

Not All Queries Are Equal: Query-Type-Aware Mixture-of-Experts for Temporal Knowledge Graph Forecasting

Rong Ma^{1,2,3}, Bo Ma^{1,2,3*}, Yating Yang^{1,2,3*}, Xi Zhou^{1,2,3}, Lei Wang^{1,2,3},
Rui Dong^{1,2,3}, Ahtamjan Ahmat^{1,2,3}, Zhen Wang^{1,2,3}

¹Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Sciences, Urumqi, China ²University of Chinese Academy of Sciences, Beijing, China

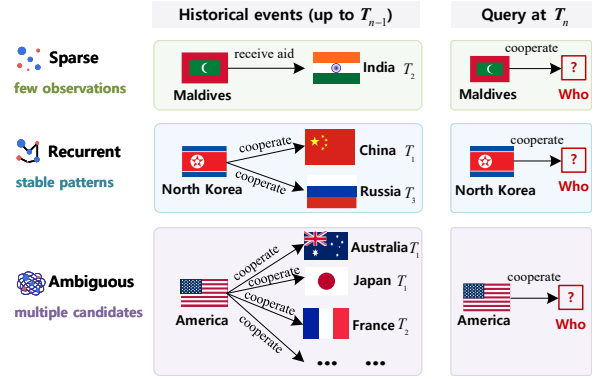
³Xinjiang Laboratory of Minority Speech and Language Information Processing, Urumqi, China
{marong, mabo, yangyt, zhoxi, wanglei, dongrui, ahtmj, wang_zhen}@ms.xjb.ac.cn

Abstract

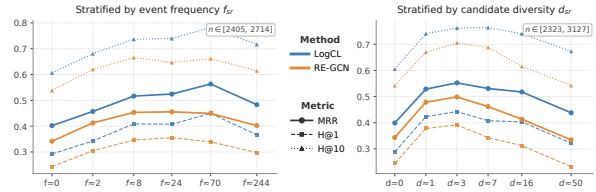
Temporal knowledge graph forecasting predicts future facts from historical snapshots, and its accuracy hinges on how well a model exploits the historical evidence available to each query. Existing methods route every query through a single reasoning pipeline, yet queries with the same predicate form can rest on very different histories and call for different inductive biases. We observe that query histories fall into three regimes: sparse histories with few prior events, recurrent histories dominated by stable repetition, and ambiguous histories with many competing candidates. Building on this observation, we propose QTI-MoE, a Query-Type-aware Inductive-bias Mixture-of-Experts framework. QTI-MoE characterises each query with a soft type label derived from its historical prefix and forwards it to one of three experts whose memory pathways carry complementary biases: recency-weighted aggregation for sparse queries, count-based repetition for recurrent queries, and diversity-normalised scoring for ambiguous queries. QTI-MoE achieves state-of-the-art performance on four standard benchmarks, and a type-stratified analysis confirms that each expert peaks on its target regime¹.

1 Introduction

Temporal Knowledge Graphs (TKGs) encode real-world events as timestamped relational facts, capturing how entities interact and how those interactions change over time (Cai et al., 2024a). TKG forecasting, the task of predicting missing entities in future queries from historical snapshots (Wang et al., 2023), supports applications including foreign-affairs analysis, social-event anticipation, and conflict early warning (Mirtaheri et al., 2023; Xu et al., 2025). Unlike static knowledge graph



(a) TKG forecasting examples: queries with the same predicate form may have *sparse*, *recurrent*, or *ambiguous* histories.



(b) RE-GCN and LogCL on ICEWS14 perform notably worse in low-frequency (f) and high-diversity (d) regions, where f denotes the number of historical occurrences and d the number of distinct candidate entities.

Figure 1: Motivation for query-type-aware forecasting.

completion, TKG structures evolve over time, making query predictability depend on historical evidence. A query with a sparse event record poses a very different challenge from one supported by frequent, recurrent patterns (Zhao et al., 2026).

Faced with this evidence-dependent nature of TKG forecasting, prior work has advanced along two complementary lines. The first line is structure-based (Li et al., 2022b; Chen et al., 2024; Dong et al., 2025). It models graph evolution with relational graph convolutions and recurrent dynamics, with later refinements sharpening the temporal encoder through contrastive learning, entity-aware attention, and multi-scale encoding (Ma et al., 2025; Dong et al., 2025). The second line is LLM-based (Xia et al., 2024; Zhang et al., 2025). It casts forecasting as semantic reasoning over linearised event

*Corresponding author.

¹The code and data are available at <https://github.com/wooden070/QTI-MoE>.

histories (Lee et al., 2023), and is subsequently extended with retrieval-augmented and parameter-efficient adaptations that inject scenario-specific knowledge into the LLM (Liao et al., 2024). More recently, hybrid multi-expert frameworks such as MESH (Deng et al., 2025) attempt to bring the two signals together by assigning a structural expert and a semantic expert to handle different aspects of the same query. These lines of work differ in machinery, but converge on a shared assumption: one reasoning pipeline is expected to serve every query equally well. Even when multiple experts are introduced, routing operates within a query, deciding which signal to trust rather than which strategy fits the query at hand. Yet queries with the same predicate form can rest on very different histories (Figure 1a). Such differences in evidence shape are not cosmetic; they change what kind of reasoning a query actually demands. A uniform pipeline cannot match this variation, suggesting that not all queries are equal.

To check whether this gap is observable in practice, we evaluate two representative forecasters, RE-GCN (Li et al., 2021) and LogCL (Chen et al., 2024), on ICEWS14 (García-Durán et al., 2018), stratifying their predictions by the historical evidence available to each query. As shown in Figure 1b, both follow an inverted-U pattern, with accuracy peaking in the mid-range and dropping at the low-frequency and high-diversity ends. Despite their architectural differences, both models suffer a sharp drop across all metrics when queries fall into the few-observation or many-candidate regions. This is not an artefact of any single architecture. It indicates that different query regimes call for different inductive biases, and that no monolithic pipeline can host all of them at once. A natural response is to assign each regime its own expert and inductive bias, with a router that selects by the shape of a query’s history rather than its content. This shifts routing from operating within a query to operating across queries.

Building on this observation, we propose QTI-MoE, a Query-Type-aware Inductive-bias Mixture-of-Experts method for TKG forecasting. QTI-MoE reads the historical shape of each query and forwards it to the expert best suited to that regime. The method has three components. We first construct soft type labels for each forecasting query from its historical prefix statistics, calling on a frozen LLM only to refine boundary cases. We then design three experts that share a structural backbone but

carry distinct inductive biases in their memory pathways: recency for sparse histories, stable repetition for recurrent histories, and diversity-aware aggregation for ambiguous histories. A query-aware router, conditioned on subject and relation embeddings, finally mixes the three expert predictions according to the historical shape of the query. Soft labels and the LLM enter only at training time. Inference relies on query embeddings alone, without invoking the LLM. We evaluate QTI-MoE on four standard benchmarks and obtain state-of-the-art performance. A type-stratified analysis further shows that each expert reaches its highest accuracy on the query type it is designed for, confirming the effectiveness of our type-specialised expert design.

Our contributions can be summarised as below.

- We propose a soft query-type labelling scheme that characterises each forecasting query across sparse, recurrent, and ambiguous regimes for query-type-aware reasoning.
- We design QTI-MoE, in which three experts carry complementary inductive biases matched to the three regimes, and a query-aware router mixes them by historical shape with no LLM involvement at inference.
- QTI-MoE achieves state-of-the-art performance on four standard benchmarks, and a type-stratified analysis confirms that each expert peaks on its target regime.

2 Problem Formulation

A Temporal Knowledge Graph (TKG) is a time-ordered sequence of snapshots $\mathcal{G} = \{G_1, \dots, G_T\}$, where each $G_t = (\mathcal{V}, \mathcal{R}, \mathcal{F}_t)$ shares an entity set $\mathcal{V}$ and a relation set $\mathcal{R}$, and $\mathcal{F}_t = \{(s, r, o, t)\}$ collects the facts at time t with $s, o \in \mathcal{V}$ and $r \in \mathcal{R}$. We write $\mathcal{G}_{<t_q} = \{G_1, \dots, G_{t_q-1}\}$ for the strict prefix before the query time t_q . Given $\mathcal{G}_{<t_q}$, *TKG forecasting* ranks candidate entities $e \in \mathcal{V}$ for a candidate object query $(s, r, ?, t_q)$. Subject queries $(?, r, o, t_q)$ are converted to object queries on the inverse relation r^{-1} , i.e., $(o, r^{-1}, ?, t_q)$, which are then scored using the same prediction pipeline. No fact or temporal information at time t_q or later may be accessed during prediction.

