Designing Annotation Guidelines for Trait-Based Arabic Automated Essay Scoring: A Systematic Methodology

Walid Massoud[1] **Houda Bouamor**[2]
Abdelrahman Abdel Latif Hussein[3] **Abdullah Mohamed Mohamed Zekri**[4]

[1]Qatar University
[2]Carnegie Mellon University in Qatar
[3]Ministry of Eduction, Egypt
[4]National Center for Examinations and Educational Evaluation, Egypt

Abstract

Automated Essay Scoring (AES) fundamentally depends on high-quality annotated data, yet systematic approaches to developing annotation guidelines remain largely undocumented, especially for Arabic. We present a comprehensive methodology for trait-based Arabic AES annotation, applied to build a dataset of 7,859 essays by high school students annotated across seven writing traits, achieving substantial inter-annotator agreement (QWK: 0.66–0.75). Our methodology encompasses: (1) a seven-dimensional scoring framework grounded in Arabic linguistic and rhetorical conventions; (2) over 25 pages of Arabic-language guidelines with terminology unification, text-type-specific scoring descriptors, and annotated student examples; (3) a multi-stage training protocol that raised annotator agreement before production began; and (4) quality assurance mechanisms, including dual annotation and supervisor adjudication. We release all materials publicly, providing both a validated foundation for Arabic AES research and a replicable template for annotation guideline development in other morphologically complex, under-resourced languages

1 Introduction

Automated Essay Scoring (AES) has emerged as a critical NLP application enabling scalable writing assessment. While significant progress has been made for English (Mathias and Bhattacharyya, 2018; Crossley et al., 2023b), its efficiency fundamentally depends on the quality of the human-annotated data used for model training. In practice, this quality is shaped by the extent to which human evaluators achieve objectivity, accuracy, and inter-rater reliability. Essay assessment, by its nature, involves subjective judgment and interpretation, making it inherently difficult to ensure consistency and precision across raters. These challenges are further amplified in multi-trait scoring settings, where evaluators must simultaneously assess multiple dimensions of writing quality, increasing cognitive load and the potential for inconsistency. Ensuring the reliability and validity of the resulting scores remains a persistent challenge. This problem is especially pronounced for Arabic, an under-resourced language with distinct morphological, syntactic, and rhetorical characteristics that complicate both writing assessment and annotation guideline design. In such contexts, annotation is not merely a labeling task, but a structured reasoning process that must be explicitly supported to achieve consistent and reliable outcomes.

Arabic AES research faces particular challenges: scarcity of annotated datasets, the lack of established annotation protocols (Bashendy et al., 2024), and the inherent linguistic complexity of Modern Standard Arabic (MSA). MSA exhibits complex morphological agreement, flexible word order, and a rich system of cohesive devices that do not map onto frameworks developed for European languages. Existing Arabic AES datasets provide limited annotation documentation (Habash and Palfreyman, 2022), making replication and adaptation difficult. Annotators must distinguish subtle gradations - such as the difference between surface Arabic cohesive devices (روابط خطية) and deep lexical cohesion (ترابط معجمي), while maintaining consistency across thousands of essays. Without systematic, Arabic-specific guidelines, this leads to low inter-annotator agreement and unreliable training data.

This paper presents a comprehensive methodology for developing annotation guidelines for trait-based Arabic AES, applied to build a large-scale Arabic essay dataset comprising 7,859 essays from 4,372 highschool students across 24 schools in an Arab country (Bashendy et al., 2025b). While the individual components of our approach (multi-stage annotator training, dual annotation, calibra-

Proceedings of the 20th Linguistic Annotation Workshop (LAW XX), pages 146–157

tion sessions) reflect established best practices in annotation methodology (Artstein and Poesio, 2008; Williamson et al., 2012), their systematic adaptation to Arabic presents non-trivial challenges. Arabic-specific morphology, cohesive conventions, and rhetorical norms require decisions that cannot be derived from existing frameworks developed for English or other languages. Our core contribution is therefore the principled, documented *operationalization* of these practices for Arabic academic writing. Concretely, we contribute:

1. A seven-dimensional scoring framework grounded in Arabic linguistic and rhetorical conventions, with observable, Arabic-specific indicators replacing vague criteria (e.g., topic sentences الجمل الموضوعية as a proxy for organization);

2. Over 25 pages of annotation guidelines written entirely in Arabic, including a terminology unification section, text-type-specific scoring descriptors, and annotated student writing examples at each score level, none of which exist for Arabic AES in prior work;

3. A multi-stage training protocol with calibration providing evidence for each training decision rather than reporting protocol as a black box;

4. Quality assurance mechanisms adapted to the Arabic context, including adjudication protocols for code-switching and AI-generated text specific to MSA student writing; and

5. Full public release of all materials to enable direct replication and adaptation.

2 Related Work

AES Annotation Frameworks. English AES benefits from well-established datasets such as ASAP (Mathias and Bhattacharyya, 2018), ELLIPSE (Crossley et al., 2023a), and PERSUADE (Crossley et al., 2023b), but these provide limited documentation of annotation development. Cross-lingual frameworks such as TCFLE-8 for French (Wilkens et al., 2023), MERLIN for European languages (Boyd et al., 2014), and ACEA for Chinese (He et al., 2022) offer holistic assessments but lack trait-level granularity. None addresses Arabic academic writing, its morphological richness, or Arabic rhetorical conventions.

Annotation Guideline Development. Best practices for subjective NLP annotation emphasize clear definitions, concrete examples, and iterative refinement (Artstein and Poesio, 2008). Detailed rubrics, borderline-case training, and multiple pilot rounds are essential (Fort et al., 2011). In educational assessment, high inter-rater reliability requires extensive training and calibration (Williamson et al., 2012), with feedback and group discussion significantly improving agreement (Landis and Koch, 1977).

