdelrahman El-Sheikh,<sup>7</sup> Muhammad Elmallah,<sup>7</sup> Haonan Li,<sup>3</sup> Hamdy Mubarak,<sup>2</sup>  
Mohamed Anwar,<sup>3</sup> Zaid Alyafeai,<sup>8</sup> Ahmed Abdelali,<sup>9</sup> Nora Altwairesh,<sup>9</sup>  
Maram Hasanain,<sup>2</sup> Abdulmohsen Al Thubaity,<sup>9</sup> Shady Shehata,<sup>10</sup> Bashar Alhafni,<sup>3</sup>  
Injy Hamed,<sup>6</sup> Go Inoue,<sup>3</sup> Khalid Elmadani,<sup>6</sup> Ossama Obeid,<sup>6</sup> Fatima Haouari,<sup>11</sup>  
Tamer Elsayed,<sup>11</sup> Emad Alghamdi,<sup>12</sup> Khalid Almubarak,<sup>13</sup> Saied Alshahrani,<sup>14</sup>  
Ola Aljarrah,<sup>1</sup> Safa Alajlan,<sup>1</sup> Areej Alshaqarawi,<sup>1</sup> Maryam Alshihri,<sup>1</sup>  
Sultana Alghurabi,<sup>1</sup> Atikah Alzaghayer,<sup>1</sup> Afrah Altamimi,<sup>1</sup>  
Abdullah Alfaifi,<sup>1</sup> Abdulrahman AlOsaimy<sup>1</sup>

<sup>1</sup>King Salman Global Academy For Arabic Language, <sup>2</sup>Qatar Computing Research Institute,  
<sup>3</sup>MBZUAI, <sup>4</sup>Sdaia, <sup>5</sup>King Saud University, <sup>6</sup>NYU Abu Dhabi, <sup>7</sup>aiXplain,  
<sup>8</sup>King Abdullah University of Science and Technology, <sup>9</sup>Humain,  
<sup>10</sup>University of Waterloo, <sup>11</sup>Qatar University, <sup>12</sup>King Abdulaziz University,  
<sup>13</sup>Prince Sattam bin Abdulaziz University, <sup>14</sup>University of Bisha

## Abstract

The impressive advancement of Large Language Models (LLMs) in English has not been matched across all languages. In particular, LLM performance in Arabic lags behind, due to data scarcity, linguistic diversity of Arabic and its dialects, morphological complexity, etc. Progress is further hindered by the quality of Arabic benchmarks, which typically rely on static, publicly available data, lack comprehensive task coverage, or do not provide dedicated platforms with blind test sets. This makes it challenging to measure actual progress and to mitigate data contamination. Here, we aim to bridge these gaps. In particular, we introduce BALSAM, a comprehensive, community-driven benchmark aimed at advancing Arabic LLM development and evaluation. It includes 78 NLP tasks from 14 broad categories, with 52K examples divided into 37K test and 15K development, and a centralized, transparent platform for blind evaluation. We envision BALSAM as a unifying platform that sets standards and promotes collaborative research to advance Arabic LLM capabilities.<sup>1</sup>

## 1 Introduction

Arabic is a prominent language with more than 400 million speakers (Boulesnam and Boucetti, 2025) and major religious significance for two billion Muslims. This has translated into significant demand for robust Arabic Natural Language Processing (NLP) systems, resulting in the development

of multiple Arabic-centric Large Language Models (LLMs), such as Jais (Sengupta et al., 2023) and Fanar (Fanar Team et al., 2025), and in improved Arabic support in multilingual models such as Gemini (Gemini Team et al., 2023), GPT-4o (OpenAI et al., 2024). Despite recent progress, LLMs still underperform in Arabic compared to English. This stems from limited training data, the linguistic diversity of Modern Standard Arabic (MSA) and regional dialects, and Arabic’s complex morphology.

Robust benchmarking is crucial to quantify the gaps and guide future improvements in Arabic capabilities of LLMs. Yet, existing Arabic benchmarking initiatives, such as LARA-Bench (Abdelali et al., 2024), have primarily focused on standard natural language generation and understanding tasks. A more recent effort, AraGen (El Filali et al., 2024), introduced a leaderboard-based framework that evaluates LLM performance across multiple dimensions, including correctness, completeness, conciseness, helpfulness, honesty, and harmlessness, in an LLM-as-a-judge setup. In parallel, several datasets have been developed to assess LLM capabilities across different dimensions: ArabicMMLU (Koto et al., 2024) targets world knowledge, AraDICE (Mousi et al., 2025) focuses on dialects with cognitive and cultural understanding, Palm (Alwajih et al., 2025) addresses cultural comprehension, and Ashraf et al. (2025) focus on safety. However, existing efforts address limited LLM capabilities, lack comprehensive coverage, and have no dedicated *platforms* for community

<sup>1</sup><https://benchmarks.ksaa.gov.sa>

collaboration. Critically, measuring progress in a consistent and reliable manner requires a standardized, community-driven framework with blind test datasets, an aspect that remains largely lacking.

Here, we aim to bridge this gap. In particular, we present the *Benchmark for Arabic Language Models (BALSAM)*, which is a comprehensive community-driven initiative designed to advance benchmarking efforts for Arabic LLMs. *BALSAM* includes a collection of 78 tasks across 14 categories, with a total of 52K examples divided into 37K test and 15K dev. These tasks span a wide range of natural language understanding and generation tasks, including summarization, question answering, information extraction, machine translation, and text classification, among others.

*BALSAM* further provides an integrated *evaluation platform* featuring an Arabic LLM Leaderboard. This enables the research community to systematically assess the performance of Arabic LLMs, to monitor progress over time, and to access up-to-date benchmark results for the top-performing LLMs. The *BALSAM* platform goes beyond a traditional leaderboard, serving as a collaborative effort for leading academic and governmental institutions across the Middle East and beyond. Its core mission is to drive the creation of domain-specific test datasets and to establish robust benchmarks for evaluating Arabic LLMs. By promoting transparency and cooperation, *BALSAM* aims to unify the Arabic NLP community around shared datasets and standards. Further, we investigate a variety of automated metrics and measure their correlation with human evaluation. We show that using LLM-as-a-Judge highly correlates with human judgments while other measures such as BLEU, ROUGE, and BertScore don't.

The contributions of *BALSAM* and this paper are summarized as follows:

- *BALSAM* is a community driven consortium that provides a centralized evaluation platform with an associated leaderboard.
- *BALSAM* provide diverse dev/test sets based on 78 tasks, where the test sets are blind.
- We compare the efficacy of using automated evaluations based on BLEU, ROUGE, BERTScore, and LLM-as-a-judge compared to human judgments.

## 2 Related Work

This section reviews prior work across four dimensions: Arabic-centric benchmarks developed to evaluate LLMs in MSA and dialects, English and multilingual benchmarks providing broader frameworks but with limited Arabic coverage, tools and leaderboards enabling systematic model comparison, and a concluding Challenges and Gaps subsection that synthesizes the main limitations of earlier efforts

### 2.1 Arabic-Centric Benchmarks

Recent efforts have focused on benchmarking LLMs for Arabic, targeting tasks such as natural language understanding, generation, and speech processing (Abdelali et al., 2024; Elmadany et al., 2023; Nagoudi et al., 2023). While LLMs have demonstrated remarkable capabilities across various domains, including solving graduate-level mathematical problems and passing medical examinations, these achievements have been predominantly assessed using English-language benchmarks. Thus, in order to evaluate and advance the performance of LLMs for Arabic, there is a critical need for the development of dedicated Arabic benchmarks. Koto et al. (2024) developed ArabicMMLU, an Arabic version of the MMLU benchmark constructed from authentic school exam questions sourced from Arabic-speaking countries, without relying on translation. Similarly, Mousi et al. (2025) created resources for MSA and dialectal Arabic, aiming to assess linguistic, cognitive, and cultural competencies. Alwajih et al. (2025) introduced datasets to evaluate the cultural and dialectal capabilities of LLMs. Almazrouei et al. (2023) adopted and restructured existing datasets to create benchmarks for evaluating LLMs in MSA and dialectal Arabic. Moreover, resources have been developed to assess domain-specific knowledge, e.g., ArabLegalEval (Hijazi et al., 2024) focuses on legal knowledge, while Qiyas (Al-Khalifa and Al-Khalifa, 2024) targets mathematical reasoning. Finally, Ashraf et al. (2025) developed an Arabic dataset for safety.

### 2.2 English/Multilingual Benchmarks

Several prominent benchmarks remain focused on English-centric evaluations, including MMLU (Hendrycks et al., 2021), HELM (Liang et al., 2023), and BIG-bench (Srivastava et al., 2022). MMLU is designed to assess reasoning and knowl-

edge in real-world contexts, while HELM evaluates LLMs across a variety of metrics and scenarios. BIG-bench offers an extensive evaluation framework comprising 214 tasks, some of which include coverage of low-resource languages. Additionally, a range of multilingual benchmarks have been developed to assess model performance across diverse languages, including morphologically complex and low-resource languages such as Arabic.

