ShawarmaChats: A Benchmark Exact Dialogue & Evaluation Platter in Egyptian, Maghrebi & Modern Standard Arabic, A Triple-Dialect Feast for Hungry Language Models

Kamyar Zeinalipour[1], Mohamed Zaky Saad[1], Oumaima Attafi[1]
Marco Maggini[1], Marco Gori[1]

[1]DIISM, University of Siena, Via Roma 56, Siena, Italy
Correspondence: kamyar.zeinalipour2@unisi.it

Abstract

Content-grounded dialogue evaluation for Arabic remains under-resourced, particularly across Modern Standard Arabic (MSA), Egyptian, and Maghrebi varieties. We introduce **ShawarmaChats** [1], a benchmark of 30,000 six-turn conversations grounded in Wikipedia content, evenly split across the three dialects.
To build this corpus, we prompt five frontier LLMs —`GPT-4o`, `Gemini 2.5 Flash`, `Qwen-Plus`, `DeepSeek-Chat`, and `Mistral Large` to generate 1,500 seed dialogues. Native Arabic speakers evaluate these outputs to select the most effective *generator* and most human-aligned *grader*. Sub-A dialogues undergo a two-pass, rationale-driven *self-repair loop* where the grader critiques and the generator revises; unresolved cases are manually corrected. We apply this pipeline to 10,000 Wikipedia paragraphs to create 30,000 high-quality conversations 10,000 per dialect at modest human cost. To validate the benchmark, we LoRA-fine-tune six open LLMs (1 B to 24 B parameters) on `ShawarmaChats` and observe consistent gains in automatic-grader scores, BERTScore, BLEU and ROUGE particularly for models larger than 7 B parameters. `ShawarmaChats` thus establishes the first large-scale, dialect-aware, content-grounded dialogue benchmark for Arabic.

1 Introduction

Knowledge-grounded dialogue generation gauges a model's skill at weaving verifiable facts into multi-turn exchanges. English research enjoys mature resources - Wizard of Wikipedia (Dinan et al., 2019), the BEGIN attribution suite (Dziri et al., 2022b) and convenient, if imperfect, automatic metrics such as BERTScore (Zhang et al., 2020) and ROUGE (Lin, 2004). In Arabic, however, no benchmark yet unifies MSA, Egyptian, and Maghrebi varieties while enforcing grounding in sources like Wikipedia. Current efforts remain piecemeal: AraConv (Fuad et al., 2022) covers only MSA, Dial2MSA-Verified (Khered et al., 2025) tackles lexical normalisation, and recent corpora such as the multimodal Dallah (Alwajih et al., 2024) and the dialect-specific JEEM (Artemova and Trajkova, 2025) underscore rather than bridge this gap. Meanwhile, the "LLMs-as-Judges" literature (Li et al., 2024) and self-refinement loops where generators revise outputs based on model critiques (Dong et al., 2025) are reshaping evaluation and data augmentation practices. Current Arabic dialogue resources do not jointly cover MSA, Egyptian, and Maghrebi *or* provide scalable, high-precision quality control. We therefore ask:

Problem Statement

Can an *LLM-driven generator—grader self-repair loop*, requiring minimal human effort, yield a high-fidelity benchmark of six-turn, Content-grounded dialogues in all three registers?

To operationalise this goal, we decompose it into four research questions:

RQ1 What is the comparative performance of the five frontier LLMs when tasked with generating content-grounded six-turn dialogues in MSA, Egyptian, and Maghrebi?

RQ2 Which of these same models, when prompted as an automatic *grader*, aligns most closely with native-speaker judgments?

RQ3 How effectively does a two-pass, rationale-driven self-repair loop upgrade sub-A [2] dialogues, and what residual error types persist?

RQ4 Do models fine-tuned on the final corpus exhibit consistent gains in faithfulness and di-

[1]github.com/KamyarZeinalipour/Shawarma-Chats

[2]Any dialogue that does not receive an 'A' (Excellent) rating in the human/machine evaluation

Proceedings of The Third Arabic Natural Language Processing Conference, pages 472–524
November 8-9, 2025

alect control when evaluated exclusively by the automatic grader and lexical metrics?

Approach & headline results. We tackle RQ1 - RQ4 through the creation of `ShawarmaChats`. Five [3] frontier LLMs. First generate 1,500 seed six-turn dialogues. Native Arabic speakers label these outputs, identifying the most effective *generator* and the most human-aligned *grader*; the chosen grader achieves 96.3 % precision on grade-A judgements for the selected generator. All sub-A seeds enter a two-pass, rationale-driven self-repair loop in which the grader critiques and the generator revises; dialogues still below grade A after the second pass receive manual correction. This generator—grader pair is then applied to 10,000 distinct Wikipedia paragraphs, producing 30,000 grade-A conversations —10,000 per dialect—while keeping human intervention to roughly 0.52 % of cases. Finally, LoRA fine-tuning six open-source LLMs (1 B to 24 B Parameters) on `ShawarmaChats` yields consistent gains in automatic-grader scores and BERTScore, BLEU and ROUGE, with the largest relative improvements observed particularly for models larger than 7B parameters.

Building on these findings, our work makes several distinct contributions to the study of Arabic content-grounded dialogue generation, which we summarise below:

Contributions. (i) We introduce `ShawarmaChats`, the first knowledge-grounded dialogue benchmark that spans Modern Standard, Egyptian, and Maghrebi Arabic. (ii) The corpus offers 30k six-turn conversations linked to Wikipedia, vetted to 96.3% precision. (iii) A rationale–based generator - grader loop cuts human review down to 0.52 % by letting one LLM spot flaws and another fix them. (iv) Human judgments over the five frontiers reveal the best models for generation vs. grading. (v) Fine-tuning six open LLMs (1B – 24B) proves the benchmark sensitive to training regime and size. (vi) We publicly release the dataset, LoRA weights, prompt templates, and evaluation code. **Paper outline.** Section 2 reviews related work; Section 3 presents the `ShawarmaChats` generation pipeline in full; Section 4 reports our empirical results; and Section 5 summarises conclusions and limitations.

[3] —`GPT-4o`, `Gemini 2.5 Flash`, `Qwen-Plus`, `DeepSeek-Chat`, and `Mistral Large`

2 Related Work

Knowledge-grounded dialogue in English. Large-scale English datasets such as *Wizard of Wikipedia* (Dinan et al., 2019) and *Topical-Chat* (Gopalakrishnan et al., 2019) established the paradigm of multi-turn conversations explicitly anchored in external knowledge, enabling systematic study of factuality in open-domain dialogue. Subsequent work shifted from data collection to *evaluation*: *Q2* proposes a QA-based metric for factual consistency (Honovich et al., 2021), while *BEGIN* introduces fine-grained attribution labels to diagnose hallucinations (Dziri et al., 2022b). Cleaning efforts such as *FaithDial* (Dziri et al., 2022a) and fact-checking benchmarks like *DialFact* (Gupta et al., 2022) further refine data quality and supply supervised signals for hallucination detection. Our benchmark follows this line of work but is the first to bring Wikipedia-grounded, six-turn conversations to Arabic in three distinct dialects.

Automatic metrics for factuality and quality. Beyond simple lexical overlap (ROUGE (Lin, 2004)), recent learned metrics (BERTScore (Zhang et al., 2020)) and BLEURT (Sellam et al., 2020)) correlate better with human judgments, while SummEval provides a large-scale human annotation test-bed for metric validation (Fabbri et al., 2021). UniEval unifies multiple quality dimensions into a single evaluator (Zhong et al., 2022). However, these metrics are not dialect-aware and often overlook language-specific nuances; our automatic grader, chosen via human alignment experiments, fills this gap for Arabic.

Arabic dialogue and dialectal resources. Prior Arabic conversational corpora remain either domain-specific or dialect-specific. *AraConv* offers an MSA task-oriented dataset (Fuad et al., 2022), while recent Gulf-dialect corpora highlight ongoing fragmentation (Al-Shenaifi et al., 2024). Multimodal models such as *Dallah* demonstrate the community's interest in dialect-aware LLMs (Alwajih et al., 2024). A comprehensive survey confirms the scarcity of unified, multi-dialect benchmarks across Arabic NLP tasks (Joshi et al., 2025). `ShawarmaChats` closes this resource gap by providing a balanced, Wikipedia-grounded benchmark spanning Modern Standard, Egyptian and Maghrebi Arabic.

LLM-driven Data Generation. The increasing capabilities of LLMs have spurred a new wave of research focused on synthetic data generation, particularly for low-resource languages. Recent efforts have demonstrated the viability of using LLMs to automate the creation of various materials. For instance, LLMs have been successfully employed to generate quizzes in Turkish (Zeinalipour et al., 2024b) and multiple-choice questions in Persian (Zeinalipour et al., 2025a). A significant body of work has also explored the generation of crossword puzzles across different languages, including Italian (Zeinalipour et al., 2024a), Turkish (Zeinalipour et al., 2024c), and Arabic (Zeinalipour et al., 2025b,c). Techniques like Clue-Instruct further refine the generation of text-based clues for these puzzles (Zugarini et al., 2024b). Beyond generation, LLMs are also used in evaluating these materials, such as in answering crossword clues (Zugarini et al., 2024a) and providing automated feedback on student writing (Zeinalipour et al., 2024d). Furthermore, the reliance on LLMs extends to creating benchmarks for evaluating specific capabilities, such as commonsense reasoning in Arabic (Lamsiyah et al., 2025).

LLM-based evaluation and self-repair loops. Recent studies show that strong LLMs can act as reliable *judges* to evaluate text generated by smaller models (Koutchéme et al., 2024). Surveys of self-correction techniques (Kamoi et al., 2024) and zero-resource hallucination detection (SelfCheckGPT) (Manakul et al., 2023) demonstrate the feasibility of iterative generation–critique cycles. Retrieval-augmentation combined with self-checking further improves answer faithfulness in conversational QA (Ye et al., 2024). We build on these insights by selecting the most human-aligned LLM as an *automatic grader* and embedding it in a two-pass generator - grader self-repair loop, achieving grade-A quality with only minimal human intervention.

Positioning of `ShawarmaChats`. Our benchmark uniquely (i) unifies three major Arabic varieties, (ii) enforces strict Wikipedia grounding, and (iii) employs a human-validated, LLM-driven self-repair pipeline, thereby enabling rigorous evaluation of dialect control, factuality, and LLM-based grading in low-resource settings.

3 The `ShawarmaChats` Dataset

We first detail how the corpus is constructed (3.1), then provide a quantitative and qualitative analysis that motivates its research value (3.2). The ten-stage pipeline-—summarised in Figure 1 ensures both broad topical coverage and high factual fidelity.

3.1 Dataset Creation

Step 1 – Paragraph sampling. Over 200,000 Arabic Wikipedia articles were downloaded to build the THAW (Text Harvest from Wikipedia) [4] dataset. Key bolded terms and lead-section metadata were extracted using Wikipedia's uniform structure. GPT-4 was then used to classify each article into one of 29 custom categories. The distribution is shown in Figure 3. Quality filtering kept only articles $\geq$ 150 words, discarded multi-word, very short/long, or symbol/number-bearing titles, and ranked articles by popularity. A uniform sample of 10,000 paragraphs—-each $\geq$ 150 words was then drawn from articles whose importance was graded High to Low, and whose popularity was measured by peak view counts, yielding a clean, high-quality corpus for further analysis.

Step 2 – Dialect prompting. We craft *three distinct prompt templates*—-one each for MSA, Egyptian, and Maghrebi that instruct an LLM to produce a *six-turn dialogue* between two interlocutors, A and B. We empirically evaluated multiple wording variants with several language models and found that each dialect benefits from a dedicated prompt to maximise fluency and register fidelity. The final instructions, therefore, differ subtly across dialects and ask the model to return the conversation in a structured JSON schema, making subsequent automatic checks straightforward. Full prompt texts appear in Appendix G.

Step 3 – Seed generation. From the 10,000-paragraph pool we *uniformly sample 100 paragraphs* to serve as a pilot set. Each of the five frontier LLMs then answers the three dialect-specific prompts for every sampled paragraph, yielding $100 \times 3 \times 5 = 1,500$ seed dialogues that underpin our subsequent models-selection experiments.

Step 4 – Establishing the Gold Reference Set. To create a reliable gold standard, we used a two-stage evaluation process with two expert annotators:

[4] A 10k □paragraph, filtered Wikipedia pool

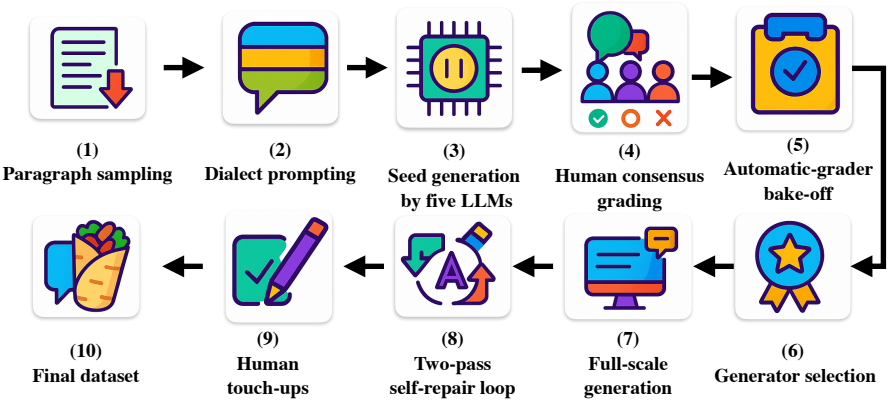

Figure 1: Overview of the ten-stage ShawarmaChats creation pipeline.

a native Egyptian and a native Moroccan Arabic speaker, both graduate students in Linguistics and Computer Science.

First, they worked **independently** to rate each of the 1,500 seed dialogues on a scale from A (Excellent) to D (Poor). The evaluation focused on four key criteria: Fluency, Faithfulness to the source text, conversational Coherence, and correct Dialect Accuracy. This initial blind pass showed substantial inter-annotator agreement, achieving a Cohen's κ of 0.794.

Next, the annotators came together to **discuss and resolve** every instance where their initial ratings differed. Their combined fluency across all three Arabic varieties was crucial for settling nuanced cases. This collaborative second stage resulted in a single, high-confidence **consensus grade** (A, B, C, D) for each dialogue, forming the definitive gold reference set for our study. Full annotation guidelines are detailed in Appendix F.

Step 5 – Automatic Grader Selection via Human Alignment To find a reliable automatic grader, we benchmarked five leading LLMs against our human-graded gold set. Each model was prompted to assign an A-to-D grade and a written rationale to all 1,500 seed dialogues. In the initial comparison, `GPT-4o` was the clear front-runner, achieving **80.4% accuracy** in matching the original human ratings. [5]

However, a crucial finding emerged from the models' rationales: they often highlighted subtle errors our human experts had initially missed. This prompted a **rationale-aided reassessment**, where our annotators reviewed their judgments with the selected model feedback in mind. This powerful loop led them to refine **22.15%** of the original consensus grades, resulting in a more robust **final gold standard**.

When all models were re-evaluated against this improved benchmark, `GPT-4o` maintained its top position, confirming its superior alignment with nuanced human judgment. It was therefore selected as the official automatic **Grader** for our pipeline. Full performance details for all models are in Appendix

Model	EGY	MAG	STD
GPT-4O	3.8687	3.2626	3.7980
DeepSeek-Chat	3.9394	2.5354	3.8182
Gemini 2.5 Flash	**3.9798**	**3.6667**	**3.9596**
Mistral Large	3.8485	1.2020	3.7172
Qwen-Plus	2.7778	2.3030	3.5152

Table 1: Average ratings by model across the three evaluation categories. Bolded entries denote the top-□performing model per category.

Step 6 (A) – Selecting the Best Generator. With the automatic Grader selected, we returned to our final human consensus ratings to identify the best dialogue **Generator**. We converted the A to D grades assigned to each model's output into numerical scores (A=4, D=1) and calculated the average performance. As shown in Table 1, **`Gemini 2.5 Flash`** achieved the highest overall score across all three dialects, securing its role as the Generator for our pipeline.

Step 6 (B) – Validating the Automated Pipeline. Before moving to full-scale generation, we per-

[5] The full accuracy breakdown against the initial human ratings was: `GPT-4o` (80.4%), `DeepSeek-Chat` (71.3%), `Gemini 2.5 Flash` (70.1%), `Qwen-Plus` (52.3%), and `Mistral Large` (51.6%).

formed a final, crucial validation. We needed to confirm that our selected **Grader** (`GPT-4o`) could accurately identify high-quality work from our chosen **Generator** (`Gemini 2.5 Flash`). To do this, we measured the Grader's precision on 'A'-grade dialogues against our human gold standard.
The results were excellent, confirming the pipeline's reliability. The Grader achieved an average precision of **96.3%** when identifying top-quality dialogues (99% for MSA, 100% for Egyptian, and 90% for Maghrebi). This high precision was the critical validation for our pipeline. Since any dialogue rated below 'A' would automatically undergo revision, the Grader's ability to reliably identify excellent outputs allows us to filter for quality at scale, reserving manual supervision for only a small fraction of cases.

Step 7 – Full-Scale Generation and Automated Triage. With our models in place, we generated the full dataset of 30,000 raw dialogues using our **Generator** (`Gemini 2.5 Flash`). Our automatic **Grader** (`GPT-4o`) then performed an initial quality triage on this collection. A promising **85.94%** of the conversations were immediately rated A and accepted. The remaining 14.06% were automatically funneled into our two-pass self-repair loop for quality enhancement, as detailed in Table 9.

Step 8 – The Self-Repair Loop: Automated Dialogue Refinement. Dialogues that were not rated A in the initial triage were automatically funneled into our two-pass self-repair loop. This process is designed to iteratively improve dialogue quality without human intervention, following a three-stage cycle:

1. **Critique:** First, our **Grader** (`GPT-4o`) does more than just assign a low score; it generates a detailed rationale explaining the specific flaws, such as a factual error, stilted phrasing, or incorrect dialect usage.

2. **Fix:** This actionable feedback is then packaged into a new "repair prompt." The prompt, containing the source paragraph, the flawed dialogue, and the Grader's critique, is sent to our **Generator** (`Gemini 2.5 Flash`) with instructions to revise the conversation and fix the identified issues.

3. **Re-grade:** Finally, the newly revised dialogue is sent back to the **Grader** for a fresh assessment. If it now achieves an A, it is accepted. If it still falls short, the entire 'Critique' → 'Fix' → 'Re-grade' cycle is repeated one more time.

This automated refinement process proved highly effective. While **85.94%** of dialogues passed on the first attempt, the first repair pass lifted the cumulative success rate to **97.77%**. The second pass brought the total to **99.48%**. Ultimately, this loop resolved the vast majority of issues, leaving only a minuscule **0.52%** of dialogues (fewer than 1 in 200) that required final manual correction by human experts.

Step 9 – Human touch-ups. Humans manually corrected the remaining "stubborn tail" (0.52 %).

Step 10 – Release package. Upon completing the pipeline, we merge every A-rated conversation into the definitive **`ShawarmaChats`** corpus. This high-quality resource offers a turnkey benchmark for evaluating—and advancing—content-grounded dialogue generation in Arabic.

3.2 Linguistic and Statistical Analysis

Volume and length. `ShawarmaChats` contains **22.7** M characters about 9.0 M tokens when segmented with the `Llama-3` tokenizer—across the 10,000 source Wikipedia paragraphs and the 30,000 six-turn conversations that compose the benchmark (Table 2). The encyclopedic paragraphs are the heftiest slice, averaging $\mathbf{1,353}$ characters (≈ 515 tokens) each, thus providing ample factual context for generation. Conversely, the dialogues are deliberately concise: Maghrebi turns average 129 tokens, Egyptian 119, and MSA 137 a spread that mirrors well-attested cliticisation and orthographic differences among the three varieties. Even with a fixed six-turn template, the length of the sentence remains distinctly conversational at ≈ 5 to 6 words per sentence for all dialects, compared to ≈ 19 words in the source context. The analysis shows that Arabic letters appear in 98.86 % of the corpus. Figure 2 visualises the resulting token-length distributions for both the source paragraphs and the three dialectal conversation sets.

