Validation of a large language models-based socio-political sentiment analysis system with a retrieval-augmented generation approach

Sidharta Sidharta [a,1,*], Hady Pranoto [a,2], Frederik Masri Gasa [b,3], Nur Kholis [b,4], Ardvin Kester S. Ong [c,5]

[a] Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia
[b] Communication Science Department, Faculty of Digital Communication and Hotel & Tourism, Bina Nusantara University, Jakarta, Indonesia
[c] School of Industrial Engineering and Engineering Management, Mapua University, Philippines
[1] sidharta@binus.ac.id; [2] hadypranoto@binus.ac.id; [3] frederik.gasa@binus.edu; [4] nur.kholis001@binus.ac.id; [5] aksong@mapua.edu.ph
* corresponding author

ARTICLE INFO

Article history
Received January 1, 2026
Revised March 11, 2026
Accepted March 12, 2026
Available online August 31, 2026

Keywords
Large Language Model (LLM)
Sentiment analysis
Retrieval-Augmented Generation
Socio-political discourse
System-level validation

ABSTRACT

Socio-political discourse on digital platforms generates large volumes of unstructured textual data, posing challenges for reliable sentiment analysis and contextual interpretation. Conventional sentiment analysis approaches often lack contextual grounding, while direct use of Large Language Models (LLMs) may introduce hallucinations, bias, and excessive generalization. To address these issues, this study proposes an LLM-based sentiment analysis system that integrates multi-source data acquisition, zero-shot sentiment classification, Retrieval-Augmented Generation (RAG), and generative reasoning for contextual interpretation. Unlike prior studies that mainly focus on improving model accuracy, this research emphasizes system-level validation of the entire analytical pipeline. The evaluation includes sentiment calibration using a Golden Dataset, data ingestion performance analysis, retrieval quality assessment based on semantic distance, and generative evaluation using an LLM-as-a-Judge framework. Experimental results indicate that the system provides stable data acquisition and produces coherent analytical reports. However, retrieval noise and neutrality bias introduced by LLM alignment mechanisms still affect the groundedness of generated outputs. These findings highlight the importance of retrieval quality and demonstrate the potential of LLMs as reasoning and evaluation components for analyzing complex socio-political discourse.

© 2026 The Author(s).
This is an open access article under the CC–BY-SA license.

1. Introduction

The rapid development of public discourse on social media has significantly changed the dynamics of shaping public opinion, particularly in social and political issues that are evolving rapidly, emotionally, and often ambiguously in the expression of sentiment. Sensitive topics such as policy reform, corruption, and public demands such as "17+8 people's demands" often generate complex and emotionally charged discourse, where sentiment polarity is not always expressed explicitly. This complexity poses challenges in conducting automated sentiment analysis, especially when the text contains irony, sarcasm, or implicit criticism. Therefore, sentiment analysis systems that are able to understand the semantic context and nuances of discourse are becoming increasingly crucial.

Deep learning-based approaches such as CNN, LSTM, and transformers have been widely used in traditional sentiment analysis [1], [2]. However, these methods generally rely on human-labeled datasets and is domain-specific. The advent of Large Language Models (LLMs) opens new opportunities through zero-shot and few-shot classification approaches, which allow models to infer without domain-specific retraining [3], [4]. In addition, the Retrieval-Augmented Generation (RAG) integration allows the system to combine semantic retrieval with generative capabilities to generate context-based reports [5], [6].

Nonetheless, the use of LLMs in sentiment analysis presents new challenges. First, many systems utilize labels generated by other LLMs as pseudo-ground-truth, which has the potential to result in teacher-model bias and inter-model instability. Studies show that LLMs can provide different interpretations of the same data, especially in evaluative tasks and opinion classification [3], [4]. Second, alignment through Reinforcement Learning from Human Feedback (RLHF) may affect the model's sensitivity to political content [7], [8]. RLHF training often penalizes outputs that appear politically aggressive or controversial, encouraging models to produce safer responses. As a result, in sentiment classification tasks involving political discourse, the model may prefer neutral interpretations rather than explicitly negative labels. Third, in RAG architecture, retrieval quality greatly determines the level of groundedness of the final generation, so the distribution of semantic similarity needs to be analyzed explicitly [5].

Although various studies have focused on improving the accuracy of sentiment classification or optimizing model architecture, relatively few studies have evaluated LLM-based sentiment analysis systems holistically at the pipeline level, particularly those that include data crawling, zero-shot classification, RAG-based semantic retrieval, and integrated generative output evaluation [2]. As a result, understanding of the limits of system reliability and inter-component interactions in the LLM-based sentiment analysis system is still limited in the current literature. In addition, the concept of a performance ceiling due to the use of system ground-truth labels and the impact of divergence between LLMs on the interpretation of evaluation metrics are still rarely explicitly discussed. Thus, there is a research gap in terms of system-level evaluation frameworks for integrated LLM-based sentiment analysis systems.

This study proposes a system-level validation framework for an LLM-based sentiment analysis system that includes four main components: (1) model calibration using independent golden datasets, (2) multi-source data ingestion performance evaluation, (3) retrieval quality analysis based on L2 semantic distance distribution, and (4) generative evaluation using the LLM-as-a-Judge approach. This approach measures not only classification accuracy but also architectural bias, realistic performance ceilings, and the relationship between retrieval dispersion and generative groundedness.

Based on these challenges, this study addresses the following research questions:

RQ1: How reliable is zero-shot sentiment classification for Indonesian socio-political discourse when evaluated using cross-LLM calibration?

RQ2: How does semantic retrieval quality in the RAG pipeline influence the groundedness of generative analytical reports?

RQ3: To what extent can LLM-as-a-Judge evaluation provide a reliable indicator of generative reasoning quality in socio-political sentiment analysis systems?

RQ4: What architectural limitations emerge in an end-to-end LLM-based sentiment analysis pipeline for dynamic socio-political discourse?

Unlike conventional studies that evaluate sentiment models or RAG pipelines separately, this study introduces a system-level validation framework that explicitly analyses the interaction between sentiment classification, retrieval quality, and generative reasoning. In particular, the framework links semantic distance distribution in the retrieval stage with groundedness scores produced by the LLM-as-a-Judge

evaluation, enabling an integrated analysis of how retrieval noise propagates into generative interpretation. In particular, the contributions of this research are as follows:

- Propose a system-level validation framework for an LLM-based sentiment analysis system integrated with the RAG pipeline.
- Identify and analyze neutrality bias as an architectural effect in NLI-based zero-shot classifications.
- Introduce the concept of performance ceiling in LLMs' pseudo-ground-truth-based evaluation.
- Link semantic distance distribution of retrieval with the quality of generative groundedness.
- Provide an integrated evaluation of the crawling–classification–generation pipeline in the context of Indonesian-language socio-political discourse.

With this approach, the research not only contributes to the implementation aspect of the system but also provides a methodological perspective on the validation of LLM-based sentiment analysis system models in a dynamic and ambiguous data environment.

2. Related Works

2.1. Sentiment Analysis on Social Media and Socio-Political Discourse

Sentiment analysis has long been used to understand public opinion on social media and online platforms. Early studies have made extensive use of classical machine learning and statistical approaches to classify sentiment on Twitter data and online forums, particularly in political and public policy contexts [9]–[11]. This approach has since been developed with the use of deep learning models such as CNN, LSTM, and hybrid architectures to improve sentiment classification performance on large-scale data [2], [12], [13].

Nonetheless, most of those studies rely on static datasets and offline analysis processes. In socio-political contexts, texts are often ambiguous, contain sarcasm, or use informal language, thus decreasing the accuracy of conventional models and making it difficult to generalize across issues and time [10], [14]. These limitations encourage the exploration of more flexible and adaptive approaches.

2.2. Large Language Models for Sentiment Analysis

The emergence of LLMs introduced a new approach in sentiment analysis through zero-shot and few-shot learning capabilities, allowing models to be applied without retraining on specific datasets [3], [4], [15]–[18]. A number of studies have shown that LLMs are able to capture more complex emotional nuances and contexts than traditional deep learning models, including in the domains of finance and long-dimensional texts [4], [19]–[22]. However, a number of studies have also highlighted the limitations of LLMs in sentiment analysis, especially regarding bias, consistency, and interpretability. LLMs are known to exhibit systemic biases, including a bias towards neutrality in political texts, which are influenced by safety alignment mechanisms and RLHF [7], [8]. In addition, the direct application of LLMs without external context risks producing hallucinations, so the output of the model is not always tied to the data being analyzed [23], [24]. These findings suggest that the use of LLMs for sentiment analysis requires additional strategies to be more reliable in real-world applications.

