Toward Culturally-Aware Arabic Debate Platforms with NLP Support

Khalid Al-Khatib
University of Groningen
khalid.alkhatib@rug.nl

Mohammad Khader
QatarDebate Center
mkhader@qatardebate.org

Abstract

Despite the growing importance of online discourse, Arabic-speaking communities lack platforms that support structured, culturally grounded debate. Mainstream social media rarely fosters constructive engagement, often leading to polarization and superficial exchanges. This paper proposes the development of a culturally aware debate platform tailored to the values and traditions of Arabic-speaking users, with a focus on leveraging advances in natural language processing (NLP). We present findings from a user survey that explores experiences with existing debate tools and expectations for future platforms. Besides, we analyze 30,000 English-language debate topics using large language models (LLMs) to assess their cultural relevance and appropriateness for Arab audiences. We further examine the ability of LLMs to generate new culturally resonant debate topics, contributing to the emerging tasks of culture-aware topic assessment and generation. Finally, we propose a theoretical and technical framework for building an NLP-supported Arabic debate platform. Our work highlights the urgent need for culturally sensitive NLP resources that foster critical thinking, digital literacy, and meaningful deliberation in Arabic.

1 Introduction

Online debate platforms foster structured argumentation and the exchange of diverse viewpoints. They allow users to present claims, support them with evidence, and engage in critical dialogue. By encouraging deliberation and reasoned disagreement, such platforms strengthen public discourse, offering an alternative to the fragmented and polarized interactions typical of social media (Kriplean et al., 2012; Frappier et al., 2024). These problems are also evident in Arabic social media, where false information, conspiracy theories, and divisive content are widespread (Milli et al., 2025; Fawzi et al., 2026; Abouzied et al., 2025).

Despite their potential, structured debate platforms are largely unavailable or ill-suited for Arabic-speaking communities. Most existing platforms are designed for English-speaking users and fail to reflect the linguistic and cultural norms of the Arab world. As a result, Arabic online discussions often lack structure, critical engagement, and depth, leading instead to polarization, misinformation, and unproductive dialogue. This gap hinders meaningful civic discourse and the development of argumentation skills in Arabic digital spaces.

This paper takes a first step toward addressing this gap by proposing AI-powered debate platforms tailored for Arabic-speaking users. Such platforms must go beyond translation: they should reflect Arabic cultural values, discourse traditions, and social norms. They should also support structured interaction, encourage evidence-based reasoning, and foster cross-cultural understanding by treating argumentation as both a civic and cultural practice.

Recent advances in NLP offer promising tools to support such platforms. NLP can help users build arguments, find supporting evidence, identify counterpoints, and follow the flow of debate. It can also support cultural alignment by generating or filtering content that reflects Arab values. These capabilities make NLP a key component in developing culturally aware debate technologies.

To guide the development of Arabic debate platforms, we investigate four interrelated questions:

(1) *Do Arabic speakers see a need for culturally grounded debate platforms, and what features do they expect?* We address this through a preliminary survey that explores Arabic speakers' experiences with online debate platforms, their expectations for core functionalities, and their openness to AI-assisted interactions.

(2) *Do existing English-language debate platforms include topics that are culturally relevant or appropriate for Arabic audiences?* To contextualize user expectations, we analyze 30,000 debate

Proceedings of The Third Arabic Natural Language Processing Conference, pages 359–374

topics from prominent English-language platforms. Using three LLMs, we assess the cultural specificity and appropriateness of these topics, identifying whether they resonate with Arabic values or reflect mismatches that highlight the need for dedicated platforms.

(3) *Can LLMs generate culturally specific and resonant debate topics for Arabic users?* Given the limitations of existing content, we examine whether LLMs can generate debate topics that align with Arabic cultural and social contexts. This preliminary exploration considers the potential of LLMs and informs future strategies for culturally aware content creation.

(4) *What are the essential components of such a platform, and how can Arabic NLP contribute to its development?* Based on the insights gained from addressing the previous questions, we outline the technical and conceptual foundations needed for a culturally aware Arabic debate platform. We determine core NLP tasks, such as argument mining, human value detection, and topic generation, and consider the readiness of current Arabic NLP resources to support them.

Our findings reveal both societal demand and technical opportunity for culturally grounded Arabic debate platforms. Arabic speakers express strong interest in structured, AI-supported tools for meaningful discourse, while also emphasizing the importance of cultural sensitivity and minimizing AI bias. We show that existing platforms rarely address Arabic-specific topics, yet LLMs demonstrate promising capabilities in generating culturally relevant debate topics. This work lays the foundation for a new research direction at the intersection of Arabic NLP, computational argumentation, and culturally aware AI systems. All resources developed in this paper are available online[1].

2 Related Work

We review related work across four key areas: online debate and argumentation platforms, computational modeling of debate structures and content, AI integration in online communication, and Arabic argument mining.

Online Debate and Argumentation Platforms Several structured platforms have emerged to support public debate, argument exchange, and educational discussion. *Kialo*[2] is widely recognized for its graph-based interface that organizes debates into tree structures of pro and con arguments, enabling visual navigation and collaborative reasoning. *iDebate*'s Debatabase offers structured pro–con arguments for hundreds of debate motions and is widely used in formal debate training. Similarly, the *Kialo Edu Topics Library* and the *Kialo Edu Blog* provide classroom-ready prompts and debate templates for educators, covering domains such as ethics, technology, and education.

ChangeMyView[3] offers a more informal but constructive setting, where users post opinions and invite challenges. Persuasive responses are rewarded with "deltas," making it a valuable resource for studying persuasion strategies.

Beyond these structured platforms, repositories such as *DebateData* curate thousands of competitive debate motions used in tournaments, providing a rich source of real-world argumentative topics. *Britannica ProCon* supplements this with balanced summaries, evidence, and statistics on controversial public-policy issues, while *I Side With* offers issue quizzes and opinion research with ideology breakdowns and analytics. These platforms serve as debate resources and also as empirical foundations for argumentation research.[4]

Argument Structure and Computational Modeling Structured platforms such as *Kialo* and *ChangeMyView* have been instrumental in computational argumentation research. (Agarwal et al., 2022) used Kialo data to model argument polarity, while (Boschi et al., 2021) developed graph-based sampling strategies to extract high-quality arguments. The moderation and structure of Kialo discussions enable detailed modeling of argument relations, positions, and discourse flow (Mezza et al., 2024; Ghafouri et al., 2023). *iDebate*'s database has been used to train models for identifying argumentative roles such as claims and premises (Al-Khatib et al., 2016a; Hua and Wang, 2017), and *ChangeMyView* has supported research into persuasion analysis (Al-Khatib et al., 2020).

AI Integration in Communication Platforms LLMs have increasingly been integrated into online platforms for moderation, guidance, and content enrichment. (Ye et al., 2023) introduce a multilingual Reddit moderation dataset and analysis,

[1] https://github.com/Arabic-Argument-Mining/ArabicNLP25

[2] https://www.kialo.com

[3] https://www.reddit.com/r/changemyview

[4] URLs of the debate platforms can be found in Table 1.

while (Lee et al., 2024) survey and systematize AI writing assistants that provide real-time suggestions for tone, evidence, and argumentative clarity. Although such systems improve online discourse, most lack cultural sensitivity or adaptability to non-Western norms.

