Supplementary Information

Evidence-constrained railway fault question answering with adaptive graph-vector retrieval and conditional answerability control

Rongrong Fu[1], Qiyun Su[1], Yong Yang[2] & Qizheng Hu[3,*]

[1] Hebei Key Laboratory of Measurement Technology and Instrumentation, Yanshan University, Qinhuangdao 066004, China

[2] Center of National Railway Intelligent Transportation System Engineer and Technology, Beijing 100081, China

[3] Signal & Communication Research Institute, China Academy of Railway Sciences Corporation Limited, Beijing 100081, China

* Correspondence: Qizheng Hu (email: hqzcars@163.com).

Contents

Supplementary Tables S1–S14: knowledge representation, extraction instructions, retrieval parameters and routes, ablations, detailed results, evidence audits and timing.

Supplementary Figures S1–S3: graph topology, dictionary-anchored query analysis, and performance/source complementarity.

Supplementary table and figure numbers are independent of the numbering in the main manuscript.

Supplementary Table S1. Five-field representation used to construct the railway fault knowledge resources.

Field	Provenance	Role in the knowledge unit
Alarm type	Observed	Identifies the reported alarm category and provides the primary graph entry point.
Maintenance response	Observed	Records the field operation or maintenance recommendation associated with the alarm.
Trigger condition	Generated by the LLM and manually reviewed	Specifies the event or state under which the alarm is activated.
Direct cause	Generated by the LLM and manually reviewed	Provides an immediate diagnostic explanation rather than an estimated causal effect.
Potentially faulty component	Generated by the LLM and manually reviewed	Lists one or more components prioritized for inspection.

Supplementary Table S2. Main elements of the extraction instruction and an illustrative example.

Element	Content
Role	Act as an experienced railway field technician who identifies entities and relations accurately and selects the most logically consistent extraction.
Task	Extract all triples from the text. Each entity is represented by its name and type, and each relation by its value and type, in the form (entity, entity type; relation, relation type; entity, entity type).
Constraints	Extract explicit and context-supported relations as completely as possible and follow the prescribed triple format strictly.
Example input	Alarm: CCS interface alarm. Response: restart during a maintenance window. Trigger: failure of data synchronization between the signalling system and the dispatching centre. Direct cause: temporary interruption of the CCS communication link. Components: CCS interface and communication link.
Example output	(CCS interface alarm, Alarm; has response, Relation; restart during a maintenance window, Maintenance response); (restart during a maintenance window, Maintenance response; has trigger, Relation; data-synchronization failure, Trigger condition); (restart during a maintenance window, Maintenance response; suggests cause, Relation; temporary CCS-link interruption, Direct cause); (restart during a maintenance window, Maintenance response; suggests component, Relation; CCS interface/communication link, Component).

Supplementary Table S3. Fixed models, knowledge resources, and online retrieval limits.

Item	Setting	Role or scope
Extraction model	GLM-4-9B with LoRA	Extraction of typed entities and directed relations
Answer generator	GLM-4-9B with LoRA	Generation of Chinese answers from retrieved evidence
Embedding model	BAAI/bge-large-zh-v1.5	Normalized 1,024-dimensional embeddings
Graph resource	Neo4j	6,838 nodes; 12,849 relations; 2,345 retrieval units
Vector resource	FAISS	4,651 structured and rewritten records
Retrieval limits	Cypher top-k = 10; vector top-k = 10.	Keyword branch disabled; at most six final evidence records
Generation	Temperature = 0	Conversation history omitted in controlled evaluation

Supplementary Table S4. Trigger conditions, operations, and final evidence composition of the adaptive retrieval routes.

Retrieval route	Trigger condition	Retrieval operation and final evidence
Graph Verification	Verification syntax is detected, or a quoted alarm or early-warning term requires strict exact validation because it is not registered.	Query registered alarms exactly and serialize the returned graph chains. When no registered alarm is present, use the alarm or early-warning term parsed from the quoted claim as the exact lookup term. Pure vector evidence is prohibited, and the generation prompt limits the answer to requested fields. An unregistered term or absent direct relation leads to the sentinel or Guard-insufficiency path rather than vector fallback.
Graph Direct	A registered alarm or faulty component is identified in a non-scenario question.	Expand an alarm forward through response and diagnostic fields, or trace a component backward to responses and alarms. The final context contains only graph chains.
Graph-only Route	Preliminary graph evidence for a scenario reaches the sufficiency threshold.	Deduplicate and rerank the preliminary graph candidates without calling the vector index. Retain no more than six graph evidence chains.
Vector-guided Graph Route	Preliminary graph evidence for a scenario is below the sufficiency threshold.	Use FAISS to propose at most two registered alarms, validate them against the domain dictionary, and expand them exactly in Neo4j. Remove the source vector narratives before generation; only verified graph chains remain.
Hybrid Fallback	A non-verification, non-scenario question contains no registered alarm or component anchor.	Run domain-aware fuzzy graph retrieval and vector retrieval in parallel, normalize and fuse their evidence, and retain no more than six records, including at most two pure-vector cases marked as pending graph verification.

