Evaluating Large Language Model Versus Human Performance in Islamophobia Dataset Annotation

Rafizah Daud[1], Nurlida Basir[2]*, Nur Fatin Nabila Mohd Rafei Heng[3],
Meor Mohd Shahrulnizam Meor Sepli[4], Melinda Melinda[5]
Faculty of Science and Technology, Universiti Sains Islam, Malaysia[1, 2, 3]
National Digital Department, The Ministry of Digital, Malaysia[4]
Department of Electrical Engineering and Computer-Engineering Faculty, Universitas Syiah Kuala, Banda Aceh, Indonesia[5]

***Abstract*—Manual annotation of large datasets is a time-consuming and resource-intensive process. Hiring annotators or outsourcing to specialized platforms can be costly, particularly for datasets requiring domain-specific expertise. Additionally, human annotation may introduce inconsistencies, especially when dealing with complex or ambiguous data, as interpretations can vary among annotators. Large Language Models (LLMs) offer a promising alternative by automating data annotation, potentially improving scalability and consistency. This study evaluates the performance of ChatGPT compared to human annotators in annotating an Islamophobia dataset. The dataset consists of fifty tweets from the X platform using the keywords Islam, Muslim, hijab, stopislam, jihadist, extremist, and terrorism. Human annotators, including experts in Islamic studies, linguistics, and clinical psychology, serve as a benchmark for accuracy. Cohen's Kappa was used to measure agreement between LLM and human annotators. The results show substantial agreement between LLM and language experts (0.653) and clinical psychologists (0.638), while agreement with Islamic studies experts was fair (0.353). Overall, LLM demonstrated a substantial agreement (0.632) with all human annotators. ChatGPT achieved an overall accuracy of 82%, a recall of 69.5%, an F1-score of 77.2%, and a precision of 88%, indicating strong effectiveness in identifying Islamophobia-related content. The findings suggest that LLMs can effectively detect Islamophobic content and serve as valuable tools for preliminary screenings or as complementary aids to human annotation. Through this analysis, the study seeks to understand the strengths and limitations of LLMs in handling nuanced and culturally sensitive data, contributing to broader discussion on the integration of generative AI in annotation tasks. While LLMs show great potential in sentiment analysis, challenges remain in interpreting context-specific nuances. This study underscores the role of generative AI in enhancing human annotation efforts while highlighting the need for continuous improvements to optimize performance**.

***Keywords*—Large Language Model; generative AI; human intelligence; automatic data annotation; sentiment analysis; islamophobia; ChatGPT**

I. INTRODUCTION

Data annotation is the process of tagging raw data with relevant information to enhance the performance of machine learning models. The terms "data annotation" and "data labeling" are often used interchangeably, referring to the assignment of predefined labels to data points to create training datasets for machine learning algorithms [1]. Traditionally, this task is performed by human annotators following established rules and standards. For instance, in sentiment analysis, sentences or documents are classified as "positive", "negative", or "neutral". However, manual annotation is both time-consuming and labor-intensive, limiting its scalability for various natural language processing (NLP) applications [2].

Employing human annotators or outsourcing to specialized platforms can be costly, making large-scale annotation challenging [3], [4]. Additionally, human annotation is prone to inconsistencies, particularly when dealing with complex or ambiguous data, as interpretations may vary among annotators, impacting the reliability and reproducibility of datasets [5], [6]. This issue is especially pronounced in subjective tasks like sentiment analysis or hate speech detection, where annotator disagreement is common [5], [7]. Furthermore, the reliance on experts in fields such as linguistics, Islamic studies, or clinical psychology restricts the availability of qualified annotators, further complicating manual annotation efforts [8].

Despite these challenges, human annotation remains an essential component of machine learning and NLP. It goes beyond simple label assignment by incorporating contextual and supplementary information. Crowdsourcing has emerged as an effective approach for constructing large-scale datasets, particularly for subjective or culturally sensitive tasks [9]. It plays a crucial role in training machine learning models for applications such as hate speech detection [10], reading comprehension [11], sentiment analysis [12],[13], and bot detection [14]. However, the process remains resource-intensive, requiring domain expertise, significant time investment, and extensive labor, particularly for large datasets [15]. As dataset sizes continue to expand, the scalability of manual annotation becomes increasingly impractical, leading to delays and higher costs in data processing and analysis [16], [17].

This study is organized as follows: Section I introduces the research background, motivation, and objectives of the study. Section II presents a comprehensive review of related literature. Section III details the research methodology, including dataset selection, annotation protocols, and validation metrics. Section IV reports and analyzes the results of the comparative annotation task. Section V discusses the findings in relation to prior studies. Section VI identifies the limitations of the study, while Section VII offers recommendations for enhancing LLM-based annotation frameworks. Finally, Section VIII concludes with a summary of the key insights and suggests directions for

future research in the automated annotation of culturally sensitive content.

A. Challenges in Annotating Islamophobia Datasets

Islamophobia, defined as prejudice and discrimination against Muslims [18], [19] has become increasingly prevalent on social media and online platforms [20], [21]. Its manifestations range from over hate speech to subtle biases, making its detection and mitigation a challenging task. Research on Islamophobia dataset annotation reveals significant gaps in existing methods, particularly in consistency and accuracy. Annotating social media content for Islamophobia is complex, requiring cultural awareness, linguistic expertise, and standardized methodologies. This section critically examines the limitations of current text annotation approaches while identifying potential areas for improvement.

One major challenge is the ability of models to interpret nuanced and ambiguous content, especially in Islamophobic narratives [22] ,[23]. Many existing approaches fail to account for cultural and linguistic diversity, underrepresented languages and dialects [24] leading to misclassifications. Transformer-based models such as BERT and GPT offer a potential solution by enhancing contextual understanding. However, applying these models to low-resource languages [25] requires extensive fine-tuning and pre-training on diverse corpora.

Another critical issue is bias and fairness in model outputs, which is particularly problematic when classifying sensitive topics like religion [26] ,[27] . Labeling discrepancies often arise due to subjective interpretations by human annotators, introducing unintended biases into machine learning models. This problem is especially evident in hate speech detection, where definitions and interpretations vary across studies [28]. Manual annotation can worsen these inconsistencies, as annotators' personal and cultural perspectives influence labeling decisions, leading to a lack of standardization [29],[30], [31] . Addressing these biases requires the development of standardized annotation frameworks that promote fairness and consistency. Multi-annotator systems and consensus-based labeling methods can help mitigate subjectivity, improving dataset reliability and validity [32] , [33].

Despite these challenges, recent advancements present new opportunities for improvement. Hybrid approaches that combine human expertise with large language models (LLMs) leverage the semantic understanding capabilities of LLMs to enhance annotation consistency [7], [33] . Techniques such as zero-shot and few-shot learning, where pre-trained models classify data with minimal labeled examples, offer potential solutions for handling ambiguous content. Additionally, integrating auxiliary tasks such as sentiment analysis and emotion detection can provide deeper insights, improving classification accuracy in Islamophobia-related research. Addressing labeling inconsistencies, limited contextual awareness, and challenges in interpreting ambiguous language requires a combination of hybrid models, context-aware architectures, ethical annotation frameworks, and advanced AI methodologies. Future research should prioritize these developments to enhance the robustness, reliability, and fairness of Islamophobia detection systems.

B. The Role of Large Language Models in Annotation

The emergence of advanced Large Language Models (LLMs), such as ChatGPT, has revolutionized the data annotation landscape. Developed by OpenAI, ChatGPT can generate human-like text responses, making it a valuable tool for automating labor-intensive annotation tasks. Its ability to understand context, produce coherent text, and adapt to different styles and tones makes it a promising alternative to manual data labeling.

