Injecting Knowledge from Social Science Journals to Improve Indonesian Cultural Understanding by LLMs

Anonymous ACL submission

Abstract

Recently there have been intensifying efforts to improve the understanding of Indonesian cultures by large language models (LLMs). An attractive source of cultural knowledge that has been largely overlooked is local journals of social science, which likely contain substantial cultural studies from a native perspective. We present a novel text dataset of journal article passages, created from 151 open-source Indonesian social science journals, called IndoSoSci. We demonstrate an effective recipe for injecting Indonesian cultural knowledge therein into LLMs: extracting the facts related to Indonesian culture, and apply retrieval-augmented generation (RAG) with LLM-generated hypothetical documents as queries during retrieval. The proposed recipe yields strong performance gains over several strong baselines on the IndoCulture benchmark. Additionally, by combining IndoSoSci with Indonesian Wikipedia, we set a new state-of-the-art accuracy on the IndoCulture benchmark.

1 Introduction

Most large language models today are trained with text in predominantly Western languages, which may have created a Western bias in these models (Cao et al., 2023; Adilazuarda et al., 2024; Lovenia et al., 2024; Pawar et al., 2025). When interacting with users from underrepresented regions such as South-East Asia (SEA), the LLMs may generate responses that are insensitive, irrelevant, or otherwise premised on Western cultural norms. Additionally, the Western bias presents the risk of flattening global cultural diversity. Therefore, improving the cultural awareness and understanding of LLMs has gained increasing research attention.

As the fourth populous country in the world, Indonesia has been historically under-represented in NLP research (Aji et al., 2022). Indonesia is also one of the most ethnically and culturally diverse countries, with 600 to 1200 ethnic groups in the country, depending on the classification method (BPS, 2024). In recent years, there has been an intensifying effort to improve the availability of NLP resources for Indonesia. This includes development of benchmarks, such as Koto et al. (2023) and Koto et al. (2024), as well as a consolidation and standardization of disparate Indonesian datasets, as part of the SEACrowd initiative to facilitate usage of the datasets in research (Lovenia et al., 2024).

One attractive source of cultural knowledge that has been largely overlooked is social science[1] publications produced locally, which likely contain studies into local cultures from a native perspective. We present a novel text dataset of journal article passages, IndoSoSci, which is created from Indonesian social science journals indexed in the Directory of Open Access Journals[2], and demonstrate its effectiveness on the Indonesian cultural benchmark, IndoCulture (Koto et al., 2024).

On top of the dataset, the present study devises an effective technique to inject the cultural knowledge into LLMs. Inspired by previous research that retrieval may be more suitable for injecting specialized knowledge into LLMs than finetuning (Ovadia et al., 2024; Soudani et al., 2024), we propose to employ IndoSoSci in retrieval-augmented generation (RAG). First, we extract the factual statements regarding Indonesian culture in the journal articles. This is to prevent other types of text in the articles from interfering with the RAG process. During retrieval, for each question the LLM is prompted to generate a hypothetical answer, which is then used as an informative key for retrieval.

We evaluate on the IndoCulture benchmark (Koto et al., 2024), which covers diverse cultures in eleven Indonesian provinces. The proposed recipe results in strong performance gains over baselines. The best model achieves an accuracy of 79.5%, 3.1

[1] For brevity, in this paper "social science" refers to both humanities and social sciences.

[2] https://doaj.org/

percentage points higher than the previous SOTA. We carefully verify the effectiveness of all components of the proposed recipe using ablation experiments.

The contributions of this paper are as follows:

1. We present a novel text dataset created from carefully parsed Indonesian social science publications, IndoSoSci, which contains Indonesian cultural knowledge.
2. We demonstrate a technique for injecting Indonesian cultural knowledge from the journal articles into the LLMs. Facts extracted from the papers serves as values to be retrieved whereas hypothetical documents generated from the target question serve as the query.
3. We set a new SOTA of 81.4% on IndoCulture by using RAG with a mixed corpus of Indonesian Wikipedia and extracted facts from journals, highlighting the potential of our corpus in complementing more common sources of knowledge.

2 Related Work

NLP Corpora of Academic Papers. Academic publications in science and engineering fields have been included in NLP corpora. For example, papers from ArXiv and PubMed have been included in open-source datasets such as The Pile (Gao et al., 2020). Beyond the STEM fields, S2ORC (Lo et al., 2020) covered more academic disciplines by collecting papers from Semantic Scholar. A small minority of papers contained in S2ORC are from social studies subjects such as Sociology, History, and Art; however, the dataset is limited to English-language papers.

OpenMSD (Gao et al., 2024a) is a multilingual dataset of scientific papers, including papers from Southeast Asia (SEA). Nevertheless, the dataset is primarily intended for scientific document similarity measurement, and the data distribution is still heavily skewed towards STEM subjects. To our knowledge, there have been no previous corpora that focus on social science journal articles designed for the task of cultural understanding in Southeast Asia.

Retrieval-Augmented Generation. Pioneered by Lewis et al. (2020), retrieval-augmented generation (RAG) is a technique for augmenting the internal knowledge of an LLM by retrieving documents from an external database and placing the retrieved documents in the LLM decision context. Current implementation of RAG generally involves three steps: indexing, retrieval, and generation (Gao et al., 2024b). In the indexing step, text chunks are encoded into vector representations using an embedding model, and the vector representations are collected into a vector database. In the retrieval step, the same embedding model is used to encode a user query, and the similarity scores between the query vector and the vectors in the database are calculated. A predefined number D of documents with the highest similarity scores is subsequently added to the prompt that will be given to the LLM, as context to the user query. In the generation step, the LLM is instructed to answer the new prompt containing both the context and the original query.

Much research has been conducted on RAG since the technique's introduction. There has been research into improving the retrieval stage (Gao et al., 2023; Zhu et al., 2025b; Laitenberger et al., 2025), instruction finetuning of the LLM to make more effective use of the retrieved documents (Liu et al., 2025; Bhushan et al., 2025), and develop new evaluation frameworks (Zhu et al., 2025a). Other works investigate interleaving RAG with multi-step reasoning to improve the LLM's reasoning capability (Li et al., 2025a; Jiang et al., 2025), and apply RAG to improve LLM performance in various domains (Li et al., 2025b; Wu et al., 2025).

RAG for LLM Cultural Understanding. There have been few studies into the use of RAG in cultural context. In the review of Pawar et al. (2025) regarding prior works on LLM cultural awareness, it is reported that training-free methods to improve LLM cultural awareness have historically focused on prompting techniques, such asking the LLM to adopt personas belonging to a particular cultural group or sociodemographic background (for example AlKhamissi et al. (2024); Cheng et al. (2023)).

