Tom Strube*, Tom Nowak, Mariia Pokotylo, and Bernd Kuhlenkötter

Reliable and Content-specific Support for Keyword Selection through AI and Statistics

Characterising Educational Content with Large Language Models & Agreement Analyses

https://doi.org/10.1515/cdbme-2024-2154

Abstract: Due to the recent popularity and availability of Large Language Models (LLMs), creators of educational materials can more efficiently extract keywords for use in personalised learning recommendations than ever before. However, due to the LLMs' probabalistic nature, the automation of the otherwise labour-intense keyword extraction inherits the risk of biased and non-explainable results. In this research, we present an original framework to enhance keyword selection based on content title and description through a novel, reliability-sensitive, keyword selection algorithm. For this, we collected 38 potential keywords (together with their definitions) for five topics on dementia care from previous studies, together with two contents per topic. To assess the new method's support in extracting keywords, we then prompted 5 human experts and 3 LLMs (using Retrieval Augmented Generation (RAG) for the keyword definitions) to select keywords to include and exclude for each content. Using Krippendorf's α metric, we then were able to adapt to the present agreement, and to reliably select keyword sets for inclusion and exclusion for each content individually. Last, we compared these LLM-based keyword sets with those selected by humans to assess the impact of the adaptive keyword selection algorithm. Overall, the results suggest that LLMs generally struggle with the task (66% of extraction attempts either contained hallucinated or did not return any keywords), and topic-wise internal agreement is low (α=0.59 (0.42) for model 3 (using RAG) on average; α=0.68 for human raters). Due to this, the reliable keyword selection resulted in a median set of 6|27 keywords for inclusion|exclusion per topic, with many of those keywords being within the benchmark keyword sets selected by human raters. To conclude, this approach shows effective in adapting to different levels of agreement in extracting keywords.

***Corresponding author: Tom Strube,** Fraunhofer Institute for Software and Systems Engineering, Dortmund, Germany; E-Mail: tom.strube@isst.fraunhofer.de
Tom Nowak, Bernd Kuhlenkötter, Ruhr University Bochum, Bochum, Germany
Mariia Pokotylo, Fraunhofer Institute for Software and Systems Engineering, Dortmund, Germany

Keywords: Education, Reliability, Keyword Extraction, Content Analysis, Generative AI

1 Introduction

Nowadays, continuous education has become more accessible and relevant than ever before. Especially in hospitals and care homes, user-centric learning experiences through interactive systems are increasingly popular due to social dynamics such as rising nursing needs and the CoVID-19 pandemic. These primarily knowledge-based systems, such as recommender systems and keyword search engines, thereby strongly rely on content metadata for providing learners with personalised learning experiences [1]. Yet, manually extracting suitable keywords is costly and highly subjective, posing the risk of recommendation biases.

To address this issue, many statistical, linguistic, machine-learning-based or hybrid methods were developed [2]. In addition, with the recent rise of Large Language Models (LLMs), a new and yet nontransparent alternative with a general understanding of language and meaning emerged. Still, their popularity within research and industry poses the risk of unreliable keyword extraction, while trying to resolve the previously outlined need for support in content metadata enhancement and user-centric education and training.

In our study, we investigate the degree to which statistical analysis and LLMs can be merged to support humans in reliably selecting keywords for content metadata. To do so, we propose a content-sensitive approach to reliable keywords selection using Krippendorff's α and results from studies conducted in the research project *MINDED.Ruhr* by Malek et al [3]. Using this, we analyse the performance of pre-trained LLMs for extracting keywords from content descriptions of dementia-focused learning materials and discuss the effects of keyword definitions in Retrieval Augmented Generation (RAG) for keyword extraction, adapting LLMs to the content domain. Overall, we tested and compared three different approaches to keyword selection to show the support and flexibility our statistics-based framework for keyword extraction provides.