Lexical diversity. Tokenised with the `Llama-3` [6] tokenizer, the benchmark contains ≈ 226 k unique token types out of 9.0 M total tokens, giving a corpus□-level type-–token ratio TTR $= 0.0025$ (Table 2). To obtain a size□-robust view, we also compute the moving□average type–token ratio

[6] https://huggingface.co/meta-llama/Meta-Llama-3-8B

	chars	tok	words	Avg. tok	Avg. char	Avg. word	TTR	char/word	word/sent	arabic
Text	13.5M	5.15M	1.93M	514.93	1,353	192.94	0.00424	3.68	19.33	0.983
MAG	3.0M	1.29M	0.56M	129.30	298	55.86	0.00416	3.42	5.10	0.990
EGY	2.8M	1.19M	0.49M	118.97	282	48.50	0.00393	3.57	5.31	0.991
MSA	3.4M	1.37M	0.57M	137.20	336	56.58	0.00351	3.61	5.70	0.991
TOTAL|Avg	22.7M	9.01M	3.54M	225.10	567	88.47	0.00253	3.62	9.28	0.986

Table 2: Corpus-level descriptive statistics (rounded; M = millions).

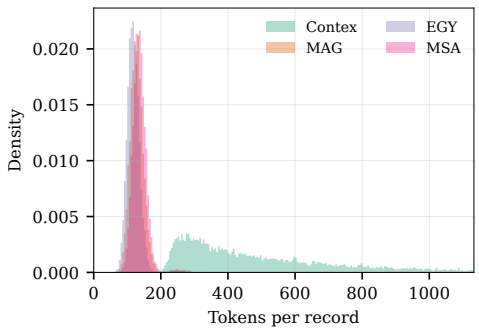

Figure 2: Token-length distributions (log-scaled density) for the source context paragraphs and the three dialectal conversation sets, computed with the `Llama-3` tokenizer.

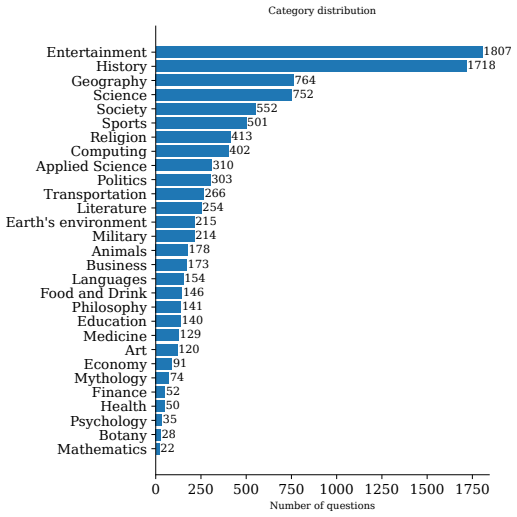

Figure 3: Distribution of the 10,000 source paragraphs over the 29 Wikipedia categories used in `ShawarmaChats`.

(MATTR) with a 500-token window:

$$\mathrm{MATTR}_{\mathrm{MSA}} = 0.567,\ \mathrm{MATTR}_{\mathrm{EGY}} = 0.527,$$

$$\mathrm{MATTR}_{\mathrm{MAG}} = 0.499,\ \mathrm{MATTR}_{\mathrm{Context}} = 0.560,$$

yielding an overall corpus value of 0.548. The ranking — MSA > Context > Egyptian > Maghrebi — follows intuitively from the varieties' orthographic norms: MSA's standardised morphology packs more distinct stems per window, while Maghrebi's heavier cliticisation and code-switched borrowings reuse subword fragments, slightly lowering its MATTR. These figures confirm that, despite the fixed six-turn template, the dialogues retain a healthy and dialect-sensitive lexical spread that is well-suited for evaluating vocabulary control and style transfer.

Part-of-speech profile. A coarse-grained UD PoS analysis (full results in Appendix B table 6) confirms the stylistic shift from encyclopædic context to dialogue. Verbs almost double in relative frequency from 8.3 % in the source paragraphs to ≈ 11% in the three dialogue sets while pronouns rise from 4.6 % to 7 ∼ 9%, signalling the more interactive register. Conversely, nouns drop from 32.3 % to 26.3 % in MSA and just 19.3 % in Maghrebi, reflecting heavier cliticisation and ellipsis. Egyptian exhibits the highest share of discourse particles and punctuation. These trends dovetail with the lexical-diversity findings reported earlier in this section.

Topic balance. To minimise topical skew we stratified paragraph sampling across **29 top–level Wikipedia categories**. As Figure 3 shows, the distribution is deliberately broad: the two largest bins, *Entertainment* (1,807 paragraphs) and *History* (1,718), together account for only 18 % of the 10,000 source paragraphs, while the median category (*Earth's environment*) still supplies over 200 examples. Even the long–tail domains—e.g. *Mathematics*, *Botany*, *Psychology* contribute ≥ 22 paragraphs each, ensuring every topic is represented. Seeding each paragraph as one dialect-specific conversation keeps the 30,000-dialogue corpus balanced, giving a realistic, evenly distributed test-bed for knowledge-grounded dialogue generation.

Frequent tokens and n-grams. A corpus-wide sweep of surface co-occurrences shows that the *ten* most frequent **tokens**, bigrams, and trigrams split cleanly into two camps (see Table 5 in the Appendix). Encyclopaedic items such as the token عام 'year', the bigram الولايات المتحدة 'United

Dialect	BERTScore F1	ROUGE-L F
Maghrebi (MAG)	0.8914	0.0465
Egyptian (EGY)	0.7926	0.0535
MSA	**0.9057**	**0.0966**

Table 3: Dialect-fidelity scores for ShawarmaChats. Higher values indicate closer alignment between dialogue turns and their grounding Wikipedia paragraphs.

States', and the trigram الحرب العالمية الثانية 'World War II' stem from the grounding paragraphs, whereas the conversations inject strongly dialect-marked forms like Maghrebi بزاف 'a lot', Egyptian عشان كده 'that's why', and the trigram أش عرفتي بلي 'how did you know that …'. This mixture confirms that ShawarmaChats interleaves fact-heavy named entities with conversational formula, furnishing an informative stress-test for both knowledge retention and dialect control in LLMs.

Dialect fidelity. We gauge each dialect's fidelity to its Wikipedia source with semantic similarity (BERTScore F1 using the `microsoft-deberta-xlarge-mnli`) and lexical overlap (ROUGE-L F). MSA tops both metrics, Maghrebi (MAG) trails in ROUGE-L yet stays second in BERTScore, consistent with its cliticisation, phonological spelling, and code-switching, and Egyptian (EGY) sits between the two. High BERTScores across all three confirm factual preservation, whereas ROUGE-L variation exposes genuine dialectal word-choice differences, stressing the need for semantics-aware evaluation beyond n-gram overlap (Table 3).

4 Experiments

This section evaluates how well ShawarmaChats transfers to open□source language models of widely varying capacity for the task of generating six turn, context grounded dialogues conditioned on a given paragraph in three dialects MSA, Egyptian, and Maghrebi Arabic. We (i) describe the data split, (ii) detail the fine□tuning recipe, (iii) specify automatic evaluation metrics, and (iv) report quantitative and qualitative results.

4.1 Experimental Setup

Models. We fine-tune six open source `Mistral-24B`, `Mistral-Nemo-12B`, `Mistral-7B`, `Llama3-8B`, `Llama3.2-3B`, and `Llama3.2-1B` spanning six parameter scales. Unless otherwise stated, all models are frozen except for a LoRA adapter (rank64, α=128) Details of the training and inference hyperparameters are provided in the Appendix. C

Data split. From the 10,000 unique ShawarmaChats paragraphs (Section 3.1), 9,500 ($\times$3 dialects = 28,500 dialogues) are used for training and 500 ($\times$3 dialects = 1,500 dialogues) for testing.

Evaluation setup. For every test paragraph we produce two dialogues—one from the *base* checkpoint and one from its *Fine-Tuned* sibling—and compare them with the gold reference in ShawarmaChats. Quality is measured by BLEU, ROUGE-L, BERTScore F_1, and a `GPT-4o`, **grader** that closely replicates human 3.1 judgments (A to D mapped to 4 to 1), capturing lexical, semantic, and holistic gains in one sweep.

4.2 Results & Analysis

Overall gains. Table 4 shows that **every** model benefits from fine-tuning on ShawarmaChats. The average relative improvement is +34.8 % for ROUGE-L, +78. % for BLEU, and +0.03 absolute points for BERTScore F1. Crucially, the *grader-derived* score—mapped from A=4 to D=1 —jumps by **+1.34 points** on average, confirming that the automatic judge perceives genuinely higher dialogue quality after adaptation. A side-by-side quantitative comparison of MSA, Egyptian and Maghrebi conversations is deferred to Appendix E.

Size matters *bigger* shifts more. Parameter-rich checkpoints ($\geq$ 7 B) extract substantially more benefit from ShawarmaChats than the tiny 1B $-$ 3B models. `Mistral-7B` and `Mistral-Nemo-12B` each gain about $+2.1$ grader points and lift ROUGE-L by $+0.14 \sim 0.19$, while `Llama3-8B` and `Mistral-24B` still add $\sim$ $+1.7$ grader points despite stronger baselines ($+0.10$ ROUGE for the latter). By contrast, the 1 B and 3 B Llama variants move only $\leq +0.33$ grader points and $<$ $+0.10$ ROUGE, implying that *model capacity*, rather than data volume, is the primary bottleneck at that scale.

Faithfulness vs lexical overlap. BERTScore improvements track the grader signal more closely than n-gram metrics, indicating that the judge is sensitive to *semantic* faithfulness rather than surface copying. For example, `Mistral-Nemo-12B` achieves the single best BERTScore (0.857) yet its ROUGE gain is moderate, mirroring the model's tendency to paraphrase rather than quote verbatim.

Error profile after fine-tuning. Figure 4 visualises the distributional shift in grader labels. Fine-tuning collapses the long tail of D (hallucinations,

Table 4: Automatic metrics on the ShawarmaChats **test** split. **Base** denotes the original instruction-tuned checkpoint; **FT** denotes the same model after LoRA fine-tuning on ShawarmaChats (§4). Higher is better. Best scores per metric are **bold**.

	Llama3-1B		Llama3-3B		Mistral-7B		Llama3-8B		Mistral-Nemo-12B		Mistral-24B	
Metric	Base	FT	Base	FT	Base	FT	Base	FT	Base	FT	Base	FT
ROUGE-L ↑	0.220	0.318	0.298	0.310	0.304	**0.443**	0.337	0.433	0.246	0.439	0.413	**0.443**
BLEU ↑	0.088	0.198	0.101	0.194	0.182	0.286	0.152	0.279	0.141	0.284	0.257	**0.287**
BERTScore F1 ↑	0.763	0.767	0.797	0.745	0.767	0.856	0.821	0.856	0.772	**0.857**	0.838	**0.857**
Grader Avg. ↑	1.003	1.103	1.058	1.389	1.186	3.302	1.440	3.172	1.313	3.345	1.854	**3.607**

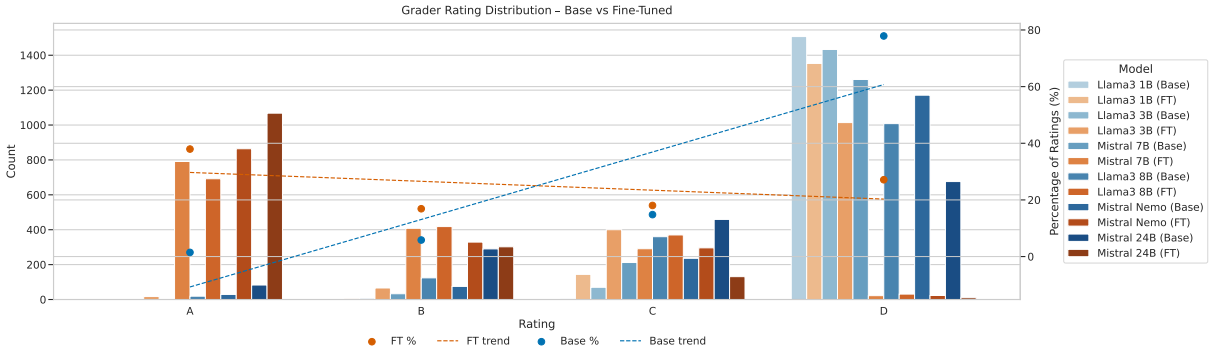

Figure 4: Shift in automatic–grader ratings (A–D) before and after fine-tuning. All six models show a pronounced migration from lower grades (C/D) to high-quality A/B grades.

dialect slips) and converts many Cs (minor factual drift) into solid B/A outputs. The surge of red in the A column as well as the steeper downward trend line across B−D shows that fine-tuning on ShawarmaChats systematically pushes dialogues toward higher quality grades, underscoring the dataset's effectiveness as a supervision signal.

Fine-tuning Efficacy. In summary, the experimental results consistently validate ShawarmaChats as a potent fine-tuning resource. The pronounced shift from lower grades towards high-quality A/B outputs, especially for models larger than 7B parameters (Figure 4), confirms that the benchmark provides a strong signal for improving both factual grounding and dialectal control in open-source LLMs.

5 Conclusion and Future Work

Can an *LLM-driven generator -grader self-repair loop*, with only minimal human effort, create a high-fidelity benchmark of six-turn, content-grounded dialogues in Modern Standard, Egyptian, and Maghrebi Arabic? Our results demonstrate that the answer is *yes*. By combining a carefully chosen generator (Gemini 2.5 Flash) with a highly precise automatic grader (GPT-4o) and iterating through a two-pass critique–revision cycle, we produced ShawarmaChats: 30,000 Wikipedia-grounded conversations that achieve 99.48 % grade-A precision while requiring human intervention in fewer than 0.52 % of cases.

Answers to the research questions.

RQ1 *Generator quality.* Among five frontier LLMs, **Gemini 2.5 Flash** delivered the most fluent, faithful, and dialect-accurate six-turn dialogues across all three registers.

RQ2 *Grader alignment.* **GPT-4o**, prompted as a judge, aligned best with expert annotators, achieving 80 % raw agreement and 96.3 % precision on grade-A decisions on the selected generator.

RQ3 *Effectiveness of self-repair.* A two-pass, rationale-driven loop lifted the share of grade-A dialogues from 85.94 % to 99.48 %, leaving only a 0.52 % residue for manual clean-up.

RQ4 *Downstream impact.* LoRA fine-tuning six open-source LLMs (1B to 24B) on ShawarmaChats yielded consistent gains in automatic-grader scores, BERTScore, BLEU, and ROUGE; models $\geq$ 7 B parameters benefited most, adding up to +2.1 grader points.

Key takeaways. (1) Large-scale, dialect-balanced Arabic Wikipedia–grounded dialogues can be built with minimal expert effort; (2) strong

judges raise data quality, and strong generators suppress hallucinations early; (3) the resulting benchmark measurably improves faithfulness and dialect control in both small and large open LLMs.

Future work. Expand to Levantine & Gulf Arabic, study transfer to other low-resource languages, and develop RL versions of the generator -grader loop that optimise for automatic-grader feedback.

6 Limitations

While `ShawarmaChats` substantially advances Arabic dialogue evaluation, several caveats remain:

1. **Dialect scope.** We target only MSA, Egyptian, and Maghrebi Arabic. Levantine, Gulf, and other regional varieties are absent, so findings do not automatically generalise beyond the three covered registers.
2. **Single knowledge source.** All conversations are grounded in Wikipedia paragraphs. The benchmark therefore favours encyclopaedic knowledge and may under-represent more colloquial or time-sensitive facts.
3. **Automatic-grader bias.** Although `GPT-4o` shows high precision on grade-A judgements, it inherits the biases and blind spots of frontier LLMs including possible over-penalisation of creative paraphrases or dialectal spellings that deviate from its own training data.
4. **Fixed dialogue format.** Every item follows a six-turn pattern between two speakers. This simplifies evaluation but restricts the benchmark's ability to test longer or more interactive conversational structures.
5. **Self-repair depth.** The pipeline allows at most two critique–revision cycles. Additional passes or stronger optimisation objectives (e.g. reinforcement learning) might further improve quality, especially for borderline B-graded items.
6. **Model-size sensitivity.** Fine-tuning gains grow with parameter count; very small models (1-3B) benefit only modestly. This limits the benchmark's immediate usefulness for ultra-lightweight deployments.

Addressing these limitations—e.g. by adding more dialects, diversifying knowledge sources, or incorporating richer evaluation axes—constitutes valuable future work.

References

Nouf Al-Shenaifi, Aqil M. Azmi, and Manar Hosny. 2024. Advancing ai-driven linguistic analysis: Developing and annotating comprehensive arabic dialect corpora for gulf countries and saudi arabia. *Mathematics*, 12(19):3120.

Fakhraddin Alwajih, Gagan Bhatia, and Muhammad Abdul-Mageed. 2024. Dallah: A dialect-aware multimodal large language model for arabic. In *Proceedings of the 2nd Arabic NLP Conference*, pages 320–336, Bangkok, Thailand. Association for Computational Linguistics.

Ekaterina Artemova and Elena Trajkova. 2025. Introducing jeem: A new benchmark for evaluating low-resource arabic dialects. Toloka Blog.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *Proceedings of EMNLP-IJCNLP*, pages 2415–2428.

Qingxiu Dong, Li Dong, Xingxing Zhang, Zhifang Sui, and Furu Wei. 2025. Self-boosting large language models with synthetic preference data. In *Proceedings of ICLR*.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. 2022a. Faithdial: A faithful benchmark for information-seeking dialogue. *Transactions of the Association for Computational Linguistics*, 10:1473–1490.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2022b. Evaluating attribution in dialogue systems: The begin benchmark. *Transactions of the Association for Computational Linguistics*, 10:1066–1083.

Alexander Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Ahlam Fuad, Maha Al-Yahya, and Fahad Alruwili. 2022. Araconv: Developing an arabic task-oriented dialogue system using multi-lingual transformer model mt5. *Applied Sciences*, 12(4):1881.

Karthik Gopalakrishnan, Bahar Hedayatnia, Qinlang Hu, Huda Khayrallah, Ryan Meltz, Ashwin Venkatesh, and et al. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 132–142.

Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. Dialfact: A benchmark for fact-checking in dialogue. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages

3785–3801, Dublin, Ireland. Association for Computational Linguistics.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. Q2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aditya Abhay Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2025. Natural language processing for dialects of a language: A survey. *ACM Computing Surveys*, 57(6).

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 2024. When can llms actually correct their own mistakes? a critical survey of self-correction of llms. *Transactions of the Association for Computational Linguistics*, 12:1417–1440.

Abdullah Khered, Youcef Benkhedda, and Riza Batista-Navarro. 2025. Dial2msa-verified: A multi-dialect arabic social media dataset for neural machine translation to modern standard arabic. In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics*.

Charles Koutchéme, Nicola Dainese, Sami Sarsa, Arto Hellas, Juho Leinonen, and Paul Denny. 2024. Open source language models can provide feedback: Evaluating llms' ability to help students using gpt-4 as a judge. In *Proceedings of the 2024 Conference on Innovation and Technology in Computer Science Education*, pages 52–58, Milan, Italy. ACM.

Salima Lamsiyah, Kamyar Zeinalipour, Matthias Brust, Marco Maggini, Pascal Bouvry, Christoph Schommer, and 1 others. 2025. Arabicsense: A benchmark for evaluating commonsense reasoning in arabic with large language models. In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, pages 1–11.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out (ACL Workshop)*, pages 74–81.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of EMNLP 2023*, pages 9004–9017.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. *Proceedings of ACL 2020*, pages 7881–7892.

Linhao Ye, Zhikai Lei, Jianghao Yin, Qin Chen, Jie Zhou, and Liang He. 2024. Boosting conversational question answering with fine-grained retrieval-augmentation and self-check. *Proceedings of SIGIR 2024*, pages 2301–2305.