Arabic AES and Annotation. Arabic AES has been constrained by limited annotated data. ZAEBUC (Habash and Palfreyman, 2022) provides linguistic annotations but lacks trait-specific labels. QAES (Bashendy et al., 2024) introduced trait annotations at small scale (195 essays). The TAQEEM shared task (Bashendy et al., 2025a) focuses on dataset release rather than annotation methodology. To the best of our knowledge, no prior work has published Arabic annotation guidelines accounting for the full range of Arabic-specific writing features.

3 Annotation Framework Design

We designed a scoring framework that reflects the linguistic and rhetorical characteristics of Arabic academic writing, rather than adapting existing frameworks developed for other languages.

3.1 Design Principles

Our framework was guided by six principles: (1) **trait-based assessment**; (2) **text-type differentiation** between expository (نص تفسيري) and persuasive (نص إقناعي) writing; (3) **clear Arabic operationalization**, replacing vague criteria like "good organization" with observable indicators such as topic sentences (الجملة الموضوعية); (4) **granular but manageable scales**; (5) **evidence-based scoring** requiring textual justification; and (6) **cultural and linguistic appropriateness** for MSA norms and Arabic cohesive mechanisms, including reference (الإحالة), ellipsis/deletion (الحذف), substitution (الإبدال), and lexical cohesion (الترابط المعجمي).

3.2 Seven-Dimensional Scoring Framework

Table 1 summarizes our seven traits, adopted from the Core Academic Skills Test rubric (Bashendy

Trait	Scale	Description
REL	0–2	Relevance to Arabic prompt topic
ORG	0–5	Structure per Arabic essay conventions
VOC	0–5	Arabic lexical range, precision, MSA use
STY	0–5	Arabic cohesive devices & discourse patterns
DEV	0–5	Idea clarity, evidence, argument quality
MEC	0–5	Arabic spelling (إملاء), punctuation (ترقيم)
GRA	0–5	Syntactic variety & Arabic grammatical accuracy
HOL	0–32	Sum of all trait scores

Table 1: Seven-dimensional scoring framework. REL uses a 3-point scale; all other traits use 6-point scales (0–5). HOL is the sum of all traits.

et al., 2025b) and adapted for Arabic academic writing at the high-school level (grades 10–12).

VOC distinguishes limited vocabulary (المدى المحدود) from rich MSA usage, penalizing three error types: semantic mismatch (اختيار الكلمة غير المناسبة دلاليًا), collocation errors (أخطاء التراكيب المعجمية), and filler words (الكلمات الحشوية), while rewarding idiomatic expressions (التعبيرات الاصطلاحية) and implicit meaning (المعنى الضمني).

STY evaluates Arabic linear cohesion (الروابط الخطية) via four mechanisms: referencing (الإحالة), connective tools (الأدوات), deletion (الحذف), and substitution (الإبدال); and lexical cohesion (الترابط المعجمي) through repetition, synonymy, antonymy, and semantic field. The guidelines enumerate six Arabic organizational patterns (أنماط عرض الأفكار): cause-effect, compare-contrast, classification, chronological, interpretation, and pros-cons.

MEC covers Arabic-specific orthographic errors: hamza confusion (همزة القطع وألف الوصل), *ta marbuta* confusion (التاء المربوطة والهاء), and *ya* vs. *alif maqsura* (الياء والألف المقصورة).

GRA distinguishes simple Arabic structures(التراكيب البسيطة) from syntactic variety (التنوع التركيبي) including conditional, relative, and parenthetical constructions. If REL $= 0$, all other traits are automatically scored 0.

3.3 Expository vs. Persuasive Differentiation

The most critical design decision was distinguishing evaluation criteria for the two text types. Arabic expository writing demands neutrality and objectivity (حياد وموضوعية), while persuasive writing requires a clear position (موقف واضح) supported by evidence and rhetorical technique.

For **DEV**, expository essays are evaluated on clarity of explanation, depth of analysis, and objectivity (الحياد التفسيري). The guidelines warn annotators that first-person phrases "I believe" (أنا أرى), and "we should" (يجب علينا) constitute a deficiency in expository writing. Persuasive essays are evaluated on consistency of position, argument strength, evidence quality, persuasive techniques (الأساليب الإقناعية), and acknowledgment plus refutation of counterarguments (عرض الآراء المختلفة).

4 Guideline Development

Translating the scoring framework into a usable annotation instrument required an iterative development process grounded in authentic Arabic student writing, expert consensus, and empirical piloting.

4.1 Development Process

Guideline development proceeded through six phases over four months: (1) Rubric adoption (Weeks 1 - 2); (2) Arabic-specific adaptation for Style and Development (Weeks 3 - 4); (3) Expert exemplar development with two Arabic educators independently annotating 60 essays to consensus (Weeks 5 - 8); (4) Guidebook drafting in Arabic (Weeks 9 - 12); (5) External blind review by an Arabic language pedagogy expert, who evaluated the guidelines and provided structured evaluation leading to iterative revisions (Weeks 13 - 14); and (6) Pilot testing and refinement with three annotators on a subset of 20 essays (Weeks 15 - 16).

4.2 Guideline Components

The final guidebook, written entirely in Arabic to eliminate translation ambiguity and to ensure that annotators engage with scoring criteria in the same language as the essays they evaluate, comprises

four components: Conceptual Foundations, Scoring Descriptors, Annotated Arabic Examples, and Edge Case Protocols.

Conceptual Foundations The guidebook opens with a terminology unification section (توحيد المفاهيم والمصطلحات). **Text type definitions** provide full Arabic characterizations of expository and persuasive writing with a comparison table and multiple Arabic student writing examples. **Vocabulary terminology** (المفردات) defines all error types, idiomatic expressions, and implicit meaning, each illustrated with authentic student Arabic. **Style and cohesion terminology** (الأسلوب والتماسك البنائي) defines all four linear cohesion mechanisms and all six organizational patterns with Arabic examples and characteristic connective words. **Grammar terminology** (البناء والتراكيب) distinguishes simple from complex Arabic structures.