More recently, Utami et al. (2025) utilized RAG in a chatbot for mental health of Aboriginal mothers in Australia. Closer to the present work, Lee et al. (2025) developed a benchmark for Hakka culture, intended to test an LLM's capability across the six aspects of Bloom's Taxonomy: remembering, understanding, applying, analyzing, evaluating, and creating. They showed that RAG with a corpus constructed primarily from Hakka-language Wikipedia leads to higher performance on their benchmark over a no-retrieval baseline. Nevertheless, their study is focused on creating their bench-

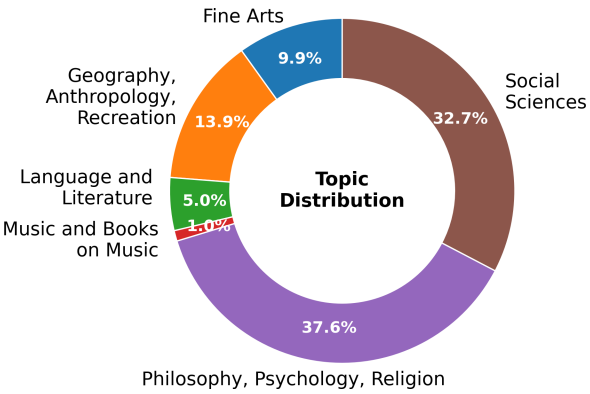

Figure 1: The proportions of social science topics covered by the IndoSoSci dataset.

Figure 2: A word cloud of the frequent phrases in the academic text extracted from the regions labeled with main title, section title, abstract, text, and list.

mark instead of developing a dataset for improving the LLM performance, and they focused a single cultural group. Our work is intended to cover diverse cultural groups within Indonesia.

Computational Understanding of Indonesian Culture. Early works on this topic focuses on LLMs 'understanding of Indonesian language (Wilie et al., 2020; Mahendra et al., 2021). More recently, research has focused more on LLM's reasoning ability related to Indonesian culture. New benchmarks include testing the LLMs on Indonesian exam questions from primary school to university entrance levels (Koto et al., 2023), Indonesian terminology, language nuances, and culture of Jakarta (Wibowo et al., 2024), as well as on human- and LLM- generated questions about general Indonesian culture (Putri et al., 2024). The current most comprehensive benchmark, IndoCulture, evaluates LLMs' understanding of Indonesian culture across eleven provinces, hence capturing the regional cultural diversity (Koto et al., 2024).

Figure 3: A word cloud of the frequent phrases in the text of the cultural facts extracted.

3 Methodology

3.1 Creating the IndoSoSci Dataset

We crawled Indonesian journals with Creative Commons licenses[3] indexed in the Directory of Open Access Journals. From January to February 2025, we downloaded the pdf articles from all available online issues, yielding a total of 21,500 pdf files. The topic of journals, covering various social science topics, such as anthropology, ethics, and linguistic theory. We show a donut chart of the main category proportions in Figure 1 and leave the complete ontology of topics to Appendix G.

To facilitate downstream application of the dataset, we need to convert the collected pdf articles into plain text. However, one challenge we face is the complex layouts of academic publications, containing headers, footnotes, diagrams, tables, and single- or double-column text; simply extracting text line by line will mix text from different regions and disrupt the semantic meaning. To prevent this, we computationally identify the page layout, dividing each page into text regions and classifying them by function.

Empirically, we find that off-the-shelf page layout detection systems to be insufficient for our purposes, as their label set is not designed for social science publications and they are not trained on Indonesian text. As a result, we redesigned the label space, annotated our own training data, and finetuned the network. The new label space contains four region labels from PubLayNet (Zhong et al., 2019): text, list, table, and figure, as well as four additional labels that we created: main title, section title, abstract, and caption.

We finetune an object detection network of LayoutLMv3 (Huang et al., 2022). The network employs the LayoutLM backbone (Xu et al., 2020),

[3]CC-BY, CC-BY-SA, CC-BY-NC, and CC-BY-NC-SA

which is a BERT-like text-vision bimodal Transformer, and a feature pyramid network (Lin et al., 2017) for feature extraction, and Cascade R-CNN (Cai and Vasconcelos, 2018) for region detection. The network is trained on medical publications in English from PubMed Central, creating distribution shift from Indonesian social science publications. Thus, we manually annotated the layout of 500 pages from IndoSoSci, including bounding boxes and classes of each box, and finetuned the Cascade R-CNN classification and regression heads accordingly. After finetuning, we attain a mean average precision (mAP) of 91.8%.

For our purposes, we keep text from the following detected regional bounding boxes: main title, abstract, text, section title, or list were selected. We extract the text from the bounding boxes using the library PyMuPDF[4] and assemble them in the same order as they appear in the pdf file. To exclude the bibliography, we remove text after the section title "bibliography", "references", or their equivalent in Bahasa Indonesia. This yields about 212 million tokens from 21,374 articles from 151 journals.

3.2 Cultural Facts Extraction from Academic Text

For computational understanding of Indonesian cultural practices, we are primarily interested in facts widely recognized among social scientists. However, social science publications often contain idiosyncratic opinions of authors or novel insights that have not yet reached consensus. During retrieval and generation, the existence of those text may mislead the RAG system. Therefore, we use an LLM (Sailor2-20B-Chat) to extract the facts related to Indonesian culture from the journal text. The LLM prompt used and an example output can be found in Appendices B and C, respectively.

Before fact extraction, we split the academic text into approximately 650,000 chunks of roughly three paragraphs each. All facts extracted from one chunk are merged together, forming one textual entry. The resulting dataset contains approximately 102,000 entries of Indonesian cultural facts and a total of 15 million tokens. The token yield ratio is 7.1%. The relatively low ratio stems from the fact that much academic text does not describe cultural facts. For example, some journal articles may discuss statistical procedure at length.

We visualize the text before and after the fact extraction step using word clouds in Fig. 2 and 3. Before fact extraction, the most frequent phrases are generic expressions like "orang tua" (parent) and "laki laki" (man). After fact extraction, phrases like "kearifan lokal" (local wisdom) and "nilai nilai" (values) become more frequent, suggesting we indeed capture local cultural values and practices.

3.3 RAG with Hypothetical Documents

We now describe the RAG pipeline. As the result of the fact extraction step, we have access to a number of textual entries regarding Indonesian culture. Given a question about Indonesian culture and a few answer choices, we retrieve at the level of textual entries. We recognize the existence of a distributional gap between the question, and the facts that can be used to answer the question. Thus, instead of using the question as the query, we use as the query a hypothetical document that is more similar to the factual statements to be retrieved.

More specifically, we prompt the LLM being tested to generate a synthetic document that might provide the answer. Note that the synthetic document is not used to answer the question, only to retrieve relevant facts. Therefore, we do not expect the synthetic document to be factually correct, only that it is distributionally similar to the correct factual entry that we want to retrieve. After that, we apply an embedding model to convert the synthetic document to a query vector. We then retrieve the textual entries with the highest cosine similarity and place them in the LLM context, from which the LLM answers the target question.

4 Experiments

In this section we present two sets of experiments:

1. RAG with our corpus of extracted facts from social science journals
2. RAG with a mixed corpus of Indonesian Wikipedia + extracted facts from social science journals

For each set of experiments we present the main results of RAG performance on the IndoCulture benchmark, the ablation studies, and additional experiments.

4.1 Setup

Indonesian Culture Benchmark. Our proposed method was tested on the IndoCulture benchmark (Koto et al., 2024). This commonsense reasoning

[4] https://github.com/pymupdf/PyMuPDF

benchmark contains 2,429 questions designed to test an LLM's understanding of various cultural topics, ranging from food to religious holidays, across eleven Indonesian provinces. IndoCulture is currently the most comprehensive benchmark on Indonesian culture. The multiple-choice question (MCQ) format with province context was chosen our experiments. The prompt format used is provided in Appendix E.