Arabic Argument Mining and Culture-Aware Argumentation Arabic argument mining remains an underexplored area, with only a limited number of high-quality resources available. Notable examples include the Munazarat 1.0 corpus (Khader et al., 2024), which compiles roughly 50 hours of recordings from 73 debates at QatarDebate-recognized tournaments; the hybrid annotation model (Al-Sharafi et al., 2025), which extends this work by introducing debate-specific labels; and the computational benchmark study (Al-Zawqari et al., 2025), which evaluates a range of models on the enriched corpus to establish strong baselines for argument mining in Arabic debates. Another notable resource is QCAW 1.0 (Zaghouani et al., 2024), a bilingual corpus of 195 argumentative essays by Qatari students. While these corpora provide valuable foundations, existing efforts rarely account for the cultural and linguistic nuances unique to Arabic-speaking communities. Most debate platforms and argumentation tools are designed for Western audiences, often overlooking religious, regional, and social sensitivities. This paper introduces the task of culturally grounded topic generation and evaluation as a step toward developing NLP tools tailored to Arabic discourse and public debate.

3 Arabic Users and Online Debate: Survey Insights

To understand the needs and expectations of Arabic-speaking users regarding online debate platforms, we conducted a survey. The survey was designed to accommodate varying levels of user familiarity with debating platforms by tailoring questions based on prior exposure and participation. It covered a broad spectrum of topics, from frequency of use and user motivations to preferences for debate topics and expectations for platform features.

In addition to exploring past experiences, the survey placed particular emphasis on users' expectations for AI-supported features. It examined attitudes toward technologies such as automated argument generation, summarization, and moderation, and aimed to identify the ideal balance between human control and AI assistance. The survey also evaluated the perceived cultural and linguistic appropriateness of existing platforms, helping assess the need for culturally tailored debate environments for Arabic-speaking communities.

The 62-question survey was organized into four main sections:

1. **General Usage:** Questions addressing whether and how participants have used debating platforms in the past.

2. **Engagement and Participation:** Divided into two tracks depending on user experience, this section explored motivations, usage frequency, and perceived benefits or challenges.

3. **Expectations and AI Integration:** Focused on users' views toward AI tools, particularly their utility, cultural fit, and potential drawbacks.

4. **Open Feedback:** Offered space for detailed user input beyond fixed responses.

The survey was administered via Prolific[5] to 50 native Arabic speakers. All participants completed the questionnaire, with an average response time of approximately 10 minutes and 40 seconds.

Findings reveal a strong interest in structured debate platforms tailored to Arabic users. Many participants had previously interacted with forums such as Reddit's *r/ChangeMyView*, citing motivations like expanding their perspectives, learning from diverse opinions, and entertainment. Popular discussion themes included politics, education, culture, and religion. Participants valued features such as topic discovery, voting mechanisms, and AI-generated arguments, though they preferred moderate AI involvement, favoring tools that aid rather than replace human reasoning. Concerns were raised around AI accuracy, potential cultural insensitivity, and the risk of diminishing human agency. Desired features included real-time fact-checking, exposure to diverse perspectives, and inclusive engagement. Participants also emphasized such a platform's potential to enhance Arabic literacy, reduce misinformation, and foster open dialogue. At the same time, they expressed concern over hostility, bias, and judgmental tones in debates. These insights offer a user-informed roadmap for designing AI-powered, culturally sensitive Arabic debate platforms. Selected charts from the survey are included in the appendix.

[5] www.prolific.com

Platform	# Topics
DebateData	27,393
Kialo Edu Blog	1,047
iDebate Debatabase	683
Kialo Edu Topics Library	531
I Side With	250
Britannica ProCon	101
Total (raw)	30,005
Total (deduplicated)	29,965

Table 1: Debate topics collected from different English debate platforms.

4 Cultural Relevance of Topics in Existing Debate Platforms

To assess the suitability of existing English-language debate platforms for Arabic-speaking users, we analyze how well their topics reflect Arabic culture, traditions, and social norms. This evaluation informs whether such platforms can be effectively adapted or if there is a need for culturally specific alternatives. We focus on two key aspects: (1) the representation of topics that are relevant to Arabic contexts, and (2) the inclusion of content that may be culturally inappropriate or misaligned.

4.1 Debate Topic Collection

We collected debate topics from six well-known English-language online debate platforms. These platforms were selected based on their popularity, diversity of subject areas, and use of structured debate formats. Table 1 shows the platforms along with the number of topics extracted from each.

The collected topics vary in specificity and stance expression. Some are framed as open-ended questions or discussion prompts (e.g., *"Is homeschooling better than traditional schooling?"*), while others present clear argumentative claims (e.g., *"The death penalty deters crime"*). In total, we gathered around 30,000 unique topics spanning domains such as politics, ethics, religion, education, technology, and gender. This dataset serves as the foundation for the cultural alignment analysis in the following sections.

4.2 LLM-Based Cultural Analysis

To conduct a large-scale cultural assessment of debate topics, we employed three diverse LLMs: *Fanar-1-9B-Instruct*, *DeepSeek R1*, and *Claude Sonnet 4*. These models were selected for their differing training backgrounds and cultural priors. Fanar is designed to align with Arabic cultural norms, DeepSeek is developed in a Chinese context and reflects a non-Western worldview, while Claude is a general-purpose Western model. This diversity enables examining how various cultural lenses assess topic relevance and appropriateness for the Arab world.

Each LLM was prompted to classify every topic along two dimensions:

- **Cultural Specificity:** Whether the topic is fundamentally tied to Arab cultural, historical, or religious contexts.
- **Debate Suitability:** Whether the topic is appropriate and constructive for public discourse in Arabic-speaking societies.

To ensure consistent evaluation, we designed a structured and culturally grounded prompt. The models were asked to produce:

`Specificity`	`Specific` or `General`
`Debate Fit`	`Inappropriate`, `Neutral`, or `Resonant`
`Explanation`	A concise 2–3 sentence justification referencing Arab culture and norms.

The prompt positioned the model as an expert cultural analyst and included detailed classification criteria and examples. All models received the same prompt structure, with only minor adjustments to match input formatting requirements. The complete prompt is provided in the Appendix.

By using models with distinct cultural foundations, we aim to uncover not only which topics are flagged as culturally aligned or misaligned, but also how different LLMs reason about cultural fit. High agreement across models suggests the presence of shared cultural cues, while divergence highlights the cultural assumptions embedded in each model's training data.

4.3 Human Validation

To assess the reliability and cultural reasoning of LLM outputs, we conducted a stratified human evaluation of model classifications. Rather than using random sampling, we employed a *case-based sampling strategy* to ensure coverage across all combinations of model-generated labels. This choice

reflects our hypothesis that not all classification scenarios are equally challenging or informative: for example, culturally 'General' topics labeled as 'Inappropriate' may reveal over-sensitivity, while 'Specific' and 'Resonant' combinations may reflect culturally grounded debate material. Stratified evaluation allows us to probe both model strengths and failure modes across the full labeling space. For each LLM, we attempted to sample 50 debate topics from each of the six possible *(Specificity, Debate Fit)* combinations, yielding 725 topics, as the Claude and Fanar models produced fewer than 50 instances in some combinations. The distribution of labels across models is reported in Table 2.

