Journal of Future Artificial Intelligence and Technologies

E-ISSN: 3048-3719

Research Article

LLM-Compiled Wikis for Cross-Institutional Policy Retrieval: A Benchmark on Chinese University Grant Policies

Bailing Zhang *

School of Computer Science and Data Engineering, NingboTech University, Ningbo 315100, China;
e-mail : bailing.zhang@nit.zju.edu.cn
* Corresponding Author : Bailing Zhang

Abstract: We investigate whether pre-compiling a document corpus into an LLM-compiled wiki can improve cross-institutional policy-conflict queries that naturally exhibit a tabular comparison structure. We define a task (T3a) in which a system must identify, for a focal institution and policy dimension, both the focal rule and the institutions that disagree with it, supported by article-level citations. We construct a 28-query benchmark from 22 Chinese university grant policy documents across 19 institutions. Gold citations are extracted automatically using an LLM-based pipeline with mechanically auditable post-checks and validated through a human audit of six representative queries (172 gold citations audited, with 93.0% manually verified as correct or partially supporting; see Section 4.2.1). We compare three systems that share the same language-model generator: no-retrieval, BM25 retrieval-augmented generation, and a wiki-based retrieval system whose pages are compiled offline from the same corpus. The wiki system achieves a citation recall of 0.242 ± 0.21, compared with 0.095 ± 0.12 for BM25 and 0.023 ± 0.03 for no-retrieval; all pairwise differences are statistically significant ($p < 10^{-3}$, paired Wilcoxon test). Crucially, the wiki system achieves this performance while consuming only approximately 29% of BM25's input-context tokens, indicating that the observed improvement is not driven by larger context volume. A failure-mode analysis shows that hallucinated citations decrease monotonically across the three systems (75% → 54% → 39%). We further identify two LLM-induced failure modes in the compiled wiki that are not detected by citation-level validation. We frame these findings as a design insight into task-representation alignment for cross-document comparison queries—a proof-of-concept for compile-then-query arrangements within this task family rather than a general claim of superiority over modern retrieval systems.

Keywords: BM25; Chinese NLP; Citation Faithfulness; Cross-Document Comparison; Knowledge Compilation; Large Language Models; Policy Documents; Retrieval-Augmented Generation.

Received: April, 27th 2026
Revised: May, 24th 2026
Accepted: May, 27th 2026
Published: June, 1st 2026
Curr. Ver.: June, 1st 2026

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY SA) license (https://creativecommons.org/licenses/by-sa/4.0/)

1. Introduction

Many real-world applications of language models require synthesizing information across multiple documents that describe the same class of object—a portfolio of contracts, a set of regulatory filings, or the grant-management policies of a collection of universities. When a user asks how one institution's rule compares with those of others, a language model supported by retrieval-augmented generation (RAG) [1]–[3] must implicitly perform cross-document inference at each query: retrieve the relevant articles, align their meaning, identify the differences, and cite the supporting evidence. This inference is not persisted; each new query repeats the same process.

An alternative arrangement, recently proposed by Karpathy [4], is to compile the corpus once, offline, into a persistent structured wiki—a collection of Markdown pages with pre-resolved cross-references and explicit comparison tables—and then query the wiki at inference time. While the original proposal describes the pattern, it does not evaluate it empirically. In this paper, we test the approach on a controlled task and frame the result as a design insight

into task-representation alignment rather than a general claim of superiority over retrieval-based approaches.

We define a cross-institutional policy-conflict query task, which we call T3a: given a focal institution and a policy dimension, the system must identify the focal institution's rule on that dimension, the institutions whose rules differ, and the supporting article-level citations. The task is naturally tabular—the most direct human-readable answer is a comparison table with one row per institution—which is a structural property that we explicitly exploit and discuss as a scope condition in Section 5. We construct a benchmark of 28 such queries from a corpus of 22 Chinese university grant-management policy documents spanning 19 institutions. We evaluate three matched systems: a parametric-only baseline without retrieval, a BM25-based RAG baseline using a character n-gram tokenizer appropriate for Chinese text, and a Wiki RAG system that retrieves a pre-compiled comparison page. All three systems use the same underlying language model (glm-4-plus). Outputs are evaluated at the (doc_id , article_number) level against LLM-extracted gold citations, providing a mechanically auditable metric that does not depend on matching natural-language paraphrases or on an LLM-as-judge protocol [5], whose reliability for long-form factuality remains an open question.

On the 28-query benchmark, the Wiki RAG system achieves a mean citation recall of 0.242, compared with 0.095 for BM25 RAG and 0.023 for the no-retrieval baseline. All pairwise differences are statistically significant under a paired Wilcoxon signed-rank test ($p < 10^{-3}$), with bootstrap 95% confidence intervals excluding zero. Three secondary analyses help explain the source of this improvement. First, a retrieval-only analysis shows that the compiled wiki contains more gold citations than BM25 retrieval even at $k = 20$, suggesting that the advantage arises from structural pre-compilation rather than retrieval breadth alone. Second, a context-budget analysis shows that Wiki RAG achieves these results using only approximately 29% of the prompt-input tokens required by BM25 RAG, ruling out the alternative explanation that the system simply receives more information. Third, a failure-mode analysis of all 84 predictions shows that hallucinated citations decline monotonically (75% → 54% → 39%) as retrieval support improves. To further strengthen validity, we conduct a manual audit of six representative queries (one per policy dimension) and report the resulting agreement rate. Our contributions are as follows:

- A defined T3a task and a 28-query benchmark on Chinese higher-education grant-management policies, released with article-level gold citations and a human-audited sub-benchmark;
- An implementation of the LLM-wiki compilation pattern with a citation-checking post-condition, together with a side-by-side comparison against a matched BM25 RAG baseline using the same language-model generator and a quantified input-context budget;
- Statistical evidence, supported by retrieval-only, context-budget, and failure-mode analyses, that pre-compiled wiki pages outperform flat retrieval for the specific task family of cross-institutional policy-conflict queries considered in this benchmark;
- Explicit documentation of two LLM-induced failure modes in the compiled wiki that are not captured by citation-level validation—one at the display-name level and one at the focal-attribution level—thereby motivating future work on semantic citation verification.

The remainder of the paper is organized as follows. Section 2 reviews related work on retrieval-augmented generation, knowledge compilation, and citation faithfulness. Section 3 formalizes the T3a task and describes the wiki-compilation pipeline, the three baseline systems, and the evaluation metrics. Section 4 presents the experimental results, including corpus statistics, the human audit of benchmark gold citations, context-budget comparisons, the main evaluation metrics, retrieval-only analysis, failure-mode analysis, and a detailed case study. Section 5 discusses why structured wikis outperform flat retrieval for this task family, outlines the scope conditions of the findings, and examines threats to validity. Section 6 concludes the paper.

2. Related Work

The technical foundation of our work rests on the transformer architecture [6] and the family of large language models built upon it, including BERT [7], T5 [8], GPT-3 [9], GPT-4 [10], and LLaMA-2 [11]. Three strands of prior work are particularly relevant to this study.

2.1. Retrieval-Augmented Generation

Retrieval-augmented generation (RAG) [1] extends a language model with a non-parametric memory that is accessed at inference time, addressing the inherent limitations of storing knowledge solely within model parameters. A comprehensive survey of the resulting ecosystem of retrievers, fusion strategies, and generation techniques is provided by Gao et al. [2]. A complementary survey by Sharma and Bhattarai [3] focuses on architectures, research challenges, and emerging frontiers, including retrieval-design trade-offs, hallucination behavior, modular RAG architectures, and the challenges of grounding evaluation in modern RAG systems. BM25 [12] remains the canonical sparse-retrieval baseline against which new retrieval techniques are commonly evaluated, and we adopt it for that reason. The limitation of query-time RAG targeted in this work is well recognized in the literature: the language model must repeatedly reconstruct cross-document structure at inference time without access to a persistent artifact that captures previously inferred relationships.

2.2. LLM-Based Knowledge Compilation

In April 2026, Karpathy [4] published an idea file describing a pattern he found useful in practice: rather than retrieving from raw documents at query time, a language model incrementally compiles a structured, interlinked Markdown wiki from the source corpus and subsequently answers queries using the compiled wiki. The resulting wiki serves as a persistent artifact in which cross-references and contradictions are resolved during compilation rather than rediscovered for each query. The proposal itself is conceptual rather than empirical. It motivates the central question addressed in this paper: whether a compile-then-query arrangement can outperform a matched query-time RAG system on a measurable task. To the best of our knowledge, this work provides the first quantitative evaluation of that pattern using a controlled benchmark.