MCQ Evaluation. Following Koto et al. (2023) and Koto et al. (2024), for each question in IndoCulture we obtain the probabilities for the first generated token and select the probabilities that correspond to the answer choices (A, B, C). The answer choice with the highest probability is taken as the model's answer for that question.

LLMs Employed. We applied our proposed method on recent models that are specifically developed for Southeast Asian languages, including SeaLLMs-v3 (Zhang et al., 2025), Sailor2 (Dou et al., 2025), and SEA-LION v4 (Ng et al., 2025). We experimented with both the base pretrained models and finetuned chat models where available.

The state-of-the-art (SOTA) performance reported previously for IndoCulture is 76.4, using Sailor2-20B model (Dou et al., 2025). The regular versions of Sailor2 models have relatively short context lengths of 4096; to accommodate the large number of tokens from all the retrieved passages, for Sailor2 models, we conducted RAG experiments with the long-context variants.

RAG Details. The raw journal articles were chunked using the recursive text splitter from LangChain[5] with a chunk size of 1600. To encode texts into vector representations, BGE-M3 (Chen et al., 2024) was chosen as the embedding model due to its multilingual capabilities. The resulting vector embeddings from the text chunks are indexed using FAISS on GPU (Johnson et al., 2019).

In generating both the facts from journal article chunks and the hypothetical documents that may answer the benchmark question, we used a temperature of 0.5, top-p sampling with p = 0.9, and no top-k sampling. The vLLM library (Kwon et al., 2023) was used in both generation tasks for efficiency.

[5] https://docs.langchain.com/oss/python/integrations/splitters/recursive_text_splitter

Base LLM	No RAG	D=20
SEALLMs-v3-7B	54.6	61.3
SEALLMs- v3-7B-Chat	60.6	65.3
Sailor2-L-8B	64.2	74.5
Sailor2-L-8B-Chat	70.5	73.9
Sailor2-L-20B	72.1	75.7
Sailor2-L-20B-Chat	75.4	**79.3**
Qwen-SEA-LION-v4-32B-IT	70.9	75.5

Table 1: Zero-shot accuracy on IndoCulture using RAG with extracted facts and hypothetical document queries. The results in the No RAG column were obtained by directly prompting the model with the benchmark questions, without any additional context. We use 20 retrieved passages in the LLM decision context.

4.2 RAG Extracted Facts from Social Science Journals

Table 1 presents the performance of RAG with the corpus of extracted facts from social science journals. On all the LLMs tested, retrieval from IndoSoSci results in considerable performance improvement over the no-retrieval baseline. In particular, the best score of 79.5 achieved by Sailor2-L-20B-Chat with RAG is better than the previously reported SOTA of 76.4 (Dou et al., 2025).

The performance gain starts to be observed when the number of retrieved passages D is one. The improvement in performance generally increases with increasing D, although as noted by Ovadia et al. (2024) the optimal number of retrieved passages may be both model- and task-dependent. To avoid tuning the hyperparameter on the test set, we simply report the performance at D=20.

Ablation: Fact Extraction from Scientific Texts In this ablation study, we analyze the effectiveness of the fact extraction step. Table 2 demonstrates that for most models, RAG using the corpus of journal extracted facts yields better performance than RAG with the corpus of raw journal texts.

The observed performance gain suggests that presenting the cultural knowledge in as a collection of facts is indeed important. The training data for the tested models include Wikipedia articles (Zhang et al., 2025; Dou et al., 2025; Ng et al., 2025); the format of a Wikipedia article can be seen as a series of facts. As such, converting the academic style of the social science journal articles into a format the LLMs are already familiar with

Base case : RAG with journal extracted facts Ablation case : RAG with raw journal text chunks			
	Base	**Ablation**	**B-A**
SEALLMs-v3-7B	60.3	60.4	-0.1
SEALLMs- v3-7B-Chat	63.9	62.4	+1.5
Sailor2-L-8B	73.4	70.7	+2.7
Sailor2-L-8B-Chat	73.5	74.5	-1.0
Sailor2-L-20B	75.1	73.1	+2.0
Sailor2-L-20B-Chat	78.6	78.0	+0.6
Qwen-SEA-LION-v4-32B-IT	74.7	73.5	+1.2

Table 2: Average change in RAG performance on IndoCulture when using the corpus of journal extracted facts, over an ablation case of using the corpus of raw journal texts. The result shown for each model is the average of the numbers $D = \{1, 2, 5, 10, 12, 15, 20\}$ of retrieved passages.

Base case : RAG with hypothetical documents Ablation case : RAG with no hypothetical documents			
	Base	**Ablation**	**B-A**
SEALLMs-v3-7B	60.3	59.6	+0.7
SEALLMs- v3-7B-Chat	63.9	64.0	-0.1
Sailor2-L-8B	73.4	72.1	+1.3
Sailor2-L-8B-Chat	73.5	71.5	+2.0
Sailor2-L-20B	75.1	73.2	+1.9
Sailor2-L-20B-Chat	78.6	77.2	+1.4
Qwen-SEA-LION-v4-32B-IT	74.7	73.4	+1.3

Table 3: Average change in RAG performance on IndoCulture when using model-generated hypothetical documents as the retrieval queries, over an ablation case of using the IndoCulture benchmark questions as the queries.

could help the LLMs in utilizing the knowledge content.

Ablation: Hypothetical Documents as Query. We conducted an ablation study to investigate the impact of using the model-generated hypothetical answers as the retrieval queries. The results (Table 3) show that in most models tested, using hypothetical documents as the retrieval queries outperforms using the IndoCulture questions as the queries. This is in line with the result of Gao et al. (2023). The observed trend in Table 3 that hypothetical document generation is more applicable for the stronger models is also in line with their observation.

Ablation: Alternative Textual Units for Retrieval. The raw journal texts includes text chunks that contain no cultural knowledge, such as discussion of statistical procedures. That is why we specifically extract cultural facts from the journal text before using them in RAG. However, it is possible that the cultural fact extraction step is overly aggressive and remove necessary context for the facts, which may mislead the RAG system.

Base case : RAG with journal extracted facts Ablation case : with raw texts of the extracted facts			
	Base	**Ablation**	**B-A**
SEALLMs-v3-7B	60.3	60.4	-0.1
SEALLMs- v3-7B-Chat	63.9	63.4	+0.5
Sailor2-L-8B	73.4	71.0	+2.4
Sailor2-L-8B-Chat	73.5	74.2	-0.7
Sailor2-L-20B	75.1	73.3	+1.8
Sailor2-L-20B-Chat	78.6	78.5	+0.1
Qwen-SEA-LION-v4-32B-IT	74.7	73.9	+0.8

Table 4: Average change in RAG performance on IndoCulture when using the extracted facts from the journal text chunks as the retrieval corpus, over an ablation case of using the corresponding raw text chunks as the corpus.

In this ablation study, we try to retain the immediate context of the extracted cultural facts by keeping the entire raw text chunks around the facts. If a textual chunk does not contain any fact, it is discarded. We call the resulting corpus the filtered raw corpus. We present a comparison of RAG performance with the extracted facts corpus and this filtered raw corpus in Table 4. The observed result indicates that for most models, the additional context confuses more than it clarifies.

