DeepWeaver: Bridging the Evidence Synthesis Gap in Open-Ended Question Answering

Anonymous ACL submission

Abstract

Retrieve-then-generate pipelines are commonly used to produce deep-research answers for open-ended questions, but retrieval alone is insufficient: LLMs must organize noisy and fragmented evidence into comprehensive, well-cited answers. We refer to this process as **evidence synthesis**. However, direct generation often underuses evidence, misaligns citations, and collapses diverse information into shallow summaries, exposing an evidence synthesis gap between retrieval and generation. Thus, we propose **DeepWeaver**[1], a novel framework that weaves noisy retrieved evidence into comprehensive answers by maintaining Thought Block Chains (TBCs), a structured representation that groups claims, salient information, keywords, and supporting evidence. DeepWeaver uses subordinate TBCs to inspect residual evidence, commit TBC revisions, and discover new claims before final generation. We evaluate DeepWeaver on open-ended QA over both knowledge bases and the web, and introduce **LoQA**, a high-density benchmark for evidence synthesis. Across multiple LLMs, DeepWeaver improves content sufficiency, citation grounding, and detail preservation on LoQA, while achieving deeper insights and higher citation quality on DeepResearch Bench. These results show that evidence weaving is an effective mechanism for bridging retrieval and generation in open-ended QA.

1 Introduction

Large language models (LLMs) are increasingly being used to generate research-style answers grounded in external evidence (Menick et al., 2022; Asai et al., 2024a). For domain-specific open-ended questions and deep-research-style tasks, answering a query may require integrating evidence from hundreds of retrieved passages, web pages, books, or technical documents. As a result, evidence retrieval and synthesis are central to both

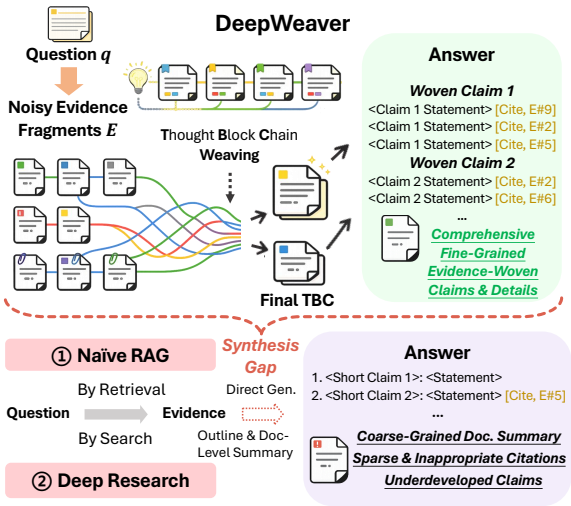


Figure 1: DeepWeaver bridges the evidence synthesis gap between retrieval and generation by weaving noisy evidence into structured Thought Block Chains (TBCs). Unlike naïve retrieval-augmented generation, DeepWeaver synthesizes noisy evidence fragments into answers with well-supported woven claims, fine-grained details, and grounded citations.

retrieval-augmented generation (RAG) pipelines (Lewis et al., 2021; Izacard and Grave, 2021) and web-based deep-research agents (Nakano et al., 2021; Yao et al., 2022). These systems share a common workflow: retrieve evidence, place it into the context, and prompt the model to synthesize a comprehensive answer (Huang and Huang, 2024).

Although prior work has extensively studied how to improve evidence retrieval quality (Han et al., 2024; Guo et al., 2024; Sun et al., 2025), retrieval itself is merely the first step: **enabling LLMs to effectively use *extensive, noisy, and fragmented* evidence to generate comprehensive answers remains a key challenge**. To better integrate upstream evidence, context-refinement methods (Xu et al., 2023; Jin et al., 2025b) compress, filter, or restructure passages before generation, while divide-and-conquer methods (Zhang et al., 2024b; Zhou et al., 2024) split long contexts into segments and

[1]The source code and data will be publicly available.

aggregate intermediate outputs. End-to-end deep-research systems (Wang et al., 2024c; Tao et al., 2025; Li et al., 2025b) go further by automating evidence organization and long-form answer writing with agent memory and outline-guided generation.

However, prior methods overestimate LLMs' ability to fully exploit input evidence to derive in-depth, diverse claims. By primarily optimizing evidence presentation and outline-guided writing (Xu and Peng, 2025), these approaches often fall short in deeply synthesizing retrieved evidence into well-developed answers. When faced with noisy, fragmented, and knowledge-dense evidence, relying solely on the LLM's internal reasoning often obscures fine-grained details and produces under-developed claims. This disconnect between raw retrieved evidence and comprehensive generation is what we define as **the evidence synthesis gap**.

We argue that *an intermediate module for orchestrating evidence and weaving evidence-grounded claims* is essential for bridging the gap between retrieval and final answer generation (Figure 1). The central idea is to structure noisy retrieved evidence into fine-grained thought units, rather than synthesizing answers in a single pass.

To this end, we propose **DeepWeaver**, a novel framework for open-ended answer generation grounded in noisy evidence. DeepWeaver maintains a **T**hought **B**lock **C**hain (TBC) structure, where each thought block corresponds to a discovered aspect of the answer and stores its claim, keywords, salient information, and supporting evidence fragments. DeepWeaver weaves TBCs by identifying overlooked evidence in the current TBC and generating subordinate TBCs to inspect residual evidence that has not been sufficiently covered. Newly discovered information and claims are then committed back into the main TBC. This design produces evidence-woven claims and links these claims to relevant fragments, thereby improving answer comprehensiveness and citation quality while alleviating context-window pressure.

We evaluate DeepWeaver on our **LoQA** benchmark (§2.2), an open-ended answer generation benchmark with extensive noisy evidence. LoQA contains 100 water-environment research questions, each paired with evidence retrieved from a Chinese expert knowledge base comprising 500 books and over 100M characters. It stress-tests whether QA systems can comprehensively use evidence, preserve domain-specific details, and generate well-cited answers. We also evaluate DeepWeaver on web-based deep-research questions using **DeepResearch Bench** (Du et al., 2025), which contains PhD-level research tasks across diverse fields. Experiments show that DeepWeaver substantially improves content sufficiency, citation grounding, and detail preservation on LoQA, while achieving deeper insights and higher citation quality than advanced agents on DeepResearch Bench.

Our main contributions include:

- We identify the evidence synthesis gap in open-ended QA, highlighting the need for an explicit synthesis stage beyond retrieval.
- We propose DeepWeaver, a framework that orchestrates noisy retrieved evidence into structured Thought Block Chains and generates open-ended answers via evidence weaving.
- We introduce LoQA, a benchmark for evaluating evidence-grounded answer generation. Experiments show that DeepWeaver improves answer comprehensiveness and citation quality on both LoQA and DeepResearch Bench.

2 Task Formulation

We evaluate whether a QA system can synthesize a large set of evidence fragments into comprehensive open-ended answers. This setting reflects realistic information-seeking needs: answering a query may involve multiple claims, while the supporting evidence is scattered across sources (*e.g.*, books, technical documents, and passages). Although RAG techniques can retrieve such evidence, whether QA systems can fully exploit it is unclear.

2.1 Definition

Given an open-ended research-style question q and a retrieved evidence pool $E = \{e_1, e_2, \ldots, e_N\}$ collected from relevant documents, a QA system is required to generate an answer y that thoroughly resolves q while being grounded in the evidence. Answers are evaluated according to three aspects that characterize evidence synthesis quality:

- **Content sufficiency**: the answer should cover the major content and claims implied by the question and synthesized from the evidence.
- **Detail preservation**: the answer should retain fine-grained information rather than collapse it into generic summaries.
- **Citation grounding**: the answer should ground its claims in relevant evidence.

2.2 The LoQA Benchmark

LoQA is built from a Chinese water-environment domain knowledge base containing 500 books and over 100M characters. We choose this domain because it contains dense technical knowledge, complex causal relations, and practical open-ended questions, which involve evidence across multiple subtopics, such as pollution control, water quality assessment, and watershed management.

Construction. LoQA contains 100 open-ended research-style questions. We generate 3,000 candidate questions using multiple state-of-the-art LLMs based on book contexts, and then filter them using LLM-based scores for retrieval comprehensiveness, concreteness, domain depth, and expression completeness. We retrieve evidence for each candidate and select the final questions with the highest average similarity to the retrieved fragments, ensuring that each question involves rich evidence (§A.1.2).

About 100 evidence fragments (chunked into 1,024-token segments) are retrieved for each question. To simulate a noisier context, we add another 100 random evidence fragments[2].

Gold answers for long-context, open-ended, domain-specific questions are difficult to produce and evaluate. Inspired by prior work (Shao et al., 2024), we therefore evaluate evidence synthesis quality by assessing answer comprehensiveness along several controlled dimensions. Specifically, LoQA decomposes this subjective notion into three dimensions: (1) **citation grounding**, (2) **content sufficiency**, and (3) **detail preservation**.

Citation grounding. We use a strong LLM judge, DeepSeek-V3.2, to determine whether each evidence fragment is relevant to the question. This splits the evidence pool E into a relevant set E_R and an irrelevant set E_I. Given a generated answer y, let C denote the set of evidence fragments cited by y. We compute three citation metrics: (1) **Citation Count (CC)**: $|C|$, (2) **Relevant Count (RC)**: $|C \cap E_R|$, and (3) **Relevant Ratio (RR)**: $|C \cap E_R|/|C|$. These metrics measure whether the system actively uses evidence and properly avoids citing or integrating irrelevant evidence.

Content sufficiency. This dimension evaluates whether the answer sufficiently covers the key terms and arguments synthesized from the evidence. To obtain reference answers, we use DeepSeek-V3.2 to generate five answers from different partitions of the relevant evidence set E_R:

$$E_R^{0\text{-}25\%},\ E_R^{25\text{-}50\%},\ E_R^{50\text{-}75\%},\ E_R^{75\text{-}100\%},\ E_R^{0\text{-}100\%}.$$

This yields five reference answers y_1, y_2, y_3, y_4, y_5.

We compute two lexical-level metrics:

(1) **Recall**, the average of the longest common subsequence (LCS) ratio:

$$\frac{1}{5}\sum_{i=1}^{5}\frac{\mathrm{LCS}(y_i, y)}{|y_i|}.$$

(2) **Word Alignment (WA)**, the average BLEU-4 score, adopting the generated answer y as the target and each reference answer y_i as the source:

$$\frac{1}{5}\sum_{i=1}^{5}\text{BLEU-4}(y_i, y).$$

We also compute the **Argument Sufficiency (AS)** metric, using an LLM to extract a set of atomic arguments A from reference answers $y_1, \ldots, y_5$. Then we judge whether each argument in A is covered by the generated answer y via LLM:

$$\frac{1}{|A|}\sum_{a\in A}\mathbb{I}(a \text{ is covered by } y).$$

Detail preservation. We evaluate whether the generated answer preserves fine-grained details from the evidence pool. We first use an LLM to create cloze-style questions from the answer y_5. The blanks mainly target technical terms, concepts, and short phrases. We judge whether each blank can be correctly recovered by providing the generated answer y as the reference document. **Detail Preservation (DP)** is the proportion of cloze blanks in B that can be correctly recovered:

$$\frac{1}{|B|}\sum_{b\in B}\mathbb{I}(b \text{ can be recovered from } y).$$

3 Method

3.1 Thought Block Chain

We define the **T**hought **B**lock **C**hain (TBC) as an explicit data structure that bridges retrieved evidence and final answer generation. The TBC serves as the core information structure maintained by DeepWeaver, decomposing the answer into a sequence of thought blocks:

$$\mathcal{T} = [b_1, b_2, \ldots, b_K].$$

[2]Each question is paired with about $200 \times 1024 = 205$K tokens of evidence, approaching the context window of many LLMs.

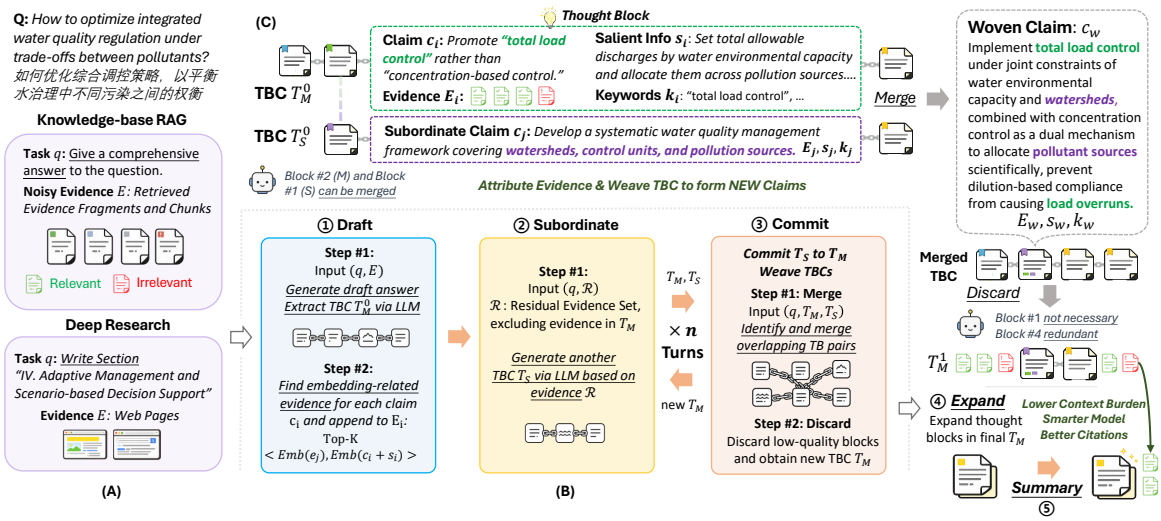


Figure 2: **Overview of DeepWeaver.** (A) We apply DeepWeaver to two typical retrieve-then-generate scenarios. (B) DeepWeaver maintains and refines the TBC through three stages: Draft, Subordinate, and Commit. (C) A concrete example of commitment shows how DeepWeaver synthesizes new claims from residual evidence by TBC weaving.

Each block b_i encapsulates a candidate claim and stores the corresponding keywords, salient information, and supporting evidence fragments:

$$b_i = (c_i, k_i, s_i, E_i),$$

where c_i, k_i and s_i denote the claim, keywords, and salient information, respectively, and $E_i \subseteq E$ is the subset of evidence fragments linked to the claim.

Unlike a plain outline that tends to provide a survey rather than a question-resolving answer, the TBC organizes detailed evidence into fine-grained woven claims, while maintaining an explicit mapping between claims and relevant evidence and reducing the context burden during generation.

3.2 Evidence Weaving

The core mechanism of DeepWeaver (Figure 2(B)) lies in the maintenance of a main TBC $\mathcal{T}_M$. It updates $\mathcal{T}_M$ through multi-stage refinement. Overall, DeepWeaver consists of three ordered weaving stages: Draft, Subordinate, and Commit.

Draft. Given the question q and the retrieved evidence pool E, the LLM first produces an answer draft via direct generation. DeepWeaver then extracts the **initial main TBC** $\mathcal{T}_M^0$ from this draft. This step leverages the LLM's global view of the evidence pool, allowing it to form an initial set of claims from a broad perspective.

Subordinate. The initial TBC $\mathcal{T}_M^0$ may be incomplete, as direct generation can overlook useful evidence and omit important claims. The subordinate stage is designed to identify evidence fragments that are neglected, weakly covered, or insufficiently reflected in the current TBC. We define the **residual evidence set** $\mathcal{R}_0$ of the initial TBC $\mathcal{T}_M^0$ as:

$$\mathcal{R}_0 = \{e_j \in E \mid e_j \text{ is not covered by } \mathcal{T}_M^0\}.$$

We say an evidence fragment is covered by the initial TBC $\mathcal{T}_M^0$ if it satisfies either: (1) it is directly mentioned in $\mathcal{T}_M^0$; or (2) it is among the top-k fragments most similar to the string $c_i + s_i$ in the embedding space, and the similarity is higher than the average similarity of the evidence E_i in the corresponding block $b_i \in \mathcal{T}_M^0$. The complement of the covered evidence set is defined as $\mathcal{R}_0$.

DeepWeaver generates a **subordinate TBC** $\mathcal{T}_S^0$ (Figure 2 ②) from the evidence in $\mathcal{R}_0$. This procedure acts as a local inspector of overlooked evidence. It discovers missing aspects, additional details, and alternative supporting evidence that are not captured by the main TBC.

Commit. The useful information in $\mathcal{T}_S^0$ is then committed back to the main TBC $\mathcal{T}_M^0$ through two operations: **Merge** and **Discard** (Figure 2 ③).

The merge operation identifies overlapping claim pairs between the two TBCs and merges their claims c_i and keywords k_i, while their salient information s_i and evidence sets E_i are combined to form a merged block. *This produces woven claims that integrate multi-dimensional explanations.*

After merging, the discard operation removes irrelevant, redundant, or weakly supported claims with respect to the question, preventing TBC from over-expanding. This yields a **refined TBC** $\mathcal{T}_M^1$:

$$\textsc{Discard}\big(q, \textsc{Merge}(\mathcal{T}_M^0, \mathcal{T}_S^0)\big) \rightarrow \mathcal{T}_M^1.$$

DeepWeaver repeats the subordinate and commit stages for n rounds to refine the woven claims and incorporate additional supporting evidence[3]:

$$\mathcal{T}_M^0 \rightarrow \mathcal{T}_M^1 \rightarrow \cdots \rightarrow \mathcal{T}_M^t \rightarrow \cdots \rightarrow \mathcal{T}_M^n.$$

Furthermore, to reduce the context burden during TBC construction, the draft and subordinate stages randomly sample r ($r < |E|$) evidence fragments from E and $\mathcal{R}_t$ (the residual evidence set at refinement turn t), respectively. Across multiple revision rounds, the model can inspect the full evidence pool while avoiding the burden of processing an excessively long context in one pass.

3.3 Evidence-Grounded Answer Generation

After evidence weaving, DeepWeaver generates the final answer from the refined TBC $\mathcal{T}_M^n$ (Figure 2 ④–⑤). For each block b_i, the model receives its metadata and linked evidence subset E_i as a focused local context, and then generates an **answer section** S_i grounded in the corresponding evidence:

$$S_i = \textsc{Generate}(q, c_i, s_i, E_i).$$

The final answer[4] is composed of all generated sections using the LLM to build links among S_i:

$$y_i = \textsc{Append}(y_{i-1}, S_i),\ y = y_K$$

This block-wise generation strategy has two advantages: (1) it decomposes context pressure into claim-level generation, and (2) it improves citation grounding because each section is generated from a smaller and more relevant evidence subset.

4 Experiments

We evaluate DeepWeaver on both knowledge-base open-ended question answering and end-to-end web-based open-ended deep research.

4.1 Datasets

LoQA. A high-density evidence-based QA benchmark (§2.2) with 100 research-style practical water-environment questions. Each question is paired with long-context evidence retrieved from a 500-book knowledge base, evaluating whether systems can deeply synthesize fragmented retrieved evidence to generate comprehensive answers.

DeepResearch Bench. It contains 100 PhD-level deep-research tasks across 22 fields, where evaluated agents search for web information and write long-form answers. The **RACE** split measures answer quality using LLM-judged criteria, including comprehensiveness, insight, instruction following, and readability[5], while **FACT** measures citation accuracy and the number of effective citations.

4.2 Baselines

We compare DeepWeaver with representative evidence-based QA baselines on LoQA.