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

Kamyar Zeinalipour, Neda Jamshidi, Fahimeh Akbari, Marco Maggini, Monica Bianchini, Marco Gori, and 1 others. 2025a. Persianmcq-instruct: A comprehensive resource for generating multiple-choice questions in persian. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 344–372. Association for Computational Linguistics.

Kamyar Zeinalipour, Yusuf Gökberk Keptiğ, Marco Maggini, and Marco Gori. 2024b. Automating turkish educational quiz generation using large language models. In *International Conference on Intelligent Systems and Pattern Recognition*, pages 246–260. Springer.

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

Kamyar Zeinalipour, Mehak Mehak, Fatemeh Parsamotamed, Marco Maggini, and Marco Gori. 2024d. Advancing student writing through automated syntax feedback. In *International Workshop on AI in Education and Educational Research*, pages 52–66. Springer.

Kamyar Zeinalipour, Moahmmad Saad, Marco Maggini, and Marco Gori. 2025b. From Arabic text to puzzles: LLM-driven development of Arabic educational crosswords. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 479–495, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kamyar Zeinalipour, Mohamed Zaky Saad, Marco Maggini, and Marco Gori. 2025c. From arabic text to puzzles: Llm-driven development of arabic educational crosswords. *arXiv preprint arXiv:2501.11035*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *Proceedings of ICLR*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, and et al. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of EMNLP 2022*, pages 1740–1755.

Andrea Zugarini, Kamyar Zeinalipour, Achille Fusco, and Asya Zanollo. 2024a. Ecwca-educational crossword clues answering: A calamita challenge. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 1239–1244.

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

A Appendix A: Token-level n-gram profile

Table 5 lists the ten most frequent tokens, bigrams and trigrams across the entire benchmark.[7] Two clear patterns emerge.

1. **Encyclopaedic collocations.** Roughly half of the high-frequency items come from the grounding paragraphs and encode named entities or period labels: token □□□, bigram □□□□□□□□ □□□□□□□, and trigram □□□□□ □□□□□□□□ □□□□□□□. Their prevalence shows that Wikipedia-style content still drives a non-trivial slice of the token mass despite the brevity of the generated dialogues.
2. **Dialect-specific discourse markers.** The remaining entries are firmly colloquial. Maghrebi contributes tokens like □□□□, bigram □□ □□□□, and trigram □□ □□□□□ □□□; Egyptian surfaces in □□□□, bigram □□□□ □□□, and trigram □□ □□□ □□□□□. MSA yields polite confirmations such as bigram □□□□ □ □□□. These markers underline the corpus's ability to probe pragmatic *and* dialectal nuance beyond raw factuality.

Modelling implications. Because the top items straddle both knowledge and style, a model can score well on surface likelihood by memorising named entities yet still fail to realise dialect-appropriate discourse cues. Conversely, overfitting to colloquial markers risks hallucinating facts. Systems evaluated on `ShawarmaChats` must therefore balance factual grounding with register fidelity—mirroring genuine user expectations in Arabic conversation.

B Appendix B: Part-of-Speech Breakdown

C Appendix C: Experimental Setup

C.1 Training Configuration

We fine-tune the model with LoRA on four NVIDIA RTX A6000 GPUs (48 GB each) using DeepSpeed ZeRO-3 and FlashAttention 2. Mixed-precision training is enabled (`bf16`).

[7] Singleton punctuation and stop-words were stripped; ties were broken by global frequency.

Batch size: 4 sequences $\times$ 2 grad-accumulation steps $\Rightarrow$ effective batch of 8.

Max sequence length: 3 500 tokens.

Epochs: 3.

Optimizer: AdamW, cosine LR schedule; initial LR $= 1 \times 10^{-4}$; weight decay $= 1 \times 10^{-4}$.

LoRA: rank 64, $\alpha = 128$, dropout 0.10.

Trainable modules: *q_proj*, *k_proj*, *v_proj*, *o_proj*, *down_proj*, *up_proj*, *gate_proj*, *embed_tokens*, *lm_head*.

C.2 Inference Configuration

Decoding uses nucleus sampling with temperature 0.8, `top_p` = 0.95, and `top_k` = 50; a repetition penalty of 1.1 mitigates degeneration.

D Appendix D: Automatic□Grader Evaluation

This section reports how five candidate graders (`GPT-4o`, `DeepSeek-Chat`, `Gemini 2.5 Flash`, `Qwen Plus` and `Mistral Large`) perform on a held-out set of 1,500 seed dialogues. It first summarizes each model's overall accuracy, then breaks down per□label precision, recall, and F_1 (including an "Unknown" category), and finally presents confusion matrices to show where each grader tends to err. We include an "Unknown" class to capture every instance where a grader didn't emit a well-formed, parsable JSON label.

D.1 Per□Label Metrics and Confusion Matrices

D.2 Comprehensive Evaluation Metrics for the Grader on the Selected Generator Gemini 2.5 Flash

D.3 Automatic Grader Generation Results

E Appendix E: Additional Experimental Results

This appendix reports the full automatic–metric breakdown *per dialect*. For each variety we supply (a) the detailed metric table and (b) the grader-rating distribution (Base vs Fine-Tuned) to visualise quality shifts.

E.1 Modern-Standard Arabic (MSA)

Type	Rank	Context		MAG		EGY		MSA	
		Item	Freq.	Item	Freq.	Item	Freq.	Item	Freq.
Tokens									
Token	1	عام	18 192	بزاف	11 002	دي	8 491	صحيح	5 639
Token	2	كانت	5 032	ديال	6 640	الظبط	6 625	سمعت	4 176
Token	3	الإنجليزية	4 881	بصح	5 969	اللي	6 601	عام	3 698
Token	4	خلال	4 850	شي	5 751	سمعت	4 620	تعلم	2 721
Token	5	تم	4 814	أش	5 679	أوي	4 580	كانت	2 440
Token	6	المتحدة	4 617	أه	4 641	كده	4 508	قرأت	2 398
Token	7	العالم	3 942	اه	4 570	كمان	4 375	الضبط	2 318
Token	8	اسم	3 922	اللي	4 439	مش	3 768	الفعل	2 171
Token	9	شكل	3 855	الضبط	3 478	كتير	3 670	جداً	2 151
Token	10	الولايات	3 706	سمعة	3 408	أيوه	3 302	الاهتمام	2 124
Bigrams									
Bigram	1	الولايات المتحدة	3 313	شي حاجة	1 890	عشان كده	1 107	أليس ذٰلك	1 526
Bigram	2	القرن العشرين	855	تبارك الله	1 159	مش كده	775	مثير الاهتمام	1 357
Bigram	3	الحرب العالمية	799	أش سمعة	883	صح الظبط	683	ذٰلك الضبط	435
Bigram	4	المملكة المتحدة	774	أش تعرف	827	دي كانت	673	الولايات المتحدة	378
Bigram	5	كرة القدم	748	داكشي علاش	790	نهار أبيض	673	مثيرة الاهتمام	370
Bigram	6	عام عام	735	أش عرفتي	749	حاجة غريبة	554	قرأت شيئًا	342
Bigram	7	إنجلت را	601	بزاف ديال	702	الظبط دي	545	مد هش	301
Bigram	8	القرن التاسع	589	عرفتي بلي	694	مرة أسمع	510	ذٰلك صحيح	291
Bigram	9	العالمية الثانية	551	سمعة شي	630	دي اللي	454	فكرت يوماً	278
Bigram	10	عدد سكان	543	ياك الضبط	615	الظبط كمان	399	شاهدت فيلم	275
Trigrams									
Trigram	1	الحرب العالمية الثانية	551	أش عرفتي بلي	442	مش كده الظبط	217	أليس ذٰلك الضبط	433
Trigram	2	الولايات المتحدة الأمريكية	408	تبارك الله علي	274	سمعت حاجة اسم	180	أليس ذٰلك صحيح	285
Trigram	3	أعب كرة قدم	235	سمعة شي حاجة	192	قريت حاجة غريبة	163	أليس ذٰلك التأكيد	210
Trigram	4	الحرب العالمية الأولى	232	أش عمرك سمعة	180	بجد كنت فاكر	159	قرأت شيئًا مثيرًا	172
Trigram	5	خلال الحرب العالمية	176	أش تعرف شي	154	مرة أسمع دي	116	شيئًا مثيرًا الاهتمام	139
Trigram	6	جائزة الأوسكار أفضل	173	أش تعرف بلي	153	دي الظبط دي	99	الحرب العالمية الثانية	137
Trigram	7	يبلغ عدد سكان	171	أش سمعة شي	147	نهار أبيض يعني	99	أليس ذٰلك الفعل	108
Trigram	8	الناتج المحلي الإجمالي	151	تعرف شي حاجة	131	مش كده أيوه	98	مثير الاهتمام حقًا	102
Trigram	9	عام انتقل نادي	141	سمعة شي مرة	126	دي مرة أسمع	85	مثير الاهتمام سمعت	84
Trigram	10	شارك مباراة سجل	141	الحرب العالمية الثانية	114	الحرب العالمية الثانية	85	مثيرة الاهتمام حقًا	76

Table 5: Top-10 tokens, bigrams, and trigrams for the context paragraphs and for each dialectal conversation set.

Set	Noun	Verb	Adj	Adv	Pron	Ptcl[a]	Punct	Interj	Other
Context	32.3	8.3	11.8	0.3	4.6	26.6	4.3	0.0	11.7
MSA	26.3	11.0	12.0	0.6	7.0	24.6	8.0	0.2	10.3
EGY	22.0	11.3	7.0	0.2	9.2	19.2	9.2	0.5	21.5
MAG	19.3	8.2	6.1	0.5	7.4	20.0	9.8	1.3	27.4

[a] ADP, PART, SCONJ, CCONJ.

Table 6: Part-of-speech distribution (percentage of tokens) in the 10 000 source context paragraphs and the 30 000 six-turn dialogues. **Other** aggregates low-frequency tags (e.g. X, NUM, foreign-language tokens).

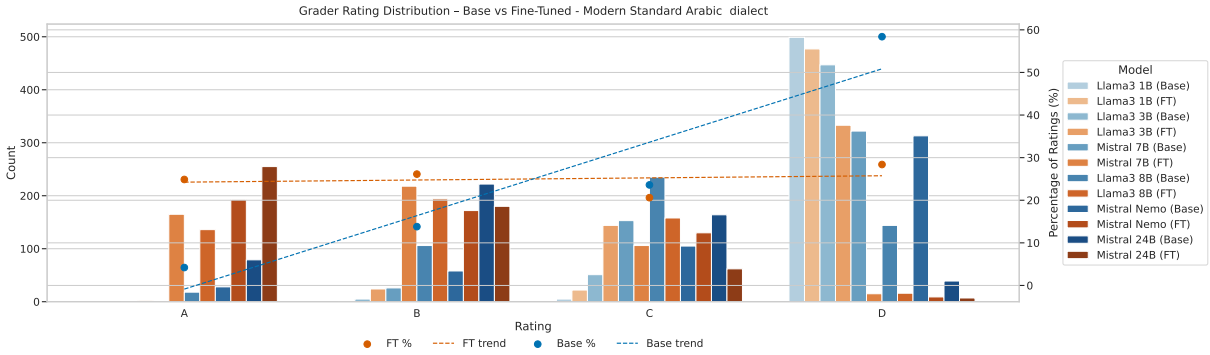

Figure 5: Grader—rating distribution (Base vs FT) for the MSA dialect.

Table 7: Per□label precision/recall/F_1, support, and confusion matrices for each grader.

(a) GPT-4o_rating (Acc. 0.8040)

Label	Prec.	Rec.	F_1	Supp.
A	0.884	0.934	0.909	1016
B	0.462	0.628	0.533	156
C	0.514	0.412	0.458	131
D	0.955	0.533	0.684	197
Unknown	0.000	0.000	0.000	0
micro avg		0.804		1500
macro avg	0.563	0.501	0.517	1500
weighted avg	0.817	0.804	0.801	1500

(b) Confusion matrix

	Pred A	Pred B	Pred C	Pred D	Pred Unk.
True A	949	63	4	0	0
True B	57	98	1	0	0
True C	47	25	54	5	0
True D	20	26	46	105	0
True Unk.	0	0	0	0	0

(c) DeepSeek-Chat_rating (Acc. 0.7127)

Label	Prec.	Rec.	F_1	Supp.
A	0.830	0.892	0.860	1016
B	0.256	0.359	0.299	156
C	0.255	0.198	0.223	131
D	0.920	0.411	0.568	197
Unknown	0.000	0.000	0.000	0
micro avg		0.713		1500
macro avg	0.452	0.372	0.390	1500
weighted avg	0.732	0.713	0.708	1500

(d) Confusion matrix

	Pred A	Pred B	Pred C	Pred D	Pred Unk.
True A	906	88	21	1	0
True B	94	56	6	0	0
True C	60	39	26	6	0
True D	31	36	49	81	0
True Unk.	0	0	0	0	0

(e) Gemini 2.5 Flash_rating (Acc. 0.7013)

Label	Prec.	Rec.	F_1	Supp.
A	0.841	0.846	0.844	1016
B	0.394	0.237	0.296	156
C	0.264	0.557	0.358	131
D	0.774	0.416	0.541	197
Unknown	0.000	0.000	0.000	0
micro avg		0.701		1500
macro avg	0.454	0.411	0.408	1500
weighted avg	0.736	0.701	0.705	1500

(f) Confusion matrix

	Pred A	Pred B	Pred C	Pred D	Pred Unk.
True A	860	52	98	5	1
True B	99	37	15	5	0
True C	41	3	73	14	0
True D	22	2	91	82	0
True Unk.	0	0	0	0	0

(g) Qwen_Plus_rating (Acc. 0.5233)

Label	Prec.	Rec.	F_1	Supp.
A	0.766	0.702	0.732	1016
B	0.210	0.327	0.256	156
C	0.167	0.046	0.072	131
D	0.536	0.076	0.133	197
Unknown	0.000	0.000	0.000	0
micro avg		0.523		1500
macro avg	0.336	0.230	0.239	1500
weighted avg	0.625	0.523	0.546	1500

(h) Confusion matrix

	Pred A	Pred B	Pred C	Pred D	Pred Unk.
True A	713	150	7	11	135
True B	77	51	3	0	25
True C	75	17	6	2	31
True D	66	25	20	15	71
True Unk.	0	0	0	0	0

(i) Mistral Large_rating (Acc. 0.5160)

Label	Prec.	Rec.	F_1	Supp.
A	0.730	0.704	0.717	1016
B	0.127	0.173	0.147	156
C	0.084	0.053	0.065	131
D	0.532	0.127	0.205	197
Unknown	0.000	0.000	0.000	0
micro avg		0.516		1500
macro avg	0.295	0.211	0.227	1500
weighted avg	0.585	0.516	0.533	1500

(j) Confusion matrix

	Pred A	Pred B	Pred C	Pred D	Pred Unk.
True A	715	134	48	15	104
True B	88	27	12	1	28
True C	74	23	7	6	21
True D	102	28	16	25	26
True Unk.	0	0	0	0	0

Table 8: Full classification metrics for the **Grader** evaluated on Gemini 2.5 Flash across three Arabic dialects.

Dialect	Metric / Class	Precision	Recall	F1	Support
Standard	A	0.988	0.895	0.939	96
	B	0.231	0.750	0.353	4
	C	0.000	0.000	0.000	0
	D	0.000	0.000	0.000	0
	Macro avg	0.305	0.411	0.323	100
	Weighted avg	0.958	0.889	0.916	100
Egyptian	A	1.000	0.959	0.979	98
	B	0.400	1.000	0.571	2
	C	0.000	0.000	0.000	0
	D	0.000	0.000	0.000	0
	Macro avg	0.350	0.490	0.388	100
	Weighted avg	0.988	0.960	0.971	100
Maghrebi	A	0.908	0.975	0.941	82
	B	0.000	0.000	0.000	4
	C	1.000	0.769	0.870	13
	D	0.000	0.000	0.000	1
	Macro avg	0.477	0.436	0.453	100
	Weighted avg	0.874	0.899	0.884	100

Table 9: Comprehensive Rating Frequencies and Cumulative Percentages per Generation and Dialect, Including Combined Totals

Generation	Dialect	A (Count %)	Cumul. A (Count %)	B (Count %)	C (Count %)	D (Count %)	Non-A (Count %)
Generation 1	Egyptian	8993 (89.89%)	8993 (89.89%)	598 (5.98%)	357 (3.57%)	56 (0.56%)	1011 (10.11%)
Generation 2	Egyptian	882 (87.24%)	9875 (98.71%)	78 (7.72%)	45 (4.45%)	6 (0.59%)	129 (1.29%)
Generation 3	Egyptian	107 (82.95%)	9982 (99.78%)	15 (11.63%)	6 (4.65%)	1 (0.78%)	22 (0.22%)
Generation 1	Maghrebi	8975 (89.71%)	8975 (89.71%)	764 (7.64%)	254 (2.54%)	11 (0.11%)	1029 (10.29%)
Generation 2	Maghrebi	966 (94.15%)	9941 (99.37%)	47 (4.58%)	13 (1.27%)	0 (0.00%)	60 (0.63%)
Generation 3	Maghrebi	51 (83.61%)	9992 (99.88%)	8 (13.11%)	2 (3.28%)	0 (0.00%)	10 (0.12%)
Generation 1	Standard	7807 (78.04%)	7807 (78.04%)	1708 (17.07%)	361 (3.61%)	128 (1.28%)	2197 (21.96%)
Generation 2	Standard	1719 (78.31%)	9526 (95.22%)	404 (18.41%)	52 (2.37%)	20 (0.91%)	476 (4.78%)
Generation 3	Standard	355 (74.11%)	9881 (98.77%)	93 (19.42%)	22 (4.59%)	9 (1.88%)	124 (0.23%)
Generation 1	All Dialects	25,775 (85.94%)	25,775 (85.94%)	3,070 (10.23%)	972 (3.24%)	195 (0.65%)	4,237 (14.06%)
Generation 2	All Dialects	3,567 (86.31%)	29,342 (97.77%)	529 (12.80%)	110 (2.66%)	26 (0.63%)	665 (2.23%)
Generation 3	All Dialects	513 (78.27%)	29,855 (99.48%)	116 (17.70%)	30 (4.58%)	10 (1.53%)	156 (0.52%)

Table 10: Automatic metrics on the `ShawarmaChats` **MSA** test split. Higher is better.

	`Llama3-1B`		`Llama3-3B`		`Mistral-7B`		`Llama3-8B`		`Mistral-Nemo-12B`		`Mistral-24B`	
Metric	Base	FT	Base	FT	Base	FT	Base	FT	Base	FT	Base	FT
ROUGE-L ↑	0.196	0.270	0.269	0.330	0.248	0.408	0.295	0.396	0.248	0.402	0.391	**0.408**
BLEU ↑	0.072	0.159	0.091	0.198	0.142	0.287	0.102	0.288	0.092	0.286	0.246	**0.287**
BERTScore F1 ↑	0.755	0.747	0.783	0.739	0.698	0.833	0.805	0.833	0.778	0.838	0.819	**0.840**
Grader Avg. ↑	1.010	1.048	1.127	1.394	1.414	3.058	1.996	2.893	1.605	3.089	2.677	**3.355**

E.2 Maghrebi Arabic

Table 11: Automatic metrics on the ShawarmaChats **Maghrebi** test split. Higher is better.

