International Journal of Smart Computing and Artificial Intelligence
International Institute of Applied Informatics
2025, Vol. 9, No. 1, IJSCAI865

Zero-Shot Text Classification Using Large Language Models for Key Audit Matters in Japanese Audit Reports

Nobushige Doi [*], Yusuke Nobuta [†],
Takeshi Mizuno [*]

Abstract

Japanese-listed companies are required to submit audit reports to the Prime Minister of Japan. In principle, these reports must include "Key Audit Matters" (KAMs), which are matters that the auditors, as professional experts, have judged as particularly important when auditing financial statements. A previous study proposed an automatic classification method called zero-shot text classification for KAMs. We examine whether zero-shot text classification with large language models (LLMs) such as ChatGPT can automatically classify KAMs. We also examine how the following three approaches contribute to the accuracy of zero-shot text classification by LLMs: definition refinement, majority decision-making based on LLM outputs, and use of state-of-the-art models. The experimental results confirm that definition refinement and majority decision-making based on more than three results are useful to some extent. Furthermore, the latest ChatGPT model, gpt-4-1106-preview of the Generative Pre-trained Transformer 4 (GPT-4) model, achieved a classification accuracy of up to 87.2%.

Keywords: Auditing, Financial disclosure, Key Audit Matters, Text classification, Large Language Models, ChatGPT.

1 Introduction

The International Auditing and Assurance Standards Board (IAASB) of the International Federation of Accountants began its audit reporting project in 2011. After publishing two consultation papers and an exposure draft of the standards, the IAASB published a series of International Standards on Auditing (ISA) for audit report reform in January of 2015 [1]. The new ISA 701, "Communicating Key Audit Matters in the Independent Auditor's Report," requires that new reports include the auditor's report on the general-purpose financial statements of a listed company. Audit report reform was also discussed in the United States (U.S.). In June of 2017, the Public Company Accounting Oversight Board issued its Audit Standard 3101 [2] "The Auditor's Report on an Audit of Financial Statements When the Auditor Expresses an

[*] Japan Exchange Group, Inc., Tokyo, Japan
[†] Tokyo Stock Exchange, Inc., Tokyo, Japan

Unqualified Opinion," which introduces the Critical Audit Matter (CAM) to the audit report. The auditing standards in Japan were revised in July of 2018. The new standards introduce Key Audit Matters (KAMs), which improve the transparency of the Japanese auditing process while considering international trends [3]. The KAMs in Japan were applied to audit reports for the fiscal year ending March 31 of 2021, and have continued until the present; various countries and regions have introduced similar systems to KAMs and CAMs.

In Japan, listed companies must submit security reports detailing the company's general situation and financial status to the Prime Minister of Japan within three months of the fiscal year end. The financial statements attached to the financial section of the security report must also be audited. The security report and the auditor's report, including KAMs, are disclosed through the EDINET (Electronic Disclosure for Investors' NETwork) system operated by the Financial Services Agency (FSA).

The KAMs of Japanese-listed companies comprise a Title, a major description of the audit along with reasons for the decision (Description and Reasons), and the Auditor's Responses. As the text contents are not attached with tags indicating the topics or other additional content; the dynamics of KAMs can be understood only by reading and understanding the auditor's reports. Automatic topic-classification methods would alleviate the burden of manually analyzing the auditor's reports of all listed companies. However, traditional machine learning-based methods incur substantial costs in building training data. Moreover, proper nouns and terminologies in financial documents can change over time, adding maintenance costs to machine-learning models used in practice.

To solve this problem, Doi et al. [4] proposed a topic-classification method for KAMs based on zero-shot text classification (hereafter, we refer to Doi et al.'s work as *the existing research*). Zero-shot text classification is a natural language processing task that inputs a text and a list of potential classes and outputs the class to which the inputted text belongs without requiring training data. The existing research reported that large language models (LLMs) such as OpenAI's Generative Pre-trained Transformer (ChatGPT) gave highly accurate results, but further accuracy improvement might be required for a detailed analysis. In addition, LLMs have remarkably improved in recent times and the latest LLMs might offer more sufficient accuracy.

