ChunQiuTR: Time-Keyed Temporal Retrieval in Classical Chinese Annals

Anonymous ACL submission

Abstract

Retrieval shapes how language models access and cite knowledge in retrieval-augmented generation (RAG). In historical research, the goal is often to locate the exact record for a specific regnal month, where temporal alignment matters as much as topical relevance. This is especially challenging for Classical Chinese annals: time is encoded in terse, implicit, non-Gregorian reign phrases that are context-dependent, 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. It organizes records by month-level reign keys and includes chrono-near confounders that mimic real retrieval failures. We propose **CTD** (Calendrical Temporal Dual-encoder), a time-aware dual-encoder combining Fourier-based absolute context with relative offset biasing. Experiments show consistent gains over semantic dual-encoder baselines under time-keyed evaluation. We will release ChunQiuTR and code after the anonymity period.

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 can 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).

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. Given such confounders, a retrieval-augmented model can produce a fluent yet temporally incorrect answer. This motivates the following question:

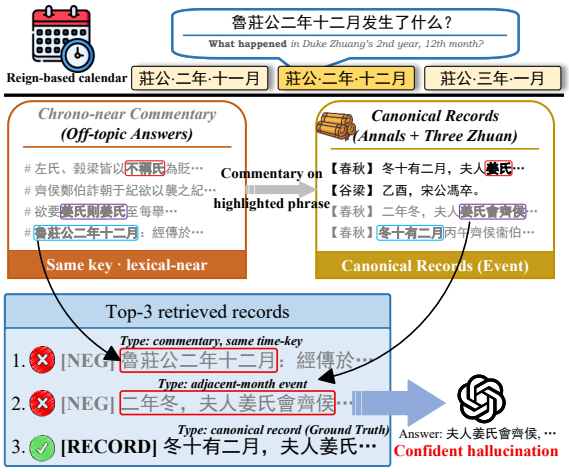

Figure 1: A query about a specific month can retrieve same-month commentary that repeats the date phrase, or adjacent-month near-miss events with confusable wording, so a retrieval-augmented LLM answers fluently but at the wrong time.

> **(Q)** *Can modern large language models—and the retrievers that feed them—faithfully track temporal information in historical records that rely on* ***non-Gregorian*** *chronologies?*

Directly probing this question is challenging be-

cause 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 becomes 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** (*Temporal Retrieval*), 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 represents each query and record with a soft location on the stitched into a single ordered 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.

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.

3 Dataset Construction

To study temporal retrieval under non-Gregorian dating systems, we construct **ChunQiuTR**, a benchmark based on the *Spring and Autumn Annals* and its classical commentaries. The benchmark evaluates whether models can identify both *events* and their *temporal positions* when time is

expressed using reign-year references rather than explicit Gregorian dates. Fig. 2 outlines the data construction process. We next explain why the *Chunqiu* is a natural testbed and how the benchmark is instantiated.

3.1 Chunqiu Corpus and Temporal Scheme

Why the *Chunqiu*? The *Chunqiu* (《春秋》, *Spring and Autumn Annals*) is a terse chronicle of the state of Lu (722–481 BCE) with a long exegetical and historiographical tradition built around it. Its entries are extremely compact and date events only using a reign-based temporal language. Brief formulae such as "元年春" (first year, spring) or "夏五月" (summer, fifth month) omit explicit absolute years and often leave the ruler implicit. Because many landmark events occur only once, mentioning the event together with such a relative month phrase often suffices to pin down a unique point on the timeline. These compressed records are then expanded and re-interpreted by the three classical *zhuan* (*Zuo*, *Gongyang*, *Guliang*) and later commentarial and historiographical works.

Taken together, this layered structure makes the *Chunqiu* corpus an ideal testbed for temporal retrieval: all layers align to the same ruler-centric timeline without explicit Gregorian dates, and different layers often describe the same event or nearby months with overlapping phrasing, naturally producing realistic "near-miss" hard negatives. Similar combinations of a shared chronological backbone and dense, overlapping commentary recur in many pre-modern annalistic corpora (e.g., dynastic histories such as the *Hanshu*), so we use the *Chunqiu* as a compact starting point for methods that can later be generalized to broader non-Gregorian historical collections. Additional details on sources and preprocessing are provided in Appendix A.1.

Reign-based time keys. Unlike modern texts that use absolute year numbering (e.g., "709 BCE"), our sources use a ruler-centric, reign-based dating scheme in which regnal years restart for each new Lu duke. A regnal year may begin with a full formula such as "元年春王正月" (first year, spring, royal first month), but later entries are often shortened to month phrases like "夏五月" (summer, fifth month), with the duke and year supplied by context—effectively "in the fifth month of summer of this duke's current regnal year." Because many key events in the *Chunqiu* are unique, the event description plus such a relative month phrase often pins down a single point on the reign-based timeline. This compact temporal language contrasts with standard time-IR corpora, where documents carry explicit Gregorian dates that can be treated as fixed metadata rather than temporal expressions to be interpreted.

For our benchmark, we normalize this temporal language into month-level time keys $\tau = (\text{gong}, \text{year}, \text{month})$ to index all records, where gong denotes the ruling duke title (Chinese character "公", glossed as "Duke" for readability). Month-level is the finest temporal unit consistently recorded in the *Chunqiu*; finer-grained dates (e.g., day-level) are largely absent or sporadic and thus cannot be normalized systematically. We assign a time key to every month, including months with no annals entry; we later instantiate them as synthetic `no_event` records. We then align each annals entry and its associated historiographical passages to a unique τ, forming a time-keyed gallery of short records (Section 3.2, Appendix A.2.1). Although in the *Chunqiu* this appears as a concrete gong–year–month scheme, the pipeline itself only assumes an ordered set of time keys with aligned texts and thus applies to other non-Gregorian chronologies.

3.2 Record Alignment

Record–time-key alignment. Building on the reign-based month keys $\tau = (\text{gong}, \text{year}, \text{month})$ defined above, we define the record units used for retrieval as shown in Fig. 2 (1). We treat each *record* as the atomic retrieval unit: a short event-level passage aligned to a single month key τ. For each time key τ, we gather all snippets from the annals and the three classical *zhuan* and refine them into event-level record sets $\mathcal{D}_\tau$. This is non-trivial: the annals often compress several events into a single sentence, while the commentaries may spread one event across multiple fragments, so naive sentence or paragraph boundaries do not match historical events. We therefore use a lightweight LLM prompt to propose splits and groupings under each τ, then manually review and correct them so that each final record corresponds to one coherent event with its aligned commentarial material. This yields a high-precision, time-keyed collection of event-level records for the entire *Chunqiu* period (Appendix A.2.2).

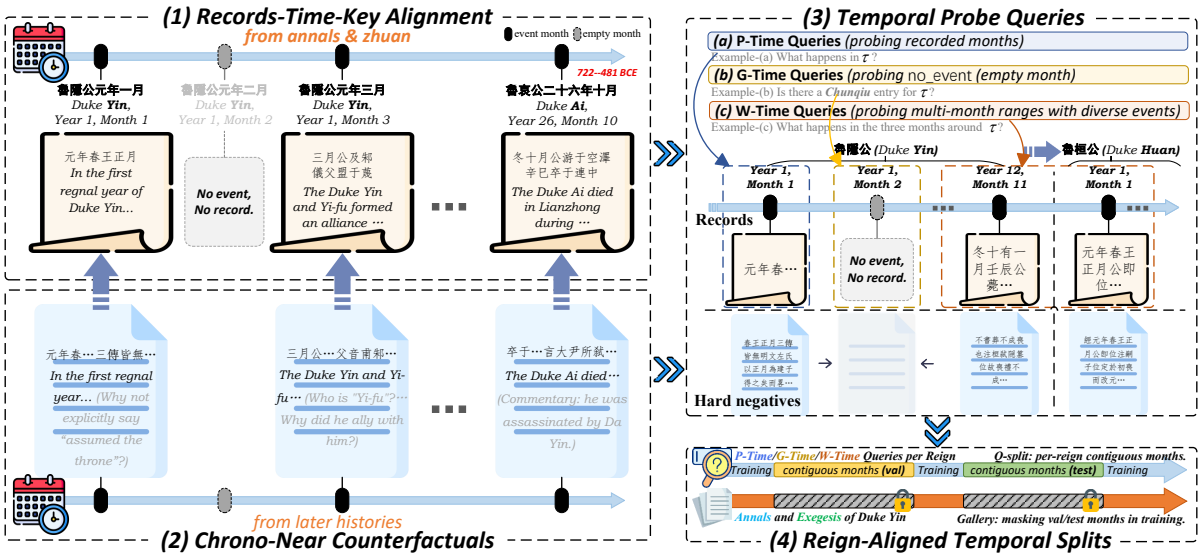

Figure 2: Overview of the ChunQiuTR construction. (**Left**) Time-key alignment yields event-level records, augmented with chrono-near counterfactual hard negatives from later histories. (**Right**) We define P/G/W temporal queries and leak-free, reign-aligned splits.

Chrono-near counterfactual negatives. Later historiographical layers, such as Gu Donggao's *Chronological Tables*, re-group and paraphrase the same Spring and Autumn events instead of introducing new ones, creating naturally occurring, easy-to-confuse off-topic variants. As shown in Fig. 2 (2), we align these sources to our reign-based time keys using LLM-assisted reverse matching and fuzzy string matching (Appendix A.3). For each time key τ, we define $\mathcal{D}_\tau^{\text{cf}}$ as records from these layers that that share the same time key and describe the same situation in later paraphrase, but are *not used as ground-truth retrieval targets* for queries targeting τ. These historically grounded near-miss variants serve as hard negatives for time-aware retrieval.

