

ArabicNLP 2025

The Third Arabic Natural Language Processing Conference

Volume 2. Proceedings of the ArabicNLP 2025 Shared Tasks

November 8-9, 2025

The ArabicNLP organizers gratefully acknowledge the support from the following sponsors.

Platinum



Gold



©2025 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-356-2

Introduction

Welcome to the proceedings of the Shared Tasks at the **Third Arabic Natural Language Processing Conference (ArabicNLP 2025)**, co-located with **EMNLP 2025** in Suzhou, China, November 8–9, 2025. This volume represents a major achievement in collaborative Arabic NLP research, bringing together a total of **138 papers from 11 shared tasks** — including **11 overview papers** and **127 system description papers** from participating teams worldwide. This substantial collection reflects the growing vitality and maturity of the Arabic NLP research community and showcases the field’s expansion into increasingly diverse and sophisticated challenges.

This collection highlights the community’s growing interest toward evaluating, aligning, and extending these models for Arabic — across text, speech, and multimodal domains, as well as culturally grounded and ethically sensitive applications. This year marks a record milestone for ArabicNLP, featuring the largest number of shared tasks to date — **11 in total**. The process began with **26 submitted shared task proposals**, from which **11 shared tasks were accepted**, including **6 formed through successful mergers**. This outcome underscores the community’s strong spirit of collaboration and its collective effort to design impactful, high-quality shared tasks.

Organization and Participation of the Shared Tasks

The 11 shared tasks at ArabicNLP 2025 are organized into three thematic tracks, each addressing critical needs and emerging priorities in Arabic language technology. **Together, they represent the community’s most comprehensive shared task effort to date, covering speech, multimodal processing, text quality, and cultural and ethical evaluation in the era of large language models (LLMs).**

Track 1: Speech and Multimodal Processing This track features four shared tasks that advance Arabic language processing beyond traditional text:

- **ImageEval (Arabic Image Captioning)** — 8 papers
- **Iqra’Eval (Qur’anic Pronunciation Assessment)** — 5 papers
- **NADI 2025 (Multidialectal Arabic Speech Processing)** — 7 papers
- **MAHED 2025 (Multimodal Detection of Hope and Hate Emotions in Arabic Content)** — 23 papers

These tasks address the pressing need for technologies that can process Arabic across different modalities and dialectal variations.

Track 2: Text Quality and Generation Assessment This track comprises four shared tasks focused on evaluating and enhancing Arabic text quality:

- **AraGenEval (Arabic Authorship Style Transfer and AI-Generated Text Detection)** — 16 papers
- **TAQEEM 2025 (Arabic Quality Evaluation of Essays in Multi-dimensions)** — 4 papers
- **BAREC 2025 (Arabic Readability Assessment)** — 17 papers
- **AraHealthQA 2025 (Comprehensive Arabic Health Question Answering)** — 15 papers

These tasks tackle essential challenges in automated assessment, generation evaluation, and domain-specific question answering.

Track 3: Cultural and Ethical Evaluation of LLMs for Arabic This track introduces three groundbreaking shared tasks designed to evaluate large language models’ understanding of Arabic culture and Islamic knowledge:

- **IslamicEval (Capturing LLMs’ Hallucination in Islamic Content)** — 7 papers
- **PalmX 2025 (Benchmarking LLMs on Arabic and Islamic Culture)** — 9 papers
- **QIAS 2025 (Islamic Inheritance Reasoning and Knowledge Assessment)** — 16 papers

These tasks represent a critical step toward ensuring that AI systems appropriately and accurately represent the cultural and religious contexts of Arabic-speaking communities.

Community Impact and Participation

The remarkable response to this year’s shared tasks — with **127 system submissions across the 11 tasks** — demonstrates the Arabic NLP community’s continued growth and dynamism. Participating teams represent a diverse range of institutions across multiple continents, including universities, research centers, and industry partners, reflecting both the global interest in Arabic NLP and the real-world relevance of these challenges.

The breadth of methodological approaches presented in these proceedings is particularly noteworthy, spanning traditional machine learning techniques, state-of-the-art transformer models, retrieval-augmented generation systems, and multimodal architectures. This methodological diversity not only reflects the field’s technical maturity but also highlights researchers’ creativity in addressing the unique challenges posed by Arabic language processing.

The breadth of methodological approaches presented in these proceedings is particularly noteworthy, spanning traditional machine learning techniques, state-of-the-art transformer models, retrieval-augmented generation systems, and multimodal architectures. This methodological diversity not only reflects the field’s technical maturity but also highlights researchers’ creativity in addressing the unique challenges posed by Arabic language processing.

Acknowledgments

We extend our sincere gratitude to the dedicated organizers of each shared task. Their expertise in task design, dataset development, evaluation framework creation, and ongoing participant support has been fundamental to the success of these shared tasks. The quality and impact of this volume are a direct reflection of their commitment and hard work.

We thank all participating teams whose innovative research pushes the boundaries of Arabic NLP. Their contributions address not only technical challenges but also important societal needs in education, healthcare, cultural preservation, content moderation, and ethical AI development.

We also acknowledge the program committee members, reviewers, and the ArabicNLP 2025 organizing committee for their invaluable support throughout the shared task cycle.

Finally, we gratefully acknowledge our platinum and gold sponsors whose support has made this conference and these proceedings possible.

We hope this volume serves as both a comprehensive record of the current state of Arabic NLP research and a foundation for future innovations that will continue to advance language technologies for Arabic-speaking communities worldwide.

Wajdi Zaghouni, Northwestern University in Qatar, Qatar.

Sakhar Alkhereyf, Humain, Saudi Arabia.

Shared Tasks Chairs

Website of the shared Tasks: <https://arabicnlp2025.sigarab.org/shared-tasks>

Organizing Committee

General Chair

Kareem Darwish, Qatar Computing Research Institute, Qatar

Program Chairs

Ahmed Ali, Humain, KSA

Ibrahim Abu Farha, Alsun AI, UK

Samia Touileb, University of Bergen, Norway

Imed Zitouni, Google, USA

Publication Chairs

Ahmed Abdelali, Humain, KSA

Sharefah Al-Ghamdi, King Saud University, KSA

Shared Tasks Chair

Sakhar Alkhereyf, Humain, KSA

Wajdi Zaghouani, Northwestern University in Qatar, Qatar

Publicity Chairs

Salam Khalifa, Stony Brook University, USA

Badr AlKhamissi, EPFL, Switzerland

Rawan Almatham, King Salman Global Academy for Arabic Language, Saudi Arabia

Scholarships and Awards Chairs

Injy Hamed, New York University Abu Dhabi, UAE

Zaid Alyafeai, KAUST, KSA

Sponsorship Chairs

Areeb Alowisheq, Humain, KSA

Imed Zitouni, Google, USA

Social Chairs

Go Inoue, MBZUAI, UAE

Khalil Mrini, TikTok, USA

Waad Alshammari, King Salman Global Academy for Arabic Language, Saudi Arabia

Program Committee

Reviewers

Abdelkader El Mahdaouy, Mohammed VI Polytechnic University
Abdellah El Mekki, University of British Columbia
AbdelRahim A. Elmadany, University of British Columbia
Abdelrhman Ahmed Yousry Elnainay, Alexandria University
Abdulaziz Alhamadani
Abdulkareem Alsudais, Prince Sattam bin Abdulaziz University
Abdalmohsen Al-Thubaity, Humain
Abdurahman Khalifa AAlAbdulsalam, Sultan Qaboos University
Abed Qaddoumi, State University of New York at Stony Brook
Abul Hasnat
Ahamed Rameez Mohamed Nizzad, British College of Applied Studies
Ahmad M Mustafa, Jordan University of Science and Technology
Ahmed Abdelali, Humain
Ahmed Taha, Whiterabbit.AI
Ahmed Wasfy
Ahmed Cherif Mazari, University of Médéa
Ahmed Oumar El-Shangiti, Mohamed Bin Zayed University of Artificial Intelligence
Alaa Aljabari, Birzeit University
Alexis Nasr, Aix Marseille University
Ali Al-Laith
Ali S. Al-Zawqari
Almoataz B. Al-Said
Aloulou Chafik, Univeristy of Sfax
Amel Muminovic, International Balkan University
Amir Hussein
Amr El-Gendy
Amr Keleg, University of Edinburgh, University of Edinburgh
Ann Bies, Linguistic Data Consortium, University of Pennsylvania
Ashraf Elnagar, Google
Ashwag Alasmari, King Khaled University
Attia Nehar
Badr M. Abdullah
Baraa Hikal
Bashar Alhafni, Mohamed bin Zayed University of Artificial Intelligence
Bashar Talafha, University of British Columbia
Caroline Sabty, German International University
Chaima Ben Rabah, weill cornell Medicine
Claudia Borg, University of Malta
David Corney, Full Fact
David M. Palfreyman, United Arab Emirates University
Duygu Altinok
El Moatez Billah Nagoudi, University of British Columbia
Elisa Gugliotta, CNR-Istituto di Linguistica Computazionale A. Zampolli"
Elsayed Issa, Purdue University
Enas Albasiri, NVIDIA
Eyob Nigussie Alemu, Addis Ababa University

Fadhl Eryani, Eberhard-Karls-Universität Tübingen
 Fadi Zaraket, Arab Center for Research and Policy Studies and American University of Beirut
 Fatima Haouari, University of Sheffield
 Fethi Bougaes, elyadata
 Firoj Alam, Qatar Computing Research Institute
 Ghassan Mourad, Lebanese University
 Go Inoue, Mohamed bin Zayed University of Artificial Intelligence
 Hadda Cherroun, Université Amar Telidji
 Hamzah Luqman, King Fahad University of Petroleum and Minerals
 Hoda Zaiton, Pharos University in Alexandria
 Hossam Ahmed, Leiden University
 Houda Bouamor, Carnegie Mellon University
 Ibrahim Bounhas
 Injy Hamed, Mohamed bin Zayed University of Artificial Intelligence
 Irfan Ahmad, King Fahad University of Petroleum and Minerals
 Ismail Berrada, Mohammed VI Polytechnic University
 Kamel Gaanoun, Institut National de Statistiques et d'Economie Appliquées
 Kedir Yassin Hussen
 Khalil Hennara
 Khloud Al Jallad, HIAST
 Kurt Micallef, University of Malta
 Maged Al-shaibani, SDAIA-KFUPM Joint Research Center for Artificial Intelligence
 Majd Hawasly
 Malik H. Altakrori, IBM TJ Watson Research Center
 Marwan Torki, Alexandria University
 Mayar Nassar, Ain Shams University
 Minh Ngoc Ta
 Mohamed Lichouri, Université des Sciences et de la Technologie Houari Boumediène
 Mohamed Nabih, Fondazione Bruno Kessler
 Mohamed Bayan Kmainasi, University of Qatar
 Mohamed Motasim Hamed
 Mohammed Attia, Google
 Mohammed Salah Al-Radhi, Budapest University of Technology and Economics
 Mona Abdelazim, Ain Shams University
 Mouath Abu Daoud
 Moustafa Wassel
 Muhammad Shakeel, Honda Research Institution Japan Co., Ltd.
 Muhammed AbuOdeh, New York University, Abu Dhabi
 Mustafa Jarrar, Birzeit University
 Nada Ghneim
 Nada Sharaf, The German International University
 Nizar Habash, New York University Abu Dhabi
 Omar Trigui, Institut Supérieur de Gestion de Sousse
 Omer Goldman, Bar Ilan University
 Omer Nacar
 Pagon Gatchalee
 Panigrahi Srikanth
 Pavel Denisov, Fraunhofer IAIS and University of Stuttgart
 Peter Sullivan, University of British Columbia
 Petr Zemánek, Charles University Prague
 Preslav Nakov, Mohamed bin Zayed University of Artificial Intelligence

