

A Preliminary Study of RAG for Taiwanese Historical Archives

Claire Lin^{1*}, Bo-Han Feng^{2*}, Xuanjun Chen^{3*}, Te-Lun Yang⁴
Hung-yi Lee³, Jyh-Shing Roger Jang^{2,4}

¹Department of Information Management, National Taiwan University

²Department of Computer Science and Information Engineering, National Taiwan University

³Graduate Institute of Communication Engineering, National Taiwan University

⁴Graduate Institute of Networking and Multimedia, National Taiwan University

{b10705004, b10902031, d12942018, d12944007}@ntu.edu.tw

hungyilee@ntu.edu.tw, jang@mirlab.org

Abstract

Retrieval-Augmented Generation (RAG) has emerged as a promising approach for knowledge-intensive tasks. However, few studies have examined RAG for Taiwanese Historical Archives. In this paper, we present an initial study of a RAG pipeline applied to two historical Traditional Chinese datasets, Fort Zeelandia and the Taiwan Provincial Council Gazette, along with their corresponding open-ended query sets. We systematically investigate the effects of query characteristics and metadata integration strategies on retrieval quality, answer generation, and the performance of the overall system. The results show that early-stage metadata integration enhances both retrieval and answer accuracy while also revealing persistent challenges for RAG systems, including hallucinations during generation and difficulties in handling temporal or multi-hop historical queries.

Keywords: Retrieval-Augmented Generation, Humanities Data, Large Language Model

1 Introduction

Recent advances in large language models have substantially improved open-domain question answering and knowledge-intensive tasks. Retrieval-Augmented Generation (RAG) (Lewis et al., 2020), which combines document retrieval with text generation, has shown promise in mitigating hallucination and improving factuality. Prior research has primarily focused on English (Bajaj et al., 2018; Kwiatkowski et al., 2019; Yang et al., 2024) or Simplified Chinese datasets (Lyu et al., 2024; Li et al., 2024a) and general-purpose domains such as Wikipedia or web-collected questions.

However, much less attention has been given to RAG performance on underrepresented languages

and culturally specific corpora, particularly in the humanities. Historical contexts in Traditional Chinese pose unique challenges, including unstructured documents, time-sensitive content, and linguistic differences between queries and archival sources. These factors complicate both retrieval and generation, making it unclear how well current RAG systems handle such materials.

To address this gap, we propose two Taiwanese historical datasets, Fort Zeelandia and Taiwan Provincial Council Gazette (TPCG), along with their associated query sets, as case studies for historical open-ended question answering. The datasets are annotated with query-level and document-level metadata, enabling fine-grained experiments on how query types and metadata integration strategies affect RAG performance. Through systematic evaluation across multiple retrieval methods and query characteristics, we demonstrate that early-stage metadata integration substantially improves system effectiveness. Furthermore, our findings reveal persistent challenges: hallucinations remain a recurring issue during generation, and questions involving temporal reasoning exhibit notable difficulty. Furthermore, our analysis of retrieval performance reveals that multi-hop and time-sensitive queries tend to yield lower recall, whereas early-stage metadata integration consistently delivers the strongest overall retrieval effectiveness.

2 Related Work

RAG Lewis et al. (2020) improves language model performance on knowledge-intensive tasks by incorporating relevant external information during generation. By grounding outputs in retrieved evidence, RAG reduces hallucinations when models encounter unfamiliar topics and alleviates the substantial cost of continuously retraining models to incorporate new knowledge.

*Equal contribution.

Dataset	Language	Humanities	Query-Passage Pairs	Metadata
MS MARCO	English	✗	✓	Limited
Natural Questions	English	✗	✓	✗
MMLU	English	✓	✗	✗
CMMLU	Simplified Chinese	✓	✗	✗
Fort Zeelandia Query Set (Our)	Traditional Chinese	✓	✓	✓
TPCG Query Set (Our)	Traditional Chinese	✓	✓	✓

Table 1: Comparison of datasets by language, domain knowledge, structure, and metadata. Fort Zeelandia and TPCG Query sets stand out for their rich metadata and grounding in historical or contextual knowledge.

Early benchmarks of RAG mainly relied on general-purpose datasets such as MS MARCO (Bajaj et al., 2018) and Natural Questions (Kwiatkowski et al., 2019). More recently, researchers have introduced domain-specific datasets in areas including biomedicine (Xiong et al., 2024; Li et al., 2024b; He et al., 2025), law (Pipitone and Alami, 2024; Zheng et al., 2025; Wahidur et al., 2025), and non-English languages such as Traditional Chinese (Yang et al., 2025). However, RAG applications in the humanities are underexplored, particularly for Taiwanese historical materials.

Table 1 compares the key differences of existing benchmarks with the query sets from our newly introduced Fort Zeelandia and TPCG datasets. Firstly, in terms of humanities coverage, MS MARCO and Natural Questions primarily target general-purpose or factual QA and contain little to no humanities material, whereas MMLU (Hendrycks et al., 2021) and CMMLU (Li et al., 2024a) include partial coverage through their broader topical scope. By contrast, our Fort Zeelandia and TPCG query sets are explicitly designed around humanities data, with a particular emphasis on historical materials. Secondly, with respect to query–passage alignment, MS MARCO and Natural Questions are constructed around paired queries and passages, a design we also adopt for Fort Zeelandia and TPCG query sets to support retrieval-based evaluation. MMLU and CMMLU, in contrast, rely on multiple-choice formats. Finally, in terms of metadata, our proposed datasets provide rich query- and document-level annotations, enabling more fine-grained retrieval experiments and analysis than existing resources.

3 Dataset

We introduce two Traditional Chinese datasets from Taiwanese historical archives: Fort Zeelandia and Taiwan Provincial Council Gazette (TPCG). We refer to the associated queries as the Fort Zeelandia Query Set and the TPCG Query Set, and to Fort Zeelandia and TPCG themselves as the document datasets in this paper.

landia Query Set and the TPCG Query Set, and to Fort Zeelandia and TPCG themselves as the document datasets in this paper.

3.1 Fort Zeelandia

Entity	Single-hop	Multi-hop	Total
Event	32	18	50
Item	14	2	16
People	19	4	23
Place	16	6	22
Time	19	4	23
Multi-entity	0	39	39
Total	100	73	173

Table 2: Fort Zeelandia Query Set Entity Focus Distribution across Question Complexity

This dataset is constructed from historical diaries¹ documenting Dutch colonization of Taiwan in the 17th century. We collected 5,443 passages and collaborated with students from the Department of History, who created 173 queries and annotated the relevant passages for each query.

Query-level Metadata. Each QA pair is annotated with query-level metadata, including:

- **Question complexity:** Single-hop or multi-hop question. A multi-hop question requires combining information from multiple passages to determine the answer, whereas a single-hop question can be answered using just one passage.
- **Entity focus:** Whether the question centers on a person, item, time, event, or location.

An example from the Fort Zeelandia dataset is demonstrated in Appendix A.1.

3.2 Taiwan Provincial Council Gazette

The TPCG dataset comprises official meeting records from the Taiwan Provincial Council As-

¹<https://taco.ith.sinica.edu.tw/tdk/>

sembly², spanning the mid to late 20th century, totaling 228,135 documents. To build the question answering benchmark, history students manually crafted 56 question-passage pairs based on selected gazette excerpts. The resulting dataset captures realistic information needs and research scenarios commonly encountered in historical inquiry.

Document-level Metadata. TPCG is characterized with well-defined document-level metadata, enabling experiments on how structured context can be used to improve system performance. Each document is associated with:

- **Time/Event Information:** Includes time information such as the start and end dates, volume and published date.
- **Person/Organization Information:** Covers participating members, agencies, decree, presiding officials and president at that time.
- **Content/Document Information:** Includes document title, abstract, content type, category, subject, keywords, attachments, references, and remarks.

An example from the TPCG dataset is demonstrated in Appendix A.2.

4 Methods

The RAG pipeline in Figure 1 comprises four stages: Input, Retrieval, Generation, and Evaluation. Throughout the pipeline, we (a) construct datasets and annotate query–passage pairs, (b) retrieve candidate passages using lexical, dense, and hybrid methods with optional metadata integration and reranking, (c) prompt a generator LLM with the query, retrieved passages, and metadata to generate an answer, and (d) assess answer quality with an LLM-as-judge protocol.

4.1 Input

The input stage in Figure 1 (a) covers data acquisition and annotation. We first crawl and normalize raw materials into document collections for Fort Zeelandia and TPCG datasets. Domain experts (Taiwanese history students) then author queries and annotate the associated gold passages, yielding high-quality query–passage pairs for RAG experimentation. To enable controlled analysis, we further annotate (i) question complexity (single-hop

²<https://drtpa.th.gov.tw/index.php?act=Archive>

vs. multi-hop) and entity focus (people, event, time, place, item, or multi-entity) for Fort Zeelandia, and (ii) document-level metadata for TPCG, grouped into Time/Event, Person/Organization, and Document/Content categories.

4.2 Retrieval

Given a user query, the retrieval stage in Figure 1 (b) identifies a small set of passages most likely to support grounded answer generation. This stage is essential in a RAG pipeline because it (i) grounds the generator in verifiable evidence to reduce hallucinations, (ii) filters a large corpus into a compact candidate set that fits the context window, and (iii) adapts to lexical, semantic information, and structured metadata in Fort Zeelandia and TPCG. The stage comprises two parts: retrieval models (sparse, dense, hybrid) that score query–passage relevance, and retrieval strategies that optionally use document-level metadata and a second-stage reranker. Together, these components return top- k passages for the generation stage.

4.2.1 Retrieval Models

We instantiate three families of retrieval models:

Sparse retrieval. We adopt BM25 (Robertson and Zaragoza, 2009), which retrieves documents based on term-matching style term-frequency and inverse document frequency (TF-IDF) weighting (Salton and Buckley, 1987), together with sparse embeddings derived from a BGE-M3-based model (Chen et al., 2024).

Dense retrieval. A BGE-M3-based dense encoder maps queries and passages into a shared embedding space for semantic matching, which is helpful when relevant evidence is phrased differently from the query.

Hybrid retrieval. To leverage both lexical and semantic signals, we fuse the sparse and dense ranked lists using Reciprocal Rank Fusion (RRF) (Cormack et al., 2009):

$$\text{RRF}(d) = \sum_{i=1}^n \frac{1}{k + r_i(d)} \quad (1)$$

where d is the document, n is the number of ranked lists, $r_i(d)$ is the rank of document d in the i -th ranked list, and k is a constant that dampens the contribution of the lower-ranked documents.

4.2.2 Retrieval Strategies

Beyond first-stage retrieval, we integrate document-level metadata and a second-stage reranker to im-

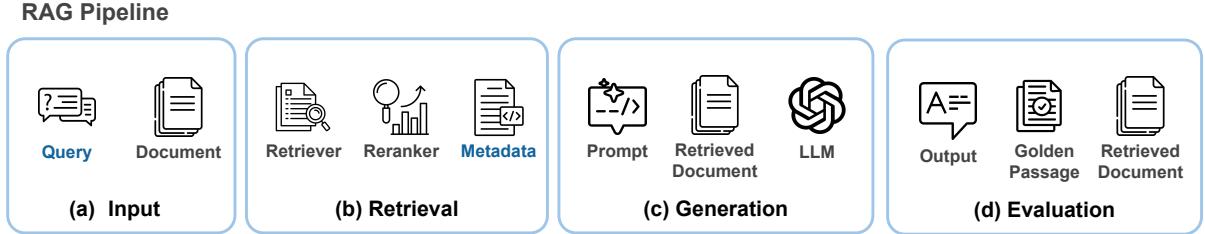


Figure 1: Overview of RAG pipeline and components in each stage. The two highlighted elements: **Query** and **Metadata** are the key factors that impact RAG system performance we focused on in this paper. The details of these factors are discussed in Section 3.1 and Section 3.2, respectively. Section 6.2 and Section 6.3 elaborates how these factors impact retrieval and generation performance.

prove ranking. Metadata in TPCG is grouped into Time/Event, Person/Organization, and Content/Document fields; these fields capture signals (e.g., publication dates, presiding officials, content categories) that are often only weakly expressed in raw text but crucial for precise matching in civic or historical domains. We adopt four strategies, illustrated in Figure 2.