3.3 Temporal Queries and Evaluation Splits

Temporal query design. Building on the time-keyed records $\{\mathcal{D}_\tau\}$, we design three families of temporal queries that mirror common ways historians ask about time: point queries (**P-Time**), gap queries (**G-Time**), and local-window queries (**W-Time**) as shown in Fig. 2 (3). Each query is mapped to a target interval Q_i on the reign-based timeline; span (single month vs. short multi-month window) and eventfulness (whether Q_i contains empty months) are vary independently, so any family can in principle target either eventful or empty periods.

To make gaps and empty intervals queryable, months with no annals entry are instantiated as standardized `no_event` records keyed by τ and included in the retrieval gallery. G-Time queries are gap-oriented: they ask which month(s) within a reign or specified range lack recorded events and are answered by these `no_event` records.

P-Time queries explicitly target a single time key (e.g., "What happens in 魯隱公元年三月?") and are answered by the corresponding event-bearing record, or by an explicit `no_event` record when that month is empty. W-Time queries target a local temporal window, such as "around the time when X occurs" or "in the months before/after Y", and are mapped to short contiguous ranges of time keys (see Fig. 7 for representative patterns). All query templates and concrete examples are provided in Appendix A.4.

Reign-aligned splits. We partition the month-level timeline into train/validation/test splits in a reign-aware, leak-free way. For each gong's reign, we allocate disjoint contiguous blocks of months to train, validation, and test, roughly in an 80/10/10 ratio, and assign all records and queries in those months to the corresponding split. No time key, record, or query ever appears in more than one split. At evaluation time, validation and test queries are drawn from their held-out reign segments, while retrieval is always performed over the full time-keyed record gallery, so models must generalize temporal reasoning from seen parts of a reign to months they have *never* observed during training. 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

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

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

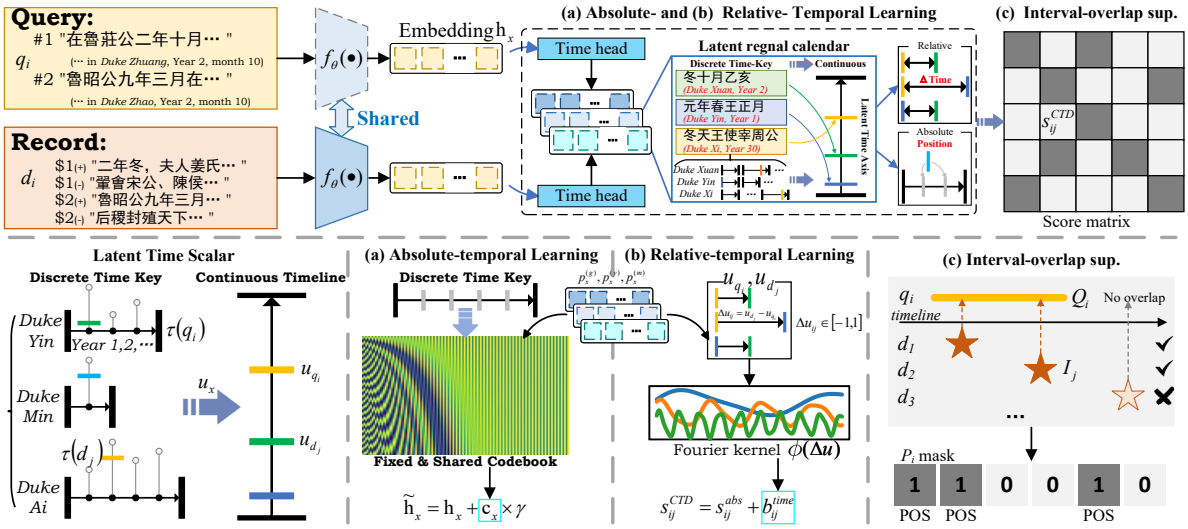

Figure 3: Overview of our Calendrical Temporal Dual-encoder (CTD). A shared Transformer dual-encoder encodes queries and records into embeddings. Temporal heads place each text on a unified regnal calendar axis (latent time scalar), supporting (a) absolute context injection and (b) relative biasing to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

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

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

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

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

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}:

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

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

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

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

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

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-EMBEDDING-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.

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.

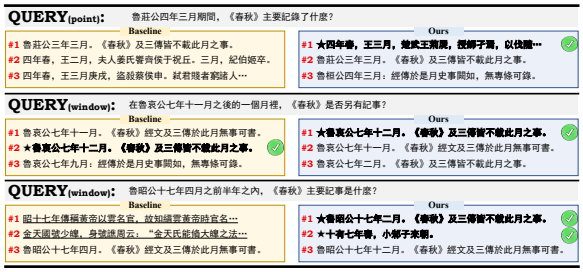

Figure 4: Visualization of Qualitative Examples.

5.3 Analysis

5.3.1 Impact of Query Type

Table 2 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 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 $\mathbf{c}_x$. 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 ei-

Method	Pub	Type	R@1	R@5	R@10	MRR@10	nDCG@10
Sparse retrieval							
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-$\ell_0{}_{(ZS)}$	SIGIR'25	Sparse	0.0006	0.0309	0.0587	0.0132	0.0143
Fusion / late interaction							
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
Dense retrieval (encoder-based)							
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 $_{\text{BERT-base}}$ (Ours)	This work	Dense	**0.5826**	**0.6721**	**0.7090**	**0.6193**	**0.4575**
Dense retrieval (LM-based embeddings)							
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 $_{\text{Qwen3-Embed-0.6B}}$ (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.

Method	Single-month ($\lvert Q\rvert = 1$)		Multi-month ($\lvert Q\rvert > 1$)	
	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 $_{\text{BERT-base}}$ (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 $_{\text{Qwen3-Embed-0.6B}}$ (Ours)	0.491	0.513	0.733 ↑+0.242	0.767 ↑+0.255

Table 2: 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.

Variant	$\mathcal{L}_{\text{multi}}$	b_{ij}^{time}	$\mathbf{c}_x$	R@1	MRR@10
FT baseline	–	–	–	0.5771 —	0.6044 —
+ $\mathcal{L}_{\text{multi}}$	✓	–	–	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 3: 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.

ther 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.

7 Limitations

ChunQiuTR is constructed from the *Chunqiu* annals and its major commentaries, and uses a reign-based month-level time key. 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 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 and time alignment, and we leave such expansions to future work.

Moreover, a month-level timeline cannot represent finer-grained temporal relations, and remaining errors indicate that (i) near-duplicate records in neighboring months and (ii) genuinely ambiguous historiographical cases are still challenging even with temporal supervision. Future work includes extending the benchmark to broader historical corpora and finer temporal granularity, incorporating stronger reranking or evidence-checking for borderline confusions, and evaluating downstream impacts in end-to-end RAG pipelines.

8 Ethical Considerations

Our data are derived from pre-modern Chinese historical texts (the *Chunqiu* annals and commentarial tradition) and do not contain information about living individuals. We use publicly available digital editions and licensed scholarly resources; when redistribution is permitted, we release processed splits and annotations, otherwise we provide code and metadata sufficient to reproduce the benchmark from legitimate sources.

These chronicles reflect historical norms and power structures (e.g., elite-centric perspectives and occasional descriptions of conflict or punishment). We preserve such content for historical and linguistic research rather than endorsement. We caution against deploying models trained on this benchmark in high-stakes decision-making, and recommend that any downstream RAG or narrative generation make the historical context explicit and include appropriate uncertainty signaling.

We follow the ACL Code of Ethics and complete the Responsible NLP checklist to the best of our knowledge.

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.

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, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2022. From distillation to hard negative sampling: Making sparse neural ir models more effective. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2353–2359, New York, NY, USA. Association for Computing Machinery.

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.

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 Mastrapas, 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.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

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.

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.

Bin Li, Bolin Chang, Zhixing Xu, Minxuan Feng, Chao Xu, Weiguang Qu, Si Shen, and Dongbo Wang. 2024. Overview of EvaHan2024: The first international evaluation on Ancient Chinese sentence segmentation and punctuation. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 229–236, Torino, Italia. ELRA and ICCL.

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.

Xiaomeng Pan, Hongfei Wang, Teruaki Oka, and Mamoru Komachi. 2022. Zuo zhuan Ancient Chinese dataset for word sense disambiguation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 129–135, Hybrid: Seattle, Washington + Online. 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.

Jon Saad-Falcon, Omar Khattab, Keshav Santhanam, Radu Florian, Martin Franz, Salim Roukos, Avirup Sil, Md Sultan, and Christopher Potts. 2023. UDAPDR: Unsupervised domain adaptation via LLM prompting and distillation of rerankers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11265–11279, Singapore. Association for Computational Linguistics.

Keshav Santhanam, Jon Saad-Falcon, Martin Franz, Omar Khattab, Avi Sil, Radu Florian, Md Arafat Sultan, Salim Roukos, Matei Zaharia, and Christopher Potts. 2023. Moving beyond downstream task accuracy for information retrieval benchmarking. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11613–11628, Toronto, Canada. Association for Computational Linguistics.

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.

Liquid AI Team. 2025. Lfm2 technical report. *Preprint*, arXiv:2511.23404.

Qwen Team and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3).

Qwen Team and 1 others. 2025a. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Qwen Team and 1 others. 2025b. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Huishuang Tian, Kexin Yang, Dayiheng Liu, and Jiancheng Lv. 2021. Anchibert: A pre-trained model for ancient chinese language understanding and generation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

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.

Yuting Wei, Qi Meng, Yuanxing Xu, and Bin Wu. 2025. TEACH: A contrastive knowledge adaptive distillation framework for classical Chinese understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3537–3550, Vienna, Austria. Association for Computational Linguistics.

