Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Auditable Multi-LLM Evidence Synthesis for Contested Heritage Origins: Comparing Competing Hypotheses in the Turpan Karez Debate

Xiangping Qi
Ningxia University
Muhammad Irfan Malik
Tsinghua University
Zhongjing Wang
z.j.wang@tsinghua.edu.cn

Ningxia University
Jingping Zhang
Lanzhou University
Shaolan Wang
Tsinghua University
Hongjian Gan
Tsinghua University
Kanghui Li
Ningxia University
Yixuan Shi
Ningxia University

Article

Keywords: Retrieval-augmented generation, multi-LLM, Karez, Origin of cultural heritage

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11469328/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Auditable Multi-LLM Evidence Synthesis for Contested Heritage Origins: Comparing Competing Hypotheses in the Turpan Karez Debate

Xiangping Qi[1], Muhammad Irfan Malik[2,3], Zhongjing Wang[1,2*], Jingping Zhang[4], Shaolan Wang[5], Hongjian Gan[2], Kanghui Li[1], Yixuan Shi[1]

1. School of Civil and Hydraulic Engineering, Ningxia University, Yinchuan, 750021, China,
2. School of Civil Engineering, Tsinghua University, Beijing 100084, China.
3. Department of Water Resources Management, Lasbela University of Agriculture, Water, and Marine Sciences, Uthal 90150, Pakistan.
4. School of History and Culture, Lanzhou University, Lanzhou 730000, China
5. Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

Xiangping Qi and Muhammad Irfan Malik contributed equally to this work

Corresponding author. E-mail: zj.wang@tsinghua.edu.cn

Abstract

The origin of Karez in Turpan, China remains contested among three hypotheses: Central Asia Introduced (CAI), Central Plains Introduced (CPI), and Local People Innovated (LPI). Fragmented multilingual literature, circular citation practices, and inconsistent evidentiary standards hinder this debate. An auditable multilingual, multi-LLM evidence-synthesis hypothesis evaluation framework was proposed, comprising progressive literature retrieval, RAG-based candidate-passage localization, parallel multi-LLM evidence discovery, source-constrained validation, Evidence Family consolidation, and robustness analysis. From 275 full-text publications, 141 text chunks were retrieved and 727 candidate evidence records generated by DeepSeek V4 Pro, GLM-5.2, and Qwen3.7-Plus. Integrating evidence sufficiency, net evidence balance, scholarly consensus, and evidence-type diversity yielded scores of 0.5735, 0.5125, and 0.3321 for CAI, LPI, and CPI, respectively. CAI ranked first in 82.6% of 10,000 family-level bootstrap iterations. The framework enhances the traceability and auditability of historical evidence synthesis and offers a new paradigm for investigating heritage origins where decisive evidence is unavailable.

Keywords: Retrieval-augmented generation, multi-LLM, Karez, Origin of cultural heritage

Introduction

The history of water-resource development and utilization provides an important perspective for understanding the connection between human societies and their environments[1]. In arid regions, karez is a groundwater-utilization technology that has persisted for millennia and remains in use across many parts of the world. A typical karez system comprises a mother well, underground channels, vertical shafts, an outlet, and surface open channels[2] (Fig. 1a). Underground channels convey groundwater from aquifers in the mountains to lower plains for domestic water supply and agricultural irrigation[3]. This technology has played a vital role in the formation and development of oases and has become a distinctive cultural symbol of arid regions[4].

The origin and diffusion of karez technology have long been debated. The prevailing view holds that underground water-collection tunnels first emerged around 1000 BCE on the Iranian Plateau and in adjacent parts of West Asia and the Arabian Peninsula[5,6]. These processes produced a global distribution spanning Asia, Africa, Europe, and the Americas, with the Iranian Plateau serving as a major early center and diffusion hub (Fig. 1b)[5].

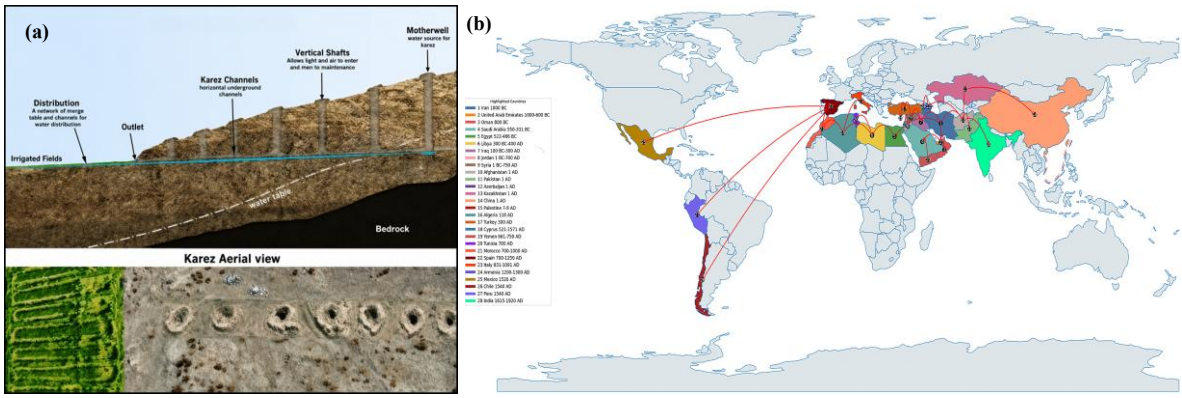


Fig. 1 | Karez and its global distribution. (a) Typical Karez system. **(b)** Global distribution and hypothesized diffusion routes.

Against this background of global diffusion and regional evolution, the Turpan Basin provides a representative setting for investigating the eastward transmission and local adaptation of karez technology because of its distinctive natural environment and strategic location[7]. Situated in eastern Xinjiang, Turpan connects the Central Plains with Central and West Asia. The basin is surrounded by high mountains and has a low, relatively flat interior with extensive piedmont alluvial–proluvial fans. Its surface gradients and aquifer structure facilitate gravity-driven groundwater conveyance[8,9] (Fig. 2a–c). Remote-sensing data from the past five years indicate that mean annual evapotranspiration ranges from 365.0 to 1,420.9 mm, whereas mean annual precipitation ranges from 9.5 to 402.8 mm[10]. These data reveal an arid, low-precipitation basin interior surrounded by comparatively wetter mountainous areas[8,10]. Existing karez systems are concentrated primarily along the northern mountain front, forming an east–west distribution across Toksun, Gaochang, and Shanshan, with local clusters around towns and traditional agricultural areas. They are comparatively sparse in the southern basin and at higher elevations[9] (Fig. 2d, e). This spatial pattern is closely associated with mountain water recharge, groundwater flow, suitable terrain gradients, and water demand from oasis agriculture[4,8,9].

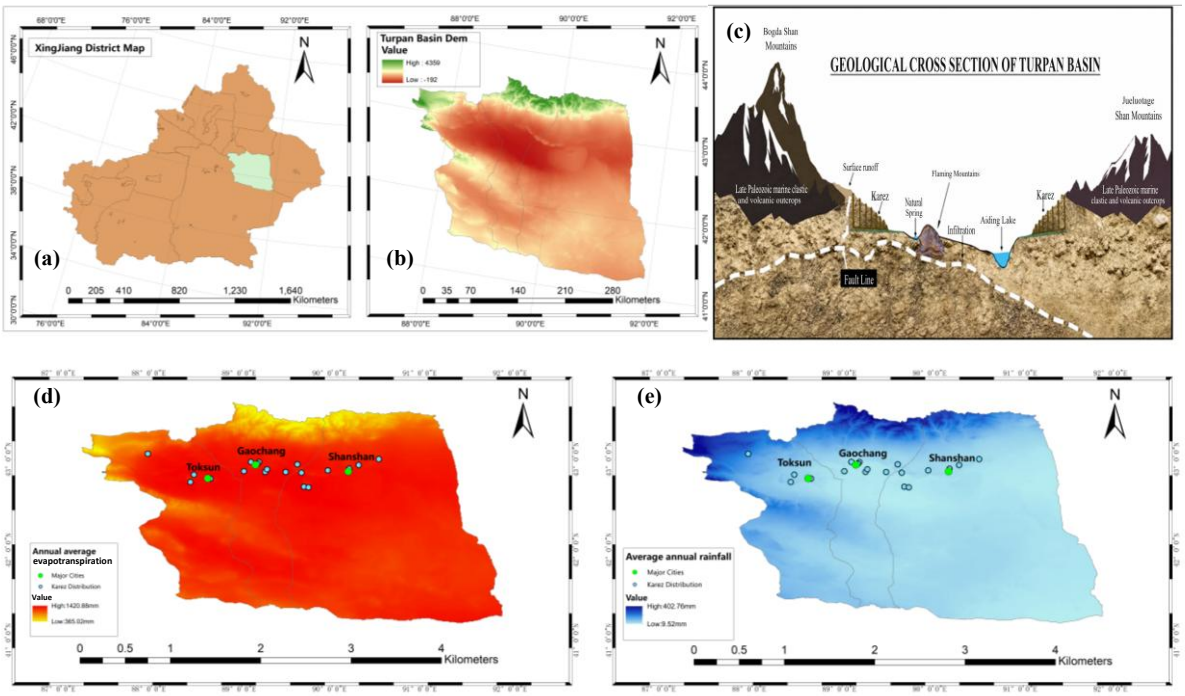


Fig. 2 | Environmental setting of the Turpan Basin. (a) Location of Turpan Basin. **(b)** Elevation of Turpan. **(c)** Hydrogeological section. **(d)** Mean annual evapotranspiration. **(e)** Mean annual precipitation and karez distribution.

Turpan's location at an important crossroads suggests that its karez systems may be associated with the eastward diffusion of underground water-conveyance technology from West and Central Asia, while also reflecting influences from the Central Plains well-channel tradition and local hydraulic practices[5,11,12]. Consequently, broader debates over single-origin models, technological diffusion, and polycentric development of karez systems are particularly complex in Turpan. The initial construction date, technological origins, transmission processes, and local adaptation remain central points of contention regarding Xinjiang karez systems[7,11,13]. Scholars have long debated the origin of the Turpan karez; however, no golden evidence has been identified; that is, there is no single piece of evidence that can provide either a "one-vote veto" or "one-vote confirmation"[11]. The available evidence neither decisively rules out competing explanations nor independently confirms any single interpretation. Accordingly, the debate centers on three competing hypotheses: Central Plains Introduced (CPI), Central Asia Introduced (CAI) and Local People Innovated (LPI) become the debate center.

The CPI hypothesis can be traced to Tao Baolian (1862-1938), a historical geographer of the late Qing Dynasty, in his book *Xinmao Shixing Ji*[14] (《辛卯侍行记》Travel Notes of the Xinmao Year) published in 1897, in which Tao linked the Turpan karez to the well-channel project constructed during the Western Han Dynasty. The CPI hypothesis was formally proposed by Wang Guowei (1877–1927), a leading Chinese historian of the late Qing and early Republican eras, in his treatise *Xiyu Jingqu Kao*[15] (《西域井渠考》, A Study on the Canal-Wells in the Western Regions), published in 1919. In this treatise, Wang argued that the Turpan karez originated in the Central Plains and bore structural similarities to the Longshou Canal of Shaanxi Province built in the Han Dynasty. His view was subsequently supported by numerous scholars in subsequent studies[16,17].