Rania Al-Sabbagh
 Rania Azad M. San Ahmed, Sulaimani Polytechnic University, Sulaymaniyah, Iraq
 Saddam Al-Azani, King Fahad University of Petroleum and Minerals
 Saeed Ahmadnia, University of Illinois at Chicago
 Saied Alshahrani, University of Bisha
 Sakhar Alkhereyf, Humain
 Salam Albatarni
 Salam Khalifa, New York University and State University of New York, Stony Brook
 Salima Harrat
 Salima Mdhaffar, Université d'Avignon
 Samhaa R. El-Beltagy
 Sanaa Kaddoura
 Seid Muhie Yimam, Universität Hamburg
 Serry Sibae, Prince Sultan University
 Shah Nawaz, Johannes Kepler Universität Linz
 Sharif Ahmed, University of Central Arkansas
 Simran Tiwari, Mendel Health Inc
 Slimane Bellaouar
 Sohaila Eltanbouly, University of Qatar
 Sultan Alrowili, IBM Research
 Suveyda Yeniterzi, GenAIus Technologies
 Taha Zerrouki
 Tamer Elsayed, Qatar University
 Usman Nawaz, University of Palermo, Italy
 Vincent Koc, Comet ML, The University of Queensland and Massachusetts Institute of Technology
 Violetta Cavalli-Sforza
 Waad Thuwaini Alshammari, King Salman Global Academy for Arabic Language
 Wajdi Zaghouani, Northwestern University
 Waseem Safi, Damascus University
 Wasif Feroze
 Watheq Mansour
 Wissam Antoun
 Yassine El Kheir
 Youssef Al Hariri, Edinburgh University, University of Edinburgh
 Yuchen Zhang, University of Essex
 Zahra Bokaei
 Zaid Alyafeai, King Abdullah University of Science and Technology
 Ziani Amel, Chadli Benjedid University
 Ömer Tarik Özyilmaz

Invited Speaker

Houda Bouamor
 Areeb Alowisheq

Table of Contents

<i>The AraGenEval Shared Task on Arabic Authorship Style Transfer and AI Generated Text Detection</i> Shadi Abudalfa, Saad Ezzini, Ahmed Abdelali, Hamza Alami, Abdessamad Benlahbib, Salmane Chafik, Mo El-Haj, Abdelkader El Mahdaouy, Mustafa Jarrar, Salima Lamsiyah and Hamzah Luqman	1
<i>MISSION at AraGenEval Shared Task: Enhanced Arabic Authority Classification</i> Thamer Maseer Alharbi	14
<i>Nojoom.AI at AraGenEval Shared Task: Arabic Authorship Style Transfer</i> Hafsa Kara Achira, Mourad Bouache and Mourad Dahmane	18
<i>LMSA at AraGenEval Shared Task: Ensemble-Based Detection of AI-Generated Arabic Text Using Multilingual and Arabic-Specific Models</i> Kaoutar Zita, Attia Nehar, Abdelkader Khelil, Slimane Bellaouar and Hadda Cherroun	26
<i>Amr&MohamedSabaa at AraGenEval shared task: Arabic Authorship Identification using Term Frequency – Inverse Document Frequency Features with Supervised Machine Learning</i> Amr Sabaa and Mohamed Sabaa	32
<i>NLP_wizard at AraGenEval shared task: Embedding-Based Classification for AI Detection and Authorship Attribution</i> Mena Hany	37
<i>PTUK-HULAT at AraGenEval Shared Task: Fine-tuning XLM-RoBERTa for AI-Generated Arabic News Detection</i> Tasneem Duridi, Areej Jaber and Paloma Martínez	42
<i>ANLPers at AraGenEval Shared Task: Descriptive Author Tokens for Transparent Arabic Authorship Style Transfer</i> Omer Nacar, Mahmoud Reda, Serry Sibae, Yasser Alhabashi, Adel Ammar and Wadii Boulila	49
<i>Athership at AraGenEval Shared Task: Identifying Arabic Authorship with a Dual-Model Logit Fusion</i> Mohamed Amin, Mahmoud Rady, Mariam Hossam, Sara Gaballa, Eman Samir, Maria Bassem, Nisreen Hisham and Ayman Khalafallah	54
<i>Sebaweh at AraGenEval Shared Task: BERENSE - BERT based ENSEmbler for Arabic Authorship Identification</i> Muhammad Helmy, Batool Najeh Balah, Ahmed Mohamed Sallam and Ammar Sherif	59
<i>CUET-NLP_Team_SS306 at AraGenEval Shared Task: A Transformer-based Framework for Detecting AI-Generated Arabic Text</i> Sowrav Nath, Shadman Saleh, Kawsar Ahmed and Mohammed Moshiul Hoque	65
<i>BUSTED at ARATECT Shared Task: A Comparative Study of Transformer-Based Models for Arabic AI-Generated Text Detection</i> Ali Zain, Sareem Farooqui and Muhammad Rafi	72
<i>CIOL at AraGenEval shared task: Authorship Identification and AI Generated Text Detection in Arabic using Pretrained Models</i> Sadiah Tasnim Meem and Azmine Tushik Wasi	77

Osint at AraGenEval shared task: Fine-Tuned Modeling for Tracking Style Signatures and AI Generation in Arabic Texts

Shifali Agrahari, Hemanth Prakash Simhadri, Ashutosh Kumar Verma and Ranbir Singh Sanasam

82

MarsadLab at AraGenEval Shared Task: LLM-Based Approaches to Arabic Authorship Style Transfer and Identification

Md. Rafiul Biswas, Mabrouka bessghaier, Firoj Alam and Wajdi Zaghouani 88

REGLAT at AraGenEval shared task: Morphology-Aware AraBERT for Detecting Arabic AI-Generated Text

Mariam Labib, Nsrin Ashraf, Mohammed Aldawsari and Hamada Nayel 94

Jenin at AraGenEval Shared Task: Parameter-Efficient Fine-Tuning and Layer-Wise Analysis of Arabic LLMs for Authorship Style Transfer and Classification

Huthayfa Malhis, Mohammad Tami and Huthaifa I. Ashqar 99

AraHealthQA 2025: The First Shared Task on Arabic Health Question Answering

Hassan Alhuzali, Farah E. Shamout, Muhammad Abdul-Mageed, Chaimae Abouzahir, Mouath Abu Daoud, Ashwag Alasmari, Walid Al-Eisawi, Renad Al-Monef, Ali Alqahtani, Lama Ayash, Nizar Habash and Leen Kharouf 107

NYUAD at AraHealthQA Shared Task: Benchmarking the Medical Understanding and Reasoning of Large Language Models in Arabic Healthcare Tasks

Nouar AlDahoul and Yasir Zaki 119

MedLingua at MedArabiQ2025: Zero- and Few-Shot Prompting of Large Language Models for Arabic Medical QA

Fatimah Mohamed Emad Elden and Mumina Ab. Abukar 126

Sakinah-AI at MentalQA: A Comparative Study of Few-Shot, Optimized, and Ensemble Methods for Arabic Mental Health Question Classification

Fatimah Mohamed Emad Elden and Mumina Ab. Abukar 140

MindLLM at AraHealthQA 2025 Track 1: Leveraging Large Language Models for Mental Health Question Answering

Nejood Abdulaziz Bin Eshaq 149

Quasar at AraHealthQA Track 1 : Leveraging Zero-Shot Large Language Models for Question and Answer Categorization in Arabic Mental Health

Adiba Fairouz Chowdhury and Md Sagor Chowdhury 155

Binary_Bunch at AraHealthQA Track 1: Arabic Mental Health Q&A Classification Using Data Augmentation and Transformer Models

Sajib Bhattacharjee, Ratnajit Dhar, Kawsar Ahmed and Mohammed Moshiul Hoque 164

!MSA at AraHealthQA 2025 Shared Task: Enhancing LLM Performance for Arabic Clinical Question Answering through Prompt Engineering and Ensemble Learning

Mohamed Younes, Seif Ahmed and Mohamed Basem 176

Sindbad at AraHealthQA Track 1: Leveraging Large Language Models for Mental Health Q&A

AbdulRahman A. Morsy, Saad Mankarious and Ayah Zirikly 184

Arabic Mental Health Question Answering: A Multi-Task Approach with Advanced Retrieval-Augmented Generation

Abdelaziz Amr AbdelAziz, Mohamed Ahmed Youssef, Mamdouh Mohamed Koritam, Marwa Eldeeb and Ensaf Hussein 192