Yuting Wei, Yuanxing Xu, Xinru Wei, Simin Yang, Yangfu Zhu, Yuqing Li, Di Liu, and Bin Wu. 2024. AC-EVAL: Evaluating Ancient Chinese language understanding in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1600–1617, Miami, Florida, USA. Association for Computational Linguistics.

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.

Guo Zhipeng, Xiaoyuan Yi, Maosong Sun, Wenhao Li, Cheng Yang, Jiannan Liang, Huimin Chen, Yuhui Zhang, and Ruoyu Li. 2019. Jiuge: A human-machine collaborative Chinese classical poetry generation system. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 25–30, Florence, Italy. Association for Computational Linguistics.

A Details of Dataset

A.1 Annals vs. Exegesis

The *Chunqiu* (《春秋》, *Spring and Autumn Annals*) itself is extremely terse: each entry records only a handful of characters for a given year and month in the state of Lu. Later readers often summarized its hermeneutic ideal with slogans such as "无事不书" ("nothing trivial is recorded") and "微言大义" ("subtle words, weighty meanings"), highlighting the assumption that even small lexical choices encode moral and political judgment.

By contrast, the three classical *zhuan* (《左传》, 《公羊传》, 《谷梁传》) are extended exegetical traditions that unpack the annals line by line. For a single cryptic record in the annals, they may provide several paragraphs of historical narrative, ritual debate, or moral argument. Our dataset construction treats the annals sentence as the *anchor* and the three exegesis traditions as aligned *expansions* that supply additional temporal and semantic context.

Example: "郑伯克段于鄢". A canonical illustration is the entry "郑伯克段于鄢" in the Duke Yin's first year. In the annals, this appears as a single laconic record:

> 《春秋》： 夏/五月/郑伯克段于鄢。
> *(In the fifth month of summer, the earl of Zheng "subdued" Duan at Yan.)*

The three exegesis traditions greatly expand on this one line:

- **Zuo zhuan** (《左传》) narrates the full backstory: the struggle between Zheng Zhuanggong (earl of Zheng) and his younger brother Duan, their mother's partiality, and the military campaign at Yan.
- **Gongyang zhuan** (《公羊传》) asks why the verb "克" ("to subdue") is used instead of "杀" ("to kill"), and interprets this as a moral judgment on Zheng Zhuanggong's behavior and on proper fraternal conduct.
- **Guliang zhuan** (《谷梁传》) further probes questions of rank, rite, and political legitimacy, treating the same event as a case study in correct ritual and naming.

This contrast between a *single* annals line and *multiple* rich exegetical expansions is typical. In our corpus, we use the terse annals sentence as the primary temporal anchor (e.g., "鲁隐公元年五月"), while the corresponding *Zuo*, *Gongyang*, and *Guliang* passages provide fine-grained descriptions of *what* happened at that time key, enabling temporally grounded retrieval over both the annals and its exegesis, as shown in Table 14.

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

A.2.1 Details of Reign-based time keys.

Our normalized time axis is defined at the granularity of *gong–year–month*. Each time key τ takes the canonical form

$$\tau = [\text{gong}] + [\text{year}] + [\text{month}],$$

such as 「鲁隐公元年正月」 or 「鲁桓公元年三月」. This allows us to place all annals and commentary snippets on a single, consistent timeline while preserving the traditional regnal structure.

To construct these time keys, we scan the annals sequentially and maintain a triple (g, y, m) for the current *ruler* (g), *regnal year* (y), and *month* (m). At each step, we update this triple according to simple pattern-based rules:

- **Start of a reign-year.** Phrases such as 「元年，春，王正月」 initialize a new triple: we set the current ruler g (e.g., 隐公), the regnal year y = 元年, and the month m = 正月, yielding the key 「鲁隐公元年正月」.

- **New year under the same ruler.** When we encounter a bare year marker such as 「二年，春，公会戎于潜。」 without a new ruler name, we interpret it as a new regnal year under the current ruler. We update y to 「二年」, reset the month m to 正月, and obtain the key 「鲁隐公二年正月」.

- **New month under the same reign-year.** Month expressions such as 「三月，公及邾仪父盟于蔑。」 or 「夏，五月，郑伯克段于鄢。」 provide a new month but no new ruler or year. In this case we keep the current g and y, update m to the detected month (e.g., 三月 or 五月), and assign keys like 「鲁隐公元年三月」 or 「鲁隐公元年五月」.

- **Implicit inheritance.** Sentences that do not contain any explicit ruler, year, or month simply inherit the previous triple (g, y, m). Thus, multiple consecutive events may share the same time key if the chronicle does not introduce a new temporal cue.

- **Change of ruler.** When the text signals a new ruler, for example 「元年，春，王正月，公即位。」 for 桓公, we switch the ruler g to the new gong, reset the year to 元年 and the month to 正月, yielding 「鲁桓公元年正月」. Subsequent month markers such as 「三月，公会郑伯于垂，郑伯以璧假田。」 then update only m, giving 「鲁桓公元年三月」.

Applying these rules produces a deterministic mapping from raw chronicle phrases to normalized time keys, as illustrated in Table 4. The same normalized keys are then used in our record–time-key alignment step to group all annals and commentary passages that belong to the same (g, y, m) cell on the timeline.

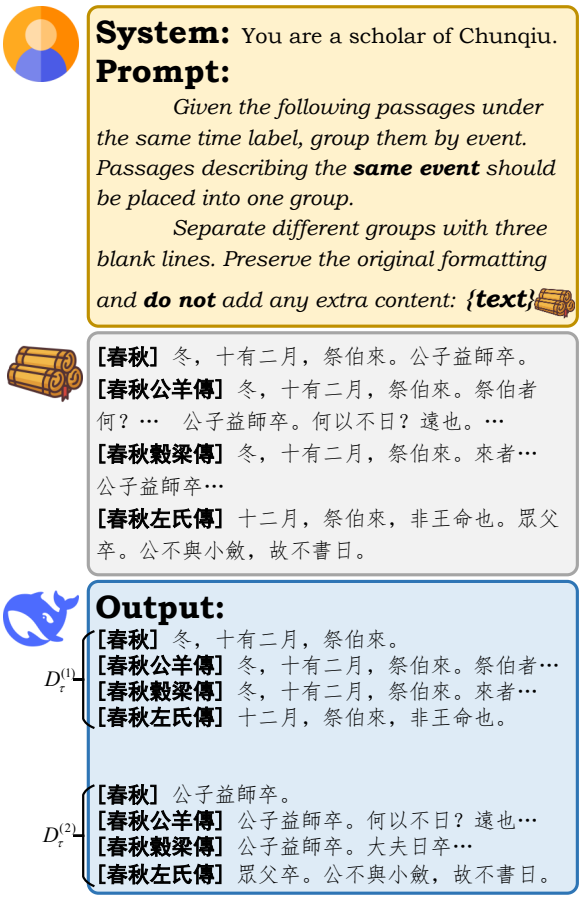

Figure 5: Text-only prompt and example of event-level grouping under the reign-based time label "鲁隐公元年十二月". (a) Instruction given to the language model. (b) Raw passages from different sources under the same time label. (c) Target output in which passages are grouped by event ($D_\tau^{(1)}$ and $D_\tau^{(2)}$), with different groups separated by three blank lines.

A.2.2 Details of Record–time-key alignment

This appendix illustrates how multiple event-level records are aligned under a single normalized time key. A lightweight LLM prompt is used to suggest splits for mixed sentences, and all suggestions are then manually verified.

In the original annals, these two events are compressed into a single sentence:

> 冬，十有二月，祭伯来。// 公子益师卒。
> *In winter, the twelfth month, Zhai Bo came; Prince Yishi passed away.*

To obtain event-level records under a fixed time key, we first use a lightweight LLM prompt to suggest a grouping, and then manually verify and correct the result. The core prompt is shown in Fig. 5.

For the time key 「鲁隐公元年十二月」, the input to the LLM is the concatenation of the snippets from the annals and the three *zhuan* under this

Original chronicle snippet	Detected cue & update rule	Normalized time key
元年，春，王正月。	New reign-year detected (隐公元年)，initialize month to 正月	鲁隐公元年正月
三月，公及邾仪父盟于蔑。	New month (三月)，no new reign or year, inherit previous reign-year	鲁隐公元年三月
夏，五月，郑伯克段于鄢。	New month (五月)，no new reign or year, inherit previous reign-year	鲁隐公元年五月
...	...	...
二年，春，公会戎于潜。	New year (二年) under the same reign (隐公)，month reset to 正月	鲁隐公二年正月
夏，五月，莒人入向。	New month (五月)，inherit current reign-year (隐公二年)	鲁隐公二年五月
...	...	...
元年，春，王正月，公即位。	New reign detected (鲁桓公)，year reset to 元年，month reset to 正月	鲁桓公元年正月
三月，公会郑伯于垂，郑伯以璧假田。	New month (三月)，inherit current reign-year (桓公元年)	鲁桓公元年三月
...	...	...

Table 4: Examples of mapping raw chronicle phrases to reign-based normalized time keys by propagating and updating reign, year, and month cues.

time key. The model suggests splitting the mixed sentence "[春秋] 冬，十有二月，祭伯来。公子益师卒。" into two groups corresponding to 「祭伯来」 and 「公子益师卒」, which are then aligned with the respective commentary passages, as shown in Fig. 5. The final records are therefore:

- $D_\tau^{(1)}$: arrival of Zhai Bo and its three commentarial interpretations;
- $D_\tau^{(2)}$: death of Prince Yishi and its three commentarial interpretations.

All such LLM suggestions are manually checked to ensure that each record corresponds to a coherent historical event.

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, as shown in Table 14:

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, 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), and layered annotations (注疏). By aligning them

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}=隱公

{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，虢叔死焉，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 "祭仲谏郑庄封叔段" 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.