Annotator Profile Six native Arabic speakers (*3 males, 3 females*), all with a background in debate practice or argumentation research, served as annotators. All annotators were fluent in Modern Standard Arabic and familiar with cultural, traditional, and social sensitivities relevant to public discourse in the Arab world.

Annotation Procedure Annotators followed detailed guidelines adapted from the LLM prompt. Prior to annotation, a calibration session was held to align interpretations and resolve potential ambiguities. During the task, each topic was independently annotated by two annotators, who assigned *Specificity* and *Debate Fit*. Also, a binary judgment on whether the topic is suitable for inclusion in the well-known Arabic debate organization QatarDebate[6]. Disagreements were adjudicated by a third senior annotator with expertise in cultural argumentation, who resolved them by selecting one of the two labels already assigned.

4.4 Results

We report results along four dimensions: inter-annotator agreement, the adjudicated gold dataset, model–human agreement, and model output distributions over about 30,000 debate topics.

Inter-Annotator Agreement We report inter-annotator agreement (IAA) using Cohen's κ and overall agreement rate on the full stratified annotation sample (Table 3). Agreement was calculated separately for the two classification dimensions: *Cultural Specificity* and *Debate Fit*.

For *Specificity*, annotators reached 89.52% agreement with a Cohen's κ of 0.50, reflecting moderate consistency in identifying whether topics were culturally grounded in Arab contexts. *Debate Fit* showed lower agreement, with 44.55% agreement and $\kappa = 0.19$ ("fair"). When reframed as suitability for an Arabic debate organization, agreement improved ($\kappa = 0.29$), suggesting that institutional framing can help reduce ambiguity.

These findings support our hypothesis that not all tasks are equally clear-cut: cultural specificity tends to yield more stable judgments, while appropriateness is shaped by individual and regional sensitivities within the diverse Arab world.

Gold Standard Dataset After resolving disagreements through senior adjudication, we obtained a gold dataset of 725 debate topics. For *cultural specificity*, the data is highly imbalanced: 660 topics (91%) were labeled as *General*, while only 65 (9%) were labeled as *Specific*. In contrast, *Debate Fit* shows a more balanced distribution, with 310 *Neutral*, 220 *Resonant*, and 195 *Inappropriate* topics. For *organizational suitability*, 275 topics were judged *Resonant*, 254 *Inappropriate*, and 196 *Neutral*. Table 4 summarizes these distributions. This dataset provides a valuable benchmark for evaluating culturally-aware NLP systems and future models for Arabic debate platforms.

Human-Model Agreement In order to assess how closely model predictions aligned with human judgments, we compared each model's output to the final adjudicated annotations. Table 5 reports agreement rates and Cohen's κ across tasks.

Across all models, reliability was highest on the *Debate Fit* task ($\kappa = 0.32$, 53% agreement) and *Organizational Suitability* ($\kappa = 0.29$, 54%), with lower consistency on *Cultural Specificity* ($\kappa = 0.21$, 68%). Among the individual models, Claude showed the strongest alignment with human annotations, reaching $\kappa = 0.54$ (81% agreement) on specificity and $\kappa = 0.39$ (59%) on debate fit. DeepSeek achieved fair reliability on organizational suitability ($\kappa = 0.31$, 55%), while Fanar lagged behind overall, particularly on specificity ($\kappa = 0.04$, 62%). These results indicate that Claude provides the most consistent judgments, whereas Fanar is less reliable despite its cultural orientation, and DeepSeek performs moderately across tasks.

Model Output Distribution Table 6 presents the distribution of combined cultural alignment labels assigned by each model: Fanar, DeepSeek, and Claude to the 30,000 English-language de-

[6] https://qatardebate.org

	Cultural Specificity			Debate Fit			
Model	General	Specific	Total	Resonant	Neutral	Inappropriate	Total
Fanar	150	95	245	100	50	95	245
DeepSeek	150	100	250	100	50	100	250
Claude	150	80	230	100	54	76	230

Table 2: Distribution of sampled debate topics by model and label, after stratified selection.

Task	**Agreement (%)**	κ
Cultural Specificity	89.52	0.50
Debate Fit	44.55	0.19
Org. Suitability	52.83	0.29

Table 3: Inter-annotator agreement across tasks.

Task	**Label**	**Count**	**%**
Cultural Specificity	General	660	91
	Specific	65	9
Debate Fit	Neutral	310	43
	Resonant	220	30
	Inappropriate	195	27
Org. Suitability	Resonant	275	38
	Inappropriate	254	35
	Neutral	196	27

Table 4: Adjudicated gold label distributions.

bate topics [7]. The majority of topics were labeled as *General* across all models. Fanar classified over 24,000 topics as *General-Resonant*, showing a tendency to view general content as culturally relevant, while assigning very few topics to the *Specific* categories. In contrast, DeepSeek had a more critical stance: it labeled nearly 7,000 topics as *General-Inappropriate* and 668 as *Specific-Inappropriate*, suggesting a stricter interpretation of cultural fit. Claude offered a more balanced distribution, with significant counts in both *General-Neutral* (20,873) and *General-Resonant* (4,511), and modest allocations across *Specific* labels. Notably, only Claude produced *Specific-Neutral* labels (5), and all models showed relatively low counts for *Specific-Resonant* topics, highlighting a shared perception that few English debate topics directly address Arab cultural contexts.

Insights from Model Explanations Since the LLMs were prompted to explain their labeling decisions, their responses provide a lens into how they assess relevance and appropriateness. These justifications may offer indications of the models' reasoning and implicit judgments about culture and values. Inspecting them across the models, we noted patterns that suggest how they might frame cultural sensitivity, traditions, and regional context.

For example, consider the topic "This House would allow adoption agencies to guarantee to biological parents that their child will not be adopted by a same-sex couple." Fanar labeled it General-Resonant, citing "varying legal frameworks across Arab countries regarding adoptions and same-sex relationships." DeepSeek, by contrast, labeled it General-Inappropriate, arguing that "public discussion of LGBTQ+ matters violates cultural-religious taboos in most Arab societies." These contrasts illustrate possible differences in how the models respond to the intersection of legal considerations, cultural norms, and religious values.

5 Culturally Grounded Topic Generation

To explore the capabilities of LLMs in culturally grounded debate, we prompted the three models used in the previous study: *Fanar-1-9B-Instruct*, *DeepSeek R1*, and *Claude Sonnet 4* to generate new debate topics tailored to Arabic-speaking culture. This complements our earlier classification study by experimenting whether models can deliver topics aligned with Arab cultural values, traditions, and discourse norms.

Each model was asked to generate 50 debate topics rooted in Arab cultural, religious, or historical contexts, with relevance to public discourse. The prompt framed the model as a cultural expert and debate strategist, instructing it to draw from diverse regional traditions (e.g., Gulf, Levant, North

[7] To ensure label validity, we excluded topics with invalid `specificity` or `debate fit` labels due to API or parsing errors. These malformed cases were rare: 2.6% for Fanar, 3.9% for Claude, and 6.7% for DeepSeek.

