Q-Evolve : A self-evolving knowledge assistant for superconducting quantum computing

Anonymous ACL submission

Abstract

The rapid growth of research literature in superconducting quantum computing has made it increasingly difficult to synthesize fragmented findings into actionable insights. While retrieval-augmented LLMs expand coverage beyond static parametric memory, existing pipelines remain constrained by the absence of logical dependence modeling and prohibitive preprocessing costs, limiting their practical utility in this domain. We introduce **Q-Evolve**, a knowledge assistant that integrates semantic retrieval with a dynamically expanding logical dependence graph, constructed lazily from retrieval outputs without text-to-graph preprocessing. It employs bidirectional self-evolution: retrieval-augmented outputs inform dependence reasoning, and refined dependence structures guide subsequent retrieval. On a domain-specific quantum computing benchmark, Q-Evolve consistently outperforms the evaluated general-purpose LLM baselines and the NVIDIA Ising quantum calibration model. These findings indicate that Q-Evolve keeps pace with the evolving literature, extracting cross-study regularities and producing informed responses that bring automated scientific discovery closer to practical use.

1 Introduction

Recent advances in large language models (LLMs) have accelerated literature review workflows. These models rely on static parametric knowledge that can become outdated or incomplete (Anthropic, 2026; OpenAI, 2026). Retrieval-Augmented Generation (RAG) mitigates this limitation by grounding LLM outputs in dynamically retrieved external evidence (Lewis et al., 2020). When deployed over large scholarly corpora, however, conventional RAG pipelines remain insufficient for high-fidelity scientific synthesis.

Work on enhancing RAG through iterative retrieval-generation cycles and graph-based indexing (Shao et al., 2023; Trivedi et al., 2023; Edge et al., 2024; Guo et al., 2025) has demonstrated progress, with lightweight variants reducing cost via NER-based structural indexing (Song et al., 2026; Zhuang et al., 2026). Despite these advances, these methods do not model logical dependencies between concepts. They link documents that share surface terms or entity mentions but cannot capture how one physical mechanism governs another. Retrievers therefore fail to distinguish functionally dependent evidence from merely topically related evidence, a distinction that scientific synthesis requires.

Attempts to incorporate such logical dependencies, however, introduce prohibitive preprocessing costs. Methods like CausalRAG (Wang et al., 2025) and HGMEM (Zhou et al., 2026) rely on LLM-driven construction of causal graphs or hypergraphs from every ingested document. In rapidly evolving domains such as superconducting quantum computing, this overhead creates two compounding challenges: the computational cost of processing continuously growing corpora becomes unsustainable, and portions of the pre-constructed graph risk obsolescence before deployment. These constraints render such methods impractical for domains demanding both temporal currency and interactive responsiveness.

No existing approach simultaneously achieves deep cross-document logical dependence reasoning, continuous knowledge accumulation, and scalable preprocessing. This gap motivates our work.

We propose **Q-Evolve** (overall architecture shown in Figure 1), a knowledge assistant that resolves this trilemma through a **dual-pathway architecture** unifying semantic dense retrieval with a **lazily expanding dependence graph**. Unlike static RAG or offline memory-intensive paradigms, Q-Evolve engages in **bidirectional self-evolution**: semantic retrieval informs graph construction, and the enriched dependence graph in turn guides subse-

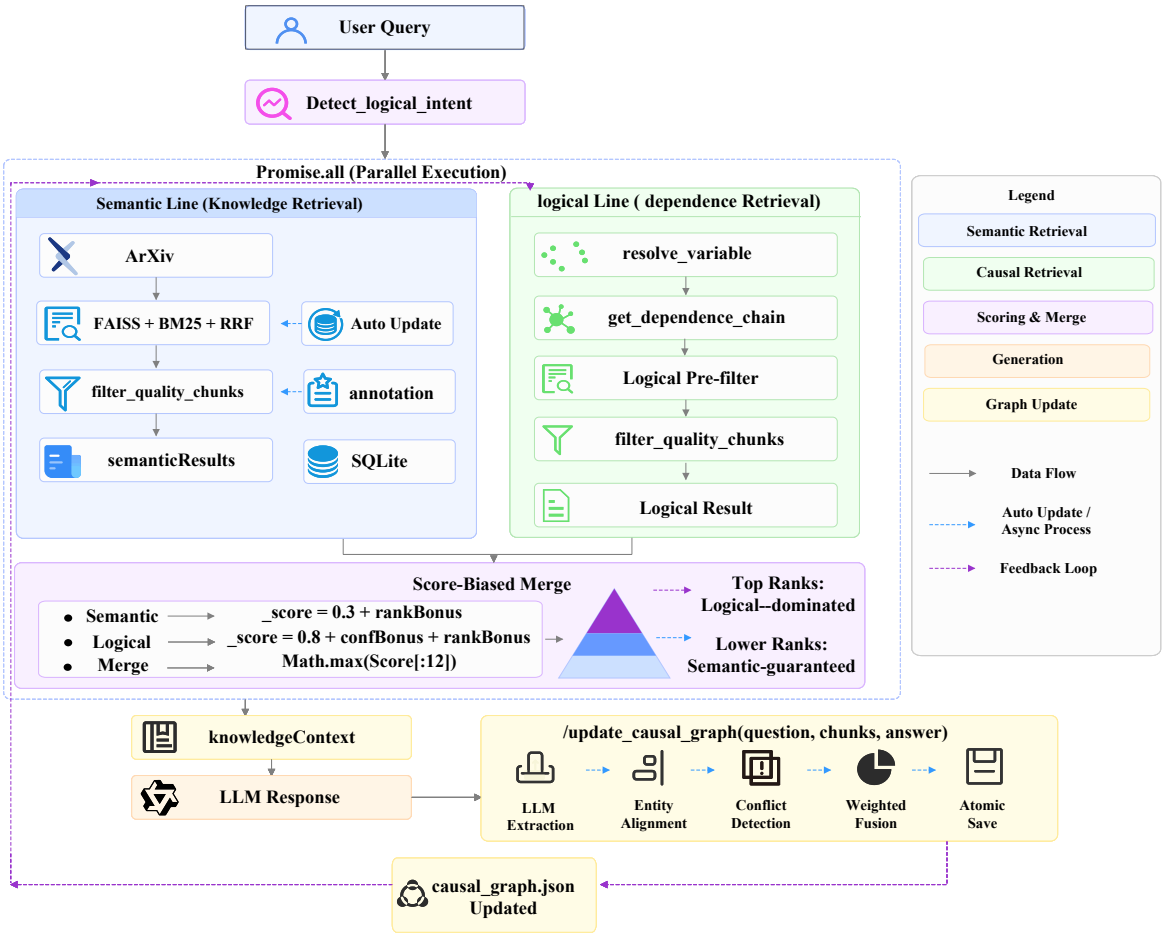

Figure 1: The overall workflow of Q-Evolve includes four stages: hybrid document retrieval, score aggregation, LLM response, and logical dependence updates.

quent retrieval, forming a self-reinforcing cycle that improves response quality and regularity discovery without prohibitive preprocessing. The graph grows incrementally through ordinary research interactions, capturing cross-domain dependencies invisible to human readers or conventional LLM assistants.

We evaluate Q-Evolve on a domain-specific benchmark comprising 330 questions in superconducting quantum computing. Under blinded evaluation, Q-Evolve consistently outperforms state-of-the-art general-purpose LLMs and the NVIDIA Ising quantum calibration model (NVIDIA, 2026), while domain experts verify the reliability of the logical dependence chains it constructs. The principal contributions of this work are:

- **A dual-pathway architecture with bidirectional self-evolution**, the implementation is open-sourced at `https://github.com`.
- An open knowledge base specialized for superconducting quantum computing and a benchmark comprising 330 domain-specific questions with standardized evaluation criteria.
- The resulting knowledge assistant for superconducting quantum computing is continuously accessible at `http://edaq.org.cn:3000`.

2 Related Work

2.1 Retrieval-Augmented Generation and Structured Indexing

Conventional RAG systems retrieve text chunks via semantic similarity. To improve retrieval quality, recent studies have explored iterative retrieval-generation cycles (Shao et al., 2023; Trivedi et al., 2023) and modular architectures that refine queries dynamically. Another line of work constructs structured external knowledge indices to support multi-

hop reasoning. Tree-structured approaches (Sarthi et al., 2024; Fatehkia et al., 2024; Tao et al., 2025) build hierarchical document representations for progressive abstraction. Graph-based indices (Edge et al., 2024; Guo et al., 2025; Xu et al., 2024; Li et al., 2025) organize documents into entity graphs with community-level summaries, enhancing global reasoning over long contexts. Hypergraph variants (Feng et al., 2026; Luo et al., 2025; Wang and Han, 2025) extend this paradigm by capturing higher-order relationships among entities. Lightweight alternatives such as EHRAG (Song et al., 2026) and LinearRAG (Zhuang et al., 2026) replace expensive LLM-driven extraction with NER-based indexing or semantic hyperedges, substantially reducing computational costs.

Yet these methods share a fundamental limitation: they retrieve semantically or structurally related text, not logical structure. In scientific domains, many important quantities are governed by interacting mechanisms rather than isolated terms or local co-occurrence. A retrieval system that scores relevance by similarity or connectivity therefore surfaces passages that are topically related but weak in explanatory value, making it poorly suited for questions that require reasoning about how factors influence one another.

2.2 Causal Graphs and Memory Mechanisms for RAG

Beyond static structured indices, a separate line of research has explored **causal graphs** and **dynamic memory mechanisms** to enhance RAG reasoning.

For causal integration, CausalRAG (Wang et al., 2025) constructs causal graphs from the corpus to guide retrieval and improve context precision. On the memory side, structured mechanisms support multi-step RAG through contextual memory (Chen et al., 2023; Lee et al., 2024; Gutierrez et al., 2024; Li et al., 2024) and parametric memory (Qian et al., 2025). Hypergraph-based approaches such as HGMEM (Zhou et al., 2026) capture higher-order correlations through dynamic updates. Yet these methods share two critical limitations. First, they lack mechanisms for incremental knowledge integration: as new literature emerges, existing memory structures cannot absorb fresh insights without costly re-indexing. Second, they require exhaustive LLM-driven preprocessing across entire document corpora, regardless of whether users will ever query those relationships. We term this the preprocessing paradox: substantial computational investment in relationships that may never serve an actual query.

Thus, no existing approach achieves three goals simultaneously: deep cross-document causal-semantic reasoning, continuous knowledge accumulation without reconstruction, and practical scalability without exhaustive preprocessing. This gap motivates Q-Evolve.