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.

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

Table 5: Query template groups used in our experiments. We distinguish point vs. window queries (by span) and whether a group primarily targets detailed content vs. existence-oriented intents (including gap/no-event answers).

A.4 Details of Queries

For the time-aware retrieval experiments, we instantiate a small set of natural-language templates over our reign-based time keys. All templates are written in Traditional Chinese and are filled with a normalized time expression (e.g., "公元年二月"). We distinguish: (i) *point* queries that refer to a single month, and (ii) *window* queries that refer to a span of months or years around a given reference month. Note that "no-event" (gap-style) is not a third structural type: empty months may be the correct answer for both point queries and window queries, depending on the target interval.

Within point queries, we further separate templates that explicitly ask for detailed historical content from those that only ask whether the month has any entry in the *Chunqiu* canon (i.e., existence-

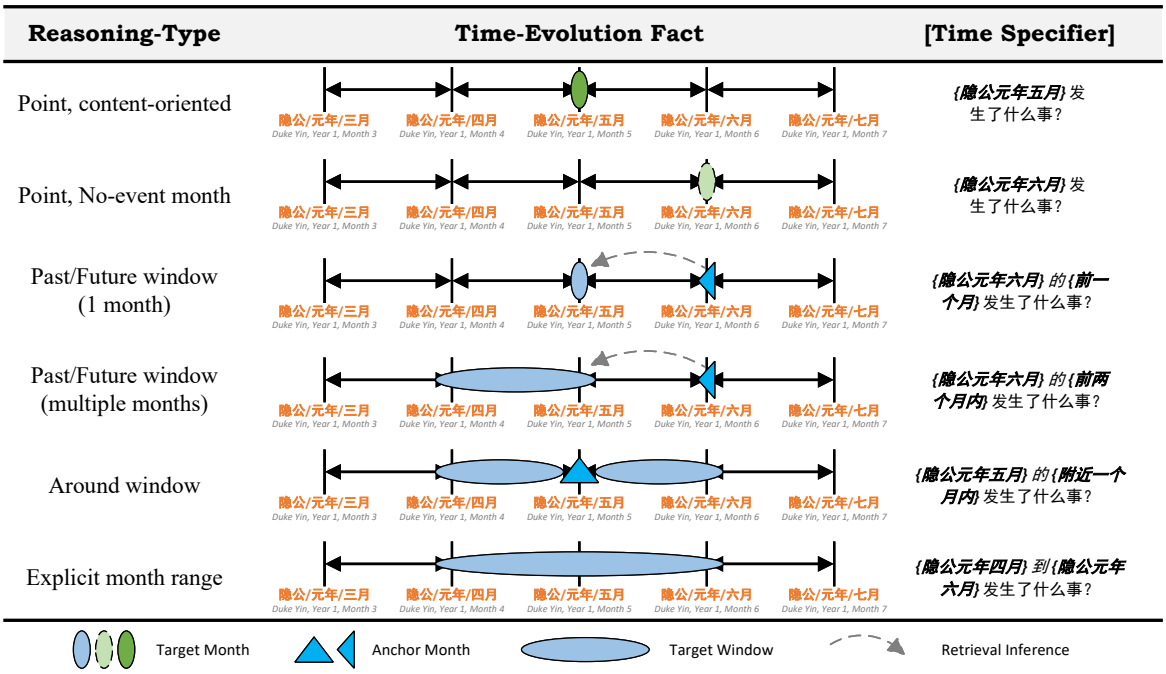

Figure 7: Temporal reasoning patterns in the Spring and Autumn benchmark. We visualize representative time-evolution facts over reign-based year–month timelines, including content-oriented point queries (with and without recorded events), past/future windows of different spans, symmetric around windows, and explicit month ranges. Each row links a Classical Chinese time specifier to its normalized anchor month and the target retrieval window used during evaluation.

or gap/no-event–oriented intents). During construction, we deliberately mix months with and without recorded events under the same template pool, so that models cannot exploit superficial lexical cues to distinguish "gap" months from eventful months.

Table 5 summarizes the query template groups and reasoning types, and Fig. 7 illustrates representative temporal interpretations; the full natural-language templates are enumerated below. Window queries can likewise cover empty months; when the target span includes such months, the ground-truth set contains the corresponding `no_event` placeholders for those keys.

Point queries (single-month). Content-oriented templates (BASE 1–12). Here `{time_text}` is replaced with a normalized month-level time expression, e.g., "鲁隐公元年二月".

1. `{time_text}` 发生了什么事?
2. 在 `{time_text}`，《春秋》记录了哪些史事?
3. `{time_text}`，鲁国有何记事?
4. `{time_text}`，《春秋》经文书何事?
5. `{time_text}`，《春秋》经传有何记载?
6. `{time_text}` 期间，《春秋》主要记录了什么?
7. 请列出 `{time_text}` 在《春秋》中的记事。
8. `{time_text}`，《春秋》记了几件事?
9. 查询《春秋》，`{time_text}` 这一月有什么重要事件?
10. 《春秋》在 `{time_text}` 这条记录说了什么内容?
11. `{time_text}` 在《春秋》原文中的记事内容是什么?
12. 如果只看《春秋》，`{time_text}` 这段时间发生了哪些变化?

Existence / gap-oriented templates (BASE 13–20).

13. 在 `{time_text}`，《春秋》是否有记事?
14. `{time_text}`，《春秋》经传有无记载?
15. 请问 `{time_text}` 这一月，《春秋》有没有书写事件?
16. `{time_text}`，《春秋》有记事还是阙如?
17. `{time_text}`，《春秋》经文对此月是记事还是无事可书?
18. 《春秋》在 `{time_text}` 是否另立一条记事?
19. 请检索 `{time_text}` 在《春秋》中是记事月份还是空月?
20. `{time_text}` 在《春秋》里究竟有无条目?

Window queries (month-level). Here `{delta_desc}` is a verbalized span description (e.g., "前后三个月").

Past windows (MONTH_PAST).

21. 在 `{time_text}` 之前的 `{delta_desc}` 内，《春秋》记录了哪些事件?
22. 自 `{time_text}` 向前推 `{delta_desc}`，《春秋》经传有哪些记事?
23. 回顾 `{time_text}` 之前 `{delta_desc}`，《春秋》对鲁国有何记载?
24. `{time_text}` 之前 `{delta_desc}` 之内，《春秋》主要记事是什么?

25. 请检索 {time_text} 之前 {delta_desc} 内，《春秋》中所有相关记事。
26. 在 {time_text} 以前的 {delta_desc} 里，《春秋》是否记录了重要事件?

Future windows (MONTH_FUTURE).

27. 自 {time_text} 起的 {delta_desc} 内，《春秋》记录了哪些事件?
28. 从 {time_text} 开始，接下来 {delta_desc} 内《春秋》经传有哪些记事?
29. {time_text} 之后的 {delta_desc} 之内，《春秋》主要记录了什么史事?
30. 在 {time_text} 之后的 {delta_desc} 里，《春秋》是否另有记事?
31. 以 {time_text} 为起点，后续 {delta_desc} 内《春秋》对鲁国有何记载?

Around windows (MONTH_AROUND).

32. 在 {time_text} 前后 {delta_desc} 内，《春秋》记录了哪些事件?
33. 以 {time_text} 为中心，前后 {delta_desc} 内《春秋》经传有哪些记事?
34. {time_text} 附近 {delta_desc} 之内，《春秋》主要记事是什么?
35. 请检索 {time_text} 前后 {delta_desc} 内，《春秋》中所有相关的记事。
36. {time_text} 前后 {delta_desc} 之内，《春秋》是否有重要事件被记录?

Explicit month ranges (MONTH_RANGE). Here {start_text} and {end_text} are month-level endpoints, and {range_text} is a verbalized range such as "鲁隐公元年正月至二年春".

37. 从 {start_text} 到 {end_text} 期间，《春秋》记录了哪些事件?
38. {range_text} 之间，《春秋》经传有哪些记事?
39. 请列出 {range_text} 这几个月内，《春秋》所载的史事。
40. 在 {range_text} 这段时间里，《春秋》主要记录了哪些事件?
41. {range_text} 期间，《春秋》有无重要记事?

Window queries (year-level). Here {year_text} is a normalized reign-year expression, e.g., "鲁隐公元年".

Current year (YEAR_CURRENT).

42. {year_text} 一年内，《春秋》记录了哪些事件?
43. 在 {year_text}，《春秋》经传共有哪些史事?
44. 请列出 {year_text} 这一年，《春秋》中所有记事。
45. {year_text}，《春秋》主要记事是哪些事件?
46. {year_text} 一年内，《春秋》是否记录了重要事件?

Previous year (YEAR_PAST).

47. 相对于 {year_text}，在前一年《春秋》记录了哪些事件?
48. {year_text} 的上一年，鲁国在《春秋》中有哪些主要记事?
49. 在 {year_text} 之前一年内，《春秋》是否有重要事件被记录?

Next year (YEAR_FUTURE).

50. 相对于 {year_text}，在下一年《春秋》又记录了哪些事件?
51. {year_text} 的下一年，鲁国在《春秋》中有哪些记事?
52. 在 {year_text} 之后一年内，《春秋》是否有新的重要记事?

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 zhuanshuo* and *Chunqiu Zhengyi* (negative pool). Table 6 reports basic character-length statistics over these raw records.

In this raw snapshot, the positive pool comprises 6641 records with a total of 293.7K characters (average length 44.22), whereas the negative pool comprises 9227 records with 1014.9K characters (average length 109.99). Later exegetical texts are thus substantially longer and more discursive than the terse annalistic records, which motivates our subsequent sentence-level segmentation to reduce length bias and format heterogeneity.