A recent study [34] explored the use of ChatGPT as a zero-shot learning model for annotating financial sentiment datasets. The study found that when ChatGPT was integrated with machine learning models such as pre-trained BERT and Support Vector Machines, it achieved an average accuracy of 90%. This research highlights ChatGPT's potential to identify emotional tone and sentiment in textual data, facilitating annotation for sentiment analysis tasks.

To address the growing need for scalable and consistent annotation of Islamophobia-related content, this study investigates the performance of a Large Language Model (ChatGPT) in comparison to domain-expert human annotators. In doing so, the research places strong emphasis on validation measures such as inter-rater reliability (Cohen's Kappa) and classification performance metrics including accuracy, precision, recall, and F1-score, which are widely used to assess model performance in annotation tasks [67], [68]. These measures are critical for ensuring the credibility and reproducibility of automated annotation efforts. Furthermore, the study situates its findings within the broader context of related work by comparing the model's annotation performance to outcomes from prior studies using both human and LLM-based approaches [34], [36], [38], [41]. This comparative perspective highlights not only the capabilities and limitations of ChatGPT but also informs the design of hybrid human-AI annotation frameworks.

While human annotators, particularly those with specialized knowledge, remain indispensable, their involvement poses challenges related to scalability and consistency. This study investigates the feasibility of using LLMs for text annotation in the context of sentiment analysis related to Islamophobia. The objectives of the research are:

1) Assess the agreement level between LLM-generated annotations and human-labeled data.

2) Evaluate the accuracy of LLMs in annotation tasks.

To guide the investigation and align with the study's objectives, the following research questions are proposed:

1) To what extent do LLM-generated annotations agree with human-labeled data in the context of Islamophobia detection?

2) How accurate are Large Language Models in annotating Islamophobia-related content compared to expert human annotators?

By comparing LLM performance with human experts in Islamic studies, linguistics, and clinical psychology, the study seeks to determine whether LLMs can effectively replace human annotators in this domain. This analysis will provide insights into the strengths and limitations of LLMs in handling nuanced and culturally sensitive data, contributing to broader discussions on the integration of generative AI in annotation workflows.

II. Literature Review

A. The Role of Large Language Models in Data Annotation

LLMs like ChatGPT have recently gained traction as promising tools for automating the labor-intensive process of manual data annotation. More than just tools, these models play a crucial role in improving the accuracy and efficiency of data labeling. Since its release, ChatGPT has drawn significant attention from researchers, leading to its application across diverse fields, including social computing [35], natural language processing [36],[37], sentiment analysis [9],[38], and medical science [39].

Advancements in LLMs have reshaped the data annotation landscape, offering both opportunities and challenges for researchers. Studies indicate that ChatGPT-4 surpasses human experts in identifying political messages, demonstrating higher accuracy and reliability than crowd workers and subject matter experts, while maintaining equal or lower bias [3]. This advantage extends to sentiment analysis, where ChatGPT has achieved an impressive 98.9% sentiment recognition accuracy, outperforming traditional lexicon-based methods [34],[38]. Additionally, the development of specialized LLMs, such as BloombergGPT a 50-billion-parameter financial language model, highlights the potential for domain-specific applications, including specialized annotation tasks [40].

Despite these advancements, the performance of LLMs varies across different contexts and languages. While ChatGPT performs well in sentiment analysis, its accuracy differs across languages such as Turkish, Indonesian, and Minangkabau, where human annotators demonstrate superior context awareness and nuanced interpretation [41]. Studies show that while median accuracy across tasks reaches 85%, one-third of tasks exhibit lower precision or recall [42]. Similarly, GPT-4 achieves up to 95% accuracy for short text classification but struggles with longer texts and non-English content [43]. These performance disparities are particularly relevant to specialized fields like Islamophobia research, where cultural context and linguistic intricacies significantly impact annotation quality.

LLM-driven annotation provides significant cost and efficiency benefits. Studies show that GPT-3 reduces labeling costs by 50% to 96% compared to human annotation, with some in-house models outperforming GPT-3 when trained on labeled data [37]. Additionally, open-source LLMs such as HuggingChat and FLAN have demonstrated superior performance in specific tasks, offering cost-effective alternatives to proprietary models [44]. However, quality management remains a major concern, with 30% of studies reporting poor quality control and a lack of transparency in annotation methodologies [45].

B. Human versus LLM Hybrid Approaches for Enhanced Annotation

Research supports hybrid annotation strategies that combine LLM capabilities with human expertise. The CoAnnotating framework enhances collaboration between humans and LLMs using uncertainty measures, improving annotation efficiency by up to 21% compared to random allocation [46]. Similarly, the AnnoLLM system demonstrates that LLMs can function as guided annotators, particularly when using an explain-then-annotate approach [47]. MEGAnno+ underscores the necessity of human validation to ensure reliable labels, acknowledging inherent biases and errors in LLM-generated annotations [48].

Despite their capabilities, LLMs still face technical limitations. ChatGPT struggles with sarcasm, fragmented sentences, and often misclassifies high-polarity tweets as neutral [14], [46]. Literature suggests that ChatGPT's NLP performance may fall short of supervised baselines due to token limitations and task mismatches, though optimization techniques can significantly enhance outcomes [49]. Additionally, adversarial annotation studies reveal that more advanced models sometimes perform worse when faced with stronger adversarial inputs, emphasizing the need for robust validation procedures [11].

Quality assurance remains a critical concern in LLM annotation. Research highlights the importance of human validation in improving LLM-generated labels, with optimized workflows significantly enhancing annotation accuracy [42]. Active learning methods can reduce manual annotation efforts, with studies showing that ChatGPT's annotations closely match human-labeled data when properly evaluated [39]. The construction of gold-standard datasets is essential for maintaining annotation reliability, particularly in cases where human annotators achieve high intercoder agreement [41].

C. Optimizing LLM Performance Through Prompt Engineering

Prompt engineering plays a crucial role in maximizing LLM efficiency. The APT-Pipe framework demonstrates that customized prompts can improve F1-scores by an average of 7.01% across multiple text classification datasets [50]. Different prompting strategies significantly impact annotation quality, with GPT-4 exhibiting greater variability than GPT-3.5 [51].

A recent study [52] developed binary classification prompts using GPT-3.5 Turbo, GPT-4, and DepGPT to categorize texts as "Non-Depressive" or "Depressive", focusing on performance and cost-effectiveness, particularly in the Bangla language. Similarly, [36] explored three GPT-3-based approaches: prompt-guided unlabeled data classification, synthetic training data generation, and dictionary-assisted annotation. Findings suggest that GPT-3 can generate labeled data from scratch or convert structured knowledge into natural language, reducing the need for human annotation. Unlike human annotators, who require extensive training and work at a slower pace, GPT-3 enables rapid annotation at scale.

While LLMs can generate high-quality labels, human oversight remains essential for ensuring annotation accuracy

and reliability [36] [53]. A study by [13] found that GPT-3 significantly improves text classification by generating precise pseudo-labels across multiple languages while reducing manual workload. This adaptability makes it particularly effective for domain-specific, multilingual datasets. Implementing a verification system that assesses LLM-generated labels, coupled with manual review for low-confidence outputs, presents a promising solution [54]. Such approaches are especially critical in Islamophobia research, where cultural sensitivity and accurate interpretation are paramount.

D. Future Directions for LLM-Driven Annotation

To enhance the performance of large language models (LLMs), future research should prioritize advancements in training methodologies for low-resource languages and improvements in prompt engineering [14],[41],[55],[56]. Striking a balance between automation efficiency and human expertise is essential for ensuring accurate and contextually relevant annotations.