Scoring Descriptors For each trait and score level, we provide detailed Arabic descriptors derived from the original rubrics, organized into three complementary components: descriptor, performance characteristics, and illustrative example. As shown in Table 2, this structure specifies observable and measurable criteria while grounding them in representative examples of expected student responses, enabling annotators to anchor their judgments in concrete textual evidence. In addition, targeted guidance notes are included following each score level to support consistent interpretation of the criteria and to resolve potential ambiguities during scoring. This layered design; combining rubric-based descriptors, explicit performance features, illustrative examples, and practical annotation notes; aims to enhance clarity, reduce subjectivity, and improve inter-rater consistency by operationalizing abstract scoring criteria into interpretable and actionable annotation decisions.

Annotated Arabic Examples The guidebook includes multiple expert-annotated Arabic essays at each score level, covering both text types. Each example provides the complete Arabic text, all seven trait scores, and Arabic-language justifications with highlighted key features. Figure 1 in the Appendix shows a real annotated essay from the dataset for an expository prompt (P7: "Staying Up Late"), illustrating how the scoring descriptors are applied in practice.

Edge Case Protocols The guidebook provides explicit protocols for: **copied or AI-generated text** (score as written); **code-switching** between MSA and colloquial Arabic or English (score MSA portions; heavy non-Arabic content may affect REL and VOC); **very short essays** (<50 words; score on what is present); and **error-dense text** (score on what can be understood). Annotators are reminded that traits are independent: correct Arabic spelling does not compensate for weak content.

A recurring challenge is the use of Arabic enumeration markers (أولاً، ثانياً، وأخيراً) alone. Through calibration, guidelines were clarified to specify that ORG requires both structural elements *and* meaningful thematic connections, as shown in the Score 5 justification in Figure 1.

4.3 Addressing Subjectivity

We employed four Arabic-specific strategies to minimize annotator bias: (1) evaluate each trait independently to avoid halo effects from ornate Arabic style; (2) avoid over relying on essay length or classical Arabic expressions in isolation; (3) apply opposite standards for DEV depending on text type, rewarding objectivity in expository and position-taking in persuasive; and (4) provide diverse Arabic exemplars at each score level to illustrate multiple paths to the same score.

5 Annotator Training Protocol

Our five-stage protocol was implemented over four weeks before production annotation began.

Stage 1 – Conceptual Training (Week 1): Joint review of the Arabic guidebook, discussion of text-type distinctions with Arabic examples, examination of trait descriptors, and group exercises identifying Arabic features in sample essays.

Stage 2 – Expert Model Exposure (Week 2): Review of 8 expert-annotated Arabic essays (4 expository, 4 persuasive). Supervisors explained score assignments; discussion focused on adjacent score levels and borderline cases, e.g., distinguishing surface connectors (أدوات الربط) from genuine structural cohesion.

Stage 3 – Independent Practice (Week 3): Each annotator independently scored 12 Arabic practice essays (6 per text type) without access to expert scores, generating diagnostic data on individual understanding of Arabic-specific criteria.

Descriptor	Performance Characteristics	Illustrative Example	Score
Intro and conclusion absent. No organization or logical sequence. Ideas random or disconnected.	No clear introduction or conclusion. The text consists of a single paragraph or several unstructured sentences. Ideas are presented without sequence or connectors.	The student begins directly with: "People buy things online" (الناس يشترون من النت), then lists benefits in a scattered manner, with no ending.	1
Either intro or conclusion absent. Paragraphs lack logical progression. Weak organization attempt.	Only one of the introduction or conclusion is present; the other is completely absent or non-functional. The body consists of weakly connected or non-sequential paragraphs.	The student writes a general introduction such as: "Online shopping is beneficial" (الشراء الإلكتروني مفيد), then presents one or two ideas without clear connection or progression, ending with an incomplete phrase such as: "and so on" (وهكذا).	2
Both intro and conclusion present, but 1–2 body paragraphs lack coherence or connection to main idea.	Both introduction and conclusion are present, but not highly effective. The body includes one or two paragraphs with some coherence, though sequencing is weak or sometimes absent.	The introduction presents the topic in general terms; the paragraphs discuss two reasons for the phenomenon, but the connection between them is unclear; the conclusion repeats the initial idea without drawing a conclusion.	3
Appropriate intro and conclusion. 2–3 sequential, coherent body paragraphs. Minor transition issues only.	The introduction effectively introduces the topic and presents the main idea. The body is divided into clear paragraphs with logical progression and gradual development of ideas. The conclusion is appropriate and summarizes the points without repetition.	The student begins with an introduction such as: "With technological advancement, e-commerce has become a necessity" (مع تطور التكنولوجيا، أصبحت التجارة الإلكترونية ضرورة) then presents multiple reasons, each in a separate paragraph, and concludes with a sentence emphasizing the importance of the phenomenon.	4
Effective intro, strong conclusion, 2–3 body paragraphs with clear topic sentences and smooth transitions.	The introduction is direct and engaging (e.g., starts with a question, fact, or quotation). The body presents fully developed and interconnected ideas using cohesive devices. The conclusion provides a general insight or recommendation that highlights the discussion.	The student begins with: "Can we imagine a world without online stores?" (هل يمكن أن نتصوّر عالمًا بلا متاجر إلكترونية؟), develops paragraphs explaining the reasons for its spread in an organized manner, and concludes with: "Therefore, e-commerce is no longer an option, but a necessity" (لذلك، فإن التجارة الإلكترونية لم تعد خيارًا، بل ضرورة).	5

Table 2: Organization (ORG) scoring descriptors (translated from Arabic).

