

# When Retrieval Can Hurt: A Multi-Role 2×2 Human Evaluation of Retrieval and Self-Refinement for Arabic Multiple-Choice Question Generation

Abderrahman AZI<sup>1\*</sup>, Abdessamad JARRAR<sup>2</sup>, Moulay AMZIL<sup>3</sup>, Ahmed EL GHAZI<sup>4</sup>

Faculty of Sciences and Techniques, Sultan Moulay Slimane University, Beni-Mellal, Morocco<sup>1,3,4</sup>

Faculty of Sciences and Techniques, Hassan I University, Settat, Morocco<sup>2</sup>

**Abstract**—Retrieval-Augmented Generation (RAG) and iterative self-refinement are often added to large language model (LLM) generation pipelines, but their contribution to the quality of Arabic educational items remains uncertain. This study evaluates Modern Standard Arabic multiple-choice question (MCQ) generation in a road-safety domain with a complete 2×2 design crossing retrieval and refinement. One item was generated for each of 31 topics under four conditions, and six raters (students, driving instructors, and licensing examiners) evaluated all 124 items on six criteria, yielding 744 blinded ratings. In a mixed-effects model with crossed random effects for evaluator and topic, retrieval was associated with lower composite ratings ( $\beta = -0.94$  on a 6–30 scale, 95% CI  $[-1.45, -0.44]$ ,  $p = .0003$ ; standardized  $d = -0.28$ ). The decrease was small, about 0.16 Likert points per criterion, but persisted with evaluator-specific random slopes and after adjustment for item length. Question clarity decreased most consistently. Refinement showed no detectable effect ( $\beta = -0.07$ , 95% CI  $[-0.57, 0.43]$ ) and was statistically equivalent to zero within  $\pm 1$  composite point. ROUGE-L and BERTScore computed between conditions showed that retrieval-grounded items differed substantially in wording from baseline items, but similarity scores do not indicate the direction of quality change. Two domain experts judged the retrieved passages predominantly relevant and all audited retrieval-plus-refinement answers correct. The findings are bounded to one Llama-family model, one corpus, and one fixed top-k retrieval configuration; they indicate that retrieval should be tested empirically with human evaluation before being adopted by default for Arabic MCQ generation.

**Keywords**—Arabic NLP; multiple-choice question generation; retrieval-augmented generation; self-refine; human evaluation

## I. INTRODUCTION

The automatic generation of assessment content has emerged as an increasingly important area within educational natural language processing (NLP). Multiple-choice questions (MCQs) in particular play a central role in standardised testing, professional certification, and large-scale e-learning, where the manual authoring of high-quality items is costly and slow. Recent developments in large language models (LLMs) have heightened interest in automating this process, promising the scalable production of examination content from existing textual resources [1, 2]. For languages that are comparatively under-resourced in NLP, such as Arabic, the potential impact is considerable: reliable automatic MCQ generation could substantially reduce the burden of producing assessment

material for large numbers of learners. This potential, however, depends on whether the generation techniques developed and validated for high-resource languages transfer effectively to Arabic educational settings — a question that remains largely open.

Two techniques are now routinely combined in state-of-the-art generation pipelines. Retrieval-Augmented Generation (RAG) conditions the model on passages retrieved from an external corpus, and often improves factual grounding in knowledge-intensive question answering and generation; however, this benefit may depend on retrieval quality, task structure, and the model’s prior competence, and it has not been established for the generation of assessment items [3]. Iterative self-refinement, exemplified by the Self-Refine framework, prompts a model to repeatedly critique and revise its own output, and has been reported to yield consistent gains across a range of tasks [4]. The prevailing assumption in the literature is that both techniques are broadly beneficial and that their benefits transfer across languages and domains. This assumption, however, has rarely been subjected to rigorous human evaluation, and rarely for Arabic educational content. A systematic understanding of how retrieval and refinement contribute to the quality of generated Arabic MCQs is therefore still lacking.

A second and related gap concerns evaluation itself. Much of the generation literature relies on automatic similarity metrics, such as ROUGE-L and BERTScore, which score a generated text against a reference text as proxies for quality. Yet these metrics measure textual overlap and semantic similarity rather than the pedagogical validity of an item, and their agreement with expert human judgment in the context of Arabic MCQ generation has not been established. What is not yet clear is whether a pipeline that appears successful by automatic metrics also produces questions that domain experts regard as high in quality. Where these two forms of evidence diverge, the choice of evaluation method could determine the conclusion drawn — a possibility with direct consequences for how the field validates generation systems.

A separate line of work uses large language models themselves as evaluators (LLM-as-a-judge); such methods can align with human ratings on some tasks but exhibit biases, including a tendency to favour LLM-generated text [17]. LLM-based judging is not used here — it would be circular when the object of study is an LLM generator — but it is noted as a possible secondary evaluation in future work, to be validated

against expert human ratings rather than used as a substitute for them.

The present study addresses these gaps directly. Working in Modern Standard Arabic (MSA) on a road-safety knowledge domain, it constructs a complete  $2 \times 2$  factorial pipeline crossing retrieval (on/off) with iterative refinement (on/off) and evaluates it with a purposive multi-role human panel spanning the three stakeholders of a licensing examination: candidates, instructors, and official examiners. Contrary to the common expectation, retrieval grounding did not improve MCQ quality; it was associated with a small but statistically reliable decrease, and iterative refinement produced no measurable effect. ROUGE-L and BERTScore, computed between the outputs of different conditions, showed that retrieval-grounded items differed substantially in wording from the baseline items while remaining topically related; such between-condition similarity scores describe how much the items changed, but not whether they improved or degraded, which only the human ratings revealed.

The study addresses three research questions:

RQ1. Does retrieval grounding improve the human-rated quality of generated Arabic MCQs, and does the effect differ by evaluator role?

RQ2. Does iterative self-refinement improve quality beyond retrieval alone?

RQ3. Can between-condition similarity metrics (ROUGE-L, BERTScore) indicate the quality differences perceived by human raters?

The contributions of this study are: (i) a complete  $2 \times 2$  factorial evaluation isolating retrieval and refinement for Arabic MCQ generation; (ii) a purposive multi-role human protocol with two raters per stakeholder role, released as a reusable resource; (iii) a bounded negative result under the evaluated configuration — retrieval was associated with a small but statistically reliable reduction in ratings, and refinement showed no detectable improvement, for one model, corpus, domain, and retrieval setting — a finding of the kind that is systematically under-reported in the generation literature yet essential for calibrating expectations of these techniques; and (iv) a case in which between-condition similarity metrics registered substantial textual change but not its direction, illustrating why such metrics should not replace human evaluation of generated Arabic assessment content.

## II. RELATED WORK

### A. Automatic MCQ Generation

Automatic question generation has received considerable critical attention and has been surveyed extensively [2]. Early systems relied largely on rule-based transformations and syntactic templates, which offered control but limited fluency and generalisability. The advent of transformer architectures and, more recently, large language models has shifted the field decisively toward neural generation, with prompt-based and instruction-tuned methods now dominant [1]. A growing body of work has applied these models to domain-specific MCQ generation from regulatory, medical, and educational corpora,

reporting that LLMs can produce pedagogically plausible items at scale.

For Arabic specifically, however, question generation remains comparatively underdeveloped. The morphological richness and orthographic characteristics of Arabic pose well-documented challenges for NLP, and much of the existing work on Arabic question generation has relied on automatic or rule-based measures rather than expert human judgment [2]. More recent research has begun to treat question quality as a first-class, measurable objective, combining automatic evaluation with expert ratings of question relevance and quality [11], although such quality-centred human evaluation remains rare for Arabic. The present study extends this quality-centred direction to Arabic in two respects: it employs a multi-role panel rather than a single expert type, and it directly contrasts human judgment with automatic metrics rather than treating them as interchangeable. Native Arabic benchmarks such as ArabicMMLU show that Arabic MCQs are a demanding testbed and that strong models remain imperfect on them, motivating careful Arabic evaluation, though such benchmarks measure model answering rather than generation quality [12]. Recent Arabic question-generation work reports transformer systems evaluated largely with overlap metrics on open-ended question generation, a task and evaluation setting distinct from multi-role MCQ quality assessment [14]; and LLM-based MCQ generation with automatic LLM-based validation has been studied in regulatory domains, which the present study contrasts with expert human evaluation [13].

### B. Retrieval-Augmented Generation

Retrieval-Augmented Generation, introduced by Lewis et al. [3], conditions a generative model on passages retrieved from an external document store, and has been instrumental in reducing hallucination and improving factual grounding across knowledge-intensive tasks. The technique has been widely adopted in educational applications, where grounding in authoritative source material is particularly attractive. The bulk of the supporting evidence, however, concerns question answering rather than question generation, and English-language settings rather than Arabic ones.

This distinction matters. In question answering, retrieval supplies facts the model may lack; in question generation, its role is less clear, because a competent model may already be able to produce well-formed items without external support. Whether retrieval grounding improves the quality of generated Arabic questions — particularly in a domain the base model already handles well — has not been systematically established. The generalisability of the positive RAG findings from English question answering to Arabic question generation therefore remains an open and empirically untested question, which the present factorial design is constructed to address.

Recent surveys further decompose RAG into distinct pre-retrieval, retrieval, post-retrieval, and generation stages, each of which — including chunking, indexing, passage relevance, reranking, and filtering — can materially affect output quality [15, 16]. The present study evaluates one specific, widely used RAG implementation (AraBERT embeddings with a flat FAISS index and fixed top-k retrieval, without reranking or relevance filtering); it therefore tests this configuration rather than retrieval

augmentation in the abstract, and does not apply retrieval-faithfulness diagnostics such as those proposed for RAG evaluation [21]. Recent Arabic work shows that noisy or partially relevant passages can degrade multiple-choice performance and that candidate-aware reranking can help [19], underscoring that a fixed-k pipeline without relevance control is not a best-case RAG condition.

### C. Iterative Self-Refinement

Iterative self-refinement has emerged as a training-free strategy for improving LLM output. The Self-Refine framework [4] prompts a model to generate feedback on its own output and then revise it, reporting average improvements across a range of English tasks. The appeal of the approach is that it requires no additional data or fine-tuning, making it readily applicable to low-resource settings. Its behaviour in Arabic, however, and in particular its interaction with retrieval grounding, remains unexamined.

