

Supplementary Material for EviGraphRAG: Constraint-Verified Retrieval-Augmented Question Answering over Event Knowledge Graphs

Yuteng Sun, Yang Su* and Xu An Wang

Contents

1	S1. Dataset subsets, task definitions, and split identifiers	1
2	S2. Natural-question construction and parser prompt	2
3	S3. Parser extraction results by field and subset	3
4	S4. Oracle/parser end-to-end results	3
5	S5. NLI, ScoreGate, and soft-verification details	3
6	S6. Full per-constraint and single-constraint ablations	4
7	S7. Constraint firing and overlap matrices	4
8	S8. Threshold, M, gamma, and cap sensitivity	5
9	S9. Hard-positive construction audits and results	5
10	S10. Predicted-cluster per-subset and purity-bin results	6
11	S11. Remaining attribute and false-support taxonomies	6
12	S12. Matched generation details	7
13	S13. Reproducibility commands, versions, prompts, and release scope	8

1 S1. Dataset subsets, task definitions, and split identifiers

The controlled benchmark contains 18,791 mention-level source records, 15,510 public-ID event nodes, 14,524 predicted event-cluster nodes, and 4,800 questions. It comprises six subset-year combinations: Afghanistan 2020, Afghanistan 2021, Iraq 2014, Iraq 2015, Syria 2016, and Syria 2017. Each subset contributes 200 questions to each of four tasks: event retrieval, evidence tracing, attribute QA, and insufficient-evidence detection.

Event retrieval evaluates the returned event identifier set; evidence tracing evaluates linked source-record identifiers; a ttribute QA evaluates whether the selected target event supports the requested structured attribute; insufficient-evidence questions require abstention when no candidate satisfies every represented constraint. The original structured robustness analysis used the saved seed-42 20%/80% validation/test split. Natural-question verifier settings used a separate deterministic MD5 (global_id) 40%/60% partition, giving 218 validation and 382 test questions. The two split conventions are not mixed.

Table S1: Controlled benchmark composition.

Subset	Source records	Questions	Questions per task
Afghanistan 2020	2,542	800	200
Afghanistan 2021	2,372	800	200
Iraq 2014	825	800	200
Iraq 2015	652	800	200
Syria 2016	6,325	800	200
Syria 2017	6,075	800	200
Total	18,791	4,800	–

2 S2. Natural-question construction and parser prompt

The natural-language set contains 600 unique subset::qid questions, 100 per subset. Questions were produced by machine-assisted drafting followed by deterministic construction and semantic-consistency checks; all 600 final questions passed deterministic semantic-consistency checks. The deterministic parser received only the final question and the fixed instruction below. IDs, task labels, gold fields, target events, and candidate events were never interpolated into the parser input.

```
You extract explicit event constraints from one natural-language question.
Return exactly one compact JSON object with exactly these keys:
{"time": null, "location": null, "participants": []}
Rules:
- time: use YYYY-MM-DD for one explicit date; use
  {"start": "YYYY-MM-DD", "end": "YYYY-MM-DD"} for an explicit range;
  otherwise null.
- location: copy only the explicit location text from the question as one
  string; otherwise null.
- participants: list only explicitly named organizations, armed groups,
  governments, or actor groups; do not list victims, counts, weapons,
  report titles, or locations.
- Never infer unstated facts. Do not explain. Do not use Markdown.

USER MESSAGE TEMPLATE:
{final_question_text}
```

The placeholder denotes the final question string supplied to the parser; the parser receives only the substituted question text, not the placeholder name.

The pinned local model was Qwen/Qwen2.5-1.5B-Instruct, local snapshot SHA-256 8808b40e9d5b491f3d9ea200c731bbc283a7cf825d7d836c02ddf791169bbb93. Decoding used seed 42, greedy output, a 96-token limit, batch size 16, and float16 on an NVIDIA GeForce RTX 4070. Raw model outputs were stored before normalization.