3 Method

This section presents QTI-MoE. As shown in Figure 2, the framework proceeds in three steps, each

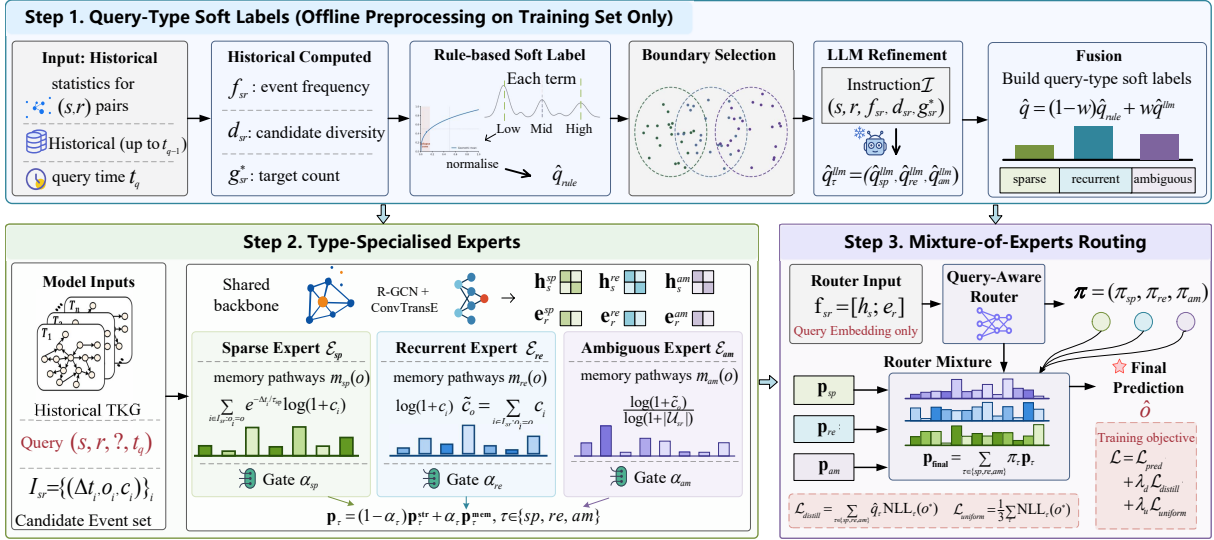


Figure 2: Overview of QTI-MoE. The framework consists of three modules: (1) offline query-type soft label construction that fuses rule-based statistics with LLM refinement on a boundary subset; (2) type-specialised experts that share a structural backbone but carry complementary memory pathways for sparse, recurrent, and ambiguous queries; and (3) a query-aware router that mixes expert predictions from query embeddings at inference.

developed in one subsection below. We first construct query-type soft labels that characterise the historical behaviour of each (s, r) pair (§3.1). We then design three type-specialised experts whose memory pathways implement complementary inductive biases, one matched to each query type (§3.2). Finally, a query-aware router combines the expert predictions and is trained jointly with the experts (§3.3). Soft labels and the LLM are used only during training; at inference, the router operates on query embeddings alone.

3.1 Query-Type Soft Labels

For each query pair (s, r) , we assign a soft label $\hat{q}(s, r) = (\hat{q}_{sp}, \hat{q}_{re}, \hat{q}_{am})$ over three query types that capture how the historical prefix supports the query: **sparse** (*sp*), where the prefix contains few events and the target entity has rarely appeared; **re-current** (*re*), where the same targets repeat across the prefix and the historical pattern is stable; and **ambiguous** (*am*), where a rich history involves many competing candidates and the most frequent one is often not the correct answer. These abbreviations are used throughout the rest of the paper.

Rule-based construction. We summarise the history of (s, r) with three prefix scalars: the event frequency f_{sr} , the candidate diversity d_{sr} , and the target occurrence count g_{sr}^* . Each scalar is passed individually through three sigmoid memberships indicating whether it is low, middle, or high relative to dataset-level percentile thresholds. Each type

score is then the geometric mean of the memberships defining it: sparse requires all three signals to be jointly low, recurrent requires reliable target recurrence under moderate frequency and diversity, and *ambiguous* requires frequency and diversity to be jointly high. The geometric form makes a type score collapse whenever any defining condition fails, and the three scores are ℓ_1 -normalised across the types to obtain $\hat{q}_{rule}$. We treat g_{sr}^* as a privileged training signal: it is strictly used only during offline labelling and is never accessed during model updates or inference.

LLM-based refinement. We query the frozen LLM only on a boundary subset: pairs whose top-two rule probabilities differ by less than ϵ , or whose statistics lie within a class threshold tolerance. For each pair, we provide an instruction $\mathcal{I}$ containing the relation surface form, the statistics $(f_{sr}, d_{sr}, g_{sr}^*)$ with dataset-level reference scales, and a request to choose one of the three type labels. The LLM type distribution is read from next-token probabilities over the label tokens:

$$\hat{q}_{\tau}^{llm} = P_{llm}(y = \tau \mid \mathcal{I}(s, r, f_{sr}, d_{sr}, g_{sr}^*)) \quad (1)$$

for $\tau \in \{sp, re, am\}$. The two distributions are then fused together as follows:

$$\hat{q}(s, r) = (1 - w) \hat{q}_{rule} + w \hat{q}^{llm} \quad (2)$$

where w is zero outside the boundary subset and, within it, is the product of the LLM’s confidence

and its disagreement with the rule (both in $[0, 1]$). The rule therefore remains dominant unless the LLM is both confident and points to a different type. The LLM is invoked only during this offline refinement step. The definition of w , the prompt template, the boundary criteria, and the factor definitions are provided in Appendix C.2.

3.2 Type-Specialised Experts

We instantiate three experts $\{\mathcal{E}_{sp}, \mathcal{E}_{re}, \mathcal{E}_{am}\}$, one for each query type in Section 3.1. Each expert $\mathcal{E}_\tau$ produces an output distribution $\mathbf{p}_\tau$ over $\mathcal{V}$ by combining a structural prediction with a type-specific memory score.

The three experts share the same backbone design but learn independent parameters. Each expert instantiates its own copy of a recurrent R-GCN (Schlichtkrull et al., 2018) with relation-conditioned message passing and a GRU cell, following the recurrent encoding scheme of Chen et al. (2024), followed by a ConvTransE decoder (Shang et al., 2019). Each expert therefore maintains its own entity and relation embeddings and encoder-decoder weights, and produces a type-specific structural distribution $\mathbf{p}_\tau^{str}$ from its own subject embedding $\mathbf{h}_s^\tau$ and relation embedding $\mathbf{e}_r^\tau$. Decoupling parameters lets each expert align its representations with its target type, while sharing the architecture keeps the encoder, parameter budget, and history length consistent across experts. The structural backbone is adopted from existing recurrent TKG forecasters rather than newly designed; our contribution lies in the type-specific memory pathways and the query-aware routing built on top of it. The backbone specification is provided in Appendix B.

Type-specific memory pathways. For a query $(s, r, ?, t_q)$, every expert reads the same evidence from the prefix: the (s, r) -conditioned event set $I_{sr} = \{(\Delta t_i, o_i, c_i)\}_i$, where $\Delta t_i = t_q - t_i$, o_i is the observed object, and c_i its per-snapshot count. The experts differ only in how they aggregate this evidence into a memory score $\mathbf{m}_\tau \in R^{|\mathcal{V}|}$.

Sparse expert: Sparse histories contain few events, and recent matches carry disproportionately strong signals. We aggregate per-event contributions with exponential recency decay and logarithmic count compression:

$$m_{sp}(o) = \sum_{i \in I_{sr}: o_i=o} e^{-\Delta t_i / \tau_{sp}} \log(1 + c_i) \quad (3)$$

where $\tau_{sp} > 0$ is a learnable time constant. When no event matches, the score vanishes and the structural pathway takes over via the gate below.

Recurrent expert: Repeating co-occurrence patterns make the total count more reliable than any single occurrence:

$$m_{re}(o) = \log(1 + \tilde{c}_o), \quad \tilde{c}_o = \sum_{i \in I_{sr}: o_i=o} c_i \quad (4)$$

This pathway has no expert-specific learnable parameters; stable repetition is fully captured by count aggregation.

Ambiguous expert: When many distinct objects appear, a raw count score concentrates probability on the most frequent candidate, which is often a near-miss. The ambiguous expert divides the count score by the log-diversity of the candidate pool $\mathcal{U}_{sr} = \{o_i : i \in I_{sr}\}$:

$$m_{am}(o) = \frac{\log(1 + \tilde{c}_o)}{\log(1 + |\mathcal{U}_{sr}|)} \quad (5)$$

where $\tilde{c}_o$ is defined as in Eq. 4. For a fixed query, the denominator is a single constant and acts as a query-adaptive temperature, preserving the count-induced ranking but flattening the resulting memory distribution as the pool grows.

Structure-memory mixture. Each memory score is turned into a probability distribution $\mathbf{p}_\tau^{mem} = \text{softmax}(\mathbf{m}_\tau)$ and combined with the expert’s structural distribution $\mathbf{p}_\tau^{str}$ through a learned, query-conditioned scalar gate $\alpha_\tau \in (0, 1)$:

$$\mathbf{p}_\tau = (1 - \alpha_\tau) \mathbf{p}_\tau^{str} + \alpha_\tau \mathbf{p}_\tau^{mem} \quad (6)$$

where $\alpha_\tau = \sigma(MLP_\tau([\mathbf{h}_s^\tau; \mathbf{e}_r^\tau]))$ is produced by a small expert-specific MLP from its own subject and relation embeddings. The gate lets each expert rely on its memory pathway when the query matches its type and fall back to the structural backbone otherwise; in particular, when the memory score vanishes the gate naturally yields $\mathbf{p}_\tau \approx \mathbf{p}_\tau^{str}$.

