

Structured Feedback for Arabic Competitive Debates Using a Small Language Model

Ali Al-Zawqari¹, Ali Safa² and Gerd Vandersteen^{1,3}

¹Department of Fundamental Electricity and Instrumentation, Vrije Universiteit Brussel, Elsenne, 1050, Belgium

²College of Science and Engineering, Hamad Bin Khalifa University, Qatar

³Interuniversity Microelectronics Centre (imec), Kapeldreef 75, Leuven, 3001, Belgium

Abstract

Competitive debating is a non-formal educational activity that strengthens critical thinking, argumentation, and persuasive communication. Yet providing structured and consistent feedback remains difficult because speeches are delivered orally and adjudicators must evaluate them under time constraints. We present a rubric-aligned feedback framework for Arabic competitive debates that combines task-specific argument mining with an open-weight 9B Arabic small language model. Building on the Munazarat corpus and its hybrid annotation scheme, we convert segment-level predictions into interpretable speech-level indicators, including claim–support coverage, rebuttal density, repetition rate, and rhetorical balance. These indicators, together with selected transcript spans, are used to generate constrained, evidence-grounded feedback for adjudicators and coaches. To improve internal consistency, we compute debate winners deterministically from the same structured signals and return an *unclear* outcome when the score margin is below a fixed threshold. This conservative abstention mechanism reduces overconfident debate-level judgments while preserving human oversight. An expert study with debate coaches and adjudicators suggests that the framework can provide scalable and pedagogically aligned feedback, while also highlighting limitations in uncertainty communication.

Keywords

Debate education, Argument mining, Automated feedback, Small language models, Arabic NLP

1. Introduction

Argument mining provides practical methods for extracting structured argumentative components, such as claims, evidence, and rebuttals, from unstructured discourse [1]. One important application of these methods is automated feedback. Competitive debating is a particularly valuable setting for such work because it combines explicit speaker roles, strict turn-taking, and clear evaluation criteria [2, 3]. In real tournaments, however, the educational value of a round depends heavily on post-round feedback: adjudicators must assign speaker grades, justify the winner, and offer actionable guidance under strict time constraints. Because speeches are spoken and fast-paced, adjudicators often have limited ability to revisit where key claims were made, whether they were properly supported, and how effectively rebuttals engaged the opposing case.

Recent work on the Munazarat corpus introduced a hybrid annotation scheme for Arabic competitive debates that combines Toulmin-inspired argument elements with Aristotle’s rhetorical appeals (Logos, Ethos, and Pathos), enabling segment-level labeling of transcribed speeches [4, 5, 6]. However, segment-level predictions do not directly map to the rubric-aligned questions that adjudicators care about in practice, such as whether claims are sufficiently supported or whether rebuttals are targeted and proportional. At the same time, prior work on this domain shows that prompt-only zero- and few-shot language models perform poorly on Arabic debate classification, substantially trailing supervised Arabic encoder-based models [5]. This makes unconstrained end-to-end large language model (LLM) judging a weak foundation for reliable feedback in this setting.

For educational deployment, we focus on an open-weight 9B Arabic small language model (SLM), which offers practical advantages for reproducibility and controlled deployment. The SLM does not infer debate structure from scratch. Instead, Arabic task-specific classifiers recover argumentative elements,

SLM4ED’26: The 1st Workshop of Small Language Models for Education (SLM4ED), June 28, 2026, Seoul, Republic of Korea

✉ aalzawqa@vub.be (A. Al-Zawqari)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

and the SLM converts structured indicators and selected transcript spans into concise pedagogical feedback.

The main contribution of this paper is an educator-facing feedback framework for Arabic competitive debates. The framework maps supervised argument-mining predictions into rubric-aligned, role-aware indicators and uses an open-weight 9B Arabic SLM to generate concise feedback grounded in selected transcript spans. At the debate level, it uses a deterministic aggregation rule with an abstention region to avoid strong winner judgments for close scores. We also report a small pilot expert evaluation with debate coaches and adjudicators.