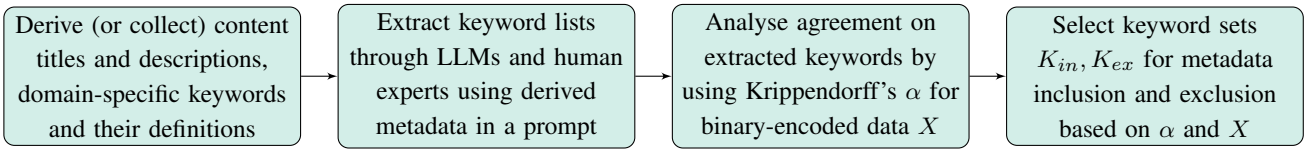

Fig. 1: A flowchart of the methodology for recommending keyword sets K_{in}, K_{ex} for inclusion/exclusion in content metadata combining qualitative methods, domain expertise & AI, statistics, and algorithms for keyword derivation, extraction, analysis and selection each.

2 Methods

To address the need of supported automated keyword selection for learning materials, previous work by Malek et al. [3] on metadata derivation, and conventional methods for keyword extraction and agreement analysis were combined with a novel approach to rater-sensible keyword selection (cf. Figure 1).

Keyword Derivation and Extraction

Collecting or deriving learning content titles and descriptions is oftentimes simple, whereas domain-specific keywords and their definitions regularly lack. In the context of the joint MINDED.Ruhr project work, Malek et al. tackled this problem by analysing interprofessional educational needs on dealing with people with dementia, resulting in five exemplary topics to create learning materials for: "Principles of Dementia" (DEF), "Forms of Dementia: Definition and Symptoms of Alzheimer's" (ALZ), "Behavioural changes of people with dementia" (BEH), "Principles of Communication with people with dementia" (COM) and "Communication techniques: The ABC-Method" (ABC). Titles were derived as combinations of topic and presentation mode, e.g. "Principles of Dementia: Video". Furthermore, 38 domain-specific keywords, together with their definitions were derived from their study and the definitions from the German Duden dictionary.

For keyword extraction, we chose *gpt-3.5-turbo-0125* (GPT) by OpenAI as it is the most wide-spead model, and compared its results with the open-source models *WizardLM-13B-V1.2* (WIZ) and *Mixtral-8x7B-Instruct-v0.1* (MIX) [4, 5], each selected for their benchmark performance and multilingual capabilities, as many other open-source models (e.g. those based on the LLAMA 2 foundation) are not fluent in German. All models also allow for general, free-of-charge access via API, removing the need for expensive hardware. Our designed prompt states the task of keyword extraction to each LLM, and is engineered to present each model with a list of Malek et al.'s predefined keywords, an exemplary output, together with content titles and description. The resulting query has roughly 2000 tokens (1500 words) on average, meaning no model's context length was exceeded. If this were the case, the content could be split into chunks, and resulting keywords for each chunk could then be collected in a set of keywords for the entire document. To further analyse the effects of contextualising the task of keyword extraction with keyword definitions via Retrieval Augmented Generation (RAG), we then used the FAISS [6] retrieval method and the OpenAI embedding model *text-embedding-ada-002* to create a vector database with all keyword definitions, enabling the LLMs to regard the keyword definitions for extraction. The following prompt (translated into English) was used for all LLMs (with and without RAG) twice/for two different titles to later assess each model's intra-rater reliability before comparing results to human raters.

Prompt (also used as instruction for human raters)

You are an expert in the field of dementia. You have developed educational materials for further training. Your task is to extract the relevant keywords from these educational materials. Only the following keywords may be used: {relevant keywords}
The output is a list of keywords relevant to the educational material.
Example output: ["Candidate1", "Candidate8", "Candidate5"]
#Used content metadata
Title of the educational material: {content title}
Content of the learning material: {content description}

#The following is only used for human raters and LLMs using RAG
Here are the relevant keyword definitions: {keyword definitions}

To compare LLM-based keyword extraction with human performance, we presented the prompt as instruction to five independent domain-experts for each topic in two runs two months apart. In the end, each extracted keyword was coded in a binary matrix X consisting of $K = 10 \times 38$ keyword codes by $R = 5 + 6$ human and LLM raters (cf. Table 1 for an exemplary layout of the coding data X for two keywords).