The CAI hypothesis was first proposed by Huntington (1876–1947), an American geographer. In his 1907 travelogue, which drew largely on local oral narratives, Huntington inferred that Xinjiang karez technology had been introduced from Persia or the Trans-Caspian region around 1780[18]. Stein (1862–1943), a Hungarian-born British archaeologist and explorer, endorsed this viewpoint and linked Xinjiang karez to the Persian-Central Asian tradition. In his 1933 book, a comprehensive report on his three Asian expeditions, he argued that the technology had arrived in Turpan no earlier than the 18th century[19]. Huang[20] and Yang[21] similarly supported the CAI hypothesis, whose core arguments rest on parallels in engineering structure and operational principles, alongside etymological evidence.

Unlike the CPI and CAI hypotheses, the LPI hypothesis was not formally advanced by any individual or institution but emerged gradually from accumulating evidence. The Xinjiang Tuzhi (《新疆图志》, Illustrated Gazetteer of Xinjiang), published in 1911, noted that "this method is sometimes said to have originated under Emperor Wu of the Han Dynasty … yet the method had already existed in the Western Regions."[22] Liu Hongliang (1986) further attributed the Qing-era expansion of karez chiefly to local communities' adoption and adaptation of indigenous techniques[23]. Citing the piedmont alluvial–proluvial fans, groundwater recharge, topographic gradients, and agricultural water demand, Zhang Xiru and co-workers held that Turpan met the environmental prerequisites for karez to emerge locally[24]. While the Panjier Mountain petroglyphs have been interpreted as early karez depictions, their date and cultural context remain archaeologically unresolved[25].

However, none of the three hypotheses possesses definitive supporting evidence. The ongoing debate is constrained not merely by the scarcity of direct archaeological finds, but also by fragmented evidence and reliance on heterogeneous sources[7,12]. Relevant records are scattered across multilingual, multidisciplinary literature, including Chinese, English, Russian, French, Arabic, and other languages. Earlier scholarship frequently drew upon limited source materials, in which researchers' own arguments, paraphrases of prior viewpoints, and primary historical evidence are often intertwined. A single archaeological find or historical document may also be repeatedly cited across numerous publications, such that the number of publications does not correspond to the number of independent lines of evidence. Accordingly, consistent evaluation of these competing origin hypotheses, with respect to evidentiary basis, scholarly support and robustness, remains challenging. Large-scale manual literature screening, stance classification and evidence extraction are further hampered by issues of efficiency, consistency and reproducibility[26].

Against such methodological constraints, existing computational tools have partially eased the workload of large-scale literature processing. CiteSpace and VOSviewer map research trends and knowledge structures via co-citation analysis, keyword co-occurrence analysis, and cluster analysis[27,28]. Latent Dirichlet allocation {LDA} enables large-scale text clustering and topic detection[29], whereas tools such as Abstrackr and ASReview facilitate literature screening for systematic reviews[26,30]. However, these approaches mainly serve document retrieval, sorting and filtering; they are poorly equipped to extract substantive scholarly claims and underlying evidence, or to clarify how such evidence relates to competing hypotheses. Subsequent advances in natural language processing (NLP) extended scientific text analysis beyond document-level bibliometric organization. SciBERT generates domain-specific semantic

representations for academic texts[31], SciCite performs citation-intent classification[32], and SciFact supports the validation of scientific claims against source documents[33]. Even so, these models are trained and validated on task-specific annotated datasets, mostly in English, limiting their direct applicability across languages and disciplines, particularly for heterogeneous historical corpora. Furthermore, they lack dedicated functions to separate researchers' original interpretations from cited viewpoints and critical citations.

Large language models (LLMs) open new avenues for extracting structured evidence from intricate texts and have seen growing adoption in literature screening, data extraction and evidence synthesis for systematic reviews, as exemplified by TrialMind and Evidence Triangulator[34]. Still, parametric knowledge embedded within LLMs is limited by weak source traceability, temporal constraints, and the risk of hallucinating content unsupported by source texts[35]. Retrieval-augmented generation (RAG) mitigates this issue by grounding model judgements in retrieved source-text excerpts[36], and platforms such as OpenScholar have demonstrated the feasibility of large-scale scientific literature retrieval and evidence synthesis[35]. Even with RAG, challenges persist: incomplete corpora, OCR artifacts, contextual degradation during text chunking, and retrieval inaccuracies. Hence, the retrieved snippets still require independent validation to confirm whether they constitute valid supporting evidence.

Accordingly, this study developed a human–LLM collaborative framework tailored for contested historical questions. This framework integrates multilingual literature collection, bibliometric screening, RAG-driven evidence retrieval, multi-model structured information extraction, and human verification. We compiled a multilingual karez corpus covering Chinese, English, Russian, French, Persian and Arabic sources. RAG was deployed to retrieve candidate evidence passages, after which multiple LLMs extracted key information, including claimed origin locations, chronological estimates, evidentiary statements, and the relationships between evidence and competing hypotheses. Source-text cross-checking and human expert review were implemented to govern evidence inclusion and attribution. This framework was designed to build a searchable, comparable and traceable evidence repository to evaluate the three competing hypotheses regarding the origin of the Turpan karez.

Methods

We developed a multilingual, traceable, multi-model evidence-synthesis framework for investigating the origin of the Turpan karez (Fig. 3). First, a full-text corpus was established through a three-stage progressive multilingual search, with study identification and screening conducted and reported in accordance with PRISMA 2020[37]. The full texts were then extracted, cleaned, and divided into hierarchical text chunks. BM25[38] and TF-IDF[39] retrieval results were integrated using reciprocal rank fusion (RRF)[40] to produce a traceable and cost-controlled candidate-passage set. Under identical input conditions, three LLMs from different model families independently extracted the proposed place and date of origin, evidentiary content, authorial stance, and provenance information[34,41]. The candidate outputs were merged and deduplicated, verified against the source text across models, and subjected to tiered human review to produce the validated evidence set.

The verified evidence records were subsequently consolidated into Evidence Families based on their shared primary sources, historical events, and citation lineages, thereby

controlling evidence inflation caused by repeated citations. Competing origin hypotheses were then compared across four dimensions: evidence sufficiency, net evidence balance, scholarly consensus, and evidence-type diversity. Ranking robustness was evaluated through threshold perturbation, leave-one-model-out analysis, weight perturbation, evidence-deletion tests, and family-level bootstrap analysis[42,43]. All analytical outputs retained provenance information, including the original publication, page number, text chunk, and Evidence Family, and were used to support hypothesis comparisons and to construct an evidence timeline and evidence graph.

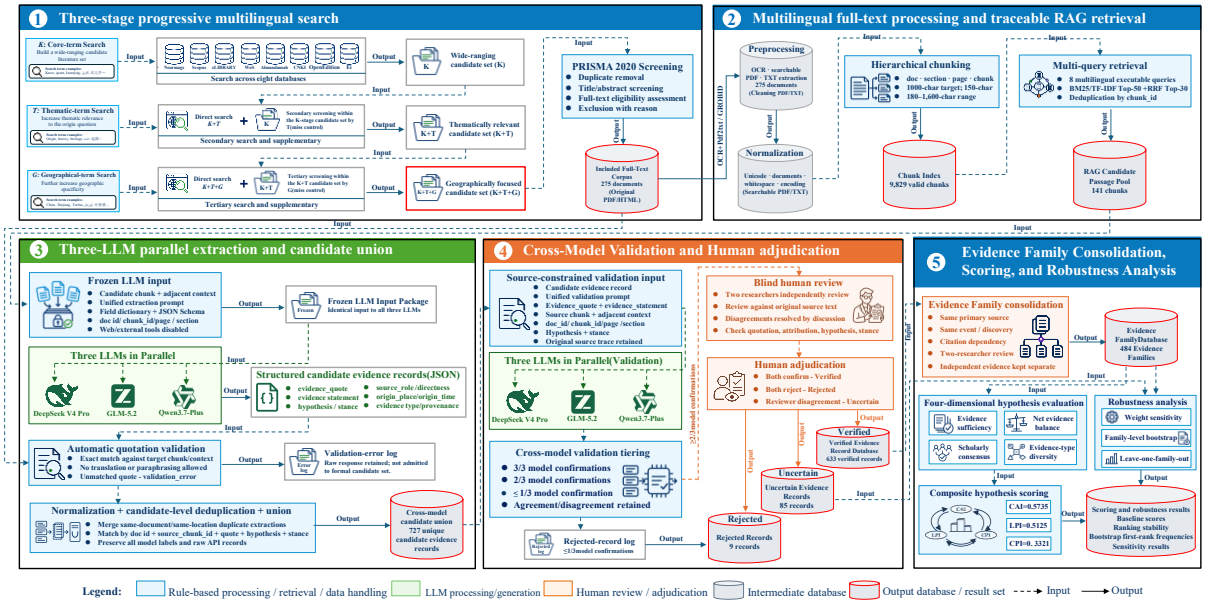


Fig. 3 | Overall workflow of the auditable multi-LLM evidence synthesis framework.

Multilingual corpus construction

We conducted a three-stage progressive multilingual search across eight databases: Web of Science Core Collection, Scopus, Ei Compendex, CNKI, eLIBRARY.RU, OpenEdition, Almandumah and Noormags. After candidate records were retrieved, deduplication, screening, full-text retrieval, and eligibility assessment were performed in accordance with the PRISMA 2020 framework. We defined K as the core-term set comprising qanat, karez, 坎儿井, and their multilingual variants; T as the thematic-term set comprising terms related to origin, history, antiquity, heritage, and diffusion; and G as the geographical-term set comprising China, Xinjiang, Turpan, Shanshan, and the Western Regions. Terms within each set were combined using the Boolean operator "OR", whereas the three sets were combined using "AND". The search strategies are presented in Table 1 and the three-stage search strategy is formulated in Eq. (1):

$$Q_1 = K,\ Q_2 = K \wedge T,\ Q_3 = K \wedge T \wedge G \tag{1}$$

The three stages served different retrieval functions rather than constituting a single progressively restrictive query path. The first-stage K search established a broad candidate literature set. At the second stage, direct $K+T$ database searches were supplemented by thematic screening of the broader K-stage candidate set, thereby retaining potentially relevant records that might not have been retrieved by the more restrictive combined query. At the third stage,

direct *K+T+G* searches were similarly supplemented by geographical screening of the preceding candidate set. The resulting geographically focused candidate set was then subjected to PRISMA 2020 screening. This staged strategy was intended to progressively increase thematic and geographical specificity while reducing omission associated with reliance on a single high-constraint *K+T+G* query. Here, "progressive" denotes increasing thematic and geographical specificity rather than the automatic exclusion of records not returned by subsequent higher-constraint queries.

Records retained through the direct-search and supplementary-screening paths were consolidated at each stage, and the final geographically focused candidate set was merged across databases and deduplicated using DOIs, database accession numbers, normalized titles, authors, and publication years. Potential cross-linguistic duplicates were further reviewed by comparing translated titles, transliteration variants, and semantic title similarity. Publications that passed full-text eligibility assessment, together with their normalized metadata, constituted the traceable RAG corpus used in the subsequent analyses.

