Chronos: Temporal Evidence Graphs for Verifiable Web Question Answering

Anonymous Author(s)

Abstract

Answers grounded in the Web can become incorrect even when their cited pages remain accessible: facts change, pages are silently revised, and multiple sources may copy the same unsupported claim. We present *Chronos*, a framework for temporal, citation-aware Web question answering. Chronos constructs an evidence graph whose nodes are timestamped passages and whose edges encode agreement, contradiction, provenance, and near-duplicate relations. A time-conditioned selector retrieves a small evidence set, while an attribution objective requires every answer claim to be supported by a visible source. We specify a leakage-resistant evaluation protocol that separates source publication time, crawl time, and question time, and report answer correctness, citation completeness, citation precision, temporal validity, and abstention quality. A worked benchmark configuration illustrates how the protocol exposes trade-offs hidden by answer accuracy alone. The numerical results in this self-contained template are explicitly illustrative; they demonstrate reporting conventions and are not claims from a completed empirical study.

CCS Concepts

• **Information systems** → **Information retrieval**; *Web applications*; • **Computing methodologies** → *Natural language generation.*

Keywords

Web question answering, retrieval-augmented generation, temporal information retrieval, citations, provenance, evaluation

ACM Reference Format:

Anonymous Author(s). 2026. Chronos: Temporal Evidence Graphs for Verifiable Web Question Answering. In *Proceedings of the ACM Web Conference 2026 (WWW '26), June 29–July 3, 2026, Dubai, United Arab Emirates.* ACM, New York, NY, USA, 4 pages.

1 Introduction

Retrieval-augmented language models connect generated answers to external collections rather than relying only on parametric memory [4, 6]. When the collection is the open Web, this connection offers both reach and risk. A search result can be current when it is indexed but obsolete when the answer is read. Two apparently independent pages can derive from a common source, and a highly ranked page can contradict a primary source published later. A fluent answer with several hyperlinks is therefore not necessarily a verifiable answer.

Natural Questions emphasizes answer accuracy in open-domain question answering [5]. KILT studies knowledge-intensive tasks [9]. Browser-assisted systems also show that models can gather and cite Web evidence [8]. These advances do not eliminate a central evaluation ambiguity: which version of a fact should be considered correct at a specified time? A benchmark that ignores time may reward a system for retrieving information published after the question, or penalize a system for returning an answer that was correct when the question was asked.

We frame verifiable Web question answering as a temporally scoped prediction problem. Given a question q, a question time t_q, and a Web snapshot whose documents carry publication and crawl metadata, a system must produce an answer a and a set of citations C. Each atomic claim in a should be supported by at least one source that was available by t_q. When the available evidence is insufficient or contradictory, abstention is a valid output.

This paper makes four contributions:

- We formalize temporal, citation-aware Web question answering and distinguish publication, crawl, revision, and question times.
- We introduce an evidence-graph architecture that represents agreement, contradiction, provenance, and duplication before answer generation.
- We prevent using evidence published after the question and measure citation quality and selective prediction alongside answer correctness.
- We provide a fully specified worked example showing how to report data, baselines, ablations, uncertainty, limitations, and reproducibility details in an ACM Web Conference manuscript.

2 Background and Related Work

2.1 Retrieval-Augmented Question Answering

Dense passage retrieval learns question and passage representations in a shared space [4]. Retrieval-augmented generation conditions a generator on retrieved documents [6], while Fusion-in-Decoder aggregates evidence from multiple passages inside the decoder [3]. KILT evaluates retrieval and generation jointly across knowledge-intensive tasks [9]. Chronos builds on this retrieve–then–generate pattern but makes temporal eligibility and claim-level attribution first-class objects.

2.2 Web-Grounded Generation and Attribution

WebGPT showed that a model can use a text-based browser to collect references for long-form answers [8]. Self-reflective retrieval can teach a generator when to retrieve and how to critique its own output [1]. Citation presence alone, however, does not guarantee

that a cited passage entails the adjacent claim. FActScore responds to a related problem by decomposing long-form text into atomic facts and verifying them against a knowledge source [7]. We adopt atomic claims as the unit of attribution, but evaluate them against time-eligible Web evidence rather than a single static knowledge base.