2. Background

2.1. Competitive Debates in Educational Settings

Competitive debating tournaments follow a standardized format intended for consistent adjudication and skill-focused training [6]. In the 3×3 Australia–Asia Debate format, two teams debate a single motion: proposition defends the motion and opposition contests it, with positions assigned for the round rather than reflecting personal beliefs. Each team consists of three speakers, and the debate proceeds in alternating order starting with the first proposition speech, followed by the first opposition speech, and so on until the first six speeches are completed; each of these speeches is limited to seven minutes. During these speeches, the opposing team may request short Points of Information (POIs), which are capped at 15 seconds. After the six main speeches, each team delivers a reply speech of three to four minutes, given by the second or third speaker. Following that, a panel with an odd number of adjudicators (e.g., 1–9) discusses and determines the winning team and assigns individual speaker scores, typically on a bounded scale (e.g., 59–81). The decision, based on majority voting, reflects which side more successfully establishes its case relative to the opponent under the format’s constraints.

2.2. Related Work

Argument mining has evolved from early rule-based and feature-driven systems [7] to neural approaches supported by larger corpora and standardized benchmarks [8]. Survey work reports strong progress in recovering claims, premises, and argumentative relations across genres, while also highlighting persistent challenges in robustness and cross-domain generalization [9]. Although contextualized representations and transfer learning have improved component classification and clustering, many pipelines still stop at span-level outputs rather than producing decision-level or feedback-oriented signals [10].

Debate settings sharpen these challenges because argumentative force depends on interaction: arguments are advanced, attacked, and reframed across turns. Work on online debates and political exchanges shows that dialogic context is central for stance and evidence modeling [11], motivating structured approaches that represent relations and dynamics in multi-participant argumentation [12]. For Arabic, progress has been constrained by the limited availability of annotated argumentative resources, since widely used corpora often emphasize speech or general NLP rather than argument structure [13], and student-focused argumentation datasets frequently target written production [14]. Recent work addresses this gap through Munazarat, which provides competitive debate transcripts and metadata [6], alongside a hybrid annotation scheme that combines Aristotle’s rhetorical appeals with an expanded Toulmin model adapted to debate discourse [4]. On top of these resources, prior work has examined encoder-based models for identifying persuasion styles in Arabic competitive debates as a non-formal educational activity [15], while subsequent studies report strong performance for persuasion-mode and argument-element classification in this domain [16, 5]. Importantly, recent results also show that zero- and few-shot 7B language models perform substantially worse than supervised Arabic encoders on Arabic debate classification, suggesting that prompt-only end-to-end LLM approaches are not yet a reliable basis for this task [5]. Another recent line of work emphasizes the need for cultural alignment in Arabic debate technologies, arguing that effective debate support should reflect

local communicative norms and values rather than directly reusing models or content designed for English-speaking contexts [17].

In parallel, artificial intelligence in education (AIEd) has long pursued scalable assessment and feedback, spanning psychometric foundations [18] and data-driven teacher support [19]. Automated evaluation has been explored for diverse educational artifacts, yet many settings still require expert interpretation and rubric alignment [20]. With the rise of language models, recent surveys highlight the promise of natural-language feedback alongside risks of ungrounded or inconsistent recommendations [21, 22]. These concerns, together with fairness requirements in AIEd, motivate reliability-aware and equity-aware feedback designs [23, 24].

Building on these threads, our work targets a gap between segment-level Arabic debate analysis and the kind of rubric-aligned feedback that adjudicators and coaches need in practice. We address this gap by mapping argument-mining outputs into interpretable speech-level indicators and using an open-weight 9B Arabic SLM in a constrained role to generate evidence-grounded feedback that remains coherent under uncertainty.

3. Methodology

Figure 1 summarizes the proposed pipeline. Given an Arabic debate transcript, we 1) decompose the transcript into speeches, 2) segment and classify spans into argumentative labels using encoder-based Arabic models, 3) compute rubric-aligned indicators at the speech level, 4) generate per-speech feedback with evidence grounding using an open-weight 9B Arabic SLM, and 5) produce a debate-level winner decision and brief justification from the same structured signals. This design assigns different roles to different model classes: task-specific models recover argumentative structure, while the SLM is used only for constrained pedagogical synthesis.

