BENCHMARKING GEOGRAPHICAL TEMPORAL KNOWLEDGE GRAPH REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Temporal knowledge graphs (TKGs) built from event data carry rich latent geography, yet existing TKG question answering (TKGQA) benchmarks test only structural and temporal reasoning. We present GeoTKGQA, a question-answer generator whose questions provably require geographical reasoning by construction. We ground the locative anchors of a TKG to a shared geographical knowledge layer and design twelve schema families that compose structural, temporal, and spatial operators. We then execute every logical form symbolically and keep a question only if dropping the constraints of any one axis would change its gold answer. Applied to four standard event graphs, the pipeline yields 67,565 triaxially entangled questions with difficulty bands and four generalization splits, without human annotation or large language models. We evaluate ten baselines from four paradigms and analyze all 160 runs with a generator-backed error attribution protocol. The strongest system wins ten out of sixteen benchmark cells, but resolves only 0.4116 of mean-pooled questions, while 19.94% of the questions defeat all ten methods. GeoTKGQA's code and data are anonymously available at `https://anonymous.4open.science/r/GeoTKGQA-6FB8/`.

1 INTRODUCTION

Question answering over temporal knowledge graphs (TKGQA) has matured around two reasoning axes: the structural axis, which composes relational facts, and the temporal axis, which handles ordering, windows, and granularity (Jia et al., 2018; Saxena et al., 2021; Jia et al., 2021; Chen et al., 2023; Sun et al., 2025). The event graphs behind this progress, the ICEWS releases and GDELT, record interactions between political actors whose names expose a deterministic geographical anchor for most facts, and geography is central to what analysts ask of such data: who in a region acted first, which neighbor a state turned to. A right actor from the wrong region is thus a wrong answer, yet no TKGQA resource requires spatial reasoning jointly with the other two axes; geographical question answering has so far targeted static geo-spatial bases (Punjani et al., 2018; Kefalidis et al., 2023), and merely naming a region does not close the gap, because such a question may still be answerable without any reasoning about that region at all.

This paper introduces GeoTKGQA, a question-answering (QA) benchmark generator that turns the latent geography of TKGs into a third, certified reasoning axis, as illustrated in Figure 1. GeoTKGQA builds a geographical knowledge layer that grounds entities to countries and organizes countries into regions, borders, centroids, and hierarchies, instantiates typed logical forms from twelve schema families, obtains gold answers by deterministic symbolic execution, and keeps a question only if deleting the constraints of any one axis provably changes its answer. The questions are thus correct by construction and entangled by certification, with no annotation and no LLM in the loop. We generate benchmarks from four widely used sources and evaluate ten baselines from four paradigms, chosen so that pairwise differences isolate question decomposition and retriever training. Our contributions are (i) a geographical knowledge layer and a generate-then-verify pipeline that produce provably triaxially entangled questions from any point-timestamped TKG; (ii) a capacity analysis that estimates how many such questions a source graph can support and makes the released size an explicit, parameterized choice; (iii) four benchmarks totaling 67,565 entangled questions with difficulty bands, diagnostic, unanswerable, and multiple-choice supplements, and four generalization splits; and (iv) an evaluation of ten baselines with a generator-backed attribution protocol that re-executes each answer under relaxations of its logical form.

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

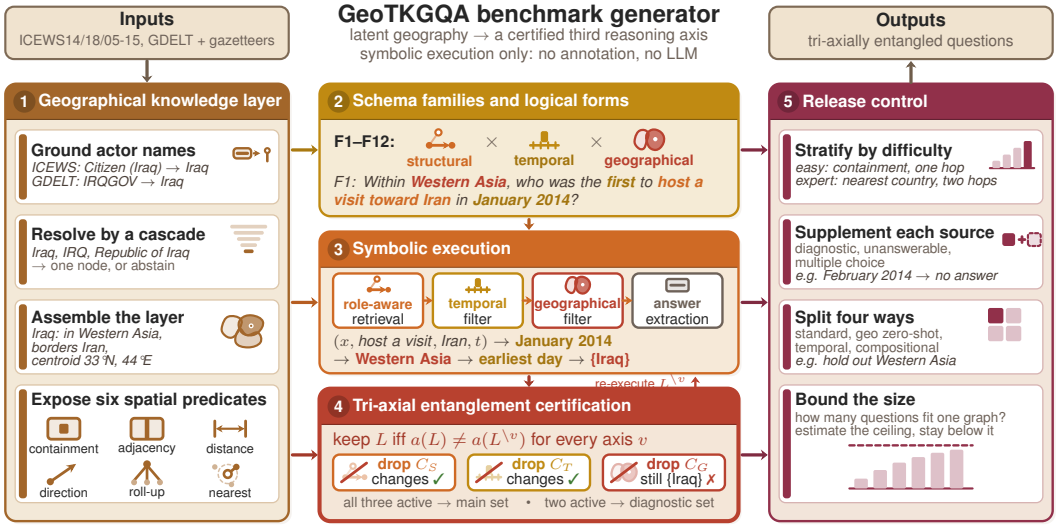

Figure 1: Overview of GeoTKGQA. A geographical knowledge layer (1) grounds actor names to countries and exposes six spatial predicates; twelve schema families compose structural, temporal, and geographical operators into typed logical forms (2), which a four-stage symbolic engine executes to gold answers (3); a form is released only if dropping the constraints of any one axis changes its answer (4); release control stratifies, supplements, splits, and bounds each benchmark (5). The running example (family F1) is followed through panels 2 to 4.

2 Related Work

Temporal and geographical QA. TempQuestions (Jia et al., 2018), CronQuestions (Saxena et al., 2021), TimeQuestions (Jia et al., 2021), MultiTQ (Chen et al., 2023), and TimelineKGQA (Sun et al., 2025) test structural composition and temporal logic over a TKG, multi-hop KGQA resources (Zhang et al., 2018; Talmor & Berant, 2018) fix the structural axis alone, and TimeQA (Chen et al., 2021) asks time-sensitive questions over text; systems have progressed from temporal embedding scorers (Saxena et al., 2021; Mavromatis et al., 2022; Chen et al., 2023) to retrieval-augmented pipelines with question rewriting or decomposition (Qian et al., 2024; Gong et al., 2025; Qian et al., 2025) and to reinforcement-learned retrieval agents (Jin et al., 2025; Gong et al., 2026). Geographical QA has been studied over static geo-spatial knowledge bases (Punjani et al., 2018; Kefalidis et al., 2023), surveyed as a problem class of its own (Mai et al., 2021), and probed as parametric knowledge of language models (Bhandari et al., 2023). No resource on either side constrains a temporal answer spatially, and none certifies that its spatial mention is needed to answer the question.

Programmatic benchmarks and shortcuts. Templated generation with executable gold answers underlies diagnostic resources for compositional (Sinha et al., 2019) and mathematical (Mirzadeh et al., 2025) reasoning, and audits have shown how answer priors and lexical overlap let models bypass the intended reasoning (Kaushik & Lipton, 2018). GeoTKGQA adds a certificate to this tradition: every released question is re-executed with each axis removed, so the necessity of each axis is measured on every question rather than assumed.

3 Problem Formulation

A temporal knowledge graph (TKG) is a tuple $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T}, \mathcal{F})$ with a finite set of entities $\mathcal{E}$, a finite set of relations $\mathcal{R}$, a finite, totally ordered set of time points $\mathcal{T}$ equipped with the granularity lattice $day \prec week \prec month \prec year$, and a finite fact set $\mathcal{F} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E} \times \mathcal{T}$ of quadruples $f = (s, r, o, t)$ with projections $\mathrm{sub}(f)$, $\mathrm{rel}(f)$, $\mathrm{obj}(f)$, and $\tau(f)$. We instantiate GeoTKGQA on four point-timestamped TKGs, three ICEWS releases (Boschee et al., 2015) and one GDELT slice (Leetaru & Schrodt, 2013), the standard graphs of temporal knowledge graph reasoning (García-Durán et al., 2018; Jin et al., 2020; Lacroix et al., 2020). Their actor names carry a deterministic

locative anchor, a final parenthetical in ICEWS and the CAMEO actor code in GDELT (Gerner et al., 2002), so a country can be recovered for at least one entity of most facts without annotation (coverage g_f in Table 1); a knowledge layer normalizes these anchors onto one shared spatial hierarchy.

Definition 1 (Geographical knowledge layer) *A geographical knowledge layer (GKL) over $\mathcal{G}$ is a tuple $\mathcal{H} = (N, A, \rho, \kappa, \gamma)$ in which N is a finite set of geographical nodes comprising country-level nodes (countries and territories) and region-level nodes $Reg \subset N$ (UN sub-regions, continents, and named code lists) organized in the containment hierarchy $subregion \subseteq continent$; $A \subseteq N \times N$ is a symmetric adjacency relation encoding shared land borders; $\rho\colon N \to \mathbb{R}^2$ maps each node to a (latitude, longitude) centroid; $\kappa\colon N \to 2^{Reg}$ assigns each node to the region-level nodes that contain it; and $\gamma\colon \mathcal{E} \to N \cup \{\perp\}$ is a partial grounding function mapping entities to their primary geographical node, or $\perp$ if unresolvable.*

For a fact f the participant locations are $\lambda_s(f) = \gamma(\mathrm{sub}(f))$ and $\lambda_o(f) = \gamma(\mathrm{obj}(f))$; f is *geographically grounded* if at least one of the two is not $\perp$, and $(\mathcal{G}, \mathcal{H})$ is a *geo-augmented TKG*. A GeoTKGQA question is a triple $q = (\sigma, L, a)$: σ is a natural language surface string, L a typed logical form fixing structural constraints C_S on entity roles and relation types, temporal constraints C_T on windows, orderings, and granularity (Jia et al., 2021), and geographical constraints C_G over the spatial relations of $\mathcal{H}$, and a is the gold answer set obtained by symbolically executing L over $(\mathcal{G}, \mathcal{H})$; the answer type is one of **entity**, **location**, **time**, **number**, or **region_choice**. Every released question is meant to be *tri-axially entangled*: removing the constraints of any single axis provably changes the gold answer, so no solver can answer it without reasoning on all three axes at once.

Two challenges. Existing benchmarks evaluate structural composition and temporal logic (Allen, 1983); the *geographical (G) axis*, reasoning over containment, adjacency, metric proximity, cardinal direction, and hierarchical roll-up, is absent even though most source facts involve an anchored entity. Mentioning geography, however, does not make geographical reasoning necessary; our generator itself produces such cases: for "Within Western Asia, who was the first to take the action 'host a visit' toward Iran in January 2014?" the answer, Iraq, is unchanged once the region constraint is dropped, because the earliest host that month is Western Asian anyway. A second failure mode is the shortcut, answer priors or lexical overlap that bypass reasoning (Mavromatis et al., 2022; Gong et al., 2025; Qian et al., 2025; Gong et al., 2026; Kaushik & Lipton, 2018). Section 5 answers both with a generate-then-verify pipeline: every question is produced by symbolic execution, certified at generation time to require all three axes, which demotes the example above to the diagnostic set, and screened for shortcuts and answer leakage before release.

4 Geographical Knowledge Layer

Gazetteers and grounding. Country-level nodes come from the ISO 3166 standard (ISO, 2006); UN M49 (Office, 1982) supplies regions and sub-regions nested within continents, supplementary nodes cover territories absent from the base snapshot, and land-border adjacency and centroids come from a public country-information snapshot cross-referenced with public gazetteers (Ahlers, 2013). One deterministic adapter per source family extracts the anchor token of each entity name, the final parenthetical in ICEWS and the leading ISO 3166-1 alpha-3 code of the CAMEO actor encoding in GDELT, and passes it through a five-stage cascade with decreasing confidence levels $c_1 \geq \cdots \geq c_5$: exact code, exact canonical name, unique normalized alias, unique normalized name, and population-based tie-breaking among ambiguous matches. A grounding is accepted iff its confidence reaches the registry threshold θ, and the grounding file records, for every entity, the resolved node, its level, and the confidence reached.