RAG. An LLM directly generates the answer from evidence. We consider three context settings: the full evidence pool E, the relevant evidence set E_R, and a randomly sampled subset E_{120} containing 120 chunks (Lewis et al., 2021).

Skeleton-of-Thought (SoT). The model first generates a thought outline and then writes the answer section by section, with the full evidence pool provided in the input context (Ning et al., 2024).

Plan-and-Solve (PaS). The model generates a plan to produce the answer (Wang et al., 2023a).

Chain-of-Agents (CoA). A divide-and-conquer framework that splits the evidence pool E into smaller subsets, processes them batch by batch, and gradually refines the answer (Zhang et al., 2024b).

LongRefiner. A method that preprocesses the evidence pool E by extracting structured articles and sentences from noisy evidence (Jin et al., 2025b).

For DeepResearch Bench, baselines include advanced agents[6] and WebWeaver (Li et al., 2025b), a dual-agent framework that uses a planner to search the web and write outlines, and a writer to fulfill the outline using a memory bank.

4.3 Implementation Details

We evaluate DeepWeaver under multiple backbones, including Qwen3-30B-A3B-Instruct-2507, Qwen3-30B-A3B-Thinking-2507 (Yang et al., 2025), DeepSeek-V3.2 (Liu et al., 2024a), DeepSeek-V4-Flash (DeepSeek-AI, 2026), and Qwen3.5-122B-A10B. We set refinement rounds to $n = 2$, the top-k covered evidence retrieval size to $k = 5$, and the number of randomly sampled

[3]Note that TBC maintenance is lightweight, adding minor output-token overhead compared with answer generation.

[4]In this work, we focus on evidence synthesis quality and answer comprehensiveness. In settings such as DeepResearch Bench, where basic readability is also required, we add simple optional steps including (1) grouping TBC blocks into article sections, (2) polishing the generated sections, and (3) adding introduction and conclusion paragraphs.

[5]Here, "comprehensiveness" and "insight" also correspond to two aspects of evidence synthesis quality under our definition (§ 2.1).

[6]doubao-research, kimi-research, Claude-research, and openai-deepresearch.

Method	► *Qwen3-30B-A3B-Instruct-2507*							► *Qwen3-30B-A3B-Thinking-2507*						
	Content S.			Citation G.			Detail	Content S.			Citation G.			Detail
	Recall	WA	AS	CC	RC	RR	DP	Recall	WA	AS	CC	RC	RR	DP
E_{120}-RAG	20.4	4.0	70.6	17.7	13.6	78.9	56.4	24.0	5.6	64.8	12.8	10.8	85.0	51.0
E-RAG	19.8	4.0	68.0	11.9	8.8	74.3	52.9	21.9	5.0	59.9	10.6	9.0	75.3	53.9
E_R-RAG	21.8	4.5	76.5	21.4	21.4	100	58.8	24.6	5.9	68.0	16.3	16.3	100	55.9
SoT	20.3	4.2	73.4	13.8	7.7	53.6	58.7	30.1	**8.2**	61.8	24.2	15.4	64.5	29.5
PaS	13.8	2.0	58.4	9.9	4.8	48.4	44.8	10.9	2.0	33.2	6.5	4.0	35.8	46.3
Chain-of-Agents	14.7	2.4	70.3	19.2	15.3	81.2	60.2	8.2	1.1	34.4	8.6	7.8	91.0	30.3
LongRefiner	21.0	4.0	78.2	21.1	16.7	79.7	57.6	23.9	5.4	69.0	16.7	14.7	87.5	36.4
DeepWeaver	**27.4**	**6.1**	**83.5**	**26.6**	**23.5**	**88.8**	**69.5**	**34.5**	8.0	**85.1**	**30.9**	**28.6**	**92.3**	**69.1**
Δ over E-RAG	+7.6	+2.1	+15.5	+14.7	+14.7	+14.5	+16.6	+12.6	+3.0	+25.2	+20.3	+19.6	+17.0	+15.2

Table 1: Comprehensiveness of answers generated by evidence-based QA systems on LoQA, reflecting evidence synthesis quality. Results are shown in terms of content sufficiency (Content S.), citation grounding (Citation G.) and detail preservation (Detail). Metric definitions are provided in §2.2. The best results are highlighted in **bold**.

chunks to $r = 120$. We use GLM-Embedding-3 (Team et al., 2025) to compute text embeddings. The maximum generation length is 32,768 tokens.

For LoQA evaluation, we use DeepSeek-V3.2 as the judge for Argument Sufficiency (AS) and Detail Preservation (DP), with temperature set to 0. For DeepResearch Bench, we follow the official setting and use Gemini-2.5-Pro (Comanici et al., 2025) for RACE and Gemini-2.5-Flash for FACT evaluation. More implementation details for DeepResearch Bench are provided in Appendix §A.2.

4.4 Main Results

Table 1 reports the main results on LoQA. DeepWeaver consistently improves evidence synthesis quality across both instruct and thinking backbone models. **DeepWeaver substantially outperforms direct generation and evidence-based QA baselines.** With Qwen3-30B-A3B-Instruct-2507, DeepWeaver surpasses E-RAG by 15.5% in Argument Sufficiency, +14.7 Relevant Citations, 14.5% in Relevant Citation Ratio, and 16.6% in Detail Preservation. These gains show that DeepWeaver improves answer quality across multiple dimensions, including content sufficiency, citation grounding, and detail preservation.

We further highlight the evidence synthesis gap: direct generation is insufficient for using noisy RAG evidence. For both backbones, E-RAG does not outperform E_{120}-RAG, despite receiving more evidence. This suggests that adding more retrieved chunks can increase the context burden and cause the model to overlook useful information. The oracle E_R-RAG setting receives only relevant evidence, yet it still lags behind DeepWeaver. This indicates that removing noise alone is insufficient to solve the evidence synthesis problem.

Among the reasoning and divide-and-conquer baselines, SoT struggles to preserve fine-grained details and identify relevant evidence under heavy context burden. PaS is less effective because open-source models often struggle to produce reliable writing plans for complex, evidence-heavy questions. CoA and LongRefiner improve citation usage in some cases by reducing context burden or refining evidence, but they still suffer from insufficient claims and missing details. These results support our main claim: **comprehensive answer generation requires not only retrieving or filtering evidence, but also an effective evidence synthesis process** that weaves fine-grained evidence into coherent claim-level structures.

4.5 Ablation Study

We ablate two core mechanisms of DeepWeaver on Qwen3-30B-A3B-Instruct-2507: TBC subordination and commitment (§3.2). In *w/o subordinate*, the system directly expands the initial TBC $\mathcal{T}_M^0$ into the final answer. In *w/o commit*, the system appends the subordinate TBC $\mathcal{T}_S^0$ to $\mathcal{T}_M^0$, without the merge and discard operations.

	AS	RC	RR	DP	#Blocks
DeepWeaver	83.5	**23.5**	88.8	**69.5**	6.63
w/o TBC subordinate	78.6	21.2	**89.2**	62.4	5.76
w/o TBC commit	**84.4**	22.3	85.8	67.9	10.2

Table 2: Ablation study on TBC subordination and commitment. Scores in red highlight variants with severe performance degradation.

Table 2 shows that removing subordination hurts all three dimensions of evidence synthesis quality, indicating that the initial TBC is often incomplete and requires further refinement to recover overlooked evidence. Removing commitment makes answers less concise and weakens claim weaving: it achieves comparable AS and DP while producing more blocks, suggesting that it introduces overlapping and redundant claims.

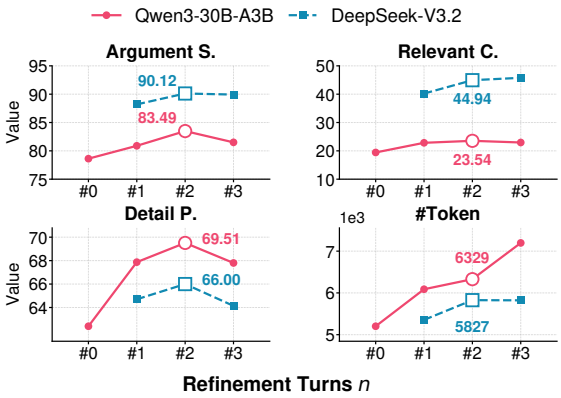


Figure 3: Effect of refinement rounds on final performance. Increasing the number of refinement rounds from $n = 1$ to $n = 2$ improves answer comprehensiveness, while a third round brings no clear additional gain. $n = 2$ provides the best cost–performance trade-off.

We also study the optimal refinement rounds n. As shown in Figure 3, DeepWeaver remains stable across different values of n, with $n = 2$ achieving the best overall performance. Increasing the number of rounds from one to two improves content sufficiency, citation grounding, and detail preservation. This indicates that the second subordinate-and-commit step helps enrich the claims and final TBC. A third round yields no substantial improvements, possibly because excessive refinement introduces marginally useful information. Considering both performance and efficiency, we set $n = 2$ by default. **With only two rounds of refinement over the initial TBC, DeepWeaver brings large gains.** This demonstrates the efficiency of the evidence weaving mechanism.

4.6 Cross-Model Generalizability

Table 3 evaluates whether DeepWeaver generalizes across different LLM backbones. The results show that DeepWeaver brings consistent gains across backbone models. For example, on DeepSeek-V4-Flash, DeepWeaver improves Recall from 24.2 to 31.6, AS from 75.4% to 86.5%, RC from 31.0 to 44.6, and DP from 60.0% to 67.4%.

These results indicate that the **benefit of DeepWeaver does not depend on a specific backbone model**. Instead, evidence weaving provides a general mechanism for improving the use of noisy evidence. Notably, DeepWeaver with Qwen3-30B-A3B-Instruct-2507 outperforms direct generation with DeepSeek-V3.2, even though DeepSeek-V3.2 is a strong long-context model. This shows that explicitly weaving evidence is more effective than relying solely on stronger long-context capabilities.

Method	Content S.			Citation G.			Detail
	Recall	WA	AS	CC	RC	RR	DP
▶ *DeepSeek-V4-Flash*							
w/o DW	24.2	6.1	75.4	35.4	31.0	88.2	60.0
w DW	**31.6**	**10.6**	**86.5**	**50.5**	**44.6**	**88.4**	**67.4**
▶ *Qwen3.5-122B-A10B*							
w/o DW	21.7	4.6	77.7	32.4	28.9	89.2	53.5
w DW	**29.0**	**7.3**	**87.7**	**49.9**	**44.4**	**89.3**	**64.6**
▶ *DeepSeek-V3.2*							
w/o DW	24.5	5.8	78.5	27.0	21.7	81.7	59.6
w DW	**32.4**	**10.3**	**90.1**	**51.8**	**44.9**	**86.7**	**65.7**
▶ *Qwen3-30B-A3B-Instruct-2507*							
w DW	27.4	6.1	83.5	26.6	23.5	88.8	69.5

Table 3: Generalizability of DeepWeaver across recent advanced LLMs. DeepWeaver consistently yields substantial gains across different backbone models. Notably, DeepWeaver with Qwen3-30B-A3B-Instruct-2507 substantially outperforms direct generation with DeepSeek-V3.2, a strong long-context model.

4.7 Extension to Web-Based Deep Research

Web-based deep-research agents also follow a retrieve-then-generate paradigm: search the web, collect evidence, and synthesize long-form answers from open-domain sources. We study whether DeepWeaver can serve as an evidence-weaving module to improve deep-research answer quality.

Since retrieval is not the focus of DeepWeaver, we reuse the web pages and top-level section titles collected by WebWeaver's planner for simplicity and fair comparison[7]. Each question is decomposed into section-level writing tasks. For each section, DeepWeaver treats the collected web pages as evidence fragments and generates the corresponding top-level answer section (§A.2).

Unlike WebWeaver, which summarizes evidence into coarse-grained outline titles at the webpage level, DeepWeaver operates on fine-grained evidence chunks and weaves them into detailed claim-level structures. As shown in Table 4, DeepWeaver achieves the best overall RACE score. **It weaves evidence more deeply into insightful and comprehensive claims**. Notably, our method also **substantially improves effective citation and citation accuracy** over WebWeaver on FACT, indicating that it organizes web evidence into reliable support for answer claims. †However, denser citations and more in-depth claims may affect readability, and the metric may favor an outline-driven writing style. Since readability optimization is not the primary focus of this paper, we leave it to future work.

[7]In terms of cost, DeepWeaver avoids using a summary model to extract webpage-level evidence during evidence collection, substantially reducing the planning-stage cost bottleneck. See §A.3 for the cost comparison.

Agent	RACE					FACT	
	Comp.	Insight	Inst.	Read.†	Overall	Eff. c.	C. acc.
doubao-research	44.84	40.56	47.95	44.69	44.34	52.62	52.86
kimi-research	44.96	41.97	47.14	45.59	44.64	-	-
Claude-research	45.34	42.79	47.58	44.66	45.00	-	-
openai-deepresearch	**46.46**	43.73	49.39	47.22	46.45	39.79	**75.01**
WebWeaver (Qwen3-30B-A3B-Instruct-2507)	45.15	45.78	49.21	**47.34**	46.77	26.74	25.00
DeepWeaver (Qwen3-30B-A3B-Instruct-2507)	45.99	**47.59**	**49.41**	42.54	**47.05**	**60.13**	62.02

Table 4: Results on DeepResearch Bench. DeepWeaver improves comprehensiveness (Comp.) and insight, and achieves substantially better effective citation (Eff. C.) and citation accuracy (C. Acc.) than WebWeaver. Better scores are highlighted in green. All scores are assigned by LLM judges according to detailed expert-level criteria.

Despite this trade-off, the overall performance gains confirm that **evidence weaving also benefits web-based deep research** as a general downstream module beyond knowledge-base QA.

5 Related Work

Retrieval-Augmented and Attributed Generation. Retrieval-augmented generation is a common way to incorporate external knowledge in question answering (Guu et al., 2020; Ram et al., 2023; Xu et al., 2024). Sparse vector retrieval (Robertson and Zaragoza, 2009; Wang et al., 2023b; Schick et al., 2023) retrieves relevant text through term-frequency or keyword matching, while dense vector retrieval (Lewis et al., 2021; Hofstätter et al., 2022; Liu et al., 2025) encodes the query and text into vectors. Recent advances further improve retrieval through self-reflective retrieval control, graph-structured evidence organization, interleaved retrieval and reasoning, and search-based planning (Asai et al., 2024b; Han et al., 2024; Wang et al., 2024d; Hu et al., 2025).

Long-Context Evidence Integration. Larger context windows allow models to include more retrieved evidence, but long-context models may still overlook relevant information in long or noisy inputs (Liu et al., 2024b; Hsieh et al., 2024; Modarressi et al., 2025). A line of work improves how upstream retrieved evidence is presented to the model. Context-refinement methods compress, filter, or restructure retrieved passages before generation, including RECOMP, Chain-of-Note, and LongRefiner (Xu et al., 2023; Yu et al., 2024; Jin et al., 2025b). Divide-and-conquer methods (Zhang et al., 2024b; Zhou et al., 2024; Guo et al., 2025) split long contexts into smaller segments and aggregate intermediate outputs. These methods reduce context burden by refining input with coarser evidence representations (Jiang et al., 2024), but leave the evidence synthesis gap underexplored.

Long-Form Grounded Writing and Deep-Research Agents. Long-form grounded writing has been studied through outline-guided writing and reflection-driven generation, mainly for survey generation (Yan et al., 2025; Skarlinski et al., 2024; Shao et al., 2024; Wu et al., 2025), but these methods are not primarily designed to answer open-ended questions. Citation-oriented work further improves attribution and verifiability of long-form answers (Huang et al., 2024; Ye et al., 2024). Recent deep-research agents extend long-form writing to end-to-end QA systems: one line of work focuses on search timing and search-reasoning interaction (Li et al., 2025a; Jin et al., 2025a), another combines evidence collection with outline-based answer writing (Li et al., 2025b), and others improve agent capabilities through agentic supervised fine-tuning and reinforcement learning (Wu et al., 2026; Zheng et al., 2025).

6 Conclusion

We present DeepWeaver, a framework for bridging the evidence synthesis gap in open-ended question answering. Rather than treating retrieved evidence as a flat long-context input, DeepWeaver organizes noisy evidence into Thought Block Chains (TBCs), which connect fine-grained evidence fragments with evidence-grounded woven claims. Through iterative evidence weaving, DeepWeaver identifies overlooked evidence, refines incomplete claims, and improves detail preservation and citation grounding. We also introduce LoQA, a high-density evidence benchmark for evaluating evidence-grounded answer generation. Experiments on knowledge-base QA and web-based deep research show that DeepWeaver consistently improves answer quality across multiple LLM backbones. These findings suggest that effective open-ended QA requires explicit evidence synthesis mechanisms between retrieval and generation, especially under noisy and fragmented evidence.

7 Limitations

In this work, we study evidence weaving as a way to advance retrieval-augmented open-ended question answering. While DeepWeaver improves evidence synthesis quality, several limitations remain. First, LoQA focuses on Chinese water-environment questions. Although this domain provides dense technical evidence and complex practical problems, it does not cover all languages, domains, or document types. We therefore evaluate DeepWeaver on DeepResearch Bench as a complementary setting, and future work will further enrich LoQA by expanding its metrics and assessment protocols. Second, DeepWeaver mainly improves the evidence synthesis stage after evidence collection. In the web-based setting, we reuse web pages and section titles collected by WebWeaver's planner to isolate the effect of evidence weaving. Designing web-search agents better aligned with DeepWeaver remains an open problem. Third, DeepWeaver currently focuses on text-only evidence, while more complete answers may involve figures, charts, images, or structured data. Extending evidence weaving to multimodal and structured evidence remains future work. Finally, although DeepWeaver improves evidence synthesis, model-generated content still requires careful verification.

References

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2024. L-eval: Instituting standardized evaluation for long context language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14388–14411.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, and 1 others. 2024a. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. *arXiv preprint arXiv:2411.14199*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. 2024b. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations*, volume 2024, pages 9112–9141.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, and 1 others. 2024. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 3119–3137.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

DeepSeek-AI. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. 2025. Deepresearch bench: A comprehensive benchmark for deep research agents. *arXiv preprint arXiv:2506.11763*.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488.

Jiani Guo, Zuchao Li, Jie Wu, Qianren Wang, Yun Li, Lefei Zhang, Hai Zhao, and Yujiu Yang. 2025. Tom: Leveraging tree-oriented mapreduce for long-context reasoning in large language models. *Preprint*, arXiv:2511.00489.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tian Ao, and Chao Huang. 2024. Lightrag: Simple and fast retrieval-augmented generation. *arXiv preprint arXiv:2410.05779*, 2(3).

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

Haoyu Han, Yu Wang, Harry Shomer, Kai Guo, Jiayuan Ding, Yongjia Lei, Mahantesh Halappanavar, Ryan A Rossi, Subhabrata Mukherjee, Xianfeng Tang, and 1 others. 2024. Retrieval-augmented generation with graphs (graphrag). *arXiv preprint arXiv:2501.00309*.

Sebastian Hofstätter, Jiecao Chen, Karthik Raman, and Hamed Zamani. 2022. Fid-light: Efficient and effective retrieval-augmented text generation. *Preprint*, arXiv:2209.14290.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*.

Yunhai Hu, Yilun Zhao, Chen Zhao, and Arman Cohan. 2025. Mcts-rag: Enhancing retrieval-augmented generation with monte carlo tree search. *arXiv preprint arXiv:2503.20757*.