Supplementary Table S5. Adaptive graph-vector retrieval with evidence control.

Step	Operation
1	Normalize q; extract registered alarms, components, and requested fields with the domain dictionary; identify Verification, Scenario, Multi-element, and Fact cues.
2	Apply the route priority: Graph Verification or strict quoted-alarm validation, scenario sufficiency gate, Graph Direct, and Hybrid Fallback.
3	In Graph Verification, query registered alarms exactly; only when none is registered, use the alarm or early-warning term parsed from the quoted claim as the exact lookup term. Serialize returned graph chains; an unregistered term or absent direct relation reaches the sentinel or Guard-insufficiency path.
4	In Graph Direct, expand an alarm forward or retrieve a component backward according to the anchor type.
5	For a Scenario question, retrieve preliminary graph candidates, deduplicate them by alarm-response pair, and compute $G_s(q)$.
6	Select Graph-only Route when $G_s(q) \geq \tau_G$; otherwise discard preliminary candidates and search FAISS.
7	Retain the leading registered alarm and retain the second only when its similarity gap is no greater than 0.2.
8	Expand retained alarms by exact Cypher queries; remove the source vector narratives after graph evidence is obtained.
9	In Hybrid Fallback, retrieve fuzzy graph and vector candidates and map both sources to the common five-field evidence schema.
10	Deduplicate by normalized alarm-response key, merge source labels, and filter candidates violating exact anchor constraints.
11	Compute the query-mode-specific fusion score, including source and dictionary bonuses and verbosity and aggregation penalties.
12	Apply source quotas and field-level compression, retain at most six records, pass the evidence-sufficiency gate, then conditionally invoke the Guard or proceed to constrained generation.

Supplementary Table S6. Fixed online evaluation parameters.

Parameter	Value
Cypher top-k	10
Vector top-k	10
Keyword top-k	0
Maximum final context records	6
Minimum vector score	0.20
Adaptive routing	Enabled
Graph-sufficiency gate	Enabled
Graph-sufficiency threshold	0.25
Answerability Guard	Enabled with conditional invocation
Generation temperature	0

Supplementary Table S7. Module-level ablation conditions and the method component tested by each condition.

Condition	Main change relative to Full system	Tested method component
Full system	Entity Dictionary, Fusion Control, and conditionally invoked Answerability Guard enabled	Complete method
w/o Entity Dictionary	Disable dictionary parsing, typed anchors, and the unregistered-concept gate	Entity Dictionary
w/o Fusion Control	Preserve route and source-authority policies; keep graph candidates in original rank for graph-only policies, otherwise alternate graph/vector within-source ranks to the six-record limit; no Hybrid Fallback occurred	Fusion Control
w/o Answerability Guard	Disable the conditional answerability audit and its controlled fallback decisions while preserving retrieval and generation settings	Answerability Guard

Supplementary Table S8. Rewrite generation and semantic-consistency screening outcomes.

Statistic	Value	Statistic	Value
Source records	2,345	Usable rewrites	2,306 (98.34%)
Automatic passes	2,299 (99.70%)	Manual completions	7
Records without a rewrite	39	Indexed texts	4,651
Mean similarity	0.9367	Median similarity	0.9400
P25	0.9261	P75	0.9505
Minimum	0.8135	Maximum	0.9739
Acceptance threshold	0.80	Automatic-pair count	2,299

Supplementary Table S9. Entity and relation extraction performance on the held-out set.

Level	Precision	Recall	F1	TP / predicted / gold
Entity	99.04%	98.64%	98.84%	2,692 / 2,718 / 2,729
Relation	99.01%	98.61%	98.81%	2,691 / 2,718 / 2,729

Supplementary Table S10. Question-level paired effects of Full system relative to each named module ablation.