Fig. 8 visualizes month-level coverage and gaps across the full *Chunqiu* timeline. Each panel corresponds to either the overall timeline or a single Lu ruler and shows three quantities: the total span in months, the number of gap months with no recorded event, and the proportion of gaps. Overall, the normalized timeline covers 3036 months from Duke Yin to Duke Ai, with roughly half of all months lacking any recorded event. Across individual reigns, the gap ratio typically stays close to 40–60%, but some reigns such as Zhuang, Xuan and Ai exhibit both long spans and relatively high gap proportions, including stretches of up to around a dozen consecutive empty months under Duke Ai. This pattern confirms that even within a fixed chronological frame, the annalistic record is highly uneven, providing a natural source of gap months for our time-aware retrieval tasks.

A.5.2 Record-level segmentation.

To avoid length-induced biases and heterogeneous formatting across sources, we convert all raw records in Table 6 into a corpus of sentence-level records, which are used as the retrieval units (gallery entries) in our experiments. Raw classical Chinese does not come with modern punctuation,

Pool	Source	# Raw units	Total chars (K)	Avg. len.	Median	Min	Max
Positive	All (annals + three *zhuan*)	6641	293.7	44.22	–	–	–
Positive	*Chunqiu* annals	1532	19.2	12.52	10	2	71
Positive	*Gongyang* zhuan	1776	44.5	25.05	10	2	671
Positive	*Zuo* zhuan	1547	189.1	122.20	48	3	2658
Positive	*Guliang* zhuan	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 6: 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.

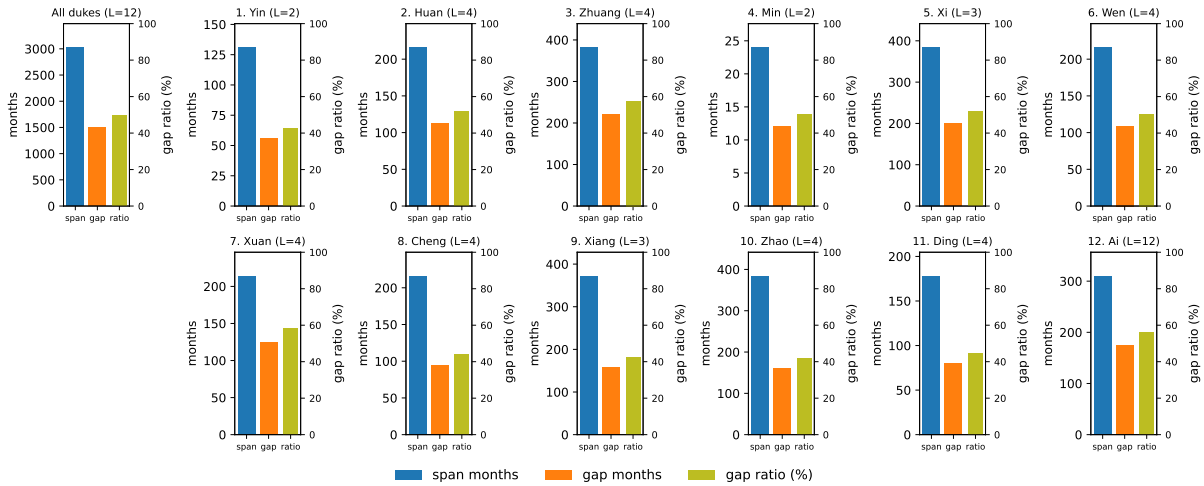

Figure 8: Month-level coverage, gap counts, and gap ratios for the normalized *Chunqiu* timeline (overall and per Lu ruler).

and commentary passages in particular can be extremely long and discursive. We therefore adopt a two-stage segmentation procedure.

First, we use a large language model to propose punctuation and sentence boundaries (句读) for each raw record, and then apply a light rule-based splitter over classical markers such as “曰”, “云”, and “传曰”. During this step we also perform minimal normalization to reduce stylistic boilerplate, for example stripping framing markers like “正义曰” or formulaic quotation headers that do not carry substantive event content, and discarding extremely short fragments that only contain a few characters.

Second, each resulting record inherits the reign-based time key and source metadata of its parent raw record, and we assign a coarse type label indicating whether it is an event description, an explicit “no-event” statement, or a counterfactual negative drawn from later exegetical layers. For months where the annals and the three *zhuan* agree that there is nothing to record, we additionally synthesize a standardized no-event record for each 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 a no-event case still requires matching the correct reign and month rather than a single global “nothing happened” entry.

This process yields 20,172 records in total, which constitute the retrieval gallery in our time-aware experiments (Table 7).

A.5.3 Final benchmark splits.

The final benchmark is constructed over record-level retrieval units paired with normalized reign-based time keys and natural-language

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 7: 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`).

queries. Month-level segments are partitioned into train/validation/test splits at the month granularity using a roughly 80/10/10 ratio, and all records and queries inherit the split of their associated month key to avoid temporal leakage. As summarized in Table 7, this yields 2424/295/317 months in train/validation/test (3036 months in total).

On the document side, the record corpus contains 16,027 records for training, 2,049 for validation, and 2,096 for test (20,172 in total). Most records come from later exegetical layers and are labeled as `neg_comment` (9,458/1,271/1,165 for train/validation/test), with 5,360/626/782 `event` records derived from the annals and the three *zhuan*, and 1,209/152/149 synthetic or explicit `no_event` records that represent documented silent months.

On the query side, we obtain 13,053 training queries, 1,520 validation queries, and 1,653 test queries (16,226 in total), covering point, gap, and window families. Each query is associated with one or more ground-truth relevant records drawn from the record corpus. Averaged over all queries, we obtain approximately 7.3, 6.8, and 7.2 ground-truth records per query in train/validation/test, respectively, providing a reasonably dense supervision signal for time-aware retrieval and reasoning models.

A.5.4 Example JSON entries.

For completeness, we show one concrete example of how the released files are structured and linked. The benchmark is stored in three parts: (i) a month-level metadata file `chunqiu_meta.json`, (ii) a query file `queries_all_labeled.jsonl`, and (iii) a month-level split file `time_splits_by_month_v1.json`. All three are keyed by the same reign-based sort key `[gong, year, month]`.

A representative entry from `chunqiu_meta.json` is:


```
{
 "time_mapping": {
  "鲁隐公元年正月": {
   "new_id": 1,
   "time_text": "鲁隐公元年正月",
   "sort_key": [1, 1, 1],
   "versions": [
   {
    "春秋公羊传": [
    { "sid": 0, "idx": 0,
     "text": "元年春，王正月。元年者何？君之始年也。..." },
    { "sid": 1, "idx": 1,
     "text": "隐长又贤，诸大夫扳隐而立之。..." }
    ],
    "春秋谷梁传": [
    { "sid": 2, "idx": 0,
     "text": "元年春，王正月。虽无事，必举正月，谨始也。..." }
    ],
    "春秋左氏传": [
    { "sid": 4, "idx": 0,
     "text": "元年春，王周正月。不书即位，摄也。" }
    ],
    "春秋": [
    { "sid": 5, "idx": 0,
     "text": "元年，春，王正月。" }
    ],
    "neg_sample": [
    {
     "work": "春秋大事表",
     "source": "三传异同表",
     "compiler": "顾栋高",
     "origin_event": "元年春王正月",
     "parsed_comment": [
     { "sid": 6, "idx": 0,
      "text": "春王正月，三传皆无明文。左氏以正月为建子，..." }
     ]
    }
    ]
   }
   ]
  }
 }
}
```


Here the time key "鲁隐公元年正月" is normalized to `"sort_key": [1, 1, 1]`, and all records under this month (identified by `sid`) inherit this key.

The corresponding point query in `queries_all_labeled.jsonl` is stored as:


```
{
 "qid": "point-1",
 "query": "鲁隐公元年正月发生了什么事？",
 "time_text": "鲁隐公元年正月",
 "sort_key": [1, 1, 1],
 "type": "point",
 "no_event": false,
 "template_id": 1,
 "target_sort_keys": [[1, 1, 1]],
 "pos_sids": [0, 1, 2, 3, 4, 5],
 "orig_qid": 1,
 "source": "point"
}
```


This line shows a single point query anchored at the same month; the field `pos_sids` lists all record IDs that are labeled as relevant answers.

Finally, the month-level split file `time_splits_by_month_v1.json` specifies which global split this reign month belongs to:


```
{
```

```
  "split_by_sort_key": {
   "1-1-1": "train",
   "1-1-2": "train",
   "1-1-3": "train",
   ...
  }
 }
```

At training time, we map between these three files via the shared sort keys and record IDs: all records under a given month inherit its split, and each query uses `pos_sids` to reference its ground-truth relevant records within the record-level corpus.

A.5.5 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 the Wikisource platform content is provided under CC BY-SA 4.0 and governed by the Wikimedia Terms of Use. To facilitate compliant reuse, we (i) record the Wikisource page revision IDs (`oldid`) for every document, and (ii) release the benchmark as derived metadata (time keys, alignments, query/qrels, indices) with scripts for re-downloading the raw texts from Wikisource.

A.6 Additional Related Work of Classical Chinese NLP.

Classical Chinese NLP has advanced via domain-specific pretraining (e.g., AnchiBERT (Tian et al., 2021)), core analyses (e.g., segmentation/POS), and downstream understanding tasks such as WSD (Pan et al., 2022) and translation. Recent benchmarks further assess LLMs on Ancient/Classical Chinese understanding across diverse task formats (Wei et al., 2024; Li et al., 2024). Relatedly, constrained generation in the classical domain (e.g., traditional-form poem generation) studies producing fluent text under strict formal constraints (Wei et al., 2025; Zhipeng et al., 2019). Despite these efforts, resources remain heterogeneous and temporally grounded reasoning is often implicit, motivating retrieval benchmarks that couple classical-language variation with explicit fine-grained time structure.