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuxuan Gu, Weihong Zhong, Xiachong Feng, Weijiang Yu, Weihua Peng, Duyu Tang, Dandan Tu, and 1 others.

2024. Learning fine-grained grounded citations for attributed large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14095–14113.

Yizheng Huang and Jimmy Xiangji Huang. 2024. A survey on retrieval-augmented text generation for large language models. *ACM Computing Surveys*.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. *Preprint*, arXiv:2007.01282.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2024. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1658–1677.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025a. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.

Jiajie Jin, Xiaoxi Li, Guanting Dong, Yuyao Zhang, Yutao Zhu, Yongkang Wu, Zhonghua Li, Ye Qi, and Zhicheng Dou. 2025b. Hierarchical document refinement for long-context retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3502–3520.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025a. Search-o1: Agentic search-enhanced large reasoning models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5420–5438.

Zijian Li, Xin Guan, Bo Zhang, Shen Huang, Houquan Zhou, Shaopeng Lai, Ming Yan, Yong Jiang, Pengjun Xie, Fei Huang, and 1 others. 2025b. Webweaver: Structuring web-scale evidence with dynamic outlines for open-ended deep research. *arXiv preprint arXiv:2509.13312*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

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

Zheng Liu, Chaofan Li, Shitao Xiao, Yingxia Shao, and Defu Lian. 2025. Llama2vec: Unsupervised adaptation of large language models for dense retrieval. *Preprint*, arXiv:2312.15503.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and 1 others. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A Rossi, Seunghyun Yoon, and Hinrich Schütze. 2025. Nolima: Long-context evaluation beyond literal matching. *arXiv preprint arXiv:2502.05167*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, and 1 others. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Xuefei Ning, Zinan Lin, Zixuan Zhou, Zifu Wang, Huazhong Yang, and Yu Wang. 2024. Skeleton-of-thought: Prompting llms for efficient parallel generation. In *International Conference on Learning Representations*, volume 2024, pages 917–967.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*, volume 4. Now Publishers Inc.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551.

Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. 2024. 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)*, pages 6252–6278.

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. 2024. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740*.

Jiashuo Sun, Xianrui Zhong, Sizhe Zhou, and Jiawei Han. 2025. Dynamicrag: Leveraging outputs of large language model as feedback for dynamic reranking in retrieval-augmented generation. *Preprint*, arXiv:2505.07233.

Zhengwei Tao, Jialong Wu, Wenbiao Yin, Junkai Zhang, Baixuan Li, Haiyang Shen, Kuan Li, Liwen Zhang, Xinyu Wang, Yong Jiang, and 1 others. 2025. Webshaper: Agentically data synthesizing via information-seeking formalization. *arXiv preprint arXiv:2507.15061*.

GLM Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, and 152 others. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *Preprint*, arXiv:2508.06471.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, and 1 others. 2024a. Mineru: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839*.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023a. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 2609–2634.

Liang Wang, Nan Yang, and Furu Wei. 2023b. Query2doc: Query expansion with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9414–9423.

Minzheng Wang, Longze Chen, Cheng Fu, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, Yunshui Li, Min Yang, Fei Huang, and Yongbin Li. 2024b. Leave no document behind: Benchmarking long-context llms with extended multi-doc qa.

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Min Zhang, Qingsong Wen, and 1 others. 2024c. Autosurvey: Large language models can automatically write surveys. *Advances in neural information processing systems*, 37:115119–115145.

Zheng Wang, Shu Teo, Jieer Ouyang, Yongjun Xu, and Wei Shi. 2024d. M-rag: Reinforcing large language model performance through retrieval-augmented generation with multiple partitions. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1966–1978.

Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhenglin Wang, Zhengwei Tao, Ding-Chu Zhang, Zekun Xi, Robert Tang, and 1 others. 2026. Webdancer: Towards autonomous information seeking agency. *Advances in Neural Information Processing Systems*, 38:120957–120985.

Yuhao Wu, Yushi Bai, Zhiqiang Hu, Juanzi Li, and Roy Ka-Wei Lee. 2025. Superwriter: Reflection-driven long-form generation with large language models. *Preprint*, arXiv:2506.04180.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2023. Recomp: Improving retrieval-augmented lms with compression and selective augmentation. *arXiv preprint arXiv:2310.04408*.

Renjun Xu and Jingwen Peng. 2025. A comprehensive survey of deep research: Systems, methodologies, and applications. *Preprint*, arXiv:2506.12594.

Shicheng Xu, Liang Pang, Mo Yu, Fandong Meng, Huawei Shen, Xueqi Cheng, and Jie Zhou. 2024. Unsupervised information refinement training of large language models for retrieval-augmented generation. *Preprint*, arXiv:2402.18150.

Xiangchao Yan, Shiyang Feng, Jiakang Yuan, Renqiu Xia, Bin Wang, Bo Zhang, and Lei Bai. 2025. Surveyforge: On the outline heuristics, memory-driven generation, and multi-dimensional evaluation for automated survey writing. *Preprint*, arXiv:2503.04629.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Xi Ye, Ruoxi Sun, Sercan Arik, and Tomas Pfister. 2024. Effective large language model adaptation for improved grounding and citation generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6237–6251.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Peixin Cao, Kaixin Ma, Jian Li, Hongwei Wang, and Dong Yu. 2024. Chain-of-note: Enhancing robustness in retrieval-augmented language models. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 14672–14685.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, and 1 others. 2024a. ∞ bench: Extending long context evaluation beyond 100k tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15262–15277.

Yusen Zhang, Ruoxi Sun, Yanfei Chen, Tomas Pfister, Rui Zhang, and Sercan Ö Arık. 2024b. Chain of agents: Large language models collaborating on long-context tasks. *Advances in Neural Information Processing Systems*, 37:132208–132237.

Zhiyuan Zhao, Hengrui Kang, Bin Wang, and Conghui He. 2024. Doclayout-yolo: Enhancing document layout analysis through diverse synthetic data and global-to-local adaptive perception. *Preprint*, arXiv:2410.12628.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. 2025. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 414–431.

Zihan Zhou, Chong Li, Xinyi Chen, Shuo Wang, Yu Chao, Zhili Li, Haoyu Wang, Rongqiao An, Qi Shi, Zhixing Tan, and 1 others. 2024. Llm × mapreduce: Simplified long-sequence processing using large language models. *arXiv preprint arXiv:2410.09342*.

A Appendix

A.1 The LoQA Benchmark

A.1.1 Data Statistics

As shown in Table 5, LoQA comprises 100 questions, each accompanied by an average of 206K evidence tokens and 91.79 relevant chunks. For evaluation, LoQA includes 2,045 atomic scoring points and 4,779 cloze blanks, corresponding to 20.45 scoring points and 47.79 blanks per question on average. The reference answers contain 9,858.64 tokens per question on average. These statistics indicate that LoQA is built on extensive noisy evidence and is designed to stress-test evidence synthesis quality, evaluating whether QA systems can comprehensively use evidence, preserve fine-grained domain-specific details, and generate well-cited answers.

Statistic	Value
# Questions	100
Avg. evidence tokens / question	206,550.41
Avg. relevant chunks / question	91.79
Avg. scoring points / question	20.45
Avg. cloze blanks / question	47.79
Avg. reference-answer tokens / question	9,858.64

Table 5: **Statistics of LoQA.** Each question is paired with a large, noisy evidence pool of approximately 200K tokens, enabling fine-grained evaluation of evidence synthesis quality.

A.1.2 Construction Details

Figure 4 illustrates the construction pipeline of LoQA. We construct LoQA from a large-scale expert corpus in the water-environment domain. Starting from approximately 30,000 professional PDF books in our hydraulic engineering book list, we first select 4,484 books with relevant tags and then use an LLM to further identify books related to water ecology, water environment, and water circulation based on their titles, keywords, and descriptions. This filtering process yields 800 relevant books, from which we randomly select 500 as the knowledge base and use the remaining books as seed documents for practical question generation. We extract plain text from the PDFs using PDF-Extract-Kit (Wang et al., 2024a; Zhao et al., 2024), resulting in approximately 360 MB of clean text.

Question generation and filtering. To construct professional, open-ended questions, we randomly sample fragments from the seed documents and prompt GLM-4-Plus, DeepSeek-V3.2, and Qwen3-235B-A22B-Thinking-2507 to generate questions with diverse styles and perspectives. Each model produces three questions for each seed fragment, yielding 2,700 candidate questions in total. We require the questions to be decontextualized, concise, generally answerable, and dependent on multi-source domain knowledge. We then use DeepSeek-V3.2 to score each question along four dimensions: retrieval comprehensiveness, concreteness, domain depth, and expression completeness. Questions that are vague, overly narrow, insufficiently extensible, or ambiguously expressed are removed. We further select the 100 questions with the highest embedding-based cosine similarity to their top-100 retrieved chunks. These questions are retained as the final test set. (Figure 4 ②)

Evidence collection. We segment the 500-book knowledge base into 1,024-token fragments and encode each fragment using GLM-Embedding-3. For each question, we retrieve approximately 90–100 fragments with the highest cosine similarity to the question embedding:

$$E_{\text{rag}} = \text{TopK}_{e_i \in E} \left(\langle \text{emb}(e_i), \text{emb}(q) \rangle \right)$$

To simulate realistic noisy retrieval, we additionally sample about 100 random fragments from the corpus. As a result, each question is paired with 200 evidence fragments, yielding an evidence pool of approximately 200K tokens. (Figure 4 ③) We

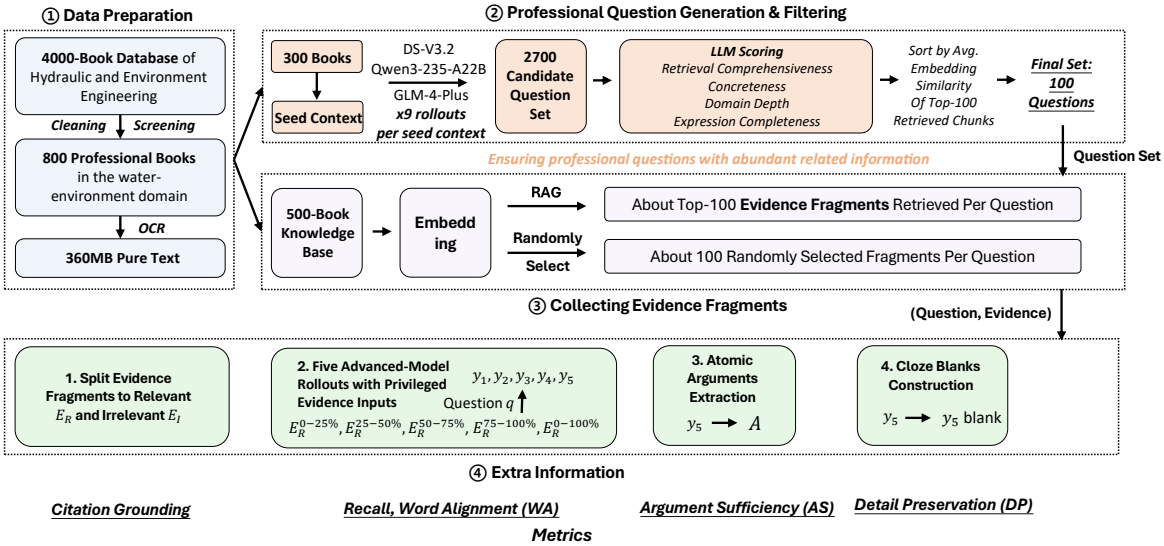


Figure 4: **Overview of LoQA construction and evaluation.** LoQA is built from a water-environment expert knowledge base through data preparation, professional question generation and filtering, and evidence-fragment collection. For evaluation, the evidence pool is split into relevant and irrelevant fragments, multiple reference answers are generated from different evidence partitions, and atomic arguments and cloze blanks are constructed to measure citation grounding, content sufficiency, and detail preservation.

further use DeepSeek-V3.2 to assess the relevance of each fragment to the question, providing annotations for citation-grounding evaluation.

Evaluation preparation. Because open-ended professional questions typically do not admit a single deterministic gold answer, LoQA constructs multiple possible reference answers synthesized from privileged evidence inputs. For each question, we shuffle the relevant evidence fragments E_R and split them into four disjoint partitions, together with the full relevant-evidence set:

$$E_R^{0\text{-}25\%},\ E_R^{25\text{-}50\%},\ E_R^{50\text{-}75\%},\ E_R^{75\text{-}100\%},\ E_R^{0\text{-}100\%}$$

DeepSeek-V3.2 generates one reference answer from each evidence subset, yielding five reference answers in total. The partition-based answers promote local, fine-grained coverage while being generated under a lower context burden, whereas the full-evidence answer provides a global synthesis.

From these reference answers, we construct more evaluation targets. First, we extract atomic arguments from the five reference answers via LLM. These arguments capture key evidence-supported answer aspects and are used to assess whether a generated answer covers the major content implied by the question and evidence. Second, we construct cloze-style blanks from the full-evidence reference answer. These blanks primarily target domain-specific terms, concepts, and short phrases, and are used to assess whether the generated answer preserves fine-grained professional details. Together with evidence relevance annotations, these targets enable LoQA to evaluate the answer along three dimensions: citation grounding, content sufficiency, and detail preservation. (Figure 4 ④)

Although LoQA uses DeepSeek-V3.2 for annotation and verification, it does not rely on holistic subjective scoring of long-form answers. Instead, we decompose evaluation into short-context extraction and verification tasks, including evidence relevance judgment, atomic argument matching, and cloze-blank recovery judgment. These localized decisions are substantially more constrained than open-ended preference scoring and fall within the reliable capability range of recent advanced LLMs, making DeepSeek-V3.2 suitable for constructing LoQA evaluation signals.

A.1.3 Comparison with Prior Benchmarks

Table 6 compares LoQA with representative benchmarks for long-form QA, citation-grounded generation, long-context understanding, and deep-research agents. Existing benchmarks primarily emphasize explanatory QA, verifiable short-form answering, long-context comprehension, or end-to-end web research. In contrast, LoQA is designed to isolate the evidence-utilization stage: each professional open-ended question is paired with a fixed, extensive, and noisy evidence pool, and systems are evaluated by their ability to generate comprehensive, well-cited, and detail-preserving answers that directly address the question. LoQA is com-

Benchmark	Task Focus	Evidence Setting	Main Difference from LoQA
ALCE (Gao et al., 2023)	Citation-grounded text generation	Retrieval corpora built from ASQA, QAMPARI, and ELI5	Focuses on fluency, correctness, and citation quality. Its tasks primarily involve producing short answers or entity lists from retrieved passages, whereas LoQA targets comprehensive answers grounded in knowledge-intensive evidence pools of approximately 200K tokens.
FreshWiki (Shao et al., 2024)	Wikipedia-like article generation from scratch	Web sources collected during the pre-writing stage	Evaluates whether systems can research a topic, construct an outline, and generate Wikipedia-like articles. In contrast, LoQA uses a noisy evidence pool and emphasizes comprehensive evidence synthesis in question answering, rather than outline construction or survey writing.
LongBench (Bai et al., 2024)	Long-context understanding	Single-document QA, multi-document QA, summarization, few-shot learning, synthetic and code tasks	Assesses general long-context capability, but most tasks require short-form outputs or automatically verifiable answers rather than open-ended long-form answers.
L-Eval (An et al., 2024)	Long-context model evaluation	Long documents across diverse domains and task types	Covers long-context QA and generation, but many tasks focus on short factual verification, extractive QA, or summarization rather than comprehensive open-ended answering.
InfiniteBench (Zhang et al., 2024a)	Extremely long-context processing	100K+ token contexts across retrieval, code, math, novels, and dialogue	Stresses context length and reasoning capacity, but its targets are mainly verifiable short answers instead of long-form citation-grounded answers.
Loong (Wang et al., 2024b)	Extended multi-document QA	Multiple relevant documents where missing any document may hurt the answer	Emphasizes reasoning over all relevant documents, while LoQA further introduces noisy irrelevant evidence and evaluates citation grounding, content sufficiency, and detail preservation.
DeepResearch Bench (Du et al., 2025)	End-to-end deep-research agents	Web-scale research tasks across diverse fields	Evaluates complete deep-research agents, including search and answer generation. LoQA provides a controlled setting to isolate evidence synthesis under a given noisy evidence pool.
LoQA (Ours)	Professional open-ended answer generation	About 200K evidence tokens per question with both relevant and noisy fragments	Evaluates evidence synthesis quality, assessing whether QA systems can transform extensive, noisy, and fragmented evidence into comprehensive, well-cited, and detail-preserving answers.

Table 6: **Comparison with prior benchmarks.** Unlike benchmarks that emphasize short-form answers, survey-style article generation, general long-context understanding, or end-to-end web research, LoQA isolates evidence synthesis quality for open-ended QA by pairing each question with a fixed, extensive, and noisy evidence pool and evaluating citation grounding, content sufficiency, and detail preservation.

plementary to prior benchmarks and suitable for studying evidence synthesis quality in open-ended question answering.

A.2 Details of Extending DeepWeaver to DeepResearch Bench

DeepResearch Bench primarily evaluates end-to-end Open-Ended Deep-Research (OEDR) systems, in which agents search the web, collect evidence, and generate open-ended long-form answers. Because DeepWeaver does not include a search module, we study a controlled setting: given the same search system and comparable inference cost, whether DeepWeaver can improve the evidence-to-answer generation stage.

To this end, we use the planner of WebWeaver (Li et al., 2025b) as the searcher. WebWeaver is an advanced dual-agent deep-research framework, where the planner iteratively searches web pages,

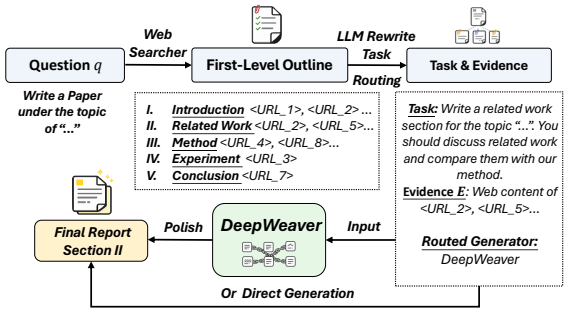


Figure 5: Overview of DeepWeaver's Extension to Web-Based Deep-Research Scenarios

refines a hierarchical multi-level outline, assigns web pages to outline nodes, and extracts evidence from each page to build a memory bank. The writer then fills the outline based on the collected evidence. In our setting, however, DeepWeaver does not rely on WebWeaver's detailed outline or extracted evidence memory. We only use the first-level outline titles O_1 and the corresponding raw web pages W under each title as the input to DeepWeaver. This keeps the search layer fixed while allowing us to isolate the effect of evidence weaving during long-form answer generation.

For each first-level outline node $o \in O_1$, we first rewrite o into a section-level DeepWeaver task q. The model also judges (routes) whether the section is critical to answering the original question or contains extensive webpage information. If so, we run DeepWeaver to generate the section; otherwise, we use direct generation. This rewriting and routing step only consumes a few hundred tokens and has negligible cost. For evidence construction, we truncate the raw webpage content associated with o to at most 35K tokens and split it into 1,024-token fragments, which are then regarded as the evidence pool for DeepWeaver.

