Enhancing Topic Interpretability with ChatGPT: A Dual Evaluation of Keyword and Context-Based Labeling

Mashael M. Alsulami[1], Maha A. Thafar[2]
Department of Information Technology-College of Computers and Information Technology,
Taif University, Taif, Saudi Arabia[1]
Department of Computer Science-College of Computers and Information Technology,
Taif University, Taif, Saudi Arabia[2]

***Abstract*—Accurate topic labeling is essential for structuring and interpreting large-scale textual data across various domains. Traditional topic modeling methods, such as Latent Dirichlet Allocation (LDA), effectively extract topic-related keywords but lack the capability to generate semantically meaningful and contextually appropriate labels. This study investigates the integration of a large language model (LLM), specifically ChatGPT, as an automatic topic label generator. A dual evaluation framework was employed, combining keyword-based and context-based assessments. In the keyword-based evaluation, domain experts reviewed ChatGPT-generated labels for semantic relevance using LDA-derived keywords. In the context-based evaluation, experts rated the alignment between ChatGPT-assigned topic labels and actual content from representative sample posts. The findings demonstrate strong agreement between AI-generated labels and human judgments in both dimensions, with high inter-rater reliability and consistent contextual relevance for several topics. These results underscore the potential of LLMs to enhance both the coherence and interpretability of topic modeling outputs. The study highlights the value of incorporating context in evaluating automated topic labeling and affirms ChatGPT's viability as a scalable, efficient alternative to manual topic interpretation in research, business intelligence, and content management systems.**

***Keywords*—Automatic label generation; topic modeling; Large Language Models (LLMs); topic labeling; semantic relevance**

I. Introduction

The exponential growth of digital content has necessitated advanced mechanisms for topic modeling and document classification, particularly in domains requiring structured knowledge extraction. Traditional approaches, such as Latent Dirichlet Allocation (LDA) [1] and Non-Negative Matrix Factorization (NMF) [2], have been widely employed in topic modeling. However, these statistical techniques often struggle with contextual understanding and semantic relevance due to their reliance on word co-occurrence patterns rather than intrinsic meaning representation. In contrast, recent advances in natural language processing (NLP) and deep learning have led to the proliferation of transformer-based models, which demonstrate a remarkable ability to capture nuanced linguistic structures and contextual dependencies [3].

Large language models (LLMs), including generative systems like OpenAI's ChatGPT, have garnered increasing attention for their potential applications in text classification, summarization, and topic labeling. Beyond topic modeling, LLMs have been integrated into various NLP tasks, such as named entity recognition (NER) [4], sentiment analysis [5], machine translation [6], and information retrieval [7], showcasing their ability to generalize across multiple domains. These models leverage deep contextual embeddings to understand syntactic and semantic nuances, enabling them to outperform traditional methods in tasks that require complex linguistic reasoning. While previous studies have explored the efficacy of deep learning models in topic modeling [8], limited research has examined the robustness and consistency of LLM-generated topic labels against traditional methodologies. This gap is particularly significant in high-stakes domains such as academia and industry, where the accuracy of topic classification directly impacts knowledge organization and retrieval [9].

In this study, we investigate two key research questions: (1) To what extent can ChatGPT reliably generate accurate and consistent topic labels when compared to domain experts? (2) What is the potential of ChatGPT in assisting domain experts for more efficient and accurate topic labeling in large-scale text datasets? To address these questions, we evaluate ChatGPT's ability to generate semantically meaningful topic labels by incorporating multiple similarity measures, including Jaccard Similarity and Cosine Similarity, combined with a Majority Voting mechanism to systematically assess labeling accuracy. The results demonstrate that the combination of these measures provides strong evidence of ChatGPT's effectiveness in generating accurate and semantically relevant topic labels.

The implications of this research extend beyond topic modeling, contributing to the broader discourse on the interpretability and reliability of generative AI models in structured classification tasks. This study sheds light on the evolving role of LLMs in automated knowledge management and retrieval, and how they may assist domain experts in more efficient knowledge categorization.

II. Related Work

The rapid advancement of large language models (LLMs), such as OpenAI's ChatGPT, has sparked significant interest across various fields, particularly in the areas of natural language processing (NLP) and automated text analysis [10][11][12]. Traditional techniques in topic modeling, such as Latent Dirichlet Allocation (LDA) and Non-Negative Matrix

Factorization (NMF), have been instrumental in identifying hidden semantic structures in text corpora. However, these methods often face challenges in capturing deeper contextual and semantic nuances in texts. Recent studies have explored the potential of transformer-based models, particularly ChatGPT, to bridge these gaps, offering new possibilities for interpreting and labeling topics. Scheepers et al. [13] conducted an initial study on interpreting topic models with ChatGPT, demonstrating that the model could assist in generating human-readable summaries for topics identified by traditional topic modeling techniques. Their findings revealed that ChatGPT's ability to describe topics accurately could be leveraged to enhance the interpretability of topic modeling outputs, particularly when prompted correctly.