	All Models		Fanar		DeepSeek		Claude	
Task	Agr.	κ	Agr.	κ	Agr.	κ	Agr.	κ
Cultural Specificity	68	0.21	62	0.04	63	0.09	81	0.54
Debate Fit	53	0.32	52	0.32	49	0.26	59	0.39
Org. Suitability	54	0.29	52	0.29	55	0.31	53	0.29

Table 5: Agreement (%) and Cohen's κ between model predictions and final human annotations across tasks.

Label	**Fanar**	**DeepSeek**	**Claude**
Specific-Resonant	309	442	288
Specific-Neutral	0	0	5
Specific-Inappropriate	45	668	58
General-Resonant	24,146	12,727	4,511
General-Neutral	2,833	7,204	20,873
General-Inappropriate	1,891	6,916	3,071

Table 6: Label distribution across Fanar, DeepSeek, and Claude LLMs.

Africa), Islamic values, family and gender dynamics, and tensions between tradition and modernity.

A key design element was a *domain-diversity constraint*: the topics had to span distinct societal domains such as governance, gender, religion, tribal customs, technology, and media. This requirement encouraged broader coverage across culturally salient but underrepresented areas of debate.

The required output format was a numbered list of 50 concise debate topics, each phrased as a clear, single-sentence proposition (e.g., "This House believes that..."), with no additional explanation or formatting. This structure ensured comparability across models and suitability for subsequent evaluation. The full prompt is included in the Appendix.

This generation task allows assessing how well different LLMs internalize Arabic cultural discourse norms and whether they can produce high-quality, debate-worthy content that is both culturally specific and socially relevant. Manual inspection[8] confirmed that nearly all generated topics adhered to the prompt constraints, producing culturally grounded and debate-appropriate content.

Model Output Analysis A close inspection of the 150 topics (50 per model) shows that all three LLMs successfully follow the required format and produce culturally specific debate topics. Claude offers the broadest domain coverage, touching on governance, technology, gender, tribal affairs, environmental policy, and bioethics. DeepSeek is similarly diverse but skews toward bolder, reform-oriented topics (e.g., revising Islamic inheritance laws, ending Wasta, modernizing awqaf), suggesting a tendency for more provocative framing. Fanar generates the most diplomatic set: many topics are phrased in supportive language ("This House supports..." or "argues that...") and often grounds proposals in Islamic principles. All three lists avoid overtly Western-centric references and include culturally salient constructs such as tribal mediation, multilingual education, and Sharia-compliant finance, indicating that the prompt effectively steers generation toward Arab contexts.

The models differ in stance nuance and sensitivity. Claude's topics often present a clear, assertive proposition (*"This House believes that daughters should inherit equally..."*), inviting direct clash. Fanar's topics tend to balance modernization with tradition ("...within the framework of Islamic law"), arguably lowering the risk of cultural offense. DeepSeek produces the highest share of potentially controversial items (e.g., critique of honor codes, social media's impact on family norms), which could spark richer but also more polarizing debate. Minor style issues appear: Fanar occasionally embeds evaluative adjectives ("positive aspects"), while DeepSeek includes a few topics that combine multiple ideas or comparisons. Overall, Claude offers the greatest topical breadth, Fanar the most culturally deferential tone, and DeepSeek the most reform-minded edge. These insights are beneficial and can guide model selection or ensemble strategies for seeding Arabic debate platforms.

6 Arabic Debate Platform Development

Developing a culturally grounded, AI-enhanced debate platform for Arabic-speaking users requires

[8] An annotation study was deemed unnecessary due to the consistency of model outputs with the prompt criteria.

the integration of argumentation theory, Arabic NLP, and human-centered AI design. This section presents a multi-layered proposal aimed at enabling structured, substantive, and culturally resonant debates. Our architecture balances theoretical depth with practical AI capabilities, ensuring that argument quality, cultural alignment, and user experience remain central to the platform's design. Insights from our survey highlight user demand for structured debates, real-time fact-checking, culturally sensitive moderation, and moderate AI involvement. In parallel, our cultural relevance study showed that English-language debate topics are, as expected, overwhelmingly classified as *General* rather than *Specific* to Arab contexts, stressing the need for dedicated, culturally grounded topic generation. Similarly, our topic generation experiments with LLMs revealed that while models can generate debate-worthy content for Arabic contexts, they differ in breadth, tone, and risk of controversy, confirming the importance of expert-guided curation.

Theoretical Foundations We propose grounding the debate platform in well-established models of argumentation that guide both the structure and interpretation of user contributions. Toulmin's model (Toulmin, 1958), which identifies core components such as claims, grounds, warrants, and rebuttals, provides a robust framework for decomposing arguments into meaningful elements. Similarly, Freeman's theory (Freeman, 2011) represents arguments as networks of interconnected claims and premises, making it well-suited for scalable, web-based implementation. Together, these models ensure that arguments are represented and visualized in ways that are both logically rigorous and accessible to users.

Seed Content and Expert Engagement To launch the platform and guide its development, we propose seeding it with a curated collection of debate threads authored by expert debaters. These debates will cover culturally salient and controversial domains, including politics, religion, education, and ethics, ensuring thematic diversity consistent with both survey findings and our cultural relevance analysis. Since our study showed that existing English debate platforms rarely address Arab-specific contexts, expert-crafted debates will provide culturally grounded exemplars for users while also serving as high-quality training data for NLP models. In addition, our topic generation study demonstrated that AI systems can produce culturally specific debate topics with varying breadth, tone, and sensitivity. Curating and combining these outputs with expert input offers a scalable strategy for bootstrapping high-quality, culturally appropriate debate content.

Arabic NLP Support A culturally aware Arabic debate platform depends on robust NLP models that support argument mining, evidence classification, topic and counterargument generation, human value detection, and cultural alignment evaluation.

For *argument mining*, models must identify components such as claims, premises, and rebuttals across diverse genres. This capability is essential for structuring debates, enabling argument maps, and providing users with transparent breakdowns of reasoning. We recommend constructing a cross-domain Arabic corpus spanning televised debates, editorials, and online forums. Annotation by native speakers, guided by established frameworks, enables fine-tuning of transformer-based models such as AraBERT[9] and MARBERT[10]. Multi-task setups and span-based labeling approaches can improve performance for complex or nested structures.

Evidence classification enhances informativeness by identifying support types (e.g., testimony, anecdotal evidence). This is critical for helping users evaluate the strength and credibility of arguments, encouraging reliance on robust support rather than weak or biased claims. Resources from prior work (Rinott et al., 2015; Al-Khatib et al., 2016b) can be adapted to Arabic, enabling models to guide users in strengthening arguments with appropriate evidence.

Debate topic and argument generation provides a scalable way to supply culturally resonant debate prompts in low-resource settings. This is especially important since our cultural analysis revealed that existing English debate platforms rarely include Arab-specific issues. Instruction-tuned LLMs can therefore be used to bootstrap content, with expert curation ensuring cultural appropriateness and thematic diversity.