3.3 Mixture-of-Experts Routing

This subsection formalises the fusion of the three expert distributions into a unified prediction and the subsequent end-to-end training of the model.

Query-aware router. The router maps each query to a mixing distribution $\pi = (\pi_{sp}, \pi_{re}, \pi_{am})$ over the three experts from the concatenated subject–relation embedding $\mathbf{f}_{sr} = [\mathbf{h}_s; \mathbf{e}_r]$:

$$\pi = \text{softmax}(MLP_{rt}(\mathbf{f}_{sr}) / T) \quad (7)$$

where $T > 0$ is a learnable temperature and the MLP output bias is initialised to $\log \bar{\mathbf{q}}$, the corpus-level prior. The router takes no soft label or gold-aware statistic as input. The final prediction is the routed mixture, implemented in log space for numerical stability:

$$\mathbf{p}_{final} = \sum_{\tau \in \{sp, re, am\}} \pi_{\tau} \mathbf{p}_{\tau} \quad (8)$$

Training objective. QTI-MoE is trained with a single objective that combines a prediction term with two specialisation terms:

$$\mathcal{L} = \mathcal{L}_{pred} + \lambda_d \mathcal{L}_{distill} + \lambda_u \mathcal{L}_{uniform} \quad (9)$$

The three terms supervise the model at different levels: the routed mixture, the type-conditioned per-expert prediction, and the per-expert prediction averaged uniformly across experts.

The prediction term $\mathcal{L}_{pred} = -\log p_{final}(o^*)$ is the standard NLL of the gold answer on the routed mixture; it is the direct task signal that the router and the experts must jointly produce a correct prediction. The distillation term $\mathcal{L}_{distill} = \sum_{\tau} \hat{q}_{\tau} \text{NLL}_{\tau}(o^*)$ is the only place where the soft label $\hat{\mathbf{q}}$ shapes training: it routes gradient mass to the expert responsible for each query, with $\text{NLL}_{\tau}(o^*) = -\log p_{\tau}(o^*)$ the per-expert NLL of the gold answer under $\mathbf{p}_{\tau}$. This term drives type-conditioned specialisation, so that each expert improves most on the queries matching its target type. The uniform term provides a small baseline supervision to every expert regardless of the soft label, preventing experts with persistently low responsibility from degrading on out-of-type queries:

$$\mathcal{L}_{uniform} = \frac{1}{3} \sum_{\tau} \text{NLL}_{\tau}(o^*) \quad (10)$$

The weights λ_d and λ_u balance soft-label-guided specialisation and uniform supervision, with their sensitivity analysed in Appendix A.4.

4 Experiments

4.1 Datasets

We evaluate QTI-MoE on four widely used TKG forecasting benchmarks: ICEWS14 (García-Durán et al., 2018), ICEWS18 (Jin et al., 2020), ICEWS05-15 (García-Durán et al., 2018), and GDELT. The first three are derived from the Integrated Crisis Early Warning System (Boschee et al., 2015), which records international political events

in the form of (s, r, o, t) quadruples. ICEWS14 covers events from the year 2014; ICEWS18 covers events from January to October 2018; and ICEWS05-15 spans an eleven-year period from 2005 to 2015. GDELT is constructed from the Global Database of Events, Language, and Tone, and contains geopolitical events recorded at a finer temporal resolution of 15 minutes. Following standard protocol (Li et al., 2021; Chen et al., 2024), each dataset is split chronologically into training, validation, and test sets. The key statistics of the four datasets are summarised in Table 1.

Statistic	ICEWS14	ICEWS18	ICEWS05-15	GDELT
Entities	7,128	23,033	10,488	7,691
Relations	230	256	251	240
Training	74,845	373,018	368,868	1,734,399
Validation	8,514	45,995	46,302	238,765
Test	7,371	49,545	46,159	305,241
Time granularity	24h	24h	24h	15m
Number of snapshots	365	304	4,017	2,975

Table 1: Statistics of TKG forecasting benchmarks.

4.2 Experimental Settings

We evaluate QTI-MoE on four standard TKG forecasting benchmarks: ICEWS14, ICEWS18, ICEWS05-15, and GDELT. Following prior work (Chen et al., 2024; Zhang et al., 2025), we report MRR and Hits@{1, 3, 10} under the standard time-aware filtered setting, and select the test checkpoint by validation MRR. The three experts share a backbone of a two-layer R-GCN encoder and a ConvTransE (Shang et al., 2019) decoder, with hidden dimension 200, 100 relation bases, and dropout 0.2; each expert keeps its own parameters. History lengths for ICEWS14, ICEWS18, ICEWS05-15, and GDELT are 7, 7, 9, and 7. We train for 30 epochs with Adam at learning rate $5e-4$, setting $\lambda_d = 0.6$, $\lambda_u = 0.2$, and the rule-LLM fusion weight $w = 0.2$. The Adam optimiser (Kingma and Ba, 2015) is applied for all parameters. InternLM2-7B serves as the frozen LLM for soft label refinement (temperature 1.0), and is invoked only at the offline labelling stage. All experiments run on eight 48GB NVIDIA GPUs.

4.3 Baselines

To comprehensively evaluate QTI-MoE, we compare it against twelve representative baselines covering two technical paradigms that have dominated TKG forecasting in recent years:

(1) Structure-based TKG Forecasting Methods, which model the structural evolution of the tempo-

Model	ICEWS14				ICEWS05-15				ICEWS18				GDELT			
	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
RE-GCN (Li et al., 2021)	41.50	30.86	46.60	62.47	46.41	35.17	52.76	67.64	30.55	20.00	34.73	51.46	19.31	11.99	20.61	33.59
HisMatch (Li et al., 2022b)	46.42	35.91	51.63	66.84	52.85	42.01	59.05	73.28	33.99	23.91	37.90	53.94	22.01	14.45	23.80	36.61
LogCL (Chen et al., 2024)	48.90	37.80	54.70	70.26	57.04	46.07	63.72	77.87	35.67	24.53	40.32	57.74	23.75	14.64	25.60	42.33
DiffuTKG (Cai et al., 2024b)	48.50	36.40	49.40	72.70	52.70	40.30	60.20	75.90	36.70	25.70	38.80	57.80	25.10	16.30	27.50	42.30
DiMNet (Dong et al., 2025)	45.72	34.41	51.37	67.99	58.93	48.55	65.16	78.33	34.13	23.29	38.42	55.80	21.93	14.03	23.57	37.49
TiRGN-ICL (Lee et al., 2023)	43.27	32.28	49.04	64.17	49.15	37.25	55.73	70.90	32.18	21.07	36.98	54.04	-	-	-	-
TiRGN-COH (Xia et al., 2024)	43.94	33.07	49.64	64.90	49.71	38.01	56.40	71.25	32.98	21.83	37.79	54.92	-	-	-	-
GenTKG (Liao et al., 2024)	/	36.85	47.95	53.50	-	-	-	-	/	24.25	37.25	42.10	/	13.90	22.55	30.45
LLM-DA (Wang et al., 2024)	47.10	36.90	52.60	67.10	52.10	41.60	58.60	72.80	35.70	25.50	40.30	57.00	-	-	-	-
MESH (Deng et al., 2025)	44.36	-	49.81	64.21	48.66	-	54.26	68.57	33.96	-	38.37	54.12	-	-	-	-
ANRE (Tang et al., 2025)	47.40	36.90	51.10	65.70	50.90	39.10	58.00	69.60	35.50	26.00	39.20	56.70	24.30	16.60	26.60	37.50
SRM-LLM (Zhang et al., 2025)	50.84	39.80	56.82	71.92	57.97	46.98	64.91	78.54	37.23	25.92	42.01	59.67	25.23	15.92	27.21	-
QTI-MoE	51.53	40.48	57.68	72.86	59.27	48.82	65.75	78.89	38.45	27.07	43.39	61.07	27.53	17.46	30.34	48.36

Table 2: Performance comparison on ICEWS14, ICEWS05-15, ICEWS18 and GDELT. All metrics are reported in percentage (%). ‘/’ indicates non-computable results, while ‘-’ denotes unavailable public results.

ral graph through relational graph encoders and recurrent temporal dynamics, often augmented with historical contrastive learning, diffusion-based generation, or structural matching mechanisms: RE-GCN (Li et al., 2021), HisMatch (Li et al., 2022b), LogCL (Chen et al., 2024), DiffuTKG (Cai et al., 2024b), and DiMNet (Dong et al., 2025).