Building on the potential of ChatGPT in content analysis, Guo et al. [14] investigated how ChatGPT compares to human experts in terms of response quality across a range of domains, including financial, medical, legal, and psychological areas. Their analysis, which included a large dataset (the Human ChatGPT Comparison Corpus, HC3), found that while ChatGPT could produce valuable insights, there were notable gaps when compared to human expertise. This highlighted the need for further evaluation of the model's performance and its alignment with expert judgment. Similarly, Wang et al. [15] explored ChatGPT's viability as an evaluator for natural language generation (NLG) models, comparing its assessments to human judgments. They concluded that ChatGPT achieved competitive correlations with human evaluators, particularly excelling in tasks that involve summarization, underscoring its potential as a reliable evaluator for NLG systems.

Another study by Alyafeai et al. [16] assessed the performance of ChatGPT-based models on various Arabic NLP tasks, such as sentiment analysis, machine translation, and diacritization. Their findings showed that while ChatGPT's performance in some tasks, such as summarization, exceeded that of state-of-the-art approaches, the model still faced challenges with certain language-specific tasks. This study underscored the importance of contextual understanding and the limitations of generalized models like ChatGPT in highly specialized linguistic environments.

In the context of software engineering, Ronanki et al. [17] examined the use of ChatGPT for evaluating the quality of user stories in agile development. Their work demonstrated that ChatGPT could be an effective tool for evaluating user stories, aligning closely with human evaluations in terms of consistency and accuracy. This study also emphasized the need for a reliable "best of three" strategy to improve the stability of ChatGPT's evaluations, ensuring that its output could be trusted for practical applications.

Despite the promise shown by these studies, ChatGPT is not without its limitations. Wu [5] evaluated ChatGPT's problem-solving abilities across various NLP tasks, revealing that while the model performed well in areas like question answering and arithmetic, it struggled with tasks that required commonsense reasoning or complex understanding. This highlights a significant gap in ChatGPT's capabilities, suggesting the need for further improvements, especially in handling more sophisticated linguistic challenges. Additionally, Koubaa et al. [18] provided a critical review of ChatGPT, emphasizing its technical innovations while also pointing out areas where the model needs further refinement, particularly in handling tasks that require deep reasoning and context-specific expertise.

Overall, while ChatGPT has demonstrated considerable potential as an evaluator across different NLP tasks, its performance remains inconsistent across domains, necessitating further research to refine its abilities and improve its reliability. These studies collectively suggest that while ChatGPT can assist domain experts in interpreting topics and evaluating content, it must be adapted and fine-tuned for specific applications to reach its full potential. As more research is conducted, it is likely that ChatGPT and similar LLMs will continue to evolve, offering new possibilities for automating complex tasks traditionally performed by human experts.

III. Methodology

This section presents the framework developed in this study, shown in Fig. 1, which integrates keyword-based and context-based evaluations of topic labels generated by ChatGPT. The workflow begins with Latent Dirichlet Allocation (LDA) to extract topic keywords from a large corpus, followed by automated labeling using ChatGPT. Both the coherence and interpretability of these labels are assessed through a mixed-methods evaluation involving domain experts.

The study evaluates topic modeling performance across two dimensions: (1) keyword-based labeling, where ChatGPT is prompted to assign topic labels based on LDA-extracted keywords, and experts assess the semantic relevance of those labels; and (2) context-based labeling, where representative sample posts for each topic based on how accurately they reflect the assigned label using a Likert scale of contextual relevance.

This dual assessment framework enables a comprehensive evaluation of ChatGPT's effectiveness in supporting topic interpretability. A comparative analysis of expert agreement with AI-generated labels highlights the potential of large language models to enhance topic modeling tasks in research, business intelligence, and content management systems.

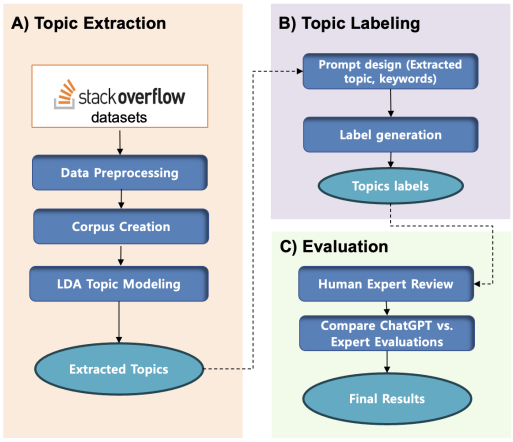

Fig. 1. The workflow of integrating LLMs in topic labeling.