Therefore, this study reexamines the classification accuracy of the topic-classification method using zero-shot text classification for KAMs with LLMs. It also investigates how the following three approaches contribute to the classification accuracy: refinement of the definition of KAM topics, majority vote in the output results of LLMs, and use of the latest models. The proposed method is evaluated on 100 correct data samples of KAM topics from three workers.

Our study makes two major contributions to topic classification of KAMs. First, it re-verifies the usefulness of zero-shot text classification using LLMs for KAM topic classification. Second, it explores how refining the definitions, determining majority vote results, and applying the latest models improve the accuracy of the proposed method.

One limitation of this study is the small dataset (only 100 sets) for evaluating the proposed model. Although this dataset is high-quality, the small sample size potentially limits the results to some KAM topics or the industries of listed companies. Furthermore, the details of ChatGPT (the LLM used in this study) are privacy-protected, which limits the reproducibility of the evaluation results.

Section 2 introduces the related studies. Section 3 explains Japan's Key Audit Matters system, highlighting its introduction and specific characteristics. Sections 4 and 5 detail the dataset used in this study and the proposed methodology for zero-shot text classification of KAM topics using Large Language Models. Section 6 presents the classification strategy for these KAMs, while Section 7 discusses the proposed method's experimental setup, results, and evaluations. Finally, Section 8 concludes the study, summarizing its findings and outlining potential directions for future research.

2 Related Studies

This section describes the existing research on topic classification for KAMs, CAMs, and zero-shot text classification.

As of April 2023, the Audit Analytics database of Audit Analytics Inc. provides CAM topics for U.S. listed companies and KAM topics for listed companies in Switzerland and member countries in the European Economic Area [5]. Using Audit Analytics' Topic Classification, Huang [6] analyzed the CAM trends of U.S.-listed companies in different industries and distinguished five broad categories of CAM topics: Intangibles, Revenue recognition, Operating accruals, Taxes, and Others. Lynch et al. used tax-related topics from the Audit Analytics database of KAMs of companies listed on the London Stock Exchange to analyze the determinants and consequences of tax-related KAMs [7].

The Japanese Institute of Certified Public Accountants (JICPA) analyzes the KAM topic classifications of Japanese-listed companies. The case analysis of the first year of mandatory KAM application (fiscal year ending March 31, 2021) analyzed trends by industry sector, classified into 14 topics [8]. In the case study analysis of April 2021–March 2022, nine topics were selected for qualitative research [9]. As these classifications are not necessarily generic, designing low-cost classifications and KAM topics suitable for Japanese-listed companies is a challenging task.

Zero-shot text classification can eliminate subjectivity and reduce the human effort of abstract screening in systematic reviews [10]. Wang et al. reported that GPT-3 [11] reduces the costs associated with label assignment [12]. Kuzman et al. [13] used ChatGPT for zero-shot text classification of nine types of sentences, such as news and promotions, on English and Slovenian datasets. They reported that ChatGPT outperforms finely tuned models in terms of accuracy.

3 Key Audit Matters in Japan

The "Opinion on Revisions to Auditing Standards," issued by the Business Accounting Council of the FSA on July 5 of 2018, determined the introduction of KAMs in Japan. Early KAM applications were available in the audit report of the fiscal year ending March 31, 2020. At that time, KAMs were disclosed by 48 listed companies [14]. Since the fiscal year ending March 31, 2021, all listed companies in Japan have been required to include KAMs in their audit reports.

Two types of audit reports exist in Japan: the "Audit Report on the Current Consolidated Financial Statements" and the "Audit Report on the Current Financial Statements." In the present paper, the KAMs in the "Audit Report on the Current Consolidated Financial Statements" are called *consolidated KAMs*, whereas those in the "Audit Report on the Current Financial

Statements" are called *nonconsolidated KAMs*. According to the JICPA, the average number of consolidated KAMs for listed companies ending 31 March of 2022 was 1.29 per company, and the average number of characters per KAM is 1,248 [9].

As mentioned above, all Japanese KAMs comprise a title, description and reasons, and auditor's responses. Because Japanese audit reports are disclosed as HTML documents and each item is tagged in XBRL format, this study adopted an XML parser that can process XBRL tags.