3 Methods

Q-Evolve is a retrieval-augmented generation system for superconducting quantum computing. It constructs and continuously refines a logical dependence graph through LLM-driven extraction from each interaction, eliminating manual specification of dependence relationships. The architecture comprises three components: (1) a **dynamic data acquisition module** that ingests new publications monthly from arXiv; (2) a **dual-line parallel retrieval engine** that combines semantic search and dependence-guided graph traversal via score-biased merging; and (3) a **self-improving logical dependence graph** that updates its edges from each inquiry, closing the feedback loop from retrieval to knowledge enrichment.

3.1 Dataset and Dynamic Data Acquisition

We curated a corpus of 5,194 superconducting quantum computing papers from arXiv to date, covering preprints submitted through April 2026. Topics span qubit coherence mechanisms, noise characterization, gate fidelity optimization, qubit–resonator coupling, and material loss analysis.

An **incremental update mechanism** automatically queries the arXiv API on the first of each calendar month for relevant papers (categories `quant-ph`, `cond-mat.supr-con`, `cond-mat.mes-hall`) published in the preceding month. Newly discovered papers are downloaded, parsed, and appended to the corpus without manual intervention. Details of the corpus composition and crawling pipeline appear in Appendix A.

3.2 Dual-Line Parallel Retrieval with Score-Biased Merging

Q-Evolve employs a dual-line retrieval architecture that executes semantic retrieval and dependence-guided retrieval in parallel, merging results through score-biased ranking. Semantic search surfaces lexically relevant content not yet encoded in the logical dependence graph, whereas dependence-guided retrieval employs a bidirectional depth-first

search strategy to retrieve structurally grounded evidence aligned with the graph topology. The two branches are assigned asymmetric scores and merged into a unified ranked list with duplicate removal by score maximization. The result is truncated to the context budget and passed to the LLM. A lightweight fallback ensures graceful degradation under branch-level failures. Implementation details of both pipelines and the merge procedure appear in Appendix B.

3.3 Self-Improving Logical Dependence Graph

Q-Evolve constructs and continuously enriches a logical dependence graph through each question-answering interaction, forming a feedback loop in which every query improves subsequent retrieval. The graph supports two initialization modes. In cold-start mode, it contains domain-specific physical quantities as nodes but no edges (enabling asynchronous data annotation), bootstrapping dependence structures entirely from interaction-driven extraction. In warm-start mode, expert-validated relations from established physics are added as seed edges. Both modes are iteratively refined through LLM-driven extraction.

Formalization We model the logical dependence structure as a directed graph $G = (V, E)$. Each node $v \in V$ corresponds to a normalized physical quantity, such as `T1`, `Gamma_1`, or `omega_q`. Every node is associated with a small alias set $A(v)$ and a short textual description Ω_v.

Each edge $e \in E$ captures a directed logical dependence relation between two variables, formalized as $e = (u, v, s, c, \phi, \psi)$. The polarity $s \in \{\text{positive}, \text{negative}\}$ indicates whether the effect is reinforcing or inhibitory, and $c \in [0, 1]$ measures the confidence of the extracted relation. When available, ϕ records the underlying physical expression and ψ points to the supporting textual evidence.

The initial graph G_0 supports two modes. In a *cold-start* setting, G_0 is empty. In a *warm-start* setting, G_0 is additionally seeded with expert-validated relations from well-established physics, such as the Purcell relation connecting T_1 to resonator decay. These seed edges are assigned $c = 1.0$ and marked as `expert_knowledge`. In either case, G_0 is a scaffold, not a closed-world knowledge base.

Dynamic updating After each Q&A round, an asynchronous pipeline enriches the graph. Denote the graph at time t as $G^{(t)} = (V^{(t)}, E^{(t)})$ and the updated graph as $G^{(t+1)}$.

Stage 1: Logical dependence relation extraction. The LLM receives a structured prompt (detailed in Appendix C) containing the user's question, retrieved knowledge chunks, and the model's reasoning output. It extracts explicit or strongly implied logical dependence relationships as valid JSON. Each relation specifies `source`, `target`, `sentiment` (`positive` or `negative`), confidence $c \in [0.5, 1.0]$, an `evidence_quote`, and an optional `equation`. The prompt enforces ≤ 10 relations per extraction and requires $c \geq 0.5$. A two-retry mechanism handles JSON parse failures, with a regex-based fallback for malformed outputs. The candidate edge set is E_{cand}.

Stage 2: Entity alignment and conflict detection. LLM-extracted variable names are often non-canonical (e.g., "qubit lifetime" and "decay time" may both refer to `T1`). The `EntityResolver` applies a four-stage alignment procedure: (i) exact node-ID match; (ii) case-insensitive alias match against $\bigcup_{v \in V} A(v)$; (iii) fuzzy matching using character-level Jaccard similarity

$$\text{sim}(n, a) = \frac{|n \cap a|_{\text{char}}}{\min(|n|, |a|)}, \tag{1}$$

where a mention is aligned if $\max_{a \in A(v)} \text{sim}(n, a) \geq 0.8$. Mentions that fail all three stages enter a candidate evaluation pipeline rather than being discarded outright.

Unmatched mentions are evaluated by the `CandidateNodeScorer`. For each candidate mention m, a composite score is computed as

$$\begin{aligned} S(m) = 0.20\, S_{\text{type}} + 0.35\, S_{\text{ctx}} \\ + 0.20\, S_{\text{freq}} + 0.25\, S_{\text{struct}}. \end{aligned} \tag{2}$$

where S_{type} assesses whether m resembles a physical quantity, S_{ctx} checks for definitional or formulaic support in the source chunk, S_{freq} counts cross-session occurrences of m, and S_{struct} evaluates whether m forms a physically reasonable connection with existing nodes.

$$\text{disp}(m) = \begin{cases} \text{ACCEPT}, & S(m) \geq 0.70, \\ \text{PENDING}, & 0.40 \leq S(m) < 0.70, \\ \text{REJECT}, & S(m) < 0.40. \end{cases} \tag{3}$$

Accepted candidates are added as new nodes to V; pending candidates are held for future accumulation; rejected candidates are discarded. Edges involving unresolved mentions are excluded from the current update cycle. The resolved edge set is E_{res}.

Each $e \in E_{\text{res}}$ is checked against $E^{(t)}$ for conflicts. A conflict is declared if an edge already exists for the same ordered pair (u, v) with opposite sentiment, i.e., $u = u'$, $v = v'$, and $s \neq s'$. Edges with reverse direction but identical sentiment are not treated as conflicts, as physical dependencies need not be symmetric. Conflicting edges are logged to a conflict journal for later inspection. The non-conflicting set is

$$E_{\text{ok}} = \{\, e \in E_{\text{res}} \mid \neg\text{conflict}(e, E^{(t)}) \,\}.$$

Stage 3: Weighted graph fusion. Non-conflicting relations are merged into the graph using exponential moving average (EMA) updates. For each $e \in E_{\text{ok}}$:

- If $e \notin E^{(t)}$, it is added to the graph with initial confidence $c_{\text{graph}}(e) = c_{\text{LLM}}(e)$.
- If $e \in E^{(t)}$, the confidence is updated with learning rate $\gamma = 0.3$:

$$c_{\text{graph}}^{(t+1)}(e) = \gamma \cdot c_{\text{LLM}}(e) + (1-\gamma) \cdot c_{\text{graph}}^{(t)}(e), \tag{4}$$

  the occurrence count $n(e)$ is incremented, and the current evidence quote is appended to the edge attributes.

Relations with $c_{\text{LLM}}(e) < 0.7$ are discarded at the extraction stage and do not enter the fusion process.

After processing, the updated edge set is

$$E^{(t+1)} = \left(E^{(t)} \setminus E_{\text{replace}}\right) \cup E'_{\text{ok}}, \tag{5}$$

where E_{replace} contains edges whose confidence values changed, and E'_{ok} carries the updated confidence and weight values.

Stage 4: Persistent update The final stage closes the feedback loop: the updated graph $G^{(t+1)}$ is committed with versioned backup and in-place replacement, ensuring crash-safe persistence. Details of the commit mechanism, batching strategy, and evidence chains appear in Appendix D.

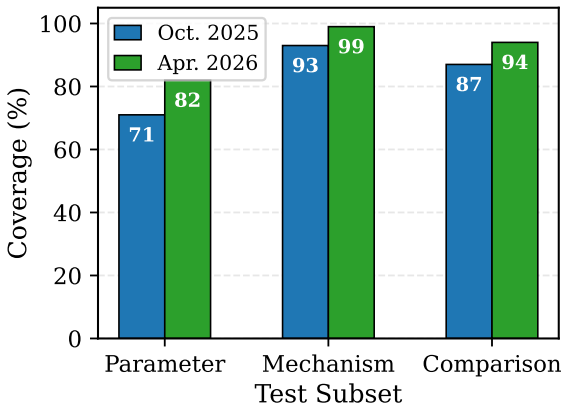

Figure 2: Coverage comparison of two knowledge bases with different cutoff dates on the Test Subset.

4 Experimental Settings

4.1 Datasets and Baselines

To evaluate the knowledge assistant across different cognitive demands, we constructed a benchmark of 330 questions in superconducting quantum computing. The questions are organized into four subsets that progress from simple fact retrieval to logical reasoning and experimental judgement. The parameter-oriented subset (100 questions) requests specific numerical values or material properties. The mechanism-oriented subset (100 questions) focuses on physical causes and decoherence pathways, requiring explanatory logical chains. The comparison-oriented subset (100 questions) asks the assistant to weigh trade-offs among architectures, materials, or fabrication processes. The calibration-oriented subset (30 questions) simulates real experimental tasks. Each task is accompanied by synthetic or measured data and labeled by difficulty. Appendix E provides detailed statistics and examples.

Our assistant is implemented on Qwen-3.5-Plus. Using the first three subsets (300 questions), we compared it against two baselines that share the same backbone: the bare LLM in a zero-shot setting (Model A) and the same model augmented with real-time web search (Model B). These baselines represent a plain LLM and a standard, non-curated retrieval practice, respectively. For the calibration subset (30 questions), we compared against the NVIDIA Ising quantum calibration model, which is built on Qwen-3.5-35B-A3B. To ensure a fair comparison, we switched our assistant's backbone to Qwen-3.5-35B-A3B for this evaluation.

Table 1: Per-dimension comparison across domain-specific and calibration subsets.

