DISTILLING TACIT PROCEDURAL KNOWLEDGE: FROM SCHOLARLY LITERATURE TO SKILL-GUIDED MULTIMODAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) possess extensive factual knowledge but often struggle with expert problems that require coordinated use of heterogeneous evidence. We argue that a key missing component is tacit procedural knowledge—the knowledge of selecting, integrating and revising evidence. We propose Explicit-to-Tacit-to-Skill (E2T2S), a framework that extracts structured knowledge from scholarly papers, aggregates it across each scholar's research to recover reusable procedural patterns, and consolidates them into an Expert Skill Library for dynamic skill-guided reasoning. We evaluate E2T2S on Chinese Chu bamboo-slip character interpretation, where models must infer a modern Chinese character from its glyph image, textual context, external evidence, and scholarly literature. On a controlled held-out evaluation set, our method outperform both generic chain-of-thought and retrieval-augmented baselines. Our results demonstrate that scholarly literature can serve not only as a source of knowledge, but also as a source of reusable expert procedures for LLM reasoning.

1 INTRODUCTION

Large language models (LLMs) have acquired extensive factual and linguistic knowledge through large-scale pretraining, yet knowledge access does not necessarily translate into expert-level problem solving. In specialized tasks, models may retrieve relevant information while still failing to determine which evidence matters, how heterogeneous evidence should be combined, or when an initial hypothesis should be revised. Recent evidence suggests that procedural knowledge plays an important role in reasoning and generalization, beyond simple retrieval of answer-bearing information Ruis et al. (2025). This raises a fundamental question: what is missing when an LLM knows the relevant information but still fails to solve a specialized problem?

We argue that an important missing component is procedural expert knowledge—knowledge about how to use knowledge. Consider a problem with several plausible alternatives and heterogeneous evidence. A knowledgeable solver may know all the relevant facts, but expertise lies in deciding which evidence to examine first, which sources provide stronger or independent support, how conflicting evidence should update competing hypotheses, and when a previous judgment should be revised. Thus, the critical component is not the individual facts themselves, but the conditional procedures that connect evidence to decisions: when evidence should be invoked, how it should affect competing hypotheses, and under what conditions a conclusion should be rejected or revised.

Existing approaches largely address knowledge access or generic reasoning rather than this domain-specific procedural layer. Retrieval-Augmented Generation (RAG) provides relevant documents to support knowledge-intensive generation Lewis et al. (2020), but retrieving additional information does not specify how that information should be organized and used. Indeed, models may struggle to effectively use relevant information in long contexts Liu et al. (2024). Reasoning methods such as Chain-of-Thought, Tree-of-Thoughts, and SELF-DISCOVER improve intermediate reasoning through generic decomposition, exploration, or composition Wei et al. (2022); Yao et al. (2023); Zhou et al. (2024a), while process supervision improves the reliability of intermediate steps Lightman et al. (2024). However, these approaches do not

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

directly address how reusable expert procedures can be induced from accumulated domain practice.

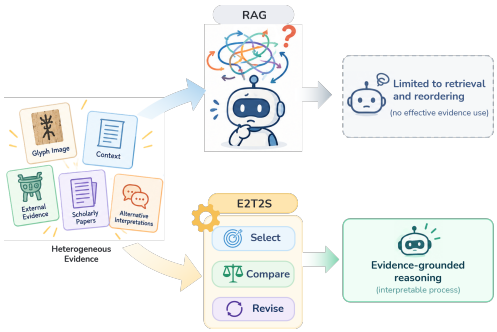

Figure 1: Whereas RAG augments LLMs by retrieving and reordering heterogeneous evidence, E2T2S equips LLMs with reusable procedures to select, compare, and revise evidence, enabling grounded and interpretable reasoning

We propose that scholarly literature provides a natural source of such expert procedures. Academic papers record not only conclusions, but also how experts compare evidence, evaluate alternatives, handle exceptions, and revise interpretations. Research systems such as PaperQA demonstrate the use of scholarly literature for retrieval and synthesis Lála et al. (2023), while recent work explicitly highlights distinctions between knowledge and skills in specialized LLM tasks Wang et al. (2025). Building on these observations, we propose E2T2S (Explicit-to-Tacit-to-Skill), which first extracts structured explicit knowledge from scholarly papers, then aggregates this knowledge across an individual scholar's research to recover recurring tacit procedural patterns, and finally consolidates these procedures into a reusable Expert Skill Library. At inference time, relevant skills are dynamically selected according to the current evidence state and injected into the model as procedural guidance.

We study E2T2S in Chinese Chu bamboo-slip character interpretation, a challenging multimodal task at the intersection of ancient writing research and natural language processing. Computational understanding of excavated writing is important for reconstructing historical writing systems and recovering linguistic and textual information from early manuscripts. Recent multimodal benchmarks and models have begun to explore ancient Chinese writing, but primarily emphasize recognition, character evolution, or broad classical-studies tasks Shi et al. (2025a); Song et al. (2026). In contrast, Chu-slip interpretation requires a model to infer a modern Chinese character from its glyph image, textual context, external evidence, and relevant scholarly literature. The difficulty arises because no single evidence source is generally sufficient: glyph form, context, phonological relationships, parallel or transmitted texts, variant readings, and scholarly interpretations may provide complementary or conflicting signals. Successful interpretation therefore requires procedures for selecting, combining, and revising evidence.