4 Dataset

The proposed method was evaluated on the dataset made from the KAM dataset used in the existing research. The KAM dataset contains the most recently consolidated KAMs for each listed company in Japan as of March 2023 for the fiscal years ending April 2021 through March 2022. Included in the KAM dataset are domestic companies listed on the Prime, Standard, or Growth market of the Tokyo Stock Exchange as of December 2022. Audit reports were collected through the EDINETAPI. The KAMs in each audit report for each listed company were extracted by referring to the XBRL tags corresponding to the "Title," "Description and Reasons," and "Auditor's Responses" sections. Following the previous study of JICPA [10, 11], the present study excludes the nonconsolidated KAMs. In addition, the KAM text was normalized through NFKC normalization. The final KAM dataset consisted of 3,928 consolidated KAMs.

The proposed method was evaluated on an evaluation dataset of 100 KAMs extracted from the KAM dataset, consistent with the existing research. However, as mentioned in subsection 5.1, the manually assigned topics of KAM in the evaluation dataset were partially changed to refine the topic definitions and review the overall allocation results.

5 Proposed Method

Consistent with the existing research, the topic classes defined in this study are suitable for KAMs in Japan and the topic-classification method for KAMs was evaluated using zero-shot text classification. Furthermore, this study examined the effects of refining the definition of KAM topics, majority voting in the LLM outputs, and using the latest models on classification accuracy.

5.1 Refinement of Definition of KAM Topics

Table 1 lists the KAM topics defined in the existing research. Five items listed in Table 1— "Impairment of fixed assets," "Revenue recognition," "Valuation of deferred tax assets," "Accounting for software," and "Valuation of inventory"—appear to be vaguely outlined and their definitions have been re-summarized in Table 2.

Accordingly, the present study reviewed the evaluation dataset used in the existing research. The evaluation dataset for the evaluation experiment (described later) was composed of 100 consolidated KAMs classified into one or more topics (Table 1) extracted from the KAM dataset. The same 100 targeted KAMs were extracted in the existing research, but the present study revisits the topic allocation to review the overall allocation results. Although Table 2 revises the definitions of Table 1, eight topics were found to correspond to both "Impairment of fixed assets"

and "Others." Consequently, the number of KAMs corresponding to "Impairment of fixed assets" increased by 8 KAMs from that of the existing research. It should be noted that no corrections in topic allocation followed the revised definitions in Table 2.

Table 3 presents the breakdown of topics included in the evaluation dataset and the statistical information in the text. The seven KAMs in the "Others" topic include Investment evaluations (2 cases), Organizational restructuring (2 cases), Evaluation of IT systems (1 case), Transfer pricing taxation (1 case), and Unclassifiable (1 case).

In the subsequent experiment, prompts will be tested on both the pre- and post-refined definitions.

Table 1: List of KAM Topics with Pre-refined Definitions

#	Topic Name	Outline
1	Impairment of fixed assets	Related to the valuation of tangible or intangible fixed assets (excluding goodwill)
2	Impairment of goodwill	Related to the valuation of goodwill
3	Revenue recognition	Related to revenue recognition
4	Valuations of deferred tax assets	Related to the valuation of deferred tax assets
5	Accounting for the software	Related to accounting for software
6	Going concern assumption	Related to the premise when circumstances exist that may cast significant doubt on the assumption of a going concern; however, no material uncertainty is recognized at this point
7	Valuation of the inventory	Related to the valuation of inventories
8	Valuations of trade receivables	Related to the valuation of trade receivables, such as accounts receivable and estimation of allowance for doubtful accounts for trade receivables
9	Estimated liabilities	Related to estimating liabilities, including reserves unidentified in other audit areas
10	Others	Issues other than those stated above

5.2 Use of Majority Voting in LLM Outputs

The output results of ChatGPT and other LLMs can randomly change with input time. To enhance the accuracy of topic classification, this study proposes a majority voting process that inputs the same prompt into the LLM multiple times and selects the topic classified by more than half of the outputs as the final prediction. This method minimizes the risk of relying on deviant output results and achieves more rational outcomes.

In the subsequent experiment, this study performs both single trials (as in the existing research) and majority voting trials with three and five output results.