2.3 Freshness, Provenance, and Duplication

Web content changes at heterogeneous rates. Some claims remain stable for years; others, such as office holders, prices, and safety notices, can change within hours. The original Web search literature already treated link structure as evidence about authority [2]. For generative systems, authority must be combined with time and provenance. Chronos therefore separates corroboration from repetition: near-identical passages connected to a shared upstream source should not count as independent support.

3 Task Definition

Let a Web snapshot be $\mathcal{D} = \{d_1, \ldots, d_n\}$. Document d_i has a publication time t_i^{pub} when available, a crawl time t_i^{crawl}, an optional revision time t_i^{rev}, and a provenance record p_i. A document is *eligible* for a question issued at t_q only if

$$t_i^{\text{avail}} = \min\left(t_i^{\text{pub}}, t_i^{\text{crawl}}\right) \le t_q, \tag{1}$$

where missing publication times default to crawl time. The benchmark retains the raw fields so alternative policies can be audited.

Write $\mathcal{K}(a) = \{k_1, \ldots, k_m\}$ for the atomic claims in answer a. A citation assignment maps each claim to zero or more documents:

$$C : \mathcal{K}(a) \longrightarrow 2^{\mathcal{D}}. \tag{2}$$

A valid prediction should satisfy three conditions:

(1) **Correctness:** each verifiable claim agrees with the reference answer scoped to t_q;
(2) **Attribution:** at least one cited passage entails each supported claim; and
(3) **Temporal validity:** no cited evidence becomes available after t_q.

Questions with no conclusive eligible evidence are labeled $y_q^{\text{ans}} = 0$. A calibrated system should abstain on these items rather than manufacture a definitive answer.

4 Chronos

Figure 1 summarizes the proposed pipeline. Its interfaces are model-agnostic: each component can be replaced without changing the temporal split or evaluation protocol.

4.1 Temporal Retrieval

Chronos first removes documents that violate Eq. (1). A hybrid retriever combines lexical score b_i, dense similarity r_i, and a freshness prior. For decay rate λ_q determined by the question type, the freshness of document d_i is

$$f_i(q) = \exp\left[-\lambda_q \max\left(0, t_q - t_i^{\text{avail}}\right)\right]. \tag{3}$$

The retrieval score is

$$s_i^{\text{ret}} = \alpha b_i + \beta r_i + \gamma f_i(q). \tag{4}$$

Stable questions use $\lambda_q \approx 0$; rapidly changing questions use a larger value. The coefficients are tuned only on the development period.

4.2 Evidence Graph Construction

For the top K passages, Chronos forms a typed graph $G_q = (V_q, E_q)$. Each node stores passage text, source domain, timestamps, retrieval scores, and source type. Edges take one of four labels: *supports*, *contradicts*, *derived-from*, or *near-duplicate*. A lightweight pairwise classifier proposes semantic edges; URL canonicalization, content hashes, and outbound links propose provenance edges. Thresholds are fixed on development data.

Let $z_i^{(0)}$ be the representation of passage i. A relation-aware update aggregates neighboring evidence:

$$z_i^{(\ell+1)} = \sigma\left(W_0^{(\ell)} z_i^{(\ell)} + \sum_{r \in \mathcal{R}} \frac{1}{|N_r(i)|} \sum_{j \in N_r(i)} W_r^{(\ell)} z_j^{(\ell)}\right). \tag{5}$$

An empty neighborhood contributes zero. Near-duplicate clusters are pooled before selection, preventing a copied claim from receiving multiple votes.

4.3 Evidence Selection and Claim Planning

The selector assigns passage i the score

$$s_i^{\text{sel}} = u^\top z_i^{(L)} + \eta s_i^{\text{ret}} + \rho a_i - \kappa c_i, \tag{6}$$

where a_i estimates source authority and c_i measures contradiction with higher-authority evidence. A greedy budgeted selector retains at most B tokens while encouraging domain diversity. A planner then predicts atomic claims and links each planned claim to selected passages. If the best support score is below threshold τ, the planner marks the claim unsupported; if a required claim is unsupported, the system abstains.

4.4 Attribution-Constrained Generation

The generator receives the question, question time, planned claims, and passage identifiers. Training minimizes

$$\mathcal{L} = \mathcal{L}_{\text{ans}} + \mu \mathcal{L}_{\text{cite}} + \nu \mathcal{L}_{\text{time}} + \xi \mathcal{L}_{\text{abstain}}, \tag{7}$$

combining answer likelihood, correct claim–citation alignment, penalties for future evidence, and calibrated abstention. At inference time, citations are emitted immediately after the claims they support. Unsupported sentences are removed rather than presented without evidence.

