Halluscoring 2026: The First Shared Task on LLM Hallucination Detection and Answer Verification

Aisha Alansari[1], Abdessalam Bouchekif[2], Ahmed Hasanaath[1], Salah Eddine Bekhouche[3], Malak Alkhorasani[4], Mohammed-En-Nadhir Zighem[2], Saad Ezzini[1], Hichem Telli[5], Hend Al-Khalifa[8], Muhammad Abdul-Mageed[6], Hadid Abdenour[7], Hamzah Luqman[1]

[1] King Fahd University of Petroleum and Minerals, Saudi Arabia
[2] Hamad Bin Khalifa University, Qatar
[3] University of the Basque Country, Spain
[4] Imam Abdulrahman bin Faisal University, Saudi Arabia
[5] University of Biskra, Algeria
[6] University of British Columbia, Canada
[7] Universiti Malaysia Kelantan, Malaysia
[8] King Saud University, Saudi Arabia

{aisha.ansari, g202302610, hluqman}@kfupm.edu.sa

Figure 1: Overview of the *HalluScoring 2026* shared task. Task 1 evaluates binary hallucination detection under two generalization settings: unseen questions (Subtask 1.1) and unseen LLMs (Subtask 1.2). Task 2 extends the evaluation to two stages, first detecting if a generated response is hallucinated and then selecting the correct factual answer from six candidate options, covering Islamic knowledge (Subtask 2.1) and general knowledge (Subtask 2.2).

Abstract

We present *HalluScoring 2026*, a shared task for evaluating hallucination detection and factual verification in Arabic question answering under challenging generalization settings. Its four subtasks are organized into two tasks. Task 1 evaluates binary hallucination detection, considering generalization to unseen questions (Subtask 1.1) and responses generated by unseen LLMs (Subtask 1.2). Task 2 extends the evaluation beyond detection by requiring the systems to additionally identify the correct factual answer from six related candidates, covering Islamic knowledge (Subtask 2.1) and general knowledge (Subtask 2.2). The shared task is based on two Arabic datasets: HalluScore and HalluTruthQA. A total of 13 teams participated in the shared task, ten of which submitted system description papers. The results of Task 1 demonstrate that hallucination detection remains challenging. On the Task 1 test sets, the top-ranked systems achieved AUC-ROC scores of 0.7717 for Subtask 1.1 (REGLAT) and 0.7670 for Subtask 1.2 (NAMAA). Under assisted evaluation, the highest combined detection and answer-selection scores for Subtasks 2.1 and 2.2 were 0.8824 and 0.8565, respectively.

1 Introduction

Large language models (LLMs) are increasingly being used to generate and process Arabic content in a wide range of applications, including question answering (QA), information retrieval, and summarization (Hasanaath et al., 2025; Alzubaidi et al., 2025; Mashabi et al., 2026). Recent Arabic-specific and multilingual LLMs can produce fluent and contextually appropriate Arabic text, supporting their use in both general-purpose and domain-specific applications (Alam et al., 2024; Abdaljalil et al., 2025; Bouchekif et al., 2025). However, fluent responses are not necessarily factually reliable. LLMs may present fabricated, unsupported, or factually incorrect claims as true; we refer to such claims as hallucinations (Maynez et al., 2020; Ji et al., 2023). Such hallucinations may involve entities, dates, numerical values, quotations, source attributions, or supporting explanations (Li et al., 2023; Alansari and Luqman, 2025; Bouchekif et al., 2026).

Proceedings of The Fourth Arabic Natural Language Processing Conference (ArabicNLP 2026) - Volume 2 Shared Tasks, pages 1106–1119
October 28-29, 2026 ©2026 Association for Computational Linguistics

Hallucinations may arise from incomplete or noisy training data, limited coverage of domain-specific knowledge, ambiguous prompts, or insufficient contextual grounding. They may also be influenced by LLMs' next-token prediction objective, which optimizes for likely continuations without explicitly verifying the factual correctness of each generated claim (Huang et al., 2025; Alansari and Luqman, 2026b). This challenge is particularly relevant to Arabic, where high-quality, diverse, and domain-specific evaluation resources remain comparatively limited.

The evaluation of hallucinations in Arabic remains underexplored, particularly with respect to generalization beyond the conditions observed during training. A detector that performs well on questions similar to those observed during training may not generalize to new questions (Alansari et al., 2026). Likewise, a system trained on responses generated by a fixed set of LLMs may capture generator-specific characteristics rather than general hallucination signals, potentially limiting its ability to generalize to responses produced by unseen models (Wu et al., 2025). Moreover, correctly detecting a hallucination does not necessarily imply that a system can identify the correct factual information (Bouchekif et al., 2026). These challenges motivate the evaluation of both the robustness of hallucination detection and the ability to identify the correct answer.

To address these challenges, we present *HalluScoring 2026*, a shared task on hallucination detection and factual verification in Arabic QA. As shown in Figure 1, the shared task aims to evaluate the robustness of hallucination detection under two complementary generalization settings: Task 1 evaluates binary hallucination detection on unseen questions (Subtask 1.1) and on responses from unseen LLMs to unseen questions (Subtask 1.2). Task 2 requires systems to detect hallucinations and select the correct answer from six candidates. Subtask 2.1 focuses on Islamic knowledge, while Subtask 2.2 covers broader domains, including history, science, and geography. The shared task is based on two carefully annotated Arabic resources, *HalluScore* (Alansari and Luqman, 2026a) and *HalluTruthQA* (Bouchekif et al., 2026; Bekhouche et al., 2026), enabling evaluation across multiple domains and model generators using expert-verified factual information. We also provide a comparative evaluation of participating systems, highlighting methodological approaches, performance trends, and remaining challenges across different evaluation settings.

2 Literature Review

Hallucination in LLMs has motivated several shared tasks aimed at developing standardized evaluation settings for hallucination detection. SemEval-2024 Task 6, *SHROOM* (Mickus et al., 2024), focused on detecting fluent but semantically incorrect outputs in machine translation, paraphrase generation, and definition modeling, considering both model-aware and model-agnostic settings. Its multilingual successor, SemEval-2025 Task 3, *Mu-SHROOM* (Vázquez et al., 2025), extended this work to 14 languages, including Arabic, and evaluated the localization of hallucinated spans in LLM outputs. These shared tasks established standardized settings for evaluating hallucination detection in language and generation tasks.

Arabic research on hallucination has also progressed through both shared tasks and dedicated datasets. The *Halwasa* shared task (Mubarak et al., 2024) addressed hallucination detection in Arabic LLM-generated content, building on a dataset of 10,000 Arabic sentences annotated for factual and linguistic correctness. *IslamicEval 2025* (Mubarak et al., 2025) subsequently targeted hallucinations within the specialized domain of Islamic content, including the detection and correction of hallucinated Qur'anic verses and Hadiths grounded in authoritative sources. Beyond shared tasks, *AraHalluEval* (Alansari and Luqman, 2025) introduced fine-grained hallucination annotations for Arabic generative QA and summarization, distinguishing different factuality and faithfulness errors.