Table 2: List of KAM Topics with Post-refined Definitions (Highlighted in Bold Font)

#	Topic Name	Outline
1	Impairment of fixed assets	Related to the valuation of **tangible fixed assets** or intangible fixed assets (excluding goodwill)
2	Impairment of goodwill	Related to the valuation of goodwill
3	Revenue recognition	Related to revenue recognition **is the calculation of the amount of revenue, starting with the period attribution of sales excluding software and the existence and accuracy of sales**
4	Valuations of deferred tax assets	Related to the valuation of deferred tax assets**, including their recoverability and reasonableness**
5	Accounting for the software	Related to accounting for software **used to provide business processing services to third parties or for efficiently or effectively conducting the company's operations**
6	Going concern assumption	Related to the premise when circumstances exist that may cast significant doubt on the assumption of a going concern; however, no material uncertainty is recognized at this point
7	Valuation of the inventory	Related to the valuation of inventories**, such as merchandize, products, and raw materials held in stock**
8	Valuations of trade receivables	Related to the valuation of trade receivables, such as accounts receivable and estimation of allowance for doubtful accounts for trade receivables
9	Estimated liabilities	Related to estimating liabilities, including reserves unidentified in other audit areas
10	Others	Issues other than those stated above

5.3 Use of the Latest Models

The existing research used the most recent models at that time, ChatGPT-3.5 [15] and ChatGPT-4 [16]. The GPT-3.5 model was gpt-3.5-turbo-0301 (snapshot of March 1, 2023) and the GPT-4 model was based on the web browser as of March 23, 2023. The GPT-4 model at that time is presumed to approximate the gpt-4-0613 model (snapshot of June 13, 2023). Considering the advancements in LLM technology, the latest models are expected to improve the accuracy of topic classification. Therefore, this study compares the accuracies of the latest models (at the time of writing this paper) with those of the models used in the existing research.

In this experiment, we employed GPT-3.5 versions gpt-3.5-turbo-0301 (snapshot of March 1, 2023) and gpt-3.5-turbo-1106 (snapshot of November 6, 2023), and GPT-4 versions gpt-4-0613 (snapshot of June 13, 2023) and gpt-4-1106-preview (snapshot of November 6, 2023).

Table 3: Breakdown of the Evaluation Dataset

#	Topic Name	No.	Average number of characters		
			Title	Description and Reasons	Auditor's Responses
1	Impairment of fixed assets	29	30.4	721.0	501.3
2	Impairment of goodwill	16	17.8	595.9	514.3
3	Revenue recognition	22	22.8	614.8	558.9
4	Valuations of deferred tax assets	13	18.5	565.0	480.8
5	Accounting for the software	4	29.0	796.0	525.0
6	Going concern assumption	3	28.3	735.3	592.3
7	Valuation of the inventory	8	20.3	526.5	428.8
8	Valuations of trade receivables	3	31.0	968.5	769.5
9	Estimated liabilities	3	30.3	494.3	389.7
10	Others	7	24.1	636.6	528.9
	Whole	100	24.69	645.62	514.95

6 Zero-shot Text Classification of KAMs using ChatGPT

The outputs generated by GPTs naturally follow the inputted sequence of tokens. Therefore, this study inputs the list of topics (except "Others") to the GPTs in prompt-format text and classifies the output text into one or more topics. The prompt-format text is explicitly marked to classify the input text as "Others" if it cannot be classified into any topics. The prompts include not only the list of topics, but also the definition of each topic as outlined in Tables 1 and 2.

The GPT output does not necessarily comprise defined topics alone. Therefore, these ChatGPT methods classify the topics of the input sentence based on whether the output sentence contains the characters of defined topics.

The input prompt consists of a preamble, a list of topics, the topic details, information on the input sentence, and sentences indicating the task. We implemented these experiments in Japanese, and an English-translated example of this prompt is shown below.

Example Prompt (reference translation)

KAMs in the Annual Securities Report are classified as follows:

Category list

Impairment of fixed assets, Impairment of goodwill, Revenue recognition, Valuations of deferred tax assets, Accounting for the software, Going concern assumption, Valuation of the inventory, Valuations of trade receivables, Estimated liabilities, Others

Category list details

Impairment of fixed assets: Related to the valuation of tangible fixed assets or intangible fixed assets (ex-cluding goodwill)

…

{Abbreviation}

Title of the input statement

{Title of KAMs}

Content of the input statement

{Description and Reasons of KAMs}

Task

Classify the input statement into one or more categories and output only their name. If the input statement belongs to no categories, the KAMs are categorized as “Others.”