Statistic	Value
Avg. first-level outline node $\lvert O_1 \rvert$ / question	4.41
Avg. routed outline node / question	2.35
Avg. routed-item ratio / question	59.39%
Avg. evidence fragments / item	88.92
Avg. evidence fragments / routed item	98.00
Avg. raw tokens of evidence / routed item	92,752.79

Table 7: **Routing statistics of DeepWeaver on DeepResearch Bench.** Each question contains 4.41 first-level outline nodes on average, among which 2.35 nodes are routed to DeepWeaver, accounting for 59.39% of the items. Routed items are evidence-heavy, containing 98 chunks and 92.75K raw tokens on average. In LoQA, the evidence length is much larger, with about 200K tokens per question, indicating that **DeepWeaver is sufficient to handle evidence-context lengths in realistic applications, such as multi-webpage reading.**

After DeepWeaver generates the section-level answers, we apply a lightweight polishing stage to improve fluency and readability. We then add an introduction and a conclusion to produce a complete end-to-end answer. These polishing, introduction, and conclusion stages are auxiliary components introduced solely to match the answer format required by DeepResearch Bench. The core comparison focuses on whether DeepWeaver can more effectively synthesize web evidence into comprehensive, insightful, and well-cited answer content.

A.3 Cost Analysis

We analyze the cost of DeepWeaver on DeepResearch Bench by reusing WebWeaver's Searcher (Planner) and replacing the downstream writer with DeepWeaver. To reduce variance introduced by API providers, network latency, and webpage accessibility, we exclude the cost and runtime of web search and page access, and measure only LLM-side computation. Local runtime is measured with Qwen3-30B-A3B-Instruct-2507 services on local 2×A100 GPUs.

As shown in Table 8, DeepWeaver incurs a monetary cost comparable to WebWeaver ($0.313 vs. $0.303), while reducing local inference time from 18.5 to 14.2 minutes. Although DeepWeaver consumes more input tokens (3.16M vs. 2.35M), it substantially reduces output tokens from 307K to 94K by avoiding explicit webpage-level evidence extraction and memory-bank construction. This extraction-free design reallocates the budget from generating intermediate summaries to reading more original evidence.

Under comparable cost, DeepWeaver can pass up to 35K tokens of original content per webpage to the writer, compared with 24K tokens in WebWeaver, reducing the cost per observed token from 1.26×10^{-5} to 8.94×10^{-6}. Figure 6 further shows that DeepWeaver introduces only lightweight internal overhead. The main additional cost comes from the *Subordinate* stage, which inspects residual evidence and enriches the TBC. However, its output-token usage is only slightly higher than that of the *Draft* and *Summary* stages, and its total token usage remains comparable to the *Draft* stage. Other maintenance stages, including *Commit*, *Discard*, *Polish*, *Intro*, and *Outro*, account for only a small fraction of the total cost. Overall, DeepWeaver improves evidence utilization not by substantially increasing cost, but by reallocating the same budget: it avoids expensive intermediate output generation and explicit evidence extraction, and instead uses lightweight TBC metadata to route more original evidence into focused answer generation.

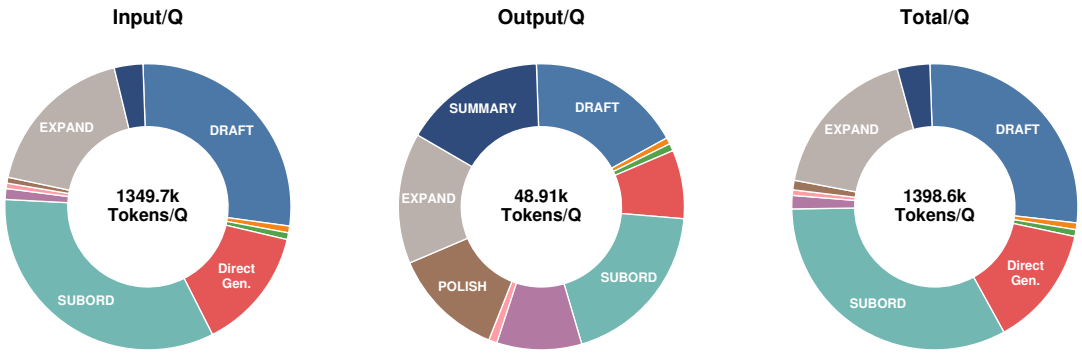
A.4 Prompt Templates

This section presents the English prompt templates used for each stage of DeepWeaver.

Agent	Evidence-Extraction-Free*	# Input Tokens	# Output Tokens	Local Time (min)	Cost	Answer Len.
WebWeaver	✗	**2,348K**	307K	18.5	**$0.303**	11,889.3
DeepWeaver	✓	3,161K	**94K**	**14.2**	$0.313	12,303.8

Agent	# Tokens Seen by Writer / Web Page	# Input Tokens /Token Seen	# Output Tokens /Token Seen	Cost /Token Seen
WebWeaver	24,000	97.83	12.79	$\$1.26 \times 10^{-5}$
DeepWeaver	**35,000** †	**90.31**	**2.69**	$\$8.94 \times 10^{-6}$

Table 8: **Average cost comparison between WebWeaver and DeepWeaver (including the searcher).** Under comparable cost and with the same searcher, replacing the downstream writer with DeepWeaver improves DeepResearch Bench performance. To reduce evaluation variance caused by API providers, network latency, and web access, we exclude the cost and time of web search and page access. Local Time only measures inference time of Qwen3-30B-A3B-Instruct-2507 on local 2×A100 GPUs. * DeepWeaver avoids extracting evidence content and maintaining a webpage-level memory bank, substantially reducing output-token overhead. † Under comparable cost, DeepWeaver can pass more original webpage content to the answer writer, up to 35K tokens, whereas WebWeaver relies on evidence extracted from 24K-token webpage inputs.



Stage	Input/Q	Output/Q	Total/Q
DRAFT	374.0k (27.7%)	8.59k (17.6%)	382.6k (27.4%)
INTRO	10.30k (0.8%)	358.4 (0.7%)	10.66k (0.8%)
OUTRO	10.31k (0.8%)	407.6 (0.8%)	10.72k (0.8%)
Direct Gen.	187.2k (13.9%)	3.79k (7.8%)	190.9k (13.7%)
SUBORD	449.2k (33.3%)	9.35k (19.1%)	458.5k (32.8%)
COMMIT	16.50k (1.2%)	4.70k (9.6%)	21.20k (1.5%)
DISCARD	8.21k (0.6%)	459.5 (0.9%)	8.67k (0.6%)
POLISH	8.92k (0.7%)	6.17k (12.6%)	15.09k (1.1%)
EXPAND	241.4k (17.9%)	7.23k (14.8%)	248.6k (17.8%)
SUMMARY	43.71k (3.2%)	7.86k (16.1%)	51.57k (3.7%)

Figure 6: **Token usage of DeepWeaver (excluding the searcher) across stages.** We report token consumption on DeepResearch Bench. POLISH, INTRO, and OUTRO denote polishing, introduction generation, and conclusion generation, respectively (§A.2). Although the *Subordinate* stage is the main additional cost, its output tokens are only slightly higher than the *Summary* output, and its total tokens are comparable to the *Draft stage*, showing that TBC-based evidence weaving is lightweight beyond evidence input and answer generation.

Thought Block Chain Structure Example in Original Source Code

```
{
    // Thought Block Chains b_i
    "logic_blocks": [
        {
          // Claims c_i
          "argument": "Netflix's adaptation of 'One Hundred Years of Solitude' succeeded by
          prioritizing cultural and linguistic authenticity through filming in Spanish and in
          Colombia.",

          // Keywords k_i
          "keywords": [
                "cultural authenticity",
                "original language",
                "Colombian production"
            ],

          // Salient Information s_i
          "key_info": "The series was filmed in Spanish and exclusively in Colombia, the author's
          native country, to preserve the novel's rhythm and regional accent, and to draw from
          the nation's real geography and history. This decision was made to honor the author's
          conditions and to avoid the 'white-washed Hollywood' approach that had previously
          failed other adaptations.",

          // Evidence Fragments E_i
          "reference_chunks": [5,22,44,82,119],

          // Covered Evidence Fragments, Top-k Embedding Similarity Chunks to c_i+s_i
          // Available only for main TBCs
          "retrieval_reference_chunks": [19,27,15,16,120]
        }, ...
    ]
}
```

Draft Stage Prompt Template

```
//Step 1: Draft Answer Generation Prompt Template
<|im_start|>system
You are a professional research assistant and need to answer the question based on the following
document passages.
<|im_end|>
<|im_start|>user

Question: {question}

Chunk #i: {evidence fragments}

Please generate a comprehensive answer report based on the following document passages. If you
cite any passage, place the cited passage number in [] at the end of the corresponding sentence,
such as [1,3] or [2]:
<|im_end|>
```

```
//Step 2: Extracting TBC Prompt Template
<|im_start|>system
You are a professional logic-organization expert. Your task is to break the "generated answer"
into a set of logic blocks.

Please output JSON according to the following requirements:
1. Each logic block should revolve around a clear subtopic and reflect the internal logical
structure of the answer.
2. Each logic block should contain:
   - argument: a one-sentence summary of the logic block's argument
   - keywords: 3-5 keywords
   - key_info: a summary of the key information corresponding to the argument
   - reference_chunks: the reference passages involved in this argument
3. The reference_chunks indices may be parsed from the displayed citations in the "generated
answer", and may also be expanded with other highly relevant reference passages. Each argument
should have 3-8 reference_chunks
4. Ensure that the JSON format is correct and can be parsed.

Output format:
{
  "logic_blocks": [
    {
      "argument": "...",
      "keywords": ["...", "..."],
      "key_info": "...",
      "reference_chunks": [1,2,3]
    },

    {
      "argument": "...",
      "keywords": ["...", "..."],
      "key_info": "...",
      "reference_chunks": [6,9]
    },
    ...
  ]
}<|im_end|>

<|im_start|>user
Please generate a structured JSON response based on the following references and question:
Question: {question}

Chunk #i: {evidence fragments}

Generated answer:
{answer}
<|im_end|>
```

Subordinate Stage Prompt Template

```
//Step 1: Supplementary Answer Generation Prompt Template
<|im_start|>system
You are a professional research assistant and need to answer the question based on the
following document passages.
<|im_end|>
<|im_start|>user

Chunk #i: {uncovered evidence fragments}

Question: {question}

Below are the logic blocks that have already been identified as discarded for this question:
{discarded logic blocks}

Please generate a comprehensive answer report based on the following document passages. If you
cite any passage, place the cited passage number in [] at the end of the corresponding sentence,
such as [1,3] or
[2]:
<|im_end|>
```

```
//Step 2: Extracting Supplementary TBC Prompt Template
<|im_start|>system
You are a professional logic-organization expert. Your task is to break the "generated answer"
into a set of supplementary logic blocks.

Please output JSON according to the following requirements:
1. Each logic block should revolve around a clear subtopic and highlight its supplementary
value to the existing logic.
2. Each logic block should contain:
   - argument: a one-sentence summary of the logic block's argument
   - keywords: 3-5 keywords
   - key_info: a summary of the key information corresponding to the argument
   - reference_chunks: the reference passages involved in this argument
3. The reference_chunks indices may be parsed from the displayed citations in the "generated
answer", and may also be expanded with other highly relevant reference passages. Each argument
should have 3-8
reference_chunks
4. Ensure that the JSON format is correct and can be parsed.

Output format:
{
  "logic_blocks": [
    {
      "argument": "...",
      "keywords": ["...", "..."],
      "key_info": "...",
      "reference_chunks": [1,2,3]
    },

    {
      "argument": "...",
      "keywords": ["...", "..."],
      "key_info": "...",
      "reference_chunks": [6,9]
    },
    ...
  ]
}
<|im_end|>

<|im_start|>user
Please generate a structured JSON response based on the following references and question:

Chunk #i: {uncovered evidence fragments}
```

```
Question: {question}

Below are the logic blocks that have already been identified as discarded for this question:
{discarded thought blocks}

Generated answer:
{answer}

JSON response:
<|im_end|>
```

Commit Stage, Merge Operation Prompt Template

```
//Step 1: Merge Candidate Screening Prompt Template
<|im_start|>system
You are a professional logic-comparison expert. Your task is to find only the pairs of logic
blocks between the main and subord groups that must be merged.

Requirements:
1. Consider only cross-group pairings, and do not consider main-main or subord-subord.
2. Suggest a merge only when the viewpoints of two logic blocks clearly overlap; otherwise do
not output it.
3. The reference_blocks in each suggestion must be a pair, and it must consist of exactly one
positive main index and one negative subord index.
4. What you output are candidate suggestions, not the final result; when in doubt, leave it out.
5. If there is no candidate pair, please return an empty array [].
6. Output at most {k} of the most important overlaps.

Output format:
[
  {
    "reference_blocks": [2, -1],
    "rationale": "Both discuss ......, and the difference between them is ..., described in a
    neutral tone",
    "overlap_ratio": 0.xx
  }
]
<|im_end|>

<|im_start|>user
Please screen the must-merge candidate pairs according to the following main and subord logic
block information:

Please organize and merge the main and subord logic blocks based on the following information:

**Question:**
{question}

**Main logic blocks (numbered 1, 2, 3, ...):**
**Main Logic Block i:**
Argument: {main argument}
Keywords: {main keywords}
Key information: {main key information}

**Subord logic blocks (numbered -1, -2, -3, ...):**
**Subord Logic Block -i:**
Argument: {subord argument}
Keywords: {subord keywords}
Key information: {subord key information}

Notes:
- Do not output a single index.
- Do not output one-to-many or many-to-one.
- If there is no candidate pair that meets the criteria, return [] directly.

JSON response:
<|im_end|>
```

```
//Step 2: Commit Thought Block Merging Prompt Template
<|im_start|>system
You are a professional expert in logical organization and information integration. Your task is
to organize and merge logic blocks based on the provided main and subord logic block
information, forming a
clearer and more complete logical structure.

Please analyze and merge the logic blocks according to the following requirements:
1. **Input format**: There are two lists of logic blocks: the main logic block list and the
subord logic block list.
2. **Identify repeated or similar arguments**: Analyze the main and subord logic blocks and
identify pairs of logic blocks with repeated or similar arguments. It is already known that
there are no repeated
viewpoints within main or within subord, so you only need to consider whether there are two
viewpoints across groups that are highly similar or repeated.
3. **Organize the logical order reasonably**: Organize the logic blocks in a reasonable order
to form merged logic blocks.
4. **Extract merged information**: For each merged logic block, extract the merged argument,
keywords, and the indices of the original logic blocks it cites. Main indices are 1,2,3,...,
and subord indices are
-1,-2,-3,....
5. **Citation format and requirements**: reference_blocks can only be a single number or a
**pair**, and a pair **must** consist of one main logic block (positive index) and one subord
logic block (negative
index).
6. **Overall requirements**: Ensure that every logic block is covered, **and ensure that no two
merged logic blocks contain the same reference_blocks.** Each index may appear only once in all
reference_blocks
dicts.

Output format requirements:
Please output in JSON format, as shown below:

[
    {
        "reference_blocks": 1,
        "argument": "Merged argument 1",
        "keywords": ["Keyword 1", "Keyword 2"]
    },
    {
        "reference_blocks": [-1, 2],
        "argument": "Merged argument 2",
        "keywords": ["Keyword 3", "Keyword 4"]
    },
    {
        "reference_blocks": 3,
        "argument": "Merged argument 3",
        "keywords": ["Keyword 5", "Keyword 6"]
    },
    {
        "reference_blocks": [-2, 4],
        "argument": "Merged argument 4",
        "keywords": ["Keyword 6", ...]
    },
    ...
]
<|im_end|>

<|im_start|>user
Please organize and merge the main and subord logic blocks based on the following information:

**Question:**
{question}

**Main logic blocks (numbered 1, 2, 3, ...):**
```

```
**Main Logic Block i:**
Argument: {main argument}
Keywords: {main keywords}
Key information: {main key information}

**Subord logic blocks (numbered -1, -2, -3, ...):**
**Subord Logic Block -i:**
Argument: {subord argument}
Keywords: {subord keywords}
Key information: {subord key information}

Below are the pre-generated merge candidate suggestions:
{candidate merge pairs and rationales}

Notes:
- Ensure that the JSON format is correct and can be parsed.
- Keep the logic clear and well-structured. Each merged argument should be concise and clear
rather than verbose. If it is verbose, it must be split into two arguments and described
separately rather than
merged into one logic block.
- Ensure that every logic block is covered, **and ensure that no two logic blocks contain the
same reference_blocks.** Each index may appear only once in all reference_blocks dicts.

JSON response:
<|im_end|>
```

Commit Stage, Discard Operation Prompt Template

```
//Thought Block Discard Prompt Template
<|im_start|>system
You are a professional logic-analysis expert. Your task is to analyze logic blocks and identify
the logic blocks that should be discarded.

Please analyze the logic blocks according to the following criteria:

1. Off-topic: logic blocks unrelated to the theme of the question.
2. Redundant information: logic blocks whose content is obviously consistent with, or seriously
repetitive of, other existing logic blocks.
3. Mixed incorrect information: logic blocks containing obvious errors or information
inconsistent with facts.
4. Low-impact information: peripheral information that does not directly help solve the
question.

Output format requirements:
Please output in JSON format, containing the two fields block_id and rationale. The rationale
field should provide a brief reason for discarding the logic block.
If no logic block needs to be discarded, please return an empty array [].

Notes:
- Return only the logic blocks that need to be discarded.
- Please judge the degree of discarding carefully. For example, if the question includes
expressions such as "as comprehensive as possible", then fewer logic blocks should be discarded,
or none at all.
- Ensure that the JSON format is correct and can be parsed.
<|im_end|>

<|im_start|>user
Please generate a structured JSON response according to the following information:

Please analyze the following logic blocks and identify which ones should be discarded:

**Question:**
{question}

**Logic Blocks:**
**Logic Block i:**
Argument: {argument}
Keywords: {keywords}
Key information: {key information}
```

```
**Discard Criteria:**
1. Off-topic: logic blocks that are unrelated to the question.
2. Redundant information: logic blocks that repeat or closely overlap with other logic blocks.
3. Mixed or incorrect information: logic blocks that contain clearly incorrect information or
factual errors.
4. Weak-impact information: peripheral information that does not directly help solve the
question.

**Task:**
Identify all logic blocks that should be discarded, and provide a one-sentence rationale for
each discarded block.

**Output Format:**
Return the result in JSON format, following this example:

[
    {
        "block_id": 1,
        "rationale": "This logic block is unrelated to the question and therefore off-topic."
    },
    {
        "block_id": 3,
        "rationale": "This logic block overlaps with logic block 2 and is therefore redundant."
    }
]

If no logic blocks need to be discarded, return an empty array `[]`.

JSON response:
<|im_end|>
```

Expand Stage Prompt Template

```
//Logic Block Expansion Prompt Template
<|im_start|>system
You are a professional report-writing expert. Your task is to discuss the argument of a
specified logic block among a series of logic blocks, based on the provided literature passages
and while staying closely focused on the question itself.

Requirements:
1. The discussion must stay closely aligned with the argument and the question, while also
maintaining coherence with the already generated preceding answer.
2. Refer to and cite information from the provided chunk pool comprehensively and extensively.
3. Use concise and clear language.
4. Maintain objectivity and accuracy.
5. Staying close to the question is the first priority. Remove material that is only loosely
related to the question.
6. When a sentence uses information from the chunk pool, cite the local chunk ids at the end of
the sentence in the format [1] or [1,2].
7. Use only the chunk ids that appear in the prompt. Do not invent citation ids.

Please generate a complete discussion and output the content directly with inline citations.
<|im_end|>

<|im_start|>user
Please write a detailed discussion for the current logic block while keeping coherence with the
overall answer.

Question:
{question}

Task:
Write the discussion for logic block. The current block must stay tightly focused on the
question and on its own argument.

Chunk pool for the current block:
Chunk #i:
{evidence fragments}
```

```
Logic-block overview:
Current logic block {i}:
Argument: {current argument}
Keywords: {current keywords}

Logic block {j}:
Argument: {other argument}
Keywords: {other keywords}

Output only the discussion for the current logic block, with inline citations.
<|im_end|>
```

