

ChunQiuTR: Time-Keyed Temporal Retrieval in Classical Chinese Annals

Yihao Wang¹, Zijian He¹, Jie Ren², Keze Wang¹,

¹Sun Yat-Sen University, ²Shaanxi Normal University,

Correspondence: kezewang@gmail.com

Abstract

Retrieval shapes how language models access and ground knowledge in retrieval-augmented generation (RAG). In historical research, the target is often not an arbitrary relevant passage, but the exact record for a specific regnal month, where temporal consistency matters as much as topical relevance. This is especially challenging for Classical Chinese annals, where time is expressed through terse, implicit, non-Gregorian reign phrases that must be interpreted from surrounding context, so semantically plausible evidence can still be temporally invalid. We introduce **ChunQiuTR**, a time-keyed retrieval benchmark built from the *Spring and Autumn Annals* and its exegetical tradition. ChunQiuTR organizes records by month-level reign keys and includes chrono-near confounders that mirror realistic retrieval failures. We further propose **CTD** (Calendrical Temporal Dual-encoder), a time-aware dual-encoder that combines Fourier-based absolute calendrical context with relative offset biasing. Experiments show consistent gains over strong semantic dual-encoder baselines under time-keyed evaluation, supporting retrieval-time temporal consistency as a key prerequisite for faithful downstream historical RAG. Our code and datasets are available at github.com/xbdxyh/ChunQiuTR.

1 Introduction

Retrieval is increasingly the interface between language models and the world’s knowledge, most visibly in retrieval-augmented generation (RAG) and search-augmented assistants (Gao et al., 2023; Lewis et al., 2020). In such systems, models ground responses in retrieved evidence rather than relying on parametric memory alone. This evidentiary role is central to expert workflows—literature survey, legal and policy analysis, and scientific claim verification—where users care not only about *what* an answer is, but also *where* it comes from (Menick et al., 2022).

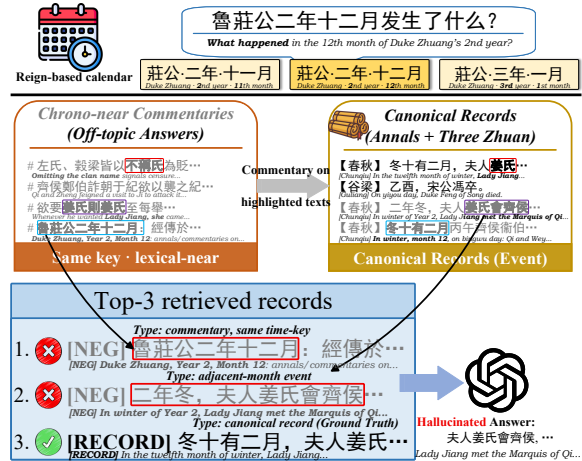


Figure 1: A query about a specific month can retrieve same-month commentary that repeats the date phrase, or adjacent-month near-miss events with confusable wording, so a retrieval-augmented model answers fluently but at the wrong time.

Historical research on pre-modern Chinese sources is a canonical example of evidence-centric retrieval (Cao et al., 2024; Zhang et al., 2024; Liu et al., 2025). Digitized annals, commentaries, and later annotations are now searchable, but the target is rarely an arbitrary topical snippet: it is the passage that records what happened in a particular month of a particular duke’s reign. As Fig. 1 illustrates, a query such as “What happened in Duke Zhuang’s 2nd year, 12th month?” can easily retrieve (i) exegetical commentary that repeats the same date phrase without answering the event, or (ii) near-duplicate events from adjacent months with highly confusable wording. In this setting, semantic relevance is insufficient without verifying temporal alignment to the queried month. Once retrieval binds a downstream generator to temporally incorrect but semantically plausible evidence, the final answer may still sound fluent while being wrong about *when* the event happened.

This motivates a more focused question that is

central to faithful historical RAG:

(Q) *How can a retriever select **time-consistent evidence** for queries expressed under **non-Gregorian, reign-based chronologies**?*

Studying this problem is already challenging because pre-modern records typically do not provide explicit, globally comparable (Gregorian) timestamps (Chen et al., 2021, 2025). Instead, they employ a ruler-centric regnal chronology: time is expressed relative to the current ruler and his regnal year and month, so temporal reference effectively resets across reigns and must be interpreted on a corpus-specific timeline rather than a monotonic calendar. Moreover, temporal phrases are often underspecified or written in shorthand—for example, “in summer, in the fifth month” may omit the absolute year and only become interpretable given the surrounding reign context. Crucially, time is not a clean metadata field separated from content: in annalistic writing, distinctive one-off events can implicitly function as temporal anchors, tightly coupling *when* with *what*. As a result, retrieval cannot rely on semantic similarity or timestamp ordering alone; it must identify evidence that is both topically relevant and temporally consistent with the intended regnal point or window.

To tackle this challenge, we ground our study in a demanding case: the *Spring and Autumn Annals* and its commentarial–exegetical corpus. We introduce **ChunQiuTR**, a time-keyed benchmark built on this material, where queries and records are expressed in a ruler-centric, non-Gregorian chronology rather than modern timestamps. Building on this benchmark, we propose the **Calendrical Temporal Dual-encoder (CTD)**, a time-aware dual-encoder retriever that augments semantic matching with learned calendrical structure. CTD places each query and record at a soft location on a unified ordered calendar axis and favors pairs that agree not only in meaning but also in calendrical position. Concretely, it injects an absolute calendrical context into embeddings and adds a relative temporal bias to similarity based on signed calendar offsets, improving robustness to adjacent-month and lexical-near confounders.

Our contributions are threefold: (i) we introduce **ChunQiuTR**, a non-Gregorian, reign-keyed temporal retrieval benchmark with point/gap/window queries and leak-free splits; (ii) we propose **CTD**,

a calendrically time-aware dual-encoder that combines absolute context injection with relative offset biasing; and (iii) we show consistent improvements over strong semantic dual-encoder baselines, especially under chrono-near and adjacent-month confounders, supporting the view that retrieval-time temporal consistency is a key prerequisite for faithful downstream historical RAG.

2 Related Work

2.1 Neural Information Retrieval

Lexical retrievers such as BM25 (Robertson and Zaragoza, 2009) remain strong and interpretable, but are limited by surface-form overlap. Neural IR methods are often grouped into neural sparse expansion (e.g., SPLADE (Formal et al., 2021)), dense dual-encoders trained with contrastive learning (e.g., DPR (Karpukhin et al., 2020), ANCE (Xiong et al., 2021), Contriever-style (Lei et al., 2023)), and late-interaction token matching (e.g., ColBERT (Khattab and Zaharia, 2020)). General-purpose embedding models (e.g., GTR (Ni et al., 2022b), E5 (Wang et al., 2024b), Qwen3-Embedding (Zhang et al., 2025b)) further enable plug-and-play retrieval. However, relevance is typically modeled as semantic similarity, which can still confuse chrono-near near-duplicates or fail to enforce fine-grained temporal constraints without explicit temporal structure.

2.2 Temporal Information Retrieval

Temporal information retrieval (TIR) incorporates time into ranking, ranging from timestamp-aware priors (e.g., time-based language models Li and Croft (2003)) to modern formulations emphasizing temporal focus/intent (Piryani et al., 2025). Recent work explores neural retrieval for time-sensitive settings by injecting temporal signals into retrieval or generation pipelines (Rajapakse, 2023; Zhang et al., 2025a), as well as mechanisms that encode time specifiers into model behaviors (Han et al., 2025). Related lines include time-aware language models with document-dating objectives and temporal label-smoothing schemes that smooth supervision over neighboring time steps, both of which we echo in our auxiliary temporal heads (Wang et al., 2023; Yèche et al., 2023; Dhingra et al., 2022). Most TIR studies target modern timestamped collections under open retrieval, whereas our setting is a micro-granular, *time-keyed* chronicle with dense chrono-near near-duplicates, leading to different

supervision and evaluation objectives.

Temporal-expression extraction and normalization are also related to our setting, including work on historical texts and cross-lingual temporal expression extraction/normalization (Korchagina, 2016; Cao et al., 2022; Su et al., 2025; Castro et al., 2025; Graciotti et al., 2025). However, in many Chunqiu passages, the ruling duke and/or regnal year is omitted, so the target month key must be recovered from annalistic structure and discourse context rather than extracted as a standalone temporal mention.

3 Dataset Construction

We construct **ChunQiuTR**, a benchmark for temporal retrieval under non-Gregorian, reign-based dating, using authentic historical texts centered on the *Chunqiu* and its classical commentarial tradition. The retrieval gallery is derived from source texts rather than AI-generated content, while queries are instantiated from a small set of manually written templates. LLMs are used only during curation to propose candidate splits or alignments; only human-verified results enter the final benchmark. Fig. 2 summarizes the construction pipeline.

3.1 Chunqiu Corpus and Temporal Scheme

Why the *Chunqiu*? The *Chunqiu* (*Spring and Autumn Annals*) is a terse chronicle of the state of Lu (722–481 BCE) whose entries record events using compact, ruler-centric temporal expressions such as “元年春” or “夏五月,” often without explicit absolute years. Its layered commentarial tradition, including the three classical *zhuan*, expands and reinterprets the same events along a shared reign-based timeline. This combination of implicit temporal reference, aligned historiographical layers, and overlapping descriptions makes the corpus a compact and natural testbed for temporal retrieval with realistic near-miss confusions. Additional source and preprocessing details are provided in Appendix A.1.

Reign-based time keys. Our sources use a ruler-centric, reign-based dating scheme in which regnal years restart for each new Lu duke. A full form may appear as “元年春王正月,” while later entries are often shortened to expressions such as “夏五月,” with the duke and regnal year supplied by context. Because these omitted elements often must be recovered from annalistic structure and discourse continuity rather than extracted as stan-

dalone temporal expressions, we manually verify the final time-key assignment for all records.

We normalize this temporal language into month-level time keys $\tau = (\text{gong, year, month})$, where gong denotes the ruling duke title. Month level is the finest temporal unit consistently recoverable in the *Chunqiu*; finer-grained dates are largely absent or too sparse to normalize systematically. We assign a time key to every month, including months with no annals entry, which are later instantiated as standardized no_event placeholders.

3.2 Record Alignment

Record-time-key alignment. We treat each *record* as the atomic retrieval unit: a short event-level passage aligned to a single month key τ . For each τ , we gather snippets from the annals and the three classical *zhuan* and refine them into event-level record sets \mathcal{D}_τ . Because sentence or paragraph boundaries do not reliably match historical events, we use a lightweight LLM prompt only to propose candidate splits and groupings, and then manually review and correct them so that each final record corresponds to one coherent event with its aligned commentarial material.

Chrono-near counterfactual negatives. Later historiographical layers provide naturally confusing, chronologically adjacent paraphrastic variants. We align these sources to the reign-based time keys using LLM-assisted candidate matching together with fuzzy string matching (Appendix A.3), and retain only human-verified alignments. For each time key τ , we define $\mathcal{D}_\tau^{\text{cf}}$ as records aligned to the same time key that describe the same situation in later paraphrase but are not used as ground-truth retrieval targets. These historically grounded near-miss variants serve as hard negatives for time-aware retrieval.

Audit and reliability. We manually verify time-key normalization because many passages omit an explicit duke or regnal year and must be resolved from annalistic structure and discourse continuity. Among 1,533 non-empty months, 558 contain multiple events; after LLM candidate grouping, only 63 required additional human correction, while the remaining 495 were accepted without change. For later-commentary alignment, LLMs only propose candidate matches, and the final human acceptance rate ranges from 93.33% to 100% across sources (Appendix B.5).

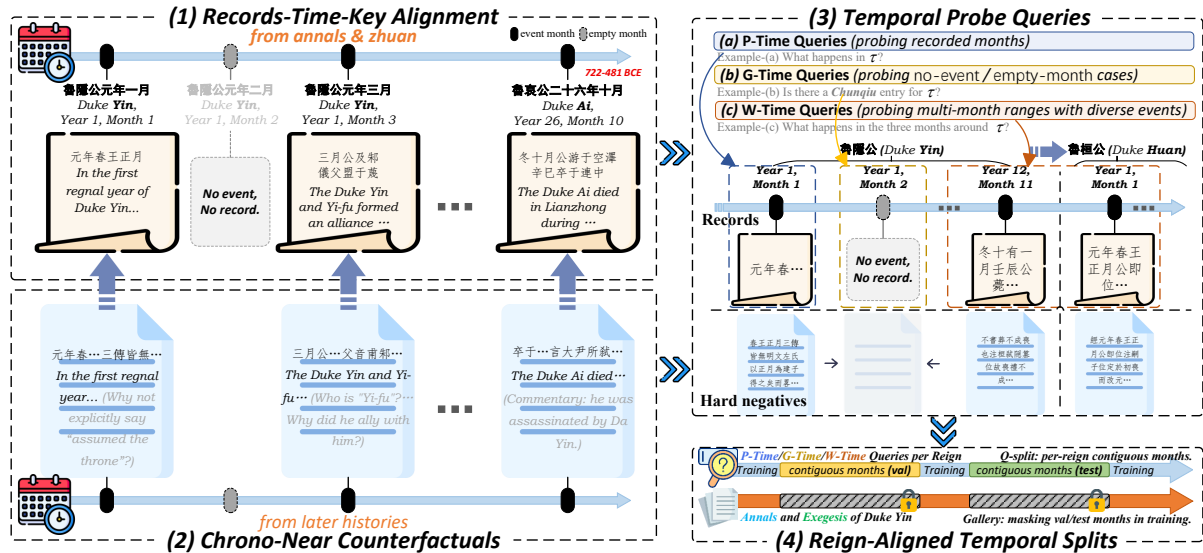


Figure 2: Overview of ChunQiuTR construction. Left: time-key alignment produces event-level records and chrono-near hard negatives. Right: P/G/W temporal queries and leak-free reign-aligned splits.

