DCPAPERS

Assessing the Effectiveness of LLMs (Large Language Models) for Extracting Topics and Themes in Survey Responses

Ying-Hsang Liu[1,*], Xin Yang[2] and Junzhi Jia[2]

[1]*Professorship of Predictive Analytics, Chemnitz University of Technology, 09126 Chemnitz, Germany*
[2]*School of Information Resources Management, Renmin University of China, Beijing, China*

Abstract

Artificial intelligence (AI) has the potential to automate metadata tasks, such as identifying key topics and analyzing recurring themes in text. Topic extraction focuses on recognizing dominant subjects, whereas theme extraction examines patterns of meaning within the text. This study evaluated DeepSeek R1 (8b), DeepSeek R1 (14b), and Gemma3 (12b) on extracting topics and themes from 50 qualitative survey comments. Using standard information retrieval methods and metrics, we found that Gemma3 (12b) consistently outperformed the DeepSeek models. Topic detection was handled with reasonable effectiveness (both DeepSeek R1 (8b) and Gemma3 (12b) global F1 0.31). However, theme detection was significantly more challenging, particularly for DeepSeek models (global F1s 0.02, 0.08), with Gemma3 (12b) achieving F1 0.26. Significant document-level variability was also observed. Standard information retrieval (IR) metrics can be applied to assess AI performance in metadata tasks, but achieving accuracy comparable to human experts in abstract thematic analysis remains a significant challenge. Developing AI systems that can better capture the subtleties of abstract meaning needs human oversight since these capabilities are critical for supporting complex analytical tasks.

Keywords

Large Language Models (LLMs), Information Retrieval Evaluation Information Extraction, Thematic Analysis, Survey Data

1. Introduction

Artificial intelligence (AI) has been recognized as a transformative force that could significantly impact metadata creation and management within libraries and information services [1]. According to the International Standard ISO/IEC 22989 [2], an AI system is defined as an ”engineered system designed to generate outputs such as content, forecasts, recommendations, or decisions for a given set of human-defined objectives.” This transformation arises from AI's potential to automate time-consuming tasks like metadata extraction, tagging and classification, which may lead to potentially increased efficiency and accuracy in workflows. However, since the implementation of AI technologies may also bring significant challenges, such as

DCMI-2025 International Conference on Dublin Core and Metadata Applications

*Corresponding author.
ying-hsang.liu@hsw.tu-chemnitz.de (Y. Liu); yangxin371@ruc.edu.cn (X. Yang); junzhij@ruc.edu.cn (J. Jia)
https://www.tu-chemnitz.de/hsw/pva/professur/mitarbeiter/ying.html.en (Y. Liu);
http://sirm.ruc.edu.cn/jszy/JunzhiJia/index.htm (J. Jia)
0000-0001-6504-4598 (Y. Liu); 0009-0007-9440-4848 (X. Yang); 0000-0002-3720-3405 (J. Jia)

https://doi.org/10.23106/dcmi.952535736

funding sources and up-skilling professionals, it is important to consider how the AI tools can be integrated into existing workflows.

Subject indexing as applied in library systems involves organizing resources based on the aboutness of text, in order to facilitate information retrieval (IR), is a traditionally labor-intensive manual process performed by subject specialists and assisted by automated systems [3, 4]. This task has become increasingly challenging with the growing number of digitized information resources. Automated methods using Natural Language Processing (NLP) and machine learning (ML) have been developed and implemented for tasks, such as assigning MeSH (Medical Subject Headings) terms for biomedical research literature, with human reviews [5]. One of the major challenges of automated subject indexing is the quality of extracted terms for representing the topics of information content and how they can be mapped into a list of controlled vocabularies. In NLP applications, the extraction of key topics, namely keyword/topic extraction refers to identifying the key subjects of text, which can be conceptualized as metadata enrichment tasks [6]. Extraction of themes involves identifying patterns of meaning within the text by adopting an inductive (bottom-up) approach for qualitative data analysis [7].

More recently, the potential application of Large Language Models (LLMs) with extensive knowledge representation capabilities for topic extraction and subject classification has received significant attention. For example, the initiative of comparing the system performance of automating subject tagging for scientific and technical library records with the records created by experts has gained some traction [8, 9]. Systems that have been developed to use LLMs or LLM-assisted techniques for automated subject indexing have demonstrated varying system performance, as measured by F1 scores, ranging between 0.3 and 0.5 [10] and approximately 0.65 [9]. Despite these promising developments, it is still challenging to achieve expert-level performance, particularly when dealing with specialized domains that require a more nuanced understanding of technical terminologies and complex conceptual relationships within a domain that may not have been captured in LLMs' training data.