2.3. Citation Faithfulness and Long-Context Inference

A parallel line of research investigates the reliability of citations in LLM-generated outputs. Menick et al. [13] use reinforcement learning from human preferences to train a model that cites supporting quotations and argue that explicit citation is a key component of trustworthy language-model outputs. Liu et al. [14] demonstrate that language models do not reliably utilize information distributed across long contexts; performance degrades substantially when relevant information appears in the middle of lengthy prompts, even for models designed to support long contexts.

Building on these directions, recent work has expanded the landscape of citation-aware and retrieval-aware generation. ALCE [15] introduces an automatic benchmark for evaluating LLM citations across fluency, correctness, and citation quality. Self-RAG [16] introduces reflective tokens that allow the model to determine when retrieval is necessary. In-Context RAG [17] frames retrieval as prompt augmentation without retraining, while Graph RAG [18] organizes documents into graph-structured summaries to support query-focused reasoning. Our citation-based evaluation builds upon the first line of work, while the mechanistic interpretation developed in Section 5.1 relates to the second: compiling a wiki can be viewed as relocating relevant information from the middle of long source documents into a focused comparison structure where adjacency makes the required inference more local.

2.4. Policy-Text NLP in Chinese

To the best of our knowledge, no prior benchmark targets cross-institutional policy-conflict queries in Chinese at the article-level citation granularity considered in this study. Our corpus consists of 22 grant-management policy documents collected from 19 Chinese universities and assembled from publicly available official sources specifically for this task. The corpus is released as part of the evaluation archive.

Our work is complementary to domain-specific pre-trained models such as LEGAL-BERT [19] and to broader information-retrieval benchmarks such as BEIR [20]. Whereas these approaches focus on training or evaluating general-purpose retrievers and encoders, our study fixes the language-model generator and compares retrieval substrates with different structural commitments.

3. Proposed Method

3.1. Problem Setting

We define a cross-institutional policy-conflict query task, denoted as T3a. The input consists of a corpus $D = \{d_1, d_2, \ldots, d_n\}$ of policy documents, with one or more documents associated with each institution in a finite institution set I. Each document d_i is segmented into articles $a_{i,n}$, indexed by article number n, following the standard "第 N 条" numbering convention commonly used in Chinese administrative documents. We further define a small closed set of policy dimensions Δ of interest. In our benchmark, these dimensions are indirect cost, labor cost, equipment, amount thresholds, percent thresholds, and horizontal-versus-vertical funding.

A T3a query is a tuple $q = (i^*, \delta)$ with $i^* \in I$ the focal institution and $\delta \in \Delta$ the policy dimension. The expected response has three components:

1. a natural-language statement of focal institution i^*'s rule on dimension δ;
2. the set F of other institutions whose rule on δ differs from i^*'s; and
3. a set of supporting citations C, where each citation (i, n) is a reference to article n of the policy document of institution i.

For evaluation we compare a predicted citation set C against a gold citation set C^* using tuple-level precision, recall, and F1. We designate citation recall as our primary metric; Section 3.4 explains why the alternative conflict-based metric is structurally unsuited for significance testing on this benchmark.

3.2. Wiki Compilation Pipeline

Our wiki-compilation pipeline transforms a raw policy corpus into a small collection of comparison pages, one for each policy dimension δ. Each page contains a comparison table with one row per institution and article-level citations. Figure 1 illustrates how this compilation stage is positioned between the raw corpus and the retrieval substrate used during inference. Algorithm 1 summarizes the per-dimension compilation process. The six pipeline components are described below.

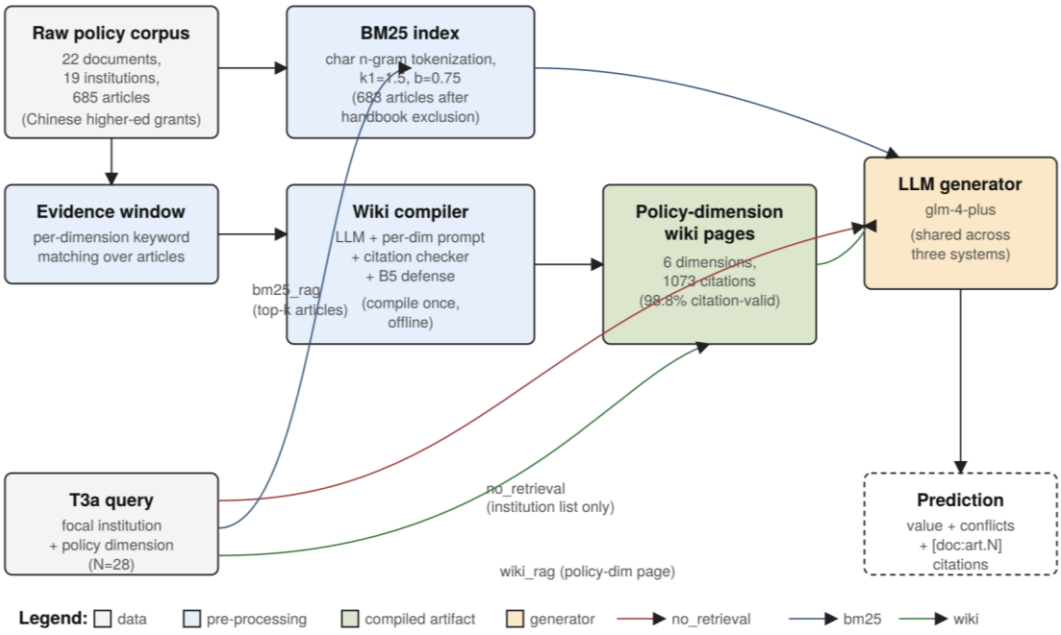

Figure 1. System architecture. The raw policy corpus is processed along two paths: (i) into a BM25 index used at query time by the BM25 RAG baseline, and (ii) through the wiki compiler, which uses an LLM to produce policy-dimension comparison pages stored as Markdown and retrieved at query time by Wiki RAG. The no-retrieval baseline receives only the institution list and no source text. All three systems share the same underlying language-model generator (glm-4-plus).

Article Segmentation and Handbook Exclusion

Each document is segmented at article boundaries by locating the canonical "第 N 条" heading pattern. The segmenter supports both Arabic and Chinese numerals and produces a document-local sequence of articles $a_{i,1}, a_{i,2}, \ldots$. One document in the corpus uses a Q&A handbook format rather than numbered articles. Because its structure is incompatible with article-level citation, it is excluded from both the BM25 index and the wiki-compilation pipeline. This exclusion is controlled through a `doc_format` field in the institution registry and results in 683 of 685 articles being indexed.

Evidence Window

For a given policy dimension δ and document d_i, the evidence window selects up to eight articles that are most relevant to δ using keyword matching. Each dimension is associated with a small manually curated keyword list. For example, the `indirect_cost` dimension includes keywords such as "间接费", "管理费", and "提取". The keyword lists were developed iteratively during corpus exploration with the goal of achieving complete coverage of the canonical Chinese terminology associated with each dimension. The evidence window acts as a per-dimension, per-document filter. It reduces the amount of text presented to the LLM from hundreds of articles to a small set of highly relevant candidates. This not only lowers computational cost but also reduces contextual load during compilation, thereby mitigating the long-context degradation documented by Liu et al. [14].

Per-Dimension Comparison-Table Prompt

For each policy dimension, we maintain a YAML-defined prompt that instructs the LLM to generate a Markdown comparison table using a fixed schema: institution, rule, specific value, source citation. Each citation must follow the format: `[doc_id: art. N]`, where the referenced article must appear within the evidence window. The prompt explicitly instructs the model to mark institutions that do not regulate the target dimension as silent rather than fabricating a rule. During compilation, the temperature is fixed at 0.2 to allow limited recovery from lexical errors while preserving table structure and formatting consistency. The complete prompt template, evidence-window configuration, and YAML schema are released with the code archive (see the Data Availability Statement).

Citation Checker

Each generated page is parsed to extract all citation tokens of the form: `[doc: art. N]`. For every citation, we verify (i) whether `doc_id` corresponds to a known institution in the registry; and (ii) whether article number N falls within the valid range for that document. The checker is intentionally conservative. It detects the two most common hallucination categories—unknown institution identifiers and out-of-range article numbers—but does not verify whether the cited article actually supports the associated claim. Detecting such semantic citation errors requires a different mechanism. Section 4.7 presents one example of a failure that bypasses the current checker, while Section 5.2 discusses future directions for semantic citation verification.

B5 Defense

If the citation checker detects any error, the generated page is returned to the LLM together with a structured error report requesting a corrected version. We allow up to $r_{max} = 2$ correction attempts and cache only the final accepted output. Across the corpus, the process converges within at most one retry for all pages, producing a final citation-validation pass rate of: 1060/1073=98.8%. The remaining thirteen failures correspond to unknown-article references in two institutions whose actual article counts are close to the extrapolated values selected by the LLM. Because both values appear plausible, the defense loop does not reliably eliminate these cases. We therefore retain them and document their behavior in the failure-mode analysis.