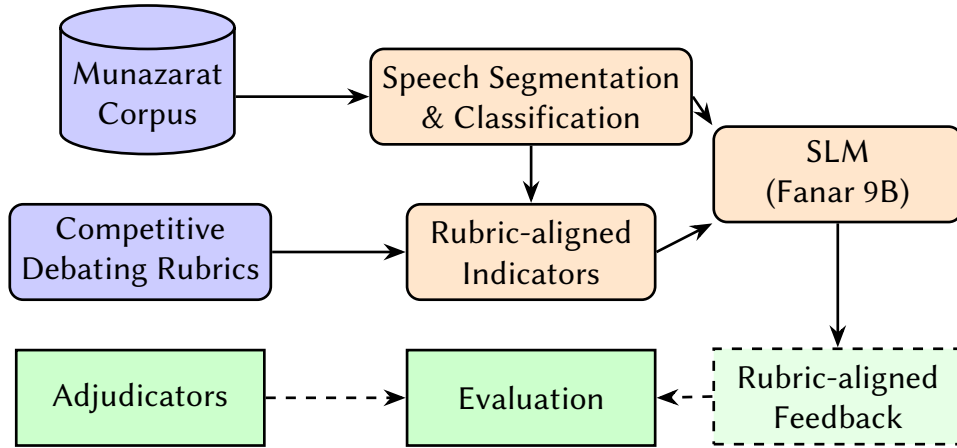


Figure 1: Pipeline of the proposed rubric-aligned feedback framework. Blue denotes data/rubric resources, orange denotes automated processing modules, and green denotes human involvement and outputs (dashed arrows indicate evaluation).

3.1. Dataset and Transcript Structure

We build on the Munazarat corpus of Arabic competitive debates [6], which provides transcribed debates with a fixed multi-speaker structure and span-level annotations¹. Each debate is loaded from a plain-text file. We identify speech boundaries using transcript speaker headers and extract exactly six speeches (i.e., we do not process reply speeches in the current implementation). Speech segmentation assigns roles in the canonical order:

$$\mathcal{R} = \{1st_prop, 1st_opp, 2nd_prop, 2nd_opp, 3rd_prop, 3rd_opp\}.$$

¹https://github.com/NoobleonM/annotated_munazarat

Table 1

Label groupings used to compute rubric-aligned indicators.

Group	Labels
Claims $\mathcal{L}_{claim}$	CLAIM, SUPER CLAIM
Supports $\mathcal{L}_{supp}$	WARRANT, RATIONAL BACKING, EVIDENTIAL BACKING, GROUND, QUALIFIER
Rebuttals $\mathcal{L}_{reb}$	REBUTTAL
Repetition $\mathcal{L}_{rep}$	SELF-REPETITION, TEAM REPETITION, RECITATION
Logos $\mathcal{L}_{logos}$	POSITION, GROUND, QUALIFIER, SUPER CLAIM, CLAIM, WARRANT, RATIONAL BACKING, EVIDENTIAL BACKING
Ethos $\mathcal{L}_{ethos}$	ETHOS
Pathos $\mathcal{L}_{pathos}$	PATHOS $\cup \mathcal{L}_{rep}$

Table 2

Speech-level indicators used in feedback generation and debate summarization.

Indicator	Interpretation
Cov(s)	Fraction of claims with nearby support
Reb(s), Rep(s)	Rebuttals per 1,000 words, Repetition/recitation per 1,000 words
(Logos, Ethos, Pathos)	Proportions over labeled span length

For each speech s , we store its role $r(s) \in \mathcal{R}$, a speech index, the extracted speech text T_s , and the debate title.

3.2. Span-level Segmentation and Classification

For each speech, we apply the span-level segmentation and classification models from [5]. Model outputs are represented as tuples (b_i, e_i, ℓ_i) , where (b_i, e_i) are character offsets in the transcript and ℓ_i is a label from the hybrid scheme used in Munazarat [25]. We denote the set of labeled spans within a speech s by $A_s = \{(b_i, e_i, \ell_i)\}$. This section of the pipeline provides the structured input required to compute speech-level indicators and to ground feedback in explicit transcript spans.