To conduct a systematic evaluation of LLM-assisted systems in analyzing qualitative data, we developed a test collection to assess their comprative performance based on information retrieval methodologies [11, 4, 12]. This collection includes 50 records of survey responses, originally in simplified Chinese to assess the effectiveness of LLMs for extracting key topics and themes from text, as potential candidates for mapping these terms to a controlled vocabulary list.

2. Research Question

Within the IR evaluation methodology, we proposed the research question: How effective are Large Language Models (LLMs) at extracting key topics and themes from qualitative survey responses, compared with human analysis of text?

3. Methods

Drawing from the responses from the international Survey on Metadata and AI conducted by the DCMI Education Committee, we created a test collection of 50 qualitative comments originally

https://doi.org/10.23106/dcmi.952535736

in simplified Chinese. We compared the performance of three state-of-the-art LLMs, namely DeepSeek R1 (8b), DeepSeek R1 (14b) and Gemma3 (12b) by using a multi-model inference approach with prompting strategies. We established ground truth data through the analysis of expert human annotators, and conducted inter-coder reliability assessments to enahnce the quality of the test collection.

Our evaluation metrics, which aligns with the evaluation objectives, included precision, recall, and F1 scores that are calculated at both document and corpus levels. It allows us to conduct a systematic comparison between machine and human performance in the analysis of qualitative data. Precision, calculated as the proportion of correctly identified topics among those the system predicted, whereas recall assessed the proportion of actual topics successfully identified by the system. The F1-score provides a single metric to represent overall performance by combining precision and recall measures. Figure 1 provides an overview of the evaluation process.

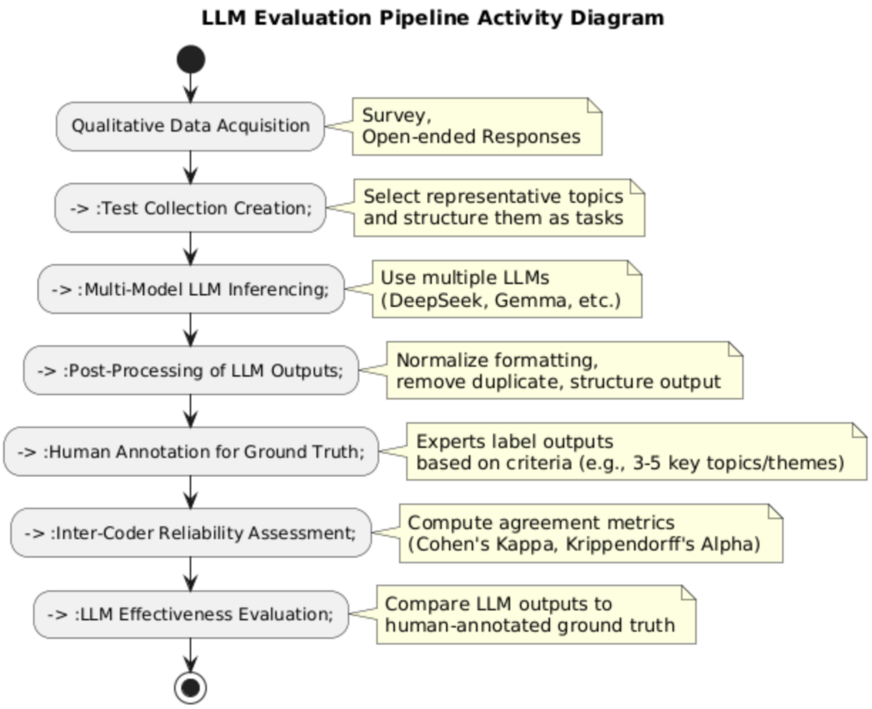
LLM Evaluation Pipeline Activity Diagram
Qualitative Data Acquisition
Survey,
Open-ended Responses
-> :Test Collection Creation;
Select representative topics
and structure them as tasks
-> :Multi-Model LLM Inferencing;
Use multiple LLMs
(DeepSeek, Gemma, etc.)
-> :Post-Processing of LLM Outputs;
Normalize formatting,
remove duplicate, structure output
-> :Human Annotation for Ground Truth;
Experts label outputs
based on criteria (e.g., 3-5 key topics/themes)
-> :Inter-Coder Reliability Assessment;
Compute agreement metrics
(Cohen's Kappa, Krippendorff's Alpha)
-> :LLM Effectiveness Evaluation;
Compare LLM outputs to
human-annotated ground truth

Figure 1: Activity diagram of LLM evaluation pipeline, outlining the overall evaluation process.