3.3 Temporal Queries and Evaluation Splits

Temporal query design. We define three families of temporal queries: point queries (**P-Time**), gap queries (**G-Time**), and local-window queries (**W-Time**), each mapped to a target interval Q_i on the reign-based timeline. To make empty intervals queryable, months with no annals entry are instantiated as standardized no_event records and included in the gallery. P-Time queries target a single time key and are answered by the corresponding event-bearing record or by an explicit no_event record when that month is empty. G-Time queries target missing-event months within a specified range, while W-Time queries target short contiguous windows around an event. All query templates and examples are provided in appendix A.4.

Reign-aligned splits. We partition the month-level timeline into reign-aware, leak-free train/validation/test splits. For each duke’s reign, we allocate disjoint contiguous month blocks to train, validation, and test in an approximate 80/10/10 ratio, and assign all records and queries in those months to the corresponding split. No time key, record, or query appears in more than one split. Evaluation is performed over the full time-keyed gallery, while validation and test queries come only from held-out reign segments. Overall, the benchmark contains 20,172 records and 16,226 queries (13,053 train / 1,520 validation / 1,653 test); detailed statistics and split visualizations are

provided in Appendix A.5.

4 Methods

We first formalize ChunQiuTR as a time-keyed retrieval task in Sec. 4.1. Sec. 4.2 then presents our Calendrical Temporal Dual-encoder (CTD): starting from a semantic dual-encoder score, CTD learns a latent regnal calendar scalar and incorporates an **absolute** calendrical context and a **relative** temporal bias to form the final score, as illustrated in Fig. 3. Finally, Sec. 4.3 describes our interval-overlap multi-positive supervision and joint training objective.

4.1 Task Formulation

We cast our ChunQiuTR benchmark as a temporal retrieval task over a discrete, reign-based month timeline. Building on the time-keyed records in Sec. 3.2, we formalize all aligned historical material as a fixed retrieval gallery

$$\mathcal{D} = \{d_j\}_{j=1}^N,$$

where each short Classical Chinese record d_j is associated with a reign-based month key $\tau(d_j)$; we write $\mathcal{D}_\tau = \{d_j \in \mathcal{D} : \tau(d_j) = \tau\}$ for the subset under time key τ .

For each query q_i constructed in Section 3.3, the benchmark specifies a target interval Q_i on the same month axis and a small multi-positive ground-truth set

$$\mathcal{G}_i \subseteq \mathcal{D}.$$

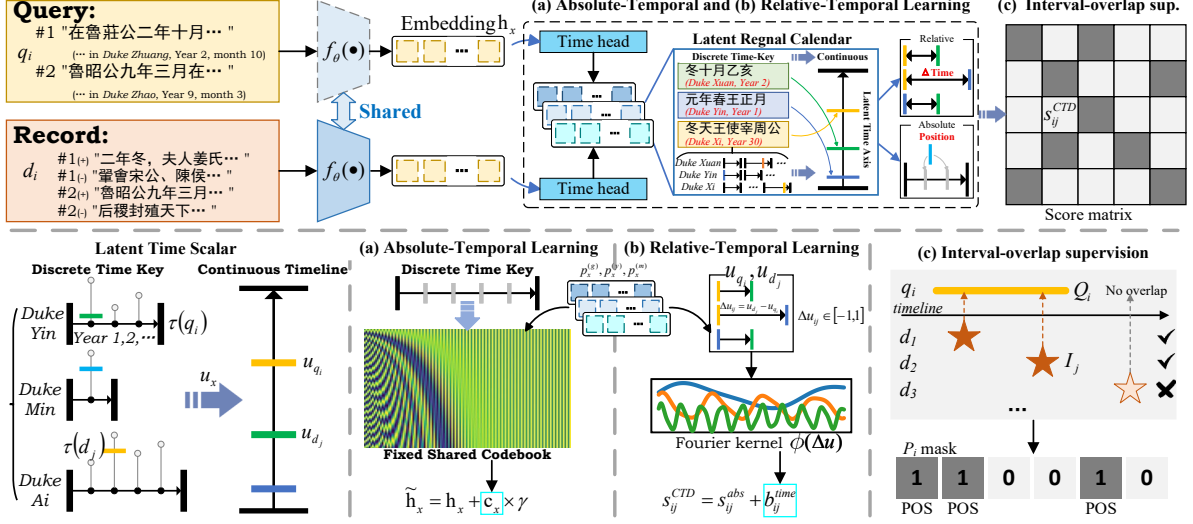


Figure 3: Overview of our Calendrical Temporal Dual-encoder (CTD). A shared Transformer dual-encoder encodes queries and records into embeddings. Temporal heads place each text on a unified regnal calendar axis (latent time scalar), supporting (a) absolute context injection and (b) relative biasing to form s_{ij}^{CTD} . (c) Interval-overlap supervision marks in-batch multi-positives by query–record overlap and trains a multi-positive contrastive loss.

Ground-truth records $d_j \in \mathcal{G}_i$ are exactly those that describe events or explicit non-events recorded during the queried interval Q_i , i.e., their time keys $\tau(d_j)$ fall within Q_i . The learning objective is to train a scoring function $S_\theta(q_i, d_j)$ that, for each query, ranks its ground-truth set \mathcal{G}_i ahead of the remaining elements of \mathcal{D} .

4.2 CTD: Calendrical Temporal Dual-encoder

We instantiate S_θ with a standard dual-encoder retriever. A shared Transformer encoder $f_\theta(\cdot)$ maps both temporal queries and candidate records into a common embedding space, producing pooled embeddings $\mathbf{h}_{q_i}, \mathbf{h}_{d_j} \in \mathbb{R}^H$. As a *purely semantic* baseline, we compute temperature-scaled dot-product similarities $s_{ij}^{\text{sem}} = s^{\text{sem}}(q_i, d_j) = \mathbf{h}_{q_i}^\top \mathbf{h}_{d_j} / \alpha$, for a mini-batch of B queries $\{q_i\}_{i=1}^B$ and B records $\{d_j\}_{j=1}^B$. Building on this semantic score, CTD augments the retriever with (i) an *absolute* calendrical context injected into the embeddings and (ii) a *relative* temporal bias added to the similarity, so that matches must agree in both meaning and calendrical position.

4.2.1 Latent calendar scalar

Reign-based month keys are discrete identifiers and do not directly provide a metric notion of *position* or *distance* across the stitched regnal calendar. To support both absolute positioning (for context injection) and relative offsets (for biasing), we therefore learn a continuous calendar axis where temporal

relations become measurable.

For any text x (either a query q_i or a record d_j), let $\mathbf{h}_x \in \mathbb{R}^H$ denote its pooled embedding. On top of \mathbf{h}_x , we attach three lightweight prediction heads for *gong*, *year*, and *month*. Each head produces logits over its discrete index set, which we normalize into distributions $\mathbf{p}_x^{(g)}, \mathbf{p}_x^{(y)}, \mathbf{p}_x^{(m)}$. Taking expectations yields soft calendrical coordinates g_x, y_x, m_x , which locate x on the ruler–year–month grid.

We then linearize this grid in calendar order and normalize it to $[0, 1]$, defining a shared latent time scalar

$$u_x = \frac{g_x \cdot (Y \cdot M) + y_x \cdot M + m_x}{G \cdot Y \cdot M - 1} \in [0, 1].$$

Here G, Y , and M denote the (padded) maximum numbers of gongs, years-per-gong, and months-per-year used to index the unified calendar. Texts from earlier dukes, years, or months receive smaller u_x than those later in the chronicle, enabling both relative distances Δu and absolute positions to be modeled on the same axis.

4.2.2 Absolute-temporal learning

We first exploit this signal in an *absolute* manner (Fig. 3 (a)): instead of feeding discrete (gong, year, month) indices as hard metadata, we convert the heads’ *soft* predictions into a continuous context vector and inject it into the embedding.

Reusing $\mathbf{p}_x^{(g)}, \mathbf{p}_x^{(y)}, \mathbf{p}_x^{(m)}$, we map each calendrical index to a fixed Fourier-style code and build

sinusoidal codebooks

$$E^{(g)} \in \mathbb{R}^{G \times D_t}, E^{(y)} \in \mathbb{R}^{Y \times D_t}, E^{(m)} \in \mathbb{R}^{M \times D_t}.$$

This fixed sinusoidal codebook provides a smooth, non-parametric absolute-position signal, avoiding a large learned embedding table for sparse calendrical indices. Taking expectations under $\mathbf{p}_x^{(\cdot)}$ yields a mixture representation that naturally reflects the model’s uncertainty instead of committing to a single hard index.

We obtain soft absolute-time contexts by taking expectations:

$$\mathbf{c}_x^{(g)} = \mathbf{p}_x^{(g)} E^{(g)}, \mathbf{c}_x^{(y)} = \mathbf{p}_x^{(y)} E^{(y)}, \mathbf{c}_x^{(m)} = \mathbf{p}_x^{(m)} E^{(m)}.$$

Concatenating and projecting yields

$$\mathbf{c}_x = W_{\text{ctx}}[\mathbf{c}_x^{(g)}; \mathbf{c}_x^{(y)}; \mathbf{c}_x^{(m)}] \in \mathbb{R}^H,$$

which we inject via a scalar-gated residual

$$\tilde{\mathbf{h}}_x = \mathbf{h}_x + \gamma \mathbf{c}_x,$$

where γ is learned.

We compute similarities with the context-enriched representations,

$$s_{ij}^{\text{abs}} = \tilde{\mathbf{h}}_{q_i}^\top \tilde{\mathbf{h}}_{d_j} / \alpha,$$

which reduces to the semantic baseline s_{ij}^{sem} when $\gamma = 0$.

4.2.3 Relative-temporal learning.

Building on the absolute similarity s_{ij}^{abs} , we further use the learned calendar axis to bias matching by relative offsets (Fig. 3 (b)). Given the latent coordinates u_{q_i} and u_{d_j} for a query–record pair (q_i, d_j) , we form the temporal offset

$$\Delta u_{ij} = u_{d_j} - u_{q_i} \in [-1, 1],$$

so that distances along the learned timeline can modulate how easily two texts should match. We embed this scalar with Fourier-style features

$$\phi(\Delta u_{ij}) \in \mathbb{R}^{D_\phi},$$

and apply a small MLP to produce an additive temporal bias

$$b_{ij}^{\text{time}} = \epsilon \text{MLP}(\phi(\Delta u_{ij})).$$

The final retrieval score is

$$s_{ij}^{\text{CTD}} = s_{ij}^{\text{abs}} + b_{ij}^{\text{time}},$$

where the learnable scale ϵ (initialized near zero) keeps the bias lightweight: when $\epsilon = 0$, CTD reduces to the absolute-only scorer s_{ij}^{abs} , and more generally the model can downweight this term if the learned calendar signal is unreliable.

4.3 Learning Objectives

We train the purely semantic dual-encoder baseline with a symmetric single-positive InfoNCE (Chen et al., 2020) objective over s_{ij}^{sem} . For CTD, we instead optimize a temporally aware multi-positive retrieval loss using the final scores s_{ij}^{CTD} .

Interval-overlap multi-positive retrieval. As shown in Fig. 3 (c), we treat temporal overlap as weak supervision: each query q_i targets an interval $Q_i = [\tau_i^{\min}, \tau_i^{\max}]$, and each record d_j carries a single month key $\tau(d_j)$ (i.e., $I_j = [\tau(d_j), \tau(d_j)]$). We mark in-batch positives by overlap,

$$P_i = \{j \mid Q_i \cap I_j \neq \emptyset\},$$

and optimize a multi-positive InfoNCE loss using the final scores s_{ij}^{CTD} :

$$\mathcal{L}_q^{\text{multi}} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\sum_{j \in P_i} \exp(s_{ij}^{\text{CTD}})}{\sum_{k=1}^B \exp(s_{ik}^{\text{CTD}})}.$$

The remaining in-batch records serve as negatives. We define $\mathcal{L}_d^{\text{multi}}$ symmetrically by transposing (s_{ij}^{CTD}) and use

$$\mathcal{L}_{\text{multi}} = \frac{1}{2} (\mathcal{L}_q^{\text{multi}} + \mathcal{L}_d^{\text{multi}}).$$

Auxiliary calendrical classification. To stabilize the absolute calendrical signal, we supervise the gong/year/month heads on passages with cross-entropy:

$$\mathcal{L}_{\text{time}} = \mathbb{E}_{d \sim \text{batch}} [\sum_{r \in \{g, y, m\}} \text{CE}(\mathbf{p}_d^{(r)}, y_d^{(r)})]$$

where $y_d^{(r)}$ are the ground-truth calendrical labels from the aligned time keys (queries are unlabeled).

Overall objective. We jointly optimize the retrieval and auxiliary temporal losses with a small weight λ_{time} :

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{multi}} + \lambda_{\text{time}} \mathcal{L}_{\text{time}}.$$

5 Experiments

5.1 Experiment Setting

We fine-tune two dual-encoder backbones, BERT-BASE-CHINESE and QWEN3-EMBED-0.6B, on ChunQiuTR. Model details are deferred to Appendix B.1.1, training cost and compute settings to Appendix B.1.2, and the full list of compared methods to Appendix B.3.