An evaluation of LLM capabilities through Bloom's Taxonomy suggests that ChatGPT-4 excels in lower-order cognitive processes such as Remembering, Understanding, and Applying [57], [58]. Similar to human memory, the model effectively retrieves and categorizes information. However, studies indicate that GPT-4 may struggle with transferring learned concepts to new contexts, leading to occasional misinterpretations or omissions of critical details [59]. These challenges often arise from inherent model biases and a tendency to generate responses that maximize probabilistic likelihood rather than maintaining strict logical coherence [60] [61].

A hybrid approach that combines AI-generated pseudo-labels with human annotations could enhance both accuracy and cost efficiency in annotation tasks [37], [62]. While LLMs significantly reduce the workload associated with manual labeling, challenges related to consistency, bias mitigation, and contextual awareness persist. Research highlights the importance of human-in-the-loop validation to uphold annotation quality [63]. Active learning techniques, where human annotators review and refine low-confidence instances identified by LLMs, have shown promise in improving dataset reliability. This synergy between AI-driven efficiency and human intuition could establish a more robust annotation framework, particularly in sensitive areas such as Islamophobia detection.

III. METHODOLOGY

This study employs a comparative methodology to evaluate the alignment between human annotations and ChatGPT-generated annotations on a primary Islamophobia dataset. As illustrated in Fig. 1, the research framework includes data collection, and annotation by both human experts and ChatGPT, followed by an assessment phase utilizing Cohen's Kappa analysis and various performance metrics.

A. Workflow for Data Annotation

Fig. 1 illustrates a data processing workflow for analyzing content related to sensitive topics. It begins with data crawling from a platform (referred to as "X platform") using specific keywords like "Islam", "Muslim", "women", "hijab" "stopislam", and "terrorist". This collected data forms a research dataset, which then branches into two parallel processing paths. On the left path, a validation survey form is created, followed by human-based dataset labeling. On the right path, prompt engineering is developed, followed by dataset labeling using Generative Artificial Intelligence (ChatGPT). Both labeling approaches converge to create a dataset consisting of comments and labels. The final step involves evaluating inter-rater agreement between the human and AI labeling methods using Cohen Kappa analysis to measure consistency and reliability of the classifications

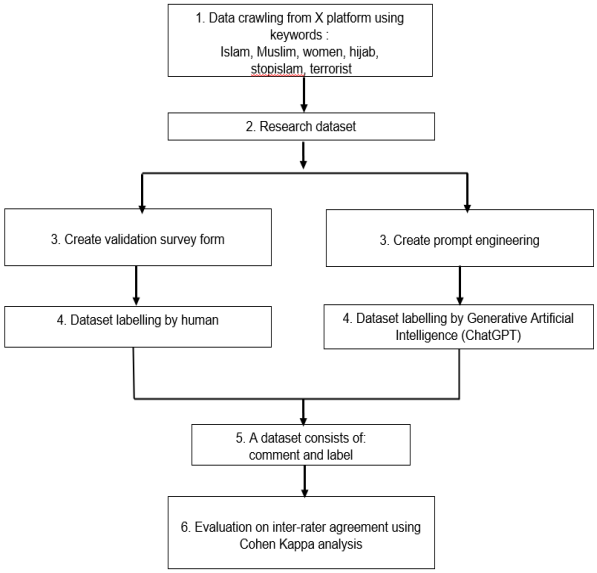

Fig. 1. Workflow for data annotation by human and LLM.

B. Dataset

The dataset consists of fifty publicly available tweets, manually collected from the X platform (formerly known as Twitter). The tweets were retrieved using a set of Islamophobia-related keywords, including Islam, Muslim, hijab, stopislam, jihadist, Islamic extremist, and terrorism, adapted from prior research on Islamophobia detection[64],[65],[66]. The selection of these keywords is justified by their relevance to the study of Islamic beliefs, practices, and the worldview of over a billion people. Each tweet was selected to represent a range of sentiments (positive, negative, and neutral) and was manually reviewed for relevance. The dataset was then annotated as either Islamophobia or Non-Islamophobia, as shown in Table I.

TABLE I. THE DATASETS WITH THEIR APPROPRIATE LABEL

Tweet ID	Tweet	Label
1	The best part of living in #Malaysia as a #Muslim majority country is being able to pray anywhere and at anytime. Alhamdulillah.	NON-ISLAMOPHOBIA
2	#Indonesia, #Malaysia and other #Asian countries often criticize the West's hypocrisy, citing lack of criticism on #Israel as main example	NON-ISLAMOPHOBIA
3	CCP #China people shitting and peeing in #Malaysia again. This time in #Islam's holiest place	ISLAMOPHOBIA

C. Human Annotators

Experts annotated the dataset to determine whether it contained Islamophobic content. Three specialists (i.e., an Islamic scholar, a language expert, and a clinical psychologist) conducted independent evaluations based on their respective areas of expertise. Their assessments established a baseline for accuracy and reliability, against which the performance of the LLM was compared. This diverse panel was selected to account for different perspectives on the topic.

The Islamic scholar provided deep insight into Islamic teachings, cultural nuances, and religious sensitivities, ensuring an accurate and contextually appropriate identification of Islamophobia. Their expertise was crucial in detecting subtle forms of discrimination and bias that might go unnoticed by those less familiar with Islamic culture and theology. The language expert analyzed linguistic structures, semantics, and pragmatics to ensure the sentiment analysis accurately captured the intended meaning and tone of the tweets. Their role was essential in identifying nuanced expressions of prejudice or bias embedded in language. The clinical psychologist contributed an understanding of human behavior, emotions, and the psychological impact of Islamophobic content, helping to assess the potential harm or distress it could cause to individuals and communities. Their expertise in bias and discrimination added depth to the evaluation process.

Due to their high-ranking positions within their institutions and other professional commitments, the panelists required three months to complete the annotation process for just fifty tweets.

D. LLM Annotation

This study utilized the ChatGPT 3.5 API to annotate the dataset, following OpenAI's official prompt examples for classification tasks. The prompt strategy was based on the structured approach outlined in OpenAI's documentation, where most prompts are framed as imperative sentences starting with action verbs like "classify" or "give". To ensure efficient processing within ChatGPT's token constraints, the dataset was fed into the model in batches of ten lines per prompt. This batch processing method was designed to align with ChatGPT's optimized token window size of 16,385 tokens. Table II provides an example of the prompt used in this study.

TABLE II. EXAMPLE OF TWEET, PROMPT, AND CHATGPT RESPONSE

Tweet	"CCP #China people shitting and peeing in #Malaysia again. This time in #Islam's holiest place"
Prompt	Assess the classification label of the following sentences either islamophobia or non-islamophobia.\nFormat of output: ID, label. "CCP #China people shitting and peeing in #Malaysia again. This time in #Islam's holiest place"
ChatGPT's response	ID: 1 Label: Islamophobia

E. Inter-Rater Analysis

The statistical measure Cohen's Kappa was utilized to evaluate the reliability and agreement between LLM and human annotators. Introduced by Cohen in 1960 [67], the Kappa coefficient quantifies chance-corrected agreement on a nominal scale between two raters. This measure is widely employed to assess inter-rater reliability, offering insights into the consistency and agreement among different annotators. Table III presents the formula for Cohen's Kappa statistical technique.

TABLE III. INTER-RATER AGREEMENT (COHEN KAPPA)

Statistical Techniques	Variable	Formula	Program and Tools
Inter-rater agreement measure of how reliably two raters measure the same	Nominal variable i. Islamophobia ii. Non-Islamophobia	$k = \frac{p_o - p_e}{1 - p_e}$ Po = observed agreement Pe expected agreement if a random agreement	Python

Table IV provides the interpretation of Cohen's Kappa agreement [68] which is used in this study.