3.1. Data Source

To understand the impact of AI on metadata creation and management, the Metadata and AI task group of the DCMI Education Committee developed a survey targeting library and information professionals. The survey was informed by recent research on the application of AI tools to metadata creation and management tasks, potential benefits, challenges and concerns of the anticipated impact of AI and ethical considerations. The survey went through expert reviews and was translated into 15 languages. It was intended to ensure global access, covering the languages across the regions of Asia (Chinese—Simplified & Traditional), Hindi, Japanese, Korean and Tamil), Europe (Finnish, French, German, Italian, Polish, Portuguese and Spanish), and South America (Brazilian Portuguese and Spanish). We adopted a purposive sampling strategy to distribute the survey. We received a total of 752 complete responses between October 2024 and March 2025. A key component of the survey was a qualitative question, ”In your opinion, how can AI tools be utilized to improve the creation and management of metadata? Please provide specific examples or insights based on your experience”. We received a total of 414 responses from this question item. A subset of the responses (n=66) in simplified Chinese, was selected for the present study. Finally, a test collection was created by selecting 50 representative topics after excluding short (e.g., ’NA,’ ’none’), excessively long, and incomplete or poorly formed responses from the survey and from the LLM’s outputs.

3.2. Technical Specifications of LLMs

As shown in Table 1, the DeepSeek R1 (8b) model utilizes the LLaMA architecture with 8 billion parameters and 4096-dimensional embeddings. The DeepSeek R1 (14b) employs the Qwen2 architecture with 14.8 billion parameters and expanded 5120-dimensional embeddings. Meanwhile, Gemma3 (12b) provides 12.2 billion parameters with 3840-dimensional embeddings. All three models feature a large 131,072 token context length and implement Q4_K_M quantization for efficient deployment. While the 8B variant provides a more parameter-efficient solution, the 14B model provides greater representational capacity for complex analytical tasks. Gemma3 serves as a balanced middle option. The differences in architecture and parameter scales may influence model performance in tasks related to topic/theme extraction, such as identifying abstract patterns or complex conceptual relationships in qualitative text. For instance, higher parameter counts (e.g., 14B) could enhance the ability to capture complex thematic structures, while lower-parameter models (e.g., 8B) may prioritize speed or resource efficiency at the cost of representational depth.

Table 1
Comparison of Evaluated Large Language Models

Feature	DeepSeek R1 (8b)	DeepSeek R1 (14b)	Gemma3 (12b)
Architecture	LLaMA	Qwen2	Gemma3
Parameters	8.0B	14.8B	12.2B
Context Length	131,072	131,072	131,072
Embedding Dimension	4,096	5,120	3,840
Quantization	Q4_K_M	Q4_K_M	Q4_K_M

3.3. Qualitative Comments Analysis

The analysis framework implements a multi-model inference approach, using five local language models (DeepSeek R1 variants, llama3.1 (8b), and Gemma3 (12b)) accessed through Ollama’s REST API. The system employs a tiered prompting strategy with three specialized templates

https://doi.org/10.23106/dcmi.952535736

(standard, sentiment-focused, and category-focused) to ensure a comprehensive analysis, falling back to alternative strategies when initial attempts generate insufficient results. The input processing is able to accommodate multilingual text through language-specific tokenization (using jiebaR for Chinese). We also implemented output validation to ensure the adherence to a predefined schema for sentiment classification, category assignment (using a fixed taxonomy of codes, corresponding to the survey question groups), and extraction of relevant topics and themes.

3.4. Post-Processing Phase

The post-processing phase used a Python script to consolidate multi-model outputs through ensemble aggregation techniques. For each comment, we aggregated sentiment classifications, category assignments, topics, and thematic elements into unified collections and removed duplications. This approach was intended to reflect the breadth of model interpretations rather than forcing consensus among the models. It was similar to the pooling method in IR evaluation initiatives [13, 14]. Further analysis of the data focused on assessing the three LLMs (DeepSeek R1 (8b), DeepSeek R1 (14b), and Gemma3 (12b)), partly due to the coverage and performance of these models.