Subset Type	Model	Phys. Logic	Prob. Relevance	Arg. Chain	Lit. Evidence	Boundary Cond.	Mag. & Units	Consensus /Controv.	Meta. Halluc.
	Qwen-3.5-plus								
Domain-specific	Base	6.54	6.94	5.36	1.45	4.26	6.85	4.86	5.05
	Base+websearch	6.91	7.36	6.31	5.49	5.26	7.13	5.51	5.63
	Q-Evolve	**8.41**	**8.41**	**8.37**	**7.79**	**7.99**	**8.23**	**7.78**	**7.94**
	Qwen-3.5-35B-A3B								
Calibration	Ising-Calib.	7.34	8.24	7.00	5.22	6.38	8.62	6.41	6.64
	Q-Evolve	**8.96**	**8.95**	**8.95**	**8.45**	**8.38**	**9.73**	**7.95**	**7.81**

4.2 Metrics and Implementation Details

To quantitatively assess response quality, we designed a multi-dimensional Scientific Credibility Metric (SCM) tailored to superconducting quantum computing. The metric evaluates each answer across eight dimensions: physical logic consistency, question relevance, argument chain completeness, literature evidence quality, boundary condition specification, dimensional and unit self-consistency, distinction between consensus, controversy, and hypothesis, and metacognitive hallucination control. Responses were primarily evaluated by LLM-based judges under blinded conditions. A subset of questions, including the logical dependence chain assessments(Babakov et al., 2025), were evaluated by domain experts who applied the SCM rubrics while also verifying answer correctness against ground truth. The weighting scheme, detailed rubrics, and inter-rater agreement statistics are provided in Appendix F.

The system retrieves 10-15 text chunks per query. A logical-dependence-intent detector first analyses the user question. When a logical dependence intent is identified (such as "why" or "how does ... affect"), the retrieved set consists of 60%–80% logical-dependence-retrieved chunks and 20%–40% semantic-retrieved chunks. For queries without logical dependence intent, the ratio is reversed. This hybrid strategy keeps logical dependence evidence dominant for mechanistic questions while still allowing semantic exploration. The backbone language model (Qwen-3.5-Plus for main experiments, Qwen-3.5-35B-A3B for calibration tasks) was configured with temperature 0.6, top-p 0.8, and maximum output tokens 4,096. No explicit reasoning-chain prompting was enabled.

5 Results and Analysis

5.1 Baseline Comparison

We evaluated the assistant using a two-tier blinded assessment. First, three LLM judges (Claude-Sonnet-4-6, GLM-5.1, and Qwen-3.6-Plus) independently scored all responses under blinded conditions. Second, three domain experts in superconducting quantum computing independently evaluated a subset of 10 questions within their respective areas of expertise, applying the SCM rubrics described in Section 4.1 and verifying answer correctness against ground truth. To improve response quality, the logical-dependence graph was warm-started with expert-validated prior knowledge.

The LLM-judged results are summarized in Table 1. On the domain-specific subset, the assistant achieved a mean score of 8.78, compared with 5.16 for the base model and 5.97 for the web-search-augmented variant. On the calibration subset, it attained 9.39 versus 7.27 for the Ising baseline. The largest gains appeared in literature evidence and boundary conditions, where domain-specific knowledge is most critical. In the calibration setting, the Ising baseline triggered six safety vetoes (errors in anharmonicity definition, circuit-QED interpretation, and unit consistency), none of which occurred in the assistant's outputs. The expert-evaluated subset yielded a mean score of 8.5, confirming the reliability of the logical-dependence chains and the physical plausibility of the responses. Overall, these results demonstrate that structured knowledge augmentation significantly narrows the gap between general-purpose language models and expert-level responses in superconducting quantum computing.

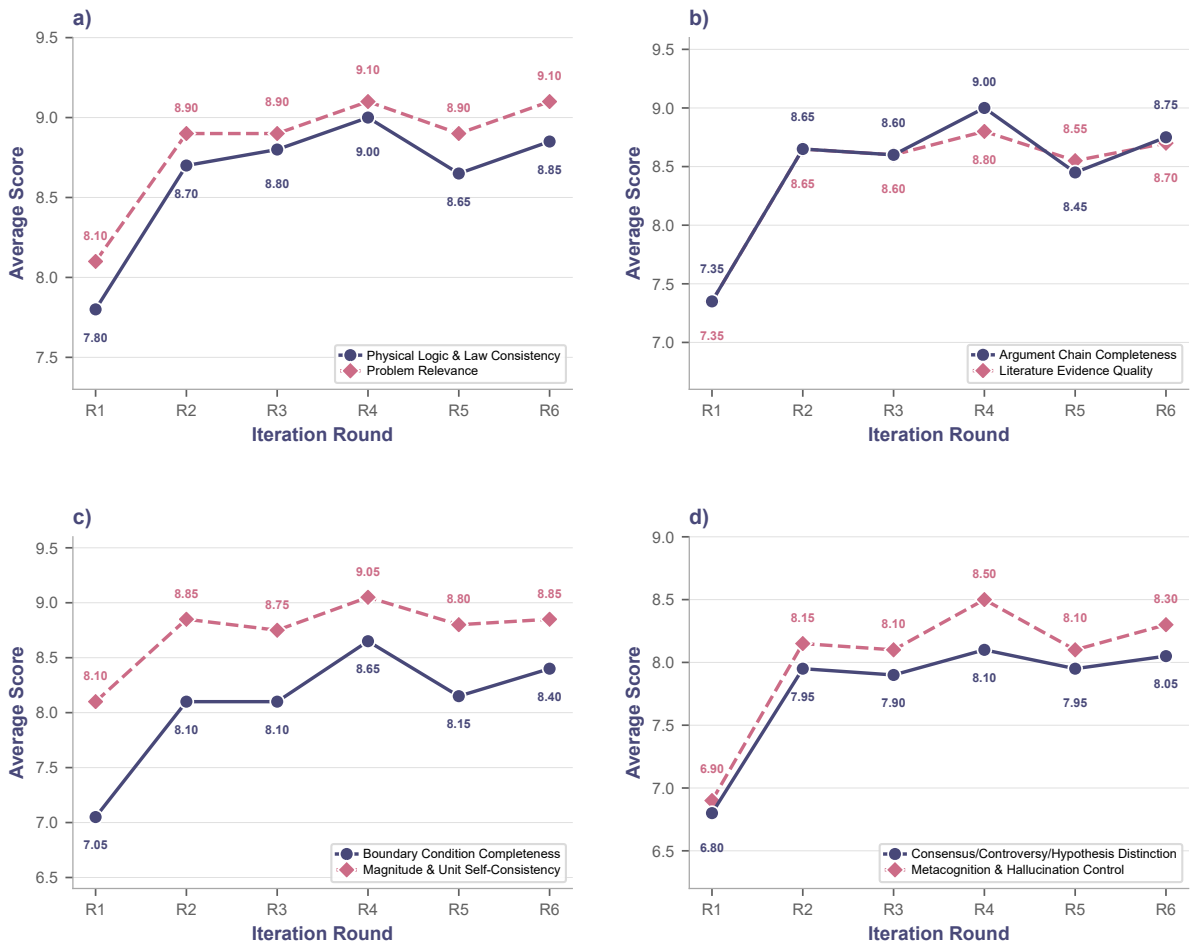

Figure 3: Score comparison of each turn's output across eight dimensions in a six-turn dialogue.

5.2 Dynamic Knowledge Update and Coverage Expansion

To evaluate whether temporal extension of the knowledge base expands system coverage, we constructed two corpus versions: one with a knowledge cut-off at October 2025 and another updated through April 2026. From the full 300-question benchmark, we identified 49 questions that the October-2025 baseline could not answer (29 parameter-oriented, 7 mechanism-oriented, and 13 comparison-oriented) and re-evaluated them using the April-2026 knowledge base.

Coverage gains varied across the three question categories (Fig. 2), with improvements of 11%, 6%, and 7% for parameter-oriented, mechanism-oriented, and comparison-oriented questions, respectively. Only one response triggered a safety veto due to a hallucinated critical temperature value. In the remaining unresolved cases, the system appropriately declined to generate unsupported answers and instead indicated the knowledge boundary. The notably stronger recovery on mechanism-oriented questions suggests that literature injection is most effective at filling causal reasoning gaps, whereas many parameter-oriented questions remain constrained by the specificity of the numerical values sought. Overall, the substantial conversion of previously unanswerable queries into well-supported responses underscores the value of temporally extended knowledge curation in sustaining the system's capacity to address emerging scientific questions.

5.3 Logical Dependence Graph Construction

To examine whether the logical dependence graph could be enriched through dialogue, we analyzed graph growth following the question-answering session in Section 5.1. Starting from a warm-start configuration of approximately 28 nodes and 35 expert-validated edges, the graph expanded to 50 nodes and 74 edges, a gain of roughly 19 nodes and 39 edges. Among these new edges, 36 achieved high confidence (≥ 0.90) while 38 fell in the 0.70–0.90

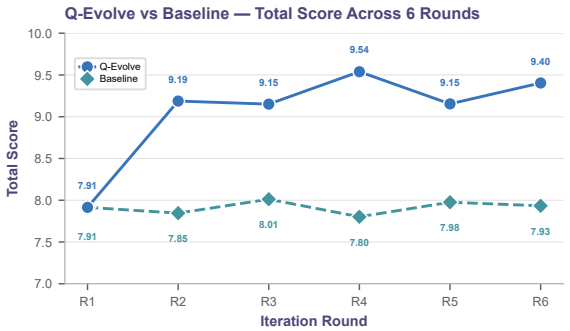

Figure 4: Performance comparison of Q-Evolve and the semantic retrieval framework in a six-turn dialogue.

range. Importantly, the graph was not static after initialization; it continued to self-enrich and update as the dialogue progressed, incorporating newly supported nodes and edges when warranted by the evidence. The final graph spans five domains: relaxation and decoherence, microscopic mechanisms of noise-induced relaxation, qubit–resonator coupling networks, gate-fidelity pathways, and material and fabrication loss.

To verify the reliability of graph evolution, three domain experts independently evaluated newly added nodes and edges across three sub-domains, following the evaluation protocol described in Section 4.2. Expert agreement varied systematically with extraction confidence: relations above 0.90 achieved 94.4% agreement with expert annotation, those between 0.70 and 0.90 reached 89.5%, and none fall in the 0.5–0.7 range. This confidence-dependent accuracy gradient suggests that the graph can accumulate and refine knowledge through interaction, supporting practical deployment. However, the declining agreement at lower confidence levels underscores the need for continued expert oversight: extractions below 0.80 should be flagged for review before incorporation, ensuring that self-evolution remains grounded in physically valid dependencies.