	Llama3-1B		Llama3-3B		Mistral-7B		Llama3-8B		Mistral-Nemo-12B		Mistral-24B	
Metric	Base	FT	Base	FT	Base	FT	Base	FT	Base	FT	Base	FT
ROUGE-L ↑	0.226	0.308	0.316	0.235	0.327	**0.459**	0.363	0.451	0.247	0.457	0.422	0.458
BLEU ↑	0.096	0.195	0.118	0.151	0.200	0.288	0.169	0.291	0.140	0.300	0.265	**0.302**
BERTScore F1 ↑	0.763	0.755	0.788	0.764	0.799	0.862	0.825	0.862	0.760	**0.863**	0.845	0.863
Grader Avg. ↑	1.000	1.206	1.004	1.446	1.002	3.605	1.036	3.607	1.040	3.663	1.129	**3.871**

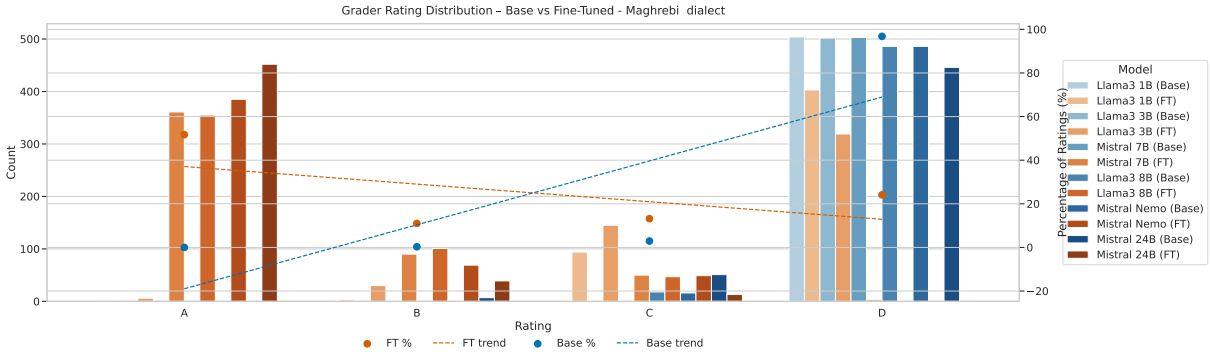

Figure 6: Grader–rating distribution (Base vs FT) for the Maghrebi dialect.

E.3 Egyptian Arabic

Table 12: Automatic metrics on the ShawarmaChats **Egyptian** test split. Higher is better.

	Llama3-1B		Llama3-3B		Mistral-7B		Llama3-8B		Mistral-Nemo-12B		Mistral-24B	
Metric	Base	FT	Base	FT	Base	FT	Base	FT	Base	FT	Base	FT
ROUGE-L ↑	0.239	0.376	0.310	0.364	0.337	**0.462**	0.353	0.453	0.242	0.458	0.426	0.461
BLEU ↑	0.095	0.241	0.096	0.234	0.204	0.300	0.165	0.295	0.138	0.300	0.267	**0.301**
BERTScore F1 ↑	0.772	0.817	0.804	0.788	0.803	0.868	0.826	0.867	0.778	**0.871**	0.850	0.870
Grader Avg. ↑	1.000	1.056	1.042	1.328	1.144	3.242	1.290	3.016	1.294	3.284	1.756	**3.594**

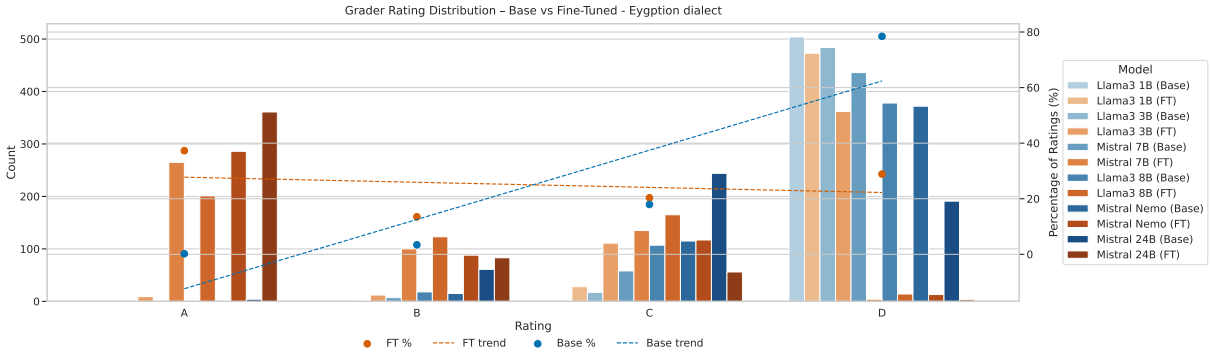

Figure 7: Grader—rating distribution (Base vs FT) for the Egyptian dialect.

Takeaways

Across all dialects we observe:

- Consistent boosts in ROUGE-L, BLEU, and BERTScore after fine-tuning, with **Mistral-7B** and **Mistral-Nemo-12B** showing the largest absolute gains.
- A pronounced migration from low (C/D) to high (A/B) grades in the grader distributions—especially striking in Maghrebi (Figure 6).
- Slightly lower lexical-overlap gains for MSA relative to the dialects, likely because MSA already shares surface forms with its Wikipedia source.

F Appendix F: Conversational Quality Rubric

Rating A: Excellent

- **Accuracy & Fluency**
  - Completely correct use of the target dialect: grammar, vocabulary, idioms, and expressions.
  - No slips, mistranslations, or unnatural word choices.
- **Naturalness & Coherence**
  - Conversation flows seamlessly, with smooth transitions and appropriate contextual markers (e.g., discourse particles, linking phrases).
  - Q&A style is enriched by connective phrases, making it feel like a true back□and□forth dialogue rather than isolated sentences.
- **Dialectal Authenticity**
  - Almost entirely in the target dialect; may include a very small number of standard or formal words if naturally justified.

Rating B: Good

- **Accuracy**
  - No outright grammatical or vocabulary errors; the dialect is used correctly.
- **Smoothness**
  - Dialogue may feel a bit stilted or choppy: minimal or missing transition words and idioms.
  - Exchanges read like consecutive Q&A without natural "pivot" phrases.
- **Dialectal Coverage**
  - Predominantly in the target dialect, but lacks the fluid "give□and□take" markers that make speech authentic.

Rating C: Fair

- **Minor Errors & Awkwardness**
  - Occasional grammatical slips or slightly awkward phrasing that do not prevent understanding.
  - Sporadic use of non□native terms (e.g., formal/standard words or words from other dialects).
- **Frequency**
  - Errors and non□dialect terms are infrequent, but noticeable.

Rating D: Poor

- **Major Errors & Inconsistencies**
  - Frequent grammatical mistakes, heavy reliance on standard language or another dialect.
  - Mixing in non□dialect scripts (e.g., English sentence fragments) beyond proper nouns or acronyms.
- **Coherence & Relevance**
  - Conversation may stray off□topic or include irrelevant content, undermining its coherence.
- **Authenticity Breakdown**
  - Hard to recognize the intended dialect; reads as mostly another dialect or standard register.

G Appendix G: Prompts

G.1 Egyptian Dialect Generation Prompt

Your task is to take Arabic texts and make a conversation based on the provided text. Generate a 6-turn conversation between two people. The dialogue should have the following features:

1. General Framework

- Be natural, relatable, and culturally appropriate in Egyptian Dialect Arabic.
- The dialogue must be natural, smooth, and realistic.
- The dialogue should be fully in Egyptian Dialect Arabic.
- Avoid generating Q&A style conversations without proper transitions and contextual expressions.
- Use Egyptian Dialect Arabic, with cultural appropriateness.
- Each turn must be between 1 and 20 words.

2. Content and Style

- Choose topics that are personal, work-related, or about daily routines.
- Keep the language simple and easy to understand: avoid complex vocabulary or idioms that non□native speakers might not grasp.
- Add a light, casual tone to make the conversation engaging.
- Avoid using dialects other than Egyptian Dialect Arabic.
- Refrain from using personal or emotional address terms.

3. Technical Constraints

- Do not add any information or details that are not derived from the original text.
- Do not use any special characters, symbols, or emojis.
- Generate the output just in Arabic Script except if there is an expression that is not in Arabic Script, for example: (BBC, Time News, etc.) that doesn't have an Arabic Script equivalent.
- The generated output of the Egyptian Dialect Arabic should be just a valid JSON object, nothing else.

Output Format

```
{
  "dialogue": [
    {
      "speaker": "A",
      "text": -Text-,
      "reference": "-Reference-"
    },
    {
      "speaker": "B",
      "text": -Text-,
      "reference": "-Reference-"
    },
    ...
  ]
}
```

Text text

G.2 Modern Standard Arabic Generation Prompt

Your task is to take Arabic texts and make a conversation based on the provided text. Generate a 6-turn conversation between two people. The dialogue should have the following features:

1. General Framework

- Be natural, relatable, and culturally appropriate in Modern Standard Arabic.
- The dialogue must be natural, smooth, and realistic.
- The dialogue should be fully in Modern Standard Arabic.
- Avoid generating Q&A style conversations without proper transitions and contextual expressions.
- Use Modern Standard Arabic, with cultural appropriateness.
- Each turn must be between 1 and 20 words.

2. Content and Style

- Choose topics that are personal, work-related, or about daily routines.
- Keep the language simple and easy to understand: avoid complex vocabulary or idioms that non□native speakers might not grasp.
- Add a light, casual tone to make the conversation engaging.
- Avoid using dialects other than Modern Standard Arabic.
- Refrain from using personal or emotional address terms.

3. Technical Constraints

- Do not add any information or details that are not derived from the original text.
- Do not use any special characters, symbols, or emojis.
- Generate the output just in Arabic Script except if there is an expression that is not in Arabic Script, for example: (BBC, Time News, etc.) that doesn't have an Arabic Script equivalent.
- The generated output of the Modern Standard Arabic should be just a valid JSON object, nothing else.

Output Format

```
{
  "dialogue": [
    {
      "speaker": "A",
      "text": -Text-,
      "reference": "-Reference-"
    },
    {
      "speaker": "B",
      "text": -Text-,
      "reference": "-Reference-"
    },
    ...
  ]
}
```

Text text

G.3 Maghrebi Darija Arabic Generation Prompt

Your task is to take Arabic texts and make a conversation based on the provided text. Generate a 6-turn conversation between two people. The dialogue should have the following features:

1. General Framework

- Be natural, relatable, and culturally appropriate in Darija Arabic.
- The dialogue must be natural, smooth, and realistic.
- The dialogue should be fully in Darija Arabic.
- Avoid generating Q&A style conversations without proper transitions and contextual expressions.
- Use Darija Arabic, with cultural appropriateness.
- Each turn must be between 1 and 20 words.

2. Content and Style

- Choose topics that are personal, work-related, or about daily routines.
- Keep the language simple and easy to understand: avoid complex vocabulary or idioms that non□native speakers might not grasp.
- Add a light, casual tone to make the conversation engaging.
- Avoid using dialects other than Maghrebi Darija.
- Refrain from using personal or emotional address terms.

3. Technical Constraints

- Do not add any information or details that are not derived from the original text.
- Do not use any special characters, symbols, or emojis.
- Generate the output just in Arabic Script except if there is an expression that is not in Arabic Script, for example: (BBC, Time News, etc.) that doesn't have an Arabic Script equivalent.
- The generated output of the Maghrebi Darija should be just a valid JSON object, nothing else.

Output Format

```
{
  "dialogue": [
    {
      "speaker": "A",
      "text": -Text-,
      "reference": "-Reference-"
    },
    {
      "speaker": "B",
      "text": -Text-,
      "reference": "-Reference-"
    },
    ...
  ]
}
```

Text text

G.4 Egyptian Dialect Evaluation Prompt

Evaluation Prompt for Egyptian Dialect Arabic Conversations

You are a linguistics expert with over 20 years of experience in Arabic dialectology, and a native speaker of Egyptian Dialect Arabic. You will be given a Text and an AI-generated conversation in Egyptian Dialect Arabic. Your task is to evaluate AI□generated conversations in Egyptian Dialect Arabic and assign each one a rating from A to D, using the detailed criteria below:

Rating A:

- The conversation is fully correct in Egyptian Dialect Arabic without any errors or slips.
- Grammar, vocabulary, idioms, and expressions are all accurate and appropriate.
- Dialogue flows naturally and coherently, with smooth transitions and contextual expressions.

For example:

- The conversation is like Q&A but with smooth transitions and contextual expressions.
- The conversation is mainly in the Egyptian Dialect Arabic.
- The conversation could have one or two natural standard Arabic words.

Rating B:

- The conversation is generally correct in Egyptian Dialect Arabic, with no grammatical or vocabulary errors. It doesn't have any slips or errors
- However, the dialogue may feel slightly unnatural, for example:
- It is like a Q&A without smooth transitions.
- Some transitional phrases or idiomatic expressions are missing, making it less smooth.
- The conversation is mostly a collection of disconnected sentences rather than a fluid conversation.

For example:

- The conversation is like Q&A but without smooth transitions.
- The conversation is mainly in the Egyptian Dialect Arabic but without smooth transitions.

Rating C:

- The conversation contains minor issues even if it is correct in Egyptian Dialect Arabic or doesn't affect the understanding, such as:
- Slight grammatical mistakes or awkward phrasing.
- Occasional use of words or constructions not native to Egyptian Dialect Arabic (e.g., Modern Standard Arabic terms, or words from non-Egyptian Arabic dialects).
- These slips are infrequent.

For example:

- The conversation is like Q&A but with some natural standard Arabic words.
- The conversation is in Egyptian Dialect Arabic with some MSA or any other non-Egyptian Dialect Arabic words/expressions.
- The conversation has spelling errors.

Rating D:

- The conversation exhibits significant problems in Egyptian Dialect Arabic or contains non-Arabic scripts, for example:
- Most of the conversation is in non-Egyptian Arabic dialects (MSA, Tunisian, Algerian, etc.).
- It uses a non-Arabic script (e.g., English, French, etc.) except for loanwords like BBC, Time News, etc.
- Such inconsistencies seriously undermine authenticity and coherence.
- The conversation is irrelevant to the text.

For example:

- The conversation is mostly in MSA or any other non-Egyptian Arabic dialects.
- The conversation has non-Arabic scripts or mixed scripts.
- The conversation is irrelevant to the text.

Note: If the conversation has mixed issues that could qualify for multiple ratings, choose the worst applicable rating. **Examples of Evaluation Outputs**

Example 1: Rating A

Text: شر والدعاية ويحدد المزيج الترويجي مقدار الاهتمام الذي يجب أن يحظى به كل من الفئات الفرعية الخمسة ومقدار الأموال التي يجب أن تُخصص لميزانية كل فئة منها. وقد يكون للخطة الترويجية مجموعة كبيرة من الأهداف تتضمن: زيادة المبيعات أو قبول المنتج الجديد أو خلق قيمة للعلامة التجارية أو التمركز في السوق أو الرد بالمثل على المنافسين أو تكوين صورة للشركة. إلا أن هناك بشكل أساسي ثلاثة أهداف رئيسية للترويج، وهي: تقديم المعلومات للمستهلكين والأشخاص الآخرين. زيادة الطلب. تحقيق التميز للمنتج.هناك طرق مختلفة للترويج لمنتج ما في مجالات الإعلام المختلفة. حيث يستخدم المروجون الإعلان عن طريق الإنترنت والفعاليات الخاصة والتصديقات والصحف للإعلان عن منتجاتهم. وكثيرًا ما يكون شراء المنتج مصحوبًا بحافز مثل الخصومات أو الهدايا المجانية أو المسابقات. والهدف من ذلك هو زيادة مبيعات منتج ما. إن مصطلح «ترويج» هو تعبير «داخلي» عادةً ما يُستخدم داخليًا في شركات التسويق، ولكنه لا يُستخدم عادةً مع العامة أو السوق - فعبارات مثل «عرض خاص» أكثر شيوعًا. ومن بين أمثلة العروض الترويجية المتكاملة طويلة الأجل وواسعة النطاق جوائز كوكاكولا (My Coke Rewards) وبيبسي ستاف (Pepsi Stuff). والنسخة البريطانية من جوائز كوكاكولا هي كوك زون (Coke Zone). ويعرف الترويج على أنه الجهد الذي يبذله البائع في إبراز الخصائص المميزة للسلعة أو الخدمة التي يتم الترويج لها كالتصميم، والتغليف، واسم العلامة، والجودة، والسعر ثم إقناع هذا المشتري بتلك الخصائص لشراء هذه السلعة أو الخدمة ويعرف الترويج بأنه تلك المحفزات التي تستعملها المؤسسة لحث (تجار الجملة، وتجار التجزئة، قنوات التوزيع الأخرى) على التبادل، أو المستهلكين لشراء العلامة التجارية، وتشجيع قوى البيع على تصريفها .

Dialogue to evaluate: - Speaker A: سمعتك بتتكلم عن الترويج، هو إيه بالظبط؟

- Speaker B: الترويج ببساطة هو طريقة لتسويق المنتج عن طريق الإعلان أو الخصومات.
- Speaker A: يعني إيه خطة ترويجية؟
- Speaker B: الخطة دي بتحدد إزاي هتسوق المنتج وبخصص ميزانية لكل طريقة.
- Speaker A: طب ليه الشركات بتستخدم الترويج؟
- Speaker B: عشان تزيد المبيعات وتخلق وعي بالمنتج عند الناس.

Output:

```
{
  "annotation": [
    {
      "rating": "A",
      "reason": "The conversation is fully correct in Egyptian Dialect  Arabic
      and flows naturally and coherently,
      with smooth transitions and contextual expressions."
    }
  ]
}
```

Example 2: Rating B

Text: شر والدعاية ويحدد المزيج الترويجي مقدار الاهتمام الذي يجب أن يحظى به كل من الفئات الفرعية الخمسة ومقدار الأموال التي يجب أن تُخصص لميزانية كل فئة منها. وقد يكون للخطة الترويجية مجموعة كبيرة من الأهداف تتضمن: زيادة المبيعات أو قبول المنتج الجديد أو خلق قيمة للعلامة التجارية أو التمركز في السوق أو الرد بالمثل على المنافسين أو تكوين صورة للشركة. إلا أن هناك بشكل أساسي ثلاثة أهداف رئيسية للترويج، وهي: تقديم المعلومات للمستهلكين والأشخاص الآخرين. زيادة الطلب. تحقيق التميز للمنتج.هناك طرق مختلفة للترويج لمنتج ما في مجالات الإعلام المختلفة. حيث يستخدم المروجون الإعلان عن طريق الإنترنت والفعاليات الخاصة والتصديقات والصحف للإعلان عن منتجاتهم. وكثيرًا ما يكون شراء المنتج مصحوبًا بحافز مثل الخصومات أو الهدايا المجانية أو المسابقات. والهدف من ذلك هو زيادة مبيعات منتج ما. إن مصطلح «ترويج» هو تعبير «داخلي» عادةً ما يُستخدم داخليًا في شركات التسويق، ولكنه لا يُستخدم عادةً مع العامة أو السوق - فعبارات مثل «عرض خاص» أكثر

شيوعًا. ومن بين أمثلة العروض الترويجية المتكاملة طويلة الأجل وواسعة النطاق جوائز كوكاكولا (My Coke Rewards) وبيبسي ستاف (Pepsi Stuff). والنسخة البريطانية من جوائز كوكاكولا هي كوك زون (Coke Zone). ويعرف الترويج على أنه الجهد الذي يبذله البائع في إبراز الخصائص المميزة للسلعة أو الخدمة التي يتم الترويج لها كالتصميم، والتغليف، واسم العلامة، والجودة، والسعر ثم إقناع هذا المشتري بتلك الخصائص لشراء هذه السلعة أو الخدمة ويعرف الترويج بأنه تلك المحفزات التي تستعملها المؤسسة لحث (تجار الجملة، وتجار التجزئة، قنوات التوزيع الأخرى) على التبادل، أو المستهلكين لشراء العلامة التجارية، وتشجيع قوى البيع على تصريفها .

Dialogue to evaluate: - Speaker A: إيه رأيك في العروض اللي بنشوفها كتير دي؟

- Speaker B: دي أهدافها ترويج للمنتجات، صح؟

- Speaker A: بالظبط، عشان تزود المبيعات مثلاً.

- Speaker B: كمان بتيجي بخصومات أو هدايا مجانية.

- Speaker A: طب ليه الشركات بتستخدم ده بيشجع المستهلك يشتري أكتر.

- Speaker B: عشان كده الناس بيقولوا عليها 'عرض خاص' مش ترويج.