3.5. Ground Truth Data Preparation

To establish a reliable benchmark for evaluating the performance of LLMs, we constructed a ground truth dataset through expert annotations. Native-language experts annotated multilingual comments generated by the LLM-based systems. Topic and theme labels were derived through an aggregation and filtering process of multi-model outputs. Initial filtering removed irrelevant and duplicate entries to ensure language consistency. Subsequently, we excluded weakly related terms were excluded and domain-specific terminologies were examined. Each comment was assigned 3–5 representative topics/themes for the purposes of evaluation. To evaluate inter-coder reliability, we prepared a subset of 11 records, all previously annotated by two experts and extracted from the full dataset.

3.6. Inter-Coder Reliability Assessment

We assessed the inter-rater reliability of two raters with doctoral-level expertise in information organization when selecting topics and themes to each response from LLM-generated outputs. To quantify this reliability score, we used Cohen's kappa statistic, which adjusts for agreement that could occur by chance [15]. The inter-coder reliability analysis for topic identification showed fair agreement (κ = 0.39, SD = 0.37). The median kappa value was 0.33, and scores ranged from -0.33 to 1.00. This indicated some inconsistency in how raters labeled topics in the data. Similarly, theme extraction demonstrated fair agreement between coders (κ = 0.31, SD = 0.33), with a median of 0.28 and range from -0.20 to 1.00. It reflected the inherent subjectivity involved in thematic analysis. These results suggest that while the annotation framework captured meaningful patterns in the data, both topic and theme identification involved considerable interpretive judgment. The presence of some negative agreement values indicated some instances where coders disagreed more than would be expected by chance alone.

https://doi.org/10.23106/dcmi.952535736

These reliability levels are consistent with those observed in previous studies of indexing consistency [3, 16]. To ensure data quality, we excluded records that were too short/long, ambiguous, or irrelevant, or those for which no topics or themes were extracted by the LLMs. This resulted in a final dataset of 50 records, a size commonly used in IR research for evaluating the effectiveness of retrieval techniques [12] for addressing the issue of search topic variability.

3.7. Evaluation of LLMs Effectiveness

Our evaluation framework assesses the performance of three large language models (DeepSeek R1 (8b), DeepSeek R1 (14b), Gemma3 (12b)) on document-level information extraction tasks. The evaluation includes multilingual text processing and calculates precision, recall, and F1 scores at the document and corpus levels. We visualized the results by using precision-recall plots and presented the model capabilities at different levels of textual granularity (i.e. at the document and corpus levels).

4. Results

Our results demonstrate that Gemma3 (12b) consistently outperformed the DeepSeek models. Both DeepSeek R1 (8b) and Gemma3 (12b) showed moderate success in topic detection (global F1 0.31), but theme detection was considerably more difficult. We found that DeepSeek models struggle to achieve meaningful results (global F1s of 0.02 and 0.08, compared to 0.26 for Gemma3). The results of substantial document-level variability suggests the limitations of current LLMs-based systems in achieving the performance of thematic interpretations at the expert level.

4.1. Model Performance by Metrics for Topic Extraction

As shown in Figure 2, Gemma3 (12b) consistently outperforms the other models, achieving a well-balanced profile with precision, recall, and F1 scores of 0.31, 0.32, and 0.31 respectively. DeepSeek R1 (8b) results achieve high precision (0.31) but limited recall (0.19). DeepSeek R1 (14b) shows the lowest scores overall, with weaker topic evaluation performance. Overall, Gemma3 (12b) provides the most consistent results, while DeepSeek R1 (8b) achieves high precision but struggles with recall.

A comparative evaluation of DeepSeek R1 and Gemma3 (12b) for topic extraction (Table 2) reveals that Gemma3 (12b) consistently outperformed DeepSeek R1. To characterize the model performance, we used both *global metrics* (evaluating overall topic extraction accuracy) and *whole term metrics* (assessing individual term correctness within those topics). The discrepancies between these metrics can reveal a model that selects relevant topics but struggles with identifying key terminologies. We found that Gemma3 (12b) had a more balanced performance across both sets of metrics. These results suggest a capability of accurately identifying relevant document contents and their relevant terms.