To evaluate procedural effectiveness, we construct a scholarly-literature-based dataset of 644 Chu-slip interpretation instances and adopt a paper–character double-isolation split, ensuring that neither the target character nor its associated paper in the evaluation set is available to the corpus used for skill extraction or retrieval. We compare Base, CoT, RAG, RAG+CoT, and our E2T2S method under identical multimodal inputs and generation settings; RAG-based method and E2T2S method use the same corpus. Our method achieves 51.57% accuracy and 56.16% Macro-F1, substantially outperforming the conventional reasoning and retrieval baselines. Evidence-aware evaluation and stage-wise ablation further examine when expert skills are useful and the contribution of each E2T2S stage.

Our contributions are summarized as follows:

- We identify procedural expert knowledge as an important missing layer between knowledge access and expert-level LLM reasoning, emphasizing reusable procedures for evidence selection, integration, and revision.
- We propose E2T2S, a framework that transforms scholarly literature into reusable expert skills by progressively extracting explicit knowledge, recovering scholar-level tacit procedural knowledge, and constructing an Expert Skill Library for dynamic reasoning.
- We establish a challenging multimodal evaluation setting for Chinese Chu bamboo-slip character interpretation and demonstrate that dynamically selected expert skills substantially improve performance over generic reasoning and retrieval-based approaches, with evidence-aware and ablation analyses providing further support for the proposed procedural reasoning mechanism.

2 Related Work

Knowledge augmentation and structured reasoning are two major approaches for improving LLM problem solving. Retrieval-based methods such as REALM and RAG augment models with external information retrieved from documents Guu et al. (2020); Lewis et al. (2020), while systems such as Atlas extend retrieval augmentation to few-shot knowledge-intensive tasks Izacard et al. (2023). Scientific literature has increasingly been used as a knowledge source for LLMs, with systems such as PaperQA, STORM, PaperQA2, and OpenScholar focusing on retrieving, synthesizing, and organizing information from research papers Lála et al. (2023); Shao et al. (2024); Skarlinski et al. (2024); Asai et al. (2024). In parallel, Chain-of-Thought, Tree-of-Thoughts, SELF-DISCOVER, and process supervision improve reasoning through explicit reasoning structures or intermediate-step supervision Wei et al. (2022); Yao et al. (2023); Zhou et al. (2024a); Lightman et al. (2024). Recent studies further highlight the role of procedural knowledge and reusable skills in reasoning, including work on procedural knowledge in pretraining Ruis et al. (2025), procedural knowledge models for planning Brahman et al. (2024), expert decision models Wang et al. (2024), and reusable skill libraries for agents Wang et al. (2023); Zheng et al. (2025). However, existing approaches generally treat scholarly literature as a source of declarative knowledge or derive procedures from interaction, demonstrations, or predefined reasoning structures. Our work instead explores how tacit procedural knowledge can be systematically recovered from naturally occurring scholarly analyses and transformed into reusable expert skills.

AI research on ancient Chinese writing has primarily focused on character recognition, glyph understanding, decipherment, and multimodal evaluation. Existing resources and methods include benchmarks for Oracle Bone Inscriptions Li et al. (2024); Zhou et al. (2026); Diao et al. (2026), automatic oracle-bone character recognition Zhou et al. (2024b); Shi et al. (2025a;b), diffusion-based decipherment Guan et al. (2024), multimodal benchmarks for Chinese Classical Studies Liu et al. (2025), and recent work on ancient Chinese character evolution Song et al. (2026). These studies demonstrate the value of multimodal models for ancient writing, but generally emphasize visual recognition or character-level understanding rather than the expert procedures required to integrate heterogeneous evidence. In contrast, we formulate Chinese Chu bamboo-slip character interpretation as an expert reasoning problem in which glyph images, textual context, external evidence, and scholarly literature must be jointly used to resolve competing interpretations. Our work therefore complements existing ancient-writing research by focusing on the missing procedural layer between access to scholarly knowledge and expert-level evidence-based interpretation.

3 Proposed Method

3.1 Task Formulation

We formulate Chu bamboo-slip character interpretation as a simulation of the actual scholarly interpretation process. Given a Chu bamboo-slip character to be interpreted, along with associated information as evidence, the model is required to determine the corresponding modern Chinese character and provide the reasoning process supporting the interpretation. The associated evidence consists of four complementary sources:

$$x = (I, C, E, P),$$

where I denotes image evidence, C denotes contextual evidence, E denotes external evidence, and P denotes the scholarly literature corpus.

Specifically, I consists of the complete Chu bamboo-slip image and a cropped image of the target character, providing both the visual context of the manuscript and the detailed form of the target glyph. C consists of the modern transcription of the corresponding Chu bamboo-slip text with the target character removed, providing the linguistic context surrounding the target position. E consists of external evidence relevant to the interpretation, such as archaeological provenance, excavation information, parallel texts, and other textual or historical evidence. P consists of relevant scholarly papers concerning the target character, related characters, corresponding texts, or competing interpretations.