Two questions are especially pertinent. First, it is unclear whether self-refinement produces genuine quality gains in a morphologically rich language, or whether it merely paraphrases without improving. Second, it is unknown whether refinement contributes independently of retrieval, or whether the two interact. Because prior work has typically applied refinement only in combination with other components, its independent effect has been difficult to isolate. The present factorial design addresses this directly by applying refinement both with and without retrieval, allowing its independent contribution and any interaction to be estimated.

### D. Automatic Versus Human Evaluation

The evaluation of generated text has long been a contested issue. Reference-based similarity metrics such as ROUGE-L [5] and BERTScore [6], which score a generated text against a reference text, are convenient, reproducible, and inexpensive, and are consequently widely used in generation studies. It is well documented, however, that such metrics correlate only imperfectly with human judgments of quality [1], since they capture lexical overlap and semantic similarity rather than the pedagogical or communicative adequacy of a text. For MCQ generation, where quality depends on subtle properties such as the plausibility of distractors and the clarity of the stem, the adequacy of automatic metrics is particularly doubtful.

Despite these known limitations, automatic metrics continue to serve as the primary or sole evaluation in many generation studies, partly because rigorous human evaluation is expensive to organise. This reliance is problematic when similarity metrics are read as quality signals, because they carry no information about whether a change improved or degraded an item. The present study provides a concrete case in which substantial textual difference coincided with lower human-rated quality in Arabic MCQ generation and, in doing so, reinforces the argument that human evaluation remains indispensable in high-stakes assessment contexts.

## III. METHOD

### A. Corpus and Data

The generation corpus consisted of Modern Standard Arabic road-safety instructional material — lesson texts covering traffic

regulations, road signs, and related driving topics — drawn from official Moroccan driving-education lesson PDFs covering the licensing curriculum and organised into 31 topics. For each source, a stable identifier, the originating lesson file and page, jurisdiction (Morocco), language (MSA), and approximate length were recorded. All text was normalised (Unicode normalisation, whitespace and markup cleanup) and indexed at page level, with each PDF page forming one retrieval unit (referred to below as a passage). The source lesson text is third-party copyrighted material used for non-commercial academic research and is not redistributed; the data release comprises the model-generated items and derived data (see Data availability). Because road-safety regulations are jurisdiction- and version-dependent, the factual reference frame is that of the Moroccan licensing curriculum at the time of collection.

### B. Generation Pipeline and the 2×2 Design

Fig. 1 presents the overall generation and evaluation pipeline. Two binary factors are crossed — retrieval (off/on) and iterative refinement (off/on) — giving four conditions. 1) Baseline: the LLM generates one MCQ per topic with no retrieval and no refinement. 2) Refine-no-RAG: the baseline question is passed through three Self-Refine iterations without retrieved context. 3) RAG t0: the model generates grounded in the top-k ( $k=3$ ) retrieved passages, without refinement. 4) RAG + refine (t3): the grounded question is passed through three Self-Refine iterations. Generation used the llama-3.3-70b-versatile model accessed through the Groq API, with a generation temperature of 0.7 and a refinement temperature of 0.5, and no fixed random seed (so exact regeneration is subject to provider-side sampling variability); embeddings used AraBERT [7] and retrieval used FAISS [8]. Refinement was guided by Arabic-specific criteria covering morphological correctness, MSA fluency, and, in the retrieval branch, factual consistency with the retrieved passages; the change-notes preamble emitted by the model was stripped so that all four conditions were presented in identical format, preserving blinding.

1) *Prompt differences between conditions*: The factors were not manipulated through prompt context alone. The retrieval-grounded generation prompt (Appendix A.1) inserts the retrieved pages and asks for an explanation grounded in the lesson content, whereas the baseline prompt asks only for a question on the topic in the same output format. In refinement, the sixth check of the critique prompt (Appendix A.2) is content consistency with the retrieved pages in the retrieval branch and explanation clarity in the non-retrieval branch. The retrieval factor therefore denotes the full retrieval-grounded configuration, retrieved context together with its grounding instructions, rather than the inserted context in isolation, and the retrieval × refinement interaction compares refinement under two rubrics that differ in one of six checks. Separating the contribution of the context from that of the instructions would require prompts that differ only in the inserted context, which is left to future work (Section V-E).

A Llama-family model [20] was selected as a strong, reproducible open-weight-family choice available through the selected provider at the time of study; it is not claimed to be the optimal model for Moroccan road-safety Arabic, which was not

benchmarked. This choice serves reproducibility and scientific control: it allows other researchers to audit the full pipeline without reliance on closed, black-box commercial systems whose underlying weights may change without notice and

whose terms often preclude the kind of controlled, isolated factorial analysis presented here. Whether the negative result observed here extends to larger or proprietary models is an empirical question left to future work.

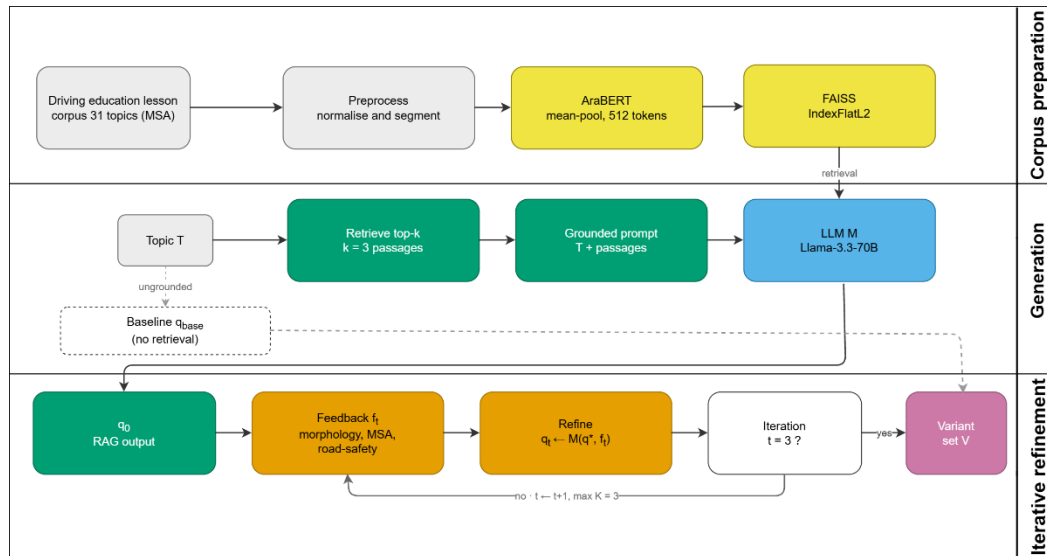


Fig. 1. The generation and evaluation pipeline. Corpus passages (31 topics) are normalised, embedded with AraBERT, and indexed with FAISS. For each topic, the baseline generates without retrieval, while the RAG conditions retrieve the top  $k = 3$  passages to form a grounded prompt for the generator (Llama-3.3-70B); refinement conditions apply a fixed sequence of three feedback-and-revise iterations.

### C. Multi-Role Human Evaluation

Six evaluators were recruited by purposive sampling to represent the three stakeholders of the licensing examination, two per role: candidates (Student), certified driving instructors (Instructor), and official licensing examiners (Examiner). Each evaluator rated all 31 topics in all four conditions — 124 items per evaluator, 744 ratings in total — on six Likert criteria (1–5): factual accuracy, question clarity, distractor quality, difficulty level, Arabic fluency, and explanation clarity. The panel size reflects a design choice: rather than many raters each judging a subset of items, a small panel of stakeholder representatives each rated every item in every condition, so that every retrieval and refinement contrast is estimated within each rater. This maximises the precision of within-rater contrasts but limits generalisation across raters and precludes confirmatory comparison between roles; the consequences of this choice are examined in Sections IV-A and IV-B. A composite score was formed by summing the six criteria (range 6–30). The six criteria (factual accuracy, question clarity, distractor quality, difficulty, Arabic fluency, and explanation clarity) follow established multiple-choice item-writing guidance [22]; the full Arabic rubric with anchors for scores 1, 3, and 5, the rater instructions, and the interface text are provided in the supplementary material (Appendix).

1) *Blinding and randomisation*: Evaluation was delivered through a custom web application. Condition identity was hidden, and the four versions of every topic were scattered independently through each evaluator’s session by a per-evaluator deterministic shuffle, so that raters judged each question in isolation rather than by direct comparison, and any fatigue was spread evenly across conditions rather than tied to

one. The order was stable across sessions, allowing evaluators to pause and resume. Because every evaluator rated every condition of every topic, the design is a complete within-subject factorial, licensing repeated-measures analysis.

### D. Automatic Metrics

Two metrics were computed for each condition pair: ROUGE-L (longest-common-subsequence F1) over Arabic-normalised tokens (removal of diacritics and tatweel; alef, ya, and ta-marbuta normalisation), and BERTScore F1 using the multilingual bert-base-multilingual-cased model with the language set to Arabic and without baseline rescaling (the library default); unrescaled multilingual-BERT scores are compressed toward the upper range, so values around 0.7 indicate only moderate similarity. These measure, respectively, lexical difference and semantic similarity between conditions; neither measures pedagogical quality, and low ROUGE-L is expected when two items differ substantially in wording and is not intrinsically indicative of lower quality [18].

### E. Statistical Analysis

Because the 744 ratings are not independent — they are clustered within six evaluators and 31 topics — the confirmatory analysis is a linear mixed-effects model, which correctly propagates this dependency structure. The composite score is modelled as a function of retrieval, refinement, and their interaction (fixed effects, treatment coding with the baseline condition as the reference cell), with crossed random intercepts for evaluator and topic. The model was fitted by REML using the statsmodels implementation; coefficient estimates, 95% confidence intervals, and variance components are reported, and fits were checked for singularity. The primary estimand is the

average within-rater condition contrast over the six sampled raters and 31 topics; because the evaluator factor has only six levels, population-level generalisation is treated as limited. To check that this inference is not anti-conservative, the model was also refitted with evaluator-specific random slopes for retrieval and refinement and with topic-specific random slopes for retrieval, and the retrieval effect was re-tested on per-evaluator ( $n = 6$ ) and per-topic ( $n = 31$ ) aggregated contrasts, which treat raters and items, rather than ratings, as the independent units. The random-intercept model is the primary analysis. The Wilcoxon signed-rank and Friedman tests reported alongside are transparent nonparametric sensitivity analyses; for these, effect size  $r = Z / \sqrt{N}$  is computed over non-tied pairs following Rosenthal [9], with magnitude interpreted against Cohen's conventions [10], with  $N = 186$  evaluator  $\times$  topic pairs, and per-criterion Wilcoxon tests use Holm correction across the six criteria, a family defined before analysis.  $N = 186$  denotes six evaluators  $\times$  31 topics, not 186 independent participants; because the Wilcoxon tests treat these clustered cells as independent pairs, they are reported as descriptive checks, and inference rests on the mixed-effects models and the aggregated contrasts. Normality was rejected by Shapiro–Wilk tests for every condition (all  $p < .01$ ), motivating the nonparametric sensitivity analyses. The retrieval and refinement main effects are designated as the primary (confirmatory) analyses, the per-criterion and interaction effects as secondary, and the per-role comparisons as exploratory. Inference is based on estimated effect sizes and their confidence intervals rather than on achieved-power calculations. All analyses were performed in Python using SciPy and statsmodels.