The precision-recall relationship for the three models is visualized in Figure 3. Each point represents the performance of an individual document, which provides a detailed comparison of the models. Gemma3 (12b) consistently demonstrates a good balance of precision (P=0.38) and recall (R=0.32), with better topic coverage and relevance. The DeepSeek models display a

https://doi.org/10.23106/dcmi.952535736

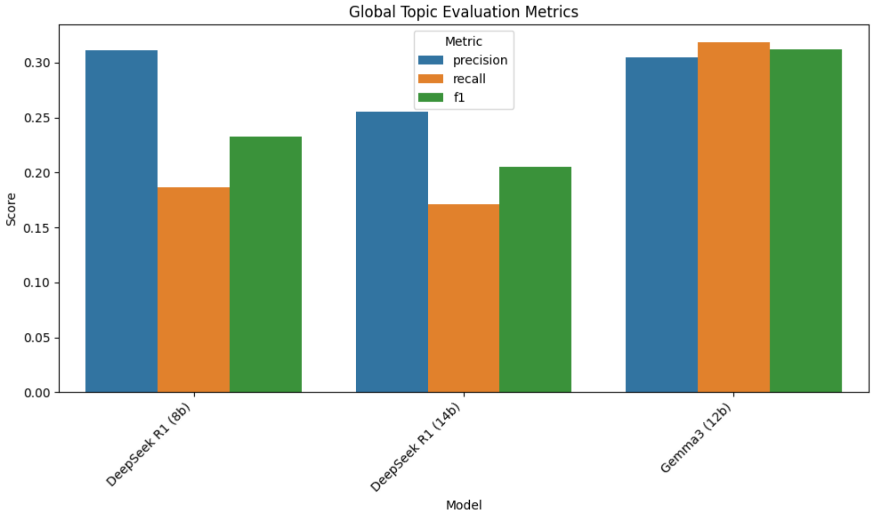

Figure 2: Performance comparison of LLM models on topic extraction tasks. The results show that Gemma3 (12b) achieved the highest global F1 score (0.31) among the tested models.

Table 2
Model Performance Comparison of Topic Extraction

Model	Global Precision	Global Recall	Global F1	Avg Whole Term F1	Avg Whole Term Precision	Avg Whole Term Recall
DeepSeek R1 (8b)	0.31	0.19	0.23	0.24	0.33	0.19
DeepSeek R1 (14b)	0.26	0.17	0.21	0.22	0.30	0.18
Gemma3 (12b)	0.31	0.32	0.31	0.34	0.38	0.32

trade-off between precision and recall. DeepSeek R1 (8b) achieves high precision (P=0.33) at the expense of low recall (R=0.19). DeepSeek R1 (14b) achieves comparable precision (P=0.30) but has the lowest recall (R=0.18). It is worth noting that the iso-F1 curves show that Gemma3 (12b) achieves the highest overall F1 score. We also observed the considerable document-level variation that needs to be further considered when assessing system performance.

4.2. Model Performance by Metrics for Theme Extraction

Table 3 reveals that Gemma3 (12b) consistently outperforms the other models. We found the most balanced profile with precision (0.29), recall (0.23), and F1 score (0.26) of the Gemma3 (12b). DeepSeek R1 (8b) presents the lowest scores across all metrics, which suggests substantially weaker theme extraction capabilities. Although DeepSeek R1 (14b) also shows low scores, Gemma3 (12b) demonstrates greater consistency in model performance.

To assess the effectiveness of various LLMs for theme extraction, we evaluated their performance using both global and whole-term metrics (Table 3). Similar to the evaluation of topic extraction, our evaluation included both 'Global' and 'Whole Term' metrics to provide a comprehensive understanding of the LLMs' performance.