Method	Pub	Type	R@1	R@5	R@10	MRR@10	nDCG@10
<i>Sparse retrieval</i>							
BM25	–	Sparse	0.3962	0.5209	0.5620	0.4487	0.3404
BM25+TimeKDE	–	Sparse	0.4943	0.6086	0.6709	0.5456	0.4222
SPLADE-IDF _(ZS)	arXiv’24	Sparse	0.1361	0.2765	0.3569	0.1971	0.1596
SPLADE- ℓ_0 _(ZS)	SIGIR’25	Sparse	0.0006	0.0309	0.0587	0.0132	0.0143
<i>Fusion / late interaction</i>							
ColBERT-JINA _(ZS)	MRL’24	Fusion	0.2498	0.4102	0.4743	0.3167	0.2569
ColBERT-LFM2 _(ZS)	arXiv’25	Fusion	0.3345	0.4567	0.4894	0.3865	0.2691
<i>Dense retrieval (encoder-based)</i>							
mE5-Large _(ZS)	arXiv’24	Dense	0.2916	0.3969	0.4574	0.3389	0.2441
mE5-Large-ins _(ZS)	arXiv’24	Dense	0.2359	0.3545	0.4162	0.2862	0.2358
GTE-Large _(ZS)	arXiv’23	Dense	0.2293	0.3527	0.3890	0.2826	0.2188
BGE-Large-v1.5 _(ZS)	arXiv’23	Dense	0.2208	0.3430	0.4144	0.2775	0.2280
BGE-m3 _(ZS)	Findings ACL’24	Dense	0.2698	0.3775	0.4253	0.3135	0.2299
BERT-base _(FT)	NAACL’19	Dense	0.5088	0.6279	0.6727	0.5597	0.4283
BERT-base + TempDate _(FT)	SIGIR’23	Dense	0.5027	0.6165	0.6691	0.5508	0.4243
BERT-base + TempDate-Smooth _(FT)	PMLR’23	Dense	0.5051	0.6152	0.6673	0.5519	0.4244
CTD BERT-base (Ours)	This work	Dense	0.5826	0.6721	0.7090	0.6193	0.4575
<i>Dense retrieval (LM-based embeddings)</i>							
GTE-Qwen2-1.5B _(ZS)	arXiv’23	Dense	0.2783	0.4453	0.5009	0.3501	0.2613
E5-mistral-7B _(ZS)	ACL’24	Dense	0.2196	0.3212	0.3684	0.2619	0.2359
PQR (Qwen2.5-7B) _(re)	ACL’25	Dense	0.1585	0.3134	0.3805	0.2226	0.1712
PQR (Qwen3-8B) _(re)	ACL’25	Dense	0.0901	0.2184	0.3152	0.1481	0.1184
Qwen3-Embed-0.6B _(ZS)	arXiv’25	Dense	0.3376	0.4852	0.5354	0.3973	0.3107
Qwen3-Embed-4B _(ZS)	arXiv’25	Dense	0.4410	0.5783	0.6013	0.4985	0.3793
Qwen3-Embed-0.6B _(FT)	arXiv’25	Dense	0.5771	0.6376	0.6818	0.6045	0.4460
Qwen3-Embed-0.6B + TempDate _(FT)	SIGIR’23	Dense	0.5523	0.6425	0.6630	0.5924	0.4391
Qwen3-Embed-0.6B + TempDate-Smooth _(FT)	PMLR’23	Dense	0.5638	0.6346	0.6727	0.5942	0.4396
CTD Qwen3-Embed-0.6B (Ours)	This work	Dense	0.5923	0.6485	0.6927	0.6194	0.4575

Table 1: Test-set retrieval performance on our ChunQiuTR benchmark under the official evaluation protocol.

5.2 Main Results

From Table 1, ChunQiuTR is clearly non-trivial and strongly time-sensitive: most zero-shot sparse, fusion, and dense retrievers lag behind tuned BM25, while a simple temporal prior (BM25+TimeKDE) yields a large gain over BM25 and nearly matches supervised dense models. On encoder-based dense retrievers, in-domain fine-tuning already outperforms BM25+TimeKDE, generic dating auxiliaries (TempDate / TempDate-Smooth) give little benefit, and adding our CTD objectives on the same backbone yields a clear boost in early precision (around +7–8 points on R@1). LM-based dense retrievers show a similar pattern: zero-shot LM encoders and PQR pipelines underperform BM25+TimeKDE, lightly fine-tuned Qwen3-Embed-0.6B is strong, and the CTD-enhanced variant further improves early precision and achieves the best overall scores (R@1, MRR@10, nDCG@10), indicating that explicit time-key supervision adds fine-grained temporal structure beyond simple priors or auxiliary dating heads.

Cross-corpus pilot on Zizhi Tongjian. As an out-of-domain probe, we further evaluate ChunQiuTR-

Subset	Records	Queries	FT baseline		CTD (ours)	
			MRR	R@1	MRR	R@1
Qi Ji (part)	268	92	0.2081	0.1848	0.2304	0.2065
Jin Ji (part)	820	119	0.1598	0.1345	0.1751	0.1597

Table 2: Cross-corpus pilot on processed *Zizhi Tongjian* subsets. No target-corpus training is performed.

trained retrievers on two processed subsets from *Zizhi Tongjian*, an annalistic general history that also records events under non-Gregorian, reign-based temporal expressions (Appendix B.4). For each subset, we build a month-level gallery from event-bearing lines and automatically instantiate one point-style query for each unique month key using traditional reign-year expressions, without any additional training on the target corpus. Unlike the full ChunQiuTR benchmark, this pilot does not reconstruct explicit no_event months, commentary-derived hard negatives, or the full point/gap/window query families, and is therefore intended as a lightweight cross-corpus transfer probe rather than a second benchmark.

As shown in Table 2, CTD consistently improves both MRR and R@1 on the two *Zizhi Tongjian* subsets without any target-corpus fine-tuning. Al-

Method	Single-month ($ Q = 1$)		Multi-month ($ Q > 1$)	
	R@1	MRR@10	R@1	MRR@10
BM25	0.397	0.413	0.396 ↓-0.001	0.499 ↑+0.086
ColBERT-LFM2 _(ZS)	0.298	0.312	0.386 ↑+0.088	0.490 ↑+0.178
mE5-Large _(ZS)	0.259	0.279	0.337 ↑+0.078	0.422 ↑+0.143
BERT-base _(FT)	0.497	0.516	0.525 ↑+0.027	0.621 ↑+0.106
CTD BERT-base (Ours)	0.509	0.530	0.685 ↑+0.176	0.744 ↑+0.214
Qwen3-Embed-0.6B _(ZS)	0.353	0.385	0.317 ↓-0.036	0.415 ↑+0.031
Qwen3-Embed-0.6B _(FT)	0.481	0.495	0.711 ↑+0.230	0.757 ↑+0.262
CTD Qwen3-Embed-0.6B (Ours)	0.491	0.513	0.733 ↑+0.242	0.767 ↑+0.255

Table 3: Impact of query span on retrieval performance on the test set. We compare single-month ($|Q| = 1$) and multi-month ($|Q| > 1$) queries; for multi-month queries, the numbers in parentheses give absolute changes relative to single-month queries.

Figure 4 shows three examples of query results. Each example has a 'QUERY' and a 'Baseline' section. The 'Ours' section shows results that are more accurate and relevant than the 'Baseline' results. For example, in the first example, the 'Ours' result correctly identifies the event as '魯莊公三年三月' (Spring 3rd year of Duke Zhuang), while the 'Baseline' results are incorrect or less relevant.

Figure 4: Visualization of Qualitative Examples.

though this pilot is intentionally lighter than the full ChunQiuTR setup, the trend suggests that the temporal-consistency bias learned on ChunQiuTR transfers beyond the source corpus and continues to help distinguish chrono-near but temporally mismatched evidence.

5.3 Analysis

5.3.1 Impact of Query Type

Table 3 compares single-month ($|Q| = 1$) and multi-month ($|Q| > 1$) queries. Across most methods, multi-month queries substantially boost MRR@10 and give small gains or no change in R@1, reflecting the fact that it is easier to hit any correct month within a span than a single target month. Our CTD models achieve the best performance in both regimes, with especially large gains on multi-month queries for the BERT backbone (roughly +0.16 R@1) and consistent improvements for Qwen3-Embed, indicating better temporal ordering under chrono-near confounds.

5.3.2 Qualitative Examples

Fig. 4 illustrates the contrast between a BERT-based baseline and our time-aware retriever via two queries. For a point query, the baseline is distracted by frequent no_event templates and events from neighboring months, while our model correctly locates the target chronicle entry. For a

Variant	$\mathcal{L}_{\text{multi}}$	b_{ij}^{time}	c_x	R@1	MRR@10
FT baseline	—	—	—	0.5771	—
+ $\mathcal{L}_{\text{multi}}$	✓	—	—	0.5820 ↑+0.0049	0.6107 ↑+0.0063
+ Bias	✓	✓	—	0.5898 ↑+0.0127	0.6135 ↑+0.0091
+ Ctx	✓	—	✓	0.5850 ↑+0.0079	0.6134 ↑+0.0090
Full (Ours)	✓	✓	✓	0.5923 ↑+0.0152	0.6194 ↑+0.0150

Table 4: Ablation study on the test set under the same evaluation protocol as Table 1. For each metric, the second line reports the change relative to the FT baseline.

broader window query, the baseline’s results are dispersed across later exegetic discussions, whereas our model concentrates probability on the correct local window, retrieving both the pertinent event and an explicit no_event record. Overall, these examples show that our retriever couples temporal reasoning with semantic matching beyond surface cues; moreover, temporal errors are often *confident* rather than uncertain, motivating retrieval-time temporal constraints over downstream generation fixes (see Appendix B.2.4 and Appendix B.2.5).

5.3.3 Ablation study

We ablate three components: the retrieval objective $\mathcal{L}_{\text{multi}}$, the relative-time logit bias b_{ij}^{time} , and the soft absolute temporal context c_x . Starting from the FT baseline, adding $\mathcal{L}_{\text{multi}}$ leads to a modest but consistent improvement. This indicates that explicit retrieval supervision enhances time-key discrimination beyond standard fine-tuning. Adding either temporal signal further improves performance. The logit bias yields a larger gain in R@1, suggesting that it effectively reshapes in-batch matching toward chronologically plausible candidates. In contrast, injecting temporal context achieves comparable improvements in MRR, indicating better overall ranking quality. Combining the bias and the context produces the best results, with additive gains over each individual component. This supports their complementary roles: the bias calibrates pairwise similarities, while the context enriches representations with absolute-time distributional cues.

6 Conclusion

We presented **ChunQiuTR**, a time-keyed temporal retrieval dataset built on the *Spring and Autumn Annals* and its commentarial tradition. ChunQiuTR operationalizes a non-Gregorian, reign-based month timeline (gong-year-month) and evaluates retrieval under realistic historical confounders—lexical-near same-key materials, adjacent-month near-misses, and explicit

no_event months—making temporal fidelity a first-class requirement beyond topical relevance. We further proposed **CTD**, a calendrically time-aware dual-encoder that augments semantic matching with absolute context injection and relative offset biasing. Against strong semantic dual-encoder baselines, CTD consistently improves retrieval quality and reduces chrono-near confusions that can mislead evidence-grounded systems.

Limitations

ChunQiuTR is constructed from the *Chunqiu* annals and their major commentaries, and is organized around reign-based month-level time keys. This narrow scope limits the generality of our findings: retrieval behaviors and error patterns may differ in other pre-modern corpora with different calendrical conventions, narrative styles, and editorial traditions, and our results therefore do not directly imply the same gains under those settings. Extending the same construction procedure to other dynastic corpora would require additional source-specific normalization, temporal alignment, and curation, which we leave to future work.

Moreover, a month-level timeline cannot represent finer-grained temporal relations within a month, and the remaining errors suggest that two cases are still particularly challenging even with temporal supervision: near-duplicate records in neighboring months, and genuinely ambiguous historiographical cases that are difficult to disambiguate from local textual evidence alone. Future work includes extending the benchmark to broader historical corpora and finer temporal granularity, incorporating stronger reranking or evidence-checking modules for borderline confusions, and evaluating downstream impacts in end-to-end RAG pipelines.

Potential Risks. We consider the overall risk of this work to be relatively low, since ChunQiuTR is intended as a research benchmark for temporal retrieval over classical Chinese historical texts. However, retrieval errors or temporal misalignment may still mislead downstream historical interpretation if model outputs are treated as authoritative evidence rather than retrieval support. In addition, benchmark construction choices, such as source selection, commentary filtering, and month-key normalization, may introduce biases inherited from the source materials and from our curation pipeline. Accordingly, the benchmark and released models

should be used with human verification in scholarly or educational settings.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62276283 and 62372281, in part by the China Meteorological Administration’s Science and Technology Project under Grant CMA-JBGS202517, in part by Guangdong-Hong Kong-Macao Greater Bay Area Meteorological Technology Collaborative Research Project under Grant GHMA2024Z04, in part by Fundamental Research Funds for the Central Universities, Sun Yat-sen University under Grant 23hytd006 and 23hytd006-2, in part by Guangdong Provincial High-Level Young Talent Program under Grant RL2024-151-2-11, in part by the Key Development Project of the Artificial Intelligence Institute, Sun Yat-sen University, and in part by The Major Key Project of PCL (Grant No. PCL2025A17).