Tab. 1: Exemplary layout of the coding data X for one descriptive text, two keywords and each two human/artificial raters. If rater r extracted keyword k from the text, then $X_{k,r} = 1$ (0 otherwise).

	Human Raters		Artificial Raters	
Keywords	**Expert A**	**Expert B**	**LLM A**	**LLM B**
Symptoms	1	0	0	0
Communication	1	1	0	1

Keyword Analysis and Selection

Analysing intra-rater and inter-rater agreement is at the heart of the eventual keyword selection for content metadata, and will be measured using Krippendorff's α statistic [7]:

$$\alpha(X) = \frac{p_a(X) - p_e(w_{r_1,r_2})}{1 - p_e(w_{r_1,r_2})}, \quad w_{r_1,r_2} = \begin{cases} 1 & \textit{if } r_1, r_2 \textit{ agree} \\ 0 & \textit{otherwise.} \end{cases}$$

In this definition, $p_a(X)$ refers to the percent agreement observed in the coding data X. To accurately adjust for the scaling of the data (here: nominal/binary), the weight function w_{r_1,r_2} for the percent agreement by chance $p_e(w_{r_1,r_2})$ needs to be chosen accordingly (cf. [7] for further weight functions w_{r_1,r_2}). As the theoretical distribution of α is generally undefined [8], we used the bootstrapping method with $n = 1.000$ data sets $\{X_1, \dots X_{1000}\}$ to estimate the lower boundary of the 95% confidence interval (referred to as α_5) as the fifth percentile of all then computed set of statistics $\{\alpha(X_1), \dots \alpha(X_{1000})\}$. In addition, as α is a correlation coefficient of the raters' assessment on existent keywords in a content description, its squared α^2 poses an estimate of the percentage of reliably extractable keywords (see [9]; Table 3 for the extensive argument laid out for another agreement statistic). Now assuming that $\alpha_5 > 0$, α_5^2 describes an estimate of reliably selectable keywords with 95% certainty sensitive to the presented content and used extractors. Combining this percentage estimate with percent agreement in the coding data X, we designed Algorithm 1 for enhanced keyword selection[1]. In this, K_{in} and K_{ex} describe keyword sets to include in and exclude from the content metadata. The exclusion of keywords via K_{ex} is necessary, as the raters' strong agreement on their absence in the treated content is reflected in $\alpha(X)$.

Algorithm 1 Computing $K_{in}; K_{ex}$: keywords to in-/exclude

Require: Coding data X: *K keywords* $\times$ *N raters*
1: Create $\{X_1, \dots X_n\}$: *n bootstrapped data sets*
2: Compute α_5: *5th percentile of* $\{\alpha(X_1), \dots \alpha(X_n)\}$
3: Estimate α_5^2: % of reliable keywords in X
4: $K_{in} \leftarrow \emptyset; K_{ex} \leftarrow \emptyset$
5: $perc \leftarrow 0;\ i \leftarrow N$
6: **while** $perc \leq \alpha_5^2$ **do**
7: $\quad K_{in}(i) \leftarrow \{k\ keyword \mid$ *i ratings: «k in content»*$\}$
8: $\quad K_{ex}(i) \leftarrow \{k\ keyword \mid$ *i ratings: «k not in content»*$\}$
9: $\quad K_{in} \leftarrow K_{in} \cup K_{in}(i); K_{ex} \leftarrow K_{ex} \cup K_{ex}(i)$
10: $\quad perc \leftarrow perc + |K_{in}(i) \cup K_{ex}(i)|/K;\ i \leftarrow i - 1$
11: **end while**

1 If $perc > \alpha_5^2$ with the last loop-iteration, no keyword is removed as all are equally likely to (not) appear in the content.