3 S3. Parser extraction results by field and subset

Table S2: Field-level parser results on all 600 natural-language questions. Values are reproduced from the frozen experimental results.

metric	value	n
Temporal normalized accuracy	0.8933	600
Location normalized accuracy	0.4833	600
Participant precision (micro)	0.7794	600
Participant recall (micro)	0.7183	600
Participant F1 (micro)	0.7476	600
Participant set exact accuracy	0.6617	600
Full T/L/A tuple exact match	0.3267	600
Invalid JSON rate	0.0017	600
Schema missing-field rate	0.0317	600
Schema extra-field rate	0.0050	600
Gold-required value missing rate	0.1833	600

Table S3: Parser results by subset.

subset	n	temporal_accuracy	location_accuracy	participant_precision	participant_recall	participant_f1	full_tla_exact	invalid_json_rate
Afghanistan_2020	100	0.9600	0.4600	0.8298	0.7800	0.8041	0.3500	0.0100
Afghanistan_2021	100	0.9100	0.3700	0.8523	0.7500	0.7979	0.2700	0.0000
Iraq_2014	100	0.9000	0.6500	0.8090	0.7200	0.7619	0.4300	0.0000
Iraq_2015	100	0.8000	0.4600	0.5978	0.5500	0.5729	0.1900	0.0000
Syria_2016	100	0.9400	0.4800	0.7449	0.7300	0.7374	0.3600	0.0000
Syria_2017	100	0.8500	0.4800	0.8478	0.7800	0.8125	0.3600	0.0000

The sole invalid JSON output corresponds to 1/600; schema-missing outputs correspond to 19/600. Location accuracy (0.4833) is the largest field-level weakness.

4 S4. Oracle/parser end-to-end results

On the full 600-question set, parser extraction lowers Macro Score from 0.8717 to 0.7317, lowers positive answer coverage from 0.9533 to 0.8022, and increases false support from 0.0267 to 0.0600.

5 S5. NLI, ScoreGate, and soft-verification details

The natural-question NLI verifier used MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli, revision 6f5cf0a2b59cabb106aca4c287eed12e357e90eb, and a validation-selected threshold of

Table S4: Paired EviGraphRAG-Struct results with oracle-normalized versus parser-extracted constraints.

Condition	Split	n	Event	Evid.	Attr.	Abst.	False	Macro
Oracle	Test	382	.8506	.9053	.7476	.9794	.0206	.8707
Parser	Test	382	.6322	.7684	.5631	.9485	.0515	.7280
Oracle	Full	600	.8533	.9000	.7600	.9733	.0267	.8717
Parser	Full	600	.6600	.7267	.6000	.9400	.0600	.7317

Table S5: Oracle/parser paired results by subset. Values are reproduced from the frozen by-subset experimental results.

Condition	Subset	Event	Evid.	Attr.	Abst.	False	Macro
Oracle	Afghanistan 2020	1.0000	.9600	.7200	.9200	.0800	.9000
Oracle	Afghanistan 2021	1.0000	.8800	.8000	1.0000	.0000	.9200
Oracle	Iraq 2014	.9200	1.0000	.8400	.9200	.0800	.9200
Oracle	Iraq 2015	.8800	.9200	.9200	1.0000	.0000	.9300
Oracle	Syria 2016	.5600	.7200	.6000	1.0000	.0000	.7200
Oracle	Syria 2017	.7600	.9200	.6800	1.0000	.0000	.8400
Parser	Afghanistan 2020	.8800	.6800	.6000	.9200	.0800	.7700
Parser	Afghanistan 2021	.9200	.8000	.6800	.9600	.0400	.8400
Parser	Iraq 2014	.7600	.8800	.5600	.7600	.2400	.7400
Parser	Iraq 2015	.4800	.4800	.8000	1.0000	.0000	.6900
Parser	Syria 2016	.2800	.6400	.4800	1.0000	.0000	.6000
Parser	Syria 2017	.6400	.8800	.4800	1.0000	.0000	.7500

0.75. The validation-selected soft verifier uses equal T/L/A/D weights and threshold 1.0, so it operationally reduces to exact conjunction. Its Macro Score (0.728246) is essentially tied with EviGraphRAG-Struct (0.728041); no meaningful difference is claimed.