(2) LLM-based and LLM-Augmented Forecasting Methods, which cast forecasting as semantic reasoning over linearised event histories or augment structural models with LLM-derived priors, retrieval, or multi-expert designs: TiRGN-ICL (Lee et al., 2023), TiRGN-COH (Xia et al., 2024), GenTKG (Liao et al., 2024), LLM-DA (Wang et al., 2024), MESH (Deng et al., 2025), ANRE (Tang et al., 2025), and SRM-LLM (Zhang et al., 2025).

4.4 Performance Comparison

Table 2 reports the comparison of QTI-MoE against twelve baselines on four benchmarks. QTI-MoE consistently attains the best score on every metric across all four benchmark datasets. On the relatively dense ICEWS14 and ICEWS05-15, QTI-MoE leads the strongest prior result by 0.69 and 0.34 MRR; on the more challenging ICEWS18 and GDELT, the MRR margin widens to 1.22 and 2.30, and the largest single gain reaches 6.03 points on GDELT Hits@10. The advantage of QTI-MoE grows as the data shifts toward sparser and more diverse query histories, consistent with a model that adapts its inductive bias to the shape of each query. Against MESH, the most recent multi-expert baseline, QTI-MoE improves MRR by 7.17 points on ICEWS14 and 4.49 points on ICEWS18, show-

ing that routing across queries by historical shape captures information that within-query evidence routing does not. Overall, QTI-MoE achieves state-of-the-art performance on all four benchmarks.

4.5 Ablation Study

We ablate one component at a time to verify its contribution. Table 3 reports MRR and Hits@1; full results are in Table 6.

Model	ICEWS14		ICEWS05-15		ICEWS18		GDELT	
	MRR	H@1	MRR	H@1	MRR	H@1	MRR	H@1
QTI-MoE	51.53	40.48	59.27	48.82	38.45	27.07	27.53	17.46
w/o MoE	50.79	39.78	58.49	47.63	37.57	26.14	24.24	14.85
w/o sparse	50.09	38.89	58.14	47.39	37.03	25.67	23.74	14.47
w/o recurrent	47.73	36.30	56.03	44.78	35.62	24.41	23.13	14.09
w/o ambiguous	49.43	38.22	56.95	45.81	35.84	24.37	23.28	14.10
w/o $\mathcal{L}_{distill}$	49.24	37.95	57.81	46.96	37.76	26.32	27.15	16.97
w/o $\mathcal{L}_{uniform}$	50.27	39.14	56.28	45.58	37.12	25.92	26.58	16.46
w/o LLM	50.38	38.98	57.65	46.85	37.75	26.39	24.81	15.36

Table 3: Ablation study results across four datasets (%).

(1) **Effect of the MoE framework.** Replacing the three-expert mixture with a single shared backbone (w/o MoE) degrades performance on all datasets, with MRR dropping by 0.74 on ICEWS14 and 3.29 on GDELT. The gap is largest on GDELT, where finer temporal resolution yields more diverse historical shapes, supporting our premise that routing across queries pays off when query variation is pronounced.

(2) **Effect of the three experts.** Removing any expert causes a larger drop than removing the MoE itself, showing that each occupies a distinct functional slot. The *recurrent* expert is the most critical (MRR -3.80 on ICEWS14, -4.40 on GDELT), re-

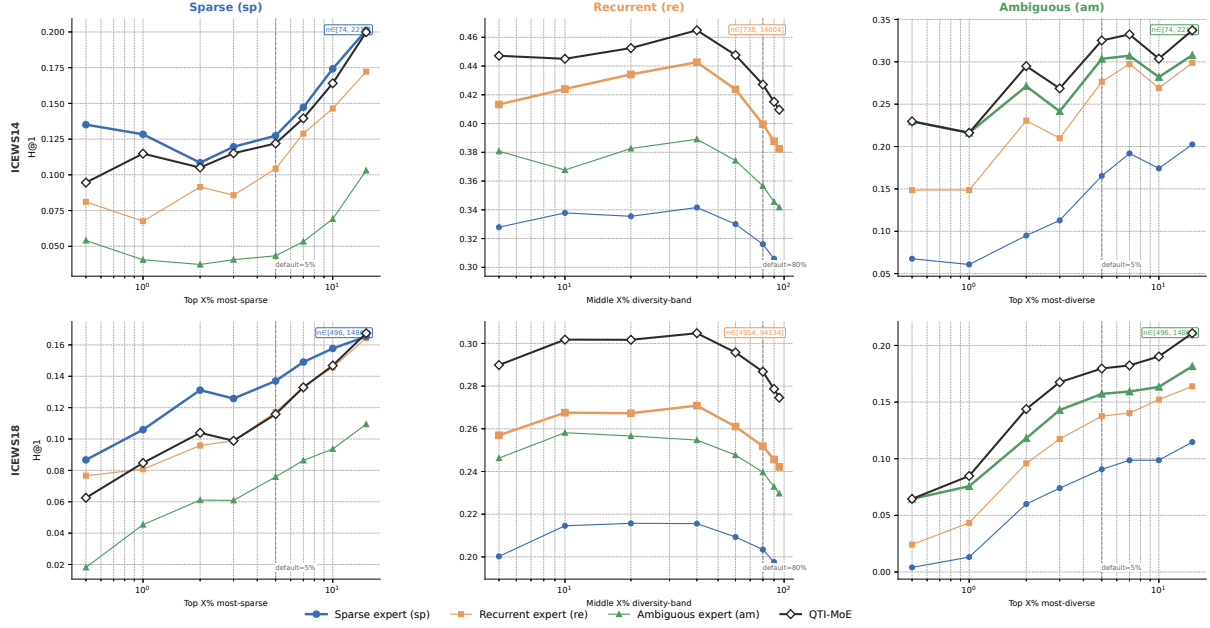


Figure 3: H@1 robustness sweep across three query scenarios (columns) and two TKG datasets (rows). The focal expert for each scenario is drawn in bold; the dashed vertical line marks the default threshold.

ICEWS14	MRR	Hit@1	Hit@3	Hit@10
QTI-MoE	51.53	40.48	57.68	72.86
Ens-3	48.04	36.81	53.85	69.94
Ens-6	47.95	36.96	53.64	69.26
ICEWS18	MRR	Hit@1	Hit@3	Hit@10
QTI-MoE	38.45	27.07	43.39	61.07
Ens-3	35.84	24.82	40.32	57.69
Ens-6	35.13	24.04	39.67	57.20

Table 4: Parameter-budget controls (%). Ens- k is a uniform ensemble of k independently trained backbones without query-type awareness. Ens-3 matches the parameter budget of QTI-MoE; Ens-6 doubles it.

flecting the prevalence of repeating patterns. The *ambiguous* expert matters most on ICEWS18 (-2.61), where rich but conflicting histories cannot be resolved by count alone, while the *sparse* expert yields a smaller yet consistent drop, peaking at -3.79 on GDELT. The three inductive biases are thus complementary.

(3) Effect of training signals and LLM refinement. Ablating $\mathcal{L}_{distill}$ reduces ICEWS14 MRR by 2.29, confirming its role in type-conditioned specialisation; ablating $\mathcal{L}_{uniform}$ reduces ICEWS05-15 MRR by 2.99, demonstrating its importance in preserving competence among low-responsibility experts. Removing LLM refinement consistently degrades performance, most notably on GDELT (-2.72), where boundary queries are more frequent, while the refinement incurs no inference cost.

(4) Effect of the parameter budget. QTI-MoE runs three backbones, so a natural question is whether its gains simply follow from a larger model. We build two parameter-only controls that keep the backbones but drop the query-type-aware design: *Ens-3* averages the predictions of the same three encoders with uniform weights, and *Ens-6* does the same with six. Table 4 shows that QTI-MoE outperforms Ens-3 on every metric at an identical parameter budget, and that Ens-6 brings no further gain despite twice the parameters. The advantage of QTI-MoE therefore comes from how experts are specialised and routed, rather than from how many of them are available.