A. Topic Modeling Using LDA

For this research, we utilize the Stack Overflow Posts dataset [19], a diverse source of user-generated questions,

answers, and discussions on software development topics. As one of the largest online communities for developers, Stack Overflow provides rich technical content covering programming languages, frameworks, and common challenges.

The dataset consists of 44,000 posts, each containing a Title and Body. The Title offers a concise description of the issue or question, while the Body provides detailed explanations, discussions, or responses. This structure enables an analysis of both high-level problem descriptions and in-depth technical details.

Stack Overflow is an ideal choice for this study due to its specialized terminology, programming jargon, and references to specific frameworks. The complexity of its content makes it well-suited for topic modeling, where generated topics often contain detailed subtopics and technical nuances. Evaluating topic coherence in this dataset is challenging due to the specialized vocabulary and potential ambiguity of terms across contexts. For instance, words like "Java," "Python," "API," and "debugging" may appear in multiple topics with different meanings.

ChatGPT plays a crucial role in assessing topic coherence by interpreting technical jargon and providing consistent evaluations, addressing challenges that human annotators might face in maintaining consistency across complex discussions. Topic modeling is conducted using Latent Dirichlet Allocation (LDA) [20], an unsupervised machine learning technique that identifies latent topics in large text collections. The process involves several steps:

1) Data preprocessing: The text data undergoes preprocessing to remove irrelevant content, such as URLs and common stopwords. Words are tokenized and stemmed to normalize different word forms into their base form.

2) Corpus creation: A dictionary is constructed, mapping each unique word to an ID, and a corpus is generated, representing the text as a bag of words. This corpus serves as input for the LDA model.

3) Model training: The LDA model is trained using the preprocessed corpus. The number of topics is set to five, and the model runs through 15 iterations to ensure convergence [21].

4) Topic extraction: After training, the top topics are extracted from the model, each represented by a set of keywords. These keywords define the core themes of the topics, providing insights into the underlying structure of the text.

Topic coherence measures the semantic similarity within a topic, indicating how related the words within each topic are [22]. In this study, coherence is measured using the c_v metric, which computes the degree to which the top words of each topic frequently appear together in the text corpus. A higher coherence score suggests that the words in the topic are more likely to form a meaningful and interpretable theme. The c_v score is derived by integrating statistical co-occurrence metrics with semantic similarity, ensuring a balance between data-driven insights and interpretability [23]. The coherence score is computed as illustrated in Eq. (1).

$$c\_v = \frac{1}{|T|} \sum_{t \in T} \frac{1}{|t|} \sum_{1 \le i < j \le |t|} \text{NPMI}(w_i, w_j) \tag{1}$$

Where:

noitemsep, topsep=0pt

- T is the set of topics,
- t is an individual topic containing a set of words,
- w_i, w_j are word pairs within a topic,
- $\text{NPMI}(w_i, w_j)$ is the normalized pointwise mutual information between word pairs.

To evaluate the coherence of topics, the CoherenceModel from Gensim is used [24]. This model computes the coherence score based on the tokenized text and dictionary, providing an essential metric for assessing the quality of the topics generated by the LDA model.

B. ChatGPT-Based Topic Labeling

Interpretability refers to the extent to which the identified topics are meaningful and understandable in the context of the original text [25]. In this study, interpretability is assessed through both keyword-based and context-based evaluations involving ChatGPT-generated labels and expert reviews.

To enhance the accuracy and consistency of topic labeling, ChatGPT is prompted using a role-playing approach, where it assumes the role of an NLP expert specializing in topic modeling. This method ensures that ChatGPT evaluates topics with a structured, expert-like perspective rather than relying solely on statistical patterns. Recent studies have shown that role-playing prompts improve AI performance by guiding the model to adopt domain-specific reasoning strategies, leading to more contextually relevant and accurate outputs [26].

In the keyword-based phase, after the LDA model generates topics, ChatGPT systematically analyzes their coherence and interpretability by assessing the relevance of extracted keywords and their alignment with real-world themes. In its expert role, ChatGPT follows a structured decision-making process: it critically examines the provided keywords, determines the most precise and semantically meaningful topic label, and justifies its selection. This approach reduces ambiguity and enhances the semantic clarity of topic assignments. Additionally, ChatGPT suggests refined labels or descriptions for each topic based on its expert-level analysis.