Positioning of HalluScoring 2026. *HalluScoring 2026* extends Arabic hallucination shared tasks by evaluating both the robustness of hallucination detection and the ability to recover correct factual information. Building on *HalluScore* (Alansari and Luqman, 2026a) and *HalluTruthQA* (Bouchekif et al., 2026; Bekhouche et al., 2026), the shared task evaluates generalization to unseen questions and LLMs, as well as joint hallucination detection and correct-answer selection. Moreover, the evaluation covers both Islamic and general-knowledge domains, with the latter spanning areas such as history, science, and geography. Reliable hallucination detection is particularly important in these settings because factually incorrect responses can misrepresent religious teachings, historical events,

scientific facts, or geographical information, potentially misleading users who rely on LLMs as sources of information.

3 HalluScoring 2026

3.1 Task 1: Hallucination Detection

The goal of Task 1 is to detect hallucinations in Arabic QA in diverse domains. Given a question and one or more model-generated responses, the objective is to automatically determine whether each response contains hallucinated content. The questions span multiple domains and topics, exposing models to different types of knowledge and varying levels of factual specificity, which can make generated responses particularly prone to hallucination. To encourage the development of robust detection methods that generalize across different questions and LLMs, the shared task is organized into two complementary subtasks.

Subtask 1.1: Generalize Across Questions In this subtask, participants are required to develop hallucination detection systems that can generalize across different questions. The training and test splits are constructed such that all models appear in both sets, but the questions are disjoint. Each instance consists of a question and a generated response, in which the system must predict whether the response is hallucinated or non-hallucinated.

Subtask 1.2: Generalize Across Models This subtask evaluates the ability of the proposed hallucination detection techniques to generalize across unseen models. The dataset is split such that the training set contains responses generated by one subset of LLMs, while the test set contains responses from different LLMs that are entirely unseen during training. The questions are also partitioned between the training and test sets to evaluate generalization across both models and questions. The goal remains to classify hallucinated versus non-hallucinated outputs. This setting is particularly challenging, as it requires learning model-invariant hallucination signals that transfer across different architectures.

3.2 Task 2: From Hallucination Detection to Truth

The goal of Task 2 is to evaluate whether a system can both detect hallucinations and identify the correct information. Given an Arabic question and a generated answer, the system must first determine whether the answer contains a hallucination. Then, the correct answer must be selected from six candidate options, including a correct answer and five plausible distractors. The candidate options are designed to be close to one another, making it more challenging to distinguish correct from incorrect answers. Task 2 consists of two subtasks.

Subtask 2.1: Islamic Knowledge. This subtask focuses on Islamic knowledge. The questions cover topics such as Prophetic biography, Islamic law, religious concepts, and source-sensitive information, including Qur'anic verses, Hadith, scholarly opinions, and source attributions. This domain is particularly challenging for hallucination detection because a generated answer may provide the correct main answer while still containing incorrect details, such as a verse number, a quotation, an attribution, or a supporting statement.

Subtask 2.2: General Culture. This subtask focuses on general knowledge across a wide range of topics, including history, geography, science, culture, and Islamic knowledge. It evaluates the system's ability to detect factual errors and identify the correct answer across diverse domains. The questions may contain plausible but incorrect information, including errors in dates, entities, locations, events, names, numerical facts, or other factual details, requiring the system to carefully distinguish between accurate and misleading information. Subtask 2.1 focuses on Islamic knowledge, while Subtask 2.2 extends the evaluation to three additional domains: history, science, and geography.

3.3 Task Setup

HalluScoring 2026 was organized by hosting four separate CodaBench competitions, one for each subtask. This separation allowed each task to have its own data format, submission pipeline, and evaluation procedure. The training and development sets for all tasks were released in 16th May 2026. For Task 1, the training and development data were provided in `Excel` format, whereas for Task 2, they were provided in `JSON` format. Appendix B illustrates some samples given to the participants.

Participants used these released splits to develop and validate their systems before submission. To preserve the integrity of the final evaluation, we did not publicly release the test sets to participants. Instead, we provided a Jupyter Notebook template that specified the required procedures for model loading, inference, and output. Participants were asked to implement their systems using this template and submit the completed notebook to the

organizers. The organizers executed the submitted notebooks on the hidden test sets.

3.4 Shared Task Restrictions

To ensure a fair and controlled evaluation setting, we provided clear restrictions on participating systems via the CodaBench platform. All submissions were required to be based on a fine-tuned generative language model whose base model is open-source or openly available for research use. We restricted submitted models to a maximum of 13 billion parameters to maintain comparable computational settings between participants. During evaluation, participants were not allowed to use retrieval-augmented generation (RAG), external search tools, or other retrieval-based systems. This restriction was introduced to ensure that performance primarily reflects the capabilities of the submitted model rather than its access to external information during inference. However, participants were allowed to use external datasets for model fine-tuning and were requested to disclose any external datasets. These constraints were designed to provide a standardized evaluation setting while still allowing teams flexibility in model selection, training data, and fine-tuning strategies.

4 Datasets and Evaluation Metrics

4.1 HalluScore Dataset

The *HalluScore* (Alansari and Luqman, 2026a) dataset is proposed for the evaluation, detection, and mitigation of Arabic hallucination in LLMs. It is a structured Arabic QA benchmark comprising 827 carefully curated QA pairs across multiple domains, including health, science, finance, religion, and geography. The dataset is constructed through a rigorous pipeline involving quality assurance, filtering for clarity and factual validity, and model-driven selection to retain questions that consistently trigger hallucinations. For each question, responses were generated by 17 Arabic, multilingual, and reasoning LLMs and manually annotated as either hallucinated or non-hallucinated, resulting in a total of 14,059 annotated QA-response instances. For the shared task, only the information required for the corresponding prediction tasks was provided to participants, while the additional annotations and metadata available in the original *HalluScore* benchmark were not included as prediction targets. The test sets and their corresponding labels were kept hidden from participants and were used exclusively by the organizers for the final evaluation. The annotation guidelines are provided in Appendix A.1.