4.6 Type-Stratified Expert Analysis

We compare the three experts and the routed QTI-MoE on type-specific subsets defined by percentile thresholds over the statistics in Section 3.1. The sparse and ambiguous subsets take the top $X\%$ of queries by fewest historical events and highest candidate diversity respectively, and the recurrent subset takes the middle $X\%$ of the diversity band. Sweeping X traces each expert as the subset narrows toward the extreme end of each regime. Figure 3 reports H@1 on ICEWS14 and ICEWS18; full results on the four datasets are in Figures 8–11. Two observations emerge:

(1) Per-regime leadership with MoE consolidation. The matched expert consistently leads its

Scenario	Method	Gate	Top-5 prediction	Rank
Sparse (2018-10-18)	Sparse	1.00	Nyamdorj* , Ministry (Vietnam), Ministry (South Korea), Matia Kasaija, Guinea	1
Ministry (Vietnam)	Recurrent	0.00	Mongolia, Ministry (Vietnam), Vietnam, Nigeria, Government (Indonesia)	48
–[Sign formal agreement (inv)]– ?	Ambiguous	0.00	Vietnam, Ministry (Vietnam), Government (Indonesia), European Union, Botswana	20
True tail: Nyamdorj	QTI-MoE	–	Nyamdorj* , Ministry (Vietnam), Ministry (South Korea), Matia Kasaija, Mongolia	1
Recurrent (2018-10-03)	Sparse	0.00	Rodrigo Duterte, Government..., Other Authorities..., Police (Philippines), City Mayor...	12
President of the...	Recurrent	0.99	Senate (Philippines)* , Congress (Philippines), President of the..., Liberal Party..., Ruling Parties...	1
–[Make statement]– ?	Ambiguous	0.01	Rodrigo Duterte, Government..., Philippines, City Mayor..., Police (Philippines)	6
True tail: Senate (Philippines)	QTI-MoE	–	Senate (Philippines)* , Congress (Philippines), Rodrigo Duterte, President of the..., Liberal Party...	1
Ambiguous (2018-10-30)	Sparse	0.00	Central Committee of..., Walid Muallem, Head of Government..., UN General Assembly, Scholar (United...	21
China	Recurrent	0.35	Head of Government..., Cui Tiankai, Foreign Affairs... , Huanqiu Shibao, Li Keqiang	3
–[Make statement (inv)]– ?	Ambiguous	0.65	Foreign Affairs... , Donald Trump, Media Personnel..., United States, Xi Jinping	1
True tail: Foreign Affairs...	QTI-MoE	–	Foreign Affairs... , Head of Government..., Media Personnel..., Donald Trump, Xi Jinping	1

Table 5: Case study on three representative queries. The ground-truth tail is marked with **bold***. Gate columns report the soft assignment produced by the gating network.

own column across all thresholds on both datasets, confirming that the three memory pathways faithfully carry their intended inductive biases, while the QTI-MoE curve stays at least on par with the focal expert in every focal column and occasionally surpasses all single experts, showing that the router accurately identifies suitable experts and exploits their complementarity under mixed regimes.

(2) **Specialization sharpens at the extreme ends.** The gap between the focal expert and the other two widens steadily as the subset narrows; on the ICEWS14 ambiguous column the focal lead is about 0.08 H@1 at $X=0.5\%$ but gradually converges at $X=15\%$, and on ICEWS18 the sparse expert is the only one with non-trivial accuracy at the tightest sparse threshold, mirroring the inverted-U observation in Figure 1b and showing that the three experts recover accuracy at exactly the low-frequency and high-diversity ends where a monolithic pipeline visibly degrades.

4.7 Case Study

To illustrate how QTI-MoE handles different regimes, we examine three queries from ICEWS18, one per type. Table 5 reports the gate, top-5 predictions, and gold rank; a further set of cases with full histories is in Figure 7 of the appendix. Three observations follow: (1) **experts behave differently across regimes**, since only the matched expert places the gold at rank 1, while the other two push it to ranks 20–48 by over-relying on frequent entities such as *Vietnam*; (2) **the router selects the right expert**, with gate values reaching 1.00, 0.99, and 0.65 on the focal expert of the sparse, recurrent, and ambiguous query, indicating that historical shape is read from $(\mathbf{h}_s, \mathbf{e}_r)$ alone; (3) **the mixture refines rather than duplicates**, as QTI-MoE matches the rank-1 of the focal expert in all cases and reorganises the rest, e.g. keeping the gold

from the ambiguous expert while inheriting *Head of Government* from the recurrent one, yielding a list no single expert produces.

5 Related Work

Structure-based TKG Forecasting: Early work on TKG forecasting models the structural evolution of the graph by combining relation-aware graph convolutions with recurrent networks (Jin et al., 2020; Li et al., 2022a), and later studies further strengthen the encoder through historical contrastive learning (Chen et al., 2024; Xu et al., 2023) or structural matching paradigms (Li et al., 2022b; Liu et al., 2023).

A few of these methods implicitly recognize that queries differ in their historical behaviour, for instance by separating repeating facts from newly emerging ones (Xu et al., 2023; Ma et al., 2025). The same encoder and decoder are still used for every query at inference, so the model has limited capacity to adjust when query patterns vary across the event stream.

LLM-based Generative Forecasting: With the rapid progress of Large Language Models (LLMs), a second line of work introduces LLMs into TKG forecasting. TKG-ICL (Lee et al., 2023) casts forecasting as in-context learning over linearised event histories, and CoH (Xia et al., 2024) further reasons over chains of historical events to expose temporal dependencies to the LLM. To reduce inference cost and adapt the LLM to scenario-specific knowledge, retrieval-augmented and parameter-efficient frameworks such as GenTKG (Liao et al., 2024) and G2S (Bai et al., 2025) have also been developed. These methods benefit from the LLM’s reasoning ability, but often underperform on queries that rely heavily on graph topology (Chen et al., 2025), and applying a single generative paradigm to all inputs leads to

unstable behaviour on structure-dominated cases.

A third line of work attempts to fuse structural and semantic signals within a single architecture, either by using LLMs to extract logical priors for graph-based scoring (Zhang et al., 2025) or through Mixture-of-Experts (MoE) designs (Mu and Lin, 2025). The closest example is MESH (Deng et al., 2025), which pairs a structural expert with a semantic one and routes between them according to the type of evidence each query carries. Such routing operates within a query, choosing which signal to trust, but treats queries as interchangeable. Our work routes across queries instead: each query is associated with a soft distribution over three historical regimes, and three experts with regime-matched inductive biases are mixed accordingly.

6 Conclusion

This paper studies TKG forecasting and observes that queries with the same predicate form can rest on very different histories, falling into three regimes: *sparse*, *recurrent*, and *ambiguous*. A stratified analysis of RE-GCN and LogCL on ICEWS14 shows an inverted-U accuracy curve across event frequency and candidate diversity, motivating regime-specific inductive biases. Based on this analysis, we propose **QTI-MoE**, which assigns each query a soft label from rule-based statistics with offline LLM refinement on boundary cases, and routes it to one of three experts with complementary memory pathways: recency-weighted aggregation, count-based repetition, and diversity-normalised scoring. QTI-MoE achieves state-of-the-art results on four standard benchmarks, with the largest margins on sparser and more diverse datasets, and type-stratified analyses confirm that each expert peaks on its target regime.

Limitations

While QTI-MoE delivers consistent gains across four benchmarks, several limitations remain. First, our method discretises a continuous spectrum of historical shapes into three query types. Soft labels and LLM refinement smooth the boundaries between sparse, recurrent, and ambiguous queries, but other historical patterns are absorbed into the closest of the three; extending the type set or moving towards a continuous typing scheme is left to future work. Second, although the LLM is invoked only offline and the inference stage carries no LLM cost, soft label construction still requires

running a frozen LLM over the boundary subset of the training set, which may be less convenient in resource-limited settings. Third, our evaluation follows the standard transductive protocol on four widely used benchmarks in which test entities are observed during training; how regime-specialised experts behave under inductive settings with emerging entities is left for future investigation.

Ethics Statement

This work complies with the ACL Ethics Policy. We use ICEWS and GDELT, two public datasets of coded political events involving public actors rather than private individuals, so no additional anonymisation was needed. Their event coding follows news coverage and inherits its biases, and forecasting over conflict-related events carries dual-use potential, so our released code is intended for benchmark evaluation only. AI assistants were used solely for language polishing.

Acknowledgement

This work is mainly supported by the Tianshan Talent Training Program (Grant No. 2023TSY-CCX0041 and 2023TSYCCX0044), in part by the Key Research and Development Program of Xinjiang Uyghur Autonomous Region (Grant No. 2023B03024, 2024B03026 and 2025B03042), in part by the Science and Technology Innovation Program of the Chinese Academy of Sciences (Grant No. KGFZD-145-26-44), in part by the Natural Science Foundation of Xinjiang Uyghur Autonomous Region (Grant No. 2023D01D17 and 2024D01D29), in part by the Guidance on Local Science and Technology Development Fund of Xinjiang Uygur Autonomous Region (Grant No. ZYYD2026JD03).

References

- Long Bai, Zixuan Li, Xiaolong Jin, Jiafeng Guo, Xueqi Cheng, and Tat-Seng Chua. 2025. [G2S: A general-to-specific learning framework for temporal knowledge graph forecasting with large language models](#). In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, volume ACL 2025 of *Findings of ACL*, pages 20927–20938. Association for Computational Linguistics.
- Elizabeth Boschee, Jennifer Lautenschlager, Sean O’Brien, Steve Shellman, James Starz, and Michael Ward. 2015. [ICEWS Coded Event Data](#).