<i>AraMinds at AraHealthQA 2025: A Retrieval-Augmented Generation System for Fine-Grained Classification and Answer Generation of Arabic Mental Health Q&A</i>	
Mohamed Zaytoon, Ahmed Mahmoud Salem, Ahmed Sakr and Hossam Elkordi	198
<i>Fahmni at AraHealthQA Track 1: Multi-Agent Retrieval-Augmented Generation and Multi-Label Classification for Arabic Mental Health Q&A</i>	
Caroline Sabty, Mohamad Rasmy, Mohamed Eyad Badran, Nourhan Sakr and Alia El Bolock	204
<i>MedGapGab at AraHealthQA: Modular LLM Assignment for Gaps and Gabs in Arabic Medical Question Answering</i>	
Baraa Hikal	213
<i>Egyhealth at General Arabic Health QA (MedArabiQ): An Enhanced RAG Framework with Large-Scale Arabic Q&A Medical Data</i>	
Hossam Amer, Rawan Tarek Taha, Gannat Elsayed and Ensaf Hussein Mohamed	222
<i>mucAI at AraHealthQA 2025: Explain–Retrieve–Verify (ERV) Workflow for Multi-Label Arabic Health QA Classification</i>	
Ahmed Abdou	226
<i>MarsadLab at AraHealthQA: Hybrid Contextual–Lexical Fusion with AraBERT for Question and Answer Categorization</i>	
Mabrouka bessghaier, Shima Ibrahim, Md. Rafiul Biswas and Wajdi Zaghouani	233
<i>BAREC Shared Task 2025 on Arabic Readability Assessment</i>	
Khalid N. Elmadani, Bashar Alhafni, Hanada Taha and Nizar Habash	239
<i>Syntaxa at BAREC Shared Task 2025: BERTnParse - Fusion of BERT and Dependency Graphs for Readability Prediction</i>	
Ahmed Bahloul	253
<i>GNNinjas at BAREC Shared Task 2025: Lexicon-Enriched Graph Modeling for Arabic Document Readability Prediction</i>	
Passant Elchafei, Mayar osama, Mohamad Rageh and Mervat Abu-Elkheir	261
<i>ZAI at BAREC Shared Task 2025: AraBERT CORAL for Fine Grained Arabic Readability</i>	
Ahmad M. Nazzal	266
<i>ANLPers at BAREC Shared Task 2025: Readability of Embeddings Training Neural Readability Classifiers on the BAREC Corpus</i>	
Serry Sibae, Omer Nacar, Yasser Alhabashi, Adel Ammar and Wadii Boulila	269
<i>MarsadLab at BAREC Shared Task 2025: Strict-Track Readability Prediction with Specialized AraBERT Models on BAREC</i>	
Shima Ibrahim, Md. Rafiul Biswas, Mabrouka Bessghaier and Wajdi Zaghouani	274
<i>SATLab at BAREC Shared Task 2025: Optimizing a Language-Independent System for Fine-Grained Readability Assessment</i>	
Yves Bestgen	280
<i>MorphoArabia at BAREC Shared Task 2025: A Hybrid Architecture with Morphological Analysis for Arabic Readability Assessment</i>	
Fatimah Mohamed Emad Elden	286
<i>!MSA at BAREC Shared Task 2025: Ensembling Arabic Transformers for Readability Assessment</i>	
Mohamed Basem, Mohamed Younes, Seif Ahmed and Abdelrahman Moustafa	297

<i>Qais at BAREC Shared Task 2025: A Fine-Grained Approach for Arabic Readability Classification Using a pre-trained model</i>	
Samar Ahmad	306
<i>mucAI at BAREC Shared Task 2025: Towards Uncertainty Aware Arabic Readability Assessment</i>	
Ahmed Abdou	312
<i>AMAR at BAREC Shared Task 2025: Arabic Meta-learner for Assessing Readability</i>	
Mostafa Saeed, Rana Waly and Abdelaziz Ashraf Hussein	320
<i>Noor at BAREC Shared Task 2025: A Hybrid Transformer-Feature Architecture for Sentence-level Readability Assessment</i>	
Nour Rabih	331
<i>PalNLP at BAREC Shared Task 2025: Predicting Arabic Readability Using Ordinal Regression and K-Fold Ensemble Learning</i>	
Mutaz Ayesb	343
<i>Pixels at BAREC Shared Task 2025: Visual Arabic Readability Assessment</i>	
Ben Sapirstein	350
<i>Phantoms at BAREC Shared Task 2025: Enhancing Arabic Readability Prediction with Hybrid BERT and Linguistic Features</i>	
Ahmed Alhassan, Asim Mohamed and Moayad Elamin	357
<i>STBW at BAREC Shared Task 2025: AraBERT-v2 with MSE-SoftQWK Loss for Sentence-Level Arabic Readability</i>	
Saoussan Trigui	362
<i>LIS at BAREC Shared Task 2025: Multi-Scale Curriculum Learning for Arabic Sentence-Level Readability Assessment Using Pre-trained Language Models</i>	
Anya Amel Nait Djoudi, Patrice Bellot and Adrian-Gabriel Chifu	367
<i>ImageEval 2025: The First Arabic Image Captioning Shared Task</i>	
Ahlam Bashiti, Alaa Aljabari, Hadi Khaled Hamoud, Md. Rafiul Biswas, Bilal Mohammed Shalash, Mustafa Jarrar, Fadi Zaraket, George Mikros, Ehsaneddin Asgari and Wajdi Zaghoulani ..	376
<i>Codezone Research Group at ImageEval Shared-Task 2: Arabic Image Captioning Using BLIP and M2M100: A Two-Stage Translation Approach for ImageEval 2025</i>	
Abdulkadir Shehu Bichi	390
<i>BZU-AUM@ImageEval2025: An Arabic Image Captioning Dataset for Conflict Narratives with Human Annotation</i>	
Mohammed Alkhanafseh, Ola surakhi and Abdallah Abedaljalill	395
<i>ImpactAi at ImageEval 2025 Shared Task: Region-Aware Transformers for Arabic Image Captioning – A Case Study on the Palestinian Narrative</i>	
Rabee Al-Qasem and Mohannad Hendi	401
<i>VLCAP at ImageEval 2025 Shared Task: Multimodal Arabic Captioning with Interpretable Visual Concept Integration</i>	
Passant Elchafei and Amany Fashwan	408
<i>PhantomTroupe at ImageEval 2025 Shared Task: Multimodal Arabic Image Captioning through Translation-Based Fine-Tuning of LLM Models</i>	
Muhammad Abu Horaira, Farhan Amin, Sakibul Hasan, Md. Tanvir Ahammed Shawon and Muhammad Ibrahim Khan	414

<i>NU_Internship team at ImageEval 2025: From Zero-Shot to Ensembles: Enhancing Grounded Arabic Image Captioning</i>	
Rana Gaber, Seif Eldin Amgad, Ahmed Sherif Nasri, Mohamed Ibrahim Ragab and Ensaf Hussein Mohamed	419
<i>Averroes at ImageEval 2025 Shared Task: Advancing Arabic Image Captioning with Augmentation and Two-Stage Generation</i>	
Mariam Saeed, Sarah Elshabrawy, Abdelrahman Hagrass, Mazen Yasser and Ayman Khalafallah	432
<i>AZLU at ImageEval Shared Task: Bridging Linguistics and Cultural Gaps in Arabic Image Captioning</i>	
Sarah Yassine	438
<i>Iqra'Eval: A Shared Task on Qur'anic Pronunciation Assessment</i>	
Yassine El Kheir, Amit Meghanani, Hawau Olamide Toyin, Nada Almarwani, Omnia Ibrahim, Yousseif Ahmed Elshahawy, Mostafa Shahin and Ahmed Ali	443
<i>Hafs2Vec: A System for the IqraEval Arabic and Qur'anic Phoneme-level Pronunciation Assessment</i>	
Ahmed Ibrahim	453
<i>Phoneme-level mispronunciation detection in Quranic recitation using ShallowTransformer</i>	
Mohamed Nadhir Daoud and Mohamed Anouar Ben Messaoud	457
<i>ANPLers at IqraEval Shared task: Adapting Whisper-large-v3 as Speech-to-Phoneme for Qur'anic Recitation Mispronunciation Detection</i>	
Nour Qandos, Serry Sibae, Samar Ahmad, Omer Nacar, Adel Ammar, Wadii Boulila and Yasser Alhabashi	464
<i>AraS2P: Arabic Speech-to-Phonemes System</i>	
Bassam Mattar, Mohamed Fayed and Ayman Khalafallah	469
<i>Metapseud at Iqra'Eval: Domain Adaptation with Multi-Stage Fine-Tuning for Phoneme-Level Qur'anic Mispronunciation Detection</i>	
Ayman Mansour	475
<i>IslamicEval 2025: The First Shared Task of Capturing LLMs Hallucination in Islamic Content</i>	
Hamdy Mubarak, Rana Malhas, Watheq Mansour, Abubakr Mohamed, Mahmoud Fawzi, Majd Hawasly, Tamer Elsayed, Kareem Mohamed Darwish and Walid Magdy	480
<i>NUR at IslamicEval 2025 Shared Task: Retrieval-Augmented LLMs for Qur'an and Hadith QA</i>	
Serag Amin, Ranwa Aly, Yara Allam, Yomna Eid and Ensaf Hussein Mohamed	494
<i>BurhanAI at IslamicEval 2025 Shared Task: Combating Hallucinations in LLMs for Islamic Content; Evaluation, Correction, and Retrieval-Based Solution</i>	
Arij Al Adel, Abu Bakr Soliman, Mohamed Sakher Sawan, Rahaf Al-Najjar and Sameh Amin	503
<i>HUMAIN at IslamicEval 2025 Shared Task 1: A Three-Stage LLM-Based Pipeline for Detecting and Correcting Hallucinations in Quran and Hadith</i>	
Arwa Omayrah, Sakhar Alkhereyf, Ahmed Abdelali, Abdulmohsen Al-Thubaity, Jeril Kuriakose and Ibrahim AbdulMajeed	509
<i>TCE at IslamicEval 2025: Retrieval-Augmented LLMs for Quranic and Hadith Content Identification and Verification</i>	
Mohammed ElKoumy, Khalid Allam, Ahmad Tamer and Mohamed Alqablawi	515