Table 1 | Database-specific implementation of the multilingual three-stage progressive search strategy

Platform	Coverage	Stage 1: system-term block	Stage 2: thematic block	Stage 3: geographic block
Scopus	1999-2026	TITLE-ABS-KEY(qanat OR karez OR foggara OR khettara OR falaj OR aflaj OR kanerjing)	AND (basin OR ancient OR origin OR history OR heritage)	AND (China OR Xinjiang OR Turfan OR Turpan)
Web of Science	2000-2026	TS=(qanat OR karez OR foggara OR khettara OR falaj OR aflaj OR kanerjing)	AND TS=(basin OR ancient OR origin OR history OR heritage)	AND TS=(China OR Xinjiang OR Turfan OR Turpan)
Ei Compendex	1884-2026	(qanat OR karez OR foggara OR khettara OR falaj OR aflaj OR kanerjing) WN KY	AND (basin OR ancient OR origin OR history OR heritage) WN KY	AND (Asian OR China OR Xinjiang OR Turfan OR Turpan)WN KY
CNKI	1959-2026	SU=‘坎儿井’ OR SU=‘Karez’	AND SU=('文化遗产' OR '丝绸之路' OR '起源' OR '井渠' OR '来源')	AND SU=(‘中亚’ OR ‘'吐鲁番' OR '鄯善县' OR '西域' OR '亚尔村'OR'新疆')
eLIBRARY.RU	1989-2026	кяриз OR кяризы OR кариз OR каризы OR кяризная система OR karez OR qanat	AND (бассейн OR древний OR происхождение OR история OR наследие)	AND(Синьцзян OR Турфан OR Китай)
OpenEdition Search	1900-2026	karez OR qanat OR foggara OR khettara OR falaj OR aflaj OR kanerjing	AND (basin OR ancient OR origin OR history OR heritage)	AND (Asian OR Turfan OR Tourfan OR Turpan OR Xinjiang OR "Turkestan oriental")
Almandumah	1968-2026	كاريز OR كهريز OR قناة جوفية OR الأفلاج OR الفقارة OR القنوات الجوفية OR karez OR qanat	AND (الأصل OR المنشأ OR النشأة OR التأريخ OR الانتقال OR الانتشار OR التاريخ OR "علم الآثار")	AND(تورفانOR توربان OR شينجيانغ OR "تركستان الشرقية")
Noormags	1913-2026	"كاريز" OR "قنات" OR "كهريز" OR karez OR qanat	AND ("منشأ" OR "خاستگاه" OR "پيدايش"OR "تاريخ" OR "باستانى" OR "ميراث"OR "گسترش" OR "انتقال")	AND ("تورپان" OR "تورفان" OR "تركستان شرقى" OR "سين‌كيانگ" OR "چين")

RAG retrieval and evidence extraction

Native text was preferentially extracted from machine-readable PDF files using pdf2txt parsing tools. For PDFs requiring structured document parsing, GROBID was used to extract textual content and document structure, whereas optical character recognition (OCR) was applied to pages without a usable text layer or with text of insufficient quality[41,44]. The resulting text outputs were converted into machine-readable TXT representations. During preprocessing, repeated headers and footers, anomalous line breaks, invisible characters, and clearly identifiable OCR noise were removed without rewriting the source text. The extracted texts were subsequently normalized with respect to Unicode representation, whitespace, and character encoding. Original files, OCR outputs, tool versions, and processing logs were retained to preserve traceability. The normalized texts then entered hierarchical chunking for subsequent retrieval.

Full texts were divided into chunks following a hierarchical publication–section–page–chunk structure while retaining provenance metadata, including the document identifier, section, page number, language, and file path. The target chunk length was 1,000 characters, with an overlap of 150 characters between adjacent chunks and minimum and maximum lengths of 180 and 1,600 characters, respectively. Residual text shorter than 180 characters at the end of a section or document was retained when it could not be merged with the preceding chunk. Chunk boundaries were preferentially placed at the ends of sentences or paragraphs. Eight retrieval targets were defined to cover origin-related claims and evidence for the CPI, CAI, and LPI hypotheses, together with dating evidence and authorial stance. One multilingual query was constructed for each target by combining Chinese and English search terms, together with additional language variants where relevant, into a single query string. This produced eight executable queries in total. The complete query strings, applicable hypotheses, query identifiers, and version information were retained as part of the audit data[45].

The relevance of query q to chunk d was calculated independently using BM25 and term frequency–inverse document frequency (TF–IDF)[39]. The BM25 score is defined in Eq. (2), and the inverse document frequency of each term is defined in Eq. (3).

$$S_{\mathrm{BM25}}(q,d) = \sum_{t \in q} \mathrm{IDF}(t) \frac{f(t,d)(k_1+1)}{f(t,d) + k_1\left(1 - b + b\frac{|d|}{\overline{dl}}\right)} \quad (2)$$

$$\mathrm{IDF}(t) = \ln\left(1 + \frac{N - n_t + 0.5}{n_t + 0.5}\right) \quad (3)$$

Here, q denotes the query, d the text chunk, and t a query term. $f(t,d)$ is the frequency of term t in chunk d, d is the length of the chunk, and $|d|$ is the mean length of all chunks. The term-frequency saturation parameter k_1 was set to 1.5, and the document-length normalization parameter b was set to 0.75[38]. $\mathrm{IDF}(t)$ denotes the inverse document frequency of term t.

In parallel with BM25, a TF–IDF lexical representation was constructed[39]. Non-zero term frequencies were logarithmically scaled to reduce the disproportionate influence of frequently repeated technical terms and duplicated text introduced by OCR. The TF–IDF weight of term t in chunk d is defined in Eq. (4).

$$w_{t,d} = \mathrm{TF}(t,d) \times \left[\ln\left(\frac{N+1}{n_t+1}\right) + 1\right] \quad (4)$$

Logarithmic term-frequency scaling was applied when term t occurred at least once in chunk d; otherwise, its weight was set to zero. After L2 normalization of the TF–IDF vectors, the relevance between a query and a chunk was calculated using cosine similarity[39], as defined in Eq. (5).

$$S_{\mathrm{TF-IDF}}(q,d) = \frac{v_q \cdot v_d}{\|v_q\|\|v_d\|} \quad (5)$$

Here, the relevant terms denote the TF–IDF vectors of query q and chunk d and their Euclidean norms. Cosine similarity was defined as zero when either vector was a zero vector.

For each query, the k highest-scoring chunks were retained independently by BM25 and TF–IDF. The Top-k results returned by an individual retriever are defined in Eq. (6).

$$L_a^{(k)}(q) = \mathrm{TopK}_{d\in D}(S_a(q,d),k) \quad a \in \{\mathrm{BM25}, \mathrm{TF-IDF}\} \quad (6)$$

Here, a denotes the retriever, D is the complete set of text chunks, and $S_a(q,d)$ is the relevance score assigned by retriever a to query q and chunk d. Each retriever returns $k = 50$ results.

Because raw BM25 and TF–IDF scores are not directly comparable, reciprocal rank fusion (RRF) was used to integrate and rerank the two result sets[40]. The RRF score is defined in Eq. (7).

$$S_{\mathrm{RRF}}(d,q) = \sum_{a:\in L_a^{(k)}(q)} \frac{1}{c + r_a(d,q)} \quad (7)$$

A retriever contributed to the RRF score only when the chunk appeared among its Top-k results; otherwise, its contribution was set to zero. The resulting RRF score was determined jointly by the rank of each chunk across the two retrievers, with the rank-smoothing constant set to 60[40].

For each query, the 30 highest-ranked chunks after RRF were retained as candidate passages, as defined in Eq. (8).

$$C(q) = TopK_d(S_{\mathrm{RRF}}(d,q), K_{\mathrm{RRF}}) \quad (8)$$

The resulting set represents the RRF-ranked candidate chunks associated with query q. The post-fusion retention threshold was 30 chunks, and the RRF scores were used to determine their order.

Results from the eight multilingual queries were subsequently combined and deduplicated by chunk_id, as defined in Eq. (9).

$$S_{sum}(d,t) = \sum_{q\in Q_t} \mathbb{I}[d \in C(q)] S_{\mathrm{RRF}}(d \mid q) \quad (9)$$

This procedure produced a traceable candidate-passage set for subsequent parallel multi-model extraction and cross-model validation. The complete query set, retrieval parameters, and candidate results from every stage were frozen and retained to support reproducible reruns and audits of retrieval errors[41,44].

Parallel Evidence Extraction and Aggregation

Candidate passages retrieved through RRF were independently processed for evidence extraction using DeepSeek V4 Pro, GLM-5.2, and Qwen3.7-Plus. Each model received identical target chunks, document metadata, section headings, retrieval intents, field definitions, label schemes, and task instructions, and operated without access to the outputs of the other models. To restore context potentially truncated during text chunking, each target chunk was accompanied by one valid preceding and one valid following chunk from the same document[36,46]; at document boundaries, only available adjacent content was provided. Adjacent chunks served solely as contextual information and were not treated as additional processing units. Each model was required to identify the actual location of the extracted evidence using source_chunk_id.

The models were instructed to identify evidence only from the supplied text concerning the origin, dating, transmission routes, technological sources, or origin hypotheses of karez systems in Turpan or Xinjiang. They were prohibited from using the internet, external tools, or parametric knowledge encoded in model parameters to supplement information not explicitly stated in the source text[34,47]. Passages concerning only general hydraulic engineering, modern conservation, ecological conditions, or management policies were labeled has_relevant_evidence = false.The structured output fields included input_id, target_chunk_id, evidence_quote, evidence_statement, source_chunk_id, origin_place, origin_time, hypothesis, stance, source_role, directness, the primary evidence category and subtype, original source, and certainty. The hypothesis field comprised CAI, CPI, LPI, MULTI, and NONE; stance comprised support, oppose, qualified_support, neutral, uncertain, and not_applicable; and source_role distinguished primary materials (primary_material), reported views (reported_view), authorial inferences (author_inference), and explicit authorial conclusions (author_conclusion). Evidence was classified into three predefined primary categories—material, documentary, and legendary evidence—following Huang Xianfan's "Triple Evidence Method," an extension of Wang Guowei's Dual Evidence Method[48,49].

All evidence_quote values were required to be copied verbatim from the target chunk or its adjacent context. Translation, paraphrasing, correction of OCR errors, and reconstruction of missing content were prohibited. After model inference, each quotation was mechanically matched against the source text. Quotations that could not be located were labelled validation_error, and the original responses were retained for auditing but excluded from the formal candidate evidence set.

All three models used the same frozen evidence-extraction prompt, field dictionary, and JSON Schema; only the API wrapper differed among providers. The temperature was set to 0.0, Top-p was left at the provider default, the maximum output length was 3,000 tokens, and the request timeout was 180 s. Failed requests were retried up to three times using exponential backoff, with external tools and internet access disabled (Table 2). For each call, we recorded the experiment version, actual model ID, input identifier, target_chunk_id, prompt and schema versions, generation parameters, token usage, response time, raw response, parsed output, and error logs. Results were written incrementally, and checkpoint-based resumption was implemented to ensure that the experiments were recoverable and auditable.

Table 2 | Configuration parameters of the three large language models used for parallel evidence extraction.

Model ID	Provider	Temperature	Top-p	Max tokens	Output format	Web/External tools	API timeout	API retries
DeepSeek V4 Pro	DeepSeek	0.0	Default	3000	JSON	Disabled	180s	3 times
GLM-5.2	Zhipu AI	0.0	Default	3000	JSON	Disabled	180s	3 times
Qwen3.7-Plus	Alibaba	0.0	Default	3000	JSON	Disabled	180s	3 times

Let E_m denote the set of candidate evidence records extracted by model m. The union of candidate records is defined in Eq. (10).

$$E_{candidate} = \bigcup_{m=1}^{3} E_m \quad (10)$$

The candidate union aggregated evidence independently identified by all models. A union rather than an intersection was used at this stage to retain as much potentially relevant evidence as possible. Both agreement and disagreement among the models were preserved for subsequent cross-model verification against the source text. The number of models identifying an evidence record was used neither to determine its validity nor to establish the final ranking of the three origin hypotheses.