Summary Stage Prompt Template (for deep-research style answers)

```
//Subsection Summary Generation Prompt Template
<|im_start|>system
You are a professional answer-writing expert.
You are writing {one numbered subsection / one subsection} inside a larger report section.
Your task is to produce the complete subsection, including the subsection title line and the
subsection body, while preserving the factual support and inline citation markers already
present in the source
material.

Requirements:
1. Preserve valid inline citation markers such as [5] and [5,9] for claims that remain in the
text.
2. Do not invent new citation ids.
3. Improve clarity, professionalism, and readability while keeping the content tightly focused
on the question.
4. If numbering is used, you must generate the subsection title yourself. If numbering is not
used, do not generate an extra inner title.
5. Output format is mandatory:
{subsection label}. Subsection Title  Subsection body starts on the next line.
6. If earlier content already covers part of the same information, reduce repetition and keep
only what is still necessary here.
7. Every sentence must contribute directly to answering the question. Avoid open-ended
expansion, tangents, and over-elaboration.
8. {first-line format requirement}
9. Output only the complete subsection with inline citations preserved.
<|im_end|>

<|im_start|>user
Refine the current logic block into {one numbered subsection / one subsection} inside its
parent section.

Question:
{question}

Current subsection label:
{subsection label}.

Current logic block information:
Argument:
{argument}

Keywords:
{keywords}

Key information:
{key information}

Detailed content with inline citations:
{expanded discussion with citations}

Previously generated content:
{previous section headings and summaries}

Requirements:
```

```
1. Preserve all valid citation markers already present in the detailed content, such as [5] or
[5,9].
2. Do not invent new citation ids, and do not delete citations that support retained claims.
3. Improve clarity, density, and readability while keeping the content tightly focused on the
question.
4. If numbering is used, you must generate the subsection title yourself based on the current
logic block. If numbering is not used, do not generate an extra inner title.
5. Output format is mandatory and must be exactly:
{subsection label}. <your subsection title>  <subsection content starts on the next line>
6. If the previously generated content already covers some information from the current logic
block, reduce repeated discussion and keep only the incremental value needed here.
7. Every part of the subsection must directly help answer the question. Stay problem-oriented
and do not expand indefinitely.
8. {first-line format requirement}
9. Do not output any extra preface, explanation, or alternative formats.
10. Make the subsection directly appendable after the previously generated content while
maintaining coherence.
11. Make the language clear, easy to understand, and highly readable, while avoiding excessive
elaboration.
12. The language you use should keep same with that of the question.

Please generate the complete subsection for the current logic block:
<|im_end|>
```

Auxiliary Generation Prompt Templates (for deep-research style answers)

```
//Direct Generation Prompt Template
<|im_start|>system
You are a professional report-writing expert. Your task is to discuss the argument of a
specified logic block among a series of logic blocks, based on the provided literature passages
and while staying
closely focused on the question itself.

Requirements:
1. The discussion must stay closely aligned with the argument and the question, while also
maintaining coherence with the already generated preceding answer.
2. Refer to and cite information from the provided chunk pool comprehensively and extensively.
3. Use concise and clear language.
4. Maintain objectivity and accuracy.
5. Staying close to the question is the first priority. Remove material that is only loosely
related to the question.
6. When a sentence uses information from the chunk pool, cite the local chunk ids at the end of
the sentence in the format [1] or [1,2].
7. Use only the chunk ids that appear in the prompt. Do not invent citation ids.

Please generate a complete discussion and output the content directly with inline citations.
<|im_end|>

<|im_start|>user
Please write a detailed discussion following instruction.

Instruction:
{instruction}

Available evidence chunk pool for citation:
Chunk #i:
{evidence fragments}

Requirements:
1. Stay tightly focused on both the question and the argument.
2. Use the chunk pool extensively when it is relevant.
3. When a sentence uses information from a chunk, cite it with the provided local chunk id, such
as [1] or [1,3].
4. Use only the chunk ids shown in the chunk pool above. Do not invent citation ids.
5. Place citations at the end of the corresponding sentence.
6. Output only the discussion text.
7. You must keep the language you use same with the question.

Additional requirement for this direct-generation task:
```

```
- Organize the answer using exactly two heading levels when structure is needed.
- Use top-level headings in the form `A. ...`, `B. ...`, `C. ...`, `D. ...`.
- Under each top-level heading, use second-level headings in the form `1. ...`, `2. ...`, `3.
...`, `4. ...` when there are multiple subpoints.
- Keep the structure clear and reader-friendly.
- Do not add any heading level deeper than the `1. / 2. / 3. / 4.` level.
- Preserve inline citations in the generated content.
- Keep the language of the discussion be the same with instruction (Chinese / English).
<|im_end|>
```

```
//Polish Prompt Template
<|im_start|>system
You are a careful editor for report sections.
Polish the section for readability and concision while preserving all information and all valid
inline citations.
Return only the polished section text.
<|im_end|>

<|im_start|>user
Instruction for this section:
{instruction}

Current section draft:
{section draft}

Editing requirements:
1. Preserve all information that is already present in the draft.
2. Preserve all valid inline citations such as [5] and [5,9]. Do not invent, remove, or renumber
citations.
3. Reduce useless discussion and repeated phrasing.
4. Improve clarity, readability, and logical flow while preserving the original meaning,
technical content, and conclusions.
5. Remove redundancy and excessive elaboration. Merge overlapping ideas and keep the response
focused on the target question.
6. Strengthen the structure of the answer so that the argument follows a clear progression from
background, to relationship/mechanism analysis, to modeling framework, and then to conclusions.
7. Ensure that technical and domain-specific terms are accurate, consistent, and professionally
used. Briefly clarify complex or ambiguous concepts when necessary.
8. When discussing models, methods, or frameworks, clearly explain their purpose, assumptions,
inputs, outputs, and how they connect to the main problem.
9. Keep the writing tightly aligned with the instruction for this section.
10. Keep the existing heading and numbering structure whenever it is already useful.
11. Output only the polished section text. You don't need to add the section title.
12. The language of your polished output should remain consistent with the draft.
<|im_end|>
```

```
//Answer Introduction Generation Prompt Template
<|im_start|>user
Question:
{question}

Report body:
{assembled section contents}

Write the opening section of a report that answers the question.
Requirements:
1. Use 1-2 paragraphs.
2. Introduce the question and briefly preview the report.
3. Do not use bullet points.
4. Do not include a heading.
5. Do not include citation markers such as [1].
6. Keep the language same with the report.
<|im_end|>
```

```
//Answer Conclusion Generation Prompt Template
<|im_start|>user
Question:
{question}

Report body:
{assembled section contents}

Write the ending section of the report.
Requirements:
1. Use 1-2 paragraphs.
2. Synthesize the key points and close the report.
3. Do not introduce new section headings or bullet points.
4. Do not include a heading.
5. Follow the language of the question.
6. Do not include citation markers such as [1].
7. Keep the language same with the report.
<|im_end|>
```

A.5 Case Study

We present two examples to illustrate how DeepWeaver composes new claims by weaving related thought blocks across refinement steps.

LoQA The question is how industrial water treatment should be redefined when shifting from "pollution control" to "resource recovery." The initial claims discuss water reuse, energy recovery, material recovery, ecological evaluation, and economic feasibility separately. DeepWeaver links these claims with subordinate claims about circular resource use and multidimensional evaluation. The final result reframes wastewater treatment as a resource recovery system: wastewater is no longer treated only as pollution to remove, but as a resource pool for recovering water, energy, and valuable materials. This case shows DeepWeaver's ability to turn scattered technical points into a higher-level conceptual claim.

DeepResearch Bench The topic is about the applications and functional capabilities of a sports intelligent tutoring system driven by multimodal data fusion. The draft claims cover real-time posture correction, performance evaluation, personalized training, injury prediction, and tactical analysis. DeepWeaver further connects these functions with signal-level, feature-level, and decision-level fusion. For example, posture correction is refined into a claim about fusing IMU and video data to build a 3D motion model for immediate feedback; performance evaluation is refined into feature-level fusion of biomechanical and physiological signals. This case shows how DeepWeaver converts separate application claims into a coherent set of system capabilities.

B Ethics Statement

The source books used in this work are publicly published books and copyrighted works owned by their respective authors and publishers. They were accessed through our institution's paid and licensed electronic-book collections. We do not claim ownership over the original book text, and our access to these books does not imply any right to re-license or redistribute the original works.

For each book, we release only a limited excerpt of approximately 1,000 OCR-derived tokens, together with our derived annotations and metadata, for non-commercial academic research, reproducibility, and evaluation purposes. The release is neither intended nor sufficient to substitute for the original books, enable reconstruction of the original works, or circumvent the access restrictions of the licensed e-book collections.

We rely on fair-use-style research use for the limited excerpts, considering the non-commercial scholarly purpose, the limited amount of text released, and the absence of a market substitute for the original books. Users are responsible for ensuring that their use complies with applicable copyright laws and the terms of the original sources and institutional licenses. Upon request from copyright holders, we will review and remove disputed excerpts.

The data used in this work are derived from publicly published books and does not involve private user data, personal communications, or sensitive personal information.

C Dataset License and Terms of Use

Annotations, metadata, and evaluation code in this work are released under Apache-2.0. Data are released under CC BY-NC 4.0, while the copyrighted book excerpts remain the property of their respective rights holders.

Case Study on LoQA

Inference Trajectory of DeepWeaver on LoQA

Q：工业水处理系统从"污染控制"向"资源回收"转型过程中，如何重新定义处理工艺的技术目标与评价体系，使其既符合生态循环原理又能满足经济可行性?

Draft


```
[
    {
      "argument": "工业水处理系统转型的核心是将技术目标从单一的达标排放，升级为以水资源高效循环回用为核心的多层次循环经济目标。",
      "keywords": ["技术目标", "水资源循环", "再生水", "梯级利用", "深度处理"],
      "key_info": "技术目标需从“无害化排放”转向生产满足不同用途（如工业冷却、锅炉补给、市政杂用、农业灌溉）的高品质再生水。实现路径包括企业内部或园区内的清污分流、分质处理、一水多用和串级使用，以及采用以“双膜法”为代表的深度处理技术。",
      "reference_chunks": [17, 21, 34, 51, 59, 70, 90],
      "retrieval_reference_chunks": [189, 79, 27, 76, 48]
    },
    {
      "argument": "技术目标应涵盖从废水中回收能源与有价值物质，实现污染物资源化转化。",
      "keywords": ["能源回收", "物质回收", "厌氧处理", "营养物质回收", "副产品"],
      "key_info": "将废水中的有机物和营养物质视为资源进行转化回收。具体包括：对高浓度有机废水采用厌氧技术（如UASB）回收沼气能源；回收氮、磷等营养物质；以及从特定工业废水中回收原料、贵重金属或有价值的副产品（如电镀废水中的金属、造纸黑液中的碱、味精废水中的菌体蛋白）。",
      "reference_chunks": [17, 23, 42, 54, 153, 188],
      "retrieval_reference_chunks": [77, 86, 84, 5, 196]
    },
    ...,
    {
      "argument": "新的评价体系需建立以生态循环绩效为核心的多维度综合评估框架。",
      "keywords": ["评价体系", "生态循环绩效", "水资源循环指标", "物质回收指标", "生态足迹"],
      "key_info": "评价指标应超越单一的处理效率，涵盖水资源循环（如重复利用率、再生水回用率）、物质循环与回收（如能源转化率、营养盐回收率）、生态足迹（能耗、温室气体排放、污泥资源化率）以及系统稳健性（抗冲击负荷能力、出水稳定性）。",
      "reference_chunks": [17, 21, 23, 51, 76, 153, 176],
      "retrieval_reference_chunks": [178, 48, 122, 44, 53]
    },
    {
      "argument": "经济可行性评价需基于全生命周期成本分析和成本-效益分析，并考虑市场化因素。",
      "keywords": ["经济可行性", "全生命周期成本", "成本效益分析", "市场化", "第三方治理"],
      "key_info": "评价需计算包括建设投资、运行费用、设备折旧、资源回收收益、节省的水费与排污费等在内的全生命周期成本。进行成本-效益分析，确保总效益大于总成本。同时评估技术的规模化适用性、再生水及回收产品的市场竞争力，并推动“第三方治理”等市场化运营模式。",
      "reference_chunks": [9, 34, 64, 106, 113],
      "retrieval_reference_chunks": [116, 197, 62, 27, 178]
    },
    ...
]
```


Subordinate (Refinement Turn #1)


```
[
    {
      "argument": "转型后的技术目标需构建多层次体系，核心是从末端无害化处理升级为水、能、物的协同回收与闭环循环。",
      "keywords": ["技术目标", "资源回收", "水资源回用", "能源回收", "物质回收"],
      "key_info": "技术目标应超越传统的污染物去除，扩展为水资源高效回用与闭环循环、能源与资源的同步回收、污染物的最小化与无害化、以及工艺过程的生态化与低碳化。这要求深度处理以满足回用标准，并通过厌氧消化、生物脱氮除磷、膜分离等技术回收沼气、营养物质和有价物质。",
      "reference_chunks": [3, 16, 31, 36, 45, 52, 69, 138]
    },
    {
      "argument": "评价体系需从单一环境指标转向涵盖资源效率、生态影响、经济可行性与系统智能性的多维度综合指标。",
      "keywords": ["评价体系", "资源回收效率", "生态影响", "全生命周期成本", "智能管控"],
      "key_info": "新型评价体系应包含资源回收效率指标（如水回收率、能源净产出率）、生态与环境影响指标（如碳足迹、生态毒性削减）、技术经济可行性指标（如全生命周期成本、投资回收期）以及系统协同与智能管控指标（如物质能量流集成度、自动化水平），以全面衡量转型成效。",
```

```
      "reference_chunks": [16, 45, 57, 95, 104, 124, 138]
    },
    {
      "argument": "实现转型需要革新工艺路线，发展集成化的“资源化工艺包”，并推动源头分类与分散式处理以提升回收针对性。",
      "keywords": ["工艺革新", "集成工艺", "分散式处理", "工艺包", "源头分离"],
      "key_info": "工艺路径应从单一处理链转向集预处理回收、厌氧产沼、主流脱氮除磷、深度处理与膜分离、浓水处理于一体的“资源化工艺包”。同时，借鉴分散式处理与再循环（DESAR）理念，在源头对黑水、灰水等不同性质污水进行分类，以实现更高效、有针对性的资源回收。",
      "reference_chunks": [16, 43, 52, 95, 104]
    },
    {
      "argument": "转型面临经济性挑战，需通过技术创新降低关键单元工艺成本，并依赖政策驱动与市场机制突破经济瓶颈。",
      "keywords": ["经济可行性", "技术成本", "政策驱动", "市场机制", "绿色金融"],
      "key_info": "高级氧化、膜分离、蒸发结晶等深度处理与资源回收技术成本较高，是转型的主要经济瓶颈。突破需要依靠技术创新降低单元工艺能耗与物耗，同时通过建立资源化产品标准、提供政策补贴、发展绿色金融等手段，提升资源回收项目的市场竞争力与投资回报率。",
      "reference_chunks": [16, 43, 95, 104, 130]
    },
    ...,
    {
       "argument": "跨学科融合与智能管控是提升资源回收系统稳定性、适应性与整体效益的关键支撑技术。",
      "keywords": ["跨学科融合", "智能管控", "系统优化", "稳定性", "适应性"],
      "key_info": "实现高效资源回收需要环境工程、过程工程、材料科学、信息技术等多学科交叉，以开发高性能材料（如膜、吸附剂）和高效反应器。同时，通过智能监测与控制系统对工艺参数进行优化调控，能有效应对进水波动，保障回收过程的稳定高效，并降低运行成本，是实现系统协同优化的关键。",
      "reference_chunks": [16, 30, 69, 104, 124, 141]
    }
]
```


Commit (Refinement Turn #1)


```
{
    "trash": [
      {
        "reference_blocks": 3,
        "argument": "转型需通过清洁生产、系统集成和生态技术应用，实现处理过程的生态化与源头融合。将水处理系统视为模拟自然生态的单元，通过改革生产工艺、加强管理实现源头减污；在工业园区构建“工业生态园”，实现企业间废物与资源的交换共生；结合人工湿地、稳定塘等生态技术进行深度净化，降低能耗并提升生态效益。",
        "keywords": ["清洁生产", "工业生态园", "源头减量", "生态处理技术", "系统集成"],
        "key_info": "将水处理系统视为模拟自然生态的单元。通过改革生产工艺、加强管理实现源头减污；在工业园区构建“工业生态园”，实现企业间废物与资源的交换共生；结合人工湿地、稳定塘等生态技术进行深度净化，降低能耗并提升生态效益。",
        "reference_chunks": [99, 5, 168, 9, 137, 108, 17],
         "discard_rationale": "该逻辑块关于清洁生产和生态技术，其核心信息（源头减污、系统集成）已包含在逻辑块2、4、9的“生态化”、“系统集成”、“实现路径”等相关论述中，属于冗余信息。"
      },
      {
        "reference_blocks": [-5],
        "argument": "针对难降解与有毒污染物的强化预处理与安全处置，是保障资源回收系统生态安全的前提。在资源回收导向下，仍需确保残留有毒有害物质（如抗生素、持久性有机污染物）被高效去除。这要求开发并应用针对性的强化预处理技术（如强化水解、高级氧化）以消除其生物抑制性，控制抗性基因扩散，防止回收过程产生二次污染，保障出水回用与副产物资源化的生态安全性。",
        "keywords": ["难降解有机物", "有毒污染物", "预处理", "高级氧化", "生态安全"],
        "key_info": "在资源回收导向下，仍需确保残留有毒有害物质（如抗生素、持久性有机污染物）被高效去除。这要求开发并应用针对性的强化预处理技术（如强化水解、高级氧化）以消除其生物抑制性，控制抗性基因扩散，防止回收过程产生二次污染，保障出水回用与副产物资源化的生态安全性。",
        "reference_chunks": [66, 43, 81, 151, 57],
         "discard_rationale": "该逻辑块聚焦有毒污染物预处理与生态安全，虽然重要，但相对于问题核心的“重新定义技术目标与评价体系”而言，属于更具体的工艺保障措施，对直接解答问题帮助有限，可视为弱影响信息。"
      },
      "..."
    ],
    "golden": [
      {
        "reference_blocks": 1,
```