In the context-based phase, the interpretability of the assigned topic labels is further evaluated using real sample posts most strongly associated with each topic. These posts are presented to human experts, who rate how well the content aligns with the topic label generated by ChatGPT using a 5-point Likert scale. This step ensures that the labeling process is not only linguistically appropriate but also contextually accurate in practical usage scenarios.

An example of the instructions given to ChatGPT is shown in Fig. 2.

Expert human reviewers manually assess both the coherence and interpretability of each topic. In the keyword-based

You are an expert in topic modeling evaluation. Your task is to assess the coherence and interpretability of topics generated by a Latent Dirichlet Allocation (LDA) model.

Each topic consists of a list of keywords. Evaluate how well these keywords collectively represent a meaningful and interpretable real-world theme. Consider whether the terms are semantically related and if they align with a coherent subject matter (e.g., a specific programming concept, a software development topic, etc.).

For each topic, provide:

1. **Interpretability Score (0 to 1):** Assign a score indicating how well the keywords represent a clear, real-world topic. A score of 1 indicates perfect interpretability, while a score of 0 indicates no meaningful coherence.
2. **Suggested Topic Label:** Based on the keywords, suggest a concise and descriptive label for the topic.

Here are the topics and their associated keywords:

Topic 0: gt, lt, android, div, p, code, pre, amp, fals, true, td, button, view, import, counti
Topic 1: int, public, string, new, class, void, return, null, privat, main, char, static, std, els, amp
Topic 2: p, code, error, file, pre, use, run, tri, app, work, noreferr, li, import, version, instal
Topic 3: p, data, imag, id, user, tabl, name, select, use, text, php, enter, button, tri, work
Topic 4: p, data, imag, id, user, tabl, name, select, use, text, php, enter, button, tri, work

Evaluate each topic individually and provide your assessment in the following format:

Topic X:

- **Interpretability Score:** (0 to 1)
- **Suggested Label:** (Provide a descriptive label for the topic)

Fig. 2. An example of the prompt used to interact with ChatGPT.

evaluation, they are provided with the top keywords and the corresponding ChatGPT-generated label, and asked to assess the semantic similarity of the terms and their relevance to the label. In the context-based evaluation, they are presented with representative posts and asked to rate how accurately each post reflects its assigned topic label. This dual evaluation allows for a more comprehensive understanding of the effectiveness and reliability of ChatGPT-generated topic labels.

C. Expert Evaluation Setup

This section describes the user study designed to evaluate the reliability of ChatGPT in labeling topics generated from text datasets. The study assesses ChatGPT's labeling performance through a dual approach, comparing its automated topic labels with human expert evaluations in both keyword-based and context-based contexts.

In the keyword-based evaluation, experts assess the semantic appropriateness of ChatGPT-generated labels based on the top keywords extracted from each topic. In the context-based evaluation, experts rate the relevance of representative sample posts to their corresponding ChatGPT-assigned topic labels using a 5-point Likert scale.

The primary objective is to investigate whether ChatGPT can reliably generate topic labels that align with human expert judgment, not only in abstract keyword interpretation but also in practical, real-world content alignment. This comprehensive evaluation explores ChatGPT's potential as a robust and scalable tool for enhancing topic modeling tasks in research, content analysis, and knowledge discovery.

1) Participants: The study involved 39 expert participants recruited through the Prolific platform [27], a widely used online research tool known for its diverse and high-quality participant pool. Prolific enables targeted recruitment based on specific criteria, ensuring that participants meet predefined qualifications. In this study, only individuals with demonstrated computer programming skills were selected, allowing them to accurately distinguish content and assess the compatibility of topics. To uphold a high standard of evaluation, all participants were required to be graduate students pursuing an M.Sc. or Ph.D., ensuring their expertise aligned with the study's objectives.

2) Task and evaluation metrics: The user study evaluated the reliability of ChatGPT's topic labeling by comparing it to expert judgments across both keyword-based and context-based dimensions. Five topics were selected from a dataset processed using LDA, with each topic represented by a set of ten keywords. These topics were pre-labeled by ChatGPT.

In the keyword-based evaluation, experts were presented with the top keywords and ChatGPT-generated labels, and asked to indicate whether they fully agreed with the label, disagreed, or had suggestions for improvement. In the context-based evaluation, two representative posts were selected for each topic, and experts were asked to rate how well the content of each post aligned with the assigned topic label using a 5-point Likert scale.

This dual evaluation approach allowed for a more comprehensive assessment of the semantic appropriateness and contextual accuracy of ChatGPT's labels.