https://doi.org/10.23106/dcmi.952535736

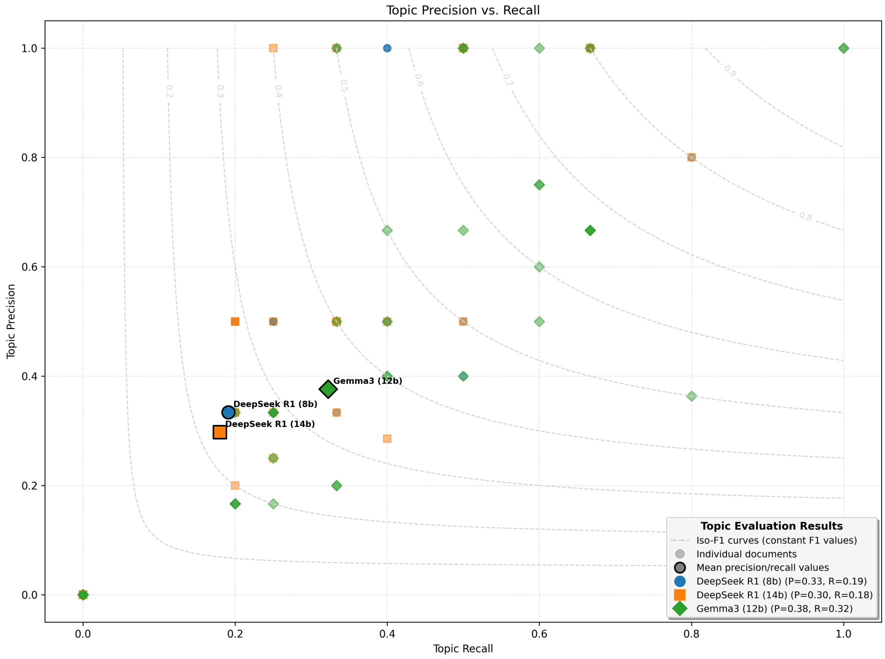

Figure 3: Precision-recall plot comparing DeepSeek R1 (8b & 14b) and Gemma3 (12b) language models, with Gemma3 showing marginally better performance (P=0.38, R=0.32) on topic evaluation tasks.

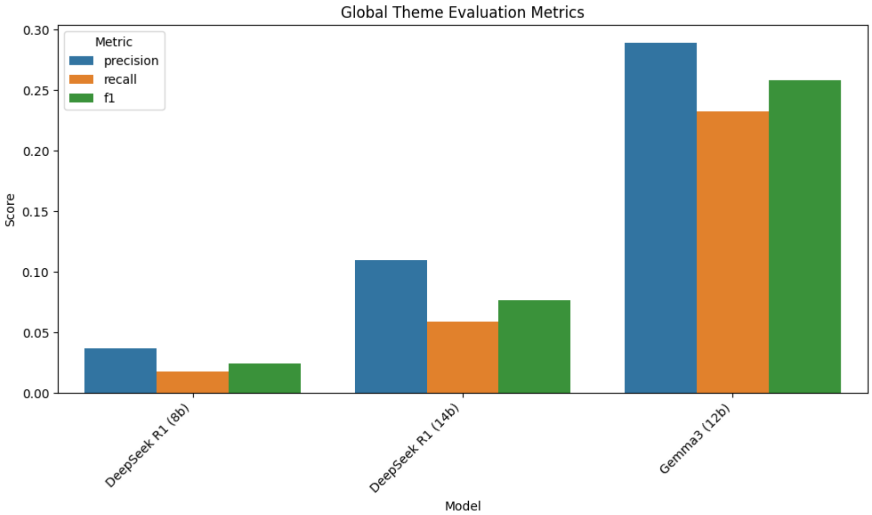

Figure 4: Performance comparison of LLM models on theme extraction tasks. The results show that Gemma3 (12b) achieved the highest global F1 score (0.258) among the tested models.

The results indicate that Gemma3 (12b) consistently outperformed the alternative models. We foudn that Gemma3 (12b) achieved significantly better performance compared to the DeepSeek models, with a balanced profile of respectable precision, recall, and F1 scores. While Gemma3

https://doi.org/10.23106/dcmi.952535736

(12b) showed consistent performance, DeepSeek models demonstrated a more substantial gap between global and term-level accuracy. This suggests a potential weakness in term-level identification of themes.

Table 3

Comparison Model Performance of Theme Extraction

Model	Global Precision	Global Recall	Global F1	Avg Whole Term F1	Avg Whole Term Precision	Avg Whole Term Recall
DeepSeek R1 (8b)	0.04	0.02	0.02	0.02	0.05	0.02
DeepSeek R1 (14b)	0.11	0.06	0.08	0.08	0.11	0.07
Gemma3 (12b)	0.29	0.23	0.26	0.28	0.32	0.26

The theme precision-recall plot (Figure 5) provides a comparison of the three models' capabilities to identify the themes from text. Gemma3 (12b) substantially outperforms the other models in theme extraction. It achieves moderate precision (P=0.32) and recall (R=0.26) for identifying and capturing thematic elements in documents.