Spatial relations. GeoTKGQA consumes six relation families derived from $\mathcal{H}$. *Containment* maps each region name to its member countries and *adjacency* maps each country to its border neighbors. *Distance* between two country centroids is the Haversine great-circle distance $d(n_i, n_j) = 2R \arcsin\sqrt{u_{ij}}$ with $u_{ij} = \sin^2(\Delta\varphi/2) + \cos\varphi_i \cos\varphi_j \sin^2(\Delta\psi/2)$, where $(\varphi_i, \psi_i) = \rho(n_i)$ and $R \approx 6{,}371$ kilometers; this one quantity drives both the distance-threshold and the nearest-neighbor predicates. *Direction* takes two forms under a fixed no-wraparound longitude convention: the *directional extremum* of a candidate set is the country maximizing the coordinate of a cardinal direction, and the *directional half-plane* of an anchor country is the set of countries lying strictly

further in that direction. Finally, *co-membership under roll-up* maps an anchor country to its enclosing UN sub-region and returns the remaining members, a containment predicate whose surface string names no region, so the solver must infer the region itself.

5 Question Generation and Certification

5.1 Evidence Consistency and Schema Families

Definition 2 (Tri-axially consistent evidence) *Let $q = (\sigma, L, a)$ have constraints (C_S, C_T, C_G) with a temporal window $W \subseteq \mathcal{T}$, an optional ordinal operator, and a target granularity ℓ. An evidence set $E \subseteq \mathcal{F}$ is tri-axially consistent with q if (i) every $f \in E$ matches the role and relation bindings of C_S; (ii) $\tau(f) \in W$ for every $f \in E$ and, under an ordinal operator, every $f \in E$ attains the minimal (first) or maximal (last) admissible day, coerced to granularity ℓ; and (iii) every participant e of f that C_G constrains is resolvable ($\gamma(e) \neq \perp$) and satisfies its spatial predicate.*

We extract gold answers exclusively from such sets. We design a library of twelve *schema families*, F1 through F12 (Table A1 in Appendix A.1), each a parameterized template fixing one composition of structural, temporal, and geographical operators; twelve compositions exercise every spatial predicate and every answer type at least once, and the registry admits further families. A *typed logical form* L instantiates a family's parameters, anchor entity and role, relation, geographic constants, window, and operator settings, with constants drawn from $(\mathcal{G}, \mathcal{H})$, and is released with each question. The library covers ordinal reasoning under region containment (F1, F2, and the two-hop chain F8), selection and counting over the border graph (F3, F4, F12), metric proximity (F5, F6), comparative direction (F9), hierarchical roll-up (F10), and counting, comparison, and precedence over named regions (F7, F11); together the families realize all five answer types.

5.2 Symbolic Execution and Gold Answers

The execution engine turns a typed logical form L into the gold answer set $a(L)$ with a supporting evidence pool $E(L) \subseteq \mathcal{F}$ in four deterministic stages (Figure 1, panel 3). Stage 1 (role-aware retrieval) fetches all facts matching C_S through indices keyed on the anchor's role and relation, enforcing structural consistency by construction. Stage 2 (temporal filtering) retains facts with $\tau(f) \in W$. Stage 3 (geographical filtering) applies the spatial predicate of L or, for the selection operators, computes the nearest neighbor and the directional extremum over the candidate countries. Stage 4 (answer extraction) projects the answer from the filtered pool: ordinal selection at the target granularity, the set of qualifying countries, distinct-country counts, the earliest-day precedence between two regions, or, for the chain family F8, intermediate-entity resolution followed by a second-hop retrieval (index layout in Appendix A.2).

Proposition 1 (Correctness by construction) *If γ correctly maps entities to geographical nodes, then for every generated question $q = (\sigma, L, a)$ the stored answer satisfies $a = a(L)$ and $E(L)$ is tri-axially consistent with q: Stages 1 to 3 enforce clauses (i) to (iii) of Definition 2, the answer is extracted from their intersection, and under correct grounding every membership test of Stage 3 is sound, so the stored answer set is exact.*

5.3 Tri-Axial Entanglement Certification

A correct answer is necessary but not sufficient: the question must *require* all three axes.

Definition 3 (Tri-axial entanglement) *For each axis $v \in \{S, T, G\}$ let $L^{\setminus v}$ denote the form obtained from L by removing the constraints of axis v, with answer set $a(L^{\setminus v})$. Axis v is* active *in L if $a(L) \neq a(L^{\setminus v})$, and L is* tri-axially entangled *if all three axes are active:*

$$\bigwedge_{v \in \{S,T,G\}} a(L) \neq a(L^{\setminus v}). \tag{1}$$

The implemented certification of Equation 1 is stricter in two ways: each single-axis relaxation must also grow the projected candidate pool or the answer set, guarding against vacuously satisfied

constraints, and no single-axis execution alone may reproduce the full answer. Tuple-valued comparisons (the F7 comparison mode and F11) relax to a degenerate pair so that all comparisons remain well-typed, and any form whose full execution yields a tie, a zero count, or a missing-region sentinel is rejected. Forms that pass exactly two of the three tests are retained, up to a small registry quota, as *diagnostic* items, separated from the entangled main set.

5.4 Difficulty, Supplementary Sets, and Splits

Definition 4 (Programmatic difficulty) *Given a typed logical form L with evidence pool $E(L)$ and answer set $a(L)$, its programmatic difficulty is*

$$d_{\mathrm{prog}}(L) = h_S + h_T + h_G + \tfrac{1}{2}\ln\big(1 + |E(L)|\big) + \tfrac{1}{2}\ln\big(1 + |a(L)|\big). \tag{2}$$

Here h_S is the number of relational hops (2 for F8, 1 otherwise); h_T starts at 0 and is incremented for a non-span window, an ordinal operator, a pivot or chain ordering, a precedence comparison, and coarse granularity; h_G sums the operator weights along the composition, with containment 2, roll-up 3, adjacency 4, direction 5, distance threshold 6, and nearest neighbor 7; and the two damped logarithmic terms grow with the evidence-pool and answer-set sizes. Within each source, main-set values are binned at their empirical quartiles into four bands, L1_Easy to L4_Expert, whose cut-points ship with the release, in the spirit of item response theory (Lord, 2012).

Three *supplementary sets* accompany the entangled main set of each source. *Unanswerable* questions perturb an entangled question, by replacing the target region with a disjoint one or shifting the window by its own width, so that its logical form provably has an empty result, verified by re-execution. *Multiple-choice* items pair single-answer entity or location questions with three distractors drawn from the geographically relaxed candidate pool, so every distractor satisfies the structural and temporal constraints but fails the geographical one. *Diagnostic* items are the two-axis forms above. Each source finally receives four training, validation, and test *splits*: a *standard* split stratified by family and band; a *geographical zero-shot* split that holds out a fraction η_G of the geographic constants and routes every question mentioning a held-out constant to the test partition; a *temporal* split that routes to the test partition every question whose window or pivot lies in the final fraction η_T of the source span; and a *compositional* split that holds out the family-operator combinations (F1, last), (F4, east or west), (F7, compare), and (F9, east or west).

6 Benchmark Capacity and Release

How many entangled questions can a source TKG support, and how much of that capacity should a release use? Let g_f be the fraction of facts with at least one grounded participant; over the multi-fact groups, the entity-relation groups in either role orientation that contain at least two facts, let p_G be the fraction whose grounded counterpart entities span at least two countries and p_T the fraction whose facts span at least two distinct days of the source.

Definition 5 (Entanglement density) *The entanglement density of a TKG is*

$$\delta = g_f \cdot p_G \cdot p_T, \qquad g_f = \big|\{f \in \mathcal{F} : \lambda_s(f) \neq \bot \vee \lambda_o(f) \neq \bot\}\big| \,/\, |\mathcal{F}|. \tag{3}$$

Equation 3 multiplies the three requirements a fact group must meet to support a discriminative constraint on every axis, so δ estimates the probability that a uniformly drawn group can anchor an entangled question. The number N^* of distinct certifiable forms is finite and scales to first order as $N^* \propto m \cdot \delta$ with $m = |\mathcal{F}|$, the size of the admissible window family $\mathcal{W}$ acting as a secondary multiplier. A sampling budget of B attempts yields the observed unique count $U(B)$ of certified forms and the numbers f_1 and f_2 of forms seen once and twice, from which the Chao1 estimator (Chao, 1984) extrapolates the ceiling. The released size N_{rel}, the observed count, and the estimated ceiling are then linked by the chain of inequalities

$$\widehat{N^*} = S_{\mathrm{obs}} + \frac{f_1^2}{2f_2}, \qquad N_{\mathrm{rel}} \leq U(B) \leq \widehat{N^*}, \tag{4}$$

with $S_{\mathrm{obs}} = U(B)$ and the bias-corrected variant applied when $f_2 = 0$. The screen retains the four event graphs and excludes the YAGO (Suchanek et al., 2007) and Wikidata (Vrandečić & Krötzsch, 2014) subsets, whose density collapses because their facts concentrate on a few biographical relations.

Table 1: The released benchmarks: m facts, span in days, coverage g_f, the entangled main set by gold answer type, and the total with the three supplementary sets (1,500 diagnostic, 1,000 unanswerable, and 2,000 multiple-choice items per source).

Source	m	span	g_f	**entity**	**location**	**time**	**number**	**region_choice**	**Main**	**Total**
ICEWS14	90,730	365	0.850	5,536	3,109	2,125	308	289	11,367	15,867
ICEWS18	468,558	304	0.855	9,307	6,449	2,500	468	1,069	19,793	24,293
ICEWS05-15	461,329	4,017	0.864	6,500	4,356	2,500	348	506	14,210	18,710
GDELT	2,278,405	31	0.836	10,000	8,369	1,798	506	1,522	22,195	26,695
Total				31,343	22,283	8,923	1,630	3,386	67,565	85,565

The gap between N_{rel} and $\widehat{N^*}$ is deliberate and fully parameterized by one hyperparameter registry whose resolved snapshot ships with each release (Appendix A.3). Volume knobs (per-family quota, sampling budget B, shards) set how much of the observed richness is released; balance and anti-shortcut knobs (per-anchor, per-answer, per-constant, and per-relation caps, an answer-set size cap, and a leakage policy rejecting any question whose surface string contains one of its answers) reshape the released distribution without touching the ceiling; and content-shape knobs (window and granularity registries, the distance ladder of F6, the grounding threshold θ, split fractions, and the master seed) fix what the questions look like. Releasing well below the ceiling is what keeps the balance caps affordable: a run that exhausted the observed richness would have to admit the long tail of anchors, answers, and constants that those caps exist to exclude. Table 1 reports the outcome of one deterministic run per source under the default registry snapshot.