SQLite Result Cache

Every LLM call is identified by the tuple: (model, messages, temperature) and its response is stored in a local SQLite cache. The cache follows a conservative policy: outputs that fail any post-check are not persisted. During repeated experimental runs, cache hit rates typically exceed 90%, making iterative evaluation feasible at minimal additional cost. The final product of the pipeline is a collection of Markdown comparison pages, one per policy

dimension. Together, these pages contain 1,073 article citations spanning 19 institutions and 6 policy dimensions, forming a compact and auditable retrieval substrate that can be queried either by human readers or, as in our experiments, by a downstream LLM acting as a question-answering agent.

Algorithm 1. Wiki Compilation for a Single Policy Dimension δ

INPUT: corpus D, policy dimension δ, LLM M, maximum retries r_{max}
OUTPUT: comparison page for δ (Markdown)

$E_\delta \leftarrow \text{EvidenceWindow}(D, \delta)$
page $\leftarrow M(\text{prompt}_\delta, E_\delta)$
for $r = 1$ to r_{max} do
    $(ok, errs) \leftarrow \text{CheckCitations}(\text{page}, D)$
    if ok then
        return page
    page $\leftarrow M(\text{prompt}_{\delta^+}, E_\delta, errs)$ // B5 defense
return page, flag unresolved

3.3. Baseline Systems

We compare three systems. All three share the same underlying language model (glm-4-plus from the Zhipu AI family), the same parsing and scoring pipeline, and the same benchmark queries. They differ only in the context supplied to the LLM at inference time. This matched design isolates the effect of the retrieval substrate from that of the generator. Section 4.3.1 quantifies the resulting context budgets and shows that the Wiki RAG system, in fact, uses substantially fewer input tokens than the BM25 RAG system.

no_retrieval

The LLM receives the query together with the closed list of 19 institutions (represented by 21 institution identifiers, including version suffixes for institutions associated with multiple documents), but no source text. The purpose of this baseline is not to compete with retrieval-based systems but to establish a lower bound on performance. It measures what the generator can produce using only its parametric knowledge, without any document grounding. We provide the institution list rather than relying on the model to recall institution names from pre-training so that all three systems operate within a common identifier space. This makes the predicted conflict sets directly comparable across systems.

bm25_rag

The LLM receives the query together with the top-k articles retrieved from a BM25 index constructed over the policy corpus [12]. Our BM25 implementation uses a character-level n-gram tokenizer ($n \in \{2, 3\}$), which is appropriate for Chinese text, where word boundaries are not explicitly marked, and is consistent with prior work on Chinese pre-trained language models [21]. Standard BM25 parameters ($k_1 = 1.5, b = 0.75$) are used throughout. The index is built over the 683 articles retained after excluding the handbook-format document. Unless otherwise stated, the default value $k = 10$ is used in the main experiments; the effect of varying k is examined in Section 4.5. We choose BM25 rather than dense retrievers such as DPR [22], ColBERT [23], or BGE-M3 [24] because BM25 requires no task-specific fine-tuning and provides a well-established sparse-retrieval baseline. A direct comparison with dense retrieval methods is left as future work and discussed in Section 6.

wiki_rag

The LLM receives the query together with the complete policy-dimension wiki page generated by the compilation pipeline described in Section 3.2. The page corresponds to the policy dimension δ associated with the query. No additional retrieval is performed at inference time. The wiki page is a single pre-compiled artifact generated once offline and subsequently reused for all relevant queries. This implementation represents a direct operationalization of the compile-then-query pattern proposed by Karpathy [4].

All three systems are instructed to produce outputs in the same format: a short free-text statement describing the rule, followed by a JSON block containing a value field and a conflicts list of institution identifiers. Citations appear inline within the free-text portion using the [doc: art. N] format.

3.4. Evaluation Metrics

Given a predicted citation set C and a gold citation set C^*, we compute tuple-level citation precision, recall, and F1 using the standard definitions shown in Equations. (1) – (3):

$$\text{precision} = \frac{|C \cap C^*|}{|C|} \tag{1}$$

$$\text{recall} = \frac{|C \cap C^*|}{|C^*|} \tag{2}$$

$$\text{F1} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{3}$$

Citations are matched at the $(\text{doc_id}, \text{article_number})$ level. No partial credit is awarded for matching the institution while failing to match the corresponding article.

Primary Metric: Citation Recall

Our central hypothesis is that Wiki RAG retrieves and surfaces more gold-supporting articles than the flat-retrieval baseline BM25 RAG. Citation recall provides the most direct measure of this capability. Citation recall is also robust to minor variations in the LLM's natural-language rendering of policy rules. Such variations may affect value-matching metrics but do not alter the cited articles themselves.

Why We Do Not Use Conflict-Based Metrics

A seemingly natural alternative is to evaluate the predicted conflict set F against the gold conflict set F^* using set-level precision, recall, and F1 over institution identifiers. However, during pilot evaluation we found this metric to be structurally biased in our benchmark. The gold extraction procedure produces institution-level conflict labels, and because Chinese universities frequently differ on at least one policy sub-rule, the gold conflict set for a typical query contains between 12 and 18 of the remaining institutions. As a consequence, a system can achieve a high conflict-based F1 score simply by predicting nearly all institutions as conflicts, even when it provides no meaningful evidence. We observe precisely this behavior in the no-retrieval baseline, which has no access to source documents yet achieves a conflict F1 score of 0.49, exceeding the 0.20 achieved by BM25 RAG. Citation-based metrics do not admit this trivial strategy. To receive credit for a citation match, a system must correctly identify both a specific document and a specific article number. Such information cannot be inferred from the institution list alone and therefore provides a substantially stronger grounding signal than institution-level conflict prediction.

4. Results and Discussion

4.1. Corpus

Our corpus consists of 22 grant-management policy documents collected from the official websites of 19 institutions, where these documents are publicly available. Throughout this paper, institution refers to the 19 distinct universities, whereas document refers to the 22 source files. Three institutions contribute multiple documents and are represented using version-suffixed identifiers such as `zjdx_01`, `zjdx_02`, and `zjdx_03`, yielding a total of 21 institution identifiers. This identifier space is shared by all three systems at inference time. The documents cover the standard spectrum of Chinese university research-funding regulations, including budget categories, reimbursement procedures, project lifecycle management, and special-case handling. After article segmentation, the corpus contains 685 articles.

One document, a grant-management Q&A handbook, does not follow the numbered-article structure used by the remaining 21 documents. Because our citation scheme relies on article numbers as atomic reference units, this document is excluded from both the BM25 index and the wiki-compilation pipeline. After exclusion, 683 of the 685 articles remain indexed. Corpus statistics are summarized in Table 1.

Table 1. Corpus statistics.

Item	Count
Source documents	22
Distinct institutions	19
Institution identifiers (including version suffixes)	21
Documents excluded (handbook format)	1
Articles (segmented)	685
Articles indexed by BM25	683
Wiki pages (policy dimensions)	6
Citations emitted by wiki compilation	1,073
Citation-checker pass rate	98.8%
T3a benchmark queries (PARSE_ERROR filtered)	28
Benchmark coverage (policy dimensions)	6

The citation-checker pass rate is computed over all citations emitted by the wiki compiler; a single citation may appear in multiple wiki pages.

4.2. Benchmark Construction

Because manually labeling citation gold for a multi-document policy corpus is expensive, we employ an automated extraction pipeline to construct the benchmark and explicitly acknowledge the implications of this choice. For each (focal institution, policy dimension) pair, we perform a dedicated extraction pass that reads only the articles associated with that institution within the corresponding evidence window. The LLM is then asked to generate:

- a short value summary;
- a structured list of rule components; and
- the supporting article-level citations.

The extractor is constrained to cite only articles that appear within the supplied evidence window, eliminating hallucinated article numbers by construction.

The benchmark initially contains 29 queries spanning six policy dimensions. One query contains a gold-value extraction parse error and is removed, leaving the final 28-query benchmark used throughout this study. We recognize that LLM-extracted gold is inherently weaker than manually annotated gold and explicitly acknowledge this limitation. We mitigate the associated risks in three ways. First, our primary metric—citation recall at the (doc_id, article_number) level—is mechanically auditable. Any predicted citation can be checked directly against the source corpus and judged independently of the benchmark gold.

Second, the extraction pipeline operates on a strictly smaller input than the systems under evaluation, processing one institution at a time using only that institution's evidence window. As shown in the audit presented in Section 4.2.1, the extractor exhibits a recall-oriented bias rather than a precision-oriented bias. In particular, it tends to include topically adjacent articles, especially within the equipment and amount-threshold dimensions. Such behavior uniformly depresses measured recall across all systems and therefore does not affect the relative comparisons that form the basis of our conclusions. Third, the same audit shows that the extracted gold achieves 93.0% precision when partial matches are counted and 73.3% precision under a strict correctness criterion.