7 Evaluation Experiment

7.1 Experimental Environment

To verify the usefulness of zero-shot text classification with LLMs in KAM classification and the proposed method, we estimated the topics of 100 KAMs from the evaluation dataset using each method. Following the existing research, we adopted the ChatGPT models GPT-3.5 and GPT-4 as the LLMs. The estimation result of each model was compared with the classification results

estimated by three workers (Ground Truth). The procedure was performed five times for each method and the classification accuracies were averaged to give the final result of each method.

Following the existing research, each proposed method was input with two patterns of KAM input text: the KAM title alone and both the KAM title and its description and reasons. To evaluate the usefulness of refining the definitions, experiments were also conducted on the pre- and post-refined outlines. The effectiveness of majority voting was evaluated for different numbers of majority voted results: one, three, and five. The usefulness of the latest models was evaluated on two GPT-3.5 versions (gpt-3.5-turbo-0301 and gpt-3.5-turbo-1106) and two GPT-4 versions (gpt-4-0613 and gpt-4-1106-preview). Consequently, the classification accuracy was evaluated for 48 input patterns. The topic-specific classification accuracy of the model giving the highest overall accuracy was also evaluated.

The evaluation metric in this study was the Accuracy, defined as the percentage of KAMs for which the classification results matched the Ground Truth. The Precision, Recall, and F-1 Score were also computed as measures of the topic-specific classification accuracy.

Defining the evaluation metrics is crucial to enhance the experimental results' clarity and robustness. We provide clear definitions for Accuracy, Precision, Recall, and F-1 Score. First, the elements of the mixing matrix needed for the definition of these evaluation indicators are defined as follows:

TP = True Positives: KAMs correctly classified as a specific topic.

TN = True Negatives: KAMs correctly not classified as a specific topic.

FP = False Positives: KAMs incorrectly classified as a specific topic.

FN = False Negatives: KAMs incorrectly not classified as a specific topic.

Accuracy is the ratio of correctly predicted observations to the total observations as formula (1). It's a measure of the overall correctness of the model. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations as formula (2). It measures the accuracy of positive predictions. Recall is the ratio of correctly predicted positive observations to all observations in the actual class as formula (3). It measures the model's ability to detect positive samples. The F-1 Score is the weighted average of Precision and Recall as formula (4). It is beneficial when seeking a balance between Precision and Recall.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F\text{-}1\ Score = 2 \times \frac{Precison \times Recall}{Precision + Recall} \quad (4)$$

7.2 Experimental Results and Evaluation

7.2.1 Overall Trend

The experimental results are presented in Table 4. In general, GPT-4 yielded higher Accuracy scores than GPT-3.5. The Accuracy was improved to 87.2% by inputting the KAM title along with its description and reasons, refining the definition, setting three or five majority votes, and applying gpt-4-1106-preview. In contrast, the Accuracy was lowest (74.2%) after inputting the KAM title along with its description and reasons, refining the definition, excluding majority voting, and applying the gpt-3.5-turbo-0301 model.

Table 4: Evaluation Results of Accuracy by the Proposed Method

Model	No. of Majority Voting	Title		Title + Description and Reasons	
		Pre-refined	Post-refined	Pre-refined	Post-refined
gpt-3.5-turbo-0301	1	83.6%	82.6%	82.6%	78.6%
	3	83.4%	84.2%	84.2%	78.4%
	5	83.4%	83.0%	83.0%	79.4%
gpt-3.5-turbo-1106	1	79.0%	78.8%	78.8%	74.2%
	3	78.8%	79.0%	79.0%	78.0%
	5	78.2%	78.8%	78.8%	77.2%
gpt-4-0613	1	83.6%	83.2%	83.2%	81.8%
	3	83.8%	84.6%	84.6%	82.2%
	5	83.8%	84.0%	84.0%	82.0%
gpt-4-1106-preview	1	84.0%	85.2%	84.4%	86.6%
	3	85.4%	86.0%	86.0%	87.2%
	5	86.4%	86.4%	86.4%	87.2%