### 2.3 Tools and Leaderboards

As LLMs continue to advance rapidly, it has become essential to compare their performance across various capabilities and domains. Over time, numerous tools and leaderboards have been developed to facilitate such evaluations. This includes [LLMeBench](#), a comprehensive benchmarking platform with a primary focus on Arabic NLP, speech, and multimodal tasks (Dalvi et al., 2024). Moreover, tools such as [LM-Evaluation-Harness](#), [OpenCompass](#), and [BigCode-Evaluation-Harness](#) provide standardized frameworks for assessing model performance across a wide range of tasks and datasets, facilitating more robust and comprehensive comparisons, as well as signaling to LLM developers areas in which their models need improvement. Several open-source leaderboard initiatives have emerged to benchmark Arabic language models, including the [Open Arabic LLM Leaderboard](#), the [Arabic-MMMLU-Leaderboard](#) (Nacar et al., 2025), and [AraGen](#) (El Filali et al., 2024). Each of them serves a specific purpose. For example, the Arabic-MMMLU-Leaderboard is based on the MMMLU OpenAI benchmark, while AraGen focuses on a diverse set of tasks such as question answering, summarization, and reasoning.

### 2.4 Challenges and Gaps

Existing evaluation benchmarks rely on static, publicly available datasets, enabling rapid community assessment. Yet, as LLMs advance rapidly, static benchmarks struggle to capture their evolving capabilities. The growing size of LLMs and their increasingly extensive training data heighten the risk of test data contamination, which is difficult to detect due to opaque training data and widespread use of synthetic data (Dong et al., 2024). Hence, leaderboards with rigorous contamination checks and adaptive benchmarks that reflect the latest model capabilities are needed (Deng et al., 2023).

The LMSYS Chatbot Arena (Zheng et al., 2023; Chiang et al., 2024) enables robust evaluation of

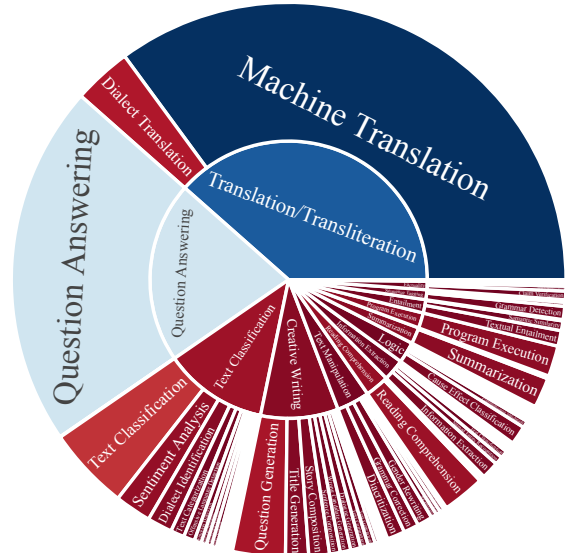


Figure 1: *BALSAM* data distribution across general categories and tasks in these categories.

LLMs through conversational interactions and Elo-based rankings (Bai et al., 2022), but relies on human evaluation, which is time-consuming and limits scalability (Luo et al., 2024). The LLM-as-a-judge approach was introduced to reduce human involvement on platforms such as Chatbot Arena and MT-bench (Zheng et al., 2024), but it requires careful handling to avoid biases such as verbosity, position, and self-enhancement. Moreover, this method struggles with assessing reasoning and math tasks. Indeed, several popular leaderboards, including MT-bench and OpenLLM, face issues of saturation and inconsistent alignment with real-world chatbot performance (Luo et al., 2024).

Despite significant progress in developing English benchmarks and LLM leaderboards, there remains much work to be done for languages such as Arabic. This includes the creation of new datasets to address emerging capabilities and the establishment of sustainable leaderboards that integrate human and LLM-based evaluation approaches.

## 3 BALSAM Dataset

### 3.1 Dataset Creation

The *BALSAM* benchmark is composed of 78 tasks from 14 coarse-grained categories, with a total of 52K examples divided 37K test and 15K development, and a centralized, transparent platform for blind evaluation. We made the design decision to have many datasets, but only have 10–100 test examples per dataset. For most datasets, we also have up to 50 development examples.

Figure 1 shows the data distribution across general categories and tasks in these categories. We can see that the main categories are multiple-choice questions (MCQ), text generation, translation, and transliteration. Table 7 and Table 8 in the Appendix gives the complete list of tasks in *BALSAM* along with the sizes of their development and test sets. The number of examples varies widely between tasks, with some tasks containing thousands of samples and others only a few. Figure 2 in the Appendix shows sample entries for different categories. Note that we converted some tasks to MCQ or text generation, e.g. Part-of-Speech (POS) tagging and Named Entity Recognition (NER), which have been traditionally addressed as sequence labeling tasks. The aim was to ease evaluation as we currently cannot handle sequence labeling tasks (we plan support for this in the future).

**Reusing Public Datasets** Some of the datasets are subsampled from publicly available test sets with preexisting prompts and ground-truth answers. This includes datasets from the Arabic subset of the xP3 dataset (Muennighoff et al., 2023), from which we subsampled 68 datasets, covering 12 tasks, to include 25 development and 50 test examples. We further reformatted AraMath (Alghamdi et al., 2022) to MCQ format, as an additional dataset.

**Prompting Existing NLP Datasets** We created natural language prompts based on publicly available Arabic NLP datasets using the PromptSource tool (Bach et al., 2022). We developed 2–8 different prompt templates per dataset, resulting in an equal number of sub-datasets. Figure 3 in the Appendix shows four different prompt templates developed for one of the datasets.

**Translating English Datasets to Arabic** Some of our datasets were created by translating existing English datasets to Arabic. We have a total of 483 such datasets, covering 29 different tasks, sampled from PromptSource (Bach et al., 2022), SuperNaturalInstructions (Wang et al., 2022; Mishra et al., 2022), and TruthfulQA (Lin et al., 2022). The translations were evaluated both automatically and manually as described in (El-Sheikh et al., 2024).

**Developing New Datasets** We further developed 16 brand new datasets with 1,755 prompts, covering specialized, structured, and rare examples to better test model generalization, e.g., to tasks such as grammatical error detection and factuality.

**Augmenting with Synthetic Examples** Our target was to have 10–100 test examples per dataset. However, for 14 datasets, we had less than 10 examples; we thus used GPT-4o to generate synthetic examples, which we checked manually.

### 3.2 Quality Assurance

To ensure data quality, we conducted extensive quality checks in three iteratively repeated stages:

- *Completeness:* We ensured that all required fields in all datasets were fully populated, with no missing or null values. We found that 1% of our test examples contained null values, which we removed; we further found that 7% of the datasets included duplicates, which we also removed.
- *Consistency:* We established a standardized format to maintain consistency across the datasets. We found that approximately 17% of the datasets exhibited format-related issues, such as improper structure, or incorrect labels, which we fixed.
- *Reliability* We asked 16 annotators to conduct a manual review of random samples from each dataset checking that each *instruction*, *input*, and *output* were clear and cultural appropriate. We found issues for 10% of the datasets; to fix them, we edited some specific examples or excluded entire datasets.

### 3.3 Mitigating Data Leakage

A primary goal of the *BALSAM* initiative is to establish a fair, unbiased, and trusted benchmark for evaluating LLMs in Arabic. Thus, it is critical to prevent test set leakage and to minimize the risk of contamination of LLM training data.

In order to protect the integrity and reliability of the benchmark, we restricted the access to the test sets to a small group of individuals responsible for quality assessment and platform development: in fact, the vast majority of members of the *BALSAM* team only know the part of the raw test data candidates they contributed initially, but they have no access to the final test data.

## 4 Evaluation Setup

### 4.1 Benchmarking Phases

The *BALSAM* benchmark comprises a total of 37,419 test and 15,742 development examples and runs in two phases:

- *Phase 1.* This phase includes 54 tasks across 13 categories focusing on text generation. It contains 13,121 test and 6,434 dev examples. The largest categories are *creative writing* and *translation*, which cover tasks such as *story composition* and *dialect translation*, respectively. A complete breakdown of the categories and associated tasks in this phase is given in Table 7 in the Appendix.
- *Phase 2:* This phase includes 50 tasks across 13 categories and contains 24,298 test examples and 9,308 development examples. The focus of this phase is on multiple-choice question answering and specific generation tasks (Diacritization, Translation/Transliteration).

The two phases share 12 categories in common, with the remaining categories being *translation* (unique to Phase 1) and *factuality* (unique to Phase 2). A complete breakdown of all categories and tasks is provided in Table 8 in the Appendix.

## 4.2 Evaluation Framework

We adopted the LM-Evaluation-Harness (Gao et al., 2024) framework, henceforth *LM-Harness*, for several reasons: (i) it supports evaluation of both open-source LLMs with accessible weights as well as commercial LLMs that are only available via API calls, (ii) it allows flexible customization of tasks and benchmarks through YAML files, and (iii) it has been used in various leaderboards on Hugging Face and as part of various LLM development pipelines, e.g., by Fanar (Fanar Team et al., 2025).

## 4.3 Evaluation Platform

We enhanced the schema of *LM-Harness*<sup>2</sup> to standardize the input data. Each dataset file is assigned a unique ID, and its JSON content is pre-processed into the YAML format required by *LM-Harness*, which includes task metadata and dataset split paths. The evaluation jobs on the platform are organized into categories, tasks, and datasets. Categories group related tasks for visualization purposes. Tasks represent specific objectives such as summarization, sequence tagging, title generation, and transliteration, while datasets contain data split by prompts and data items for each task.