<i>ThinkDrill at IslamicEval 2025 Shared Task: LLM Hybrid Approach for Qur'an and Hadith Question Answering</i>	
Eman Elrefai, Toka Khaled and Ahmed Soliman	528
<i>Burhan at IslamicEval: Fact-Augmented and LLM-Driven Retrieval for Islamic QA</i>	
Mohammad Basheer, Watheq Mansour, Abdulhamid Touma and Ahmad Qadeib Alban	534
<i>Isnad AI at IslamicEval 2025: A Rule-Based System for Identifying Religious Texts in LLM Outputs</i>	
Fatimah Mohamed Emad Elden	540
<i>MAHED Shared Task: Multimodal Detection of Hope and Hate Emotions in Arabic Content</i>	
Wajdi Zaghouani, Md. Rafiul Biswas, Mabrouka Bessghaier, Shima Ibrahim, George Mikros, Abul Hasnat and Firoj Alam	560
<i>NYUAD at MAHED Shared Task: Detecting Hope, Hate, and Emotion in Arabic Textual Speech and Multi-modal Memes Using Large Language Models</i>	
Nouar AlDahoul and Yasir Zaki	575
<i>NguyenTriet at MAHED Shared Task: Ensemble of Arabic BERT Models with Hierarchical Prediction and Soft Voting for Text-Based Hope and Hate Detection</i>	
Nguyen Minh Triet and Thìn Đăng Văn	585
<i>ANLPers at MAHED2025: From Hate to Hope: Boosting Arabic Text Classification</i>	
Yasser Alhabashi, Serry Sibae, Omer Nacar, Adel Ammar and Wadii Boulila	590
<i>LoveHeaven at MAHED 2025: Text-based Hate and Hope Speech Classification Using AraBERT-Twitter Ensemble</i>	
Nguyễn Thiên Bảo and Dang Van Thin	595
<i>CIC-NLP at MAHED 2025 TASK 1: Assessing the Role of Bigram Augmentation in Multiclass Arabic Hate and Hope Speech Classification</i>	
Tolulope Olalekan Abiola, Oluwatobi Joseph Abiola, Ogunleye Temitope Dasola, Tewodros Achamleh and Obiadoh Augustine Ekenedilichukwu	599
<i>TranTranUIT at MAHED Shared Task: Multilingual Transformer Ensemble with Advanced Data Augmentation and Optuna-based Hyperparameter Optimization</i>	
Trinh Tran Tran and Thìn Đăng Văn	603
<i>YassirEA at MAHED 2025: Fusion-Based Multimodal Models for Arabic Hate Meme Detection</i>	
Yassir El Attar	608
<i>AAA at MAHED Text-based Hate and Hope Speech Classification: A Systematic Encoder Evaluation for Arabic Hope and Hate Speech Classification</i>	
Ahmed Khalil Elzainy, Mohamed Amin, Ahmed Samir and Hazem Abdelsalam	615
<i>CUET-823 at MAHED 2025 Shared Task: Large Language Model-Based Framework for Emotion, Offensive, and Hate Detection in Arabic</i>	
Ratnajit Dhar and Arpita Mallik	620
<i>AraMinds at MAHED 2025: Leveraging Vision-Language Models and Contrastive Multi-task Learning for Multimodal Hate Speech Detection</i>	
Mohamed Zaytoon, Ahmed Mahmoud Salem, Ahmed Sakr and Hossam Elkordi	626
<i>DesCartes-HOPE at MAHED Shared task 2025: Integrating Pragmatic Features for Arabic Hope and Hate Speech Detection</i>	
Leila MOUDJARI, Hacène-Cherkaski Mélissa and Farah Benamara	632

<i>ANLP-UniSo at MAHED Shared Task: Detection of Hate and Hope Speech in Arabic Social Media based on XLM-RoBERTa and Logistic Regression</i>	
Yasmine El Abed, Mariem Ben Arbia, Saoussen Ben Chaabene and Omar Trigui	639
<i>REGLAT at MAHED Shared Task: A Hybrid Ensemble-Based System for Arabic Hate Speech Detection</i>	
Nsrin Ashraf, Mariam Labib, Tarek Elshishtawy and Hamada Nayel	645
<i>HTU at MAHED Shared Task: Ensemble-Based Classification of Arabic Hate and Hope Speech Using Pre-trained Dialectal Arabic Models</i>	
Abdallah Saleh and Mariam M Biltawi	651
<i>SmolLab_SEU at MAHED Shared Task: Do Arabic-Native Encoders Surpass Multilingual Models in Detecting the Nuances of Hope, Hate, and Emotion?</i>	
Md Abdur Rahman, Md Sabbir Dewan, Md. Tofael Ahmed Bhuiyan and Md Ashiqur Rahman	658
<i>Baoflowin502 at MAHED Shared Task: Text-based Hate and Hope Speech Classification</i>	
Nguyen Minh Bao and Dang Van Thin	665
<i>AyahVerse at MAHED Shared Task: Fine-Tuning ArabicBERT with Preprocessing for Hope and Hate Detection</i>	
Ibad-ur-Rehman Rashid and Muhammad Hashir Khalil	670
<i>MultiMinds at MAHED 2025: Multimodal and Multitask Approaches for Detecting Emotional, Hate, and Offensive Speech in Arabic Content</i>	
Riddhiman Debnath, Abdul Wadud Shakib and Md Saiful Islam	677
<i>joy_2004114 at MAHED Shared Task : Filtering Hate Speech from Memes using A Multimodal Fusion-based Approach</i>	
Joy Das, Alamgir Hossain and Mohammed Moshiul Hoque	683
<i>Quasar at MAHED Shared Task : Decoding Emotions and Offense in Arabic Text using LLM and Transformer-Based Approaches</i>	
Md Sagor Chowdhury and Adiba Fairouz Chowdhury	692
<i>CUET_Zahra_Duo@Mahed 2025: Hate and Hope Speech Detection in Arabic Social Media Content using Transformer</i>	
Walisa Alam, Mehreen Rahman, Shawly Ahsan and Mohammed Moshiul Hoque	700
<i>AraNLP at MAHED 2025 Shared Task: Using AraBERT for Text-based Hate and Hope Speech Classification</i>	
Wafaa S. El-Kassas and Enas A. Hakim Khalil	706
<i>Thinking Nodes at MAHED: A Comparative Study of Multimodal Architectures for Arabic Hateful Me-me Detection</i>	
Itbaan Safwan	712
<i>NADI 2025: The First Multidialectal Arabic Speech Processing Shared Task</i>	
Bashar Talafha, Hawau Olamide Toyin, Peter Sullivan, AbdelRahim A. Elmadany, Abdurrahman Juma, Amirbek Djanibekov, Chiyu Zhang, Hamad Alshehhi, Hanan Aldarmaki, Mustafa Jarrar, Nizar Habash and Muhammad Abdul-Mageed	720
<i>Munsit at NADI 2025 Shared Task 2: Pushing the Boundaries of Multidialectal Arabic ASR with Weakly Supervised Pretraining and Continual Supervised Fine-tuning</i>	
Mahmoud Salhab, Shameed Sait, Mohammad Abusheikh and Hasan Abusheikh	734
<i>Lahjati at NADI 2025 A ECAPA-WavLM Fusion with Multi-Stage Optimization</i>	
Sanad Albawwab and Omar Qawasmeh	740

<i>Saarland-Groningen at NADI 2025 Shared Task: Effective Dialectal Arabic Speech Processing under Data Constraints</i>	
Badr M. Abdullah, Yusser Al Ghussin, Zena Al-Khalili, Ömer Tarik Özyilmaz, Matias Valdenegro-Toro, Simon Ostermann and Dietrich Klakow	745
<i>MarsadLab at NADI Shared Task: Arabic Dialect Identification and Speech Recognition using ECAPA-TDNN and Whisper</i>	
Md. Rafiul Biswas, Kais Attia, Shima Ibrahim, Mabrouka bessghaier and Wajdi Zaghouani	752
<i>Abjad AI at NADI 2025: CATT-Whisper: Multimodal Diacritic Restoration Using Text and Speech Representations</i>	
Ahmad Ghannam, Naif Alharthi, Faris Alasmay, Kholood Al Tabash, Shouq Sadah and Lahouari Ghouti	757
<i>ELYADATA & LIA at NADI 2025: ASR and ADI Subtasks</i>	
Haroun Elleuch, Youssef Saidi, Salima Mdhaifar, Yannick Estève and Fethi Bougares	762
<i>Unicorn at NADI 2025 Subtask 3: GEMM3N-DR: Audio-Text Diacritic Restoration via Fine-tuning Multimodal Arabic LLM</i>	
Mohamed Lotfy Elrefai	767
<i>PalmX 2025: The First Shared Task on Benchmarking LLMs on Arabic and Islamic Culture</i>	
Fakhraddin Alwajih, Abdellah El Mekki, Hamdy Mubarak, Majd Hawasly, Abubakr Mohamed and Muhammad Abdul-Mageed	774
<i>Hamyar at PalmX2025: Leveraging Large Language Models to Improve Arabic Multiple-Choice Questions in Cultural and Islamic Domains</i>	
Walid Al-Dhabyani and Hamzah A. Alsayadi	790
<i>ISL-NLP at PalmX 2025: Retrieval-Augmented Fine-Tuning for Arabic Cultural Question Answering</i>	
Mohamed Gomaa and Nouredin Elmadany	797
<i>ADAPT-MTU HAI at PalmX 2025: Leveraging Full and Parameter-Efficient LLM Fine-Tuning for Arabic Cultural QA</i>	
Shehenaz Hossain and Haithem Afli	802
<i>CultranAI at PalmX 2025: Data Augmentation for Cultural Knowledge Representation</i>	
Hunzalah Hassan Bhatti, Youssef Ahmed, Md Arid Hasan and Firoj Alam	809
<i>MarsadLab at PalmX Shared Task: An LLM Benchmark for Arabic Culture and Islamic Civilization</i>	
Md. Rafiul Biswas, Shima Ibrahim, Kais Attia, Firoj Alam and Wajdi Zaghouani	818
<i>Star at PalmX 2025: Arabic Cultural Understanding via Targeted Pretraining and Lightweight Fine-tuning</i>	
Eman Elrefai, Esraa Khaled and Alhassan Ehab	825
<i>AYA at PalmX 2025: Modeling Cultural and Islamic Knowledge in LLMs</i>	
Jannatul Tajrin, Bir Ballav Roy and Firoj Alam	830
<i>Cultura-Arabica: Probing and Enhancing Arabic Cultural Awareness in Large Language Models via LoRA</i>	
Pulkit Chatwal and Santosh Kumar Mishra	837
<i>Phoenix at Palmx: Exploring Data Augmentation for Arabic Cultural Question Answering</i>	
Houdaifa Atou, Issam AIT YAHIA and Ismail Berrada	843