7 Experiments

7.1 Baselines and Protocol

We evaluate ten systems from four paradigms, chosen so that several pairs differ in exactly one component and no modelling assumption is shared by all; a property of the whole panel is then a property of the questions. **Embedding readers.** MultiQA (Chen et al., 2023) scores candidates with TComplEx temporal embeddings (Lacroix et al., 2020) and multi-granularity time aggregation; CronKGQA (Saxena et al., 2021) scores entity and time candidates over the same embeddings with one timestamp embedding per question; EmbedKGQA (Saxena et al., 2020) is deliberately time-blind, scoring atemporal ComplEx embeddings of the static collapse of the graph. **Retrieval-augmented pipelines.** In-Context RALM (Ram et al., 2023) performs one BGE-M3 retrieval (Chen et al., 2024) of 50 facts and one reader call, instantiated as the "without decomposer" ablation of RTQA with the same corpus, prompt, and DeepSeek-V4-Flash reader; RTQA (Gong et al., 2025) recursively decomposes a question into a tree, answers each node with the same LLM over the same retrieval, and aggregates bottom-up, so the two differ only in decomposition; TimeR^4 (Qian et al., 2024) runs a retrieve, rewrite, retrieve, rerank loop whose retriever is fine-tuned per cell (Reimers & Gurevych, 2019) for the same frozen reader, isolating trained retrieval. **Reinforcement-learned agent.** Temp-R1 (Gong et al., 2026) is a Search-R1-style agent (Jin et al., 2025): Llama-3.1-8B-Instruct (Grattafiori et al., 2024) is cold-started on teacher trajectories and trained with two-stage GRPO (Shao et al., 2024) under a reverse curriculum, with e5-base-v2 retrieval (Wang et al., 2022) of 30 facts per query for at most three turns. **Closed-book control.** DeepSeek-V4-Flash, Gemma4-31B-IT, and Qwen3.6-35B-A3B answer from parametric knowledge alone under eight-shot prompting (Brown et al., 2020), reasoning modes disabled and decoding greedy; this group calibrates the floor that evidence access must clear and never counts as a winner.

Protocol. Every (source, split) cell receives its own trained model, 16 per trained baseline, because the four splits repartition one shared question pool. Each system is scored by its own official Hits@1 protocol: rank-based Hits@1 for the readers, the lenient string match of RTQA (Rajpurkar et al., 2016) for In-Context RALM, RTQA, and the closed-book group, TimeR^4's containment-based hit, and Temp-R1's lenient exact match; a uniform strict exact-match score is stored for every generative run and is reported in Table A5. Configurations are in Appendix B; all per-question predictions are released, and the analysis pipeline reproduces every headline number from them.

Table 2: Main results: test Hits@1 of the ten baselines on all 16 (source, split) cells, each system scored by its own official protocol (Section 7.1); n is the test-set size and bold marks the best evidence-eligible system per row; the closed-book control group is never bolded. RALM abbreviates In-Context RALM; DS-V4F, G4-31B, and Q3.6-35B abbreviate DeepSeek-V4-Flash, Gemma4-31B-IT, and Qwen3.6-35B-A3B. Splits: std = standard, geo = geographical zero-shot, tmp = temporal, cmp = compositional; confidence intervals in Appendix C.6.

			Embedding			**Retrieval-augmented**			**RL**	**Closed-book (control)**		
Source	**Split**	n	Multi QA	Cron KGQA	Embed KGQA	RALM	RTQA	TimeR^4	Temp-R1	DS-V4F	G4-31B	Q3.6-35B
ICEWS14	std	2,285	0.2508	0.2416	0.3107	0.3514	0.3037	0.3383	**0.5133**	0.0902	0.1046	0.1107
	geo	2,154	0.1346	0.1272	0.1518	0.3250	0.2669	0.2906	**0.4475**	0.1133	0.1207	0.1184
	tmp	1,165	0.1004	0.0343	0.1923	0.4618	0.3906	0.3948	**0.4961**	0.1227	0.1365	0.1348
	cmp	1,301	0.2590	0.2129	0.2006	0.4420	0.3359	0.4412	**0.5919**	0.0884	0.0999	0.1115
ICEWS18	std	3,973	0.1540	0.1447	0.2071	0.3781	0.3111	0.3793	**0.4908**	0.1631	0.1651	0.1472
	geo	1,530	0.1869	0.1464	0.1739	**0.4477**	0.4111	0.4248	0.4124	0.1425	0.1634	0.1418
	tmp	2,396	0.1402	0.0071	0.1924	0.4232	0.3606	0.4278	**0.4925**	0.2116	0.2199	0.1949
	cmp	3,170	0.1013	0.0681	0.1435	0.4388	0.3297	0.4662	**0.4681**	0.1691	0.1852	0.1612
ICEWS05-15	std	2,858	0.1407	0.1382	0.2057	0.3810	0.3223	0.3429	**0.4843**	0.1211	0.1148	0.1085
	geo	1,279	0.1603	0.1658	0.1861	0.4292	0.3714	0.3495	**0.5004**	0.0993	0.1134	0.1126
	tmp	1,188	0.0791	0.0825	0.1507	0.4259	0.3645	0.3653	**0.4638**	0.1279	0.1397	0.1296
	cmp	1,898	0.1607	0.1117	0.1396	**0.5116**	0.3915	0.4636	0.4673	0.1054	0.1365	0.0864
GDELT	std	4,457	0.1710	0.1750	0.2387	0.3094	0.2704	**0.3350**	0.3090	0.1014	0.1521	0.1479
	geo	3,106	0.0840	0.1217	0.1851	0.2888	0.2511	**0.3303**	0.2598	0.0702	0.0889	0.0976
	tmp	2,823	0.1187	0.0680	0.2586	**0.3900**	0.3333	0.3829	0.2692	0.1321	0.1831	0.1807
	cmp	3,653	0.1163	0.0903	0.1566	0.3589	0.3052	**0.4303**	0.2776	0.1298	0.1793	0.1831

7.2 Main Results

Table 2 reports all 160 results; four observations organize the rest of the section. First, the benchmark is far from saturated and the evidence channel is what separates systems from the floor: the best single cell reaches 0.5919 (Temp-R1 on the ICEWS14 compositional split), the strongest pooled accuracy is 0.4116, and every cell of every evidence-using system exceeds the best closed-book cell (0.2199), while the embedding readers span 0.0071 to 0.3107. The control group is at its best on the temporal split, the highest of its four cells in eleven of twelve source-model pairs. Second, no paradigm dominates: the agent wins 10 of 16 cells, by up to 26 points over the training-free pipelines on the ICEWS graphs over the training-free pipelines on the ICEWS graphs, yet on GDELT, the largest and most retrieval-hostile source, TimeR^4's fine-tuned retriever takes three cells and In-Context RALM the temporal cell, with Temp-R1 falling to third or lower as its share of unanswered test questions climbs to 27.1%. Third, recursive question decomposition is a measurable liability: RTQA trails its own ablation In-Context RALM in all 16 cells, an effect Section 7.3 traces to the certified multi-fact structure of the questions. Fourth, the ranking depends on the scoring policy: under a uniform strict exact-match view the winner changes in 6 of 16 cells, always in favor of Temp-R1, because the official lenient protocols of the other systems credit answers that merely contain a gold string; lenient-only hits reach 29.33% of all answers for In-Context RALM, 21.60% for RTQA, and 33.27% for TimeR^4, against 1.32% for Temp-R1 and at most 1.14% for the closed-book group. We release both views (Appendix C.4) and treat metric sensitivity as a finding rather than a footnote.

7.3 Error Analysis

Because every question carries its logical form, a standalone protocol re-executes each form with the generator under a fixed set of relaxations and attributes every error to the constraint axes the predicted answer violates: a prediction is checked against the full answer set, then against the three single-axis relaxations that drop the temporal, geographical, or structural constraint, then against the views that keep only one axis (Appendix C.1). A prediction first accepted when the temporal constraint is dropped is a T violation, and symmetrically for G and S; predictions accepted only under joint relaxations are multi-axis; predictions of the wrong answer type or accepted under no relaxation are off-target; empty or unparseable outputs count as no answer. Over all 392,360 pooled (question, system) decisions, only 0.07% are own-metric hits the generator rejects and 0.22% the converse, so the attribution rests on the same ground truth as the official scores.

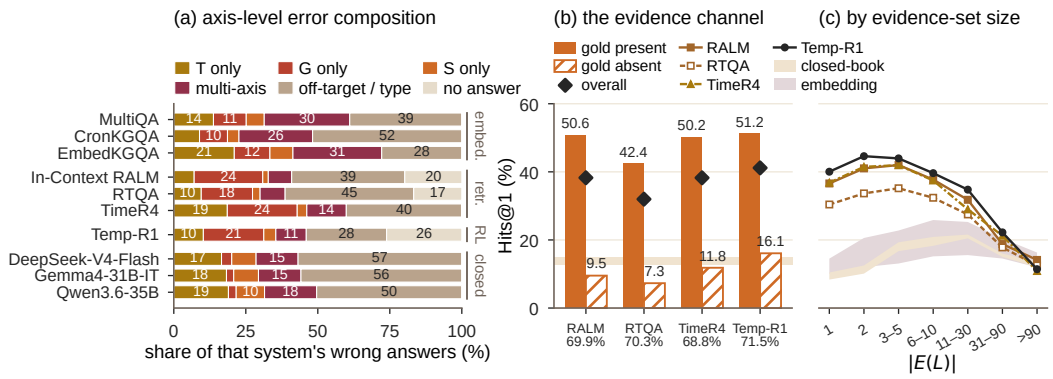

Figure 2: Where the errors are, pooled over all 16 cells. (a) Each bar partitions one system's wrong answers by the first relaxation of the logical form that accepts the prediction: one violated axis (T, G, or S), two or more, none (off-target or wrong type), or no parsable answer. (b) Hits@1 of each evidence-using system when a gold answer was in its observed context and when it was not, with its overall score and, below its name, its gold-in-context rate h. (c) Pooled Hits@1 of the same four systems against the certified evidence-set size $|E(L)|$; in both panels the bands span the three closed-book models and, in (c), the three embedding readers.

Errors concentrate on the certified axes, and differently per paradigm. Figure 2(a) shows that for every evidence-using system the geographical axis is the largest single-axis error source: pure G violations account for 17.80% to 24.27% of the wrong answers of In-Context RALM, RTQA, TimeR[4], and Temp-R1, answers that satisfy the structural and temporal constraints yet name an actor from the wrong region, border, distance band, or direction. The closed-book models show the complementary signature, a G share of 2.52% to 3.55% with off-target answers above 50%: they rarely get close enough to the constraints for one axis to decide. The embedding readers sit in between, with 25% to 31% of errors violating two or more axes at once, the signature of scoring functions that never represent the composition explicitly. That the dominant single-axis failure is geographical, precisely the axis this benchmark certifies, indicates that the questions measure what they are designed to measure rather than a proxy for it.

The evidence channel is load-bearing, and reasoning over it is the bottleneck. Figure 2(b) splits each evidence-using system by whether a gold answer actually appeared in its observed context. The closed-book floor is low and flat, 0.1264 to 0.1486 pooled, and the best closed-book cell (0.2199) lies below the worst cell of any retrieval-augmented system, so priors do not answer these questions at any useful rate. Retrieval is not the primary residual bottleneck on the ICEWS graphs: pooled gold-in-context rates sit between 68.84% and 71.50% (48% to 62% per cell on GDELT), and accuracy on retrieval misses falls to between 7.30% and 16.08%. The dominant loss is post-retrieval: even when the gold answer is present in context, the best system converts only 51.16% of such questions, so roughly half of the certified reasoning demand survives ideal evidence access. The closed-book group's few wins concentrate where guessing is cheap: its pooled region_choice accuracy of 0.6281 to 0.6481 barely clears the 0.5 floor of a two-option type. Figure 2(c) adds the certified evidence demand: the four evidence-using systems peak at three to five supporting facts and decay beyond ten (Spearman -0.75 to -0.89 against the bucket index), falling to 0.1087 to 0.1413 above ninety facts, inside the closed-book band, whereas the embedding readers and the closed-book models stay flat because they never ingest the evidence; the programmatic bands of Definition 4 instead order those non-retrieving systems (Spearman 0.80 to 1.00), so the two difficulty notions are complementary rather than redundant (Appendices C.3 and C.7).