- Li Cai, Xin Mao, Yuhao Zhou, Zhaoguang Long, Changxu Wu, and Man Lan. 2024a. [A survey on temporal knowledge graph: Representation learning and applications](#). *CoRR*, abs/2403.04782.
- Yuxiang Cai, Qiao Liu, Yanglei Gan, Changlin Li, Xueyi Liu, Run Lin, Da Luo, and JiayeYang JiayeYang. 2024b. [Predicting the unpredictable: Uncertainty-aware reasoning over temporal knowledge graphs via diffusion process](#). In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, volume ACL 2024 of *Findings of ACL*, pages 5766–5778. Association for Computational Linguistics.
- Wei Chen, Huaiyu Wan, Yuting Wu, Shuyuan Zhao, Jiayao Cheng, Yuxin Li, and Youfang Lin. 2024. [Local-global history-aware contrastive learning for temporal knowledge graph reasoning](#). In *40th IEEE International Conference on Data Engineering, ICDE 2024, Utrecht, The Netherlands, May 13-16, 2024*, pages 733–746. IEEE.
- Wei Chen, Yuting Wu, Shuhan Wu, Zhiyu Zhang, Mengqi Liao, Youfang Lin, and Huaiyu Wan. 2025. [Cogntke: A cognitive temporal knowledge extrapolation framework](#). In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 14815–14823. AAAI Press.
- Yimin Deng, Yuxia Wu, Yejing Wang, Guoshuai Zhao, Li Zhu, Qidong Liu, Derong Xu, Zichuan Fu, Xian Wu, Yefeng Zheng, Xiangyu Zhao, and Xueming Qian. 2025. [A multi-expert structural-semantic hybrid framework for unveiling historical patterns in temporal knowledge graphs](#). In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, volume ACL 2025 of *Findings of ACL*, pages 20553–20565. Association for Computational Linguistics.
- Hao Dong, Ziyue Qiao, Zhiyuan Ning, Qi Hao, Yi Du, Pengyang Wang, and Yuanchun Zhou. 2025. [Disentangled multi-span evolutionary network against temporal knowledge graph reasoning](#). In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, volume ACL 2025 of *Findings of ACL*, pages 12697–12707. Association for Computational Linguistics.
- Alberto García-Durán, Sebastijan Dumancic, and Mathias Niepert. 2018. [Learning sequence encoders for temporal knowledge graph completion](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4816–4821. Association for Computational Linguistics.
- Woojeong Jin, Meng Qu, Xisen Jin, and Xiang Ren. 2020. [Recurrent event network: Autoregressive structure inference over temporal knowledge graphs](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6669–6683. Association for Computational Linguistics.
- Diederik P. Kingma and Jimmy Ba. 2015. [Adam: A method for stochastic optimization](#). In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Dong-Ho Lee, Kian Ahrabian, Woojeong Jin, Fred Morstatter, and Jay Pujara. 2023. [Temporal knowledge graph forecasting without knowledge using in-context learning](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 544–557. Association for Computational Linguistics.
- Yujia Li, Shiliang Sun, and Jing Zhao. 2022a. [Tirgn: Time-guided recurrent graph network with local-global historical patterns for temporal knowledge graph reasoning](#). In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 2152–2158. ijcai.org.
- Zixuan Li, Zhongni Hou, Saiping Guan, Xiaolong Jin, Weihua Peng, Long Bai, Yajuan Lyu, Wei Li, Jiafeng Guo, and Xueqi Cheng. 2022b. [Hismatch: Historical structure matching based temporal knowledge graph reasoning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, volume EMNLP 2022 of *Findings of ACL*, pages 7328–7338. Association for Computational Linguistics.
- Zixuan Li, Xiaolong Jin, Wei Li, Saiping Guan, Jiafeng Guo, Huawei Shen, Yuanzhuo Wang, and Xueqi Cheng. 2021. [Temporal knowledge graph reasoning based on evolutionary representation learning](#). In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 408–417. ACM.
- Ruotong Liao, Xu Jia, Yangzhe Li, Yunpu Ma, and Volker Tresp. 2024. [Gentkg: Generative forecasting on temporal knowledge graph with large language models](#). In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, volume NAACL 2024 of *Findings of ACL*, pages 4303–4317. Association for Computational Linguistics.
- Kangzheng Liu, Feng Zhao, Guandong Xu, Xianzhi Wang, and Hai Jin. 2023. [RETIA: relation-entity twin-interact aggregation for temporal knowledge graph extrapolation](#). In *39th IEEE International Conference on Data Engineering, ICDE 2023, Anaheim, CA, USA, April 3-7, 2023*, pages 1761–1774. IEEE.
- Rong Ma, Lei Wang, Yating Yang, Bo Ma, Rui Dong, Fengyi Yang, Ahtamjan Ahmat, Kaiwen Lu, and

- Xinyue Wang. 2025. [Mining the past with dual criteria: Integrating three types of historical information for context-aware event forecasting](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 8148–8163. Association for Computational Linguistics.
- Mehrnoosh Mirtaheri, Mohammad Rostami, and Aram Galstyan. 2023. [History repeats: Overcoming catastrophic forgetting for event-centric temporal knowledge graph completion](#). In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, volume ACL 2023 of *Findings of ACL*, pages 7740–7755. Association for Computational Linguistics.
- Siyuan Mu and Sen Lin. 2025. [A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications](#). *CoRR*, abs/2503.07137.
- Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. [Modeling relational data with graph convolutional networks](#). In *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, volume 10843 of *Lecture Notes in Computer Science*, pages 593–607. Springer.
- Chao Shang, Yun Tang, Jing Huang, Jinbo Bi, Xiaodong He, and Bowen Zhou. 2019. [End-to-end structure-aware convolutional networks for knowledge base completion](#). In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3060–3067. AAAI Press.
- Guo Tang, Zheng Chu, Wenxiang Zheng, Junjie Xiang, Yizhuo Li, Weihao Zhang, Ming Liu, and Bing Qin. 2025. [Anre: Analogical replay for temporal knowledge graph forecasting](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 4632–4650. Association for Computational Linguistics.
- Jiayu Wang, Kai Sun, Linhao Luo, Wei Wei, Yongli Hu, Alan Wee-Chung Liew, Shirui Pan, and Baocai Yin. 2024. [Large language models-guided dynamic adaptation for temporal knowledge graph reasoning](#). In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Jiayu Wang, Boyue Wang, Meikang Qiu, Shirui Pan, Bo Xiong, Heng Liu, Linhao Luo, Tengfei Liu, Yongli Hu, Baocai Yin, and Wen Gao. 2023. [A survey on temporal knowledge graph completion: Taxonomy, progress, and prospects](#). *CoRR*, abs/2308.02457.
- Yuwei Xia, Ding Wang, Qiang Liu, Liang Wang, Shu Wu, and Xiaoyu Zhang. 2024. [Chain-of-history reasoning for temporal knowledge graph forecasting](#). In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, volume ACL 2024 of *Findings of ACL*, pages 16144–16159. Association for Computational Linguistics.
- Yi Xu, Luoyi Fu, and Xinbing Wang. 2025. [Rethinking temporal knowledge graph extrapolation: prioritizing historical events over graph](#). *Frontiers Comput. Sci.*, 19(11):1911366.
- Yi Xu, Junjie Ou, Hui Xu, and Luoyi Fu. 2023. [Temporal knowledge graph reasoning with historical contrastive learning](#). In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 4765–4773. AAAI Press.
- Fu Zhang, Panfeng Zhang, and Jingwei Cheng. 2025. [SRM-LLM: semantic relationship mining with llms for temporal knowledge graph extrapolation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 16007–16021. Association for Computational Linguistics.
- Ze Zhao, Yuhui He, Lyuwen Wu, Gu Tang, Bin Lu, Xiaoying Gan, Luoyi Fu, Xinbing Wang, and Chenghu Zhou. 2026. [Inductive reasoning for temporal knowledge graphs with emerging entities](#). *CoRR*, abs/2604.10164.

A Appendix

The appendix provides supplementary experiments, specifications, and analyses. Sections A.1–A.4 extend the experimental results. Section A.1 gives the full ablation under MRR and Hits@{1, 3, 10}. Section A.2 varies the refinement stage, replacing InternLM2-7B with two other LLMs and with a lightweight MLP classifier. Section A.3 compares training and testing cost against representative baselines. Section A.4 sweeps the fusion weight w and the two loss weights λ_d and λ_u . Sections B and C specify the model: the shared R-GCN, GRU, and ConvTransE encoder held by each expert, and the three-stage training and inference procedure. The remaining sections analyse the trained model. Section C.1 presents an extended case study with the historical events behind each routing decision. Section C.2 details the soft label construction and shows that the labels remain effective without the privileged count g_{sr}^* . Section C.3 reports the complete type-stratified sweep. Section C.4 verifies

that each of the three query types is necessary, and Section C.5 examines whether the router responds to historical shape rather than to pair identity.

A.1 Full Ablation Results

Table 6 reports the full ablation results on the four benchmarks under MRR and Hits@{1, 3, 10}, extending the simplified view in Table 3. The three patterns identified in Section 4.5 hold consistently across Hits@3 and Hits@10. The *recurrent* expert is again the most critical component, with Hits@10 falling by 2.54 on ICEWS14 and 6.82 on GDELT once it is removed. The *ambiguous* expert still produces the largest drop on ICEWS18 among the three experts, and the *sparse* expert peaks on GDELT with a Hits@10 drop of 5.71. On the training signals, $\mathcal{L}_{distill}$ shows its strongest effect on ICEWS14 with a Hits@3 drop of 2.36, while $\mathcal{L}_{uniform}$ matters most on ICEWS05-15 with a Hits@3 drop of 3.00, both in line with their assigned roles. Removing LLM refinement produces a 4.19 Hits@10 drop on GDELT, the largest among the four benchmarks, matching the higher prevalence of boundary queries on that dataset. The only cell where QTI-MoE is not the best is Hits@10 on ICEWS05-15, where *w/o MoE* exceeds the full model by 0.10. This margin is well within the variation observed across other settings and does not affect the overall trend.