Baseline Retrieval. Retrieve using only the query and original document text without metadata. This provides a clean reference that relies purely on text similarity.

Metadata-Augmented Retrieval. Append selected metadata fields to each document chunk before embedding, treating metadata as part of the content. This allows the retriever to encode, for instance, dates, roles, or categories directly into passage representations so they influence similarity at retrieval time. The retriever returns top- k passages given the embeddings of query and metadata-augmented document chunks.

Metadata-Only Reranking. Incorporate metadata at the reranking stage rather than directly appended to the documents. We first retrieve the top-100 candidate passages using the original documents. Then, compute the similarity between the query and the available document-level metadata of each candidate passage. The passages are reranked based on this similarity score, and the final top- k passages are returned for generation.

Metadata-Augmented Reranking. Append metadata to the original document text before computing similarity for reranking. After retrieving the candidate passages, we concatenate each document’s metadata with its original content, and then measure the similarity between this augmented text and the query to rerank the candidates. The top- k passages are returned for generation.

By comparing these strategies, we aim to quantify the contribution of metadata at both embedding and reranking stages, and to better understand how different integration points influence retrieval effectiveness for historical information retrieval.

4.3 Generation

We use GPT-4o (OpenAI et al., 2024) to produce answers conditioned on the retrieved passages. The goal is to leverage an LLM to aggregate information dispersed across multiple relevant passages into a fluent natural-language response.

At inference time, each query is paired with the top-5 retrieved passages and any available metadata, which together serve as the external knowledge context for generation. The model is instructed to ground its answer strictly in the provided materials and to avoid introducing external knowledge not mentioned in the documents. When multiple passages support the same fact, the model is encouraged to prioritize such corroborated information. If none of the provided materials is relevant to the query, the model is instructed to respond with “I don’t know”. The full generation prompt is detailed in Appendix A.3.

4.4 Evaluation

We evaluate both retrieval performance and end-to-end RAG quality. For retrieval evaluation, we report Recall@ k , which measures the ratio of relevant passages that appear in the top- k retrieved results for each query:

$$\text{Recall}@k = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(\text{Relevant}_i \in \text{Top-}k) \quad (2)$$

where N is the number of relevant passages for the query, $\mathbb{I}(\cdot)$ is the indicator function, Relevant_i is

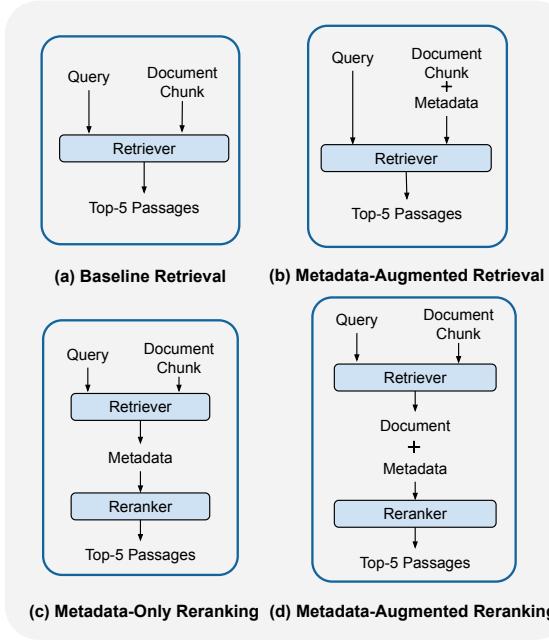


Figure 2: Overview of four retrieval strategies with different metadata integration stages explored in this work. (a) **Baseline Retrieval** retrieves top passages using only the query and document content. (b) **Metadata-Augmented Retrieval** integrates metadata into the document representation during retrieval. (c) **Metadata-Only Reranking** uses only metadata during the reranking stage after initial retrieval. (d) **Metadata-Augmented Reranking** incorporates both document content and metadata in the reranking stage.

the i^{th} relevant passage, and $\text{Top-}k$ denotes the top- k retrieved passages. The average $\text{Recall}@k$ across all queries yields the overall retrieval performance.

For generation quality, we employ Gemini-2.5-Pro (Comanici et al., 2025) as an evaluator following (Chiang and Lee, 2023). The evaluator is given the golden passage, the retrieved top-5 passages, and the answer from GPT-4o. The complete evaluation prompt is provided in Appendix A.4. It consists of three scoring dimensions: groundedness, relevance, and hallucination.

Groundedness. Assesses whether the generated answer correctly incorporates information directly supported by the golden passage. Each distinct atomic fact from the golden passage that appears correctly in the answer receives one point.

Relevance. Evaluates whether the answer includes additional information present in other retrieved passages consistent with the golden passage. Each relevant atomic fact receives one point.

Hallucination. Penalizes content that is unsupported or irrelevant. For each hallucinated state-

ment or extraneous detail that is neither aligned with the golden passage nor substantiated by the retrieved passages, one point is deducted.

5 Experimental Setup

In our experiments, each document is segmented into chunks of 512 tokens with an overlap of 128 tokens to preserve contextual continuity. For direct retrieval methods, where reranking is not applied, both BM25 and BGE-M3-based approaches are configured to return the top 5 most relevant passages (i.e., $\text{top-}k = 5$). The hybrid method independently retrieves 5 passages using both the sparse and dense retrievers, then combines the two ranked lists using RRF, setting $k = 60$, to produce the final top-5 results. For experiments involving reranking, we first retrieve the top-100 candidate passages and then apply reranking using BGE-reranker (Xiao et al., 2023) to select the final top-5 results. In the reranking scenario, the hybrid approach similarly retrieves 100 passages from each retriever before merging and reranking. We do not perform any retriever and reranker tuning; all retrievers and reranker are used off-the-shelf.

For Fort Zeelandia and its query set, we use passages retrieved by a hybrid retriever with baseline retrieval. For TPCG and the associated query set, we fix the retriever to BM25 and evaluate the impact of different metadata integration stages and types on answer quality. GPT-4o is used to generate answers with the retrieved passages, and Gemini 2.5 Pro is used as an independent evaluator.

6 Results

Figure 1 illustrates the RAG pipeline and its key components at each stage. To evaluate the applicability of the RAG system on historical materials, we conduct experiments using Fort Zeelandia, TPCG, and their query sets. Our study examines how different retrieval strategies, query characteristics, and metadata integration approaches affect overall system performance. The evaluation focuses on multiple dimensions, including the ability to leverage accurate context and the extent of hallucinations.

6.1 Overall RAG Results

Tables 3 and 4 show the overall RAG results on the Fort Zeelandia and TPCG datasets. In Table 4, Metadata-Augmented Retrieval with early Document/Content metadata achieves the highest groundedness, with a significant increase of

Question Type	Subcategory	Groundedness \uparrow	Relevance \uparrow	Hallucination \uparrow
All Questions	-	2.9769	1.0578	-0.6821
Question Complexity	Single-hop	2.8600 (-0.1169)	0.8700 (-0.1878)	-0.5600 (+0.1221)
	Multi-hop	3.1370 (+0.1601)	1.3151 (+0.2573)	-0.8493 (-0.1672)
Entity Focus	People	3.2174 (+0.2405)	1.0870 (+0.0292)	-0.5217 (+0.1604)
	Event	3.4600 (+0.4831)	1.2200 (+0.1622)	-0.5800 (+0.1021)
	Time	1.3478 (-1.6291)	0.4783 (-0.5795)	-0.9565 (-0.2744)
	Place	1.8636 (-1.1133)	1.2273 (+0.1695)	-0.7727 (-0.0906)
	Item	2.5625 (-0.4144)	0.1875 (-0.8703)	-0.5625 (+0.1196)
	Multi-entity	3.9744 (+0.9975)	1.4359 (+0.3781)	-0.7436 (-0.0615)
All Questions (Oracle)	-	4.4104	0.2312	-0.2601

Table 3: RAG evaluation by Query Type on the Fort Zeelandia dataset. The table reports average scores for three evaluation metrics: **Groundedness** (incorporates gold passage information), **Relevance** (integrates relevant passages information), and **Hallucination** (including irrelevant information). For all three metrics, higher values indicate better performance. Since Hallucination scores are negative, a value closer to zero reflects fewer hallucinations. All values are compared against the "All Questions" row. Colored deltas in parentheses indicate the difference from the average: green for improvement and red for decline. The Oracle row denotes the upper bound of the LLM's performance when directly given the gold passages. An evaluation example can be found in Appendix A.5.

Integration Stage	Metadata Type	Groundedness \uparrow	Relevance \uparrow	Hallucination \uparrow
Baseline	-	0.7321	0.8571	-0.2500
Metadata-Augmented Retrieval	Time/Event	1.0893 (+0.3572)	1.0000 (+0.1429)	-0.2857 (-0.0357)
	Person/Organization	1.1786 (+0.4465)	0.7321 (-0.1250)	-0.2679 (-0.0179)
	Document/Content	2.1429 (+1.4108)	1.2500 (+0.3929)	-0.3214 (-0.0714)
Metadata-Only Reranking	Time/Event	0.3393 (-0.3928)	1.0000 (+0.1429)	-0.4821 (-0.2321)
	Person/Organization	0.5714 (-0.1607)	0.6071 (-0.2500)	-0.2857 (-0.0357)
	Document/Content	1.5893 (+0.8572)	1.8571 (+1.0000)	-0.3393 (-0.0893)
Metadata-Augmented Reranking	Time/Event	1.2679 (+0.5358)	1.0357 (+0.1786)	-0.6250 (-0.3750)
	Person/Organization	0.9821 (+0.2500)	1.1071 (+0.2500)	-0.6250 (-0.3750)
	Document/Content	1.3750 (+0.6429)	1.0536 (+0.1965)	-0.5357 (-0.2857)
Oracle	-	3.6964	0.0179	-0.0714

Table 4: RAG evaluation by Metadata Integration Strategies on the TPCG dataset. The table reports average scores across the three evaluation metrics. All rows are compared to the Baseline Retrieval, values in the parentheses indicate the improvement or decline. The Oracle row denotes the upper bound of the LLM's performance when directly given the gold passages. Two evaluation examples can be found in Appendix A.6.

1.4108 over the baseline. Appendix A.7 details the significance tests for various retrieval methods. Performance also varies by query type: event-related queries benefit most, with groundedness up 0.4831, relevance by 0.1622, and hallucinations reduced 0.1021. These findings indicate that RAG effectiveness depends on query characteristics and is strengthened by metadata-augmented retrieval, though hallucinations persist even with oracle passages, highlighting a key limitation.