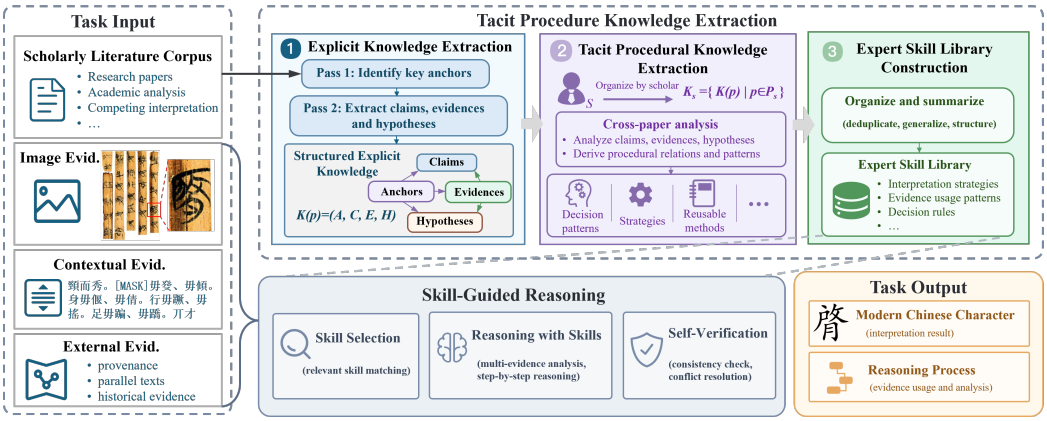

Figure 2: Overview of the E2T2S framework. Scholarly literature is progressively transformed into structured explicit knowledge, scholar-level tacit procedural knowledge, and a reusable Expert Skill Library. For a new Chu bamboo-slip interpretation task, relevant expert skills are dynamically selected and injected together with multimodal character evidence to guide evidence-grounded reasoning and produce the final character interpretation and reasoning process.

Given the four types of evidence, the model produces the modern Chinese character corresponding to the target Chu bamboo-slip character together with a reasoning process:

$$(\hat{y}, \hat{r}) = M_\theta(I, C, E, P),$$

where $\hat{y}$ denotes the interpreted modern Chinese character and $\hat{r}$ denotes the reasoning process supporting the interpretation.

This formulation reflects the actual practice of expert paleographic interpretation, in which the interpretation of a character is based not on its visual form alone, but on the joint consideration of glyph, context, external evidence, and accumulated scholarly research.

3.2 Explicit-to-Tacit-to-Skill Framework

Figure 2 illustrates the overall framework, which contains three stages. The first stage converts scholarly papers into structured explicit knowledge. The second stage infers the procedural relations underlying those knowledge units. The third stage consolidates these procedures into an expert skill library and dynamically selects relevant skills for reasoning on a new problem.

3.2.1 Explicit Knowledge Extraction

For all scholarly papers, given a paper p, we first extract its explicit knowledge through two successive passes. The first pass identifies the key anchors that organize the scholarly discussion:

$$A(p) = \{a_1, a_2, \ldots, a_n\},$$

including glyph instances, variants, inscriptions, artifacts, citations, scholars, dates, and other relevant references. Each anchor is linked to its surface form, normalized representation, and source provenance.

In the second pass, conditioned on the extracted anchors, we identify the paper's claims, hypotheses, and evidences, forming a structured representation

$$K(p) = \big(A(p), C(p), H(p), E(p)\big).$$

Claims encode explicit scholarly relations, such as read-as, variant-of, supports, and refutes. Hypotheses capture the interpretations proposed or evaluated by scholars together with their stance and confidence. Evidences record the basis for these claims and hypotheses, including glyph comparison, chronology, parallel texts, phonetic arguments, semantic analogy, negative evidence, and contextual fit.

This two-pass design converts unstructured scholarly writing into a grounded, claim-centric, and hypotheses-centric knowledge representation, preserving the relationships among entities, claims, evidence, and interpretations as well as their source provenance. It captures what experts claim and what evidence they use, providing the structured basis for the subsequent extraction of tacit procedural knowledge.

3.2.2 TACIT PROCEDURAL KNOWLEDGE EXTRACTION

The second stage extracts tacit procedural knowledge from the structured knowledge obtained in first stage. Unlike explicit knowledge, which records individual scholarly claims and their supporting evidence, tacit procedural knowledge captures the reusable methods, strategies, and decision patterns that experts develop through sustained research. Such knowledge is rarely stated as a complete procedure within a single paper and may instead emerge consistently across an expert's research over multiple studies.

We therefore organize the structured knowledge $K(p)$ from the scholarly corpus by scholar. For a scholar s, let

$$\mathcal{K}_s = \{K(p) \mid p \in \mathcal{P}_s\},$$

where $\mathcal{P}_s$ denotes the set of papers authored by s. We then jointly analyze the claims, evidences, hypotheses, and their relations across this scholar-specific corpus. This scholar-centric aggregation allows us to distinguish a one-time observation or paper-specific argument from a recurring research practice that reflects the expert's accumulated experience.

The resulting tacit procedural knowledge is represented through four complementary components:

$$T_s = \big(V_s, D_s, B_s, R_s\big),$$

where V_s is a core stance vector, D_s is a decision template set, B_s contains boundary and exception rules, and R_s records revision traces.