Question decomposition is a measurable liability on entangled questions. The RALM and RTQA columns form a controlled experiment, and the outcome is one-sided: In-Context RALM is better in all 16 cells (exact McNemar $p \leq 3.59 \times 10^{-3}$ per cell, $p = 1.76 \times 10^{-164}$ pooled; Appendix C.5). Figure 3(a, b) locates the damage: on questions RTQA leaves undecomposed the gap is 3.13 points, but it grows to 19.50 points on 2-to-3-node trees and 37.60 points on 4-to-6-node trees, and on 38.80% of the 2-to-3-node misses a sub-question node had already produced a gold

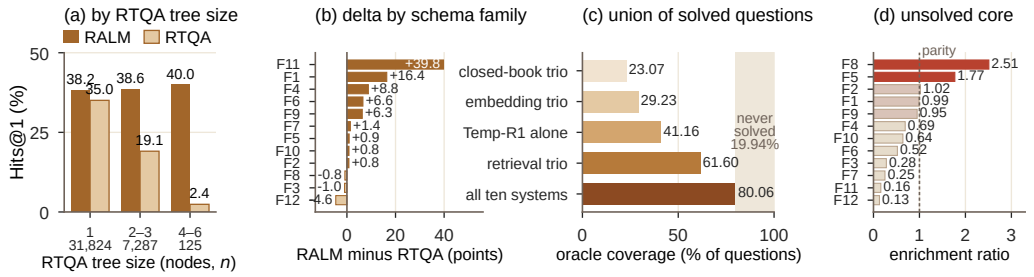

Figure 3: Decomposition forensics and headroom. (a, b) The controlled pair In-Context RALM against RTQA, which share corpus, retriever, reader, prompt, and metric: Hits@1 by the size of RTQA's decomposition tree, and the accuracy delta by schema family, positive bars favoring the ablation without decomposition. (c) Fraction of the 39,236 pooled test questions solved by at least one member of each group; the shaded remainder is solved by no system. (d) Enrichment of each schema family within that never-solved core.

answer that aggregation then discarded. The family profile matches the mechanism: the largest deltas arise for regional precedence (F11, $+39.75$ points) and regional ordinal entity (F1, $+16.40$), whose certified constraints must be evaluated jointly, while decomposition helps only for border counting (F12, -4.59) and two-hop chains (F8, -0.84), the shapes RTQA targets. Both systems degrade as certified entanglement grows, the gap narrowing from 9.03 to 3.78 points from the lowest to the highest tercile. A benchmark property, not a solver accident, does the work: the sub-questions of an entangled question are individually under-determined, so answering them one at a time compounds error instead of removing it, whatever the solver.

The paradigms are complementary, and a fifth of the test questions defeats all of them. Cross-paradigm agreement is low: Cohen's κ between systems of different paradigms never exceeds 0.184, and the strongest single system contributes 2,662 uniquely solved questions with TimeR^4 adding 2,148. Figure 3(c, d) aggregates this complementarity: the union of all ten systems solves 80.06% of pooled questions, so an oracle over paradigms would add almost 39 points over the best single system, yet 19.94% of questions defeat every system evaluated. This never-solved core is enriched 2.51-fold in two-hop chains (F8) and 1.77-fold in nearest-neighbor selection (F5), and 2.59-fold among questions requiring more than 90 supporting facts, while the families where parametric priors or enumeration suffice (F3, F7, F11, F12) are depleted by factors of 4 to 8. The chain family F8 also produces the panel's most instructive inversion: the retrieval pipelines collapse to 0.0380 and 0.0463 there because a single retrieval anchored on the first hop cannot surface second-hop evidence, the agent recovers to 0.2928 through iterated search, and the closed-book models, at 0.1044 to 0.1510, outscore the retrieval pipelines with no evidence at all, so the core marks a capability boundary rather than an artifact of one architecture or one training recipe.

8 Conclusion

GeoTKGQA turns the latent geography of event TKGs into a certified third reasoning axis: a geographical knowledge layer, twelve operator compositions, symbolic execution, and tri-axial certification yield 67,565 provably entangled questions over four sources, with capacity estimates that make the released size a controlled choice. Ten baselines from four paradigms show that the benchmark is unsaturated and discriminative: the closed-book floor is low, errors concentrate on the certified geographical axis, question decomposition fails because the questions are entangled, no paradigm dominates, and a never-solved core of 19.94%, enriched in chains and evidence-heavy questions, marks the open problem of geographical reasoning. Its limits point to next steps: grounding stops at the country level and relies on the naming conventions of ICEWS and CAMEO, the benchmark inherits the errors of its sources, and the certificate guarantees necessity rather than naturalness, so the surface strings stay templated and a paraphrase layer is left to future work. Finer-grained gazetteers, interval-timestamped facts, and solvers that plan over the three axes jointly are the natural extensions, and the never-solved core is their test to be evaluated and compared.

AI USE STATEMENT

In this work, we used Claude Opus 5 (Anthropic, 2026b) and Claude Fable 5.1 (Anthropic, 2026a) as generative AI tools to correct grammar and improve readability. We have not used generative AI tools to propose hypotheses, design the generator or the experiments, or implement the generator or the baselines; translation and qualitative analysis are not applicable to this work. Additionally, we used generative AI tools to reorganize the released code into modules and document it, and to re-serialize the released JSON files to ASCII without changing their content. We have reviewed all AI-assisted work: every modified paragraph was checked against the ideas or results it describes, and the reorganized code was verified to be a line-for-line rearrangement of the code that produced the released benchmarks. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

GeoTKGQA derives its benchmarks from publicly released event datasets, three ICEWS releases and a GDELT slice, that record interactions between political actors as reported in news media. The work involves no human subjects and no annotators, collects no personal data beyond the public actor names already present in those sources, and releases only derived questions, logical forms, gold answers, and splits. Grounding assigns a country to an actor name as that name appears in the source; it is a statement about the naming convention of the source, not about any individual. The benchmark measures reasoning over political event data and could in principle be used to evaluate systems deployed for political analysis, so we release it together with the certificates, the resolved registry, and the attribution protocol, so that any system scored on it can be audited against the same evidence the generator used. The authors declare no conflicts of interest.

REPRODUCIBILITY STATEMENT

Sections 3 to 6 define every object of the generation pipeline, Definition 3 and Equation 1 state the certificate that every released question passes, and Appendix A documents the execution engine and the complete hyperparameter registry whose resolved snapshot ships with the release. Appendix B gives the configuration, faithfulness-preserving adaptations, and hardware of every baseline, Appendix C defines the attribution protocol and reports the statistics behind the figures, and Appendix D collects every symbol. The anonymous repository at `https://anonymous.4open.science/r/GeoTKGQA-6FB8/` contains the generator with its registry snapshot, the four released benchmarks with logical forms, certificates, difficulty bands, supplementary sets, and splits, the reproduction code of all ten baselines, and the analysis protocol that recomputes every number in the paper from the prediction files those scripts produce. Generation is deterministic given the registry and the seed, therefore a rerun reproduces the released benchmarks.

REFERENCES

Dirk Ahlers. Assessment of the accuracy of geonames gazetteer data. In *Proceedings of the 7th workshop on geographic information retrieval*, pp. 74–81, 2013.

James F Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26 (11):832–843, 1983.

Anthropic. Claude: Fable 5.1 and mythos 5.1, 2026a. URL `https://www.anthropic.com/claude-fable-and-mythos-5-1`. Accessed: 2026-09-22.

Anthropic. Introducing claude opus 5, 2026b. URL `https://www.anthropic.com/news/claude-opus-5`. Accessed: 2026-09-22.

Prabin Bhandari, Antonios Anastasopoulos, and Dieter Pfoser. Are large language models geospatially knowledgeable? In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, pp. 1–4, 2023.

Elizabeth Boschee, Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, James Starz, and Michael Ward. ICEWS Coded Event Data, 2015. URL `https://doi.org/10.7910/DVN/28075`.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Anne Chao. Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of statistics*, pp. 265–270, 1984.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the association for computational linguistics: ACL 2024*, pp. 2318–2335, 2024.

Wenhu Chen, Xinyi Wang, and William Yang Wang. A dataset for answering time-sensitive questions. *arXiv preprint arXiv:2108.06314*, 2021.

Ziyang Chen, Jinzhi Liao, and Xiang Zhao. Multi-granularity temporal question answering over knowledge graphs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11378–11392, 2023.

J Cohen. A co-efficient of agreement for norminal scales. *Educ Psychol Meas*, 20:111–116, 1960.

Alberto García-Durán, Sebastijan Dumančić, and Mathias Niepert. Learning sequence encoders for temporal knowledge graph completion. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 4816–4821, 2018.

Deborah J Gerner, Philip A Schrodt, Omür Yilmaz, and Rajaa Abu-Jabr. Conflict and mediation event observations (cameo): A new event data framework for the analysis of foreign policy interactions. *International Studies Association, New Orleans*, 35, 2002.

Zhaoyan Gong, Juan Li, Zhiqiang Liu, Lei Liang, Huajun Chen, and Wen Zhang. Rtqa: Recursive thinking for complex temporal knowledge graph question answering with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9864–9881, 2025.

Zhaoyan Gong, Zhiqiang Liu, Songze Li, Xiaoke Guo, Yuanxiang Liu, Xinle Deng, Zhizhen Liu, Lei Liang, Huajun Chen, and Wen Zhang. Temp-r1: A unified autonomous agent for complex temporal kgqa via reverse curriculum reinforcement learning. *arXiv preprint arXiv:2601.18296*, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

ISO ISO. 3166-1: Codes for the representation of names of countries and their subdivisions–part 1: Country codes. *ISO: Geneva, Switzerland*, 2006.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. Tempquestions: A benchmark for temporal question answering. In *Companion Proceedings of the The Web Conference 2018*, pp. 1057–1062, 2018.

Zhen Jia, Soumajit Pramanik, Rishiraj Saha Roy, and Gerhard Weikum. Complex temporal question answering on knowledge graphs. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pp. 792–802, 2021.

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

Woojeong Jin, Meng Qu, Xisen Jin, and Xiang Ren. Recurrent event network: Autoregressive structure inferenceover temporal knowledge graphs. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 6669–6683, 2020.

Divyansh Kaushik and Zachary C Lipton. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 5010–5015, 2018.

Sergios-Anestis Kefalidis, Dharmen Punjani, Eleni Tsalapati, Konstantinos Plas, Mariangela Pollali, Michail Mitsios, Myrto Tsokanaridou, Manolis Koubarakis, and Pierre Maret. Benchmarking geospatial question answering engines using the dataset geoquestions1089. In *International semantic web conference*, pp. 266–284. Springer, 2023.

Timothée Lacroix, Guillaume Obozinski, and Nicolas Usunier. Tensor decompositions for temporal knowledge base completion. *arXiv preprint arXiv:2004.04926*, 2020.

Kalev Leetaru and Philip A Schrodt. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, volume 2, pp. 1–49. Citeseer, 2013.

Frederic M Lord. *Applications of item response theory to practical testing problems*. Routledge, 2012.

Gengchen Mai, Krzysztof Janowicz, Rui Zhu, Ling Cai, and Ni Lao. Geographic question answering: challenges, uniqueness, classification, and future directions. *AGILE: GIScience series*, 2:8, 2021.