Stage 4 – Calibration and Discussion (Week 4): Calculation of individual agreement with expert scores; group discussion of high-disagreement Arabic cases; re-scoring of 4 essays together. A concrete calibration finding was the early confusion about enumeration markers: annotators initially awarded ORG $= 5$ for essays using "first, second, finally" (أولاً، ثانياً، أخيراً) without thematic paragraph development; calibration established that such essays should receive ORG $=$ 3–4 depending on paragraph coherence.

Stage 5 – Monitored Production (ongoing): Annotation production with dual annotation, weekly calibration sessions on challenging cases, and random control essay insertion (see Section 6).

5.1 Annotators

The annotation team consisted of six annotators and three supervisors. All members of the team were Arabic language teachers or lecturers with formal training in Arabic language education. Five members of the team held advanced degrees (MSc or PhD) in Arabic language or linguistics.

Annotators were responsible for the primary essay scoring tasks, while supervisors oversaw annotator training, quality assurance procedures, and dispute resolution during the annotation process. Supervisors also conducted periodic calibration sessions and monitored annotation quality through control essays and adjudication reviews.

5.2 Annotation Guidelines

All essays in the dataset were evaluated using the Core Academic Skills Test rubric developed by the Qatar University Testing Center (QUTC)[1]. The rubric evaluates seven writing traits: Relevance (REL), Organization (ORG), Vocabulary (VOC), Style (STY), Development (DEV), Mechanics (MEC), and Grammar (GRA). In addition, a Holistic score (HOL) was computed as the sum of the individual trait scores.

Six traits (ORG, VOC, STY, DEV, MEC, GRA) were rated on a 6-point scale (0 = lowest, 5 = highest), while Relevance (REL) was rated on a 3-point scale (0 = not relevant, 1 = partially relevant, 2 = fully relevant). If an essay received a REL score of 0, all remaining trait scores were automatically set to 0, since responses that do not address the prompt are not subject to further evaluation.

To ensure consistent interpretation of the rubric, two supervisors developed a comprehensive annotation guidebook containing detailed scoring terminology, annotated examples, and practice exercises for each prompt type.[2] Annotators were required to review the guidebook and complete structured training sessions before beginning annotation production.

Following training, moderation sessions were conducted in which annotators jointly reviewed sample essays, discussed scoring discrepancies, and harmonized interpretations of the rubric. These sessions ensured consistent application of the rubric across the annotation team before large-scale annotation began.

6 Quality Assurance Mechanisms

To ensure annotation reliability, we implemented several quality assurance procedures supervised by the senior annotation team.

Dual Annotation. Every essay was independently scored by two annotators (R1 and R2) under blind conditions using the Assessment Gourmet Platform[3], which anonymized student identity and prevented annotators from seeing each other's scores.

Discrepancy Resolution. If the difference between the two holistic scores (HOL) was less than 6 points (approximately 19% of the maximum possible score of 32), the scores were averaged. Larger discrepancies ($\geq$ 6 points) were escalated to a supervising annotator (R3), who performed adjudication and provided written feedback to the original annotators. This process served both as conflict resolution and as an ongoing learning mechanism.

Overall, 15.7% of essays required supervisor adjudication (range: 10.4%–23.3% depending on the prompt), as reported in Table 3.

Control Essay Monitoring. Approximately 5% of the essays assigned to each annotator were prescored expert control essays inserted without the annotator's knowledge. Inter-annotator agreement was monitored using Quadratic Weighted Kappa (QWK). If an annotator's QWK dropped below 0.60 over any 20-essay window, annotation was paused and additional calibration training was conducted with supervisors.

P#	Type	Essays	Avg Len	R3%
P1 Sports	EXP	1,122	162	10.4
P2 Social Media	PER	1,168	175	15.5
P3 Technology	PER	521	159	9.4
P4 Communication	PER	500	152	15.0
P5 Heritage	EXP	1,181	157	23.3
P6 Homework	PER	1,162	160	20.3
P7 Staying Up	EXP	1,143	202	10.6
P8 Video Games	PER	1,062	186	15.7
Total		**7,859**	**171**	**15.7**

Table 3: Dataset statistics. EXP/PER = expository/persuasive. R3% =percentage of essays requiring supervisor adjudication.

7 Results

We report results across two dimensions: the operational characteristics of the annotation process itself, and the inter-annotator agreement achieved under the finalized methodology.

7.1 Application Context

The methodology was applied to build an Arabic Automatic Essay Scoring dataset (Bashendy et al., 2025b): 7,859 essays by 4,372 high school students across 8 prompts (3 expository, 5 persuasive), annotated over one academic year by 6 annotators under 3 senior supervisors. Table 3 summarizes prompt-level statistics.

Adjudication patterns. Overall, 15.7% of essays required supervisor adjudication (R3), ranging from 10.4% (P1,Sports) to 23.3% (P5, Her-

[1] https://www.qu.edu.qa/sites/en_US/testing-center/TestDevelopment/cast

[2] The full annotation guidelines are available at: https://gitlab.com/bigirqu/laila/-/raw/main/rubrics/annotation_guidebook.pdf

[3] https://g-assess.com/

Trait	P1	P2	P3	P4	P5	P6	P7	P8
REL	.79	.60	.67	.59	.58	.68	.75	.77
ORG	.78	.72	.83	.77	.74	.78	.77	.78
VOC	.74	.71	.76	.69	.71	.75	.79	.80
STY	.74	.71	.76	.72	.60	.71	.68	.72
DEV	.79	.64	.58	.66	.70	.72	.71	.55
MEC	.69	.65	.77	.72	.58	.65	.69	.70
GRA	.71	.65	.77	.70	.70	.73	.67	.61
Avg	**.75**	**.67**	**.73**	**.69**	**.66**	**.72**	**.72**	**.70**