7.2.2 Evaluation of the Proposed Method

First, we discuss the usefulness of refining the definition. Refining the definition improved the classification accuracy of the GPT-4 model gpt-4-1106-preview, suggesting that models with more advanced interpretive abilities can extract more accurate information from detailed topic definitions. However, the accuracy tended to decrease in the other models, possibly because models with inferior interpretive abilities cannot easily process refined definitions, increasing the likelihood of misclassification. In contrast, models with sufficient interpretive abilities can exploit detailed topic definitions to improve the accuracy of zero-shot text classification. In particular, when the topics include ambiguous expressions, refining the definition will likely enhance the classification accuracy.

We next discuss the usefulness of majority voting. Overall, the classification accuracy of multiple output results exceeded that of one output result because majority voting of multiple outputs accounts for the inherent variability in ChatGPT outputs, which is not considered in single outputs. However, the accuracy improvement was limited to a few percentage points and the accuracies did not noticeably differ between three and five outputs. This finding suggests that the effect of majority voting has certain limits.

Finally, we discuss the usefulness of applying the latest models. Comparing the performances of the GPT-4 models, the newer gpt-4-1106-preview tended to achieve higher accuracy than the older gpt-4-0613 model, reflecting the technical improvements in the performance of ChatGPT-4. Conversely, the accuracy of the older GPT-3.5 model (gpt-3.5-turbo-0301) tended to exceed that of the newer model (gpt-3.5-turbo-1106), indicating that the classification accuracy is not necessarily improved by updating a model but depends on the individual characteristics of the model.

In summary, the proposed method achieved high classification accuracy when combining refined definitions with majority voting and (usually) applying models with high interpretive abilities. However, the model selection and prompt design must be carefully considered, as refined definitions are potentially counterproductive, especially when applying models with limited interpretive abilities. In addition, majority voting has limited ability to improve the classification accuracy and excessively many repeats can reduce the model's effectiveness. The latest models will likely improve the classification accuracy as they are technologically advanced. However, the latest model is not necessarily the best choice, and selecting a suitable model for the purpose is crucial.

7.2.3 *Trends in Classification Accuracy Organized by Topic*

Table 5: Evaluation results of different topics

#	Topic Name	Precision	Recall	F-1 Score
1	Impairment of fixed assets	95.8%	77.9%	85.9%
2	Impairment of goodwill	94.1%	100.0%	97.0%
3	Revenue recognition	94.6%	95.5%	95.0%
4	Valuations of deferred tax assets	100.0%	100.0%	100.0%
5	Accounting for the software	65.3%	85.0%	73.8%
6	Going concern assumption	100.0%	100.0%	100.0%
7	Valuation of the inventory	88.9%	100.0%	94.1%
8	Valuations of trade receivables	100.0%	73.3%	82.0%
9	Estimated liabilities	69.0%	100.0%	81.4%
10	Others	96.0%	54.3%	69.0%
	Whole	92.7%	88.9%	90.7%

Table 5 tabulates the topic-specific classification evaluation results of the model giving the highest accuracy; specifically, the GPT-4 model gpt-4-1106-preview input with the KAM title along with its description and reasons, the refined definition, and five outputs for majority voting. Note that the values represent the average of the five output values.

Overall, the F-1 scores were variable, with some topics having high scores and others having low scores. High F-1 scores were obtained for "Impairment of goodwill," "Revenue recognition," "Valuation of deferred tax assets," "Going concern assumption," and "Inventory valuation." In these topics, the points to be considered by auditors are similar across samples, so the model can easily infer accurate classifications based on specific keywords or contexts. In contrast, topics such as "Impairment of fixed assets," "Accounting for the software," and "Valuations of trade receivables" are sometimes misclassified into other KAMs. The points for auditors to consider on these topics often vary across samples and will likely include generic words or ambiguous expressions within KAMs.

The classification accuracy of the low F-1-scoring topics might be improved by subdividing these topics. For example, "Impairment of fixed assets" could be subdivide into "Impairment of tangible fixed assets" and "Impairment of intangible fixed assets excluding goodwill". Similarly to the existing research, the classification accuracy is lowered for the "Other" category. As a solution, the definition of new topics is suggested.