Human Audit of LLM-Extracted Gold

To address concerns regarding the reliability of LLM-extracted gold citations, we conducted a manual audit of six representative queries, one from each policy dimension. The selected queries were chosen to span the full range of observed system performance. Specifically, the audit includes:

- the documented failure case discussed in Section 4.7 (t3a_indirect_cost_fddx_01);
- the best-performing case study examined in detail (t3a_labor_cost_hnkjxy);
- the best-performing query from each of the remaining four dimensions; and
- the weakest-performing dimension, in which Wiki RAG achieves zero recall across all five queries (t3a_horizontal_vs_vertical).

The audit was performed by the author using the original policy documents and a structured audit workbook released as supplementary material. For each gold citation, the cited article was inspected and assigned one of three verdicts:

- Correct: the article directly supports the claim;
- Partial: the article is relevant but supports only part of the claim or supports it narrowly;
- Incorrect: the article does not support the claim.

The audit results are summarized in Table 2.

Table 2. Human audit of LLM-extracted gold citations across six representative queries (one per policy dimension).

Audit Dimension	Citations Audited	Correct	Incorrect	Partial	Strict Precision	Correct + Partial
Indirect cost (t3a_indirect_cost_fddx_01)	13	9	3	1	69.2%	76.9%
Labor cost (t3a_labor_cost_hnkjxy)	19	16	1	2	84.2%	94.7%
Equipment (t3a_equipment_hnkjxy)	29	8	3	18	27.6%	89.7%
Amount thresholds (t3a_amount_thresholds_xxxy)	33	18	2	13	54.5%	93.9%
Percent thresholds (t3a_percent_thresholds_zjdx_01)	42	40	2	0	95.2%	95.2%
Horizontal vs Vertical (t3a_horizontal_vs_vertical_hnkjxy)	36	35	1	0	97.2%	97.2%
Overall	172	126	12	34	73.3%	93.0%

All 172 gold citations were manually verified against the source documents. Strict precision counts only "Correct" verdicts, whereas the final column counts both "Correct" and "Partial" verdicts.

Across all 172 audited citations, strict precision (counting only Correct verdicts) reaches 73.3% (126/172). When Partial verdicts are also counted, precision increases to 93.0% (160/172). Only 12 citations (7.0%) are judged outright incorrect. Per-query strict precision ranges from 27.6% in the equipment dimension to 97.2% in the horizontal-versus-vertical dimension. The equipment and amount-threshold dimensions exhibit particularly high rates of partial matches (18/29 and 13/33, respectively), indicating a recall-oriented bias in the extraction process. The extractor often includes articles related to asset management, procurement procedures, or fixed-asset attribution even when those articles do not explicitly define an equipment-cost or amount-threshold rule. Importantly, this bias affects all systems equally because every system is evaluated against the same gold set. Consequently, the bias reduces absolute recall values but does not alter the relative ranking of systems, which is the focus of our analysis.

The 12 incorrect citations are concentrated in five source documents, with five of them originating from fddx_01 (Fudan University, STEM document). Inspection reveals a recurring pattern: the extractor occasionally cites administrative or compliance-oriented articles—for example, fddx_01 Article 21 (reward-and-punishment mechanisms), fddx_01 Article 16 (misuse-prohibition clauses), or hbkjxy Article 6 (departmental responsibilities)—as evidence for substantive policy rules that they do not actually define. This behavior represents the same class of semantic-attribution failure illustrated by the focal-confusion case discussed in Section 4.7 for Wiki RAG. The audit independently confirms that similar failures also occur, albeit less frequently, within the gold-extraction pipeline itself. We discuss the implications of these findings for future semantic-verification mechanisms in Sections 5.2 and 6. Because the audit was conducted by a single author, inter-rater agreement statistics are not reported. To improve transparency, the complete audit workbook is released alongside the benchmark, allowing independent reviewers to verify every judgment against the publicly available source documents.

4.3. Implementation Details

All experiments were conducted on a single workstation equipped with an NVIDIA RTX 3090 GPU, running Windows 11 and Python 3.10 within a clean Anaconda environment. The language model used across all systems and experimental phases is glm-4-plus from

the Zhipu AI family, accessed through the zai-sdk Python package. We use the model version available through the Zhipu API during the first quarter of 2026; the corresponding API snapshot date for each experimental batch is recorded in the released results archive. Temperature is set to 0 during evaluation to ensure reproducibility and to 0.2 during wiki compilation to allow limited recovery from over-constrained prompts while maintaining stable table structure.

Every LLM call is cached using the tuple (model, messages, temperature) as a key and stored in a local SQLite database. Consequently, once the cache has been populated, re-running experiments incurs negligible additional cost. Output stability under cache hits is exact by construction. The complete codebase, benchmark gold, compiled wiki pages, raw evaluation outputs, prompt templates, evidence-window keyword lists, and analysis scripts are released as a unified archive (see the Data Availability Statement).

Context Budget Comparison

Because the three systems differ in both the structure and volume of context supplied to the LLM, a natural concern is that any observed performance difference may simply reflect unequal access to information. To address this possibility, we measure the prompt input-token budgets of all three systems. Token counts are estimated using a conservative Chinese-aware tokenization heuristic, in which each Chinese character is approximated as one token and English words as approximately 1.3 tokens, together with a fixed system-prompt overhead for each configuration. The results are summarized in Table 3 and visualized in Figure 2.

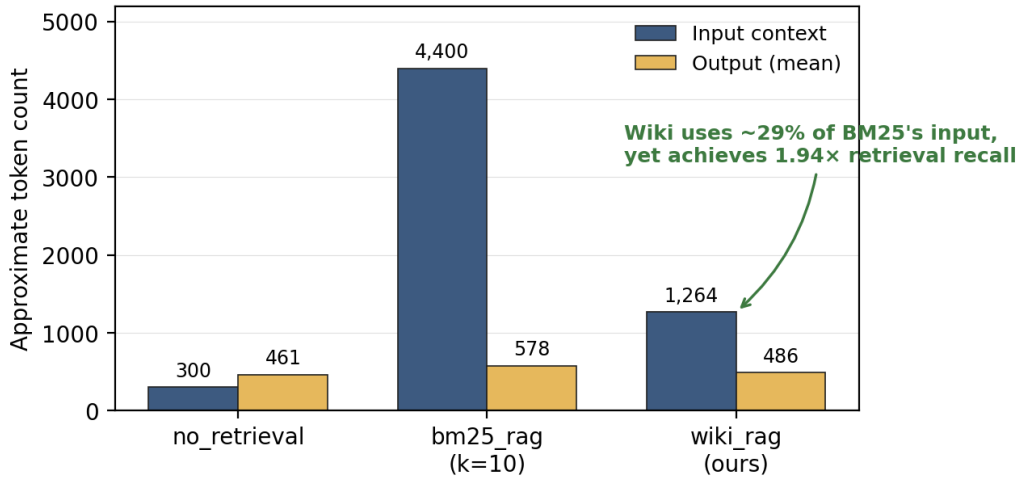

Figure 2. Approximate input-context and output-token counts per query across the three systems. The BM25 RAG baseline (k = 10) consumes approximately 4,400 input tokens (ten retrieved articles), compared with approximately 1,264 tokens for Wiki RAG (one pre-compiled comparison page). Thus, Wiki RAG operates using roughly 29% of the input context required by BM25 RAG while achieving substantially higher retrieval-only recall (see Figure 4 and Section 4.5).

Table 3. Prompt input-token and output-token budgets measured across 28 queries.

System	Input Tokens (Mean)	Output Tokens (Mean / Min / Max)	Citations per Query (Mean)
no_retrieval	~300	461 / 112 / 790	7.1
bm25_rag ($k = 10$)	~4,400	578 / 80 / 1,027	5.7
wiki_rag (ours)	~1,264	486 / 177 / 828	6.8
Ratio (wiki / bm25)	~0.29	–	–

Output sizes are reported as mean/min/max. Input sizes are fixed for no-retrieval and Wiki RAG and determined by retrieval depth k for BM25 RAG.

The context-budget analysis helps sharpen the interpretation of the main findings. Any advantage achieved by Wiki RAG over BM25 RAG cannot be attributed to the injection of more raw information into the prompt. In fact, Wiki RAG operates with roughly one-third

of BM25 RAG's input context. The wiki page serves as a structurally compact representation in which cross-institutional alignment has already been performed offline. As shown in the following sections, this compact representation also covers more gold-supporting citations than retrieval over ten raw articles.