Continuing this line of inquiry, we investigate the relative importance of the extracted facts during the retrieval step and the generation step. In the retrieval step, it may be easier to discriminate the relevant passages from the less relevant ones when the embeddings are made from the extracted facts rather than the raw text chunks. Alternatively, in the generation step, the format of the passages added as context to the LLM prompt may be important. We conducted an experiment in which the embeddings of the extracted facts were used for similarity calculations with the hypothetical answers, but the passages added to the LLM context were the corresponding raw texts of the extracted facts.

From Table 5, for most models using the extracted facts for both embedding similarity calculation and as context passages still outperforms using the extracted facts only for embedding similarity

Base case : Ext. facts for retrieval and generation Ablation case : Ext. facts for retrieval only			
	Base	**Ablation**	**B-A**
SEALLMs-v3-7B	60.3	60.5	-0.2
SEALLMs- v3-7B-Chat	63.9	63.1	+0.8
Sailor2-L-8B	73.4	71.5	+1.9
Sailor2-L-8B-Chat	73.5	74.6	-1.1
Sailor2-L-20B	75.1	73.1	+2.0
Sailor2-L-20B-Chat	78.6	78.5	+0.1
Qwen-SEA-LION-v4-32B-IT	74.7	73.9	+0.8

Table 5: Average change in RAG performance on IndoCulture when the extracted facts are used for both the embedding similarity calculation and as passages added to LLM context, over an ablation case of using the raw text chunks of the extracted facts as the context passages.

Base LLM	**No RAG**	**D=20**
SEALLMs-v3-7B	54.6	64.1
SEALLMs- v3-7B-Chat	60.6	66.3
Sailor2-L-8B	64.2	74.3
Sailor2-L-8B-Chat	70.5	77.2
Sailor2-L-20B	72.1	78.0
Sailor2-L-20B-Chat	75.4	**81.4**
Qwen-SEA-LION-v4-32B-IT	70.9	79.5

Table 6: Zero-shot accuracy on IndoCulture using RAG on both cultural facts extracted from IndoSoSci and Indonesian Wikipedia. The results in the column labeled "No RAG" were obtained by directly prompting the model with the benchmark questions, without any additional context.

calculation. This indicates that using the extracted facts throughout the RAG pipeline is advantageous.

4.3 RAG with Mixture of Extracted Journal Facts and Wikipedia

To further explore the potential of our extracted facts corpus for RAG application, we propose to append our corpus to Indonesian Wikipedia text. We hypothesize that scholarly publications from social science journal may contain different kinds of knowledge that complement Wikipedia. For example, the knowledge contained in Wikipedia may be more widely known, while the journals may incorporate more knowledge about cultural minorities or ancient practices.

The Wikipedia corpus was created from a dump of Indonesian-language Wikipedia dated 20 August 2025. To improve the relevance of the articles for downstream retrieval application, the articles included in the corpus are restricted to those containing "Indonesia" or the name of an Indonesian province in the main text. The articles are chunked with the same settings as those used for the journal articles. The resulting corpus contains 184,000 passages and 103 million tokens. The chunked Wikipedia articles were combined with the extracted facts from social science journals, and the mixed corpus was indexed as a single vector database.

Main Result. The results of RAG with the mixed corpus shown in Table 6 show even stronger performance gains over the no-retrieval baseline. With the help of retrieval from the mixed corpus, the best results from four models outperform the previous SOTA of 76.4 on IndoCulture. The score of 81.4 obtained using Sailor2-L-20B-Chat sets a new SOTA for the benchmark.

Base case : RAG with Wikipedia + journal ext. facts Ablation case : RAG with Wikipedia only			
	Base	**Ablation**	**B-A**
SEALLMs-v3-7B	63.6	62.1	+1.5
SEALLMs- v3-7B-Chat	66.3	65.4	+0.9
Sailor2-L-8B	73.8	72.9	+0.9
Sailor2-L-8B-Chat	76.5	76.0	+0.5
Sailor2-L-20B	77.2	76.9	+0.3
Sailor2-L-20B-Chat	80.4	79.4	+1.0
Qwen-SEA-LION-v4-32B-IT	78.2	76.7	+1.5

Table 7: Average change in RAG performance on IndoCulture when using the mixed corpus of journal extracted facts and Wikipedia texts, over an ablation case of using only the Wikipedia texts.

Ablation: Effects of Journal Text. The goal of this ablation study is to evaluate the impact of adding our corpus of extracted facts to a corpus of Wikipedia texts.

Table 7 shows that RAG using the mixed corpus outperforms RAG using the corpus of only Wikipedia texts for all models tested. This observation suggests that our specialized corpus of extracted facts from social science journals can well complement a corpus created from common sources such as Wikipedia. Correspondingly, the observed results also support our hypothesis that social science journals may contain cultural knowledge that is distinct from that already captured in Wikipedia. Exactly how the knowledge content of

Case 1 : RAG with Wikipedia extracted facts Case 2: RAG with raw Wikipedia			
	Case 1	**Case 2**	**1 - 2**
SEALLMs-v3-7B	62.6	62.1	+0.5
SEALLMs- v3-7B-Chat	66.1	65.4	+0.7
Sailor2-L-8B	73.4	72.9	+0.5
Sailor2-L-8B-Chat	73.8	76.0	-2.2
Sailor2-L-20B	76.1	76.9	-0.8
Sailor2-L-20B-Chat	78.3	79.4	-1.1
Qwen-SEA-LION-v4-32B-IT	76.2	76.7	-0.5

Table 8: Average change in RAG performance on IndoCulture when using extracted facts from Wikipedia chunks as the retrieval corpus, over a case of using the raw Wikipedia text chunks.

the two sources differ is an interesting avenue for investigation in future research.

Ablation: Fact Extraction on Wikipedia Text. To test whether the fact extraction can help regardless of original text format, we apply the fact extraction prompt on Wikipedia text chunks. We conduct RAG experiments using a corpus of extracted facts from Wikipedia and using a corpus of raw Wikipedia texts.

Table 8 shows that for the stronger models, RAG using a corpus of extracted facts from Wikipedia leads to worse results than using the raw Wikipedia corpus. However, the weaker models such as SEALLMs-v3-7B and Sailor2-L-8B benefit from the additional fact extraction step.

A possible reason is that, as the result of its editing process, Wikipedia is already quite clean and contains mostly widely recognized facts. The stronger models are already capable of utilizing cultural knowledge from raw Wikipedia. Further trimming it down could lose contextual information or introduce errors. This result is similar to the finding of Laitenberger et al. (2025), who reported that retrieving original passages leads to better RAG performance than retrieving generated summaries.

This experiment therefore highlights the pertinence of applying the fact extraction step to our corpus of Indonesian social science journal articles. As an illustration, fact extraction results from two passages about traditional Indonesian snacks are shown in Appendix C (from journal article) and Appendix D (from Wikipedia article). In Appendix C, the argumentative style of the original journal passage regarding the history of the dish is converted into shorter factual statements regarding the origin and ingredients of the dish. In contrast, as shown in Appendix D, the extracted factual statement from the Wikipedia passage is remarkably similar to the original passage. Some information regarding the ingredients of the dish has also not been included in the extracted factual statement.