Cross-model source-text validation and standardized evidence inclusion

The candidate union was first subjected to field normalization and candidate-level deduplication. Duplicate extractions from the same publication and text location were normalized and merged when their quotations and core meanings were consistent. Matching was based on the document identifier, source_chunk_id, verbatim quotation, hypothesis, and stance, while all model labels and original API-call records were retained. Repeated accounts of the same historical source across different publications were not removed at this stage but were subsequently consolidated into Evidence Families.

A second round of blind, source-constrained cross-validation was then conducted using the same three LLMs employed during evidence extraction. Each model received the normalized candidate evidence record, corresponding source passage, section heading, page number, and necessary adjacent context, but was not shown the identity of the model that identified the candidate or the assessments made by other models. Based exclusively on the supplied source text, each candidate was classified as supported by the source (entailed), contradicted by the source (contradicted), or insufficiently supported (insufficient). The models also verified the proposed place and date of origin, evidence quotation, authorial stance, and hypothesis assignment. When a field was incorrect, the model was required to return both a corrected value and the supporting source text. Candidates for which no supporting passage could be located were not admitted to the validated evidence set.

Let y_{im} denote the validation result assigned to candidate evidence record by model m, taking a value of 1 when the model classifies the candidate as entailed and 0 otherwise. The number of valid model confirmations received by candidate e_i is defined as:

$$V_i = \sum_{m=1}^{3} y_{im} \tag{11}$$

where i denotes the candidate-evidence index, I is an indicator function. V_i represents the number of valid model confirmations received by candidate evidence record.

A traceability indicator, T_i, was also defined. $T_i = 1$ when candidate evidence record could be traced to a valid document identifier, page number, chunk identifier, and verbatim source quotation; otherwise, $T_i = 0$. The assessments of the two researchers were denoted by z_{i1} and z_{i2}, respectively, where $z_{ij} = 1$ indicates that Researcher j confirmed that the candidate evidence was supported by the source text, and $z_{ij} = 0$ indicates rejection. The standardized evidence-inclusion rule is defined in Eq. (12).

$$status(e_i)\begin{cases} Verified,\ T_i = 1 \wedge V_i \geq 2 \wedge z_{i1} = z_{i2} = 1, \\ Rejected,\ T_i = 0 \vee V_i \leq 1, \\ Rejected,\ T_i = 1 \wedge V_i \leq 2 \wedge z_{i1} = z_{i2} = 0, \\ Uncertain,\ T_i = 1 \wedge V_i \geq 2 \wedge z_{i1} \neq z_{i2}. \end{cases} \tag{12}$$

Under this rule, candidate evidence records were classified as *Verified*, *Rejected*, or *Uncertain*. Verified records were included in the formal evidence set, Rejected records were excluded, and Uncertain records were retained only for sensitivity analyses. Records that met the traceability requirements and received confirmation from all three models were reviewed against the source text by Researcher 1 and subsequently checked by Researcher 2. Disputed records receiving confirmation from two models were independently reviewed by both researchers. Records confirmed by both researchers were included in the formal evidence set, those rejected by both were excluded, and records for which their assessments differed were labelled *Uncertain*. Model consensus was used solely to triage records for review and was not interpreted as the probability that the evidence was authentic or that the historical claim was true. All model assessments, human adjudications, and field revisions were independently versioned and retained.

Evidence records that passed review were further consolidated into Evidence Families based on shared primary sources, historical events, archaeological discoveries, and citation dependencies. Records from multiple publications were assigned to the same Evidence Family when they cited the same primary historical source, described the same site or archaeological discovery, or merely repeated an earlier study without contributing new independent material. Separate families were established when records were based on distinct primary materials, independent dating results, or additional empirical evidence. The two researchers independently assigned family membership using normalized source names, verbatim quotations, event locations and dates, bibliographic relationships, and citation lineages. Disagreements were resolved through discussion. Evidence records for which dependency relationships could not be established were retained as separate families and marked as having uncertain family membership. The reasons for and versions of all merges, splits, and revisions were documented. Individual evidence records were used for source tracing and timeline visualization, whereas Evidence Families served as the fundamental analytical units for hypothesis scoring, evidence-deletion tests, and bootstrap analysis.

Hypothesis and Robustness Assessment

Following cross-model validation and human adjudication, the three principal hypotheses—CAI, CPI, and LPI—were compared. The Other category was used solely to retain interpretations that did not constitute independent hypotheses and was excluded from the primary ranking. Only Verified evidence records meeting the source-traceability and human-adjudication criteria described above were included in the primary analysis; Uncertain records were retained only for sensitivity analyses.

To prevent evidence inflation caused by repeated accounts of the same primary historical source, archaeological discovery, or historical event across multiple publications, Evidence Families were used as the fundamental scoring units. Let F denote the total number of Evidence Families, F_f the set of evidence records contained in Evidence Family f, e_i the evidence record i, and s_{ih} the stance of evidence record e_i towards hypothesis h, where $s_{ih} \in \{-1,0,+1\}$.

Here, $s_{ih} = +1$indicates support, $s_{ih} = 0$ indicates neutrality or no explicit relationship, and $s_{ih} = -1$ indicates opposition. Regardless of how many publications repeat the evidence contained within a family, each Evidence Family contributes at most one supporting count and one opposing count to a given hypothesis. Accordingly, the number of supporting Evidence Families for hypothesis h, denoted by P_h, is defined in Eq. (13):

$$P_h = \sum_{f=1}^{F} I(\exists e_i \in F_f : s_{ih} = +1) \quad (13)$$

The number of Evidence Families opposing hypothesis h, denoted by O_h, is defined as:

$$O_h = \sum_{f=1}^{F} I(\exists e_i \in F_f : s_{ih} = -1) \quad (14)$$

where I is an indicator function that equals 1 when the specified condition is satisfied and 0 otherwise, and $\exists$ indicates that at least one qualifying evidence record exists. If an Evidence Family contains independent materials that both support and oppose the same hypothesis, it may contribute once to both P_h and O_h. Such cases were labelled as within-family evidence conflicts.

Based on the numbers of supporting and opposing Evidence Families, the standardized net evidence balance, B_h, was used to characterize the overall direction of the evidence:

$$B_h = \begin{cases} \dfrac{P_h - O_h}{P_h + O_h}, & P_h + O_h > 0 \\ 0, & P_h + O_h = 0 \end{cases} \quad (15)$$

Here, $B_h \in [-1,1]$. A positive value indicates that supporting evidence outweighs opposing evidence, whereas a negative value indicates that opposing evidence outweighs supporting evidence. $B_h = 0$ indicates equal numbers of supporting and opposing Evidence Families. When $P_h + O_h = 0$, B_h was defined as zero and the hypothesis was labelled as having "no scorable evidence", thereby preventing this value from being misinterpreted as a balance between supporting and opposing evidence.

Scholarly consensus and evidence sufficiency were calculated separately. Each

independent publication contributed no more than one authorial position towards each hypothesis. Duplicate versions of the same study and records clearly dependent on the same citation lineage were consolidated. Let A_h^+ and A_h^- denote the numbers of independent scholarly positions supporting and opposing hypothesis h, respectively. The direction of scholarly consensus, C_h, is defined as:

$$C_h = \begin{cases} \frac{A_h^+ - A_h^-}{A_h^+ + A_h^-}, & A_h^+ + A_h^- > 0 \\ 0, A_h^+ + A_h^- = 0 \end{cases} \quad (16)$$

A value of $C_h > 0$ indicates that more independent scholarly positions support the hypothesis than oppose it, whereas $C_h < 0$ indicates that opposing positions predominate. Neutral publications were excluded from the denominator but were reported separately. When no explicit supporting or opposing positions were identified, C_h was defined as zero and labelled as indicating "no directional scholarly position", thereby avoiding its misinterpretation as an exact balance between support and opposition.

To evaluate whether the evidence supporting a given hypothesis was derived from multiple independent material types, evidence was classified into three primary categories—documentary, material, and legendary evidence—following the Triple Evidence Method. Evidence-type diversity, D_h, was measured using a normalized Shannon index. Let n_{hj} denote the number of Evidence Families in category j that support hypothesis h, $p_{hj} = n_{hj} / \sum_j n_{hj}$ the proportion of supporting Evidence Families belonging to that category, and J the predefined total number of evidence categories. Evidence-type diversity was calculated as:

$$D_h = -\frac{\sum_{j=1}^{J} p_{hj} \ ln(p_{hj})}{ln(J)} \quad (17)$$

where $D_h \in [0,1]$, and terms for which $p_{hj} = 0$ were defined as zero. Values of D_h approaching 1 indicate that the supporting evidence is more diverse and more evenly distributed across evidence categories, whereas $D_h = 0$ indicates that all supporting evidence is concentrated in a single category. If no supporting Evidence Families existed for hypothesis h, D_h was defined as zero and labelled as indicating "no supporting evidence type".

The supporting volume, opposing volume, net evidence balance, scholarly consensus, and evidence-type diversity were reported separately and then integrated into a composite score. To obtain the baseline ranking of the three hypotheses, S_h^* is the evidence-sufficiency measure obtained by min–max normalizing the number of supporting Evidence Families, S_h, across the three principal hypotheses. In the baseline analysis, equal weights were assigned to the four dimensions, such that $w_S = w_B = w_C = w_D = 0.25$. The composite score was defined as:

$$Q_h = w_S S_h^* + w_B B_h + w_C C_h + w_D D_h, \ \sum_j w_j = 1 \quad (18)$$

where Q_h is the composite score for hypothesis h, S_h^* is normalized evidence sufficiency, B_h is the net evidence balance, C_h is the direction of scholarly consensus, and D_h is evidence-type diversity.

Multiple sensitivity analyses were prespecified to evaluate the effects of the model-

validation threshold, model selection, human adjudication, uncertain evidence, Evidence Family composition, and scoring weights on the hypothesis ranking. Model-confirmation thresholds were set to at least 1/3, at least 2/3, and 3/3 models. The 1/3 threshold was used only as a permissive boundary scenario and did not represent the formal evidence-inclusion criterion. In every scenario, candidate evidence was required to be traceable to the source text. Three scenarios were considered for Uncertain records: exclusion; inclusion after human confirmation of the relevant fields; and inclusion in the direction least favourable to the currently leading hypothesis.

A family-level non-parametric bootstrap used the F Evidence Families as the sampling units[42,43]. In each iteration, F families were sampled with replacement, with the same family-level sampling weights applied to all hypotheses. Under the baseline scoring rules, P_h, O_h, B_h, C_h, D_h, S_h and Q_h were then recalculated. This procedure was repeated 10,000 times[50], and 95% bootstrap intervals were reported using the 2.5th and 97.5th percentiles[43]. When two or more hypotheses tied for first place, the first-place weight for that iteration was divided equally among the tied hypotheses. The complete-ranking preservation rate was defined as the proportion of bootstrap iterations in which the full hypothesis ranking was identical to the baseline ranking. The bootstrap analysis quantified uncertainty arising from Evidence Family composition, whereas the effects of model choice, validation thresholds, human adjudication, and scoring weights were evaluated through separate sensitivity scenarios.

Results

The 275 publications were qualified to the corpus in 6 languages

The three-stage search progressively increased thematic and geographical specificity. The first-stage search used qanat, karez, 坎儿井, and their terminological variants as core search terms and retrieved 4909 records. The second-stage retrieval procedure, combining direct K+T searches with supplementary thematic screening of the first-stage candidate set, yielded 726 records, corresponding to 14.79% of the first-stage total. The third-stage procedure, combining direct K+T+G searches with supplementary geographical screening, yielded 289 records, corresponding to 39.81% of the second-stage results.