References

- Jiahuan Cao, Dezhi Peng, Peirong Zhang, Yongxin Shi, Yang Liu, Kai Ding, and Lianwen Jin. 2024. [TongGu: Mastering classical Chinese understanding with knowledge-grounded large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4196–4210, Miami, Florida, USA. Association for Computational Linguistics.
- Yuwei Cao, William Groves, Tanay Kumar Saha, Joel Tetreault, Alejandro Jaimes, Hao Peng, and Philip Yu. 2022. [XLTime: A cross-lingual knowledge transfer framework for temporal expression extraction](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1931–1942, Seattle, United States. Association for Computational Linguistics.
- Alejandro Sánchez de Castro, Lourdes Araujo, and Juan Martínez-Romo. 2025. [A novel methodology for enhancing cross-language and domain adaptability in temporal expression normalization](#). *Computational Linguistics*, 51(4):1303–1335.
- Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. [M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. [A simple framework for](#)

- contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.
- Wenhu Chen, Xinyi Wang, William Yang Wang, and William Yang Wang. 2021. [A dataset for answering time-sensitive questions](#). In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.
- Ziyang Chen, Erxue Min, Xiang Zhao, Yunxin Li, Xin Jia, Jinzhi Liao, Jichao Li, Shuaiqiang Wang, Baotian Hu, and Dawei Yin. 2025. [A question answering dataset for temporal-sensitive retrieval-augmented generation](#). *Scientific Data*, 12(1):1855.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. [Time-aware language models as temporal knowledge bases](#). *Transactions of the Association for Computational Linguistics*, 10:257–273.
- Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. [Splade: Sparse lexical and expansion model for first stage ranking](#). In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21*, page 2288–2292, New York, NY, USA. Association for Computing Machinery.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. [Retrieval-augmented generation for large language models: A survey](#). *arXiv preprint arXiv:2312.10997*, 2(1).
- Zhichao Geng, Yiwen Wang, Dongyu Ru, and Yang Yang. 2025. [Towards competitive search relevance for inference-free learned sparse retrievers](#). *Preprint*, arXiv:2411.04403.
- Arianna Graciotti, Leonardo Piano, Nicolas Lazzari, Enrico Daga, Rocco Tripodi, Valentina Presutti, and Livio Pompianu. 2025. [KE-MHISTO: Towards a multilingual historical knowledge extraction benchmark for addressing the long-tail problem](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20316–20339, Vienna, Austria. Association for Computational Linguistics.
- SeungYoon Han, Taeho Hwang, Sukmin Cho, Soyeong Jeong, Hoyun Song, Huije Lee, and Jong C. Park. 2025. [Temporal information retrieval via time-specifier model merging](#). In *Proceedings of the 3rd Workshop on Towards Knowledgeable Foundation Models (KnowFM)*, pages 1–13, Vienna, Austria. Association for Computational Linguistics.
- Rohan Jha, Bo Wang, Michael Günther, Georgios Mastrotras, Saba Sturua, Isabelle Mohr, Andreas Koukounas, Mohammad Kalim Wang, Nan Wang, and Han Xiao. 2024. [Jina-ColBERT-v2: A general-purpose multilingual late interaction retriever](#). In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 159–166, Miami, Florida, USA. Association for Computational Linguistics.
- Junfeng Kang, Rui Li, Qi Liu, Yanjiang Chen, Zheng Zhang, Junzhe Jiang, Heng Yu, and Yu Su. 2025. [PQR: Improving dense retrieval via potential query modeling](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13455–13469, Vienna, Austria. Association for Computational Linguistics.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.
- Omar Khattab and Matei Zaharia. 2020. [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, SIGIR '20*, page 39–48, New York, NY, USA. Association for Computing Machinery.
- Natalia Korchagina. 2016. [Building a gold standard for temporal entity extraction from medieval german texts](#). In *"2016 Conference on Language Technologies and Digital Humanities"*.
- Yibin Lei, Liang Ding, Yu Cao, Changtong Zan, Andrew Yates, and Dacheng Tao. 2023. [Unsupervised dense retrieval with relevance-aware contrastive pre-training](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10932–10940, Toronto, Canada. Association for Computational Linguistics.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive nlp tasks](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.
- Xiaoyan Li and W. Bruce Croft. 2003. [Time-based language models](#). In *Proceedings of the Twelfth International Conference on Information and Knowledge*

- Management*, CIKM '03, page 469–475, New York, NY, USA. Association for Computing Machinery.
- Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. [Towards general text embeddings with multi-stage contrastive learning](#). *Preprint*, arXiv:2308.03281.
- Yang Liu, Lan Lan, Jiahuan Cao, Hiuyi Cheng, Kai Ding, and Lianwen Jin. 2025. [Large-scale corpus construction and retrieval-augmented generation for Ancient Chinese poetry: New method and data insights](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 779–817, Albuquerque, New Mexico. Association for Computational Linguistics.
- Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and 1 others. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*.
- Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022a. [Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874, Dublin, Ireland. Association for Computational Linguistics.
- Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. 2022b. [Large dual encoders are generalizable retrievers](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Bhawna Piryani, Abdelrahman Abdallah, Jamshid Mozafari, Avishek Anand, and Adam Jatowt. 2025. [It's high time: A survey of temporal question answering](#). *Preprint*, arXiv:2505.20243.
- Thilina C. Rajapakse. 2023. [Dense passage retrieval: Architectures and augmentation methods](#). In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 3494, New York, NY, USA. Association for Computing Machinery.
- Stephen Robertson and Hugo Zaragoza. 2009. [The probabilistic relevance framework: Bm25 and beyond](#). *Found. Trends Inf. Retr.*, 3(4):333–389.
- Xinjie Shen, Zhichao Geng, and Yang Yang. 2025. [Exploring 0 sparsification for inference-free sparse retrievers](#). In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 2572–2576, New York, NY, USA. Association for Computing Machinery.
- Xin Su, Phillip Howard, and Steven Bethard. 2025. [Transformer-based temporal information extraction and application: A review](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28822–28841, Suzhou, China. Association for Computational Linguistics.
- Liquid AI Team. 2025. [Lfm2 technical report](#). *Preprint*, arXiv:2511.23404.
- Jiexin Wang, Adam Jatowt, Masatoshi Yoshikawa, and Yi Cai. 2023. [Bitimebert: Extending pre-trained language representations with bi-temporal information](#). In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 812–821, New York, NY, USA. Association for Computing Machinery.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. [Improving text embeddings with large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, Bangkok, Thailand. Association for Computational Linguistics.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. [Multilingual e5 text embeddings: A technical report](#). *Preprint*, arXiv:2402.05672.
- Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. [C-pack: Packaged resources to advance general chinese embedding](#). *Preprint*, arXiv:2309.07597.
- Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. [Approximate nearest neighbor negative contrastive learning for dense text retrieval](#). In *International Conference on Learning Representations*.
- Hugo Yèche, Alizée Pace, Gunnar Ratsch, and Rita Kuznetsova. 2023. [Temporal label smoothing for early event prediction](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 39913–39938. PMLR.
- Siyue Zhang, Yuxiang Xue, Yiming Zhang, Xiaobao Wu, Anh Tuan Luu, and Chen Zhao. 2025a. [Mrag: A modular retrieval framework for time-sensitive question answering](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 3080–3118.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025b. [Qwen3 embedding: Advancing text embedding and reranking through foundation models](#). *Preprint*, arXiv:2506.05176.

Yuqing Zhang, Baoyi He, Yihan Chen, Hangqi Li, Han Yue, Shengyu Zhang, Huaiyong Dou, Junchi Yan, Zemin Liu, Yongquan Zhang, and Fei Wu. 2024. *PhiloGPT: A philology-oriented large language model for Ancient Chinese manuscripts with dunhuang as case study*. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2784–2801, Miami, Florida, USA. Association for Computational Linguistics.

A Details of Dataset

A.1 Annals and Exegetical Layers

The *Chunqiu* corpus combines two tightly coupled layers. The annals themselves are extremely terse month-level records, which provide the primary temporal anchors on the Lu-state regnal timeline. By contrast, the three classical *zhuan* (*Zuo*, *Gongyang*, and *Guliang*) expand the same entries into narrative, interpretive, or doctrinal prose. In *ChunQiuTR*, we treat the annals line as the anchor and the aligned *zhuan* passages as semantically richer event descriptions under the same reign-based time key.

Table 5 illustrates this structure with the canonical case “郑伯克段于鄆” (*Duke Yin, Year 1, Month 5*). A single annals line is expanded by the three *zhuan* in different ways, while later commentarial and historiographical sources further paraphrase or reinterpret the same event. This layered organization is central to our benchmark design: it provides both aligned event records and naturally occurring chrono-near but non-target passages.

A.2 From Parallel Texts to Reign-Based Time Keys

A.2.1 Reign-based time keys

Our normalized time axis uses month-level keys of the form


$$\tau = (\text{gong, year, month}),$$

for example 「鲁隐公元年正月」 or 「鲁桓公元年三月」. We scan the annals sequentially while maintaining the current triple (g, y, m) . Full reign cues initialize a new key, bare year markers start a new regnal year under the current ruler, month markers update only m , and sentences without new temporal cues inherit the current triple. Table 6 shows representative cases.

A.2.2 Record–time-key alignment

We align event-level records under each normalized time key using lightweight LLM suggestions

System: You are a scholar of Chunqiu.
Prompt:
 Given the following passages under the same time label, group them by event. Passages describing the **same event** should be placed into one group. Separate different groups with three blank lines. Preserve the original formatting and **do not** add any extra content: {text}

 **[春秋]** 冬，十有二月，祭伯來。公子益師卒。
[春秋公羊傳] 冬，十有二月，祭伯來。祭伯者何？… 公子益師卒。何以不日？遠也。…
[春秋穀梁傳] 冬，十有二月，祭伯來。來者… 公子益師卒…
[春秋左氏傳] 十二月，祭伯來，非王命也。眾父卒。公不與小斂，故不書日。

Output:

$D_r^{(1)}$ **[春秋]** 冬，十有二月，祭伯來。
[春秋公羊傳] 冬，十有二月，祭伯來。祭伯者…
[春秋穀梁傳] 冬，十有二月，祭伯來。來者…
[春秋左氏傳] 十二月，祭伯來，非王命也。

$D_r^{(2)}$ **[春秋]** 公子益師卒。
[春秋公羊傳] 公子益師卒。何以不日？遠也…
[春秋穀梁傳] 公子益師卒。大夫日卒…
[春秋左氏傳] 眾父卒。公不與小斂，故不書日。

Figure 5: Example of event-level grouping under the reign-based time key 「鲁隐公元年十二月」. The model is prompted to split mixed passages and group aligned commentary snippets by event.

followed by manual verification. This is necessary because a single annals line may compress multiple events, while commentary passages may expand one event across several fragments. Figure 5 shows a representative case under the time key 「鲁隐公元年十二月」.

In this case, the mixed annals line is split into two records, one for 「祭伯来」 and one for 「公子益师卒」, each grouped with its aligned *zhuan* passages. All such suggestions are manually checked before entering the benchmark.

A.3 Details of Chrono-near counterfactual negatives

A.3.1 Classical sources and perspectives

Our chrono-near counterfactual negatives draw on several classical works that reorganize or reinterpret *Chunqiu* events from distinct perspectives:

Gu Donggao’s Chronological Tables (顾栋高《春秋大事表》). Gu’s Qing-dynasty compilation systematically re-orders the *Chunqiu* and the three *zhuan* into explicit chronological tables. Each entry typically specifies the state, the reigning gong, the year, and a short prose summary of the event,

Text (excerpt + short gloss)	Source	Label
夏，五月，郑伯克段于鄢。	《春秋》 <i>Chunqiu</i>	
<i>Annals anchor: Zheng subdued Duan at Yan in month 5.</i>	经文 / <i>annals</i>	event
夏，五月，郑伯克段于鄢。克之者何？杀之也。杀之则曷为谓之克？大郑伯之恶也。段者何？郑伯之弟也。何以不称弟？当国也。	《春秋公羊传》 <i>Gongyang zhuan</i>	
<i>Gongyang reads ke as killing and amplifies moral blame.</i>	传文 / <i>zhuan</i>	event
夏，五月，郑伯克段于鄢。克者何？能也。何能也？能杀也。何以不言杀？见段之有徒众也。段，郑伯弟也。……段失子弟之道矣，贱段而甚郑伯也。	《春秋谷梁传》 <i>Guliang zhuan</i>	
<i>Guliang stresses armed revolt, conduct, and intensified censure.</i>	传文 / <i>zhuan</i>	event
书曰：“郑伯克段于鄢。”段不弟，故不言弟；如二君，故曰克；称郑伯，讥失教也，谓之郑志。不言出奔，难之也。	《春秋左传》 <i>Zuo zhuan</i>	
<i>Zuo treats wording itself as a signal of layered blame.</i>	传文 / <i>zhuan</i>	event
案左氏云段出奔共，而公、谷皆曰杀。据隐十一年传，庄公曰：“寡人有弟不能和协，使糊其口于四方”，则未杀明矣，公、谷之说非是。	《春秋大事表》 <i>Chronological</i>	
<i>Gu Donggao rejects the reading that Duan was killed.</i>	顾栋高 / <i>Qing</i>	neg
不称国讨而言郑伯，讥失教也。段不弟，故不言弟，明郑伯虽失教而段亦凶逆。	《春秋左传注》 <i>Zuo zhuan zhu</i>	
<i>Du Yu explains how wording encodes both political and personal blame.</i>	杜预 / <i>Jin</i>	neg
以“国讨”“得隳曰克”等例，说明称郑伯乃罪君，不称弟乃罪段，兼示兄虽失教而弟为乱首。	《春秋左传疏》 <i>Zuo zhuan shu</i>	
<i>Kong Yingda systematizes the case through doctrinal categories.</i>	孔颖达 / <i>Tang</i>	neg

Table 5: Aligned materials under the reign-based time key “鲁隐公元年五月” for “郑伯克段于鄢”. The annals give the anchor, the three *zhuan* provide aligned expansions, and later sources yield chrono-near non-target paraphrases.

sometimes highlighting cross-state interactions or disagreements among the Zuo, Gongyang, and Guliang traditions. Compared to the terse annals, these tables provide a more “modernized” timeline and condensed paraphrases of events, which we reuse as temporally grounded, paraphrastic negatives.