2.3. Retrieval-Augmented Generation (RAG) in LLM-Based Sentiment Analysis Systems

Retrieval-Augmented Generation (RAG) is proposed as a solution to increase the attachment of LLM output to external data sources by combining retrieval and text generation processes [25]–[28]. This approach has been applied to a variety of domains, such as document-based chatbots, decision support systems, and clinical data analysis [6], [27], [29]. The studies show that RAGs can reduce hallucinations and improve the relevance of answers by providing a more structured context for LLMs. Although RAG is growing in popularity, most research focuses on question answering tasks or specific non-political domains. Evaluation of RAG in the context of sentiment analysis, particularly in real-time systems based on multi-source social media and online news, is still relatively limited. In addition, few

studies have quantitatively evaluated the quality of retrieval and attributed it to the quality of generative output of LLMs [14], [26], [27].

2.4. LLM System Evaluation and the LLM-as-a-Judge Approach

Evaluating LLM outputs is challenging, as traditional metrics such as BLEU or ROUGE are less able to represent the quality of reasoning, coherence, and attachment to context [26], [30]. Therefore, the LLM-as-a-Judge approach began to be used to assess the quality of model output automatically and more holistically [14], [16], [31]. This approach utilizes independent LLMs as evaluators to reduce the model's judgment bias against its own outputs. Evaluating the operational performance of LLM-based systems has also attracted increasing attention, particularly regarding inference efficiency and reliability in real-time applications [32]. Related research also emphasizes the importance of systemic evaluation that includes aspects of groundedness, robustness against noise retrieval, and potential generative bias [16], [23], [24]. However, the integration of this kind of evaluation in an end-to-end sentiment analysis system is still rare, especially in sensitive and dynamic socio-political contexts.

2.5. Position and Contribution of This Research

Based on the above literature review, it can be concluded that previous studies have generally focused on improving model performance or LLM/RAG applications in specific domains, with limitations to the overall evaluation of the system. This research positions itself as an intelligent informatics study that integrates real-time data crawling, zero-shot sentiment analysis, and RAG-based reasoning in one end-to-end pipeline. In contrast to previous research, the main focus of this study is component-by-component evaluation and analysis of the practical limitations of LLMs in real applications, particularly in multi-source-based socio-political discourse. This approach expands the perspective of evaluation from simply measuring classification accuracy to architecture-level system validation.

3. System Architecture

The proposed system architecture is designed as an end-to-end intelligent informatics system that integrates real-time data acquisition, sentiment analysis, and reasoning based on LLMs. Pipeline-based architectures are increasingly used in LLM-based sentiment analysis system to integrate information retrieval, analysis, and text generation processes in an integrated manner [27], [33]. As illustrated in Fig. 1, the architecture integrates three main processes: multi-source data acquisition, zero-shot sentiment classification, and RAG-based reasoning. This pipeline structure enables the system to transform raw socio-political discourse into contextual analytical reports through a combination of retrieval and generative reasoning. This architecture allows for integration between sentiment classification processes and generative reasoning to produce a more contextual interpretation of public discourse.

3.1. User Input Layer

The system begins with Input Keywords provided by the user to represent the topic or issue to be analyzed. This keyword triggers the entire analysis process in the system and is used in parallel by two main modules, namely the sentiment analysis module and the reasoning module, thus ensuring the consistency of the context in both processing paths.

3.2. Data Acquisition Layer

This layer is responsible for collecting text data from various online platforms through a keyword-based scraping and crawling process. The main data sources include online news, discussion forums, and social media comments. The data obtained is then processed through the initial cleanup stage to eliminate duplication, text noise, and non-linguistic elements before being used in the next stage of analysis

4. Method

This section describes the methods used to validate the proposed LLM-based sentiment analysis system. Unlike approaches that emphasize model development or optimization, this study evaluates the

performance of the entire end-to-end system, including sentiment analysis modules, data acquisition, RAG-based retrieval, and system generative outputs. The evaluation framework consists of four main components, each of which represents a critical stage in the system architecture.

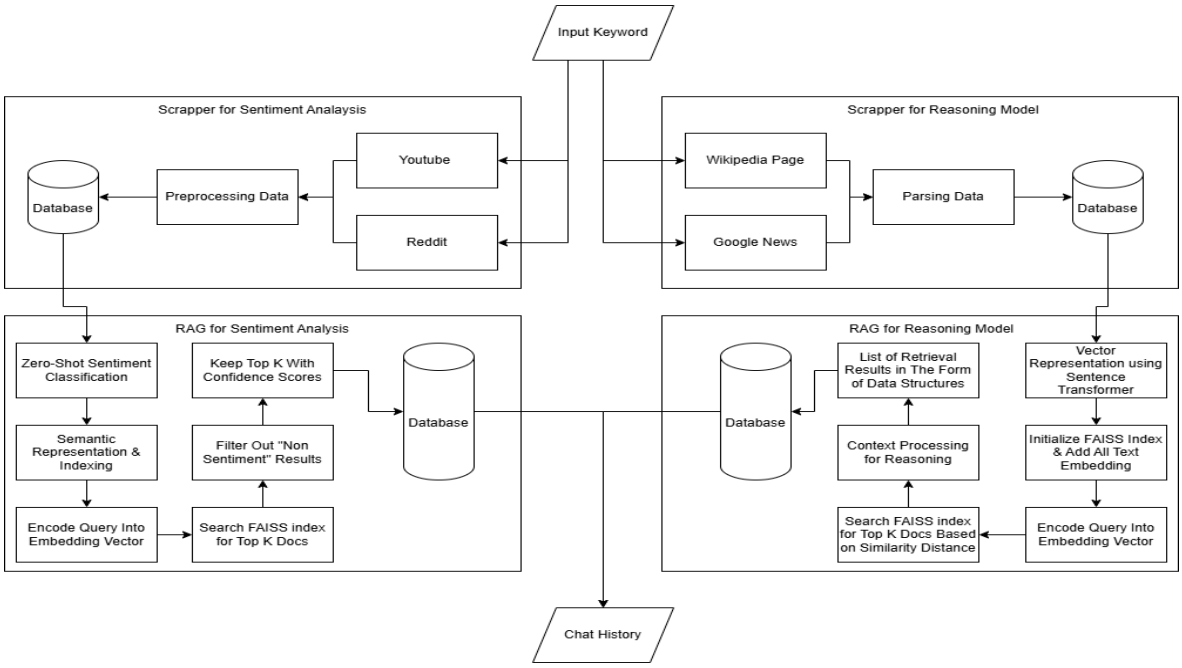


Fig. 1. System architecture of the proposed LLM-based sentiment analysis system integrating data scraping, sentiment analysis, and RAG-based reasoning

4.1. Sentiment Model Calibration Using a Golden Dataset

The sentiment calibration stage in this study aims to evaluate the reliability of the XLM-RoBERTa-Large-XNLI multilingual zero-shot model in analyzing Indonesian-language socio-political discourse. In contrast to the previous approach that used a small golden dataset, this study builds a large-scale evaluation corpus derived from two LLMs developed for the Southeast Asian context, namely Sahabat AI (SAI) and SeaLLM (SEA). The initial dataset was obtained through an exact-text matching process between the labeling results of the two models, resulting in 4,489 integrated entries.

The initial distribution of data showed a significant class imbalance, with about 76% of the sample labeled negative and 24% labeled positive. This inequality has the potential to produce biased accuracy estimates due to the dominance of the majority class. Therefore, a merge-first, sample-second strategy is applied, where all data from the merger is maintained first, then balanced sampling is carried out based on the SAI label. A total of 1,000 negative samples and 1,000 positive samples were randomly selected (random_state = 42) to form a controlled evaluation corpus of 2,000 samples. This class-balancing strategy ensures that evaluation metrics such as precision, recall, and F1-score are not distorted by class imbalance, as has been emphasized in the modern sentiment analysis evaluation literature [2], [34].