5.4 Iterative Self-Improvement and Ablation Analysis

To isolate the contribution of dynamic logical-dependence reasoning and assess cold-start bootstrapping, we ran two controlled comparisons: a full Q-Evolve pipeline vs. a semantic-only ablation across six rounds, and a cold-start iteration evaluation.

Contribution of dynamic logical-dependence retrieval. To disentangle the effect of logical-dependence reasoning from that of semantic retrieval alone, we disabled the dynamic logical-dependence chain and compared the semantic-only variant against the full Q-Evolve pipeline across six rounds (Figure 4). Both conditions began at an identical mean score of 7.91 at R1, confirming that the backbone model and retriever were held constant. From R2 onward, the full pipeline consistently outperformed the ablation, with the gap widening to 1.53 points at R4 (9.54 vs. 8.01). The semantic-only variant remained essentially flat (range: 7.80–8.01), indicating that repeated semantic retrieval without logical-dependence guidance does not yield cumulative benefit. These results establish that the dynamic logical-dependence chain, rather than increased semantic exposure or iterative prompting alone, drives the quality gains observed in Q-Evolve.

Cold-start iteration. Under cold-start conditions, we tracked eight evaluation dimensions across six rounds (R1–R6), as summarized in Figure 3. All eight dimensions increased sharply from R1 to R2, coinciding with the construction of an initial logical-dependence graph from the first round of interaction. The largest gains (+1.3 points) appeared in Argument Chain Completeness and Literature Evidence Quality, indicating that improvements in retrieved literature quality propagated through the logical-dependence structure and lifted the other metrics. Performance stabilized after R2, peaked at R4, and then oscillated within a narrow band. These results show that Q-Evolve's dynamic logical-dependence reasoning bootstraps a functional dependence structure within just two dialogue rounds, yielding response quality sufficient for practical deployment.

6 Conclusion

We introduced Q-Evolve, a self-evolving knowledge assistant for superconducting quantum computing that combines semantic retrieval with a lazily constructed logical dependence graph. Experiments show that Q-Evolve outperforms both general-purpose LLMs and domain-specific models such as the NVIDIA Ising calibration model. By balancing answer quality with deployment cost, avoid expensive offline preprocessing while staying current with evolving literature. It offers a practical approach that we believe can extend to other fast-moving scientific domains requiring structured knowledge synthesis.

Limitations

The current corpus draws exclusively from arXiv preprints, which limits coverage of peer-reviewed literature from high-impact venues. Incorporating published articles from journals such as *Nature Physics*, *Physical Review Letters*, and *npj Quantum Information* would improve both the quality and temporal stability of retrieved evidence. Additionally, the self-evolving logical dependence graph requires periodic expert annotation to prevent error accumulation from low-confidence LLM extractions. While this introduces a human-in-the-loop requirement, it also creates an opportunity: the graph may surface non-obvious cross-study dependencies that merit expert attention, turning verification into a productive discovery workflow rather than a purely corrective one.

References

Anthropic. 2026. Claude-sonnet-4-6.

Nikolay Babakov, Ehud Reiter, and Alberto Bugarín. 2025. CausalGraphBench: a benchmark for evaluating language models capabilities of causal graph discovery. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 240–258, Vienna, Austria. Association for Computational Linguistics.

Howard Chen, Ramakanth Pasunuru, Jason Weston, and Asli Celikyilmaz. 2023. Walking down the memory maze: Beyond context limit through interactive reading. *arXiv preprint arXiv:2310.05029*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph RAG approach to query-focused summarization. ArXiv:2404.16130 [cs].

Masoomali Fatehkia, Ji Kim Lucas, and Sanjay Chawla. 2024. T-rag: Lessons from the llm trenches. *arXiv preprint arXiv:2402.07483*.

Y. Feng, H. Hu, S. Ying, and et al. 2026. Hyperrag: combating llm hallucinations using hypergraph-driven retrieval-augmented generation. *Nature Communications*.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2025. LightRAG: Simple and fast retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10746–10761, Suzhou, China. Association for Computational Linguistics.

Bernal Jimenez Gutierrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. Hipporag: Neurobiologically inspired long-term memory for large language models. In *Proceedings of the Neural Information Processing Systems*.

Kuang-Huei Lee, Xinyun Chen, Hiroki Furuta, John F. Canny, and Ian Fischer. 2024. A human-inspired reading agent with gist memory of very long contexts. In *Proceedings of the International Conference on Machine Learning*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Rui Li, Zeyu Zhang, Xiaohe Bo, Zihang Tian, Xu Chen, Quanyu Dai, Zhenhua Dong, and Ruiming Tang. 2025. Cam: A constructivist view of agentic memory for llm-based reading comprehension. *CoRR*, abs/2510.05520.

Shilong Li, Yancheng He, Hangyu Guo, Xingyuan Bu, Ge Bai, Jie Liu, Jiaheng Liu, Xingwei Qu, Yangguang Li, Wanli Ouyang, Wenbo Su, and Bo Zheng. 2024. Graphreader: Building graph-based agent to enhance long-context abilities of large language models. In *Findings of the Association for Computational Linguistics: Empirical Methods in Natural Language Processing*, pages 12758–12786.

Haoran Luo, Haihong E, Guanting Chen, Yandan Zheng, Xiaobao Wu, Yikai Guo, Qika Lin, Yu Feng, Zemin Kuang, Meina Song, Yifan Zhu, and Anh Tuan Luu. 2025. Hypergraphrag: Retrieval-augmented generation via hypergraph-structured knowledge representation. 38:152206–152234.

NVIDIA. 2026. ising-calibration-1-35b-a3b.

OpenAI. 2026. Gpt-5.5.

Hongjin Qian, Zheng Liu, Peitian Zhang, Kelong Mao, Defu Lian, Zhicheng Dou, and Tiejun Huang. 2025. Memorag: Boosting long context processing with global memory-enhanced retrieval augmentation. In *Proceedings of the 2025 Web Conference*, pages 2366–2377.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. 2024. Raptor: Recursive abstractive processing for tree-organized retrieval. In *International Conference on Learning Representations*.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models

with iterative retrieval-generation synergy. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9248–9274. Association for Computational Linguistics.

Yifan Song, Xingjian Tao, Zhicheng Yang, Yihong Luo, and Jing Tang. 2026. Ehrag: Bridging semantic gaps in lightweight graphrag via hybrid hypergraph construction and retrieval. *Preprint*, arXiv:2604.17458.

Wenyu Tao, Xiaofen Xing, Yirong Chen, Linyi Huang, and Xiangmin Xu. 2025. Treerag: Unleashing the power of hierarchical storage for enhanced knowledge retrieval in long documents. In *Findings of the Association for Computational Linguistics*, pages 356–371.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 10014–10037. Association for Computational Linguistics.

Jingjin Wang and Jiawei Han. 2025. PropRAG: Guiding retrieval with beam search over proposition paths. pages 6212–6227.

Nengbo Wang, Xiaotian Han, Jagdip Singh, Jing Ma, and Vipin Chaudhary. 2025. CausalRAG: Integrating causal graphs into retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22680–22693, Vienna, Austria. Association for Computational Linguistics.

Liyan Xu, Jiangnan Li, Mo Yu, and Jie Zhou. 2024. Fine-grained modeling of narrative context: A coherence perspective via retrospective questions. In *Proceedings of the Association for Computational Linguistics*, pages 5822–5838.

Chulun Zhou, Chunkang Zhang, Guoxin Yu, Fandong Meng, Jie Zhou, Wai Lam, and Mo Yu. 2026. Improving multi-step rag with hypergraph-based memory for long-context complex relational modeling. *Preprint*, arXiv:2512.23959.

Luyao Zhuang, Shengyuan Chen, Yilin Xiao, Huachi Zhou, Yujing Zhang, Hao Chen, Qinggang Zhang, and Xiao Huang. 2026. LinearRAG: Linear Graph Retrieval Augmented Generation on Large-Scale Corpora. In *International Conference on Learning Representations*.

A ArXiv Crawling and Download Pipeline

Algorithm 1 ArXiv crawling and download pipeline

Require: CSV metadata files, optional date limit, JSON mapping file
Ensure: Downloaded PDFs and updated arXiv ID mapping

```
Locate the latest CSV file in the current directory
Load existing JSON mapping if available; otherwise initialize an empty mapping
Parse the user-specified date limit, if provided
for each paper record in the CSV file do
    Extract title, submission date, and PDF link
    Convert the PDF link into an arXiv time code
    if paper time code is earlier than the limit then
        Skip this paper
        continue
    end if
    Clean the title to form a valid filename
    Extract the arXiv ID from the PDF link
    Update the JSON mapping with filename → arXiv ID
    Add the paper to the download list
end for
Save the updated JSON mapping
for each paper in the download list do
    Download the PDF using wget
    Verify the file integrity
end for
Return the download summary
```

B Dual-Line Parallel Retrieval with Score-Biased Merging

B.1 Semantic Retrieval Line

The semantic retrieval line combines dense vector search and sparse lexical matching.

1. **Dense retrieval** is performed using FAISS approximate nearest-neighbor search over `text-embedding-v4` vectors.
2. **Sparse retrieval** is performed using BM25 over the chunk text.
3. The two signals are fused using reciprocal rank fusion (RRF):

$$\mathrm{RRF}(d) = \sum_{r \in R} \frac{1}{k + \mathrm{rank}_r(d)}, \qquad R = \{\mathrm{FAISS}, \mathrm{BM25}\}, \tag{6}$$

where $R = \{\mathrm{FAISS}, \mathrm{BM25}\}$ and $k = 60$ is the smoothing constant.

Retrieved chunks are subsequently passed through a quality filter to remove low-quality content such as very short fragments or reference-section artifacts. Before merging, an asynchronous annotation trigger is fired for chunks lacking `related_variables` metadata. The semantic branch assigns a base score of 0.3, with an additional rank-based bonus to preserve within-source ordering.

B.2 Logical Dependence Retrieval Line

The logical dependence retrieval line runs in parallel with semantic retrieval. It transforms a user query into a set of documents ranked by how many logical edges they co-mention, following the eight steps detailed below.