#### F. Implementation and Reproducibility

For transparency, the full computational configuration is detailed here. Question generation and self-refinement were performed by a Meta Llama-family instruction-tuned model served through the Groq API; the model identifier was llama-3.3-70b-versatile. An earlier endpoint, meta-llama/llama-4-maverick-17b-128e-instruct, was deprecated by the provider during the study and is not used in the reported results. Each item was produced by a single chat completion with the prompt supplied as one user-role message (no system message) and a maximum of 1000 output tokens. Initial generation used a sampling temperature of 0.7 to balance lexical variety and fidelity, and the refinement passes used 0.5 to favour conservative, targeted revision; all other decoding parameters were left at their API defaults, and no fixed random seed was set. The refinement loop ran for a fixed three iterations per item with no convergence or early-stopping criterion, so every refined item is the output of exactly three sequential feedback-and-revise passes. A two-second pause separated successive API calls, and requests that hit provider rate limits were retried with exponential back-off. Because the model was accessed remotely at temperature 0.7 without a seed, each condition's item is a single stochastic realisation rather than a deterministic function of the treatment; the estimand is the effect of the pipeline under provider sampling, and exact regeneration is not claimed. Generating multiple independent completions per topic and condition is a worthwhile extension that was not run here (Section V-E).

Dense retrieval used AraBERT (aubmindlab/bert-base-arabertv02) [7], applied off the shelf without retrieval-specific fine-tuning. Each PDF page was embedded as a single vector by mean-pooling the encoder's final hidden states, with inputs truncated to the 512-token limit, and the page vectors were indexed with FAISS [8] using a flat L2 index (IndexFlatL2). The retrieval query was the Arabic topic title (e.g., التقابل والتجاوز, “meeting and overtaking”), embedded in the same way; the  $k = 3$  pages nearest to the query by exact (brute-force) L2 search were concatenated, separated by blank lines, to form the retrieval context inserted into the generation prompt. A value of  $k = 3$  was used as a common default that balances sufficient grounding against the risk of introducing irrelevant passages;  $k$  was not tuned on the study data, and a  $k$ -sensitivity analysis remains future work (Section V-E). Automatic metrics were computed with a longest-common-subsequence ROUGE-L over Arabic-normalised tokens (removal of diacritics and tatweel, alef/ya/ta-marbuta normalisation) and BERTScore F1 using a multilingual BERT model.

The embedding, indexing, and metric computations were run on a desktop workstation (Intel Core i5-4590S CPU at 3.0 GHz, 4 cores, 16 GB RAM, with an NVIDIA GeForce GTX 1050 Ti GPU, 4 GB VRAM), with LLM generation handled remotely through the Groq API. Generation and evaluation were performed using Python 3.11 with transformers 5.2.0, sentence-transformers 5.2.0, faiss-cpu 1.15.1, numpy 2.0.2, scipy 1.16.3, statsmodels 0.14.6, krippendorff 0.8.2, bert-score 0.3.13, rouge-score 0.1.2, torch 2.6.0, and groq 1.7.0. No fixed random seed was set for generation. The generated items, anonymised ratings, and the metric-computation and figure scripts are available in the repository (see Data availability).

## IV. RESULTS

### A. Effect of Retrieval

1) *Primary analysis (mixed-effects factorial)*: Table I reports the mixed-effects model with crossed random effects for evaluator and topic. The retrieval main effect was negative and statistically significant after accounting for clustering ( $\beta = -0.94$ , 95% CI  $[-1.45, -0.44]$ ,  $p = .0003$ ). The refinement main effect was negligible ( $\beta = -0.07$ ,  $p = .79$ ), and the retrieval  $\times$  refinement interaction was not significant ( $\beta = +0.20$ ,  $p = .57$ ), so refinement did not recover the retrieval decrease. The evaluator random-effect variance (4.66) substantially exceeded the topic variance (0.49), indicating that raters differed markedly in their overall scoring level. The estimated marginal means coincide with the observed cell means.

2) *Effect size and practical significance*: Expressed as a standardized effect, the retrieval coefficient corresponds to  $d = -0.28$  when divided by the total model-implied standard deviation of the composite ( $\sqrt{4.66 + 0.49 + 6.15} = 3.36$ ), and to  $d = -0.38$  relative to the residual standard deviation alone; both fall in the small range by conventional benchmarks [10]. On the original scale,  $-0.94$  composite points correspond to about 0.16 Likert points per criterion, or about 4% of the 24-point range of the composite, and the confidence interval spans roughly 0.07–0.24 Likert points per criterion. Part of this interval lies within the  $\pm 1$ -point margin used for the refinement

equivalence test (Section IV-C), and an equivalence test against that margin was not significant (TOST  $p = .41$ ). The data therefore establish neither that the retrieval effect is negligible nor that it exceeds one composite point. The retrieval effect is best described as statistically reliable but small: retrieval did not improve quality and lowered ratings slightly, rather than producing a large practical loss.

3) *Robustness to rater and item heterogeneity*: Because raters disagreed substantially (Section IV-B), the primary model was refitted with evaluator-specific random slopes for retrieval and refinement. The retrieval estimate was unchanged ( $\beta = -0.94$ ), but its uncertainty widened (95% CI  $[-1.62, -0.26]$ ,  $p = .007$ ). The between-rater variance of the retrieval slope was small (0.33) relative to the intercept variance (5.33), and the refinement slope variance was estimated at zero. Adding topic-specific random slopes for retrieval widened the interval further (95% CI  $[-1.72, -0.16]$ ,  $p = .019$ ), indicating that the retrieval effect varied more across topics (slope variance 1.27) than across raters. The refinement effect remained null in both models ( $\beta = -0.07$ ,  $p = .78$ ). As a further check that treats raters and items, rather than ratings, as the independent units, the retrieval effect averaged over both refinement levels ( $-0.84$  composite points) was computed for each rater and each topic. It was negative for five of the six raters ( $t(5) = -2.72$ ,  $p = .042$ , 95% CI  $[-1.63, -0.05]$ ) and for 23 of the 31 topics ( $t(30) = -2.68$ ,  $p = .012$ ; Wilcoxon  $p = .002$ ). The retrieval effect is therefore not an artefact of the random-intercept specification, although its  $p$ -value under the fuller models (.007–.019) is less extreme than under the primary model ( $p = .0003$ ), and the six-rater panel limits generalisation to wider rater populations.

TABLE I. PRIMARY LINEAR MIXED-EFFECTS MODEL: COMPOSITE SCORE  $\sim$  RETRIEVAL  $\times$  REFINEMENT, WITH CROSSED RANDOM INTERCEPTS FOR EVALUATOR (VARIANCE 4.66) AND TOPIC (VARIANCE 0.49); RESIDUAL VARIANCE 6.15. EMM = ESTIMATED MARGINAL MEAN

| Effect                        | $\beta$<br>(estimate) | 95% CI         | p      |
|-------------------------------|-----------------------|----------------|--------|
| Intercept (baseline EMM)      | 23.32                 | [21.54, 25.10] | <0.001 |
| Retrieval (main effect)       | -0.94                 | [-1.45, -0.44] | 0.0003 |
| Refinement (main effect)      | -0.07                 | [-0.57, +0.43] | 0.786  |
| Retrieval $\times$ Refinement | +0.20                 | [-0.51, +0.92] | 0.574  |

TABLE II. SENSITIVITY ANALYSIS: COMPOSITE SCORE (6–30) BY CONDITION WITH WILCOXON SIGNED-RANK TESTS AGAINST THE BASELINE (N = 186 EVALUATOR  $\times$  TOPIC PAIRS PER CONTRAST; PAIRS ARE CLUSTERED WITHIN RATERS AND TOPICS, SO THESE TESTS ARE DESCRIPTIVE)

| Condition         | Mean $\pm$ SD    | $\Delta$ vs base | W      | p          | r     |
|-------------------|------------------|------------------|--------|------------|-------|
| Baseline          | 23.32 $\pm$ 3.00 | —                | —      | —          | —     |
| Refine-no-RAG     | 23.25 $\pm$ 3.50 | -0.07            | 5128.5 | 0.968 (ns) | 0.003 |
| RAG t0            | 22.38 $\pm$ 3.36 | -0.94            | 4157.5 | 0.002 **   | 0.249 |
| RAG + refine (t3) | 22.52 $\pm$ 3.10 | -0.81            | 4352.0 | 0.005 **   | 0.228 |

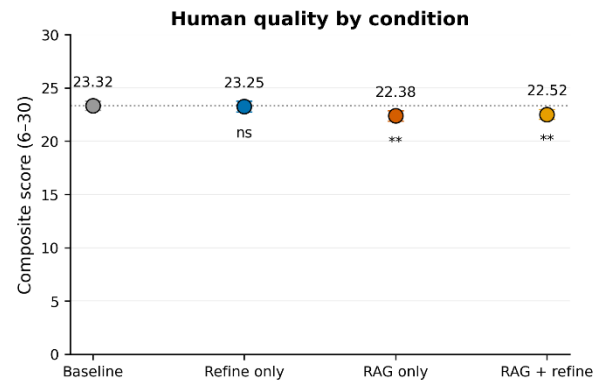


Fig. 2. Human quality by condition: composite score (6–30) with 95% confidence intervals; significance shown against the baseline (Wilcoxon; \*\*  $p < .01$ ), computed over 186 evaluator–topic pairs per contrast that are clustered within raters and topics, so the markers are descriptive (inference in Table I). Points are estimated means; the dotted line marks the baseline.