Wei Liaoweng’s *Chunqiu Zuozhuan Yaoyi* (魏了翁《春秋左传要义》). Wei’s Southern Song work focuses on extracting the “essential meanings” of *Zuo zhuan* episodes. His prose often paraphrases the underlying narrative, emphasizes moral and ritual judgments, and occasionally re-groups several *Zuo* passages into a single didactic unit. From our perspective, these are high-level, discursive restatements of the same historical events, written in a style that differs noticeably from the base corpus.

Zuoshu annotations and sub-commentaries (注疏). In addition, we employ the *Siku Quanshu* edition of *Zuozhuan* annotations, which combines multiple layers: Lu Deming’s *yin yi* (音义), Du Yu’s Jin-dynasty commentary, Kong Yingda’s Tang-dynasty *Chunqiu Zhengyi*, and later Song-dynasty notes such as Lü Zuqian’s *Zuozhuan shuō*. These texts embed glosses, philological notes, and exegetical reformulations around the same events. While they do not always restate the full narrative, they frequently echo key phrases, name important actors, or re-frame the event in ritual or moral terms.

Taken together, these sources provide us with multiple “views” on the same historical episodes: terse annal entries, narrative expansions in the three *zhuan*, tabular re-organizations (Gu Donggao), moral-didactic summaries (Wei Liaoweng),

Original chronicle snippet	Cue / update	Normalized time key
元年，春，王正月。	Initialize a new reign-year (隐公元年), month = 正月	鲁隐公元年正月
三月，公及邾仪父盟于蔑。	New month (三月), inherit current reign-year	鲁隐公元年三月
二年，春，公会戎于潜。	New year (二年) under the same ruler, reset month to 正月	鲁隐公二年正月
元年，春，王正月，公即位。	New ruler detected (桓公), reset year to 元年 and month to 正月	鲁桓公元年正月

Table 6: Representative mappings from raw chronicle phrases to normalized reign-based time keys.

and layered annotations (注疏). By aligning them into the reign-based time-key space defined in the main text, we obtain chrono-near passages that are temporally co-located with our ground truth records D_τ but often differ in wording, emphasis, or even event granularity.

A.3.2 LLM-assisted reverse matching and fuzzy alignment

For sources like Lü Zuqian’s *Chunqiu ZuoZhuan Shuos* (《春秋左氏传说》) and certain annotation layers, the text is often organized as short titled sections (e.g., event summaries or topic headings) rather than direct quotations of the base corpus. To map these paraphrastic units back to concrete passages in the annals and *Zuo zhuan*, we adopt an LLM-assisted “reverse matching” strategy, illustrated in Fig. 6.

Given a candidate item with a ruler name, an approximate year range, and a short event title (e.g., from Lü’s *Chunqiu ZuoZhuan Shuos*), we query a classical-Chinese LLM (DeepSeek) as a virtual *Zuo zhuan* expert. When the model can make a judgment, it must return only the original *Zuo zhuan* text segment; when it is uncertain, it must output the sentinel token NONE. In all cases, these LLM suggestions are further filtered and manually checked before being accepted into our aligned record set.

Once a plausible *Zuo zhuan* span has been suggested and validated, it can be matched back to the digitized base text with simple fuzzy string matching, which uniquely anchors the passage to its canonical location and the corresponding reign-based time key τ .

A.4 Query Types and Templates

We instantiate a small set of Traditional-Chinese natural-language templates over normalized reign-based month keys (e.g., “公元年二月”). Queries are divided into *point* queries, which target a single

Template groups (index range)	Type
BASE (1–12)	point, content-oriented
BASE (13–20)	point, existence / no-event
MONTH_PAST (21–26)	window, past
MONTH_FUTURE (27–31)	window, future
MONTH_AROUND (32–36)	window, around
MONTH_RANGE (37–41)	window, range
YEAR_CURRENT (42–46)	window, current year
YEAR_PAST (47–49)	window, previous year
YEAR_FUTURE (50–52)	window, next year

Table 7: Query template groups used in ChunQiuTR.

month, and *window* queries, which target a span around a reference month. Point queries include both content-oriented and existence-oriented formulations, while window queries cover past, future, around, and explicit-range retrieval. Figure 7 illustrates representative temporal interpretations, and Table 7 summarizes the template groups used in experiments. Empty months are handled through the same point/window formulations via `no_event` placeholders when the target month or span contains no recorded event.

A.5 Statistics of Dataset

A.5.1 Raw unit-level length statistics (before sentence splitting)

Before sentence-level segmentation, we collect raw records from the *Chunqiu* annals and the three traditional *zhuan* (positive pool), as well as later exegetical layers such as *Zuoshi zhuan* and *Chunqiu Zhengyi* (negative pool). Table 8 reports basic character-length statistics over these raw records.

The raw source pools differ substantially in length and discourse style, especially between canonical records and later exegetical materials, which motivates sentence-level segmentation before retrieval construction.

Fig. 8 summarizes month-level coverage and gap months over the normalized *Chunqiu* timeline. The benchmark spans 3036 reign-based months from

Pool	Source	# Raw units	Total chars (K)	Avg. len.	Median	Min	Max
Positive	All (annals + three <i>zhuan</i>)	6641	293.7	44.22	–	–	–
Positive	<i>Chunqiu</i> annals	1532	19.2	12.52	10	2	71
Positive	<i>Gongyang zhuan</i>	1776	44.5	25.05	10	2	671
Positive	<i>Zuo zhuan</i>	1547	189.1	122.20	48	3	2658
Positive	<i>Guliang zhuan</i>	1786	41.0	22.93	11	2	467
Negative	All exegetical layers	9227	1014.9	109.99	–	–	–
Negative	Lü Zuqian	241	99.0	410.70	365	14	1715
Negative	Kong Yingda	3653	483.5	132.36	83	0	2243
Negative	Du Yu	3652	218.1	59.72	49	1	714
Negative	Gu Donggao	481	39.3	81.80	62	1	597
Negative	Wei Liaoweng	1200	174.9	145.79	106	10	2877

Table 8: Raw record-level character-length statistics before sentence splitting. Character counts are reported in thousands (K). Positive records come from the *Chunqiu* annals and the three traditional *zhuan*, while negative records come from later exegetical layers.

Duke Yin to Duke Ai, with a substantial proportion of months containing no recorded event, confirming that gap months are a pervasive property of the corpus rather than an edge case.

A.5.2 Record-level segmentation.

To construct retrieval units, we convert heterogeneous raw source materials into sentence-level records. We first use an LLM to propose punctuation and sentence boundaries (句读) for each raw passage, and then apply a light rule-based splitter over classical discourse markers such as “曰”, “云”, and “传曰”. During this step, we also perform minimal normalization to reduce stylistic boilerplate, for example by stripping framing markers such as “正义曰” or formulaic quotation headers that do not contribute substantive event content, and by discarding extremely short fragments that contain only a few characters.

Each resulting record inherits the reign-based time key and source metadata of its parent unit, and is assigned a coarse type label: event, no_event, or neg_comment. For months where the annals and the three *zhuan* jointly indicate that nothing was recorded, we additionally synthesize a standardized no_event record for that time key (e.g., “鲁隐公元年二月：《春秋》经文及三传于此月无事可书。”), so that retrieving an empty-month case still requires matching the correct reign and month, rather than collapsing all such queries to a single global “nothing happened” entry.

This process yields 20,172 records in total, which constitute the retrieval gallery used in our time-aware experiments (Table 9).

A.5.3 Final benchmark splits.

The benchmark is split at the month level using an approximate 80/10/10 partition, and all records and queries inherit the split of their associated time key to avoid temporal leakage. As shown in Table 9, the final benchmark contains 3036 month keys, 20,172 record-level retrieval units, and 16,226 queries, with explicit breakdowns by split and record type.

A.5.4 Data sources and licensing.

Source & license. All digitized texts used in ChunQiuTR are retrieved from Chinese Wikisource (*Siku Quanshu* editions). Individual work pages are tagged as public domain (e.g., PD-old), while platform content is provided under CC BY-SA 4.0 and the Wikimedia Terms of Use. To facilitate compliant reuse, we record page revision IDs (oldid) and release the benchmark as derived metadata together with scripts for re-downloading the raw texts.

B Details of Methods

B.1 Details of Experiment Setting

B.1.1 Model and Training Details

We fine-tune two dual-encoder backbones: BERT-BASE-CHINESE and QWEN3-EMBED-0.6B. For BERT-BASE, we use [CLS] pooling; for QWEN3-EMBED-0.6B, we use last-token pooling.

Both backbones are trained with a contrastive retrieval objective using multi-positive supervision and explicit hard-negative training. We additionally apply an auxiliary time classification loss over the three discrete factors (gong/year/month), with weight 0.1 and label smoothing $\epsilon = 0.2$. For CTD, we enable both the relative temporal bias and the

Split	# months	# records	# queries	Avg. ground-truth recs/query	# event recs	# no-event recs	# neg. comments
Train	2424	16027	13053	7.3	5360	1209	9458
Validation	295	2049	1520	6.8	626	152	1271
Test	317	2096	1653	7.2	782	149	1165
Total	3036	20172	16226	7.2	6768	1510	11894

Table 9: Final benchmark statistics and splits over month-level time keys, record-level retrieval units, and queries. The ‘‘Avg. ground-truth recs/query’’ column reports the average number of labeled relevant records per query in each split, and the last three columns break down records by type (event, no_event, and neg_comment).

Work / layer	Role in ChunQiuTR	Digital source (edition)	License note / release plan
<i>Chunqiu</i> (春秋)	base corpus	Chinese Wikisource (<i>Siku Quanshu</i> edition; page revision oldid recorded)	Work pages: tagged PD-old. Platform text: CC BY-SA 4.0 (Wikimedia Terms of Use). We release derived metadata (time keys, alignments, queries/qrels, indices) and scripts to re-fetch raw texts from recorded oldid revisions.
<i>Zuo zhuan</i> (左氏)	base corpus		
<i>Gongyang zhuan</i> (公羊)	base corpus		
<i>Guliang zhuan</i> (梁)	base corpus		
Gu Donggao, <i>Chunqiu Dashibiao</i> (栋高《春秋大事表》)	chrono-near paraphrastic negatives		
Wei Liaoweng, <i>Chunqiu Zuozhuan Yaoyi</i> (魏了翁《春秋左要》)	chrono-near discursive negatives		
<i>Zuozhuan</i> annotations / sub-commentaries (注疏; e.g., 音/杜注/正)	lexical-near annotation negatives		

Table 10: Text sources and licensing. All digitized texts are retrieved from Chinese Wikisource (*Siku Quanshu* editions), with page revision IDs (oldid) recorded for traceability. See Appendix A for alignment and preprocessing.

soft absolute temporal context derived from predicted time distributions.

We optimize with AdamW (weight decay 0.01) and a linear learning-rate schedule with warmup ratio 0.1. For BERT-BASE, we train for 5 epochs with batch size 64 and learning rate 2×10^{-5} , using maximum query/passage lengths of 64/196. For QWEN3-EMBED-0.6B, we train for 3 epochs with effective batch size 16 and learning rate 3×10^{-6} , using maximum query/passage lengths of 128/256; we also enable global in-batch negatives.

We select checkpoints by validation **Recall@1** and report **Recall@K** and **MRR@10** on the test split under the same evaluation protocol. During evaluation, both commentary negatives and explicit no_event records are included in the candidate gallery.

B.1.2 Training Cost and Computational Resources

We report the computing infrastructure and approximate GPU-hours for representative fine-tuning runs.

Hardware. BERT-BASE-CHINESE is trained on a single GPU (either $1 \times$ NVIDIA RTX A6000 or $2 \times$ RTX 3090, depending on availability). QWEN3-EMBED-0.6B uses multi-GPU distributed

Backbone	Variant	#GPU	Time / run	GPU-hours
BERT-BASE-CHINESE	FT baseline	1	≈ 15 min	≈ 0.25
BERT-BASE-CHINESE	CTD (full)	1	≈ 19 min	≈ 0.32
QWEN3-EMBED-0.6B	FT baseline	2	≈ 45 min	≈ 1.50
QWEN3-EMBED-0.6B	CTD (full)	2	≈ 45 min	≈ 1.50

Table 11: Compute cost for representative fine-tuning runs on ChunQiuTR. Time is wall-clock per run; GPU-hours are computed as (#GPU) \times (time in hours).

training (either $2 \times$ RTX A6000 or $4 \times$ RTX 3090) to support global in-batch negatives.

Training cost. Table 11 reports representative wall-clock time per run and the corresponding GPU-hours. These numbers cover end-to-end fine-tuning with periodic validation; sparse baselines and non-parametric time priors incur negligible training cost.

Hyperparameters. We do not perform large-scale hyperparameter sweeps. Instead, we adopt standard fine-tuning settings for each backbone and select the best checkpoint by validation **Recall@1**.

B.2 Details of Analysis

B.2.1 Details of Ablation Study

We further replicate the ablation with bert-base-chinese to examine backbone sensitivity. As shown in Table 12, adding multi-

System: You are a scholar familiar with the Zuozhuan.