3.3. Rubric-Aligned Indicator Design

We bridge span-level labels to rubric-aligned signals by grouping labels into functional sets (Table 1). These sets are designed to capture core judging dimensions: 1) whether claims are supported, 2) whether the speech engages the opposing case through rebuttal, 3) whether it relies on repetition or recitation, and 4) the rhetorical distribution across Logos, Ethos, and Pathos. For each speech s , we also compute a word count W_s from T_s (via whitespace tokenization) to normalize span counts by speech length.

We compute four speech-level indicators (Table 2). *Claim-support coverage* estimates whether argumentative claims are accompanied by nearby support within the speech. *Rebuttal density* and *repetition/recitation density* capture how often the speaker performs refutation or relies on repetitive content, normalized per 1,000 words. Finally, *rhetorical balance* reports the distribution of labeled span length across Logos, Ethos, and Pathos.

3.4. Role-Aware Normalization

Because speaking roles differ in competitive debates, raw indicator values can systematically favor certain roles (e.g., the first proposition speaker is not expected to refute). To account for these structural differences, we compute role-conditioned z-scores for each metric in Table 2 using corpus-wide statistics for each role in $\mathcal{R}$. These role-normalized values are used in two places: 1) as compact contextual signals in the per-speech feedback prompt, and 2) as the basis for aggregating performance when producing debate-level outputs.

3.5. SLM-Based Feedback Generation and Deterministic Winner Output

For each speech, we attach short evidence excerpts to ground the generated rationale. We select a small number of representative spans per category (claim, support, rebuttal, repetition) and include a short surrounding context, yielding evidence identifiers such as [CLAIM_1] and [REBUTTAL_2]. We then prompt *Fanar Prime 9B* [26] with: 1) the debate title and speaker role, 2) the computed indicators together with their role-normalized values, 3) the evidence snippets, and 4) the full speech text. We intentionally restrict the role of the SLM to feedback generation rather than argument classification, because prior work on this domain shows that prompt-only zero- and few-shot language models perform poorly on Arabic debate classification compared with supervised Arabic encoder-based models [5].

The output is constrained to JSON with keys {*rubric_level*, *rubric_rationale*}. The rubric level is selected from six ordered categories (Very weak, Weak, Below average, Average, Above average, Professional), and the rationale is limited to a small number of short lines. Each rationale line must reference at least one indicator and at least one evidence identifier, which encourages grounded feedback and reduces generic statements.

At the debate level, we compute the winner deterministically from the same speech-level outputs to avoid contradictory judging. Specifically, we aggregate rubric levels and role-normalized indicators within matched speaker positions (e.g., first proposition vs. first opposition) and combine these differences into a single debate score. The winner label is then derived from this score, with an explicit *unclear* outcome used when the margin is below a threshold. Finally, the SLM is used only to generate a brief justification that is constrained to reference the computed score summary, ensuring that the textual rationale remains consistent with the deterministic decision.

3.6. Reliability Gate and Uncertainty Handling

The reliability gate operates at the debate level. After aggregating the matched speaker-position differences into a single debate score D , the framework returns Proposition when $D > 0.75$, Opposition when $D < -0.75$, and *Unclear* otherwise. The interval $[-0.75, 0.75]$ acts as an abstention region for close decisions. This mechanism is intentionally conservative: it prevents the system from presenting a strong debate-level judgment when the aggregate margin is small. It does not replace adjudicator judgment and does not fully solve the communication of uncertainty in the generated feedback.

3.7. Expert Evaluation Protocol

For evaluation, we construct debate-level files that include: 1) the predicted winner with a short justification, and 2) per-speech rubric levels with evidence-grounded rationales. Experts independently adjudicate selected debates and then review the system output as a potential aid for adjudication and coaching. They answer a small set of closed questions covering a) perceived correctness, b) usefulness/actionability, and c) appropriateness of uncertainty expression. We summarize responses using descriptive statistics, as the study is small and the questionnaire uses ordinal response categories.