4.2 HalluTruthQA Dataset

HalluTruthQA (Bekhouche et al., 2026) contains 4,000 Arabic question–answering instances covering four domains: Islamic knowledge, history, science, and geography. The corpus is balanced, with 1,000 instances per domain. Each instance includes a question, a model-generated answer, an expert-verified reference answer, a binary hallucination label, and six candidate answers consisting of one correct answer and five plausible distractors. The candidate answers were manually constructed to be semantically related and stylistically similar, reducing the possibility of identifying the correct answer through superficial cues. The model-generated answers were produced using `QCRI/Fanar-1-9B-Instruct` (Team et al., 2025) with a maximum of 1,024 new tokens, a temperature of 0.0, top-p set to 1.0, and a batch size of 4, using `bfloat16` precision.[1]

Beyond the HalluScoring task's objectives of binary hallucination detection and correct-option selection, the original HalluTruthQA corpus provides richer fine-grained annotations. In particular, hallucinated responses include character-level erroneous spans and human-written explanations describing why each span is incorrect. This supports error localization, explanation, and fine-grained analysis beyond the official shared-task objectives. The annotation guidelines are provided in Appendix A.2.

4.3 Data Splits

The numbers of LLMs, questions, and instances in the train, development, and test sets are reported in Table 1. Subtask 1.1 and 1.2 combine *HalluScore* and *HalluTruthQA* to increase the number and diversity of instances available for training and evaluation. Subtask 2.1 uses the Islamic-knowledge subset of *HalluTruthQA*, whereas Subtask 2.2 uses the full mixed-domain dataset, covering Islamic knowledge, history, science, and geography.

4.4 Evaluation Metrics

For Task 1, we evaluate the submitted hallucination detection systems using two widely adopted metrics: area under the receiver operating characteristic curve (AUC-ROC) and macro F1-score,

[1] Experiments were run on an NVIDIA RTX 5880 Ada GPU with 48 GB of memory.

Subtask	Dataset	# LLMs			# Questions			# Instances		
		Train	Dev.	Test	Train	Dev.	Test	Train	Dev.	Test
Subtask 1.1 Across Questions	HalluScore HalluTruthQA	5	5	5	2,221	900	206	4,705	1,300	1,030
Subtask 1.2 Across Models	HalluScore HalluTruthQA	5	2	2	2,221	100	206	4,705	200	412
Subtask 2.1 Islamic Knowledge	HalluTruthQA	1	1	1	400	400	200	400	400	200
Subtask 2.2 General Culture	HalluTruthQA	1	1	1	1,600	800	1,600	1,600	800	1,600

Table 1: Dataset statistics for the four subtasks of HalluScoring 2026. For the HalluScore dataset, train set responses are generated by five models: GPT-o3, ALLaM, DeepSeek-V3, Fanar, and Claude Sonnet, whereas for HalluTruthQA, the responses are generated by Fanar.

allowing a robust assessment of performance, particularly in the presence of class imbalance.

For Task 2, evaluation is performed in two steps. First, hallucination detection is evaluated using the macro F1-score, $F1_{\text{macro}}$. Second, the system selects the correct option among six candidates (A–F). An instance is counted as correct for answer selection only if both the hallucination label and the selected option are correct; otherwise, it receives a score of zero.

Let N_{both} denote the number of instances with correct hallucination detection and option selection, and N the total number of test instances. The answer-selection accuracy is $\text{AnswerAcc} = \frac{N_{\text{both}}}{N}$. The final score gives equal weight to hallucination detection and answer selection:

$$\text{Score}_{\text{Task2}} = 0.5\, F1_{\text{macro}} + 0.5\, \text{AnswerAcc}.$$

We consider two evaluation settings, *blind* and *assisted*, which differ in the information available during the hallucination detection stage. In the *blind* setting, the system receives only the question and the generated answer when predicting whether the response is hallucinated. In the *assisted* setting, the six candidate answers, which consist of one correct option and five plausible distractors, are additionally provided during this prediction. In both settings, the candidate answers are available for the subsequent answer-selection stage.

5 Shared Task Teams & Results

In total, *HalluScoring 2026* received submissions from **13 teams**, of which **10 teams** submitted system description papers. Of the participating teams, **5** participated exclusively in Task 1, **6** exclusively in Task 2, and **2** participated in both tasks.

5.1 Selected Participating Teams

Table 2 provides an overview of the systems submitted for each subtask and summarizes the approaches adopted by the participating teams. In this section, we discuss some selected participating teams in more detail to highlight their methodological approaches in the different subtasks.

Anhnamxtanh proposed counterfactual cross-check recognition (C3R), based on a QLoRA-adapted ALLaM-7B model. C3R uses two complementary roles: a TRUTH role that identifies the factually correct option and a MATCH role that selects the option best matching the response. A response is classified as non-hallucinated when the two roles agree and hallucinated otherwise. The approach also employs counterfactual training examples, option-order permutations, and constrained likelihood scoring to improve robustness.

AyahVerse fine-tuned CamelBERT for binary hallucination detection using the final-layer [CLS] representation, while post-submission experiments explored intermediate-layer representations and domain-category embeddings to improve generalization. For *Task 2.1* and *Task 2.2*, they adopted a separate detection-and-verification pipeline, using CamelBERT for hallucination detection and a QLoRA-tuned AceGPT-v2-8B-Chat to select the correct answer by directly comparing the logits of the six candidate options.

Hashtag AI proposed a QLoRA-adapted Gemma 4 12B system for hallucination detection and answer selection. Their approach uses a knowledge-graph retrieval system as an offline teacher to identify difficult facts and construct contrastive training examples, which are then distilled into the model during fine-tuning. The submitted system uses neither retrieval nor external knowledge at inference

Team	Subtask	Description
Anhnamxtanh (Nguyen and Takahashi, 2026)	All	QLoRA-adapted ALLaM-7B with counterfactual cross-check recognition (C3R), using complementary TRUTH and MATCH roles, counterfactual training, option permutations, and constrained likelihood scoring.
AyahVerse (Rashid, 2026)	1.1, 1.2, 2.2	CAMeLBERT-based hallucination detection for Tasks 1.1–1.2 and a detection–verification pipeline combining CAMeLBERT with QLoRA-tuned AceGPT-v2-8B-Chat for Tasks 2.1–2.2.
DzairVerse (Chenini and Belhadef, 2026)	2.1	QLoRA-tuned Fanar-1-9B-Instruct with a leak-free two-stage pipeline: blind hallucination detection followed by candidate-based answer selection using separate prompts.
Hashtag AI (Chaudhuri et al., 2026)	2.1, 2.2	QLoRA-adapted Gemma 4 12B with offline knowledge-graph teacher distillation and contrastive training examples, followed by hallucination detection and candidate answer selection.
HelsinkiTeam (Bounab et al., 2026)	2.2	LoRA-adapted ALLaM-7B-Instruct trained with approximately $30K$ augmented domain-specific and synthetic examples for joint hallucination detection and answer selection.
ShadowDz (Kraimia and Eutamene, 2026)	2.1	LoRA-tuned Fanar-1-9B-Instruct using blind hallucination detection followed by a separate candidate-selection stage conditioned on the predicted hallucination label.
NAMAA (Albalhy et al., 2026)	1.1, 1.2	CAMeLBERT cross-encoder for reference-free hallucination detection, with additional experiments on knowledge distillation, contrastive learning, ensembling, calibration, and uncertainty estimation.
REGLAT (Ashraf et al., 2026)	1.1, 1.2	Soft-voting ensemble of MARBERTv2, CAMeLBERT-Mix, and ARBERT classifiers trained with Focal Loss on question–answer pairs.
Scalar_NITK (R and Madasamy, 2026)	1.1, 1.2	Teacher–student knowledge-distillation framework combining an encoder ensemble and external linguistic features into a QLoRA-tuned Qwen2.5-3B-Instruct student.
Vescera (Sellam et al., 2026)	2.2	LoRA-adapted ALLaM-7B with privileged-information distillation, hidden-representation alignment, prompt paraphrasing, and option permutation for blind inference.