Before evaluating the performance of XLM-RoBERTa, an analysis of inter-LLM agreement was conducted to determine the limits of empirical reliability. Although human annotation is commonly considered the gold standard in sentiment analysis, this study adopts cross-LLM calibration due to the large scale and dynamic nature of socio-political discourse data. Previous research has shown that multiple independent LLM annotators can provide consistent pseudo-ground-truth labels when inter-model agreement is explicitly measured. Therefore, the agreement between SAI and SEA is used as a practical empirical reliability bound in this system-level evaluation framework. Of the 2,000 balanced samples, 1,467 samples (73.35%) showed label conformity between SAI and SEA, while 533 samples (26.65%) showed disagreement. Analysis of the direction of disagreement showed an asymmetrical pattern, with 407 cases in which SAI labeled positive while SEA labeled negative, and 126 cases with the opposite pattern. These findings indicate a difference in calibration tendencies between the two models,

where SeaLLM tends to be more conservative in interpreting socio-political expressions as negative sentiments. Cohen's Kappa coefficient is calculated to measure agreement beyond random chance and yields a value $\kappa = 0.467$ which belongs to the category of moderate agreement according to the standard interpretation scale [35]. Thus, the 73.35% agreement rate can be seen as an empirical performance ceiling in the context of LLM-based sentiment analysis system automated annotation. This variation in alignment between models has also been reported in LLMs evaluation and benchmarking studies that show differences in interpretation between models in text classification and assessment tasks [3], [4].

Because XLM-RoBERTa-Large-XNLI produces three sentiment classes (negative, neutral, positive), while SAI and SEA use binary (negative/positive) schemes, two evaluation strategies were carried out to address label space mismatches. The first strategy maps neutral predictions into negative classes as a conservative approach, assuming that in political discourse, neutral expression often represents implicit criticism. The second strategy excludes all neutral predictions and calculates performance metrics only on binary predictions. Adaptations of these kind of evaluation strategy are commonly used in Natural Language Inference-based model benchmarking studies when the model label space is not fully aligned with the dataset's annotation scheme [4], [15]. The metrics used in the evaluation included accuracy, precision, recall, macro F1-score, weighted F1-score, and Cohen's Kappa.

To ensure the statistical robustness of the evaluation results, a 95% confidence interval is calculated using the standard binomial proportion formula. In the XLM (filtered strategy) evaluation of SAI with an accuracy of 0.6918 and $N = 1.408$, a confidence interval of around [0.667, 0.716] was obtained. Meanwhile, in the comparison of SAI and SEA with an accuracy of 0.7335 and $N = 2.000$, the confidence interval was in the range [0.713, 0.753]. Relatively narrow intervals indicate the stability of performance estimation under balanced sampling conditions. In addition, the McNemar test was applied to evaluate the consistency of paired predictions between evaluation strategies. The test results showed a statistically significant difference ($p < 0.05$), so the observed performance improvement could not be considered a mere random variation. The use of confidence intervals and paired significance tests is a common practice in the evaluation of modern LLM classification and benchmarking systems [16], [36]. With this approach, sentiment calibration not only assesses the suitability of the model to a single annotation source, but also measures cross-model reliability and empirical agreement limits in the context of Indonesian socio-political discourse.

4.2. Data Ingestion Performance Evaluation

Evaluation at the data ingestion layer aims to measure the stability and responsiveness of the system in conducting multi-source data acquisition in semi-real time. This module is responsible for crawling and preprocessing text from several public platforms, including Google News, Reddit, and YouTube comments, based on user-entered keywords. This kind of web crawling-based data acquisition approach has been widely used in information extraction systems from social media to support the analysis of emotions and public opinion [37]. Because the system is designed to support sentiment analysis and reasoning based on Retrieval-Augmented Generation (RAG), the performance of this layer is crucial in determining the overall operational feasibility of the system.

The main metric used in this evaluation is latency, which is defined as the total time (in seconds) it takes from the time the query is entered until the scraped text data is ready for use at the next stage of analysis. Latency measurement covers the entire acquisition process, including document retrieval, content parsing, and initial text cleanup. Latency was selected as the primary metric because the system is intended to monitor dynamic socio-political issues, so response time becomes a more relevant indicator of performance than just the volume of data obtained.

To test the system's resilience to variations in topic complexity, evaluations were conducted on three query categories with different characteristics, namely: (1) niche topics with limited scope and low volume, (2) general topics with wide scope and high volume, and (3) specific issue topics with high public discourse intensity. This approach allows for comparative analysis of latency based on the variation in the number and heterogeneity of documents that must be processed by the system.

All measurements were performed in a consistent system configuration with no changes to model parameters or infrastructure during the experiment. Thus, the observed latency variation can be attributed to topic characteristics and content volume, rather than to changes in the system's technical configuration. This evaluation is not intended to optimize crawling performance, but rather to ensure that the data acquisition layer has sufficient stability and responsiveness as the foundation for LLM-based sentiment analysis systems and reasoning pipelines.

4.3. Maintaining the Integrity of the Specifications

To ensure experimental validity and prevent evaluation bias, this study strictly applies the principle of component isolation and integrity of system specifications. The entire evaluation was conducted in a zero-shot scenario, with no additional fine-tuning processes on the XLM-RoBERTa-Large-XNLI model or on the generative model used in the pipeline. As such, the reported performance reflects the model's built-in generalization capabilities, rather than the result of adaptation to a specific evaluation dataset. This kind of zero-shot approach is commonly used in LLM benchmarking to maintain evaluation objectivity and avoid bias due to retraining on evaluation datasets [4], [15].

In addition, there is no data leakage between the data used in the sentiment calibration evaluation and the corpus used in the retrieval stage in the Retrieval-Augmented Generation (RAG) architecture. The evaluation dataset is built from the annotation results of the SAI and SEA models, while the retrieval system takes context based on dynamic crawling results from public sources (YouTube and Reddit) through user queries. This separation ensures that the model does not evaluate the same text that is used as a generative context, thus avoiding performance inflation due to information redundancy.

The integrity of the specification is also maintained by keeping the model configuration consistent throughout the experiment. Model parameters, architectural versions, and evaluation strategies are not changed between experiments. The class balancing strategy is applied before the evaluation without modifying the model's internal distribution. Thus, the observed performance change can be attributed to the evaluation strategy, not to the system configuration change.

In the generative stage, a separation of roles between the generator model and the evaluator model is also applied to avoid self-preference bias, i.e., the tendency of the model to assess its own output more positively. The Mistral 7B model is used as a report generator, while the Llama-3-70B-Instruct model is used as an independent evaluator in the LLM-as-a-Judge framework. This cross-evaluation approach has been recommended in the modern LLM evaluation literature to improve the objectivity of assessment [16], [36].

Finally, all experiments were carried out with reproducible procedures, including the use of fixed random seeds in the sampling process and systematic recording of metrics. Although annotation is performed automatically by LLMs, inter-LLM agreement analysis and statistical significance testing are applied to mitigate potential systematic bias. With this approach, the research not only evaluates the performance of the model, but also ensures that the evaluation process takes place within a controlled and transparent experimental framework.

4.4. RAG Retrieval Quality Evaluation using Semantic Distance

The retrieval stage in the Retrieval-Augmented Generation (RAG) architecture acts as a context-providing mechanism for generative models. The quality of the context captured directly affects the quality of generative output, because the principle of "garbage in, garbage out" applies strongly to LLM-based sentiment analysis systems. Therefore, retrieval evaluation is not only carried out descriptively, but also quantitatively using embedding-based semantic similarity measurements.

In this study, a text vector representation was generated using the all-MiniLM-L6-v2 sentence embedding model, and the document search process was carried out using FAISS IndexFlatL2 based on Euclidean distance (L2 distance). This kind of semantic embedding-based retrieval approach is widely used in modern dense retrieval systems to improve the relevance of context in LLM and RAG pipelines [38]. For each user query, the embedding query is compared to the document embedding in the database, and the number of documents with the lowest distance is taken as context. Retrieval quality is measured

through the distribution of L2 distance values between the query and the retrieved document. Conceptually, smaller L2 distance values indicate higher semantic similarity between the query and the retrieved document. The distance distribution was then analyzed to identify three main indicators: (1) the mean value of the distance, (2) the distribution (standard deviation), and (3) the presence of outliers with high distances that could potentially represent semantic noise.