TABLE IV. COHEN KAPPA LEVEL AGREEMENT

Cohen Kappa	Level of agreement
<0.00	Poor
0.00 - 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 - 1.00	Almost Perfect

F. Majority Voting Rule

In this study, we enlisted three experts to evaluate whether each tweet is Islamophobic or not. The experts represent different fields: Islamic studies, language studies, and clinical psychology. Each expert provides their classification for the tweets. To determine the final classification for each tweet, a majority vote was used due to its superior performance compared to other linear and metaclassifier combiners (Raza, 2018). The majority voting rule stipulates that the class with the highest number of votes is selected as the final decision, provided that the total exceeds 50%. The steps of the majority voting process are as follows:

- Collect Votes: Gather the classifications from all experts for each tweet.
- Count Votes: Count the number of votes for each category (Islamophobia or Non-Islamophobia).
- Determine Majority: The category with the most votes is chosen as the final classification for the tweet.

G. Performance Metrics

A confusion matrix is a table used to evaluate the performance of a classifier on a binary dataset. Table V presents the confusion matrix utilized in calculating the accuracy performance.

TABLE V. CONFUSION MATRIX

Actual	Prediction	
	Positive	**Negative**
Positive	TP	FP
Negative	FN	TN

The performance metrics used in this study are derived from the counts of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), as outlined below:

- True Positives (TP) – both the prediction and actual are yes.
- True Negatives (TN) – both the prediction and actual are no.
- False Positives (FP) – prediction is yes and actual is no.
- False Negatives (TN) – prediction is no and actual is yes.

Table VI shows the validation performance metrics used in this study, including Precision, Recall, F1-Score, and Accuracy. The validation performance analysis was implemented using Google Colab and Python programming.

TABLE VI. VALIDATION PERFORMANCE METRICS

Statistical Techniques	Explanation	Program and Tools
Precision	Positive predictive value in classifying the data instances.	Google Colab and Python Programming
Recall	Recall is also known as sensitivity or true positive rate	
F1-Score	An F1-score is a combination of the precision and the recall, providing a single score.	
Accuracy	Accuracy represents the number of correctly classified data instances over the total number of data instances.	

IV. RESULT

This section provides a detailed study of the data annotation summary, focusing on comparing the classification findings between three human annotators with different backgrounds (Islamic, Linguistic, and Psychological), as well as a Large Language Model (ChatGPT) and performance matrix.

A. Data Annotation Summary

Table VII presents a comparison of classification outcomes between three human annotators with different backgrounds (Islamic, Linguistic, and Psychological) and a Large Language Model (ChatGPT) in categorizing content into Islamophobic and Non-Islamophobic classifications. The data reveals varying levels of identification across the annotators. Human 1 (Islamic background) identified twenty-four instances of Islamophobia and twenty-six cases of non-Islamophobia. Human 2 (Linguist) classified eighteen cases as Islamophobic and thirty-two as non-Islamophobic. Human 3 (Psychologist) detected twenty-sseven instances of Islamophobia and twenty-three cases of non-Islamophobia. The LLM (ChatGPT) categorized eighteen cases as Islamophobic and thirty-two as non-Islamophobic, showing identical results to Human 2's annotations. Notably, there appears to be some variance in the identification of Islamophobic content among human annotators, with the psychologist identifying the highest number of Islamophobic instances (twenty-seven) while the linguist and LLM identified the lowest (eighteen each). This variation might reflect the different professional backgrounds and perspectives of the annotators in interpreting the content.

TABLE VII. DATA ANNOTATION RESULTS

Classification label	Human 1 (Islamic)	Human 2 (Linguist)	Human 3 (Psychologist)	LLM (ChatGPT)
Islamophobia	24	18	27	18
Non-Islamophobia	26	32	23	32

B. Agreement Levels Between LLM and Human Annotators in Data Annotation Tasks

Table VIII shows the level of agreement between LLM and the human annotators based on the classification of the tweets. The analysis of inter-rater agreement between ChatGPT and human annotators reveals notable variations in classification consistency across different domains of expertise. The findings indicate that the linguist demonstrated the highest concordance with the LLM, achieving a Kappa coefficient of 0.653, while the psychologist showed moderate agreement at 0.648, and the Islamic expert exhibited the lowest agreement level at 0.353. This hierarchical pattern of agreement can be attributed to several underlying factors.

TABLE VIII. INTER-RATER FINDINGS

Human annotator	LLM (ChatGPT)
Human 1 (Islamic)	0.353
Human 2 (Linguist)	0.653
Human 3 (Psychologist)	0.648
Human (Average)	0.632

The Cohen's Kappa analysis reveals varying levels of agreement between different human annotators and ChatGPT (LLM) in detecting Islamophobic content. The Islamic expert showed fair agreement ($\kappa = 0.353$), which was notably lower than other annotators, suggesting that ChatGPT may have limitations in capturing the subtle nuances and cultural contexts that an Islamic expert would recognize. In contrast, both the linguist and psychologist demonstrated substantial agreement with ChatGPT, scoring $\kappa = 0.653$ and $\kappa = 0.648$ respectively. The strong agreement with the language expert indicates that ChatGPT effectively aligns with linguistic patterns and markers of Islamophobia, while the high agreement with the psychologist suggests competency in recognizing psychological aspects of discriminatory language. The average agreement across all human annotators ($\kappa = 0.632$) falls within the substantial agreement range, indicating that ChatGPT performs well in Islamophobia detection. However, the variation in agreement levels, particularly the lower agreement with the Islamic expert, highlights areas for improvement in ChatGPT's understanding of cultural and religious nuances. This suggests that while ChatGPT is reliable for detecting linguistic and psychological patterns of Islamophobia, it may benefit from enhanced cultural-religious context understanding to match human expert judgment more closely.

Below is an example of the calculation for the Cohen Kappa analysis.

C. Human 2 (Linguist) and LLM

Step 1: Create a Confusion Matrix

The confusion matrix between the actual labels provided by Human Annotator 2 and the anticipated labels produced by the Large Language Model (LLM) in the Islamophobia detection test is shown in Table IX. The classification results are divided into four main categories in the table:

- True Positives (TP): Both the human annotator and the LLM accurately identified cases of Islamophobia.
- False Negatives (FN): LLM misclassified Islamophobic as non-Islamophobia.
- False Positives (FP): When the LLM mistakenly classifies non-Islamophobic as Islamophobic.
- True Negatives (TN): Cases that the human annotator and the LLM both appropriately categorized as non-Islamophobic.

The matrix calculation used to assess the model's classification performance uses this table as an example. When evaluating the accuracy of automatic annotation compared to human judgment, the confusion matrix offers valuable information on how well the model separates Islamophobic from non-Islamophobic content. The numbers in each cell indicate the count of instances for each classification outcome: True Positives (TP): 14, False Positives (FP): 4, False Negatives (FN): 4, and True Negatives (TN): 28.

Step 2: Calculate Observed Agreement (Po)

$$Po = (\text{Number of agreements}) / (\text{Total cases}) \quad (1)$$

Agreements = 14 + 28 = 42

Po = 42/50 = 0.84

Step 3: Calculate Expected Agreement by Chance (Pe)

$$Pe = (\text{Pe for Islamophobia}) + (\text{Pe for non-Islamophobia}) \quad (2)$$

For Islamophobia:

Expert 2 proportion: 18/50 = 0.36

ChatGPT proportion: 18/50 = 0.36

Pe for Islamophobia label = 0.36 × 0.36

= 0.1296

For non-Islamophobia:

Expert 2 proportion: 32/50 = 0.64

ChatGPT proportion: 32/50 = 0.64

Pe for non-Islamophobia label = 0.64 × 0.66

= 0.4096

Pe = 0.1296 + 0.4096 = 0.5392

Step 4: Calculate Cohen's Kappa

$$\kappa = (Po - Pe) / (1 - Pe) \quad (3)$$

$\kappa = (0.84 - 0.5392) / (1 - 0.5392)$

$\kappa = 0.3008/0.4608$

$\kappa = 0.653$

A score of 0.653 suggests that the agreement between Linguist and ChatGPT is substantial.