B Details of Methods

B.1 Details of Experiment Setting

B.1.1 Model and Training Details

Backbones and pooling. We fine-tune two dual-encoder backbones: BERT-BASE-CHINESE and QWEN3-EMBEDDING-0.6B. For BERT-BASE, we use `[CLS]` pooling; for QWEN3-EMBEDDING-0.6B, we use last-token pooling, following the default embedding setup of the model.

Training objective and time supervision. Both backbones are trained with a contrastive retrieval objective with multi-positive supervision and explicit hard-negative training. In addition, we include an auxiliary time classification loss over the three discrete factors (gong/year/month) with weight 0.1. We apply label smoothing to the time heads with $\epsilon = 0.2$. For the time-aware variants, we enable (i) the relative temporal bias and (ii) the soft absolute temporal context derived from predicted time distributions (used for context injection), consistent with the main model description.

Optimization and lengths. 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/passages lengths of 64/196. For QWEN3-EMBEDDING-0.6B, we train for 3 epochs with an effective batch size 16 and learning rate 3×10^{-6}, using maximum query/passages lengths of 128/256; we also enable global in-batch negatives.

Model selection and evaluation. We select the checkpoint with the best **Recall@1** on the validation split and report **Recall@K** and **MRR@10** on the test split under the same evaluation protocol. During evaluation, we include both commentary negatives and explicit `no_event` records in the candidate gallery.

B.1.2 Training Cost and Computational Resources

Following the ARR Responsible NLP checklist, we report the computing infrastructure and the total computational budget (GPU hours) for our main fine-tuning runs, and summarize our experimental setup and chosen hyperparameters in Appendix A.[1]

Hardware. We fine-tune BERT-BASE-CHINESE on a single GPU (either 1× NVIDIA RTX A6000 or 2× RTX 3090, depending on availability). For QWEN3-EMBEDDING-0.6B, we use multi-GPU distributed training (either 2× RTX A6000 or 4×

[1] The checklist explicitly asks for computing infrastructure and total budget (e.g., GPU hours), as well as a discussion of experimental setup and hyperparameters.

Work / layer	Role in ChunQiuTR	Digital source (edition)	License note / release plan
Chunqiu (春秋)	base corpus	Chinese Wikisource (*Siku Quanshu* 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.
Zuo zhuan (左氏)	base corpus		
Gongyang zhuan (公羊)	base corpus		
Guliang zhuan (梁)	base corpus		
Gu Donggao, *Chunqiu Dashibiao* (栋高《春秋大事表》)	chrono-near paraphrastic negatives		
Wei Liaoweng, *Chunqiu Zuozhuan Yaoyi* (魏了翁《春秋左要》)	chrono-near discursive negatives		
Zuozhuan annotations / sub-commentaries (注疏; e.g., 音/杜注/正/)	lexical-near annotation negatives		

Table 8: 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.

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-EMBEDDING-0.6B	FT baseline	2	≈45 min	≈1.50
QWEN3-EMBEDDING-0.6B	CTD (full)	2	≈45 min	≈1.50

Table 9: Compute cost for representative fine-tuning runs on ChunQiuTR. Time is wall-clock per run (including periodic validation); GPU-hours are computed as (#GPU)×(time in hours).

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

Table 10: Ablation study on the test set under the same evaluation protocol of BERT-base-chinese.

RTX 3090), which is required to enable global in-batch negatives under our effective batch size.

Wall-clock time and GPU hours. Table 9 reports the wall-clock time per run for representative fine-tuning settings, together with the resulting GPU-hours (GPU count × wall-clock hours). These costs cover end-to-end fine-tuning with periodic validation; sparse baselines and non-parametric time priors have negligible training cost.

Hyperparameter search. 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** (Appendix A).

B.2 Details of Analysis

B.2.1 Details of Ablation Study

We further replicate the ablation in Table 10 using `bert-base-chinese` to examine backbone sensitivity. Adding the retrieval objective $\mathcal{L}_{\text{multi}}$ already yields a consistent improvement over the FT baseline, indicating that explicit contrastive supervision helps BERT better separate time-keyed records. Incorporating temporal modeling brings substantially larger gains: the relative-time bias b_{ij}^{time} improves top-rank accuracy, suggesting that it effectively re-calibrates pairwise similarities toward chronologically plausible candidates, while the soft absolute temporal context $\mathbf{c}_x$ yields a stronger boost on both R@1 and MRR@10, reflecting improved ranking quality across the list. Combining both components achieves the best performance, with additive gains over each single module, supporting their complementarity—bias reshapes the matching landscape at the logit level, whereas context enriches embeddings with distributional absolute-time cues. Notably, the overall improvements on BERT are larger than those on stronger embedding backbones, consistent with a larger "headroom" where temporal supervision and structured negatives can correct near-duplicate and chrono-adjacent confusions more effectively.

B.2.2 No-event and Hard-negative Behavior

Table 11 probes two protocol switches: whether to include *pure no-event queries* (`dq`) and whether to inject `neg_comment` passages as chrono-near hard negatives (`neg`). On the left, keeping pure no-event queries (`dq=0`) yields a large and highly consistent uplift across all retrievers (roughly +0.11 to +0.18 on both R@1 and MRR@10; e.g., BM25 +0.178/+0.173 and Qwen3-Embed$_{\text{(FT)}}$ +0.181/+0.171). This indicates that empty-month

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	↑+0.178	↑+0.173	0.240	0.296	0.228	0.282	↓-0.012	↓-0.014
ColBERT-LFM2$_{(ZS)}$	0.220	0.277	0.333	0.386	↑+0.113	↑+0.109	0.220	0.277	0.222	0.278	↑+0.002	↑+0.001
BERT-base$_{(FT)}$	0.378	0.439	0.560	0.606	↑+0.182	↑+0.167	0.378	0.439	0.306	0.374	↓-0.072	↓-0.065
CTD $_{\text{BERT-base}}$ (Ours)	0.407	0.456	0.583	0.619	↑+0.175	↑+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	↑+0.163	↑+0.177	0.062	0.093	0.054	0.082	↓-0.008	↓-0.011
Qwen3-Embed-0.6B$_{(FT)}$	0.396	0.434	0.577	0.605	↑+0.181	↑+0.171	0.396	0.434	0.396	0.433	0.000	↓-0.001
CTD $_{\text{Qwen3-Embed-0.6B}}$ (Ours)	0.420	0.457	0.594	0.621	↑+0.174	↑+0.164	0.420	0.457	0.418	0.455	↓-0.002	↓-0.002

Table 11: 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/keeps *pure no-event queries* whose ground-truth set contains only no_event records. **Right:** neg=0 vs. neg=1 (fix ne=1, dq=1), where neg injects neg_comment passages into the gallery as chrono-near hard negatives. Δ denotes the within-method change (higher is better).

questions are intrinsically easier under our retrieval formulation and can inflate aggregate scores if mixed with event-seeking queries without stratification. Notably, CTD models exhibit a similar uplift magnitude to their backbones, suggesting that explicit no_event supervision is useful but does not solely account for CTD's overall gains.

On the right, injecting neg_comment distractors makes performance changes more model-dependent. BERT-base$_{(FT)}$ drops substantially (-0.072 R@1 / -0.065 MRR), whereas CTD $_{\text{BERT-base}}$ is essentially unaffected (0.000/0.000), indicating markedly improved robustness to exegesis-style chrono-near confounds. For LM-based encoders, Qwen3-Embed$_{(FT)}$ is already stable (0.000/-0.001) and CTD $_{\text{Qwen3-Embed}}$ changes only marginally (-0.002/-0.002). Overall, the benchmark contains an "easy" no-event subset that should be controlled by protocol, and a genuinely harder regime induced by exegetical hard negatives, where CTD improves stability without removing all difficulty.

B.2.3 Full Protocol Grid Results

Protocol switches and mode naming. To make evaluation conditions explicit and to pre-empt questions about alternative settings, we report a full grid over three evaluation switches. We denote an evaluation mode as neg{0/1}_ne{0/1}_dq{0/1}, following our implementation: (i) neg indicates whether exegetical neg_comment sentences are injected into the gallery as chrono-near hard negatives; (ii) ne indicates whether explicit no_event placeholders are included in the gallery; (iii) dq equals drop_no_event_q, i.e., dq=1 drops queries whose gold set contains only no_event records, while dq=0 keeps them. When ne=0, no_event records are absent from the gallery, so queries with *only* no_event gold records are not well-defined for retrieval; therefore ne=0 is reported only with dq=1.

Unless otherwise stated, our **main-paper evaluation uses dq=0 (no dropping)** to reflect the intended benchmark that includes both event-bearing and explicitly empty months. We additionally report dq=1 as a diagnostic setting that isolates event-seeking behavior without the potentially easier pure no-event subset.

Full-grid results. Tables 12 and 13 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).

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 misses the correct passage at rank 1 but still retrieves at least one *ground-truth* passage within the top–5, while the baseline fails to surface any ground-truth evidence. In

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 12: Full protocol grid results for a single model (**CTD**-QWEN3-EMBEDDING-0.6B). Validation has 1,520 queries and test has 1,653 queries. Mode names follow neg/ne/dq as defined in Section B.2.3.

both cases, a key confounder is the highly reusable *no-event* style template (e.g., "无事可书") and its chrono-near reoccurrences in adjacent months, which can dominate lexical cues and lead to rank-1 swaps. For the point query, our model places multiple ground-truth passages about establishing the *Wu Palace* within the top–5, indicating that the miss is largely an ordering issue. For the window query, boundary-adjacent no_event records may be over-preferred, pushing event-bearing months downward; this suggests that window queries may benefit from query-type-aware calibration and coverage-oriented ranking objectives.