IV. Results

This section outlines the evaluation framework developed in this study, which utilizes ChatGPT as an automated tool to label and assess the quality of topics generated from text datasets. The research explores how ChatGPT can be leveraged to evaluate the topics discovered by Latent Dirichlet Allocation (LDA) models. The study aims to determine whether ChatGPT can effectively label topics based on their keywords and how these evaluations align with human judgment. The analysis of the topics reveals a coherence score (c_v) of 0.584, indicating a moderate level of coherence across the dataset. This score suggests that while the topics are relevant, there is some variability in their consistency. The identified topics are shown in Table I.

TABLE I. Topics Identified by LDA Model

Topic ID	ChatGPT Generated Label	Top 10 Keywords
0	Programming Errors and File Operations	gt, lt, android, div, p, code, pre, amp, fals, true
1	Programming Fundamentals and Data Structures	int, public, string, new, class, void, return, null, privat, main
2	HTML/XML Encoding and Syntax	p, code, error, file, pre, use, run, tri, app, work
3	UI/UX Design and Functionality	p, data, imag, id, user, tabl, name, select, use, text
4	Database Operations and Queries	p, data, imag, id, user, tabl, name, select, use, text

The application of ChatGPT for topic labeling involved analyzing the keyword sets for each topic and categorizing them based on common patterns and relationships. For example, keywords such as "public," "new," "class," "static" and "void" were interpreted by ChatGPT as related to object-oriented programming and Android development, resulting in the label "Programming Fundamentals and Data Structures."

ChatGPT's role as an automated labeling tool provided a first-pass categorization, facilitating the identification of relevant themes and reducing the need for manual intervention.

Despite the moderate coherence score, the results suggest that ChatGPT can be an effective tool for evaluating topic models. To evaluate the quality of the LDA model, ChatGPT was prompted to assess the topics based on their keywords, focusing on how well the terms align with real-world themes. The findings from the comparison of ChatGPT's interpretability scores and the LDA model's c_v coherence scores reveal valuable insights into the coherence and interpretability of the generated topics. Topics with higher interpretability scores, such as Topic 4 (0.85), also exhibit higher c_v coherence scores (0.667), suggesting a strong alignment between human understanding and statistical coherence. These topics are not only easy to interpret but also show that the keywords are semantically related and form a clear, meaningful theme. Conversely, topics with lower interpretability scores, such as Topic 0 (0.6), also tend to have lower coherence scores (0.515), indicating weaker semantic alignment and making them harder to interpret or connect to a coherent real-world theme. Overall, the results suggest that when both the human evaluation and model-based coherence agree, the topics are well-defined and semantically robust. However, discrepancies between the two metrics indicate areas where the model may require refinement to produce more coherent and interpretable topics. Fig. 3 illustrates a comparison between ChatGPT's interpretability scores and the LDA coherence scores.

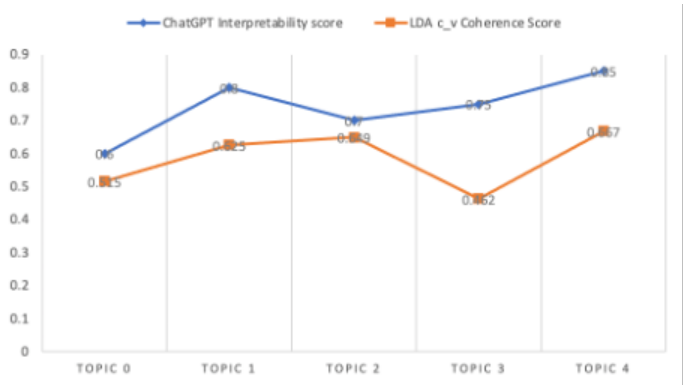

Fig. 3. Comparison of ChatGPT interpretability and LDA coherence scores across topics.

To assess the quality of topic labels generated through ChatGPT, we evaluated their alignment with human expert interpretations. The analysis compared ChatGPT's interpretability scores with the LDA model's c_v coherence scores, offering insights into the reliability of topic labeling. Higher interpretability scores generally corresponded with stronger coherence, indicating well-defined and semantically robust topics.

In addition, Jaccard similarity [28] and cosine similarity [29] were used to measure the levels of agreement between the expert responses and ChatGPT labels. Jaccard Similarity quantifies the overlap between sets of labels assigned by different annotators, while Cosine Similarity assesses the semantic consistency of label assignments using TF-IDF vectorization. TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical measure that reflects the importance of a word in a document relative to a collection of documents [30]. By converting labels into TF-IDF vectors, Cosine Similarity can effectively capture their semantic relationships, even when different terms are used to describe similar topics.