To ensure *cultural fit*, classifiers should assess both generated and user-submitted content for specificity and appropriateness. This safeguards against culturally insensitive or irrelevant debates and maintains user trust. Techniques such as adapter fusion or instruction tuning on multilingual backbones (e.g., mT5, Falcon-Instruct) can help

[9] https://github.com/aub-mind/arabert
[10] https://github.com/UBC-NLP/MARBERT

models recognize subtle cultural cues and prevent harmful misalignment.

Human value detection allows debates to be anchored in ethical and societal considerations that resonate with Arabic-speaking communities. By identifying which values (e.g., humility, hedonism, tradition) are being appealed to, platforms can better surface value-sensitive debates, guide moderation, and promote inclusivity. Structured taxonomies, such as those from the Touché shared task[11], can be localized for this purpose.

Finally, *culture-aware counterargument generation* supports balanced and critical discussions by automatically suggesting respectful and contextually appropriate challenges to user claims. This reduces the risk of echo chambers, promotes exposure to diverse perspectives, and enhances critical thinking. Techniques such as contrastive decoding and evidence-informed generation (Lin et al., 2023) can be adapted to Arabic contexts to ensure cultural alignment in counterargumentation.

This integrated NLP pipeline supports debates that are linguistically robust, culturally aligned, and socially responsible. It enables Arabic-speaking users to construct persuasive, respectful, and well-grounded arguments, advancing both civic discourse and computational argumentation.

User Experience and Interaction The platform interface should prioritize usability, reflection, and constructive engagement. Visual argument maps help users navigate debate structures, while real-time AI feedback assists in improving clarity, coherence, and relevance. Community-driven features such as voting and content rating promote high-quality input and collaborative norms. At the same time, moderation systems must address the risks revealed by our analyses: models occasionally generated sensitive or polarizing topics (e.g., inheritance laws, honor codes), and survey respondents voiced concerns about hostility, bias, and judgmental tones in debates. This motivates a hybrid approach where AI-generated topics and user contributions are filtered and contextualized by expert review, ensuring debates remain culturally resonant, inclusive, and socially constructive.

7 Conclusion

Despite the growing importance of online discourse, Arabic-speaking communities remain underserved by platforms that support structured, culturally grounded debate. This paper proposed a vision for a culturally aware Arabic debate platform and presented a multi-layered investigation to support its development. Through a user survey, we identified key shortcomings in existing platforms and outlined user expectations for culturally appropriate deliberation. We also analyzed around 30,000 English-language debate topics using LLMs to assess their cultural relevance and explored the capability of LLMs to generate new, resonant controversial topics. Our findings show both the limitations of current debate content for Arab audiences and the potential of prompt-guided LLMs to support culturally sensitive topic generation and evaluation.

As part of this work, we introduce the new task of assessing *topic relevance to culture*, a new perspective for NLP research at the intersection of argumentation, content generation, and cultural alignment. Further, we present a technical and theoretical framework for building AI-supported debate platforms that reflect Arabic communication styles, social values, and linguistic norms.

Rather than isolating a single technical contribution, our approach integrates survey insights, prompt engineering, LLM evaluation, and dataset construction, laying the foundation for future research in culturally aligned Arabic NLP applications aimed at civic discourse and public reasoning. Ultimately, our work identifies a critical but underexplored direction in Arabic NLP: designing language technologies that not only process Arabic text effectively, but also support meaningful engagement, critical thinking, and digital literacy.

In future work, we plan to deepen our focus on culture-aware generation and assessment tasks, and to develop computational models for argument mining, moderation, and content evaluation tailored to the sociocultural realities of the Arab world. We argue that culturally sensitive NLP tools are essential to enabling inclusive, thoughtful, and constructive online debate for Arabic-speaking communities.

Acknowledgements

This work was supported by the QD Fellowship award [QDRF-2025-02-020] from QatarDebate Center. We would like to thank Baraa Alahmar, Batool Alnobani, Beshr Alsioufy, Esraa Afifi, Manar Khabaz, and Nahla Basiouni for their valuable contributions in conducting the annotation study.

[11] https://touche.webis.de/clef24/touche24-web/human-value-detection.html

References

Azza Abouzied, Firoj Alam, Raian Ali, and Paolo Papotti. 2025. Combating misinformation in the arab world: Challenges & opportunities. *arXiv preprint arXiv:2506.05582*.

Vibhor Agarwal, Sagar Joglekar, Anthony P. Young, and Nishanth Sastry. 2022. GraphNLI: A graph-based natural language inference model for polarity prediction in online debates. In *Proceedings of the ACM Web Conference 2022 (WWW '22)*, pages 2729–2737, New York, NY, USA. Association for Computing Machinery.

Khalid Al-Khatib, Michael Völske, Shahbaz Syed, Nikolay Kolyada, and Benno Stein. 2020. Exploiting personal characteristics of debaters for predicting persuasiveness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7067–7072, Online. Association for Computational Linguistics.

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, and Benno Stein. 2016a. Cross-domain mining of argumentative text through distant supervision. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1395–1404, San Diego, California. Association for Computational Linguistics.

Khalid Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016b. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3433–3443, Osaka, Japan. The COLING 2016 Organizing Committee.

Abdul Gabbar Al-Sharafi, Mohammad Majed Khader, Mohamed Ahmed, Mohamad Hamza Al-Sioufy, Wajdi Zaghouani, and Ali Al-Zawqari. 2025. A hybrid annotation model for arabic argumentative debate corpus. In *Arabic Language Processing: From Theory to Practice*, pages 97–113, Cham. Springer Nature Switzerland.

Ali Al-Zawqari, Mohamed Ahmed, Abdul Gabbar Al-Sharafi, Mohammad M. Khader, Ali Safa, and Gerd Vandersteen. 2025. Neural classification of argument elements and styles in arabic competitive debates. *IEEE Access*, 13:115944–115959.

Gioia Boschi, Anthony P. Young, Sagar Joglekar, Chiara Cammarota, and Nishanth Sastry. 2021. Who has the last word? understanding how to sample online discussions. *ACM Transactions on the Web*, 15(3):12:1–12:25.

Mahmoud Fawzi, Björn Ross, and Walid Magdy. 2026. Fabricating holiness: Characterizing religious misinformation circulators on arabic social media. 20.

Tallullah Frappier, Nathalie Bressa, and Samuel Huron. 2024. Jumping to conclusions: A visual comparative analysis of online debate platform layouts. In *Proceedings of the 13th Nordic Conference on Human-Computer Interaction*, NordiCHI '24, New York, NY, USA. Association for Computing Machinery.

James B. Freeman. 2011. *Argument Structure: Representation and Theory*. Springer, Dordrecht, Netherlands.

Vahid Ghafouri, Vibhor Agarwal, Yong Zhang, Nishanth Sastry, Jose Such, and Guillermo Suarez-Tangil. 2023. Ai in the gray: Exploring moderation policies in dialogic large language models vs. human answers in controversial topics. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23)*, pages 556–565, New York, NY, USA. Association for Computing Machinery.

Xinyu Hua and Lu Wang. 2017. Understanding and detecting supporting arguments of diverse types. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 203–208, Vancouver, Canada. Association for Computational Linguistics.