<i>QIAS 2025: Overview of the Shared Task on Islamic Inheritance Reasoning and Knowledge Assessment</i>	
Abdessalam Boucekif, Samer Rashwani, Emad Soliman Ali Mohamed, Mutaz Alkhatib, Heba Sbahi, Shahd Gaben, Wajdi Zaghouani, Aiman Erbad and Mohammed Ghaly	851
<i>NYUAD at QIAS Shared Task: Benchmarking the Legal Reasoning of LLMs in Arabic Islamic Inheritance Cases</i>	
Nouar AlDahoul and Yasir Zaki	861
<i>SHA at the QIAS Shared Task: LLMs for Arabic Islamic Inheritance Reasoning</i>	
Shatha Altammami	867
<i>ANLPers at QIAS: CoT for Islamic Inheritance</i>	
Serry Sibae, Mahmoud Reda, OMER NACAR, Yasser Alhabashi, Adel Ammar and Wadii Boulila	873
<i>N&N at QIAS 2025: Chain-of-Thought Ensembles with Retrieval-Augmented framework for Classical Arabic Islamic</i>	
Nourah Alangari and Nouf AlShenaifi	878
<i>HIAST at QIAS 2025: Retrieval-Augmented LLMs with Top-Hit Web Evidence for Arabic Islamic Reasoning QA</i>	
Mohamed Motasim Hamed, Nada Ghneim and Riad Sonbol	883
<i>QU-NLP at QIAS 2025 Shared Task: A Two-Phase LLM Fine-Tuning and Retrieval-Augmented Generation Approach for Islamic Inheritance Reasoning</i>	
Mohammad AL-Smadi	892
<i>Transformer Tafsir at QIAS 2025 Shared Task: Hybrid Retrieval-Augmented Generation for Islamic Knowledge Question Answering</i>	
Muhammad Abu Ahmad, Mohamad Ballout, Raia Abu Ahmad and Elia Bruni	899
<i>PuxAI at QIAS 2025: Multi-Agent Retrieval-Augmented Generation for Islamic Inheritance and Knowledge Reasoning</i>	
Nguyen Xuan Phuc and Thìn Đăng Văn	905
<i>Athar at QIAS2025: LLM-based Question Answering Systems for Islamic Inheritance and Classical Islamic Knowledge</i>	
Yossra Noureldien, Hassan Suliman, Farah Attallah, Abdelrazig Mohamed and Sara Abdalla	914
<i>ADAPT-MTU HAI at QIAS2025: Dual-Expert LLM Fine-Tuning and Constrained Decoding for Arabic Islamic Inheritance Reasoning</i>	
Shehenaz Hossain and Haithem Afli	923
<i>CVPD at QIAS 2025 Shared Task: An Efficient Encoder-Based Approach for Islamic Inheritance Reasoning</i>	
Salah Eddine Bekhouche, Abdellah Zakaria Sellam, Telli Hichem, Cosimo Distanto and Abdenour Hadid	929
<i>CIS-RG at QIAS 2025 Shared Task: Approaches for Enhancing Performance of LLM on Islamic Legal Reasoning and its Mathematical Calculations</i>	
Osama Farouk Zaki	935
<i>SEA-Team at QIAS 2025: Enhancing LLMs for Question Answering in Islamic Texts</i>	
Sanaa Alowaidi	940

<i>MorAI at QIAS 2025: Collaborative LLM via Voting and Retrieval-Augmented Generation for Solving Complex Inheritance Problems</i>	
Jihad R'baiti, Chouaib El Hachimi, Youssef Hmamouche and Amal Seghrouchni	947
<i>Gumball at QIAS 2025: Arabic LLM Automated Reasoning in Islamic Inheritance</i>	
Eman Elrefai, Mohamed Lotfy Elrefai and Aml Hassan Esmail	953
<i>Tokenizers United at QIAS-2025: RAG-Enhanced Question Answering for Islamic Studies by Integrating Semantic Retrieval with Generative Reasoning</i>	
Mayar Boghdady	960
<i>TAQEEM 2025: Overview of The First Shared Task for Arabic Quality Evaluation of Essays in Multi-dimensions</i>	
May Bashendy, Salam Albatarni, Sohaila Eltanbouly, Walid Massoud, Houda Bouamor and Tamer Elsayed	966
<i>ARxHYOKA at TAQEEM2025: Comparative Approaches to Arabic Essay Trait Scoring</i>	
Mohamad Alnajjar, Ahmad Almoustafa, Tomohiro Nishiyama, Shoko Wakamiya, Eiji Aramaki and Takuya Matsuzaki	977
<i>912 at TAQEEM 2025: A Distribution-aware Approach to Arabic Essay Scoring</i>	
Trong-Tai Dam Vu and Thìn Đăng Văn	983
<i>Taibah at TAQEEM 2025: Leveraging GPT-4o for Arabic Essay Scoring</i>	
Nada Almarwani, Alaa Alharbi and Samah Aloufi	989
<i>MarsadLab at TAQEEM 2025: Prompt-Aware Lexicon-Enhanced Transformer for Arabic Automated Essay Scoring</i>	
Mabrouka bessghaier, Md. Rafiul Biswas, Amira Dhouib and Wajdi Zaghouani	998

Program

Friday, November 8, 2024

08:45 - 08:30	<i>Welcome & SIGARAB Update</i>
09:30 - 08:45	<i>Beyond Resources: Building an Arabic NLP Ecosystem Rooted in Representation, Collaboration, and Responsibility, by Dr. Houda Bouamor</i>
09:15 - 10:30	<i>LLM Benchmarking & Development (1)</i>
10:30 - 11:00	<i>Coffee Break</i>
11:00 - 12:30	<i>LLM Benchmarking & Development (2)</i>
12:30 - 14:00	<i>Lunch Break</i>
14:00 - 14:30	<i>Multimodality</i>
14:30 - 15:30	<i>Shared Tasks (1)</i>

The AraGenEval Shared Task on Arabic Authorship Style Transfer and AI Generated Text Detection

Shadi Abudalfa, Saad Ezzini, Ahmed Abdelali, Hamza Alami, Abdessamad Benlahbib, Salmane Chafik, Mo El-Haj, Abdelkader El Mahdaouy, Mustafa Jarrar, Salima Lamsiyah and Hamzah Luqman

AraHealthQA 2025: The First Shared Task on Arabic Health Question Answering

Hassan Alhuzali, Farah E. Shamout, Muhammad Abdul-Mageed, Chaimae Abouzahir, Mouath Abu Daoud, Ashwag Alasmari, Walid Al-Eisawi, Renad Al-Monef, Ali Alqahtani, Lama Ayash, Nizar Habash and Leen Kharouf

BAREC Shared Task 2025 on Arabic Readability Assessment

Khalid N. Elmadani, Bashar Alhafni, Hanada Taha and Nizar Habash

ImageEval 2025: The First Arabic Image Captioning Shared Task

Ahlam Bashiti, Alaa Aljabari, Hadi Khaled Hamoud, Md. Rafiul Biswas, Bilal Mohammed Shalash, Mustafa Jarrar, Fadi Zaraket, George Mikros, Ehsaneddin Asgari and Wajdi Zaghoulani

Iqra'Eval: A Shared Task on Qur'anic Pronunciation Assessment

Yassine El Kheir, Amit Meghanani, Hawau Olamide Toyin, Nada Almarwani, Omnia Ibrahim, Yousseif Ahmed Elshahawy, Mostafa Shahin and Ahmed Ali

IslamicEval 2025: The First Shared Task of Capturing LLMs Hallucination in Islamic Content

Hamdy Mubarak, Rana Malhas, Watheq Mansour, Abubakr Mohamed, Mahmoud Fawzi, Majd Hawasly, Tamer Elsayed, Kareem Mohamed Darwish and Walid Magdy

Friday, November 8, 2024 (continued)

16:00 - 15:30 *Coffee Break*

14:30 - 15:30 *Shared Tasks (2)*

MAHED Shared Task: Multimodal Detection of Hope and Hate Emotions in Arabic Content

Wajdi Zaghoulani, Md. Rafiul Biswas, Mabrouka Bessghaier, Shima Ibrahim, George Mikros, Abul Hasnat and Firoj Alam

NADI 2025: The First Multidialectal Arabic Speech Processing Shared Task

Bashar Talafha, Hawau Olamide Toyin, Peter Sullivan, AbdelRahim A. Elmada-ny, Abdurrahman Juma, Amirbek Djanibekov, Chiyu Zhang, Hamad Alshehhi, Hanan Aldarmaki, Mustafa Jarrar, Nizar Habash and Muhammad Abdul-Mageed

PalmX 2025: The First Shared Task on Benchmarking LLMs on Arabic and Islamic Culture

Fakhraddin Alwajih, Abdellah El Mekki, Hamdy Mubarak, Majd Hawasly, Abubakr Mohamed and Muhammad Abdul-Mageed

QIAS 2025: Overview of the Shared Task on Islamic Inheritance Reasoning and Knowledge Assessment

Abdessalam Boucekif, Samer Rashwani, Emad Soliman Ali Mohamed, Mutaz Alkhatib, Heba Sbahi, Shahd Gaben, Wajdi Zaghoulani, Aiman Erbad and Mohammed Ghaly

TAQEEM 2025: Overview of The First Shared Task for Arabic Quality Evaluation of Essays in Multi-dimensions

May Bashendy, Salam Albatarni, Sohaila Eltanbouly, Walid Massoud, Houda Bouamor and Tamer Elsayed

17:00 - 18:00 *Poster presentations + Shared Task posters*

Saturday, November 9, 2024

08:45 - 08:30	<i>Welcome</i>
09:30 - 08:45	<i>From Benchmarks to the Real-World Impact: Arabic LLMs in Production, by Dr. Areeb Alowisheq</i>
09:30 - 10:30	<i>Round Table (1)</i>
10:30 - 11:00	<i>Coffee Break</i>
11:00 - 12:30	<i>Education and Speech</i>
12:30 - 14:00	<i>Lunch Break</i>
14:00 - 14:30	<i>Legal & Agents</i>
14:30 - 15:30	<i>Arab Culture & Retrieval</i>
16:00 - 15:30	<i>Coffee Break</i>
16:00 - 17:00	<i>Discussion Roundtable (2)</i>

Athar at QIAS2025: LLM-based Question Answering Systems for Islamic Inheritance and Classical Islamic Knowledge

Yossra Nouredien¹ Hassan Suliman² Farah Attallah¹

Abdelrazig Mohamed¹ Sara Abdalla³

¹University of Khartoum ²African Institute for Mathematical Sciences

³International Islamic University Malaysia

{yossra.nouredien, farah.hassan, abdelrazig.mohamed}@uofk.edu

hassan.suliman017@gmail.com ssa_abdalla@iium.edu.my

Abstract

The intersection of Arabic linguistic complexity and specialized reasoning presents a key challenge for Islamic question-answering systems, particularly in the under-addressed area of inheritance law. This paper presents our methodology for the QIAS2025 shared task, assessing LLM capabilities in Islamic knowledge through two subtasks: Inheritance Reasoning (‘ilm al-mawārith) and General Islamic Assessment. A zero-shot, prompt-based approach with DeepSeek-R1 (deepseek-reasoner) addresses the former, while a three-stage RAG pipeline handles the latter. Our approaches achieved competitive results, with an accuracy of 0.704 for inheritance reasoning (10th place/15 teams) and 0.9272 for general Islamic assessment (2nd place/10 teams), demonstrating the efficacy of tailored model strategies for religious QA. These insights pave the way for more culturally and linguistically adaptive AI systems in Islamic scholarly applications.