B.2.1 Intent Detection

The module `detect_logical_dependence_intent` computes a confidence score $\gamma \in [0, 1]$ based on keyword matching and variable alias matching:

$$\gamma = 0.4 \cdot s_{\text{kw}} + 0.6 \cdot s_{\text{var}},$$

where s_{kw} is the fraction of logical keywords (e.g., "why", "decrease", "affect") present in the query, and s_{var} is the fraction of query tokens that appear in a pre-built alias map $\mathcal{A}$ mapping surface forms to canonical variable names. If $\gamma > 0.35$, the query is flagged as logical and the set of matched variables $V_m \subset \mathcal{V}$ is recorded.

B.2.2 Graph Traversal

For each seed $v \in V_m$, the logical dependence graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is traversed both upstream (causes) and downstream (effects) up to depth $d_{\max} = 3$. Let $\text{Back}(v, d)$ be the set of nodes reachable from v by following edges in reverse direction (i.e., inputs to v), and $\text{Forw}(v, d)$ the set reachable by following edges forward (outputs from v). The edges collected for seed v are:

$$\mathcal{E}_v = \underbrace{\{e \in \mathcal{E} \mid \text{target}(e) \in \text{Back}(v, 3)\}}_{\text{upstream edges}} \cup \underbrace{\{e \in \mathcal{E} \mid \text{source}(e) \in \text{Forw}(v, 3)\}}_{\text{downstream edges}}.$$

Duplicate edges across seeds are removed. For example, the seed "T1" typically yields about 30 unique edges.

B.2.3 Edge–Document Co-occurrence via SQL

For each seed v, the system issues one SQL query per edge $e \in \mathcal{E}_v$. Each variable name is expanded to its alias set from $\mathcal{A}$ (e.g., "T1" $\rightarrow$ {"T1", "T_1"}). For an edge $e : A \rightarrow B$, the query condition is:

$$(\text{content LIKE ANY(aliases}(A))) \ \wedge \ (\text{content LIKE ANY(aliases}(B))).$$

A document receives one vote for each edge it satisfies. The count for document d with respect to seed v is:

$$\text{edge_count}_v(d) = \big|\{e \in \mathcal{E}_v \mid d \text{ contains both ends of } e\}\big|.$$

All queries are executed over a single SQLite connection; results are aggregated into $\text{chain_doc_scores}[v] = \{d \mapsto \text{edge_count}_v(d)\}$.

B.2.4 Intra-chain Ranking and Round-Robin Interleaving

For each seed v, documents are sorted descending by $\text{edge_count}_v(d)$, yielding ranked lists L_v. To balance coverage across different logical chains, we perform round-robin interleaving: starting with the seed having the largest $|\mathcal{E}_v|$, we iteratively take the top remaining document from each active chain, skipping already selected documents, until a total of $k = 5$ documents are collected. The resulting merged list D_{logical} retains each document's associated seed and edge count.

B.2.5 Document Fetching and Metadata Enrichment

For each $d \in D_{\text{logical}}$, the full record (content, metadata, images) is retrieved from the document store. Two additional fields are attached:

- `_logical_edge_count` $= \max_{v \in V_m} \text{edge_count}_v(d)$,
- `_logical_chains` $= \{v \in V_m \mid \text{edge_count}_v(d) > 0\}$.

A document that appears in multiple chains (e.g., d_{17} covering both T1 and temperature) is marked accordingly. Additionally, a `retrieved_logical_context` field is generated by scanning the content for variable mentions and returning the incident logical edges from the graph.

B.2.6 Quality Filtering

Documents are discarded if:

- content length < 100 characters,
- content consists predominantly of bibliographic references (e.g., starting with "[1] arXiv:"),
- the document's `related_variables` field does not intersect with the traversed chain set $C = \bigcup_{v \in V_m} \text{Forw}(v, 3) \cup \text{Back}(v, 3)$ (this acts as a pre-filter during retrieval, reapplied here for safety).

Surviving documents may be enriched with image captions if `has_image` is true.

B.2.7 Response Assembly

The logical dependence line returns a JSON object containing the logical flag, confidence γ, matched variables V_m, a debug list of formatted logical chains, and the filtered documents with all metadata.

B.3 Merging with Semantic Retrieval

The front-end merges the logical dependence results with the semantic (BM25 / dense) results. For each document d, the final score is:

$$s_{\text{final}}(d) = \max\big(s_{\text{semantic}}(d),\ s_{\text{logical_dep}}(d)\big),$$

with

$$s_{\text{logical_dep}}(d) = 0.80 + \gamma \cdot 0.20 + r(d),$$

where $r(d)$ is a rank bonus (e.g., $0.01 \times (k - \text{rank})$). This scheme ensures that documents with strong logical dependence evidence dominate the top positions, while purely semantic documents still appear lower. The merged list is truncated to 10 documents, formatted as `knowledgeContext`, and sent to the LLM.

B.4 Parallel Execution and Fault Tolerance

The two retrieval branches are executed in true parallelism. For a given query, let L_{sem} and $L_{\text{logical_dep}}$ denote the latencies of the semantic and logical dependence branches, respectively. The overall retrieval latency is

$$L_{\text{total}} = \max(L_{\text{sem}}, L_{\text{logical_dep}}) + L_{\text{merge}}, \tag{7}$$

where L_{merge} is the score-biased merging cost and is typically negligible.

Each branch includes an independent fallback handler that returns an empty result set upon failure, ensuring that the system degrades gracefully to single-branch retrieval rather than failing entirely.

C LLM Extraction Prompt

This appendix describes the prompt used for logical dependence relation extraction in superconducting quantum computing.

C.1 Logical dependence Relation Extraction Prompt

```
You are a Knowledge Graph Extractor specializing in Superconducting Quantum Computing.

Analyze the provided LLM reasoning and evidence chunks. Extract explicit logical dependence
↪  relationships
BETWEEN TWO DISTINCT PHYSICAL VARIABLES mentioned or strongly implied in the context.

## Critical: What is NOT a dependence relation
The following patterns are frequently misinterpreted — do NOT extract them:
- Intrinsic properties of a single variable (e.g., "TLS has a broad distribution of energies"
↪  describes a property of TLS, not a logical effect)
- Symmetry or statistical behavior descriptions (e.g., "singlet is symmetric -> bunching" is a
↪  property of the state, not a logical link between variables)
```

```
- Scaling relations where one variable expresses another without causation (e.g., "nu(t) \propto T"
↪  is just a thermal scaling)
- A performance metric described in terms of itself (e.g., "enhanced coherence" without naming the
↪  logical driver)

## Output Format
Output STRICTLY valid JSON matching the schema below. Do NOT hallucinate relations.
Only extract relations that are clearly stated or physically well-established.

{
  "relations": [
    {
      "source": "Variable A (canonical physics name)",
      "target": "Variable B (canonical physics name)",
      "sentiment": "positive or negative",
      "confidence": 0.0-1.0,
      "evidence_quote": "Exact text from chunks or reasoning supporting this relation",
      "equation": "Relevant formula if mentioned (optional)"
    }
  ]
}

## Rules
1. source and target MUST be DISTINCT variables. Do NOT output relations where source == target.
2. Use standard physics variable names (e.g., T1, T2, Gamma_1, omega_q, C_g, kappa_r, Q_F, g).
3. "positive" sentiment means: source increase -> target increase (direct correlation).
4. "negative" sentiment means: source increase -> target decrease (inverse correlation).
5. Confidence should reflect how strongly the evidence supports the relation:
   - 0.9-1.0: Explicitly stated with formula or direct evidence
   - 0.7-0.9: Clearly implied by context
   - 0.5-0.7: Plausible but indirect evidence
   - <0.5: Do NOT include (too uncertain)
6. Maximum 10 relations per extraction.
7. If you are unsure whether text describes a logical relation (X impacting Y) vs. a
↪  definition/conversion/property, do NOT extract it.

## Physical Reference Ranges (for sanity checking, not for fabrication)
When evaluating whether an extracted relation is physically plausible, use these reference values.
Do NOT include these values in the output unless the context explicitly states them.

- Fundamental constants: hbar = 1.0546e-34 J.s, k_B = 1.3806e-23 J/K, e = 1.6022e-19 C, Phi_0 = h/(2e)
↪  = 2.0678e-15 Wb.
- Qubit frequency omega_q/(2*pi) ~ 4-8 GHz (transmon); anharmonicity alpha ~ -200 to -300 MHz; E_J/E_C
↪  ~ 50-100.
- T1 ~ 10-500 us; must satisfy T2 <= 2*T1 (fundamental limit).
- T2* ~ 0.5-30 us; T2 (echo) ~ 2-80 us; pure dephasing: 1/T2* = 1/(2*T1) + 1/T_phi.
- Coupling g/(2*pi) ~ 50-300 MHz (readout), ~ 5-30 MHz (two-qubit gates).
- Resonator kappa_r/(2*pi) ~ 1-10 MHz; omega_r/(2*pi) ~ 5-9 GHz.
- C_g ~ 0.1-5 fF; C_q ~ 50-100 fF; C_r ~ 100-500 fF.
- Quality factor Q ~ 1e5-1e7; loss tangent tan_delta ~ 1e-6 to 1e-3.
- Superconducting gap Delta ~ 0.6-0.7 meV (alpha-Ta, T_c ~ 4.5 K), ~ 1.4-1.5 meV (Nb, T_c ~ 9.2 K).
- Operating temperature ~ 10-50 mK (dilution fridge).
- Josephson relation: I_c = (2e/hbar)*E_J; R_N ~ 1-10 kOhm.

## Context
Question: {question}
Retrieved Chunks: {chunks}
LLM Reasoning/Answer: {llm_output}

## Extracted Relations (JSON only, no other text):
```

D Persistent Graph Update

D.1 Persistent Update of the dependence Graph

The final stage of Q-Evolve commits newly resolved logical knowledge to persistent storage. After entity alignment and conflict filtering, the updated graph

$$G(t+1) = (V(t+1), E(t+1))$$

is serialized to a temporary file. To ensure crash safety and avoid partial writes, the system creates a versioned backup of the current graph file and then performs an atomic file replacement using `os.replace()`. This mechanism guarantees that the graph is either fully updated or left unchanged in the event of failure.

After persistence, internal indices are rebuilt, including the node index, edge index, and alias map, so that subsequent retrieval and resolution steps operate on a consistent in-memory representation.

To prevent hallucinated or low-quality extractions from corrupting the graph, an anti-hallucination circuit breaker rejects updates that contain more than 20 relations or whose average confidence is below 0.5.

D.2 Evidence traceability.

Each edge in the logical dependence graph retains a list of evidence quotes from source documents, including arXiv identifiers and page numbers. This design serves two purposes: first, it enables experts to verify the extracted relations; second, it allows the system to cite specific passages when generating responses, thereby reducing the risk of hallucinated claims. When multiple documents support the same relation, all evidence quotes are preserved, allowing the system to present consensus views with appropriate attribution.