4.4. Main Results

The main citation-based results are reported in Table 4 and visualized in Figure 3. The ranking is consistent across all three metrics: Wiki RAG performs best, followed by BM25 RAG, with no-retrieval performing worst. The effect on our primary metric, citation recall, is substantial. Wiki RAG achieves a recall of 0.242, compared with 0.095 for BM25 RAG and 0.023 for no-retrieval. Relative to BM25 RAG, Wiki RAG improves recall by +0.147 absolute (+155% relative). Relative to the parametric-only baseline, the improvement is +0.219 absolute, corresponding to roughly an order-of-magnitude increase. Citation precision follows the same ordering (0.709 > 0.422 > 0.091), indicating that Wiki RAG does not achieve higher recall simply by emitting more citations indiscriminately; rather, it retrieves and surfaces a larger proportion of relevant citations.

Table 4. Main results on 28 T3a queries (mean ± standard deviation).

System	Citation Recall	Citation Precision	Citation F1
no_retrieval	0.023 ± 0.03	0.091 ± 0.11	0.035 ± 0.04
bm25_rag	0.095 ± 0.12	0.422 ± 0.33	0.141 ± 0.15
wiki_rag	0.242 ± 0.21	0.709 ± 0.40	0.338 ± 0.26
Δ (wiki − bm25)	+0.147 (+155%)	+0.287 (+68%)	+0.197 (+140%)
Δ (wiki − no_retrieval)	+0.219 (10.5×)	+0.618 (7.8×)	+0.303 (9.6×)

Bold values indicate the best result in each column. All citation-based differences are statistically significant (Table 5).

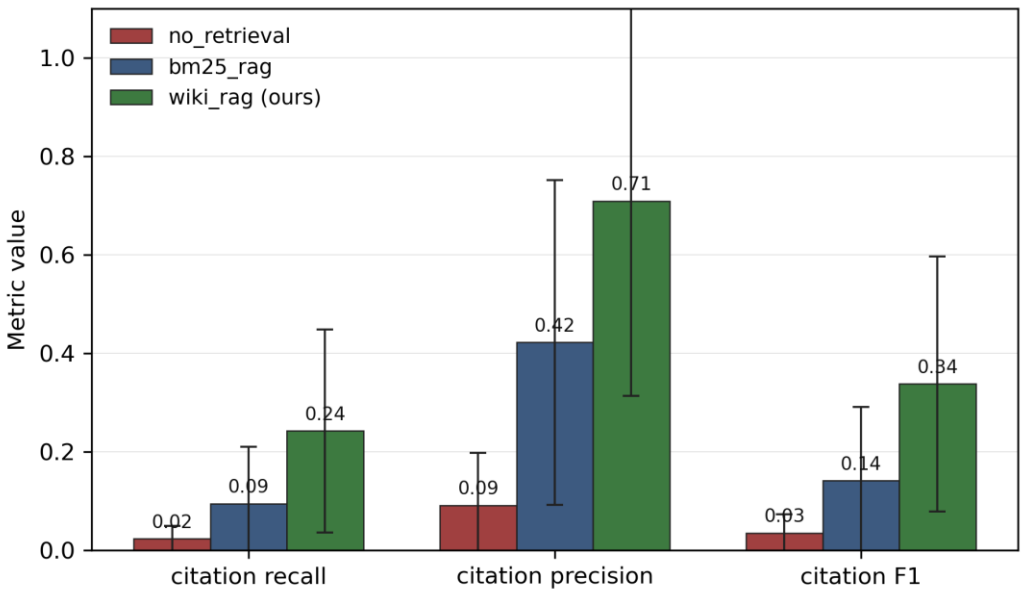

Figure 3. Citation-based metrics across the three systems. Error bars indicate ±1 standard deviation over 28 queries. The ordering Wiki RAG > BM25 RAG > no-retrieval is consistent across recall, precision, and F1; all pairwise differences are statistically significant (Table 5).

Statistical significance results are summarized in Table 5. All pairwise differences in citation recall are significant under a paired Wilcoxon signed-rank test ($p < 10^{-3}$). Paired bootstrap 95% confidence intervals (10,000 resamples over query indices) exclude zero in every comparison. Cliff's δ indicates a medium effect size for Wiki RAG versus BM25 RAG and for BM25 RAG versus no-retrieval, and a large effect size for Wiki RAG versus no-retrieval.

The absolute citation-recall values remain modest across all systems. A typical T3a query contains between 5 and 40 gold citations distributed across multiple institutions, making high recall challenging within a single generation step. Accordingly, these results should not be interpreted as demonstrating end-to-end sufficiency. Rather, they establish a consistent relative ordering within this task family: retrieval over the raw corpus improves substantially over the parametric-only baseline, while replacing flat retrieval with a compiled wiki provides an additional gain. The implications of these absolute performance levels for practical deployment are discussed in Section 5.2.

Table 5. Statistical significance results for citation recall.

Comparison	N	Δ (95% CI)	Wilcoxon p	Cliff's δ	Effect
wiki vs. bm25	28	+0.147 [+0.091, +0.205]	0.0001	+0.364	Medium
wiki vs. no_retrieval	28	+0.219 [+0.147, +0.292]	$<10^{-4}$	+0.597	Large
bm25 vs. no_retrieval	28	+0.072 [+0.034, +0.115]	0.0003	+0.466	Medium

Pairwise comparisons are evaluated using a paired Wilcoxon signed-rank test and bootstrap 95% confidence intervals (10,000 resamples). Cliff's δ is reported as a non-parametric effect-size measure.

4.5. Retrieval-Only Analysis

The results reported above combine two distinct factors: the quality of the retrieval substrate (which evidence is presented to the LLM) and the quality of the LLM's use of that evidence (which citations are ultimately selected and surfaced). To isolate the retrieval component, we compute, for each query, the intersection between the pre-LLM retrieval set and the gold citation set while varying BM25 retrieval depth across $k \in \{1, 3, 5, 10, 20\}$. For Wiki RAG, the retrieval set is the complete citation set contained in the corresponding policy-dimension page and is therefore independent of k.

Throughout this subsection, retrieval-only recall is defined as the fraction of gold-supporting citations contained within a system's pre-LLM retrieval substrate—either the top-k BM25 articles or the complete wiki page—regardless of whether those citations are subsequently selected by the LLM. This metric isolates evidence coverage from downstream citation selection. The results are shown in Figure 4.

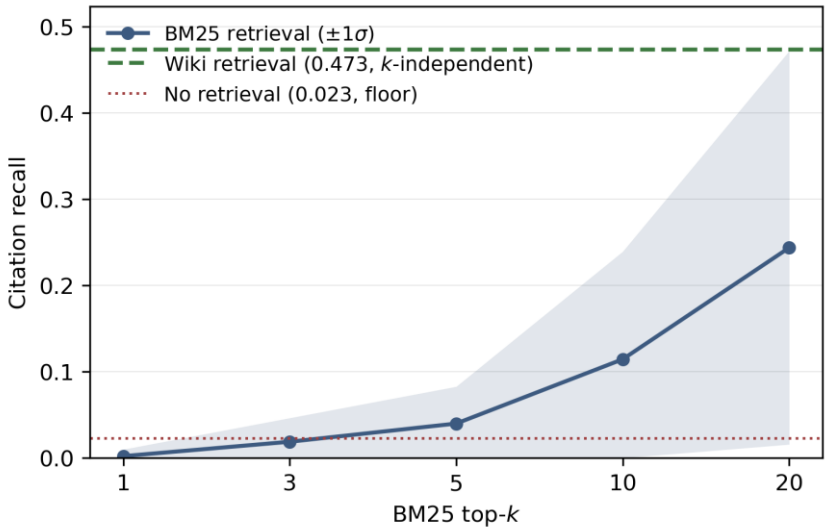

Figure 4. Retrieval-only citation recall across 28 T3a queries. BM25 retrieval recall increases monotonically with retrieval depth k but does not reach the Wiki RAG reference level within the tested range. Wiki retrieval is k-independent by construction because each query receives the complete policy-dimension page. The shaded region represents ±1 standard deviation across queries.

BM25 retrieval-only recall increases monotonically from 0.002 at $k = 1$ to 0.244 at $k = 20$. The Wiki RAG retrieval substrate achieves a retrieval-only recall of 0.473, shown as a horizontal reference line, while the no-retrieval baseline remains at 0.023. Within the range

tested, the wiki substrate consistently exceeds the strongest BM25 configuration, achieving 0.473 versus 0.244, a ratio of approximately 1.94×. When considered together with the context-budget analysis in Section 4.3.1, the evidence suggests that Wiki RAG provides a retrieval substrate that is simultaneously more compact (approximately 29% of BM25 RAG's input-token budget) and more comprehensive with respect to gold-supporting citations (approximately 1.94× higher retrieval-only recall).