D. Human (Average) and LLM

Step 1: Determine the majority voting value (Refer Table X).

- Collect Votes: Gather the classifications from all experts for each tweet.
- Count Votes: Count the votes for each category (Islamophobia or Non-Islamophobia).
- Determine Majority value: The category with the most votes is chosen as the final classification for the tweet.

TABLE IX. CONFUSION MATRIX

		Predicted Label (LLM)		**TOTAL**
		Islamophobia	**Non-Islamophobia**	
Actual Label (Human 2)	Islamophobia	14 (TP)	4 (FN)	18
	Non-Islamophobia	4 (FP)	28 (TN)	32
TOTAL		18	32	50

TABLE X. MAJORITY VOTING VALUE CALCULATION

Tweet	**Human 1**	**Human 2**	**Human 3**	**Count Vote (Phobia)**	**Count Vote (Non)**	**Majority Voting Value**
1	Non	Non	Non	0	3	Non
2	Non	Non	Non	0	3	Non
3	Phobia	Non	Phobia	2	1	Phobia
4	Phobia	Phobia	Phobia	3	0	Phobia
5	Phobia	Non	Phobia	2	1	Phobia

*non = non -islamophobia, Phobia = Islamophobia

*Human 1 = Islamic

*Human 2 = Linguist

*Human 3 =Psychologist

Step 2: Create a new Confusion Matrix

Table XI shows the confusion matrix between the actual labels provided by the Human average and the anticipated labels produced by the Large Language Model (LLM) in the Islamophobia detection test. The numbers in each cell indicate the count of instances for each classification outcome: True Positives (TP): 16, False Positives (FP): 2, False Negatives (FN): 7, and True Negatives (TN): 25.

TABLE XI. CONFUSION MATRIX

		Predicted Label (LLM)		TOTAL
		Islamophobia	**Non-Islamophobia**	
Actual Label (Human-average)	Islamophobia	16 (TP)	7 (FN)	23
	Non-Islamophobia	2 (FP)	25 (TN)	27
TOTAL		18	32	50

Step 3: Calculate Observed Agreement (Po)

$$Po = (\text{Number of agreements}) / (\text{Total cases}) \quad (1)$$

Agreements = 16 + 25 = 41

Po = 41/50 = 0.82

Step 4: Calculate Expected Agreement by Chance (Pe)

$$Pe = (Pe \text{ for Islamophobia}) + (Pe \text{ for non-Islamophobia}) \quad (2)$$

For Islamophobia:

Expert 2 proportion = 23/50 = 0.46

ChatGPT proportion = 18/50 = 0.36

Pe for Islamophobia label = 0.46 × 0.36

= 0.1656

For non-Islamophobia:

Expert 2 proportion = 27/50 = 0.54

ChatGPT proportion = 32/50 = 0.64

Pe for non-Islamophobia label = 0.54 × 0.64

= 0.3456

Pe = 0.1656 + 0.3456 = 0.5112

Step 5: Calculate Cohen's Kappa

$$\kappa = (Po - Pe) / (1 - Pe) \quad (3)$$

κ = (0.82 - 0.5112) / (1 - 0.5112)

κ = 0.3088 / 0.4888

κ = 0.632

E. LLM Performance Based on the Approximation Accuracy

Refer to Table XI for the confusion matrix in the form of a heatmap. The heatmap represents the LLM's performance, with actual labels from humans (average) on the vertical axis and predicted labels from the LLM on the horizontal axis. The value in the table is used to calculate the performance based on the approximation accuracy. Below is the equation to calculate the performance:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (4)$$

$$\text{Precision} = TP / (TP + FP) \quad (5)$$

$$\text{Recall} = TP / (TP + FN) \quad (6)$$

$$\text{F1 score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (7)$$

Table XII presents the performance evaluation metrics for the classification model using approximation accuracy. The table includes four key metrics: accuracy, precision, recall, and F1-Score.

TABLE XII. APPROXIMATION ACCURACY

Accuracy = (16 + 25) / (16 + 25 + 2 + 7) × 100 = 41 / 50 × 100 = 0.82	**Precision** = 16 / (16 + 2) = 16 / 18 = 0.88
Recall = 16 / (16 + 7) = 16 / 23 = 0.695	**F1 Score** = 2 × (0.88 × 0.695) / (0.88+ 0.695) = 2 × (0.6116) / (1.583) = 0.772

V. DISCUSSION

Based on the result in Table XII, the analysis of ChatGPT's classification performance in identifying Islamophobic content reveals both strengths and limitations in its capabilities. Based on the classification metrics analysis, the model demonstrates strong overall performance with 82% accuracy across all predictions, correctly classifying forty-one out of fifty cases. Its precision score of 88.8% was particularly impressive, indicating high reliability when content was flagged as Islamophobic, with only two false positives out of sixteen positive predictions.

However, the model's recall performance was considerably an average at 69.5%, suggesting a significant limitation in its ability to identify all instances of Islamophobia. Of the sixteen actual Islamophobic cases in the dataset, the model only successfully identified fourteen, missing cases. This difference between precision and recall resulted in an F1 score of 77.2%, which indicates good overall performance, though there is room for improvement. The lower F1 score compared to precision suggests that recall could be improved. These findings indicate that ChatGPT adopts a conservative approach in its classification of Islamophobic content, prioritizing precision over recall. While this cautious stance minimizes false accusations of Islamophobia, it comes at the cost of failing to identify a substantial number of genuine cases. This behavior pattern suggests a deliberate design choice for handling sensitive content, though it raises important considerations about the model's effectiveness in comprehensive content moderation.

The data indicates that there were seven "Total Missed Cases", which represent instances, where ChatGPT failed to identify Islamophobic content when it was present. Additionally, there were two "Total False Cases", which indicates situations, where ChatGPT incorrectly flagged content as Islamophobic when it was not. These findings suggest potential limitations in ChatGPT's ability to accurately detect and classify Islamophobic content, with a notably higher rate of False Negatives (missed cases) compared to False positives (incorrect flags). This data could be valuable for understanding the model's current capabilities and areas for improvement in content moderation related to religious bias and discrimination.

The tweet "For now, it is status quo for #Christians in #Malaysia on the escalating row over the use of the word 'Allah' as a translation for the Christian God in the #Muslim-majority nation", is an example of the classification disagreement between human annotators and ChatGPT. The content of this tweet refers to an ongoing interfaith issue in Malaysia, specifically surrounding the contested use of "Allah" by non-Muslims, which has been a sensitive topic in the country given its implications on religious identity and freedoms in a Muslim-majority context. Human annotators may have identified this tweet as Islamophobic due to its potential to highlight religious tension or imply a critique of policies perceived as biased in favor of the Muslim majority. The phrasing could be interpreted as subtly presenting Muslims or Muslim-majority policies as restrictive towards Christians, thus indirectly invoking a stereotype of Islam as intolerant or limiting religious freedom. ChatGPT, however, may have classified this tweet as non-Islamophobic due to the absence of explicit negative language or hostile sentiment directed towards Islam or Muslims. The tweet is largely informational, stating the current situation without clearly insulting language, which could lead the model to overlook the potentially implicit bias or underlying critique that human annotators detected.