Fig. 2 shows the composite human-quality scores for the four conditions with 95% confidence intervals. As can be seen from the figure, retrieval grounding reduced composite quality relative to the ungrounded baseline. The ungrounded baseline scored 23.32, whereas RAG t0 scored 22.38 ( $\Delta = -0.94$ ,  $p = 0.002$ ,  $r = 0.249$ ) and RAG + refine 22.52 ( $\Delta = -0.81$ ,  $p = 0.005$ ,  $r = 0.228$ ; Table II). The direction is the opposite of the common expectation that retrieval improves generation quality.

TABLE III. EXPLORATORY PER-ROLE COMPARISON, BASELINE VS. RAG + REFINEMENT (N = 62 PAIRS PER ROLE); CONFIDENCE INTERVALS INDICATIVE GIVEN TWO RATERS PER ROLE

| Role       | Baseline | RAG+refine | $\Delta$ | p          | r     |
|------------|----------|------------|----------|------------|-------|
| Student    | 25.08    | 24.19      | -0.89    | 0.267 (ns) | 0.164 |
| Instructor | 22.34    | 21.76      | -0.58    | 0.116 (ns) | 0.220 |
| Examiner   | 22.55    | 21.60      | -0.95    | 0.050 *    | 0.262 |

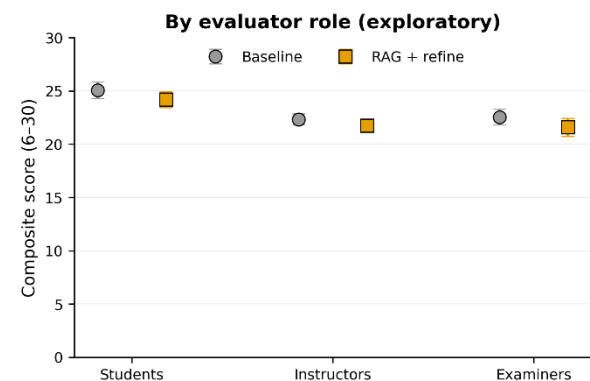


Fig. 3. Exploratory per-role comparison: baseline vs. RAG + refine composite means by evaluator role, with 95% confidence intervals (62 evaluator–topic observations per role).

As illustrated in Fig. 3, the reduction held across all three evaluator roles (Table III). The role-level analysis is exploratory: with only two evaluators per role and low within-role agreement, the study was not powered to compare stakeholder groups. The confidence intervals reflect within-panel variability rather than population-level role uncertainty, so these estimates are indicative rather than confirmatory, and no

formal role $\times$ condition interaction test is reported. The reduction was numerically largest for the examiner ( $\Delta = -0.95$ ); the associated  $p = 0.050$  is borderline and would not survive correction for the three role-level tests, and it is therefore not treated as definitive evidence of a stakeholder difference. No role showed any improvement under retrieval.

TABLE IV. PER-CRITERION CHANGE, BASELINE  $\rightarrow$  RAG + REFINE, WITH HOLM-ADJUSTED PAIRED WILCOXON TESTS ACROSS THE SIX CRITERIA. \*  $P_{\text{ADJ}} < .05$ , \*\*  $P_{\text{ADJ}} < .01$

| Criterion          | Base | RAG+ref | $\Delta$ | $P$ (raw) | $p$ (adj)  |
|--------------------|------|---------|----------|-----------|------------|
| Factual accuracy   | 4.16 | 4.01    | -0.16    | 0.019     | 0.074 (ns) |
| Question clarity   | 4.18 | 3.92    | -0.26    | 0.0004    | 0.002 **   |
| Distractor quality | 3.68 | 3.42    | -0.26    | 0.002     | 0.008 **   |
| Difficulty level   | 3.02 | 3.09    | +0.06    | 0.379     | 0.735 (ns) |
| Arabic fluency     | 4.04 | 3.99    | -0.05    | 0.367     | 0.735 (ns) |
| Explanation        | 4.23 | 4.09    | -0.14    | 0.058     | 0.174 (ns) |

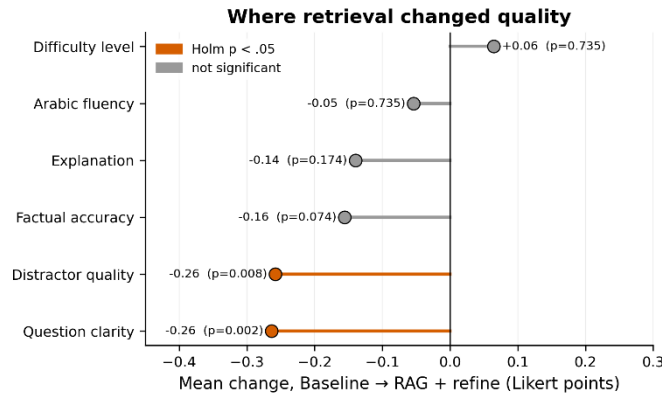


Fig. 4. Per-criterion change from baseline to RAG + refine, sorted by magnitude. Labels show Holm-adjusted paired Wilcoxon results; coloured points mark Holm  $p < .05$ .

Fig. 4 and Table IV report the per-criterion changes from baseline to the retrieval-plus-refinement condition, ordered by magnitude. In this contrast, the reductions surviving Holm correction were in question clarity ( $\Delta = -0.26$ ,  $p_{\text{adj}} = 0.002$ ) and distractor quality ( $\Delta = -0.26$ ,  $p_{\text{adj}} = 0.008$ ), with factual accuracy also declining though not surviving correction ( $\Delta = -0.16$ ). Difficulty level (+0.06) and Arabic fluency (-0.05) were essentially unchanged. The per-criterion mixed models for the retrieval-only contrast (Section IV-B) show a partly different profile, discussed in Section V-A. Retrieval therefore did not even improve the factual dimension it is designed to support, in this domain.

4) *Length control*: Retrieval-plus-refinement items were longer than baseline items (mean 76.9 vs 66.4 words; Wilcoxon  $p = .002$ ), whereas retrieval-only items were of similar length to the baseline (69.3 words;  $p = .37$ ). Length is therefore unlikely to explain the retrieval effect in Table I, which compares retrieval-only items with the baseline. To test this directly, the primary model was refitted with item length (in

words, centred) as a covariate. The retrieval estimate was essentially unchanged ( $\beta = -0.96$ , 95% CI [-1.47, -0.46],  $p = .0002$ ), and length itself did not predict ratings ( $\beta = +0.08$  per 10 words, 95% CI [-0.08, 0.23],  $p = .31$ ). Consistent with this, the per-topic quality change was not associated with the length difference (Spearman  $\rho = -0.13$ ,  $p = .50$ ). The quality decrease is therefore not accounted for by the greater length of some retrieval-grounded items.

### B. Inter-Rater Agreement

To characterise agreement between the two evaluators of each role, Spearman correlations of their composite scores were first computed; these were low ( $\rho = -0.07$  for students, 0.11 for instructors, 0.12 for examiners over 124 paired items each). To provide a standardised measure suitable for ordinal Likert data, Krippendorff's alpha [23] was also computed (ordinal level, using the algorithm's standard difference function) on the composite score, yielding  $\alpha = -0.54$  (Student), -0.16 (Instructor), and -0.12 (Examiner), and  $\alpha = -0.02$  across all six raters pooled. These values confirm low inter-rater reliability in absolute terms.

Two considerations bear on this. First, low agreement across roles was expected from the purposive design, which selected evaluators to represent distinct stakeholder perspectives — a candidate's tolerance for complexity versus an examiner's demand for regulatory precision — rather than to achieve consensus. This does not, however, explain the low agreement within roles, which was lowest between the two student raters ( $\alpha = -0.54$ ). Negative values indicate systematic rather than random disagreement, most plausibly reflecting individual differences in how raters used the scale and weighted the criteria; with two raters per role, such individual differences cannot be separated from role effects. Low agreement is nonetheless informative: it indicates that the composite does not reflect a single shared construct across raters, and the six raters are therefore not treated as interchangeable measurements of one latent quality. Second, the primary analyses rely on paired, within-rater comparisons, so each evaluator serves as their own control across the four conditions. This supports the observed within-rater contrast in this panel while limiting generalisation beyond these raters; it reduces dependence on between-rater level differences but does not, by itself, extend the result to the wider stakeholder populations. Per-role effects are accordingly treated as exploratory (Section V-E).

1) *Composite reliability*: Although between-rater agreement was low, the six criteria cohere as an internal scale: Cronbach's  $\alpha$  for the six-item composite was 0.729, which is acceptable and supports summing the criteria into a single composite. Reliability of the composite (internal consistency) and agreement between raters (consensus) are distinct properties, and only the former is required to justify the composite outcome.

2) *Per-criterion mixed models*: To confirm that the criterion-level pattern is not an artifact of pooling clustered observations, each criterion was refitted within the mixed-effects framework (crossed random effects for evaluator and topic). Note that this model estimates the retrieval main effect

(baseline vs. RAG-only, treatment-coded), which is a different contrast from Table IV (baseline  $\rightarrow$  RAG + refine); the two therefore need not coincide. Under the retrieval main effect, the coefficient was negative and significant for Arabic fluency ( $\beta = -0.29$ ,  $p < .001$ ), question clarity ( $\beta = -0.27$ ,  $p < .001$ ), explanation clarity ( $\beta = -0.27$ ,  $p < .001$ ), and factual accuracy ( $\beta = -0.13$ ,  $p = .037$ ), negative but non-significant for distractor quality ( $\beta = -0.13$ ,  $p = .094$ ), and slightly positive for difficulty ( $\beta = +0.15$ ,  $p = .055$ ). The direction is broadly consistent with the clarity-related degradation seen in Table IV, but because the contrasts differ, the two analyses are not claimed to estimate the same quantity; the per-criterion main effects are reported as a clustered-data robustness check on direction only. Some per-criterion models produced boundary/singular fits given the small number of evaluators, so these criterion-level estimates are treated as descriptive.

### C. Effect of Refinement

Iterative refinement produced no detectable quality change. The mixed-effects refinement main effect was negligible ( $\beta = -0.07$ , 95% CI  $[-0.57, 0.43]$ ,  $p = .786$ ; Table I). Descriptively, applied without retrieval refinement left the baseline essentially unchanged (paired  $\Delta = -0.07$ , 95% CI  $[-0.46, 0.32]$ ), and applied within RAG it did not recover the retrieval decrease ( $\Delta = +0.13$ , 95% CI  $[-0.30, 0.56]$ ). Because a non-significant test does not by itself establish equivalence, a two-one-sided-tests (TOST) equivalence analysis was run against a smallest effect of interest of  $\pm 1$  composite point (equivalent to one of the six

criteria shifting by one Likert level, or about 0.17 Likert points per criterion): both refinement contrasts were statistically equivalent to zero within this bound (both TOST  $p < .001$ ).