"argument": "工业水处理系统转型的核心是将技术目标从单一的达标排放，升级为以水资源高效循环回用为核心的多层次循环经济目标。技术目标需从“无害化排放”转向生产满足不同用途的高品质再生水，实现路径包括企业内部或园区内的清污分流、分质处理、一水多用和串级使用，以及采用以“双膜法”为代表的深度处理技术。",
"keywords": ["技术目标", "水资源循环", "再生水", "梯级利用", "深度处理"],
"key_info": "技术目标需从“无害化排放”转向生产满足不同用途（如工业冷却、锅炉补给、市政杂用、农业灌溉）的高品质再生水。实现路径包括企业内部或园区内的清污分流、分质处理、一水多用和串级使用，以及采用以“双膜法”为代表的深度处理技术。",
"reference_chunks": [34, 70, 17, 51, 21, 90, 59]
},
{
"reference_blocks": [2, -1],
"argument": "技术目标应涵盖从废水中回收能源与有价值物质，实现污染物资源化转化，构建从末端无害化处理升级为水、能、物协同回收与闭环循环的多层次体系。具体包括：对高浓度有机废水采用厌氧技术回收沼气能源；回收氮、磷等营养物质；从特定工业废水中回收原料、贵重金属或有价值的副产品。这要求通过厌氧消化、生物脱氮除磷、膜分离等技术，实现水资源高效回用与闭环循环、能源与资源的同步回收、污染物的最小化与无害化，以及工艺过程的生态化与低碳化。",
"keywords": ["能源回收", "物质回收", "厌氧处理", "营养物质回收", "副产品", "技术目标", "资源回收", "水资源回用"],
"key_info": "将废水中的有机物和营养物质视为资源进行转化回收。具体包括：对高浓度有机废水采用厌氧技术（如UASB）回收沼气能源；回收氮、磷等营养物质；以及从特定工业废水中回收原料、贵重金属或有价值的副产品（如电镀废水中的金属、造纸黑液中的碱、味精废水中的菌体蛋白）。\n技术目标应超越传统的污染物去除，扩展为水资源高效回用与闭环循环、能源与资源的同步回收、污染物的最小化与无害化、以及工艺过程的生态化与低碳化。这要求深度处理以满足回用标准，并通过厌氧消化、生物脱氮除磷、膜分离等技术回收沼气、营养物质和有价物质。",
"reference_chunks": [3, 36, 69, 42, 138, 45, 16, 17, 52, 54, 23, 153, 188, 31]
},
{
"reference_blocks": [-3],
"argument": "实现转型需要革新工艺路线，发展集成化的“资源化工艺包”，并推动源头分类与分散式处理以提升回收针对性。工艺路径应从单一处理链转向集预处理回收、厌氧产沼、主流脱氮除磷、深度处理与膜分离、浓水处理于一体的“资源化工艺包”。同时，借鉴分散式处理与再循环（DESAR）理念，在源头对黑水、灰水等不同性质污水进行分类，以实现更高效、有针对性的资源回收。",
"keywords": ["工艺革新", "集成工艺", "分散式处理", "工艺包", "源头分离"],
"key_info": "工艺路径应从单一处理链转向集预处理回收、厌氧产沼、主流脱氮除磷、深度处理与膜分离、浓水处理于一体的“资源化工艺包”。同时，借鉴分散式处理与再循环（DESAR）理念，在源头对黑水、灰水等不同性质污水进行分类，以实现更高效、有针对性的资源回收。",
"reference_chunks": [104, 43, 16, 52, 95]
},
{
"reference_blocks": [4, -2],
"argument": "新的评价体系需建立以生态循环绩效为核心的多维度综合评估框架，从单一环境指标转向涵盖资源效率、生态影响、经济可行性与系统智能性的综合指标。评价指标应超越单一的处理效率，涵盖水资源循环（如重复利用率、再生水回用率）、物质循环与回收（如能源转化率、营养盐回收率）、生态足迹（能耗、温室气体排放、污泥资源化率）、系统稳健性（抗冲击负荷能力、出水稳定性）、资源回收效率指标（如水回收率、能源净产出率）、生态与环境影响指标（如碳足迹、生态毒性削减）、技术经济可行性指标（如全生命周期成本、投资回收期）以及系统协同与智能管控指标（如物质能量流集成度、自动化水平）。",
"keywords": ["评价体系", "生态循环绩效", "水资源循环指标", "物质回收指标", "生态足迹", "资源回收效率", "生态影响", "全生命周期成本", "智能管控"],
"key_info": "评价指标应超越单一的处理效率，涵盖水资源循环（如重复利用率、再生水回用率）、物质循环与回收（如能源转化率、营养盐回收率）、生态足迹（能耗、温室气体排放、污泥资源化率）以及系统稳健性（抗冲击负荷能力、出水稳定性）。\n新型评价体系应包含资源回收效率指标（如水回收率、能源净产出率）、生态与环境影响指标（如碳足迹、生态毒性削减）、技术经济可行性指标（如全生命周期成本、投资回收期）以及系统协同与智能管控指标（如物质能量流集成度、自动化水平），以全面衡量转型成效。",
"reference_chunks": [104, 138, 76, 45, 176, 17, 16, 51, 21, 23, 153, 124, 95, 57]
},
{
"reference_blocks": [5, -4],
"argument": "经济可行性评价需基于全生命周期成本分析和成本-效益分析，并考虑市场化因素，同时需应对高级氧化、膜分离等技术成本较高的经济性挑战。评价需计算包括建设投资、运行费用、设备折旧、资源回收收益、节省的水费与排污费等在内的全生命周期成本，进行成本-效益分析，确保总效益大于总成本。同时评估技术的规模化适用性、再生水及回收产品的市场竞争力，并推动“第三方治理”等市场化运营模式。突破经济瓶颈需要依靠技术创新降低单元工艺能耗与物耗，同时通过建立资源化产品标准、提供政策补贴、发展绿色金融等手段，提升资源回收项目的市场竞争力与投资回报率。",
"keywords": ["经济可行性", "全生命周期成本", "成本效益分析", "市场化", "第三方治理", "技术成本", "政策驱动", "市场机制", "绿色金融"],
"key_info": "评价需计算包括建设投资、运行费用、设备折旧、资源回收收益、节省的水费与排污费等在内的全生命周期成本。进行成本-效益分析，确保总效益大于总成本。同时评估技术的规模化适用性、再生水及回收产品的市场竞争力，并推动“第三方治理”等市场化运营模式。\n高级氧化、膜分离、蒸发结晶等深

1268

度处理与资源回收技术成本较高，是转型的主要经济瓶颈。突破需要依靠技术创新降低单元工艺能耗与物耗，同时通过建立资源化产品标准、提供政策补贴、发展绿色金融等手段，提升资源回收项目的市场竞争力与投资回报率。",
        "reference_chunks": [64, 34, 130, 104, 9, 106, 43, 16, 113, 95]
      },
      ...
    ]
}

Subordinate (Refinement Turn #2)

[
    {
      "argument": "技术目标应从单一污染物去除转向水资源、能源与物质的多层次协同回收。",
      "keywords": ["资源回收", "多层次目标", "水资源回用", "能源回收", "物质回收"],
      "key_info": "技术目标体系需进行根本性重构，其核心是实现水资源的分质、分级、梯级利用（如工艺用水、冷却水、生态补水），并协同回收废水中的化学能（如沼气）及高价值物质（如重金属、营养盐），将废水视为稳定的替代水源和资源库。",
      "reference_chunks": [20, 60, 86, 129]
    },
    {
      "argument": "评价体系需构建涵盖技术效能、生态效益、经济可行性和系统稳健性的多维综合指标。",
      "keywords": ["评价体系", "多维指标", "技术效能", "生态效益", "经济可行性"],
      "key_info": "新评价体系需突破传统以‘去除率’和‘达标排放’为核心的框架，建立包括资源回收效率（如水回收率、能源净产量）、过程效能（如综合能耗、药耗）、生态效益（如碳足迹、生态毒性削减）、经济可行性（如全生命周期成本、投资回报率）及系统稳健性（如抗冲击负荷能力、智能管控水平）在内的综合指标体系。",
      "reference_chunks": [18, 44, 49, 93, 100, 114, 195]
    },
    {
      "argument": "转型过程需强化风险控制，确保有毒有害物质被有效去除以保障资源回收的生态安全性。",
      "keywords": ["生态安全", "风险控制", "有毒污染物", "强化预处理", "二次污染"],
      "key_info": "在追求资源回收的同时，必须确保有毒有害物质（如持久性有机物、重金属、抗生素）被高效去除，防止其在回收链条中迁移富集或产生二次污染。这要求开发并应用针对性的强化预处理技术（如高级氧化、强化水解），以消除其生物抑制性，保障后续处理单元稳定运行及回收副产物的安全利用。",
      "reference_chunks": [61, 86, 140]
    },
    {
      "argument": "系统生态化与源头融合是转型的重要过程目标，旨在模拟自然并实现低能耗运行。",
      "keywords": ["生态化", "源头融合", "清洁生产", "工业生态园", "低能耗"],
      "key_info": "将水处理系统设计为模拟自然生态的单元，强调低能耗和生态融合。通过贯彻清洁生产理念从源头减少污染产生和用水量，并在工业园区层面构建‘工业生态园’，实现企业间废物与资源的交换共生。结合人工湿地、稳定塘等生态技术进行深度净化，可降低能耗并提升生态效益。",
      "reference_chunks": [1, 44, 133, 150, 163]
    },
    {
      "argument": "智慧化管控与跨学科融合是提升系统稳定性、适应性与整体效益的关键技术支撑。",
      "keywords": ["智慧化管控", "跨学科融合", "系统稳定性", "智能监测", "工艺优化"],
      "key_info": "通过建立基于物联网和大数据的智慧水厂管理系统，实现从源头到回用终端的全过程精细化、智能化管控，这是保障资源回收产品（水、能、物）质量稳定、提升系统抗冲击能力、实现生态与经济双重目标的关键技术支撑。",
      "reference_chunks": [11, 67, 73, 127, 194]
    },
    {
      "argument": "政策标准、商业模式与技术创新共同构成推动转型的支撑体系。",
      "keywords": ["政策驱动", "标准体系", "商业模式创新", "技术创新", "市场机制"],
      "key_info": "转型需要政策与标准驱动，如制定鼓励资源回收的法规、经济激励政策和完善的再生水水质标准体系。同时，需通过商业模式创新（如合同节水管理、能源合同管理）降低业主风险并创造市场动力。持续的技术创新与集成（如高效厌氧技术、精准膜分离）则是实现经济可行性的根本保障。",
      "reference_chunks": [18, 20, 60, 114, 181]
    }
]

Commit (Refinement Turn #2)

{
    "trash": [
      {
        "reference_blocks": 3,

"argument": "转型需通过清洁生产、系统集成和生态技术应用，实现处理过程的生态化与源头融合。将水处理系统视为模拟自然生态的单元，通过改革生产工艺、加强管理实现源头减污；在工业园区构建“工业生态园”，实现企业间废物与资源的交换共生；结合人工湿地、稳定塘等生态技术进行深度净化，降低能耗并提升生态效益。",
"keywords": ["清洁生产", "工业生态园", "源头减量", "生态处理技术", "系统集成"],
"key_info": "将水处理系统视为模拟自然生态的单元。通过改革生产工艺、加强管理实现源头减污；在工业园区构建“工业生态园”，实现企业间废物与资源的交换共生；结合人工湿地、稳定塘等生态技术进行深度净化，降低能耗并提升生态效益。",
"reference_chunks": [99, 5, 168, 9, 137, 108, 17],
"discard_rational": "该逻辑块关于清洁生产和生态技术，其核心信息（源头减污、系统集成）已包含在逻辑块2、4、9的“生态化”、“系统集成”、“实现路径”等相关论述中，属于冗余信息。"
},
{
"reference_blocks": [-5],
"argument": "针对难降解与有毒污染物的强化预处理与安全处置，是保障资源回收系统生态安全的前提。在资源回收导向下，仍需确保残留有毒有害物质（如抗生素、持久性有机污染物）被高效去除。这要求开发并应用针对性的强化预处理技术（如强化水解、高级氧化）以消除其生物抑制性，控制抗性基因扩散，防止回收过程产生二次污染，保障出水回用与副产物资源化的生态安全性。",
"keywords": ["难降解有机物", "有毒污染物", "预处理", "高级氧化", "生态安全"],
"key_info": "在资源回收导向下，仍需确保残留有毒有害物质（如抗生素、持久性有机污染物）被高效去除。这要求开发并应用针对性的强化预处理技术（如强化水解、高级氧化）以消除其生物抑制性，控制抗性基因扩散，防止回收过程产生二次污染，保障出水回用与副产物资源化的生态安全性。",
"reference_chunks": [66, 43, 81, 151, 57],
"discard_rational": "该逻辑块聚焦有毒污染物预处理与生态安全，虽然重要，但相对于问题核心的“重新定义技术目标与评价体系”而言，属于更具体的工艺保障措施，对直接解答问题帮助有限，可视为弱影响信息。"
},
...
],
"golden": [
{
"reference_blocks": 1,
"argument": "工业水处理系统转型的核心技术目标是将单一的达标排放升级为以水资源高效循环回用为核心的多层次循环经济目标。具体而言，技术目标需从“无害化排放”转向生产满足不同用途（如工业冷却、锅炉补给、市政杂用、农业灌溉）的高品质再生水。实现路径包括在企业内部或园区内实施清污分流、分质处理、一水多用和串级使用，并采用以“双膜法”为代表的深度处理技术，以实现水资源的梯级利用。",
"keywords": ["技术目标", "水资源循环", "再生水", "梯级利用", "深度处理"],
"key_info": "技术目标需从“无害化排放”转向生产满足不同用途（如工业冷却、锅炉补给、市政杂用、农业灌溉）的高品质再生水。实现路径包括企业内部或园区内的清污分流、分质处理、一水多用和串级使用，以及采用以“双膜法”为代表的深度处理技术。",
"reference_chunks": [34, 70, 17, 51, 21, 90, 59]
},
{
"reference_blocks": [2, -1],
"argument": "技术目标应超越传统的污染物去除，升级为从废水中协同回收水资源、能源与有价值物质的多层次体系。这包括：将废水视为稳定的替代水源和资源库，实现水资源的分质、分级、梯级利用；对高浓度有机废水采用厌氧技术（如UASB）回收沼气能源；回收氮、磷等营养物质；以及从特定工业废水中回收原料、贵重金属或有价值的副产品（如电镀废水中的金属、造纸黑液中的碱）。这要求通过厌氧消化、生物脱氮除磷、膜分离等技术，实现水资源高效回用与闭环循环、能源与资源的同步回收、污染物的最小化与无害化，以及工艺过程的生态化与低碳化。",
"keywords": ["技术目标", "资源回收", "水资源回用", "能源回收", "物质回收", "厌氧处理", "营养物质回收", "副产品"],
"key_info": "将废水中的有机物和营养物质视为资源进行转化回收。具体包括：对高浓度有机废水采用厌氧技术（如UASB）回收沼气能源；回收氮、磷等营养物质；以及从特定工业废水中回收原料、贵重金属或有价值的副产品（如电镀废水中的金属、造纸黑液中的碱、味精废水中的菌体蛋白）。\n技术目标应超越传统的污染物去除，扩展为水资源高效回用与闭环循环、能源与资源的同步回收、污染物的最小化与无害化、以及工艺过程的生态化与低碳化。这要求深度处理以满足回用标准，并通过厌氧消化、生物脱氮除磷、膜分离等技术回收沼气、营养物质和有价物质。\n技术目标体系需进行根本性重构，其核心是实现水资源的分质、分级、梯级利用（如工艺用水、冷却水、生态补水），并协同回收废水中的化学能（如沼气）及高价值物质（如重金属、营养盐），将废水视为稳定的替代水源和资源库。",
"reference_chunks": [129, 3, 36, 69, 42, 138, 45, 60, 16, 17, 52, 20, 54, 23, 86, 153, 188, 31]
},
{
"reference_blocks": 3,
"argument": "实现转型需要革新工艺路线，发展集成化的“资源化工艺包”，并推动源头分类与分散式处理以提升回收针对性。工艺路径应从单一处理链转向集预处理回收、厌氧产沼、主流脱氮除磷、深度处理与膜分离、浓水处理于一体的“资源化工艺包”。同时，借鉴分散式处理与再循环（DESAR）理念，在源头对黑水、灰水等不同性质污水进行分类，以实现更高效、有针对性的资源回收。",
"keywords": ["工艺革新", "集成工艺", "分散式处理", "工艺包", "源头分离"],

"key_info": "工艺路径应从单一处理链转向集预处理回收、厌氧产沼、主流脱氮除磷、深度处理与膜分离、浓水处理于一体的“资源化工艺包”。同时，借鉴分散式处理与再循环（DESAR）理念，在源头对黑水、灰水等不同性质污水进行分类，以实现更高效、有针对性的资源回收。",
"reference_chunks": [104, 43, 16, 52, 95]
},
{
"reference_blocks": [4, -2],
"argument": "新的评价体系需建立以生态循环绩效为核心的多维度综合评估框架，从单一环境指标转向涵盖技术效能、生态效益、经济可行性与系统稳健性的综合指标。具体指标应超越传统的处理效率，涵盖资源回收效率（如水回收率、能源净产出率、营养盐回收率）、过程效能（如综合能耗、药耗）、生态与环境影响（如碳足迹、生态毒性削减、污泥资源化率）、技术经济可行性（如全生命周期成本、投资回收期）以及系统协同与智能管控（如物质能量流集成度、自动化水平、抗冲击负荷能力）。",
"keywords": ["评价体系", "生态循环绩效", "多维指标", "资源回收效率", "生态影响", "全生命周期成本", "技术效能", "系统稳健性"],
"key_info": "评价指标应超越单一的处理效率，涵盖水资源循环（如重复利用率、再生水回用率）、物质循环与回收（如能源转化率、营养盐回收率）、生态足迹（能耗、温室气体排放、污泥资源化率）以及系统稳健性（抗冲击负荷能力、出水稳定性）。\n新型评价体系应包含资源回收效率指标（如水回收率、能源净产出率）、生态与环境影响指标（如碳足迹、生态毒性削减）、技术经济可行性指标（如全生命周期成本、投资回收期）以及系统协同与智能管控指标（如物质能量流集成度、自动化水平），以全面衡量转型成效。\n新评价体系需突破传统以‘去除率’和‘达标排放’为核心的框架，建立包括资源回收效率（如水回收率、能源净产量）、过程效能（如综合能耗、药耗）、生态效益（如碳足迹、生态毒性削减）、经济可行性（如全生命周期成本、投资回报率）及系统稳健性（如抗冲击负荷能力、智能管控水平）在内的综合指标体系。",
"reference_chunks": [138, 16, 17, 18, 21, 23, 153, 44, 45, 176, 49, 51, 57, 195, 76, ..., 124]
},
{
"reference_blocks": [5, -6],
"argument": "经济可行性评价需基于全生命周期成本分析和成本-效益分析，并考虑市场化因素与政策标准、商业模式、技术创新的协同支撑。评价需计算包括建设投资、运行费用、设备折旧、资源回收收益、节省的水费与排污费等在内的全生命周期成本，进行成本-效益分析。同时，需评估技术的规模化适用性、再生水及回收产品的市场竞争力，并推动“第三方治理”、合同节水管理等市场化运营模式以降低业主风险。突破高级氧化、膜分离等技术成本较高的经济瓶颈，需要依靠持续的技术创新降低单元工艺能耗与物耗，同时通过制定鼓励资源回收的法规、经济激励政策、完善再生水水质标准、提供政策补贴、发展绿色金融等手段，共同提升项目的市场竞争力与投资回报率。",
"keywords": ["经济可行性", "全生命周期成本", "成本效益分析", "市场化", "政策驱动", "标准体系", "商业模式创新", "技术创新", "绿色金融"],
"key_info": "评价需计算包括建设投资、运行费用、设备折旧、资源回收收益、节省的水费与排污费等在内的全生命周期成本。进行成本-效益分析，确保总效益大于总成本。同时评估技术的规模化适用性、再生水及回收产品的市场竞争力，并推动“第三方治理”等市场化运营模式。\n高级氧化、膜分离、蒸发结晶等深度处理与资源回收技术成本较高，是转型的主要经济瓶颈。突破需要依靠技术创新降低单元工艺能耗与物耗，同时通过建立资源化产品标准、提供政策补贴、发展绿色金融等手段，提升资源回收项目的市场竞争力与投资回报率。\n转型需要政策与标准驱动，如制定鼓励资源回收的法规、经济激励政策和完善的再生水水质标准体系。同时，需通过商业模式创新（如合同节水管理、能源合同管理）降低业主风险并创造市场动力。持续的技术创新与集成（如高效厌氧技术、精准膜分离）则是实现经济可行性的根本保障。",
"reference_chunks": [64, 34, 130, 104, 9, 106, 43, 16, 113, 18, 114, 20, 181, 60, 95]
},
{
"reference_blocks": -3,
"argument": "在追求资源回收的同时，必须强化风险控制，确保有毒有害物质（如持久性有机物、重金属、抗生素）被高效去除，以保障资源回收的生态安全性。这要求开发并应用针对性的强化预处理技术（如高级氧化、强化水解），防止其在回收链条中迁移富集或产生二次污染，消除其生物抑制性，从而保障后续处理单元稳定运行及回收副产物的安全利用。",
"keywords": ["生态安全", "风险控制", "有毒污染物", "强化预处理", "二次污染"],
"key_info": "在追求资源回收的同时，必须确保有毒有害物质（如持久性有机物、重金属、抗生素）被高效去除，防止其在回收链条中迁移富集或产生二次污染。这要求开发并应用针对性的强化预处理技术（如高级氧化、强化水解），以消除其生物抑制性，保障后续处理单元稳定运行及回收副产物的安全利用。",
"reference_chunks": [140, 61, 86]
},
...
]
}