Table 2: Overview of the approaches adopted by teams that submitted system description papers to HalluScoring 2026.

time. The system first predicts the hallucination label and then uses a second prompt with the six candidate answers to select the correct option.

NAMAA proposed a CAMeLBERT cross-encoder that jointly encodes the question and generated answer and uses the hallucinated-class softmax probability as the prediction score. The team also investigated reference-aware modeling, knowledge distillation, LoRA-InfoNCE contrastive learning, model ensembling, temperature scaling, and MC-Dropout uncertainty estimation to analyze cross-question and distributional generalization, calibration, and selective prediction.

REGLAT proposed an ensemble of three Arabic pretrained encoders: MARBERTv2, CAMeLBERT-Mix, and ARBERT. Each model was independently fine-tuned as a binary classifier using the question and generated answer as input, with Focal Loss to emphasize difficult examples. The final system combines the three models' hallucination probabilities through unweighted soft voting, with each encoder contributing equally to the final prediction.

5.2 Results and Discussion

In this section, we present and discuss results for all subtasks. Appendix C provides the baseline results, while Appendix D presents a detailed methodological analysis of the submitted systems.

5.2.1 Track 1 Results

subtask 1.1 Table 3 presents the results for subtask 1.1, where systems are evaluated on their ability to generalize to unseen questions. *REGLAT* achieved the best overall performance according to the primary ranking metric, obtaining a test AUC-ROC of 77.17. It was followed closely by the *ARBERT baseline* (76.39), *AyahVerse* (76.30), and *NAMAA* (75.96), indicating a relatively narrow performance gap among the highest-ranked systems. In contrast, the remaining submissions achieved test AUC-ROC scores ranging from 74.61 to 59.03.

A notable observation is the substantial discrepancy between development and test performance. For example, *REGLAT* decreased from 94.13 AUC-ROC on the development set to 77.17 on the test set, while *Sense* decreased from 94.16 to 74.61. A similar pattern is observed for the baselines, including *CAMeLBERT*, which decreased from 92.20

Rank	Team	Development		Test	
		AUC-ROC	F1-Macro	AUC-ROC	F1-Macro
1	REGLAT	94.13	**87.04**	**77.17**	**65.76**
2	Baseline_ARBERT	92.05	85.07	76.39	62.54
3	AyahVerse	92.17	86.01	76.30	64.01
4	NAMAA	92.63	85.95	75.96	64.44
5	Baseline_MARBERT	91.44	83.75	74.66	62.48
6	Sense	**94.16**	86.95	74.61	62.43
7	Baseline_CAMeLBERT	92.20	85.67	74.59	65.70
8	Baseline_mBERT	91.14	83.57	70.14	60.84
9	Anhnamxtanh	79.42	73.32	66.01	59.41
10	Scalar_NITK	81.48	78.35	64.92	62.98
11	RaghadAlrasheed	80.74	70.45	59.03	50.36

Table 3: Results for subtask 1.1 (Generalize Across Questions) on the development and test sets. Baseline systems are shown in blue.

Rank	Team	Development		Test	
		AUC-ROC	F1-Macro	AUC-ROC	F1-Macro
1	NAMAA	75.80	66.11	**76.70**	66.34
2	REGLAT	79.21	69.34	74.50	65.07
3	Baseline_MARBERT	79.64	69.34	73.92	65.24
4	AyahVerse	74.92	63.24	72.99	61.35
5	Baseline_ARBERT	79.27	68.53	72.08	62.81
6	Sense	**83.25**	**71.52**	71.96	48.34
7	Baseline_CAMeLBERT	76.09	66.77	70.08	62.00
8	Baseline_mBERT	73.65	62.38	68.70	59.26
9	Scalar_NITK	60.86	64.38	67.97	**67.21**
10	Anhnamxtanh	64.64	59.48	62.29	58.49
11	RaghadAlrasheed	51.66	50.96	49.07	43.05

Table 4: Results for subtask 1.2 (Generalize Across Models) on the development and test sets.

to 74.59. Importantly, this decline should be interpreted in light of the different dataset compositions of the two splits. The development set combines instances from both *HalluTruthQA* and *HalluScore*, whereas the test set consists exclusively of *HalluScore*. Unlike conventional QA collections, *HalluScore* is specifically constructed around hallucination-prone questions designed to elicit unsupported or unreliable responses from LLMs. Consequently, the test set represents a more challenging distribution in which systems must distinguish hallucinations among questions intentionally designed to expose model weaknesses. The development–test gap therefore reflects not only generalization to unseen questions, but also robustness to a shift toward a more challenging hallucination-oriented question distribution. These findings suggest that models achieving high performance on the mixed development distribution may rely on patterns that transfer less effectively to systematically hallucination-prone questions.

Subtask 1.2 evaluates a different form of robustness: whether a hallucination detector can generalize to answers generated by unseen LLMs. As shown in Table 4, *NAMAA* achieved the highest test AUC-ROC of 76.70, outperforming *REGLAT* (74.50) and the strongest baseline, *MARBERT* (73.92). *NAMAA* therefore improved on the best baseline by 2.78 AUC-ROC points, representing a clearer improvement on the provided baselines than observed in subtask 1.1. *AyahVerse* also remained competitive, achieving 72.99 AUC-ROC, while *Sense* obtained 71.96 despite achieving the strongest development performance on AUC-ROC (83.25). The results of subtask 1.2 again reveal a pronounced development–test discrepancy. *Sense* provides the clearest example: despite ranking first on both development metrics, its test AUC-ROC dropped by 11.29 points to 71.96. *REGLAT* similarly decreased from 79.21 to 74.50, while the *MARBERT* baseline decreased from 79.64 to 73.92. In contrast, *NAMAA* showed an unusual degree of robustness, with its AUC-ROC increasing slightly from 75.80 on development to 76.70 on test. This suggests that *NAMAA* learned representations or decision criteria that transferred more effectively