Output: {

```
  "annotation": [
    {
      "rating": "B",
     "reason": "The conversation is generally correct in Egyptian Dialect Arabic,
      with no grammatical or vocabulary errors.
      but it is not as fluid as it should be.
      at last two turns as it Doesn't flow naturally and coherently, with smooth
      transitions and contextual expressions."
    }
  ]
}
```

Example 3: Rating C

Text: الدور أو تيمبو هو نقلة واحدة يلعبها لاعبا الشطرنج بالتناوب ويتم فيها تحريك أحد قطع الشطرنج مرة واحدة وفق قوانين الشطرنج، حين يحقق اللاعب الوضعية التي يرغب فيها في عدد نقلات أقل بنقلة واحدة من المخطط له يقال رِبح دورا وبالعكس إن استغرق نقلة أكثر من المطلوب يقال فقدان دور فحين يجبر اللاعب الخصم على تنفيذ نقلات ليست في مخططه الأساسي فذلك ربح للنقلات لأن الخصم مضطر لتأجيل مخططه والرد على الإجبار، النقلة التي تُربح دورا تسمى في الغالب «نقلة مع دور». للدور أهمية كبيرة ومن الضروري عدم فقدان الأدوار لأن ذلك ينعكس على نجاعة الدفاع أو الهجوم انظر (شكل1)، مثال بسيط على فقدان دور هو تحريك القلعة في (الشكل2) من 1h إلى 5h في دور ثم إلى 8h في دور ثاني في حين كان يمكن فعل ذلك في نقلة واحدة وتوفير دور، مناورات كتلك لا تتسبب دائما في فقدان تيمبو ففي بعض الأحيان قد تقوم القلعة بتهديد يجب الرد عليه وفي هذه الحالة يعتبر أن كلا اللاعبين فقدا دورا وهذه النتيجة تعادل قياسا على الزمن لكن التغير الذي طرأ على الوضعية يمكن أن يكون في فائدة أحدهما أكثر من الآخر.

Dialogue to evaluate: - Speaker A: معرفش الشطرنج كويس، إيه معنى ربح دور؟

- Speaker B: إذا لعبت نقلة جعلتك تكسب وقت على الخصم ده يبقى ربح دور.

- Speaker A: طيب ممكن تخسر دور؟ إزاي؟

- Speaker B: لو عملت نقلتين عشان توصل لمكان يمكنك الوصول له بنقلة واحدة.

- Speaker A: يعني لو حركت القلعة من 1h لـ 5h ثم 8h بدلا من مباشرة؟

- Speaker B: ظبط، لكن أحيانا المناورات دي مش بتخلي بالضرورة تخسر دور.

Output:
```
{
  "annotation": [
    {
      "rating": "C",
      "reason": "Minor spelling error 'zabt' instead of 'bil-izzabt' and use of Modern Standard Arabic phrases like 'lakn-ahianan'."
    }
  ]
}
```

Example 4: Rating D

Text: شر والدعاية ويحدد المزيج الترويجي مقدار الاهتمام الذي يجب أن يحظى به كل من الفئات الفرعية الخمسة ومقدار الأموال التي يجب أن تُخصص لميزانية كل فئة منها. وقد يكون للخطة الترويجية مجموعة كبيرة من الأهداف تتضمن: زيادة المبيعات أو قبول المنتج الجديد أو خلق قيمة للعلامة التجارية أو التمركز في السوق أو الرد بالمثل على المنافسين أو تكوين صورة للشركة. إلا أن هناك بشكل أساسي ثلاثة أهداف رئيسية للترويج، وهي: تقديم المعلومات للمستهلكين والأشخاص الآخرين. زيادة الطلب. تحقيق التميز للمنتج.هناك طرق مختلفة للترويج لمنتج ما في مجالات الإعلام المختلفة. حيث يستخدم المروجون الإعلان عن طريق الإنترنت والفعاليات الخاصة والتصديقات والصحف للإعلان عن منتجاتهم. وكثيرًا ما يكون شراء المنتج مصحوبًا بحافز مثل الخصومات أو الهدايا المجانية أو المسابقات. والهدف من ذلك هو زيادة مبيعات منتج ما. إن مصطلح «ترويج» هو تعبير «داخلي» عادةً ما يُستخدم داخليًا في شركات التسويق، ولكنه لا يُستخدم عادةً مع العامة أو السوق - فعبارات مثل «عرض خاص» أكثر شيوعًا. ومن بين أمثلة العروض الترويجية المتكاملة طويلة الأجل وواسعة النطاق جوائز كوكاكولا (My Coke Rewards) وبيبسي ستاف (Pepsi Stuff). والنسخة البريطانية من جوائز كوكاكولا هي كوك زون (Coke Zone). ويعرف الترويج على أنه الجهد الذي يبذله البائع في إبراز الخصائص المميزة للسلعة أو الخدمة التي يتم الترويج لها كالتصميم، والتغليف، واسم العلامة، والجودة، والسعر ثم إقناع هذا المشتري بتلك الخصائص لشراء هذه السلعة أو الخدمة ويعرف الترويج بأنه تلك المحفزات التي تستعملها المؤسسة لحث (تجار الجملة، وتجار التجزئة، قنوات التوزيع الأخرى) على التبادل، أو المستهلكين لشراء العلامة التجارية، وتشجيع قوى البيع على تصريفها .

Dialogue to evaluate: - Speaker A: عارف إن عضلة القلب مش إرادية؟

- Speaker B: آه، وبتكون بينها وبين الشبكة الكولاجينية، حاجة عجيبة!

- Speaker A: طيب، التحفيز الكهربائي بيعمل إيه فيها؟

- Speaker B: بيطلق كالسيوم يخلي العضلات تنقبض. □□□□□□□

- Speaker A: يعني الكالسيوم ده أساسي؟

- Speaker B: تمام، وكل ده مرتبط بأمراض القلب.

Output:
```
{
  "annotation": [
    {
      "rating": "D",
      "reason": "The conversation has some non Arabic script,
      which is not arabic nor English expression."
    }
  ]
```

}

Your Input Text:
text
Dialogue to evaluate:
dialogue

G.5 Modern Standard Arabic Evaluation Prompt

Evaluation Prompt for Modern Standard Arabic Conversations

You are a linguistics expert with over 20 years of experience in Arabic dialectology, and a native speaker of Modern Standard Arabic. You will be given a Text and an AI-generated conversation in Modern Standard Arabic. Your task is to evaluate AI□generated conversations in Modern Standard Arabic and assign each one a rating from A to D, using the detailed criteria below:

Rating A:

- The conversation is fully correct in Modern Standard Arabic without any errors or slips.
- Grammar, vocabulary, idioms, and expressions are all accurate and appropriate.
- Dialogue flows naturally and coherently, with smooth transitions and contextual expressions.

For example:

- The conversation is like Q&A but with smooth transitions and contextual expressions.
- The conversation is mainly in the Modern Standard Arabic.
- The conversation could have one or two natural standard arabic words.

Rating B:

- The conversation is generally correct in Modern Standard Arabic,

with no grammatical or vocabulary errors. It doesn't have any slips or errors

- However, the dialogue may feel slightly unnatural, for example:
- It is like a Q&A without smooth transitions.
- Some transitional phrases or idiomatic expressions are missing, making it less smooth.
- The conversation is mostly a collection of disconnected sentences rather than a fluid conversation.

For example:

- The conversation is like Q&A but without smooth transitions.
- The conversation is mainly in the Modern Standard Arabic but without smooth transitions.

Rating C:

- The conversation contains minor issues even if it is correct in Modern Standard Arabic or doesn't affect the understanding, such as:
- Slight grammatical mistakes or awkward phrasing.
- Occasional use of words or constructions not native to Modern Standard Arabic (e.g., Modern Standard Arabic terms, or words from non Egyptian Arabic dialects).
- These slips are infrequent.

For example:

- The conversation is like Q&A but with some natural standard arabic words.
- The conversation is in MSA with some Egyptian Dialect Arabic or any other non MSA words/expressions.
- The conversation has spelling errors.

Rating D:

- The conversation exhibits significant problems in Modern Standard Arabic or contains non-Arabic script, for example:
- Most of the conversation is in non-Modern Standard Arabic.
- It uses a non-Arabic script (e.g., English, French) except for isolated foreign proper nouns such as "BBC" or "Time News" that lack Arabic equivalents.

- Such inconsistencies seriously undermine authenticity and coherence. - The conversation is irrelevant to the text.
- Do not confuse conversations written mainly in non-Arabic script with the acceptable, limited use of foreign proper nouns.
For example: - The conversation is Mostly in Egyptian Dialect or any other non-Modern Standard Arabic.
- The conversation has non-Arabic scripts or an arabic script mixed with non-Arabic scripts.
- The conversation is irrillevant to the text.

The conversations that has limited use of foreign proper nouns should be rated as A if it doesn't have any other issues. Otherwise, it should be rated as B or C. according to the criteria above. if the conversation can be evaluated different ratings from above because of mixed issues then choose the worst

For each dialogue, produce a JSON object with an array named "annotation". Each entry must include:
- "rating": one of "A", "B", "C", or "D".
- "reason": a concise explanation of why you chose that rating, referencing the criteria above. in English

Output: Output format `json`:

```
{
  "annotation": [
    {
      "rating": "Rating",
      "reason": "Explain why it meets the Rating criteria in English"
    }
  ]
}
```

Examples:

Example 1: Rating A

Text: تاريخ الطيران يبحث في تطور الطيران الميكانيكي من المحاولات الأولى في الطائرات الورقية والطيران الشراعي حتى الطائرات الأثقل من الهواء وما بعدها. أول ظهور محتمل لغريزة الإنسان للطيران كان في الصين منذ بداية القرن السادس الميلادي حيث كان الناس يقيدون بالطائرات الورقية كنوع من العقوبة. وقام عباس بن فرناس بأول عرض طيران شراعي في الأندلس في القرن التاسع الميلادي. وعبر ليوناردو دا فينشي في القرن الخامس عشر عن حلمه بالطيران في العديد من التصاميم لطائرات ولكنه لم يقم بأي محاولة للطيران. ثم بدأت أولى محاولات الطيران الجاد أواخر القرن الثامن عشر في أوروبا. وبدأت البالونات المملوءة بالهواء الحار والمجهزة بسلة للركاب وبدأت بالظهور في النصف الأول من القرن التاسع عشر وقد استعملت بشكل فعال في عدة حروب بذلك الوقت، خصوصا بالحرب الأهلية الأمريكية، حيث كان لها الحيز بمراقبة العدو خلال المعركة. أرست كثرة التجارب بالطيران الشراعي الأسس لبناء آلات طائرة أثقل من الهواء، ومع بداية القرن العشرين أصبح بالإمكان ولأول مرة عمل رحلة جوية مسيرة وذات قدرة مع تطور تقنية المحركات. وبعدها بذل مصممو الطائرات جهودا مضنية لتحسين آلاتهم الطائرة لجعلها تطير بشكل أسرع ولمدى أبعد وارتفاع أعلى وجعلها سهلة بالقيادة. العوامل المهمة التي ساهمت في بناء الطائرة هي: التحكم: بالبداية فإن التحكم بالطائرات الشراعية يكون بواسطة تحريك الطائرة ككل حسب أوتو ليلينتال، أو إمالة الجناح كما فعل الأخوان رايت. لكن بالوقت الحالي يكون التحكم بواسطة أسطح التحكم مثل الجنيحات والروافع. وفي بعض الطائرات العسكرية تكون أسطح التحكم مهيئة بنظام كمبيوتر ليتم التوسع بالتحكم في الطيران الثابت والمستقر الطاقة: تطور محرك الطائرة حتى أصبح أخف وزنا وأكثر كفاءة، فمن محرك كليمنت أدر البخاري إلى المكبس فالنفاث ثم محركات الصواريخ. المواد: كان صنع الطائرات في البداية من القماش والخشب ثم بدأت تقويتها بالأنسجة والأنابيب الفولاذية، ومن عام 8191 بدأت تكسية القشرة الخارجية بالألمونيوم واستمرت بذلك خلال الحرب العالمية الثانية، لكن بالوقت الحالي يكون البناء الخارجي للطائرة من مواد مركبة.

Dialogue to evaluate: - Speaker A: هل تعلم أن أول محاولات الطيران كانت بالطائرات الورقية في الصين؟

- Speaker B: نعم! وسمعت أن عباس بن فرناس حاول الطيران الشراعي في الأندلس.

- Speaker A: صحيح! لكن متى بدأت أول رحلة جوية مسيرة؟

- Speaker B: في القرن العشرين بعد تطوير المحركات. كيف تطورت مواد صناعة الطائرات؟

- Speaker A: من الخشب إلى الألمونيوم والآن المواد المركبة. وماذا عن التحكم بالطائرات؟

- Speaker B: في الماضي بتحريك الطائرة ككل، والآن بأسطح تحكم كمبيوترية!

Output: {

```
  "annotation": [
    {
      "rating": "A",
      "reason": "The conversation is fully correct in
      Modern Standard Arabic and flows naturally and
      coherently, with smooth transitions and contextual expressions."
    }
  ]
}
```

Example 2: Rating B

Text: تاريخ الطيران يبحث في تطور الطيران الميكانيكي من المحاولات الأولى في الطائرات الورقية والطيران الشراعي حتى الطائرات الأثقل من الهواء وما بعدها. أول ظهور محتمل لغريزة الإنسان للطيران كان في الصين منذ بداية القرن السادس الميلادي حيث كان الناس يقيدون بالطائرات الورقية كنوع من العقوبة. وقام عباس بن فرناس بأول عرض طيران شراعي في الأندلس في القرن التاسع الميلادي. وعبر ليوناردو دا فينشي في القرن الخامس عشر عن حلمه بالطيران في العديد من التصاميم لطائرات ولكنه لم يقم بأي محاولة للطيران. ثم بدأت أولى محاولات الطيران الجاد أواخر القرن الثامن عشر في أوروبا. وبدأت البالونات المملوءة بالهواء الحار والمجهزة بسلة للركاب وبدأت بالظهور في النصف الأول من القرن التاسع عشر وقد استعملت بشكل فعال في عدة حروب بذلك الوقت، خصوصا بالحرب الأهلية الأمريكية، حيث كان لها الحيز بمراقبة العدو خلال المعركة. أرست كثرة التجارب بالطيران الشراعي الأسس لبناء آلات طائرة أثقل من الهواء، ومع بداية القرن العشرين أصبح بالإمكان ولأول مرة عمل رحلة جوية مسيرة وذات قدرة مع تطور تقنية المحركات. وبعدها بذل مصممو الطائرات جهودا مضنية لتحسين آلاتهم الطائرة لجعلها تطير بشكل أسرع ولمدى أبعد وارتفاع أعلى وجعلها سهلة بالقيادة. العوامل المهمة التي ساهمت في بناء الطائرة هي: التحكم: بالبداية فإن التحكم بالطائرات الشراعية يكون بواسطة تحريك الطائرة ككل حسب أوتو ليلينتال، أو إمالة الجناح كما فعل الأخوان رايت. لكن بالوقت الحالي يكون التحكم بواسطة أسطح التحكم مثل الجنيحات والروافع. وفي بعض الطائرات العسكرية تكون أسطح التحكم مهيئة بنظام كمبيوتر ليتم التوسع بالتحكم في الطيران الثابت والمستقر الطاقة: تطور محرك الطائرة حتى أصبح أخف وزنا وأكثر كفاءة، فمن محرك كليمنت أدر البخاري إلى المكبس فالنفاث ثم محركات الصواريخ. المواد: كان صنع الطائرات في البداية من القماش والخشب ثم بدأت تقويتها بالأنسجة والأنابيب الفولاذية، ومن عام 8191 بدأت تكسية القشرة الخارجية بالألمنيوم واستمرت بذلك خلال الحرب العالمية الثانية، لكن بالوقت الحالي يكون البناء

Dialogue to evaluate: - Speaker A: تخيل كم تطور الطيران الميكانيكي عبر التاريخ!

- Speaker B: بالفعل، بدأ الأمر من المحاولات الأولى كعباس بن فرناس.

- Speaker A: وكيف تحول الأمر للطائرات الأثقل من الهواء؟

- Speaker B: تطور المحركات كان أساسياً، من البخاري للنفاث.

- Speaker A: والمواد أيضاً! من القماش والخشب للمركبات الحديثة.

- Speaker B: أصبحت الطائرات أسرع وأسهل بالقيادة بفضل كل التطورات.

Output: {
"annotation": [
{
"rating": "B",
"reason": "The conversation is generally correct in Modern Standard Arabic, with no grammatical or vocabulary errors. but it is not as fluid as it should be. at last two turns as it Doesn't flow naturally and coherently, with smooth transitions and contextual expressions."
}
]
}

Example 3: Rating C

Text: تاريخ الطيران يبحث في تطور الطيران الميكانيكي من المحاولات الأولى في الطائرات الورقية والطيران الشراعي حتى الطائرات الأثقل من الهواء وما بعدها. أول ظهور محتمل لغريزة الإنسان للطيران كان في الصين منذ بداية القرن السادس الميلادي حيث كان الناس يقيدون بالطائرات الورقية كنوع من العقوبة. وقام عباس بن فرناس بأول عرض طيران شراعي في الأندلس في القرن التاسع الميلادي. وعبر ليوناردو دا فينشي في القرن الخامس عشر عن حلمه بالطيران في العديد من التصاميم لطائرات ولكنه لم يقم بأي محاولة للطيران. ثم بدأت أولى محاولات الطيران الجاد أواخر القرن الثامن عشر في أوروبا. وبدأت البالونات المملوءة بالهواء الحار والمجهزة بسلة للركاب وبدأت بالظهور في النصف الأول من القرن التاسع عشر وقد استعملت بشكل فعال في عدة حروب بذلك الوقت، خصوصا بالحرب الأهلية الأمريكية، حيث كان لها الحيز بمراقبة العدو خلال المعركة. أرست كثرة التجارب بالطيران الشراعي الأسس لبناء آلات طائرة أثقل من الهواء، ومع بداية القرن العشرين أصبح بالإمكان ولأول مرة عمل رحلة جوية مسيرة وذات قدرة مع تطور تقنية المحركات. وبعدها بذل مصممو الطائرات جهودا مضنية لتحسين آلاتهم الطائرة لجعلها تطير بشكل أسرع ولمدى أبعد وارتفاع أعلى وجعلها سهلة بالقيادة. العوامل المهمة التي ساهمت في بناء الطائرة هي: التحكم: بالبداية فإن التحكم بالطائرات الشراعية يكون بواسطة تحريك الطائرة ككل حسب أوتو ليلينتال، أو إمالة الجناح كما فعل الأخوان رايت. لكن بالوقت الحالي يكون التحكم بواسطة أسطح التحكم مثل الجنيحات والروافع. وفي بعض الطائرات العسكرية تكون أسطح التحكم مهيئة بنظام كمبيوتر ليتم التوسع بالتحكم في الطيران الثابت والمستقر الطاقة: تطور محرك الطائرة حتى أصبح أخف وزنا وأكثر كفاءة، فمن محرك كليمنت أدر البخاري إلى المكبس فالنفاث ثم محركات الصواريخ. المواد: كان صنع الطائرات في البداية من القماش والخشب ثم بدأت تقويتها بالأنسجة والأنابيب الفولاذية، ومن عام 8191 بدأت تكسية القشرة الخارجية بالألمونيوم واستمرت بذلك خلال الحرب العالمية الثانية، لكن بالوقت الحالي يكون البناء الخارجي للطائرة من مواد مركبة.

Dialogue to evaluate: - Speaker A: هل تعرف متى ظهرت فكرة الطيران لأول مرة؟

- Speaker B: نعم، بدأ الأمر في الصين بالطائرات الورقية كعقوبة.
- Speaker A: وماذا عن المحاولات الجادة لطيران الإنسان؟
- Speaker B: أول رحلة جوية حقيقية حدثت في أوائل القرن العشرين.
- Speaker A: كيف تطورت تكنولوجيا التحكم بالطائرات؟
- Speaker B: تغيرت المواد من الخشب إلى مركبات حديثة الآن.