The core stance vector captures relatively stable preferences or methodological orientations that characterize how an expert evaluates different forms of evidence. The decision template set captures reusable reasoning procedures, specifying their trigger conditions, required inputs, operation steps, expected output patterns, and applicable scope. The boundary and exception rules specify situations in which a procedure should not be directly applied, or conditions that require alternative treatment. The revision traces capture how an expert's interpretation or methodological position changes across studies in response to new evidence, revealing when and why an established judgment is revised.

Formally, the extraction process can be expressed as

$$\mathcal{K}_s \xrightarrow{f_{\text{tacit}}} T_s,$$

where f_{tacit} identifies recurring evidence-use and decision patterns across the scholar's publications. For example, if an expert repeatedly resolves competing interpretations by comparing parallel texts, evaluating their phonological consistency, and revising an earlier reading when the comparison reveals contradictory evidence, these recurring patterns can be abstracted into a reusable decision procedure rather than remaining as paper-specific arguments.

Importantly, tacit knowledge is extracted at the scholar level rather than the individual-paper level. This enables the framework to capture methodological regularities accumulated over an expert's research trajectory while retaining differences among scholars. Each latent item is linked to the claims and evidence from the source papers that support its induction, allowing the extracted procedural knowledge to remain grounded and auditable.

The output of Stage II is therefore a scholar-specific representation of how an expert reasons, rather than merely what the expert knows. These reusable procedural patterns constitute the basis for the subsequent construction of a unified Expert Skills Library.

3.2.3 EXPERT SKILL LIBRARY CONSTRUCTION

The third stage consolidates scholar-specific tacit procedural knowledge into a reusable Expert Skill Library. Since different scholars may describe similar reasoning practices with different formulations, we use a set of canonical skill patterns as a shared organizational space and align the extracted decision templates from different scholars to these patterns.

For each canonical skill s_j, we aggregate the corresponding tacit knowledge from multiple scholars, including decision templates, boundary and exception rules, core stance priors, and revision traces. Decision templates provide the operational procedures of a skill, while boundary rules specify when these procedures should be weakened or avoided. Core stance priors capture relatively stable expert preferences over relevant evidence, and revision traces record how scholarly interpretations change in response to new evidence. All aggregated components retain provenance to the source scholars and papers.

Formally, the construction process can be written as

$$\mathcal{T} \xrightarrow{f_{\text{align}}} \mathcal{S} = \{s_1, s_2, \ldots, s_J\},$$

where $\mathcal{T}$ denotes the collection of scholar-specific tacit procedural knowledge and f_{align} aligns and aggregates semantically consistent procedures into canonical skills. A resulting skill can be represented as

$$s_j = (D_j, B_j, V_j, R_j, \Pi_j),$$

where D_j denotes decision templates, B_j boundary and exception rules, V_j stance priors, R_j revision traces, and Π_j provenance.

After aggregation, we further refine the operation steps and instruction templates of each skill to remove redundancy and produce a compact executable representation. The resulting library therefore contains reusable evidence-use procedures grounded in multiple expert analyses, rather than isolated paper-specific instructions or answer-bearing knowledge. This library serves as the procedural interface between the tacit knowledge extracted from scholarly literature and the subsequent reasoning process of the LLM.

3.3 Dynamic Skill Inject and Reasoning

Given the Expert Skill Library $\mathcal{S}$ constructed in Section 3.2, the final stage dynamically selects skills that are relevant to the current interpretation problem and injects them into the LLM as procedural guidance. The key idea is to avoid exposing the entire skill library to the model. Instead, the reasoning process is conditioned on a small set of skills selected according to the evidence available in the current instance.

For a new interpretation instance

$$x = (I, C, E, P),$$

we first construct an evidence-state representation

$$q_x = f_{\text{signal}}(I, C, E),$$

which summarizes the reasoning conditions of the current problem, such as the available evidence types, contextual constraints, candidate alternatives, and potential evidence conflicts. Each skill $s_j \in \mathcal{S}$ is then evaluated against q_x, and the most relevant skills are retrieved:

$$S_x = \text{TopK}\left(\mathcal{S}, q_x\right) = \{s_{j_1}, \ldots, s_{j_K}\}.$$

The retrieved skills are not additional answer-bearing evidence. Instead, they provide compact procedural instructions specifying when the procedure should be applied, what evidence it operates on, how the evidence should be processed, and under what conditions its conclusion should be reconsidered.

We dynamically inject the selected skills into the user-side reasoning context together with the original task evidence:

$$\mathcal{M}_x = \text{Inject}\left(x, S_x\right),$$

where $\mathcal{M}_x$ contains the current multimodal inputs and the detailed instructions of the selected skills. In this process, each retrieved skill contributes its operational steps and applicable boundary conditions, while the original evidence remains unchanged. Thus, skill injection changes how the model uses the evidence, rather than what evidence the model receives.

The multimodal LLM then reasons over the complete input and the dynamically selected procedures:

$$(\hat{y}, \hat{r}) = M_\theta\left(\mathcal{M}_x\right).$$

Method	Acc.(%)	Macro-P(%)	Macro-R(%)	Macro-F1(%)
Base	38.74	49.51	39.07	42.43*
CoT Wei et al. (2022)	39.79	47.24	39.59*	42.09
RAG Lewis et al. (2020)	38.29	51.47*	37.64	42.05
RAG+CoT	41.03*	48.52	38.54	41.61
E2T2S (ours)	**51.57**	**66.50**	**51.16**	**56.16**

Table 1: Main results on Chinese Chu bamboo-slip character interpretation. All methods use the same model, test instances, candidate sets, images, contextual evidence, and external evidence. RAG-based methods use the same retrieved scholarly evidence. Bold indicates the highest score, while "*" denotes the highest score for baselines.