5 Conclusion

In this paper we explore the utilization of Indonesian social science journals to inject cultural knowledge into LLMs in the understanding of Indonesian culture. We present a novel text dataset of journal article passages, created from 151 open-source Indonesian social science journals. We use a strong LLM to extract facts related to Indonesian culture from the raw journal text passages. We subsequently use the resulting corpus of extracted facts for retrieval-augmented generation. We show that our proposed method results in strong performance gains over the no-retrieval baseline on the IndoCulture benchmark. Additionally, by combining our corpus with Indonesian Wikipedia, our best RAG performance on IndoCulture sets a new SOTA accuracy of 81.4%.

Limitations

The journal articles that we collected are written exclusively in Indonesian or English. Meanwhile, Indonesia has more than 700 spoken languages (Aji et al., 2022). As such, our journal corpus may not fully capture the richness of Indonesian cultural traditions.

Furthermore, this paper focuses on improving an LLM's knowledge of Indonesian cultural practices. We have not evaluated whether our method can allow an LLM to understand "deeper" aspects of culture, such as nuanced understanding of Indonesian language or culturally appropriate responses in conversational contexts.

References

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards Measuring and Modeling "Culture" in LLMs: A Survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Ma-

hendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249, Dublin, Ireland. Association for Computational Linguistics.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating Cultural Alignment of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Kushagra Bhushan, Yatin Nandwani, Dinesh Khandelwal, Sonam Gupta, Gaurav Pandey, Dinesh Raghu, and Sachindra Joshi. 2025. Systematic Knowledge Injection into Large Language Models via Diverse Augmentation for Domain-Specific RAG. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5922–5943, Albuquerque, New Mexico. Association for Computational Linguistics.

BPS. 2024. Profile of Ethnic Groups and Regional Language Diversity Results of the 2020 Population Census Long Form. `https://www.bps.go.id/en/publication/2024/12/12/6feb932e24186429686fb57b/profile-of-ethnic-groups-and-regional-language-diversity-results-of-the-2020-population-census-long-form.html`.

Zhaowei Cai and Nuno Vasconcelos. 2018. Cascade R-CNN: Delving Into High Quality Object Detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6154–6162.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia. Association for Computational Linguistics.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada. Association for Computational Linguistics.

Longxu Dou, Qian Liu, Fan Zhou, Changyu Chen, Zili Wang, Ziqi Jin, Zichen Liu, Tongyao Zhu, Cunxiao Du, Penghui Yang, Haonan Wang, Jiaheng Liu, Yongchi Zhao, Xiachong Feng, Xin Mao, Man Tsung Yeung, Kunat Pipatanakul, Fajri Koto, Min Si Thu, and 22 others. 2025. Sailor2: Sailing in South-East Asia with Inclusive Multilingual LLMs. *Preprint*, arXiv:2502.12982.

Kezia Elsty and Zayyini Nahdlah. 2020. PENELUSURAN SEJARAH, FILOSOFI DAN BUDAYA MAKAN KUE GEPLAK KHAS BETAWI. *Jurnal Pariwisata Pesona*, 5(2):69–75.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *Preprint*, arXiv:2101.00027.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise Zero-Shot Dense Retrieval without Relevance Labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777, Toronto, Canada. Association for Computational Linguistics.

Yang Gao, Ji Ma, Ivan Korotkov, Keith Hall, Dana Alon, and Donald Metzler. 2024a. OpenMSD: Towards Multilingual Scientific Documents Similarity Measurement. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12467–12480, Torino, Italia. ELRA and ICCL.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024b. Retrieval-Augmented Generation for Large Language Models: A Survey. *Preprint*, arXiv:2312.10997.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, pages 4083–4091, New York, NY, USA. Association for Computing Machinery.

Jinhao Jiang, Jiayi Chen, Junyi Li, Ruiyang Ren, Shijie Wang, Wayne Xin Zhao, Yang Song, and Tao Zhang. 2025. RAG-Star: Enhancing Deliberative Reasoning with Retrieval Augmented Verification and Refinement. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7064–7074, Albuquerque, New Mexico. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-Scale Similarity Search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023. Large Language Models Only Pass Primary School Exams in Indonesia: A Comprehensive Test on IndoMMLU. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12359–12374, Singapore. Association for Computational Linguistics.

Fajri Koto, Rahmad Mahendra, Nurul Aisyah, and Timothy Baldwin. 2024. IndoCulture: Exploring Geographically Influenced Cultural Commonsense Reasoning Across Eleven Indonesian Provinces. *Transactions of the Association for Computational Linguistics*, 12:1703–1719.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, pages 611–626, New York, NY, USA. Association for Computing Machinery.

Alex Laitenberger, Christopher D Manning, and Nelson F. Liu. 2025. Stronger Baselines for Retrieval-Augmented Generation with Long-Context Language Models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32547–32557, Suzhou, China. Association for Computational Linguistics.

Hung-Shin Lee, Chen-Chi Chang, Ching-Yuan Chen, and Yun-Hsiang Hsu. 2025. Evaluating cultural knowledge processing in large language models: A cognitive benchmarking framework integrating retrieval-augmented generation. *The Electronic Library*, pages 1–22.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, pages 9459–9474, Red Hook, NY, USA. Curran Associates Inc.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025a. Search-o1: Agentic Search-Enhanced Large Reasoning Models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5420–5438, Suzhou, China. Association for Computational Linguistics.

Yuyang Li, Pjm Kerbusch, Rhr Pruim, and Tobias Käfer. 2025b. Evaluating the Performance of RAG Methods for Conversational AI in the Airport Domain. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 794–808, Albuquerque, New Mexico. Association for Computational Linguistics.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125.

Wanlong Liu, Junying Chen, Ke Ji, Li Zhou, Wenyu Chen, and Benyou Wang. 2025. RAG-Instruct: Boosting LLMs with Diverse Retrieval-Augmented Instructions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3865–3888, Suzhou, China. Association for Computational Linguistics.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James V. Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno P. Kampman, Joel Ruben Antony Moniz, Muhammad Ravi Shulthan Habibi, Frederikus Hudi, Railey Montalan, Ryan Ignatius, Joanito Agili Lopo, William Nixon, Börje F. Karlsson, James Jaya, and 42 others. 2024. SEACrowd: A Multilingual Multimodal Data Hub and Benchmark Suite for Southeast Asian Languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5155–5203, Miami, Florida, USA. Association for Computational Linguistics.

Rahmad Mahendra, Alham Fikri Aji, Samuel Louvan, Fahrurrozi Rahman, and Clara Vania. 2021. IndoNLI: A Natural Language Inference Dataset for Indonesian. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10511–10527, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Raymond Ng, Thanh Ngan Nguyen, Yuli Huang, Ngee Chia Tai, Wai Yi Leong, Wei Qi Leong, Xianbin Yong, Jian Gang Ngui, Yosephine Susanto, Nicholas Cheng, Hamsawardhini Rengarajan, Peerat Limkonchotiwat, Adithya Venkatadri Hulagadri, Kok Wai Teng, Yeo Yeow Tong, Bryan Siow, Wei Yi Teo, Wayne Lau, Choon Meng Tan, and 12 others. 2025. SEA-LION: Southeast Asian Languages in One Network. *Preprint*, arXiv:2504.05747.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. 2024. Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 237–250, Miami, Florida, USA. Association for Computational Linguistics.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2025. Survey of Cultural Awareness in

Language Models: Text and Beyond. *Computational Linguistics*, 51(3):907–1004.