1 Introduction

Arabic Islamic question-answering (QA) systems face dual challenges of linguistic complexity and specialized domain knowledge requirements. Inheritance law (‘ilm al-mawārith) and classical Islamic scholarship remain computationally under-explored, despite growing demand for accessible religious knowledge through digital platforms.

Historically, Islamic QA relied on symbolic systems such as rule-based expert systems and ontology-driven frameworks (Alshahad and Abutiheen, 2015; Zouaoui and Rezeg, 2021) or traditional information retrieval, effective in structured domains like inheritance law, but limited in handling linguistic variation and complex reasoning.

Modern large language models (LLMs) (e.g., GPT series (Radford et al., 2018)) and Arabic-centric models (e.g., ALLaM (Bari et al., 2024)) offer greater flexibility and cultural alignment, yet

their evaluation in specialized domains like inheritance and multi-disciplinary Islamic studies remains scarce, motivating the need for dedicated benchmarks.

The QIAS2025 shared task (Boucekif et al., 2025a) establishes a benchmark for evaluating LLMs across two domains: SubTask 1, *Islamic Inheritance Reasoning*, uses multiple-choice questions (MCQs) to test rule application, proportional reduction (‘awl), exclusion (hajb), and precise share allocation; and SubTask 2, *Islamic Studies Assessment*, comprises MCQs derived from 23 classical Islamic texts spanning Qur’anic studies, ḥadīth, fiqh, uṣūl al-fiqh, and sīrah

This paper presents our approach to the QIAS2025 shared task, addressing the two subtasks:

- SubTask 1: Zero-shot DeepSeek-R1 pipeline for inheritance reasoning, with output-constrained prompting and regex-based label extraction.
- SubTask 2: Three-stage hybrid RAG pipeline for general Islamic assessment, combining dense and BM25 retrieval with LLM reranking.
- Results: Competitive leaderboard rankings, 10th/15 for inheritance reasoning and 2nd/10 for general Islamic assessment.

2 Related Work

Recent advancements in transformer-based architectures and fine-tuning methodologies have significantly shaped Arabic Islamic question-answering systems. The field has seen significant progress through shared tasks such as Qur’an QA 2022 (Malhas et al., 2022), with notable contributions including Basem et al. (2025) expanding the Qur’an QA dataset to 1,895 question-answer pairs, achieving MAP@10 of 0.36 and 75% success in zero-

answer detection, and Abdallah et al. (2024) introducing ArabicaQA with over 89,000 questions for comprehensive Arabic QA benchmarking.

Domain-specific approaches have emerged for Islamic knowledge processing. For instance, Adel et al. (2023) developed AraQA for authentic religious texts with careful dataset curation to reduce misleading answers, while Alan et al. (2025) proposed MufassirQAS, a RAG-based system outperforming ChatGPT through vector databases and fact-checking mechanisms. Additionally, Qamar et al. (2024) developed a large-scale dataset with 73,000+ QA pairs for Tafsir and Ahadith, revealing limitations in automatic evaluation metrics and emphasizing the need for human expert assessment in religious QA contexts. Sibae et al. (2025) have also addressed Arabic language model assessment challenges, with comprehensive studies revealing significant performance variations across cultural and specialized domains.

Despite these advances, significant gaps remain particularly in computational approaches to Islamic inheritance law (ʿilm al-mawārith), which requires precise numerical calculations. While Alshammary et al. (2024) demonstrated promising results with their RFPG RAG model, most prior work focuses on extractive QA or general Islamic content. Most recently, Bouchekif et al. (2025b) conducted a large-scale evaluation of seven LLMs on Islamic inheritance, finding strong results for reasoning-oriented models but major errors in open-source Arabic ones.

Our participation in QIAS2025 explores both specialized inheritance reasoning and broader Islamic knowledge assessment through domain-specific MCQs. By tackling these distinct challenges, our work contributes novel empirical insights into the capabilities and limitations of LLMs in religious question answering.

3 Data

The QIAS2025 shared task provided two datasets from distinct domains, summarized in Table 1.

For SubTask 1, the dataset comprises Arabic multiple-choice questions in Islamic inheritance (ʿIlm al-Mawārith), each with six options (A–F). It includes 20,000 training, 1,000 development, and 1,000 test examples, plus 3,165 IslamWeb fatwas as extra data. For SubTask 2, the dataset consists of multiple-choice questions with four options (A–D) drawn from classical Islamic texts spanning fiqh,

ḥadīth, tafsīr, and other disciplines. The development set has 700 questions from 21 books, and the test set has 1,000 questions from 23 books (including two unseen in the development set).

Task	Train	Dev	Test	Extra Data
SubTask 1	20,000	1,000	1,000	3,165 fatwas
SubTask 2	–	700	1,000	23 classical texts

Table 1: Dataset statistics and additional resources for the QIAS2025 subtasks.

4 System Overview

Our proposed solution addressed the QIAS2025 shared task through two distinct pipelines, each tailored to the requirements of its respective subtask.

For Subtask 1, we tested several reasoning-capable models via in-context prompting and selected DeepSeek-R1 for its strong Arabic reasoning and cost-effective API. For Subtask 2, we normalized the provided corpus of 23 classical books spanning HTML and DOCX formats, enabling the construction of a unified hybrid index. We experimented with several retrieval strategies, including retrieving surrounding passages and hybrid fusion, and found that applying a LLM reranker yielded the best approach. Across both subtasks, the design emphasizes robustness to varied encodings, domain specificity, and consistent answer formatting.

4.1 SubTask 1: Islamic Inheritance Reasoning

To handle the complex reasoning required in Islamic inheritance (ʿIlm al-Mawārith), we employed a zero-shot, prompt-based approach with the `deepseek-reasoner` model (DeepSeek-R1-0528) via API (DeepSeek-AI et al., 2025). No fine-tuning was performed; instead, the model was directly evaluated in zero-shot mode, leveraging its Arabic reasoning capability. The domain-specific Arabic prompt is illustrated in Figure 1, and the English translation is provided in Appendix A.

Prompt Design. The prompt included the question, six answer options, and a strict instruction to output only the correct choice in the format `<answer> X </answer>`, producing deterministic, machine-readable results. This format eliminated ambiguity and avoided the mixing of Arabic text with answer labels. We did not experiment with alternative prompt formats, as our primary ob-

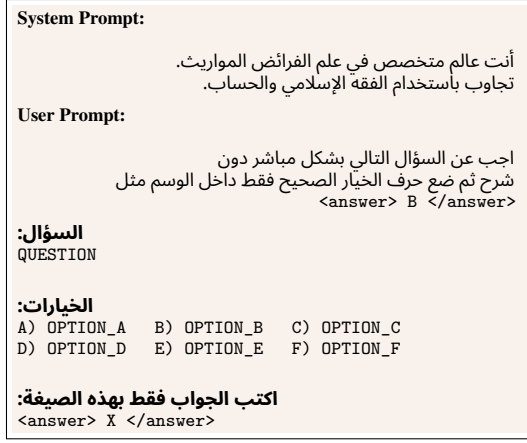


Figure 1: Zero-shot prompt used in SubTask 1

jective was to suppress free-form “thinking” outputs and enforce consistent, extractable answers.

Pipeline Execution. Following prompt construction, each instance was submitted to the DeepSeek API using fixed decoding parameters. Model responses were then parsed using a regular expression to extract the predicted label. All model outputs and extracted answers were logged per instance to ensure reproducibility and support error analysis. The pipeline operated in a CPU-only environment via the paid API tier, ensuring stable latency and no token constraints.

4.2 Subtask 2: Islamic Studies Assessment

For this task, a RAG pipeline was adopted to manage the semantic diversity of questions, heterogeneous text formats, and the need for source-grounded reasoning. The pipeline was inspired by methodologies from the RAG-Challenge-2 repository¹, and the overall workflow is shown in Figure 2. Translation of Arabic text is available in Appendix A.

Corpus Ingestion and Indexing. After normalizing the corpus into plain text for consistency across formats, each book was segmented into semantically coherent passages using LangChain’s RecursiveCharacterTextSplitter, configured with a chunk size of 500 characters and a 50 character overlap to preserve contextual continuity. This overlap mitigates semantic fragmentation across chunk boundaries, a technique commonly used in multilingual and Arabic NLP. Each chunk was embedded using OpenAI’s

¹<https://github.com/Ilyarice/RAG-Challenge-2>

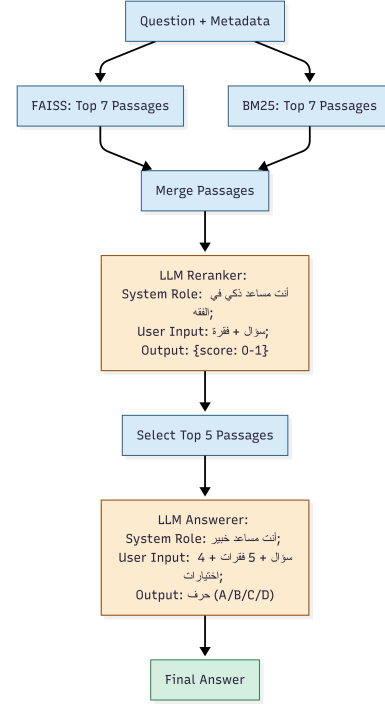


Figure 2: Pipeline used in SubTask 2

text-embedding-3-large model, producing dense semantic vectors. These were indexed using FAISS’s IndexFlatIP for dense similarity search (Johnson et al., 2021). In parallel, sparse representations were computed using the Okapi BM25 algorithm (Robertson and Zaragoza, 2009) to support lexical-level retrieval. Each chunk was also stored with metadata such as book title to support traceability and analysis.

Hybrid Retrieval and Reranking. To leverage both semantic and lexical retrieval, we adopted a hybrid strategy without score fusion. Instead of α -weighted interpolation, we performed parallel top- k retrieval: the top 7 passages were retrieved independently from FAISS and BM25, producing a 14-passages candidate set that maintained both semantic relevance and lexical precision. Our methodology prioritized demonstrating the hybrid approach’s optimal performance for Islamic inheritance QA, with individual retrieval component analysis considered beyond the current work’s scope and suitable for future comparative studies. A lightweight reranking stage using GPT-4o-mini was then applied to semantically compare the question with each of the 14 retrieved passages and select the 5 most relevant ones for the final answer generation stage.