The reasoning process is guided by the selected skills rather than by a fixed global reasoning template. The model first identifies evidence relevant to the current target, applies the corresponding skill operations, compares the resulting support for competing interpretations, and revises a provisional interpretation when the evidence conflicts with the skill's boundary conditions. The model is also constrained to produce the final interpretation together with its supporting reasoning.

4 Experiments

4.1 Datasets

We construct our evaluation dataset from scholarly literature on Chinese Chu bamboo-slip character interpretation, a challenging setting in which target characters often require coordinated use of contextual, textual, phonological, and other external evidence beyond glyph form alone.

We collect papers on excavated Chu manuscripts and use an LLM to extract target characters, character images, modern transcriptions, and scholarly external evidence. These elements are organized into interpretation instances following our task formulation, and a domain expert further verifies character identification, image-text alignment, transcription, and evidence associations.

After quality control, the dataset contains 644 instances. To prevent answer leakage, we adopt a 7:3 scholar-character double-isolation split: the corpus set is used for procedural knowledge extraction and RAG retrieval, while the evaluation set contains target characters and scholarly papers absent from the corpus. This ensures that models cannot directly retrieve the scholarly papers containing the gold interpretation.

4.2 Baselines and Evaluation Metrics

We use Qwen3-VL-8B-Instruct as the common multimodal base model, with all methods sharing the same evaluation instances, candidate sets, images, contextual and external evidence, and generation settings. We compare Base, which uses only the task inputs; CoT, which adds generic chain-of-thought prompting Wei et al. (2022); RAG, which retrieves relevant passages from the training-paper corpus Lewis et al. (2020); and RAG+CoT, which combines both. Our primary metric is Accuracy, defined as the proportion of correctly interpreted instances. We additionally report Macro-Precision, Macro-Recall, and Macro-F1, computed over the global class space formed by the unique gold characters in the evaluation set, since the same character may occur in multiple distinct interpretation instances.

4.3 Implementation

We formulate Chu bamboo-slip character interpretation as a four-way candidate selection task, where the model receives the full bamboo-slip image, target-glyph image, masked transcription, external evidence, and four candidate modern Chinese characters, and selects the corresponding interpretation. All conditions use the same Qwen3-VL-8B-Instruct model, inputs, and generation

Evidence Condition	Samples	Base	CoT	RAG	RAG+CoT	E2T2S
No External Evidence	66	42.42	40.90	42.42	43.93*	**53.03**
Single Evidence Type	82	36.58	47.56*	42.68	41.46	**54.87**
Multiple Evidence Types	120	30.83	28.33	32.20	38.98*	**49.16**
Supportive Evidence	120	36.66	34.16	35.29	41.66*	**50.00**
Contrasting Evidence	6	16.66	33.33	33.33	66.66*	**66.66**
Mixed Evidence	137	30.65	31.38	32.84	37.95*	**48.33**
Rich Evidence	113	32.74	32.74	33.62	39.82*	**50.44**

Table 2: Accuracy for the Evidence-Aware experiment. Bold text indicates the highest score, while '*' denotes the highest score achieved by each baseline.

settings. For RAG-based methods, we retrieve the top-3 scholarly evidence chunks from the corpus and reuse the identical cached retrieval results across RAG, RAG+CoT, and Ours.

4.4 Main Results

Table 1 presents the main results on the Chinese Chu bamboo-slip character interpretation task. All methods are evaluated under the same model, test set, multimodal inputs, candidate sets, external evidence, and generation settings. The results show a clear advantage of the proposed E2T2S approach over conventional reasoning and retrieval-based baselines.

First, generic chain-of-thought reasoning provides only a marginal improvement over the base model. CoT increases accuracy from 38.74% to 39.79%, while its Macro-F1 slightly decreases from 42.43% to 42.09%. This suggests that generic reasoning instructions alone are insufficient for this task, where successful interpretation requires domain-specific procedures for organizing and evaluating heterogeneous evidence.

Second, directly retrieving scholarly information does not consistently improve interpretation accuracy. RAG achieves 38.29% accuracy, slightly below the Base model, although its Macro-Precision increases to 51.47%. RAG+CoT further decreases accuracy to 41.03% and Macro-F1 to 41.61%. These results indicate that providing additional scholarly information, or combining such information with generic reasoning, does not necessarily translate into more accurate expert interpretation.

In contrast, our E2T2S approach achieves 51.57% accuracy and 56.16% Macro-F1, substantially outperforming all comparison methods. Compared with the Base model, E2T2S improves accuracy by 12.83 percentage points and Macro-F1 by 13.73 points. More importantly, under the same retrieved scholarly evidence, E2T2S outperforms RAG by 13.28 points in accuracy and 14.11 points in Macro-F1, and outperforms RAG+CoT by 10.54 points in accuracy and 14.55 points in Macro-F1. Since RAG, RAG+CoT, and E2T2S share the same retrieved scholarly evidence, these differences provide evidence that the gain is associated with the additional procedural guidance supplied by the Expert Skill Library rather than simply with access to more scholarly information. The improvement in Macro-F1 is also notable. E2T2S achieves 56.16% Macro-F1, compared with 42.43% for Base, indicating that the improvement is not limited to overall sample-level accuracy but is also reflected across different target-character classes. This is particularly relevant to our task because the same modern character may appear in multiple interpretation instances with different contextual and scholarly evidence. The higher Macro-F1 therefore suggests more consistent performance across the heterogeneous set of character interpretations represented in the evaluation set.