The findings highlight how agreement metrics further validate the interpretability and coherence of generated topics, identifying areas where refinements may enhance alignment between human understanding and statistical modeling. The computed Jaccard Similarity score of 0.5175 indicates a moderate degree of agreement among expert evaluations regarding ChatGPT's labels. Meanwhile, the Cosine Similarity score of 0.6237 reflects a relatively higher degree of semantic similarity, suggesting that the topic labels generated by ChatGPT align well with human opinions in terms of conceptual meaning. To further assess the validity of ChatGPT-generated labels, a Majority Voting approach [31] was applied, where expert responses were binarized (1 = Agree, 0 = Disagree). A threshold of 50% agreement was used to determine whether a label was accepted by the majority of participants. The results of the Majority Voting analysis indicate that ChatGPT's topic labels were largely accepted, as shown in Table II.

TABLE II. TOPICS IDENTIFIED BY LDA MODEL

Topic	Majority Vote	ChatGPT Label	Agreement
Programming Errors and File Operations	1	1	TRUE
Programming Fundamentals and Data Structures	1	1	TRUE
HTML/XML Encoding and Syntax	1	1	TRUE
UI/UX Design and Functionality	1	1	TRUE
Database Operations and Queries	1	1	TRUE

These findings indicate a full agreement between ChatGPT's assigned labels and the majority of expert responses. The combination of Jaccard Similarity, Cosine Similarity, and Majority Voting results provides strong evidence of ChatGPT's effectiveness in generating accurate and semantically relevant topic labels. These findings highlight the potential for leveraging ChatGPT in automated topic classification tasks based on LDA-generated keyword distributions.

The results of this evaluation revealed distinct differences in label performance across the posts as shown in Figure 4. For instance, Posts 3, 6, 8, and 10 achieved high mean scores above 4.0 with low standard deviations, suggesting strong agreement among reviewers regarding the contextual relevance of the ChatGPT-assigned labels. These findings affirm that the labels for these topics were not only semantically valid at the keyword level but also contextually robust when applied to actual post content.

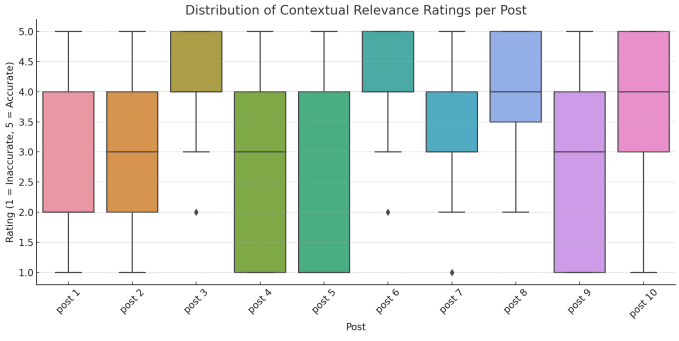

Fig. 4. Distribution of contextual relevance ratings per post.

To measure inter-rater reliability, Kendall's W was computed and yielded a value of 0.99, indicating excellent agreement among expert reviewers. Furthermore, the average pairwise Spearman correlation was 0.23, reflecting consistent ranking tendencies across raters despite some variance. These findings reinforce the need for context-based assessments in evaluating topic modeling outputs. While keyword-based labeling offers a foundational view of topic coherence, context-based ratings provide practical validation of the interpretability and usability of topic labels in real-world scenarios. Together, these perspectives offer a more comprehensive understanding of how well large language models such as ChatGPT perform in automated topic labeling.

V. Discussion

The discussion now turns to the ethical implications of AI-based labeling, particularly in the context of ChatGPT's role in automated topic classification. While the results indicate that ChatGPT-generated labels align well with human evaluations, certain ethical concerns related to bias, transparency, and accountability must be addressed to ensure responsible deployment.

One primary concern is bias in topic representation. The findings suggest that ChatGPT accurately labeled structured programming-related topics but exhibited inconsistencies in areas with lower coherence scores. This variability raises concerns about the model's potential to reinforce systematic biases in topic classification, leading to overrepresentation of dominant themes or misclassification of ambiguous content. Previous research has demonstrated that AI models trained on large-scale datasets can inadvertently inherit biases present in the data, impacting the fairness of generated outputs [32].

Another key issue is transparency in AI-generated classifications. While similarity metrics, such as Jaccard and Cosine Similarity, provide insights into how closely ChatGPT's labels align with human interpretations, the lack of explainability in AI decision-making remains a challenge. Unlike human experts who can articulate their reasoning, ChatGPT's topic assignments rely on statistical correlations rather than explicit contextual understanding. This limitation can make it difficult to assess the reliability of AI-generated labels and detect systematic misclassifications [33].