Expand

[
{
"reference_blocks": 1,
"argument": "工业水处理系统转型的核心技术目标是将单一的达标排放升级为以水资源高效循环回用为核心的多层次循环经济目标。具体而言，技术目标需从“无害化排放”转向生产满足不同用途（如工业冷却、锅炉补给、市政杂用、农业灌溉）的高品质再生水。实现路径包括在企业内部或园区内实施清污分流、分质处理、一水多用和串级使用，并采用以“双膜法”为代表的深度处理技术，以实现水资源的梯级利用。",

"keywords": ["技术目标", "水资源循环", "再生水", "梯级利用", "深度处理"],
"key_info": "技术目标需从“无害化排放”转向生产满足不同用途（如工业冷却、锅炉补给、市政杂用、农业灌溉）的高品质再生水。实现路径包括企业内部或园区内的清污分流、分质处理、一水多用和串级使用，以及采用以“双膜法”为代表的深度处理技术。",
"reference_chunks": [34, 70, 17, 51, 21, 90, 59],
"retrieval_reference_chunks": [189, 79, 76, 27, 48],
"response": "在工业水处理系统从“污染控制”向“资源回收”转型的背景下，其核心技术目标必须从过去单一的达标排放，升级为以水资源高效循环回用为核心的多层次循环经济目标。这意味着技术目标的重心需从“无害化排放”转向生产能够满足不同用途的高品质再生水。具体而言，转型要求将废水不再视为“废物”进行末端处置，而是作为稳定的替代水源，通过企业内部或园区内的清污分流、分质处理，实现一水多用和串级使用。例如，在钢铁、电力等用水大户中，可根据不同工序的水质要求，将轻度污染的废水作为锅炉水力排渣用水，或将高炉煤气洗涤废水经沉淀、冷却后循环使用。这种梯级利用模式能显著提高水资源利用效率，减少新鲜水取用量。\n\n为实现这一目标，以超滤（UF）和反渗透（RO）为代表的“双膜法”深度处理技术成为关键路径。实践表明，此类技术能有效去除废水中的有机物、悬浮物和各种离子，产出满足高品质工业用水标准的再生水。例如，内蒙古托克托发电公司、北京首钢、新疆独山子石化公司等大型工业企业，均成功应用超滤-反渗透工艺处理循环排污水或化工废水，产水回用于锅炉补给水或循环冷却水补充水，实现了水资源在企业内部的闭环循环。这种深度处理与分质回用的结合，正是将技术目标从单纯处理转向资源化利用的体现，它要求处理工艺不仅能去除污染物，更能根据回用用途（如工业冷却、锅炉补给、市政杂用、农业灌溉）定制产出水质，从而支撑起一个多层次的水资源循环体系。",
"citation": [51, 27, 17, 21, 70, 79]
},
{
"reference_blocks": [2, -1],
"argument": "技术目标应超越传统的污染物去除，升级为从废水中协同回收水资源、能源与有价值物质的多层次体系。这包括：将废水视为稳定的替代水源和资源库，实现水资源的分质、分级、梯级利用；对高浓度有机废水采用厌氧技术（如UASB）回收沼气能源；回收氮、磷等营养物质；以及从特定工业废水中回收原料、贵重金属或有价值的副产品（如电镀废水中的金属、造纸黑液中的碱）。这要求通过厌氧消化、生物脱氮除磷、膜分离等技术，实现水资源高效回用与闭环循环、能源与资源的同步回收、污染物的最小化与无害化，以及工艺过程的生态化与低碳化。",
"keywords": ["技术目标", "资源回收", "水资源回用", "能源回收", "物质回收", "厌氧处理", "营养物质回收", "副产品"],
"key_info": "将废水中的有机物和营养物质视为资源进行转化回收。具体包括：对高浓度有机废水采用厌氧技术（如UASB）回收沼气能源；回收氮、磷等营养物质；以及从特定工业废水中回收原料、贵重金属或有价值的副产品（如电镀废水中的金属、造纸黑液中的碱、味精废水中的菌体蛋白）。\n技术目标应超越传统的污染物去除，扩展为水资源高效回用与闭环循环、能源与资源的同步回收、污染物的最小化与无害化、以及工艺过程的生态化与低碳化。这要求深度处理以满足回用标准，并通过厌氧消化、生物脱氮除磷、膜分离等技术回收沼气、营养物质和有价物质。\n技术目标体系需进行根本性重构，其核心是实现水资源的分质、分级、梯级利用（如工艺用水、冷却水、生态补水），并协同回收废水中的化学能（如沼气）及高价值物质（如重金属、营养盐），将废水视为稳定的替代水源和资源库。",
"reference_chunks": [129, 3, 36, 69, 42, 138, 45, 60, 16, 17, 52, 20, 54, 23, 86, 153, 188, 31],
"retrieval_reference_chunks": [48, 79, 32, 196, 27],
"response": "承接上文对技术目标从单一排放转向多层次循环利用的讨论，要实现“资源回收”的转型，技术目标必须进一步深化为构建一个从废水中协同回收水资源、能源与有价值物质的多层次体系。\n\n首先，技术目标的核心是将废水视为稳定的替代水源和资源库，实现水资源的分质、分级、梯级利用。这意味着处理工艺需根据回用对象（如工业冷却、锅炉补给、市政杂用）对水质的不同要求，进行精细化处理。例如，工业冷却水对水质要求相对较低，可采用“二级处理-混凝沉淀-过滤-消毒”等组合工艺；而作为锅炉补给水等更高品质用水，则可能需要采用反渗透（RO）等深度脱盐工艺。这种分质供水模式是实现水资源高效循环利用的基础。\n\n其次，技术目标应拓展至能源回收。对于高浓度有机废水（如酿造、食品加工废水），采用厌氧生物处理技术（如UASB反应器）不仅能高效去除有机物，还能回收沼气能源。例如，处理味精废水时，通过前段气浮回收菌体蛋白，再经厌氧处理产沼，可实现经济效益与环境效益的双重提升。这要求工艺设计超越单纯的污染物去除，将能源净产出作为重要考量。\n\n再者，有价值物质的回收是转型的另一关键维度。这包括：1) **营养物质回收**：从废水中回收氮、磷等元素，既可减轻水体富营养化压力，又可作为肥料资源。例如，可采用空气吹脱、膜分离等技术回收氨氮。2) **原料与副产品回收**：针对特定工业废水，回收其中的原料、半成品或贵重金属。例如，从造纸废液中回收碱和木质素，从电镀废水中回收金属。这要求处理工艺具备针对性的分离与富集能力，如采用化学沉淀、离子交换、膜技术等。\n\n综上所述，转型后的技术目标体系要求通过厌氧消化、生物脱氮除磷、膜分离等多种技术的耦合，最终实现**水资源的高效回用与闭环循环**、**能源与资源的同步回收**、**污染物的最小化与无害化**，以及**工艺过程的生态化与低碳化**。这不仅提升了处理系统的综合效益，也使其更紧密地契合生态循环原理。",
"citation": [48, 17, 20, 79, 16, 69, 42, 23, 153, 54, 196, 31, 45, 60]
},
...
]

Summary

{
"summary": "### 核心目标：从达标排放转向多层次循环回用\n\n工业水处理系统转型的核心技术目标，是摒弃单一的“无害化排放”范式，构建以水资源高效循环回用为核心的多层次循环经济目标体系。这一转型要求将废水从“需处置的废物”重新定义为具有特定价值的“替代水源”，其技术目标聚焦于产出满足不同用途水质标准的高品质再生水。\n\n#### 1. 目标内涵：从末端处理到资源化定向生产\n技术目标的重心从确保排放达

标，转变为根据具体回用场景定制化生产再生水。这意味着处理工艺需具备精准的污染物去除能力和水质稳定调控能力，以满足工业冷却、锅炉补给、市政杂用、农业灌溉等不同用途的差异化水质要求。这直接驱动了从“一刀切”的末端处理向“分质处理、按需供水”模式的转变。\n\n#### 2. 实现路径：企业内部/园区的梯级循环体系\n实现上述目标的关键路径在于构建企业内部或园区尺度的水循环网络，其核心操作包括：\n* **清污分流与分质处理：** 对不同污染性质和程度的废水进行源头分离与分类处理，为后续的精准回用创造条件。\n* **一水多用与串级使用：** 根据各生产工序对水质要求的差异，实现水资源的梯级利用。例如，在钢铁、电力等行业，轻度污染的废水可作为锅炉水力排渣用水，而高炉煤气洗涤废水经沉淀、冷却后即可循环回用于原工序，大幅提升单次取水的利用效率。\n* **深度处理技术支撑：** 以超滤（UF）和反渗透（RO）为代表的“双膜法”深度处理技术是实现高品质再生水生产的工程保障。该技术组合能有效去除溶解性有机物、悬浮物、盐分及各类离子，产水水质可满足高端工业用水标准。实践案例表明，如内蒙古托克托发电公司、北京首钢、新疆独山子石化公司等企业，通过应用超滤-反渗透工艺处理循环排污水或化工废水，成功将产水回用于锅炉补给水或循环冷却水系统，实现了水资源在企业内部的闭环循环。\n\n这一技术目标的重定义，本质上是将处理工艺的功能从“污染消减”升级为“资源工厂”，其评价体系也需相应地从“排放浓度合规”转向“水资源回收率、再生水水质达标率及循环经济贡献度”等综合性指标。\n\n### 技术目标深化：构建水资源、能源与物质协同回收的多层次体系\n\n在确立以“分质处理、按需供水”为核心的循环回用目标基础上，技术目标需进一步深化为构建一个从废水中协同回收水资源、能源与有价值物质的多层次资源回收体系。这要求处理工艺超越传统功能，将废水整体视为稳定的替代水源和资源库，实现水、能、物的综合回收，以契合生态循环原理并提升系统经济性。\n\n#### 1. 水资源的高效回用与闭环循环\n水资源回收是实现循环经济的基础，其技术目标核心是实现从“回用”到“梯级循环”的升级。\n* **分质分级与梯级利用：** 工艺设计需严格匹配回用对象的水质标准，实现精细化、差异化的处理路径。例如，对水质要求较低的工业冷却水，可采用“二级处理-混凝沉淀-过滤-消毒”组合工艺；而对锅炉补给水等高品质用水，则需集成反渗透（RO）等深度脱盐单元。这种分质供水模式最大化提升了水资源的利用效率。\n* **闭环循环技术支撑：** 实现企业内部或园区尺度的水闭环，依赖于膜分离（如超滤、反渗透）、高级氧化等深度处理技术的可靠性与经济性，确保再生水水质长期稳定满足高端工业回用要求。\n\n#### 2. 能源的主动回收与净产出\n将能源回收纳入核心目标，是提升处理系统经济可行性与低碳属性的关键。\n* **高浓度有机废水的能源化：** 针对酿造、食品加工等行业产生的高浓度有机废水，优先采用厌氧生物处理技术（如上流式厌氧污泥床，UASB）。该技术不仅能高效去除有机物，更能将有机质转化为沼气（主要成分为甲烷）进行能源回收。例如，在处理味精废水时，可结合前段气浮回收菌体蛋白，再经厌氧消化产沼，实现环境效益与能源收益的双重提升。\n* **工艺能耗的优化与抵消：** 技术目标应包含通过工艺优化（如厌氧氨氧化技术）降低处理过程本身的能耗，并力求使回收的能源（沼气、热能）能够部分或全部抵消系统运行能耗，甚至实现能源净输出。\n\n#### 3. 有价值物质的定向分离与回收\n从废水中回收特定物质，可直接创造经济价值并减轻环境负荷，是资源回收转型的深化体现。\n* **营养物质的回收：** 从市政及农业废水中回收氮、磷等营养物质，可转化为缓释肥料等产品。技术路径包括通过空气吹脱、磷酸铵镁（MAP）结晶、膜分离等技术回收氨氮和磷。\n* **工业原料与贵重金属的回收：** 针对特定工业废水，工艺目标需具备高效的分离与富集能力。例如，从造纸黑液中回收碱和木质素；从电镀、冶金废水中通过化学沉淀、离子交换、电渗析等技术回收铜、镍、铬等金属。这不仅减少了有毒有害物质的排放，也实现了原料的循环利用。\n\n#### 4. 工艺过程的生态化与低碳化整合\n最终的技术目标要求上述多层次回收过程实现系统整合，并导向生态化和低碳化。\n* **技术耦合与协同：** 通过厌氧消化、生物脱氮除磷（如A^2/O）、膜分离等技术的有机耦合，在一个处理框架内同步实现水质净化、能源回收与物质提取，减少处理环节和二次污染。\n* **污染物最小化与无害化：** 资源回收本身即是最有效的污染物削减路径。通过将污染物转化为资源，从根本上实现了污染物的源头减量与无害化。\n* **全生命周期低碳设计：** 将碳足迹评估纳入工艺设计与评价，优先选择能耗低、资源回收率高、温室气体排放少的工艺组合，使处理系统本身成为低碳循环的组成部分。\n\n综上，转型后的技术目标是一个集成“水资源高效回用与闭环循环”、“能源与资源同步回收”、“污染物最小化与无害化”及“工艺过程生态化与低碳化”四位一体的多层次体系。这标志着处理工艺从被动的“污染控制终端”彻底转变为主动的“资源回收与循环工厂”。...",
    "citation": [51, 27, 17, 21, 70, 79, 48, 20, 16, 69, 42, 23, 153, 54, 196, 31, 45, 60,
    77, 130, 138, 52, 43, 95, 7, 178, 44, 76, 104, 124, 9, 34, 64, 113, 106, 66, 169, 86, 142,
    1, 133, 173, 189, 163, 73, 194, 96, 67, 167, 19, 139, 143]
}

Case Study on DeepResearch Bench

Inference Trajectory of DeepWeaver on DeepResearch Bench

Q: Please write the section titled "III. Applications and Functional Capabilities" for a comprehensive report on the 'Construction and Application of a Sports Intelligent Tutoring and Learning Guidance System Driven by Multimodal Data Fusion.' This section must detail the practical applications and the specific functional capabilities of the proposed system. Explain how the system leverages multimodal data fusion to deliver intelligent tutoring and learning guidance in sports contexts, outlining its core features, use cases, and the value it provides to users such as athletes, coaches, or learners.

Draft


```
  [
      {
        "argument": "The Sports Intelligent Tutoring and Learning Guidance System (SITLS)
leverages multimodal data fusion to deliver real-time posture and movement correction by
integrating diverse sensor inputs for precise biomechanical analysis.",
        "keywords": ["real-time correction", "multimodal fusion", "posture analysis", "inertial
measurement units", "biomechanical feedback"],
```

"key_info": "The SITLS uses a multi-sensor network including IMUs, pressure sensors, and vision-based cameras to capture high-resolution kinematic data. Through feature-level and decision-level fusion, the system combines these inputs to create a unified real-time representation of movement. This allows for the detection of deviations from optimal biomechanical patterns, such as knee valgus or inefficient swing planes, and provides immediate corrective feedback via audio, visual, or haptic signals.",
"reference_chunks": [2, 4, 8, 9, 11, 14, 16, 21],
"retrieval_reference_chunks": [97, 3, 23, 62, 63]
},
{
"argument": "Comprehensive performance evaluation in the SITLS is enabled by fusing biomechanical, physiological, and visual data to assess athletic performance across multiple dimensions.",
"keywords": ["performance evaluation", "multi-dimensional assessment", "biomechanical data", "physiological monitoring", "visual analysis"],
"key_info": "The system integrates data from IMUs (capturing speed and acceleration), physiological sensors (monitoring heart rate variability and muscle oxygenation), and video analysis (for visual assessment of technique). This fusion allows for a holistic evaluation that goes beyond simple motion tracking to analyze the 'what', 'how', and 'why' of athletic performance, identifying subtle bottlenecks that single-modality systems would miss.",
"reference_chunks": [2, 4, 10, 11, 14, 16, 19, 21],
"retrieval_reference_chunks": [97, 104, 106, 109, 105]
},
{
"argument": "Personalized training and learning pathways are dynamically generated by the SITLS through continuous updates to a student model based on individual performance data and learning preferences.",
"keywords": ["personalized training", "adaptive learning", "student model", "machine learning", "dynamic adaptation"],
"key_info": "The SITLS maintains a dynamic student model that updates in real-time based on performance data and feedback. By combining this model with a domain model of sports skills and a tutor model using machine learning algorithms, the system can identify individual strengths, weaknesses, and preferences to generate personalized training plans that adapt in real-time to ensure optimal challenge and relevance.",
"reference_chunks": [2, 4, 8, 10, 14, 16, 19, 21],
"retrieval_reference_chunks": [6, 3, 23, 20, 26]
},
{
"argument": "Predictive injury risk assessment is achieved by analyzing long-term biomechanical patterns and real-time physiological signals to identify early warning signs of potential injury.",
"keywords": ["injury prevention", "predictive analytics", "biomechanical load", "fatigue monitoring", "early warning system"],
"key_info": "The SITLS integrates long-term biomechanical data (from IMUs and pressure sensors) with real-time physiological data (e.g., heart rate variability, muscle fatigue) and structural data (like CT scans) to detect deviations from an athlete's personal baseline. This enables the system to flag potential injury risks, such as asymmetry in joint loading or spikes in fatigue, long before a physical injury occurs, allowing for proactive intervention.",
"reference_chunks": [10, 11, 14, 16, 19, 21, 23, 24],
"retrieval_reference_chunks": [92, 61, 93, 75, 64]
},
{
"argument": "In precision sports like golf and gymnastics, the SITLS provides unparalleled technical guidance by combining real-time sensor data with edge AI for immediate fault detection and correction.",
"keywords": ["skill acquisition", "technical mastery", "golf training", "gymnastics analysis", "edge AI"],
"key_info": "For golf, the system uses IMUs in gloves or clubs and cameras to analyze the entire swing sequence, detecting faults like early extension or reverse pivot in real-time. In gymnastics, it tracks rotation speed and body position during complex skills, providing feedback on air awareness and landing mechanics to ensure safety and technical accuracy. The use of edge AI enables immediate processing and feedback, allowing athletes to correct technique on the spot.",
"reference_chunks": [8, 9, 14, 16, 19, 21, 24, 27],
"retrieval_reference_chunks": [96, 97, 22, 95, 61]
},
{
"argument": "In team sports, the SITLS enables collaborative dynamics analysis by fusing data from wearable sensors and positioning systems to quantify team coordination and tactical effectiveness.",