E Superconducting Qubit Calibration and Benchmarking Questions

Due to space limitations, this appendix presents only the 30 calibration-related questions, while the remaining test questions are made publicly available with the released code.

1. I performed an experiment on a superconducting qubit: fixing the microwave drive frequency at the qubit transition frequency f_{01}, gradually changing the microwave pulse duration (from 0 to 200 ns), and measuring the excited state probability after each pulse. The resulting curve shows periodic oscillations with a frequency of about 25 MHz. What is this experiment? What is its main purpose?

2. I measured the relaxation process of the qubit from state $|1\rangle$ to $|0\rangle$: first applying a π pulse to excite the qubit to $|1\rangle$, then waiting for different durations before readout. The measured excited state probability decays exponentially with waiting time, and the fitted time constant is about 40 μs. What is this experiment? What is the significance of this parameter for quantum computing?

3. I performed an experiment between two qubits: initializing one qubit to $|1\rangle$ and the other to $|0\rangle$, applying a CR (cross-resonance) pulse, and measuring the population evolution on both qubits as a function of pulse duration. The results show oscillations on the target qubit, with small oscillations also appearing on the control qubit. What is this experiment? What is the physical mechanism of the CR pulse?

4. I performed an experiment: first preparing the qubit in a superposition state with a $\pi/2$ pulse, waiting for some time, then applying a second $\pi/2$ pulse followed by readout. By varying the waiting time, the readout probability shows oscillatory decay. The fitted oscillation frequency is 3.2 MHz, and the decay time constant is about 12 μs. What is this experiment? What is the difference between this decay time constant and T_1?

5. I performed a 2D sweep experiment: the horizontal axis is the qubit bias voltage (controlling frequency), the vertical axis is the microwave drive power, and the color represents the excited state probability. A V-shaped resonance line appears in the figure, but in the high-power region, a second parallel resonance line appears next to the V-shaped line. What is this experiment? What does the appearance of the second line indicate?

6. During a Rabi oscillation experiment, the measured curve shows no obvious periodic oscillations; the excited state probability fluctuates around 50%, without showing a complete oscillation from 0 to 1. What are the possible causes? What is the most likely cause?

7. In a Ramsey experiment, the expected interference fringe contrast should be close to 80–90%, but the actual measured contrast is only about 30%. The T_1 measurement result is 50 μs (normal), and the fitted T_2^* value is 5 μs (much shorter than the expected 15 μs). Analyze the possible causes.

8. During single-shot readout calibration, the IQ plane distributions for states $|0\rangle$ and $|1\rangle$ were measured and found to have significant overlap. The calculated assignment fidelity is 82% (target $> 95\%$). Meanwhile, the measured Q factor of the readout resonator is about 30% higher than the design value. Analyze possible hardware causes and improvement solutions.

9. In a T_1 measurement, the fitted relaxation curve shows a clear rapid drop within the first 100 ns (from 1.0 to ~ 0.8), followed by normal exponential decay (time constant $\sim 35\mu$s). Using a single exponential fit over the entire data range yields an R^2 of only 0.85. Is this result reliable? What are the possible causes of the rapid drop?

10. In a qubit spectroscopy experiment, the expected f_{01} resonance peak should be near 5.2 GHz, but no peak is seen at 5.2 GHz. Instead, weaker peaks are observed at 4.8 GHz and 5.6 GHz. The qubit design frequency is indeed ~ 5.2 GHz. What is this phenomenon?

11. My qubit calibration results are as follows: $f_{01} = 5.10$ GHz, $f_{02} = 10.15$ GHz, $T_1 = 45\mu$s, π pulse time = 20 ns. Please determine whether this qubit is suitable for standard single-qubit gate operations, and explain why.

12. Two-qubit gate calibration results: CZ gate fidelity (interleaved RB) = 96.5%, leakage = 1.2%, ZZ coupling measured value = 8 MHz. The target fidelity is $> 99\%$. Please analyze which parameters limit the fidelity and propose improvement directions.

13. A qubit chip has 4 qubits (Q0–Q3). Single-qubit calibration is complete with good results (T_1 in the range of 30–50 μs, single-gate fidelity $> 99.5\%$). However, during two-qubit gate operations, the CZ gate fidelity between Q0-Q1 is only 94%, while Q2-Q3 reaches 98.5%. The frequencies of Q0 and Q1 are 5.1 GHz and 5.3 GHz respectively, and the coupling strength $g/2\pi = 20$ MHz. Please analyze the possible reasons for the lower fidelity of Q0-Q1.

14. Readout calibration results show: the average IQ amplitude for state $|0\rangle$ is 0.45, for state $|1\rangle$ is -0.40, and the standard deviations of the two distributions are 0.12 and 0.15 respectively. The integration time is 500 ns. Calculate the assignment fidelity and determine whether optimization is needed.

15. I have a batch of new superconducting qubit chips. After preliminary calibration, it was found that: all qubits have T_1 in the range of 10–15 μs (far below the design expectation of 50+ μs). At the same time, each qubit shows unexpected loss peaks (dip in T_1 vs frequency) at different frequency positions. Please analyze what the problem is and how to systematically troubleshoot it.

16. In a Rabi oscillation experiment, fitting with a damped sine function yields: Rabi frequency = 15.3 MHz, decay rate = 2.1 MHz^{-1}, $R^2 = 0.98$. However, the residual plot shows residuals randomly distributed between -0.05 and +0.05, without obvious structure. Is this fitting result reliable? What is the π pulse time?

17. T_1 measurement data for a qubit is as follows: waiting time from 0 to 200 μs, step size 5 μs, 41 data points total. Using a single exponential $P(t) = A \times \exp(-t/T_1) + B$, the fit gives $T_1 = 28\mu$s, $R^2 = 0.92$. The residual plot shows: for the first 40 μs, residuals are systematically negative (-0.03 to -0.08), while afterwards they are randomly distributed. What is wrong with this fit?

18. Ramsey experiment data fitting results: $T_2^* = 8\mu$s, oscillation frequency = 0.5 MHz, amplitude = 0.72, $R^2 = 0.95$. However, looking closely at the data, the oscillation amplitude seems to show a slight "rephasing" phenomenon after about 15 μs. After fitting with a standard single exponential decay envelope, the residuals in the 15–25 μs interval show a positive-negative-positive oscillation. What does this indicate? What model should be used to refit the data?

19. In a qubit spectroscopy experiment, fitting the resonance peak with a Lorentzian function yields: center frequency = 5.1234 GHz, linewidth (FWHM) = 2.3 MHz, peak height = 0.85, $R^2 = 0.99$. However, the linewidth corresponds to a T_1 upper limit ($T_{1,\max} = 1/(\pi \times \text{FWHM})$) of about 138 ns, while the directly measured T_1 is 45 μs. Are these two results contradictory?

20. During a DRAG calibration sweep, the $|2\rangle$ state leakage probability was measured for different DRAG coefficients α. The data are as follows:

α	0.0	0.5	1.0	1.5	2.0	2.5	3.0
$P(\|2\rangle)$	3.8%	2.1%	0.9%	0.4%	0.5%	1.2%	2.3%

Fitting with a quadratic function $P(\alpha) = a(\alpha - \alpha_{\text{opt}})^2 + P_{\min}$ gives $\alpha_{\text{opt}} = 1.6$, $P_{\min} = 0.38\%$, $R^2 = 0.997$. However, the experimenter suspects that α_{opt} should be more precisely determined between 1.0 and 2.0. What should be done?

21. A Rabi oscillation experiment measures a Rabi frequency of 20 MHz. What are the π pulse and $\pi/2$ pulse times of the qubit? If one wants to perform an $X(\theta)$ rotation gate (rotation angle θ), how should the pulse time be set?

22. A Ramsey experiment measures $T_2^* = 15\mu$s, and a direct T_1 measurement gives $T_1 = 60\mu$s. Calculate the pure dephasing time T_ϕ, and analyze which decoherence mechanism dominates.

23. A transmon qubit has the following parameters:

    - Josephson energy $E_J/h = 20$ GHz
    - Charging energy $E_C/h = 200$ MHz
    - Qubit frequency $f_{01} = 5.0$ GHz

    Calculate: (1) anharmonicity α; (2) effective charge dispersion number n_{zpf}; (3) the sensitivity of this qubit to charge noise (relative to a Cooper pair box).

24. CZ gate benchmarking results: interleaved RB fidelity $F = 98.7\%$, gate operation time $t_{\text{gate}} = 40$ ns. Assuming the main error sources are T_1 relaxation ($T_1 = 30\mu$s) and dephasing ($T_2^* = 15\mu$s), estimate the errors contributed by each mechanism, and analyze where the remaining error comes from.

25. On a 5-qubit chain, Readout-Error Mitigation calibration is performed. The diagonal elements of the measured readout confusion matrix (assignment matrix) are:

    - Q0: 97.2%, Q1: 95.8%, Q2: 96.5%, Q3: 94.1%, Q4: 96.0%

    (a) Calculate the average single-qubit readout fidelity.
    (b) If readout crosstalk is ignored, what is the total fidelity of the 5-qubit joint readout (probability of reading all 5 qubits correctly at the same time)?
    (c) What impact does this total fidelity have on quantum error correction?

26. Single-shot readout calibration results for a qubit: fidelity = 96.5%, assignment error = 3.5%, readout_time = 400 ns, SPAM error $\approx 1.2\%$. For a superconducting transmon qubit, what level does this readout quality represent?

27. The following is a set of qubit calibration metrics. Determine whether this qubit is suitable for running a VQE (Variational Quantum Eigensolver) algorithm:

- $f_{01} = 4.95$ GHz
- $T_1 = 25\mu$s
- $T_2^* = 6\mu$s
- Single-qubit gate fidelity (RB) = 99.1%
- Two-qubit gate fidelity (RB) = 97.8%
- Readout fidelity = 95.0%
- Leakage rate (single gate) = 0.15%

28. The following metrics for a qubit have changed (compared to last week's baseline data):

Metric	Last Week (Baseline)	This Week
f_{01}	5.10 GHz	5.08 GHz
T_1	45 μs	12 μs
T_2^*	18 μs	5 μs
Readout fidelity	96.5%	94.0%
Single-gate fidelity	99.6%	98.2%

The system underwent a thermal cycle (from room temperature to 10 mK) during the past week. Determine the system status and analyze possible causes.