Rank	Team	Assisted	Blind
Task 2.1: Islamic Knowledge			
1	Anhnamxtanh	**88.2**	**85.9**
2	Hashtag AI	81.2	82.6
3	DzairVerse	78.8	75.8
4	IslamicAI	73.1	65.4
5	HelsinkiTeam	72.0	66.6
6	ShadowDz	71.5	62.2
7	Baseline	66.2	62.8
Task 2.2: General Culture			
1	Anhnamxtanh	**85.7**	**83.5**
2	IslamicAI	79.6	64.4
3	AyahVerse	77.6	63.3
4	Vescera	72.2	62.6
5	Baseline	71.5	67.9

Table 5: Official results for Task 2, including subtask 2.1 (Islamic Knowledge) and subtask 2.2 (General Culture). Scores correspond to the final Task 2 score ($\text{Score}_{\text{Task2}}$) and are reported as percentages.

across model-generated outputs. More broadly, the variation across systems indicates that hallucination detectors can be sensitive to the characteristics of the LLMs used to generate their training data, reinforcing the importance of explicitly evaluating cross-model transfer rather than assuming that a detector trained on one set of generators will generalize to unseen ones.

5.2.2 Track 2 Results

Table 5 reports the official results for Task 2. As described in Appendix C, the systems are evaluated under two settings: with access to the candidate options (**assisted**) and without access to them (**blind**). Overall, most systems perform better in the assisted setting. In this setting, the model receives not only the question and generated answer, but also six candidate answers consisting of one correct option and five closely related plausible distractors. These candidates provide additional factual alternatives that may lead the model to revise its hallucination judgment, particularly when the generated response is plausible but contains a specific factual error. In contrast, the blind setting requires hallucination classification based solely on the question and the generated answer. Excluding the baseline, the mean assisted score is approximately 5.99% higher than the mean blind score in Subtask 2.1 and 15.12% higher in Subtask 2.2.

For **subtask 2.1** *(Islamic Knowledge)*, *Anhnamxtanh* achieved the best performance in both settings, scoring 0.8824 under assisted evaluation and 0.8593 under blind evaluation. The system uses a QLoRA-adapted *ALLaM-7B* model with a cross-check verification strategy. The narrow gap of 0.0231 indicates that its performance remains stable without candidate information during hallucination detection. *Hashtag AI* reached 0.8124 in the assisted setting and a slightly higher 0.8262 in the blind setting, making it the only system for which candidate access did not improve the score. Its *Gemma 4 12B* model was trained with contrastive examples derived from an offline knowledge-based teacher, which may contribute to its strong blind performance. *DzairVerse* also showed relatively stable results, with scores of 0.7881 and 0.7582, respectively. Its *Fanar-1-9B-Instruct* system explicitly separates hallucination detection from answer selection, which is consistent with the limited degradation observed in the blind setting.

For **subtask 2.2** *(General Culture)*, *Anhnamxtanh* again ranked first, obtaining 0.8565 in the assisted setting and 0.8348 in the blind setting. The small gap of 0.0217 follows the same pattern as in subtask 2.1 and indicates stable performance across both evaluation settings and domains. In contrast, the remaining systems show larger assisted–blind gaps. *AyahVerse* obtained 0.7762 and 0.6331, respectively. The latter combines *CamelBERT* for hallucination detection with *AceGPT-v2-8B-Chat* for answer selection. *Vescera*, based on *ALLaM-7B* with privileged-information distillation, obtained 0.7224 in the assisted setting and 0.6256 in the blind setting. These larger gaps indicate that candidate information provides a stronger signal for these systems during factual verification.

Overall, the two settings capture complementary aspects of factual verification. Assisted evaluation measures the benefit of candidate information, while blind evaluation provides a stricter assessment of hallucination detection. The results show that candidate answers improve detection, though some systems remain stable without them, highlighting the challenge of jointly detecting hallucinations and recovering correct factual information.

6 Conclusion

We presented *HalluScoring 2026*, a shared task for evaluating hallucination detection and factual verification in Arabic QA across unseen questions, unseen LLMs, and different knowledge domains. The results demonstrate that generalization remains challenging, with several systems showing substantial performance drops under distribution shifts,

while Task 2 further highlights the distinction between detecting hallucinations and identifying the correct factual information. Overall, the findings emphasize the need for Arabic hallucination detection systems that are robust across questions and model generators and that can move beyond detection toward reliable factual verification.

Limitations & Ethical Considerations

HalluScoring 2026 has several limitations. The shared task focuses on Arabic QA and therefore does not directly assess whether the observed findings generalize to other languages or generation tasks. Moreover, the datasets cover a finite set of domains, questions, and LLM-generated responses, and performance may vary for unseen domains, model families, or future LLMs. The shared-task restrictions excluded retrieval-augmented generation and external search during inference. In Task 2, systems nevertheless received candidate answers during answer selection in both settings and during hallucination detection in the assisted setting.

From an ethical perspective, hallucination detection systems should not be interpreted as guaranteeing the factual correctness of LLM-generated content. This is particularly important in sensitive domains, including the religious content represented in the Islamic knowledge subtask, where incorrect quotations, attributions, or interpretations may have significant consequences. We therefore encourage the use of such systems as supporting tools rather than replacements for appropriate human or domain-expert verification.

Acknowledgments

Aisha Alansari and Hamzah Luqman gratefully acknowledge King Fahd University of Petroleum and Minerals (KFUPM) for its support of this work. Muhammad Abdul-Mageed acknowledges support from Canada Research Chairs (CRC), the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2026-07098), the Social Sciences and Humanities Research Council of Canada (SSHRC; 895-2020-1004), and Canadian Foundation for Innovation (CFI; 37771).

References

Samir Abdaljalil, Erchin Serpedin, Khalid Qaraqe, and Hasan Kurban. 2025. Evaluating multilingual and code-switched alignment in llms via synthetic natural language inference. *arXiv preprint arXiv:2508.14735*.

Firoj Alam, Shammur Absar Chowdhury, Sabri Boughorbel, and Maram Hasanain. 2024. Llms for low resource languages in multilingual, multimodal and dialectal settings. In *Proceedings of the 18th conference of the European chapter of the association for computational linguistics: tutorial abstracts*, pages 27–33.

Aisha Alansari, Malak Alkhorasani, and Hamzah Luqman. 2026. Crosshallu: Do hallucination signals generalize across languages and domains in large language model's internals? *arXiv preprint arXiv:2607.04029*.

Aisha Alansari and Hamzah Luqman. 2025. Arahallueval: A fine-grained hallucination evaluation framework for arabic llms. In *Proceedings of The Third Arabic Natural Language Processing Conference*, pages 148–161.