For this T3a benchmark, the observed advantage therefore appears to arise from the pre-compiled cross-institutional organization of the wiki page rather than from larger context volume or broader retrieval. We emphasize, however, that this conclusion is restricted to the task family examined in this study and should not be generalized beyond the scope conditions discussed in Section 5.1.1.

4.6. Failure-Mode Analysis

To better understand how each system fails, we classify all 84 predictions (28 queries × 3 systems) into six mutually exclusive categories assigned in a top-down order:

- api_blocked (content-filter error),
- no_citation (no [doc: art. N] token emitted),
- hallucinated_cite (invalid institution identifier or article number),
- wrong_doc (all cited articles are structurally valid but none overlap with the gold citation set),
- success (citation recall ≥ 0.5), and
- partial (all remaining cases).

The resulting distribution is summarized in Table 6 and visualized in Figure 5.

Table 6. Failure-mode distribution across 28 T3a queries.

System	Success	Partial	Wrong_Doc	No_Cite	Hallucinated	Blocked	Total
no_retrieval	0	3	2	2	21	0	28
bm25_rag	0	9	0	4	15	0	28
wiki_rag	3	8	6	0	11	0	28

Categories are mutually exclusive and assigned in the order described in the text.

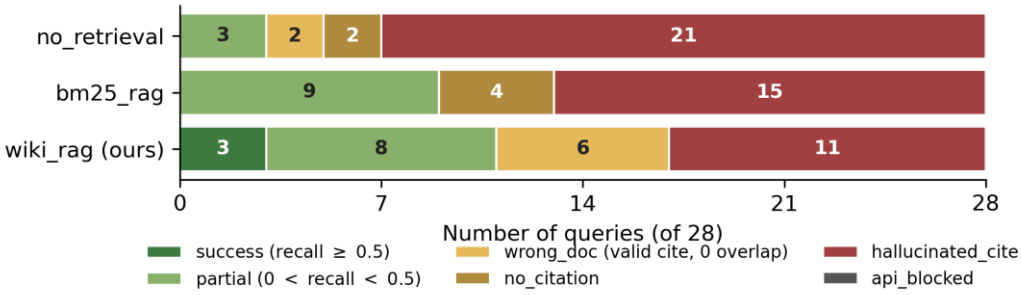

Figure 5. Failure-mode distribution visualized. The hallucinated citation rate decreases monotonically across the three systems (75% → 54% → 39%). Only wiki_rag produces any success outputs (3 of 28). The wiki_rag row also exposes a specific trade-off: 21% of its queries produce structurally valid citations that do not intersect the gold set (wrong_doc); Section 4.7 discusses one such case.

The first notable pattern is that the hallucinated-citation rate decreases monotonically across the sequence no-retrieval → BM25 RAG → Wiki RAG, with rates of 75%, 54%, and 39%, respectively. Introducing a flat retrieval substrate reduces hallucinated citations from 21 of 28 queries to 15 of 28. Replacing flat retrieval with a structured wiki further reduces this number to 11 of 28. Each retrieval enhancement eliminates approximately four hallucinated outputs, which is consistent with the hypothesis that citation generation benefits from direct access to the referenced source material and that a structured citation table provides a more accessible substrate than a flat collection of retrieved articles.

Second, Wiki RAG is the only system that produces any success cases. It achieves citation recall ≥ 0.5 on 3 of the 28 benchmark queries, whereas both comparison systems achieve

none. Third, the analysis reveals a trade-off that does not favor Wiki RAG. Approximately 21% of Wiki RAG outputs (6 of 28 queries) contain citations that are structurally valid but do not intersect the gold citation set. This wrong_doc rate is higher than that observed for the other systems, which produce 0 and 2 such cases, respectively. Section 4.7 examines one representative example in detail.

4.7. Case Study: Three Systems on a Single Query

To make the aggregate results more concrete, we trace a single query through all three systems. The query asks what Hunan University of Science and Engineering (hnkjxy) specifies regarding labor cost (劳务费) and which institutions have conflicting rules. The correct answer, verified against Article 14 of Hunan University of Science and Engineering's policy document, states that labor costs may account for up to 60% of realized project funding when the contract does not specify otherwise. Several peer institutions impose no such proportional cap.

The no-retrieval baseline, which receives only the institution list and the query, generates a generic description of Chinese and provincial labor-cost regulations without identifying the specific 60% limit. It cites five articles, only one of which overlaps with the gold citation set. Three of the five citations attribute rules to the wrong institution. Its citation recall is 0.053.

The BM25 RAG baseline, supplied with the top ten retrieved articles, correctly identifies the 60% rule and states it accurately. However, it identifies only two conflicting institutions—Beijing University (bjdx) and East China University of Science and Technology (ecust). Although both are correct, the benchmark requires identification of all institutions whose rules differ from the focal institution. Many such institutions are omitted. The resulting citation recall is 0.158, with citation precision of 1.000. The system is accurate but incomplete.

The Wiki RAG system receives the pre-compiled labor-cost comparison page shown in Table 7. Like BM25 RAG, it correctly identifies the 60% rule. More importantly, it enumerates nine institutions with differing policies and correctly classifies all nine as conflicts. The list includes, for example:

- Henan Institute of Technology (hngxy), which specifies degree-dependent monthly caps (≤1000 CNY for undergraduates, ≤2000 for master's students, and ≤3000 for doctoral students);
- Zhejiang University (zjdx_02), which imposes no proportional cap; and
- Zhengzhou University (zzdx), which likewise specifies no proportional limit.

Table 7. Excerpt from the compiled labor_cost wiki page used as context by Wiki RAG for query t3a_labor_cost_hnkjxy.

Institution	Rule	Source
Beijing University	No proportional cap; labor cost dispensed per project	[bjdx: art. 21]
East China University of Science and Technology	No proportional cap; off-campus labor costs may not be paid in cash	[ecust: art. 26]
Henan Institute of Technology	Monthly caps: undergraduate ≤1000 CNY, master's ≤2000 CNY, doctoral ≤3000 CNY	[hngxy: art. 23]
Hunan University of Science and Engineering (focal)	Up to 60% of realized funding when the contract does not specify otherwise	[hnkjxy: art. 14]
Nankai University	Up to 60% of realized funding (matches focal institution)	[nkdx: art. 26]
Wuhan University	Linked to local average wages for R&D services	[whdx: art. 13]
Zhejiang University	No specific proportional limit	[zjdx_02: art. 14]
Zhengzhou University	No proportional limit	[zzdx: art. 16]

The table is generated offline by the wiki-compilation pipeline (Section 3.2) and retrieved directly at inference time. Because the focal institution appears alongside peer institutions within the same comparison structure, the query-time LLM can perform cross-institutional comparison through local lookup rather than repeated multi-document inference.

The system also correctly recognizes that Nankai University (nkdx) shares the same 60% cap as the focal institution and therefore excludes it from the conflict set. Its citation recall reaches 0.579, while citation precision remains 1.000. This example helps explain the benchmark-wide results. Citation recall for the three systems follows the ratio: $0.053 : 0.158 : 0.579$, which simplifies approximately to: $1 : 3 : 11$. This closely mirrors the benchmark-wide recall ratio of: $0.023 : 0.095 : 0.242$.

The difference between BM25 RAG and Wiki RAG is not that one system identifies the focal rule while the other does not. Both successfully recover the focal rule. The difference lies in the retrieval substrate. BM25 retrieval primarily surfaces articles associated with the focal institution that are highly similar to the query, whereas the compiled wiki page presents a pre-aligned cross-institutional comparison organized around the same policy dimension. In effect, the wiki structure transforms the task of identifying conflicting institutions from a multi-document reasoning problem into a localized comparison task over a pre-organized table.

5. Comparison and Discussion

5.1. Why Structural Wikis Help on This Task Family

The experimental results suggest a specific mechanism underlying the performance gap between BM25 RAG and the pre-compiled wiki approach. This interpretation is supported jointly by the retrieval-only analysis (Section 4.5), the context-budget analysis (Section 4.3.1), and the case study (Section 4.7).

BM25 ranks articles according to lexical similarity between the query and individual articles. This strategy works well when the query and the relevant article share distinctive vocabulary—for example, when the query explicitly names the focal institution and asks about a specific policy dimension. Indeed, BM25 retrieves the focal institution's article for the queried dimension in most cases.

The difficulty arises from the second requirement of the T3a task. The objective is not merely to identify the focal institution's rule but also to determine which other institutions differ from it. Articles belonging to those institutions typically do not contain the focal institution's name and often share little lexical overlap with the query itself. Consequently, the task requires cross-institutional synthesis across multiple documents. This type of distributed reasoning closely resembles the long-context inference problem identified by Liu et al. [14], who showed that language models struggle when relevant information is dispersed across lengthy inputs.