6.2 RAG Results

This section takes a deeper dive into two key factors that critically influence RAG performance at the Input and Retrieval stages: query type and use of document-level metadata. Specifically, we an-

alyze how different query types affect accuracy, relevance, and hallucination. Additionally, we examine the impact of metadata integration at different stages of retrieval and reranking, considering multiple metadata types. This analysis highlights which combinations of query characteristics and metadata strategies yield the most reliable and accurate outputs for historical open-ended QA tasks.

1) Different Query Types Table 3 illustrates RAG performance across query types. Multi-hop and Multi-entity questions are high-risk: when successful, groundedness increases by 0.1601 and 0.9975, and relevance by 0.2573 and 0.3781, but hallucination worsens by -0.1672 and -0.0615, highlighting a trade-off between complexity and reliability. People- and event-focused queries are

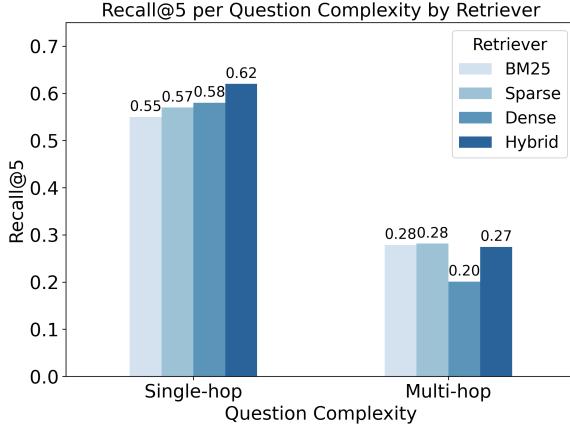


Figure 3: Fort Zeelandia Dataset Recall@5 per Question Complexity by Retriever

more stable, achieving gains in groundedness and relevance with lower hallucination. Time-focused queries are the most challenging, with groundedness and relevance decreasing by 1.6291 and 0.5795, alongside worse hallucination, indicating that temporal reasoning remains a key bottleneck.

2) Different Metadata Integration Strategies
Table 4 presents the evaluation scores across three dimensions for the open-ended question answering task, focusing on the key factor Metadata, using TPCG and its query set.

Overall, Metadata-Augmented Retrieval proves the most reliable approach, improving groundedness and relevance with minimal worsening in hallucination. By contrast, reranking strategies show mixed results: Metadata-Only Reranking underperforms the baseline, while Metadata-Augmented Reranking achieves gains in retrieval quality but at the cost of greater hallucination, making it less stable. Across all strategies, Document/Content metadata emerges as the most effective type, underscoring its importance for enhancing the system.

6.3 Ablation Study of Retrieval Results

In this section, we take a closer look at the Retrieval stage of the RAG pipeline. Since RAG fundamentally relies on retrieved documents as the foundation for generating answers, understanding retrieval effectiveness is critical to interpreting overall system performance. By analyzing how different retrieval strategies, query types, and metadata integration methods influence the quality of retrieved context, we can better identify the factors that drive successes and failures in retrieval.

1) Retrieval with Query-level Metadata We

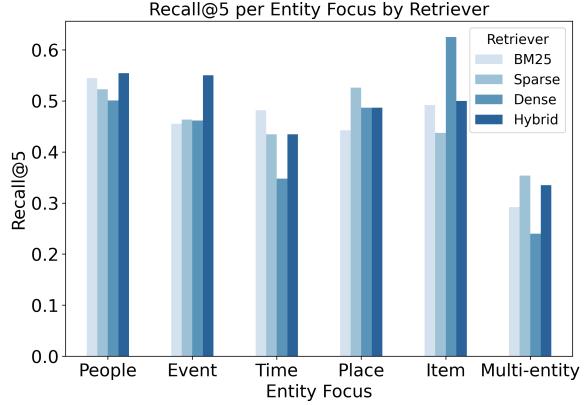


Figure 4: Fort Zeelandia Dataset Recall@5 per Entity Focus by Retriever

investigate the impact of query types on retrieval performance using query-level metadata, focusing on query complexity and entity focus.

Different Question Complexity. To gain deeper insight into RAG performance across varying query complexity, we further analyze the retrieval results on the Fort Zeelandia dataset. Figure 3 presents Recall@5 scores comparing single-hop and multi-hop questions across different retrievers. For single-hop questions, Recall@5 scores are roughly twice as high as for multi-hop questions, corresponding to a lower tendency for hallucination. In contrast, retrievers achieve Recall@5 of at most only 0.28 for multi-hop queries, increasing the likelihood of hallucinated responses.

Notably, despite the lower recall, multi-hop and multi-entity questions still achieve higher groundedness and relevance, suggesting that the LLM is capable of performing multi-step reasoning when appropriate context is provided.

Different Entity Focus. We analyze retrieval performance across different entity focuses to better understand its impact on RAG outcomes. Figure 4 presents Recall@5 scores for People, Event, Time, Place, Item, and Multi-entity questions. For the hybrid retriever used in the RAG pipeline for Fort Zeelandia, performance is notably higher for People- and Event-focused questions, with Recall@5 around 0.55, corresponding to better-controlled hallucination. In contrast, Time- and Multi-entity questions exhibit lower retrieval performance, with Recall@5 of 0.43 and 0.33, respectively, which aligns with increased hallucination.

Considering both RAG scores and retrieval results, we find that although retrieval for Time-focused questions is slightly better than for Multi-

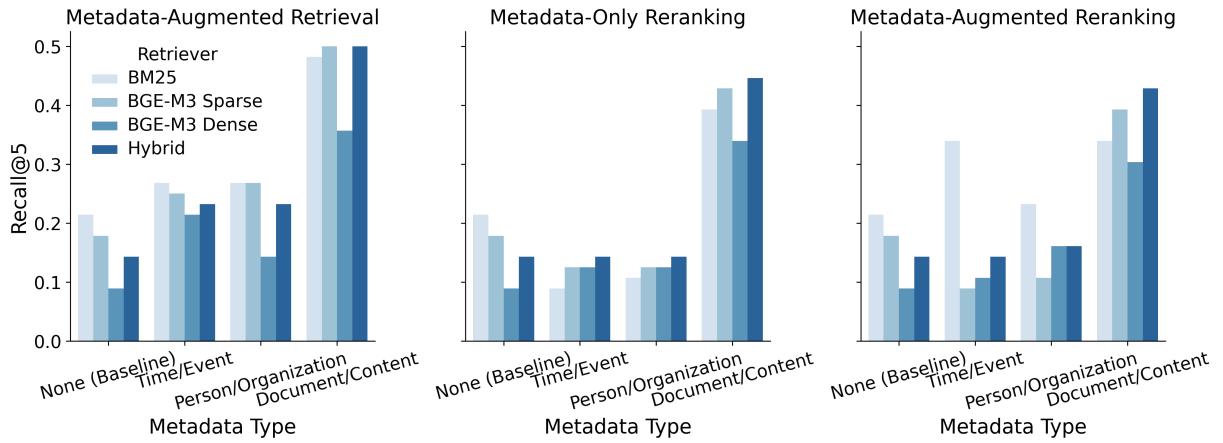


Figure 5: TPCG retrieval performance across different metadata integration stages and metadata types. Left: Metadata-Augmented Retrieval performance across different metadata types. Center: Performance of Metadata-Only Reranking across different metadata types. Right: Retrieval performance of Metadata-Augmented Reranking across different metadata types.

entity queries, the system achieves higher overall evaluation scores on Multi-entity questions. This indicates that the LLM can generate high-quality answers for Multi-entity queries even with partial or imperfect context. In contrast, despite adequate retrieval for Time-focused questions, generation performance remains poor, highlighting that time-sensitive reasoning constitutes a key limitation of the LLM rather than retrieval.

2) Retrieval with Document-level Metadata

We examine the role of document-level metadata in the retrieval process, focusing on metadata type and integration stage.

Different Metadata Type. Figure 5 compares TPCG retrieval performance across different retrievers and metadata types: Time/Event, Person/Organization, and Document/Content, at each integration stage, arguing how metadata affects RAG performance. Document/Content metadata provides the largest improvement over the baseline across all strategies, achieving recall scores roughly twice those of the other types, with the highest around 0.5 under the Metadata-Augmented Retrieval setting. This enhanced retrieval supplies essential context to the LLM, improving answer quality and boosting groundedness and relevance, as shown in Table 4. In contrast, Time/Event and Person/Organization metadata exhibit variable effectiveness across integration stages and are insufficient alone for effective reranking, a trend also reflected in the RAG evaluation scores.

Different Metadata Integration. Figure 5 also illustrates retrieval performance across different metadata integration stages. Metadata-Augmented

Retrieval consistently outperforms the baseline across all retrievers and metadata types. For BM25, which is used for TPCG, recall increases from 0.21 to 0.48, indicating that integrating metadata directly into document embeddings during retrieval enables the most effective use of structured information.

In contrast, Metadata-Only Reranking produces only modest gains and sometimes underperforms the baseline; for BM25, recall drops from 0.21 to 0.08, suggesting that metadata applied solely at the reranking stage is insufficient. Metadata-Augmented Reranking yields mixed results: while recall generally improves over the baseline, gains are smaller than those of Metadata-Augmented Retrieval, leading to greater instability in generation.

7 Conclusion

This study investigates the application of RAG to historical open-ended question answering using two Traditional Chinese historical datasets, Fort Zeelandia and TPCG, along with query sets. By examining the impact of query types and metadata integration strategies on retrieval and end-to-end RAG, we show that early-stage metadata integration substantially enhances performance. Our results also reveal persistent challenges: hallucinations are frequent during generation, and temporal or multi-hop queries are particularly difficult because of the low retrieval recall. These findings inform future humanities-focused RAG research and underscore the need for robust retrieval strategies in historical and Traditional Chinese contexts.

Acknowledgments

We would like to express our sincere thanks to the National Science and Technology Council (NSTC), Taiwan, for funding this research project under Grant No. NSTC 113-2740-H-002-001-MY3, “TAIHUCAIS: TAIwan HUmanities Conversational AI Knowledge Discovery System”. The support has been instrumental in enabling the study.

References

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, et al. 2018. **Ms marco: A human generated machine reading comprehension dataset.**

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. **Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation.**

Cheng-Han Chiang and Hung-yi Lee. 2023. **Can large language models be an alternative to human evaluations?** In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, et al. 2025. **Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities.**

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. **Reciprocal rank fusion outperforms condorcet and individual rank learning methods.** In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’09*, page 758–759. Association for Computing Machinery.

Jiawei He, Boya Zhang, Hossein Rouhizadeh, Yingjian Chen, Rui Yang, Jin Lu, Xudong Chen, Nan Liu, Irene Li, and Douglas Teodoro. 2025. **Retrieval-augmented generation in biomedicine: A survey of technologies, datasets, and clinical applications.**

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. **Measuring massive multitask language understanding.**

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, et al. 2019. **Natural questions: A benchmark for question answering research.** *Transactions of the Association for Computational Linguistics*, 7:452–466.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, et al. 2020. **Retrieval-augmented generation for knowledge-intensive nlp tasks.** *Advances in neural information processing systems*, 33:9459–9474.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024a. **Cmmlu: Measuring massive multitask language understanding in chinese.**

Mingchen Li, Halil Kilicoglu, Hua Xu, and Rui Zhang. 2024b. **Biomedrag: A retrieval augmented large language model for biomedicine.**

Yuanjie Lyu, Zhiyu Li, Simin Niu, Feiyu Xiong, Bo Tang, Wenjin Wang, Hao Wu, Huanyong Liu, Tong Xu, and Enhong Chen. 2024. **Crud-rag: A comprehensive chinese benchmark for retrieval-augmented generation of large language models.**

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, et al. 2024. **Gpt-4o system card.**

Nicholas Pipitone and Ghita Houir Alami. 2024. **Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain.**

Stephen Robertson and Hugo Zaragoza. 2009. **The probabilistic relevance framework: Bm25 and beyond.** *Foundations and Trends in Information Retrieval*, 3:333–389.