Users register models via an OpenAI-compatible API (requiring model ID and URL) or a public model (e.g., from aiXplain) with optional metadata

<sup>2</sup><https://github.com/ksaa-nlp/balsam-eval>

such as model name and training data. Evaluation requests are run in parallel for selected categories to minimize waiting times. Results are calculated as task-level macro-averages of dataset scores. Similarly, category-level results are computed as the macro-average of per-task scores. The overall score of a model is the macro-average score across all tasks. The *BALSAM* Leaderboard<sup>3</sup> summarizes the model performance, displaying average scores for all tasks. Scores, ranging from 0 to 1, reflect task-specific metrics and enable clear comparisons of model performance across tasks.

## 4.4 Evaluation Measures

Given that the focus of Phase 1 on text generation, we began evaluation using BLEU (Papineni et al., 2002) for the translation category and ROUGE-LSum for the rest of categories (Lin, 2004). For analysis purposes, we also perform manual judgments (see below).

## 5 Experiments

### 5.1 Experimental Setup

We selected a comprehensive set of LLMs that support Arabic; see Appendix D for a detailed list and description of the models we used.

- *Open-weights models:* we chose them based on public availability, relevance to Arabic NLP, and architectural diversity. We conducted all experiments using four NVIDIA A100 GPUs, each with 40G of VRAM.
- *Closed models:* we included some popular ones that support Arabic, and we accessed them via their standard APIs or by provider request.

### 5.2 Results and Discussion

**Challenges in Automatic Evaluation.** Table 1 shows the automatic evaluation results of the LLMs across 13 categories using ROUGE-LSum and BLEU. Unexpectedly, the results show that SILMA-9B is far ahead of much larger models such as Aya 32B, Qwen-2.5 32B, and DeepSeek V3. This prompted us to manually examine random output samples to better understand the underlying reasons. Our analysis revealed the following:

<sup>3</sup><https://benchmarks.ksaa.gov.sa/b/balsam>

Model	CW	ENT	FIB	IE	LOG	PE	QA	RC	ST	SUM	TC	TM	MT/TL	AVG	AVG*
SILMA-9B Instruct-v1.0	0.23	0.13	0.12	0.32	<b>0.22</b>	0.66	<b>0.31</b>	<b>0.55</b>	0.20	0.20	0.36	0.60	0.13	0.31	0.33
Nuha v2	<b>0.22</b>	0.12	0.12	0.32	0.20	0.81	0.25	0.35	0.28	0.19	<b>0.39</b>	<b>0.64</b>	0.15	0.31	0.32
Jais-family 13B-chat	<b>0.24</b>	0.08	0.10	0.25	0.17	<b>0.89</b>	0.22	0.51	0.19	<b>0.26</b>	0.15	0.48	–	–	0.30
Command R+	0.19	0.07	0.11	<b>0.33</b>	0.17	0.68	<b>0.31</b>	0.41	0.28	0.16	0.28	0.53	0.15	0.28	0.29
GPT-4o	0.22	0.10	0.20	0.28	0.16	0.21	0.23	0.29	<b>0.38</b>	0.17	0.30	0.62	<b>0.17</b>	0.26	0.27
Iron Horse GV V5a	0.20	0.10	<b>0.21</b>	0.27	0.15	0.48	0.21	0.24	0.36	0.15	0.27	0.56	0.14	0.26	0.27
Yehia 7B Preview	0.23	0.13	0.18	0.26	0.20	0.34	0.23	0.28	0.26	0.20	0.24	0.62	0.14	0.25	0.27
AceGPT-v2 8B Chat	0.19	0.11	0.14	0.29	0.18	0.49	0.25	0.36	0.19	0.19	0.23	0.51	0.11	0.25	0.26
Grok-2-latest	0.20	0.08	0.14	0.23	0.15	0.16	0.22	0.29	0.30	0.18	0.18	0.49	0.14	0.21	0.24
Gemini 2.0 Flash	0.17	0.06	0.14	0.28	0.15	0.13	0.25	0.30	0.33	0.15	0.24	0.33	0.13	0.20	0.22
Mistral-saba-latest	0.21	0.07	0.16	0.18	0.14	0.15	0.20	0.23	0.29	0.18	0.19	0.55	0.15	0.21	0.21
Claude Sonnet 3.5	0.13	<b>0.15</b>	0.07	0.19	0.09	0.24	0.18	0.20	0.35	0.15	0.12	0.38	0.12	0.18	0.19
Command-r7b 12-2024	0.17	0.07	0.15	0.15	0.13	0.26	0.15	0.22	0.19	0.16	0.13	0.41	0.13	0.18	0.19
Gemma 2 9B	0.16	0.09	0.11	0.19	0.14	0.31	0.18	0.23	0.19	0.15	0.10	0.30	0.05	0.17	0.19
Qwen 2.5 32B	0.14	0.09	0.13	0.16	0.11	0.30	0.15	0.13	0.23	0.16	0.08	0.43	0.08	0.17	0.18
DeepSeek V3	0.17	0.12	0.11	0.18	0.11	0.12	0.15	0.14	0.25	0.15	0.08	0.40	0.15	0.16	0.17
C4AI Aya Expanse 32B	0.14	0.07	0.11	0.13	0.07	0.23	0.14	0.25	0.13	0.19	0.06	0.38	0.10	0.15	0.16
Fanar-C-1-8.7B	0.14	0.09	0.07	0.16	0.11	0.36	0.14	0.15	0.11	0.14	0.11	0.33	–	–	0.16
Amazon Nova Pro	0.15	0.07	0.07	0.12	0.07	0.14	0.15	0.10	0.26	0.15	0.04	0.37	0.09	0.14	0.15
Mistral Large	0.08	0.10	0.04	0.10	0.08	0.17	0.12	0.15	0.06	0.07	0.09	0.30	0.05	0.11	0.12
DBRX-instruct	0.03	0.01	0.03	0.03	0.02	0.10	0.04	0.04	0.03	0.03	0.02	0.12	0.02	0.04	0.04
Aragpt2 mega	–	0.11	0.04	–	0.04	–	0.05	0.06	0.04	0.13	0.06	0.33	–	–	–

Table 1: **Automatic evaluation across categories.** “–” indicates that the model exceeded the token limits and did not complete the category. List of categories: CW (Creative Writing), ENT (Entailment), FIB (Fill in the Blank), IE (Information Extraction), LOG (Logic), PE (Program Execution), QA (Question Answering), RC (Reading Comprehension), ST (Sequence Tagging), SUM (Summarization), TC (Text Classification), TM (Text Manipulation), MT/TL (Machine Translation/Transliteration), AVG (Average), AVG\* (Average w/o Translation).

- SILMA-9B’s output was generally terse, while the outputs of the other models were verbose; the metrics naturally preferred shorter answers. In a Question Answering example where the correct answer was باريس (Paris), SILMA-9B gave a matching terse reply, while other models provided more detailed, verbose answers with 25 words or longer (Full example in Appendix E).
- BLEU uses the geometric mean of unigram to 4-gram precisions. Because many gold answers were short, trigram and 4-gram matches were often absent, causing BLEU scores to be zero despite matching unigrams and bigrams.
- BLEU and ROUGE rely on exact word matches, which is difficult for Arabic’s complex morphology. For example, the reference كتاب (‘book’) and the prediction الكتاب (‘the book’) do not match exactly.

**Human Evaluation.** Next, we conducted a manual evaluation on a random sample of the test set, composed of 20 questions per category, where humans would rate the outputs from all LLMs. The correctness of each output, on a 0–3 scale, was judged by three judges. Thus, the total number of performed judgments was 254 questions  $\times$  22 LLMs  $\times$  3 judges = 16,764 judgments. The de-

tailed annotation instructions we gave to the judges are given in Appendix G.

The average score per model from these judgments are reported in Table 2, where we can see that GPT-4o achieves the highest average score.

#### Human-to-Automatic Measure Correlation.

We measured the Pearson correlation of human judgments against ROUGE-LSum and BLEU. Table 3 shows the correlation between the three human judgments across categories. The average correlation between the judges is 0.75, and they correlated more with each other for some categories compared to others. For example, Creative Writing had the lowest correlation (0.636), while Reading Comprehension had the highest correlation (0.88). Table 4 lists the correlations of manual evaluation against ROUGE-LSum and BLEU. Since we had three judges, we computed the correlation between the metrics and the average judges’ scores. We can see very poor correlation between manual judgments and automatic measures.

**Beyond BLEU and ROUGE.** We explored some alternative evaluation approaches, namely:

- **Semantic Evaluation:** We used BERTScore (Zhang et al., 2020), which captures semantic similarity more effectively than surface-level  $n$ -gram overlap.