Costas Mavromatis, Prasanna Lakkur Subramanyam, Vassilis N Ioannidis, Adesoji Adeshina, Phillip R Howard, Tetiana Grinberg, Nagib Hakim, and George Karypis. Tempoqr: temporal question reasoning over knowledge graphs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 5825–5833, 2022.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Iman Mirzadeh, Keivan Alizadeh-Vahid, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. In *International Conference on Learning Representations*, volume 2025, pp. 94743–94765, 2025.

United Nations. Statistical Office. *Standard country or area codes for statistical use*. Number 49. UN, 1982.

Dharmen Punjani, Kuldeep Singh, Andreas Both, Manolis Koubarakis, Iosif Angelidis, Konstantina Bereta, Themis Beris, Dimitris Bilidas, Theofilos Ioannidis, Nikolaos Karalis, et al. Template-based question answering over linked geospatial data. In *Proceedings of the 12th workshop on geographic information retrieval*, pp. 1–10, 2018.

Xinying Qian, Ying Zhang, Yu Zhao, Baohang Zhou, Xuhui Sui, Li Zhang, and Kehui Song. Timer4: Time-aware retrieval-augmented large language models for temporal knowledge graph question answering. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 6942–6952, 2024.

Xinying Qian, Ying Zhang, Yu Zhao, Baohang Zhou, Xuhui Sui, and Xiaojie Yuan. Plan of knowledge: Retrieval-augmented large language models for temporal knowledge graph question answering. *arXiv preprint arXiv:2511.04072*, 2025.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 2383–2392, 2016.

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

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Apoorv Saxena, Aditay Tripathi, and Partha Talukdar. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4498–4507, 2020.

Apoorv Saxena, Soumen Chakrabarti, and Partha Talukdar. Question answering over temporal knowledge graphs. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6663–6676, 2021.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L Hamilton. Clutrr: A diagnostic benchmark for inductive reasoning from text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4506–4515, 2019.

Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pp. 697–706, 2007.

Qiang Sun, Sirui Li, Du Huynh, Mark Reynolds, and Wei Liu. Timelinekgqa: A comprehensive question-answer pair generator for temporal knowledge graphs. In *Companion Proceedings of the ACM on Web Conference 2025*, pp. 797–800, 2025.

Alon Talmor and Jonathan Berant. The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 641–651, 2018.

Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

Yuyu Zhang, Hanjun Dai, Zornitsa Kozareva, Alexander Smola, and Le Song. Variational reasoning for question answering with knowledge graph. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Table A1: The twelve schema families: identifier, name, gold answer type, the geographical operator composed with the structural and temporal constraints, and a sketch of the fixed-grammar surface template, with braces marking slots filled from the geo-augmented TKG.

ID	Name	Answer	Geo. operator	Template sketch
F1	Regional ordinal entity	entity	containment	Within {region}, who was the {first/last} to {rel} {object} {window}?
F2	Regional ordinal time	time	containment	When did {subject} {first/last} {rel} an actor in {region}?
F3	Adjacency	location	adjacency	{Window}, {subject} {rel} an actor from which country bordering {country}?
F4	Directional extremal	location	direction + adjacency	Among countries bordering {country} where {subject} {rel} {window}, which lies furthest {direction}?
F5	Nearest neighbor	location	metric proximity	After {date}, {subject} {rel} an actor from which country nearest to {country}?
F6	Distance threshold	entity	distance threshold	{Window}, who from within {D} km of {country} {rel} {object}?
F7	Count and compare	number / region_choice	containment × counting	{Window}, actors from how many countries in {region} did {subject} {rel}? / Did {subject} {rel} actors from more countries in {region$_1$} or {region$_2$}?
F8	Two-hop temporal chain	entity	containment	Who from {region} {rel$_2$} an actor after that actor {rel$_1$} {object} {window}?
F9	Directional field filter	entity	directional half-plane	{Window}, who from a country lying further {direction} than {country} {rel} {object}?
F10	Co-subregion	location	roll-up + co-membership	{Window}, {subject} {rel} an actor from which country in the same subregion as {country}?
F11	Regional precedence	region_choice	containment × precedence	{Window}, did {subject} {rel} an actor in {region$_1$} or in {region$_2$} first?
F12	Border count	number	adjacency × counting	{Window}, actors from how many countries bordering {country} did {subject} {rel}?

A Pipeline Details

A.1 The Twelve Schema Families

Table A1 lists the twelve schema families of Section 5.1, giving for each one its answer type, the geographical operator it composes, and the surface template from which its questions are rendered.

A.2 Execution Engine

To keep symbolic execution tractable on graphs with millions of facts, all indices of the engine of Section 5.2 follow compressed sparse row storage: fact arrays are lexicographically sorted by composite keys, (object, relation) for the families whose anchor fills the object role and (subject, relation) otherwise, with contiguous slices per key, so Stage 1 is a binary search followed by a slice. Group-level statistics used for eligibility screening, the fact count, the number of distinct days, and the number of distinct grounded counterpart countries per group, are computed in one vectorized pass over the sorted arrays, and expensive geographical artifacts such as distance rows, directional half-plane sets, and co-membership arrays are computed lazily and cached across the questions that share an anchor. Every relaxation used by certification (Section 5.3) and by the attribution protocol (Appendix C.1) re-runs the same engine with one constraint removed, so certified answers and attributed errors are produced by identical code paths.

A.3 Release-Control Registry

Every knob of the generator lives in a single hyperparameter registry, any entry can be overridden on the command line, and the resolved snapshot is written into the release summary of every merge for provenance. The volume knobs are the per-family quota Q, the per-family sampling budget B, the shard count realized through seed offsets, and the family subset. The balance and anti-shortcut knobs are per-anchor, per-answer, per-geographic-constant, and per-relation caps χ_{anc}, χ_{ans}, χ_{geo}, and χ_{rel}, a hub-degree cap $D_{\max}$ on anchors, an answer-set size cap $A_{\max}$, a diagnostic quota fraction, and an inline leakage policy rejecting any question whose surface string contains one of its answers, with tuple-valued and numeric answer types exempted. The content-shape knobs comprise the per-source window and granularity registries, the distance ladder $\mathcal{D}$ of F6, the family-internal mixtures and retry budgets, the grounding threshold θ, the split fractions of Section 5.4, the supplementary-set quotas, and the master seed from which every random stream is derived. Raising B leaves $\widehat{N^*}$ unchanged but tightens its estimate, whereas tightening the caps reshapes the released distribution without touching

the ceiling, so the three quantities of Equation 4 respond to disjoint sets of knobs. The default registry is the one whose resolved snapshot ships with the release; all counts of Table 1 refer to that snapshot and to no tuned variant of it, so that every released size is exactly reproducible.

B Reproduction and Compute Details

All GPU training, covering the three embedding readers, the TimeR^4 retriever fine-tuning, and Temp-R1, run on a single 96 GB NVIDIA Blackwell workstation GPU with 176 GB of system memory under Google Colab Pro Plus sessions capped at 24 hours; the API-backed pipelines (In-Context RALM, RTQA, and the closed-book DeepSeek-V4-Flash arm) run on MacBookPro with 18 GB M3 Pro Apple Silicon chip, and the two open-weight closed-book models are served in bfloat16 by 3×48 GB NVIDIA L40S GPUs on a Linux Ubuntu 24.04.5 LTS server. Every baseline is a single-file script that reads the benchmark release without modifying it, every training and evaluation step supports checkpointed pause and resume, and each (source, split, step) writes exactly one log file, all of which are released together with the per-question predictions and the analysis script.

MultiQA. TComplEx embeddings are trained per source over the full quadruple set, and the MultiQA question-answering model of Chen et al. (2023) is then trained once per (source, split) cell on that split's training questions. Evaluation ranks the shared closed answer vocabulary, covering entities, timestamps, countries, numbers, and regions, and reports rank-based Hits@1, the system's official metric, with month-level and year-level gold times expanded to their day identifiers so that a prediction inside the gold period counts as a hit.

CronKGQA. Our port follows the reference implementation of Saxena et al. (2021): a frozen DistilBERT question encoding is projected by a single linear layer into two parallel heads, without any activation function, yielding one relation vector for the time scorer and one for the entity scorer, both evaluated with the TComplEx scoring kernels over the same per-source embeddings as MultiQA. Training uses Adam with a cross-entropy loss on one uniformly resampled gold answer per example and selects the checkpoint by development Hits@1. Relative to MultiQA the model has no multi-granularity aggregation, so each question contributes exactly one timestamp embedding, which is the capability difference the temporal splits expose. The evaluation protocol, answer vocabulary, and gold expansion are byte-identical to MultiQA's.

EmbedKGQA. Following Saxena et al. (2020) with the temporal-benchmark setup of the CronKGQA reference code, the graph is collapsed to static (head, relation, tail) triples on which ComplEx embeddings are trained with reciprocal relations and one-against-all binary cross-entropy with label smoothing; the question-answering model projects the frozen language-model encoding through a single linear layer and scores atemporal ComplEx against a concatenated answer table whose timestamp block is randomly initialized and left trainable, so the question's time window never enters the scoring function, making this the deliberately time-blind member of the panel. Training uses Adam with exponential learning-rate decay, a Kullback-Leibler loss against the normalized multi-hot gold vector, and patience-based early stopping; the evaluation protocol, answer vocabulary, and gold expansion are again identical to MultiQA's.

In-Context RALM. The single-shot retrieve-then-read instantiation of Ram et al. (2023) is realized as the "without decomposer" ablation of our RTQA port, so the two systems form a controlled pair: it keeps the verbalized fact corpus (one sentence per quadruple with display names and ISO dates), the BGE-M3 dense retriever with exact inner-product search and a budget of 50 facts, the reader prompt and answer extraction copied byte for byte from the RTQA port, greedy decoding, and the DS-V4-Flash reader with extended reasoning disabled, while removing the question decomposer, the recursive solver, and the answer aggregator, so every test question costs exactly one retrieval and one reader call. Corpus preparations and fact embeddings are shared with RTQA per source.

RTQA. We port the official RTQA pipeline of Gong et al. (2025) with byte-identical prompt templates and its released aggregation logic: a temporal question decomposer builds a hierarchical tree, a recursive solver answers each node bottom-up with facts retrieved by BGE-M3 over the verbalized corpus (50 facts per call), and a multi-source aggregator arbitrates candidate answers.

The solver LLM is DeepSeek-V4-Flash with extended reasoning disabled, called with the original decoding settings; all API responses are cached in a request ledger for exact replay.

TimeR[4]. Our reproduction follows the retrieve, rewrite, retrieve, rerank scheme of Qian et al. (2024) at the tier that fine-tunes the retriever while keeping the reader frozen: per (source, split) cell, an all-mpnet-base-v2 sentence encoder (Reimers & Gurevych, 2019) is fine-tuned contrastively on time-aware positives and negatives mined from that cell's training questions, a first-round retrieval grounds the question's temporal expression, the frozen LLM rewrites the question accordingly, the fine-tuned encoder retrieves again, and the context is reranked by temporal relevance before the frozen DeepSeek-V4-Flash reader answers. Scoring uses the system's official Hit metric, containment of a normalized gold string in the normalized answer; base-corpus embeddings are cached per source and all repairs needed to make the released reference code runnable are disclosed as numbered notes in the reproduction script, together with the reason for each repair.