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

Figure 10 illustrates an online LLM queried *without* any supporting materials on a month-level point question from ChunQiuTR (Duke Yin, year 1, month 12; gold entries: 祭伯来 and 公子益师卒). Without a reasoning trace, the model incorrectly asserts an empty month (GT coverage 0/2). With a reasoning trace enabled, it produces a more elaborate narrative but still drops a key gold record (公子益师卒) and makes an unsupported claim that no other event is recorded (GT coverage 1/2). In contrast, Figure 11 shows that, when we provide an evidence pack containing the gold month records (along with confusable material) and enforce an evidence-selection-and-citation format, the same model recovers both gold entries (GT coverage 2/2) and grounds the final answer in the cited records, in both thinking and non-thinking modes. These cases suggest that longer "thinking" traces alone do not ensure temporal completeness, whereas explicit evidence binding substantially improves month-level faithfulness.

B.3 Details of Compared Methods

B.3.1 Sparse retrieval.

For sparse retrieval, we group a classical lexical baseline and two inference-free neural sparse retrievers (Formal et al., 2021, 2022) under a single sparse family, all evaluated under the same *sparse* retrieval protocol.

- **BM25** (Robertson and Zaragoza, 2009) is a standard term-matching model that scores documents based on exact lexical overlap and corpus-level term specificity, providing a strong and transparent lexical baseline.

- **BM25+TimeKDE.** Inspired by classical temporal priors in TIR (Li and Croft, 2003), we augment BM25 with a simple time-aware reranking step: for each query we first obtain

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 13: Full protocol grid results for BM25, reported in the same format as Table 12. Validation has 1,520 queries and test has 1,653 queries.

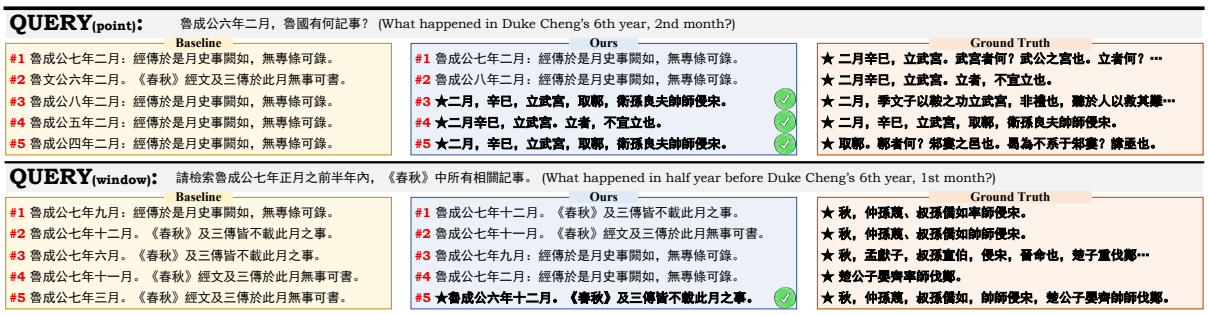

Figure 9: Top-1 near-miss cases on ChunQiuTR (test set). We show the top–5 results from a baseline (left) and ours (middle), with the ground-truth set (right). ⋆ marks ground-truth passages; ✓indicates a ground-truth hit in top–5.

BM25 scores, estimate a Gaussian kernel-density distribution over the regnal-month indices of the top-K hits, and then add the resulting query-specific log time prior to the BM25 scores for all documents. This baseline remains inference-free and non-parametric, but explicitly biases retrieval toward months whose neighbors are frequently retrieved.

- **SPLADE-IDF**$_{(ZS)}$ (Geng et al., 2025) is an inference-free neural sparse retriever that uses a masked-LM–style scoring function to expand each document into a sparse vocabulary-level vector with context-dependent term weights, while queries remain an IDF-weighted bag-of-words over the same vocabulary; this keeps query encoding purely lexical but injects learned contextual expansion on the document side.

- **SPLADE-**$\ell_{0(ZS)}$ (Shen et al., 2025) is a sparsity-controlled variant trained with an -style regularizer to further reduce the number of non-zero dimensions in document vectors, aiming for smaller indices and lower retrieval cost with minimal effectiveness loss; like SPLADE-IDF, it is used zero-shot on our Classical Chinese benchmark.

B.3.2 Fusion / late-interaction retrieval.

For fusion / late-interaction retrieval, we use two multi-vector models that keep per-token representations and score relevance via explicit query–

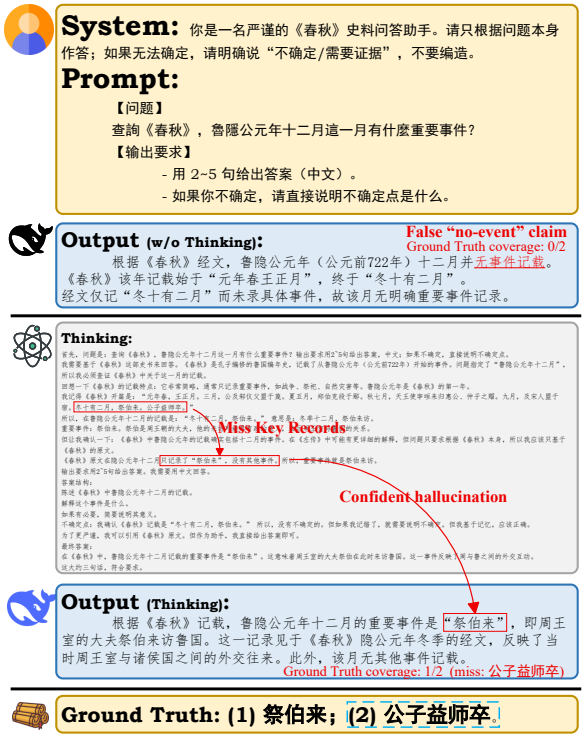

Figure 10: Online LLM without evidence grounding (No-RAG) on a month-level point query from ChunQiuTR. The query asks what is recorded in *Duke Yin, year 1, month 12*. The gold month contains two independent annal entries: 祭伯来 and 公子益师卒 (GT coverage = 2). Without a reasoning trace, the model incorrectly predicts an empty month (GT coverage 0/2). With a reasoning trace enabled, the model mentions 祭伯来 but omits 公子益师卒 and over-commits to the claim that no other event is recorded (GT coverage 1/2).

document token interactions rather than collapsing each side into a single embedding; both are applied in a zero-shot fashion without further fine-tuning.

- **ColBERT-JINA**$_{(ZS)}$ (Jha et al., 2024) follows the ColBERT-style (Khattab and Zaharia, 2020; Santhanam et al., 2023; Saad-Falcon et al., 2023) design that encodes queries and documents into token-level embeddings and aggregates their interactions with a max-sim operator, enabling fine-grained lexical and semantic matching beyond single-vector dense retrievers.

- **ColBERT-LFM2**$_{(ZS)}$ (Team, 2025) extends the ColBERT paradigm with longer input windows and multi-scale representations tailored for long-context and multi-granularity semantics, making it more suitable for capturing dispersed evidence in long historical passages.

B.3.3 Dense retrieval, encoder-based.

For encoder-based dense retrieval, we adopt a family of dual-encoder models that map queries and documents to single-vector embeddings via contrastive training; some are used off-the-shelf in a zero-shot fashion, while others are further fine-tuned on our Chunqiu data.

- **GTR-T5-Base (Ni et al., 2022b) / Sentence-T5-Base (Ni et al., 2022a)**$_{(ZS)}$ are T5-based dual-encoders trained for retrieval and sentence-level semantic similarity, representing strong English dense retrievers built on sequence-to-sequence backbones.

- **mE5-Large / mE5-Large-ins**$_{(ZS)}$ (Wang et al., 2024b) are multilingual extensions of the E5 framework trained on cross-lingual relevance data, with the *ins* variant further exposed to instruction-style inputs to improve robustness to natural-language queries.

- **GTE-Large**$_{(ZS)}$ (Li et al., 2023) is a general-purpose embedding model trained with contrastive objectives over diverse relevance pairs, targeting broad semantic coverage across retrieval and similarity tasks.

- **BGE-Large-v1.5 (Xiao et al., 2023) / BGE-M3 (Chen et al., 2024)**$_{(ZS)}$ belong to the BAAI General Embedding family, combining large-scale multi-task training and support for long contexts; we use them in their dense single-vector mode as strong multilingual encoder-based baselines.

- **BERT-base**$_{(FT)}$ (Devlin et al., 2019) fine-tunes a Chinese BERT encoder on Chunqiu query–document pairs as a standard semantic dual-encoder without explicit time modeling, while our time-aware extension **CTD**$_{\text{BERT-base}}$ is detailed separately in Section 4.2.

B.3.4 Dense retrieval, LM-based embeddings.

For LM-based dense retrieval, we consider methods that leverage large language models either as contrastively trained text encoders or as generators of intermediate representations, aiming for stronger semantic and cross-task generalization with single-vector scoring.

- **GTE-Qwen2-1.5B / E5-Mistral-7B (Wang et al., 2024a)**$_{(ZS)}$ are LLM-sized embedding models that adapt Qwen2 (Team et al., 2024)

and Mistral (Jiang et al., 2023) decoders into contrastively trained text encoders, providing strong multilingual and long-context semantic representations derived from large language models.

- **PQR (Qwen2.5-7B (Team et al., 2025a) / Qwen3-8B (Team et al., 2025b))**$_{(\text{re})}$ (Kang et al., 2025) is a training-free retrieval framework in which an LLM generates potential queries for each document and these are embedded to form a mixture-style document representation; relevance is then scored by matching user queries against this distribution, illustrating an LLM-assisted, distributional document encoding paradigm.