29. After a two-qubit system completes calibration, Bell state preparation and measurement are performed, yielding the following results:

- Target Bell state: $|\Phi^+\rangle = (|00\rangle + |11\rangle)/\sqrt{2}$
- Quantum state tomography fidelity: 89.2%
- Expectation values: $\langle XX\rangle = 0.72$, $\langle YY\rangle = -0.68$, $\langle ZZ\rangle = 0.75$

Determine the quality of this Bell state preparation and analyze the main sources of error.

30. After a 27-qubit chip completes calibration, the statistics are as follows:

- T_1 median: 35 μs, range: 15–80 μs (3 qubits $< 20\mu$s)
- Single-gate fidelity median: 99.5%, lowest: 98.1% (same qubit with lowest T_1)
- Two-qubit gate fidelity median: 98.2%, lowest: 94.5% (two pairs of connections involve those 3 low-T_1 qubits)
- Readout fidelity median: 96.0%, lowest: 91.0%
- All qubit frequencies are uniformly distributed, with no obvious frequency collisions

Please evaluate whether this chip can be used for medium-scale quantum algorithm experiments (such as 27-qubit QAOA or VQE), and provide recommendations.

F SCM Multidimensional Scientific Credibility Evaluation System

F.1 System Role Definition

You are a rigorous Superconducting Quantum Computing (SQC) physics expert and a top-tier journal reviewer (Nature Physics, PRL, APL, etc.). Based on the SCM multidimensional scientific credibility evaluation system, you will objectively, systematically, deeply, and logically compare and evaluate the answers (Answer A, Answer B, Answer C) provided by three AI models to a scientific question.

F.1.1 Core Evaluation Principles

- Prioritize "scientific trustworthiness" over "looking like an expert".
- Focus on comparing the three answers, rather than over-scrutinizing detailed data of a single answer.
- Use the same criteria for all three answers under the same question; avoid biases caused by length, terminology density, etc.

F.2 Hard Veto Checks

Before scoring the dimensions, the following veto items must be checked. Any trigger should mark the answer as "high risk" or directly deduct >20% of the total score.

F.2.1 Veto Item List

Veto Item	Definition	Judgment Criterion
Basic physics error	Violation of basic laws (quantum mechanics, thermodynamics, electromagnetism)	e.g., claiming "energy conservation can be violated" or "entanglement can transmit information faster than light"
Obvious hallucination	Fabrication of non-existent phenomena, experiments, or data	e.g., "IBM achieved 1000-qubit error-free computation in 2024"
Literature/experiment fabrication	Citing non-existent papers, authors, conferences, or experimental results	e.g., "Smith et al. Nature 2025" (future paper) or fabricated data
Major unit/magnitude error	Serious error in numerical value or unit that invalidates the conclusion	e.g., GHz as MHz, mK as K, ns as μs
Irrelevant answer	Answer unrelated to the user's question or severely off-topic	e.g., asking "How does TLS noise affect T_1?" but answering about history of quantum computing
Controversy presented as consensus	Presenting a debated issue as established fact	e.g., "TLS is the only source of noise in superconducting qubits" (still controversial)
Safety/ethics violation	Harmful, discriminatory, or unethical content	Directly disqualified

F.2.2 Handling Rules

- If any veto item is triggered, mark "High risk" in the final score.
- Basic physics error or obvious hallucination: directly deduct 30–50% of total score.
- Other veto items: deduct 15–25% of total score.

F.3 Main Scoring Dimensions (8 Dimensions)

Each dimension is scored out of 10. Weighted total score = $\sum$(dimension score $\times$ weight).

Score	Description
10	Rigorous physical logic, fully compliant with basic laws, with no gaps in the derivation; the mechanism is based on a sound physical picture (e.g., BCS theory or the TLS model).
8–9	Essentially correct and compliant with basic laws; at most 1–2 imprecise expressions that do not affect the core conclusion.
6–7	Generally acceptable, but with 3–5 vague statements or logical leaps, or insufficient depth in the mechanism explanation.
4–5	Contains obvious defects, such as logical confusion, leaps in logic, or inconsistency with some physical laws.
2–3	Contains severe errors, including multiple violations of basic laws or self-contradictions.
0–1	Completely violates physical laws or has chaotic logic.

F.3.1 Dimension 1: Physical Logic and Consistency with Laws (Weight 20%)

Definition: Whether the core argument and derivation strictly follow basic physical laws, and whether there are logical gaps, logical inversion, or self-contradictions.

Checklist:

- Logical inversion (effect treated as cause)?
- Self-contradictions?
- Interactions between variables physically logical?
- System evolution trend reasonable?
- Conceptual confusion (e.g., entanglement vs. coherence, Pauli error vs. coherent error)?

F.3.2 Dimension 2: Question Relevance (Weight 15%)

Definition: Whether the answer correctly understands the user's intent, directly responds to the core question, and avoids irrelevant elaboration.

Score	Description
10	Fully understands the intent and directly answers the core question; every sentence is targeted, with no irrelevant expansion.
8–9	Accurately understands the question and gives a direct answer; with at most 1–2 background sentences.
6–7	Basically understands the question and mainly answers the core issue; includes 3–5 irrelevant or only marginally relevant sentences.
4–5	Understands the question but not accurately; includes obvious irrelevant expansions or off-topic parts, though the main answer can still be found.
0–3	Severely off-topic or does not answer the question.

Checklist:

- Core question correctly understood?
- Distinction between "explaining a concept" and "giving a conclusion"?
- Avoided over-expansion?
- Directly responded to the user's specific query?

Score	Description
10	Complete and clear argument chain; premises are explicit, steps are fully developed without leaps, the conclusion is well supported by evidence, and the wording is appropriate.
8–9	Basically complete, with at most 1–2 points that could be clearer, but with no logical flaws.
6–7	Generally traceable, but contains 3–5 leaps or unclear statements, or uses "must" where "may" would be more appropriate.
4–5	Contains obvious defects, multiple leaps, or overly strong statements; the reasoning is hard to trace.
0–3	Severely incomplete, with many leaps or chaotic logic; the conclusion does not match the evidence, or the reasoning is completely chaotic/untraceable.

F.3.3 Dimension 3: Completeness of Argument Chain (Weight 15%)

Definition: Step-by-step derivation from premises to conclusion; whether there are leaps, whether speculation is disguised as proof, whether correlation is treated as causation.

Checklist:

- Premises explicitly stated?
- Reasoning steps complete (no leaps)?
- Correlation mistaken for causation?
- Speculation stated as proof?
- Overly strong wording ("must", "certainly", "proves") with insufficient evidence?

F.3.4 Dimension 4: Quality of Literature Evidence (Weight 18%)

Definition: Whether citations are relevant, authoritative, recent, and truly support the argument, not merely decorative.

Score	Description
10	≥ 5 relevant references, with $\geq 50\%$ from the last 3 years, including original experiments and high-quality reviews; each reference directly supports its point.
8–9	4–5 relevant references, with $\geq 40\%$ from the last 3 years; most references directly support the point, with at most 1 background citation.
6–7	3–4 references, with $\geq 30\%$ from the last 3 years; some references directly support the point, while 2–3 are background or indirect.
4–5	2–3 references, but relevance or recency is poor; or there are many citations but weak support.
0–3	1–2 references or irrelevant citations; or all citations are outdated (> 5 years); or there are no citations / fabricated citations.

Literature Quality Priority:

1. Original experimental papers (Nature, Science, PRL, APL, etc.)
2. High-quality reviews (Nature Reviews, Annual Review, etc.)
3. Conference papers (ISQC, Superconducting Qubits, etc.)
4. Preprints (arXiv, must be noted)

5. Textbook chapters
6. Press releases, blogs (reference only, not primary evidence)

Checklist:

- Citations directly relevant to the argument?
- Include recent (1–3 years) research?
- Include original experimental evidence?
- Citations decorative or supportive?
- Sources clearly marked?

F.3.5 Dimension 5: Completeness of Boundary Conditions (Weight 12%)

Definition: Whether the answer lists preconditions (substrate type, packaging, frequency range, etc.) and uses cautious academic wording instead of absolute statements.

Score	Description
10	Clearly lists ≥ 5 key constraints; uses phrases such as "under the condition of ..." or "limited by ..."; and clearly defines the scope.
8–9	Lists 3–4 key constraints; uses cautious wording; and the scope is basically clear.
6–7	Lists 2–3 constraints but the wording is unclear, or includes 1–2 overly strong expressions.
4–5	Lists 1–2 constraints or only vague statements, or contains multiple absolute statements.
0–3	Almost no constraints are mentioned, or there are many absolute statements; boundary conditions are totally ignored.

Checklist:

- Substrate type, material system specified?
- Temperature range, frequency range specified?
- Measurement conditions, packaging conditions specified?
- Cases where conclusion does not apply mentioned?
- Cautious wording: "likely", "tends to", "based on current evidence"?

F.3.6 Dimension 6: Magnitude and Unit Self-Consistency (Weight 10%)

Definition: Whether numerical values, units, and dimensions are correct; no magnitude errors; no confusion of similar units.

Score	Description
10	All values, units, and dimensions are completely correct, with no errors.
8–9	99% correct, with at most 1 minor inconsistency that does not affect the conclusion.
6–7	95% correct, with 2–3 inconsistencies or unclear expressions.
4–5	Contains 4–5 unit or magnitude errors, but the main conclusion is unaffected.
0–3	Multiple errors affecting the credibility of part of the conclusion, or numerous errors with obvious dimensional inconsistency.

Common Errors to Check:

- MHz vs GHz ($\times 1000$)
- ns vs μs ($\times 1000$)
- mK vs K ($\times 1000$)
- percentage point vs percent
- ppm vs ppb
- frequency units (Hz, GHz) vs time units (s, ns)

F.3.7 Dimension 7: Distinction of Consensus/Controversy/Hypothesis (Weight 7%)

Definition: Whether the answer clearly distinguishes established facts, ongoing hypotheses, and debated mechanistic explanations.

Score	Description
10	Clearly distinguishes "established facts", "mainstream hypotheses", "controversial views", and "still uncertain" points, and notes the strength of evidence for each.
8–9	Basically distinguishes these categories, with at most 1–2 unclear or imprecise markings.
6–7	Partial distinction, but contains 3–5 confusions or unclear markings.
4–5	Rarely distinguishes among these categories, and often presents hypotheses as facts or controversy as consensus.
0–3	Almost no distinction, with substantial confusion; or no distinction at all, presenting everything as absolute fact.