Accountability and human oversight are also critical considerations. The majority voting analysis confirms that ChatGPT's labels were widely accepted, but overreliance on AI-generated labels without human validation could lead to misclassifications in cases where expert knowledge is necessary. Ethical AI deployment should incorporate a human-in-the-loop (HITL) framework, where AI serves as a decision-support tool rather than an autonomous classifier [34]. This approach ensures that AI-generated labels remain subject to expert validation, reducing the risk of incorrect topic assignments.

To mitigate these ethical concerns, AI-based labeling should integrate bias detection techniques, fairness-aware learning algorithms, explainability mechanisms, and expert validation frameworks to ensure that AI-generated labels are transparent, unbiased, and aligned with human judgment.

VI. Conclusion

This study demonstrates the potential of ChatGPT as an effective tool for automated topic labeling and evaluation in topic modeling tasks. By comparing the labels generated by ChatGPT with those of domain experts, we found that ChatGPT can generate semantically meaningful and coherent topic labels, offering valuable insights for large-scale text datasets. The integration of multiple similarity measures, including Jaccard Similarity, Cosine Similarity, and a Majority Voting mechanism, provided a comprehensive framework for assessing labeling accuracy. The results indicate that ChatGPT's labels align well with human judgment, with moderate to strong agreement levels, particularly in terms of semantic consistency.

Despite some variability in the coherence scores, the analysis suggests that ChatGPT can reliably categorize topics based on keyword sets, facilitating the identification of relevant themes and reducing the need for manual intervention. Topics with higher interpretability and coherence scores further demonstrate the robustness of ChatGPT in providing accurate labels, while discrepancies in less coherent topics highlight areas for potential refinement.

Overall, this research underscores the viability of ChatGPT as a complementary tool for topic labeling, offering a scalable approach to streamline the evaluation and categorization of topics in large text corpora. Further work may focus on refining the model's accuracy, especially in cases of lower coherence, to further enhance the precision and consistency of automated topic labeling systems.

Acknowledgment

The authors acknowledge the Deanship of Graduate Studies and Scientific Research, Taif University, for funding this work.

References

[1] Z. Rosadi and A. Solichin, "Topic modeling tugas akhir mahasiswa menggunakan metode latent dirichlet allocation dengan gibbs sampling," *Jurnal TICOM: Technology of Information and Communication*, vol. 13, no. 1, pp. 38–44, September 2024.

[2] R. Barron, M. E. Eren, D. P. Truong, C. Matuszek, J. Wendelberger, M. F. Dorn, and B. Alexandrov, "Matrix factorization for inferring associations and missing links," *arXiv preprint arXiv:2503.04680*, 2025, manuscript submitted to ACM. [Online]. Available: https://arxiv.org/abs/2503.04680

[3] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, Minneapolis, USA, 2019, pp. 4171–4186.

[4] L. Yao, C. Mao, and Y. Luo, "Graph-based few-shot learning for named entity recognition," in *Proceedings of ACL*, 2021, pp. 3333–3345.

[5] J. Lossio-Ventura, R. Weger, A. Lee *et al.*, "A comparison of chatgpt and fine-tuned open pre-trained transformers (opt) against widely used sentiment analysis tools: Sentiment analysis of covid-19," *JMIR Mental Health*, vol. 11, no. 1, p. e50942, 2024. [Online]. Available: https://mental.jmir.org/2024/1/e50942

[6] J. Tiedemann and Y. Scherrer, "Neural machine translation with extended context," in *Proceedings of NAACL-HLT*, 2017, pp. 1256–1265.

[7] J. Lin and W. Ma, "A few-shot semantic retrieval framework using pretrained language models," in *Proceedings of SIGIR*, 2021, pp. 1983–1986.

[8] Y. Liu *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[9] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proceedings of ICLR*, 2013.