- **Qwen3-Embed-0.6B / 4B**$_{(\text{ZS})}$ (Zhang et al., 2025b) are dedicated embedding models built on the Qwen3 family and trained with large-scale contrastive and instruction-style objectives, serving as strong multilingual LM-based dense retrievers in a zero-shot setting.

- **Qwen3-Embed-0.6B**$_{(\text{FT})}$ (Zhang et al., 2025b) further fine-tunes the 0.6B variant on Chunqiu query–document pairs as a task-adapted dense dual-encoder without explicit time modeling, isolating the effect of in-domain adaptation for LM-based embeddings.

B.3.5 Time-aware auxiliary variants.

Beyond the non-parametric BM25+TimeKDE temporal prior in the sparse family, we also report lightweight temporal extensions that can be wrapped around any single-vector dense retriever.

- **TempDate.** We add a temporal key prediction head that classifies each record into its (gong, year, month) triple, using a document-dating style auxiliary objective as in time-aware language models such as BiTimeBERT (Wang et al., 2023; Dhingra et al., 2022). The head is trained jointly with the retrieval loss, while its logits are discarded at inference time so that scoring remains purely embedding-based.

- **TempDate-Smooth.** A variant of TempDate that applies neighbor-aware label smoothing over ordered time keys, assigning small probability mass to adjacent keys (e.g., $\lambda \pm 1$) in the spirit of Temporal Label Smoothing (TLS) for early event prediction (Yèche et al., 2023).

C Broader Influence

Beyond serving as a temporal retrieval benchmark, ChunQiuTR also provides a time-keyed, multi-layer Classical Chinese text collection with rich provenance and aligned exegetical structure. Each discrete reign-month is associated with multiple textual realizations (annals, major commentaries, and later annotations), and the dataset preserves segment-level metadata (e.g., work/source/compiler, note type, and parsed spans), enabling broader computational research on pre-modern historical writing.

Time-aware question answering and grounded RAG. The time-key alignment offers a natural substrate for temporal QA and evidence-grounded generation: systems can retrieve and cite time-localized passages, compare parallel commentarial accounts, and trace how later annotations rationalize or reinterpret earlier records. This is particularly useful for end-to-end RAG settings where attribution, temporal faithfulness, and near-miss confusions (adjacent months; lexical-near notes) must be explicitly controlled.

Event understanding and timeline modeling. Because records and commentarial layers are anchored to the same discrete timeline, the collection can support event extraction and timeline construction in Classical Chinese: identifying event triggers and participants, clustering cross-source paraphrases, and modeling local causality or narrative continuity across neighboring months. The explicit `no_event` months further enable studying "silence" on historical timelines and developing models that distinguish missing reports from genuinely uneventful periods. This also opens a path to modeling evaluative language and moral judgment in historical exegesis (e.g., praise/blame framing), rather than modern sentiment in the narrow sense.

Comparative exegesis and stance/argument analysis. Parallel versions (e.g., *Zuo*, *Gongyang*, *Guliang*, plus later notes) create a controlled setting for analyzing interpretive divergence: models can be evaluated on detecting agreement vs. disagreement, identifying the argumentative function of annotations (explanation, criticism, harmonization), and attributing claims to specific traditions or compilers using the preserved provenance fields.

Information extraction with historical constraints. The structured metadata and consistent

time keys can facilitate relation and tuple extraction under temporal constraints (e.g., person–polity relations, diplomatic visits, punishments, succession-related claims), and enable consistency checks such as whether extracted relations are temporally compatible with the targeted reign-month window.

Overall, we view ChunQiuTR as a reusable, time-keyed resource for computational historiography and Classical Chinese NLP, where the same aligned corpus can support multiple tasks beyond retrieval.

D Use of AI Assistants

We used OpenAI ChatGPT as an auxiliary tool during manuscript preparation and internal implementation discussion. Specifically, it was used to (i) polish and paraphrase author-written drafts for clarity and English fluency, (ii) suggest alternative phrasings for captions and short technical descriptions based on author-provided content, and (iii) provide LaTeX formatting suggestions. All scientific claims, experimental designs, results, and final text were verified and approved by the human authors, who take full responsibility for the correctness of the paper. We did not use AI assistants to generate experimental results, fabricate data, or insert hidden content intended to manipulate automatic readers. We did not rely on AI assistants to create citations; all references were selected and checked by the authors.

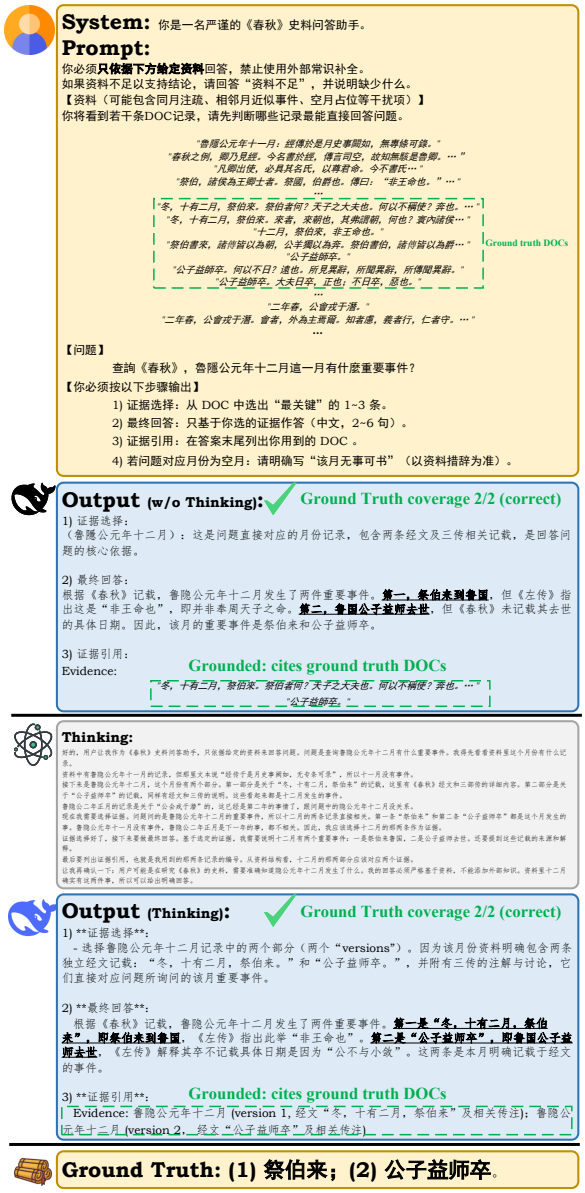

Figure 11: Evidence-bounded RAG with our annotated document pack for the same query as Fig. 10. We provide a small set of materials that includes the gold month records along with confusable supporting text, and enforce a two-step format (evidence selection + citation in the final answer). Under this constraint, the model recovers both gold entries (GT coverage 2/2) and the answer is directly supported by the cited evidence, in both without-thinking and thinking modes.

Text (excerpt + gloss)	Source	Author	Label
夏，五月，郑伯克段于鄢。 *In the fifth month of summer, the Earl of Zheng "subdued" Duan at Yan.*	《春秋》 *Chunqiu*	经文 *annals*	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。 *Gongyang focuses on why the verb* ke *("to subdue") is used and reads it as literal killing, in order to maximize moral blame on the Earl of Zheng and on Duan as a younger brother already in charge of state power.*	《春秋公羊传》 *Gongyang zhuan*	传文 *zhuan text*	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。 *Guliang interprets* ke *as "being able to subdue (kill)", stressing Duan's armed followers and his loss of proper fraternal conduct, yet still intensifying the condemnation of the Earl of Zheng's intentions.*	《春秋谷梁传》 *Guliang zhuan*	传文 *zhuan text*	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。 *Zuozhuan turns wording into a precise signal: omitting "younger brother" to blame Duan, using* ke *as if between two rulers, and explicitly naming the Earl of Zheng to mark failure of fraternal instruction, while even the omission of "fled" encodes further censure.*	《春秋左传》 *Zuo zhuan*	传文 *zhuan text*	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。 *Gu Donggao reads Zuozhuan's account that Duan fled to Gong together with a much later speech by the Earl of Zheng, arguing that Duan was not actually killed and that the Gongyang–Guliang reading of* ke *as "killed" is textually untenable.*	《春秋大事表》 *Chronological*	顾栋高 *Qing scholar*	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。 *Du Yu explains that the text does not say "Zheng (the state) punished" but names the Earl of Zheng directly, so that the wording simultaneously encodes state-level punishment and personal blame for failed instruction, while still marking Duan as a rebellious younger brother.*	《春秋左传注》 *Zuo zhuan zhu*	杜预 *Jin scholar*	neg
以"国讨""得隽曰克"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。 *Kong Yingda systematizes this case via the general norms of "state punishment" and "*ke *= gaining a notable victory", showing how the same* ke *and omission of "younger brother" are re-used across cases to redistribute blame between ruler and brother as doctrinal categories.*	《春秋左传疏》 *Zuo zhuan shu*	孔颖达 *Tang scholar*	neg

Table 14: Chrono-near interpretations aligned at the reign-based time key "鲁隐公元年五月：夏五月郑伯克段于鄢". Each row shows only a minimal classical excerpt plus an English gloss, highlighting how different traditions re-encode the same reign-based event through the verb *ke* ("subdue") and the status of Duan (younger brother vs. traitor vs. quasi-ruler). The original passages in each source are substantially longer and more layered; here we deliberately compress them into aligned snippets to expose the interpretive difficulty rather than to reproduce the full text.