In the separate mixed held-out robustness comparison, the validation-only selected NLI threshold is 0.70. Positive score is 0.847873, adversarial abstention accuracy 0.780987, false support 0.219013, balanced score 0.814430, overall UCR 0.188928, answered-conditioned UCR 0.307729, and answer coverage 0.613943. The validation-selected soft verifier reaches balanced score 0.914349 with threshold 1.0. A weaker equal-weight threshold of 0.75 raises answer coverage to 0.973931 but permits 0.946450 adversarial false support.

6 S6. Full per-constraint and single-constraint ablations

Table S7 includes every saved held-out ablation configuration. Values are reproduced from the frozen experimental results.

Removing T or L sharply increases false support. Removing A also reduces abstention accuracy. In the w/o D condition, documents are retrieved from the global document pool after event selection; the resulting Evidence F1 of 0.3222 quantifies the loss of event-record linkage. Single-constraint configurations never abstain on the original negatives.

7 S7. Constraint firing and overlap matrices

The candidate audit covers all 240,000 top-50 rows. Among 180,000 positive-candidate rows, temporal, location, and participant failures occur 159,791, 40,404, and 3,159 times; 36,137 rows have multiple simultaneous failures. Among 60,000 insufficient-evidence candidate rows, the corresponding counts are 54,955, 29,473, 13,834, and 36,470. No candidate has a provenance-link failure under the operational zero-linked-record definition, and no candidate fails the legacy 0.10 score threshold after the T/L/A mask.

For the additional adversarial candidates, wrong-actor, wrong-date, and wrong-place perturbations contain

Table S6: Natural-question comparison on the disjoint 382-question MD5 test complement. Thresholds and weights were selected only on the 218-question validation partition.

Method	Macro	False	Overall UCR	Ans. UCR	Coverage
Soft validation-selected	.728246	.051546	.1335	.2198	.607330
EviGraphRAG-Struct	.728041	.051546	.1309	.2155	.607330
Struct-GRAG+ScoreGate	.714517	.072165	.1257	.2124	.591623
NLI-Verifier	.700341	.051546	.2435	.3577	.680628
Soft equal-weight	.655012	.680412	.3298	.3739	.882199
Struct-GRAG	.595183	1.000000	.4319	.4319	1.000000

Table S7: All saved held-out ablation configurations (3,840 questions). Attr.: Attribute Accuracy; Event/Evid.: event/evidence F1; Abst.: abstention accuracy; False: false support; Macro: macro task score.

Configuration	Attr.	Event	Evid.	Abst.	False	Macro
Full T/L/A/D	.7857	.8928	.8756	.9776	.0224	.8829
T only	.7857	.8928	.8756	.0000	1.0000	.6385
L only	.7857	.8928	.8756	.0000	1.0000	.6385
A only	.7857	.8928	.8756	.0000	1.0000	.6385
D only	.7857	.8928	.8756	.0000	1.0000	.6385
w/o T	.7857	.8928	.8756	.2825	.7175	.7092
w/o L	.7857	.8928	.8756	.2783	.7217	.7081
w/o A	.7857	.8928	.8756	.6898	.3102	.8110
w/o D, global documents	.7857	.8928	.3222	.9776	.0224	.7446
$\gamma = 0$, provenance ranker	.7857	.8928	.8756	.9776	.0224	.8829
No event-score threshold	.7857	.8928	.8756	.9776	.0224	.8829
Soft equal, predefined	.7857	.8928	.8756	.0000	1.0000	.6385
Soft validation-selected	.7857	.8928	.8756	.9787	.0213	.8832

60,000, 19,850, and 60,000 candidate rows; multiple simultaneous failures occur in 54,379, 17,515, and 49,312 rows, respectively. These overlaps explain why one-dimensional score gating cannot substitute for field-wise conjunction.