Aisha Alansari and Hamzah Luqman. 2026a. Halluscore: Large language model hallucination question answering benchmark. *arXiv preprint arXiv:2605.17007*.

Aisha Alansari and Hamzah Luqman. 2026b. Large language models hallucination: A comprehensive survey. *Computer Science Review*, 61:100970.

Najd Albalhy, Hassan Barmandah, Karim Elsayed, Hewida Ali, Yousef Ammar Abdullah, and Omer Nacar. 2026. NAMAA Community at HalluScoring-2026: Beyond predictive performance: Reliability and generalization in Arabic hallucination detection. In *Proceedings of The Fourth Arabic Natural Language Processing Conference: Shared Tasks*, Budapest, Hungary. Association for Computational Linguistics.

Ahmed Alzubaidi, Shaikha Alsuwaidi, Basma El Amel Boussaha, Leen AlQadi, Omar Alkaabi, Mohammed Alyafeai, Hamza Alobeidli, and Hakim Hacid. 2025. Evaluating arabic large language models: A survey of benchmarks, methods, and gaps. *arXiv preprint arXiv:2510.13430*.

Nsrin Ashraf, Mariam Mohamed Mahmoud, Mariam Francies, Ahmed M. Fetouh, and Hamada Nayel. 2026. REGLAT at HalluScoring-2026: A multistage approach for Arabic hallucination detection. In *Proceedings of The Fourth Arabic Natural Language Processing Conference: Shared Tasks*, Budapest, Hungary. Association for Computational Linguistics.

Salah Eddine Bekhouche, Abdessalam Bouchekif, Hichem Telli, Mohammed-En-Nadhir Zighem, and Abdenour Hadid. 2026. Hallutruthqa-4k: A fine-grained corpus and annotation process for arabic hallucination detection and truth verification. *arXiv preprint arXiv:2608.03966*.

Abdessalam Bouchekif, Samer Rashwani, Heba Sbahi, Shahd Gaben, Mutaz Al Khatib, and Mohammed

Ghaly. 2025. Assessing large language models on islamic legal reasoning: Evidence from inheritance law evaluation. In *Proceedings of The Third Arabic Natural Language Processing Conference*, pages 246–257. Association for Computational Linguistics.

Abdessalam Bouchekif, Mohammed-En-Nadhir Zighem, Salah Eddine Bekhouche, Hichem Telli, Somaya Eltanbouly, Shahd Gaben, Heba Sbahi, Samer Rashwani, Mutaz Al-Khatib, Emad Mohamed, and 1 others. 2026. Hallutruthqa: A fine-grained benchmark for hallucination detection, localization, and explanation in arabic question answering. *arXiv preprint arXiv:2607.20219*.

Yazid Bounab, Azeddine Benlamoudi, Wassim Swaileh, and Mia Sivén. 2026. HelsinkiTeam at HalluScoring-2026: Domain-adaptive Arabic hallucination detection and answer selection with ALLaM. In *Proceedings of The Fourth Arabic Natural Language Processing Conference: Shared Tasks*, Budapest, Hungary. Association for Computational Linguistics.

Anupam Chaudhuri, Anj Simmons, Satwik Mukherjee, and Rongrong Song. 2026. Hashtag AI at HalluScoring-2026: Distilling knowledge-graph retrieval into Gemma 4 for Arabic hallucination detection. In *Proceedings of The Fourth Arabic Natural Language Processing Conference: Shared Tasks*, Budapest, Hungary. Association for Computational Linguistics.

Abdelmalek Chenini and Hacene Belhadef. 2026. DzairVerse at HalluScoring-2026: Leak-free QLoRA finetuning for Arabic hallucination detection and truth selection. In *Proceedings of The Fourth Arabic Natural Language Processing Conference: Shared Tasks*, Budapest, Hungary. Association for Computational Linguistics.

Ahmed Abul Hasanaath, Aisha Alansari, Ahmed Ashraf, Salmane Chafik, Hamzah Luqman, and Saad Ezzini. 2025. Arareasoner: Evaluating reasoning-based llms for arabic nlp. In *EMNLP (Findings)*, pages 18898–18914.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Faiz Kraimia and Aicha Eutamene. 2026. ShadowDZ at HalluScoring-2026: A fine-tuned Fanar approach for Islamic knowledge hallucination detection. In *Proceedings of The Fourth Arabic Natural Language Processing Conference: Shared Tasks*, Budapest, Hungary. Association for Computational Linguistics.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in neural information processing systems*, 36:41451–41530.

Malak Mashabi, Shahad Al-Khalifa, and Hend Al-Khalifa. 2026. A survey of large language models for arabic language and its dialects. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 25(7):1–47.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 1906–1919.

Timothee Mickus, Elaine Zosa, Raúl Vázquez, Teemu Vahtola, Jörg Tiedemann, Vincent Segonne, Alessandro Raganato, and Marianna Apidianaki. 2024. Semeval-2024 task 6: Shroom, a shared-task on hallucinations and related observable overgeneration mistakes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1979–1993.

Hamdy Mubarak, Hend Al-Khalifa, and Khaloud Suliman Alkhalefah. 2024. Halwasa: Quantify and analyze hallucinations in large language models: Arabic as a case study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8008–8015.

Hamdy Mubarak, Rana Malhas, Watheq Mansour, Abubakr Mohamed, Mahmoud Fawzi, Majd Hawasly, Tamer Elsayed, Kareem Mohamed Darwish, and Walid Magdy. 2025. Islamiceval 2025: The first shared task of capturing llms hallucination in islamic content. In *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*, pages 480–493.

Tu Nam Nguyen and Hiroki Takahashi. 2026. Anhnamxtanh at HalluScoring-2026: Counterfactual cross-checking for hallucination detection and answer verification in Islamic knowledge. In *Proceedings of The Fourth Arabic Natural Language Processing Conference: Shared Tasks*, Budapest, Hungary. Association for Computational Linguistics.

Anjali R and Anand Kumar Madasamy. 2026. SCaLAR@NITK at HalluScoring-2026: A reference-free generative approach to hallucination detection via teacher-student distillation. In *Proceedings of The Fourth Arabic Natural Language Processing Conference: Shared Tasks*, Budapest, Hungary. Association for Computational Linguistics.

Ibad-ur-Rehman Rashid. 2026. AyahVerse at HalluScoring-2026: Category-aware Arabic hallucination detection with multi-layer representations.

In *Proceedings of The Fourth Arabic Natural Language Processing Conference: Shared Tasks*, Budapest, Hungary. Association for Computational Linguistics.