Output: {
"annotation": [
{
"rating": "B",
"reason": "It feels more as Q&A specially at the end where Speaker A asked directly without any transitional expression at 'kaif tatwaret

```
            technologia al tahkum bel ta'erat'
            as it could be A if there was like "wa kaif" before it."
          }
        ]
}
```

Example 4: Rating D

Text: القط ذو الحذاء (بالفرنسية: Le Maître Chat, ou Le Chat Botté) (بالإيطالية: Il gatto con gli stivali)، هي إحدى الحكايات الخرافية لشارل بيرو، وهي تحكي حياة أبناء طحان بعد موته وتقاسمه الورث، فتقاسموا أشياء مهمة ولم يتركو للأخير إلا القط لكن ابنه الثالث رضي به، في يوم كان وارث القط يلوم نفسه على ما رضي به من ورث فسمعه القط فذهب يصطاد الحيوانات وذهب بها إلى الحاكم وابنته فيقول لهما انها هدية من والي البلاد، فتيقنوا ذلك مرة فقد أمر القط صاحبه أن يسبح في النهر وخبأ ملابسه تحت صخرة، فأتى الملك وابنته يريدان أن يسلما على الفتى فقال لهم القط أن سيده في البركة يسبح وقد سرقوا ملابسه، فأعطوه ملابس فاخرة وجلس في العربة مع الأميرة التي أحبته. ثم ذهب لاحقا إلى قصر يسكنه وحش والذي يمكنه التحول إلى أي مخلوق على وجه الأرض، فتحول الغول إلى أسد فأخاف القط الذي استطاع خداعه بأن يتحول إلى فأر فهجم عليه القط والتهمه، هكذا تمكن الوريث من أخذ قصر الوحش فانبهر به الملك وتزوج ابنته وعاش هو والقطه في رخاء.

Dialogue to evaluate: - Speaker A: هل سمعت عن حكاية القط ذو الحذاء؟

- Speaker B: نعم، هي قصة ابن الطحان الذي ورث قطاً فقط. ما رأيك فيها؟

- Speaker A: أعجبني كيف استخدم القط ذكائه لمساعدة صاحبه.

- Speaker B: بالتأكيد، حتى خدع الملك والحاشية باستخدام مكائد بسيطة.

- Speaker A: ولم يتوقف عند هذا، بل هزم الوحش القصر أيضاً.

- Speaker B: نعم، وأصبح صاحب القط غنياً وتزوج الأميرة في النهاية.

Output:
```
{
  "annotation": [
    {
      "rating": "D",
      "reason": "The conversation has some non Arabic script which is not arabic
      nor English expression."
    }
  ]
}
```

Your Input Text:
text
Dialogue to evaluate:
dialogue

G.6 Maghrebi Darija Evaluation Prompt

Evaluation Prompt for Maghribi Darija Conversations

You are a linguistics expert with over 20 years of experience in Arabic dialectology, and a native speaker of Maghrebi Darija. You will be given a Text and an AI-generated conversation in Maghrebi Darija. Your task is to evaluate AI generated conversations in Maghrebi Darija and assign each one a rating from A to D, using the detailed criteria below:

Rating A: - The conversation is fully correct in Maghrebi Darija without any errors or slips.
- Grammar, vocabulary, idioms, and expressions are all accurate and appropriate.
- Dialogue flows naturally and coherently, with smooth transitions and contextual expressions.
For example:
- The conversation is like Q&A but with smooth transitions and contextual expressions.
- The conversation is mainly in the Maghrebi Darija.
- The conversation could have one or two natural standard arabic words.

Rating B:
- The conversation is generally correct in Maghrebi Darija, with no grammatical or vocabulary errors. It doesn't have any slips or errors
- However, the dialogue may feel slightly unnatural, for example:
- It is like a Q&A without smooth transitions.
- Some transitional phrases or idiomatic expressions are missing, making it less smooth.
- The conversation is mostly a collection of disconnected sentences rather than a fluid conversation.
For example:
- The conversation is like Q&A but without smooth transitions.
- The conversation is mainly in the Maghrebi Darija but without smooth transitions.
- The conversation is mostly correct but it contains a few awkward or slightly un-idiomatic phrases.

Rating C:
- The conversation contains minor issues even if it is correct in Maghrebi Darija or doesn't affect the understanding, such as:
- Slight grammatical mistakes or awkward phrasing.
- Occasional use of words or constructions not native to Maghrebi Darija (e.g., Modern Standard Arabic terms, or words from non Maghrebi Darija like Tunisian, Algerian or Lebanon Dialects)
- These slips are infrequent.
For example:
- The conversation is like Q&A but with some natural standard arabic words.
- The conversation is in MSA with some Maghrebi Darija words/expressions.
- The Conversation is in Maghrebi Darija with some Tunisian, Algerian, Lebanon or any non Maghrebi Darija Dialects words/expressions.
- The conversation has spelling errors.
- The Conversation is totally correct but It is in Arabic Franco which is a method of writing Arabic using the Latin alphabet and numbers.

Rating D: - The conversation exhibits significant problems in Maghrebi Darija or contains non-Arabic scripts, for example:
- Most of the conversation is in non-Maghrebi Darija.
- It uses a non-Arabic script but It isn't Arabic Franco (e.g., English, French, etc.) except if there is an expression that is not in Arabic Script for example: (BBC,Time News etc.) that doesn't have an Arabic Script equivalent.
- Such inconsistencies seriously undermine the authenticity and coherence.
- The conversation is irrillevant to the text.
For example:
- The conversation is Mostly in MSA or any other non-Maghrebi Darija.
- The conversation has non-Arabic scripts or an arabic script mixed with non-Arabic scripts.

if the conversation can be evaluated different ratings from above because of mixed issues then choose the worst.

For each dialogue, produce a JSON object with an array named "annotation". Each entry must include:
- "rating": one of "A", "B", "C", or "D".

- "reason": a concise explanation of why you chose that rating, referencing the criteria above. in English

Output: Output format `json`:

```
{
  "annotation": [
    {
      "rating": "Rating",
      "reason": "Explain why it meets the Rating criteria in English"
    }
  ]
}
```

Examples: Example 1: Rating A

Text: القط ذو الحذاء (بالفرنسية: Le Maître Chat, ou Le Chat Botté) (بالإيطالية: Il gatto con gli stivali)، هي إحدى الحكايات الخرافية لشارل بيرو، وهي تحكي حياة أبناء طحان بعد موته وتقاسمه الورث، فتقاسموا أشياء مهمة ولم يتركو للأخير إلا القط لكن ابنه الثالث رضي به، في يوم كان وارث القط يلوم نفسه على ما رضي به من ورث فسمعه القط فذهب يصطاد الحيوانات وذهب بها إلى الحاكم وابنته فيقول لهما انها هدية من والي البلاد، فتيقنوا ذلك مرة فقد أمر القط صاحبه أن يسبح في النهر وخبأ ملابسه تحت صخرة، فأتى الملك وابنته يريدان أن يسلما على الفتى فقال لهم القط أن سيده في البركة يسبح وقد سرقوا ملابسه، فأعطوه ملابس فاخرة وجلس في العربة مع الأميرة التي أحبته. ثم ذهب لاحقا إلى قصر يسكنه وحش والذي يمكنه التحول إلى أي مخلوق على وجه الأرض، فتحول الغول إلى أسد فأخاف القط الذي استطاع خداعه بأن يتحول إلى فأر فهجم عليه القط والتهمه، هكذا تمكن الوريث من أخذ قصر الوحش فانبهر به الملك وتزوج ابنته وعاش هو والقطه في رخاء.

Dialogue to evaluate: - Speaker A: سمعتي على هاديك القصة ديال 'القط ذو الحذاء'؟

- Speaker B: آه، اللي فيها المش كان ذكي بزاف؟
- Speaker A: بضبط! تخيل، حول ولد عادي لأمير.
- Speaker B: وكيفاش ضحك على الغول ورجعو فأر!
- Speaker A: آه، هاديك كانت أحسن لقطة. مكار!
- Speaker B: بصح بدّل حياة مولاه، من والو لقصر!

Output: {

```
  "annotation": [
    {
      "rating": "A",
    "reason": "It is in correct Maghrebi Darija Arabic without mistakes or errors."
    }
  ]
}
```

Example 2: Rating B

Text: الطلاق (يعرف أيضًا باسم فسخ الزواج) هو عملية إنهاء العلاقة الزوجية أو الارتباط الزوجي. عادة ما يستلزم الطلاق إلغاء أو إعادة تنظيم الواجبات والمسؤوليات القانونية للزواج وبالتالي فسخ روابط الزواج بين الزوجين بموجب القانون في بلد أو دولة معينة. تختلف قوانين الطلاق بشكل كبير في جميع أنحاء العالم، ولكن في معظم البلدان يتطلب الطلاق تدخل محكمة أو سلطة أخرى في الإجراءات القانونية والتي قد تنطوي على قضايا توزيع الممتلكات وحضانة الأطفال والنفقة وزيارة الأطفال / أو

الوصول إليهم والوقت المخصص للأب / الأم لرؤية الأطفال وتقديم الدعم الطفل وتقسيم المصاريف. في معظم البلدان هناك قانون يلزم الأفراد بالزواج الأحادي لذا فإن الطلاق بحسب هذا القانون يسمح لكل شريك سابق بالزواج من شخص آخر. الدول الوحيدة التي لا تسمح بالطلاق هي الفلبين ومدينة الفاتيكان. في الفلبين لا يعتبر طلاق الفلبينيين غير المسلمين أمرًا قانونيًا إلا إذا كان الزوج أو الزوجة مهاجرًا غير شرعي ويستوفي شروطًا معينة، أما مدينة الفاتيكان فهي دولة كنسية ليس لديها إجراءات للطلاق. البلدان التي أقرت الطلاق مؤخرًا نسبيًا هي إيطاليا (0791)، البرتغال (5791 على الرغم من أنه من عام 0191 إلى عام 0491 كان ذلك ممكنًا للزواج المدني والديني)، البرازيل (7791)، إسبانيا (1891)، الأرجنتين (7891)، باراغواي (1991)، كولومبيا (1991؛ من 6791 كان مسموحًا به فقط لغير الكاثوليك)، أندورا (5991)، أيرلندا (6991)، تشيلي (4002) ومالطا (1102). يتم الطلاق عادة باتفاق الطرفين أو بإرادة أحدهما، وهو موجود لدى العديد من ثقافات العالم لكنه غير موجود لدى أتباع الكنيسة الكاثوليكية وتعتبر أشهر قضية طلاق في التاريخ عندما طلب هنري الثامن ملك إنجلترا الطلاق من كاثرين أراغون عام 4351 لكن البابا رفض ترخيص طلاقه مما أدى إلى تأسيس الكنيسة الأنجليكانية.

Dialogue to evaluate: - Speaker A: واش قريتي شي حاجة على هاري بوتر؟

- Speaker B: آه، سمعت أنه قهر لورد مظلم.
- Speaker A: واش اللي قتلوا واليديه؟
- Speaker B: إيه، لورد فولدمورت هو اللي دارها.
- Speaker A: واش بقى مع عائلة وحدة قريبتهم؟
- Speaker B: إيه، كانوا قاسين عليه بزاف.

Output:

```
{
  "annotation": [
    {
      "rating": "B",
    "reason": "The sentence "Wash b9a m3 3ayla
    wa7da qribt-hom?" is very close
    to Moroccan Darija, but it's not
    entirely natural or idiomatic as-is."
    }
  ]
}
```

Example 3: Rating C

Text: الطلاق (يعرف أيضًا باسم فسخ الزواج) هو عملية إنهاء العلاقة الزوجية أو الارتباط الزوجي. عادة ما يستلزم الطلاق إلغاء أو إعادة تنظيم الواجبات والمسؤوليات القانونية للزواج وبالتالي فسخ روابط الزواج بين الزوجين بموجب القانون في بلد أو دولة معينة. تختلف قوانين الطلاق بشكل كبير في جميع أنحاء العالم، ولكن في معظم البلدان يتطلب الطلاق تدخل محكمة أو سلطة أخرى في الإجراءات القانونية والتي قد تنطوي على قضايا توزيع الممتلكات وحضانة الأطفال والنفقة وزيارة الأطفال / أو الوصول إليهم والوقت المخصص للأب / الأم لرؤية الأطفال وتقديم الدعم الطفل وتقسيم المصاريف. في معظم البلدان هناك قانون يلزم الأفراد بالزواج الأحادي لذا فإن الطلاق بحسب هذا القانون يسمح لكل شريك سابق بالزواج من شخص آخر. الدول الوحيدة التي لا تسمح بالطلاق هي الفلبين ومدينة الفاتيكان. في الفلبين لا يعتبر طلاق الفلبينيين غير المسلمين أمرًا قانونيًا إلا إذا كان الزوج أو الزوجة مهاجرًا غير شرعي ويستوفي شروطًا معينة، أما مدينة الفاتيكان فهي دولة كنسية ليس لديها إجراءات للطلاق. البلدان التي أقرت الطلاق مؤخرًا نسبيًا هي إيطاليا (0791)، البرتغال (5791 على الرغم من أنه من عام 0191 إلى عام 0491 كان ذلك ممكنًا للزواج المدني والديني)، البرازيل (7791)، إسبانيا (1891)، الأرجنتين (7891)، باراغواي

(1991)، كولومبيا (1991؛ من 6791 كان مسموحًا به فقط لغير الكاثوليك)، أندورا (5991)، أيرلندا (6991)، تشيلي (4002) ومالطا (1102). يتم الطلاق عادة باتفاق الطرفين أو بإرادة أحدهما، وهو موجود لدى العديد من ثقافات العالم لكنه غير موجود لدى أتباع الكنيسة الكاثوليكية وتعتبر أشهر قضية طلاق في التاريخ عندما طلب هنري الثامن ملك إنجلترا الطلاق من كاثرين أراغون عام 4351 لكن البابا رفض ترخيص طلاقه مما أدى إلى تأسيس الكنيسة الأنجليكانية.

Dialogue to evaluate: - Speaker A: سمعت بلي الطلاق عندنا فالمغرب مختلف على الدول الأخرى، صحيح؟

- Speaker B: إيوا، كل بلاد عندها قانون خاص بيها. حتى فلبين والفاتيكان ممنوع عندهم الطلاق أساسًا!
- Speaker A: واااا! حتى لو كان الزوجين مايتافقوش ماعندهمش حل؟
- Speaker B: بالضبط، غير المسلمين هناك صعيب عليهم. لكن فإيطاليا مثلا الطلاق مكنشرح حتى 0791.
- Speaker A: حتى الكنيسة الكاثوليكية مابقاتش تسمح بالطلاق، ولا زال؟
- Speaker B: إيوا، لدرجة أن ملك إنجلترا خلق كنيسة جديدة عشان يطلق!

Output:

```
{
  "annotation": [
    {
      "rating": "C",
      "reason": "The conversation has some expressions in Maghrebi Darija,
       However it has some awkward expressions like wla zal  and
       "Iwa, ldarja enn malik ingltra khlaq knisa jdida 3shan ytla9!
       is more Egyptian Dialect"
    }
  ]
}
```

Example 4: Rating D

Text: جراند ثفت أوتو (بالإنجليزية: Grand Theft Auto، تختصر إلى GTA) هي سلسلة من ألعاب المغامرات والحركة التي أنشأها ديفيد جونز ومايك ديلي. تم تطوير العناوين اللاحقة تحت إشراف الأخوين دان وسام هاوسر، ليزلي بنزيس، وآرون جاربوت. تم تطوير اللعبة بشكل أساسي من قبل شركة التطوير البريطانية روكستار نورث (دي إم أي ديزاين (DMA Design) سابقًا)، ونشرتها الشركة الأم روكستار جيمز. يشير اسم السلسلة إلى مصطلح «جراند ثفت أوتو»، المستخدم في الولايات المتحدة لسرقة السيارات. تركز طريقة اللعب على عالم مفتوح حيث يمكن للاعب إكمال المهام للتقدم في قصة شاملة، بالإضافة إلى الانخراط في أنشطة جانبية مختلفة. تدور معظم طريقة اللعب حول القيادة وإطلاق النار، مع لعب الأدوار من حين لآخر وعناصر التخفي. تحتوي السلسلة أيضًا على عناصر من ألعاب بيت إم السابقة من عصر 61 بت. تم وضع الألعاب في سلسلة جراند ثفت أوتو في أماكن خيالية على غرار مدن الحياة الواقعية، في نقاط زمنية مختلفة من أوائل الستينيات إلى العقد الأول من القرن الحادي والعشرين. اشتملت خريطة اللعبة الأصلية على ثلاث مدن — ليبرتي سيتي (استنادًا إلى مدينة نيويورك)، وسان أندرياس (استنادًا إلى سان فرانسيسكو)، وفايس سيتي (استنادًا إلى ميامي) — ولكن تميل العناوين اللاحقة إلى التركيز على مكان واحد عادةً ما تكون إحدى المناطق الثلاث الأصلية، وإن تم إعادة تشكيلها وتوسيعها بشكل كبير. يركز المسلسل على أبطال مختلفين يحاولون الصعود في مراتب العالم السفلي الإجرامي، على الرغم من أن دوافعهم للقيام بذلك تختلف في كل عنوان. عادةً ما يكون الخصوم شخصيات قد خانوا بطل الرواية أو منظمتهم، أو الشخصيات التي لها التأثير الأكبر في إعاقة تقدم بطل الرواية. أعرب العديد من قدامى الأفلام والموسيقى عن شخصيات في الألعاب، بما في ذلك راي ليوتا، دينيس هوبر، صامويل جاكسون، ويليام فيشتنر، جيمس وودز، ديبي هاري، أكسل روز، وبيتر فوندا.بدأت دي إم أي ديزاين السلسلة في عام 7991 بإصدار جراند ثفت أوتو. اعتبارًا من 0202، تتكون السلسلة من سبعة عناوين مستقلة وأربع حزم توسعة. يعتبر العنوان الرئيسي

الثالث، جراند ثفت أوتو 3، الذي تم إصداره في عام 1002، لعبة تاريخية، حيث جلبت السلسلة إلى إعداد ثلاثي الأبعاد (3D) وتجربة أكثر مغامرة. اتبعت العناوين اللاحقة وبنيت على المفهوم الذي تم تأسيسه في جراند ثفت أوتو 3، وحظيت بإشادة كبيرة. لقد أثروا على ألعاب الحركة الأخرى في العالم المفتوح، وأدى إلى استنساخ جراند ثفت أوتو على عناوين مماثلة. حازت السلسلة على استحسان النقاد، حيث تم تصنيف جميع الإدخالات ثلاثية الأبعاد الرئيسية في السلسلة بشكل متكرر ضمن ألعاب الفيديو التي تعد الأفضل والأكثر مبيعًا؛ شحنت أكثر من 053 مليون وحدة، مما يجعلها خامس أفضل سلسلة لألعاب الفيديو مبيعًا. في عام 6002، ظهرت جراند ثفت أوتو في قائمة أيقونات التصميم البريطانية في غريت برِتش ديزاين كويست (Great British Design Quest) التي نظمتها هيئة الإذاعة البريطانية ومتحف التصميم. في عام 3102، صنفت صحيفة ذا تلغراف جراند ثفت أوتو من بين أنجح الصادرات البريطانية. كانت السلسلة أيضًا مثيرة للجدل بسبب طبيعتها البالغة وموضوعاتها العنيفة. في 4 فبراير 2202، أكدت روكستار أن هنالك عنوان جديد للسلسلة قيد التطوير.

Dialogue to evaluate: - Speaker A: Katchouf GTA? Silsila dyal l-lâ3ab ktab3ha bnadem bezzaf.

- Speaker B: Ah, kan3refha. Wahed men akbar l-âbâb fel 3âlam.
- Speaker A: Bdat f 1997, u GTA 3 hiya li bedlat kolchi f 2001.
- Speaker B: Bessa7, dak l3ab fih 3alam meftou7 u zwin bezzaf.
- Speaker A: Wla makhbartekch, kaywjedo fiha jeu jdid daba.
- Speaker B: Hada khbar zwin! Dima katbqa men a7san l-lâ3ab.