Table 4: Inter-annotator agreement (QWK) by prompt and trait. The majority of values fall in the substantial range (0.61–0.80); moderate values (<0.61) occur in nine prompt–trait cells, concentrated in REL and DEV across semantically open prompts (P2, P4, P5).

itage). This variation is not random: adjudication rates correlate with prompt framing breadth. Narrowly framed prompts with concrete phenomena (P1 Sports, P7 Staying Up Late) produced the lowest adjudication rates (10.4% and 10.6%), as the essay topic constrained valid interpretations of REL and DEV. Broadly framed prompts admitting diverse valid positions (P5 Heritage: *"balance between heritage and modernity"*) drove higher disagreement, confirming that prompt design is a significant but often overlooked source of annotation variance. We recommend pilot-testing prompts on a small essay sample and targeting adjudication rates below 15% as an indicator of adequate prompt specificity before large-scale annotation.

7.2 Inter-Annotator Agreement

We measured agreement between R1 and R2 (before adjudication) using Quadratic Weighted Kappa (QWK) following (Landis and Koch, 1977): <0.40 (poor), 0.40–0.60 (moderate), 0.61–0.80 (substantial), 0.81–1.00 (almost perfect). QWK is particularly appropriate for ordinal scales such as ours, as it penalizes disagreements proportionally to their distance on the scale, making it more sensitive than simple percent agreement for detecting systematic annotator bias.

Table 4 shows that the methodology achieved substantial agreement (≥0.61) in the large majority of prompt–trait combinations, with an overall average QWK of 0.71. Nine cells fall into the moderate range (<0.61): these are not distributed randomly but cluster in two traits (REL and DEV) and three prompts (P2, P4, P5), all of which involve persuasive writing on open-ended social topics. This pattern is theoretically coherent: both REL and DEV require annotators to exercise semantic judgment about argument quality and topical relevance, precisely the dimensions most sensitive to prompt framing.

Across traits, ORG achieved the highest average agreement (0.77), with no prompt falling below 0.72. This is consistent with our design decision to anchor ORG descriptors in structurally observable Arabic features; presence of introduction and conclusion, paragraph count, and explicit topic sentences; which leave less room for subjective interpretation. At the other end, DEV showed the highest variance across prompts (range: 0.55–0.79), reflecting the inherent difficulty of operationalizing Arabic argumentation quality, where annotators must assess not only what claims are made but whether evidence is culturally appropriate and rhetorically effective by MSA norms. The low DEV agreement on P8 (Video Games, 0.55) and P3 (Technology, 0.58) warrants attention: both are persuasive prompts on technology topics where student essays frequently blurred the boundary between anecdotal opinion and substantiated argument, a distinction our guidelines address but which remains challenging to apply consistently.

REL shows the widest prompt-level range of any trait (0.58–0.79), a finding that directly informs prompt design: prompts receiving lower REL agreement (P4: 0.59, P5: 0.58) are precisely those with broader, more abstractly framed topics, where the threshold for "partial relevance" (REL = 1) versus "full relevance" (REL = 2) is harder to determine. This suggests that REL agreement is as much a function of prompt specificity as of guideline quality, and future work should consider tightening prompt framing as a complementary strategy to guideline refinement.

7.3 Agreement Patterns

Trait-level variation. ORG achieved the highest average agreement (0.77), with no prompt falling below 0.72. This consistency reflects our design decision to anchor ORG descriptors in structurally observable Arabic features; presence of introduction and conclusion, paragraph count, and explicit topic sentences; which can be verified directly against the essay text and generalize well across both text types. DEV showed the lowest average agreement (0.66) and the highest variance across prompts (range: 0.55 - 0.79), reflecting the inherent subjectivity of judging Arabic argumentation quality - consistent with the prior educational assess-

ment literature (Williamson et al., 2012). The two lowest individual DEV values occur on persuasive prompts (P8: 0.55; P3: 0.58), where student essays frequently blurred the boundary between anecdotal opinion and substantiated argument. The correlation between DEV agreement and adjudication rate further supports this: prompts with higher R3% (e.g., P5 at 23.3%) consistently show lower DEV agreement in Table 3, suggesting that semantic trait subjectivity and prompt ambiguity are compounding rather than independent sources of annotation variance.

Prompt-level variation. Agreement ranged from 0.66 (P5 Heritage) to 0.75 (P1 Sports). The Heritage prompt's broad framing ("balance between heritage and modernity") admits diverse valid interpretations, making REL and DEV judgments harder and driving a 23.3% adjudication rate versus the project average of 15.7%.

Semantic vs. surface traits. Traits requiring deeper Arabic semantic understanding (DEV, REL) show lower agreement than surface-level traits (MEC, GRA), consistent with the challenge of codifying Arabic argumentative quality.

Learning curve. During the early stages of annotation, substantial variability was observed across annotators, reflecting differences in interpreting the scoring criteria. Following calibration, this variability decreased, and agreement improved further with continued practice. However, variability re-emerged with the introduction of new prompts, indicating the need for renewed alignment when encountering unfamiliar responses. This pattern is consistent with previous findings on the transition from guided calibration to independent scoring of unseen responses (Williamson et al., 2012). Moderation sessions between the supervisor and individual annotator helped restore consistency. Overall, these observations suggest that annotator agreement is affected by both task familiarity and the progressive clarification of the guidelines, with the final, post-calibration version supporting substantially higher and more stable agreement.

8 Discussion

Key success factors. Five elements proved most critical: (1) an Arabic-language guidebook with concrete student writing examples, especially for STY where abstract cohesion definitions required grounding; (2) the terminology unification section, which resolved early disagreements from inconsistent interpretation of Arabic linguistic terms; (3) text-type awareness training that calibrated opposite scoring mindsets for expository versus persuasive writing; (4) multi-stage training with calibration before production began; and (5) continuous quality assurance throughout the year-long annotation.