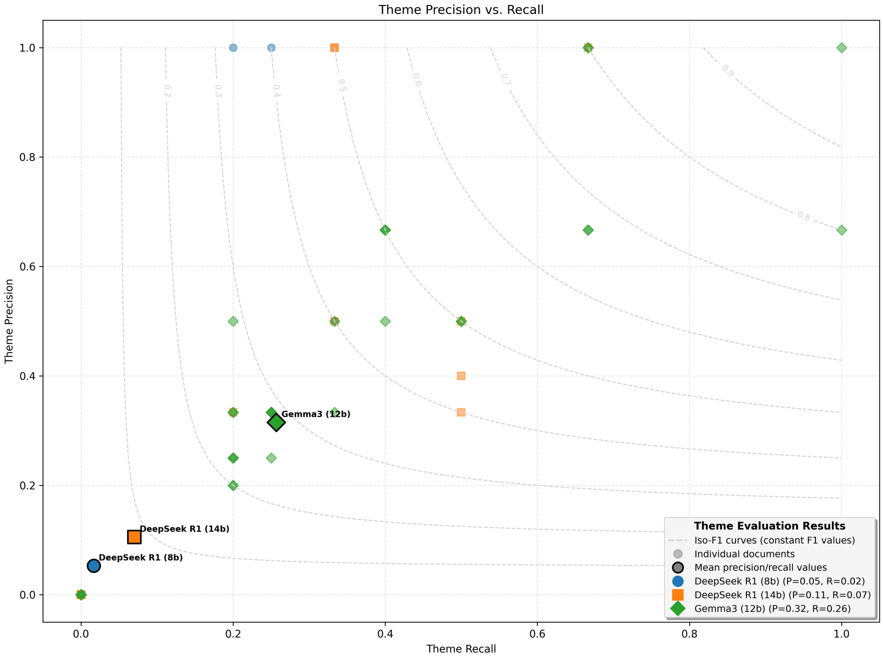

Figure 5: Precision-recall plot comparing three language models, with Gemma3 (12b) significantly outperforming (P=0.32, R=0.26) both DeepSeek models on theme evaluation tasks.

The DeepSeek models show weaker theme identification capabilities. DeepSeek R1 (14b) demonstrates very low precision (P=0.11) and recall (R=0.07), while DeepSeek R1 (8b) performs even worse with minimal precision (P=0.05) and recall (R=0.02). The distribution of individual document results also shows significant performance variability across all models. Some documents achieved perfect precision scores of 1.0, particularly with Gemma3 (12b). However,

https://doi.org/10.23106/dcmi.952535736

the widespread results show inconsistent performance across different documents. Overall, Gemma3 (12b) achieves the best performance in theme detection tasks compared to both DeepSeek variants. All models have significant room for improvement in consistently identifying patterns of meaning in text.

4.3. Discussion

Our analysis of model performance in topic and theme extraction revealed a distinction in task difficulty. The LLM-based systems were able to achieve reasonable performance for topic extraction, with relatively balanced precision-recall trade-offs. On the other hand, we foud that theme extraction was substantially more challenging, with the DeepSeek models having difficulties in extracting the more abstract elements, or patterns of meaning of theme identification. Gemma3 (12b) consistently outperformed the alternative models. Document-level variability was observed in all models, although instances of near-perfect precision scores were more common in topic extraction. As such, theme extration at the expert level requires a significantly higher level of semantic knowledge within a specific domain. While Gemma3 (12b) showed considerable performance, the DeepSeek models appeared better suited to the extraction of more concrete topics. As such, this study shows that standard quantitative IR metrics (e.g. precision, recall, and F1 score) are applicable and can be applied to evaluate the performance of LLMs on information extraction tasks from unstructured qualitative text, such as identifying topics and themes in survey comments. This evaluation methodology [12, 4] provides a framework for comparing the performance of different LLM models in extracting topics and themes from text.

The observed difficulty of the DeepSeek models and moderate performance of Gemma3 model in theme extraction suggest that it is still challenging to extract abstract thematic elements for most LLMs. In this study, we evaluated a limited set of three specific, relatively smaller LLMs: DeepSeek R1 (8b), DeepSeek R1 (14b), and gemma3 (12b) to compare these particular models on the topic and theme extraction tasks. The SemEval study evaluated a diverse range of systems developed by multiple participating teams as part of a shared task of automated subject tagging for scientific records using a controlled vocabulary [8, 9]. These systems employed various LLMs (including smaller models with 8B or fewer parameters and very large models like Qwen2.5-72B-Instruct and ChatGLM 4 (130B)) and different methodological approaches to benchmark different system designs and strategies. Both studies have successfully applied IR evaluation methodology to assess the effectiveness of LLMs. The SemEval study included the standard IR practice for ranked output evaluation and added a layer of qualitative evaluation by domain experts to assess practical utility. The present study used quantitative metrics primarily to assess overall performance and focused its qualitative assessment efforts on analyzing the inconsistencies in the human-annotated ground truth through inter-coder reliability measures, as related to the indexing consistency issues in organizing and retrieving information [16]. While this study applied quantitative IR metrics to assess model performance, the observed difficulty in theme extraction, particularly for the DeepSeek models, suggests the need for ongoing methodological refinement (cf. [17]) of the evaluation methodology of automated subject indexing. Compared to the SemEval shared task, which included both quantitative and qualitative expert evaluations, our focus on inter-coder reliability also suggests the important

https://doi.org/10.23106/dcmi.952535736

role of human assessment in understanding the capabilities of LLMs for information and theme extraction.