Gerard Salton and Chris Buckley. 1987. **Term weighting approaches in automatic text retrieval.** Technical report, Cornell University, USA.

Rahman S. M. Wahidur, Sumin Kim, Haeung Choi, David S. Bhatti, and Heung-No Lee. 2025. **Legal query rag.** *IEEE Access*, 13:36978–36994.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. **C-pack: Packaged resources to advance general chinese embedding.**

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. **Benchmarking retrieval-augmented generation for medicine.** In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251, Bangkok, Thailand. Association for Computational Linguistics.

Te-Lun Yang, Jyi-Shane Liu, Yuen-Hsien Tseng, and Jyh-Shing Roger Jang. 2025. **Knowledge retrieval based on generative ai.**

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, et al. 2024. **Crag – comprehensive rag benchmark.**

Lucia Zheng, Neel Guha, Javokhir Arifov, Sarah Zhang, Michal Skreta, Christopher D. Manning, Peter Henderson, and Daniel E. Ho. 2025. **A reasoning-focused legal retrieval benchmark.** In *Proceedings of the Symposium on Computer Science and Law on ZZZ, CSLAW ’25*, page 169–193. ACM.

A Appendix

A.1 Fort Zeelandia Dataset Example

Figure 6 gives an example from the Fort Zeelandia dataset.

Query Set	
Query: 誰在1632年被派去，再度調查中國漳州河至南澳島的海岸地形？	
Question complexity: Single-hop	
Entity focus: People	
Gold passage ID: 热蘭遮城日誌/I-C/1632-03-04	
Document dataset	
Passage ID: 热蘭遮城日誌/I-C/1632-03-04	
Passage content: 2月28，29日，3月1，2，3，4日。無事，為快艇Catwijck號卸貨。今天有一艘中國人的戎克船從漳州河來到此地，帶消息來說，有兩三艘戎克船裝著貨停泊在廈門港口，準備沒有軍門的通行證也要出航前往馬尼拉，被軍門下令逮捕，貨物全被沒收充公，因此本季前往馬尼拉的船，將不會有二十艘，最多也將不過十四到十五艘戎克船。今天也決定了關於長官普特曼閣下要搭快艇Catwijck號前往巴達維亞的事，議會重新討論以後，仍然決議，一切準備妥當以後，該快艇就要立刻出航；並決議，為要在本季還能儘快向總督閣下報告日本方面傳來的消息和公司在這邊的消息，決定要在本月底以前備好一艘戎克船，以便到時可立即出航前往巴達維亞。又決議，在5月1日到10日之間，下席商務員Jacob van Sandt要率領兩艘裝備齊全的戎克船前往中國，去重新探查從漳州河到南澳島之間的中國沿岸，然後去該南澳島以南巡航，或停泊在南澳島下面的岸邊等候巴達維亞來的船隻，以便長官閣下來回的時候，得以向他報告所有他離開以後所發生的公司的事情。	

Figure 6: A data sample of the Query Set and its relevant passage in the document dataset from the Fort Zeelandia dataset.

福利差遇，致造成他們的流向都市，我認為這就很不公平，就本省而言經常有缺額現象，而北、高二市也有同樣情形，但其遇有缺額一經招考即馬上可補足，而這些人員無非是來自本省，因而造成我們的教師缺額更形嚴重。以上年度而言本省的教師缺額僅國小即高達三、〇〇三位，本以師院畢業生及國小師資進修班的一千七百人，計三千多人應已可補足，但目前還是有缺額，而所缺的應僅止是退休人員的數目，但北、高二市也缺額，馬上的找去我們的老師，因而又造成了我們的缺額，這就是造成教育上目前很不平之處。以上這種現象，我是覺得教育行政單位是有責任，如何的在培育師資方面做到供過於求，如此的師資才能安定，減少流動性，這也是我們努力的方向。就目前估計，國小師資到明年大概已可平衡，因此以後流動就會減少。但這個政策有個缺點，即被分發到偏遠地區的自認已無法調動，屆時恐又會有意見。所以任何政策如何的能取到一個平衡點，這也是我們努力的方向。至於北高二市福利待遇較好的問題，本廳也多次向教育部反映，亦受到重視，因此，對本省偏遠地區的老師加給也特別編列一部分經費，同時其補助辦法、方式本廳也擬定送中央教育部轉報行政院，只要奉核定，則對本省偏遠地區老師的加給會有很大的改善。余議員玲雅質詢：我也是知道廳長很認真的在做，但以高雄縣市而言每年的自強活動經費高市是一千元、高縣是八百元、裝服費高市是二千元，高雄縣...
Start date: 1990-03-26
End date: 1990-03-26
Vloume: 67：第九期，（1990年）
Published date:
Members: 余玲雅
Agencies: 教育處，教育廳
Decree:
Officials:
President:
Title: 臺灣省議會第九屆第一次定期大會：質詢-教育-教育
Abstract: 本省與北、高二市所受教育的質差很多，不知感想如何？延長十二年國教 對學生數的如何分配與現行教育體系下是否足夠分配，均應慎重考慮；目前課程的僵化是值得檢討，尤其課程的一元化是適應於升學的，對於不升學的根本無法接受；如教材要有所改進時，建議應要如何落實本土化教育；如果調整教材之後老師是否要進修？對於新教材要有新的教法？
Type: 公報
Category: 質詢
Subject:
Keywords:
Attachments:
References:
Remarks: 於第九屆第三次臨時大會舉行

Figure 7: A data sample of the Query Set and its relevant passage in the document dataset from the TPCG dataset. Note that some metadata fields are missing in the raw data source, such as **Decree** and **Officials**. The second half of **Passage content** is omitted for brevity.

Query Set	
Query: 臺灣省議會在1990年3月26日的第九期中，如何評估本省與北、高二市在教育品質上的差異？	
Gold passage ID: 003-09-01OA-67-6-6-01-01120	
Document dataset	
Passage ID: 003-09-01OA-67-6-6-01-01120	
Passage content: 臺灣省議會公報 第六十七卷 第九期 九一二分校獨立或設校之權限，自去年開始已授權縣市政府。 蘇洪議員月嬌質詢：如果縣府不做呢？ 教育廳廳長俾民答復：這是不太可能，因縣長是民選的，對於民眾的需要他一定會重視。蘇洪議員月嬌質詢：他人的想法你怎可知道，請教就貴廳的立場是如何？ 教育廳廳長俾民答復：只要本廳可為助力的一定儘量支援。至於縣府不做時，我想本廳應也是有辦法制肘的，或如將補助經費減少等方式。 蘇洪議員月嬌質詢：■召集開會，請也邀請臺影公司到現場錄影 >以上謝謝。 余議員玲雅質詢：（79年8月14日）首先請教廳長，我想廳長應知道，本省與北、高二市所受教育的質差很多。據本席持有之分析資料表，如以福利、設備、退休、教育而言，本省是遠遠落在北、高二市之後，不知廳長是否了解，甚至其都有不同標準。尤其高雄市縣只有一線之隔，因此大部分只要有機會轉到高雄市服務。我們常說當兵、納稅、受教育是人的三大義務，而納稅方面而言，省民並不亞於北、高二市 >但為受教育的處遇就要差很多？就此點不知廳長的感想如何？ 教育廳廳長俾民答復：省市的教師待遇有差別是事實，但就全省二十一縣市而言，我們也不能保證各縣市教師的福利措施都完全一樣，因這是涉及到各縣市財政問題。但要以投入教育經費每位學生平均數額而言，本省是比北、高二市低 >不過要比高市略高。至於所提福利措施方面，據資料有一部分本省是較少，如北高二市有交通費補助、水電費部分補助，這部分在本省一般地區就沒有，但偏遠地區仍有補助的，所以就說是補助重點不一樣。另所提因老師	

A.3 Generation Prompt

The full prompt provided to GPT-4o for response generation, given the query, retrieved passages, and available metadata, is shown in Figure 8.

你是一個根據檢索文檔回答問題的語言模型。你會收到一個使用者的問題，以及一筆或多筆可能相關的文檔，每筆文檔都包含「內容 (content)」和「後設資料 (metadata)」。請遵循以下指引回答問題：
1. 根據文檔內容和後設資料回答問題，避免加入檢索資料中未提及的資訊。
2. 優先使用內容一致、來源可靠的文件，不使用有矛盾、錯誤或無關的資料。
3. 如果所有文件都與問題無關，請誠實地說明找不到答案。
使用者問題 {query}
檢索文檔 內容: {content_1} {metadata1_1 name} : {metadata1_1 content} {metadata1_2 name} : {metadata1_2 content} ... {metadata1_m name} : {metadata1_m content}

```

    內容 : {content_2}
    {metadata_{2-1} name} : {metadata_{2-1} content}
    {metadata_{2-2} name} : {metadata_{2-2} content}
    ...
    {metadata_{2-m} name} : {metadata_{2-m} content}
    ...

    內容 : {content_n}
    {metadata_{n-1} name} : {metadata_{n-1} content}
    {metadata_{n-2} name} : {metadata_{n-2} content}
    ...
    {metadata_{n-m} name} : {metadata_{n-m} content}

```

Figure 8: RAG generation prompt to GPT-4o. Retrieved passages are numbered from 1 to n , representing the 1st retrieved passage to the n^{th} retrieved passage. Metadata rows for each retrieved passage are numbered from 1 to m , representing the 1st type of metadata to the m^{th} type of metadata.

A.4 Evaluation Prompt

The full prompt provided to Gemini-2.5-Pro for response evaluation, given the query, golden passages, retrieved passages, available metadata, and model response of GPT-4o, is shown in Figure 9.

你是一位專業的回答品質評估員，負責根據模型回覆是否正確地整合了提供的【標準答案文檔】與【檢索文檔】來回答問題，進行評分。

請根據以下三個面向，以 0 分為每個面向的初始分數給出各個面向的分數，最後計算總分：

面向一：【是否包含標準答案文檔的內容】

* 若檢索文檔中不包含標準答案文檔，則此面向為 0 分。
* 若有包含，把標準答案文檔的內容以最小單位拆解成多個敘述，只要模型回覆每包含一個敘述，就加 1 分。

面向二：【是否整合檢索文檔中其他與標準答案文檔相關的內容】

* 標準答案文檔除外，若其他檢索文檔內容都與標準答案文檔無關，則此面向為 0 分。
* 若有與標準答案文檔相關的其他文檔，把所有相關的其他文檔的內容以最小單位拆解成多個敘述，只要模型回覆每包含一個敘述，就加 1 分。
* 出現多個內容重複的敘述只加一次分。

面向三：【是否包含幻覺或無關內容】

* 每一個包含幻覺、參考無關文檔、對回答問題沒有幫助的敘述，各減 1 分。

請嚴格依據以上標準評分，並輸出以下格式：

...

面向一：+[分數]
面向一評分理由：
[條列每個加分項]

面向二：+[分數]
面向二評分理由：
[條列每個加分項]

面向三：-[分數]
面向三評分理由：
[條列每個減分項]

總分：[分數總和]

...