A Friedman test across all four conditions was significant ( $\chi^2(3) = 10.17$ ,  $p = .017$ ); this omnibus test establishes only that at least one condition differs, and the pairwise contrasts against the baseline (Table II) locate that difference in the retrieval conditions rather than in refinement: both retrieval contrasts remain significant under Bonferroni correction for three comparisons (adjusted  $p = .006$  and  $.015$ ), whereas the refinement-only contrast does not ( $p = .968$ ).

### D. Automatic Vs. Human

TABLE V. AUTOMATIC METRICS BETWEEN CONDITION PAIRS (MEAN OVER 31 TOPICS).

| Comparison                           | ROUGE-L | BERTScore F1 | Reading                        |
|--------------------------------------|---------|--------------|--------------------------------|
| Baseline $\rightarrow$ RAG t0        | 0.205   | 0.738        | wording differs, topic related |
| Baseline $\rightarrow$ RAG t3        | 0.202   | 0.734        | same pattern after refinement  |
| Baseline $\rightarrow$ Refine-no-RAG | 0.594   | 0.864        | moderate change                |
| RAG t0 $\rightarrow$ RAG t3          | 0.803   | 0.925        | refinement changes little      |

<sup>a</sup> Lower ROUGE-L indicates greater lexical difference; higher BERTScore indicates greater semantic similarity. Neither measures pedagogical quality.

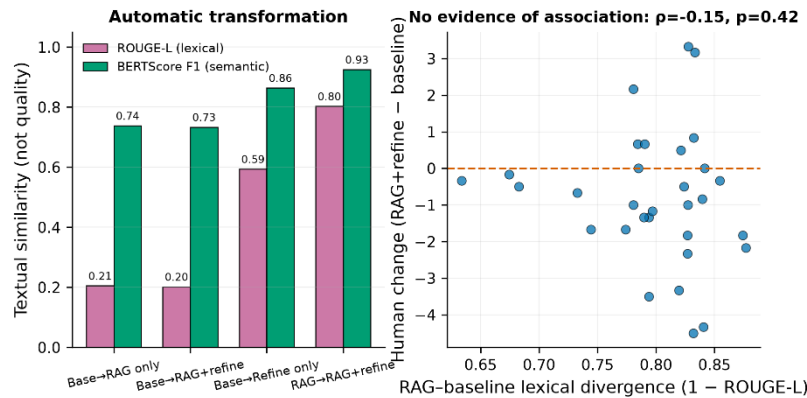


Fig. 5. Automatic textual change versus rated quality. Left: between-condition ROUGE-L (lexical overlap) and BERTScore F1 (semantic similarity), means over 31 topics — both are similarity measures between generated outputs, not quality scores. Right: per-topic lexical divergence between the RAG and baseline items ( $1 - \text{ROUGE-L}$ ) against the human quality change (mean of six evaluator composites;  $n = 31$ ).

Fig. 5 contrasts the automatic textual-change metrics with the human quality ratings. The automatic metrics describe a different quantity from the human ratings. RAG-grounded items diverged strongly in wording from the independently generated baseline items (ROUGE-L  $\approx 0.20$ ) while retaining moderate semantic similarity (BERTScore  $\approx 0.74$ ; Table V). Refinement within RAG changed the text little (ROUGE-L 0.80, BERTScore 0.93), consistent with the null human effect. These metrics documented substantial textual difference and moderate semantic similarity, but neither supplied a directional quality criterion; they therefore could not, by themselves, indicate whether the pipeline improved item quality. The human panel rated the transformed items lower. Across the 31 topics, there

was no evidence of an association between the lexical divergence of the RAG and baseline items and the human quality change (Spearman  $\rho = -0.15$ , 95% CI  $[-0.48, 0.22]$ ,  $p = .42$ ; Fig. 5); at  $n = 31$ , this test can detect only moderate or larger associations, so it does not establish the absence of a small one, but it does not indicate that the decrease is confined to the most divergent items. Textual-similarity and rated-quality signals thus measure different things: only the human ratings indicate whether the items improved or degraded.

1) *Field-level divergence*: To locate where the transformation occurred, ROUGE-L and BERTScore were computed separately for each MCQ field (stem, options, correct

answer, explanation). Relative to the baseline, the retrieval-grounded items differed from the baseline in every field while remaining in a related semantic space: ROUGE-L ranged from 0.10 (answer) to 0.20 (options), with BERTScore F1 of 0.71–0.77 across fields (Table VI). The answer field showed the lowest lexical overlap (ROUGE-L  $\approx$  0.10), indicating that RAG items often targeted a different fact from the baseline item (see the example in Section V-A); the answer-correctness audit (see below) found all audited RAG + refine answers factually correct. Within the RAG branch, refinement, by contrast, changed all fields only marginally (ROUGE-L 0.72–0.85, BERTScore 0.93–0.96). The field-level pattern thus reinforces the whole-item result: retrieval-grounded items were lexically distinct from, but topically related to, the baseline items, and their audited answers were correct; human raters nonetheless judged them lower in quality. As with the whole-item metrics, these field-level scores quantify textual change, not quality.

TABLE VI. FIELD-LEVEL TEXTUAL CHANGE BETWEEN CONDITIONS: ROUGE-L AND BERTSCORE F1 (MEAN OVER 31 TOPICS), COMPUTED SEPARATELY FOR EACH MCQ FIELD

| Field       | RAG vs base<br>ROUGE-L | RAG vs base<br>BERTScore | Refine (in<br>RAG)<br>ROUGE-L | Refine (in<br>RAG)<br>BERTScore |
|-------------|------------------------|--------------------------|-------------------------------|---------------------------------|
| Stem        | 0.19                   | 0.77                     | 0.72                          | 0.93                            |
| Options     | 0.20                   | 0.71                     | 0.77                          | 0.93                            |
| Answer      | 0.10                   | 0.71                     | 0.85                          | 0.96                            |
| Explanation | 0.12                   | 0.73                     | 0.83                          | 0.96                            |

<sup>b</sup>. Low ROUGE-L with moderate BERTScore indicates different wording with related content; neither is a quality measure.

2) *Structural-format audit*: Because human quality ratings do not by themselves establish that an item is structurally usable, a deterministic structural-format audit was applied to all 124 generated items (31 topics  $\times$  4 conditions), checking for a well-formed stem, exactly three options, a designated correct answer, an explanation, and absence of verbatim answer leakage from the stem. All four conditions were structurally compliant in every case (31/31 with three options, a correct answer, and an explanation), with a single instance of potential answer leakage in the RAG t0 condition (1/31) and none elsewhere. This establishes formatting compliance only; it does not establish factual correctness, answer uniqueness, distractor plausibility, or fidelity to the source. Answer correctness and passage relevance were assessed separately by two domain experts (see the retrieval-relevance audit below); source entailment was not audited and is left to future work. The human-rated differences should therefore be read as differences in perceived item quality rather than as gross structural defects.

3) *Retrieval-relevance audit*: To assess whether the observed decrease could be attributed to poor retrieval, two domain experts (a certified instructor and a licensing examiner) independently audited all 31 topics through a dedicated web interface. For each topic, they rated the relevance of each of the three retrieved passages (93 in total) on a three-point ordinal scale (relevant / partly relevant/irrelevant), and judged whether

the marked answer of the RAG + refine (t3) item for that topic was factually correct (correct/incorrect/unsure). Baseline and non-retrieval items were not shown and were not audited. The passages were reconstructed from the stored retrieval references; 25 of the 93 resolved only to short headings or title pages and are reported separately. Across all 93 passages (Table VII), 90% were judged at least partly relevant, with substantial inter-rater agreement (Cohen's  $\kappa$  = 0.63; ordinal Krippendorff's  $\alpha$  = 0.72; 78% exact and 100% within-one-grade agreement). Restricting to the 68 passages with full body text, every passage was rated at least partly relevant by both experts (100%;  $\kappa$  = 0.61), and none was rated irrelevant; the 25 heading-only passages had a lower relevance rate and lower Cohen's  $\kappa$  (64% at least partly relevant;  $\kappa$  = 0.55), though ordinal Krippendorff's  $\alpha$  was not correspondingly lower ( $\alpha$  = 0.73). Both experts judged all 31 audited RAG + refine answers factually correct (100% agreement; no 'unsure' responses). Retrieval therefore supplied predominantly on-topic context, and the audited RAG + refine answers were correct, indicating that the quality reduction in this condition is unlikely to be explained by retrieval failure or by factual errors in its items. Because the baseline and the other two conditions were not expert-audited, the audit does not establish whether answer correctness differed between conditions; the only cross-condition evidence on factual quality is the raters' factual-accuracy criterion (Table IV and Section IV-B).

TABLE VII. RETRIEVAL-RELEVANCE AND ANSWER-CORRECTNESS AUDIT BY TWO DOMAIN EXPERTS. " $\geq$  PARTLY" = JUDGED RELEVANT OR PARTLY RELEVANT;  $\kappa$  = COHEN'S KAPPA;  $\alpha$  = ORDINAL KRIPPENDORFF'S ALPHA; WITHIN-1 = AGREEMENT WITHIN ONE ORDINAL GRADE. ALL 31 AUDITED RAG + REFINE (T3) ANSWERS WERE JUDGED CORRECT BY BOTH EXPERTS (100% AGREEMENT)

| Passage set                  | n  | $\geq$ partly<br>relevant | Cohen<br>$\kappa$ | Krippendorff<br>$\alpha$ | Within-<br>1-grade |
|------------------------------|----|---------------------------|-------------------|--------------------------|--------------------|
| All<br>retrieved<br>passages | 93 | 90%                       | 0.63              | 0.72                     | 100%               |
| Full-text<br>passages        | 68 | 100%                      | 0.61              | 0.61                     | 100%               |
| Heading-<br>only<br>passages | 25 | 64%                       | 0.55              | 0.73                     | 100%               |

## V. DISCUSSION

### A. Why Retrieval Did Not Help

The main finding of this study is that retrieval grounding did not improve the human-rated quality of generated MCQs and lowered it slightly ( $d$  = -0.28). The domain was chosen deliberately as one likely to be familiar to the base model, to probe a boundary condition for retrieval. The observed pattern is compatible with the possibility that the base model already encoded enough information for this relatively familiar domain; however, baseline competence was not independently measured, and this account should be read as a hypothesis rather than a demonstrated conclusion. An important alternative explanation, discussed below and in Section V-E, is that retrieval quality was imperfect.