Checklist:

- "Established fact" vs "hypothesis" clearly marked?
- Different existing views mentioned?
- Evidence strength for a mechanism explained?
- Avoid presenting a single paper as field consensus?
- Note "this issue is still under investigation"?

F.3.8 Dimension 8: Metacognition and Hallucination Control (Weight 3%)

Definition: Whether terminology is precise, whether limitations are acknowledged, and whether unfounded speculation is avoided.

Score	Description
10	Rigorous terminology, no confusion; actively points out limitations and uncertainties; avoids baseless speculation.
8–9	Basically rigorous, with at most 1 imprecise expression; includes appropriate limitation statements.
6–7	Generally correct, but with 2–3 imprecise expressions; limitation statements are insufficient.
4–5	Contains 4–5 confusions or imprecisions; almost no limitation statements.
0–3	Multiple confusions, obvious baseless speculation, or severe confusion/hallucination.

Checklist:

- Confusion of entanglement vs. coherence?
- Confusion of Pauli error vs. coherent error?
- Confusion of T_1 vs. T_2?
- Confusion of detuning vs. coupling strength?
- Baseless speculation present?
- Explicitly state "beyond my knowledge" or "needs further verification"?

F.4 Bonus Items (max +20%)

After scoring the eight dimensions, check the following bonus items. Each can add up to 5%.

F.4.1 Bonus 1: Epistemic Honesty and High-Quality Refusal (+5%)

Trigger conditions:

- When a question involves unpublished data, exceeds knowledge boundaries, or lies in a scientific blind spot, explicitly states: "Due to insufficient research data or knowledge limitations, I cannot give a definitive conclusion."
- Actively points out: "This derivation is based on an early model and may not apply to the latest processes."
- Explicitly states: "This issue is still under debate in the community; I cannot give a final answer."

Scoring:

- 1 high-quality refusal: +2%
- 2–3 high-quality refusals or active limitation statements: +5%

F.4.2 Bonus 2: Scope of Applicability and Transferability (+5%)

Trigger conditions:

- Clearly states under which conditions the conclusion holds.
- Actively points out under which changed conditions it may not hold.
- Avoids extrapolating a local experiment into a universal law.

Scoring:

- 1–2 clear scope statements: +2%
- 3 or more or particularly insightful scope analysis: +5%

F.4.3 Bonus 3: Suggestions for Further Verification (+5%)

Trigger conditions:

- Proposes "To verify this conclusion, we need ..."
- Points out "Next steps to further confirm ..."
- Mentions "Related open questions include ..."

Scoring:

- 1–2 suggestions: +2%
- 3 or more or particularly insightful suggestions: +5%

F.4.4 Bonus 4: Balanced Presentation of Controversial Views (+5%)

Trigger conditions:

- For issues where the community disagrees, presents multiple representative views.
- Fairly presents supporting evidence and counterarguments for each view.
- Avoids one-sided bias.

Scoring:

- Presents 2 views in a balanced manner: +2%
- Presents 3 or more views with balanced depth: +5%

F.5 For human experts (Dimension 9) : Factual Correctness & Completeness (Weight 45%)

Definition: Whether the answer is both factually correct and sufficiently complete. Factual correctness means all statements, numerical values, formulas, and physical relationships are accurate and consistent with established knowledge in superconducting quantum computing. Completeness means the answer covers all essential aspects required by the question without omitting critical information.

Score Criteria

Score	Description
10	Completely correct and fully complete. All claims are accurate; every necessary sub-question, condition, exception, and step is included. No error or omission.
8–9	Almost correct and very complete. At most one minor factual inaccuracy (e.g., a slightly outdated parameter that does not affect core meaning) or one minor omission (e.g., a secondary calibration method not mentioned). Not both.
6–7	Moderately correct but with some omissions, or vice versa. Contains 2–3 factual errors or misses 2–3 relevant points, but the core answer remains usable. Errors and omissions do not severely mislead.
4–5	Significant flaws in both correctness and completeness. Multiple factual errors (≥ 4) and missing several important aspects (e.g., lists only one noise source when asked for multiple, and misstates its physical origin). Answer is misleading.
0–3	Almost or completely wrong, and/or extremely incomplete. Hardly any correct information related to the question; answer is essentially useless.

Checklist:

- Are all numerical values (coherence times, gate fidelities, anharmonicity, etc.) physically plausible?
- Are cause–effect relations (e.g., increased temperature $\rightarrow$ reduced T_1) correctly stated?
- Does the answer address every sub-question explicitly asked?
- Are necessary preconditions or operating regimes included (e.g., "in the dispersive regime", "for a weakly anharmonic transmon")?
- Are relevant exceptions or limitations mentioned (e.g., "this formula holds only when $\Omega \ll \Delta$")?
- For multi-step processes (e.g., a calibration routine), are all steps listed in logical order?

F.6 Comprehensive Score and Ranking

F.6.1 Total Score Calculation Formula

$$
\begin{aligned}
\text{Base Score} &= \sum(\text{Dimension}_i \times \text{Weight}_i) \\
&= \text{Dim1} \times 0.20 + \text{Dim2} \times 0.15 + \text{Dim3} \times 0.15 + \text{Dim4} \times 0.18 \\
&\quad + \text{Dim5} \times 0.12 + \text{Dim6} \times 0.10 + \text{Dim7} \times 0.07 + \text{Dim8} \times 0.03 \\
\text{Bonus} &= \text{Bonus1} + \text{Bonus2} + \text{Bonus3} + \text{Bonus4} \quad (\max\ + 20\%) \\
\text{Final Score} &= \min(\text{Base Score} + \text{Bonus},\ 10) \\
\text{Veto Deduction} &= \text{If triggered, deduct } 15\% - -50\% \text{ depending on severity} \\
\text{Final Score} &= \max(\text{Final Score} - \text{Veto Deduction},\ 0)
\end{aligned}
$$

F.6.2 Expert Scoring Total Formula

When expert scoring is used, the final score is computed by combining the score from the first subsection and the score of Dimension 9:

$$\text{Total Score} = 0.55 \times \text{Base Score} + 0.45 \times \text{Dim9}$$

F.6.3 Rating Standard

Score Range	Rating	Meaning
9.0–10.0	Excellent	Scientifically highly trustworthy; can be used as reference answer.
7.5–8.9	Good	Scientifically basically trustworthy; minor flaws.
6.0–7.4	Moderate	Scientifically acceptable but with obvious deficiencies.
4.0–5.9	Poor	Scientifically problematic; use with caution.
2.0–3.9	Very poor	Scientifically serious problems.
0–1.9	× Untrustworthy	Should not be used.

F.7 Evaluation Criteria for LLM-Extracted Logical dependence Relations (Simplified Version)

F.7.1 Evaluation Objective

This appendix presents a simplified evaluation framework for logical dependence relations, designed for human experts to assess whether dependence edges extracted by large language models (LLMs) from literature are valid.

The proposed criteria do not rely on a gold-standard graph. Instead, they are based entirely on the expert's domain knowledge, logical reasoning, and literature traceability. The framework applies to the following dependence types:

- Logical dependence
- Formula dependence
- Structural dependence
- Definitional dependence
- Constraint dependence
- Mediating dependence

F.7.2 Evaluation Principle

Experts need to answer only one question:

Can this edge be considered valid?

If not, the primary issue should be identified.

F.7.3 Three-Category Judgment Criteria

1. Valid The edge is reasonable based on domain knowledge, logical relations, or literature evidence, and can therefore be accepted.

Basis for judgment:

- The direction is reasonable;
- The semantics of the two nodes match;
- It is consistent with formulas, experiments, or literature descriptions;
- Even if it is not strictly logical, it still holds as a dependence relation.

Examples:

- $T_1 \to \Gamma_1$: if Γ_1 denotes the relaxation rate, this relation is generally valid.
- $l \to f_r$: physically reasonable, where l denotes length and f_r denotes resonant frequency.

2. Questionable The edge is not necessarily wrong, but the evidence is insufficient, the semantics are unclear, or more context is needed for judgment.

Common reasons:

- The direction may be reasonable, but the literature does not clearly state it;
- Node definitions are not clear enough;
- It is an indirect relation but was extracted as a direct edge;
- The expert cannot confirm it immediately.

Handling:

- Keep the edge for re-examination;
- Mark it as "needs manual confirmation" if necessary.

3. Invalid The edge is clearly unreasonable, or directly conflicts with domain knowledge, logic, or formulas.

Common reasons:

- Self-loop: $A \to A$
- Reversed direction
- Mismatched node semantics
- Clearly violates physical or mathematical common sense
- Conflicts with formulas or literature descriptions

Examples:

- $T_1 \to l$ (should be $l \to T_1$)
- $A \to A$
- Incompatible dimensional connection

F.7.4 Recommended Annotation Fields

For each edge, the following information should be recorded:

Field	Meaning
Edge ID	Unique identifier
Source node	Start node of the edge
Target node	End node of the edge
Edge type	Causal / Definitional / Correlational / Constraint / Mediating / Uncertain
Judgment result	Valid / Questionable / Invalid
Primary reason	Brief explanation for the judgment
Evidence source	Formula / Literature / Experiment / Domain experience
Remarks	Optional supplementary information

Expert Evaluation Procedure **Step 1: Prepare materials**

The expert receives:

- Node definition table;
- Edge list;
- Literature sources (if available);
- Edge type annotations (if available).

Step 2: Judge each edge

For each edge, the expert makes one of three judgments:

- Valid
- Questionable
- Invalid

Step 3: Record reasons for invalid edges

For each invalid edge, briefly note the reason, e.g.:

- Self-loop;
- Direction error;
- Semantic mismatch;
- Insufficient evidence;
- Violates common sense.

Step 4: Summarize results

Count:

- Number of valid edges;
- Number of questionable edges;
- Number of invalid edges.

F.7.5 Recommended Output Report Template

1. Overall Statistics

- Total edges: _____
- Valid: _____
- Questionable: _____
- Invalid: _____

Edge ID	Source $\rightarrow$ Target	Reason
e_001	$T_1 \rightarrow l$	Direction error
e_002	$A \rightarrow A$	Self-loop
e_003	$X \rightarrow Y$	Semantic mismatch

2. List of Invalid Edges

Edge ID	Source $\rightarrow$ Target	Reason
e_010	$Q \rightarrow T_1$	Insufficient evidence, needs review
e_011	$\Gamma_1 \rightarrow S$	Unclear semantic boundary

3. List of Questionable Edges

4. Overall Conclusion

☐ Can be used directly

☐ Needs manual review before use

☐ Needs re-extraction