The results of the analysis showed that the retrieval distribution had an average distance shift in the range of 0.35–0.40, with a relatively widespread distribution compared to the ideal retrieval configuration. The presence of multiple documents at a higher distance indicates that the system retrieves not only highly relevant documents, but also semantically adjacent documents; that is, they have a general topical relevance but do not fully answer the specific intent of the query. This phenomenon has the potential to cause noise injection into the context window, which then affects the quality of groundedness at the generative stage.

Previous literature has shown that RAG performance is highly sensitive to retrieval quality and embedding alignment to the target language [5], [26]. In the context of this study, the use of all-MiniLM-L6-v2, which is optimized primarily for English, has the potential to result in a mismatch of semantic representations in Indonesian texts. Cross-lingual retrieval mismatch has been reported to degrade semantic similarity performance when embedding models trained primarily on English corpora are applied to other languages [30]. This cross-lingual embedding mismatch may explain a wider distribution of distances than expected. The retrieval evaluation in this study is not intended to optimize the embedding model, but rather to identify the limits of overall system performance. By quantifying the semantic distance distribution, this study correlates retrieval quality with groundedness scores at the generative stage, thus enabling a systemic analysis of the interactions between components in the RAG architecture.

4.5. Generative Evaluation using LLM-as-a-Judge

The final stage in the system architecture is the evaluation of the quality of generative reports generated through the Retrieval-Augmented Generation (RAG) approach. After the process of crawling, sentiment classification, and context retrieval is performed, the generative model (Mistral 7B) generates an analytical report based on the synthesis of the collected data. Because generative outputs are narrative and reasoning-oriented, their evaluation cannot rely on token similarity-based metrics such as BLEU or ROUGE that are commonly used in translation or summarization tasks. Therefore, this study adopts the LLM-as-a-Judge paradigm, which is the use of an independent LLM as an evaluator of the quality of generative outputs.

The evaluator model used is Llama-3-70B-Instruct, which differs from the generator model to prevent self-preference bias, i.e., the tendency of the model to assess its own output more positively. This kind of evaluation approach has been widely used in modern LLM evaluation research to reduce assessment bias and increase the objectivity of generative evaluation [16], [26]. The evaluation was conducted using a structured prompt that asked the evaluator model to assess three main aspects of report quality, namely coherence, groundedness, and neutrality. Coherence refers to the logical consistency and readability of the report's structure, including inter-paragraph continuity and clarity of argumentation. Groundedness measures the extent to which the claims and interpretations in the report are actually supported by the documents obtained at the retrieval stage, thus minimizing the possibility of unfounded hallucinations or generalizations. Meanwhile, neutrality assesses whether the report maintains an objective academic tone without reinforcing the emotional or provocative language that may be present in the raw source text.

Each dimension was graded on an ordinal scale of 1–5 based on the evaluator's response to the prompt that had been standardized and unchanged during the experiment. No manual adjustments were made to the scores produced by the evaluator model. Although this approach allows for a more comprehensive assessment of reasoning quality and context alignment than conventional automated metrics, this study recognizes that LLM-based sentiment analysis system evaluations remain model-dependent and do not completely replace human evaluations. Therefore, the score obtained is treated as a comparative indicator

of system quality in a controlled experimental context. Although this approach allows for a more holistic evaluation of generative quality, a number of studies have shown that LLM-based sentiment analysis system evaluations remain limited by evaluator model bias, so score results should be understood as comparative indicators in specific experimental contexts [16].

5. Results and Discussion

This section presents the results of the evaluation and discussion of the proposed LLM-based sentiment analysis system. The results were compiled following four aspects of evaluation that have been described in the Methods section, namely sentiment model calibration, data acquisition performance, retrieval quality in the RAG pipeline, and LLM-based sentiment analysis system generative output quality. This approach ensures clear traceability between evaluation methods and empirical findings.

5.1. Results of Sentiment Model Calibration

The sentiment calibration evaluation was carried out using a balanced dataset of 2,000 samples built through merge-first and balanced sampling strategies (1,000 negative and 1,000 positive). This evaluation aims to measure the performance of XLM-RoBERTa in the context of Indonesia's socio-political discourse and compare it with the level of agreement between LLMs as the limit of empirical reliability. Initial analysis was carried out on the agreement between two annotator models, namely Sahabat AI (SAI) and SeaLLM (SEA). Of the 2,000 balanced samples, 1,467 samples (73.35%) showed label conformity, while 533 samples (26.65%) showed disagreement. Cohen's Kappa value of 0.467 indicates a moderate level of agreement. These findings suggest that even two models that were jointly developed for the context of Southeast Asian languages are not fully aligned in interpreting political sentiment. Therefore, this figure of 73.35% can be viewed as an empirical performance ceiling in this dataset configuration, as the level of agreement between annotators' methods limits the maximum performance that can be realistically achieved by other classification models evaluated against the same annotation reference.

The evaluation of XLM-RoBERTa was first carried out using a binary mapping strategy (neutral → negative). At this configuration, the model achieves an accuracy of 0.662. Fig. 2 presents three confusion matrices under balanced sampling. The left panel compares XLM-RoBERTa predictions with Sahabat AI labels, the middle panel compares predictions with SeaLLM labels, and the right panel illustrates the agreement between the two LLM annotators.

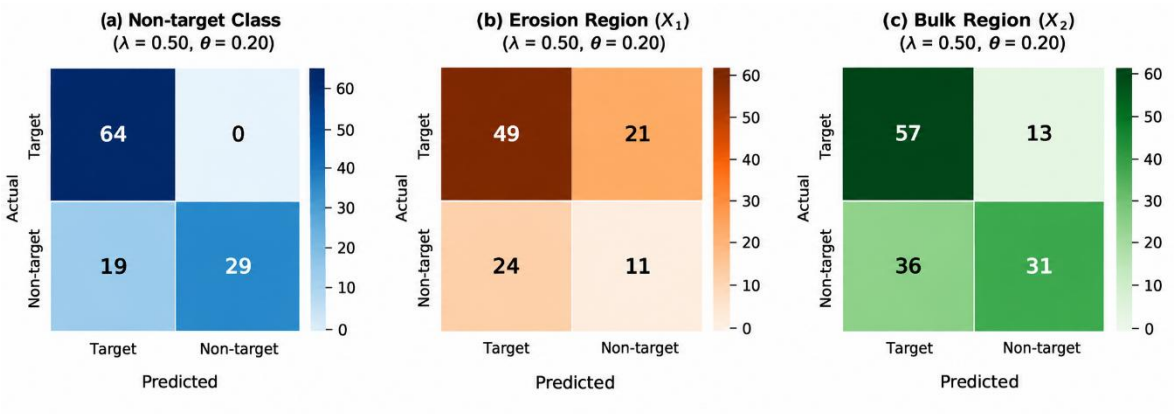


Fig. 2. Confusion matrices under balanced sampling (N = 2000). The left panel shows XLM-RoBERTa predictions compared with Sahabat AI, the middle panel shows predictions compared with SeaLLM, and the right panel illustrates inter-LLM agreement between the two reference models

Based on the XLM-RoBERTa confusion matrix on SAI (left panel), a total of 744 negative samples were correctly classified as negative, while 256 negative samples were incorrectly classified as positive. In the positive class, 580 samples were correctly predicted, while 420 samples were misclassified as negative. This pattern shows that the misclassification errors are relatively balanced between classes, but the level

of false negative in the positive class is quite significant. When compared to the right panel (SAI vs SEA), even between LLMs there are 407 cases of difference in the positive class and 126 in the negative class, which confirms the calibration variation between models.

When the filtered evaluation strategy was applied—by excluding all neutral predictions, the accuracy increased to 0.6918 in the remaining 1,408 samples. The 95% confidence interval for this configuration is in the range [0.667–0.716], which indicates the stability of the estimated performance under balanced sampling conditions. The McNemar test confirmed that the difference in performance between binary and filtered mapping strategies was statistically significant ($p < 0.05$), so the improvement could not be considered a random variation.