Method component	Set	Role	Delta	95% CI	W/T/L
Entity Dictionary	V1	Development/ test	+0.1289	[+0.0748, +0.1906]	25/20/5
Entity Dictionary	V2	Validation	+0.0495	[+0.0242, +0.0770]	26/17/7
Entity Dictionary	V3	Development/ test	+0.0286	[-0.0086, +0.0789]	22/17/11
Entity Dictionary	V4	Validation	+0.0155	[-0.0050, +0.0362]	20/16/14
Fusion Control	V1	Development/ test	-0.0019	[-0.0334, +0.0405]	8/28/14
Fusion Control	V2	Validation	+0.0013	[-0.0207, +0.0247]	16/17/17
Fusion Control	V3	Development/ test	+0.0472	[-0.0000, +0.1106]	19/13/18
Fusion Control	V4	Validation	+0.0057	[-0.0126, +0.0263]	17/16/17
Answerability Guard	V1	Development/ test	+0.0200	[+0.0000, +0.0600]	1/49/0
Answerability Guard	V2	Validation	+0.0400	[+0.0000, +0.1000]	2/48/0
Answerability Guard	V3	Development/ test	+0.0600	[+0.0000, +0.1400]	3/47/0
Answerability Guard	V4	Validation	+0.0800	[+0.0200, +0.1600]	4/46/0

Delta = Full system − ablated condition. Intervals are 95% percentile bootstrap intervals from 100,000 paired question-level resamples (seed 20260722; n = 50 per set). W/T/L = wins/ties/losses for the Full system.

Supplementary Table S11. Paired efficiency comparison of Full system and w/o Fusion Control.

Set	Latency Full/w/o Fusion (s)	Latency saving [95% CI]	Context Full/w/o Fusion (characters)	Context saving [95% CI]
V1	8.23 / 12.65	+4.43 [+1.88, +7.75]	948.4 / 1194.3	+245.9 [+177.0, +319.1]
V2	10.18 / 13.21	+3.03 [-0.35, +6.78]	1021.4 / 1224.2	+202.8 [+143.2, +264.4]
V3	7.71 / 8.71	+1.00 [+0.04, +2.25]	889.1 / 1118.3	+229.1 [+165.3, +295.9]
V4	9.73 / 12.51	+2.78 [-0.65, +6.83]	1013.0 / 1281.6	+268.6 [+188.1, +354.7]
Pooled	8.96 / 11.77	+2.81 [+1.36, +4.41]	968.0 / 1204.6	+236.6 [+201.8, +272.3]

Full/w/o Fusion lists the means for the two conditions. Savings are w/o Fusion Control − Full system. Brackets show 95% percentile bootstrap intervals based on 100,000 paired question-level resamples; n = 50 per set and n = 200 pooled.

Supplementary Table S12. Deterministic audit of serialized V1 graph evidence.

Metric	Handling-mediated	Flat alarm-star
Evidence bundles	258	80
Serialized field items	2,493	6,015
Response-cause candidates	485	19,557
Response-cause relation accuracy	52.99%	2.43%
Response-component candidates	1,282	42,853
Response-component relation accuracy	54.06%	3.95%
Complete-record candidates	5,371	741,008
Valid complete-record candidates	689	998
Complete-record candidate accuracy	12.83%	0.135%
Potential complete-record mismatch	87.17%	99.87%

Supplementary Table S13. Stage timing on 144 knowledge-dependent questions.

Metric	Handling-mediated	Flat alarm-star	Delta	95% CI	Relative change
Graph retrieval (s)	0.191	0.044	+0.147	[+0.123, +0.172]	+330.7%
LLM generation (s)	8.903	30.582	-21.679	[-26.018, -17.455]	-70.9%
End-to-end server (s)	9.101	30.633	-21.532	[-25.860, -17.262]	-70.3%
End-to-end P95 (s)	21.17	69.84	-48.67	NA	-69.7%
Context length (characters)	1,045	1,756	-711	[-917, -523]	-40.5%
Answer length (characters)	391	1,360	-969	[-1,166, -782]	-71.2%