Answer Generation. In the final stage, the top 5 passages, selected by the reranker, were used as contextual input for answer generation with GPT-4o. These passages were injected into a constrained multiple-choice question prompt, which explicitly instructed the model to return only a single answer choice, formatted within an <answer> tag. This strict output format minimized generation variability. Importantly, no model fine-tuning was performed at any stage. Both the reranking and answer generation components operated in zero-shot inference mode, relying solely on carefully crafted prompts and high-quality context to guide the model’s reasoning.

5 Results

5.1 Evaluation and Performance

Accuracy was used as the primary evaluation metric across both subtasks, defined as:

$$Accuracy = \frac{Correct_predictions}{Total_samples}$$

This metric directly reflects the proportion of correctly answered questions, making it appropriate for multiple-choice QA tasks. Our evaluation was carried out on both the development and test sets, with results summarized in Table 2.

Task	System	Devset	Testset
Subtask 1	DeepSeek API with Direct Prompting	0.885	0.704
Subtask 2	RAG with Hybrid Retrieval and LLM Reranker	0.914	0.927

Table 2: Accuracy of Development and Test Sets

Leaderboard rankings were determined using the test set. In **Subtask 1**, our system ranked 10th out of 15 teams, with the top score reaching 0.972. In **Subtask 2**, we placed 2nd out of 10 teams, with the best score recorded at 0.937.

Beyond leaderboard ranks, we provide statistical analysis using Wilson confidence intervals, which offer superior boundary handling for proportion estimates. In Subtask 1 (6 options; chance = 16.7%), our system achieved 70.4% accuracy on N=1000 with CI [67.5%, 73.2%]. In Subtask 2 (4 options; chance = 25%), we applied majority-rule deduplication yielding N=729 samples, with accuracy of 92.32% and CI [90.16%, 94.04%]. Both confidence intervals demonstrate substantial separation from chance levels, indicating robust performance

well beyond random selection. The Wilson interval methodology ensures reliable statistical inference even near boundary conditions, while the substantial sample sizes support the stability of our accuracy estimates, though formal hypothesis testing could further strengthen these findings.

5.2 Results Analysis

For **Subtask 1**, our evaluation highlights three main trends:

- **General Performance Gap:** Accuracy dropped from 88.5% on the development set to 70.4% on the test set (−18.1%). Test questions were longer (140 vs. 97 characters, +43.8%) and answer options were much longer (653 vs. 117, 5.57×), as shown in Figure 3, suggesting a possible domain shift with added lexical variety and detail that increased reasoning difficulty.

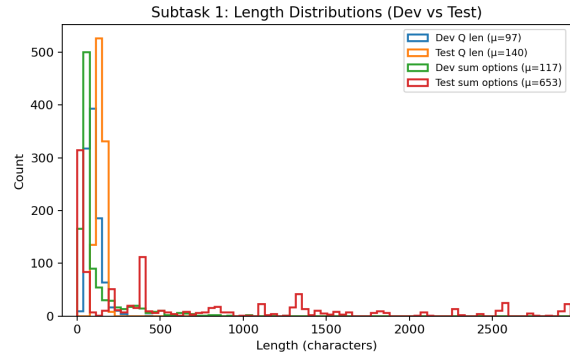


Figure 3: Subtask 1: Dev/Test sets length distributions (questions & options). Test is longer on average.

- **Level Sensitivity:** Questions were labeled Beginner or Advanced. On the development set, accuracy was 90.0% for Advanced and 87.0% for Beginner. On the test set, it dropped to 70.6% and 70.2% respectively. The similar decline across both levels indicates the drop was driven by overall complexity rather than by a specific difficulty category.
- **Error Patterns:** Accuracy varied by heir category. For example, questions mentioning أخت شقيقة/لأم/لأب (sisters) had an accuracy of 0.644, while those mentioning زوجة/زوج (spouse) reached 0.794. These results are based on *inclusive* category, meaning each question is counted under every heir it in-

volves. Appendix B contains a qualitative error example.

For **Subtask 2**, performance was consistent across development and test sets. Analysis focuses on three aspects:

- **Error Causes:** Two main issues contributed to mistakes: (i) relevant passages were not retrieved; and (ii) even when correct passages were retrieved, the LLM sometimes failed to select the right option. Examples of errors are available in Appendix B.
- **Performance by level:** Questions were labeled Beginner, Intermediate or Advanced. Accuracy was 96.7% on Beginner questions, 95.3% on Intermediate and 78.7% on Advanced. This shows that the system handled easier questions well but dropped on more challenging ones.
- **Performance by Source:** Results varied by book. Accuracy was lowest on فتح المغيث (63.6%) but reached (100%) on sources such as الرحيق المختوم and الفقه المنهجي. This indicates that differences in style, terminology, and content across sources significantly affected retrieval and answer selection. Figure 4 illustrates the top 5 lowest and highest sources by accuracy.

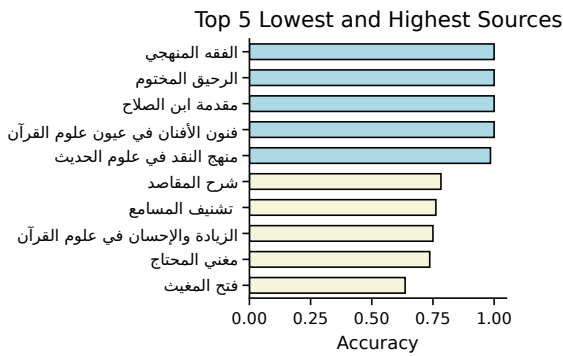


Figure 4: Top 5 lowest and highest sources by accuracy.

The findings highlights that Subtask 1 requires stronger complex reasoning, whereas Subtask 2 would benefit from enhanced retrieval and LLM comprehension to achieve more reliable answer selection.

6 Conclusion

This paper presented our systems for the QIAS2025 shared task on Islamic Inheritance

Reasoning and Classical Islamic Knowledge. We implemented a direct prompting approach for Subtask 1, achieving 70.4% accuracy, and a hybrid RAG pipeline combining FAISS and BM25 retrieval with GPT-4o-mini reranking for Subtask 2, achieving 92.72%. The analysis revealed that inheritance reasoning demands careful handling of longer and more complex scenarios, while Subtask 2 highlighted retrieval performance variation across diverse classical sources. While our systems demonstrated excellent performance, broader deployment requires addressing critical ethical and performance challenges.

Ethical Considerations

While the QA systems presented in this paper demonstrates excellent accuracy performance, the broader deployment of such systems requires careful attention to inherent challenges. These challenges include hallucination and misinformation risks (Khalila et al., 2025), algorithmic bias affecting diverse religious communities (Gupta and Giannoccaro, 2024), privacy concerns with sensitive spiritual data (Liu et al., 2025), and questions about authenticity in AI-mediated spiritual experiences (Alkhouri, 2024). Successful implementation requires inclusive algorithm design (Habib, 2025), transparent accountability measures (Sarker, 2024), and human oversight to ensure responsible and effective deployment in religious contexts. Such measures are essential for ensuring responsible AI deployment that respects the diversity and sensitivity inherent in religious contexts.

Acknowledgments

We thank the anonymous reviewers for their helpful comments.

References

- Abdelrahman Abdallah, Mahmoud Kasem, Mahmoud Abdalla, Mohamed Mahmoud, Mohamed Elkasaby, Yasser Elbendary, and Adam Jatowt. 2024. *Arabi-caQA: A Comprehensive Dataset for Arabic Question Answering*. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2049–2059, Washington DC USA. ACM.
- Yousef Adel, Mostafa Dorrah, Ahmed Ashraf, Abdallah ElSaadany, Mahmoud Mohamed, Mariam Wael, and

- Ghada Khoriba. 2023. [AraQA: An Arabic Generative Question-Answering Model for Authentic Religious Text](#). In *2023 11th International Japan-Africa Conference on Electronics, Communications, and Computations (JAC-ECC)*, pages 235–239, Alexandria, Egypt. IEEE.
- Ahmet Yusuf Alan, Enis Karaarslan, and Ömer Aydin. 2025. [A RAG-based Question Answering System Proposal for Understanding Islam: MufasssirQAS LLM](#). *arXiv preprint*. ArXiv:2401.15378 [cs].
- Khader I. Alkhouri. 2024. [The Role of Artificial Intelligence in the Study of the Psychology of Religion](#). *Religions*, 15(3):290.
- H. F. Alshahad and Z. A. Abutiheen. 2015. Computation of inheritance share in islamic law by an expert system using decision tables. *Quarterly Adjudicated Journal for Natural and Engineering Research and Studies*, 1:105–114.
- Mitha Alshammmary, Md Nahiyen Uddin, and Latifur Khan. 2024. [RFGP: Question-Answering from Low-Resource Language \(Arabic\) Texts using Factually Aware RAG](#). In *2024 IEEE 10th International Conference on Collaboration and Internet Computing (CIC)*, pages 107–116, Washington, DC, USA. IEEE.
- M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham A. Alyahya, Sultan Al-Rashed, Faisal A. Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhatran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Majed Alrubaian, Ali Alammari, Zaki Alawami, Abdulmohsen Al-Thubaity, and 6 others. 2024. [Al-lam: Large language models for arabic and english](#). *Preprint*, arXiv:2407.15390.
- Mohamed Basem, Islam Oshallah, Baraa Hikal, Ali Hamdi, and Ammar Mohamed. 2025. [Optimized Quran Passage Retrieval Using an Expanded QA Dataset and Fine-Tuned Language Models](#). In Faisal Saeed, Fathey Mohammed, Errais Mohammed, Shadi Basurra, and Mohammed Al-Sarem, editors, *Advances on Intelligent Computing and Data Science II*, volume 255, pages 244–254. Springer Nature Switzerland, Cham. Series Title: Lecture Notes on Data Engineering and Communications Technologies.
- Abdessalam Boucekif, Samer Rashwani, Mohammed Ghaly, Mutaz Al-Khatib, Emad Mohamed, Wajdi Zaghoulani, Heba Sbahi, Shahd Gaben, and Aiman Erbad. 2025a. Qias 2025: Overview of the shared task on islamic inheritance reasoning and knowledge assessment. In *Proceedings of The Third Arabic Natural Language Processing Conference, ArabicNLP 2025, Suzhou, China, November 5–9*. Association for Computational Linguistics.
- Abdessalam Boucekif, Samer Rashwani, Heba Sbahi, Shahd Gaben, Mutaz Al-Khatib, and Mohammed Ghaly. 2025b. Assessing large language models on islamic legal reasoning: Evidence from inheritance law evaluation. In *Proceedings of The Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *Preprint*, arXiv:2501.12948.
- Brij Gupta and Ivan Giannoccaro, editors. 2024. *Challenges in Large Language Model Development and AI Ethics*. Advances in Computational Intelligence and Robotics. IGI Global.
- Zainal Habib. 2025. [Ethics of Artificial Intelligence in Maqāsid Al-Sharīa’s Perspective](#). *KARSA Journal of Social and Islamic Culture*, 33(1):105–134.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. [Billion-scale similarity search with gpus](#). *IEEE Transactions on Big Data*, 7(3):535–547.
- Zahra Khalila, Arbi Haza Nasution, Winda Monika, Aytug Onan, Yohei Murakami, Yasir Bin Ismail Radi, and Noor Mohammad Osmani. 2025. [Investigating Retrieval-Augmented Generation in Quranic Studies: A Study of 13 Open-Source Large Language Models](#). *International Journal of Advanced Computer Science and Applications*, 16(2).
- Feng Liu, Jiaqi Jiang, Yating Lu, Zhanyi Huang, and Jiuming Jiang. 2025. [The ethical security of large language models: A systematic review](#). *Frontiers of Engineering Management*, 12(1):128–140.
- Rana Malhas, Watheq Mansour, and Tamer Elsayed. 2022. [Qur’an QA 2022: Overview of The First Shared Task on Question Answering over the Holy Qur’an](#). In *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur’an QA and Fine-Grained Hate Speech Detection*, pages 79–87, Marseille, France. European Language Resources Association.
- Faiza Qamar, Seemab Latif, and Asad Shah. 2024. [Techniques, datasets, evaluation metrics and future directions of a question answering system](#). *Knowledge and Information Systems*, 66(4):2235–2268.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *Preprint*.
- Stephen E. Robertson and Hugo Zaragoza. 2009. [The probabilistic relevance framework: BM25 and beyond](#). *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Iqbal H. Sarker. 2024. [LLM potentiality and awareness: a position paper from the perspective of trustworthy and responsible AI modeling](#). *Discover Artificial Intelligence*, 4(1):40.