Further analysis showed that 29.6% of the entire sample was predicted to be neutral by XLM-RoBERTa. Fig. 3 shows the distribution of XLM-RoBERTa predictions under balanced sampling, highlighting the relatively high proportion of neutral classifications. This distribution highlights the model's tendency to produce neutral predictions in texts that are political in nature, suggesting the existence of neutrality bias in zero-shot classifications.

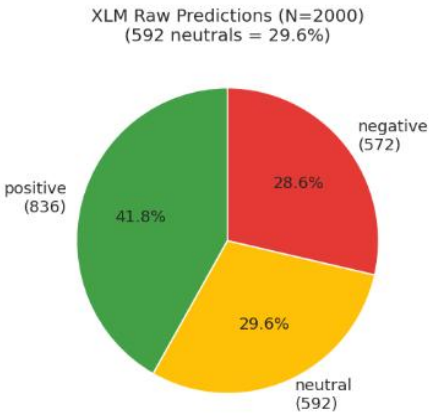


Fig. 3. Distribution of XLM-RoBERTa predictions showing 29.6% neutral classification under balanced sampling

The relatively high proportion of neutral predictions in the context of political discourse indicates a tendency for neutrality bias in Natural Language Inference-based models (NLI). Models tend to be more cautious in classifying expressions of implicit criticism as negative, especially when they do not contain explicit linguistic indicators that are confrontational in nature. This phenomenon explains why filtered strategies perform better than binary mapping, as many neutral predictions come from a sample that the LLM annotator categorizes as negative.

When compared to the inter-LLM agreement limit of 73.35%, XLM-RoBERTa's best performance (69.18%) is within about four percentage points of the empirical ceiling. This gap suggests that some of the performance differences can be attributed to the architectural limitations of NLI-based zero-shot models, particularly in addressing ambiguity and implicit criticism in socio-political discourse. Thus, these results not only represent the model's degree of accuracy but also map the limits of the reliability of automatic sentiment classifications in Indonesia's socio-political context.

The comparative summary in Table 1 shows that the best configuration was obtained in the filtered evaluation strategy of SAI with an accuracy of 0.692 and an F1-weighted of 0.689. This value is within a relatively small distance of the inter-LLM agreement level (0.734), which is positioned as the empirical performance ceiling. The slightly overlapping confidence interval between the XLM-RoBERTa best-configured configuration ([0.668–0.716]) and the inter-LLM limit ([0.714–0.753]) suggests that some of the performance gaps may be within the limits of statistical estimated variability.

Table 1. Summary of sentiment calibration performance under balanced sampling (N = 2000). Inter-LLM agreement (SAI vs SEA) represents the empirical performance ceiling, while XLM-RoBERTa results are reported under both binary (neutral→negative) and filtered evaluation strategies with corresponding 95% confidence intervals

Comparison	N	Accuracy	95% CI	Kappa	F1 (neg)	F1 (pos)	F1 (weighted)
XLM (neutral→neg) vs SAI	2000	0.662	[0.641–0.683]	0.324	0.688	0.632	0.660
XLM (neutral→neg) vs SEA	2000	0.548	[0.526–0.569]	0.051	0.630	0.418	0.554
XLM (no neutral) vs SAI	1408	0.692	[0.668–0.716]	0.375	0.645	0.728	0.689
XLM (no neutral) vs SEA	1408	0.503	[0.477–0.529]	0.054	0.523	0.482	0.508

Fig. 4 shows that the filtered evaluation strategy (accuracy 0.692) is close to the empirical reliability limit determined by inter-LLM agreement (0.734). A performance difference of 0.042 (4.2 percentage points) indicates that the available upgrade space is relatively limited in this dataset configuration. This visualization is consistent with the previous trust interval analysis, where XLM-RoBERTa's best performance estimates show a slight overlap with inter-LLM boundaries. Thus, the performance of the classifier cannot be evaluated absolutely, but must be contextualized against the level of reliability of the annotation reference used.

In contrast, the performance against SEA was relatively lower on both strategies, with a Kappa value close to zero. This indicates a significant difference in label calibration between the two LLM annotators. These findings confirm that variations between annotator models have a direct impact on classifier evaluation, so accuracy interpretations should always be contextualized against the level of reference agreement used.

These findings suggest that the performance of sentiment classification in the socio-political domain should be understood as a function of annotator reliability, not just model optimization. When two LLM annotators show a moderate level of agreement, then a classifier that is close to that level is statistically at the limit of practical reliability. As such, future performance improvements are likely to depend more on the harmonization of annotation schemes and cross-model consensus than simply refinement of the classifier architecture.

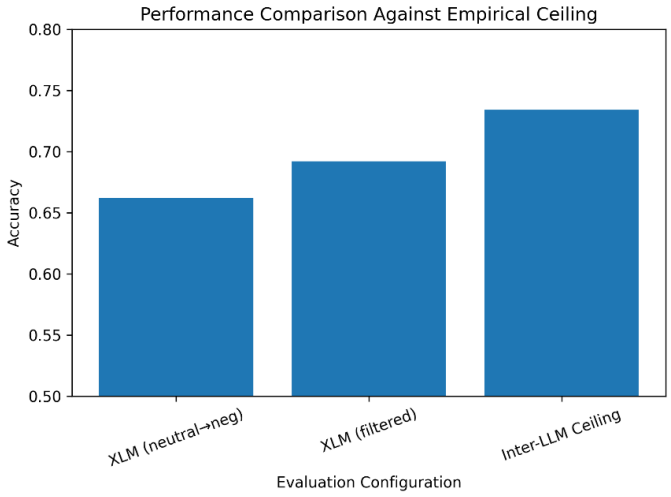


Fig. 4. Accuracy comparison between XLM-RoBERTa classification strategies and inter-LLM agreement (SAI vs SEA), which represents the empirical performance ceiling under balanced sampling

5.2. Result of Data Ingestion Performance

All latency measurements were conducted using the Google Colab cloud environment with a standard runtime configuration (Intel Xeon CPU and approximately 12 GB RAM) without GPU acceleration under stable internet connectivity. The crawling and preprocessing modules were executed using Python-based scraping scripts without GPU acceleration. The same runtime configuration was maintained throughout all experiments to ensure consistent latency measurement across different query scenarios. Fig. 5 presents the average data ingestion latency for three query categories with different topic complexities. The relatively small latency variation across queries indicates that the scraping and preprocessing modules remain stable despite differences in discourse volume and topic breadth. Latency is defined as the total time it takes the system from the time the query is entered until the entire scraping, parsing, and initial cleanup process of text is complete and the data is ready to be processed at the next stage of analysis. The test results showed that the "17+8 people's demands" niche query had an average latency of about 1.20 seconds. The topic "Reformasi" produces a latency of about 1.22 seconds, while the topic "Corruption" shows the highest latency, which is about 1.30 seconds. The latency difference between topics is relatively small (±0.1 seconds), which suggests that variations in query complexity do not result in significant performance fluctuations at the data acquisition layer.

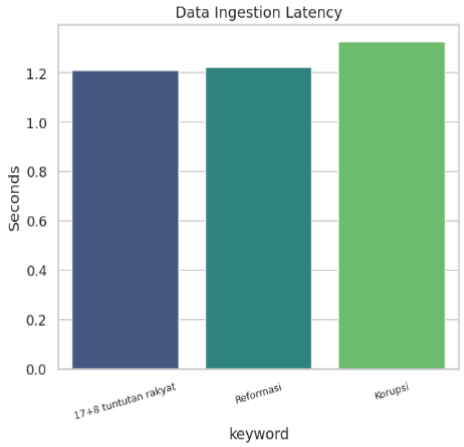


Fig. 5. Average data ingestion latency across different query complexities for the multi-source scraping module

Although issues such as "Corruption" in general have a wider and heterogeneous scope of public discourse, the observed increase in latency remains within a narrow range. This indicates that the scraping and preprocessing modules are able to adapt to the volume and diversity of content without experiencing significant performance degradation. Overall, these findings suggest that the data ingestion layer on the proposed system has sufficient stability and responsiveness to support semi-real-time sentiment analysis. With an average latency of under 1.5 seconds across all query categories, it can serve as a reliable foundation for Retrieval-Augmented Generation pipelines and LLM-based sentiment analysis system reasoning at a later stage.