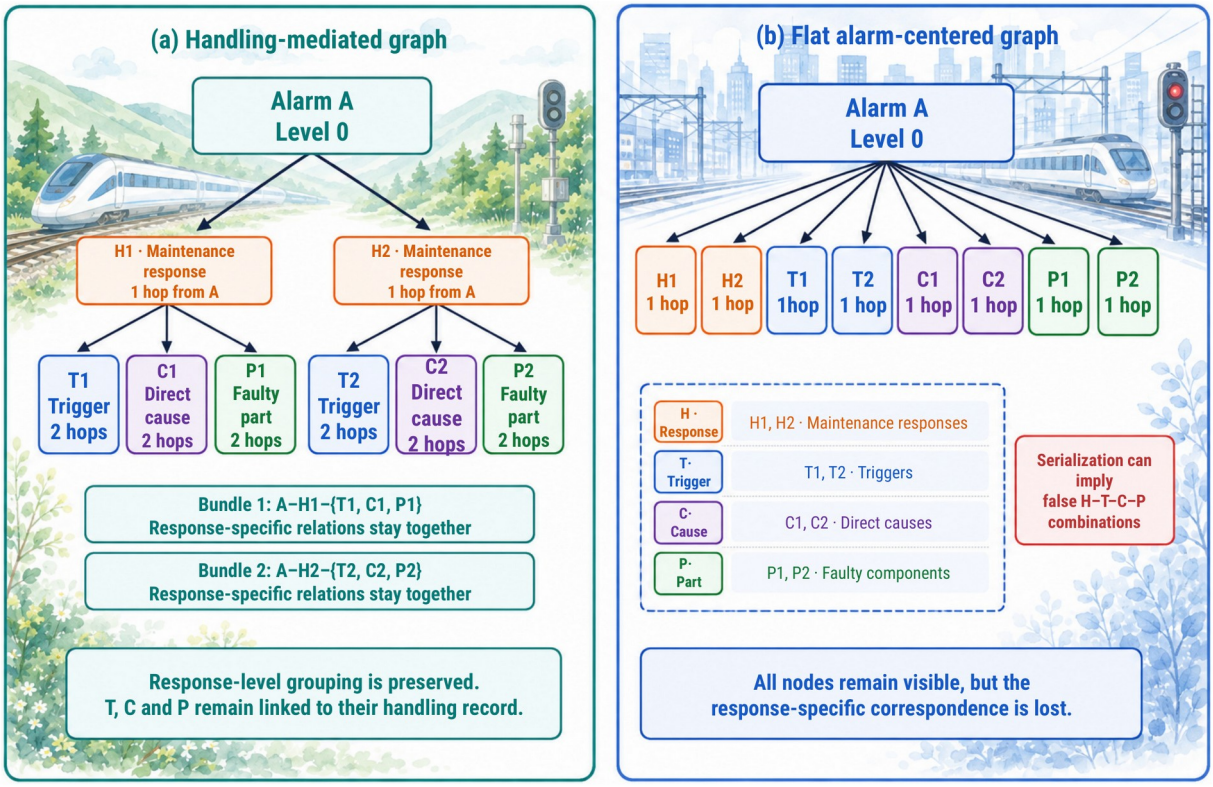
Values describe 144 knowledge-dependent questions. Delta = handling-mediated − flat alarm-star. Intervals are 95% percentile bootstrap intervals based on paired question-level differences. P95 denotes the 95th percentile; NA indicates that no interval was reported.

Supplementary Table S14. Macro-average category scores for the unoptimized source-only baselines.

Question category	Unoptimized Graph-only	Unoptimized Vector-only	Graph − Vector
Basic	0.504	0.564	-0.059
Multi-element	0.491	0.504	-0.014
Scenario	0.436	0.418	+0.018
Hallucination	0.469	0.625	-0.156
Refusal	0.208	0.292	-0.084