[10] M. Alawida, S. Mejri, A. Mehmood, B. Chikhaoui, and O. I. Abiodun, "A comprehensive study of chatgpt: Advancements, limitations, and ethical considerations in natural language processing and cybersecurity," *Information*, vol. 14, no. 8, p. 462, 2023. [Online]. Available: https://doi.org/10.3390/info14080462

[11] D. D. Torrico, "The potential use of chatgpt as a sensory evaluator of chocolate brownies: A brief case study," *Foods*, vol. 14, no. 3, p. 464, 2025. [Online]. Available: https://doi.org/10.3390/foods14030464

[12] Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu, Z. Wu, L. Zhao, D. Zhu, X. Li, N. Qiang, D. Shen, T. Liu, and B. Ge, "Summary of chatgpt-related research and perspective towards the future of large language models," *Meta-Radiology*, vol. 1, no. 2, p. 100017, 2023. [Online]. Available: https://doi.org/10.1016/j.metrad.2023.100017

[13] F. Scheepers, K. Zervanou, M. Spruit, P. Mosteiro, and U. Kaymak, "Towards interpreting topic models with chatgpt," in *The 20th World Congress of the International Fuzzy Systems Association*, 2023, available: www.tue.nl/taverne.

[14] B. Guo, M. Li, C. Wu, J. Zhao, and Y. Li, "How close is chatgpt to human experts? comparison corpus, evaluation, and detection," Jan. 2023, available: http://arxiv.org/abs/2301.07597.

[15] J. Wang, Y. Zhang, Z. Xu, and Z. Li, "Is chatgpt a good nlg evaluator? a preliminary study," Mar. 2023, available: http://arxiv.org/abs/2303.04048.

[16] Z. Alyafeai, M. S. Alshaibani, B. AlKhamissi, H. Luqman, E. Alareqi, and A. Fadel, "Taqyim: Evaluating arabic nlp tasks using chatgpt models," Jun. 2023, available: http://arxiv.org/abs/2306.16322.

[17] K. Ronanki, B. Cabrero-Daniel, and C. Berger, "Chatgpt as a tool for user story quality evaluation: Trustworthy out of the box?" *Lecture Notes in Business Information Processing*, pp. 173–181, 2024.

[18] A. Koubaa, W. Boulila, L. Ghouti, A. Alzahem, and S. Latif, "Exploring chatgpt capabilities and limitations: A critical review of the nlp game changer," Mar. 27 2023.

[19] Stack Exchange, Inc., "Stack Overflow Posts Dataset," 2023, available at https://archive.org/details/stackexchange.

[20] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, "Latent dirichlet allocation (lda) and topic modeling: Models, applications, a survey," *Multimedia Tools and Applications*, vol. 78, no. 11, pp. 15 169–15 211, 2018.

[21] P. Yang, Y. Yao, and H. Zhou, "Leveraging global and local topic popularities for lda-based document clustering," *IEEE Access*, vol. 8, pp. 24 734–24 745, 2020.

[22] Z. Chen, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, "Discovering coherent topics using general knowledge," in *Proceedings of the ACM International Conference*, 2013, pp. 209–218.

[23] S. Duraivel, Lavanya, and A. Augustine, "Understanding vaccine hesitancy with application of latent dirichlet allocation to reddit corpora," *Infolitika Journal of Data Science*, vol. 2, no. 2, 2024, original Article.

[24] N. S. M. N. Mangsor, S. A. M. Nasir, S. Abdul-Rahman, and Z. Ismail, "Identifying topic modeling technique in evaluating textual datasets," *Lecture Notes on Data Engineering and Communications Technologies*, pp. 507–521, 2023.

[25] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, 2019. [Online]. Available: https://doi.org/10.3390/electronics8080832

[26] L. Reynolds and K. McDonell, "Prompt programming for large language models: Beyond the few-shot paradigm," *arXiv preprint arXiv:2102.07350*, 2021. [Online]. Available: https://arxiv.org/abs/2102.07350

[27] Prolific, "Prolific: Online participant recruitment for research," 2025, accessed: 2025-03-01. [Online]. Available: https://www.prolific.co

[28] S. Bag, S. K. Kumar, and M. K. Tiwari, "An efficient recommendation generation using relevant jaccard similarity," *Information Sciences*, vol. 483, pp. 53–64, 2019.

[29] D. Gunawan, C. A. Sembiring, and M. A. Budiman, "The implementation of cosine similarity to calculate text relevance between two documents," *Journal of Physics: Conference Series*, vol. 978, p. 012120, 2018.

[30] A. Widianto, E. Pebriyanto, Fitriyanti, and Marna, "Document similarity using term frequency-inverse document frequency representation and cosine similarity," *Journal of Dinda: Data Science, Information Technology, and Data Analytics*, vol. 4, no. 2, pp. 149–153, 2024. [Online]. Available: http://journal.ittelkom-pwt.ac.id/index.php/dinda

[31] A. Dogan and D. Birant, "A weighted majority voting ensemble approach for classification," in *2019 6th International Conference on Computer Science and Engineering (UBMK)*, 2019.

[32] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 610–623.

[33] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[34] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint*, vol. arXiv:1702.08608, 2023.