8 Conclusion

This study proposed and validated a method for automating the classification of KAM topics in the audit reports of listed companies in Japan through zero-shot text classification. Three main approaches were considered: refinement of the definition of KAM topics, majority voting based on the output results of LLMs, and use of the latest models. The impact of each approach on the classification accuracy was evaluated.

The proposed method yielded three key results. First, refining the topic definitions improved the classification accuracy of models with high interpretive abilities. Second, majority voting of the classification results reduced the risk of relying on a single output and yielded more reliable outcomes than the single output. Finally, the latest models will be technologically advanced and enhance the classification accuracy.

However, the proposed method has several limitations. First, the small size of the evaluation dataset limits the generalizability of the results. Additionally, the interpretive abilities and consistencies differ among the LLM outputs, suggesting that the effectiveness of accuracy improvement through majority voting is limited. Moreover, the effectiveness of the proposed method largely depends on the model and the prompt design, highlighting the important of selecting the optimal method.

In future research, we should evaluate the proposed method on larger datasets to generalize the results. We should also explore the potential for accuracy improvement by combining different models and approaches. Furthermore, the proposed method must be periodically reevaluated as language models evolve. Such updates should further improve the efficiency and accuracy of

KAM topic classification.

References

[1] International Auditing and Assurance Standards Board. Handbook of International Quality Control, Auditing, Review, Other Assurance, and Related Services Pronouncements, volume 1. International Federation of Accountants, 2015.

[2] Public Company Accounting Oversight Board. As 3101: The auditor's report on an audit of financial statements when the auditor expresses an unqualified opinion. 2017. https://pcaobus.org/oversight/standards/auditing-standards/details/AS3101 (accessed on January 31 2024).

[3] The Financial Services Agency. Release of "opinion on the revision of auditing standards" (in Japanese). 2018. https://www.fsa.go.jp/news/30/sonota/20180706.html (accessed on January 31 2024).

[4] N Doi, Y Nobuta, and T Mizuno. Topic classification of key audit matters in Japanese audit reports by zero-shot text classification. In 2023 14th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), pages 540–545, Los Alamitos, CA, USA, July 2023. IEEE Computer Society.

[5] Audit Analytics, Inc. Database catalog. https://www.auditanalytics.com/product-catalog (accessed on January 31, 2024).

[6] Qian Huang. Do Critical Audit Matter Disclosures Impact Investor Behavior? PhD thesis. Columbia University, 2021.

[7] Dan Lynch, Aaron Mandell, and Linette M Rousseau. The determinants and unintended consequences of expanded audit reporting: Evidence from tax-related key audit matters. Available at SSRN 3689349, 2021.

[8] The Japanese Institute of Certified Public Accountants. Case analysis report for the first year of mandatory application of key audit matters (year ended March 31, 2021) (in Japanese). 2021. https://jicpa.or.jp/specialized_field/20211029fgf.html (accessed on January 31, 2024).

[9] The Japanese Institute of Certified Public Accountants. Case analysis report of key audit matters (April 2021–March 2022) (in Japanese). 2022. https://jicpa.or.jp/specialized_field/20221226cgi.html (accessed on January 31, 2024).

[10] Carlos Francisco Moreno-Garc´ıa, Chrisina Jayne, Eyad Elyan, and Magaly Aceves-Martins. Abstract screening for systematic reviews using machine learning and zero-shot classification. Available at SSRN 4210704.

[11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[12] Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. Want to reduce labeling cost? GPT-3 can help. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 4195–4205, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[13] Taja Kuzman, Igor Mozetic˘, and Nikola Ljubes˘ic´. Chatgpt: Beginning of an end of manual linguistic data annotation? use case of automatic genre identification, 2023.

[14] The Financial Services Agency. Key audit matters (KAMs) characteristic examples and key points for description (in Japanese). 2022. https://www.fsa.go.jp/news/r3/sonota/20220304-2/01.pdf (accessed on January 31, 2024).

[15] OpenAI. Introducing ChatGPT. https://openai.com/blog/chatgpt (accessed on January 31, 2024).

[16] OpenAI. Gpt-4 technical report. 2023. https://cdn.openai.com/papers/gpt-4.pdf (accessed on January 31 2024).