4. Results and Discussion

We conduct a descriptive expert study on a selected subset of six debates (36 speeches) from university-level competitions. Official tournament outcomes were not used to generate the feedback or the framework decisions. They were used only for post-hoc comparison and for selecting illustrative cases that cover three situations: agreement with the official record, disagreement with the official record, and abstention with an *unclear* framework decision. Because the subset is purposively selected and small, the reported proportions should be interpreted as descriptive case-study observations rather than estimates of general system accuracy.

We consulted four experts who started as competitive debaters and later became professional adjudicators and coaches, and we ensured gender diversity in recruitment. Each expert adjudicated

Table 3

Winner decisions on the expert-evaluation subset (6 debates from Munazarat). P: Proposition, O: Opposition, and U: Unclear.

Debate ID	Official Records	Framework	Expert 1	Expert 2	Comment
1	P	P	P	P	Match across all sources.
2	P	O	O	P	Experts disagree.
4	P	O	O	O	Experts agree with the framework.
20	O	O	P	O	Experts disagree.
21	O	U	O	O	Official record matches experts.
38	O	U	P	P	Official record contradicts experts.

Table 4

Speaker score-band agreement between the framework and experts, measured as overlap between framework intervals and expert scores mapped to official score intervals.

Role	Overlap (hits/total)	Median $ \Delta $ (midpoints)
1st_prop	8/12 (66.7%)	1.0
2nd_prop	5/12 (41.7%)	4.0
3rd_prop	9/12 (75.0%)	0.0
1st_opp	7/12 (58.3%)	1.0
2nd_opp	7/12 (58.3%)	1.0
3rd_opp	8/12 (66.7%)	1.0
All roles	44/72 (61.1%)	2.0

three debates, so that each debate received two independent adjudications. The estimated duration per debate was approximately one hour, including watching the debate and submitting the evaluation (12 expert-hours in total). Given the small scale of the study, we focus on descriptive analysis of two aspects: 1) agreement between the framework and experts on the debate winner and the plausibility of the produced speaker score ranges, and 2) summary statistics of the experts’ responses to the closed questions.

4.1. Winner Decisions

Table 3 reports winner decisions for the six debates. On the four debates where the framework returns a decided winner, it agrees with 6 out of 8 expert winner judgments. Two debates (Debate 2 and Debate 20) show split expert decisions, indicating that these rounds are contested even under human adjudication. Debate 4 differs from the official record but matches both experts. For the two debates where the framework outputs *Unclear* (Debate 21 and Debate 38), both experts provide a decided winner, suggesting that the system’s abstention behavior is conservative on this subset. Debate 38 also shows a mismatch between evaluation targets: both experts select Proposition while the official record indicates Opposition. This further motivates reporting expert agreement alongside official outcomes when evaluating debate-level decisions.

4.2. Speaker Score Range Plausibility

We next assess whether the framework’s rubric-aligned speaker ranges are plausible relative to expert scoring. Since expert scores are mixed (some are point scores and others are ranges), we map all expert scores to the official tournament score intervals used for speaker evaluation. Point scores are mapped to their corresponding interval, and expert-provided ranges are mapped to the minimal interval that covers their endpoints. We then measure agreement by interval overlap between the mapped expert interval and the framework interval. Across all speeches and experts (72 comparisons), the framework overlaps with expert intervals in 44 cases (61.1%). Overlap is higher when experts provide ranges (20/25, 80.0%) than when they provide point scores (24/47, 51.1%), which is expected given that the framework outputs score bands rather than exact grades. Table 4 breaks the agreement down by role.

4.3. Expert Questionnaire

Experts answered four closed questions after reviewing the framework outputs (12 total responses). For Q1 (directional correctness of the judgment), 8/12 responses indicate that the decision is correct in the overall direction (options 1–2), 3/12 judge the direction as incorrect (option 4), and 1/12 selects the uncertain option (option 3). This pattern is consistent with the winner analysis in Table 3, where some debates are contested even among experts (Debates 2 and 20), and where official outcomes can differ from expert adjudication (Debate 38).