Model	CW	ENT	FIB	IE	LOG	PE	QA	RC	ST	SUM	TC	TM	MT/TL	AVG	AVG*
GPT-4o	2.78	3.00	2.50	2.57	2.50	2.30	2.30	2.65	2.12	2.72	2.57	2.72	2.75	2.58	2.56
Iron Horse GV V5a	2.63	3.00	2.50	2.32	2.15	2.50	2.25	2.77	2.08	2.52	2.23	2.52	2.85	2.49	2.46
Claude Sonnet 3.50	2.83	2.52	2.57	2.58	2.27	2.58	2.35	2.53	2.02	2.62	1.97	2.67	2.68	2.48	2.46
DeepSeek V3	2.70	2.93	2.17	2.57	2.20	2.30	2.37	2.80	1.97	2.88	1.87	2.52	2.48	2.44	2.44
Nuha v2	2.75	2.62	2.53	2.38	1.95	2.20	2.32	2.83	2.02	2.88	1.70	2.70	2.63	2.42	2.40
Grok-2-latest	2.78	3.00	1.95	2.47	2.52	2.50	2.27	2.80	1.80	2.75	1.60	2.47	2.53	2.42	2.41
Gemini 2.0 Flash	2.73	2.74	2.42	2.53	2.43	2.13	2.12	2.77	1.95	2.60	1.55	2.37	2.72	2.39	2.36
Command R+	2.60	2.81	2.23	2.52	2.13	2.08	2.30	2.58	1.97	2.70	1.57	2.37	2.42	2.33	2.32
Fanar-C-1-8.7B	2.73	2.98	1.82	2.62	2.25	2.25	2.70	2.82	1.22	2.67	1.62	2.03	–	–	2.31
c4ai-aya-expans-32b	2.65	2.88	2.37	2.37	2.02	2.28	2.23	2.57	1.68	2.70	1.62	2.47	2.13	2.31	2.32
Mistral-saba-latest	2.60	2.86	2.15	2.55	2.00	1.25	2.38	2.82	1.90	2.78	1.43	2.53	2.50	2.29	2.27
Yehia-7B preview	2.68	2.98	1.88	2.28	2.08	1.83	2.28	2.63	1.75	2.50	1.65	2.68	2.13	2.26	2.27
Amazon Nova Pro	2.65	2.86	2.23	2.20	2.18	1.42	2.32	2.63	1.78	2.75	1.60	2.35	2.42	2.26	2.25
Gemma2 9B	2.62	2.90	1.70	2.33	2.08	1.97	2.20	2.85	1.73	2.67	1.77	1.93	2.05	2.22	2.23
Qwen-2.5 32b	2.83	2.55	1.97	2.15	1.97	2.18	2.12	2.72	1.77	2.55	1.45	2.42	2.08	2.21	2.22
Command-r7b 12-2024	2.62	2.83	1.60	2.08	1.88	2.00	2.20	2.45	1.75	2.77	1.18	2.38	1.87	2.12	2.15
Jais-family 13b-chat	2.03	2.88	1.13	2.23	1.70	2.17	1.87	2.52	1.35	2.38	1.02	2.18	–	–	1.96
SILMA-9B Instruct-v1.0	2.33	2.00	1.42	2.1	1.73	1.68	1.83	2.13	1.52	2.4	1.63	2.28	2.00	1.93	1.92
AceGPT-v2-8B-Chat	2.17	2.21	1.08	2.17	1.75	1.38	1.50	2.57	1.63	2.62	1.07	2.05	1.77	1.84	1.85
Mistral large	1.20	1.79	0.80	0.98	1.22	0.98	1.65	1.52	0.65	0.58	0.62	1.27	1.78	1.16	1.11
DBRX-instruct	0.23	0.24	0.07	0.22	0.28	0.73	0.43	0.18	0.77	0	0.22	0.12	1.28	0.37	0.29
Aragpt2-mega	–	0.14	0.13	–	0.13	–	0.13	0.42	0.1	1.63	0.05	0.37	–	–	–

Table 2: **Manual evaluation (3 evaluators; 20 examples per category).** “–” indicates that the model exceeded token limits and did not complete the category. List of categories: CW (Creative Writing), ENT (Entailment), FIB (Fill in the Blank), IE (Information Extraction), LOG (Logic), PE (Program Execution), QA (Question Answering), RC (Reading Comprehension), ST (Sequence Tagging), SUM (Summarization), TC (Text Classification), TM (Text Manipulation), MT/TL (Machine Translation/Transliteration), AVG (Average), AVG\* (Average w/o Translation).

- **LLM-Based Answer Extraction:** We used Gemini 2.5 Flash (zero-shot, no chain-of-thought) to extract concise answers from the model-generated outputs. We used the prompt reported in the Appendix, Listing 1.
- **LLM-Based Scoring:** We used Gemini 2.5 Flash to rate the extracted answers on a 0–3 scale, mirroring the manual evaluation scheme.<sup>4</sup> The scoring prompt is shown in Appendix Listing 2.

Table 5 shows the correlation of human evaluation with ROUGE-LSum, BLEU, and BERTScore (with and without extraction of answers using an LLM) and LLM as a judge. We make the following observations:

- Using an LLM to extract the answer from the LLM output generally had a positive impact on correlation for all measures (ROUGE-LSum, BLEU, and BERTScore).
- BERTScore correlated better with human judgments compared to ROUGE and BLEU.

<sup>4</sup>We also experimented with GPT-4o and GPT-4o mini as LLM judges. GPT-4 and Gemini showed nearly identical correlation with human scores, both outperforming GPT-4o mini by a sizable margin. Eventually, we selected Gemini 2.5 Flash due to its substantially lower cost.

Category	1 & 2	1 & 3	2 & 3	Avg.
Creative Writing	0.579	0.563	0.765	0.636
Entailment	0.824	0.757	0.768	0.783
Fill in the Blank	0.587	0.659	0.826	0.691
Info. Extraction	0.636	0.602	0.730	0.656
Logic	0.578	0.630	0.586	0.598
Program Execution	0.722	0.697	0.841	0.753
Q&A	0.883	0.813	0.816	0.837
Reading Compr.	0.885	0.894	0.860	0.880
Sequence Tagging	0.738	0.768	0.935	0.814
Summarization	0.828	0.774	0.754	0.785
Text Classification	0.833	0.820	0.921	0.858
Text Manipulation	0.790	0.808	0.792	0.797
Translation	0.646	0.607	0.746	0.666
<b>Average</b>	0.733	0.722	0.795	0.750

Table 3: Correlation between the three human judges (1, 2, & 3) per category.

- The correlation for ROUGE-LSum, BLEU, and BERTScore varied widely from category to category, and the average was low.
- LLM as a judge was highly correlated with human judgments for all categories, with values ranging between 0.824 and 0.977. In fact, it correlated better with the average of judges’ scores than judges correlated with each other.

Based on the above, we decided to drop ROUGE, BLEU, and BERTScore and rely solely on LLM as a Judge to evaluate the LLMs. Table 6 lists the results for all models on the entire *BALSAM* test set using LLM as a judge. When comparing the results

Category	ROUGE-Lsum	BLEU
Creative Writing	-0.509	-0.613
Entailment	-0.300	0.010
Fill in the Blank	-0.033	-0.008
Info. Extraction	0.139	0.514
Logic	0.425	0.296
Program Execution	-0.151	-0.005
Question Answering	0.339	0.316
Reading Comprehension	0.318	0.299
Sequence Tagging	0.537	0.094
Summarization	-0.393	-0.187
Text Classification	0.100	0.090
Text Manipulation	0.462	0.460
Translation	0.506	0.481
Average	0.111	0.134

Table 4: Correlation of human judgments against ROUGE-LSum and BLEU for different categories.

of using ROUGE-LSum and BLEU (Table 1) to using LLM as a judge (Table 6), we can see that the order of LLMs changes completely. In fact, the top performer in Table 1, namely SILMA-9B-IT came out in the lower third in Table 6. Given the aforementioned discussion, the LLM as a judge results are more trustworthy as they correlate much better with human judgments.

The results show that large closed models, e.g. GPT-4o, Gemini 2.0, and DeepSeek V3, significantly outperform all smaller Arabic-centric models such as Jais and Fanar. Two large models, namely Mistral large and DBRX-instruct performed poorly, trailing most others. Hence, model size is not a sufficient predictor of performance. Some of the most likely factors that come into play are Arabic tokenization, size of Arabic training set, and Arabic-centric supervised fine-tuning.

The results show some variability of how models generally perform for certain categories compared to others. For example, models overall perform better on some tasks, such as *translation* and *entailment*, and worse on others, such as *fill in the blank*. Some models are relatively more capable for some categories compared to others. For example, Grok-2 leads the pack for Logic and Iron Horse leads for Program Execution. Similarly, some models rank higher for some categories and much lower in others. For example, Jais and Fanar performed well for Summarization but poorly for Sequence Tagging. Some models performed poorly across the board, such as Aragpt2-mega and DBRX-Instruct.

## 6 Conclusion and Future Directions

We have presented *BALSAM* — a major collaborative effort to establish benchmarking standards and foster unity in LLM development and evaluation for Arabic. *BALSAM* marks a significant step forward, offering evaluation across 78 tasks from 14 categories, with 37K development and 15K test examples. It further offers an integrated platform, and Arabic LLM Leaderboard that enable effective evaluation, comparison, and progress tracking with reliable LLM-as-a-judge based evaluation. However, challenges remain in enhancing data quality, addressing Arabic’s linguistic diversity, and expanding the scope of tasks covered.

In future work, we aim to improve dataset quality (e.g., eliminate translations and any form of synthetic data generation) to add additional tasks, as well as to address the limitations listed in the next section.