Temp-R1. The backbone is Llama-3.1-8B-Instruct with e5-base-v2 retrieval over the same verbalized corpus (top 30 facts per query, at most 3 search turns, observations truncated to 2,400 tokens). Supervised cold-start uses teacher trajectories generated by DeepSeek-V4-Flash under the original prompt recipe, filtered to correct answers, and trains all parameters for 2 epochs. The two GRPO stages follow the reverse curriculum of the original (stage 1 on the two hardest difficulty bands, stage 2 on all training questions) with unchanged per-update semantics (group sizes 5 and 8, learning rate, KL and clipping settings, and reward function taken from the released code). Two adaptations keep one full cell within a 24-hour window and are disclosed as deviations: reinforcement learning updates LoRA adapters (Hu et al., 2021) of rank 32 on all seven projection matrices, merged back into the base model after each stage, and the sampling budget is reduced to 40 steps of batch 30 in stage 1 and 100 steps in stage 2 (the original uses 80 by 60 and 500). Evaluation decodes greedily with the official lengths and computes the official exact-match Hits@1; the stricter equality variant is also stored with every prediction. All runs use seed 42.

Closed-book control group. Each model receives a fixed task instruction specifying the required output format for the five answer types, followed by eight exemplars selected deterministically from the cell's training split under the quota of two entity, two location, two time, one number, and one region_choice questions, presented as bare question and answer pairs with no supporting facts, and then the test question. Decoding is greedy with at most 64 new tokens; extended reasoning modes are disabled at the API or serving layer and any residual reasoning markup is stripped before extraction. DeepSeek-V4-Flash is called through its API; Gemma4-31B-IT and Qwen3.6-35B-A3B run on a local inference server in bfloat16 with an 8,192-token context. Scoring follows the RTQA official lenient protocol, with the strict variant stored alongside, and all requests are ledgered so that repeated runs replay cached responses at no cost and with identical outputs.

C Error Analysis Protocol and Additional Results

C.1 Axis-Level Attribution

For every (question, system) pair the protocol re-executes the stored logical form with the generator under a fixed set of relaxations: the full form, the three single-axis relaxations that drop the temporal, geographical, or structural constraint, and the three complementary views that keep only one axis. A prediction is matched against each resulting answer set using the system's own normalization, by surface equality or containment, with token-level matching for counts, and the categories are assigned in a fixed check order (full, then dropped-T, dropped-G, dropped-S, then the keep-one-axis views): a prediction first accepted when the temporal constraint is dropped is a *T violation* (right actors and geography, wrong window, ordinal, or pivot), and symmetrically for *G* and *S*; predictions accepted only under joint relaxations are *multi-axis*; predictions of the wrong answer type or accepted under no relaxation are *off-target*; empty or unparseable outputs are *no answer*. Over all 392,360 pooled decisions the generator view and the official metrics disagree on only 0.29% (0.07% own-metric hits the generator rejects, 0.22% generator-accepted answers the official metric misses), so the attribution rests on the same ground truth as the headline scores. Figure A1 complements the axis view of Figure 2(a) with the full family resolution.

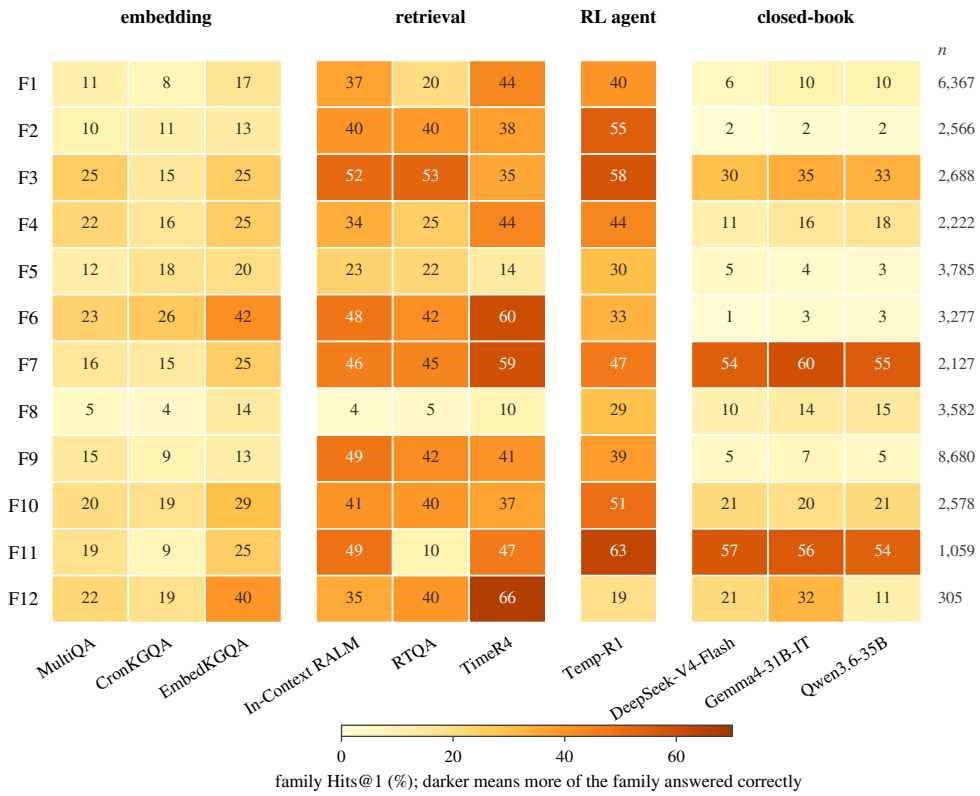

Figure A1: Per-family Hits@1 (%) of all ten systems, pooled over the 16 cells; the right margin gives the pooled family sizes. Recomposing each system's overall accuracy from these cells reproduces its pooled headline score exactly. The chain family F8 exposes the retrieval bottleneck (In-Context RALM 4%, RTQA 5%, against Temp-R1 29%), while the closed-book columns peak exactly on the two-option and enumeration families (F7, F11, F3).

The family resolution also quantifies the parametric-prior surface that the control group can exploit: the questions solved by all three closed-book models concentrate in count-and-compare (38.27% of F7) and regional precedence (34.84% of F11), the two families with tuple-valued or binary answers, and their pooled region_choice accuracy of 0.6281 to 0.6481 sits close to the 0.5 floor of a two-option type; everywhere else the all-three-solved share is at most 10.60%. The certification quotas anticipate this by capping the two families at 5.42% and 2.70% of the pooled test questions.

C.2 Oracle Coverage and the Never-Solved Core

Complementing Figure 3(c, d), the coverage ladder rises from 23.07% (closed-book trio) through 29.23% (embedding trio) and 61.60% (retrieval trio) to 80.06% for all ten systems, and the remaining 19.94% core is enriched in the two-hop chain and nearest-neighbor families and in the highest evidence-demand buckets (2.59-fold beyond 90 facts, 1.97-fold at 31 to 90), while the L1_Easy band is over-represented (1.48-fold) because its single-fact questions pose needle-in-a-haystack retrieval targets among up to 2.3 million facts. Per-cell oracle unions range from 68.13% to 88.62%.

C.3 Evidence Channel and Shift Decompositions

A gold answer counts as *in context* for a run if any gold string, after the system's own answer normalization, is a substring of the normalized observed context: for In-Context RALM the retrieved fact block, for RTQA the union of all retrieved fact blocks across the question's tree, for TimeR^4 the reranked context recorded in its ledger, and for Temp-R1 the concatenation of all observation blocks in the trajectory; the closed-book group observes no context by construction. Writing h

Table A2: Evidence channel pooled over all 16 subsets: gold-in-context rate h, conditional accuracies a^{hit} and a^{miss}, and overall accuracy. Each baseline scored by its own official protocol.

System	gold-in-ctx	acc | hit	acc | miss	acc
In-Context RALM	69.89%	50.64%	9.52%	38.26%
RTQA	70.34%	42.39%	7.30%	31.98%
TimeR[4]	68.84%	50.21%	11.83%	38.25%
Temp-R1	71.50%	51.16%	16.08%	41.16%

Table A3: Exact decomposition (Equation 5) of the geographical zero-shot accuracy delta (Δ = geo $-$ std, Hits@1 as fractions) into the retrieval-mix, hit-accuracy, and miss-accuracy terms. The last column restricts the delta to questions whose logical form names a region.

Method	Source	std	geo	Δ	Mix	Hit	Miss	Δ region-bearing
RTQA	ICEWS14	0.3037	0.2669	-0.0368	-0.0139	-0.0304	$+0.0075$	$+0.0108$
	ICEWS18	0.3111	0.4111	$+0.1000$	$+0.0274$	$+0.0754$	-0.0028	-0.0047
	ICEWS05-15	0.3223	0.3714	$+0.0491$	$+0.0177$	$+0.0391$	-0.0077	-0.0856
	GDELT	0.2704	0.2511	-0.0192	-0.0169	$+0.0065$	-0.0088	$+0.0006$
Temp-R1	ICEWS14	0.5133	0.4475	-0.0658	-0.0189	-0.0546	$+0.0078$	-0.0649
	ICEWS18	0.4908	0.4124	-0.0784	$+0.0182$	-0.0748	-0.0218	-0.0532
	ICEWS05-15	0.4843	0.5004	$+0.0161$	$+0.0034$	$+0.0358$	-0.0230	-0.2478
	GDELT	0.3090	0.2598	-0.0491	-0.0075	-0.0246	-0.0170	-0.0072

for the resulting gold-in-context rate, a^{hit} for accuracy conditional on the gold being in context, and a^{miss} for accuracy conditional on it being absent, overall accuracy satisfies the exact identity $\text{acc} = h\,a^{\text{hit}} + (1-h)\,a^{\text{miss}}$, and Table A2 reports the pooled values behind Figure 2(b). As gold region names rarely appear verbatim in fact text, containment-based context matching over-credits retrieval for the region_choice answer type, a caveat that carries over to every context-based statistic we report for that answer type, here and in the tables below.

Any two accuracy figures that share this identity admit an exact three-term decomposition. For a run x and a reference run r,

$$\text{acc}_x - \text{acc}_r = \underbrace{(h_x - h_r)\big(a_r^{\text{hit}} - a_r^{\text{miss}}\big)}_{\text{Mix}} + \underbrace{h_x\big(a_x^{\text{hit}} - a_r^{\text{hit}}\big)}_{\text{Hit}} + \underbrace{(1-h_x)\big(a_x^{\text{miss}} - a_r^{\text{miss}}\big)}_{\text{Miss}}, \quad (5)$$

where the retrieval-mix term Mix charges the difference in gold-in-context rates $h_x - h_r$ at the reference run's conditional-accuracy spread, the hit term Hit charges the change in accuracy on retrieved questions, and the miss term Miss the change on unretrieved ones; the retrieval share of a delta is $|\text{Mix}|$ over the sum of the three absolute terms. Applied across splits within one system (x the shift split, r the standard split), Tables A3 and A4 decompose every distribution-shift delta of RTQA and Temp-R1. Where genuine geographical degradation occurs (ICEWS14 for both systems, ICEWS18 for Temp-R1, GDELT for both), the dominant component is the hit term, so held-out regions primarily stress reasoning on retrieved evidence rather than retrieval itself (retrieval shares of 15% to 27%), with GDELT-RTQA the one retrieval-driven case (52.5%). The two apparent improvements reverse under composition control: restricting to region-bearing logical forms turns RTQA's $+0.1000$ on ICEWS18 and $+0.0491$ on ICEWS05-15 into -0.0047 and -0.0856, and Temp-R1's $+0.0161$ on ICEWS05-15 into -0.2478, as the geographical zero-shot test partitions contain far fewer instances of the hardest families (F8 drops from 501 to 38 questions on ICEWS18). On the temporal and compositional splits the same decomposition shows the training-free RTQA improving on every shift cell while the trained agent declines on six of eight, with its GDELT retrieval-mix share rising to 36.7% and 42.2%, the retrieval signature of learned query generation under distribution shift. Questions whose certified evidence pool exceeds the retrieval capacity a run actually deployed collapse to 6.5% to 36% Hits@1 with context hit rates of at most 52%, confirming that the evidence-size decay of Figure 2(c) is a budget bound rather than a difficulty artifact; its two rightmost buckets pool 612 and 184 questions.