8 S8. Threshold, M , gamma, and cap sensitivity

Thresholds 0.05, 0.10, and 0.15, as well as removal of the threshold, produce identical raw events and abstention decisions for all 4,800 questions. With reported $K = 1$, retained bounds $M \in \{1, 10, 30, 50, 100, \text{all}\}$ also produce zero event or support differences. Thus the threshold is not part of the conceptual default, and $M = 50$ is retained only as a matched implementation bound.

For the alternative provenance-strength ranker, $s_{\text{prov}}(e) = \min\{1, \ln(1 + n_d(e) + n_s(e)) / \ln(21)\}$ and $\ln(21) = \ln(1 + 20)$, so the term saturates at a combined count of 20. The empirical combined-count percentiles (50, 75, 90, 95, 99, maximum) are 2, 2, 3, 4, 6, and 18. Relative to cap 20, cap 10 changes 15 raw events and cap 40 changes 5; neither changes abstention. Setting $\gamma = 0$ leaves the saved aggregate unchanged. The influence is minor and non-monotonic.

9 S9. Hard-positive construction audits and results

The challenge is a separately constructed, mechanism-oriented test. Every case has a higher-ranked distractor that violates exactly one represented temporal, location, or participant constraint and a lower-ranked target

Table S8: Constraint-failure overlap counts over all 240,000 candidate rows. Diagonal entries are marginal failure counts.

	T	L	A	D	Score
T	214,746	58,628	15,059	0	0
L	58,628	69,877	2,912	0	0
A	15,059	2,912	16,993	0	0
D	0	0	0	0	0
Score	0	0	0	0	0

Table S9: Mask and score-threshold audit.

Question type	Candidate rows	Removed by T/L/A	Survive	Removed by 0.10
Positive	180,000	167,058	12,942	0
Insufficient evidence	60,000	59,955	45	0
Total	240,000	227,013	12,987	0

that satisfies all represented constraints. All 240 cases passed deterministic construction, ranking, and single-constraint consistency checks, with 80 temporal, 80 location, and 80 participant failures. The challenge is not merged into the 4,800-question benchmark.

Table S10: Construction-audited hard-positive challenge. Unver.: unverified top-1 accuracy; Ver.: verified top-1 accuracy; Removal: distractor removal rate; Recovery: lower-ranked target recovery rate; FRR: false-rejection rate.

Group	n	Unver.	Ver.	Removal	Recovery	FRR	Mean target rank
Overall	240	.0000	1.0000	1.0000	1.0000	.0000	3.8417
Temporal	80	.0000	1.0000	1.0000	1.0000	.0000	4.0500
Location	80	.0000	1.0000	1.0000	1.0000	.0000	4.0250
Participant	80	.0000	1.0000	1.0000	1.0000	.0000	3.4500

The median target rank before verification is 2, and every verified target rank is 1. The result demonstrates the intended correction mechanism under designed single-constraint distractors; it is not general benchmark performance. To avoid redistributing restricted record text, this supplement reports construction rules and aggregates rather than raw examples.

10 S10. Predicted-cluster per-subset and purity-bin results

The final predicted-cluster evaluation uses only the post-location-repair output. The final run uses quoted-title-aware actor and location extraction, each audited on a fixed 50-case sample. The full question remains unchanged for semantic ranking, while quoted title spans are masked only for structured extraction. The final run covers all 4,800 questions, with supported-question coverage 1.0 and positive false rejection 0 for both methods.