Rifki Afina Putri, Faiz Ghifari Haznitrama, Dea Adhista, and Alice Oh. 2024. Can LLM Generate Culturally Relevant Commonsense QA Data? Case Study in Indonesian and Sundanese. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20571–20590, Miami, Florida, USA. Association for Computational Linguistics.

Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. 2024. Fine Tuning vs. Retrieval Augmented Generation for Less Popular Knowledge. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 12–22, Tokyo Japan. ACM.

Made Srinitha Millinia Utami, Wai Hang Kwok, Jayne Kotz, Roz Walker, Guanjin Wang, and Rhonda Marriott. 2025. Facilitating Aboriginal Perinatal Mental Health Information Access with a Retrieval-Augmented LLM-based Chatbot. In *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–7.

Haryo Wibowo, Erland Fuadi, Made Nityasya, Radityo Eko Prasojo, and Alham Aji. 2024. COPAL-ID: Indonesian Language Reasoning with Local Culture and Nuances. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1404–1422, Mexico City, Mexico. Association for Computational Linguistics.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. 2020. IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843–857, Suzhou, China. Association for Computational Linguistics.

Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, Yueming Jin, and Vicente Grau. 2025. Medical Graph RAG: Evidence-based Medical Large Language Model via Graph Retrieval-Augmented Generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28443–28467, Vienna, Austria. Association for Computational Linguistics.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1192–1200.

Wenxuan Zhang, Hou Pong Chan, Yiran Zhao, Mahani Aljunied, Jianyu Wang, Chaoqun Liu, Yue Deng, Zhiqiang Hu, Weiwen Xu, Yew Ken Chia, Xin Li, and Lidong Bing. 2025. SeaLLMs 3: Open Foundation and Chat Multilingual Large Language Models for Southeast Asian Languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 96–105, Albuquerque, New Mexico. Association for Computational Linguistics.

Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. 2019. PubLayNet: Largest dataset ever for document layout analysis. *Preprint*, arXiv:1908.07836.

Kunlun Zhu, Yifan Luo, Dingling Xu, Yukun Yan, Zhenghao Liu, Shi Yu, Ruobing Wang, Shuo Wang, Yishan Li, Nan Zhang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2025a. RAGEval: Scenario Specific RAG Evaluation Dataset Generation Framework. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8520–8544, Vienna, Austria. Association for Computational Linguistics.

Xiangrong Zhu, Yuexiang Xie, Yi Liu, Yaliang Li, and Wei Hu. 2025b. Knowledge Graph-Guided Retrieval Augmented Generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8912–8924, Albuquerque, New Mexico. Association for Computational Linguistics.

A Risks

A cultural tradition associated with a cultural group may not be practiced by all members of that cultural group. Users of our corpus should keep this in mind to avoid stereotyping the members of a cultural group. Additionally, research findings published in the social science journals regarding particular cultural practices or social phenomena may be time-dependent. As such, future users of our dataset should take care to verify that such information are still applicable.

B Prompt for Fact Extraction

The following prompt is used to extract facts related to Indonesian culture from a chunk of journal article text. The text passage is placed in the [DOCUMENT] field.

Prompt for Fact Extraction

Extract all factual claims related to Indonesian culture from the following passage. Enclose your response within <factual_claims> and </factual_claims> tags. Write the factual claims in Indonesian. If you cannot find any factual claims related to Indonesian culture, write 'No relevant factual claims found'.

PASSAGE:

[DOCUMENT]

OUTPUT: <factual_claims></factual_claims>

C Example Fact Extraction Result from Journal

The following box provides an example of the resulting factual statements extracted from a raw text passage in our journal dataset. The text passage is taken from a paper by Elsty and Nahdlah (2020) about Kue Geplak Betawi, which is a traditional dish.

Example Fact Extraction Result from Journal

Original text passage:

Sejarah Kue Geplak Betawi

Bila dilihat dari berbagai pendekatan, setidaknya ada lima perspektif untuk mengenal asal Kue Geplak khas Betawi. Pendekatan pertama dapat dilihat dari asal kue ini tercipta. Saputra (2019) menjelaskan saat ini tidak ada dokumen tertulis dan tidak diketahui persisnya kapan kue ini tercipta. Namun keberadaan kue ini dapat dikaitkan dengan keberadaan ekosistem dengan segala hasil bumi di dalamnya. Makanan dan lingkungan merupakan hal yang tidak dipisahkan karena pembuatan makanan sebagian besar dari ketersedaan bahan pangan yang disediakan oleh lingkungan yang dalam hal ini disebut ekosistem pada suatu tempat di mana aktivitas keseharian dan kehidupan suatu masyarakat berlangsung (Untari, 2016). Ketersediaan bahan pada ekosistem yang ada mempengaruhi pemikiran nenek moyang untuk mengolah hasi bumi yang tercatat juga pada relief-relief Candi Borobudur di mana kegiatan terkait kuliner sudah ada sejak abad ke-8 yang ditandai dengan adanya relief kegiatan memasak di dapur atau masyarakat yang mengangkut padi. Kreatifitas masyarakat juga yang akhirnya melahirkan sebuah makanan khususnya kue. Asal Kue Geplak Betawi dapat ditelusuri dari membedah bahan-bahan dari kue ini salah satunya beras. Bahan lainnya adalah kelapa parut sangrai, larutan gula, daun pandan, dan daun jeruk. Kue Geplak Betawi ini dapat ditemui pada masyarakat Betawi pinggiran. Penggunaan bahan hasil bumi padi yang diolah menjadi kue Geplak Betawi juga menjadi upaya untuk menjadikan bahan tersebut untuk memiliki manfaat lebih yakni menjadi cemilan kue.

Extracted factual statements:

Kue Geplak adalah makanan khas Betawi.

Asal-usulnya terkait dengan ekosistem dan hasil bumi, terutama beras dan kelapa.

Kue Geplak Betawi menggunakan bahan seperti beras, kelapa parut sangrai, larutan gula, daun pandan, dan daun jeruk.

Kue ini ditemukan di masyarakat Betawi pinggiran, mencerminkan kreativitas lokal dalam pengolahan hasil bumi.

English translations of the original passage and the extracted facts are provided in the following box.

English translation from ChatGPT

Original text passage:

History of Betawi Geplak Cake

When viewed from various approaches, there are at least five perspectives for understanding the origins of Betawi-style Geplak cake. The first approach looks at how this cake came into existence. Saputra (2019) explains that, to date, there are no written documents, and it is not known exactly when this cake was first created. However, its existence can be linked to the presence of an ecosystem and all the natural resources within it. Food and the environment are inseparable, because food production largely depends on the availability of food ingredients provided by the environment, which in this context is referred to as the ecosystem of a place where daily activities and the life of a community take place (Untari, 2016). The availability of materials within an ecosystem influenced the thinking of ancestors in processing agricultural products, a fact also recorded in the reliefs of Borobudur Temple, where culinary-related activities have existed since the 8th century. This is indicated by reliefs depicting cooking activities in kitchens or people transporting rice. The creativity of the community ultimately gave rise to various foods, especially cakes. The origin of Betawi Geplak cake can be traced by examining its ingredients, one of which is rice. Other ingredients include toasted grated coconut, sugar syrup, pandan leaves, and kaffir lime leaves. Betawi Geplak cake can be found among Betawi communities living on the outskirts. The use of rice-based agricultural products processed into Betawi Geplak cake also represents an effort to give these ingredients added value by turning them into snack foods.