Prompt:

Ruling lord (state): **{gong}** 📖
 Reign year(s): **{time_range}** 📖
 Event subtitle in 《Chunqiu Zuozhuan Shuos》: **{event}** 📖
 Based on your knowledge of the Spring and Autumn Annals (《春秋》) and the Zuozhuan (《左傳》), infer which passage in the Zuozhuan this event subtitle is most likely to correspond to.

Instructions:

- As long as it describes the same historical event and is broadly similar in content, the wording does not need to be exactly identical.
- It can be one sentence or several sentences from either the Classic (jing) or the Commentary (zhuan), and you may extract only the key part.
- Please try to give the single passage from the Zuozhuan that you believe is most likely. Only when you have absolutely no clue and cannot judge at all should you output "NONE".

⚠️ **Format requirements (very important):**

- If you can make a judgment, ****output only the original Zuozhuan text itself****, without any explanation, analysis, or commentary.
- Do not output book titles, years, or chapter names, and do not add any prefixes such as "The original text is:".
- Do not add quotation marks, book-title marks, or brackets; only output the plain original text.
- If you cannot judge, output only: **"NONE"** (all uppercase, with no other content and no punctuation).

Please strictly follow the above format in your answer.

Input:

{gong}= 魯公
{time_range}= 元年
{event}= 祭仲諫鄭莊封叔段

Output:

初，鄭武公娶于申，曰武姜，生莊公及共叔段。莊公寤生，驚姜氏，故名曰「寤生」，遂惡之。愛共叔段，欲立之。亟請于武公，公弗許。及莊公即位，為之請制。公曰：「制，巖邑也，饒叔死焉，佗邑唯命。」請京，使居之，謂之京城大叔。祭仲曰：「都城過百雉，國之害也。先王之制，大都不過參國之一，中五之一，小九之一。今京不度，非制也，君將不堪。」公曰：「姜氏欲之，焉辟害？」對曰：「姜氏何厭之有？不如早為之所，無使滋蔓！蔓，難圖也。蔓草猶不可除，況君之寵弟乎？」公曰：「多行不義，必自斃，子姑待之。」

Figure 6: LLM-assisted reverse matching from paraphrastic event titles in Lü Zuqian’s *Chunqiu Zuozhuan Shuos* (《春秋左氏传说》) to *Zuo zhuan* passages. We show the full text-only prompt given to a classical-Chinese LLM (DeepSeek), together with one concrete example: for ruler *Yin* of Lu, year 1, and the subtitle “祭仲諫鄭莊封叔段” from *Chunqiu Zuozhuan Shuos*, the model proposes the most likely *Zuo zhuan* passage. The model is required to output only the original *Zuo* text, or the sentinel token NONE when it cannot decide.

positive retrieval supervision already improves over the FT baseline, while temporal modeling yields substantially larger gains. Among the two temporal modules, the soft absolute temporal context c_x contributes a stronger overall boost than the relative-time bias b_{ij}^{time} , and combining both gives the best performance. The larger gains on BERT than on stronger embedding backbones suggest that temporal supervision and structured negatives are especially helpful when the base retriever has more room to reduce chrono-near confusions.

Variant	\mathcal{L}_{multi}	b_{ij}^{time}	c_x	R@1	MRR@10
FT baseline	—	—	—	0.5088	0.5597
+ \mathcal{L}_{multi}	✓	—	—	0.5178 $\uparrow +0.0090$	0.5685 $\uparrow +0.0088$
+ Bias	✓	✓	—	0.5384 $\uparrow +0.0296$	0.5776 $\uparrow +0.0179$
+ Ctx	✓	—	✓	0.5620 $\uparrow +0.0532$	0.5961 $\uparrow +0.0364$
Full (Ours)	✓	✓	✓	0.5826 $\uparrow +0.0738$	0.6193 $\uparrow +0.0596$

Table 12: Ablation on the test set with bert-base-chinese.

B.2.2 No-event and Hard-negative Behavior

Table 13 probes two protocol switches: whether to keep pure no_event queries (dq) and whether to inject neg_comment passages as chrono-near hard negatives (neg). Including pure no-event queries consistently raises scores across all methods, indicating that empty-month retrieval is substantially easier and should be controlled by protocol. By contrast, injecting neg_comment passages exposes genuine robustness differences: the BERT FT baseline drops noticeably, whereas CTD $BERT_{base}$ remains stable, and both Qwen-based systems change only marginally. Overall, the benchmark contains both an easier no-event regime and a harder exegetical hard-negative regime, with CTD improving robustness especially for weaker backbones.

B.2.3 Full Protocol Grid Results

We report results under all valid combinations of the three evaluation switches: neg (whether neg_comment passages are included in the gallery), ne (whether explicit no_event records are included), and dq (whether pure no_event queries are dropped). Modes are denoted as $neg\{0/1\}_{ne\{0/1\}}_{dq\{0/1\}}$. When $ne=0$, pure no_event queries are ill-defined for retrieval, so only $dq=1$ is reported.

Tables 14 and 15 summarize validation and test results for all queries, and separately for the point and window families. Two trends are consistent across settings: keeping pure no_event queries ($dq=0$) raises aggregate scores, whereas injecting neg_comment passages ($neg=1$) yields a harder and more realistic gallery. The full grid is therefore intended mainly as a protocol reference and robustness diagnostic.

Full-grid results. Tables 14 and 15 report Recall@K, MRR@10, and nDCG@10 on validation and test under each mode. We report results on all queries, and also stratify by point and window families (corresponding to our point-/gap-like vs. window-style temporal probes in the main paper).

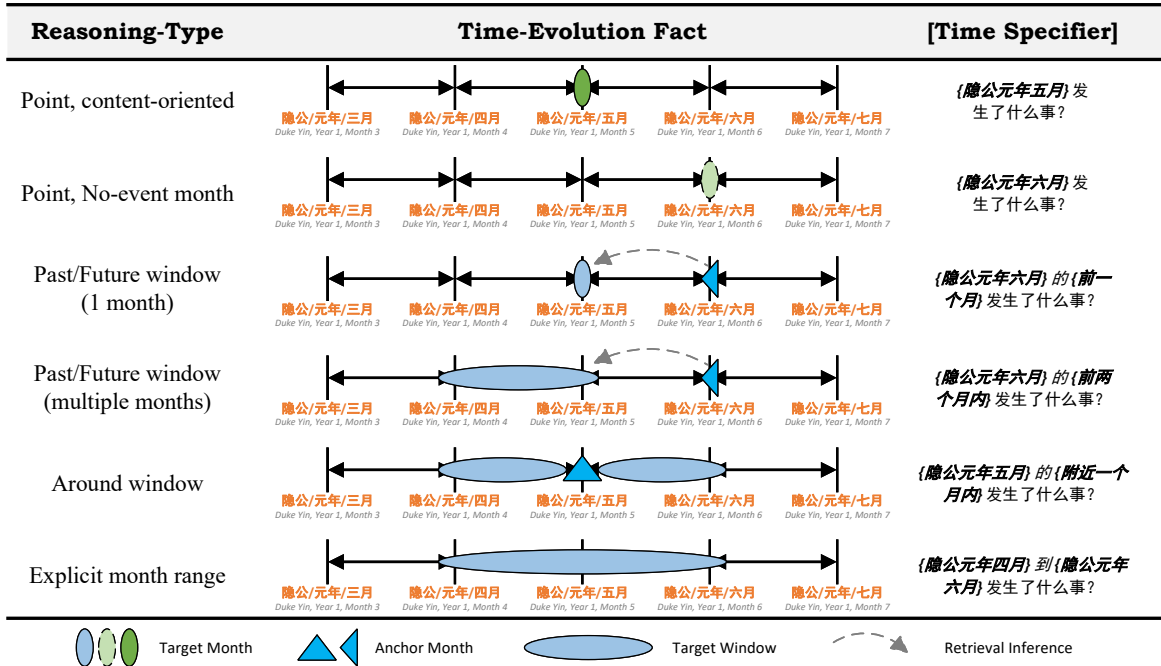


Figure 7: Representative temporal query types in ChunQiuTR, including point queries, past/future windows, around windows, and explicit ranges over reign-based month keys.

Method	Including pure no-event queries (fix neg=0, ne=1)						Hard-negative robustness (fix ne=1, dq=1)					
	dq=1		dq=0		Δ		neg=0		neg=1		Δ	
	R@1	MRR@10	R@1	MRR@10	R@1	MRR@10	R@1	MRR@10	R@1	MRR@10	R@1	MRR@10
BM25	0.240	0.296	0.418	0.470	$\uparrow+0.178$	$\uparrow+0.173$	0.240	0.296	0.228	0.282	$\downarrow-0.012$	$\downarrow-0.014$
ColBERT-LFM2 _(ZS)	0.220	0.277	0.333	0.386	$\uparrow+0.113$	$\uparrow+0.109$	0.220	0.277	0.222	0.278	$\uparrow+0.002$	$\uparrow+0.001$
BERT-base _(FT)	0.378	0.439	0.560	0.606	$\uparrow+0.182$	$\uparrow+0.167$	0.378	0.439	0.306	0.374	$\downarrow-0.072$	$\downarrow-0.065$
CTD BERT-base (Ours)	0.407	0.456	0.583	0.619	$\uparrow+0.175$	$\uparrow+0.163$	0.407	0.456	0.407	0.456	0.000	0.000
E5-mistral-7B _(ZS)	0.062	0.093	0.225	0.270	$\uparrow+0.163$	$\uparrow+0.177$	0.062	0.093	0.054	0.082	$\downarrow-0.008$	$\downarrow-0.011$
Qwen3-Embed-0.6B _(FT)	0.396	0.434	0.577	0.605	$\uparrow+0.181$	$\uparrow+0.171$	0.396	0.434	0.396	0.433	0.000	$\downarrow-0.001$
CTD Qwen3-Embed-0.6B (Ours)	0.420	0.457	0.594	0.621	$\uparrow+0.174$	$\uparrow+0.164$	0.420	0.457	0.418	0.455	$\downarrow-0.002$	$\downarrow-0.002$

Table 13: No-event and hard-negative behavior under protocol variations on the test set. Left: dq=1 vs. dq=0 (fix ne=1, neg=0), where dq drops or keeps pure no_event queries. Right: neg=0 vs. neg=1 (fix ne=1, dq=1), where neg injects neg_comment passages into the gallery. Δ denotes the within-method change.

What this grid clarifies. Two takeaways are particularly relevant for interpreting aggregate scores. First, dq=0 (keeping pure no-event queries) can noticeably inflate overall metrics compared to dq=1, motivating our practice of reporting both settings: dq=0 reflects the benchmark’s intended scope (event months *and* explicit empty months), while dq=1 isolates event-seeking behavior. Second, injecting exegetical hard negatives (neg=1) is a strictly harder and more realistic gallery setting; models that remain stable between neg=0 and neg=1 exhibit stronger robustness to chrono-near confounds from commentarial material.

B.2.4 Top-1 Near-miss Failure Cases

Figure 9 shows two representative near-miss cases on the test set: our retriever fails to place the correct passage at rank 1, but still retrieves at least one ground-truth passage within the top-5, whereas the baseline fails to surface any ground-truth evidence. In both cases, a key confounder is the highly reusable no_event-style wording and its chrono-near recurrence across adjacent months, which can trigger top-rank swaps. These examples suggest that the remaining errors are often ordering errors under strong lexical or temporal confounders, rather than complete retrieval failure.

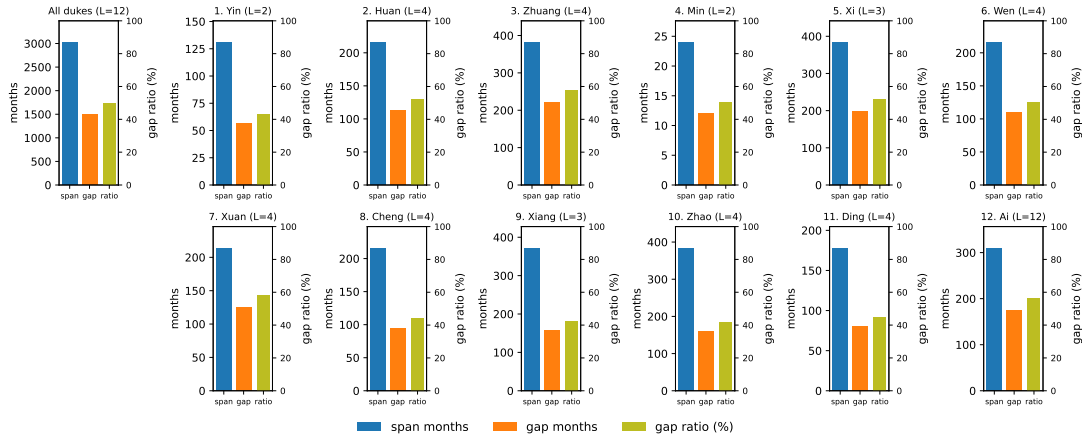


Figure 8: Month-level coverage, gap counts, and gap ratios for the normalized *Chunqiu* timeline (overall and per Lu ruler).