以下是你要評估的資料：

問題

```

{query}

#### 標準答案文檔
{golden passages}

#### 檢索文檔
{retrieved passages}

#### 模型回覆
{model response}

```

Figure 9: RAG evaluation prompt to Gemini-2.5-Pro. Formats for golden passages and retrieved passages are the same as the retrieved passages in the RAG generation prompt.

A.5 Fort Zeelandia Dataset Evaluation Example

Figure 10 gives a detailed example of the evaluation result on a single-hop question from the Fort Zeelandia dataset.

Focusing on the third scoring dimension of the evaluation result, we can observe that GPT-4o, which is used for model response generation, can still hallucinate, even when the golden passage is retrieved as the first retrieved passage. The hallucination may be attributed to the model's tendency not to include violence-related information from the golden passage, resulting in an incomplete response.

A.6 TPCG Dataset Evaluation Example

Figure 11 and 12 give two detailed examples of the evaluation results on the TPCG dataset.

In the first example, the model response from GPT-4o covers almost all the information in the golden passage, which is also the fifth retrieved document. However, the meeting session (in the **Title** metadata field) of the first retrieved document is wrongly linked to the golden passage and appears in the model response. This example suggests the limitation that hallucination may come from the integration of rich and complex metadata information.

In the second example, the evaluation result of the second scoring dimension shows that GPT-4o can still summarize related information from other retrieved passages even though the golden passage is not given for generation. Nonetheless, the model regards the requirements and questions, which are raised by council members, as implemented specific measures, introducing another type of hallucination due to the imprecise interpretation of retrieved passages.

Query

荷蘭當局決議要求牧師甘迪留斯以「甜蜜的方法」勸阻Taccaran前往日本，這反映了當局在處理此事上的態度為何？

Golden passage

Passage ID: 热蘭遮城日誌/I-F/1633-07-07

Passage content: 7月7日。長官閣下在議會提出上述昨天的意見，乃決議[1]，由於議會與長官閣下都同樣認為，中國人由於恐懼，（跟以前一樣）會再用漂亮的言談來誤導拖延我們，使我們耗費很多費用。因此，我們要遵照總督閣下與東印督議會的指令，要用最猛烈，又儘量少流血的方法，向中國作戰，去攻擊，奪取他們的船隻，為此目的，要派快艇Bredam號、Wieringen號與平底船Warmondt號及一艘戎克船一起部署在南澳下面，去執行要給他們的指令，於其餘的船隻離開他們以後一兩天，要留在那裡，把所有泊在岸邊或擱在陸地上的船隻通通燒毀，把奪得來的貨物保存下來。達成這任務之後，上述快艇要各往預定地點，即Bredam號要偕同一艘戎克船去好望角，Wier-ing號要偕同一艘戎克船去鐘灣[2]，平底船Warmondt號則要去東山，在那裡逮捕中國人的及葡萄牙人的船隻，而大船Middelburgh號、快艇Texel號、Weesp號、Couch-ercke號、Catwijck號、Zeeburgh號、Salm號、Kemphaen號及戎克船打仗號則要前往漳州河及廈門，要去那裡同樣攻擊燒毀他們的船隻。

並決議，於上個月14日捕獲那艘從薩摩來到大員的那艘chiamboey船的船長，舵工及還在我們手中的水手，都要從那裡送來此地，以儘量避免跟日本人衝突而造成各種災難。

並決議，要用各種方法阻止麻豆社的酋長Taccaran前往日本，因為他在日本的出現，從各方面看來，會造成荷蘭聯合東印度公司很大的不利。據悉，牧師甘迪留斯跟上述Taccaran有過很好的友誼，要請他用甜蜜的方法留住該Taccaran，詳情請看決議錄。

最後決議，因鑑於在奪取戰利品時，經常發生混亂無序的現象，因此要發出公告，貼在每艘快艇上，禁止任何人不得越軌傷害中國人，更不許殺死中國人，每一個人都要繼續拿著武器，靜靜地留在船裡，詳情及如何處罰違規，請參閱該告示。

Retrieved passages

Passage ID: 热蘭遮城日誌/I-F/1633-07-07

Passage content: 7月7日。長官閣下在議會提出上述昨天的意見，乃決議[1]，由於議會與長官閣下都同樣認為，中國人由於恐懼，（跟以前一樣）會再用漂亮的言談來誤導拖延我們，使我們耗費很多費用。因此，我們要遵照總督閣下與東印督議會的指令，要用最猛烈，又儘量少流血的方法，向中國作戰，去攻擊，奪取他們的船隻，為此目的，要派快艇Bredam號、Wieringen號與平底船Warmondt號及一艘戎克船一起部署在南澳下面，去執行要給他們的指令，於其餘的船隻離開他們以後一兩天，要留在那裡，把所有泊在岸邊或擱在陸地上的船隻通通燒毀，把奪得來的貨物保存下來。達成這任務之後，上述快艇要各往預定地點，即Bredam號要偕同一艘戎克船去好望角，Wier-ing號要偕同一艘戎克船去鐘灣[2]，平底船Warmondt號則要去東山，在那裡逮捕中國人的及葡萄牙人的船隻，而大船Middelburgh號、快艇Texel號、Weesp號、Couch-ercke號、Catwijck號、Zeeburgh號、Salm號、Kemphaen號及戎克船打仗號則要前往漳州河及廈門，要去那裡同樣攻擊燒毀他們的船隻。

並決議，於上個月14日捕獲那艘從薩摩來到大員的那艘chiamboey船的船長，舵工及還在我們手中的水手，都要從那裡送來此地，以儘量避免跟日本人衝突而造成各種災難。

並決議，要用各種方法阻止麻豆社的酋長Taccaran前往日本，因為他在日本的出現，從各方面看來，會造成荷蘭聯合東印度公司很大的不利。據悉，牧師甘迪留斯跟上述Taccaran有過很好的友誼，要請他用甜蜜的方法留住該Taccaran，詳情請看決議錄。

最後決議，因鑑於在奪取戰利品時，經常發生混亂無序的現象，因此要發出公告，貼在每艘快艇上，禁止任何人不得越軌傷害中國人，更不許殺死中國人，每一個人都要繼續拿著武器，靜靜地留在船裡，詳情及如何處罰違規，請參閱該告示。

Passage ID: 热蘭遮城日誌/I-G/1635-05-09

Passage content: 5月9日。有一艘戎克船出航前往〔下〕淡水，要去捕魚。有南風，好天氣。

今晚長官閣下接到牧師們從新港寄來的一封信，從那封信的內容得悉，麻豆的一個人Taccaran，他以前有一段長時間在各方面的看法都被我們當作是朋友，而被我們善予款待，最近對我方裝腔作勢，表現非常莽撞，大膽而且高傲，屢對新港及其附近村落的荷蘭人暴怒，說，荷蘭人怕他，因為他們的人殺過他們〔荷蘭人〕的士兵，所以如果要使附和我方的新港人害怕，他們也必須要這樣做；因此新港人非常恐慌起來，都想要對麻豆人作戰，而且聽說，麻豆人要來放火燒毀他們的村子，因為在那裡只不過有10到12個荷蘭人。他們〔牧師們〕也寫說，上述Taccaran很隨意地把一種他們稱之為pockon[1]的器具〔或樂器，instrument〕要拿去Topangh[2]，用以強調，將來他要保護他們。

因為這些事情如果不予及時處理，必將造成公司的侮辱和傷害，因此那些牧師們非常懇切地請求，長官閣下要親自帶領一隊約80到100個士兵〔前來新港〕，因為他們確信，這將使新港人鼓舞起來，而使麻豆人的敵氣消沈下去，並可使其他村落的人保持應有的服從與和平；而且，為要使這些大膽的民族更加害怕起來，長官閣下也要前往日加溜灣（藉口要出去散步），據上述牧師們的見解，這是一趟不必動武器的出征。

Passage ID: 热蘭遮城日誌/I-G/1635-05-18

Passage content: 5月15，16，17，18日。無特別的事，只有這幾天都很忙著從上述前來的各戎克船收購他們的絲，絲貨及其他貨物，這幾天有幾艘戎克船出航前往中國。

大部份時間吹北風。

今天長官收到牧師羅伯·尤紐斯的一封來信，信裡寫說，新港人表示抱歉，他們不知道他閣下禁止向麻豆人再表現友誼，並請他閣下對此原諒他們；要跟葡萄牙人多來往，而跟麻豆人少來往的〔想法〕，他認為那是他閣下很遇到的想法，可以把麻豆人壓制下去，而把葡萄牙人聯合到我們這邊來，但不要做得使他們看出我們跟他們友好是為要獲得某種利益的。此外又寫說，昨天有兩個麻豆的長老去過那裡，乃向他們陳述他閣下對Taccaran甚為憤怒，並責令他們，任何人都不得妨礙持有公司證件的中國人在魍港燒石灰或捕魚，也不得侮辱他們或其他傷害的舉動。對此他們承諾會遵行，並會向所有的麻豆人傳達。他們因此請求，派兩三個荷蘭人跟他們一起去〔麻豆〕，用以象徵友誼。對此上述尤紐斯予以拒絕說，要等候他閣下命令才能派遣[1]。

Passage ID: 热蘭遮城日誌/II-F/1644-09-06

Passage content: 9月6日。好天氣，吹陸風。隊長Pieter Boon率領快艇Leeuwerck號與那艘大的小艇出航前往淡水。繼續忙著裝糖桶到Dolphijn號上，要運去交給大船Haarlem號。

我們也要探訪傳道Gerrit Jansz. Hartgrinck寫一封信，用以答覆昨天收到的他的來信，要他把學校老師Caesar van Winschooten儘早送來此地，以便來為他所犯的幾樣過錯答辯。

今天有1艘小戎克船出航前往淡水載硫磺，搭31個人；也有1艘coya船出航前往澎湖，空船，搭11個人；另有2艘coya船出航前往中國，載有鹹魚，搭5個人[1]。

今天長官與議會允許並規定，首先，教會會議會得以檢查所有教會人員的工作情況，並得予以停職或降職；並規定，該教會會議會必須將那些人的犯錯資料以及他們應受處罰的意見交上來，以便於需要時〔福爾摩沙〕議會可針對有關案件進一步審議[2]。fol. 171其次，在探訪傳道權力下的所有本地學生，以及在福爾摩沙的學校擔任學校教師的居民〔指原住民〕，都要維持現狀，這些教師或學生，政務員都不得調派去做任何其他工作，除非有緊急的需要[3]。

第三，那些長老，即當地的酋長，將取消允許他們將學生帶離學校的權力（因為還有很多人是異教徒，他們妨礙神的教會的發展），唯一例外的就是新港的長老，在有進一步的決議以前，他們仍得擁有這項權力，因為他們是最虔誠神的當地人。

也決議，從所提選舉長老與執事的兩倍人選中，決定選擇下席商務員Eduard aux Brebis[4]為長老，下席商務員Wijnant Rutgers[5]為執事。

並決議，要派2個探訪傳道和6個士兵隨牧師Simon van Breen去北區，以便去那邊學習語言，並推展教會的工作。並決議，要派傳道Hans Oloff與Hendrick Veer，取代探訪傳道Gerrit Jansz. Hartgringh，去大木連工作；並派曾經在麻豆與阿猴（Acau）任職的學校教師Caesar van Winschooten[6]，fol. 171v以及曾在日加治灣任職的探訪傳道Joost Gilles，去新港工作[7]。

並決議，在那艘Dolphijn號裡，除了裝運要帶去交給大船Haarlem號的糖以外，還要裝上要運回祖國的7箱各種布料、4桶薑糖、2桶茯苓、170包瓷器等、100個圓形的大醜缸。