A.2 Effect of the Refinement LLM

To check whether QTI-MoE depends on a specific LLM for soft-label refinement, we replace InternLM2-7B² with Qwen2.5-7B-Instruct³ and Llama-3.1-8B-Instruct⁴, keeping all other settings identical. As shown in Figure 4, the three choices yield closely matched MRR on all three datasets, with the spread staying within 0.10 on ICEWS18 and 1.44 on ICEWS05-15. The variation is also smaller than the drop observed when the LLM stage is removed entirely (Table 3), which is consistent with our design where the LLM only adjusts a small portion of boundary queries while rule-based labels remain dominant.

Replacing the LLM with a lightweight classifier. We further ask whether an LLM is needed at all. We train a small MLP classifier that takes

²<https://huggingface.co/internlm/internlm2-7b>

³<https://huggingface.co/Qwen/Qwen2.5-7B-Instruct>

⁴<https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>

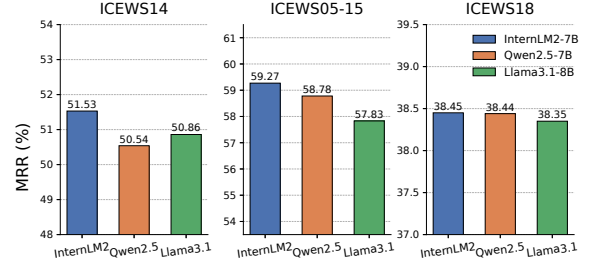


Figure 4: MRR of QTI-MoE on three ICEWS benchmarks when the refinement LLM is replaced by Qwen2.5-7B and Llama3.1-8B.

the same prefix statistics as the rule and outputs a three-way type distribution, fit on the confidently labelled non-boundary pairs, and use it to refine the same boundary subset $\mathcal{B}$ under the same fusion weight w ; all other settings are unchanged. Table 7 shows that the MLP recovers most of the benefit, so it is a practical substitute when running an LLM is inconvenient. The LLM nevertheless keeps a small and consistent margin on every metric, and together with the *w/o LLM* row of Table 6 the three settings order monotonically: no refinement, MLP refinement, LLM refinement. The gap is what one would expect from the inputs each sees. The MLP reads only the numeric statistics that the rule is already built from, whereas the LLM also reads the relation surface form and can bring semantic priors to bear on boundary pairs. Since labelling is a one-time offline cost, we keep the LLM as the default and report the MLP as the cheaper alternative.

A.3 Inference Time Analysis

We further compare per-epoch training cost on ICEWS14 against four representative baselines, with results reported in Table 8. The numbers for HTCCN, LogCL, RETIA, and SRM-LLM are taken from Zhang et al. (2025), who measured all four methods on the same hardware setup.

QTI-MoE takes about 10 minutes per training epoch on ICEWS14, which sits in the middle of the four baselines. It is roughly three times slower than LogCL, consistent with running three structural backbones in parallel, but remains faster than SRM-LLM and almost an order of magnitude faster than RETIA. The full training run finishes within a few hours on a single card, and the offline soft-label construction is a one-time cost that does not enter the per-epoch budget. Taken together with the inference-time results above, this places the overall training and deployment cost of QTI-MoE on the

Model	ICEWS14				ICEWS05-15				ICEWS18				GDELТ			
	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
QTI-MoE	51.53	40.48	57.68	72.86	59.27	48.82	65.75	78.89	38.45	27.07	43.39	61.07	27.53	17.46	30.34	48.36
w/o MoE	50.79	39.78	56.61	72.49	58.49	47.63	65.33	78.99	37.57	26.14	42.49	60.17	24.24	14.85	26.22	43.54
w/o sparse	50.09	38.89	56.00	71.90	58.14	47.39	64.65	78.43	37.03	25.67	41.87	59.58	23.74	14.47	25.64	42.65
w/o recurrent	47.73	36.30	53.72	70.32	56.03	44.78	62.89	77.41	35.62	24.41	40.19	58.03	23.13	14.09	24.85	41.54
w/o ambiguous	49.43	38.22	55.28	71.26	56.95	45.81	63.84	78.14	35.84	24.37	40.60	58.59	23.28	14.10	25.05	42.07
w/o $\mathcal{L}_{distill}$	49.24	37.95	55.32	71.32	57.81	46.96	64.75	78.06	37.76	26.32	42.51	60.26	27.15	16.97	30.04	48.26
w/o $\mathcal{L}_{uniform}$	50.27	39.14	56.47	71.54	56.28	45.58	62.75	76.62	37.12	25.92	41.93	59.36	26.58	16.46	29.39	47.87
w/o LLM	50.38	38.98	56.66	72.48	57.65	46.85	64.43	78.04	37.75	26.39	42.51	60.40	24.81	15.36	26.91	44.17

Table 6: Ablation study results on ICEWS14, ICEWS05-15, ICEWS18 and GDELТ. All metrics are reported in percentage (%).

Dataset	Refinement	MRR	H@1	H@3	H@10
ICEWS14	LLM	51.53	40.48	57.68	72.86
	MLP	51.13	39.81	57.37	72.66
ICEWS05-15	LLM	59.27	48.82	65.75	78.89
	MLP	58.40	47.76	65.02	78.33
ICEWS18	LLM	38.45	27.07	43.39	61.07
	MLP	38.11	26.60	43.10	60.98

Table 7: Refining the boundary subset with the frozen LLM against a lightweight MLP classifier over the same prefix statistics (%). All other settings are identical.

Model	Train (per epoch)	Test (per epoch)
HTCCN	5.93 min	1.05 min
LogCL	3.33 min	0.18 min
RETIA	93.68 min	8.78 min
SRM-LLM	13.1 min	1.32 min
QTI-MoE (ours)	10.15 min	0.46 min

Table 8: Per-epoch training and testing time on ICEWS14. Baseline numbers are taken from Zhang et al. (2025).

same order as recent structure-based forecasters.

A.4 Sensitivity to Key Hyperparameters

QTI-MoE has three hyperparameters that shape label construction and expert specialisation. The fusion weight w controls how strongly the LLM refines rule-based soft labels on boundary queries, while λ_d and λ_u balance type-conditioned distillation and uniform supervision in Eq. 9. We sweep each one while keeping the others at their default values, and report MRR on the four benchmarks.

Effect of w . Figure 5 reports the four metrics on ICEWS14 as w varies from 0.1 to 1.0. All curves remain nearly flat, with MRR moving within 1.5 points and Hits@10 within 1.1 points over the full range. The peak sits at $w=0.2$ on every metric.

This is in line with how the refinement stage is designed. The LLM is invoked only on boundary queries, so a small fusion weight is enough to correct ambiguous cases without overriding the rule-based labels on the remaining majority. Larger w propagates the LLM’s noise to confident queries and produces a mild but consistent drop.

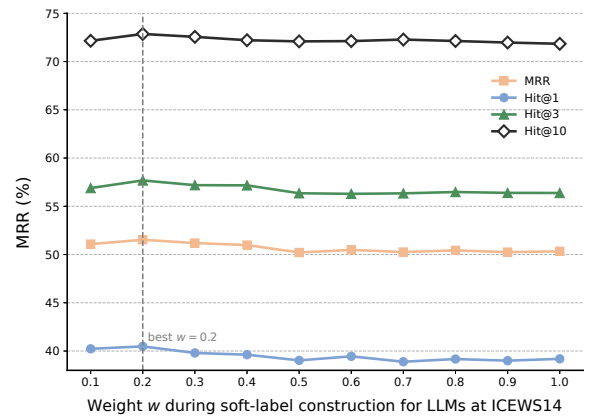


Figure 5: Sensitivity of QTI-MoE to the rule-LLM fusion weight w in Eq. 2, evaluated on ICEWS14 across MRR and Hits@{1,3,10}.

Effect of λ_d . Panel (a) of Figure 6 sweeps λ_d over $\{0.2, 0.4, 0.6, 0.8, 1.0\}$. The peak is reached at $\lambda_d=0.6$ on all four datasets, with the total spread reaching 1.63 MRR on ICEWS14 and staying within 0.5 MRR on the other three. A small λ_d weakens the soft-label signal and reduces type-conditioned specialisation, while a large λ_d overrides $\mathcal{L}_{pred}$ and hurts the joint optimisation of the router and the experts. The intermediate value therefore matches the role that $\mathcal{L}_{distill}$ is intended to play, providing strong but bounded type guidance.

Effect of λ_u . Panel (b) sweeps λ_u under the same protocol. The optimum lies at the centre of the swept range on three of the four benchmarks, and ICEWS18 stays flat across all five values. A small λ_u lets non-dominant experts gradually lose competence, since few training queries route gradient through them. A large λ_u dilutes the type-specific gradient injected by $\mathcal{L}_{distill}$ and pushes the three experts toward homogeneity. The intermediate setting therefore preserves baseline competence on every expert without erasing the specialisation that distillation builds up.