5.3. RAG Retrieval Quality Analysis

Fig. 6 illustrates the distribution of L2 semantic distances between query embeddings and retrieved documents in the RAG pipeline. While lower distances indicate highly relevant documents, the presence of higher-distance values suggests moderate retrieval noise. This distribution provides empirical evidence that some retrieved documents are only partially relevant, which may influence the groundedness of the generated reports. Analysis was performed on the top-K document of the representative query "17+8 people's demands" using the all-MiniLM-L6-v2 embedding model and FAISS IndexFlatL2-based search. L2 distance values are in the range of about 0.4 to 1.2, with the majority of documents concentrated in the interval of 0.6 to 1.0. The presence of documents with low distances (0.4–0.6) indicates that the retriever system is capable of finding semantically relevant content. However, a

distribution that widens to above 1.0 indicates that some documents in the top K have only general semantic proximity, not strong topical relevance.

This distribution reflects the presence of moderate noise retrieval. In the practice of RAG systems, a semantic distance distribution that widens to above 1.0 generally indicates the presence of a document with partial topical relevance, which in the retrieval literature is often associated with moderate noise at the generative context stage [12]. In a Retrieval-Augmented Generation architecture, noise at the retrieval stage can cause some of the less relevant context to enter the generation window. This condition has the potential to lower the level of groundedness, because generative models not only rely on documents that are truly relevant, but are also exposed to fragments of information that are peripheral. Factors contributing to this relatively wide distribution of distance include the use of embedding that is more optimized for English than Indonesian, as well as the absence of a cross-encoder-based re-ranking mechanism to refine retrieval results. Therefore, the results in Fig. 6 directly become an interpretive context for the generative evaluation in Subchapter 5.4, especially regarding groundedness scores that have not reached the optimal level.

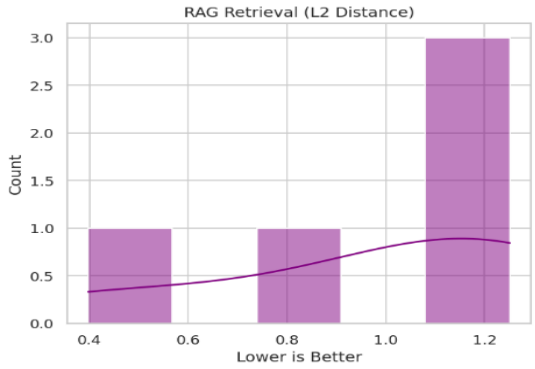


Fig. 6. Distribution of L2 semantic distances between query embeddings and retrieved document embeddings in the RAG pipeline. Lower values indicate higher semantic similarity

5.4. Generative Evaluation Results using LLM-as-a-Judge

Generative evaluation was conducted using the LLM-as-a-Judge approach, where the independent model Meta-Llama-3-70B-Instruct was used to assess the quality of the reports generated by Mistral 7B. This approach aims to reduce self-preference bias and provide a more objective assessment of the quality of the final generation. Fig. 7 summarizes the generative evaluation results using the LLM-as-a-Judge framework. The relatively high coherence and neutrality scores indicate that the system is able to produce structured and balanced analytical reports, while the groundedness score reflects the model's ability to utilize retrieved context effectively. The coherence score shows that the resulting report has a logical argumentation structure and a systematic delivery flow. The model is able to summarize the results of sentiment analysis in an organized and easy-to-understand format.

In terms of neutrality, a score of 4.0 indicates that the system manages to maintain an academic tone and reduce emotional or provocative language, even though the raw data source comes from socio-political discourses that tend to contain strong opinions. This shows that generative models are able to abstract information without reinforcing the bias or polarization contained in the source data. Interestingly, the groundedness score reached 5.0/5.0 according to the LLM-as-a-Judge evaluation, which indicates that the resulting reports consistently refer to the retrieved documents and show no strong indication of claims that are not supported by the available context. These findings suggest that although in Subchapter 5.3 there is an indication of a relatively wide semantic distance distribution, generative models are still able to selectively utilize the most relevant contexts and ignore less contextually related fragments. In other words, these results suggest that the generator is capable of utilising relevant context from the retrieval stage. However, since the groundedness evaluation is also performed by an LLM-based evaluator, the possibility remains that subtle hallucinations or unsupported claims may not be fully detected, particularly in complex socio-political contexts.

Overall, the results of this generative evaluation confirm that the quality of LLM output is not only determined by the model's ability to string text, but is also strongly influenced by the quality of the context provided by the RAG pipeline. Thus, the increase in groundedness at the generative stage is highly dependent on the optimization of retrieval mechanisms, not solely on increasing the capacity of the language models used.

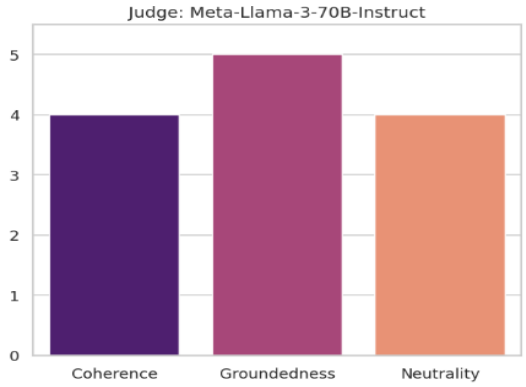


Fig. 7. Generative evaluation scores for coherence, groundedness, and neutrality assessed using an independent LLM-as-a-Judge (Llama-3-70B-Instruct)

5.5. Discussion: System-Level Validation Perspective

The results of the evaluation in Subchapters 5.1 to 5.4 show that system performance cannot be understood through accuracy metrics or generative scores alone. In contrast, performance dynamics are influenced by the interaction between model architecture, the system ground-truth labels sources, and retrieval–generation pipeline configurations. Recent research also explores the integration of LLMs with structured knowledge and reasoning frameworks [38]–[40]. Therefore, this discussion examines the implications of research results from the perspective of system-level validation.

- **Neutrality Bias as Architectural Effect :** The findings at the calibration stage show the tendency of the XLM-RoBERTa-based zero-shot model to classify implicit political criticism as neutral rather than negative. This phenomenon is related to the characteristics of Natural Language Inference-based (NLI)-based models, which map logical relationships between texts and tend to choose neutral labels when contradictory relationships are not explicit. In addition, RLHF-based alignment processes in modern models are known to affect sensitivity to political or controversial content [7], [8]. Thus, neutrality bias in this context is not just a prediction error, but an architectural and alignment consequence. Therefore, LLM-based political sentiment analysis systems must be evaluated while considering these structural biases.
- **Inter-LLM Calibration Divergence:** Comparisons between AI Friends and SeaLLM show an imperfect level of agreement, even though both act as pseudo-ground-truth label providers. This suggests that LLM-based sentiment analysis system datasets have an inherent level of uncertainty. The recent literature also highlights that LLM models can produce different interpretations of the same data, particularly in evaluative tasks and opinion classification [3], [4]. This divergence reinforces the importance of additional calibration based on independent datasets. Without such an approach, evaluation risks only measuring similarity between models, not conformity to human interpretation.
- **Performance Ceiling Implications:** The level of agreement between LLMs implicitly forms a realistic performance ceiling when evaluating LLM-based sentiment analysis system labels. If the two reference models do not reach full agreement, then the expectation that the third model will exceed that limit becomes methods irrational. This concept is in line with findings in LLM evaluation studies and generative-based sentiment systems, which show that the quality of labels and reference sources strongly influence the limits of achievable performance [4]. Therefore, the

interpretation of metrics such as accuracy and Cohen's Kappa must be placed in the context of label quality and stability.

- **Balanced Sampling Impact:** The use of balanced sampling in this study aims to reduce distortions due to class imbalances, thereby providing a more proportional performance picture to the model's ability to distinguish between positive and negative sentiments. Previous studies have shown that unbalanced class distribution can significantly affect the interpretation of evaluation metrics in sentiment analysis [2]. However, balanced sampling also results in more controlled evaluation conditions than the natural distribution of social media data. Therefore, the results of the evaluation reflect strong internal validity, although generalizations to real-world distributions need to be further tested.
- **Deployment Implication:** From an implementation perspective, the proposed RAG architecture demonstrates the ability to produce coherent, neutral, and well-grounded reports. Previous studies on RAG pipelines for LLM-based sentiment analysis system applications have also shown that the integration of retrieval and generation can improve output reliability if the retrieval mechanism is well controlled [5]. However, the existence of neutrality bias and the limitations of cross-lingual embedding suggest that this system is more appropriately positioned as a decision-support system, rather than as an automated decision-making system. The system-level validation carried out in this study provides a more comprehensive evaluation approach than measuring classification accuracy alone.