The 289 records retrieved in the third stage entered the PRISMA 2020 screening process (Fig. 4). After nine duplicate records had been removed, 280 records underwent title and abstract screening, with no further exclusions at this stage. The full texts of five publications could not be obtained. The remaining 275 publications underwent full-text eligibility assessment and all met the inclusion criteria. The final corpus comprised 148 TXT documents and 127 PDF documents with extractable text layers.

Text extraction, cleaning, and document identification were completed for all 275 publications. Text quality was classified as high for 272 publications and moderate for three. The corpus covered six languages: 150 publications in English, 76 in Chinese, 29 in French, 14 in Russian, 5 in Persian, and 1 in Arabic. The full-text processing success rate was 100%, and 98.9% of the publications were classified as having high-quality text (272/275).

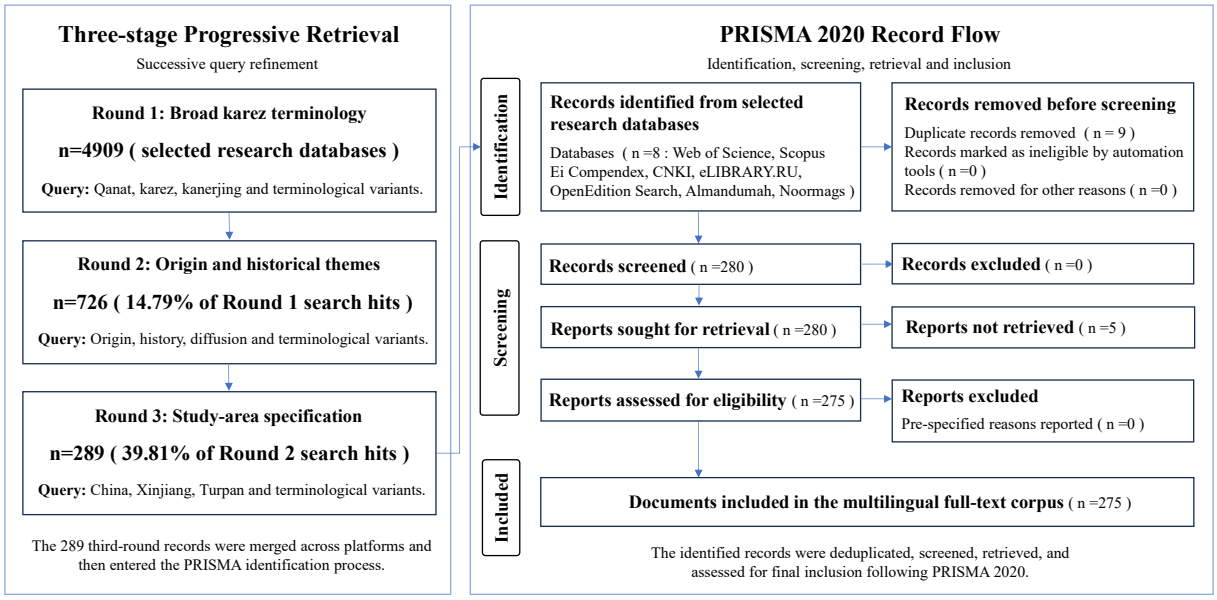


Fig. 4 | Three-stage progressive retrieval strategy and PRISMA 2020 record flow for corpus construction.

Hierarchical chunking of the 275 full-text publications produced 10,192 text chunks, with a mean of 37.06 chunks per publication. The median chunk length was 1,325 characters, the mean was 1,240.3 characters, and the observed range was 70–1,600 characters. Reference-section filtering first identified 202 chunks based on section headings and textual structure. A further 161 bibliography-like chunks were identified from numbered references, GB/T 7714-style citation markers, and bibliographic features including author names, publication years, volume and issue numbers, page ranges, and publication information. After the two sets had been merged and deduplicated, 363 chunks were excluded, representing 3.56% of all chunks. The remaining 9,829 valid chunks, accounting for 96.44% of the original total, were entered into the final retrieval index.

For every publication, a correspondence chain was established linking the source file, document identifier, cleaned full text, text chunks, and source-text locators. Each chunk retained its doc_id, chunk_id, title, source filename, language, section heading, and either a page number or TEXT_SECTION locator to support subsequent source tracing.

The final frozen retrieval configuration indexed titles, section headings, keywords, and body text. BM25 and TF–IDF retrieval results were integrated through RRF using a rank-smoothing constant of 60. Each retriever returned its 50 highest-ranked candidate chunks, and the 30 highest-ranked fused results were retained for each multilingual subquery. Eight multilingual subqueries were constructed, one for each retrieval target. After deduplication by chunk_id across queries, 141 unique target chunks were retained. After the available preceding and following context chunks were added, the inputs comprised 406 chunk occurrences in total. Deduplication of target and context chunks by chunk_id yielded 306 unique chunks. The same target-centered inputs and contextual information were provided to all three LLMs, thereby controlling for differences in retrieval input.

The 141 unique target chunks were retrieved after Top-K evaluation

To evaluate the suitability of the primary retrieval configuration and retrieval depth, we compared BM25+TF–IDF+RRF, E5-base vector retrieval, and BGE reranking under both sum

and max aggregation strategies (Fig. 5). Document Coverage@*K* generally increased across all configurations as the retrieval depth *K* increased from 10 to 100. BM25+TF–IDF+RRF maintained relatively stable document-level coverage across retrieval depths. At *K*=30, which was used in the primary analysis, sum aggregation achieved a Document Coverage@30 of 72.55%, covering 111 of the 153 Gold Standard records. E5-base produced comparatively lower coverage at most retrieval depths, whereas BGE reranking gradually approached the lexical-fusion configuration as K increased but did not consistently outperform BM25+TF–IDF+RRF across the full depth range. Based on these results, BM25+TF–IDF+RRF with sum aggregation was retained for the primary analysis.

The Top-K sensitivity analysis employed two marginal-gain metrics to assess the trade-off between increased coverage and downstream input volume. Increasing K from 20 to 30 yielded an 8.75-percentage-point increase in Document Coverage for every additional 10,000 characters of candidate text. When K increased from 30 to 40, this marginal gain declined to 4.29 percentage points, representing a decrease of approximately 51%. The Gold-hit metric showed a similar pattern: the *K*= (20,30) interval produced the highest marginal gain within the observed range, corresponding to 15.56 additional Gold Standard records per 100,000 additional characters of downstream input.

Although increasing *K* further continued to improve cumulative coverage, both metrics showed that the retrieval gain per unit of additional input declined substantially beyond *K*=30. We therefore selected *K*=30 as the cost-sensitive operating point for the primary analysis.

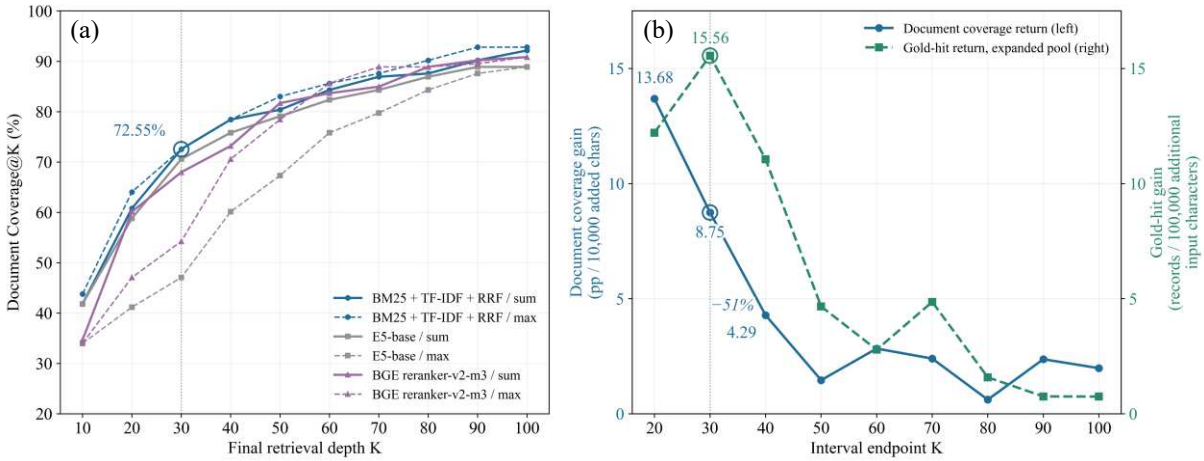


Fig. 5 | Retrieval performance and marginal returns across retrieval depths. (a) Document Coverage@*K* across six retrieval configurations. (**b)** Marginal returns of increasing retrieval depth under the selected retrieval setting.

Under this configuration, RAG outputs served only as a candidate-screening layer between the PRISMA full-text corpus and multi-LLM evidence mining and were not treated directly as historical evidence. The final set of 141 unique target chunks was frozen as the shared first-round input to the three LLMs, with the same adjacent source context provided for each target. This design controlled downstream processing volume while maintaining the retrieval coverage observed at *K*=30 and ensuring that cross-model comparisons were based on identical and traceable source materials.

The 727 unique evidences were identified parallelly across three AI models

DeepSeek V4 Pro, GLM-5.2, and Qwen3.7-Plus generated 332, 359, and 256 raw

candidate evidence records, respectively. After within-model normalization and deduplication, the 359 records generated by GLM-5.2 were consolidated into 353 unique candidate records, whereas the counts for DeepSeek V4 Pro and Qwen3.7-Plus remained unchanged. The three models therefore contributed 332, 353, and 256 normalized candidate evidence records, respectively.

The normalized outputs of the three models were subsequently merged and deduplicated across models, yielding 727 unique candidate evidence records. Of these, 37 (5.09%) were identified by all three models, 140 (19.26%) by two models, and 550 (75.65%) by only one model. Thus, 75.65% of the unique candidate evidence records were identified by only one model.

The relationship between the number of models and candidate-evidence coverage was further examined using the union of 727 unique candidates identified across the three models as the reference set. A single model identified a mean of 313.7 candidates, corresponding to 43.1% of the candidate union, with normalized candidate counts ranging from 256 to 353 across models. Combining any two models identified a mean of 543.7 candidates, corresponding to 74.8% of the union, with individual two-model combinations covering 513–585 candidates. Combining all three models yielded the complete union of 727 candidate evidence records(Fig. 6).