This case shows how ChatGPT might miss subtle cues tied to interfaith or political undertones, especially where the language is indirect, and specific negative implications about Islam are implied rather than directly stated. In terms of classification metrics, ChatGPT achieved 82% accuracy, 88% precision, 69.5% recall, and a 77.2% F1-Score. These values are comparable to results reported in other LLM annotation studies, where models performed well on general sentiment tasks but showed variability in detecting minority or sensitive expressions [36], [38], [41]. The high precision suggests that ChatGPT is conservative in its classifications, minimizing false positives—an approach consistent with OpenAI's design for handling sensitive content. However, the lower recall indicates that the model may miss instances of Islamophobia that are implicit or linguistically complex, a pattern also noted in recent LLM evaluation studies [42], [43].

These findings reinforce the importance of incorporating domain expertise in the annotation of cultural or religiously sensitive content. While ChatGPT can serve as a reliable tool for preliminary screening or large-scale annotation, human-in-the-loop systems remain essential for capturing deeper contextual meanings, particularly in domains like Islamophobia detection. Studies such as AnnoLLM [2], CoAnnotating [46], and MEGAnno+ [48] also advocate for hybrid approaches, where human validation is integrated with LLM outputs to improve reliability and reduce biases.

Moreover, this study contributes to ongoing efforts in evaluating the real-world applicability of LLMs in underrepresented language and cultural contexts, where high-quality labeled data is scarce. By benchmarking ChatGPT's annotations against experts from Islamic studies, linguistics, and psychology, the study provides a multidisciplinary evaluation framework that can inform future research on automated content moderation and hate speech detection. In particular, the use of Cohen's Kappa as a validation metric enables robust assessment of model-human agreement, addressing concerns about reproducibility and inter-rater reliability raised in earlier annotation quality reviews [5], [6], [45].

The classification challenge surrounding the tweet regarding religious terminology in Malaysia can be evaluated critically using Bloom's Taxonomy, namely its higher-order cognitive domains of analysis, evaluation, and synthesis. While basic computational models typically operate at the lower levels of Bloom's hierarchy, focusing primarily on remembering (recognition of explicit linguistic elements) and understanding (surface-level comprehension of textual content), the nuanced identification of potential Islamophobic discourse requires cognitive processes aligned with the taxonomy's more sophisticated levels. To analyze implicit bias, it must be able to break down complex linguistic structures (analysis), critically evaluate the underlying sociopolitical context and potential rhetorical implications (evaluation), and finally, synthesize multiple interpretative layers that go beyond literal textual content.

This research employed a focused methodological approach combining expert panel evaluation with a majority voting system to assess Islamophobia detection. The expert panel was strategically composed of diverse stakeholders, including Islamic scholars, sociologists, extremism researchers, linguistic experts, and social media analysts, ensuring a comprehensive evaluation perspective. The majority voting system was implemented with a structured protocol, where three to five expert evaluators assessed each case using a standardized scoring rubric. Final classifications were determined based on a threshold of greater than 60% agreement among the experts. This dual-component methodology was specifically chosen to balance the need for diverse expert insights with a quantifiable decision-making process. While this approach may have limitations, it provides a practical and systematic framework for evaluating the accuracy of Islamophobia detection in computational systems.

VI. LIMITATIONS

The limits of this study show fundamental issues in employing LLM such as ChatGPT to annotate the nuanced, culturally sensitive text. The LLM struggles to perceive and apply cultural and religious nuances consistently, as evidenced by a poorer Cohen's Kappa agreement with an Islamic studies expert ($\kappa = 0.353$) compared to linguist ($\kappa = 0.653$) and clinical psychologist experts ($\kappa = 0.648$). This shows that despite its great language capabilities, ChatGPT lacks the depth of context

required to understand subtleties that experts in Islamic studies may easily detect. As a result, the model may misclassify tweets with indirect or implicit biases, highlighting a potential limitation to its efficacy as an independent annotator in fields requiring great cultural sensitivity.

Another limitation is the LLM's conservative approach, which prioritizes precision above recall. While this may reduce false positives (when non-Islamophobic content is mistakenly categorized as Islamophobic), it also results in missing occurrences of true Islamophobia. The study found that ChatGPT's recall performance, at 69.5%, is significantly lower than its precision, suggesting its cautious approach yet resulting in missed instances of Islamophobic content. This trade-off affects its usefulness in tasks that require thorough content detection since missing harmful content can be worse than occasionally misclassifying safe content. The conservative classification method may be consistent with ChatGPT's design goals, but it implies a limited ability to handle edge circumstances or content, that, although not Islamophobic, contains more nuanced possibly destructive views.

Using a small dataset (fifty tweets) poses another limitation: the model's performance may not generalize to larger or more diversified datasets. The short sample size reduces the statistical robustness of performance measurements such as Cohen's Kappa and F1 scores, which may inflate perceived model efficacy. Furthermore, the relatively quick and informal style of tweets may not accurately represent the range of Islamophobic information seen on social media or other platforms. This constraint requires additional study using larger, more diverse datasets to validate the model's capabilities across various content, various types, and levels of implicit bias.

VII. Recommendations

To enhance the accuracy and cultural sensitivity of LLMs in annotating Islamophobia-related content, several key improvements are necessary. First, domain-specific fine-tuning on a larger and more diverse dataset can help address the model's limitations in detecting cultural variations in Islamophobic narratives. Training on datasets annotated by Islamic studies experts, with a focus on subtle and implicit forms of prejudice, can improve the model's ability to recognize nuanced biases that were previously overlooked. Additionally, continuous fine-tuning based on expert feedback will allow the model to adapt to evolving linguistic and cultural expressions of Islamophobia, making it more effective in identifying implicit bias and complex contextual cues.

A hybrid annotation approach that integrates LLM-based automation with human validation is another crucial improvement, particularly for culturally sensitive content. Human experts can review low-confidence cases flagged by the model, ensuring greater accuracy while maintaining efficiency in large-scale annotation tasks. This human-in-the-loop strategy is particularly beneficial for social media content moderation, where precise classification is essential to avoid mislabeling ambiguous or indirect expressions of Islamophobia. Furthermore, refining the model's ability to process indirect language such as passive phrasing, coded language, or ambiguous terms often found in Islamophobic discourse can help minimize False Negatives and improve annotation accuracy.

Finally, addressing the study's limitations regarding sample size and dataset diversity is critical for improving generalizability. Expanding data collection to multiple social media platforms and content formats will enable LLMs to better adapt to various linguistic styles and modes of expression. This broader dataset will enhance the model's ability to detect Islamophobic rhetoric across different online spaces. Collaborating with interdisciplinary experts in linguistics, psychology, and Islamic studies during dataset creation and analysis can further enrich the model's contextual understanding, making it a more reliable tool for detecting Islamophobia across diverse digital communities.

VIII. Conclusion

This study demonstrates both the potential and limitations of employing LLMs for annotating culturally sensitive text, addressing the challenges of manual annotation, including inconsistencies, resource intensity, and scalability constraints. ChatGPT exhibited substantial agreement with human annotators, particularly those specializing in linguistics and psychology, reinforcing its viability for automating large-scale data annotation. By reducing time and resource requirements, LLMs offer a scalable alternative to traditional manual labeling approaches. Moreover, the model's strong precision and recall scores indicate its effectiveness in identifying overt Islamophobic content, positioning it as a useful tool for preliminary screenings or as a supplementary aid in sentiment analysis tasks.

A notable strength of LLMs lies in their ability to maintain annotation consistency, minimizing variability stemming from human subjectivity, an essential factor in large-scale labeling tasks requiring uniformity. This consistency enhances the reliability of labeled datasets, providing a robust foundation for further refinement by domain experts. However, the lower agreement between ChatGPT and Islamic studies specialists highlights its shortcomings in detecting implicit and complex forms of bias, underscoring the need for greater cultural and contextual sensitivity in AI-driven annotation models.