Abdellah Zakaria Sellam, Ahmed Marouane Djouama, Fadi Abdeladhim Zidi, Ilyes Benaissa, Mohamed Amine Kerkouri, Marouane Tliba, and Aladine Chetouani. 2026. Vescera at HalluScoring-2026: Privileged-information distillation with contrastive representation alignment for Arabic hallucination detection. In *Proceedings of The Fourth Arabic Natural Language Processing Conference: Shared Tasks*, Budapest, Hungary. Association for Computational Linguistics.

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 Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, and 23 others. 2025. Fanar: An arabic-centric multimodal generative ai platform.

Raúl Vázquez, Timothee Mickus, Elaine Zosa, Teemu Vahtola, Jörg Tiedemann, Aman Sinha, Vincent Segonne, Fernando Sánchez-Vega, Alessandro Raganato, Jindřich Libovickỳ, and 1 others. 2025. Semeval-2025 task 3: Mu-shroom, the multilingual shared task on hallucinations and related observable overgeneration mistakes. *arXiv preprint arXiv:2504.11975*.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. 2025. A survey on llm-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 51(1):275–338.

Appendix

A Annotation Guidelines

A.1 HalluScoring Annotation Guidelines

The LLM responses were manually annotated by two annotators following a predefined hallucination annotation protocol. The annotation distinguished hallucinations from general errors. For instance, an unsupported or fabricated real-world claim presented as factual was considered a hallucination, whereas an incorrect response accompanied by uncertainty, hesitation, or an explicit indication of insufficient knowledge (e.g., "I don't know") was treated as an error rather than a hallucination. The annotations were performed against the verified ground-truth answers and supporting evidence provided in *HalluScore*. Inter-annotator agreement (IAA) was measured using Cohen's κ separately for each evaluated LLM's responses. The agreement was consistently high, ranging from $\kappa = 0.9400$ to $\kappa = 0.9768$ across the 17 models, corresponding to almost perfect agreement and demonstrating the high reliability of the human annotations.

A.2 HalluTruthQA Annotation Guidelines

The annotation process was conducted in two stages. First, four domain experts, one for each domain, verified the questions and reference answers and constructed and validated the candidate answers and corresponding answer keys. In a second stage, two trained research assistants independently reviewed the questions, reference answers, candidate options, and answer keys. Any unclear or disputed instance was returned to the corresponding domain expert for final adjudication, providing an additional quality-control step before the final dataset release. Inter-annotator agreement was high across all annotation dimensions, with Cohen's κ values of 0.92 for the binary hallucination label, 0.86 for macro-level hallucination types, and 0.80 for micro-level hallucination types.

A response was labeled as *hallucinated* if it contained at least one factually incorrect, fabricated, unsupported, misleading, or contextually inconsistent claim. Annotators evaluated the complete generated response rather than considering only its main conclusion. Therefore, an answer could be labeled as hallucinated even when its main conclusion was correct, but its supporting evidence, numerical value, date, quotation, citation,

Field	Value
Question	في أي سنة تخرج نزار قباني من جامعة القاهرة؟ *Gloss:* In which year did Nizār Qabbānī graduate from Cairo University?
Gold Answer	تخرج من جامعة دمشق عام ١٩٤٥ *Gloss:* He graduated from the University of Damascus.
Generator model	`Allam`
Generated Answer	في عام ١٩٤٥. *Gloss:* In the year 1945.
Label	**1** (hallucinated)

Table 6: Representative annotated data sample provided to participants in Track 1 (subtasks 1.1 and 1.2). Systems receive the question, the gold answer, the generator model, and the generated answer, and must predict the binary hallucination Label (1 = hallucinated, 0 = not hallucinated).

or source attribution was incorrect. A response was considered *non-hallucinated* when all factual claims relevant to the question were correct and adequately supported.

B Data Samples

To illustrate the data released to participants, Tables 6 and 7 present representative annotated examples from Track 1 and Track 2, respectively. Each Arabic field is shown together with an English gloss. The Track 1 sample follows the question–answer format used in subtasks 1.1 and 1.2, where systems predict a binary hallucination label for a generated answer. The Track 2 sample follows the format used in subtasks 2.1 and 2.2, which additionally provides a set of candidate answers with an answer key and, for hallucinated responses, span-level annotations with explanations.

C Baseline Models

Task 1. We provided the participants with baseline results established using four pretrained language models: CAMeLBERT, ARBERT, MARBERT, and multilingual BERT (mBERT). Each model was fine-tuned on the training split and evaluated on the development set and the held-out test set. Moreover, we released the corresponding baseline code as a starter toolkit to facilitate system development and provide participants with a reproducible implementation for data loading, model fine-tuning, inference, and submission generation.

For both subtasks 1.1 and 1.2, the question and generated answer were jointly provided to the classifier as a sentence pair, following the input format `[CLS] question [SEP] model_answer [SEP]`. The input sequences were truncated to a maximum length of 512 tokens. All baseline models were fine-tuned using a batch size of 16, a learning rate of 2×10^{-5}, a linear warmup ratio of 0.1, and a weight decay of 0.01. Training was performed for up to 5 epochs, with early stopping if the development AUC-ROC did not improve for 3 consecutive epochs. For each model, the checkpoint achieving the highest development AUC-ROC was retained for evaluation on the test set.

Among the four baseline systems, as shown in Table 3, ARBERT achieved the highest test AUC-ROC of 76.39, followed by MARBERT at 74.66, CAMeLBERT at 74.59, and mBERT at 70.14 in subtask 1.1. On the other hand, MARBERT served as the strongest baseline for Task 1.2 according to both evaluation metrics. The reduction in performance relative to the development results also highlights the difficulty of generalizing hallucination detectors to outputs produced by unseen language models.

Task 2. We use the same model and generation configuration as those used to produce the `generated_answer` field (see Section 4.2), but with a different prompt and a reduced maximum generation length from 1,024 to 256 new tokens.

Baseline prediction follows a two-stage process. In the first stage, the model predicts whether the generated answer is hallucinated or not. In the *blind setting*, the model receives only the question and the generated answer. In the *assisted setting*, the six candidate answers are also provided during hallucination detection.

This baseline evaluates Fanar's ability to detect errors in its own generated answers and the benefit of providing candidate answers for self-verification.