5. Conclusion

While standard quantitative information retrieval metrics are applicable and useful for evaluating LLM effectiveness on information and theme extraction tasks, there exists a significant performance gap between topic and theme extraction and across different models. It suggests the need for further methodological refinement and advances in LLM capabilities for modeling abstract conceptual relationships to the expert level. The challenges observed in both LLM performance and human annotation consistency suggest the important role of human assessment and understanding the nuances of information and theme extraction in a specific subject domain to inform future model design, evaluation practices and potential applications to metadata creation and management tasks.

Acknowledgments

The assistance provided by the DCMI Education Committee, the expert translations of the Metadata and AI survey, and the valuable insights from global information professionals are gratefully acknowledged.

References

[1] OCLC, OCLC annual report 2023–2024, 2024. URL: https://www.oclc.org/en/annual-report/2024/home.html.

[2] International Organization for Standardization (ISO), ISO/IEC 22989:2022: Information technology — artificial intelligence — artificial intelligence concepts and terminology, 2022. URL: https://www.iso.org/standard/74296.html.

[3] J. D. Anderson, J. Pérez-Carballo, The nature of indexing: How humans and machines analyze messages and texts for retrieval. Part I: Research, and the nature of human indexing, Information Processing & Management 37 (2001) 231–254. doi:10.1016/S0306-4573(00)00026-1.

[4] K. Sparck-Jones, Automatic indexing, Journal of Documentation 30 (1974) 393–432. doi:10.1108/eb026588.

[5] National Library of Medicine, Automated indexing FAQs, 2025. URL: https://support.nlm.nih.gov/kbArticle/?pn=KA-05326.

[6] N. Firoozeh, A. Nazarenko, F. Alizon, B. Daille, Keyword extraction: Issues and methods, Natural Language Engineering 26 (2020) 259—291. doi:10.1017/S1351324919000457.

[7] V. Braun, V. Clarke, Using thematic analysis in psychology, Qualitative Research in Psychology 3 (2006) 77–101. doi:10.1191/1478088706qp063oa.

[8] L. Kluge, M. Kähler, DNB-AI-Project at SemEval-2025 Task 5: An LLM-ensemble approach for automated subject indexing, 2025. arXiv:2504.21589.

https://doi.org/10.23106/dcmi.952535736

[9] J. D'Souza, S. Sadruddin, H. Israel, M. Begoin, D. Slawig, SemEval-2025 Task 5: LLMs4Subjects – LLM-based automated subject tagging for a National Technical Library's open-access catalog, 2025. arXiv:2504.07199.

[10] O. Suominen, J. Inkinen, M. Lehtinen, Annif and Finto AI: Developing and implementing automated subject indexing, JLIS.it 13 (2022) 265–282. doi:10.4403/jlis.it-12740.

[11] Y.-H. Liu, N. Wacholder, Evaluating the impact of MeSH (Medical Subject Headings) terms on different types of searchers, Information Processing & Management 53 (2017) 851–870. doi:10.1016/j.ipm.2017.03.004.

[12] E. M. Voorhees, C. Buckley, The effect of topic set size on retrieval experiment error, in: Proceedings of the ACM SIGIR Conference, 2002, pp. 316–323. doi:10.1145/564376.564432.

[13] C. Buckley, E. M. Voorhees, Retrieval evaluation with incomplete information, in: Proceedings of the ACM SIGIR Conference, 2004, pp. 25–32. doi:10.1145/1008992.1009000.

[14] K. Sparck-Jones, C. J. Van Rijsbergen, Information retrieval test collections, Journal of Documentation 32 (1976) 59–75. doi:10.1108/eb026616.

[15] J. Cohen, A coefficient of agreement for nominal scales, Educational and Psychological Measurement 20 (1960) 37–46. doi:10.1177/001316446002000104.

[16] T. Saracevic, Individual differences in organizing, searching and retrieval information, in: Proceedings of the ASIS Annual Meeting, volume 28, 1991, pp. 82–86.

[17] K. Golub, D. Soergel, G. Buchanan, D. Tudhope, M. Lykke, D. Hiom, A framework for evaluating automatic indexing or classification in the context of retrieval, Journal of the Association for Information Science and Technology 67 (2016) 3–16. doi:10.1002/asi.23600.

https://doi.org/10.23106/dcmi.952535736