### Limitations

Our study provides insights into LLM performance; however, several key limitations warrant consideration and will be the focus of the next iteration of the *BALSAM* benchmarking test sets.

- Token length restrictions in certain models precluded their complete participation across all evaluation tasks, particularly affecting models with restricted context windows and preventing calculation of comprehensive performance scores for these systems.
- While efforts have been made to ensure the accuracy and neutrality of the datasets, we acknowledge the potential for unintended biases, particularly those arising from translated datasets that may have translation errors or cultural misalignments. For example, certain phrases, such as “the Messenger of Islam Muhammad” were identified as potentially problematic, as they may not align with widely accepted terminologies within specific cultural and religious contexts, such as the more commonly used “Prophet Muhammad” in Arabic and Islamic discourse.
- Though *BALSAM* benchmarks LLMs across a variety of categories, some notable other functions and features of LLMs need to be considered such as fluency of the generated output, cultural alignment, ability to answer



Category	ROUGE	Ext. ROUGE	BLEU	Ext. BLEU	BERT	Ext. BERT	LLM-J
Creative Writing	-0.509	-0.476	-0.613	-0.582	-0.629	-0.392	<b>0.824</b>
Entailment	-0.300	0.227	0.010	0.546	-0.244	-0.176	<b>0.950</b>
Fill in the Blank	-0.033	0.390	-0.008	-0.502	0.386	0.696	<b>0.944</b>
Information Extraction	0.139	0.656	0.514	0.766	0.034	0.691	<b>0.824</b>
Logic	0.425	0.742	0.296	0.554	0.429	0.676	<b>0.945</b>
Program Execution	-0.151	0.715	-0.005	0.882	-0.235	-0.034	<b>0.911</b>
Question Answering	0.339	0.807	0.316	0.494	0.408	0.852	<b>0.977</b>
Reading Comprehension	0.318	0.285	0.299	0.008	0.413	0.268	<b>0.931</b>
Sequence Tagging	0.537	-0.241	0.094	-0.793	0.691	0.182	<b>0.931</b>
Summarization	-0.393	-0.754	-0.187	-0.676	0.092	-0.604	<b>0.934</b>
Text Classification	0.100	0.400	0.090	0.275	0.251	0.830	<b>0.948</b>
Text Manipulation	0.462	0.677	0.460	0.685	0.401	0.678	<b>0.919</b>
Translation	0.506	0.806	0.481	0.754	0.390	0.831	<b>0.899</b>
Average	0.111	0.326	0.134	0.186	0.184	0.346	<b>0.918</b>

Table 5: Correlation of human judgments against ROUGE-LSum, BLEU, BERTScore (and their extracted versions), and LLM-as-a-Judge (LLM-J).

Model	CW	ENT	FIB	IE	LOG	PE	QA	RC	ST	SUM	TC	TM	MT/TL	AVG	AVG*
GPT-4o	1.93	2.14	<b>1.77</b>	2.14	1.92	1.81	2.16	2.21	<b>1.99</b>	1.98	<b>2.23</b>	2.02	2.3	<b>2.05</b>	<b>2.03</b>
Gemini 2.0 Flash	<b>1.96</b>	2.00	1.55	2.15	1.91	2.18	<b>2.20</b>	<b>2.27</b>	1.85	1.99	2.03	1.98	2.24	2.02	2.01
Iron Horse GV V5a	1.90	2.14	1.35	<b>2.17</b>	1.88	<b>2.56</b>	2.12	2.05	1.82	1.89	1.90	2.02	<b>2.51</b>	2.02	1.98
DeepSeek V3	1.7	2.21	1.52	2.1	1.88	2.32	2.01	2.11	1.83	1.95	2.04	2.02	2.21	1.99	1.97
Claude Sonnet 3.5	1.85	2.07	1.32	2.08	1.8	2.42	2.09	2.18	1.88	1.95	1.79	<b>2.09</b>	2.37	1.99	1.96
Grok-2-latest	1.94	2.07	1.29	2.10	<b>2.01</b>	2.15	2.04	2.22	1.59	1.98	2.07	1.86	2.10	1.96	1.94
Nuha v2	1.86	1.86	1.39	1.99	1.84	2.37	1.91	2.20	1.59	1.95	<b>2.20</b>	1.86	1.96	1.92	1.92
Qwen-2.5 32b	1.85	1.93	1.39	1.88	1.82	1.88	1.79	2.02	1.57	1.96	1.77	1.78	1.74	1.8	1.8
Mistral-saba-latest	1.82	1.93	1.39	1.98	1.68	1.43	1.98	2.12	1.6	1.84	1.95	1.84	2.06	1.81	1.79
Gemma2 9B	1.78	<b>2.29</b>	1.26	1.94	1.67	1.61	1.72	2.15	1.41	1.96	1.72	1.62	1.67	1.75	1.76
c4ai-aya-expansive-32b	1.71	1.93	1.03	1.90	1.58	2.01	1.8	1.99	1.20	2.02	1.64	1.87	2.14	1.75	1.72
Command R+	1.76	1.79	0.94	1.96	1.54	1.7	1.85	2.03	1.41	1.74	1.57	1.82	2.35	1.73	1.68
Amazon Nova Pro	1.77	2.07	1.13	1.81	1.54	1.35	1.81	1.81	1.49	1.65	1.68	1.95	2.18	1.71	1.67
Yehia-7B preview	1.79	2.14	0.9	1.89	1.46	1.34	1.73	2.06	1.17	1.83	1.62	1.83	2.02	1.68	1.65
Fanar-C-1-8.7B	1.70	1.93	0.90	1.88	1.53	1.96	1.72	1.79	0.95	1.86	1.52	1.71	-	-	1.62
Jais-family 13b-chat	1.80	1.86	0.52	1.62	1.39	2.42	1.49	1.85	0.66	<b>2.05</b>	1.11	1.57	-	-	1.53
SILMA-9B Instruct-v1.0	1.67	1.57	0.97	1.63	1.50	1.31	1.46	2.17	1.01	1.73	1.77	1.5	1.84	1.55	1.52
Command-r7b 12-2024	1.57	1.79	0.65	1.62	1.45	1.56	1.64	1.94	1.05	1.58	1.15	1.67	2	1.51	1.47
Mistral large	1.52	1.21	1.13	1.65	1.54	1.50	1.57	1.86	1.04	1.52	1.51	1.16	1.47	1.44	1.43
AceGPT-v2-8B-Chat	1.56	1.71	0.58	1.73	1.27	0.92	1.62	1.85	0.88	1.74	0.95	1.54	1.72	1.39	1.36
DBRX-instruct	0.73	0.93	0.33	1.10	0.81	0.96	1.09	1.4	0.85	0.78	1.18	0.67	1.14	0.92	0.90
Aragpt2-mega	-	0.00	0.03	-	0.05	-	0.12	0.25	0.02	0.15	0.15	0.29	-	-	-

Table 6: **LLM-as-a-judge evaluation.** “-” indicates that the model exceeded token limits and did not complete the category. List of categories: CW (Creative Writing), ENT (Entailment), FIB (Fill in the Blank), IE (Information Extraction), LOG (Logic), PE (Program Execution), QA (Question Answering), RC (Reading Comprehension), ST (Sequence Tagging), SUM (Summarization), TC (Text Classification), TM (Text Manipulation), MT/TL (Machine Translation/Transliteration), AVG (Average), AVG\* (Average w/o Translation).

religious questions, ability to chat in a multi-turn scenario, propensity to hallucinate, tool usage, structured output generation, and many others. We plan to address many of these aspects in the next iteration of BALSAM with new test sets.

## References

Ahmed Abdelali, Hamdy Mubarak, Shammur Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Samir Abdaljalil, Yassine El Kheir, Daniel Izham, Fahim Dalvi, Majd Hawasly, Nizi Nazar, Youssef Elshahawy, Ahmed Ali, Nadir Durrani, Natasa Milic-Frayling, and Firoj Alam. 2024. **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)*, pages 487–520, St. Julian’s, Malta. Association for Computational Linguistics.

Shahad Al-Khalifa and Hend Al-Khalifa. 2024. **The qiyas benchmark: Measuring chatgpt mathematical and language understanding in arabic.**

Reem Alghamdi, Zhenwen Liang, and Xiangliang Zhang. 2022. **Armath: a dataset for solving arabic math word problems.** In *Proceedings of the Language Resources and Evaluation Conference*, pages 351–362, Marseille, France. European Language Resources Association.