One possible account, proposed here as a hypothesis, is that retrieved text did not supply missing facts but constrained generation toward the phrasing of the source passages, yielding items that adhere more closely to the lesson wording but are less clear. A systematic, though indirect, check across all 31 topics supports the first part of this account. Measured against the reconstructed page text, the proportion of an item's word bigrams that also occur in its topic's retrieved pages was 0.235 for retrieval-only items and 0.204 for retrieval-plus-refinement items, compared with 0.037 for baseline items, which never saw the passages; retrieval-only items reused more source wording than the baseline item in 28 of 31 topics (Wilcoxon  $p < .001$ ). However, the per-topic increase in reused wording was not associated with the per-topic quality change (Spearman  $\rho = 0.24$ ,  $p = .19$ ), and the non-significant correlation ran in the direction opposite to that predicted. Retrieval therefore clearly pulled items toward source wording, but the data do not show that items borrowing more wording lost more quality, so over-constraint remains a plausible but unconfirmed explanation. This interpretation is consistent with the per-criterion pattern. In the baseline  $\rightarrow$  RAG + refine contrast (Table IV), the only reductions surviving Holm correction were in question clarity and distractor quality. The per-criterion mixed models for the retrieval-only contrast (Section IV-B) also show significant decreases in Arabic fluency and explanation clarity, while the distractor-quality decrease there is not significant. Question clarity is the only criterion that decreased significantly in both analyses. The pattern is therefore best described as a broad decrease in clarity and form, with difficulty unaffected.

TABLE VIII. QUALITATIVE ERROR ANALYSIS FOR THE TOPIC "MEETING AND OVERTAKING." ENGLISH GLOSSES ARE SHOWN; THE ORIGINAL ARABIC ITEMS ARE GIVEN IN APPENDIX B

|                   | Baseline                                                                                                                                       | RAG + refine (t3)                                                                                                                 |
|-------------------|------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------|
| Stem (gloss)      | When approaching a junction, what should a driver do about another vehicle attempting to overtake?                                             | In which case must a driver overtaking on the left occupy the left half of the carriageway?                                       |
| Options (gloss)   | a) speed up to prevent overtaking; b) leave sufficient space for the overtaking vehicle; c) change direction suddenly to warn the other driver | a) always, regardless of road conditions; b) unless it inconveniences oncoming vehicles; c) only when there are two or more lanes |
| Retrieved context | — (no retrieval)                                                                                                                               | Regulatory passages on overtaking procedure (section 2, lessons 5 and 7)                                                          |
| Effect            | Clear, mutually exclusive options; unambiguous correct answer                                                                                  | Legalistic stem; options hinge on legal qualifications rather than driving concepts; reduced clarity                              |

Table VIII gives an illustrative example of the failure mode driving the clarity and distractor degradation. For the topic of meeting and overtaking, the baseline model generates a succinct, unambiguous stem asking what a driver should do when another vehicle attempts to overtake, with three cleanly contrasting options. The RAG pipeline, however, retrieves a dense regulatory passage on overtaking procedure, and the model is constrained to incorporate its legislative phrasing. The result is a convoluted stem — concerning the precise conditions under which a driver overtaking on the left must occupy the left half

of the carriageway — with distractors that hinge on legal qualifications rather than clear driving concepts. This directly instantiates the significant drop in human-rated clarity: the retrieved context did not correct a factual error but degraded the pedagogical form of an already-adequate question.

1) *Representative Arabic examples*: To show how the clarity and distractor decreases arise, Appendix B gives the original Arabic baseline and retrieval-plus-refinement items for the three topics with the largest rater-averaged decreases in question clarity or distractor quality, selected by that rule rather than by hand. Three recurring patterns appear. First, regulatory phrasing: for meeting and overtaking (clarity  $-1.33$ , distractors  $-0.67$  Likert points), the retrieval item turns a practical driving question into a legal condition on occupying the left half of the carriageway, with options that hinge on legal qualifications (Table VIII). Second, recall of list structure rather than understanding: for adapting driving to dangerous bends, hills, and slopes (clarity  $-1.17$ , distractors  $-1.33$ ), the retrieval item asks for “the stages” of negotiating a bend but offers only counts (two, three, or four stages), so the stem and options do not match, and the distractors are trivially distinguishable numbers; the item also narrows a three-part topic to bends alone. Third, answer cues and copied wording: for driving on bends (distractors  $-1.00$ ), the stem announces “the three stages”, which cues the option beginning “three stages”, and the options reproduce the lesson's list almost verbatim. Both bend items also reproduce the extraction artefact *ملائمة* (for *ملاءمة*) found in the retrieved page text, a direct trace of retrieved wording entering the item, and the meeting-and-overtaking item retains a code-switched non-Arabic token (*nhâm*) in its explanation. These examples are illustrative; failure types were not classified systematically across all 31 items (Section V.E).

This finding has a broader implication for the deployment of RAG in generation pipelines. It suggests a boundary condition that is easily overlooked: the value of retrieval is contingent on the existence of a knowledge gap that retrieval can fill. For specialised or rapidly changing domains, where the model is genuinely knowledge-limited, retrieval may well remain beneficial. For simple, high-frequency domains that are well represented in a model's pre-training, however, retrieval may add little and can slightly lower item quality, as observed here. Practitioners should therefore treat the benefit of RAG for generation as an empirical question to be tested per domain, rather than as a default that holds universally.

Two mechanisms could produce the observed decrease. Under the first, “bad retrieval,” the retrieved passages are only partially relevant and the questions degrade because the context is noisy. Under the second, “over-constraint by relevant retrieval,” the passages are on-topic but bias the generator toward source phrasing, harming clarity and distractor design even when the base model could have produced a cleaner item unaided. The retrieval-relevance audit (Section IV-D) bears directly on this: because both experts judged the retrieved passages to be predominantly relevant — unanimously so for the passages with full body text — the “bad retrieval” account is not well supported for this corpus, and over-constraint by relevant

material is the more plausible remaining account, although the lexical-reuse analysis above does not link it directly to the quality loss. Retrieval relevance was nonetheless not manipulated experimentally, and a fully controlled test would compare fixed top-k retrieval against relevance-filtered or reranked retrieval [19, 21] and vary k; both remain future work. The two accounts converge on the same practical conclusion — that retrieval as implemented here did not improve generated question quality in this domain.

### B. Why Refinement Did Not Help

The second finding is that iterative self-refinement produced no measurable improvement in quality, whether applied with or without retrieval. The human ratings showed no significant change after three refinement passes (only the initial and final versions were rated), and the effect was statistically equivalent to zero within  $\pm 1$  composite point (Section IV-C).

The textual metrics do not measure quality, but they help interpret this null result. Within the RAG branch, refinement altered the text only marginally (ROUGE-L 0.80, BERTScore 0.93 between RAG t0 and t3), so there was little change for raters to detect. Without retrieval, however, refinement changed the text moderately (ROUGE-L 0.59, BERTScore 0.86), yet ratings were still unchanged ( $\Delta = -0.07$ ). The null result therefore does not reduce to refinement leaving items untouched: even where refinement rewrote items noticeably, it did not make them better. Consistent with this, informal inspection of the model's change notes suggests that the self-critique often concluded that no substantial change was needed. Refinement also left at least one clear defect in place: the RAG + refine item in Table VIII retains a code-switched non-Arabic token in its explanation. In this setting, three refinement passes did not improve perceived quality.

### C. Textual Change Versus Rated Quality

The comparison between automatic and human evaluation, addressed by RQ3, has a clear limit: ROUGE-L and BERTScore were computed between the outputs of two conditions, not against reference items, so they measure how much an item changed, not whether it improved. Under the selected configurations, retrieval-grounded items differed substantially in lexical form from the baseline items (ROUGE-L  $\approx 0.20$ ) while remaining moderately similar in meaning (BERTScore  $\approx 0.74$ ). On this evidence alone, the RAG items would appear to be lexically distinct but topically related alternatives to the baseline items. The human ratings supplied the direction of the change: the same items were rated lower in quality than the ungrounded baseline. The per-topic association between lexical divergence and the human quality change was weak and not significant (Spearman  $\rho = -0.15$ ,  $p = .42$ ,  $n = 31$ ); with 31 topics, this test can detect only moderate or larger associations, so it shows only that the quality decrease was not confined to the most heavily rewritten items. The analysis therefore does not show that automatic metrics would miss quality losses in general; it shows that between-condition similarity scores, by construction, carry no information about the direction of quality change. For assessment content, where the operative standard is expert judgment of item quality, such scores are best used to characterise the nature and extent of a transformation, alongside

rather than instead of human evaluation, consistent with broader cautions about automatic evaluation [1], [18].

### D. Implications

The findings carry several implications for research and practice in Arabic educational NLP. For system builders, they caution against the uncritical adoption of retrieval and refinement as default components of a generation pipeline; both should be validated empirically for the specific domain and model in use, since neither improved quality here and retrieval slightly lowered it. For the evaluation of generation systems, the results show that, in this setting, between-condition similarity metrics could not replace human evaluation, because they did not indicate the direction of quality change, and that multi-role human panels, though costly, reveal information that single-metric or single-rater evaluations cannot. More broadly, the study illustrates the value of complete factorial designs in generation research: by isolating retrieval and refinement rather than evaluating them only in combination, the design made it possible to attribute the observed degradation specifically to retrieval and to establish that refinement contributed nothing measurable, a decomposition that a non-factorial design could not have provided.

### E. Limitations

1) *Generalisability*: The evidence comes from 31 topics in one domain (road-safety and driving-licence instruction), one generator (llama-3.3-70b-versatile), one off-the-shelf Arabic encoder (AraBERT v2), one page-level FAISS index with  $k = 3$ , and one blinded six-rater panel. Road-safety rules are widely documented, so the base model may already have encoded much of the required knowledge. The negative retrieval effect may therefore not hold for domains in which the model is genuinely knowledge-limited (for example, recent, local, or highly specialised curricula), for other model families or sizes, for other Arabic encoders or chunking strategies, or for other item formats. The result should be read as a bounded finding for this configuration, not as a general verdict on retrieval augmentation for Arabic MCQ generation. Replication across additional domains, models, and retrieval configurations is needed to locate that boundary.