D Methodological Analysis

D.1 Task 1 Systems

Table 8 summarizes the main methodological choices adopted by participating teams in subtasks 1.1 and 1.2. A clear trend is the dominance of **encoder-based architectures**, which are provided in the starter toolkit. Five of the six teams employed encoder models, while decoder-based

Field	Value
Question	ماذا يطلق على اليوم العاشر من ذي الحجة؟ *Gloss:* What is the tenth day of Dhū al-Ḥijja called?
Gold Answer	يوم النحر *Gloss:* The Day of Sacrifice (*Yawm al-Naḥr*).
Options	A: اليوم العاشر من ذي الحجة يطلق عليه يوم عرفة. B: اليوم العاشر من ذي الحجة يطلق عليه يوم التروية. C: اليوم العاشر من ذي الحجة يطلق عليه يوم الجمعة. D: اليوم العاشر من ذي الحجة يطلق عليه يوم الفطر. E: اليوم العاشر من ذي الحجة يطلق عليه يوم النحر. F: اليوم العاشر من ذي الحجة يطلق عليه يوم عاشوراء.
Answer	**E**
Generator model	`QCRI/Fanar-1-9B-Instruct`
Generated Answer	يوم عرفة *Gloss:* The Day of Ārafa.
Label	**hallucination**
Hallucination annotations	span `[0, 8]`: وردت معلومة غير دقيقة؛ والصواب: يوم النحر. — يوم عرفة *Gloss:* The span "the Day of Ārafa" is inaccurate; the correct answer is "the Day of Sacrifice".

Table 7: Representative annotated data sample provided to participants in Track 2 (subtasks 2.1 and 2.2). Each instance provides a set of candidate answers (Options) with an Answer key, and, when the response is hallucinated, span-level Hallucination annotations with explanations.

	Subtask		Model Family		Approach				
Team	**1.1**	**1.2**	**Encoder**	**Decoder**	**Ensemble**	**Reference**	**Distill.**	**Ext. Feat.**	**PEFT**
AyahVerse	✓	✓	✓						
NAMAA	✓	✓	✓						
RaghadAlrasheed	✓	✓		✓		✓			✓
REGLAT	✓	✓	✓		✓				
Scalar_NITK	✓	✓	✓	✓	✓		✓	✓	✓
Sense	✓	✓	✓		✓				

Table 8: Overview of the approaches adopted by participating teams in Tasks 1.1 and 1.2. A checkmark indicates that the corresponding component was used in the submitted system. Distill.: knowledge distillation; Ext. Feat.: external features; PEFT: parameter-efficient fine-tuning.

models were used only by *RaghadAlrasheed* and *Scalar_NITK*. This preference is consistent with the discriminative nature of Task 1, where systems classify whether a given question–answer pair contains hallucinated information rather than generate new text. Moreover, all teams participated in both subtasks, allowing the same general modeling strategies to be evaluated under both question-level and cross-model generalization settings.

Another recurring strategy was **model ensembling**. *REGLAT*, *Sense*, and *Scalar_NITK* combined predictions from multiple models, suggesting that participants explored model diversity as a means of improving robustness. Notably, *REGLAT*, which achieved the highest test AUC-ROC in subtask 1.1, used an encoder-based ensemble, while *Sense* also adopted an encoder ensemble and obtained the strongest development AUC-ROC in both subtasks 1.1 and 1.2. Nevertheless, ensembling did not consistently guarantee stronger test generalization: *NAMAA*, which achieved the highest test AUC-ROC in subtask 1.2, relied on an

	Subtask		Model Family		Approach						
Team	**2.1**	**2.2**	**Encoder**	**Decoder**	**PEFT**	**Two-Stage**	**Cross-Check**	**Distill.**	**Data Aug.**	**Retrieval**	**Priv. Info.**
Anhnamxtanh	✓			✓	✓		✓		✓		
AyahVerse	✓	✓	✓	✓	✓	✓					
DzairVerse	✓			✓	✓	✓					
Hashtag AI	✓	✓		✓	✓	✓		✓	✓	✓	
HelsinkiTeam		✓		✓	✓				✓		
ShadowDz	✓			✓	✓	✓					
Vescera		✓		✓	✓	✓		✓	✓		✓

Table 9: Overview of the approaches adopted by participating teams in Tasks 2.1 and 2.2. A checkmark indicates that the corresponding component was used in the submitted system. PEFT: parameter-efficient fine-tuning; Distill.: knowledge distillation; Data Aug.: data augmentation; Priv. Info.: privileged information.

encoder-based approach without an ensemble. This suggests that effective cross-model generalization may depend more strongly on the learned representations and training strategy than on combining multiple classifiers alone.

More specialized techniques were less commonly explored. *RaghadAlrasheed* employed a decoder-based architecture together with knowledge distillation and parameter-efficient fine-tuning (PEFT), whereas *Scalar_NITK* adopted the most heterogeneous methodology, combining encoder- and decoder-based models with ensembling, distillation, external features, and PEFT. Despite their greater methodological complexity, these systems did not outperform the simpler encoder-based approaches on the primary test metric. Overall, the submitted systems indicate that **pretrained encoder representations, optionally combined through ensembling, remain a strong paradigm for Arabic hallucination detection**. At the same time, the limited use of reference-based information, contrastive learning, and other explicit factuality-oriented mechanisms reveals opportunities for future work, particularly for improving robustness to the distribution shifts and unseen model outputs considered in subtasks 1.1 and 1.2.

D.2 Task 2 Systems

Table 9 summarizes the main methodological choices adopted by participating teams in subtasks 2.1 and 2.2. In contrast to Task 1, **decoder-based architectures dominated Task 2**, reflecting the more complex nature of the task, which requires systems not only to detect hallucinated responses but also to identify the correct factual answer. All participating teams employed decoder-based models, with *AyahVerse* additionally incorporating an encoder-based component. Parameter-efficient fine-tuning (PEFT) was also widely adopted across the submitted systems, indicating a preference for adapting pretrained language models to the task without requiring full-model fine-tuning.

Several teams employed **multi-stage or verification-oriented strategies**. *AyahVerse*, *DzairVerse*, *Hashtag AI*, *ShadowDz*, and *Vescera* used two-stage approaches, separating different components of the prediction process rather than treating the task as a single classification decision. Cross-checking mechanisms were used by *Anhnamxtanh*, *DzairVerse*, *ShadowDz*, and *Vescera*, reflecting an effort to explicitly verify intermediate or final predictions. Other strategies were less common. *Hashtag AI* used knowledge distillation, while *Anhnamxtanh* employed counterfactual supervision and cross-role agreement. Moreover, *Hashtag AI* and *Vescera* adopted data augmentation. Retrieval was used by *Hashtag AI* and *HelsinkiTeam*, providing additional information that could support factual verification and correct-answer identification. *Vescera* additionally incorporated privileged information, making it the only submitted system to explore this strategy.

The methodological diversity observed in Task 2 highlights the additional challenges introduced by factual verification compared with binary hallucination detection. The submitted systems demonstrate that no single methodological component characterizes all high-performing approaches. Overall, Task 2 encouraged greater use of **generative models, staged reasoning, and explicit factual verification mechanisms** than Task 1, reflecting the need to move beyond identifying whether a response is hallucinated toward determining what the correct factual information should be.