Ebtesam Almazrouei, Ruxandra Cojocaru, Michele Baldo, Quentin Malartic, Hamza Alobeidli, Daniele

- Mazzotta, Guilherme Penedo, Giulia Campesan, Mugiya Farooq, Maitha Alhammedi, Julien Launay, and Badreddine Noune. 2023. [ALGhafa evaluation benchmark for Arabic language models](#). In *Proceedings of ArabicNLP 2023*, pages 244–275, Singapore (Hybrid). Association for Computational Linguistics.
- Fakhraddin Alwajih, Abdellah El Mekki, Samar Mohamed Magdy, Abdelrahim A. Elmadany, Omer Nacar, El Moatez Billah Nagoudi, Reem Abdel-Salam, Hanin Atwany, Youssef Nafea, et al. 2025. [Palm: A culturally inclusive and linguistically diverse dataset for arabic llms](#). *arXiv preprint arXiv:2503.00151*.
- Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. [AraGPT2: Pre-trained transformer for Arabic language generation](#). In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 196–207, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.
- Yasser Ashraf, Yuxia Wang, Bin Gu, Preslav Nakov, and Timothy Baldwin. 2025. [Arabic dataset for LLM safeguard evaluation](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5529–5546, Albuquerque, New Mexico. Association for Computational Linguistics.
- Stephen Bach, Victor Sanh, Zheng Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-david, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Fries, Maged Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-jian Jiang, and Alexander Rush. 2022. [Prompt-Source: An integrated development environment and repository for natural language prompts](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 93–104, Dublin, Ireland. Association for Computational Linguistics.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.
- Ilhem Boulesnam and Rabah Boucetti. 2025. [Arabic language characteristics that make its automatic processing challenging](#). *The International Arab Journal of Information Technology*, 22(4):814–831.
- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*, Vienna, Austria.
- Fahim Dalvi, Maram Hasanain, Sabri Boughorbel, Basel Mousi, Samir Abdaljalil, Nizi Nazar, Ahmed Abdelali, Shammur Absar Chowdhury, Hamdy Mubarak, Ahmed Ali, Majd Hawasly, Nadir Durani, and Firoj Alam. 2024. [LLMeBench: A flexible framework for accelerating llms benchmarking](#). *Association for Computational Linguistics*.
- John Dang, Shivalika Singh, Daniel D’souza, Arash Ahmadian, Alejandro Salama, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawlhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. [Aya expand: Combining research breakthroughs for a new multilingual frontier](#). *arXiv preprint arXiv:2412.04261*.
- DeepSeek-AI. 2024. [Deepseek-v3 technical report](#). *arXiv preprint arXiv:2412.19437*, abs/2412.19437.
- Chunyan Deng, Yilun Zhao, Xiangru Tang, Mark Gestein, and Arman Cohan. 2023. Benchmark probing: Investigating data leakage in large language models. In *NeurIPS 2023 Workshop on Backdoors in Deep Learning-The Good, the Bad, and the Ugly*.
- Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. 2024. [Generalization or memorization: Data contamination and trustworthy evaluation for large language models](#). In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12039–12050, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.
- Ali El Filali, Neha Sengupta, Arwa Abouelseoud, Preslav Nakov, and Clémentine Fourier. 2024. [Rethinking llm evaluation with 3c3h: Aragen benchmark and leaderboard](#). [urlhttps://huggingface.co/spaces/inceptionai/AraGen-Leaderboard](https://huggingface.co/spaces/inceptionai/AraGen-Leaderboard).
- Abdelrahman El-Sheikh, Ahmed Elmogtaba, Kareem Darwish, Muhammad Elmallah, Ashraf Elneima, and Hassan Sawaf. 2024. [Creating arabic llm prompts at scale](#). *arXiv preprint arXiv:2408.05882*.
- AbdelRahim Elmadany, ElMoatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. [ORCA: A challenging benchmark for Arabic language understanding](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9559–9586, Toronto, Canada. Association for Computational Linguistics.

- Amazon AGI et al. 2025. [The amazon nova family of models: Technical report and model card](#). *arXiv preprint arXiv:2506.12103*.
- Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehka, Anastasios Fragkopoulos, Maram Hasanain, Majd Hawasly, Mus'ab Husaini, Soon-Gyo Jung, Ji Kim Lucas, Walid Magdy, Safa Messaoud, Abubakr Mohamed, Tasnim Mohiuddin, Basel Mousi, Hamdy Mubarak, Ahmad Musleh, Zan Naeem, Mourad Ouzani, Dorde Popovic, Amin Sadeghi, Husrev Taha Sencar, Mohammed Shinoy, Omar Sinan, Yifan Zhang, Ahmed Ali, Yassine El Kheir, Xiaosong Ma, and Chaoyi Ruan. 2025. [Fanar: An arabic-centric multimodal generative ai platform](#).
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. [A framework for few-shot language model evaluation](#).
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. [Gemini: a family of highly capable multimodal models](#). *arXiv preprint arXiv:2312.11805*.
- Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. [Gemma: Open models based on gemini research and technology](#). *arXiv preprint arXiv:2403.08295*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). In *International Conference on Learning Representations*.
- Faris Hijazi, Somayah Alharbi, Abdulaziz AlHussein, Harethah Shairah, Reem Alzahrani, Hebah Alshamlan, George Turkiyyah, and Omar Knio. 2024. [ArabLegalEval: A multitask benchmark for assessing Arabic legal knowledge in large language models](#). In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 225–249, Bangkok, Thailand. Association for Computational Linguistics.
- Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Song Dingjie, Zhihong Chen, Mosen Alharthi, Bang An, Juncai He, Ziche Liu, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2024. [AceGPT, localizing large language models in Arabic](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8139–8163, Mexico City, Mexico. Association for Computational Linguistics.
- Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. [ArabicMMLU: Assessing massive multitask language understanding in Arabic](#). In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5622–5640, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. [Holistic evaluation of language models](#).
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. [TruthfulQA: Measuring how models mimic human falsehoods](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.
- Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Qingwei Lin, Jianguang Lou, Shifeng Chen, Yansong Tang, and Weizhu Chen. 2024. [Arena learning: Build data flywheel for llms post-training via simulated chatbot arena](#). *arXiv preprint arXiv:2407.10627*.
- Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. [Cross-task generalization via natural language crowdsourcing instructions](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.
- Basel Mousi, Nadir Durrani, Fatema Ahmad, Md. Arif Hasan, Maram Hasanain, Tameem Kabbani, Fahim

- Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2025. [AraDiCE: Benchmarks for dialectal and cultural capabilities in LLMs](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4186–4218, Abu Dhabi, UAE. Association for Computational Linguistics.
- Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. [Crosslingual generalization through multitask finetuning](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.
- Omer Nacar, Serry Taiseer Sibae, Samar Ahmed, Safa Ben Atitallah, Adel Ammar, Yasser Alhabashi, Abdulrahman S. Al-Batati, Arwa Alsehibani, Nour Qandos, Omar Elshehy, Mohamed Abdelkader, and Anis Koubaa. 2025. [Towards inclusive Arabic LLMs: A culturally aligned benchmark in Arabic large language model evaluation](#). In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 387–401, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- El Moatez Billah Nagoudi, AbdelRahim Elmadany, Ahmed El-Shangiti, and Muhammad Abdul-Mageed. 2023. [Dolphin: A challenging and diverse benchmark for Arabic NLG](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1404–1422, Singapore. Association for Computational Linguistics.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerii Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Carrier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Lukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Lukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. [GPT-4 technical report](#).
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evalu-](#)

- ation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Soudos Mahmoud Bsharat, Alham Fikri Aji, Zhiqiang Shen, Zhengzhong Liu, Natalia Vassilieva, Joel Hestness, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Hector Xuguang Ren, Preslav Nakov, Timothy Baldwin, and Eric Xing. 2023. Jais and Jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adria Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.
- Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krma Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. *Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks*. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with bert. In *International Conference on Learning Representations*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*, Red Hook, NY, USA. Curran Associates Inc.

## A Dataset Statistics

Tables 7 and 8 present BALSAM Phase 1 and 2 benchmark dataset statistics, respectively.

No.	Category	Task	Test	Dev
1	Creative Writing	Definition Generation	22	22
		Dialogue Generation	146	65
		Explanation	64	21
		Instruction Generation	10	4
		Misc.	21	9
		News Article Generation	12	12
		Poem Generation	25	9
		Question Generation	1146	483
		Question Rewriting	48	20
		Sentence Composition	235	94
		Sentence Compression	21	10
		Story Composition	430	207
		Subject Generation	497	232
Text Completion	119	46		
Text Generation	130	92		
Wrong Candidate Generation	233	93		
2	Entailment	Textual Entailment	14	13
3	Fill in the Blank	Fill in The Blank	31	10
4	Information Extraction	Coreference Resolution	18	7
		Disease Mention Identification	10	9
		Keyword Extraction	47	43
		Named Entity Recognition	161	74
		Question Understanding	22	10
		Relation Extraction	10	9
5	Logic	Extracting Required Information	335	146
		Cause Effect Classification	39	18
		Coreference Resolution	13	6
		Misc.	69	29
		Predictive Analysis	10	10
		Riddle Solving	48	25
6	Translation/Transliteration	Sentence Ordering	18	8
		Dialect Translation	1200	600
		Machine Translation	1810	646
7	Program Execution	Transliteration	220	220
		Program Execution	646	268
		Answering Given Question	2600	1484
8	Question Answering	Question Decomposition	10	2
		Reading Comprehension	492	218
9	Reading Comprehension	Grammar Detection	277	129
		Keyword Extraction	58	20
10	Sequence Tagging	Text Summarization	618	399
		Answer Extraction	10	5
		Subject Generation	10	3
		Subject Identification	10	8
		Topic Identification	23	18
11	Summarization	Command Interpretation	23	23
		Dialect Identification	27	27
		Emotion Detection	10	9
		Intent Classification	10	4
		Offensive Language Detection	21	11
		Problem Identification	10	8
		Sarcasm Detection	17	12
		Sentiment Analysis	10	2
		Text Categorization	56	23
12	Text Classification	Gender Rewriting	347	119
		Grammar Correction	269	202
		Intent Classification	18	5
		Paraphrasing	117	58
		Question Rewriting	100	34
		Text Simplification	98	41
13	Text Manipulation	Total	13,121	6,434