4.5 Evidence-Aware Evaluation

We stratify instances by the availability, diversity, polarity, and richness of external evidence and evaluate all methods on the same subsets, testing whether expert skills are especially beneficial when heterogeneous evidence must be coordinated or reconciled. We define seven conditions: No External Evidence; Single Evidence Type; Multiple Evidence Types (combining transmitted texts, variant readings, related glyphs, or phonological evidence); Supportive Evidence; Contrasting Evidence; Mixed Evidence (both supportive and contrasting signals in one instance); and Rich Evidence (dense, heterogeneous evidence under our richness criterion).

Method	Acc.(%)	Macro-P(%)	Macro-R(%)	Macro-F1(%)
E2T2S	51.57	66.50	51.16	56.16
w/o Explicit	44.53	53.48	45.93	48.25
w/o Tacit	27.11	39.39	25.75	30.30
w/o Skill	47.89	55.33	47.65	50.09

Table 3: Ablation study on the effectiveness of each stage in the E2T2S method, where 'w/o Explicit', 'w/o Tacit' and 'w/o Skill' refer to experiments in which the first, second and third stages, respectively, have been removed.

Table 2 reports the results. E2T2S achieves or ties the best performance under all seven conditions. Even without external evidence, E2T2S reaches 53.03% accuracy versus 42.42% for Base (+10.61 points), indicating that the benefit of extracted skills is not confined to processing retrieved evidence: the procedures also guide the organization and evaluation of intrinsic and contextual information in the task itself. The advantage widens when external evidence is available. Under Single Evidence Type, E2T2S attains 54.87%, outperforming the strongest conventional baseline (CoT, 47.56%) by 7.31 points. Under Multiple Evidence Types, E2T2S reaches 49.16% versus 38.98% for RAG+CoT (+10.18 points), consistent with our hypothesis that expert skills matter most when a decision depends on coordinating heterogeneous evidence rather than exploiting an isolated factual clue. Conventional RAG, by contrast, yields only limited gains over the base model here, suggesting that retrieving more information does not itself supply an effective integration procedure.

Evidence polarity shows a similar pattern. E2T2S scores 50.00% on Supportive Evidence and 48.33% on Mixed Evidence; the latter is central to our setting because it requires distinguishing supporting from counter evidence and determining how competing signals should revise the current interpretation. On Contrasting Evidence, E2T2S reaches 66.66%, matching RAG+CoT and surpassing Base and CoT, but this subset contains only six instances and should be read as descriptive. Finally, on Rich Evidence, E2T2S attains 50.44% versus 39.82% for RAG+CoT (+10.62 points), suggesting that merely exposing the model to more evidence is insufficient as heterogeneity grows—what is needed is a procedure for selecting relevant evidence and deciding how to incorporate it.

4.6 Ablation Study

We ablate each stage of E2T2S to examine its contribution. As shown in Table 3, the E2T2S achieves 51.57% accuracy and 56.16% Macro-F1. Without explicit knowledge extraction, we directly extract tacit knowledge from scholarly papers and use it as static procedural guidance; performance drops to 44.53% accuracy and 48.25% Macro-F1. Without tacit procedural knowledge extraction, structured explicit knowledge is directly used for retrieval-based augmentation, resulting in the largest degradation to 27.11% accuracy and 30.30% Macro-F1. Without skill construction, scholar-level tacit knowledge is directly injected into the prompt, yielding 47.89% accuracy and 50.09% Macro-F1. These results demonstrate that each stage contributes to the final performance, with the particularly large drop in w/o Tacit highlighting the importance of transforming explicit scholarly knowledge into reusable procedural knowledge.

5 Conclusion

We presented E2T2S, which turns expert knowledge in scholarly literature into reusable procedural skills for LLM reasoning—by extracting explicit structure, inferring tacit procedures from research practice, and organizing them in an Expert Skill Library for dynamic use. On Chinese Chu bamboo-slip character interpretation, E2T2S consistently outperforms reasoning and retrieval baselines; evidence-aware and ablation analyses show that procedural knowledge drives the gains on heterogeneous evidence. Broadly, scholarly literature can thus supply not only facts but reusable expert procedures for specialized scenarios.

AI USE STATEMENT

In this work, we used generative AI tools to extract and organize information from published literature on Chinese Chu bamboo-slip characters, and to clean and reformat it into structured records (evidence categories, instance annotations, and evaluation subsets). Specifically, LLMs were used to (1) locate and extract candidate evidence spans—transmitted texts, variant readings, related glyphs, and phonological notes—from digitized sources, (2) normalize them into a schema aligned with our evidence taxonomy, and (3) draft tabular summaries later finalized by hand. All AI-assisted outputs were inspected by domain experts; discrepancies were corrected manually, and unverifiable records were discarded. No synthetic data were generated, and no instances, labels, or evidence items in our evaluation set were created or altered by AI without human verification.