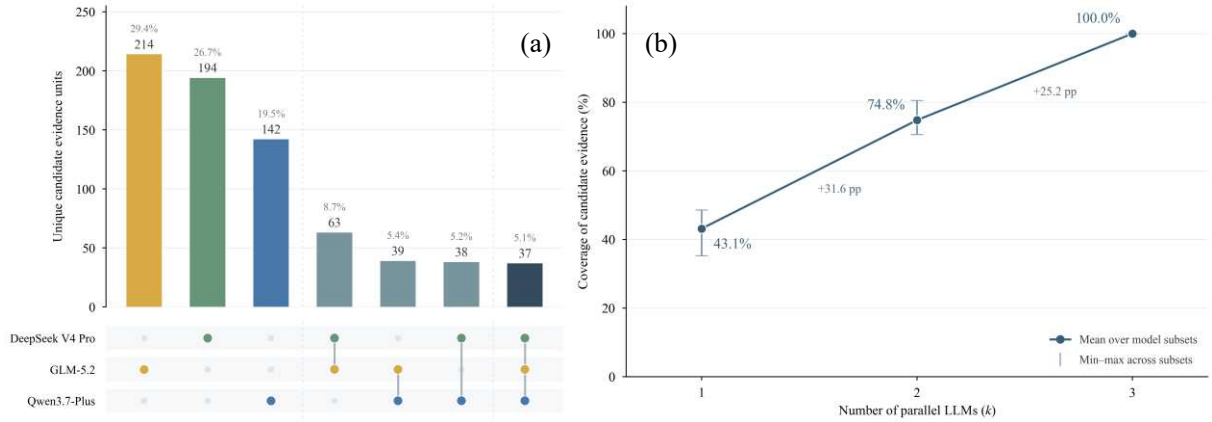


Fig. 6 | Complementarity and coverage gains across three large language models. (a) UpSet plot showing model-specific and shared candidate evidence, with percentages relative to 727 unique records. **(b)** Mean coverage across model subsets. Error bars indicate minimum–maximum ranges; pp denotes percentage points.

The 633 evidence records were cross-validated by models and human adjudication

After candidate normalization and cross-model deduplication, the first-round extractions from the three models yielded 727 unique candidate evidence records. Before the second round of blind validation, the document identifier, source_chunk_id, verbatim quotation, and text-location information of each candidate were programmatically verified. All 727 candidates were traceable to the corresponding source text ($T_i = 1$); therefore, no records were excluded because of insufficient traceability.

DeepSeek V4 Pro, GLM-5.2, and Qwen3.7-Plus subsequently performed blind, source-constrained validation of the 727 candidates without access to which model or models had originally identified each candidate or to the assessments made by the other models. Based on the number of valid entailed classifications, V_i, 613 candidates were classified as entailed by

all three models（$V_i = 3$，84.3%）, 105 by two models （$V_i = 2$，14.4%）, five by one model （$V_i = 1$，0.7%）, and four by none（$V_i = 0$，0.6%）. Seven candidates did not receive a valid assessment from GLM-5.2 because of provider-side content filtering. For these candidates, confirmation counts were calculated using only the assessments returned by DeepSeek V4 Pro and Qwen3.7-Plus. Missing assessments were counted as neither opposition nor confirmation, and the corresponding records proceeded to human review. Overall, 718 candidates (98.8%) received source-support assessments from at least two models. Among the 720 candidates for which all three models returned valid assessments, the mean pairwise observed agreement among the three models in the three-class entailed–contradicted–insufficient task was 90.19%.

Under the tiered review rules defined in Eq. (12), the 718 candidates with $V_i \geq 2$ underwent human review, whereas the nine candidates with $V_i \leq 1$ were directly classified as Rejected. The 613 candidates with $V_i = 3$ were reviewed against the source text by Researcher 1 and subsequently checked by Researcher 2. Following human adjudication, 555 records confirmed by both researchers were classified as Verified, while 58 records were classified as Uncertain because the researchers disagreed. The 105 candidates with $V_i = 2$ were independently reviewed by both researchers. Of these, 78 were jointly confirmed and classified as Verified, whereas 27 were classified as Uncertain because of disagreement or a missing review by Researcher 2. Overall, the 727 candidates comprised 633 Verified records (87.1%), 85 Uncertain records (11.7%), and nine Rejected records (1.2%). Uncertain records were excluded from the primary hypothesis scoring and retained only for sensitivity analyses. The 633 Verified records constituted the formal input for Evidence Family consolidation and hypothesis scoring using Eqs. (13)-(18). The complete workflow is shown in Fig. 7.

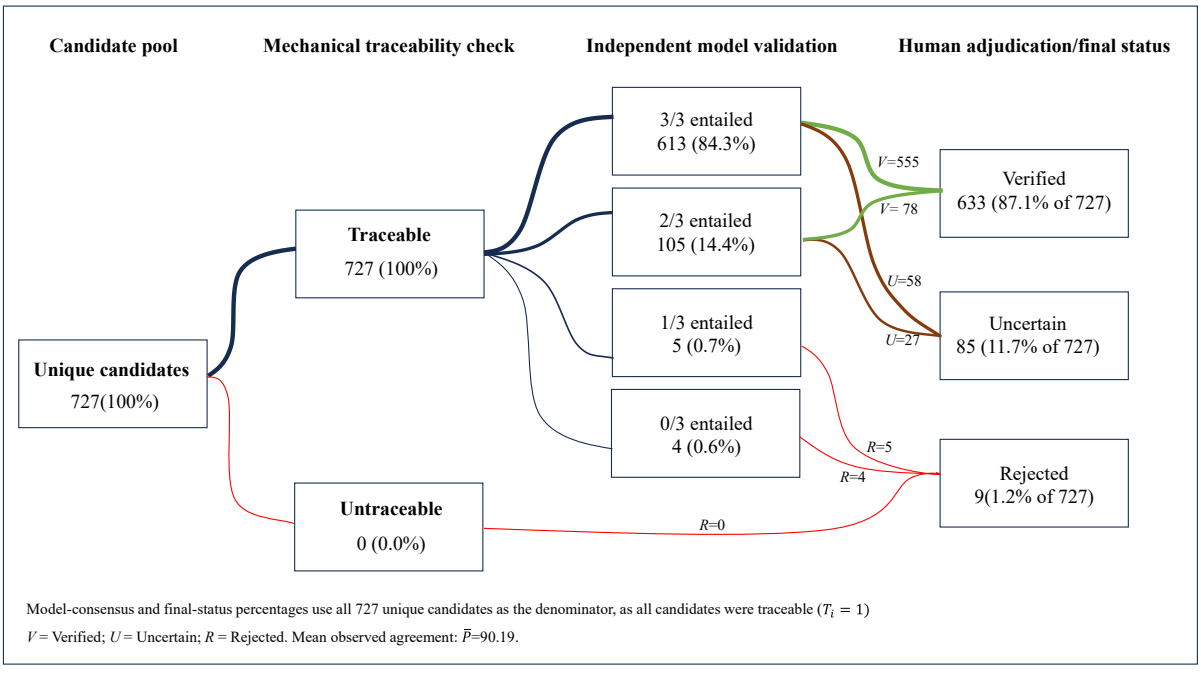


Fig. 7 | Cross-model verification, human adjudication, and final evidence-status assignment.

The 0.5735>0.5125>0.3321 points were obtained for CAI > LPI > CPI

Following cross-model validation and human adjudication, 633 *Verified* evidence records entered the formal hypothesis evaluation. To control evidence inflation caused by repeated

accounts of the same primary historical source, historical event, or archaeological material, the Verified records were further consolidated into Evidence Families based on the source-dependency criteria described in Section 2.4. All subsequent scoring used the Evidence Family as the fundamental unit. When a family contained both supporting and opposing evidence for the same hypothesis, it was counted once among both the supporting and opposing families and labelled as containing a within-family conflict.

The numbers of supporting and opposing Evidence Families and the net evidence balance for each hypothesis are presented in Table 3.

Table 3 | Evidence family counts and net evidence balance by hypothesis.

Hypothesis	Supporting families(P_h)	Opposing families(O_h)	Net evidence balance (B_h)
CAI	116	32	0.568
CPI	59	19	0.513
LPI	54	4	0.862

CAI had the largest number of supporting Evidence Families, indicating the greatest accumulation of independent evidence. Although LPI had fewer supporting families, it had only four opposing families and therefore achieved the highest net evidence balance. Thus, the volume and direction of supporting evidence were not fully aligned.

The direction of scholarly consensus was further calculated from authorial stances at the level of independent publications. Supporting and opposing evidence for the same hypothesis was aggregated within each publication. A publication was coded as support when its supporting evidence exceeded its opposing evidence, as oppose when the reverse was true, and as neutral when the two were equal. Neutral publications were excluded from the denominator of C_h. The results are presented in Table 4.

Table 4 | Distribution of independent scholarly stances and consensus direction across hypotheses.

Hypothesis	Supporting publications (A_h^+)	Opposing publications (A_h^-)	Scholarly consensus direction (C_h)
CAI	14	6	0.400
CPI	15	4	0.579
LPI	16	2	0.778

LPI had the highest net evidence balance and the strongest positive direction of scholarly consensus. Although CAI had the largest number of supporting Evidence Families, its scholarly consensus direction was lower than those of CPI and LPI. This result indicates that the accumulation of independent evidence and the distribution of scholarly positions across publications represent distinct evaluative dimensions.

Evidence-type diversity was assessed using the normalized Shannon index, D_h, across the three primary categories of documentary, material, and legendary evidence. Among the 116 Evidence Families supporting CAI, 105 were based on documentary evidence, nine on material evidence, and two on legendary evidence, yielding $D_h = 0.326$. Of the 59 families supporting CPI, 57 were based on documentary evidence, one on material evidence, and one on legendary evidence, yielding $D_h = 0.156$. Of the 54 families supporting LPI, 45 were based on documentary evidence and nine on material evidence, yielding $D_h = 0.410$. Supporting evidence for CPI was therefore the most heavily concentrated in documentary materials, whereas LPI had the greatest evidence-type diversity, with CAI occupying an intermediate

position (Fig. 8).

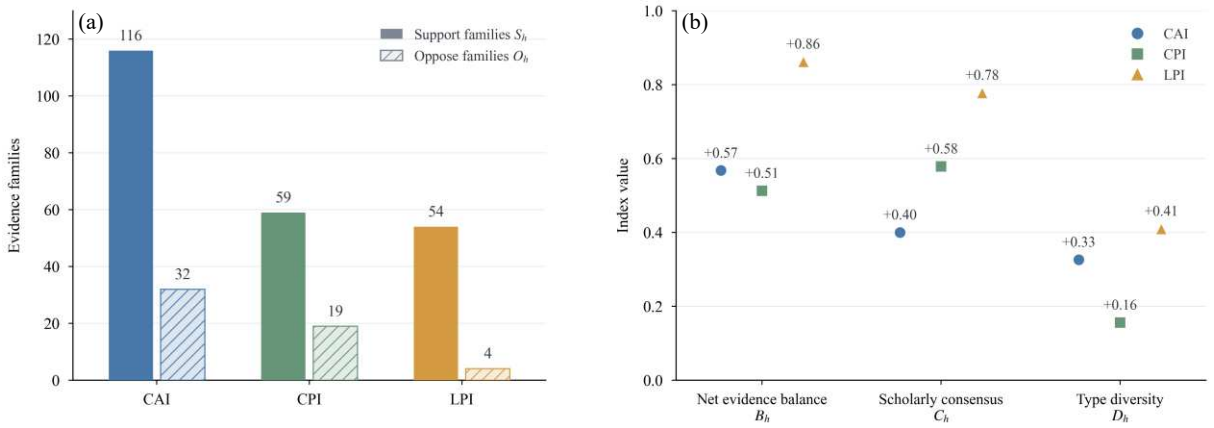


Fig. 8 | Evidence structure of the three origin hypotheses. (a) Numbers of supporting and opposing Evidence Families for CAI, CPI and LPI. **(b)** Net evidence balance, scholarly consensus and evidence-type diversity for the three hypotheses.

Under the equal-weight baseline scenario, the composite comparison incorporated four dimensions: normalized evidence sufficiency S_h^*, net evidence balance B_h, scholarly consensus direction C_h, and evidence-type diversity D_h. Composite scores Q_h were calculated using Eq. (18), and the results are presented in Table 5.

Table 5 | Evidence-based evaluation, composite scores, and ranking of the three origin hypotheses.