11 S11. Remaining attribute and false-support taxonomies

The final EviGraphRAG-Struct run has 343 attribute errors: 295 occur because the verified top-1 does not cover the gold event, and 48 because no exact target match is present in the semantic $M = 50$ pool. These are candidate retrieval/ranking limitations rather than positive verifier rejections.

Table S11: Final EviGraphRAG-Struct predicted-cluster results by subset. Cov.: supported-question coverage; FRR: positive false rejection; Abst.: abstention accuracy; Macro: macro score.

Subset	Event	Evid.	Attr.	Cov.	FRR	Abst.	False	Macro
Afghanistan 2020	.9800	.9776	.6900	1.0000	.0000	.9850	.0150	.9082
Afghanistan 2021	.9642	.9577	.7550	1.0000	.0000	.9850	.0150	.9155
Iraq 2014	.9275	.9226	.7900	1.0000	.0000	.9600	.0400	.9000
Iraq 2015	.8708	.8837	.7550	1.0000	.0000	.9850	.0150	.8736
Syria 2016	.6859	.6938	.6500	1.0000	.0000	.9700	.0300	.7499
Syria 2017	.6472	.6347	.6450	1.0000	.0000	.9600	.0400	.7217

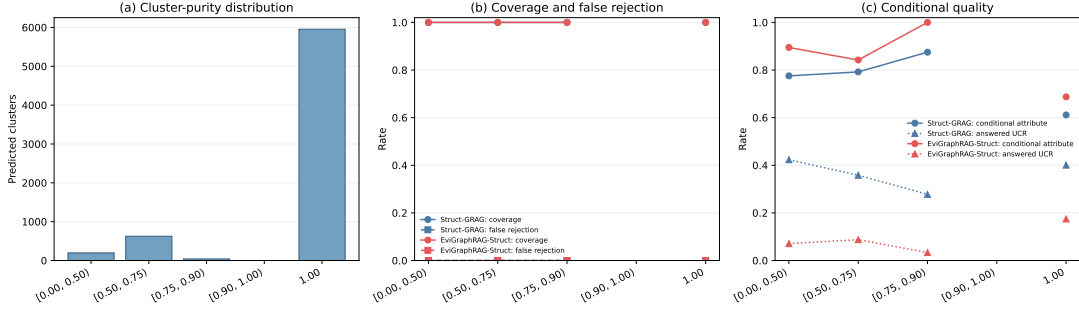


Figure S1: Final predicted-cluster performance by purity bin after the fixed-sample-audited actor and location parser repairs. Empty bins are retained in the source table but not interpreted.

In the original gold-graph benchmark, all 25 remaining false-support cases contain a title- or description-level condition not represented by T/L/A. Title and description text remains available for semantic ranking, but it is not an explicit verifier field; no post-hoc title constraint is introduced.

Among 938 held-out insufficient-evidence questions, EviGraphRAG-Prov abstains on 917 for which Struct-GRAG returns false support. Among 2,902 held-out positives, 2,278 have Text-RAG UCR at least 0.8 and EviGraphRAG-Prov UCR at most 0.2. Each query is counted once per category.

12 S12. Matched generation details

The matched generation run uses the same 600 IDs, prompt builder, context formatting, local Qwen2.5-1.5B-Instruct weights, greedy decoding, 220-token limit, seed 42, and original evaluator as the historical generation comparison.

Ret. abst. and Ver. rej. are retrieval-level abstention and verifier rejection, respectively, and are distinct from LLM refusal, which is conditioned on accepted retrieval. Eval. agg. is the original evaluator aggregate and is not conventional answered accuracy. Cond. acc. is accuracy among generated answers; UCR and Ans. UCR are overall and answered-conditioned unsupported claim rates. False support is 0, and positive generation coverage is 0.3200. Event/evidence citation credit and natural-language refusal are separate evaluator outputs: an output can refuse in natural language while retaining cited identifiers in the JSON. Structured attribute evaluation reads the selected event’s normalized attribute, whereas generation must also verbalize the value, conform to JSON, and avoid refusal; the latter interface requirements explain the lower generation-level attribute result. The No-Retrieval LLM is a prompt-behavior control, not a competitive retrieval baseline.