Serry Sibae, Omer Nacar, Adel Ammar, Yasser Al-Habashi, Abdulrahman Al-Batati, and Wadii Boulila. 2025. [From Guidelines to Practice: A New Paradigm for Arabic Language Model Evaluation](#). *arXiv preprint*.

Samia Zouaoui and Khaled Rezeg. 2021. [Islamic inheritance calculation system based on arabic ontology \(arafamonto\)](#). *Journal of King Saud University - Computer and Information Sciences*, 33(1):68–76.

Appendices

A Translations of Figures' Arabic Text

Figure 1: Zero-shot prompt used in SubTask 1

- **System Prompt:** You are a scholar specialized in inheritance law (ʿilm al-farāʿid), answer using Islamic jurisprudence and arithmetic.
- **User Prompt:** Answer the following question directly without explanation, then place only the correct option letter inside the tag e.g. `<answer> B </answer>`.
- **Label:** Write the answer only in this format: `<answer> X </answer>`

Figure 2: Pipeline used in Subtask 2

- **LLM Reranker**
System Role: You are an intelligent assistant in Islamic jurisprudence.
User Input: Question + Passage.
Output: score: 0-1
- **LLM Answerer**
System Role: You are an expert assistant.
User Input: Question + 5 Passages + 4 Options.
Output: (A/B/C/D)

B Qualitative Errors Examples

Subtask 1 (Inheritance Reasoning):

- Multiple-choice inheritance question in which **Gold = D**, **Pred = A** (Figure B1).

Subtask 2 (Islamic Studies Assessment):

- Answer-absent: the correct answer is missing from hybrid retrieval and thus absent from the reranked context; the model guessed incorrectly. **Gold = B**, **Pred = A** (Figure B2).
- Evidence-present: The correct option is supported in the reranked context, but the model chose a different option. **Gold = C**, **Pred = A** (Figure B3).

Question ID: 2173_nl7p9x5o_16

أجب عن السؤال التالي بشكل مباشر
دون شرح، ثم ضع حرف الخيار الصحيح
فقط داخل الوسم مثل

`<answer> B </answer>`

السؤال:

مات وترك: أخ لأم (2) و عم الأب لأب (3)
و ابن أخ شقيق (2) و أب أب الأب و أم
الأب و ابن ابن و أب، كم عدد الأسهم
بعد التصحيح التي يحصل عليها ابن ابن؟

الخيارات:

A) 0 سهم
B) 4 سهم
C) 6 سهم
D) 5 سهم
E) 3 سهم
F) 7 سهم

اكتب الجواب فقط بهذه الصيغة:

`<answer> X </answer>`

Model Response:

`<answer> A </answer>`

Extracted: A

Figure B1: Example Error from Subtask 1

Question: بعد رد أبي بكر الصديق جوار ابن الدعة، تعرض لابنتاه من أحد سفهاء قريش، وفي هذا الموقف الصعب، كرر قولاً يعكس صبره العظيم وإيمانه العميق. ما هو هذا القول الذي كرره ثلاث مرات؟

Options:

- A) "إحسبي الله ونعم الوكيل"
- B) "يا رب، ما أحلمك"
- C) "اللهم اجربي في مصيبتني"
- D) "الحمد لله على كل حال"

Gold: B Pred: A

Reranked top-5 passages:

- P1 (book=sira, score=0.9):
رددت إليك جوارك فقال له يا ابن أخي لعله آذاك أحد من قومي قال لا ولكني أرضى بجوار الله ولا أريد أن أستجير بغيره قال فاطلق إلى المسجد فأردد علي جوار عاتية كما أجزتك عاتية قال فاطلقا فخرجا حتى أتيا المسجد فقال الوليد هذا عثمان قد جاء برد علي جوار علي قال صدق قد وجنته وفيها كريم الجوار ولكني قد أحبيت أن لا أستجير بغير الله فقد رددت عليه جواره ثم انصرف عثمان وليد ابن ربيعة بن مالك بن جعفر بن كلاب في مجلس من قريش يشهدهم فجلس معهم عثمان فقال ليبيد ** ألا كل شيء ما خلا الله باطل **
- P2 (book=sira, score=0.9):
عَلَيْهِ وَسَلَّمَ: "تَعَمَّ، أَنَا الَّذِي أَقُولُ ذَلِكَ". قَالَ: فَلَقْتُ رَأَيْتُ رَجُلًا مِثْلَهُمْ أَخَذَ بِمَجْمَعِ رِدَائِهِ. قَالَ: فَذَامَ أَبُو بَكْرٍ رَجُلِي اللَّهُ عَنْهُ ثَوْنَةً، وَهُوَ يَبْكِي وَيَقُولُ: أَتَقْتُلُونَ رَجُلًا أَنْ يَقُولَ رَبِّيَ اللَّهُ؟ ثُمَّ انصَرَفُوا عَنْهُ، فَإِنَّ ذَلِكَ لَأَكْبَدُ مَا رَأَيْتُ قُرَيْشًا نَالُوا مِنْهُ فَط
- P3 (book=sira, score=0.9):
أَهْلَ النَّبَرَاءِ، وَأَصْحَابِي وَأَهْلَ بَيْتِي يَلْتَوُونَ مِنَ الْبَاءَةِ وَالْأَذَى فِي اللَّهِ مَا لَا يُصِيبُنِي، لِنُصْرَةِ كَبِيرٍ فِي نَفْسِي. فَخَشَى إِلَى الْوَلِيدِ بْنِ الْمُجِيرِ، فَقَالَ لَهُ: يَا أَبَا عَدْرِ شَيْئٍ، وَقَدْ نَمُتَكَ، فَذَرْنْتُ إِلَيْكَ جَوَارَكَ، فَقَالَ لَهُ: (لَمْ) [1] يَا بَنَ أَخِي؟ لَعَلَّهُ أَذَاكَ أَحَدٌ مِنْ قَوْمِي، قَالَ: لَا، وَلَكِنِّي أَرْضَى بِجَوَارِ اللَّهِ، وَلَا أُرِيدُ أَنْ أَسْتَجِيرَ بغيره؟ قَالَ: فَاطْلُقْ إِلَى الْمَسْجِدِ، فَارْجُ
- P4 (book=rwd, score=0.3):
قَالَ أَبُو خَلِيفَةَ لَأَبِي بَكْرٍ يَا بَنِي إِيَّيْكَ تَتَّقِي رَفِئًا مِثْلَهُ، فَلَوْ أَنَّكَ إِذْ فَعَلْتَ مَا فَعَلْتَ أَغْنَيْتَ رَجُلًا جَلَدًا يُفْتَحُونَكَ، وَيَقُولُونَ ثَوْنَةً؟ قَالَ فَقَالَ أَبُو بَكْرٍ رَجُلِي اللَّهُ عَنْهُ يَا أَبَتُ إِيَّيْكَ إِنَّمَا أُرِيدُ مَا أُرِيدُ إِلَهُ عَزَّ وَجَلَّ، قَالَ فَيَتَحَدَّثُ أَنَّهُ مَا نَزَلَ هَؤُلَاءِ الْأَيَّامُ إِلَّا فِيهِ وَفِيمَا قَالَ لَهُ أَبُوهُ (فَلَمَّا مِنْ أَعْطَى وَتَقَى وَصَنَّقَ بِالْحُسْنَى) [الليل 5، 6]. إِلَى قَوْلِهِ تَعَالَى: (وَمَا
- P5 (book=sira, score=0.3):
قال ابن اسحاق وحدثني هشام بن عروة عن أبيه قال كان ورقة بن نوفل يمر به وهو يحب بذلك وهو يقول أحد أحد فيقول أحد أحد والله يا باطل ثم يقبل على أمية بن خلف ومن يصنع ذلك به من بني جمح فيقول أحلف بالله لئن قتلتموه على هذا لأخذنه حدانا حتى مر به أبو بكر الصديق ابن أبي قحافة رضي الله عنه يوما وهم يصنعون ذلك به وكانت دار أبي بكر في بني جمح فقال لأمية بن خلف ألا تنقي الله في هذا المسكين حتى متى قال أنت الذي أفسدته فألفده مما ترى فقال أبو بكر أقبل عدي عادم أسود أجلد منه وأقرى على دينك أصطيك به قال قد

Figure B2: Example Error from Subtask 2 (Answer-absent)