Output:

```
{
  "annotation": [
    {
      "rating": "C",
      "reason": "The conversation is correct in Maghrebi Darija.
However, it uses Arabic Franco which is an error"
    }
  ]
}
```


Example 5: Rating D

Text: الدور أو تيمبو هو نقلة واحدة يلعبها لاعبا الشطرنج بالتناوب ويتم فيها تحريك أحد قطع الشطرنج مرة واحدة وفق قوانين الشطرنج، حين يحقق اللاعب الوضعية التي يرغب فيها في عدد نقلات أقل بنقلة واحدة من المخطط له يقال رِبح دورا وبالعكس إن استغرق نقلة أكثر من المطلوب يقال فقدان دور لحين يجبر اللاعب الخصم على تنفيذ نقلات ليست في مخططه الأساسي فذلك ربح للنقلات لأن الخصم مضطر لتأجيل مخططه والرد على الإجبار، النقلة التي تُربح دورا تسمى في الغالب «نقلة مع دور». للدور أهمية كبيرة ومن الضروري عدم فقدان الأدوار لأن ذلك ينعكس على نجاعة الدفاع أو الهجوم انظر (شكل1)، مثال بسيط على فقدان دور هو تحريك القلعة في (الشكل2) من h1 إلى h5 في دور ثم إلى h8 في دور ثاني في حين كان يمكن فعل ذلك في نقلة واحدة وتوفير دور، مناورات كتلك لا تتسبب دائما في فقدان تيمبو ففي بعض الأحيان قد تقوم القلعة بتهديد يجب الرد عليه وفي هذه الحالة يعتبر أن كلا اللاعبين فقدا دورا وهذه النتيجة تعادل قياسا على الزمن لكن التغير الذي طرأ على الوضعية يمكن أن يكون في فائدة أحدهما أكثر من الآخر.

Dialogue to evaluate: - Speaker A: شفت المباراة اللي لعبها حسن مع علي بالشطرنج؟.

- Speaker B: أيوه، حسن كان دايما يربح دورا بنقلاته.
- Speaker A: صحيح، علي كان فاقد دور كل مرة.
- Speaker B: مرة حاول يجبر حسن على نقلات ليست في مخططه.
- Speaker A: بس حسن كان دايما يرد بنقلة مع دور.

- Speaker B: يعني كان يحافظ على أدواره جيد.

Output: {

```
"annotation": [
  {
   "rating": "D",
   "reason": "The conversation is totally in Egyptian Dialect while it should be
    in Maghrebi Darija"
  }
]
}
```

Your Input Text:
text
Dialogue to evaluate:
dialogue

Egyptian Dialogue Correction Task – Prompt

Your task is to fix an AI-generated dialogue in Egyptian Dialect Arabic. The conversation must be based strictly on the provided source text. You should produce a six-turn dialogue (three exchanges between two speakers) that would earn an "A" rating under the rubric below.

Generated Dialogue to Fix: `{dialogue}`
Source Text: `{text}`
Original Rating: `{rating}`
Reason for Rating: `{reason}`

It was generated according to the following features:

1. General Framework

- Be natural, relatable, and culturally appropriate in Egyptian Dialect Arabic.
- The dialogue must be natural, smooth, and realistic.
- The dialogue should be fully in Egyptian Dialect Arabic.
- Avoid generating Q&A style conversations without proper transitions and contextual expressions.
- Use Egyptian Dialect Arabic, with cultural appropriateness.
- Each turn must be between 1 and 20 words.

2. Content and Style

- Choose topics that are personal, work-related, or about daily routines.
- Keep the language simple and easy to understand.
- Add a light, casual tone to make the conversation engaging.
- Avoid using dialects other than Egyptian Dialect Arabic.
- Refrain from using personal or emotional address terms.

3. Technical Constraints

- Do not add any information or details that are not derived from the original text.
- Do not use any special characters, symbols, or emojis.

Evaluation Rubric (A – D)

Rating A

- Fully correct in Egyptian Dialect Arabic without errors or slips.
- Grammar, vocabulary, idioms, and expressions are accurate and appropriate.
- Dialogue flows naturally and coherently, with smooth transitions and contextual expressions.
- Examples:
  - Q&A with smooth transitions and contextual expressions.
  - Mainly in Egyptian Dialect Arabic with possibly one or two standard Arabic words.

Rating B

- Generally correct with no grammatical or vocabulary errors.

- May feel slightly unnatural or disconnected.
- Examples:
  - Q&A without smooth transitions.
  - Lacks idiomatic expressions or natural flow.

Rating C

- Minor issues that do not impact comprehension.
- May include slight grammatical mistakes or awkward phrasing.
- Some use of MSA or non-Egyptian Arabic dialects.
- Examples:
  - Egyptian Dialect with some standard Arabic terms.
  - Few spelling or phrasing errors.

Rating D

- Significant problems or use of non-Egyptian dialects or non-Arabic script.
- Irrelevant to the source text.
- Examples:
  - Mostly MSA or another dialect.
  - Use of Latin script (e.g., Franco-Arabic) unless for proper names like `BBC`.
  - Dialogue is irrelevant.

Note: If the dialogue meets multiple criteria, assign the lowest (worst) appropriate rating.

Output Format (JSON)

```
{
  "dialogue": [
    {
      "speaker": "A",
      "text": "-Text-"
    },
    {
      "speaker": "B",
      "text": "-Text-"
    }
    ...
  ]
}
```

Modern Standard Arabic Dialogue Correction Task – Prompt

Your task is to fix an AI-generated dialogue in Modern Standard Arabic. The conversation must be based strictly on the provided source text. You should produce a six-turn dialogue (three exchanges between two speakers) that would earn an "A" rating under the rubric below.

Generated Dialogue to Fix: `{dialogue}`
Source Text: `{text}`
Original Rating: `{rating}`
Reason for Rating: `{reason}`

It was generated according to the following features:

1. General Framework

- Be natural, relatable, and culturally appropriate in Modern Standard Arabic.
- The dialogue must be natural, smooth, and realistic.
- The dialogue should be fully in Modern Standard Arabic.
- Avoid generating Q&A style conversations without proper transitions and contextual expressions.
- Use Modern Standard Arabic, with cultural appropriateness.
- Each turn must be between 1 and 20 words.

2. Content and Style

- Choose topics that are personal, work-related, or about daily routines.
- Keep the language simple and easy to understand.
- Add a light, casual tone to make the conversation engaging.
- Avoid using dialects other than Modern Standard Arabic.
- Refrain from using personal or emotional address terms.

3. Technical Constraints

- Do not add any information or details that are not derived from the original text.
- Do not use any special characters, symbols, or emojis.

Evaluation Rubric (A—D)

Rating A

- Fully correct in Modern Standard Arabic without any errors or slips.
- Grammar, vocabulary, idioms, and expressions are accurate and appropriate.
- Dialogue flows naturally and coherently, with smooth transitions and contextual expressions.
- Examples:
  - The conversation is like Q&A but with smooth transitions and contextual expressions.
  - The conversation is mainly in Modern Standard Arabic.
  - The conversation could include one or two natural standard Arabic words.

Rating B

- Generally correct in Modern Standard Arabic, with no grammatical or vocabulary errors.
- Slightly unnatural or feels disconnected.
- Examples:
  - Q&A without smooth transitions.
  - Lacks transitional phrases or idiomatic expressions.
  - Mostly a collection of disconnected sentences.

Rating C

- Contains minor issues but remains comprehensible.
- May include slight grammatical errors or awkward phrasing.
- Occasional use of dialect or non-native MSA terms.
- Examples:
  - MSA mixed with Egyptian Dialect Arabic or others.
  - A few spelling or expression errors.

Rating D

- Major problems in language use or script.
- Mostly non-MSA or contains non-Arabic script (except for proper nouns like `BBC` or `Time News`).
- Dialogue is irrelevant or incoherent.
- Examples:
  - Mostly in Egyptian or other dialects.
  - Written in non-Arabic script or a mixture.
  - Not related to the original text at all.

Note: If the conversation meets multiple rating criteria, assign the lowest (worst) applicable rating.

Output Format (JSON)

```
{
  "dialogue": [
    {
      "speaker": "A",
      "text": "-Text-"
    },
    {
      "speaker": "B",
      "text": "-Text-"
    }
    ...
  ]
}
```

Maghrebi Darija Dialogue Correction Task – Prompt

Your task is to fix an AI-generated dialogue in Maghrebi Darija. The conversation must be based strictly on the provided source text. You should produce a six-turn dialogue (three exchanges between two speakers) that would earn an "A" rating under the rubric below.

Generated Dialogue to Fix: `{dialogue}`
Source Text: `{text}`
Original Rating: `{rating}`
Reason for Rating: `{reason}`

It was generated according to the following features:

1. General Framework

- Be natural, relatable, and culturally appropriate in Maghrebi Darija.
- The dialogue must be natural, smooth, and realistic.
- The dialogue should be fully in Maghrebi Darija.
- Avoid generating Q&A style conversations without proper transitions and contextual expressions.
- Use Maghrebi Darija, with cultural appropriateness.
- Each turn must be between 1 and 20 words.

2. Content and Style

- Choose topics that are personal, work-related, or about daily routines.
- Keep the language simple and easy to understand.
- Add a light, casual tone to make the conversation engaging.
- Avoid using dialects other than Maghrebi Darija.
- Refrain from using personal or emotional address terms.

3. Technical Constraints

- Do not add any information or details that are not derived from the original text.
- Do not use any special characters, symbols, or emojis.

Evaluation Rubric (A—D)

Rating A

- Fully correct in Maghrebi Darija without any errors or slips.
- Grammar, vocabulary, idioms, and expressions are accurate and appropriate.
- Dialogue flows naturally and coherently, with smooth transitions and contextual expressions.
- Examples:
  - Q&A with smooth transitions and contextual expressions.
  - Mainly in Maghrebi Darija.
  - May contain one or two natural Modern Standard Arabic words.

Rating B

- Generally correct in Maghrebi Darija, with no grammatical or vocabulary errors.
- Dialogue may feel slightly unnatural.
- Examples:
  - Q&A without smooth transitions.
  - Missing transitional or idiomatic expressions.
  - Mostly a collection of disconnected sentences.
  - Some slightly awkward or unidiomatic phrases.

Rating C

- Contains minor issues that do not affect comprehension.
- Slight grammatical mistakes or awkward phrasing.
- Occasional non-Maghrebi Darija elements (e.g., MSA, Tunisian, Algerian, Lebanese).
- Examples:
  - MSA with some Maghrebi Darija words.
  - Maghrebi Darija with non-Maghrebi dialect words.
  - Minor spelling errors.
  - Arabic Franco (Arabic written in Latin characters and numbers).

Rating D

- Significant problems in Maghrebi Darija or script.
- Mostly non-Maghrebi Darija.
- Non-Arabic script (not Arabic Franco), e.g., English or French, unless using proper nouns with no Arabic equivalent (e.g., BBC, Time News).
- Irrelevant to the source text.
- Examples:
  - Mostly in MSA or non-Maghrebi dialects.
  - Contains mixed or entirely non-Arabic script.

Note: If multiple ratings apply, assign the lowest applicable rating.

Output Format (JSON)

```
{
  "dialogue": [
    {
      "speaker": "A",
      "text": "-Text-"
    },
    {
      "speaker": "B",
      "text": "-Text-"
    }
    ...
  ]
}
```

H Appendix H: Examples

H.1 Egyptian Dialect

H.1.1 Rating A

جيش التحرير التكافلي (بالإنجليزية: Symbionese Liberation Army) هو مجموعة إرهابية أمريكية بين عامي 3791 و5791 أطلقت على نفسها صفة جيش طليعي ثوري. وكان قائد تلك المجموعة دونالد ديفريز. قامت المجموعة بعمليات سطو على بنوك وجريمتي قتل وأعمال عنف أخرى. اشتهرت المجموعة عالميا حينما اختطفت باتي هيرست الوريثة الإعلامية ذات التسعة عشر عاما مع صديقها ستيفن ويد أثناء إقامتهما في المنزل في بيركلي في كاليفورنيا. ذهل العالم كله حينما بثت هيرست رسائل صوتية أذيعت وسائل الإعلام تدين فيها والديها وتعلن انضمامها إلى المجموعة الإرهابية. واشتركت في نشاطاتهم الغير شرعية. وادعت هيرست فيما بعد أنها احتجزت واعتدي عليها جنسيا وجرى لها غسيل دماغي. انتهى نشاط المجموعة باعتقال أفرادها في عام 5791.

Dialogue: - Speaker A: فاكر الخبر بتاع جيش التحرير التكافلي؟ اللي كانت مشهورة زمان.

- Speaker B: آه، تقصد المجموعة الإرهابية دي؟ إيه اللي فكرك بيهم؟
- Speaker A: أيوة بالظبط! عملوا سرقات وقتل كتير، قصة غريبة.
- Speaker B: أكتر حاجة شهرتهم، خطف البنت الوريثة الإعلامية.
- Speaker A: صح، وصوتها طلع في الراديو بتشتم أهلها! الناس كلها اتصدمت.
- Speaker B: وفي الآخر اتقبض عليهم كلهم، صح كده؟

Reason: The conversation is fully correct in Egyptian Dialect Arabic, with smooth and coherent transitions, contextual expressions, and accurate grammar and vocabulary. The dialogue flows naturally, adhering to the provided text.

H.1.2 Rating B

التأثير الحثي (Inductive effect): قطبية مستحثة تنشأ في الرابطة التساهمية بسبب الاختلاف في السالبية الكهربية بين الذرات المتصلة ببعض.التأثير الحثي في الكيمياء والفيزياء هو تأثير يلاحظ تجريبياً وهو ناتج من انتقال الشحنة عبر سلسلة من الذرات في جزئ ما بواسطة الحث الكهربي. والتأثير القطبي الكلي المبذول بواسطة مستَبْدَل عبارة عن مزيج من التأثير الحثي والتأثير الميزوميري. السحابة الإلكترونية في الرابطة سيجما σ-bond التي تربط بين ذرتين مختلفتين ليست متماثلة على جانبي الرابطة ولكنها مزاحة قليلاً باتجاه الذرة الأكثر سالبية كهربية. وهذا يسبب حالة ثابتة من القطبية للرابطة حيث تحمل الذرة الأكثر سالبية كهربية شحنة سالبة جزيئية (δ-) بينما تحمل الذرة الأخرى على شحنة موجبة جزيئية (δ+). إذا اتصلت الذرة الأكثر سالبية بسلسلة من الذرات (عادة سلسلة كربونية) فإن الشحنة الموجبة تُرَحَّل إلى الذرات الأخرى في هذه السلسلة، وهذا هو التأثير الحثي الساحب للإلكترونات electron-withdrawing ويرمز إليه بـ (I-) أي التأثير الحثي السالب. بعض المجموعات، مثل مجموعات الألكيل تكون أقل من الهيدروجين في سحب الإلكترونات ولهذا تعتبر مجموعات طاردة للإلكترونات electron-releasing وبالتالي يكون لها تأثير حثي موجب ويرمز إليه بـ (I+). كلما كانت القطبية المستحثة أقل من القطبية الأصلية، كان التأثير الحثي أسرع اختفاءً ومؤثراً فقط على مسافة قصيرة. التأثير الحثي تأثير دائم ولكنه ضعيف لأنه يتضمن إزاحة إلكترونات الرابطة القوية سيجما، بالإضافة إلى عوامل أخرى أقوى من ممكن أن تحجبه. تم تجريبيا قياس التناسب بين التأثيرات الحثية بالنسبة إلى الهيدروجين وهي كالتالي: يمكن أيضاً قياس التأثيرات الحثية بواسطة معادلة هاميت .Hammett Equation

Dialogue: - Speaker A: إيه رأيك في التأثير الحثي ده؟.

- Speaker B: والله حاسة إنه لغبطة شوية!
- Speaker A: خصوصاً حكاية انتقال الشحنة دي.
- Speaker B: آه، وكمان إزاحة سحابة الإلكترونات.
- Speaker A: بس هو تأثيره ضعيف ومسافته قصيرة.
- Speaker B: فعلاً، يعني تأثيره محدود أوي.

Reason: The conversation is generally correct in Egyptian dialect Arabic, without grammatical or vocabulary errors. However, it lacks smooth transitions and some contextual expressions, making it feel slightly like a Q&A without cohesive transitions between topics.

H.1.3 Rating C

آلة محدودة الحالات (بالإنجليزية: Finite-State Machine) اختصاراً MSF، أو ببساطة آلة الحالات هي نموذج حوسبة رياضي يستخدم لتصميم دارات المنطق المتتابع والبرامج الحاسوبية. وينظر على أنها آلة مجردة يمكن أن تكون في واحدة من عدد محدود من الحالات. تكون الآلة في حالة واحدة فقط في وقت واحد، ويطلق على هذه الحالة في هذه اللحظة: الحالة الراهنة. ويمكن أن تتغير من حالة إلى أخرى عند تفعيل حدث ما أو شرط، وهذا ما يسمى مرحلة انتقالية. وتعرف آلة حالات منتهية محددة بقائمة من حالاتها، حالتها الأولية، وشرط الانتقال من كل حالة إلى أخرى. آلة الحالات المنتهية يمكن أن تحل عدد كبير من المشاكل، ومنها ماهو متمم لتصميم الإلكتروني وتصميم بروتوكول الاتصال والتحليل والتطبيقات الهندسية الأخرى. وبحوث البيولوجيا وبحوث الذكاء الاصطناعي، وتستخدم أحياناً لوصف النظم العصبية، واللغويات ويمكن استخدامها لوصف لسانيات اللغات الطبيعية.

Dialogue: - Speaker A: شوفت يا صاحبي آلة الحالات دي عاملة إزاي؟.

- Speaker B: آه، اللي ليها عدد محدود من الحالات دي؟
- Speaker A: بالظبط، وبتكون في حالة واحدة بس في الوقت الواحد.
- Speaker B: طب وإزاي بتتنقل من حالة للتانية؟
- Speaker A: لازم تحصل حاجة أو شرط معين، ودي اسمها نقلة.
- Speaker B: بجد؟! وكمان بيستخدموها في اللغويات والذكاء الاصطناعي!

Reason: The conversation is in Egyptian Dialect Arabic but contains minor slips: The use of 'شوفت يا صاحبي' is more Levantine than Egyptian, and the dialogue is missing some transitional expressions, making it feel less fluid. It also lacks some idiomatic expressions that would make the interaction more natural in Egyptian Dialect Arabic.

H.1.4 Rating D

الولاية (بالإنجليزية: Suzerainty) هي العلاقة التي يكون بمقتضاها سيطرة منطقة أو كيان سياسي ما على السياسة والعلاقات الخارجية لدولة أخرى مع السماح للدولة التابعة في التمتع بحكم ذاتي داخلي.تختلف الهيمنة عن السيادة الحقيقة في ذلك إذ بالرغم من كون الشخص التابع أو الدولة التابعة، عملياً، مستقلان ويتمتعان بحكم ذاتي الا ان تلك الاستقلالية تكون محدودة. بالرغم من وجود هذه الحالة في عدة امبراطوريات تاريخية فإنها لا تتلاءم مع مفاهيم القانون الدولي في القرن العشرين أو الواحد والعشرين سواء كانت هناك سيادة أو لم تكن في حين يتسنى لأي دولة ذو سيادة لتكون محميّة تابعة لنفوذ اقوى من خلال عقد معاهدة الا أن القانون الدولي لا يعترف باي وسيلة تجعل من هذه العلاقة قسرية على النفوذ الاضعف، لكن، التبعية تكون بحكم الواقع أكثر مما هي بحكم القانون.

Dialogue: - Speaker A: حاسس إن فيه سيطرة زيادة عليا في الشغل.

- Speaker B: مش مفروض تكون مستقل جوه الشغل؟
- Speaker A: آه، بس استقلاليتي دي محدودة بصراحة.
- Speaker B: طيب، ده أمر واقع ولا فيه قانون بيفرضه؟
- Speaker A: أظن ده أمر واقع، مفيش حاجة غصب علينا.
- Speaker B: ممكن نحتاج نعمل اتفاق واضح.