2) *Two raters per role and inter-rater agreement*: Each role is represented by two evaluators, and within-role agreement was low (Section IV-B). The paired within-rater design supports the observed within-rater contrast in this panel, because each evaluator serves as their own control; it does not, however, license generalisation beyond these six raters to the wider populations of students, instructors, or examiners. The per-role effects — including the examiner reduction, whose  $p = .050$  would not survive correction for the three role-level tests — are exploratory rather than confirmatory, and no role-interaction conclusion is drawn without a formal interaction test. The retrieval effect was nonetheless directionally consistent across raters (negative for five of six) and remained significant with evaluator-specific random slopes (Section IV-A); larger panels per role are needed to estimate role differences and population-level rater variance.

3) *Question length*: Retrieval-plus-refinement items were longer than baseline items (Section IV-A). Adjusting for item length left the retrieval effect unchanged, and retrieval-only items were of similar length to the baseline; a generation design that constrains item length directly would nevertheless provide a stronger control.

4) *Retrieval quality*: Retrieval quality was not manipulated experimentally. A post-hoc expert audit (Section IV-D) found the retrieved passages predominantly relevant, which argues against a simple bad-retrieval explanation. However, the exact retrieved text was not stored at generation time, so the audit and the lexical-reuse analysis (Section V-A) used page text reconstructed from the stored source references, 25 of which resolved only to heading pages; relevance was assessed by two experts only; and the AraBERT encoder was used without retrieval-specific fine-tuning. Failure types were not classified systematically across the 31 retrieval-grounded items, so Table VIII and the examples in Section V-A are illustrative. The answer-correctness audit covered only the retrieval-plus-refinement condition; baseline answers were not audited, so the audit does not compare factual correctness across conditions. A stronger design would store and annotate the retrieved context prospectively, classify failure types for every item, audit answers in every condition, vary the number of retrieved passages ( $k \in \{1, 3, 5\}$ ), and compare against reranked or relevance-filtered retrieval [19].

5) *Fixed retrieval depth*: All retrieval conditions used a single setting of  $k = 3$  retrieved pages. Because  $k$  determines how much source text enters the prompt, both the size and the direction of the retrieval effect may depend on it. Fewer pages might reduce the source-conditioned constraint discussed in Section V-A, whereas more pages might add marginal or redundant content. The present result therefore applies to  $k = 3$  only. A sensitivity analysis with  $k \in \{1, 3, 5\}$ , rated under the same blinded protocol, is required before the finding can be extended to other retrieval depths.

6) *Single generation per cell*: Each of the 124 items is a single stochastic generation (temperature 0.7 for generation and 0.5 for refinement; no fixed seed). The estimated retrieval and refinement effects are therefore confounded with generation-level sampling variability: the design cannot fully separate “retrieval systematically produces lower-rated items” from “this realised sample of retrieval-grounded items was rated lower.” The topic random effect (variance 0.49) captures between-topic variation but not within-cell generation noise. The direction of the effect is consistent across topics (23 of 31; Section IV-A), which reduces this concern but does not remove it. Generating three to five independent completions per topic and condition, rated under the same blinded protocol and modelled with a generation-draw random effect, would separate the two sources of variation and is an important next step. The retrieval effect is accordingly framed as about this realised set of generations under the stated configuration, not retrieval augmentation in general.

7) *No positive control for panel sensitivity*: The design does not include a known-difference (deliberately degraded) condition rated under the same blinded protocol, so the refinement equivalence result cannot be fully distinguished from limited instrument sensitivity to a one-point composite change. Internal consistency (Cronbach’s  $\alpha = 0.729$ ) shows the criteria cohere but not that the scale tracks true quality shifts. A calibration item with an injected defect, rated blindly alongside the four conditions, would establish this sensitivity and is recommended for replication.

8) *Automatic metric choice*: ROUGE-L and BERTScore were computed between conditions rather than against expert reference items, and other metrics or reference-based designs might relate differently to human judgment. The claim is limited to these two metrics as used here: they described the extent of textual change but not its direction.

9) *Prompt differences*: The retrieval and non-retrieval conditions differed in their instructions as well as in the inserted context (Section III-B), so the retrieval effect reflects the retrieval-grounded configuration as a whole. A design with prompts identical except for the inserted context would isolate the effect of the retrieved text itself.

## VI. CONCLUSION

This study set out to test, under rigorous human evaluation, whether two techniques widely assumed to improve LLM generation — retrieval augmentation and iterative self-refinement — in fact improve the quality of automatically generated Arabic multiple-choice questions. Using a complete  $2 \times 2$  factorial design and a purposive multi-role human panel of six raters spanning candidates, instructors, and licensing examiners, 744 blinded ratings were collected across 31 topics and complemented with automatic metrics.

Within the evaluated model, corpus, domain, and protocol, the evidence supports three bounded conclusions. First, contrary to the common expectation, retrieval grounding did not improve MCQ quality; it was associated with a small but statistically reliable decrease ( $d = -0.28$ , about 0.16 Likert points per criterion), concentrated in clarity- and form-related criteria, most consistently question clarity. Second, iterative self-refinement produced no measurable change in quality, a null result supported by an equivalence test and observed even where refinement changed the text moderately. Third, between-condition similarity metrics characterised the extent of textual difference but carried no information about its direction, which only the human ratings revealed.

Taken together, these findings suggest that in this bounded setting neither retrieval nor refinement improved Arabic MCQ quality, that their benefit should be tested empirically per model, domain, and retrieval configuration rather than assumed, and that between-condition similarity metrics, which do not indicate the direction of quality change, cannot substitute for expert human evaluation of assessment content. Although the study was limited to 31 topics in a single domain, one model, one generation per item, one retrieval depth ( $k = 3$ ), and two raters per role, its complete factorial design and multi-role protocol provide a template for more general investigation. Future

research should extend the design to domains in which the base model is genuinely knowledge-limited, where retrieval may yet prove beneficial; to additional models and Arabic encoders; to several independent generations per item and several retrieval depths; and to larger, more homogeneous rater panels. To support such replication and extension, the generated questions for all four conditions, the anonymised ratings, and the analysis scripts are openly available (see Data availability), and the multi-role evaluation protocol is fully described in Section III-C. In conclusion, the evidence presented here indicates that the benefits of retrieval and refinement, far from being universal, are contingent on conditions that must be verified rather than assumed — a caution of direct relevance to the growing body of work on Arabic language models.

## VII. DECLARATIONS

1) *Data availability*: The generated multiple-choice items for all four conditions, the anonymised human ratings (evaluators identified by role only), and the derived automatic-metric, field-level, and retrieval-answer audit outputs are openly available on Figshare at <https://doi.org/10.6084/m9.figshare.33940903>. The deposit also includes the analysis and figure-generation scripts; the exact generation and refinement prompts are given in Appendix A. Retrieved passages are listed by reference (source file and page) rather than by their text. The underlying driving-education lesson material is third-party copyrighted content and is not redistributed; the release therefore comprises model-generated items and derived data, not the source text.

2) *Ethics approval and consent*: The study involved no clinical or behavioural intervention and posed no more than minimal risk, consisting of adults rating machine-generated text; no personal, sensitive, or health-related data were collected. Given this minimal-risk, non-interventional design, formal ethics-committee review was not sought. All evaluators were adults (aged 18 or over) who were informed of the purpose and voluntary nature of the study and gave verbal consent before participating; participation was unpaid, and raters could withdraw at any time. Any identifiers noted during recruitment were removed prior to analysis, and the data are reported and released in anonymised form, with evaluators identified by stakeholder role only.

3) *Competing interests*: The authors declare no competing interests.

4) *Funding*: No funding was received for this study.

5) *Use of generative AI*: An LLM was the object of study (the generation system under evaluation). AI-assisted tools were separately used for language polishing during manuscript preparation; no AI tool is an author, and the authors take full responsibility for all content.