We have not used generative AI tools to generate datasets, formulate theoretical claims or hypotheses, design the methodology or experiments, implement the core method, interpret results, or write proofs. Additionally, we used generative AI tools for recommended-disclosure tasks: summarizing and surveying existing literature, sourcing information, and improving the readability of prose. Every passage produced or modified with AI assistance was reviewed and approved by the authors.

We take full responsibility for the final content of this work, including all text, claims, data records, and artifacts produced with the aid of generative AI.

REFERENCES

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, et al. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. *arXiv preprint arXiv:2411.14199*, 2024.

Faeze Brahman, Chandra Bhagavatula, Valentina Pyatkin, Jena Hwang, Xiang Lorraine Li, Hirona Arai, Soumya Sanyal, Keisuke Sakaguchi, Xiang Ren, and Yejin Choi. Plasma: Procedural knowledge models for language-based planning and re-planning. In *International Conference on Learning Representations*, volume 2024, pp. 5555–5585, 2024.

Xiaolei Diao, Zhihan Zhou, Lida Shi, Ting Wang, Ruihua Qi, Daqian Shi, and Hao Xu. Intechar: A unified oracle bone character list for ancient chinese language modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 211–219, 2026.

Haisu Guan, Huanxin Yang, Xinyu Wang, Shengwei Han, Yongge Liu, Lianwen Jin, Xiang Bai, and Yuliang Liu. Deciphering oracle bone language with diffusion models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15554–15567, 2024.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pp. 3929–3938. PMLR, 2020.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251): 1–43, 2023.

Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G Rodriques, and Andrew D White. Paperqa: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*, 2023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Bang Li, Donghao Luo, Yujie Liang, Jing Yang, Zengmao Ding, Xu Peng, Boyuan Jiang, Shengwei Han, Dan Sui, Peichao Qin, et al. Oracle bone inscriptions multi-modal dataset. *arXiv preprint arXiv:2407.03900*, 2024.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173, 2024.

Yang Liu, Jiahuan Cao, Hiuyi Cheng, Yongxin Shi, Kai Ding, and Lianwen Jin. Mcs-bench: A comprehensive benchmark for evaluating multimodal large language models in chinese classical studies. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10435–10492, 2025.

Laura Ruis, Maximilian Mozes, Juhan Bae, Siddhartha Rao Kamalakara, Dwaraknath Gnaneshwar, Acyr Locatelli, Robert Kirk, Tim Rocktäschel, Edward Grefenstette, and Max Bartolo. Procedural knowledge in pretraining drives reasoning in large language models. In *International Conference on Learning Representations*, volume 2025, pp. 29367–29429, 2025.

Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. Assisting in writing wikipedia-like articles from scratch with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6252–6278, 2024.

Daqian Shi, Xiaolei Diao, Xu Chen, and Cédric M John. Competitive distillation: A simple learning strategy for improving visual classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2981–2990, 2025a.

Lida Shi, Fausto Giunchiglia, Hongda Zhang, Daqian Shi, Rui Song, Jian Li, Xiaolei Diao, Alan Zhao, and Hao Xu. Learn from the best: A universal self-distillation approach with historical logits. *Expert Systems with Applications*, pp. 129340, 2025b.

Michael D Skarlinski, Sam Cox, Jon M Laurent, James D Braza, Michaela Hinks, Michael J Hammerling, Manvitha Ponnapati, Samuel G Rodriques, and Andrew D White. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740*, 2024.

Rui Song, Lida Shi, Ruihua Qi, Yingji Li, and Hao Xu. Enhancing multimodal large language models for ancient chinese character evolution analysis via glyph-driven fine-tuning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22789–22804, 2026.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

Rose Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. Bridging the novice-expert gap via models of decision-making: A case study on remediating math mistakes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2174–2199, 2024.

Zhihu Wang, Shiwan Zhao, Yu Wang, Heyuan Huang, Sitao Xie, Yubo Zhang, Jiaxin Shi, Zhixing Wang, Hongyan Li, and Junchi Yan. Re-task: Revisiting llm tasks from capability, skill, and knowledge perspectives. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4925–4936, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

Boyuan Zheng, Michael Y Fatemi, Xiaolong Jin, Zora Zhiruo Wang, Apurva Gandhi, Yueqi Song, Yu Gu, Jayanth Srinivasa, Gaowen Liu, Graham Neubig, et al. Skillweaver: Web agents can self-improve by discovering and honing skills. *arXiv preprint arXiv:2504.07079*, 2025.

Pei Zhou, Jay Pujara, Xiang Ren, Xinyun Chen, Heng-Tze Cheng, Quoc V Le, Denny Zhou, Swaroop Mishra, Huaixiu S Zheng, et al. Self-discover: Large language models self-compose reasoning structures. *Advances in Neural Information Processing Systems*, 37:126032–126058, 2024a.

Yanan Zhou, Guoqi Liu, Yiping Yang, Linyuan Ru, Dong Liu, and Xueshan Li. Ufcnet: Unsupervised network based on fourier transform and convolutional attention for oracle character recognition. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*, pp. 98–106, 2024b.