Table S12: Final EviGraphRAG-Struct purity-bin results. Cov.: supported coverage; FRR: positive false rejection; Cond. values are answered-conditioned.

Purity	Queries	Cov.	FRR	Cond. Event	Cond. Evid.	Cond. Attr.	False
[0.00, 0.50)	230	1.0000	.0000	.4117	.4171	.8947	.0000
[0.50, 0.75)	516	1.0000	.0000	.6030	.5976	.8421	.0303
[0.75, 0.90)	36	1.0000	.0000	.6061	.6229	1.0000	.0000
1.00	4,018	1.0000	.0000	.9058	.9047	.6873	.0270

Table S13: Final predicted-cluster residual-error taxonomy.

Task	Residual reason	Count	Rate of all 4,800 queries
Attribute QA	Supported cluster misses attribute gold event	343	.0715
Event retrieval	Supported cluster misses gold event	104	.0217
Evidence tracing	Supported documents miss gold evidence	100	.0208
Insufficient evidence	False support on insufficient evidence	31	.0065

13 S13. Reproducibility commands, versions, prompts, and release scope

The core embedding model is `sentence-transformers/all-MiniLM-L6-v2`. Final hard-positive and predicted-cluster runs used Python 3.9.21, PyTorch 2.4.1+cu124, SentenceTransformers 5.1.2, an NVIDIA GeForce RTX 4070, batch size 128, seed 42, and a retained semantic candidate bound $M = 50$. The recorded predicted-cluster command includes legacy `-event-threshold 0.10`; Section S8 documents that this option removed no post-mask candidate and changed no output.

The following path-neutralized commands record the experiment-side settings; data and local input wiring are described in the public repository documentation.

```
python src/evaluation/run_hard_positive_final.py \
--device cuda \
--model sentence-transformers/all-MiniLM-L6-v2 \
--batch-size 128 \
--seed 42 \
--candidate-pool 50

python src/retrieval/run_predicted_graph_final_after_location_repair.py \
--device cuda \
--model sentence-transformers/all-MiniLM-L6-v2 \
--batch-size 128 \
--seed 42 \
--candidate-pool 50 \
--top-k-docs 5 \
--event-threshold 0.10
```

The public repository at <https://github.com/hulahula2333/EviGraphRAG> includes question templates, normalization rules, split identifiers, configurations, graph/evaluation scripts, parser and generation prompts, baseline implementations, redistribution-safe artifacts, and identifier-only frozen predicted-cluster assignments. The assignments support exact reconstruction of the reported predicted-graph evaluation input from a separately obtained copy of UCDP GED v25.1; they do not reproduce raw-to-cluster construction, and no recovered historical cluster producer is claimed. Raw UCDP records and restricted title/question text are not redistributed.

Table S14: Representative false-support cases from the separate seed-42 20/80 case analysis.

Case	Query claim	Struct-GRAG behavior	EviGraphRAG-Prov behavior
Wrong actor	Panjwayi district, 2021-03-11, involving Civilians.	Retrieves a same-date, same-district event whose participants are Government of Afghanistan and Taleban.	Rejects the path because the participant constraint fails and returns insufficient evidence.
Wrong place	Qarabagh district, 2021-09-18, involving IS.	Retrieves an event involving IS on the same date but in Jalalabad district.	Rejects the candidate because the location constraint fails and abstains.

Table S15: Matched EviGraphRAG-Struct generation evaluation on 600 questions.

Method	Ret. abst.	Ver. rej.	LLM refusal	Eval. agg.	Cond. acc.	Coverage	UCR	Ans. UCR
EviGraphRAG-Struct	0.2433	0.2433	0.6828	0.7700	0.5278	0.2400	0.0600	0.2500