Table 7: BALSAM Phase 1 benchmark dataset statistics

No.	Category	Task	Test	Dev		
1	Creative Writing	Dialogue Generation	72	30		
		Explanation	25	10		
		Text Completion	50	20		
		Text Continuation Evaluation	10	10		
2	Entailment	Duplicate Question Identification	20	20		
		Semantic Similarity	150	150		
		Textual Entailment	305	150		
3	Factuality	Answer Verification	50	20		
		Answerability Classification	25	10		
		Claim Verification	170	95		
		Text Classification	100	49		
3	Fill in the Blank	Fill in The Blank	25	10		
		Discourse Connective Identification	10	4		
4	Information Extraction	Disease Mention Identification	10	10		
		Named Entity Recognition	10	10		
		Entity Categorization	10	10		
		Entity Recognition and Gender Identification	30	30		
		Entity Relation Classification	25	10		
		Extracting Required Information	35	20		
		Text Classification	188	44		
5	Logic	Cause Effect Classification	350	175		
		Coherence Classification	50	20		
		Commonsense Validation	130	80		
		Evidence Evaluation	50	25		
		Logical Reasoning	30	30		
		Natural Language Inference	35	35		
6	Translation/Transliteration	Machine Translation	12890	3225		
7	Program Execution	Program Execution	25	10		
8	Question Answering	Answering Given Question	4979	2117		
9	Reading Comprehension	Answer Verification	25	10		
		Answerability Classification	75	30		
		Question Understanding	25	10		
		Reading Comprehension	350	250		
10	Sequence Tagging	Sequence Tagging	100	25		
		Dialect Identification	490	228		
		Dialogue Act Recognition	25	10		
		Emotion Detection	100	100		
		Ethics Classification	50	20		
		Hate Speech Detection	80	80		
		Offensive Language Detection	200	110		
		Query Classification	50	24		
		Question Categorization	10	10		
		Question Understanding	25	10		
		Review Rating Prediction	30	30		
		Sarcasm Detection	70	70		
		Sentiment Analysis	605	509		
		Text Categorization	235	110		
		Text Classification	1584	983		
		Topic Identification	10	10		
		13	Text Manipulation	Diacritization	300	250
		<b>Total</b>			<b>24298</b>	<b>9,308</b>

Table 8: BALSAM Phase 2 benchmark dataset statistics.

## B Examples of Samples

Figure 2 shows examples of some prompts and responses for the different categories.

Category/Dataset	Sample Input	Output
<b>Creative Writing</b> task568_circa_question_generation	في هذه المهمة ، ستعمل جونا ، ومهندتا هي اقترح سؤال مقبول لتلك الإجابة. أذهب السيتيما كثيرا	هل تحب الذهاب للسيتيما؟
<b>Entailment</b> enli_explain_the_choice	المرح لماذا يمكن وصف العلاقة بين الجنسين التاليين بأنها ناقص. الجملة 1: رجل عازي الصناد لديه وشم وسروال أسود وحذاء أسود يقف على منصة أمام مرآة. الجملة 2: الرجل مستلقي على السطح	إذا كان الرجل يقف أمام المرآة، فإنه لا يمكن أن يكون مستلقيا على السطح.
<b>Fill in the Blank</b> task1217_atomic_answer_generation	في هذه المهمة، ستلحق جملة تحتوي على كلمة مفقودة يمكن أن تكون جسيماً أو شخصاً، و / أو فعلاً. قم بملء الفراغ بكلمة مناسبة. على الرغم من وجود العديد من الإجابات الصحيحة لكل جملة، يجب عليك كتابة إجابة واحدة فقط. "بعض أكن" سيرحل قريباً	إلى القمر.
<b>Information Extraction</b> task1479_organization_entity_extraction_bt c_corpus	في هذه المهمة، سوف تعطى جملة من الموقع أن تعرف اسم المنظمة. على الرغم من أنه قد يكون هناك عدة اجوبة صحيحة ، يلزم ان تكتب واحدا منها. قطعة عظيمة لكريستين ميلر في عمارة اليوم. على الناس أن يدركوا ان الحزب الليبرالي يخطط للتسجيل على الموقع <a href="http://tco/NRCHlcv51">http://tco/NRCHlcv51</a>	الحزب الليبرالي.
<b>Logic</b> ropes.plain_no_background	كانت مياه بحيرة الارب متلظفة هذا الربيع على الرغم من ان البحيرة عادة ما تكون على مستوى كامل في اواخر الربيع. كان الحراس يخططون لترك الماء خارج البحيرة في التعريف حتى لا يتلف الجليد المرادق وغيرها من الهياكل حول البحيرة. وكانت مستويات المياه مرتفعة بما فيه الكفاية في الخريف، فبدأت في خفض مستوى المياه بعد اار 10 اقدم، وانظر الى الصورة الوردية اعلاه، يرجى الإجابة على السؤال التالي على النحو الصحيح: متى تنخفض مستويات المياه في الربيع أو خلال الصيف؟	خلال الصيف
<b>Program Execution</b> task852_synthetic_multiply_odds	في هذه المهمة سوف تحصل على قائمة من القوائم تحتوي على ارقام صحيحة. لكل قائمة داخلية في قائمة داخلية، يجب أن تتربط كل ارقام الفردية في تلك القائمة. ينبغي ان يكون الناتج قائمة بالاعداد الصحيحة بنفس طول عدد القوائم في قائمة المدخلات. إذا لم يكن هناك ارقام مفردة في قائمة داخلية يجب ان تخرج 0 لتلك القائمة. [[36, 23, 41], [27, 46, 37, 31, 24], [15, 50], [38, 2, 3]]	[3, 5, -30969, -943]
<b>Question Answering</b> arcd.qa	خضعت إسبانيا بسبب موقعها لمؤثرات خارجية كثيرة منذ صعود ما قبل التاريخ وحتى بزوغها كدولة. خرجت إسبانيا كدولة موحد في القرن الخامس عشر وذلك بعد توحيد الممالك الكاثوليكية والسجيرة على كامل شبه الجزيرة الأيبيرية في 1492. من ناحية أخرى، كانت إسبانيا مصدر نفوذ هام في مناطق أخرى خلال العصور الحديثة، عندما أصبحت إمبراطورية عالمية خلفت إرثا يضم أكثر من 500 مليون ناطق بالإسبانية، مما يجعلها ثاني اللغات الأي المستخدمة. وفقا للنص السابق، متى خرجت إسبانيا كدولة موحدة؟	في القرن الخامس عشر
<b>Reading Comprehension</b> xp3_mliqa_mlaaatar_test_cqreferrirga	رود بوي هي أفتية سجنائها المغنية البرازيلية رينا الألبومها الإيستديو الرابع، رينيد آر (2009). صدرت كأغنية منفردة تالفة من الألبوم يوم 19 فبراير، 2010، من خلال تسجيلات ديف جام. الأغنية من كتابة مايكل إس أركيس، وكور إريك هورمانسن، وأستر دين، وماكيا رينيك، وروب سوانز، ورون هيني، ومن إنتاج ستايفيت وروب سوانز. "رود بوي" هي أغنية ذات إيقاع عالي وتحتوي على عناصر من البوب، والروب، والآر أند بي.	19 فبراير، 2010.
<b>Sequence Tagging</b> spelling_error_dataset_card	بالإشارة إلى المقطع أعلاه ، فإن الإجابة الصحيحة على السؤال المعطى في لغة المقطع هي Q: متى تم إطلاق سراح رود بوي؟ A: كتب الكلمة التي تتضمن خطأ إملائي في النص التالي: وأشارت مصادر سياسية قريبة من الاجتماعات إلى أن الفجوة ما زالت كبيرة بين السياسيين الليبراليين المناهضين لسوريا والمليدين بها. يجب أن يقتصر الجواب على ذكر الخطأ من النص فقط بدون أي تفاصيل.	المليدين
<b>Summarization</b> hellaswag:Topic_of_the_context	طفل يجلس على طاولة مطعم يأكل آيس كريم القانيليا هل يمكنك تحديد موضوع الفقرة؟	تناول الآيس كريم
<b>Text Classification</b> aelineima_Saudilirony_editedCNLSirony1	هل هذه الجملة تهكم أم ليست تهكم؟ ينبغي يقول نازي القيصومه يس جاب العبد	ليست جملة تهكم
<b>Text Manipulation</b> task927_yelp_negative_to_positive_style_tr ansfer	لقد أعطيت تقييما سلبيا ومهمكم هي تحويله الى تقييم ايجابي واسطه عمل اقل تغيير ممكن. تجنب تغيير سياق التقييم.	يمكنهم ان يشكروك على تقييمك الجيد.
<b>Translation-Transliteration</b> arbnl_Arabizi_Transliteration.translation	ترابها	mazleha

Figure 2: Samples from different categories.



## C Examples of Prompts

Figure 3 shows some prompt templates that we used to create some of the datasets.