Table A4: Decomposition of the temporal (tmp) and compositional (cmp) accuracy deltas against the standard split, with the identity and columns of Table A3. RTQA, which trains nothing, improves on every shift cell, while the trained agent declines on six of eight; and on GDELT its retrieval-mix share rises to 36.7% and 42.2%, the signature of learned query generation under shift.

Split	Method	Source	std	shift	Δ	Mix	Hit	Miss	Δ region-bearing
tmp	RTQA	ICEWS14	0.3037	0.3906	+0.0868	+0.0210	+0.0602	+0.0056	−0.0326
		ICEWS18	0.3111	0.3606	+0.0495	+0.0071	+0.0365	+0.0059	−0.0136
		ICEWS05-15	0.3223	0.3645	+0.0422	+0.0085	+0.0290	+0.0047	−0.0802
		GDELT	0.2704	0.3333	+0.0630	+0.0104	+0.0431	+0.0095	+0.0212
	Temp-R1	ICEWS14	0.5133	0.4961	−0.0172	+0.0033	−0.0312	+0.0107	−0.0669
		ICEWS18	0.4908	0.4925	+0.0017	+0.0009	−0.0268	+0.0277	+0.0195
		ICEWS05-15	0.4843	0.4638	−0.0205	−0.0005	−0.0283	+0.0084	−0.0656
		GDELT	0.3090	0.2692	−0.0397	−0.0146	−0.0219	−0.0033	+0.0187
cmp	RTQA	ICEWS14	0.3037	0.3359	+0.0322	+0.0283	−0.0173	+0.0212	+0.0154
		ICEWS18	0.3111	0.3297	+0.0186	+0.0071	−0.0222	+0.0336	+0.0633
		ICEWS05-15	0.3223	0.3915	+0.0692	+0.0318	+0.0054	+0.0321	+0.0339
		GDELT	0.2704	0.3052	+0.0349	+0.0056	+0.0003	+0.0290	+0.0941
	Temp-R1	ICEWS14	0.5133	0.5919	+0.0785	+0.0138	+0.0556	+0.0091	+0.0414
		ICEWS18	0.4908	0.4681	−0.0227	−0.0012	−0.0433	+0.0218	+0.0437
		ICEWS05-15	0.4843	0.4673	−0.0169	+0.0140	−0.0434	+0.0125	+0.0264
		GDELT	0.3090	0.2776	−0.0314	−0.0133	−0.0121	−0.0060	+0.0303

Table A5: Uniform strict view of Table 2: Hits@1 under exact normalized string equality for the seven generative systems (the rank-based readers are not applicable); bold marks the best evidence-eligible system per row. Temp-R1 wins every cell under this view, so the winner changes in 6 of the 16 cells relative to the official protocols. Pooled shares of lenient-only hits: In-Context RALM 29.33%, RTQA 21.60%, TimeR^4 33.27%, Temp-R1 1.32%, closed-book 0.36%, 0.12%, and 1.14%.

Source	Split	RALM	RTQA	TimeR^4	Temp-R1	DS-V4F	G4-31B	Q3.6-35B
ICEWS14	std	0.1309	0.1431	0.0429	**0.4858**	0.0805	0.0998	0.0805
	geo	0.1486	0.1407	0.0483	**0.4127**	0.1091	0.1179	0.0961
	tmp	0.1193	0.1270	0.0489	**0.4944**	0.1133	0.1356	0.1064
	cmp	0.0584	0.0784	0.0761	**0.5895**	0.0769	0.0999	0.0515
ICEWS18	std	0.1228	0.1382	0.0410	**0.4737**	0.1606	0.1639	0.1389
	geo	0.1105	0.1490	0.0399	**0.4046**	0.1379	0.1627	0.1288
	tmp	0.1077	0.1231	0.0434	**0.4883**	0.2112	0.2199	0.1945
	cmp	0.0896	0.1063	0.0681	**0.4533**	0.1669	0.1842	0.1306
ICEWS05-15	std	0.1207	0.1288	0.0388	**0.4230**	0.1130	0.1099	0.1025
	geo	0.1196	0.1376	0.0375	**0.4980**	0.0930	0.1126	0.0907
	tmp	0.1414	0.1574	0.0505	**0.4604**	0.1254	0.1397	0.1288
	cmp	0.0743	0.0996	0.0648	**0.4626**	0.0954	0.1359	0.0785
GDELT	std	0.0534	0.0669	0.0411	**0.3081**	0.1010	0.1519	0.1479
	geo	0.0357	0.0412	0.0435	**0.2595**	0.0702	0.0889	0.0976
	tmp	0.0368	0.0599	0.0503	**0.2671**	0.1314	0.1824	0.1796
	cmp	0.0578	0.0736	0.0682	**0.2683**	0.1287	0.1788	0.1823

C.4 Metric Sensitivity

Table A5 reports the uniform strict view, exact equality of the normalized prediction and a gold answer, for the seven generative methods. The official protocols and the strict view tell one consistent story about magnitude but not always about ranking, and the six winner flips, all toward Temp-R1, are concentrated where lenient credit is largest: TimeR^4's official containment hit accepts any answer sentence mentioning a gold string, and the RTQA-family lenient match accepts answers that append dates or qualifiers to a gold string.

C.5 Paired Significance for the Decomposition Ablation

Table A6 gives the per-cell paired comparison behind Figure 3(a, b): the counts of questions only In-Context RALM solves (b) and only RTQA solves (c), with the exact two-sided McNemar probability (McNemar, 1947) on the discordant pairs; Cohen's κ (Cohen, 1960) between the two systems is 0.541 pooled. The advantage of the ablation without decomposition is significant at the 0.001 level in 15 of the 16 cells, and pooling all discordant pairs (5,351 against 2,888) gives $p = 1.76 \times 10^{-164} \ll 0.001$.

Table A6: Per-cell paired comparison of In-Context RALM and RTQA on identical test questions under their shared metric: accuracies, delta, discordant counts (b = only RALM correct, c = only RTQA correct), and the exact two-sided McNemar probability (significant difference if < 0.001).

Source	**Split**	n	RALM	RTQA	Δ	b	c	McNemar p
ICEWS14	std	2,285	0.3514	0.3037	+0.0477	273	164	$2.08 \times 10^{-7} < 0.001$
	geo	2,154	0.3250	0.2669	+0.0580	232	107	$9.63 \times 10^{-12} < 0.001$
	tmp	1,165	0.4618	0.3906	+0.0712	190	107	$1.69 \times 10^{-6} < 0.001$
	cmp	1,301	0.4420	0.3359	+0.1061	268	130	$4.02 \times 10^{-12} < 0.001$
ICEWS18	std	3,973	0.3781	0.3111	+0.0670	552	286	$2.64 \times 10^{-20} < 0.001$
	geo	1,530	0.4477	0.4111	+0.0366	207	151	$3.59 \times 10^{-3} > 0.001$
	tmp	2,396	0.4232	0.3606	+0.0626	355	205	$2.40 \times 10^{-10} < 0.001$
	cmp	3,170	0.4388	0.3297	+0.1091	584	238	$2.40 \times 10^{-34} < 0.001$
ICEWS05-15	std	2,858	0.3810	0.3223	+0.0588	297	129	$2.42 \times 10^{-16} < 0.001$
	geo	1,279	0.4292	0.3714	+0.0579	137	63	$1.79 \times 10^{-7} < 0.001$
	tmp	1,188	0.4259	0.3645	+0.0614	159	86	$3.61 \times 10^{-6} < 0.001$
	cmp	1,898	0.5116	0.3915	+0.1201	342	114	$1.60 \times 10^{-27} < 0.001$
GDELT	std	4,457	0.3094	0.2704	+0.0390	496	322	$1.27 \times 10^{-9} < 0.001$
	geo	3,106	0.2888	0.2511	+0.0377	321	204	$3.72 \times 10^{-7} < 0.001$
	tmp	2,823	0.3900	0.3333	+0.0567	399	239	$2.49 \times 10^{-10} < 0.001$
	cmp	3,653	0.3589	0.3052	+0.0537	539	343	$4.32 \times 10^{-11} < 0.001$

Table A7: 95% bootstrap confidence intervals for Table 2, first five systems.

Source	**Split**	MultiQA	CronKGQA	EmbedKGQA	RALM	RTQA
ICEWS14	std	[0.233, 0.268]	[0.224, 0.260]	[0.291, 0.330]	[0.331, 0.371]	[0.284, 0.323]
	geo	[0.120, 0.149]	[0.113, 0.142]	[0.136, 0.167]	[0.305, 0.344]	[0.248, 0.286]
	tmp	[0.083, 0.118]	[0.025, 0.045]	[0.169, 0.215]	[0.433, 0.491]	[0.361, 0.419]
	cmp	[0.235, 0.283]	[0.191, 0.235]	[0.178, 0.222]	[0.415, 0.469]	[0.310, 0.361]
ICEWS18	std	[0.143, 0.165]	[0.134, 0.156]	[0.194, 0.219]	[0.363, 0.393]	[0.297, 0.325]
	geo	[0.167, 0.207]	[0.128, 0.164]	[0.155, 0.193]	[0.422, 0.472]	[0.386, 0.435]
	tmp	[0.126, 0.154]	[0.004, 0.010]	[0.176, 0.208]	[0.403, 0.442]	[0.341, 0.379]
	cmp	[0.091, 0.112]	[0.059, 0.077]	[0.131, 0.156]	[0.421, 0.456]	[0.313, 0.346]
ICEWS05-15	std	[0.128, 0.154]	[0.125, 0.151]	[0.191, 0.221]	[0.363, 0.399]	[0.304, 0.339]
	geo	[0.140, 0.181]	[0.145, 0.186]	[0.164, 0.207]	[0.403, 0.456]	[0.345, 0.398]
	tmp	[0.063, 0.095]	[0.066, 0.099]	[0.130, 0.171]	[0.398, 0.455]	[0.337, 0.391]
	cmp	[0.144, 0.177]	[0.097, 0.126]	[0.124, 0.155]	[0.489, 0.535]	[0.369, 0.413]
GDELT	std	[0.160, 0.182]	[0.164, 0.186]	[0.226, 0.251]	[0.295, 0.323]	[0.257, 0.283]
	geo	[0.074, 0.094]	[0.110, 0.134]	[0.172, 0.199]	[0.272, 0.304]	[0.236, 0.266]
	tmp	[0.107, 0.131]	[0.058, 0.078]	[0.242, 0.275]	[0.372, 0.408]	[0.316, 0.350]
	cmp	[0.106, 0.127]	[0.081, 0.100]	[0.145, 0.168]	[0.343, 0.374]	[0.290, 0.320]

C.6 Confidence Intervals

Tables A7 and A8 give 95% bootstrap confidence intervals for every cell of Table 2, from binomial resampling of the per-question correctness indicators with 2,000 replicates. All results come from a single seeded run per cell, so these intervals quantify test-set sampling variation, not training variance.

C.7 Difficulty Bands and Split Gaps

Table A9 reports pooled accuracy per programmatic difficulty band with the Spearman correlation between band index and accuracy, quantifying the claim of Section 7.3: the bands order the non-retrieving paradigms monotonically or nearly so, while the retrieval-augmented systems peak at L3_Hard and dip at L4_Expert, whose questions combine heavy operators with the largest certified evidence pools, the regime Figure 2(c) shows they cannot ingest. Table A10 reports all split gaps as accuracy on the standard split minus accuracy on the shift split, in points, so positive values mean the shift hurts; the embedding readers are the most exposed to the geographical zero-shot split on ICEWS14 and to the temporal split everywhere, the retrieval-augmented pipelines often gain on shift splits (negative entries) through the composition effects analyzed in Appendix C.3, and the closed-book group moves little, as expected of systems that never see the partitioned training data.