6. Conclusion

This study aims to validate a sentiment analysis system based on the LLMs that integrates multi-source data acquisition, zero-shot sentiment classification, Retrieval-Augmented Generation (RAG), and generative reasoning to support socio-political public opinion analysis. This integrative approach reflects a growing trend of positioning LLMs as reasoning components in large-scale data-driven analytics systems [3], [4], [36]. In contrast to the approach that focuses on optimizing the accuracy of the model alone, this study places the LLM as a reasoning and evaluation component in a controlled and measurable end-to-end pipeline. The results of the evaluation show that the proposed system is able to operate consistently at all stages of the pipeline. The XLM-RoBERTa-based zero-shot sentiment classification module shows adequate baseline performance on the Golden Calibration Dataset, although it still faces challenges in distinguishing neutral and negative sentiment in implicit political contexts. The data ingestion module proved to be stable and responsive at various levels of topic complexity, while the evaluation of retrieval quality in the RAG pipeline showed that the system was generally capable of providing relevant context, although there were still variations in semantic proximity that could affect the groundedness level of the generative output. Generative evaluation using the LLM-as-a-Judge approach confirms that the resulting analytical reports have a good level of coherence and neutrality, but are still limited in terms of full attachment to the context of the retrieval results. However, this study also reveals some limitations. First, the use of zero-shot sentiment classifications that have been harmonized through Reinforcement Learning from Human Feedback (RLHF) leads to a tendency to neutralize political criticism that is indirect. Second, the retrieval mechanism based on semantic similarity still allows the entry of semantically adjacent but less relevant content, which has a direct impact on the quality of the LLM's reasoning. In addition, Although the calibration dataset was expanded to 2,000 balanced samples, the absence of human-annotated ground truth remains a limitation for fully generalizable evaluation. As a further direction of research, several developments can be made to improve the quality and reliability of the system. Follow-up research may explore more rigorous retrieval screening strategies or context-based reranking to improve the groundedness of generative outputs. Recent advances also show that large language models are increasingly capable of supporting complex reasoning and social event interpretation tasks, which opens opportunities for extending sentiment analysis systems toward more advanced contextual reasoning frameworks [41]. In addition, the integration of human-in-the-loop or partial annotation on calibration datasets has the potential to reduce neutralization bias in political sentiment analysis. Overall, this study shows that the evaluation of

an LLM-based sentiment analysis system needs to be done not only at the model level, but also at the level of the pipeline architecture as a whole. The proposed system-level validation approach is expected to serve as a more comprehensive evaluation framework for the development of LLM-based sentiment analysis systems in the future, especially in the context of complex and dynamic public discourse.

Acknowledgment

The authors thank colleagues and reviewers for their valuable feedback and technical insights during the development and evaluation of this research.

Declarations

Author contribution. Conceptualization & design: Sidharta Sidharta, Hady Pranoto. Experiments: Sidharta Sidharta, Hady Pranoto, Frederik Masri Gasa, Nur Kholis. Writing—review & editing: Ardvin Kester S. Ong. All authors approved the manuscript.
Funding statement. This research was funded by Universitas Bina Nusantara, Indonesia, under the Internal Research Grant Scheme [Contract No.175/VRRTT/VIII/2025].
Conflict of interest. The authors declare no conflict of interest.
Additional information. No additional information is available for this paper.

Data and Software Availability Statements

The dataset supporting the findings of this study is publicly available and can be accessed via Zenodo at the following DOI link: https://doi.org/10.5281/zenodo.17659941. This repository contains all raw and processed data used for sentiment analysis, RAG evaluation, and reasoning validation. Researchers are encouraged to reuse and cite this dataset in accordance with the terms and conditions provided on the Zenodo platform.

References

[1] J. Shen, "Sentiment Analysis of Social Media Texts for Aquatic Product Listed Companies Based on the BERT-CNN-BiLSTM-Att Hybrid Model," *Adv. Educ. Humanit. Soc. Sci. Res.*, vol. 13, no. 1, p. 182, Mar. 2025, doi: 10.56028/aehssr.13.1.182.2025.

[2] D. Guna Mandhasiya, H. Murfi, and A. Bustamam, "The hybrid of BERT and deep learning models for Indonesian sentiment analysis," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 33, no. 1, p. 591, Jan. 2024, doi: 10.11591/ijeecs.v33.i1.pp591-602.

[3] K. Ahmed, M. I. Nadeem, G. Wang, F. Zuo, and Z. Han, "LLM-infused multi-module transformer for emotion-aware sentiment analysis in few-shot scenarios," *Inf. Fusion*, vol. 126, p. 103668, Feb. 2026, doi: 10.1016/j.inffus.2025.103668.

[4] C. Liu, A. Arulappan, R. Naha, A. Mahanti, J. Kamruzzaman, and I.-H. Ra, "Large Language Models and Sentiment Analysis in Financial Markets: A Review, Datasets, and Case Study," *IEEE Access*, vol. 12, pp. 134041–134061, 2024, doi: 10.1109/ACCESS.2024.3445413.

[5] M. Khasanova Zafar kizi and Y. Suh, "Design and Performance Evaluation of LLM-Based RAG Pipelines for Chatbot Services in International Student Admissions," *Electronics*, vol. 14, no. 15, p. 3095, Aug. 2025, doi: 10.3390/electronics14153095.

[6] J. P. Nayinzira and M. Adda, "SentimentCareBot: Retrieval-Augmented Generation Chatbot for Mental Health Support with Sentiment Analysis," *Procedia Comput. Sci.*, vol. 251, pp. 334–341, 2024, doi: 10.1016/j.procs.2024.11.118.

[7] J. Zhou, J. Ji, J. Dai, and Y. Yang, "Sequence to Sequence Reward Modeling: Improving RLHF by Language Feedback," *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 26, pp. 27765–27773, Apr. 2025, doi: 10.1609/aaai.v39i26.34992.

[8] S. B. Shah *et al.*, "Navigating the Web of Disinformation and Misinformation: Large Language Models as Double-Edged Swords," *IEEE Access*, vol. 13, pp. 169262–169282, 2025, doi: 10.1109/ACCESS.2024.3406644.

[9] A. Britzolakis, H. Kondylakis, and N. Papadakis, "AthPPA: A Data Visualization Tool for Identifying Political Popularity over Twitter," *Information*, vol. 12, no. 8, p. 312, Jul. 2021, doi: 10.3390/info12080312.

[10] Y. M. Tun and M. Khaing, "A large-scale sentiment analysis using political tweets," *Int. J. Electr. Comput. Eng.*, vol. 13, no. 6, p. 6913, Dec. 2023, doi: 10.11591/ijece.v13i6.pp6913-6925.

[11] D. Valle-Cruz, R. Sandoval-Almazán, A. López-Chau, and J. I. Criado, "Unveiling political polarization on Twitter: Machine learning and sentiment analysis in presidential elections," *JeDEM - eJournal eDemocracy Open Gov.*, vol. 16, no. 1, pp. 186–212, Sep. 2024, doi: 10.29379/jedem.v16i1.846.

[12] D. Tiwari, B. S. Bhati, B. Nagpal, N. Alturki, and L. Bayisenge, "Attention-augmented hybrid CNN-LSTM model for social media sentiment analysis in cryptocurrency investment decision-making," *Sci. Rep.*, vol. 15, no. 1, p. 33201, Sep. 2025, doi: 10.1038/s41598-025-18245-x.

[13] Jinjing Zeng, Kailing Luo, Yu Lu, and Mingfen Wang, "An Evaluation Framework for Online Courses Based on Sentiment Analysis Using Machine Learning," *Int. J. Emerg. Technol. Learn.*, vol. 18, no. 18, pp. 4–22, Sep. 2023, doi: 10.3991/ijet.v18i18.42521.