## REFERENCES

- [1] H. W. Awalurahman, R. F. Aji, and I. Budi, "Transformer and large language models for automatic multiple-choice question generation: a systematic literature review," *IEEE Access*, vol. 13, pp. 127100–127112, 2025, doi: 10.1109/ACCESS.2025.3590423.
- [2] G. Kurdi, J. Leo, B. Parsia, U. Sattler, and S. Al-Emari, "A systematic review of automatic question generation for educational purposes," *Int. J. Artif. Intell. Educ.*, vol. 30, no. 1, pp. 121–204, 2020, doi: 10.1007/s40593-019-00186-y.
- [3] P. Lewis, E. Perez, A. Piktus, et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459–9474.
- [4] A. Madaan, N. Tandon, P. Gupta, et al., "Self-Refine: iterative refinement with self-feedback," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 46534–46594.
- [5] C.-Y. Lin, "ROUGE: a package for automatic evaluation of summaries," in *Text Summarization Branches Out*, 2004, pp. 74–81. [Online]. Available: <https://aclanthology.org/W04-1013/>
- [6] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: evaluating text generation with BERT," in *International Conference on Learning Representations*, 2020. [Online]. Available: <https://openreview.net/forum?id=SkeHuCVFDr>
- [7] W. Antoun, F. Baly, and H. Hajj, "ArABERT: transformer-based model for Arabic language understanding," in *Proceedings of OSACT4*, 2020, pp. 9–15, checkpoint aubmindlab/bert-base-arabertv02. [Online]. Available: <https://aclanthology.org/2020.osact-1.2/>
- [8] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Trans. Big Data*, vol. 7, no. 3, pp. 535–547, 2021, doi: 10.1109/TBDATA.2019.2921572.
- [9] R. Rosenthal, *Meta-analytic Procedures for Social Research*, rev. ed. Newbury Park, CA: Sage, 1991.
- [10] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ: Lawrence Erlbaum, 1988.
- [11] N. Scaria, S. D. Chenna, and D. Subramani, "How good are modern LLMs in generating relevant and high-quality questions at different Bloom's skill levels for Indian high school social science curriculum?," in *Proceedings of BEA 2024*, 2024, pp. 1–10. [Online]. Available: <https://aclanthology.org/2024.bea-1.1/>
- [12] F. Koto, H. Li, S. Shatnawi, et al., "ArabicMMLU: assessing massive multitask language understanding in Arabic," in *Findings of ACL 2024*, 2024, pp. 5622–5640, doi: 10.18653/v1/2024.findings-acl.334.
- [13] S. S. Mucciaccia, T. M. Paixão, F. W. Mutz, et al., "Automatic multiple-choice question generation and evaluation systems based on LLM: a case study with university resolutions," in *Proceedings of COLING 2025*, 2025, pp. 2246–2260. [Online]. Available: <https://aclanthology.org/2025.coling-main.154/>
- [14] R. A. Alhazaymeh and M. Z. Ali, "Automated question generation for Arabic language," *Cluster Comput.*, vol. 28, no. 15, Art. no. 949, 2025, doi: 10.1007/s10586-025-05626-0.
- [15] A. Gan, H. Yu, K. Zhang, et al., "Retrieval-augmented generation evaluation in the era of large language models: a comprehensive survey," *arXiv preprint arXiv:2504.14891*, 2025, doi: 10.48550/arXiv.2504.14891.
- [16] Y. Huang and J. X. Huang, "A survey on retrieval-augmented text generation for large language models," *ACM Comput. Surv.*, vol. 58, no. 12, pp. 1–38, 2026, doi: 10.1145/3805774.
- [17] Y. Liu, D. Iter, Y. Xu, et al., "G-Eval: NLG evaluation using GPT-4 with better human alignment," in *Proceedings of EMNLP 2023*, 2023, pp. 2511–2522, doi: 10.18653/v1/2023.emnlp-main.153.
- [18] T. He, J. Zhang, T. Wang, et al., "On the blind spots of model-based evaluation metrics for text generation," in *Proceedings of ACL 2023*, 2023, pp. 12067–12097, doi: 10.18653/v1/2023.acl-long.674.
- [19] Y. Bouziane, Y. Moukafih, and M. Ghogho, "Candidate-aware retrieval and reranking for multiple-choice question answering: Arabic as a case study," in *Findings of ACL 2026*, 2026, pp. 8967–8977, doi: 10.18653/v1/2026.findings-acl.435.
- [20] A. Grattafiori, A. Dubey, A. Jauhri, et al., "The Llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024, doi: 10.48550/arXiv.2407.21783.
- [21] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAs: automated evaluation of retrieval-augmented generation," in *Proceedings of EACL 2024: System Demonstrations*, 2024, pp. 150–158, doi: 10.18653/v1/2024.eacl-demo.16.
- [22] T. M. Haladyna, S. M. Downing, and M. C. Rodriguez, "A review of multiple-choice item-writing guidelines for classroom assessment," *Appl.*

Meas. Educ., vol. 15, no. 3, pp. 309–333, 2002, doi: 10.1207/S15324818AME1503\_5.

- [23] K. Krippendorff, "Computing Krippendorff's alpha-reliability," Departmental Papers, Annenberg School for Communication, University of Pennsylvania, 2011. [Online]. Available: [https://repository.upenn.edu/asc\\_papers/43/](https://repository.upenn.edu/asc_papers/43/).

## APPENDIX

### A. Generation and Refinement Prompts

For full reproducibility, the exact prompt templates used in each stage are provided below. The prompts were written in Arabic; English glosses are given in brackets. Placeholders in braces were filled at runtime.

#### 1) A.1 Baseline / RAG generation prompt

أنت خبير في تعليم قانون المرور المغربي. بناءً على المحتوى التعليمي التالي، قم بإنشاء سؤال اختبار من نوع الاختيار المتعدد. المحتوى التعليمي: {السياق المسترجع}. المطلوب: أنشئ سؤالاً واحداً بالعربية الفصحى يتضمن: سؤال محدد ودقيق؛ 3 خيارات (أ، ب، ج) من نفس الحقل الدلالي؛ الإجابة الصحيحة؛ تفسير قصير مبني على محتوى الدرس. الموضوع المطلوب: {الموضوع}. أجب بهذا ...: التنسيق فقط: السؤال: ... (أ ... (ب ... (ج ... الإجابة الصحيحة: ... التفسير

[You are an expert in Moroccan traffic-law instruction. Based on the following instructional content, create one multiple-choice test question. Instructional content: {retrieved context}. Produce exactly one question in Modern Standard Arabic containing: a specific, precise stem; three options (أ, ب, ج) from the same semantic field; the correct answer; and a short explanation grounded in the lesson content. Required topic: {topic}. Answer in this format only: Question: ... / a) ... / b) ... / c) ... / Correct answer: ... / Explanation: ... For the baseline and refine-without-retrieval conditions, the retrieved-context block was omitted; the ungrounded baseline prompt asks only for a question on {topic} in the same output format.]

#### 2) A.2 Self-refine feedback-and-revision prompt

RAG في فرع]. أنت مدقق متخصص في أسئلة امتحان رخصة القيادة المغربية [نُدرج هنا: المحتوى التعليمي المرجعي (استند إليه للتحقق من الدقة): {السياق} السؤال الحالي: {السؤال}. مهمتك: أجب تحسينات دقيقة ومحددة فقط على هذا السؤال. تحقق فقط من هذه الأخطاء الجوهرية: 1. الدقة المعلوماتية: هل الإجابة الصحيحة خاطئة أو غير دقيقة؟ 2. وضوح السؤال: هل يحتمل أكثر من تأويل؟ 3. جودة الخيارات: هل الخيارات الخاطئة من نفس الحقل الدلالي ونفس الجذر الصرفي؟ 4. مستوى الصعوبة: هل السؤال سهل جداً أو صعب جداً لمرشح رخصة القيادة؟ 5. المطابقة: RAG: فرع]. سلاسة العربية: هل هناك ركائز لغوية واضحة؟ 6 RAG: للمحتوى — هل الإجابة متوافقة مع المحتوى المرجعي؟ / فرع بدون قواعد صارمة: [وضوح التفسير — هل التفسير يبرر الإجابة دون معلومات زائدة؟ لا تُضف معلومات عامة أو تعليقات غير ضرورية؛ غير فقط ما يحتاج تغييراً فعلياً؛ إذا كان السؤال جيداً فأبقه كما هو مع ملاحظة "لا تغييرات جوهرية". أولاً: ملاحظات التغيير (جملة أو جملتان)؛ ثم: السؤال المحسن بنفس التنسيق أعلاه.

[You are a specialist reviewer of Moroccan driving-licence exam questions. (In the RAG branch, a reference block is inserted: "Reference instructional content (use it to verify accuracy): {context}"). Current question: {question}. Task: make only precise, targeted improvements. Check only these substantive errors: 1) factual accuracy — is the correct answer wrong or imprecise? 2) question clarity — is it open to more than one reading? 3) distractor quality — are the incorrect options from the same semantic field and morphological root? 4) difficulty — is it too easy or too hard for a licence candidate? 5) Arabic fluency — is there clear awkwardness? 6) In the RAG

branch: content consistency — does the answer match the reference content?; in the no-retrieval branch, this sixth check is instead explanation clarity — does the explanation justify the answer without superfluous information? Strict rules: add no general information or unnecessary commentary; change only what genuinely needs changing; if the question is already good, keep it unchanged with the note "no substantial changes." First give the change notes (one or two sentences); then give the improved question in the same format above.]

### B. Representative Arabic Items

The three topics with the largest rater-averaged decreases (baseline → RAG + refine) in question clarity or distractor quality. Changes are mean differences over the six raters, in Likert points per criterion; composite changes are on the 6 – 30 scale.

1) B.1 التقابل والتجاوز (Meeting and overtaking). Clarity –1.33; distractors –0.67; composite –3.50.

Baseline:

السؤال: عند الاقتراب من تقاطع طرق، ما الذي يجب على السائق أن يفعله تجاه مركبة أخرى تحاول تجاوزه؟

- (أ) أن يزيد من سرعته ليمنع التجاوز  
(ب) أن يترك مسافة كافية للمركبة التي تريد التجاوز  
(ج) أن يغير اتجاهه فجأة لتحذير السائق الآخر  
الإجابة الصحيحة: ب

RAG + refine:

السؤال: في أي حالة يجب على السائق الذي يقوم بالتجاوز من جهة اليسار أن يحتل النصف الأيسر من القارة؟

- (أ) دائماً، بغض النظر عن ظروف الطريق  
(ب) إلا إذا كانت هناك مضايقة للمركبات القادمة من الجهة المعاكسة  
(ج) فقط عند وجود مسلكين أو أكثر على القارة  
الإجابة الصحيحة: ب

2) B.2 ملائمة السياقة في الظروف الخاصة بالمنعرجات الخطيرة والعقبة (Adapting driving to dangerous bends, hills, and slopes). Clarity –1.17; distractors –1.33; composite –4.50.

Baseline:

السؤال: عند سيطرة مركبة على طريق منحدر شديد، ما هو الإجراء الأكثر ملاءمة لضمان السلامة المرورية؟

- (أ) تخفيض السرعة والانتقال إلى غيار أقل  
(ب) الضغط بقوة على الفرامل الرئيسية  
(ج) التسارع لتعويض تأثير الجاذبية  
الإجابة الصحيحة: أ

RAG + refine:

السؤال: ما هي المراحل التي يتم من خلالها السير في المنعرجات وتجاوزها بشكل آمن؟

- (أ) مرحلتين  
(ب) ثلاث مراحل

(ج) أربع مراحل

الإجابة الصحيحة: ب

3) B.3 (Driving on bends). Clarity -0.67; distractors -1.00; composite -4.33.

Baseline:

السؤال: ما هي السرعة المناسبة عند السياقة في المنعرجات؟

(أ) السرعة القصوى المسموح بها على الطريق

(ب) سرعة منخفضة تسمح بالسيطرة على المركبة

(ج) سرعة متوسطة تتناسب مع ظروف الطريق

الإجابة الصحيحة: ب

RAG + refine:

السؤال: ما هي المراحل الثلاث التي يجب اتباعها للسير في المنعرجات وتجاوزها بشكل آمن؟

(أ) مرحلتان: إكتشاف المنعرج وتقدير شعاعه، مألوفة وتيرة السير

(ب) ثلاث مراحل: تحليل وتقدير شعاع المنعرج وتقدير خطورته، مألوفة وتيرة السير، التحكم في المسار والتسارع

(ج) أربع مراحل: إكتشاف المنعرج، تقدير شعاعه، مألوفة وتيرة السير، والتحكم في المركبة

الإجابة الصحيحة: ب