The wiki-compilation pipeline performs much of this synthesis in advance. During compilation, the LLM reads institution-specific evidence for a given policy dimension and constructs a comparison table containing one row per institution, with cross-references already resolved. At inference time, the model no longer faces a long-context, multi-document reasoning task. Instead, it performs a focused lookup within a single comparison table where relevant information is colocated.

Conceptually, this approach is related to retrieval-augmented pre-training [25] and iterative retrieve-and-generate frameworks [26]. However, unlike those methods, the compilation process occurs offline at the document-set level rather than online at the token or interaction level. The retrieval-only analysis demonstrates this effect quantitatively: across all retrieval depths tested, the compiled wiki contains more gold-supporting citations than the corresponding BM25 retrieval results.

The failure-mode analysis (Section 4.6) provides complementary evidence. Hallucinated citations decrease from 75% to 54% when retrieval is introduced and decline further to 39% when flat retrieval is replaced by a compiled wiki. The magnitude of the reductions (approximately 21% and 15%) is broadly consistent with the interpretation that each retrieval enhancement improves grounding by increasing the availability and accessibility of supporting evidence. At the same time, the remaining 39% hallucination rate indicates that even a structured wiki does not eliminate citation errors entirely.

Task–Representation Alignment as a Scope Condition

The T3a task is inherently well aligned with a tabular representation. The most direct human-readable answer is a comparison table containing one row per institution. The wiki pages generated by our compilation pipeline are precisely such tables, constructed offline before any query is issued. Accordingly, we interpret the empirical advantage reported in Section 4.4 as evidence of task–representation alignment for this specific task family rather than as a

general claim of superiority over retrieval-based approaches. Tasks whose natural outputs are not tabular—for example, single-document question answering, narrative summarization, or problems requiring reasoning beyond the information materialized in the wiki—should not necessarily be expected to benefit from the same arrangement.

The behavior of Wiki RAG on the horizontal-versus-vertical funding dimension provides an instructive example. All five queries in that dimension achieve zero recall despite using the same retrieval substrate that performs well elsewhere. This result illustrates that compile-time design choices determine which information becomes available at inference time. Even within a task family that generally benefits from structured representation, dimensions that are not adequately materialized during compilation may remain difficult to answer. We revisit this limitation in Section 5.2.

5.2. Threats to Validity and Discussion

We summarize the principal limitations of the present study below.

LLM-Extracted Gold, Audited on a Subsample

The benchmark gold citations are generated using the LLM-based extraction pipeline described in Section 4.2 rather than by human experts. To strengthen confidence in the benchmark, we conduct a manual audit of all 172 gold citations across six representative queries (Section 4.2.1; Table 2) and release the complete audit workbook, allowing independent reviewers to verify each judgment against the publicly available source documents.

Because the audit was conducted by a single author, inter-rater agreement statistics are not available. The audit yields a strict precision of 73.3% and a precision of 93.0% when topically relevant partial matches are also counted. The results indicate that the extractor exhibits a recall-oriented bias: it tends to include articles that are topically related to the target rule, including asset-management provisions, procurement procedures, and compliance clauses, even when those articles do not explicitly define the rule under consideration. Outright errors remain relatively uncommon (12 of 172 citations, or 7.0%). Importantly, this tendency affects all evaluated systems equally because every system is scored against the same gold set. Consequently, it influences absolute performance values but does not alter the relative comparisons that form the basis of our main conclusions.

We further acknowledge that a single-auditor protocol introduces the possibility of systematic bias. A single reviewer may apply consistent subjective criteria across ambiguous cases and may systematically favor one interpretation when distinguishing between correct, partial, and incorrect citations. The reported 93.0% precision should therefore be viewed as an informed expert assessment rather than a consensus estimate. A multi-annotator validation study with inter-rater agreement statistics (e.g., Cohen's κ or Fleiss' κ on a triple-annotated subset) represents an important direction for future benchmark development and is identified as future work in Section 6.

Relationship to Modern Retrieval Systems

Our experiments compare Wiki RAG against a classical BM25 baseline. We do not evaluate dense retrieval methods such as DPR [22], ColBERT [23], or BGE-M3 [24], nor do we consider hybrid sparse–dense retrieval or query-expansion strategies. This choice was deliberate. First, the primary objective of the paper is to isolate the effect of structural pre-compilation relative to flat retrieval. Introducing multiple retrieval variants would substantially expand the experimental design without directly addressing this question. Second, dense retrieval methods typically introduce training-data dependencies that BM25 avoids, making controlled comparisons more difficult.

We expect that dense or hybrid retrieval would reduce some of the gap between BM25 RAG and Wiki RAG. Nevertheless, the task–representation alignment discussed in Section 5.1.1 remains relevant. Even an optimally tuned dense retriever returns raw documents that must be synthesized at query time, whereas the wiki provides a pre-aligned comparison structure. A related question is whether query-time table synthesis—constructing a comparison table dynamically from retrieved articles—could narrow the gap. We do not evaluate this alternative. However, such an approach would incur the full synthesis cost for every query, whereas the wiki amortizes that cost during offline compilation. Both dense-versus-wiki comparisons and synthesis-on-demand baselines are identified as future work in Section 6.

Conflict-Based Metrics Are Structurally Biased

As discussed in Section 3.4, the benchmark's gold conflict sets contain between 12 and 18 of the 20 possible institutions for a typical query. Under these conditions, a system can achieve an artificially high conflict F1 score simply by predicting nearly every institution as conflicting. For completeness, we report conflict-based metrics but exclude them from all significance testing and from our primary conclusions.

Single LLM Backbone

All experiments use glm-4-plus as the underlying language model. Although the proposed framework is not tied to a specific model family, we do not evaluate robustness across alternative LLM backbones. Replication using different language models would strengthen confidence in the generality of the observed effects.

One Corpus, One Language, One Domain

The study is restricted to Chinese higher-education grant-management policies and to a task family characterized by a naturally tabular answer structure. Generalization to other cross-institutional policy domains with similar comparison-oriented structures—such as corporate compliance frameworks or cross-jurisdiction regulatory analysis—appears plausible. By contrast, generalization to non-tabular tasks is not supported by the evidence presented here.

Small Benchmark Size

The benchmark contains 28 queries. Although the observed effect sizes are substantial and statistical significance is strong ($p < 10^{-3}$ on the primary metric), a substantially larger benchmark would provide greater confidence in the stability of the results.

Wiki Compilation Hallucinates Beyond What the Citation Checker Catches

The citation checker verifies that every [doc:art.N] token refers to an existing document and a valid article number. It does not verify semantic consistency between the surrounding claim and the cited content. Two examples from the compiled wiki illustrate this limitation.

- First, the indirect_cost page incorrectly labels institution sysfdx (Shenyang Normal University) as Huazhong University of Science and Technology. The citation itself remains structurally valid—[sysfdx:art.19] refers to a genuine article—but the displayed institution name is hallucinated.
- Second, in the indirect_cost query for Fudan University's STEM-specific document (fddx_01), Wiki RAG begins its answer by describing a rule that actually belongs to Anhui University of Technology (ahgydx) while correctly citing [ahgydx:art.12]. In this case, the model attaches a valid citation to the wrong focal institution. The citation tuple is structurally correct, but the attribution is semantically incorrect.

Both examples motivate the need for semantic validation beyond syntactic citation checking. One possible direction is an NLI-based or LLM-assisted consistency check between generated claims and the cited article text. Existing frameworks such as ALCE [15] and the reflective-retrieval mechanism of Self-RAG [16] provide promising starting points and are discussed further in Section 6.

API Content-Filter Interactions

Some passages within the Chinese grant-policy corpus contain governance-related terminology such as audit, discipline, and violation, which occasionally trigger content-safety filters in the underlying API at higher retrieval depths. No such failures occur under our default BM25 configuration ($k = 10$), so the main experimental results are unaffected. Exploratory runs at $k = 20$ occasionally encounter filtering issues. If this effect introduces any bias, it would tend to disadvantage BM25 at larger retrieval depths rather than favor Wiki RAG.

Implications for High-Stakes Deployment

The best-performing system achieves a citation recall of 0.242. In practical terms, this means that only about one quarter of the relevant supporting articles are surfaced for a typical T3a query. For research-assistance or literature-screening scenarios, where users are expected to inspect the cited evidence manually, such performance can still provide meaningful value. Each surfaced citation either supports the claim or can be rapidly rejected through human verification. For automated compliance analysis, auditing, or grant-management decision

support, however, the remaining three quarters of missing evidence represent a substantial risk. A system that confidently reports only a subset of conflicting institutional rules may create a false impression of completeness. For this reason, we view the current system as appropriate for human-in-the-loop policy exploration rather than fully automated decision-making. Substantially improving citation recall remains a prerequisite for high-stakes deployment.