也決議，要出售大量的tacabosken[8]給中國人，用以填補新開始的〔資金的〕短缺。

本月2日偕同稽查官Adriaen van der Burgh從澎湖抵達此地，並來這議會提過他們的請求的那幾個遲遲人，來回答說，他們已經考慮過我們那時回答他們的意見，即自己〔租〕用一艘戎克船去日本，因為我們的船都已經出航了，他們回答說，因為考慮到他們沒有士兵，不能使用中國人的船，現在〔要航往日本的〕季節又快要過去了，還有其他種種阻礙，因此最後認為，應當為他們的主人，遲遲的國王，再次來向長官請求，讓那一封國王的書信，以及25到30個他們的人員和那些日本的翻譯員，得以用一艘我們的船送往日本。這個請求，長官與議會聽了，並加考慮fol. 172認為這樣做會有困難，因此決定，要拒絕他們的請求。禮貌地回答他們說，除了今年〔要航往日本的季節〕已經很晚之外，我們的船也不得運外國人去日本，而且我們也不知道現在日本的和遲遲的國王互相之間關係如何，因為已經數年，遲遲的使臣似乎未被日本的皇帝陛下接見過，他們的人員，禮物和運去的貨物都被拒收，而且那些日本的翻譯員，無可懷疑地，都一定會被處死，而把他們載去日本的我們，也會因而遭遇很大的危險，我們自己也會被砍頭。我們把這些話告訴他們以後，他們感謝議會給他們的警告，說，回去遲遲以後，會這樣向他們的國王報告。於是請求，因為（如上所說）不懂如何搭用中國人船，也沒有信心將國王的書信放在中國人的船裡，因此請求我們，讓他們搭我們的船回去遲遲他們的國王那裡，並請求說，為了該書信的緣故而攜帶的他們的蘇木和其他貨物，准予在此地出售，並幫助他們出售這些貨物。對此，長官與議會決議，要簡要地回答他們說，對於第一個要求，fol. 172v即要搭我們的船回去遲遲的事情，我們將予以考慮，對於第二個要求，即要出售他們的貨物的事情，將允許他們，並在不損害公司的利益下，準備要幫助他們，詳情載於今天的決議錄裡[9]。

Passage ID: 热蘭遮城日誌/III-A/1648-11-02~1650-03-10/補充資料

Passage content: 〔補充資料〕

〔1648年11月2日至1650年3月10日〕

1648年12月裡有2艘船從大員來到巴達維亞入港。平底船Juffer號於12月5日抵達，所載貨物有1,421箱砂糖；平底船Os號於12月21日抵達，也載來一批砂糖和其他數種商品，總值28,681.18.15荷盾。

Dagregister Batavia〔荷文本《巴達維亞城日誌》〕1647-1648, 171, 188。

雖然中國大陸的內戰還沒結束，中國商人從南方的商港漳州、安海、廈門、金門跟福爾摩沙的交易還相當暢通。謠傳還會有很多暴亂發生。滿州軍隊已經攻取內陸三個城市，殺掠百姓。國家遭遇飢荒，因為農地大都荒蕪。1648年又見無數的中國人逃來福爾摩沙，其中有500個婦女和1,000個小孩。這年福爾摩沙群島上有超過20,000的成年中國人。Generale Missiven〔《總督一般報告》〕1649年1月18日函，1639-1655, Van der Lijn、Caron、Reniers.、Van Dutecom、Demmer, VIII, Batavia, 18 januari 1649, 354, 355。

荷蘭聯合東印度公司在出島和福爾摩沙的商館都賺到很多錢。這些盈餘的錢用銀送往東京和遲遲的商館去收購商品；總計運150箱去東京，20箱去遲遲。

福爾摩沙的長官與議會決議，要稍微調高糖價，用以鼓勵中國農夫種植甘蔗。過去一段時間，那些農夫到處種稻與其他糧食作物，因為中國大陸缺糧，造成糧食作物價格飛揚。公司也看到運糖去日本市場的特別好的機會。長久以來，日本的糖的市場由中國供應，現在他們的供應受到阻礙了，公司正可利用這機會出口福爾摩沙的糖去日本。熱蘭遮城堡的人員很缺乏數種需用品，例如柏油、繩索、錫等物。有兩百多個該島的駐軍契約即將屆滿，因此議長Pieter Anthonisz Overtwater請求要及時派兵來替補。

淡水的議會報告說，北方的噶瑪蘭（Cabalang）人對荷蘭人發怒[1]。直接的起因是隊長Thomas Pedel下令去處罰一個罪犯所引起的。此外，從北方地區傳來的都是好消息。那十二個龜島（Coeland）的村社，已有十一個村社的長老跟公司結盟了。位於本島東岸的哆囉滿（Tarraboaan）地區也只剩下一個反抗的村社。近期中，將派下席商務員Anthony Plockhoy去該村社勸和，並將沿途繪製該山區的地圖。

1648年春天曾有兩個福爾摩沙的村社首長從北方去熱蘭遮城堡探望。雖然他們被接待時，特別稱讚過公司的政策，但以後卻批評對他們的孩童的基督教教育。跟在城堡附近的學校的授課情形比較起來，他們那地區的教育品質就差的遠了。牧師Jacobis Vertrecht去虎尾墘附近幾個村社探訪，說服了那裡兩個地方的頭人跟公司締和。他相信，那兩個村社的人到時會派代表來參加下次的地方會議。

〈議長P. A. Overtwater致總督C. van der Lijn與巴達維亞議會函〉，大員，1649年2月1日。VOC 1172, 443-449[2]。

長官Nicolaas Verburch於1649年11月18日寫說，1649年6月18日到8月19日從巴達維亞派出，經由遲遲、占碑、東京和日本航來目的地大員的十八艘平底船，有十四艘抵達大員了。還沒看到平底船Campen號、Witte Paert號、Gulden Gans號與Salm號。希望這幾艘船已經找到避風港安全渡過這颱風季節。上述公司的奴隸那些Pampang人〔指逃走被捕捉的那些人〕，由總督府裁定釋放，因為知道福爾摩沙非常缺乏勞工。因為中國一向交易的貨物，例如絲和瓷器，缺貨，所以現在用黃金來交易。長官Verburch認為，這種商品的短缺是中國內部繼續動亂的徵候。很多來福爾摩沙的中國商人也同意這種看法。這種不穩定的中國貿易，使大員的公司當局更難於供應祖國、巴達維亞與印度沿岸通常的需求。Verburch是想要聽從巴達維亞的命令，即要將中國的黃金降價到每十兩24卡拉的黃金兌一百兩銀。不過他看不出有將商品價格按照比例降價的可能性，因為他不敢向福爾摩沙的中國商人提出降低黃金價格的事情。公司已經向他們說過，目前黃金價格不會低於每十兩24卡拉的黃金兌一百一十五兩精銀。如果這諾言不履行，黃金將可能不再運來福爾摩沙了。不過，所提每十兩24卡拉的黃金兌一百兩精銀的價格，在隔年的大員帳簿就有記載了。這樣獲得的利益可能可以補償在Coromandel和公司其他商館的損失。

想要鼓勵農夫在福爾摩沙生產絲的嘗試，沒有獲得預期的效果。長官提出相對的辦法，就是把中國絲的進價提高到每擔七百里爾。福爾摩沙的農地，因為現在不必再去種植桑樹，可以補償那提高的價格。中國購農〔pachtboeren可能是指向公司付租金種田的農夫，如同清代的墾首〕抱怨說，公司的士兵在收稅和分發人頭稅單時態度粗暴。議會答應將注意改善，以防類似事情發生。按照長官Verburch的看法，公司與這島上中國移民之間和諧的關係是非常重要的；他認為中國人是：福爾摩沙島上唯一提供蜂蜜的蜜蜂，沒有這些人，尊貴的公司是無法在此生存的[3]。

中國人湧入本島的潮流已見減緩。1650年11月登記居留的中國人為11,339人，其中有838個女人。福爾摩沙的作物，過去這季節因乾旱而歉收。同時，〔公司人員〕在出島私自交易的事情曝光，遠超過了合理的界限。以後在大員與遲遲的轉運站裝船時，對目的地為東京與日本的船將嚴格檢查。

〈長官N. Verburch致總督C. van der Lijn函〉，大員，1649年11月18日。VOC 1172, 466-491。

大員商館在1648-1649年會計年度淨賺了467,000里爾。福爾摩沙的長官報告說，養蠶業大為倒退。他建議總督府當局對養蠶業的前景要趕快做決定性的決策。在巴達維亞的人認為，在福爾摩沙島養蠶並非不適，而是被中國人的頭領暗中破壞了，他們害怕公司對這行業變成太內行。

現在中國嚴重的飢荒已經過去，逃來福爾摩沙的中國人，有些人又回去他們在中國的老家了。1649年春天繳納人頭稅的中國人為一萬兩千人，而1648年3月還有一萬四千多人繳納。駐軍有984個軍人。由於稻米和其他糧食作物在中國市場降價了，所以本島的農夫越來越多人種植甘蔗。公司預計，可自鄰近村社的農地收到作物什一稅46,000里爾。長官Verburch指出，原住民屢受中國人謀商壓榨，令人憂慮。那些中國人利用他們進口舶來品的專利，完全隨意喊價，福爾摩沙當地人為此付出過分多餘的費用，而他們提供的鹿肉和鹿皮，則必須無奈地按照通常價格交易。Verburch警告說，如果公司對這失常狀況不迅速加以處理，必將造成原住民的騷動。

今年又有很多派駐南區村社的公司人員感染地方病。不過議會還是決定不撤出該地區。佈道活動頗有進展。牧師Simon van Breen於返回祖國之際，隨身攜帶了一些有荷蘭文以及福爾摩沙語文翻譯的宗教經文的稿件。他的任務是，要在荷蘭共和國印刷這些稿件，以便將來在該島可用來作為基督教教育的教材。

巴達維亞總督府不贊成十七人董事會所提派商務代表去北京訪問滿州朝廷的意見。巴達維亞當局認為，這個新的中國的當權者不會從永久的政權基礎建立起一有力的中央政府。那些滿州人大事搶掠之後，會再退回去他們在北方的祖國，他們如此期待著。不想派正式的使臣去中國的另一理由是，南方幾個商港都還在忠於明朝的人的手裡。鄭成功是公司與滿州朝廷之間建立任何來往關係的一個直接的威脅。總督與議會判斷，公司與中國之間的商務關係，最好在正式的外交關係之外實際進行。類此政策對商務最無損害，也不致影響中國人對公司運來的日本銀的需求。

《Generale Missiven》〔《總督一般報告》〕1639-1655, Van der Lijn、Caron、Reniers、Demmer、Barendsz, X, 巴達維亞, 1649年12月31日, 391-395。

Model response

根據檢索到的文檔，荷蘭當局決議要求牧師甘迪留斯以「甜蜜的方法」勸阻麻豆社的酋長Taccaran前往日本，這反映了當局在處理此事上的態度是希望以和平和友好的方式來解決潛在的外交問題。當局認為Taccaran在日本的出現可能會對荷蘭聯合東印度公司造成不利，因此希望通過已有的友誼和柔和的方式來達成目的，而不是採取強硬或對抗的手段。這顯示出當局在處理此類國際和外交事務時，傾向於使用非暴力和協商的方式。

Evaluation result

面向一：+4

面向二評分理由：

- [+] 模型回覆提及，要勸阻麻豆社的酋長Taccaran前往日本。
- [+] 模型回覆提及，Taccaran在日本的出現可能會對荷蘭聯合東印度公司造成不利。
- [+] 模型回覆提及，要利用牧師甘迪留斯與Taccaran已有的友誼。
- [+] 模型回覆提及，要使用「甜蜜的方法」（柔的方式）來達成目的。

面向二：+0

面向二評分理由：

[0] 模型回覆未整合其他檢索文檔中的相關內容。例如，在同一份決議中，荷蘭當局才剛決議要用「最猛烈」的手段攻擊中國船隻，這與對待Taccaran的「甜蜜的方法」形成鮮明對比，能更完整地反映當局務實、依對象而調整手段的態度，但模型並未提及此點。

面向三：-1

面向三評分理由：

[−1] 模型回覆的結論「這顯示出當局在處理此類國際和外交事務時，傾向於使用非暴力和協商的方式」為不恰當的推論。根據同一份【檢索文檔】，當局同時也決議要用「最猛烈」的方式攻擊中國船隻，可見其並非普遍「傾向於使用非暴力和協商」，而是根據對象與情勢採取不同手段。此結論忽略了文檔中的其他資訊，對回答問題沒有幫助。

總分：3

Figure 10: Evaluation result on the Fort Zeelandia dataset.