Mohammad M. Khader, AbdulGabbar Al-Sharafi, Mohamad Hamza Al-Sioufy, Wajdi Zaghouani, and Ali Al-Zawqari. 2024. Munazarat 1.0: A corpus of arabic competitive debates. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation @ LREC-COLING 2024*, pages 20–30, Torino, Italia. ELRA and ICCL.

Travis Kriplean, Jonathan Morgan, Deen Freelon, Alan Borning, and Lance Bennett. 2012. Supporting reflective public thought with considerit. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, page 265–274, New York, NY, USA. Association for Computing Machinery.

Mina Lee, Katy Ilonka Gero, John Joon Young Chung, Simon Buckingham Shum, Vipul Raheja, Hua Shen, Subhashini Venugopalan, Thiemo Wambsganss, David Zhou, Emad A. Alghamdi, Tal August, Avinash Bhat, Madiha Zahrah Choksi, Senjuti Dutta, Jin L.C. Guo, Md Naimul Hoque, Yewon Kim, Simon Knight, Seyed Parsa Neshaei, and 17 others. 2024. A design space for intelligent and interactive writing assistants. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Jiayu Lin, Rong Ye, Meng Han, Qi Zhang, Ruofei Lai, Xinyu Zhang, Zhao Cao, Xuanjing Huang, and Zhongyu Wei. 2023. Argue with me tersely: Towards sentence-level counter-argument generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16705–16720, Singapore. Association for Computational Linguistics.

Stefano Mezza, Wayne Wobcke, and Alan Blair. 2024. Exploiting dialogue acts and context to identify argumentative relations in online debates. In *Proceedings of the 11th Workshop on Argument Mining (ArgMining 2024)*, pages 36–45, Bangkok, Thailand. Association for Computational Linguistics.

Smitha Milli, Micah Carroll, Yike Wang, Sashrika Pandey, Sebastian Zhao, and Anca D Dragan. 2025. Engagement, user satisfaction, and the amplification of divisive content on social media. *PNAS Nexus*, 4(3):pgaf062.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence - an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 440–450, Lisbon, Portugal. Association for Computational Linguistics.

Stephen Edelston Toulmin. 1958. *The Uses of Argument*. Cambridge University Press, Cambridge, UK.

Meng Ye, Karan Sikka, Katherine Atwell, Sabit Hassan, Ajay Divakaran, and Malihe Alikhani. 2023. Multilingual content moderation: A case study on Reddit. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3828–3844, Dubrovnik, Croatia. Association for Computational Linguistics.

Wajdi Zaghouani, Abdelhamid Ahmed, Xiao Zhang, and Lameya Rezk. 2024. 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)*, pages 13382–13394, Torino, Italia. ELRA and ICCL.

Appendix

Topic	Specificity	Debate Fit
This House believes that feminists in Muslim-majority countries should advocate for the abolition of Sharia Courts rather than for feminist reforms within them (e.g., more female judges/qadi, feminist interpretations of the Quran)	Specific	Resonant
This House supports a military attack on Israel by the belligerents of the 6-Day War.	General	Inappropriate
Should gay couples have the same adoption rights as straight couples?	General	Inappropriate
As progressive Muslims this house celebrates the decline in unity of Arab countries	Specific	Resonant
This House Regrets the abolition of the Ottoman Caliphate	General	Neutral
Should people be required to work in order to receive Medicaid?	General	Resonant
That Lebanon should abolish the confessional system	Specific	Resonant
This House regrets the Arab Spring	Specific	Resonant
This House would ban any depiction of the Prophet Mohammed.	General	Inappropriate

Table 7: Examples of annotated debate topics for specificity and debate fit.

Prompt for Cultural and Debate Fit Assessment

You are an expert cultural analyst for Arabic debate platforms. Your task has two parts:
PART 1: CULTURAL SPECIFICITY ANALYSIS
1. Analyze if the topic is uniquely rooted in Arab cultural traditions
2. Consider:
   - Regional contexts (Gulf, Levant, North Africa, etc.)
   - Islamic religious traditions and practices
   - Historical Arab customs and contemporary practices
3. Classification criteria:
   - **Specific**: Fundamentally tied to Arab cultural/religious contexts
   - **General**: Universally relevant or applies to multiple cultures
PART 2: DEBATE SUITABILITY ASSESSMENT
1. Evaluate if the topic is appropriate for public debate in Arab societies
2. Consider:
   - Compatibility with Islamic values and social norms
   - Sensitivity to cultural diversity within the Arab world
   - Potential to cause offense or social division
3. Classification criteria:
   - **Inappropriate**: Violates Islamic principles or cultural taboos
   - **Resonant**: Culturally relevant and suitable for constructive debate
   - **Neutral**: Acceptable but lacks strong cultural connection
STRICT OUTPUT REQUIREMENTS:
1. Output **exactly three lines** in this format:
   `Specificity: <Specific|General>`
   `Debate Fit: <Inappropriate|Resonant|Neutral>`
   `Explanation: <2-3 sentence concise justification>`
2. The explanation must:
   - Justify both classifications separately
   - Reference Arab cultural/religious context
   - Remain concise (2–3 sentences maximum)
3. Do not include any additional text, notes, or formatting

`Topic: {topic}`

Prompt for Generating Culturally-Specific Debate Topics

You are a cultural expert and debate strategist specializing in the Arab world. Your task is to generate 50 impactful debate topics that are both culturally specific and socially resonant within diverse Arab contexts. Follow the guidelines below:

PART 1: CULTURAL SPECIFICITY TARGET

Ensure each topic is deeply connected to Arab cultural, religious, or historical traditions.

Draw on:

- Islamic values, rituals, and scholarly discourse
- Social norms and taboos in Gulf, Levant, North Africa, etc.
- Issues involving family structures, gender roles, tribal customs, or national identities
- Tensions between traditional values and modernity

PART 2: DEBATE RELEVANCE TARGET

Design each topic to be suitable for constructive public debate in Arab societies.

Consider:

- Alignment with Islamic and cultural values, while allowing diverse opinions
- Real relevance to current Arab public discourse
- Potential to encourage critical thinking without provoking harmful polarization

DOMAIN-DIVERSITY RULE

Generate 50 topics, and each topic must belong to a distinct issue domain.

Use a wide range of domains, such as:

- Gender & family
- Religion & state
- Tribal or social structures
- Technology & AI
- Education & language
- Media & culture
- Economy & labor
- Environment & sustainability
- Health & bioethics
- Governance & political participation

. . . and more as appropriate.

Do not repeat the same domain.