For Q2 (the adjudication level reflected by the feedback), responses concentrate at the lower levels: 5/12 rate it as beginner-level and 7/12 as intermediate-level, with no responses selecting the professional level. For Q3 (usefulness/actionability), no response indicates that the feedback is directly usable as-is; most responses place it in the middle categories (4/12 usable after minor edits, 5/12 useful as a secondary reference), while 3/12 judge it not usable. Taken together, these responses suggest that the framework can support adjudication and coaching, but does not yet reach the level of polished expert feedback.

Finally, Q4 (appropriateness and clarity of uncertainty expression) shows that uncertainty is not communicated consistently: only 2/12 responses rate it as appropriate and clear, 3/12 fall in the middle categories, and most responses select the final option (7/12). The debate-level abstention region prevents strong winner predictions when the aggregate margin is small, but it does not explain which speech-level observations are tentative or why uncertainty arises.

5. Conclusion

This paper presented a rubric-aligned automated feedback framework for Arabic competitive debates. Building on span-level argument predictions, we mapped segment outputs into speech-level indicators reflecting core adjudication criteria. These indicators were used to generate evidence-linked feedback, where rubric judgments are connected to selected transcript spans. To improve internal consistency, debate-level outcomes were computed using a deterministic aggregation rule with an explicit abstention region for close scores, while Fanar Prime 9B was used to produce concise, score-consistent feedback and explanations.

Results from a pilot expert study suggest that the proposed framework can support structured feedback in competitive debates, especially as an aid for adjudicators and coaches rather than as a replacement for expert judgment. On debates where a decided winner was produced, framework decisions largely aligned with expert adjudications, and speaker score bands showed moderate overlap with expert scoring. At the same time, the questionnaire results indicate clear limitations: the feedback was generally viewed as beginner- to intermediate-level, not directly usable as-is, and insufficiently explicit in its uncertainty expression. The pilot results should be interpreted cautiously because the expert study is small and the evaluated debates were purposively selected to illustrate different framework behaviors.

Overall, these findings suggest that open-weight SLMs can support scalable and transparent feedback when grounded in structured indicators and explicit evidence, while also highlighting the need for better uncertainty communication and broader evaluation.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT for grammar and spelling checks, paraphrasing and rewording, improving writing style, and formatting assistance, including LaTeX formatting and TikZ positioning. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

Acknowledgments

The authors thank Ataa Barazi, Baraa Al-Ahmar, Fatma Sallabi, and Qamar Bader for their time and feedback during the expert evaluation. The authors also thank QatarDebate Center for providing funding and resources for this research.