[14] Y. Zhang, H. Xu, D. Zhang, and R. Xu, "A Hybrid Approach to Dimensional Aspect-Based Sentiment Analysis Using BERT and Large Language Models," *Electronics*, vol. 13, no. 18, p. 3724, Sep. 2024, doi: 10.3390/electronics13183724.

[15] A. Assiri, A. Gumaei, F. Mehmood, T. Abbas, and S. Ullah, "DeBERTa-GRU: Sentiment Analysis for Large Language Model," *Comput. Mater. Contin.*, vol. 79, no. 3, pp. 4219–4236, 2024, doi: 10.32604/cmc.2024.050781.

[16] R. Cantini, A. Orsino, M. Ruggiero, and D. Talia, "Benchmarking adversarial robustness to bias elicitation in large language models: scalable automated assessment with LLM-as-a-judge," *Mach. Learn.*, vol. 114, no. 11, p. 249, Nov. 2025, doi: 10.1007/s10994-025-06862-6.

[17] N. Christakis and D. Drikakis, "Evaluating Large Language Models in Code Generation: INFINITE Methodology for Defining the Inference Index," *Appl. Sci.*, vol. 15, no. 7, p. 3784, Mar. 2025, doi: 10.3390/app15073784.

[18] K. Combs, T. Bihl, S. Howlett, and Y. Adams, "Zero-shot Comparison of Large Language Models (LLMs) Reasoning Abilities on Long-text Analogies," in *Hawai International Conference*, 2025, p. 1610, doi: 10.24251/HICSS.2025.194.

[19] R. Chandra, A. Tiwari, N. Jain, and S. Badhe, "Large Language Models for Metaphor Detection: Bhagavad Gita and Sermon on the Mount," *IEEE Access*, vol. 12, pp. 84452–84469, 2024, doi: 10.1109/ACCESS.2024.3411060.

[20] F. Geissler, K. Roscher, and M. Trapp, "Concept-Guided LLM Agents for Human-AI Safety Codesign," *Proc. AAAI Symp. Ser.*, vol. 3, no. 1, pp. 100–104, May 2024, doi: 10.1609/aaaiss.v3i1.31188.

[21] J. Ren *et al.*, "PHM–GPT: A Large Language Model for Prognostics and Health Management," *Engineering*, vol. 63, pp. 156–172, Aug. 2026, doi: 10.1016/j.eng.2025.11.001.

[22] J. Prakash, R. Swathiramya, G. Balambigai, R. Menaha, and J.S. Abhirami, "AI-Driven Real-Time Feedback System for Enhanced Student Support: Leveraging Sentiment Analysis and Machine Learning Algorithms," *Int. J. Comput. Exp. Sci. Eng.*, vol. 10, no. 4, Dec. 2024, doi: 10.22399/ijcesen.780.

[23] M. Giuffrè *et al.*, "Expert of Experts Verification and Alignment (EVAL) Framework for Large Language Models Safety in Gastroenterology," *npj Digit. Med.*, vol. 8, no. 1, p. 242, May 2025, doi: 10.1038/s41746-025-01589-z.

[24] R. Massenon, I. Gambo, J. A. Khan, C. Agbonkhese, and A. Alwadain, ""My AI is Lying to Me": User-reported LLM hallucinations in AI mobile apps reviews," *Sci. Rep.*, vol. 15, no. 1, p. 30397, Aug. 2025, doi: 10.1038/s41598-025-15416-8.

[25] I. C. Rafidhul Haque, "Implementation of Retrieval-Augmented Generation (RAG) and Large Language Models (LLM) for a Document and Tabular-Based Chatbot System," *J. Electron. Technol. Explor.*, vol. 3, no. 1, pp. 19–23, Jul. 2025, doi: 10.52465/joetex.v3i1.588.

[26] B. Han, T. Susnjak, and A. Mathrani, "Automating Systematic Literature Reviews with Retrieval-Augmented Generation: A Comprehensive Overview," *Appl. Sci.*, vol. 14, no. 19, p. 9103, Oct. 2024, doi: 10.3390/app14199103.

[27] S.-M. Kuo, S.-K. Tai, H.-Y. Lin, and R.-C. Chen, "Automated Clinical Trial Data Analysis and Report Generation by Integrating Retrieval-Augmented Generation (RAG) and Large Language Model (LLM) Technologies," *AI*, vol. 6, no. 8, p. 188, Aug. 2025, doi: 10.3390/ai6080188.

[28] H. Lijaya, P. Ho, and H. Santoso, "Comparative Analysis of RAG-Based Open-Source LLMs for Indonesian Banking Customer Service Optimization Using Simulated Data," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 14, no. 3, pp. 330–341, Jul. 2025, doi: 10.32736/sisfokom.v14i3.2383.

[29] A. Dubey, A. Ambasta, J. Soni, P. Doshi, M. R. Rane, and P. Kanani, "A Hybrid Semantic–Rule-Based NLP Framework Integrating DFCI and MSKCC Approaches for Clinical Trial Matching Using UMLS and FAISS," *Ingénierie des systèmes d Inf.*, vol. 30, no. 9, pp. 2285–2295, Sep. 2025, doi: 10.18280/isi.300906.

[30] S. A. Haider *et al.*, "The Development and Evaluation of a Retrieval-Augmented Generation Large Language Model Virtual Assistant for Postoperative Instructions," *Bioengineering*, vol. 12, no. 11, p. 1219, Nov. 2025, doi: 10.3390/bioengineering12111219.

[31] L. Magee, V. Arora, and L. Munn, "Structured like a language model: Analysing AI as an automated subject," *Big Data Soc.*, vol. 10, no. 2, pp. 1–15, Jul. 2023, doi: 10.1177/20539517231210273.

[32] S. Sagi, "Optimizing LLM Inference: Metrics that Matter for Real Time Applications," *J. Artif. Intell. Cloud Comput.*, vol. 4, no. 1, p. 1, Jan. 2025, doi: 10.47363/JAICC/2025(4)446.

[33] Y. Guan *et al.*, "Intelligent Agents with LLM-based Process Automation," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Aug. 2024, pp. 5018–5027, doi: 10.1145/3637528.3671646.

[34] M. Azhari and P. Parjito, "Sentiment Analysis of Public Opinion on the Free Lunch Program Using Random Forest in Media," *Build. Informatics, Technol. Sci.*, vol. 6, no. 3, pp. 1932–1942, Dec. 2024, doi: 10.47065/bits.v6i3.6423.

[35] E. Wallace, M. Gardner, and S. Singh, "Interpreting Predictions of NLP Models," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, 2020, pp. 20–23, doi: 10.18653/v1/2020.emnlp-tutorials.3.

[36] A. Timashov, "Evaluating Large Language Models: A Comprehensive Guide," *Int. J. Latest Eng. Manag. Res.*, vol. 9, no. 10, pp. 5–16, Nov. 2024, doi: 10.56581/IJLEMR.9.10.05-16.

[37] X. Feng, N. Angkawisittpan, and J. Li, "Design and Performance Analysis of a Web Crawler-Based System for Emotional Feature Extraction from Social Media Data," *Teh. Vjesn. - Tech. Gaz.*, vol. 32, no. 5, p. 397, Oct. 2025, doi: 10.17559/TV-20240829001951.

[38] R. Tekumalla and J. M. Banda, "Towards automated phenotype definition extraction using large language models," *Genomics Inform.*, vol. 22, no. 1, p. 21, Oct. 2024, doi: 10.1186/s44342-024-00023-2.

[39] Y. Guo, G. Shang, and C. Clavel, "Benchmarking Linguistic Diversity of Large Language Models," *Trans. Assoc. Comput. Linguist.*, vol. 13, pp. 1507–1526, Nov. 2025, doi: 10.1162/TACL.a.47.

[40] G. Ma, Y. Ma, X. Wu, Z. Su, M. Zhou, and S. Hu, "Task-level Distributionally Robust Optimization for Large Language Model-based Dense Retrieval," *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 23, pp. 24759–24767, Apr. 2025, doi: 10.1609/aaai.v39i23.34657.

[41] A. Narasimhan Sampath, A. Thakur, and S. Krishnan, "Agentic Reasoning for Social Event Extrapolation: Integrating Knowledge Graphs and Language Models," *IEEE Access*, vol. 13, pp. 172096–172110, 2025, doi: 10.1109/ACCESS.2025.3612015.