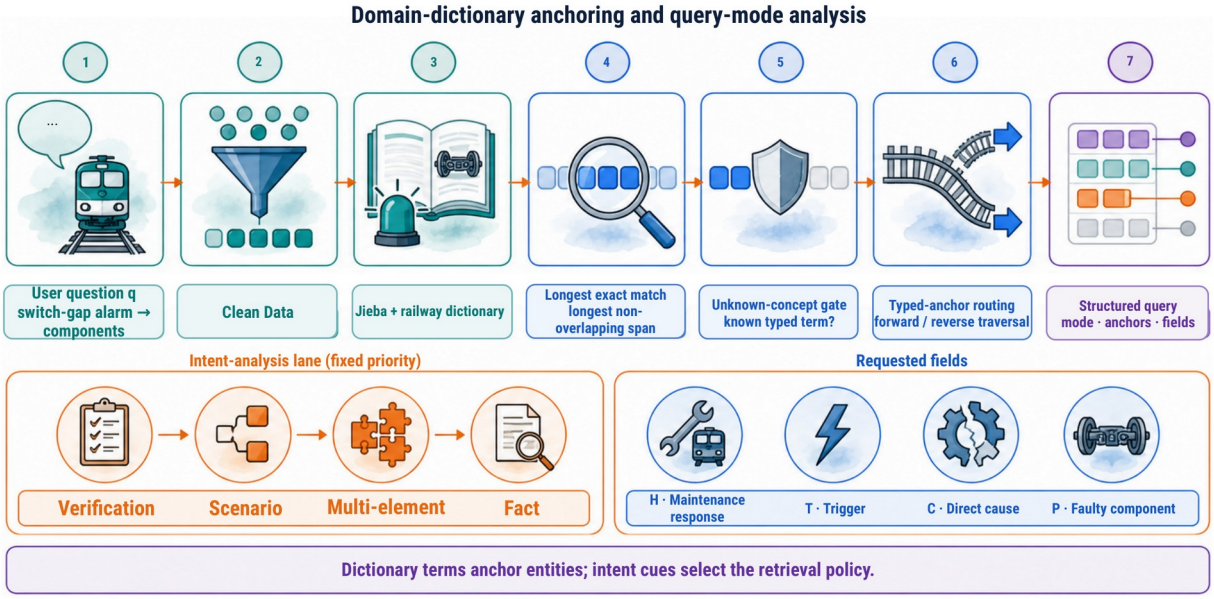
Values are category means averaged across V1–V4. Differences may not equal subtraction of the displayed rounded means.

Supplementary Figure S1



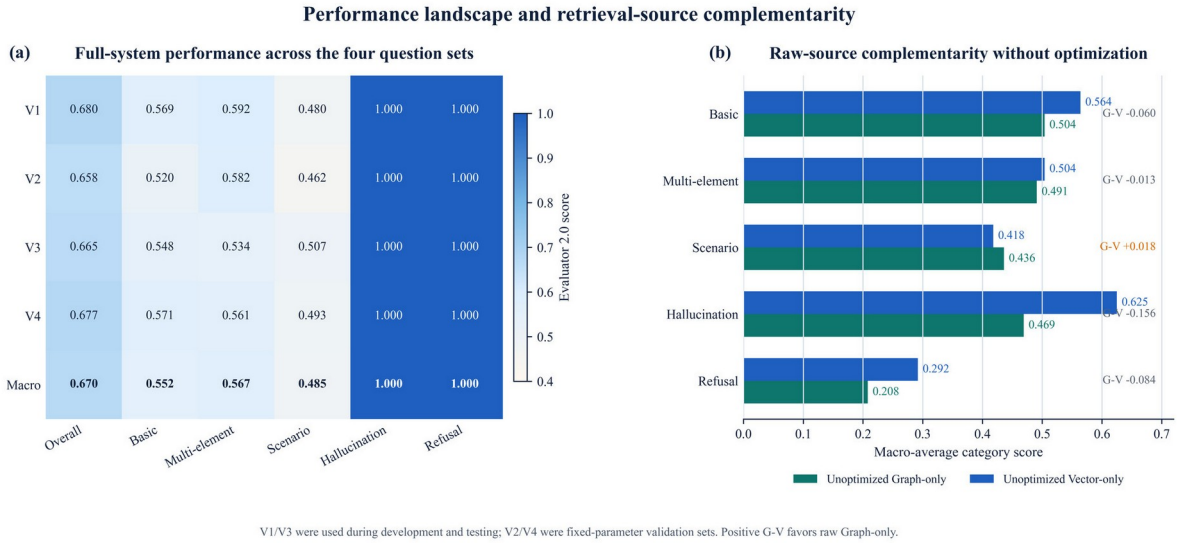
Supplementary Figure S1. Comparison of handling-mediated and flat alarm-centered graph topologies. The former preserves response-specific associations, whereas the latter loses local correspondence after serialization.

Supplementary Figure S2



Supplementary Figure S2. Domain-dictionary anchoring, longest-match entity recognition, typed graph traversal, and fixed-priority query-mode assignment. The upper sequence shows seven stages from question input to structured query construction; the lower panels show the fixed intent priority and the requested evidence fields.

Supplementary Figure S3



V1/V3 were used during development and testing; V2/V4 were fixed-parameter validation sets. Positive G-V favors raw Graph-only.

Supplementary Figure S3. Full-system performance across V1-V4 and category-level complementarity of the unoptimized Graph-only and Vector-only retrieval sources.