3 Results

In order to assess the degree to which LLMs can support keyword selection for metadata enhancement, we first investigated whether RAG and the endeavour of adding keyword definitions have an effect on keyword extraction. Then, we analysed the models' abilities for consistently extracting keywords via Krippendorff's α and contextualised these values with those of the human raters. Last, we selected the two most promising LLM candidates for keyword selection via Algorithm1 and compared their extracted keyword sets K_{in}, K_{ex} with those sets connoted with the human raters to check for agreement between the two rater categories.

Tab. 2: Topicwise analysis of intra-rater agreement of two LLM keyword extractions (Tables 2(a)-(b)), and topicwise analysis of inter-rater agreement for human experts(Table 2(c)).

	GPT	**GPT**$_{RAG}$	**WIZ**	**WIZ**$_{RAG}$	**MIX**	**MIX**$_{RAG}$
fails	0	1	10	6	3	7
keywords	78	142	93	45	98	39

(a) Overall failed runs and extracted keywords for two attempts on keyword extraction per LLM. **Fails** are runs having extracted hallucinated keywords, no keywords and/or an initial sublist of the prompt's keywords in both runs. For **Keywords**, hallucinations were not counted.

STAT	**RATER**	**DEF**	**ALZ**	**BEH**	**COM**	**ABC**
$\alpha(X_t)$	**GPT**	0.61	0.94	0.34	0.52	0.56
	GPT$_{RAG}$	0.45	0.54	0.35	0.09	0.65

(b) Topicwise intra-rater agreement $\bar{\alpha}(X_t)$ for **GPT** and **GPT**$_{RAG}$ for two runs of keyword extractions. Due to **failed runs** in for every topic, the agreement for **WIZ**, **WIZ**$_{RAG}$, **MIX**, and **MIX**$_{RAG}$ could not be computed.

STAT	**RATER**	**DEF**	**ALZ**	**BEH**	**COM**	**ABC**
$\alpha(X_t)$	**Human**	0.74	0.56	0.54	0.47	0.65
	Human*	0.83	0.71	0.55	0.86	0.45

(c) Inter-Rater agreement on extracted keywords from content descriptions for all human raters in each topic (**Human**). For comparison, **Human*** states the humans' agreement on extracted keywords for the actual content.

When observing the count of generated keywords, together with the failed keyword selection runs in Table 2a, **WIZ** and **MIX** fail to execute the task consistently, as they either hallucinate, misunderstand the prompt or give no answer at all. **GPT**, as an exception, generates keywords as present in the provided list (and as the prompt explicitly demanded the LLMs to do) with nearly twice the keywords when using RAG, and only failing the task once with the hallucinated keyword being the term "alzheimer disease". Even though this synonym for "alzheimers" (present keyword in the provided list) is correct for the context, this is still counted as an hallucination as no keyword extraction for this term has happened.

Regarding the many failed keyword extractions of **WIZ** and **MIX**, Algorithm 1 was only further applied for the keywords extracted by **GPT** and **GPT**$_{RAG}$. Table 2b thereby shows that the intra-rater agreement on the presence and absence of specific keywords is far from 1, stressing the non-deterministic output of LLMs and their inability to reliably extract keyword ad-hoc. In contrast, Table 2c shows a significantly higher agreement of human raters across all topics. To finally assess the LLMs' capabilities to support humans in keyword selection, the keyword sets K_{in}, K_{ex} reliably extracted for **GPT**, **GPT**$_{RAG}$ and **Human** were computed for each topic by applying Algorithm 1 to the extracted keyword sets.

Tab. 3: Topic-wise percent agreement $p_a(K_{in}), p_a(K_{ex})$ of **GPT** and **GPT**$_{RAG}$ with **Human** on selecting or excluding keywords. **Example**: For $p_a(K_{ex})$, **GPT** and **DEF**, 30/31 keywords selected to exclude based on the human codes are part of **GPT**'s K_{ex} list.