OUTPUT FORMAT (REQUIRED)

Generate a numbered list of exactly 50 debate topics. Each topic should:

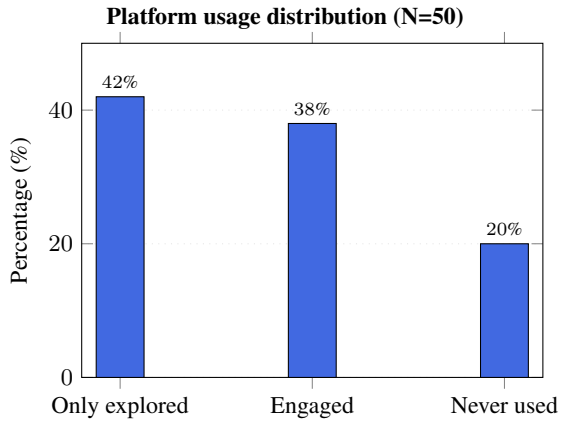
- Be one sentence
- Start with "This house believes that..." or similar phrasing
- Be clear, focused, and a maximum of 25 words

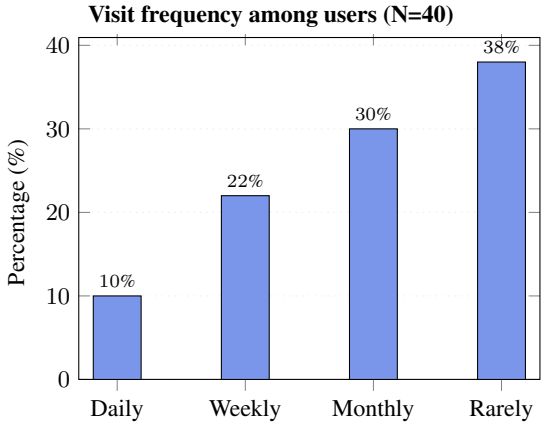
Do not include any explanation, notes, or extra formatting. Just the list.

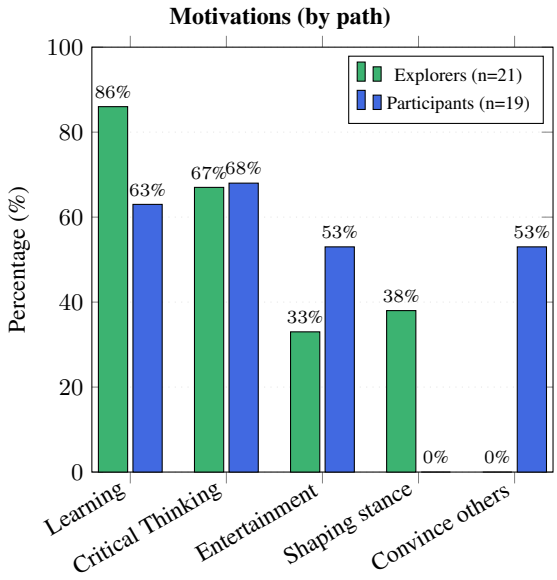
Examples of Generated Debate Topics by Three LLMs	
Fanar	
– This House believes that balancing modern education with traditional Islamic teachings is essential for youth development in Arab communities.	– This House supports integrating digital technology into educational systems while preserving Arabic language and cultural identity.
– This House argues that women's rights can be upheld within the framework of Islamic law and societal norms.	– This House believes that tribal customs should adapt to promote equal opportunities for all members of society.
– This House opposes the privatization of healthcare services over ensuring universal access based on Islamic principles of charity.	– This House believes that environmental conservation efforts should prioritize sustainable practices aligned with Islamic stewardship of Earth.
– This House supports the integration of moderate Islamic scholarship into government policies to foster social harmony.	– This House argues that media representation should reflect positive aspects of Arab cultures rather than stereotypes.
– This House believes that microfinance initiatives can empower families economically while respecting Islamic financial ethics.	– This House supports inclusive governance models that allow for greater citizen participation consistent with Islamic democracy ideals.
DeepSeek	
– This House believes that Arab media should prioritize preserving classical Arabic over colloquial dialects in programming.	– This House believes that Islamic inheritance laws should be revised to grant daughters equal shares to sons.
– This House believes tribal mediation (Sulh) undermines the rule of law in modern Arab states.	– This House believes that integrating AI into issuing Fatwas risks diluting religious scholarship.
– This House believes mandatory military service strengthens national identity over tribal affiliation.	– This House believes that Arab governments should heavily subsidize Arabic-language STEM education.
– This House believes social media platforms erode traditional Arab family communication norms.	– This House believes oil wealth should fund permanent sovereign wealth for future Arab generations.
– This House believes GCC states must adopt drastic water conservation measures for survival.	– This House believes genetic testing before marriage should be mandatory to reduce hereditary diseases.
Claude	
– This House believes that mandatory military service should be extended to women in Arab nations.	– This House believes that Islamic banking should completely replace conventional banking in Muslim-majority Arab countries.
– This House believes that tribal councils should have formal legal authority in dispute resolution.	– This House believes that artificial intelligence should be programmed with Islamic ethical principles.
– This House believes that Arabic should be the sole language of instruction in universities.	– This House believes that Arab satellite channels should prioritize local content over Western programming.
– This House believes that Gulf countries should implement universal basic income funded by oil revenues.	– This House believes that Arab nations should ban single-use plastics to protect desert ecosystems.
– This House believes that genetic engineering should be permitted when it prevents hereditary diseases.	– This House believes that Arab youth should have mandatory voting in national elections.

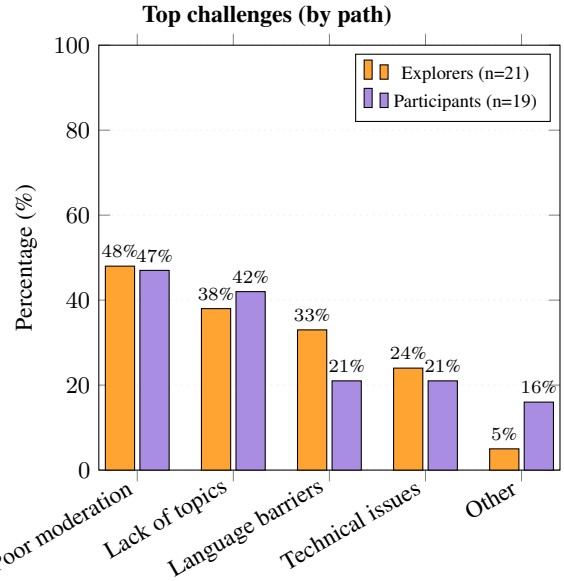
Table 8: Examples of the debate topics generated by Fanar, DeepSeek, and Claude.