References

- [1] M. Lippi, P. Torroni, Argumentation mining: State of the art and emerging trends, *ACM Transactions on Internet Technology (TOIT)* 16 (2016) 1–25.
- [2] J. Bellon, A research-based justification for debate across the curriculum, *Argumentation and Advocacy* 36 (2000) 161–175.
- [3] J. S. Huryn, Debating as a teaching technique, *Teaching Sociology* 14 (1986) 266–269.
- [4] A. G. Al-Sharafi, M. Majed Khader, M. Hamza Al-Sioufy, M. Ahmed, A. Al-Zawqari, Arabic argumentation: Presentation or proof, *Informal Logic* 45 (2025) 194–228.
- [5] A. Al-Zawqari, M. Ahmed, A. G. Al-Sharafi, M. M. Khader, A. Safa, G. Vandersteen, Neural classification of argument elements and styles in Arabic competitive debates, *IEEE Access* (2025).
- [6] M. M. Khader, A. Al-Sharafi, M. H. Al-Sioufy, W. Zaghouni, A. Al-Zawqari, Munazarat 1.0: A corpus of Arabic competitive debates, in: *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) (LREC-COLING)*, 2024, pp. 20–30.
- [7] R. M. Palau, M.-F. Moens, Argumentation mining: the detection, classification and structure of arguments in text, in: *Proceedings of the 12th international conference on artificial intelligence and law*, 2009, pp. 98–107.
- [8] L. Cheng, L. Bing, R. He, Q. Yu, Y. Zhang, L. Si, Iam: A comprehensive and large-scale dataset for integrated argument mining tasks, *arXiv preprint arXiv:2203.12257* (2022).
- [9] J. Lawrence, C. Reed, Argument mining: A survey, *Computational Linguistics* 45 (2019) 765–818.
- [10] O. Toledo-Ronen, M. Orbach, Y. Bilu, A. Spector, N. Slonim, Multilingual argument mining: Datasets and analysis, in: *Findings of the association for computational linguistics: Emnlp 2020*, 2020, pp. 303–317.
- [11] M. A. Walker, P. Anand, R. Abbott, J. E. F. Tree, C. Martell, J. King, That is your evidence?: Classifying stance in online political debate, *Decision Support Systems* 53 (2012) 719–729.
- [12] Z. Guo, M. P. Singh, Representing and determining argumentative relevance in online discussions: A general approach, in: *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, 2023, pp. 292–302.
- [13] M. Al-Fetyani, M. Al-Barham, G. Abandah, A. Alsharkawi, M. Dawas, Masc: Massive Arabic speech corpus, in: *2022 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2023, pp. 1006–1013.
- [14] W. Zaghouni, A. Ahmed, X. Zhang, L. Rezk, Qcaw 1.0: Building a qatari corpus of student argumentative writing, in: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, 2024.
- [15] A. Al-Zawqari, M. Ahmed, G. Vandersteen, Evaluation with language models in non-formal education: Understanding student’s persuasion style in competitive debating, in: *EvalLAC@ AIED*, 2024.
- [16] A. Al-Zawqari, A. G. Al-Sharafi, M. Ahmed, M. M. Khader, G. Vandersteen, Classifying persuasion modes in Arabic debates: a preliminary language model-based analysis, in: *International Conference on Arabic Language Processing*, Springer, 2024, pp. 242–252.
- [17] K. Al Khatib, M. Khader, Toward culturally-aware Arabic debate platforms with nlp support, in: *Proceedings of The Third Arabic Natural Language Processing Conference*, 2025, pp. 359–374.
- [18] R. K. Hambleton, H. Swaminathan, *Item response theory: Principles and applications*, Springer Science & Business Media, 1985.

- [19] R. Luckin, W. Holmes, M. Griffiths, L. B. Forcier, *Intelligence unleashed: An argument for ai in education* (2016).
- [20] L. Benedetto, P. Cremonesi, A. Caines, P. Buttery, A. Cappelli, A. Giussani, R. Turrin, A survey on recent approaches to question difficulty estimation from text, *ACM Computing Surveys* 55 (2023) 1–37.
- [21] A. Caines, L. Benedetto, S. Taslimipoor, C. Davis, Y. Gao, O. Andersen, Z. Yuan, M. Elliott, R. Moore, C. Bryant, et al., On the application of large language models for language teaching and assessment technology, *arXiv preprint arXiv:2307.08393* (2023).
- [22] J. Jeon, S. Lee, Large language models in education: A focus on the complementary relationship between human teachers and chatgpt, *Education and Information Technologies* 28 (2023) 15873–15892.
- [23] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, et al., Chatgpt for good? on opportunities and challenges of large language models for education, *Learning and individual differences* 103 (2023) 102274.
- [24] J. Lee, Y. Hicke, R. Yu, C. Brooks, R. F. Kizilcec, The life cycle of large language models in education: A framework for understanding sources of bias, *British Journal of Educational Technology* 55 (2024) 1982–2002.
- [25] A. G. Al-Sharafi, M. M. Khader, M. Ahmed, M. H. Al-Sioufy, W. Zaghouani, A. Al-Zawqari, A hybrid annotation model for Arabic argumentative debate corpus, in: *International Conference on Arabic Language Processing*, Springer, 2024, pp. 97–113.
- [26] F. Team, U. Abbas, M. S. Ahmad, F. Alam, E. Altinisik, E. Asgari, Y. Boshmaf, S. Boughorbel, S. Chawla, S. Chowdhury, et al., Fanar: An Arabic-centric multimodal generative ai platform, *arXiv preprint arXiv:2501.13944* (2025).