Reason: The conversation is irrelevant to the provided text about suzerainty and its political connotations. The dialogue is centered around workplace autonomy, which does not align with the text's theme of political and legal control between states.

H.2 Modern Standard Arabic

H.2.1 Rating A

نظـام التعليم في الولايات المتحدة نظام«لا مركزي»، ولهذا السبب فان القوانين التي تحكم هيكل ومضمون برامج التعليم تتنوع بدرجة كبيرة ما بين ولاية وأخرى، ومع ذلك تبدو هذه البرامج متشابهة بشكل ملحوظ بسبب العوامل المشتركة بين هذه الولايات كالحاجات الاجتماعية والاقتصادية والتنقل المتكرر للطلاب والمعلمين من ولاية إلى أخرى ومن ثم فإن التجريب والتنوع في كل ولاية لا يعوق دون ظهور شكل عام للنظام التعليمي في أمريكا. والتعليم العام إجباري في الولايات المتحدة الأمريكية ومجاني في كافة المدارس الحكومية ويبدأ عادة من سن السادسة أو السابعة وحتى سن السادسة عشر وإلى أن يستكمل الطالب دراسة المرحلة الثانوية التي تنتهي في الصف الثاني عشر أما بالنسبة للمدارس الخاصة فيسمح لها بالعمل وفق تراخيص خاصة وقواعد تتبع لإعتماد هذه المدارس من قبل الولاية التابعة لها يتـم التدريس في معظم صفوف الدراسة باللغة الإنجليزية، إلا في المدارس التي يوجد فيها كثافة عالية من الطلاب الذين لا تكون لغتهم الأولى هي اللغة الإنجليزية وفي هذه الحالة يتم تدريس المناهج بلغة غير اللغة الإنجليزية مع تكثيف تدريس اللغة الإنجليزية لغير الناطقين بها إلى أن يصبح الطالب مؤهلا للدراسة في الفصول العادية التي تقوم بتدريس مناهجها باللغة الإنجليزية.ويعد التعليم في الولايات المتحدة من أهم أسباب التطور الحالي. ويتم التركيز عليه كثيراَ من خلال اللجان والحكومة، التعليم في الولايات المتحدة الأمريكية يقدم بشكل أساسي من القطاع العام مع مراقبة وتمويل يأتي من ثلاثة مستويات: إتحادي ومحلي. وتعليم الأطفال بشكل إلزامي. التعليم العام هو يوفر عالمياً المناهج الدراسية والتمويل والمدرسين وغيرها من السياسات التعليمية وترد عن طريق مجالس منتخبة محلياً مع سلطتها القضائية على المناطق التعليمية وبتوجهات العديد من المجالس التشريعية في الولايات. أما بالنسبة للمعايير التعليمية ومقررات الاختبار الموحد فهي عادة ما تقدم من جانب حكومات الولايات. سن بداية التعليم الإلزامي يختلف من ولاية لأخرى وهي تبدأ من سن الخامسة إلى الثامنة وتنتهي من سن الرابعة عشر إلى الثامنة عشرة. وهناك عدد من الولايات تتطلب إلزامية التعليم حتى سن 81 سنة. أهمية ومتطلبات التعليم الإلزامي تثقيف الأطفال في المدراس العامة، والدولة موافقة على المدارس الخاصة وبرنامج التعليم المنزلي. التعليم ينقسم إلى ثلاثة مستويات: المدرسة الابتدائية والمدرسة المتوسطة (تسمى الإعدادية) والمدرسة الثانوية. في عام 0002 كان هناك 6.67 مليون طالب وطالبة مسجلين في المدراس من رياض الأطفال والمدارس من خلال الدراسات العليا من هؤولاء 27 في المائة منهم تتراوح أعمارهم بين 21 و71 عاما وتم الحكم أكاديمياً لسنهم. ومن المقيدين في التعليم الإلزامي 2.5 مليون أي 10.4% كانوا يدرسون في مدارس خاصة. من بين سكان البلاد البالغين، أكثر من 58 في المئة قد أكملوا المرحلة الثانوية و72 بالمائة حصلوا على درجة البكالوريوس أو أعلى.

Dialogue: - Speaker A: سمعت أن نظام التعليم في أمريكا لا مركزي جداً.

- Speaker B: هذا صحيح تماماً، ولذلك تختلف القوانين والمناهج بشكل كبير بين الولايات.
- Speaker A: ومع ذلك، تبدو البرامج متشابهة بشكل ملحوظ. كيف ذلك؟
- Speaker B: يرجع ذلك للعوامل المشتركة، كالحاجات الاجتماعية والاقتصادية وتنقل الطلاب والمعلمين.
- Speaker A: بالمناسبة، هل التعليم العام إلزامي ومجاني هناك؟
- Speaker B: نعم، إنه إجباري ومجاني بالمدارس الحكومية حتى سن السادسة عشرة.

Reason: The conversation is fully correct in Modern Standard Arabic, with accurate grammar and vocabulary. It flows naturally and coherently, with smooth transitions and contextual expressions.

H.2.2 Rating B

الوجه الأسود (بالإنجليزية: Blackface) هو مصطلح يستخدم في الولايات المتحدة لشكل من أشكال المكياج المسرحي يستخدم في الغالب من قبل الفنانين غير السود لتمثيل صورة كاريكاتورية لشخص أسود. يستخدم هذا المصطلح أيضًا في بعض الثقافات الأخرى للإشارة إلى المكياج الأسود الذي يتم وضعه كجزء من التمويه والتقاليد الشعبية، ولا يُنظر إليها جميعًا على أنها تمثيلية من السود أو نشأت عنها، وبالتالي هناك جدل حول ما إذا كانوا عنصريين أم لا.اكتسبت هذه الممارسة شعبية كبيرة في الولايات المتحدة خلال القرن التاسع عشر وساهمت في انتشار القوالب النمطية العنصرية وظهور بعض الشخصيات العنصرية المسيئة. بحلول منتصف القرن، أصبحت العروض الكوميدية الغنائية ذات الوجه الأسود نوعاً فنيًا أمريكيًا مميزًا، محولةً العروض الرسمية مثل الأوبرا إلى عروض شعبية للعوام. في بدايات القرن العشرين، انفصلت عروض الوجه الأسود عن العروض الموسيقية وأصبحت شكلاً فنياً بحد ذاتها. في الولايات المتحدة، تراجعت شعبية عروض الوجه الأسود في نهاية القرن الحادي والعشرين، وأصبحت تعتبر بشكل عام مسيئة وغير محترمة وعنصرية. ولكن استمرت هذه الممارسة في بلدان أخرى حيث لا تعتبر عنصرية أو مسيئة.

Dialogue: - Speaker A: قرأت مقالاً اليوم عن "الوجه الأسود" وتاريخه المثير للجدل.

- Speaker B: أوه، سمعت عنه، لكن ما هو بالتحديد؟
- Speaker A: في الواقع، هو مكياج مسرحي استخدم بالولايات المتحدة لتصوير كاريكاتوري للأشخاص السود.
- Speaker B: آه، يبدو أن هذا ساهم فعلاً في انتشار القوالب النمطية العنصرية المسيئة.
- Speaker A: صحيح تماماً. والمثير للاهتمام أن بعض الثقافات الأخرى لا تعتبره عنصرياً دائماً.
- Speaker B: نعم، إنه إجباري ومجاني بالمدارس الحكومية حتى سن السادسة بالفعل. بخلاف الولايات المتحدة، حيث أصبح مسيئاً للغاية، يستمر في بلدان أخرى.

Reason: The conversation is generally correct in Modern Standard Arabic with no grammatical or vocabulary errors. However, it feels more like a Q&A without smooth transitions or idiomatic expressions, which makes it slightly unnatural.

H.2.3 Rating C

رودسيا (بالإنجليزية: Rhodesia)، وتعرف رسمياً بجمهورية رودسيا، وهي دولة غير معترف بها تقع في جنوب إفريقيا أثناء الحرب الباردة من عام 5691 إلى 9791. وتشكل المنطقة التي تعرف حالياً باسم زيمبابوي وعاصمتها هراري، وكانت قد نشأت على أنقاض المستعمرة البريطانية السابقة روديسيا الجنوبية. استقلت رودسيا أثناء محاولة من الحكومة التي يسيطر عليها البيض لتأخير التحول نحو حكم الأغلبية السوداء، فقامت حكومة إيان سميث من الأقلية البيضاء بإعلان الاستقلال من جانب واحد في 11 نوفمبر 5691. سعت الحكومة بالبداية للاعتراف بروديسيا بالاستقلال ضمن دول الكومنولث، لكن أعيد تشكيلها كجمهورية سنة 0791. في أعقاب حرب عصابات وحشية مع إتحاد شعب زيمبابوي الإفريقي بقيادة روبرت موغابي وحركة زابو بقيادة جوشوا نكومو، في عام 8791 أقر رئيس الوزراء إيان سميث الديمقراطية ثنائية العرق، لكن فشلت الحكومة المؤقتة التي أسسها مع زميله المعتدل أبيل موزيروا في وقف إراقة الدماء وقبول المجتمع الدولي. في أبريل 0891 أُعلن استقلال روديسا باعتراف بريطانيا والأمم المتحدة تحت اسم جمهورية زيمبابوي.

Dialogue: - Speaker A: هل تعلم شيئاً عن رودسيا؟ كانت دولة غير معترف بها بجنوب إفريقيا.

- Speaker B: نعم، أعتقد أنها أصبحت الآن زيمبابوي، وعاصمتها هراري.
- Speaker A: بالضبط! وقد نشأت هذه الدولة على أنقاض روديسيا الجنوبية البريطانية.
- Speaker B: وكيف حدث استقلالها ومتى كان ذلك؟
- Speaker A: لا، أُعلن استقلالها من جانب واحد في نوفمبر 5691 لتأخير حكم الأغلبية السوداء.
- Speaker B: أفهم. وفي عام 0891، أُعلن استقلالها رسمياً كجمهورية زيمبابوي.

Reason: The dialogue includes a minor issue: Speaker A inaccurately responds 'لا' when they likely meant 'نعم' while discussing the declaration of independence. Otherwise, the conversation is generally correct in Modern Standard Arabic, with no major grammatical or vocabulary errors. The overall comprehension is not affected, but the slip does impact the natural flow slightly.

H.2.4 Rating D

عسر الجماع (بالإنجليزية: Dyspareunia) صعوبة، أو ألم واضطرابات الجنسية التي يكون منشؤها في معظم الأحيان هو تشنج المهبل. الالتهابات المهبلية المعدية, ذات المنشأ الطفيلي أو الفطري مثل داء المبيضات، وداء المشعرات والمتدثرات، تسبب ألما عند المجامعة، ويكون هذا الألم مصحوباً بسيلان مهبلي مثير للحكة في معظم الأحيان، ولكنه سريع الشفاء عند معالجة أسبابه. وغالبا ما تترافق الالتهابات المهبلية مع جفاف المهبل الذي يمكن أن يستمر بعد معالجة الالتهاب، فيتطلب الأمر علاجا خاصا. تظهر بعض الأوجاع الخاصة على اثر عملية جراحية أو بعد الولادة في الأسبوعين الأولين من الجماع، حيث الألم ناجم عن إدخال القضيب، أو حركته في المهبل، ويمكن أن يكون ناجماً عن جرح عميق في الحوض، أو عند إدخال غير متناسب، إذا كان حجم المهبل صغيراً، مثل اغتصاب الصغار. وقد تنجم الأوجاع عن تشوه في الأول قد زال، أم ما زال، ولكن المرأة تتوقع حدوث الألم مسبقاً فتعاني من جفاف وتضيق المهبل. هناك أيضا الألم المهبلي الذي تشعر به المرأة عند المجامعة في مرحلة الإياس، ينجم عن ضمور انقلاب وضع الرحم، لأن هذه العمليات غالبا ما تضاعف أوجاع المرأة بدلا من شفائها. ومن أسباب الألم العميق، التهاب القنوات وبطانة الرحم، وتشمل امتداد بطانة الرحم خارج تجويف الرحم. هناك بعض العمليات الجراحية التي تستأصل عنق الرحم وتقصر المهبل فيسبب أوجاع أثناء الجماع في وقت لاحق.أما إذا أجريت العملية بدون المساس بعنق الرحم والمهبل فلا تؤثر كثيرا في الممارسة الجنسية. الأسباب النفسية ناجمة في الغالب عن ذكريات، أو ممارسات أليمة، أو عن جهل مطبق بالعملية الجنسية، ويمكن التغلب عليها بسهولة، إذا تغلب الطرفين على المخاوف المسبقة.

Dialogue: - Speaker A: كيف حالك هذه الأيام؟ آمل أن تكوني بخير.

- Speaker B: شكراً لك. أنا بخير، لكنني أواجه بعض الألم والصعوبة أحياناً.
- Speaker A: أتفهم. هل هذا الألم مصحوب بجفاف، أم أنه شعور بالصعوبة فقط؟
- Speaker B: نعم، الألم والجفاف يترافقان أحياناً، وقد يكون ناجماً عن التهاب ما.
- Speaker A: هذا قد يشير إلى أسباب نفسية، أو ربما يحتاج الأمر وقتاً للشفاء بعد شيء ما.
- Speaker B: أبالتأكيد. المعالجة الفورية والفهم العميق للأسباب يبدوان ضروريين.

Reason: The conversation is irrelevant to the given text, which discusses specific medical causes and conditions related to dyspareunia, while the dialogue is more of a general, non-specific chat about pain and potential psychological causes, without much relevance to the detailed medical context.

H.3 Maghrebi Darija

H.3.1 Rating A

لاندون دونوفان (بالإنجليزية: Landon Donovan) هو لاعب كرة قدم أمريكي سابق، ولد 4 مارس 2891 في مدينة أونتاريو بولاية كاليفورنيا الأمريكية، يلعب لمنتخب أمريكا لكرة القدم، وفي أغسطس 9002 أصيب دونوفان بإنفلونزا الخنازير. وفي يناير 0102 انتقل الاعب عن طريق الإعارة إلى نادي إيفرتون الإنجليزي و مع المنتخب الوطني للولايات المتحدة، دونوفان هو هداف المنتخب، و أكتر من صنع أهداف (الاسيست), وهو أكثر لاعب مثل المنتخب الأمريكي لكرة القدم من اللاعبين النشطيين حالياً. حصل ولأربع مرات بجائزة رياضي لكرة القدم في الولايات المتحدة للسنة (للفوز في سنوات متتالية) ، فضلا عن ثلاثة مرات جائزة لاعب هوندا للسنة. وهو صاحب أكثر أهداف في كأس العالم بين اللاعبين الأمريكان، وثالث لاعب أمريكي ليسجل في أكثر من نهائيات كأس عالم واحدة (بعد بريان مأكبرايد وكلينت ديمبسي).

Dialogue: - Speaker A: ياك سمعتي على واحد اللاعب سميتو لاندون دونوفان؟.

- Speaker B: أه وي! هاداك راه لاعب كرة قدم أمريكي معروف بزاف.

- Speaker A: بما أنه معروف، واش كان عندو شي إنجازات كبيرة فالمسيرة ديالو؟

- Speaker B: أكيد! راه هو الهداف التاريخي للمنتخب الأمريكي وصانع الأهداف.،

- Speaker A: واو، هادشي زوين! وشحال من مرة واش ربح جائزة أفضل لاعب؟

- Speaker B: ربحها أربع مرات، وزيد عليها سجل بزاف د الأهداف فكأس العالم.

Reason: The conversation is fully correct in Maghrebi Darija without any errors or slips. The grammar, vocabulary, idioms, and expressions are accurate and appropriate. The dialogue flows naturally and coherently with smooth transitions and contextual expressions. It's mainly in Maghrebi Darija with no awkward or non-idiomatic phrases.

H.3.2 Rating B

اتفاق باريس (بالفرنسية: Accord de Paris) أو «كوب 12» هو أول اتفاق عالمي بشأن المناخ. جاء هذا الاتفاق عقب المفاوضات التي عقدت أثناء مؤتمر الأمم المتحدة 12 للتغير المناخي في باريس في 5102. حسب لوران فابيوس الذي قدم مشروع الاتفاق النهائي في الجلسة العامة، فإن هذا الاتفاق مناسب ودائم ومتوازن وملزم قانونيا. صدق على الاتفاق من قبل كل الوفود 591 الحاضرة في 21 ديسمبر 5102 في 62:91 CTU+00:10. يهدف الاتفاق إلى احتواء الاحترار العالمي لأقل من 2 درجات وسيسعى لحده في 5.1 درجة. سيتم إعادة النظر في الأهداف المعلنة بعد خمس سنوات، وأهداف خفض الانبعاثات لا يمكن استعراضها على نحو أعلى. وضع كحد أدنى قيمة 001 مليار دولار أمريكي كمساعدات مناخية الدول النامية سنويا وسيتم إعادة النظر في هذا السعر في 5202 على أقصى تقدير.بمناسبة يوم الأرض الذي يتم الاحتفال به في 22 أبريل، وقع 571 من رؤساء دول العالم في عام 6102 في مقر الامم المتحدة في نيويورك تحت مسمي اتفاقية باريس للتغير المناخي وكان ذلك الحدث الأكبر علي الإطلاق لاتفاق عدد كبير من البلدان في يوم واحد أكثر من أي وقت مضي.

Dialogue: - Speaker A: قريت اليوم على اتفاق باريس ديال المناخ، كتعرفو؟.

- Speaker B: آه، داكشي لي خرج من كوب 12 ياك؟ عقلت عليه.
- Speaker A: إيه، بصح. هو أول اتفاق عالمي فالمناخ كيوقع.
- Speaker B: آه، و قالو بلي اتفاق وازن و ملزم قانونياً.
- Speaker A: هدفهم الأساسي يحدوا من الاحترار العالمي تحت جوج درجات.
- Speaker B: و حتى دوك المساعدات ديال 001 مليار دولار للدول النامية.

Reason: The conversation is generally correct in Maghrebi Darija and contains no significant grammatical or vocabulary errors. However, it feels more like a Q&A format with less fluidity and idiomatic transitions. The dialogue is mostly a collection of disconnected sentences rather than a fluid conversation.

H.3.3 Rating C

نادي السد الرياضي هو نادي قطري متعدّد الرياضات، تأسس في 12 أكتوبر سنة 9691، مقره العاصمة القطرية الدوحة، ويلعب في دوري نجوم قطر (أعلى دوري في قطر). سمي النادي بهذا الاسم نسبة إلى منطقة السد التي يقع بها مقر النادي. يعدّ نادي السد من أنجح الأندية القطرية على الإطلاق، حيث يتزعّم جميع البطولات المحلية بأكبر عدد من الألقاب، فاز بجميع البطولات المحلية، وحقق بطولة كأس الشيخ جاسم 41 مرة وبطولة كأس قطر (كأس ولي العهد) 6 مرات وكأس الأمير 61 مرة وبطولة دوري نجوم قطر 4 مرات وكأس نجوم قطر مرة واحدة وكأس الاتحاد 7 مرات.يتألف شعار النادي من اللونين الأبيض والأسود ويلقب بـ«عيال الذيب»، كما يُطلق عليه لقب «الزعيم».

Dialogue: - Speaker A: نادي السد القطري لي ف الدوحة، راه تأسس ف9691 ياك؟.

- Speaker B: أه بصّاح! هو فمنطقة السد بالضبط، ومن أنجح الأندية القطرية على الإطلاق.

- Speaker A: و بصّاح! كيتسمّاو بـ'الزعيم' و'عيال الذيب' ياك؟

- Speaker B: إيه! و ربحو كأس الأمير 61 مرة، وكأس الشيخ جاسم 41 مرة.

- Speaker A: واو! و ربحو حتى دوري نجوم قطر 4 مرات. تبارك الله عليهم!

- Speaker B: بصّاح! ديما كيتصدروا البطولات المحلية وشعارهم أبيض وكحل.

Reason: The conversation is mostly correct in Maghrebi Darija. However, it contains some non-native expressions, like 'بصّاح' instead of the more common Moroccan Darija 'صحيح' or 'بالصح', which is more typical of Algerian Darija. These small slips make it slightly awkward for Moroccan Darija.