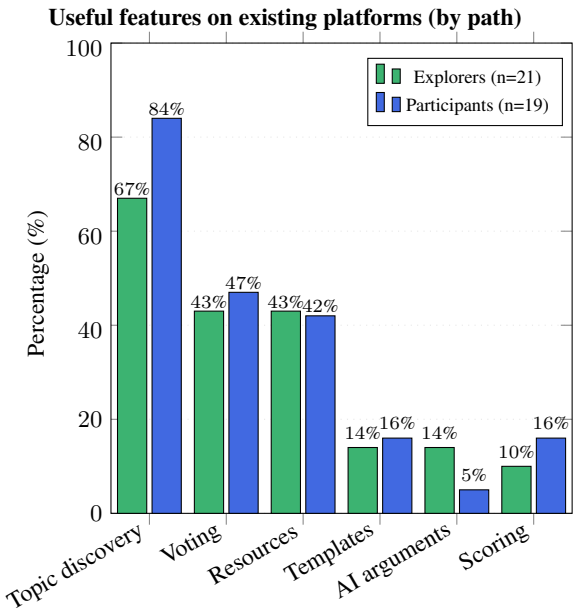
Selected Survey Charts

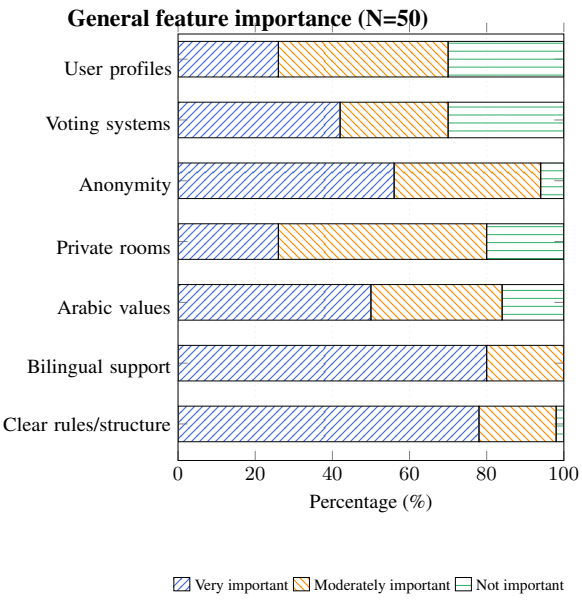


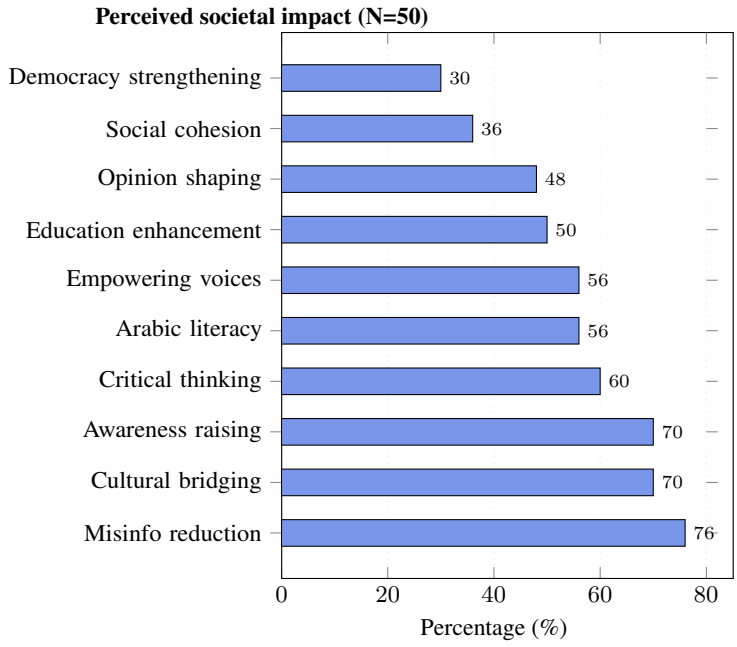


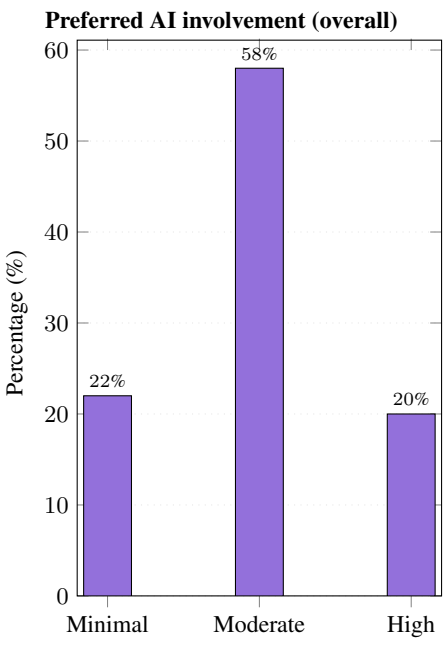


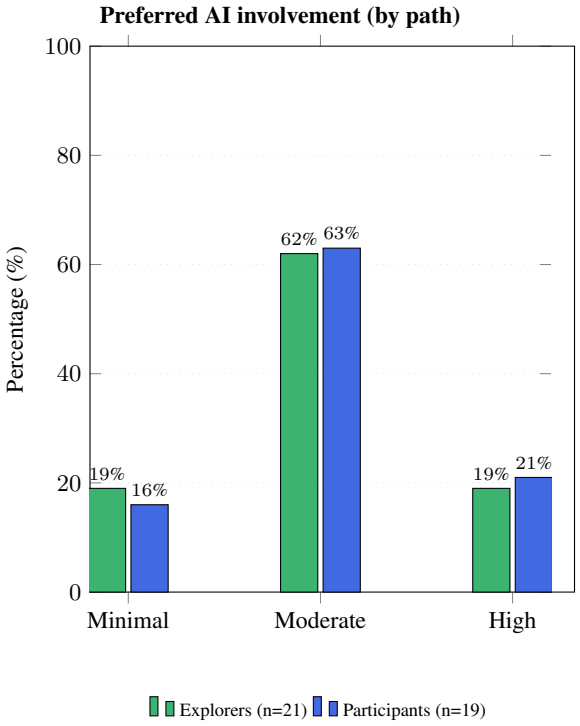


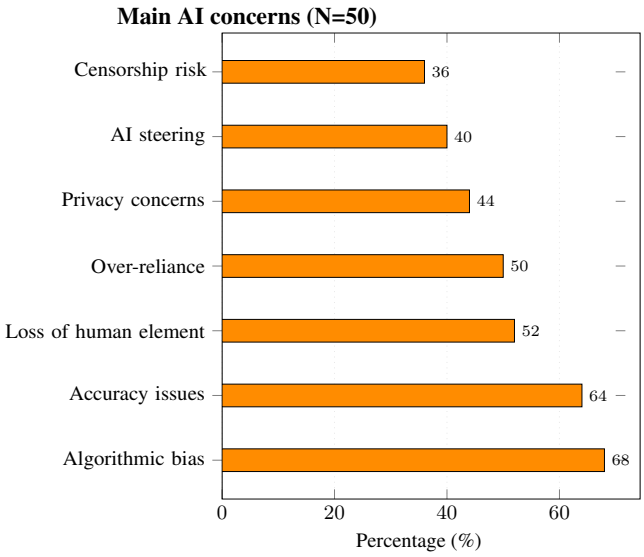





Perceived societal impact (N=50)
Democracy strengthening
Social cohesion
Opinion shaping
Education enhancement
Empowering voices
Arabic literacy
Critical thinking
Awareness raising
Cultural bridging
Misinfo reduction
30
36
48
50
56
56
60
70
70
76
0
20
40
60
80
Percentage (%)


Preferred AI involvement (overall)
Percentage (%)
0
10
20
30
40
50
60
22%
58%
20%
Minimal
Moderate
High


Preferred AI involvement (by path)
Percentage (%)
0
20
40
60
80
100
19%
16%
62%
63%
19%
21%
Minimal
Moderate
High
Explorers (n=21)
Participants (n=19)


Main AI concerns (N=50)
Censorship risk
AI steering
Privacy concerns
Over-reliance
Loss of human element
Accuracy issues
Algorithmic bias
36
40
44
50
52
64
68
0
20
40
60
Percentage (%)