QUERY(point): 魯成公六年二月，魯國有何記事？ (What happened in Duke Cheng's 6th year, 2nd month?)		
Baseline	Ours	Ground Truth
#1 魯成公七年二月：經傳於是月史事闕如，無專條可錄。	#1 魯成公七年二月：經傳於是月史事闕如，無專條可錄。	★ 二月辛巳，立武宮。武宮者何？武公之宮也。立者何？...
#2 魯文公六年二月。《春秋》經文及三傳於此月無事可書。	#2 魯成公八年二月：經傳於是月史事闕如，無專條可錄。	★ 二月辛巳，立武宮。立者，不宜立也。
#3 魯成公八年二月：經傳於是月史事闕如，無專條可錄。	#3 ★ 二月，辛巳，立武宮，取郟，衛孫良夫帥師侵宋。	★ 二月，季文子以穀之功立武宮，非禮也，闕於人以教其難...
#4 魯成公五年二月：經傳於是月史事闕如，無專條可錄。	#4 ★ 二月辛巳，立武宮。立者，不宜立也。	★ 二月，辛巳，立武宮，取郟，衛孫良夫帥師侵宋。
#5 魯成公四年二月：經傳於是月史事闕如，無專條可錄。	#5 ★ 二月，辛巳，立武宮，取郟，衛孫良夫帥師侵宋。	★ 取郟。郟者何？郟之邑也。曷為不系于郟？薛盭也。

QUERY(window): 請檢索魯成公七年正月之前半年內，《春秋》中所有相關記事。(What happened in half year before Duke Cheng's 6th year, 1st month?)		
Baseline	Ours	Ground Truth
#1 魯成公七年九月：經傳於是月史事闕如，無專條可錄。	#1 魯成公七年十二月。《春秋》及三傳皆不載此月之事。	★ 秋，仲孫蕩、叔孫僕如帥師侵宋。
#2 魯成公七年十二月。《春秋》及三傳皆不載此月之事。	#2 魯成公七年十一月。《春秋》經文及三傳於此月無事可書。	★ 秋，仲孫蕩、叔孫僕如帥師侵宋。
#3 魯成公七年六月。《春秋》及三傳皆不載此月之事。	#3 魯成公七年九月：經傳於是月史事闕如，無專條可錄。	★ 秋，孟獻子，叔孫宣伯，侵宋，晉命也，楚子重伐鄭...
#4 魯成公七年十一月。《春秋》經文及三傳於此月無事可書。	#4 魯成公七年二月：經傳於是月史事闕如，無專條可錄。	★ 楚公子嬰齊率師伐鄭。
#5 魯成公七年三月。《春秋》經文及三傳於此月無事可書。	#5 ★ 魯成公六年十二月。《春秋》及三傳皆不載此月之事。	★ 秋，仲孫蕩，叔孫僕如，帥師侵宋，楚公子嬰齊帥師伐鄭。

Figure 9: Top-1 near-miss cases on ChunQiuTR (test set). We show the top-5 results from a baseline (left) and ours (middle), with the ground-truth set (right). ★ marks ground-truth passages; ✓ indicates a ground-truth hit in top-5.

B.2.5 Qualitative demo: reasoning traces vs. evidence grounding

Figures 10 and 11 compare an online LLM on the same month-level point query with and without evidence grounding. Without retrieved evidence, the model either predicts an empty month or produces an incomplete answer even when a reasoning trace is enabled. When given an evidence pack that contains the gold month records together with confusable materials, the same model recovers both gold entries and grounds the answer in cited evidence. These examples suggest that longer reasoning traces alone do not ensure month-level completeness, whereas explicit evidence binding substantially improves temporal faithfulness.

B.3 Details of Compared Methods

B.3.1 Sparse retrieval.

We compare against a sparse family including a classical lexical retriever, a simple temporal re-ranking variant, and two inference-free neural sparse retrievers, all evaluated under the same sparse retrieval protocol.

- **BM25 (Robertson and Zaragoza, 2009)**: standard lexical term-matching baseline.
- **BM25+TimeKDE**: BM25 with a non-parametric temporal re-ranking prior over regnal-month indices, following classical TIR-style temporal priors (Li and Croft, 2003).
- **SPLADE-IDF_(ZS)** (Geng et al., 2025): inference-free neural sparse retriever used zero-shot.
- **SPLADE- ℓ_0 (ZS)** (Shen et al., 2025): sparsity-controlled neural sparse retriever used zero-shot.

B.3.2 Fusion / late-interaction retrieval.

We further compare against two multi-vector late-interaction retrievers, both used in a zero-shot setting.

- **ColBERT-JINA_(ZS)** (Jha et al., 2024): ColBERT-style token-interaction retriever.
- **ColBERT-LFM2_(ZS)** (Team, 2025): late-interaction retriever with longer-context and

Mode	Family	Validation					Test				
		R@1	R@5	R@10	MRR@10	nDCG@10	R@1	R@5	R@10	MRR@10	nDCG@10
neg0_ne0_dq1	all	0.0407	0.1200	0.2232	0.0798	0.0635	0.0654	0.1466	0.2042	0.1014	0.0677
neg0_ne0_dq1	point	0.0375	0.1148	0.2459	0.0775	0.0756	0.0710	0.1696	0.2387	0.1156	0.0934
neg0_ne0_dq1	window	0.0430	0.1239	0.2065	0.0815	0.0546	0.0610	0.1283	0.1768	0.0902	0.0472
neg0_ne1_dq0	all	0.6329	0.6697	0.7191	0.6522	0.4921	0.5935	0.6497	0.6945	0.6206	0.4588
neg0_ne1_dq0	point	0.5200	0.5440	0.6057	0.5347	0.5380	0.4922	0.5307	0.5869	0.5138	0.5103
neg0_ne1_dq0	window	0.7860	0.8403	0.8729	0.8115	0.4300	0.7341	0.8150	0.8439	0.7689	0.3871
neg0_ne1_dq1	all	0.4534	0.5050	0.5784	0.4806	0.2385	0.4197	0.4956	0.5602	0.4565	0.2223
neg0_ne1_dq1	point	0.0164	0.0656	0.1920	0.0466	0.0532	0.0375	0.1105	0.2170	0.0784	0.0719
neg0_ne1_dq1	window	0.7745	0.8279	0.8623	0.7996	0.3747	0.7230	0.8013	0.8326	0.7565	0.3417
neg1_ne0_dq1	all	0.0317	0.1101	0.2173	0.0709	0.0600	0.0593	0.1344	0.1998	0.0951	0.0647
neg1_ne0_dq1	point	0.0328	0.1077	0.2436	0.0725	0.0729	0.0690	0.1637	0.2367	0.1127	0.0912
neg1_ne0_dq1	window	0.0310	0.1119	0.1979	0.0697	0.0505	0.0516	0.1111	0.1706	0.0812	0.0437
neg1_ne1_dq0	all	0.6329	0.6691	0.7184	0.6520	0.4914	0.5923	0.6485	0.6927	0.6194	0.4575
neg1_ne1_dq0	point	0.5200	0.5429	0.6046	0.5344	0.5373	0.4912	0.5307	0.5858	0.5129	0.5094
neg1_ne1_dq0	window	0.7860	0.8403	0.8729	0.8115	0.4291	0.7327	0.8121	0.8410	0.7674	0.3854
neg1_ne1_dq1	all	0.4534	0.5040	0.5774	0.4803	0.2374	0.4180	0.4939	0.5576	0.4548	0.2205
neg1_ne1_dq1	point	0.0164	0.0632	0.1897	0.0459	0.0519	0.0355	0.1105	0.2150	0.0767	0.0701
neg1_ne1_dq1	window	0.7745	0.8279	0.8623	0.7996	0.3737	0.7214	0.7981	0.8294	0.7549	0.3399

Table 14: Full protocol grid results for CTD-QWEN3-EMBED-0.6B. Mode names follow neg/ne/dq as defined in Section B.2.3.

multi-scale representations.

B.3.3 Dense retrieval, encoder-based.

For encoder-based dense retrieval, we compare against single-vector dual-encoder models used either zero-shot or fine-tuned on ChunQiuTR.

- **GTR-T5-Base / Sentence-T5-Base**_(ZS) (Ni et al., 2022b,a): T5-based dense retrievers used zero-shot.
- **mE5-Large / mE5-Large-ins**_(ZS) (Wang et al., 2024b): multilingual E5 retrievers used zero-shot.
- **GTE-Large**_(ZS) (Li et al., 2023): general-purpose dense embedding baseline.
- **BGE-Large-v1.5 / BGE-M3**_(ZS) (Xiao et al., 2023; Chen et al., 2024): strong multilingual dense embedding baselines.
- **BERT-base**_(FT) (Devlin et al., 2019): Chinese BERT dual-encoder fine-tuned on ChunQiuTR without explicit time modeling.

B.3.4 Dense retrieval, LM-based embeddings.

We also compare against LM-based dense embedding models, including both zero-shot and task-adapted variants.

- **GTE-Qwen2-1.5B / E5-Mistral-7B**_(ZS) (Wang et al., 2024a): LLM-scale embedding baselines used zero-shot.
- **PQR (Qwen2.5-7B / Qwen3-8B)**_(re) (Kang et al., 2025): training-free retrieval framework based on LLM-generated pseudo-queries.
- **Qwen3-Embed-0.6B / 4B**_(ZS) (Zhang et al., 2025b): dedicated Qwen3 embedding models used zero-shot.
- **Qwen3-Embed-0.6B**_(FT) (Zhang et al., 2025b): task-adapted dense dual-encoder baseline without explicit time modeling.

B.3.5 Time-aware auxiliary variants.

Beyond BM25+TimeKDE, we report two lightweight temporal extensions for single-vector dense retrievers.

- **TempDate**: auxiliary time-key prediction over (gong, year, month) during training, discarded at inference time (Wang et al., 2023; Dhingra et al., 2022).
- **TempDate-Smooth**: TempDate with neighbor-aware smoothing over adjacent ordered time keys (Yèche et al., 2023).

Mode	Family	Validation					Test				
		R@1	R@5	R@10	MRR@10	nDCG@10	R@1	R@5	R@10	MRR@10	nDCG@10
neg0_ne0_dq1	all	0.0159	0.0456	0.0952	0.0331	0.0205	0.0096	0.0366	0.0672	0.0234	0.0192
neg0_ne0_dq1	point	0.0258	0.0468	0.0937	0.0382	0.0265	0.0079	0.0394	0.0690	0.0226	0.0235
neg0_ne0_dq1	window	0.0086	0.0448	0.0964	0.0294	0.0161	0.0110	0.0344	0.0657	0.0240	0.0158
neg0_ne1_dq0	all	0.4474	0.5678	0.5980	0.4989	0.3794	0.4180	0.5408	0.5735	0.4696	0.3543
neg0_ne1_dq0	point	0.4434	0.4709	0.4789	0.4556	0.4612	0.4152	0.4475	0.4506	0.4284	0.4339
neg0_ne1_dq0	window	0.4527	0.6992	0.7597	0.5576	0.2683	0.4220	0.6705	0.7442	0.5269	0.2436
neg0_ne1_dq1	all	0.2629	0.4097	0.4464	0.3261	0.1398	0.2400	0.3709	0.4127	0.2963	0.1218
neg0_ne1_dq1	point	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
neg0_ne1_dq1	window	0.4561	0.7108	0.7745	0.5658	0.2426	0.4304	0.6651	0.7402	0.5314	0.2184
neg1_ne0_dq1	all	0.0000	0.0079	0.0169	0.0043	0.0043	0.0000	0.0052	0.0122	0.0023	0.0023
neg1_ne0_dq1	point	0.0000	0.0117	0.0258	0.0073	0.0076	0.0000	0.0059	0.0178	0.0033	0.0037
neg1_ne0_dq1	window	0.0000	0.0052	0.0103	0.0021	0.0019	0.0000	0.0047	0.0078	0.0015	0.0012
neg1_ne1_dq0	all	0.4303	0.5480	0.5908	0.4809	0.3654	0.3962	0.5209	0.5620	0.4487	0.3404
neg1_ne1_dq0	point	0.4286	0.4571	0.4754	0.4415	0.4494	0.3965	0.4350	0.4495	0.4125	0.4214
neg1_ne1_dq0	window	0.4326	0.6713	0.7473	0.5343	0.2515	0.3960	0.6402	0.7182	0.4989	0.2280
neg1_ne1_dq1	all	0.2530	0.3948	0.4395	0.3137	0.1314	0.2277	0.3560	0.3988	0.2823	0.1151
neg1_ne1_dq1	point	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
neg1_ne1_dq1	window	0.4389	0.6850	0.7625	0.5442	0.2279	0.4085	0.6385	0.7152	0.5063	0.2064

Table 15: Full protocol grid results for BM25, reported in the same format as Table 14.

B.4 Cross-Corpus Pilot on *Zizhi Tongjian*

To probe whether the temporal-consistency bias learned on ChunQiuTR transfers beyond the *Spring and Autumn Annals*, we conduct a lightweight cross-corpus evaluation on two processed subsets from *Zizhi Tongjian* (*Qi Ji* and *Jin Ji*). As an annalistic general history, *Zizhi Tongjian* also records events under traditional reign-based, non-Gregorian temporal expressions, making it a natural out-of-domain probe for month-keyed retrieval.

This pilot preserves the core month-key retrieval idea of ChunQiuTR but is intentionally lighter than the full benchmark. We retain event-bearing lines as retrieval units, group them by normalized month keys derived from the available reign/year/month fields, and instantiate one point-style query for each unique month key using a traditional reign-year template. No target-corpus training is performed. Unlike the full ChunQiuTR benchmark, this pilot does not reconstruct explicit no_event placeholders, commentary-derived hard negatives, or the full point/gap/window query families, and should therefore be interpreted as a transfer probe rather than a second benchmark.