Challenges. The greatest challenge was semantic trait subjectivity: DEV and REL consistently yielded lower IAA than structural traits. We addressed this through extensive Arabic exemplars and evidence-based reasoning during training, but some residual subjectivity is unavoidable. Future work might decompose DEV into sub-traits (claim quality, evidence quality, counterargument handling). A second challenge was the steep learning curve when annotators first used the guidebook, especially for semantic traits like DEV and REL. Early calibration showed that some disagreements came from residual ambiguities in certain descriptors and edge-case protocols, not just trait subjectivity. We therefore refined the wording of these sections (for example, distinguishing superficial enumeration from genuine paragraph development in ORG and clarifying the threshold between partial and full relevance in REL) and expanded the Arabic borderline examples. These revisions increased and stabilized inter-annotator agreement across prompts, suggesting that the final guidelines were sufficiently clear despite the initial learning curve. Prompt design was also critical: the Heritage prompt's broad framing drove significantly higher adjudication (23.3%), informing our recommendation to pilot-test prompts before large-scale annotation.

Generalizability. While developed for MSA in a specific educational context, the core components - operational definitions in the target language, native-language exemplars, multi-stage training, dual annotation, continuous calibration - apply universally. Researchers adapting this approach to other Arabic varieties, proficiency levels, or languages should maintain these core components; shortcuts in any area will likely compromise annotation quality.

9 Conclusion and Future Work

We presented a systematic methodology for developing annotation guidelines for trait-based Arabic

AES, achieving substantial inter-annotator agreement (QWK: 0.66–0.75) across 7,859 Arabic essays. Our results demonstrate that reliable large-scale Arabic essay annotation is achievable when scoring criteria are operationalized in observable, Arabic-specific terms, annotators undergo structured calibrated training, and quality assurance is maintained continuously, evidenced by agreement rising before production began. Beyond the dataset, we offer a replicable template whose core components (target-language definitions, native exemplars, multi-stage calibration, and dual annotation) generalize to other morphologically complex, under-resourced languages. Several directions follow from the limitations identified above. First, decomposing DEV into sub-traits (claim quality, evidence quality, counterargument handling) may improve both reliability and model signal, given its high agreement variance (0.55–0.79). Second, our adjudication data suggest that prompt specificity is quantifiable; developing explicit prompt design criteria to target adjudication rates below 15% before deployment is a practical next step. Third, extending the framework to other Arabic varieties, narrative and descriptive genres, and university-level writing would broaden its applicability. Finally, applying AES models for automated pre-annotation could reduce annotator load while preserving reliability, particularly for surface traits where agreement is already high. Remaining challenges, particularly in semantic traits (DEV, REL), point to prompt design as a variable that guideline refinement alone cannot resolve. All materials are publicly released to lower the barrier for future annotation efforts in Arabic and comparable languages.

Limitations

Arabic variety: Guidelines target MSA in a specific educational context; adaptation may be required for other Arabic varieties. **Genre coverage:** Only expository and persuasive writing are addressed; narrative and descriptive Arabic writing require separate criteria. **Proficiency range:** Framework targets grades 10–12; university or younger students require recalibration. **Resource intensity:** Four weeks of training plus ongoing calibration may be prohibitive for low-budget contexts. **Essay length:** Essays averaged 171 words; effectiveness for longer compositions remains to be demonstrated.

Ethical Considerations

All essays were collected from high school students; participation was voluntary with consent obtained from students and guardians following institutional guidelines.[4]. No personally identifying information appears in the released dataset. Annotators were compensated fairly and workloads were monitored to prevent fatigue. The dataset is released for research use only.

Acknowledgments

We heartily thank our dedicated annotators for their contributions and express our gratitude to Qatar University Testing Center, the Ministry of Education and Higher Education in Qatar (the Arabic Section of the Department Of Educational Supervision in particular). This work was made possible by NPRP grant NPRP14S-0402-210127 from the Qatar Research Development and Innovation (QRDI) Council. The statements made herein are solely the responsibility of the authors.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

May Bashendy, Salam Albatarni, Sohaila Eltanbouly, Walid Massoud, Houda Bouamor, and Tamer Elsayed. 2025a. TAQEEM 2025: Overview of the first shared task for Arabic quality evaluation of essays in multi-dimensions. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*. To appear.

May Bashendy, Salam Albatarni, Sohaila Eltanbouly, Eman Zahran, Hamdo Elhuseyin, Tamer Elsayed, Walid Massoud, and Houda Bouamor. 2024. QAES: First publicly-available trait-specific annotations for automated scoring of Arabic essays. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 337–351, Bangkok, Thailand. Association for Computational Linguistics.

May Bashendy, Walid Massoud, Sohaila Eltanbouly, Salam Albatarni, Marwan Sayed, Abrar Abir, Houda Bouamor, and Tamer Elsayed. 2025b. LAILA: A large trait-based dataset for Arabic automated essay scoring. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics*. To appear.

Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Stindlová, and Chiara Vettori. 2014.

[4]IRB Number: QU-IRB 159/2024-EA)

The MERLIN corpus: Learner language and the CEFR. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1281–1288, Reykjavik, Iceland.

Scott Crossley, Yu Tian, Perpetual Baffour, Alex Franklin, Youngmeen Kim, Wesley Morris, Meg Benner, Aigner Picou, and Ulrich Boser. 2023a. The English language learner insight, proficiency and skills evaluation (ELLIPSE) corpus. *International Journal of Learner Corpus Research*, 9(2):248–269.