In terms of cognitive processing, LLMs demonstrate proficiency in lower-order cognitive tasks, as outlined in Bloom's Taxonomy, excelling in "Remembering" and "Understanding" by systematically categorizing explicit Islamophobic content based on predefined criteria. However, the model falls short in higher-order reasoning skills such as "Analyzing" and "Evaluating", which are crucial for discerning subtle biases and nuanced linguistic expressions. These findings suggest that while LLMs present significant advantages in efficiency and scalability, a hybrid approach integrating human expertise for complex contextual cases may offer a more balanced and culturally aware annotation framework.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, the author(s) used [Scispace/conducting literature review] to discover and analyse

the scientific study and create a matrix table for literature review. After using this tool or service, the matrix table was uploaded to ChatGPT and Claude to improve the language of the work. After that, the author(s) reviewed and edited the content as needed and took(s) full responsibility for the content of the published article.

REFERENCES

[1] H. D. Zajac, N. R. Avlona, F. Kensing, T. O. Andersen, and I. Shklovski, "Ground Truth Or Dare: Factors Affecting The Creation Of Medical Datasets For Training AI," in AIES 2023 - Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, Inc, Aug. 2023, pp. 351–362.

[2] X. He et al., "AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators," Mar. 2023, [Online]. Available: http://arxiv.org/abs/2303.16854

[3] Z. Tan et al., "Large Language Models for Data Annotation: A Survey," Feb. 2024, [Online]. Available: http://arxiv.org/abs/2402.13446

[4] D. Hovy and S. L. Spruit, "The Social Impact of Natural Language Processing," in In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2016, pp. 591–598.

[5] J. C. Klie, B. Webber, and I. Gurevych, "Annotation Error Detection: Analyzing the Past and Present for a More Coherent Future," Computational Linguistics, vol. 49, no. 1, pp. 157–198.

[6] R. Artstein and M. Poesio, "Survey Article Inter-Coder Agreement for Computational Linguistics," Computational Linguistics, vol. 34, no. 4, pp. 555–596.

[7] A. Schmidt and M. Wiegand, "A Survey on Hate Speech Detection using Natural Language Processing," Proceedings of the fifth international workshop on natural language processing for social media, pp. 1–10.

[8] K. Fort, G. Adda, and K. B. Cohen, "Last Words Amazon Mechanical Turk: Gold Mine or Coal Mine?," Computational Linguistics, vol. 37, no. 2, pp. 413–420, 2011.

[9] P. Törnberg, "ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning," arXiv preprint arXiv:2304.06588, 2023.

[10] A. A. Ahmed et al., "Arabic Text Detection Using Rough Set Theory: Designing a Novel Approach," 2023, Institute of Electrical and Electronics Engineers Inc.

[11] M. Bartolo, A. Roberts, J. Welbl, S. Riedel, and P. Stenetorp, "Beat the AI: Investigating Adversarial Human Annotation for Reading Comprehension," Transactions of the Association for Computational Linguistics, 2020 8, vol. 8, pp. 662–678.

[12] A. Saeed et al., "Topic Modeling based Text Classification Regarding Islamophobia using Word Embedding and Transformers Techniques," ACM Transactions on Asian and Low-Resource Language Information Processing.

[13] K. S. Kalyan, "A survey of GPT-3 family large language models including ChatGPT and GPT-4," Natural Language Processing Journal, vol. 6, p. 100048.

[14] Y. Zhu, P. Zhang, E.-U. Haq, P. Hui, and G. Tyson, "Can ChatGPT Reproduce Human-Generated Labels? A Study of Social Computing Tasks; Can ChatGPT Reproduce Human-Generated Labels? A Study of Social Computing Tasks," arXiv preprint arXiv:2304.10145., 2023.

[15] Z. Ashktorab et al., "AI-Assisted Human Labeling: Batching for Efficiency without Overreliance," Proc ACM Hum Comput Interact, vol. 5, no. CSCW1, Apr. 2021.

[16] Y. Naraki et al., "Augmenting NER Datasets with LLMs: Towards Automated and Refined Annotation," arXiv preprint arXiv:2404.01334, Mar. 2024.

[17] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré, "Snorkel: rapid training data creation with weak supervision," VLDB Journal, vol. 29, no. 2–3, pp. 709–730, May 2020.

[18] C. Allen, Reconfiguring Islamophobia: A Radical Rethinking of a Contested Concept. Springer Nature, 2020.

[19] E. Bleich, "What is islamophobia and how much is there? theorizing and measuring an emerging comparative concept," American Behavioral Scientist, vol. 55, no. 12, pp. 1581–1600, Dec. 2011.

[20] I. Zempi and A. Imran, The Routledge International Handbook of Islamophobia, vol. 1. London: Routledge, 2019.

[21] E. Omran, E. Al Tararwah, and J. Al Qundus, "A comparative analysis of machine learning algorithms for hate speech detection in social media," Online J Commun Media Technol, vol. 13, no. 4, Oct. 2023.

[22] T. Davidson, D. Bhattacharya, and I. Weber, "Racial Bias in Hate Speech and Abusive Language Detection Datasets," in Proceedings of the Third Abusive Language Workshop, 2019, pp. 25–35.

[23] E. W. Pamungkas, D. Galih, P. Putri, and A. Fatmawati, "Hate Speech Detection in Bahasa Indonesia: Challenges and Opportunities," IJACSA) International Journal of Advanced Computer Science and Applications, vol. 14, no. 6, 2023.

[24] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of NAACL-HLT 2019, 2019, pp. 4171–4186.

[25] B. Vidgen, T. Yasseri, and H. Margetts, "Islamophobes are not all the same! A study of far-right actors on Twitter," Journal of Policing, Intelligence and Counter Terrorism, vol. 17, no. 1, pp. 1–23, 2022.

[26] S. L. Blodgett, S. Barocas, H. D. Iii, and H. Wallach, "Language (Technology) is Power: A Critical Survey of 'Bias' in NLP," in In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 5454–5476.

[27] E. Aldreabi, J. M. Lee, and J. Blackburn, "Using Deep Learning to Detect Islamophobia on Reddit," The International FLAIRS Conference Proceedings. Florida Online Journals, vol. 36, 2023.

[28] Q. Mehmmod, A. Kaleem, Q. Mehmood, and I. Siddiqi, "Islamophobic Hate Speech Detection from Electronic Media using Deep Learning," Mediterranean conference on pattern recognition and artificial intelligence. Cham: Springer International Publishing., 2021.

[29] B. Vidgen and L. Derczynski, "Directions in abusive language training data, a systematic review: Garbage in, garbage out," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Public Library of Science, Dec. 2020, pp. 3550–3564.

[30] M. Sap et al., "Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection," in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics, 2022, pp. 4159–4175.

[31] E. Aldreabi, K. M. Harahsheh, M. D. Chhangani, C.-H. Chen, and J. Blackburn, "Beyond Binary: Revealing Variations in Islamophobic Content with Hierarchical Multi-Class Classification," Proceedings of the International Florida Artificial Intelligence Research Society Conference, vol. 30, Oct. 2024.

[32] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," ACM Comput Surv, vol. 51 (4), no. 85, Jul. 2018.

[33] A. A. Almazroi, A. A. Shah, and F. Mohammed, "Social Media and Online Islamophobia: A Hate Behavior Detection Model," International Journal of Engineering Trends and Technology, vol. 71, no. 11, pp. 27–32, 2023.

[34] M. Mathebula and A. Modupe, "ChatGPT as a Text Annotation Tool to Evaluate Sentiment Analysis on South African Financial Institutions," IEEE Access/10.1109/ACCESS.2024.3464374, 2024.