Query
臺灣省議會在第九期中，對現行教育課程的一元化問題提出了哪些檢討與改進建議？
Golden passage
Passage ID: 003-09-01OA-67-6-01-01120 Passage content: 臺灣省議會公報 第六十七卷 第九期 九一二 分校獨立或設校之權限，自去年開始已授權縣市政府。蘇洪議員月嬌質詢： 如果縣府不做呢？教育廳陳廳長偉民答復： 這是不太可能，因縣長是民選的，對於民衆的需要他一定會重視。 蘇洪議員月嬌質詢：他人的想法你怎可知道，請教就貴廳的立場是如何？教育廳陳廳長偉民答復： 只要本廳可為助力的一定儘量支援。至於縣府不做時，我想本廳應也是有辦法制肘的，或如將補助經... Start date: 1990/03/26 End date: 1990/03/26 Volume: 67：第九期，（1990年） Members: 余玲雅 Agencies: 教育處，教育廳 Title: 臺灣省議會第九屆第一次定期大會：質詢-教育-教育 Abstract: 本省與北、高二市所受教育的質差很多，不知感想如何？延長十二年國教 對學生數的如何分配與現行教育體系下是否足夠分配，均應慎重考慮；目前課程的僵化是值得檢討，尤其課程的一元化是適應於升學的，對於不升學的根本無法接受；如教材要有所改進時，建議應要如何落實本土化教育；如果調整教材之後老師是否要進修？對於新教材要有新的教法？ Type: 公報 Category: 質詢
Retrieved passages
Passage ID: 003-09-09OA-75-6-01-01058 Passage content: 臺灣省議會公報 第七十五卷 第十一期 灣省各級學校員生消費合作社改進要點」第二條規定：各級學校均應設置員生社，其有特殊情形，報經主管機關核准者得不設立。 (二)為因應實際需要，各校代訂學生餐盒自八十二學年度第二學期起由學校員生社代辦並得酌收處理費，惟至多不超過餐盒進價百分之五以內為限，並以進銷貨登帳方式辦理。至於販售之物品，應經社務會議決定並經校長同意後，方可出售。 (三)為加強督導員生社業務...
Start date: 1994/01/17 End date: 1994/12/16 Volume: 75：11，（1994年） Members: 楊文欣 Agencies: 教育廳 Title: 臺灣省議會第九屆第九次定期大會：質詢-教育-教育 Abstract: 一、全省危險教室知多少？（一）七月十日提颱風造成教室倒塌或危險程度？教育廳有無立即調查？有無補救措施？截至目前好像全無資訊。（二）多年前全省危險教室調查有三千間之多，也曾由中央及省府編列預算執行改善，但最近發表全省危險教室還是三千多間，究竟這些年間省教育廳對危險教室做了些什麼？錢用了多少？為什麼沒有效果請陳廳長詳細說明。二、最近的全國教育會議有什麼結果？（一）對教育方針有何改進意見？（二）對中、小學教材有何改進意見？（三）省教育廳對會議有什麼意見提出？三、國中國小教師爭相申請退休，以爭取在退休待遇修改之前退休獲得較好的權益，教育廳對此現象有無因應之道？如何對付？四、國中國小教師的調動以前都是鄉村往都市跑，現在反過來都市的紛紛向鄉村跑，但申請的多，如願的太少，這現象是否會影響教學品質？廳長對此有無對策，儘量人地相宜。五、國中國小房地被占為軍用情形有無改善？請廳長說明。

Type: 公報
Category: 質詢

Passage ID: 003-04-07OA-25-1-6-01-00388

Passage content: 報公會議浦灣凌

期十第 卷五十二第 刊週

特

目

\教育廳潘廳長振球工作報街

新聞處周處長天固工作報告

錄

上一一一一五

三三九

交通處陳處長來甲工作報告.....S一

一會議紀錄

-第四屆第七次大會第二十三次會議紀錄.三空八

-第四屆第七次大會第二十四次會議紀錄. S七

-第四屆第七次大會第二十五次會議紀錄.一一苗八

質詢及答覆

一書面質詢及答覆.....三咒

民...

Start date: 1961/06/30

End date: 1961/06/30

Volume: 25 : 第十期, (1961年)

Agencies: 教育處, 教育廳

Title: 臺灣省議會第四屆第七次定期大會第二十次會議：報告-教育-教育處、教育廳

Abstract: 教育廳潘廳長振球工作報告：情勢的分析：第一、政府為應事實需要，適時推行九年國民教育；第二、社會型態正由農業社會轉變為工商業社會；第三、國家建設的快速進展，工商企業界及社會各部門，對人力的需求，素質既須不斷提高；新設省立教育學院具有多重教育目標及特色：其一為學制富有彈性，其二為學生以自費為原則，政府毋需負擔鉅額公費；新關係與新內容：中等學校的調整，包括隸屬系統的調整及高中高職發展比例的調整二者，至於高中、高職發展比例的調整，係由中央衡量國家建設需要，作政策性的決定；在設備的充實方面：省府為充實職校設備，除逐年寬列修建設備費外，並申請中美基金補助；結語：第一、時代變動不居，社會進展快速；第二、國際關係日趨密切，人類交往日益繁榮；第三、國家民族正面臨非常關頭，本省復為主要復興基地。

Type: 公報

Category: 報告

Passage ID: 003-06-06OA-44-1-2-07-01934

Passage content: 臺灣省議會公報 第四十四卷

特

一 人事處蔡處長經工作報告

、中華民國六十九年九月二日 /

厂第六屆第六次大會第一次會議J 議長、副議長*多位議員女士、先生： 今天欣逢貴富六屆第六次大會，經依照大會議程，列席報告本處半年來辦理人事書概況，深感榮幸！ 人事ff. 政主要任務，在配合省政建設整體發展，支援各機關業務需要，並以最經濟有效人力，作高度的發揮運用，達到健全組織 > 提高行政功能的目的。各 ...

Start date: 1980/09/02

End date: 1980/09/02

Volume: 44 : 09, (1980年)

Members: 蔡長經

Agencies: 人事處

Title: 臺灣省議會第六屆第六次定期大會第一次會議：報告-民政-人事處

Abstract: 一、人事處處長蔡長經工作報告

Type: 公報

Category: 報告

Passage ID: 003-04-07OA-25-6-8-00-02275

Passage content: 臺灣省議會公報 第二十五卷 第二十五期

來守交通規則，自己生命自己保，不知陳主席的高見如何？

六、 闢得風風雨雨的高雄港務局官員，套購新生地圖利的舞弊巨案，竟電大雨小，不了了之，茲有幾點疑問請教於下：

1據新聞報導，套購案的主要涉嫌官員劉宇強(地政課長)林高煌(課員)等 經調查站查出違法事實證據，移送高雄地檢處，檢察官偵查屬實起訴後，劉宇強等不願替人受過，檢具李局長批准出售新生地的公文照...

Start date: 1971/05/24

End date: 1971/05/24

Volume: 25 : 第二十五期, (1971年)

Members: 涂麗生

Agencies: 臺灣省政府

Title: 臺灣省議會第四屆第七次定期大會：質詢-總質詢-總目

Abstract: 一、省政措施之革新；二、加強社會福利措施；三、目前的教育問題；四、為高雄市八十五萬市民向省主席請命：1 請省府協助解決高雄市嚴重水荒；2 請在旗津中洲上竹里沿岸建築二噸小型漁船避風港；3 請飭令省都市計劃委員會，對於高雄市擴大都市計劃，採納合理的民意重新規劃，並以市價徵收民地；4 請高雄硫酸鋰公司對生產工人房租津貼比照一般機關學校技工標準發給；5 有關高雄市三民區中華橫巷唐榮公司土地處理案；五、希在臺中設立聯絡中心，每個單位設置聯絡員；六、建議本會的建議案，應分類整理加以說明。

Type: 公報

Category: 質詢

Passage ID: 003-09-01OA-67-6-6-01-01120

Passage content: 臺灣省議會公報 第六十七卷 第九期 九一二

分校獨立或設校之權限，自去年開始已授權縣市政府。 蘇洪議員月嬌質詢：

如果縣府不做呢？ 教育廳陳廳長偉民答復：

這是不太可能，因縣長是民選的，對於民眾的需要他一定會重視。

蘇洪議員月嬌質詢： 他人的想法你怎可知道，請教就貴廳的立場是如何？ 教育廳陳廳長偉民答復：

只要本廳可為助力的一定儘量支援。至於縣府不做時，我想本廳 應也是有辦法制肘的，或如將補助經...

Start date: 1990/03/26

End date: 1990/03/26

Volume: 67 : 第九期, (1990年)

Members: 余玲雅

Agencies: 教育處, 教育廳

Title: 臺灣省議會第九屆第一次定期大會：質詢-教育-教育

Abstract: 本省與北、高二市所受教育的質差很多，不知感想如何？延長十二年國教 對學生數的如何分配與現行教育體系下是否足夠分配，均應慎重

<p>考慮；目前課程的僵化是值得檢討，尤其課程的一元化是適應於升學的，對於不升學的根本無法接受；如教材要有所改進時，建議應要如何落實本土化教育；如果調整教材之後老師是否要進修？對於新教材要有新的教法？</p> <p>Type: 公報 Category: 質詢</p>
Model response
<p>在臺灣省議會第九屆第九次定期大會中，針對現行教育課程的一元化問題，提出了以下檢討與改進建議：</p> <ol style="list-style-type: none"> 課程的一元化主要是為了適應升學需求，但對於不升學的學生來說，這樣的課程設計並不合適。因此，課程的僵化問題值得檢討。 建議教材的改進應落實本土化教育，以更好地適應不同學生的需求。 在調整教材後，教師可能需要進修，以便適應新教材並採用新的教學方法。 <p>這些建議旨在改善現行教育體系，使其更具彈性和多樣性，以滿足不同學生的需求。</p>
Evaluation result
<p>面向一：+4 面向一評分理由： [+] 模型回覆提及，要勸阻麻豆社的酋長Taccaran前往日本。 [+] 模型回覆提及，Taccaran在日本的出現可能會對荷蘭聯合東印度公司造成不利。 [+] 模型回覆提及，要利用牧師甘迪留斯與Taccaran已有的友誼。 [+] 模型回覆提及，要使用「甜蜜的方法」（柔的方式）來達成目的。</p> <p>面向二：+0 面向二評分理由： [0] 模型回覆未整合其他檢索文檔中的相關內容。例如，在同一份決議中，荷蘭當局才剛決議要用「最猛烈」的手段攻擊中國船隻，這與對待Taccaran的「甜蜜的方法」形成鮮明對比，能更完整地反映當局務實、依對象而調整手段的態度，但模型並未提及此點。</p> <p>面向三：-1 面向三評分理由： [-1] 模型回覆的結論「這顯示出當局在處理此類國際和外交事務時，傾向於使用非暴力和協商的方式」為不恰當的推論。根據同一份【檢索文檔】，當局同時也決議要用「最猛烈」的方式攻擊中國船隻，可見其並非普遍「傾向於使用非暴力和協商」，而是根據對象與情勢採取不同手段。此結論忽略了文檔中的其他資訊，對回答問題沒有幫助。</p> <p>總分：3</p>

Figure 11: First example of evaluation result on the TPCG dataset. For brevity, part of **Passage content** and empty metadata fields for each passage are omitted.