Table A8: 95% bootstrap confidence intervals for Table 2, last five systems.

Source	Split	TimeR^4	Temp-R1	DS-V4F	G4-31B	Q3.6-35B
ICEWS14	std	[0.319, 0.358]	[0.493, 0.533]	[0.078, 0.102]	[0.092, 0.117]	[0.098, 0.124]
	geo	[0.271, 0.310]	[0.426, 0.468]	[0.100, 0.127]	[0.107, 0.135]	[0.104, 0.132]
	tmp	[0.366, 0.423]	[0.467, 0.524]	[0.104, 0.142]	[0.117, 0.156]	[0.115, 0.155]
	cmp	[0.414, 0.468]	[0.565, 0.618]	[0.073, 0.105]	[0.084, 0.117]	[0.095, 0.129]
ICEWS18	std	[0.364, 0.394]	[0.475, 0.506]	[0.152, 0.175]	[0.154, 0.177]	[0.136, 0.158]
	geo	[0.400, 0.450]	[0.388, 0.437]	[0.125, 0.160]	[0.145, 0.182]	[0.124, 0.159]
	tmp	[0.407, 0.447]	[0.472, 0.512]	[0.195, 0.228]	[0.203, 0.237]	[0.179, 0.211]
	cmp	[0.449, 0.484]	[0.450, 0.486]	[0.156, 0.182]	[0.172, 0.199]	[0.148, 0.174]
ICEWS05-15	std	[0.325, 0.360]	[0.465, 0.502]	[0.109, 0.133]	[0.103, 0.127]	[0.097, 0.120]
	geo	[0.323, 0.375]	[0.473, 0.529]	[0.083, 0.116]	[0.096, 0.131]	[0.095, 0.130]
	tmp	[0.338, 0.392]	[0.436, 0.492]	[0.109, 0.147]	[0.120, 0.160]	[0.110, 0.149]
	cmp	[0.440, 0.486]	[0.444, 0.489]	[0.091, 0.119]	[0.121, 0.152]	[0.074, 0.099]
GDELT	std	[0.321, 0.349]	[0.295, 0.322]	[0.092, 0.110]	[0.141, 0.163]	[0.137, 0.158]
	geo	[0.313, 0.346]	[0.244, 0.275]	[0.061, 0.079]	[0.079, 0.099]	[0.087, 0.108]
	tmp	[0.364, 0.401]	[0.253, 0.286]	[0.119, 0.145]	[0.169, 0.198]	[0.166, 0.195]
	cmp	[0.414, 0.446]	[0.263, 0.292]	[0.119, 0.141]	[0.167, 0.192]	[0.170, 0.195]

Table A9: Pooled Hits@1 per programmatic difficulty band following Definition 4 and the Spearman correlation ρ between band index and accuracy.

System	L1_Easy	L2_Medium	L3_Hard	L4_Expert	ρ
MultiQA	0.0734	0.1290	0.2005	0.1817	+0.80
CronKGQA	0.0630	0.1131	0.1555	0.1628	+1.00
EmbedKGQA	0.1242	0.1706	0.2630	0.2398	+0.80
In-Context RALM	0.3014	0.3768	0.5049	0.3399	+0.40
RTQA	0.2237	0.3172	0.4645	0.2642	+0.40
TimeR^4	0.3395	0.3697	0.4941	0.3148	−0.20
Temp-R1	0.3681	0.4219	0.4473	0.4084	+0.40
Closed-book DeepSeek-V4-Flash	0.0487	0.1008	0.1983	0.1660	+0.80
Closed-book Gemma4-31B-IT	0.0719	0.1233	0.2278	0.1779	+0.80
Closed-book Qwen3.6-35B-A3B	0.0731	0.1096	0.2104	0.1754	+0.80

Table A10: Split gaps in points: accuracy on the standard split minus accuracy on the shift split, per source (IC14, IC18, and IC05-15 abbreviate the ICEWS sources; the closed-book models appear under their short names). Positive values mean the shift split is harder for that system.

	std minus geo				std minus tmp				std minus cmp			
System	IC14	IC18	IC05-15	GDELT	IC14	IC18	IC05-15	GDELT	IC14	IC18	IC05-15	GDELT
MultiQA	+11.61	−3.29	−1.96	+8.69	+15.03	+1.38	+6.15	+5.23	−0.83	+5.28	−2.00	+5.46
CronKGQA	+11.44	−0.17	−2.75	+5.33	+20.72	+13.76	+5.57	+10.70	+2.87	+7.66	+2.65	+8.47
EmbedKGQA	+15.89	+3.33	+1.97	+5.36	+11.84	+1.47	+5.51	−1.99	+11.01	+6.36	+6.61	+8.21
In-Context RALM	+2.64	−6.97	−4.82	+2.06	−11.04	−4.52	−4.49	−8.06	−9.05	−6.07	−13.06	−4.95
RTQA	+3.68	−10.00	−4.91	+1.92	−8.68	−4.95	−4.22	−6.30	−3.22	−1.86	−6.92	−3.49
TimeR^4	+4.77	−4.55	−0.66	+0.47	−5.66	−4.85	−2.24	−4.79	−10.29	−8.69	−12.07	−9.54
Temp-R1	+6.58	+7.84	−1.61	+4.91	+1.72	−0.17	+2.05	+3.97	−7.85	+2.27	+1.69	+3.14
DS-V4F	−2.31	+2.06	+2.18	+3.12	−3.26	−4.85	−0.69	−3.07	+0.18	−0.60	+1.57	−2.83
G4-31B	−1.61	+0.17	+0.14	+6.33	−3.19	−5.48	−2.50	−3.10	+0.47	−2.01	−2.17	−2.72
Q3.6-35B	−0.77	+0.54	−0.41	+5.03	−2.40	−4.77	−2.12	−3.28	−0.07	−1.40	+2.21	−3.53

D Notation Summary

Tables A11 and A12 collect every symbol grouped by the equation in which it first appears, with the symbols of the surrounding definitions and of the release registry listed where they are introduced.

Table A11: Notation, part one: the objects of the problem formulation and the symbols of Sections 4 to 5.4, grouped by the equation in which they first appear; symbols introduced in surrounding definitions and sections are listed at their point of introduction.

Symbol	Meaning	Introduced
Preliminaries (before Section 4)		
$\mathcal{G}$	temporal knowledge graph $(\mathcal{E}, \mathcal{R}, \mathcal{T}, \mathcal{F})$	Sec. 3
$\mathcal{E}$, $\mathcal{R}$, $\mathcal{T}$	entity set, relation set, totally ordered time-point set	Sec. 3
$\mathcal{F}$; f	fact set; one fact, the quadruple (s, r, o, t)	Sec. 3
s, r, o, t	subject, relation, object, and time point of a fact	Sec. 3
$\tau(f)$	time projection of f (likewise sub, rel, obj)	Sec. 3
$\mathcal{H}$	geographical knowledge layer $(N, A, \rho, \kappa, \gamma)$	Def. 1
N; A	geographical nodes; symmetric land-border adjacency	Def. 1
ρ	centroid map $N \to \mathbb{R}^2$, (latitude, longitude)	Def. 1
κ; *Reg*	region assignment $N \to 2^{Reg}$; set of named regions	Def. 1
γ; $\perp$	partial entity grounding $\mathcal{E} \to N \cup \{\perp\}$; unresolvable marker	Def. 1
$\lambda_s(f)$, $\lambda_o(f)$	grounded locations of the subject and object of f	Sec. 3
$q = (\sigma, L, a)$	question: surface string, typed logical form, gold answer set	Sec. 3
C_S, C_T, C_G	structural, temporal, and geographical constraints of L	Sec. 3
$c_1, \ldots, c_5$; θ	grounding-cascade confidence levels; acceptance threshold	Sec. 4
W; ℓ	temporal window $W \subseteq \mathcal{T}$; target time granularity	Def. 2
E	tri-axially consistent evidence set $E \subseteq \mathcal{F}$	Def. 2
Section 4 (Haversine distance)		
$d(n_i, n_j)$; n_i	great-circle distance between the centroids of nodes n_i and n_j	Sec. 4
R	mean Earth radius, approximately $6{,}371$ km	Sec. 4
u_{ij}	haversine of the central angle between the two centroids	Sec. 4
φ_i, ψ_i	latitude and longitude of the centroid $\rho(n_i)$	Sec. 4
$\Delta\varphi$, $\Delta\psi$	pairwise latitude and longitude differences	Sec. 4
Equation 1 (entanglement certification)		
v	one reasoning axis, $v \in \{S, T, G\}$	Eq. 1
$L^{\setminus v}$	L with the constraints of axis v removed	Def. 3
$a(L)$	answer set produced by symbolically executing L	Sec. 5.2
Equation 2 (programmatic difficulty)		
$d_{\mathrm{prog}}(L)$	programmatic difficulty score of L	Eq. 2
h_S, h_T, h_G	structural, temporal, and geographical difficulty terms	Eq. 2
$E(L)$	evidence pool produced by symbolically executing L	Sec. 5.2
$\lvert \cdot \rvert$	set cardinality	Eq. 2

Table A12: Notation, part two: the symbols of Equations 3 to 4, the release registry, and the analysis statistics of Equation 5 and Appendix C.

Symbol	Meaning	Introduced
Equation 3 (entanglement density)		
g_f	fraction of facts with at least one grounded participant	Eq. 3
δ	entanglement density $g_f \cdot p_G \cdot p_T$	Eq. 3
p_G, p_T	fractions of multi-fact groups spanning two countries / two days	Def. 5
m	fact count $\lvert\mathcal{F}\rvert$ of a source	Sec. 6
N^*	number of distinct certifiable logical forms of a source	Sec. 6
$\mathcal{W}$	admissible window family of a source registry	Sec. 6
Equation 4 (capacity estimate)		
$\widehat{N^*}$	Chao1 estimate of the certifiable-form ceiling	Eq. 4
S_{obs}	observed richness, equal to $U(B)$	Eq. 4
f_1, f_2	certified forms observed exactly once and exactly twice	Eq. 4
B; $U(B)$	per-family sampling budget; unique certified forms observed under B	Sec. 6
N_{rel}	released number of questions of a source	Eq. 4
Release registry (Section 6 and Appendix A.3)		
Q	per-family release quota	App. A.3
χ_{anc}, χ_{ans}, χ_{geo}, χ_{rel}	per-anchor, per-answer, per-constant, and per-relation caps	App. A.3
$D_{\max}$; $A_{\max}$	hub-degree cap on anchors; answer-set size cap	App. A.3
$\mathcal{D}$	distance ladder of family F6	App. A.3
η_G; η_T	held-out fraction of geographic constants; of the source span	Sec. 5.4
Equation 5 (channel decomposition) and analysis statistics		
acc_x; h_x	overall accuracy of run x; its gold-in-context rate	Eq. 5
a_x^{hit}, a_x^{miss}	accuracy of x given the gold was / was not in context	Eq. 5
Mix, Hit, Miss	retrieval-mix, hit-accuracy, and miss-accuracy terms	Eq. 5
b, c	discordant counts of a paired comparison	App. C.5
κ; ρ	Cohen's agreement coefficient; Spearman rank correlation (these two statistics reuse the letters that Definition 1 assigns to the region map and the centroid map)	Sec. 7.3