STAT	RATER	DEF	ALZ	BEH	COM	ABC
$p_a(K_{in})$	**GPT**	3/3	4/4	1/11	4/8	2/4
	GPT$_{RAG}$	1/3	1/4	5/11	6/8	3/4
$p_a(K_{ex})$	**GPT**	30/31	30/30	2/3	16/16	13/13
	GPT$_{RAG}$	26/31	24/30	3/3	6/16	13/13

Viewing Table 3, reliable keyword extraction based on **GPT** might better mimic human keyword selection as percent agreement is higher. Still, in every studied scenario, there were also further keywords extracted for both LLMs but not present in the respective human-code-based keyword set (median of 3 extra keywords for both LLM and K_{in} and K_{ex} respectively). Hence, a full replacement of human extraction by reliable, LLM-based extraction is still not feasible. Surprisingly, using RAG lead to considerably more keywords recommended for inclusion (11 to 5.6 on total average), and considerably less recommended keywords for exclusion from the data (20.8 to 28 on total average), even though the total number of recommendations is lower due to lower $\alpha(X_t)$ values. This, in turn, might be advantageous in the absence of human codes, as more reliable recommendations for keyword inclusion allows for a more efficient metadata enhancement. So, whereas the general internal agreement for **GPT**$_{RAG}$ is lower, a higher total of reliably selected keywords for inclusion were found.

4 Conclusion

To conclude this investigation, the four-step methodology for reliable keyword extraction laid out in Figure 1 and analysed in Tables 2-3 succeeds in giving content creators in digital education perspective on the reliability of open-source LLMs for automatic support of keyword selection based on their individual contents' title and description. However, the proposed methodology is limited by the assumptions of already existent, domain-specific sets of applicable keywords, and content metadata for curating the pre-defined prompt. Hence, future research should focus on improving the methodology via improved RAG and enhancing Algorithm 1. Still, comparing the keyword sets for human and LLM raters, it showed that reliable keyword extraction via **GPT** and Algorithm 1 can be used effectively in proposing suitable keywords, ultimately adressing the *cold-start problem* often faced with recommender systems and novel interactive systems. To conclude, reliable keyword selection shows to be flexible enough for general use in many further domains.

Acknowledgment: This study is financially funded by the Federal Ministry of Education and Research (BMBF), grant number 16DKZ1008, and supported by the Federal Institute for Vocational Training (BIBB), grant number 21INVI0301 (MINDED.Ruhr) and 21INVI2705 (KAINE). We thank all partners for their support.

References

[1] I. Reichow, K. Buntins, B. Paaßen, H. Abu-Rasheed, C. Weber, and M. Dornhöfer. *Recommendersysteme in der beruflichen Weiterbildung. Grundlagen, Herausforderungen und Handlungsempfehlungen. Ein Dossier im Rahmen des INVITE-Wettbewerbs.* Verlag, Berlin, 2022.

[2] S. K. Bharti and K. S. Babu. Automatic keyword extraction for text summarization: A survey. 2017.

[3] M. Malek, J. Nitsche, C. Dinand, J. P. Ehlers, V. Lissek, P. Böhm, E. M. Derksen, and M. Halek. Interprofessional needs analysis and user-centred prototype evaluation as a foundation for building individualized digital education in dementia healthcare supported by artificial intelligence: A study protocol. *Healthcare*, 05 2023.

[4] C. Xu, Q. Sun, K. Zheng, X. Geng, P. Zhao, J. Feng, C. Tao, and D. Jiang. Wizardlm: Empowering large language models to follow complex instructions. 04 2023.

[5] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, and C. Bamford. Mixtral of experts. 2024.

[6] J. Johnson, M. Douze, and H. Jégou. Billion-scale similarity search with gpus. 2017.

[7] K. Gwet. On krippendorff's alpha coefficient. 10 2015.

[8] A. Zapf, S. Castell, L. Morawietz, and A. Karch. Measuring inter-rater reliability for nominal data - which coefficients and confidence intervals are appropriate? *BMC Medical Research Methodology*, 16, 08 2016.

[9] M. McHugh. Interrater reliability: The kappa statistic. *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB*, 22:276–82, 10 2012.