[35] B. Ding et al., "Is GPT-3 a Good Data Annotator?," arXiv preprint arXiv:2212.10450., Dec. 2022

[36] F. Gilardi, M. I. Alizadeh, and M. I. Kubli, "ChatGPT outperforms crowd workers for text-annotation tasks," in Proceedings of the National Academy of Sciences, 120(30), e2305016120., PNAS, 2023. doi: 10.1073/pnas.

[37] S. Wang, Y. Liu, Y. Xu, C. Zhu, and M. Zeng, "Want To Reduce Labeling Cost? GPT-3 Can Help," Aug. 2021, [Online]. Available: http://arxiv.org/abs/2108.13487

[38] M. Belal, J. She, and S. Wong, "Leveraging ChatGPT As Text Annotation Tool For Sentiment Analysis," arXiv preprint arXiv:2306.17177, 2023.

[39] T. H. Nguyen and K. Rudra, "Human vs ChatGPT: Effect of Data Annotation in Interpretable Crisis-Related Microblog Classification," in

WWW 2024 - Proceedings of the ACM Web Conference, Association for Computing Machinery, Inc, May 2024, pp. 4534–4543.

[40] S. Wu et al., "BloombergGPT: A Large Language Model for Finance," Mar. 2023, [Online]. Available: http://arxiv.org/abs/2303.17564

[41] A. H. Nasution and A. Onan, "ChatGPT Label: Comparing the Quality of Human-Generated and LLM-Generated Annotations in Low-Resource Language NLP Tasks," IEEE Access, vol. 12, pp. 71876–71900, 2024.

[42] N. Pangakis, S. Wolken, and N. Fasching, "Automated Annotation with Generative AI Requires Validation," arXiv preprint arXiv:2306.00176., May 2023.

[43] M. Heseltine and B. Clemm von Hohenberg, "Large language models as a substitute for human experts in annotating political text," Research and Politics, vol. 11, no. 1, Jan. 2024.

[44] M. Alizadeh et al., "Open-Source LLMs for Text Annotation: A Practical Guide for Model Setting and Fine-Tuning," Jul. 2023, [Online]. Available: http://arxiv.org/abs/2307.02179

[45] J.-C. Klie, R. E. de Castilho, and I. Gurevych, "Analyzing Dataset Annotation Quality Management in the Wild," Computational Linguistics, vol. 50, no. 3, pp. 817–866, Jul. 2023.

[46] M. Li et al., "CoAnnotating: Uncertainty-Guided Work Allocation between Human and Large Language Models for Data Annotation," arXiv preprint arXiv:2310.15638., Oct. 2023.

[47] X. He et al., "AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators," arXiv preprint arXiv:2303.16854., Mar. 2023.

[48] H. Kim, K. Mitra, R. L. Chen, S. Rahman, and D. Zhang, "MEGAnno+: A Human-LLM Collaborative Annotation System," arXiv preprint arXiv:2402.18050, Feb. 2024,

[49] X. Sun et al., "Pushing the Limits of ChatGPT on NLP Tasks," arXiv preprint arXiv:2306.09719, Jun. 2023.

[50] Y. Zhu, Z. Yin, G. Tyson, E. U. Haq, L. H. Lee, and P. Hui, "APT-Pipe: A Prompt-Tuning Tool for Social Data Annotation using ChatGPT," in WWW 2024 - Proceedings of the ACM Web Conference, Association for Computing Machinery, Inc, May 2024, pp. 245–255.

[51] J. Kaikaus, H. Li, and R. J. Brunner, "Humans vs. ChatGPT: Evaluating Annotation Methods for Financial Corpora," in Proceedings - 2023 IEEE International Conference on Big Data, BigData 2023, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 2831–2838.

[52] A. K. Chowdhury et al., "Harnessing large language models over transformer models for detecting Bengali depressive social media text: A comprehensive study," Natural Language Processing Journal, vol. 7, p. 100075, Jun. 202.

[53] S. Thapa, N. Usman, and N. Mehwish, "From Humans to Machines: Can ChatGPT-like LLMs Effectively Replace Human Annotators in NLP Tasks?," in Proceedings of the Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics (ACL), 2023, pp. 11173–11195.

[54] X. Wang, H. Kim, S. Rahman, K. Mitra, and Z. Miao, "Human-LLM Collaborative Annotation Through Effective Verification of LLM Labels," in Conference on Human Factors in Computing Systems - Proceedings, Association for Computing Machinery, May 2024, pp. 1–21.

[55] Z. Al Nazi, Md. R. Hossain, and F. Al Mamun, "Evaluation of open and closed-source LLMs for low-resource language with zero-shot, few-shot, and chain-of-thought prompting," Natural Language Processing Journal, vol. 10, p. 100124, Mar. 2025.

[56] M. Son, Y. J. Won, and S. Lee, "Optimizing Large Language Models: A Deep Dive into Effective Prompt Engineering Techniques," Applied Sciences (Switzerland), vol. 15, no. 3, Feb. 2025.

[57] N. Duong-Trung, X. Wang, and M. Kravčík, "BloomLLM: Large Language Models Based Question Generation Combining Supervised Fine-Tuning and Bloom's Taxonomy," in European Conference on Technology Enhanced Learning, Cham : Springer Nature Switzerland, 2024, pp. 93–98.

[58] L. W. Anderson, D. R. Krathwohl, Bloom, and B. Samuel, A taxonomy for learning, teaching, and assessing : a revision of Bloom's taxonomy of educational objectives. Longman, 2001.

[59] S. P. Nagavalli, S. Tiwari, and W. Sarma, "Large Language Models and NLP: Investigating Challenges, Opportunities, and the Path to Human-Like Language Understanding Independent Researcher 1 Independent Researcher 2 Independent Researcher," International Research Journal of Engineering and Technology, 2024.

[60] S. Lappin, "Assessing the Strengths and Weaknesses of Large Language Models," J Logic Lang Inf, vol. 33, no. 1, pp. 9–20, Mar. 2024.

[61] A. Herrmann-Werner, T. Festl-Wietek, F. Holderried, and J. Griewatz, "Assessing ChatGPT's Mastery of Bloom's Taxonomy using psychosomatic medicine exam questions," J Med Internet Res, vol. e52113, no. 26, 2024.

[62] T. Zhang, X. Chen, C. Qu, A. Yuille, and Z. Zhou, "Leveraging Ai Predicted And Expert Revised Annotations In Interactive Segmentation: Continual Tuning Or Full Training?," in 2024 IEEE International Symposium on Biomedical Imaging (ISBI), IEEE, 2024, pp. 1–5.

[63] A. Bonet-Jover, R. Sepúlveda-Torres, E. Saquete, P. Martínez-Barco, A. Piad-Morffis, and S. Estevez-Velarde, "Applying Human-in-the-Loop to construct a dataset for determining content reliability to combat fake news," Eng Appl Artif Intell, vol. 126, p. 107152, Nov. 2023.

[64] E. Aldreabi and J. Blackburn, "Enhancing Automated Hate Speech Detection: Addressing Islamophobia and Freedom of Speech in Online Discussions," in Proceedings of the 2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2023, Association for Computing Machinery, Inc, Nov. 2023, pp. 644–651.

[65] S. Kh Hamed, M. Juzaiddin Ab Aziz, and M. Ridzwan Yaakub, "Disinformation Detection About Islamic Issues On Social Media Using Deep Learning Techniques," Malaysian Journal of Computer Science, vol. 36, no. 3, pp. 242–270, Jul. 2023.

[66] B. Vidgen and T. Yasseri, "Detecting weak and strong Islamophobic hate speech on social media," Journal of Information Technology and Politics, vol. 17, no. 1, pp. 66–78, Jan. 2020.

[67] J. Cohen, "A Coefficient of Agreement for Nominal Scales," Educational and Psychological Measurement, 20(1), 37-46., vol. 20, no. 1, pp. 37–46, 1960.

[68] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," 1977.