"keywords": ["team coordination", "tactical analysis", "multi-agent systems", "positional data", "temporal synchronization"],
"key_info": "The system integrates data from wearable sensors (IMUs, GPS) and positioning systems (LPS) to analyze collaborative dynamics across multiple players. It quantifies metrics like temporal coordination, spatial synchronization, and collective movement patterns, providing coaches with objective insights into team performance, such as identifying inconsistent defensive rotations or pressing intensity in soccer.",
"reference_chunks": [4, 9, 14, 16, 19, 21, 24, 28],
"retrieval_reference_chunks": [105, 104, 102, 106, 109]
},
...,
{
"argument": "The SITLS provides coaches and learners with a powerful, data-driven decision-support tool that offers objective performance insights and personalized feedback.",
"keywords": ["coaching support", "performance analytics", "data-driven decisions", "learner engagement", "objective feedback"],
"key_info": "The system aggregates data from multiple athletes to provide coaches with a comprehensive view of team performance, workload, and injury risk, enabling data-informed decisions. For learners, it offers a personalized, engaging tutor with continuous, objective feedback that helps them understand progress, identify improvement areas, and build confidence through incremental successes.",
"reference_chunks": [4, 9, 14, 16, 19, 21, 24, 28],
"retrieval_reference_chunks": [20, 26, 23, 8, 75]
}
]

Subordinate (Refinement Turn #1)

[
{
"argument": "The Sports Intelligent Tutoring and Learning Guidance System (SITLS) integrates multimodal data to deliver real-time, precise posture and motion correction by fusing IMU kinematics with video-based pose estimation.",
...
},
{
"argument": "Comprehensive performance evaluation in the SITLS is achieved through the integration of biomechanical, physiological, and movement efficiency metrics to provide a holistic assessment beyond traditional performance indicators.",
...
},
{
"argument": "The SITLS generates personalized training and learning paths by dynamically adapting to an athlete's unique physiological and performance data, creating customized regimens that evolve in real time.",
...
},
{
"argument": "Injury risk prediction and prevention in the SITLS is enabled by the fusion of biomechanical stress indicators with physiological fatigue signals, allowing for early detection of warning signs before injury occurs.",
...
},
{
"argument": "The SITLS accelerates skill acquisition in complex sports by breaking down technical skills into constituent parts and providing targeted, real-time feedback, significantly reducing the learning curve.",
...
},
{
"argument": "The SITLS enables advanced tactical analysis in team sports by fusing multi-player sensor data with video and game context, revealing patterns of coordination and cohesion that are invisible to traditional observation.",
...
},
...,
{
"argument": "The SITLS enhances coach effectiveness by providing data-driven, objective assessments of athletes and offering personalized professional development recommendations based on their own teaching behaviors.",

```
        ...
      }
    ]
```

Commit (Refinement Turn #1)

```
  {
      "trash": [
        {
          "argument": "The SITLS serves as a proactive health monitor by continuously tracking
biomechanical load and fatigue, automatically adjusting training loads or recommending recovery
protocols to prevent overuse injuries and optimize long-term athlete well-being.",
          ...
        },
        {
          "argument": "The SITLS enhances coach effectiveness by offering data-driven performance
assessments and analyzing teaching behaviors through multimodal deep learning, providing
personalized recommendations for professional development and improving pedagogical quality.",
          ...
        }
      ],
      "golden": [
        {
          "argument": "The Sports Intelligent Tutoring and Learning Guidance System (SITLS)
delivers real-time, precise posture and movement correction by fusing IMU kinematics with
video-based pose estimation to create a dynamic 3D model of athletic motion, enabling immediate
detection and feedback on deviations in joint angles and movement trajectories.",
          ...
        },
        {
          "argument": "Comprehensive performance evaluation in the SITLS is achieved through the
integration of biomechanical, physiological, and movement efficiency metrics, enabling a holistic
assessment that captures not only performance outcomes but also underlying efficiency and fatigue
indicators.",
          ...
        },
        {
          "argument": "Personalized training and learning pathways are dynamically generated by
the SITLS through a real-time updated student model that adapts to individual physiological
responses, performance data, and learning preferences using machine learning algorithms.",
          ...
        },
        {
          "argument": "Predictive injury risk assessment is enabled by fusing biomechanical stress
indicators with physiological fatigue signals, allowing early detection of non-linear patterns
such as asymmetrical joint loading and declining HRV before clinical injury occurs.",
          ...
        },
        {
          "argument": "In precision sports like golf and gymnastics, the SITLS provides real-time
fault detection and technical correction using edge AI and multimodal sensor data, enabling
immediate feedback on complex movement sequences such as swing mechanics and air awareness.",
          ...
        },
        {
          "argument": "In team sports, the SITLS enables advanced tactical analysis by fusing
multi-player sensor data, positioning systems, and video with game context to generate
spatiotemporal maps that quantify team coordination, synchronization, and collective movement
patterns.",
          ...
        },
        ...,
        {
          "argument": "Advanced tactical analysis in team sports is enhanced by the SITLS through
spatiotemporal mapping of multi-player dynamics, revealing hidden patterns of coordination,
cohesion, and formation weaknesses that are invisible to traditional observation.",
          ...
        },
        {
```

```
          "argument": "The SITLS supports athlete rehabilitation by providing objective,
quantifiable tracking of recovery progress through biomechanical and physiological metrics,
ensuring safe return-to-play decisions and minimizing re-injury risk.",
          ...
        }
      ]
    }
```


Subordinate (Refinement Turn #2)


```
  [
      {
        "argument": "The SITLS achieves superior motion analysis by fusing raw sensor data at the
signal level, reducing noise and motion artifacts to create a more accurate representation of
athletic movement.",
        ...
      },
      {
        "argument": "Feature-level fusion enables the SITLS to extract and integrate
high-dimensional, meaningful characteristics from diverse data streams into a unified
multi-dimensional vector for comprehensive performance analysis.",
        ...
      },
      {
        "argument": "Decision-level fusion allows the SITLS to generate actionable, personalized
recommendations by synthesizing fused data through advanced AI models like LSTM and Transformer
networks.",
        ...
      },
      {
        "argument": "The SITLS provides real-time, multimodal feedback for posture and technique
correction by combining wearable IMU data with video analysis to identify and correct movement
deviations instantly.",
        ...
      },
      {
        "argument": "The SITLS enables personalized skill acquisition by integrating
biomechanical and physiological data to evaluate performance and generate tailored training
regimens based on individual weaknesses.",
        ...
      },
      {
        "argument": "The SITLS supports adaptive training optimization by continuously monitoring
fatigue and physiological stress to adjust training intensity and recommend recovery protocols in
real time.",
        ...
      },
      {
        "argument": "The SITLS enhances injury prevention by predicting risks through the
correlation of long-term biomechanical and physiological data patterns, enabling proactive
interventions.",
        ...
      },
      {
        "argument": "The SITLS facilitates advanced tactical analysis in team sports by fusing
multi-player sensor data with video to visualize complex coordination patterns and team
dynamics.",
        ...
      }
    ]
```


Commit (Refinement Turn #2)


```
  {
      "trash": [
        {
          "argument": "The SITLS serves as a proactive health monitor by continuously tracking
biomechanical load and fatigue, automatically adjusting training loads or recommending recovery
protocols to prevent overuse injuries and optimize long-term athlete well-being.",
```

```
          ...
        },
        {
          "argument": "The SITLS enhances coach effectiveness by offering data-driven performance
assessments and analyzing teaching behaviors through multimodal deep learning, providing
personalized recommendations for professional development and improving pedagogical quality.",
          ...
        },
        ...
      ],
      "golden": [
        {
          "argument": "The SITLS achieves highly accurate real-time motion analysis by fusing raw
IMU and video data at the signal level, reducing noise and motion artifacts to create a robust,
dynamic 3D model of athletic movement for precise deviation detection and immediate feedback.",
          ...
        },
        {
          "argument": "The system enables comprehensive performance evaluation by fusing
biomechanical, physiological, and movement metrics at the feature level, extracting
high-dimensional characteristics into a unified vector that captures the interplay between
physical mechanics and biological state.",
          ...
        },
        {
          "argument": "Personalized training and learning pathways are dynamically generated
through decision-level fusion of fused data using advanced AI models like LSTM and Transformer
networks, enabling real-time, adaptive recommendations tailored to individual athlete profiles.",
          ...
        },
        {
          "argument": "Predictive injury risk assessment is enabled by analyzing long-term
patterns in biomechanical stress and physiological fatigue, detecting early warning signs such as
asymmetrical joint loading and declining HRV to support proactive intervention and injury
prevention.",
          ...
        },
        {
          "argument": "Real-time technical correction in precision sports like golf and gymnastics
is achieved by fusing IMU and video data to instantly detect faults in movement sequences—such as
swing mechanics or air awareness—and deliver immediate multimodal feedback via audio, visual, or
haptic cues.",
          ...
        },
        {
          "argument": "Advanced tactical analysis in team sports is enhanced by fusing
multi-player sensor data with video to visualize spatiotemporal coordination patterns, enabling
objective quantification of team dynamics, synchronization, and formation weaknesses.",
          ...
        },
        ...,
        {
          "argument": "The SITLS enables personalized skill acquisition by integrating
biomechanical and physiological data to evaluate performance and generate tailored training
regimens based on individual weaknesses.",
          ...
        },
        {
          "argument": "The SITLS supports adaptive training optimization by continuously
monitoring fatigue and physiological stress to adjust training intensity and recommend recovery
protocols in real time.",
          ...
        }
      ]
    }
```


Expand


```
  [
```

```
      {
        "reference_blocks": [1, -1],
        "argument": "The SITLS achieves highly accurate real-time motion analysis by fusing raw
IMU and video data at the signal level, reducing noise and motion artifacts to create a robust,
dynamic 3D model of athletic movement for precise deviation detection and immediate feedback.",
        "keywords": ["real-time correction", "multimodal fusion", "posture analysis", "IMU",
"video data", "3D motion modeling", "signal-level fusion", "motion analysis", "noise reduction",
"sensor synchronization", "data robustness"],
        "key_info": "The SITLS uses a multi-sensor network including IMUs, pressure sensors, and
vision-based cameras to capture high-resolution kinematic data. Through feature-level and
decision-level fusion, the system combines these inputs to create a unified real-time
representation of movement. ...",
        "reference_chunks": [2, 34, 4, 36, 70, 7, 8, 9, 11, 107, 14, 46, 16, 17, 78, 21, 91, 30],
        "retrieval_reference_chunks": [97, 96, 62, 106, 61],
        "response": "The Sports Intelligent Tutoring and Learning System (SITLS) achieves highly
accurate real-time motion analysis by fusing raw Inertial Measurement Unit (IMU) and video data
at the signal level, enabling a robust, dynamic 3D model of athletic movement that supports
immediate feedback and precise deviation detection. Signal-level fusion combines unprocessed
sensor streams—such as accelerometer, gyroscope, and magnetometer data from IMUs—with
synchronized video frames at the most fundamental data level, allowing for enhanced noise
reduction and motion artifact suppression before higher-level interpretation [78]. This approach
overcomes limitations inherent in data-level fusion, where disparate sensor formats and noise
profiles can lead to information conflict or redundancy, and decision-level fusion, which suffers
from low information utilization and poor pattern mining [62]. By aligning raw IMU and video data
through precise temporal synchronization, the SITLS constructs a temporally coherent 3D human
skeleton model that reflects real-time kinematics with millimeter-level accuracy [91]. This
integration enables the system to detect subtle deviations in posture, joint angles, and movement
trajectories during high-speed or complex athletic actions, such as a golf swing or gymnastic
dismount, where even minor errors can lead to performance degradation or injury risk. The
resulting model serves as the foundational input for downstream functions, including real-time
posture correction, performance evaluation, and technical feedback, ensuring that all intelligent
tutoring outputs are grounded in a highly accurate and reliable representation of the athlete's
actual movement [62]. This capability is particularly critical in dynamic, open-ended sports
domains where traditional single-modal systems fail to capture the full complexity of human
motion, thereby enabling the SITLS to deliver data-driven, context-aware guidance that surpasses
the limitations of conventional coaching and feedback mechanisms [2].",
        "citation": [78, 62, 91, 2]
      },
      ...
    ]
```


Summary


```
{
```

"summary": "A. Core Functional Capabilities\n1. Real-Time Posture Correction and Motion Analysis via Signal-Level Multimodal Fusion\nThe Sports Intelligent Tutoring and Learning System (SITLS) enables highly accurate real-time motion analysis by fusing raw Inertial Measurement Unit (IMU) and video data at the signal level, establishing a robust, dynamic 3D model of athletic movement for precise deviation detection and immediate feedback. This signal-level fusion integrates unprocessed IMU streams—including accelerometer, gyroscope, and magnetometer data—with synchronized video frames at the most fundamental data level, ensuring temporal coherence and minimizing noise and motion artifacts prior to higher-level interpretation [78]. By circumventing the limitations of data-level fusion—where mismatched sensor formats and noise profiles cause information conflict—and decision-level fusion—which suffers from reduced information utilization and weak pattern mining—the SITLS achieves superior data integration [62]. Precise temporal synchronization between IMU and video streams allows the system to construct a temporally consistent 3D human skeleton model with millimeter-level accuracy, capturing real-time kinematics during high-speed or complex movements such as a golf swing or gymnastic dismount [91]. This model facilitates the detection of subtle deviations in posture, joint angles, and movement trajectories, enabling early identification of biomechanical inefficiencies or injury risks. The resulting high-fidelity representation serves as the core input for downstream intelligent tutoring functions, including real-time posture correction, performance evaluation, and technical feedback, ensuring that guidance is both data-driven and context-aware. This capability is especially critical in dynamic, open-ended sports environments where single-modal systems fail to capture the full complexity of human motion, allowing the SITLS to deliver superior, adaptive coaching that surpasses traditional feedback mechanisms [2].\n\n2. Multi-Dimensional Performance Evaluation through Feature-Level Multimodal Fusion\nThe Sports Intelligent Tutoring and Learning System (SITLS) delivers a comprehensive, multi-dimensional performance evaluation by fusing biomechanical, physiological, and movement metrics at the feature level, transforming heterogeneous sensor data into a unified, high-dimensional vector that captures the dynamic interplay between physical mechanics and biological state. This capability is achieved through advanced feature-level fusion, which extracts modality-specific characteristics—such as joint angles, ground reaction forces, muscle activation patterns, heart rate variability (HRV), and lactate thresholds—from raw data streams before aligning them in a shared semantic space [62]. By processing features rather than raw signals or final decisions, the system avoids the pitfalls of data-level fusion—where mismatched noise profiles and sampling rates induce information conflict—and decision-level fusion, which underutilizes low-level patterns and degrades discriminative power [62,67]. This approach preserves rich, discriminative information while reducing redundancy and enhancing computational efficiency. The resulting multi-dimensional vector serves as a dynamic performance fingerprint, enabling real-time assessment of movement efficiency, energy expenditure, neuromuscular coordination, biomechanical symmetry, and fatigue onset. Integrated data from IMUs, wearable ECGs, force-sensitive insoles, and video-based motion capture allows the SITLS to detect subtle deviations in technique and early warning signs of overtraining or injury risk—capabilities unattainable by single-modality systems [91]. As demonstrated in related research, the fusion of structural and biomechanical data via deep learning significantly improves injury prediction accuracy, validating the value of integrated, multi-dimensional analysis [91]. In the SITLS, this fused representation not only informs immediate feedback but also supports longitudinal trajectory analysis, enabling coaches and athletes to optimize training loads, track skill progression, and refine long-term development strategies. The holistic assessment thus establishes a robust, data-driven foundation for intelligent tutoring, empowering users with deeper insights into execution quality, training readiness, and performance potential across all phases of athletic development.\n\n ...",
"citation": [78, 62, 91, 2, 67, 1, 35, 96, 8, 102, 104, 105, 79, 75, 97, 73, 22, 26, 64, 4],
}

Algorithm 1 DeepWeaver Pseudocode

Input: Question q, TBC refinement rounds n, retrieval parameter k
Input: Number of evidence fragments per TBC generation r.
Output: Generated answer y.

Load question q, evidence pool $E = \{e_1, \ldots, e_m\}$
$\mathbf{H} \leftarrow \text{Emb}(e_1, e_2, \ldots, e_m)$
$E_r \leftarrow \textsc{RandomSampling}(E, r)$
$\mathcal{T}_M^{(0)} \leftarrow \textsc{Draft}(q, E_r)$
    ▷ Draft: Generate the initial main Thought Block Chain (TBC).
$\mathcal{T}_{\text{discard}} \leftarrow \emptyset$
for $t \leftarrow 0$ **to** $n-1$ **do**
    $E_{\text{cover}} \leftarrow \textsc{RetrieveCovered}(E, \mathbf{H}, \mathcal{T}_M^{(t)}, k)$
    $\mathcal{R}^{(t)} \leftarrow \textsc{RandomSampling}(E \setminus E_{\text{cover}}, r)$
    ▷ Collect evidence that is neglected or only weakly covered.
    $\mathcal{T}_S^{(t)} \leftarrow \textsc{Subordinate}(q, \mathcal{R}^{(t)}, \mathcal{T}_{\text{discard}})$
    ▷ Subordinate: Construct a subordinate TBC from the residual evidence.
    $\widetilde{\mathcal{T}}^{(t)} \leftarrow \textsc{Merge}(\mathcal{T}_M^{(t)}, \mathcal{T}_S^{(t)})$
    ▷ Commit-Merge: Commit useful subordinate claims into the main TBC.
    $(\mathcal{T}_{\text{discard}}, \mathcal{T}_M^{(t+1)}) \leftarrow \textsc{Discard}(q, \widetilde{\mathcal{T}}^{(t)}, \mathcal{T}_{\text{discard}})$
    ▷ Commit-Discard: Remove redundant, irrelevant, or weakly supported blocks.
end for
$\mathcal{T}_M^{(n)} \leftarrow \textsc{LinkEvidence}(T_M^{(n)}, \textsc{RetrieveCovered}(E, \mathbf{H}, \mathcal{T}_M^{(n)}, k))$
    ▷ Link evidence with thought blocks based on top-k embedding retrieval.
$y_0 \leftarrow \emptyset$
for all $b_i = (c_i, k_i, s_i, E_i) \in \mathcal{T}_M^{(n)}$ **do**
    $S_i \leftarrow \textsc{Generate}(q, c_i, s_i, E_i)$
    ▷ Expand: Generate answer sections from each thought block.
    $y_i \leftarrow \textsc{Append}(y_{i-1}, S_i)$
    ▷ Summary: Compose evidence-grounded sections into the final answer sequentially via LLM.
end for
$y \leftarrow y_{|\mathcal{T}_M^{(n)}|}$