حدد إن كانت التغريدة الآتية: “{{text}}” عادية أم مزعجة هذه التغريدة {{answer}}
Translation: Specify if the following tweet: “{{text}}” is normal or spam The tweet is {{answer}}
أريد تصنيف التغريدة الآتية: “{{text}}” لمعرفة إذا كانت عادية أم مزعجة التغريدة التي ذكرتها {{answer}}
Translation: I want to classify the following tweet: “{{text}}” to know if it is normal or spam The tweet you provided is {{answer}}
“{{text}}” بالنسبة للرسالة السابقة، هل هي عادية أم دعائية الرسالة {{answer}}
Translation: “{{text}}” concerning the preceding message, is it normal or an advert The message is {{answer}}
وصلتني الرسالة الآتية: “{{text}}” يا ترى هل هي "عادية" أم "غير مرغوب فيها" أظن أن الرسالة {{answer}}
Translation: I received the following message: “{{text}}” I wonder if it is normal or unsolicited The think the message is {{answer}}

Figure 3: Example prompts for the Arabic tweet classification task.

## D Models

### D.1 Open-source Models

- **AceGPT-v2-8B-Chat (Huang et al., 2024)**: A fine-tuned Arabic dialogue model based on LLaMA2, designed for chat-style interactions in Arabic.
- **Aragpt2-mega (1.5B) (Antoun et al., 2021)**: A large-scale Arabic GPT-2 model designed for generating and understanding Arabic text.
- **c4ai-aya-expansive-32b (Dang et al., 2024)**: A multilingual large language model supporting 23 languages, including Arabic, with strong performance across diverse tasks.
- **Command R+ (104B):<sup>5</sup>** A multilingual model optimized for retrieval-augmented generation (RAG), reasoning, and task completion, with general Arabic support.
- **Command-r7b 12-2024:<sup>6</sup>** A compact and efficient version of the Command family of models, designed for general-purpose instruction following and language generation.
- **DeepSeek V3 (685B) (DeepSeek-AI, 2024)**: A multilingual Mixture-of-Experts model for reasoning, coding, and language understanding.
- **Gemma2 9B (Gemma Team et al., 2024)**: A multilingual language model from Google.
- **Jais-family 13b-chat (Sengupta et al., 2023)**: A bilingual Arabic-English model trained on 395B tokens, optimized for long-sequence handling.

<sup>5</sup><https://huggingface.co/CohereLabs/c4ai-command-r-plus>

<sup>6</sup><https://huggingface.co/CohereLabs/c4ai-command-r7b-12-2024>

- **qwen-2.5 32b (Yang et al., 2024):** A high-capacity language model with strong performance in Chinese and English and expanding capabilities in other languages, including Arabic.
- **SILMA-9B Instruct-v1.0:**<sup>7</sup> A 9-billion-parameter Arabic language model built on Google’s Gemma architecture, fine-tuned for instruction-following tasks.
- **Yehia-7B preview:**<sup>8</sup> A bilingual model designed for Arabic and English, capable of instruction-following and engaging in natural dialogue.
- **Fanar (Fanar Team et al., 2025):** It comes with two 7B and 9B parameter LLMs trained on nearly 1 trillion tokens. The models are designed to support both Arabic and English.
- **Mistral large :**<sup>9</sup> A multilingual model with 123B parameters by Mistral AI.
- **DBRX-instruct (132B):**<sup>10</sup> An instruction-tuned transformer developed by Databricks for high-quality reasoning and generation.

## D.2 Closed-Source Models

- **Nuha v2 :**<sup>11</sup> Nuha is an advanced, culture-aware AI assistant infused with pre-training and fine-tuning to understand Arabic nuances. With Nuha that is 40B parameter .
- **Iron Horse Gamma Velorum V5a:**<sup>12</sup> A closed-source MoE model with 1.1T - 2.3T parameters based on the request. It supports more than 25 language. .
- **Amazon Nova Pro (et al., 2025)** A multilingual model by Amazon Bedrock designed for commercial applications.
- **Mistral-saba-latest (24B):**<sup>13</sup> An Arabic fine-tuned variant of the Mistral model.
- **Grok-2-latest (314B MoE):** A closed-source model by xAI, designed for reasoning and factual recall.
- **Claude Sonnet 3.5:**<sup>14</sup> A multilingual and instruction-capable model by Anthropic, estimated at over 130B parameters.
- **Gemini 2.0 Flash:** A lightweight variant of Gemini 2.0 optimized for speed and extended context.
- **GPT-4o (OpenAI et al., 2024):** OpenAI’s model supporting multimodal and multilingual input, including Arabic.

<sup>7</sup><https://huggingface.co/silma-ai/SILMA-9B-Instruct-v1.0>

<sup>8</sup><https://huggingface.co/Navid-AI/Yehia-7B-preview>

<sup>9</sup><https://huggingface.co/mistralai/Mistral-Large-Instruct-2407>

<sup>10</sup><https://huggingface.co/databricks/dbrx-instruct>

<sup>11</sup><https://nuha.ai/>

<sup>12</sup><https://www.ironhorse.ai/>

<sup>13</sup><https://mistral.ai/news/mistral-saba>

<sup>14</sup><https://www.anthropic.com/news/claude-3-5-sonnet>

## E Example LLM Outputs with the Same Meaning

Consider the following Question Answering example where the correct answer is باريس (Paris). While SILMA 9B just answered with باريس only, the other model responses were much more verbose. Consider the answer of c4ai-aya-expanse-32b:

الجواب هو: باريس. وفقاً للفقرة، بدأت مرحلة تتابع الشعلة في فرنسا من باريس، حيث انعقدت في ٧ أبريل، وبدأت على المستوى الأول من برج إيفل، وانتهت في ملعب تشارلتي.

– **Translation:** *The answer is: Paris. According to the piece, the journey of the torch started in France from Paris on April 7 where it started from the first level of the Eiffel Tower and ended at the Charléty stadium.*

## F Prompts for LLM-Based Evaluation

Here is the prompt we used to extract the correct answer only from the LLM output:

```
""""Given the following prompt:

{prompt}

And the following automatically generated output:

{response}

Extract the answer from the automatically generated output ONLY WITHOUT any modification. Remove all non-related text from the answer. Do not put any additional text. If there are multiple answers, extract the first one only.

""""
```

Listing 1: Prompt for LLM-based answer extraction.

Here is the prompt that we used for LLM as a judge:

```
You are an impartial and expert judge evaluating the quality of text generated by another AI model. Your task is to score the generated output based on the original prompt and a provided ground truth answer, following a specific scoring rubric.

You will be provided with three pieces of information:

1. The original prompt given to the generative model.
2. The ground truth answer, representing the ideal or expected output.
3. The actual output generated by the generative model.

Evaluate the generated output by comparing it to the ground truth, considering how well it addresses the original prompt.

Scoring Rubric:

* Score 0: The automatically generated output is completely wrong, irrelevant, or unrelated to the prompt and ground truth.
* Score 1: Poor answer. The output attempts to address the prompt but contains significant errors, is largely incomplete, or is difficult to understand. It shows little resemblance to the ground truth.
* Score 2: Acceptable but different. The output is somewhat correct or addresses parts of the prompt reasonably well, but it differs significantly from the ground truth. It might be missing details present in the ground truth, include extra information not in the ground truth, or present the information in a substantially different structure or style, but it is still a valid (though not ideal) response to the prompt.
* Score 3: Perfect or almost perfect. The output is accurate, complete, and closely matches the ground truth in content and style, effectively answering the original prompt. Minor differences in wording or formatting that do not affect the meaning or quality are acceptable for a score of 3.

Output Format:

Your output must be *only* a JSON object containing two keys:

1. `score`: An integer between 0 and 3 based on the rubric above.
2. `explanation`: A brief, concise string explaining *why* you assigned that score, referencing the differences or similarities between the generated output and the ground truth in the context of the prompt.

Example Output JSON:

{
  "score": 3,
```

```

"explanation": "The generated output is accurate and complete, closely matching the ground truth."
}

[PROMPT]
{prompt}
[/PROMPT]

[GROUND TRUTH]
{reference answer}
[/GROUND TRUTH]

[GENERATED OUTPUT]
{response}
[/GENERATED OUTPUT]

```

Listing 2: LLM-as-a-Judge prompt.

## G Human Evaluation Annotation Instructions

- هل الإجابة صحيحة عند مقارنتها مع الإجابة الأصلية (٠-٣)؟
- ٠: إجابة خاطئة تماما (لا تتطابق مع الإجابة الأصلية بأي شكل).
  - ١: إجابة خاطئة جزئيا (تحتوي على بعض العناصر الصحيحة ولكن بها أخطاء جوهرية).
  - ٢: إجابة صحيحة جزئيا (تعكس بعض المعنى الصحيح ولكنها تفتقر إلى الدقة أو التفاصيل المهمة).
  - ٣: إجابة صحيحة تماما (متطابقة أو مكافئة للإجابة الأصلية دون أي أخطاء)

### Translation of instructions:

Is the answer correct when compared to the original answer (0-3)?

- 0: Completely wrong (does not match the original answer in any way).
- 1: Partially wrong (contains some correct elements but has significant errors).
- 2: Partially correct (conveys some correct meaning but lacks accuracy or important details).
- 3: Completely correct (identical or equivalent to the original answer with no errors).