Zhihan Zhou, Daqian Shi, Rui Song, Lida Shi, Xiaolei Diao, and Hao Xu. Ancientbench: Towards comprehensive evaluation on excavated and transmitted chinese corpora. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35167–35175, 2026.

A Case Study

To complement the quantitative results, we conduct qualitative case studies to examine how dynamically selected expert skills influence the model's use of external evidence. Rather than focusing only on whether the final prediction is correct, we analyze whether the model identifies relevant evidence, applies the retrieved procedural skills, compares competing candidates, and uses the available evidence to reach the final decision. We present two representative cases involving contrasting and mixed evidence, respectively.

Case 1: Contrasting Evidence. The first case contains both transmitted-text and comparative scholarly evidence. The target character is interpreted from the masked context "bù [MASK]. nǎi…", with four candidates: *zhòng*, *dì*, *jū*, and a fourth rare character. The reasoning process first inspects the target-glyph image and contextual constraint, recognizing that the masked position is likely to require a verb or adjective associated with the meaning of "avoid" or "exempt." The model then identifies the external evidence, including a transmitted-text reference from *Erya* and a comparative reading proposed by the editor.

Two skills are subsequently activated. **SK-U04 (contextual semantic constraint)** is used to infer the grammatical and semantic requirements imposed by the local context, while **SK-U06 (candidate comparison and selection)** is used to compare the four candidates against these constraints. The model progressively eliminates *zhòng* and *dì* because their meanings are incompatible with the contextual interpretation, and rejects the fourth candidate because its usage and graphical correspondence are insufficiently supported. It retains *jū* because the model interprets it as compatible with the required "avoid/exempt" meaning and with the observed glyph structure. Importantly, the model explicitly recognizes that the comparative evidence does not directly establish *jū* as the answer and treats it only as auxiliary evidence rather than as a direct answer source. The final prediction is therefore *jū*.

This case illustrates a key property of skill-guided reasoning: the retrieved scholarly evidence is not simply copied into the answer. Instead, the model first determines its relevance, applies a contextual constraint, and then uses candidate-level comparison to narrow the decision space. The resulting reasoning chain is thus structured around **evidence-to-decision operations** rather than direct retrieval of an answer-bearing statement.

Case 2: Mixed Evidence. The second case contains multiple heterogeneous evidence sources and both supportive and potentially constraining information. The target occurs in the context "sì fāng yí mò [MASK] rén miàn jiàn Tāng", with candidates including *zhòng*, *dì*, *dì*, and *jù*. The model first examines the target image and local context, identifying both a graphical correspondence with *jù* and a contextual requirement for a verbal or adjectival interpretation. It then identifies three forms of external evidence: a transmitted-text reference concerning the meaning of *jù*, a variant-reading argument that eliminates an alternative interpretation, and comparative evidence relating another character to the meaning "stop."

The model activates three skills in sequence. **SK-U04 (contextual semantic constraint)** establishes the semantic and syntactic requirement of the masked position. **SK-E047 (elimination)** removes candidates whose grammatical category or semantic interpretation is incompatible with the context. **SK-E077 (parallel-text comparison)** then examines the scholarly evidence associated with *jù* and uses the interpretation "stop" to evaluate its compatibility with the surrounding passage. As a result, *zhòng*, *dì*, and *dì* are progressively excluded, while *jù* is retained because its graphical form, contextual interpretation, and supporting scholarly evidence are mutually compatible. The model also explicitly notes the absence of a direct parallel passage and therefore uses the available transmitted-text evidence as supporting rather than conclusive evidence. The final prediction is *jù*.

This case demonstrates how multiple expert skills can operate as complementary procedures rather than as independent instructions. Contextual constraints first narrow the interpretation space, elimination rules remove incompatible candidates, and comparative evidence is subsequently used to strengthen the remaining hypothesis. The model therefore transforms heterogeneous evidence into a sequence of conditional decisions, which is the central behavior that our tacit procedural knowledge is intended to capture.

Taken together, these cases illustrate the qualitative mechanism underlying the quantitative gains of Expert Skill. In both examples, the model does not merely receive additional scholarly information; it uses dynamically selected procedures to determine **which evidence to inspect, how to apply it to competing interpretations, and how to update the current hypothesis**. Such behavior provides qualitative support for our central hypothesis that tacit procedural knowledge extracted from expert literature can serve as a reusable intermediate layer between evidence access and expert decision making.

B Implementation Details

For reproducibility, we report all experimental configurations used in the final evaluation. We conduct 3 independent runs with different random seeds and report mean standard deviation. The base model is Qwen3-VL-8B-Instruct with identical decoding settings across conditions, including temperature 0.2, top-p of 0.8, maximum new tokens of 4096, and random seed. All inference outputs, retrieval results, and selected skills are cached to support exact reruns. The evaluation pipeline uses the same candidate sets, image inputs, contextual information, and external evidence for all conditions; invalid predictions outside the candidate set are counted as incorrect. For retrieval-based conditions, the retrieval corpus contains only documents in the corpus set, and the top-k retrieval setting is fixed to 3. Retrieval is performed once per test instance and the resulting evidence chunks are shared across RAG, RAG+CoT, and Expert Skill. Expert Skill selection uses the top-6 skills selected from the Expert Skill Library according to the evidence-state matching score. Only the selected skills are injected, while the complete library is never exposed to the model.