Subset statistics. Table 16 summarizes the two processed subsets used in this pilot. *Qi Ji* contains 268 records and 92 month-level queries, while *Jin Ji* contains 820 records and 119 queries. The

two slices cover distinct reign periods and remain clearly separate from the ChunQiuTR source corpus.

Transfer results. Table 17 reports retrieval performance on the two subsets. We compare a zero-shot Qwen3-Embed-0.6B encoder, a ChunQiuTR fine-tuned dense baseline, and our CTD-enhanced retriever. Across both subsets, CTD consistently improves MRR and R@1 over the fine-tuned baseline without any target-corpus retraining; on *Qi Ji* it also improves R@5 and R@10, while on *Jin Ji* it matches the fine-tuned baseline on higher-recall metrics.

Discussion. Although this transfer setting is lighter than the full ChunQiuTR benchmark, the overall trend is consistent with our main findings: the temporal-consistency bias learned on ChunQiuTR transfers beyond in-domain fitting and continues to help distinguish chrono-near but temporally mismatched evidence. At the same time, the gains are smaller than those observed on the source benchmark, which is expected given both the domain shift and the simplified evaluation protocol. We therefore view this pilot as evidence of promising cross-corpus transfer, rather than as a replacement for a fully reconstructed *Zizhi Tongjian*-specific benchmark.

Subset	Approx. coverage	Representative reign titles	Records	Queries
Qi Ji (part)	479–489 CE	建元, 永明	268	92
Jin Ji (part)	265–279 CE	泰始, 咸宁	820	119

Table 16: Basic statistics of the processed *Zizhi Tongjian* subsets used in the cross-corpus pilot.

Model / Setting	MRR	Qi Ji (part)			MRR	Jin Ji (part)		
		R@1	R@5	R@10		R@1	R@5	R@10
Qwen3-Embed-0.6B (ZS)	0.0692	0.0217	0.0870	0.1413	0.0691	0.0420	0.0756	0.1345
Qwen3-Embed-0.6B (FT baseline)	0.2081	0.1848	0.2174	0.2391	0.1598	0.1345	0.1849	0.1849
CTD (Ours)	0.2304	0.2065	0.2391	0.2717	0.1751	0.1597	0.1849	0.1849

Table 17: Cross-corpus pilot results on two processed *Zizhi Tongjian* subsets. No target-corpus training is performed.

System: 你是一名严谨的《春秋》史料问答助手。请只根据问题本身作答；如果无法确定，请明确说“不确定/需要证据”，不要编造。

Prompt:

【问题】
查询《春秋》，鲁隐公元年十二月这一月有什么重要事件？

【输出要求】

- 用 2-5 句输出答案（中文）。
- 如果你不确定，请直接说明不确定点是什么。

Output (w/o Thinking): False "no-event" claim
Ground Truth coverage: 0/2

根据《春秋》经文，鲁隐公元年（公元前722年）十二月并元事件记录。《春秋》该年记载始于“元年春王正月”，终于“冬十有二月”。经文仅记“冬十有二月”而未录具体事件，故该月无明确重要事件记录。

Thinking:

首先，问题要求查询《春秋》中鲁隐公元年十二月这一月有什么重要事件。输出要求用中文回答，且要求简洁明了。其次，问题要求查询《春秋》中鲁隐公元年十二月这一月有什么重要事件。输出要求用中文回答，且要求简洁明了。最后，问题要求查询《春秋》中鲁隐公元年十二月这一月有什么重要事件。输出要求用中文回答，且要求简洁明了。

Miss Key Records:

根据《春秋》经文，鲁隐公元年十二月的重要事件是“祭伯来”和“公子益师卒”。

Confident hallucination:

根据《春秋》经文，鲁隐公元年十二月的重要事件是“祭伯来”和“公子益师卒”。

Output (Thinking):

根据《春秋》记载，鲁隐公元年十二月的重要事件是“祭伯来”，即周王室的大夫祭伯来聘鲁国。这一记录见于《春秋》隐公元年冬季的经文，反映了当时周王室与诸侯国之间的外交往来。此外，该月无其他事件记载。

Ground Truth coverage: 1/2 (miss: 公子益师卒)

Ground Truth: (1) 祭伯来; (2) 公子益师卒。

Figure 10: Online LLM without evidence grounding on a month-level point query from ChunQiuTR. For the query “鲁隐公元年十二月”，the gold month contains two entries (祭伯来 and 公子益师卒). Without evidence, the model either predicts an empty month or returns an incomplete answer.

B.5 Alignment Audits and Reliability

To improve the auditability of ChunQiuTR, we summarize here the two LLM-assisted curation stages and the corresponding human-verification statistics. As clarified in the revised main text, ChunQiuTR is *not* an AI-generated dataset: the retrieval gallery is derived from authentic historical sources, the queries are instantiated from a small set of manually written templates, and LLMs are used only to propose candidate splits or candidate alignments during curation. They are never used to generate, rewrite, translate, or paraphrase historical content, and only human-approved results enter the

final benchmark.

(1) **Time-key normalization and manual verification.** The corpus follows an implicit Lu-state reign calendar, normalized as month-level keys

$$\tau = (\text{gong, year, month}).$$

However, many passages do not explicitly contain a complete (gong, year, month) triple. Instead, the ruling duke and/or regnal year must often be recovered from annalistic structure, discourse continuity, and neighboring entries rather than extracted as standalone temporal mentions. For this reason, the mapping from original records to normalized time keys is manually verified during dataset construction, rather than delegated to a fully automatic temporal-expression extractor. Representative examples of reign-key propagation and normalization are provided in Appendix A.2.1.

(2) **Audit of multi-event splitting.** A first LLM-assisted step is used when a single month-level segment contains more than one historical event. In these cases, the model is asked only to propose candidate event-level groupings under a fixed time key; all such proposals are then manually reviewed and corrected if necessary.

Table 18 reports the corresponding audit statistics. Among 1,533 non-empty months, 558 contain multiple events (36.41%). After LLM candidate grouping, only 63 multi-event months required additional human correction, corresponding to 11.29% of multi-event months and 4.11% of all non-empty months. The remaining 495 multi-event months were accepted without change (88.71% direct acceptance among multi-event cases). These statistics suggest that LLM proposal is useful for reducing manual effort, while final segmentation

System: 你是一名严谨的《春秋》史料问答助手。

Prompt:
你必须只依据下方给定资料回答，禁止使用外部常识补充。
如果资料不足以支持结论，请回答“资料不足”，并说明缺少什么。
【资料（可能包含同月注疏、相邻月近似事件、空月占位等干扰项）】
你将看到若干条DOC记录，请先判断哪些记录最能直接回答问题。

“鲁隐公元年十一月：既而於是月也事始，無事特可錄。”
“春秋之例，即乃見經。今名置於經，傳言司空，故知無是魯隱。…”
“凡即出，必具其氏，以釋其位，今不置氏…”
“蔡伯，諸侯王與士者，皆稱，位釋也。傳曰：‘昭王也。’…”

“冬，十有二月，蔡伯來，蔡伯者何？天子之大夫也，何以不稱侯？弄也。…”
“冬，十有二月，蔡伯來，來者，來者何，其謂諸侯，何也？襄內諸侯…”
“十二月，蔡伯來，非王命也。”

“蔡伯來，諸侯以爲朝，公羊謂以爲葬，蔡伯來，諸侯以爲葬…”
“蔡伯來，諸侯以爲葬，公羊謂以爲葬，蔡伯來，諸侯以爲葬…”
“蔡伯來，諸侯以爲葬，公羊謂以爲葬，蔡伯來，諸侯以爲葬…”
“蔡伯來，諸侯以爲葬，公羊謂以爲葬，蔡伯來，諸侯以爲葬…”

“二年春，公會戎于潛，會者，外為主謂，知者，義者行，仁者守。…”

【问题】
查詢《春秋》，魯隱公元年十二月這一月有什麼重要事件？
【你必须按以下步骤输出】

- 1) 证据选择：从 DOC 中选出“最关键”的 1-3 条。
- 2) 最终回答：只基于你选的证据作答（中文，2-6 句）。
- 3) 证据引用：在答案末尾列出你用到的 DOC。
- 4) 若问题对应月份为空月：请明确写“该月无事可书”（以资料措辞为准）。

Output (w/o Thinking): **Ground Truth coverage 2/2 (correct)**

- 1) 证据选择：
（鲁隐公元年十二月）：这是问题直接对应的月份记录，包含两条经文及三传相关记载，是回答问题的核心依据。
- 2) 最终回答：
根据《春秋》记载，鲁隐公元年十二月发生了两件重要事件。第一，蔡伯来到鲁国，但《左传》指出这是“非王命也”，即并非奉周天子之命。第二，鲁国公子益师去世，但《春秋》未记载其去世的具体日期，因此，该月的重要事件是蔡伯来和公子益师卒。
- 3) 证据引用：
Grounded: cites ground truth DOCs
Evidence: “冬，十有二月，蔡伯來，蔡伯者何？天子之大夫也，何以不稱侯？弄也。…”
“冬，十有二月，蔡伯來，來者，來者何，其謂諸侯，何也？襄內諸侯…”
“十二月，蔡伯來，非王命也。”

Thinking:
问题：在春秋《春秋》史料问答助手，只依据给定的资料回答问题，问题是鲁隐公元年十二月有什么重要事件，我首先会梳理资料中这个月所有有记录。资料中鲁隐公元年十二月的记录，包含经文原文“冬，十有二月，蔡伯來，來者，來者何，其謂諸侯，何也？襄內諸侯…”和“十二月，蔡伯來，非王命也。”，以及三传的注释内容。第二，鲁隐公元年十二月发生了两件重要事件。第一，蔡伯来到鲁国，但《左传》指出这是“非王命也”，即并非奉周天子之命。第二，鲁国公子益师去世，但《春秋》未记载其去世的具体日期，因此，该月的重要事件是蔡伯来和公子益师卒。

Output (Thinking): **Ground Truth coverage 2/2 (correct)**

- 1) **证据选择**:
- 选择鲁隐公元年十二月记录中的两个部分（两个“versions”），因为该月资料明确包含两条独立经文记载：“冬，十有二月，蔡伯來。”和“公子益師卒。”，并附有三传的注疏与讨论，它们直接对应问题所询问的该月重要事件。
- 2) **最终回答**:
根据《春秋》记载，鲁隐公元年十二月发生了两件重要事件。第一是“冬，十有二月，蔡伯來”，即蔡伯来到鲁国。《左传》指出此事“非王命也”，即并非奉周天子之命。第二是“公子益師卒”，即鲁国公子益师去世。《左传》解释其未记载具体日期是因为“公不与小政”，这两条是本月明确记载于经文的的事件。
- 3) **证据引用**:
Grounded: cites ground truth DOCs
Evidence: 鲁隐公元年十二月 (version 1; 经文“冬，十有二月，蔡伯來”及相关注疏); 鲁隐公元年十二月 (version 2; 经文“公子益師卒”及相关注疏)。

Ground Truth: (1) 蔡伯來; (2) 公子益師卒

Figure 11: Evidence-bounded RAG for the same query as Fig. 10. With a small evidence pack containing the gold month records and confusable materials, the model recovers both gold entries and grounds the answer in cited evidence.

quality remains controlled by explicit human review.

Common correction patterns. Manual corrections in this stage mainly fall into a small number of recurrent categories: (i) *boundary shift*, where the candidate split cuts too early or too late and therefore mixes material from adjacent events; (ii) *inappropriate merge*, where two historically distinct events are grouped together because they share a compact annalistic sentence; and (iii) *inappropriate split*, where commentary fragments that should remain attached to one event are separated into different candidate groups. In all such cases, the final retained record structure is determined by human verification.

Item	Value
Total non-empty months	1,533
Months containing multiple events	558
Fraction multi-event	36.41%
Extra human corrections	63
Correction rate among multi-event months	11.29%
Corrections among all non-empty months	4.11%
Accepted without change	495
Direct acceptance rate	88.71%

Table 18: Audit statistics for multi-event splitting (LLM proposals + human verification).

Source	#Candidates	Accepted	Rejected	Acceptance
Gu Donggao	899	899	0	100.00%
Guo Yingda	5,286	5,179	107	97.98%
Du Yu	5,373	5,266	107	98.01%
Lü Zuqian	360	336	24	93.33%

Table 19: Acceptance rates for later-commentary alignments (LLM candidate proposal + human verification).

(3.) Audit of later-commentary alignment. A second LLM-assisted step is used when aligning later historiographical or commentarial materials to original Chunqiu records. These later sources often refer to canonical events through highly compressed paraphrases, lexical reformulations, or short subtitles rather than direct quotation. We therefore use an LLM only to propose candidate matched passages, after which human verification determines whether the candidate alignment is accepted into the benchmark as a chrono-near confusable negative.

Table 19 reports the human acceptance statistics for four representative source groups. Acceptance rates range from 93.33% to 100.00%, indicating that candidate proposal is generally accurate, but still benefits from manual checking to remove residual mismatches.

Typical rejection patterns. Rejected candidate alignments mainly arise from three sources. First, some later commentaries refer to the correct historical period but to an overly broad textual span, making the proposed match imprecise. Second, some candidates are semantically similar to the target event but mismatch key participants, event roles, or action focus. Third, some compressed headings or summaries are ambiguous enough that multiple canonical passages appear plausible, in which case we conservatively reject the alignment unless a human annotator can verify a unique and appropriate match.

Takeaway. Across both curation stages, the role of the LLM is restricted to efficient candidate proposal. Dataset quality is controlled by manual verification and supported by explicit audit statistics. We therefore view the resulting benchmark as a historically grounded, human-verified dataset rather than an AI-generated or synthetic resource.