5 Evaluation Protocol

5.1 Worked Benchmark Construction

The worked example uses a hypothetical collection, ChronosQA, solely to demonstrate reporting structure. Table 1 defines 6,000 questions across four volatility strata. Each question includes a timestamp, a concise reference answer, atomic reference claims, eligible and future evidence, and an answerability label. Pages are canonicalized, but all versions are retained. Annotators judge claim entailment, source independence, and whether evidence available at t_q supports a unique answer.

Splits are chronological rather than random. All training questions precede development questions, which precede test questions.

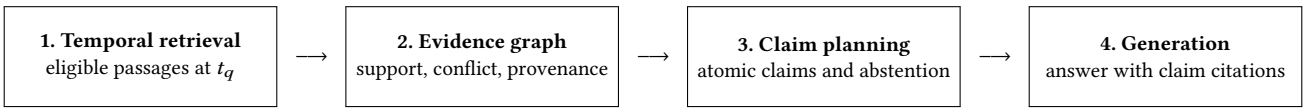

Figure 1: Chronos pipeline. Temporal filtering occurs before scoring, so evidence published after the question cannot influence retrieval, graph construction, or generation.

Table 1: Illustrative ChronosQA composition. Counts are sample data, not measurements from a released benchmark.

Volatility stratum	Train	Dev	Test
Stable facts	1,600	200	200
Annual change	1,200	150	150
Monthly change	800	100	100
Event-driven change	1,200	150	150
Total	4,800	600	600

The retrieval index for each question is reconstructed as of t_q. Pages from the same canonical URL or duplicate cluster remain within one split to reduce leakage. Annotation guidelines require two independent judgments; disagreements are adjudicated by a third annotator. A real study should report agreement, compensation, demographics relevant to the task, and institutional review requirements.

5.2 Baselines

The example compares five configurations:

- **Closed-book** answers from parametric memory without retrieval;
- **BM25+Gen** retrieves lexical matches and generates without explicit citation supervision;
- **Dense RAG** uses dense retrieval and a standard retrieval-augmented generator [6];
- **Web agent** iteratively searches and cites pages, following the browser-assisted paradigm [8]; and
- **Chronos** applies temporal filtering, evidence graphs, and attribution-constrained generation.

All retrieval systems receive the same snapshot. Generator size, maximum evidence tokens, decoding parameters, and query budget are matched where possible. Hyperparameters are selected once on the development period.

5.3 Metrics and Statistical Analysis

Answer correctness (Ans.) is the macro average of token F1 and exact match. Citation completeness (Comp.) is the fraction of verifiable output claims with at least one citation. Citation precision (Prec.) is the fraction of citations whose passages entail the associated claim. Temporal validity (Temp.) is the fraction of citations eligible under Eq. (1). Selective risk is the error rate among answered questions at a fixed coverage.

The primary utility score is declared before evaluation:

$$U = 0.40\,\text{Ans.} + 0.20\,\text{Comp.} + 0.20\,\text{Prec.} + 0.20\,\text{Temp.} \quad (8)$$

with every component scaled to $[0, 100]$. We report paired bootstrap 95% confidence intervals over questions and apply Holm correction when comparing multiple systems. Component metrics remain primary evidence because a weighted aggregate can hide qualitatively different failures.

6 Illustrative Results

Table 2 is a worked example of a complete main-results table. It intentionally emphasizes a pattern the proposed protocol is designed to reveal: a Web agent can attain strong answer accuracy and citation completeness while still using evidence that was unavailable at question time. Chronos trades a small amount of coverage for higher temporal validity and citation precision.

6.1 Ablation Reporting

Table 3 illustrates an ablation that changes one component at a time. Removing temporal filtering directly reduces temporal validity. Removing duplicate pooling primarily reduces citation precision, consistent with copied pages being mistaken for independent corroboration. Removing the abstention head increases coverage but worsens selective risk. In a real paper, these interpretations should be supported by confidence intervals and error analysis, not only point estimates.

6.2 Error Analysis

A useful error taxonomy separates failures of retrieval, temporal metadata, entailment, provenance, generation, and annotation. Three difficult cases deserve particular scrutiny. First, an undated page may be crawled after t_q even though its content existed earlier. Second, a page may be revised in place without a public version history. Third, primary sources can disagree during fast-moving events. Chronos treats these as uncertainty signals rather than forcing false precision. Real evaluations should publish category counts and representative, privacy-reviewed examples.