Scott Andrew Crossley, Perpetual Baffour, Yu Tian, Alex Franklin, Meg Benner, and Ulrich Boser. 2023b. A large-scale corpus for assessing written argumentation: PERSUADE 2.0. *Available at SSRN 4795747*.

Karen Fort, Gilles Adda, and K. Bretonnel Cohen. 2011. Amazon mechanical turk: Gold mine or coal mine? volume 37, pages 413–420.

Nizar Habash and David Palfreyman. 2022. ZAEBUC: An annotated Arabic-English bilingual writer corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 79–88, Marseille, France. European Language Resources Association.

Yaqiong He, Feng Jiang, Xiaomin Chu, and Peifeng Li. 2022. Automated Chinese essay scoring from multiple traits. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3007–3016, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Sandeep Mathias and Pushpak Bhattacharyya. 2018. Thank you for attending!: Attending to latent representation and modeling textual coherence in Automated Essay Scoring. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Rodrigo Wilkens, Alice Pintard, David Alfter, Vincent Folny, and Thomas François. 2023. TCFLE-8: A corpus of learner written productions for French as a foreign language and its application to automated essay scoring. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3447–3465, Singapore. Association for Computational Linguistics.

David M. Williamson, Xiaoming Xi, and F. Jay Breyer. 2012. A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1):2–13.

A Full Development Trait Descriptors

Expository Essays (النص التفسيري)

Score 1: Content largely unrelated to the topic. Ideas random, incoherent, no logical sequence. Main idea absent. No analysis. **Score 2:** Content somewhat related. Main idea disappears. Limited topic coverage. May contain personal opinion inappropriate for expository Arabic writing. **Score 3:** Completely on topic. Ideas mostly sequential but main idea fades. Some evidence but disorganized. Expository neutrality maintained but analytical depth insufficient. **Score 4:** Completely on topic. Ideas clear, organized, coherent. Main idea consistently connected to sub-ideas. Specific explanations and coherent supporting evidence, though not comprehensive. **Score 5:** Completely on topic throughout. Main idea strongly maintained. Comprehensive explanations. Multiple evidence forms. Full expository neutrality (حياد وموضوعية) maintained.

Persuasive Essays (النص الإقناعي)

Score 1: No clear position taken. Ideas random. No arguments or evidence. Persuasive techniques (الأساليب الإقناعية) entirely absent. **Score 2:** Position unclear or inconsistent. Limited argumentation, no evidence. Does not engage with alternative viewpoints. **Score 3:** Position stated but not consistently maintained. Arguments underdeveloped. Some evidence, not well-integrated. **Score 4:** Clear position adopted and maintained. Solid arguments with evidence. Some persuasive techniques. May not fully address counterarguments. **Score 5:** Strong, clear position throughout. Well-developed arguments with robust evidence (facts, examples, quotations, statistics). Effective persuasive techniques. Explicitly acknowledges and refutes counterarguments (عرض الآراء المختلفة والرد عليها).