Extracted factual statements:

Geplak cake is a traditional Betawi food.

Its origin is connected to the ecosystem and natural resources, especially rice and coconut.

Betawi Geplak cake uses ingredients such as rice, toasted grated coconut, sugar syrup, pandan leaves, and kaffir lime leaves.

This cake is found among Betawi communities on the outskirts, reflecting local creativity in processing natural resources.

D Example Fact Extraction Result from Wikipedia

The following box provides an example of the resulting factual statements extracted from an Indonesian Wikipedia passage. The article is titled "Geplak", included a Wikipedia dump dated 20 August 2025. Geplak is distinct from Kue Geplak Betawi in Appendix C, although they share some characteristics.

Example Fact Extraction Result from Wikipedia

Original text passage:

Geplak adalah penganan yang dibuat dari adonan kelapa parut (ampas kelapa) dicampur gula dan vanili, ada yang dicampuri durian, sirsak, atau nangka.

Geplak merupakan penganan tradisional khas Jawa yang berasal dari kabupaten Bantul, Daerah Istimewa Yogyakarta. Terdapat pula geplak yang dibuat dari waluh. Industri geplak umumnya dapat ditemui di daerah Kabupaten Bantul, Daerah Istimewa Yogyakarta, yang kebanyakan diusahakan oleh industri rumah tangga. Selanjutnya jenis penganan ini berkembang meluas akibat permintaan pasar dan diusahakan tidak hanya di sekitar Daerah Istimewa Yogyakarta akan tetapi juga di seluruh Nusantara.

Extracted factual statements:

Geplak adalah penganan tradisional khas Jawa dari Kabupaten Bantul, Daerah Istimewa Yogyakarta. Industri geplak umumnya diusahakan oleh industri rumah tangga di Bantul dan telah berkembang ke seluruh Nusantara.

English Translation from ChatGPT

Original text passage:

Geplak is a snack made from grated coconut (coconut pulp) mixed with sugar and vanilla, and sometimes flavored with durian, soursop, or jackfruit.

Geplak is a traditional snack originally from Bantul Regency, Special Region of Yogyakarta, Java. There is also a version made with pumpkin. The geplak industry is mostly found in Bantul Regency, where it is commonly produced by home industries. Over time, this type of snack has spread widely due to market demand, and is now produced not only in the Special Region of Yogyakarta but also throughout the Indonesian archipelago.

Extracted factual statements:

Geplak is a traditional Javanese snack from Bantul Regency, Special Region of Yogyakarta. The geplak industry is mostly run by home-based businesses in Bantul and has since spread throughout the Indonesian archipelago.

E Prompt for IndoCulture Benchmark

The prompt used is the Indonesian MCQ prompt with province name as the location context, taken from the IndoCulture paper (Koto et al., 2024).

IndoCulture MCQ Prompt

Untuk konteks [PROVINCE], sambungan yang tepat dari kalimat "[PREMISE]" adalah
[OPTIONS]
Jawaban:

English translation:
Given [PROVINCE] context, the correct continuation of the sentence "[PREMISE]" is
[OPTIONS]
Answer:

F RAG-related prompts

The prompt used for our RAG experiments is as follows:

Prompt for RAG

INSTRUKSI: Jawablah SOAL di bawah ini dengan bantuan BACAAN di bawah ini.
[DOCUMENT]
SOAL
[QUESTION]

English translation:
INSTRUCTION: Answer the QUESTION below with the help of the PASSAGE below.
[DOCUMENT]
QUESTION
[QUESTION]

The [QUESTION] field is replaced with an IndoCulture MCQ prompt given in Appendix E. The [DOCUMENT] field is replaced by the passages that are retrieved from the external corpus. Each passage added to the prompt is formatted as follows:

BACAAN [DOC_NUM]:
[DOC_TEXT]

The following prompt is used to generate the hypothetical document that may answer a question from IndoCulture. The [QUESTION] field is replaced with an IndoCulture MCQ prompt.

Prompt for Hypothetical Document Generation

Write a passage in Indonesian language to answer the following question in detail.
QUESTION:
[QUESTION]
PASSAGE:

G Ontology of Journal Topics from Directory of Open Access Journals

- **Fine Arts**
- **Geography. Anthropology. Recreation**
  - Anthropology
  - Environmental sciences
  - Geography (General)
  - Recreation. Leisure
    - Dancing
- **Language and Literature**
  - Literature (General)
  - Philology, Linguistics
    - Language, Linguistic theory, Comparative grammar
- **Music and Books on Music**
- **Philosophy. Psychology. Religion**
  - Ethics
  - Religions. Mythology. Rationalism
- **Social Sciences**
  - Communities. Classes. Races
  - Social history and conditions. Social problems. Social reform
  - Social pathology. Social and public welfare. Criminology
  - Social sciences (General)
  - Social sciences and state - Asia (Asian studies only)
  - Sociology (General)
  - The family. Marriage. Woman

H Model Sources

Model	Source
SEALLMs-v3-7B	SeaLLMs/SeaLLMs-v3-7B
SEALLMs-v3-7B-Chat	SeaLLMs/SeaLLMs-v3-7B-Chat
Sailor2-L-8B	sail/Sailor2-L-8B
Sailor2-L-8B-Chat	sail/Sailor2-L-8B-Chat
Sailor2-L-20B	sail/Sailor2-L-20B
Sailor2-L-20B-Chat	sail/Sailor2-L-20B-Chat
Qwen-SEA_LION-v4-32B-IT	aisingapore/Qwen-SEA-LION-v4-32B-IT

Table 9: HuggingFace sources of the models tested in this study.

I Hardware and Time Details

Our experiments were conducted using Nvidia A100 GPUs. We used up to four GPUs for one evaluation run on IndoCulture. The time taken for one evaluation run depends on the model size and the number of documents retrieved for RAG. The time taken ranges from under one minute with a 7B model and no RAG, to around seven hours with a 32B models and 20 retrieved documents per question.

J List of Journals in the Dataset

Number	Journal Name
1	ANDHARUPA Jurnal Desain Komunikasi Visual & Multimedia
2	ARISTO
3	ARSNET
4	AT-TURAS Jurnal Studi Keislaman
5	Abdihaz Jurnal Ilmiah Pengabdian pada Masyarakat
6	Absorbent Mind
7	Academic Journal of Psychology and Counseling
8	Al-Mazaahib Jurnal Perbandingan Hukum
9	Al-Misykah Jurnal Studi Al-qur'an dan Tafsir
10	Analitika Jurnal Magister Psikologi UMA
11	Anthropos Jurnal Antropologi Sosial dan Budaya
12	Arsitekno
13	Arsitektura Jurnal Ilmiah Arsitektur dan Lingkungan Binaan
14	Az-Zahra Journal of Gender and Family Studies
15	Basastra
16	Biokultur
17	Brikolase Jurnal Kajian Teori, Praktik dan Wacana Seni Budaya Rupa
18	Buddayah Jurnal Pendidikan Antropologi
19	Buletin Psikologi
20	Buletin Riset Psikologi dan Kesehatan Mental (BRPKM)
21	Bulletin of Counseling and Psychotherapy
22	CaLLs (Journal of Culture, Arts, Literature, and Linguistics)
23	Dewa Ruci Jurnal Pengkajian dan Penciptaan Seni
24	Dinamisia Jurnal Pengabdian Kepada Masyarakat
25	EL-FIKR Jurnal Aqidah dan Filsafat Islam
26	ENLIGHTEN Jurnal Bimbingan Konseling Islam
27	ETHOS Jurnal Penelitian dan Pengabdian kepada Masyarakat
28	Edudeena Journal of Islamic Religious Education
29	El-Aqwal Journal of Sharia and Comparative Law
30	Engagement Jurnal Pengabdian Kepada Masyarakat
31	GEMA TEOLOGIKA Jurnal Teologi Kontekstual dan Filsafat Keilahian
32	GUIDENA Jurnal Ilmu Pendidikan, Psikologi, Bimbingan dan Konseling
33	Gadjah Mada Journal of Professional Psychology (GamaJPP)
34	Gadjah Mada Journal of Psychology (GamaJoP)
35	Gondang Jurnal Seni dan Budaya
36	Hanifiya Jurnal Studi Agama-Agama
37	Happiness Journal of Psychology and Islamic Science
38	Harmoni Sosial Jurnal Pendidikan IPS
39	Hayula Indonesian Journal of Multidisciplinary Islamic Studies