6. Conclusions

We presented a controlled empirical comparison between a flat BM25 retrieval-augmented baseline and a pre-compiled wiki-based retrieval substrate on 28 cross-institutional policy-conflict queries derived from Chinese university grant-management documents. The results consistently indicate that, for this task family, organizing information into a structured comparison artifact before inference can be more effective than retrieving and synthesizing raw documents at query time. We interpret this finding as a design insight about the structural fit between cross-document comparison tasks and compile-then-query arrangements, supported by a proof-of-concept implementation, rather than as a general superiority claim over modern retrieval systems.

Several supporting analyses help clarify this interpretation. The retrieval-only analysis suggests that the observed advantage is associated with pre-compiled cross-document organization rather than broader retrieval coverage, a finding consistent with broader observations on retrieval design choices in LLM-augmented systems [27]. The context-budget analysis further shows that the wiki representation achieves this while operating on substantially less input context than BM25 retrieval. In addition, the failure-mode analysis demonstrates a consistent reduction in hallucinated citations as retrieval support becomes increasingly structured. A manual audit of representative benchmark queries further supports the validity of the LLM-extracted gold, while also revealing recurring semantic-attribution errors that motivate future improvements.

The scope of the result remains intentionally limited. Our study evaluates a single language, a single corpus, a single generator, and a task family whose natural answer structure is tabular. The absolute recall levels remain modest, and the compiled wiki itself is not immune to semantic errors that citation-level validation cannot detect. Accordingly, we view the contribution as a controlled investigation of one knowledge-organization strategy for LLM systems rather than a universal prescription for retrieval-augmented generation.

Several directions follow naturally from these findings. First, semantic citation-verification mechanisms—such as NLI-based validation or LLM-as-a-judge consistency checking inspired by ALCE [15] and Self-RAG [16]—could address attribution and focal-confusion failures that remain beyond the reach of syntactic citation checking. Second, comparisons against dense retrieval approaches such as bge-m3 [24], LLM-fine-tuned multi-stage retrievers [28], and query-time table-synthesis baselines would help isolate the respective contributions of retrieval quality, structural compilation, and query-time reasoning. Third, replication across additional corpora, languages, and regulated domains would test the generality of the task–representation alignment reported here. Finally, larger human-annotated benchmarks with multi-annotator validation and inter-rater agreement statistics would provide a stronger foundation for future research on cross-document reasoning and citation-grounded generation.

Author Contributions: This work is a single-author manuscript; B.Z. independently performed all aspects of the research and manuscript preparation, including conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, writing—review and editing, and visualization.

Funding: This research received no external funding.

Data Availability Statement: The benchmark (28 queries with article-level gold citations), LLM-compiled wiki pages (6 policy dimensions), raw evaluation results (84 system predictions), the complete human-audit workbook used in Section 4.2.1 (including all 172 audited gold citations with verdicts and annotations), prompt templates, evidence-window keyword lists, configuration files, source code, and analysis scripts are publicly available in a Zenodo repository at DOI: 10.5281/zenodo.20283627. The repository is released under the CC BY

4.0 license for derived artifacts and the MIT License for source code. The 22 source policy documents are publicly available from the official websites of the 19 universities from which they were collected. Institutional identifiers, segmentation hashes, and source-document metadata are provided in the released corpus_metadata.csv file.

Acknowledgments: The author thanks the two anonymous reviewers whose substantive comments directly motivated several of the analyses in this revision, including the human audit and the con-text-budget comparison. Generative AI tools were used during manuscript preparation for code generation, figure generation scripts, and language editing; the author retains full responsibility for the technical content, experimental design, and conclusions.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] P. Lewis *et al.*, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *arXiv*. Apr. 12, 2021. [Online]. Available: http://arxiv.org/abs/2005.11401

[2] Y. Gao *et al.*, "Retrieval-Augmented Generation for Large Language Models: A Survey," *arXiv*. Mar. 27, 2024. doi: 10.48550/arXiv.2312.10997.

[3] P. Sharma and S. Bhattarai, "A Review on Retrieval-Augmented Generation: Architectures, Research Challenges, and Emerging Frontiers," *J. Futur. Artif. Intell. Technol.*, vol. 2, no. 4, pp. 616–628, Jan. 2026, doi: 10.62411/faith.3048-3719-297.

[4] A. Karpathy, "LLM Wiki," *GitHub Gist*. Apr. 2026. [Online]. Available: https://gist.github.com/karpathy/442a6bf555914893e9891c11519de94f

[5] L. Zheng *et al.*, "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," *arXiv*. Dec. 24, 2023. [Online]. Available: http://arxiv.org/abs/2306.05685

[6] A. Vaswani *et al.*, "Attention Is All You Need," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017. [Online]. Available: http://arxiv.org/abs/1706.03762

[7] J. Devlin, M.-W. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North*, Oct. 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[8] C. Raffel *et al.*, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *arXiv*, Sep. 2023, [Online]. Available: http://arxiv.org/abs/1910.10683

[9] T. B. Brown *et al.*, "Language Models are Few-Shot Learners," in *arXiv*, Jul. 2020. [Online]. Available: http://arxiv.org/abs/2005.14165

[10] OpenAI *et al.*, "GPT-4 Technical Report," *ArXiv*. Mar. 04, 2024. [Online]. Available: http://arxiv.org/abs/2303.08774

[11] H. Touvron *et al.*, "Llama 2: Open Foundation and Fine-Tuned Chat Models," *arXiv*. Jul. 19, 2023. [Online]. Available: http://arxiv.org/abs/2307.09288

[12] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," *Found. Trends® Inf. Retr.*, vol. 4, no. 1–2, pp. 1–174, Sep. 2009, doi: 10.1561/1500000019.

[13] J. Menick *et al.*, "Teaching language models to support answers with verified quotes," *arXiv*. Mar. 21, 2022. [Online]. Available: http://arxiv.org/abs/2203.11147

[14] N. F. Liu *et al.*, "Lost in the Middle: How Language Models Use Long Contexts," *Trans. Assoc. Comput. Linguist.*, vol. 12, pp. 157–173, Feb. 2024, doi: 10.1162/tacl_a_00638.

[15] T. Gao, H. Yen, J. Yu, and D. Chen, "Enabling Large Language Models to Generate Text with Citations," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 6465–6488. doi: 10.18653/v1/2023.emnlp-main.398.

[16] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection," in *arXiv*, Oct. 2023. [Online]. Available: http://arxiv.org/abs/2310.11511

[17] O. Ram *et al.*, "In-Context Retrieval-Augmented Language Models," *Trans. Assoc. Comput. Linguist.*, vol. 11, pp. 1316–1331, Nov. 2023, doi: 10.1162/tacl_a_00605.

[18] D. Edge *et al.*, "From Local to Global: A Graph RAG Approach to Query-Focused Summarization," *arXiv*. Feb. 19, 2025. [Online]. Available: http://arxiv.org/abs/2404.16130

[19] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The Muppets straight out of Law School," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 2898–2904. doi: 10.18653/v1/2020.findings-emnlp.261.

[20] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, "BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models," *arXiv*. Oct. 21, 2021. [Online]. Available: http://arxiv.org/abs/2104.08663

[21] Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang, "Pre-Training With Whole Word Masking for Chinese BERT," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3504–3514, 2021, doi: 10.1109/TASLP.2021.3124365.

[22] V. Karpukhin *et al.*, "Dense Passage Retrieval for Open-Domain Question Answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 6769–6781. doi: 10.18653/v1/2020.emnlp-main.550.

[23] O. Khattab and M. Zaharia, "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Jul. 2020, pp. 39–48. doi: 10.1145/3397271.3401075.

[24] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, "M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation," in *Findings of the Association for Computational Linguistics ACL 2024*, 2024, pp. 2318–2335. doi: 10.18653/v1/2024.findings-acl.137.

[25] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang, "REALM: Retrieval-Augmented Language Model Pre-Training," in *ICML'20: Proceedings of the 37th International Conference on Machine Learning*, Feb. 2020. [Online]. Available: http://arxiv.org/abs/2002.08909

[26] Z. Shao, Y. Gong, Y. Shen, M. Huang, N. Duan, and W. Chen, "Enhancing Retrieval-Augmented Large Language Models with Iterative Retrieval-Generation Synergy," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 9248–9274. doi: 10.18653/v1/2023.findings-emnlp.620.

[27] Y. Zhu *et al.*, "Large Language Models for Information Retrieval: A Survey," *arXiv*. Sep. 17, 2025. doi: 10.1145/3748304.

[28] X. Ma, L. Wang, N. Yang, F. Wei, and J. Lin, "Fine-Tuning LLaMA for Multi-Stage Text Retrieval," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Jul. 2024, pp. 2421–2425. doi: 10.1145/3626772.3657951.