Rank	Hypothesis	Evidence sufficiency(S_h^*)	Evidence balance (B_h)	scholarly consensus direction (C_h)	Evidence type diversity(D_h)	Composite Score(Q_h)
1	CAI	1.000	0.568	0.400	0.326	0.5735
2	LPI	0.000	0.862	0.778	0.410	0.5125
3	CPI	0.081	0.513	0.579	0.156	0.3321

The baseline ranking was CAI > LPI > CPI. CAI had the largest number of supporting Evidence Families, resulting in the highest evidence-sufficiency value. Although LPI had the lowest evidence-sufficiency value, it achieved the highest net evidence balance, scholarly consensus direction, and evidence-type diversity, thereby limiting the difference between its composite score and that of CAI. CPI had a lower composite score than both CAI and LPI across the four-dimensional evaluation.

Discussion

The current results indicate that, within the materials included and the evaluation framework adopted in this study, CAI received greater overall support, while LPI remained a strong competing explanation. By contrast, CPI received relatively weaker support under the baseline evaluation and most robustness-analysis scenarios. These findings are consistent with the interpretive direction of some previous studies, although the methods and scope of the conclusions differ. Through a literature review and cross-regional technological comparison, Valipour et al.[1] examined the structures, diffusion, and sustainability of traditional underground waterworks and adopted the interpretation that karez technology was introduced into Turpan from the Middle East in successive stages. In the present study, however, support for CAI represents a relative advantage within the current evidence framework; it neither establishes specific stages of transmission nor provides sufficient grounds to rule out LPI.

Mächtle et al.[13] combined radiocarbon dating of plant remains buried in karez spoil heaps with stratigraphic relationships and oral accounts to date the construction of the oldest investigated systems to the early fifteenth century. Their recognition of a Persian technological background broadly aligns with the relative support for CAI in this study. However, their dating results primarily constrain the construction and maintenance chronology of specific systems and cannot, on their own, distinguish external transmission from independent local invention. Using hydraulic modelling of the MGK4 spate irrigation system, Li et al.[51] proposed that the relevant technologies may have developed locally or originated in regions west of Xinjiang. They emphasized the parallel development of distinct branches of ancient irrigation technology in Xinjiang, without reaching a clear conclusion favouring either CAI or LPI.

To assess whether this ranking depended on specific evidence-screening and evaluation conditions, we further examined the effects of the model-confirmation threshold, treatment of Uncertain evidence, model combination, indicator weights, and Evidence Family composition on the evaluation results, as shown in Fig 9. The baseline ranking of the three hypotheses remained unchanged when the evidence-confirmation threshold was tightened to require unanimous confirmation by all three models and when Uncertain evidence was directly included. However, when Uncertain evidence was recoded in the direction least favourable to the currently leading hypothesis, the composite score of LPI exceeded that of CAI. Leave-one-model-out analysis showed a similar pattern, with LPI ranking first after DeepSeek V4 Pro was excluded. Weight perturbation further revealed the competition between the two hypotheses: among 231 valid weight combinations, 65.8% preserved the ranking of CAI > LPI > CPI, whereas LPI ranked first under 34.2%. Leave-one-family-out analysis further showed that removing any one of the 484 Evidence Families did not alter the baseline ranking, indicating that the ranking was not driven by a single Evidence Family. Across 10,000 bootstrap iterations using Evidence Families as the sampling units, CAI and LPI ranked first in 82.6% and 17.3% of iterations, respectively. These results indicate that CAI retained the leading position more frequently under the current sampling and scoring rules, while LPI could still achieve a higher ranking as Evidence Family composition varied.

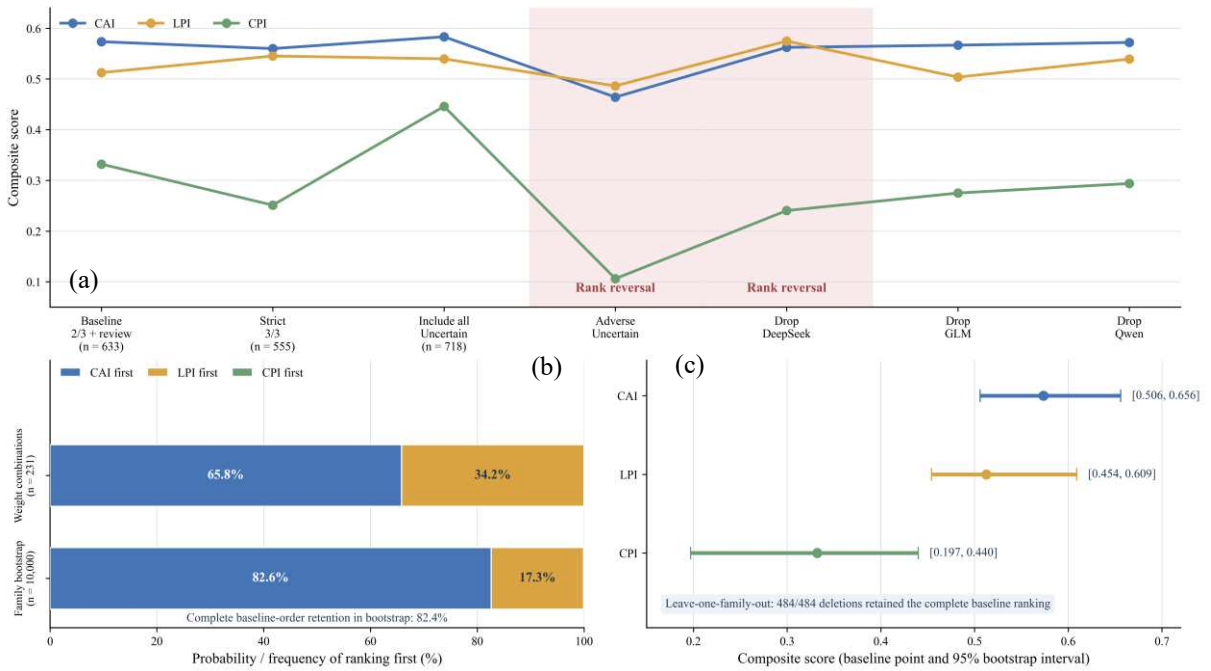


Fig. 9 | Robustness of hypothesis ranking under alternative analytical scenarios. (a) Composite scores under different evidence-admission and leave-one-model scenarios. **(b)** First-place frequencies across weight perturbation and family-level bootstrap analyses. **(c)** Baseline composite scores with 95% family-bootstrap intervals.

From the perspective of technological development, differing interpretations of the origin of the Turpan karez in previous research may also reflect differences in the historical stages addressed by the evidence[7,11]. Materials used in the origin debate may concern technical antecedents, interregional transmission, local adaptation, or later expansion, rather than a single, clearly identifiable origin event. Accordingly, some disagreements among the three hypotheses may arise not only from different interpretations of the same evidence but also from different understandings of the historical processes encompassed by "origin." External technological transmission and subsequent local adaptation are therefore not necessarily mutually exclusive within the historical process. Further direct evidence that jointly constrains chronology, engineering characteristics, and transmission relationships is needed to distinguish external introduction, independent local development, and local modification following transmission.

Evidence synthesis in debates over heritage origins requires not only processing literature across diverse languages and forms of expression but also distinguishing primary materials, subsequent interpretations, and repeated citations to avoid treating accounts with shared sources as independent support. These requirements are consistent with traditional principles of historical source criticism. However, consistently recording evidence provenance, standardizing coding, and comparing competing interpretations remain practical challenges in large multilingual corpora[30,45]. This study integrates LLM-assisted evidence discovery, source-text validation, human adjudication, Evidence Family consolidation, and hypothesis evaluation into a documented and auditable workflow. Rather than replacing source criticism and expert interpretation, this approach applies structured processing to improve the systematic organization, transparency, and reproducibility of large-scale literature-based evidence synthesis.

In the technical implementation, retrieval-augmented generation (RAG), multi-model

complementarity, and human review serve distinct functions. RAG localizes candidate passages and their surrounding context[36], linking model inputs to identifiable source texts while controlling input volume. Multiple LLMs independently extract candidate evidence from these shared inputs, using the complementarity of their outputs to broaden candidate discovery. Source-constrained cross-model validation then assesses whether the candidate statements are supported by the supplied passages, while human review examines evidence validity, authorial stance, and historical interpretation. Grouping source-dependent records reduces evidence inflation from repeated citation and allows hypothesis evaluation to account for source relationships alongside evidence volume, direction, and type. At the process level, the links among publications, page numbers or text locations, candidate statements, validation judgements, and Evidence Families provide an audit trail for review and correction.

The framework may also inform research on ancient technological diffusion, heritage origins, and other historical controversies involving dispersed and interdependent textual evidence. Its transferability lies primarily in its analytical principles and workflow, rather than in the direct replication of specific parameters or evaluation results. Research contexts may differ substantially in corpus size, historical source types, evidence preservation, and hypothesis structure. Accordingly, retrieval strategies, Evidence Family assignment rules, evidence classification schemes, and evaluation weights require adjustment to the specific research question.

This study developed a traceable multilingual evidence-synthesis framework, but its evaluation results remain influenced by corpus coverage, evidence identification, and analytical rules. Although this study assembled 275 multilingual full-text publications through a three-stage progressive multilingual search, database coverage, search terminology, full-text availability, and digitization may still affect the representativeness of the corpus. Undigitized local publications, archives, and field records, as well as materials describing relevant hydraulic structures using alternative historical terms, may be underrepresented. The corpus is also linguistically imbalanced, with English and Chinese publications predominating and relatively few Persian and Arabic sources, which may affect the representation of different scholarly traditions and interpretations. Future research should expand the corpus through historical terminology variants, local archives, and cross-language citation tracing.

The retrieval stage further constrains the range of candidate evidence. The selected configuration achieved a Document Coverage@30 of 72.55%, indicating that some known relevant materials remained uncovered under the selected retrieval depth. Moreover, the Gold Standard comprised materials previously identified by the researchers, and retrieval queries were refined in response to omissions observed in earlier stages. The reported retrieval performance therefore primarily reflects performance under the current development conditions. Future research could establish an independent evaluation set and use stratified sampling across languages, publication periods, and document types to assess omissions among unretrieved materials.

Cross-model source-text validation and human adjudication remain subject to interpretive differences. Although the three models belong to different model families, they may share training data or exhibit similar judgment biases. Human review is also influenced by reviewers' disciplinary backgrounds, language proficiency, and understanding of the coding scheme. In

this study, some records unanimously confirmed by the models were subsequently classified as Uncertain following human review, indicating that source-text support and the validity of historical interpretations need to be assessed separately. Future research should involve domain experts with relevant language expertise, implement independent coding, and document the reasons for disagreement to assess judgment consistency across evidence types and interpretive tasks. Priority should also be given to tracing records critical to distinguishing the hypotheses back to primary historical sources.

The definitions, normalization, and weighting of the four evaluation dimensions also affect the interpretation of the composite scores. Evidence sufficiency, calculated by min–max normalization of the number of supporting Evidence Families, measures the volume of supporting evidence rather than the historical evidential strength of individual items. Net evidence balance reflects the relationship between supporting and opposing evidence within the included materials; fewer opposing records do not necessarily indicate that a hypothesis has been thoroughly tested. The direction of scholarly consensus describes the distribution of stances across the included publications and cannot substitute for an assessment of historical truth. Evidence-type diversity reflects the composition of the evidence base but does not automatically establish effective corroboration among different types of evidence. Although sensitivity analyses demonstrated how selected analytical choices affected hypothesis rankings, these tests remained limited to the specified scenarios.