B Example

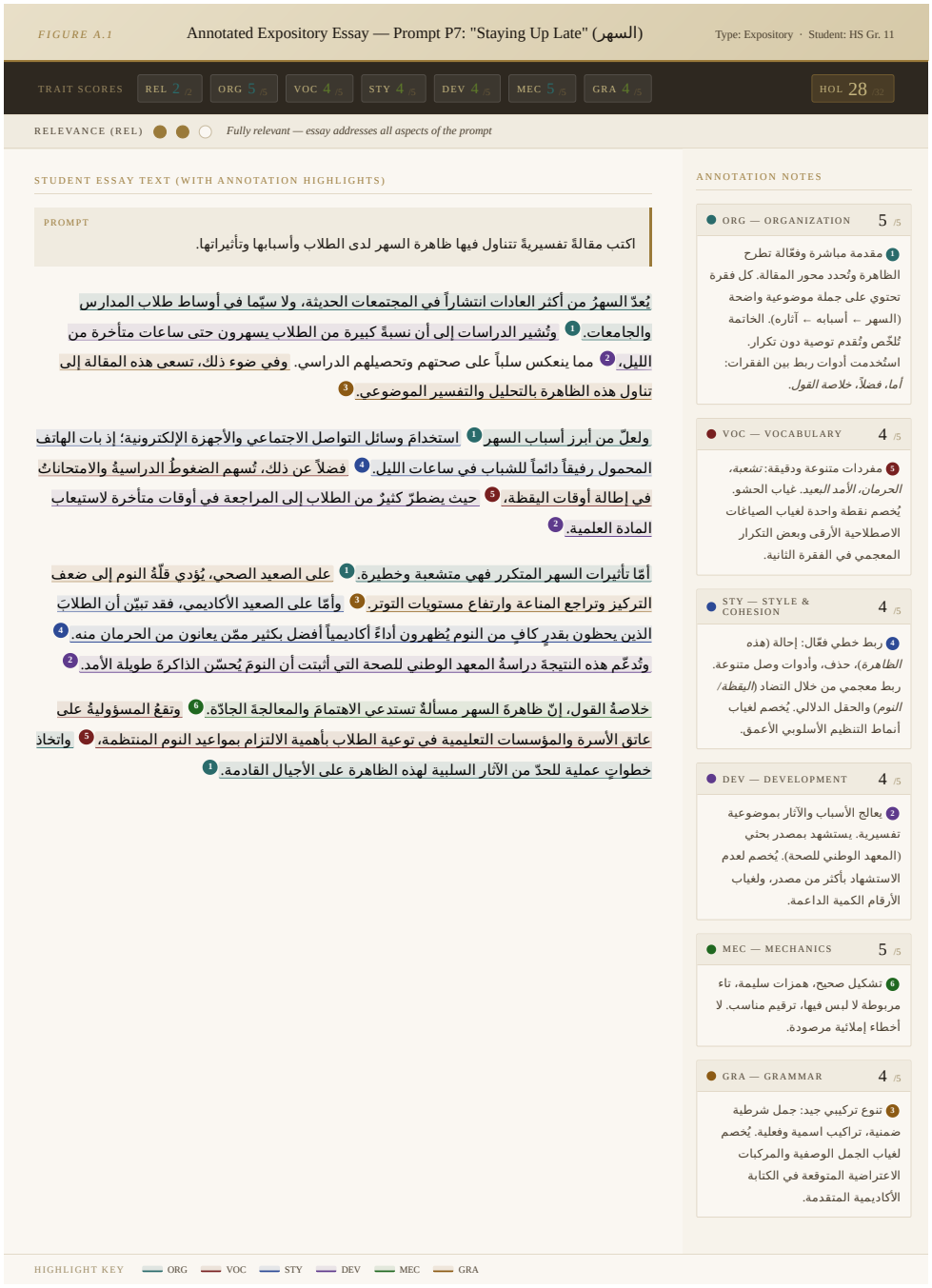

Figure 1: Expert-annotated expository student essay (P7: “Staying Up Late”, Grade 11). Colored highlights link textual evidence to the six scored traits; circled superscript numbers correspond to annotation notes in the right panel. REL = 2/2 (fully relevant); HOL = 28/32. Justifications are rendered in Modern Standard Arabic, matching the annotation protocol. ORG achieves the maximum score (5/5) owing to structurally observable features: a direct, engaging introduction; three body paragraphs each with an explicit topic sentence; and a conclusion offering a recommendation rather than mere repetition—consistent with the Score 5 descriptor in Table 2. The one-point deductions in VOC, STY, DEV, and GRA reflect specific, evidence-cited shortfalls rather than global impressions, operationalizing the evidence-based scoring principle.

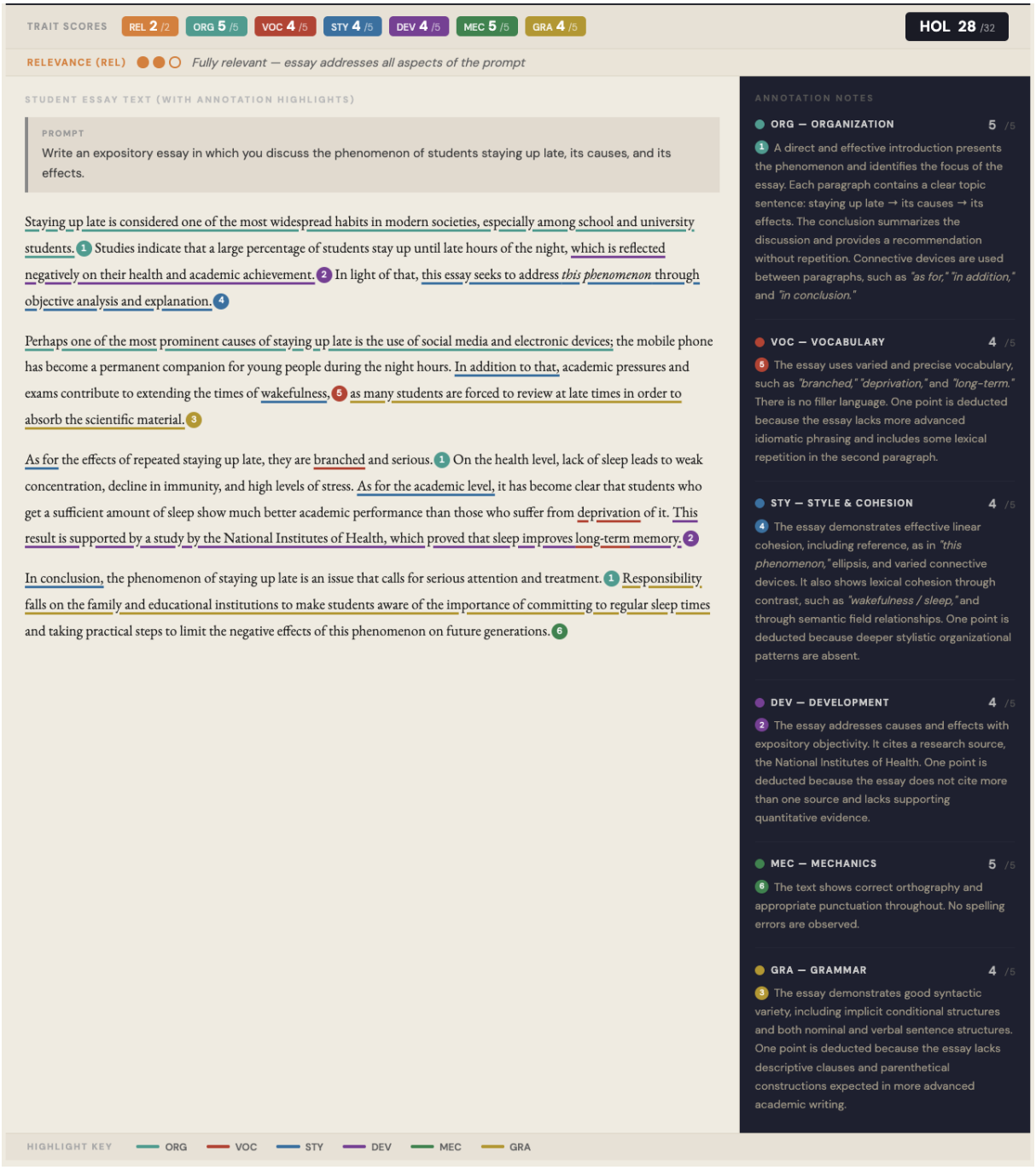

Figure 2: English translation of Figure 1: annotated expository essay with scoring rationale for each trait.