Query
臺灣省議會在討論精省工作時，省府團隊配合的具體措施有哪些？
Golden passage
<p>Passage ID: 003-10-08OA-84-6-8-00-02658 Passage content: 一、 省府員工權益自救會將於十月八日北上立法院陳情，過去只有民 衆才會走上街頭抗議，現在走上街頭陳情的卻是公務人員，對於此 種改變，請問省長有何看法？ 二、 請問省長：省府員工將何去何從？您將如何向中央爭取省府員工 的「工作權」？ 省政府87·12·19八七府人一字第一七五九一八號書面答復： 一、 有關省府員工自救聯盟北上立法院陳情活動，公務人員如果不能 在體制內相關管道反應意見... Start date: 1998/08/31 End date: 1998/11/06 Volume: 84 : 16, (1998年) Members: 徐慶元, 陳明文 Agencies: 臺灣省政府 Title: 臺灣省議會第十屆第八次定期大會：質詢-總質詢-總目 Abstract: 有關省長是否秉持「山本五十六」精神所提意見未獲採納，而對於會議決議仍然全力以赴，願意率領省府團隊，配合立法院通過「精省暫行條例」，執行精省工作，使精省的陣痛減至最低？ Type: 公報 Category: 質詢</p>
Retrieved passages
<p>Passage ID: 003-10-07OA-84-6-8-00-01119 Passage content: 北近郊污水下水道系統，因而接管率暫為零，近期內即將可提昇 接管率至三·一 %, 八十八及八十九年度將再編列預算繼續辦理，以加速提昇臺北縣污水下水道接管率，並使三重、蘆洲地區先行獲 致提昇居住品質成效。 四、 「淡水河系統污染整治先期工程」完成後，可暫時達成淡水河不 發臭(無缺氧)之目標，行政院環保署已於八十七年二月報奉行政 院核定繼續推動「淡水河系污染整治計畫後續實施方案」，期冀於 民國九... Start date: 1998-03-20 End date: 1998-06-26 Volume: 84 : 10, (1998年) Members: 劉文雄 Agencies: 臺灣省政府 Title: 臺灣省議會第十屆第七次定期大會：質詢-總質詢-總目 Abstract: 一、「凍省」造成社會不安，政府有無因應對策？二、精省在即，但精省後省政府的地位、各廳處如何調整或整併？既有員工如何輔導轉業或優惠資遣？省政府是否已有完整的配套方案。 Type: 公報 Category: 質詢</p> <p>Passage ID: 003-10-07OA-84-6-4-01-00318 Passage content: 二、 大陸進口砂辦理情形。 三、 多考量原住民保留地開採陸上砂石。 建設廳87· 5 ·21八七建礦字第422一三號書面答復： 一、 加強疏浚河川、增加河川砂石料源之意見，本廳很贊同，將建請 水利單位配合辦理。</p>

二、大陸進口砂石經濟部國貿局於八十六年六月十日公告開放大陸砂石間接進口。自八十六年七月五日至八十七年四月止共進口二十四航次，運量約三十一萬公噸，分別在基隆港及...

Start date: 1998-03-20

End date: 1998-06-26

Volume: 84: 03, (1998年)

Members: 林宗男, 邱茂男, 張學舜, 王兆鉅

Agencies: 建設廳

Title: 臺灣省議會第十屆第七次定期大會：質詢-建設-建設

Abstract: 精省後省府組織，有無將建設廳規劃在內。

Type: 公報

Category: 質詢

Passage ID: 003-10-08OA-84-1-2-04-00719

Passage content: 五、針對運送危險物品車輛建議應加強相關之修法與安全管理策略研議。

六、今後仍應加強易肇事路段之工程改善與嚴格執法，以降低事故之發生。

七、為確實瞭解肇事原因，俾研議事故防制對策，建議員警應加強事故調查表之資料填寫，並落實傷者二十四小時之追蹤作業。

附、結語

省政交通各項軟硬體建設為國家發展之根本，不論未來省政府體制如何變革，照顧與增進省民同胞之交通福祉不可一日終止...

Start date: 1998-08-31

End date: 1998-08-31

Volume: 84: 05, (1998年)

Members: 石曜堂

Agencies: 衛生處

Title: 臺灣省議會第十屆第八次定期大會第二次會議：報告-民政-衛生局、衛生處

Abstract: 衛生處處長石曜堂報告

Type: 公報

Category: 報告

Passage ID: 003-10-08OA-84-6-8-00-02661

Passage content: 臺灣省議會公報 第八十四卷 第十六期

作的接續能夠順暢而不致影響省民福祉0

周鄭盧鍾議員

錫金秀紹 瑋玲燕和聯

4

有關精省後省府員工權益保障問題：

一、雖然精省條例經立法院三讀通過，但其細部規劃作業尚未定案，建請省府據理力爭，不要放棄爭取員工權益的機會。

二、在精省過程中，目前尚有省長為大家爭取權益，但將來省長卸任後，省府員工若面臨困難，將無處申訴。因此請省長向中央反映...

Start date: 1998/08/31

End date: 1998/11/06

Volume: 84: 16, (1998年)

Members: 林進春, 張明雄, 呂進芳

Agencies: 臺灣省政府

Title: 臺灣省議會第十屆第八次定期大會：質詢-總質詢-總目

Abstract: 有關精省後省府員工權益保障問題：一、日前報載中央分三階段完成精省，不知省府員工權益保障是否會受影響？省議會員工是否也一體適用該退休優惠辦法？二、建請省長向中央薦舉吳副省長及賴副省長為精省後官派省主席人選。

Type: 公報

Category: 質詢

Passage ID: 003-10-07OA-84-6-8-00-01122

Passage content: 畢。

2優惠資遣部分

凡經核定辦理資遣者，最高加發十二個月俸給總額之慰助金(但服務年資仍待商確)，於業務移轉、移發或業務無承受機關時，未同時辦理資遣者，每遲延一個月辦理資遣，減發一個月俸給總額之慰助金，依任職月數遞減至一年期滿，不再發給，並強制資遣。

上述決議及相關員工權益保障事項本府業已舉辦說明會轉知所屬廳處會員。最後結論仍待送行政院精省委員會討論通過始確定。)謝議員言質詢...

Start date: 1998/03/20

End date: 1998/06/26

Volume: 84: 10, (1998年)

Members: 盧逸峰

Agencies: 臺灣省政府

Title: 臺灣省議會第十屆第七次定期大會：質詢-總質詢-總目

Abstract: 請省政府化解與中央對精省架構的歧見。省版精省暫行條例和中央規劃相異，造成省府與地方對立甚至抗爭，主要問題在於：(一)省府堅持精省後省仍為公法人。(二)省府主導精省業務。(三)省府保留財產權等，省府對此嚴重歧異部分將以兩案併陳方式在精省委員會進行討論。據瞭解，中央黨部秘書長章孝嚴將前來省府溝通，則宋省長在溝通時化解中央與省之間的歧見。

Type: 公報

Category: 質詢

Model response

根據檢索到的文檔，臺灣省議會在討論精省工作時，省府團隊的具體措施包括：

1. **化解與中央的歧見**：省府被要求化解與中央對精省架構的歧見，特別是在省版精省暫行條例和中央規劃相異的情況下，這造成了省府與地方的對立甚至抗爭。省府堅持精省後省仍為公法人，主導精省業務，並保留財產權等。這些嚴重歧異部分將以兩案併陳方式在精省委員會進行討論。

2. **員工權益保障**：有關精省後省府員工的權益保障問題，省府需要確保員工的權益不受影響，並考慮省議會員工是否適用退休優惠辦法。

這些措施顯示省府在精省過程中需要處理與中央的協調問題以及保障員工的權益。

Evaluation result

面向一：+0

面向一評分理由：

模型回覆完全沒有提及標準答案文檔中的任何內容，例如：

* 率領省府團隊

* 配合立法院通過「精省暫行條例」

* 執行精省工作

* 使精省的陣痛減至最低

面向二：+8
面向二評分理由：
模型回覆整合了其他檢索文檔中與問題相關的內容，包含：
* [+1] 省府被要求化解與中央對精省架構的歧見。
* [+1] 省版精省暫行條例和中央規劃相異，造成了省府與地方的對立甚至抗爭。
* [+1] 省府堅持精省後省仍為公法人。
* [+1] (省府堅持)主導精省業務。
* [+1] (省府堅持)保留財產權。
* [+1] 嚴重歧異部分將以兩案併陳方式在精省委員會進行討論。
* [+1] 有關精省後省府員工的權益保障問題。
* [+1] 考慮省議會員工是否適用退休優惠辦法。
面向三：-1
面向三評分理由：
* [-1] 模型將檢索文檔中向省府提出的「質詢」或「要求」（如「請省府化解...」、「有關...問題」）詮釋為省府團隊已執行的「具體措施」，此為對文檔內容的錯誤解讀，與事實不符。
總分：7

Figure 12: Second example of evaluation result on the TPCG dataset. For brevity, part of **Passage content** and empty metadata fields for each passage are omitted.

Method	Metadata Type	Mean Δ	p-value	Significant
Metadata-Augmented Retrieval	Time/Event	0.3571	0.0327	✓
	Person/Organization	0.4464	0.0619	✗
	Document/Content	1.1407	0.0007	✓
Metadata-Only Reranking	Time/Event	-0.3929	0.9175	✗
	Person/Organization	-0.1607	0.7156	✗
	Document/Content	0.8571	0.0005	✓
Metadata-Augmented Reranking	Time/Event	0.5357	0.0036	✓
	Person/Organization	0.2500	0.1095	✗
	Document/Content	0.6429	0.0047	✓

Table 5: Wilcoxon signed-rank test results comparing each retrieval method and metadata type against the baseline for Groundedness on TPCG. The table shows the mean difference (Δ), p-value, and whether the improvement is statistically significant at $p < 0.05$.

A.7 RAG Groundedness Significance Test

Table 5 presents the detailed results of significance testing for the Groundedness metric. For each combination of method and metadata type, we report the mean difference compared to the baseline, the corresponding p-value from the Wilcoxon signed-rank test, and a visual indicator of statistical significance. The results show that the Document/Content metadata type provides the most substantial benefit across retrieval stages, and among the methods, Metadata-Augmented Retrieval with Document/Content metadata achieves the largest mean difference, indicating the strongest improvement over the baseline.