The comparison of the three origin hypotheses has not fully distinguished among historical propositions concerning technological emergence, transmission, modification, and expansion. Although evidence extraction retained categories such as MULTI, the primary evaluation focused on CAI, LPI, and CPI. Consequently, the composite ranking may combine materials addressing different periods, locations, and technological stages. Future research should refine these broad origin hypotheses into questions testable against specific materials, such as the initial construction date of a particular system, the origin of a specific engineering feature, and whether identifiable technological transmission occurred during a given period. The transferability of the framework also requires cross-case validation. Future research could examine cases differing in source types, linguistic composition, and degree of scholarly disagreement, and compare evidence extraction, Evidence Family consolidation, and hypothesis evaluation with results independently compiled by experts.

Overall, this study provides an auditable framework for synthesizing multilingual evidence on contested heritage origins. These findings describe relative evidentiary support rather than establish a definitive historical origin. The framework's contribution lies in making evidence provenance, source dependence, validation decisions, and interpretive uncertainty explicit and open to scrutiny. The framework provides a transparent basis for reassessing these interpretations as new archaeological, technological, and documentary evidence becomes available.

Funding Declaration

This research was funded by the Water Resources Planning Institute, Department of Water Resources of the Xinjiang Uygur Autonomous Region, under the project "History, Culture, Conservation and Utilization of Karez: Research for Nomination as a World Heritage Irrigation

Structure" (Project No. 2020.A-002).

References

1. Valipour, M. et al. Sustainability of Underground Hydro-Technologies: From Ancient to Modern Times and toward the Future. Sustainability 12, 8983 (2020). https://doi.org/10.3390/su12218983
2. Megdiche-Kharrat, F., Ragala, R. & Moussa, M. Promoting a sustainable traditional technique of aquifer water acquisition common to arid lands: a case study of Ghassem Abad Qanat in Yazd Province (Iran). Water Supply 19, 527–535 (2019). https://doi.org/10.2166/ws.2018.098
3. Naghibi, S. A., Pourghasemi, H. R. & Abbaspour, K. A comparison between ten advanced and soft computing models for groundwater qanat potential assessment in Iran using R and GIS. Theor. Appl. Climatol. 131, 967–984 (2018). https://doi.org/10.1007/s00704–016–2022-4
4. Lou, H. et al. Quantitative reevaluation of the function of Karez using remote sensing technology. Ecol. Indic. 166, 112249 (2024). https://doi.org/10.1016/j.ecolind.2024.112249
5. English, P. W. The Origin and Spread of Qanats in the Old World. Proc. Am. Philos. Soc. 112, 170–181 (1968).
6. Goblot, H. Dans l'ancien Iran, les techniques de l'eau et la grande histoire. Ann. Écon. Soc. Civilis. 18, 499–520 (1963). https://doi.org/10.3406/ahess.1963.421011
7. Barbaix, S., Kurban, A., De Maeyer, P., Chen, X. & Bourgeois, J. The use of historical sources in a multi-layered methodology for karez research in Turpan, China. Water Hist. 12, 281–297 (2020). https://doi.org/10.1007/s12685-020-00259-z
8. Du, L. et al. Hydroclimatic Change in Turpan Basin under Climate Change. Water 15, 3422 (2023). https://doi.org/10.3390/w15193422
9. Li, Q., Guo, H., Luo, L., Wang, X. & Yang, S. Impact Analysis of Land Use and Land Cover Change on Karez in Turpan Basin of China. Remote Sens. 15, 2146 (2023). https://doi.org/10.3390/rs15082146
10. Wang, L., Wang, J., Ding, J. & Li, X. Estimation and Spatiotemporal Evolution Analysis of Actual Evapotranspiration in Turpan and Hami Cities Based on Multi-Source Data. Remote Sens. 15, 2565 (2023). https://doi.org/10.3390/rs15102565
11. Hu, W., Zhang, J. & Liu, Y. The qanats of Xinjiang: historical development, characteristics and modern implications for environmental protection. J. Arid Land 4, 211–220 (2012). https://doi.org/10.3724/SP.J.1227.2012.00211
12. Trombert, É. The Karez Concept in Ancient Chinese Sources: Myth or Reality? T'oung Pao 94, 115–150 (2008). https://doi.org/10.1163/008254308X367031
13. Mächtle, B. et al. The age and origin of Karez systems of Silk Road oases around Turpan, Xinjiang, P.R. of China. In Socio-Environmental Dynamics along the Historical Silk Road (eds Yang, L. E., Bork, H.-R., Fang, X. & Mischke, S.) 359–378 (Springer, 2019). https://doi.org/10.1007/978-3-030-00728-7_17
14. Tao, B. Xinmao Shixing Ji (Yangshu Shanfang, 1897).
15. Wang, G. Xiyu Jingqu Kao (1919).
16. Huang, W. & Kan, Y. Xinjiang karez: its history, present condition and future development.

Arid Land Geogr. 13, 33–37 (1990). https://doi.org/10.13826/j.cnki.cn65-1103/x.1990.03.007
17. Zhong, X. The shaft-tunnel method in the Central Plains and karez in Turpan. West. Reg. Stud. (4), 36–43 (1995).
18. Huntington, E. The Pulse of Asia: A Journey in Central Asia Illustrating the Geographic Basis of History (Houghton Mifflin Company, 1907).
19. Stein, M. A. On Ancient Central-Asian Tracks: Brief Narrative of Three Expeditions in Innermost Asia and North-Western China (Macmillan and Co., 1933).
20. Huang, S. Further discussion on the origin and diffusion of karez systems in Xinjiang. West. Reg. Stud. (1), 66–84 (1994). https://doi.org/10.16363/j.cnki.xyyj.1994.01.010
21. Yang, S. The origin and diffusion of karez and the karez culture of Turpan. In Xibei Minzu Luncong [Collected Essays on Northwestern Ethnic Groups] Vol. 1 (ed. Zhou, W.) (China Social Sciences Press, 2002).
22. Yuan, D., Wang, S. & Wang, X. Xinjiang Tuzhi (Xinjiang Tongzhi Bureau, 1911).
23. Liu, H. A review of the Turpan karez. Agric. Hist. China (4), 24–32 (1986).
24. Zhang, X. Formation conditions and theories concerning the origin of karez systems in Xinjiang. Dixiashui [Ground Water] (1), 30–32 (1985).
25. Abudureimu, A., Jiapaer, T., Kuerban, A. & Wupuer, T. The material evidence origin of Xinjiang Kaerjing. J. Xinjiang Norm. Univ. (Nat. Sci. Ed.) 30, 23–28 (2011). https://doi.org/10.14100/j.cnki.1008-9659.2011.02.006
26. Rathbone, J., Hoffmann, T. & Glasziou, P. Faster title and abstract screening? Evaluating Abstrackr, a semi-automated online screening program for systematic reviewers. Syst. Rev. 4, 80 (2015). https://doi.org/10.1186/s13643-015-0067-6
27. van Eck, N. J. & Waltman, L. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 84, 523-538 (2010). https://doi.org/10.1007/s11192-009-0146-3
28. Chen, C. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. J. Am. Soc. Inf. Sci. Technol. 57, 359–377 (2006). https://doi.org/10.1002/asi.20317
29. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent Dirichlet allocation. J. Mach. Learn. Res. 3, 993-1022 (2003).
30. van de Schoot, R. et al. An open source machine learning framework for efficient and transparent systematic reviews. Nat. Mach. Intell. 3, 125–133 (2021). https://doi.org/10.1038/s42256-020-00287-7
31. Beltagy, I., Lo, K. & Cohan, A. SciBERT: A pretrained language model for scientific text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) 3615–3620 (Association for Computational Linguistics, 2019). https://doi.org/10.18653/v1/D19-1371
32. Cohan, A., Ammar, W., van Zuylen, M. & Cady, F. Structural scaffolds for citation intent classification in scientific publications. 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) 3586–3596 (Association for

Computational Linguistics, 2019). https://doi.org/10.18653/v1/N19-1361
33. Wadden, D. et al. Fact or fiction: verifying scientific claims. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) 7534-7550 (Association for Computational Linguistics, 2020). https://doi.org/10.18653/v1/2020.emnlp-main.609
34. Wang, Z. et al. Accelerating clinical evidence synthesis with large language models. npj Digit. Med. 8, 509 (2025). https://doi.org/10.1038/s41746-025-01840-7
35. Asai, A. et al. Synthesizing scientific literature with retrieval-augmented language models. Nature 650, 857–863 (2026). https://doi.org/10.1038/s41586–025–10072–4
36. Lewis, P. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems Vol. 33 9459–9474 (Curran Associates, 2020).
37. Page, M. J. et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 372, n71 (2021). https://doi.org/10.1136/bmj.n71
38. Robertson, S. & Zaragoza, H. The Probabilistic Relevance Framework: BM25 and Beyond. Found. Trends Inf. Retr. 3, 333–389 (2009). https://doi.org/10.1561/1500000019
39. Salton, G. & Buckley, C. Term-weighting approaches in automatic text retrieval. Inf. Process. Manag. 24, 513–523 (1988). https://doi.org/10.1016/0306-4573(88)90021-0
40. Cormack, G. V., Clarke, C. L. A. & Büttcher, S. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval 758-759 (Association for Computing Machinery, 2009). https://doi.org/10.1145/1571941.1572114
41. Shi, X., Zhao, W., Chen, T., Yang, C. & Du, J. Evidence triangulator: using large language models to extract and synthesize causal evidence across study designs. Nat. Commun. 16, 7355 (2025). https://doi.org/10.1038/s41467-025-62783-x
42. Efron, B. Bootstrap methods: another look at the jackknife. Ann. Stat. 7, 1-26 (1979). https://doi.org/10.1214/aos/1176344552
43. Davison, A. C. & Hinkley, D. V. Bootstrap Methods and Their Application (Cambridge University Press, 1997).
44. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
45. Rethlefsen, M. L. et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst. Rev. 10, 39 (2021). https://doi.org/10.1186/s13643-020-01542-z
46. Liu, N. F. et al. Lost in the Middle: How Language Models Use Long Contexts. Trans. Assoc. Comput. Linguist. 12, 157-173 (2024). https://doi.org/10.1162/tacl_a_00638
47. Gartlehner, G. et al. Data extraction for evidence synthesis using a large language model: A proof-of-concept study. Res. Synth. Methods 15, 576-589 (2024). https://doi.org/10.1002/jrsm.1710
48. Huang, X., Huang, Z. & Zhang, Y. Zhuangzu Tongshi [A General History of the Zhuang People] (Guangxi Nationalities Publishing House, 1988).
49. Huang, X. Guangxi Tongzu Jianshi [A Brief History of the Zhuang People in Guangxi] (Guangxi People's Publishing House, 1957).

50. Hesterberg, T. C. What Teachers Should Know About the Bootstrap: Resampling in the Undergraduate Statistics Curriculum. Am. Stat. 69, 371-386 (2015). https://doi.org/10.1080/00031305.2015.1089789
51. Li, Y., Storozum, M. J., Jia, X., Wang, X. & Frachetti, M. D. Reconceptualizing water history of Chinese Central Asia: Hydraulic modeling of the early 1st mill. AD irrigation system at Mohuchahangoukou-4 (MGK4), Xinjiang, China. J. Archaeol. Sci. Rep. 33, 102534 (2020). https://doi.org/10.1016/j.jasrep.2020.102534