40 Hisbah Jurnal Bimbingan Konseling dan Dakwah Islam
41 Home Dynamics of Rural Society Journal
42 ICODEV Indonesian Community Development Journal
43 INFERENSI Jurnal Penelitian Sosial Keagamaan
44 INKLUSI
45 INSIGHT Jurnal Bimbingan Konseling
46 Ijtimā iyya Journal of Muslim Society Research
47 Imajinasi Jurnal Seni
48 Indonesian Journal of Earth Sciences
49 Indonesian Journal of Fundamental Sciences
50 Indonesian Journal of Religion and Society
51 Insight Jurnal Ilmiah Psikologi
52 International Journal Ihya' 'Ulum al-Din
53 International Journal Pedagogy of Social Studies
54 Islamic Counseling Jurnal Bimbingan Konseling Islam
55 JADECS (Journal of Art, Design, Art Education & Cultural Studies)
56 JAMBURA GEO EDUCATION JOURNAL
57 JAUR (JOURNAL OF ARCHITECTURE AND URBANISM RESEARCH)
58 JIP (Jurnal Intervensi Psikologi)
59 JOINS (Journal of Information System)
60 JSW (Jurnal Sosiologi Walisongo)
61 JURNAL GEOGRAFI
62 JURNAL PENELITIAN PENDIDIKAN, PSIKOLOGI DAN KESEHATAN (J-P3K)
63 JURNAL SOSIAL HUMANIORA (JSH)
64 Journal An-Nafs Kajian Penelitian Psikologi
65 Journal Fenomena
66 Journal Sampurasun
67 Journal of Community Service and Empowerment
68 Journal of Comparative Study of Religions
69 Journal of Indonesian Society Empowerment
70 Journal of Islamic Accounting and Finance Research
71 Jurnal Adabiyah
72 Jurnal Antropologi Isu-Isu Sosial Budaya
73 Jurnal Dakwah Risalah
74 Jurnal Diversita
75 Jurnal EDUCATIO Jurnal Pendidikan Indonesia
76 Jurnal Ekologi, Masyarakat dan Sains
77 Jurnal Humanitas Katalisator
78 Jurnal IPTA (Industri Perjalanan Wisata)
79 Jurnal Ilmiah Pendidikan Pancasila dan Kewarganegaraan
80 Jurnal Ilmiah Platax
81 Jurnal Kajian Seni
82 Jurnal Kawistara

83 Jurnal Layanan Masyarakat (Journal of Public Services)
84 Jurnal Litbang Provinsi Jawa Tengah
85 Jurnal Manusia dan Lingkungan
86 Jurnal Master Pariwisata (JUMPA)
87 Jurnal Pariwisata
88 Jurnal Pariwisata Pesona
89 Jurnal Pariwisata Terapan
90 Jurnal Pembangunan Wilayah dan Kota
91 Jurnal Pemberdayaan Masyarakat Madani (JPMM)
92 Jurnal Pemberdayaan Masyarakat Media Pemikiran dan Dakwah Pembangunan
93 Jurnal Psikoedukasi dan Konseling
94 Jurnal Psikogenesis
95 Jurnal Psikologi Integratif
96 Jurnal Psikologi Islam dan Budaya
97 Jurnal Psikologi Teori dan Terapan
98 Jurnal Psikologi Ulayat
99 Jurnal Riptek
100 Jurnal Sains Psikologi
101 Jurnal Sosiologi Andalas
102 Jurnal Sosiologi Pendidikan Humanis
103 Jurnal Sosiologi Reflektif
104 Jurnal Studi Agama
105 KAIBON ABHINAYA JURNAL PENGABDIAN MASYARAKAT
106 KLITIKA Jurnal Ilmiah Pendidikan Bahasa dan Sastra Indonesia
107 Kanz Philosophia A Journal for Islamic Philosophy and Mysticism
108 Khazanah Jurnal Studi Islam dan Humaniora
109 Kifah Jurnal Pengabdian Masyarakat
110 LINGUA Jurnal Bahasa, Sastra, dan Pengajarannya
111 Lamahu Jurnal Pengabdian Masyarakat Terintegrasi
112 Linguistika
113 MOZAIK HUMANIORA
114 MUHARRIK Jurnal Dakwah dan Sosial
115 Majalah Geografi Indonesia
116 Masyarakat, Kebudayaan dan Politik
117 Moderatio Jurnal Moderasi Beragama
118 Mudra Jurnal Seni Budaya
119 Musãwa Jurnal Studi Gender dan Islam
120 NALARs
121 Nurani jurnal kajian syari'ah dan masyarakat
122 POPULIKA
123 PROMUSIKA
124 Patra Widya Seri Penerbitan Penelitian Sejarah dan Budaya
125 Pelataran Seni

126	Populasi
127	Psikis Jurnal Psikologi Islami
128	Psikodimensia Kajian Ilmiah Psikologi
129	Psikoislamedia Jurnal Psikologi
130	Psikologika Jurnal Pemikiran dan Penelitian Psikologi
131	Psympathic Jurnal Ilmiah Psikologi
132	QALAMUNA Jurnal Pendidikan, Sosial, dan Agama
133	RUANG Jurnal Lingkungan Binaan (SPACE Journal of the Built Environment)
134	Religi Jurnal Studi Agama-agama
135	Religious Jurnal Studi Agama-Agama dan Lintas Budaya
136	Resital Jurnal Seni Pertunjukan
137	Riau Journal of Empowerment
138	SINTHOP Media Kajian Pendidikan, Agama, Sosial dan Budaya
139	Sawwa Jurnal Studi Gender
140	Simulacra
141	Societas Dei Jurnal Agama dan Masyarakat
142	SocioEdu Sociological Education
143	Soshum Jurnal Sosial dan Humaniora
144	Sosial Budaya
145	Sosio-Didaktika Social Science Education Journal
146	Tazkiya Journal of Psychology
147	VISIO DEI JURNAL TEOLOGI KRISTEN
148	Warta LPM
149	Wawasan Jurnal Ilmiah Agama dan Sosial Budaya
150	Zuriah Jurnal Pendidikan Anak Usia Dini
151	el Harakah Jurnal Budaya Islam