7 Responsible Deployment and Limitations

Temporal filtering does not establish truth. Publication time can be forged, crawl time only bounds when a crawler observed a page, and institutional sources can be wrong. Evidence graphs inherit errors from entailment and duplication models. Their authority features may also reproduce existing visibility and language inequities on the Web. A system that ranks established domains more highly may systematically under-represent local reporting or knowledge published outside high-resource languages.

The benchmark design introduces further limits. Atomic-claim annotation can erase legitimate ambiguity; chronological splits do not reproduce every future distribution shift; and exact-match reference answers are inadequate for many open-ended questions. Cost is another concern: versioned crawling, graph construction,

Table 2: Illustrative test results. Higher is better for all columns except selective risk (Risk). Values are fabricated for typesetting demonstration and must not be interpreted as empirical findings. Best values are bold.

System	Ans.	Comp.	Prec.	Temp.	Coverage	Risk↓	U
Closed-book	48.2	0.0	0.0	100.0	100.0	51.8	39.3
BM25+Gen	59.7	46.8	61.4	84.1	94.7	37.0	62.3
Dense RAG	64.5	63.1	69.8	86.7	93.5	31.0	69.8
Web agent	**68.9**	88.4	74.2	79.6	**96.2**	28.4	76.1
Chronos	68.1	**93.5**	**86.7**	**98.9**	89.4	**23.8**	**84.1**

Table 3: Illustrative Chronos ablation. ΔU is relative to the full worked-example system; values are not empirical claims.

Variant	Prec.	Temp.	ΔU
Full Chronos	86.7	98.9	0.0
– temporal filter	82.4	80.7	−5.2
– evidence graph	78.5	96.8	−3.7
– duplicate pooling	80.1	98.6	−2.2
– abstention objective	84.9	98.7	−1.4

and human verification are substantially more expensive than static retrieval.

Deployments should expose timestamps and quoted support, preserve the user's ability to inspect original pages, and communicate when sources conflict. High-stakes domains require qualified human review. Crawling must respect access controls, terms, privacy, and deletion obligations. Benchmark releases should avoid redistributing sensitive content and should publish a documented removal process.

8 Reproducibility Checklist

A complete study should archive or document the following:

- immutable question, document-version, and annotation identifiers;
- publication, revision, crawl, and question timestamps with time zones;
- rules for canonicalization, duplicate detection, and temporal eligibility;
- chronological split manifests and leakage audits;
- prompts, model versions, retrieval indexes, query budgets, seeds, and decoding parameters;
- annotation instructions, agreement, adjudication, and compensation;
- per-question predictions and bootstrap code for every reported metric; and
- compute, latency, API, and financial costs sufficient to compare system efficiency.

Because live pages change, reproducing a Web experiment requires more than releasing code. Where redistribution is lawful, content-addressed snapshots provide the strongest audit trail. Otherwise, a release should include hashes, timestamps, URLs, extraction rules, and clear instructions for reconstructing the closest permissible snapshot.

9 Conclusion

Chronos reframes Web-grounded question answering around three linked questions: Was the evidence available at the relevant time? Does it support the generated claim? Is apparently corroborating evidence genuinely independent? Temporal filtering, typed evidence graphs, claim-level citation constraints, and calibrated abstention provide a concrete way to address these questions. More importantly, the proposed evaluation protocol makes failures visible that answer accuracy alone can conceal. The worked results in this template are illustrative, but the task definition, reporting structure, and reproducibility checklist can be used directly to design and document a real Web Conference study.

References

[1] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*.

[2] Sergey Brin and Lawrence Page. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems* 30, 1–7 (1998), 107–117.

[3] Gautier Izacard and Edouard Grave. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 874–880.

[4] Vladimir Karpukhin, Barlas Oğuz, 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)*. 6769–6781.

[5] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics* 7 (2019), 453–466.

[6] 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*, Vol. 33. 9459–9474.

[7] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-Grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 12076–12100.

[8] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. WebGPT: Browser-Assisted Question-Answering with Human Feedback. *arXiv preprint arXiv:2112.09332* (2021).

[9] Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: A Benchmark for Knowledge Intensive Language Tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2523–2544.