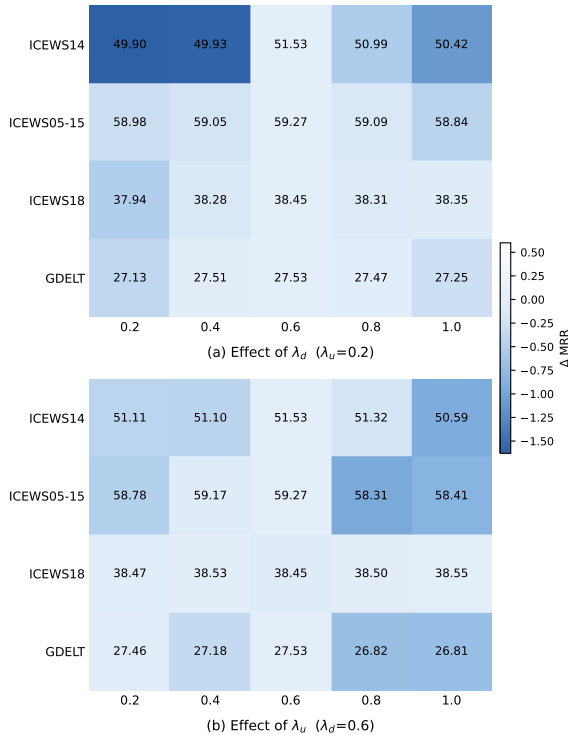


Figure 6: Sensitivity of QTI-MoE to the two loss weights in Eq. 9 on four benchmarks. (a) varies λ_d with λ_u fixed, and (b) varies λ_u with λ_d fixed. Cell values are MRR (%), and colours encode the per-row gap to the row maximum.

B Shared Structural Encoder

We provide the full specification of the structural encoder used by each expert $\mathcal{E}_\tau$ in Section 3.2. The three experts share the same encoder design but instantiate independent parameters. The description below therefore applies to each $\tau \in \{sp, re, am\}$ with its own copy of weights Θ_τ and embedding tables ($\mathbf{E}_\tau, \mathbf{R}_\tau$). Our encoder is built on LogCL (Chen et al., 2024), from which we retain the R-GCN + GRU snapshot encoder and the ConvTransE decoder (Shang et al., 2019); LogCL itself follows

the recurrent graph paradigm popularised by RE-GCN (Li et al., 2021). We reorganise these components into a per-expert pathway, so that each expert produces its own structural distribution $\mathbf{p}_\tau^{str}$ that feeds the gated mixture in Equation 6.

Per-snapshot graph encoding. At each snapshot G_t within the history window $\{G_{t_q-L}, \dots, G_{t_q-1}\}$, expert $\mathcal{E}_\tau$ applies a two-layer R-GCN (Schlichtkrull et al., 2018) to obtain entity representations $\mathbf{H}_t^{\tau,(2)}$. Messages are passed along each typed edge (s, r, o) in G_t using the relation embedding $\mathbf{R}_\tau$, so that the resulting node states encode the structure observed at that snapshot. The input to the R-GCN at snapshot t is the recurrent state $\mathbf{H}_{t-1}^\tau$ produced by the GRU update below, with $\mathbf{H}_{t_q-L-1}^\tau$ initialised to the persistent entity table $\mathbf{E}_\tau$.

Recurrent update across snapshots. The per-snapshot states are fused along the history window with a GRU cell:

$$\mathbf{H}_t^\tau = GRU_\tau(\mathbf{H}_t^{\tau,(2)}, \mathbf{H}_{t-1}^\tau) \quad t = t_q-L, \dots, t_q-1 \quad (11)$$

For a query $(s, r, ?, t_q)$, the expert reads its subject and relation representations from the last hidden state in the window:

$$\mathbf{h}_s^\tau = \mathbf{H}_{t_q-1}^\tau[s], \quad \mathbf{e}_r^\tau = \mathbf{R}_\tau[r] \quad (12)$$

and uses them throughout the gated mixture of Section 3.2.

ConvTransE decoder. The structural distribution $\mathbf{p}_\tau^{str} \in R^{|\mathcal{V}|}$ is produced by a ConvTransE decoder (Shang et al., 2019) that is also held private to the expert. Stacking $\mathbf{h}_s^\tau$ and $\mathbf{e}_r^\tau$ into a two-row matrix, the decoder applies one-dimensional convolutions followed by a feed-forward projection, and then takes an inner product against the final entity matrix $\mathbf{H}_{t_q-1}^\tau$,

$$\mathbf{p}_\tau^{str} = \text{softmax}(\text{ConvTransE}_\tau(\mathbf{h}_s^\tau, \mathbf{e}_r^\tau) \mathbf{H}_{t_q-1}^{\tau \top}) \quad (13)$$

The resulting distribution enters the gate of Equation 6, where it is mixed with the type-specific memory distribution $\mathbf{p}_\tau^{mem}$ to form the expert prediction $\mathbf{p}_\tau$.

Forward pass summary. Algorithm 1 traces the forward pass of one expert from the history window to the structural distribution. The full QTI-MoE forward pass invokes this procedure once per expert and routes the three resulting distributions through Equations 7 and 8.

Algorithm 1 Shared Structural Encoder of Expert $\mathcal{E}_\tau$

Require: Query $(s, r, ?, t_q)$; history window $\{G_{t_q-L}, \dots, G_{t_q-1}\}$; embeddings $\mathbf{E}_\tau, \mathbf{R}_\tau$; backbone parameters Θ_τ

Ensure: $\mathbf{h}_s^\tau, \mathbf{e}_r^\tau, \mathbf{p}_\tau^{str}$

- 1: $\mathbf{H}_{t_q-L-1}^\tau \leftarrow \mathbf{E}_\tau$
- 2: **for** $t = t_q - L$ **to** $t_q - 1$ **do**
- 3: $\mathbf{H}_t^{\tau, (2)} \leftarrow R-GCN_\tau(G_t, \mathbf{H}_{t-1}^\tau, \mathbf{R}_\tau)$
- 4: $\mathbf{H}_t^\tau \leftarrow GRU_\tau(\mathbf{H}_t^{\tau, (2)}, \mathbf{H}_{t-1}^\tau)$ (Eq. 11)
- 5: **end for**
- 6: $\mathbf{h}_s^\tau \leftarrow \mathbf{H}_{t_q-1}^\tau[s], \mathbf{e}_r^\tau \leftarrow \mathbf{R}_\tau[r]$ (Eq. 12)
- 7: $\mathbf{p}_\tau^{str} \leftarrow \text{softmax}(\text{ConvTransE}_\tau(\mathbf{h}_s^\tau, \mathbf{e}_r^\tau) \mathbf{H}_{t_q-1}^{\tau \top})$ (Eq. 13)
- 8: **return** $\mathbf{h}_s^\tau, \mathbf{e}_r^\tau, \mathbf{p}_\tau^{str}$

C Training and Inference of QTI-MoE

We summarise the full QTI-MoE procedure in Algorithm 2. The procedure proceeds in three stages, all specified in Section 3.1 and Section 3.3.

Stage 0 constructs the query-type soft labels offline. For each pair (s, r) in the training set, the prefix statistics $(f_{sr}, d_{sr}, g_{sr}^*)$ yield the rule distribution $\hat{\mathbf{q}}_{rule}$, which is refined by a frozen LLM on the boundary subset through Equations 1 and 2. The resulting $\hat{\mathbf{q}}$ is fixed throughout the rest of the procedure.

Stage 1 jointly trains the three experts and the router. Each expert produces its structural distribution through Algorithm 1, combines it with the type-specific memory distribution via the gate in Equation 6, and contributes to the routed prediction in Equation 8. The composite objective of Equation 9 is then minimised end-to-end.

Stage 2 performs inference on the test set. The router computes π from the query embeddings alone, each expert is evaluated on the visible history prefix, and the routed mixture is read out without any LLM call or soft-label access.

C.1 Extended Case Study with Historical Events

Figure 7 presents a second, more self-contained set of case studies on ICEWS18, with one query per regime drawn independently of Table 5. Each panel lists the historical events of the subject up to T_{n-1} , the gate vector produced by the router, the top-5 prediction of every single expert, and the rank of the gold tail under each expert as well as under the routed QTI-MoE. The history block makes the shape of the prefix directly visible, so the gate values can be read alongside the evidence that the router is actually responding to.

Algorithm 2 Training and Inference of QTI-MoE

Require: Training set $\mathcal{D}_{train}$; test set $\mathcal{D}_{test}$; frozen LLM; weights λ_d, λ_u, w

Ensure: Trained parameters Θ^* and test-set predictions

- 1: **Stage 0: Offline Soft Label Construction**
- 2: **for** each pair (s, r) in $\mathcal{D}_{train}$ **do**
- 3: Build $\hat{\mathbf{q}}_{rule}$ from $(f_{sr}, d_{sr}, g_{sr}^*)$
- 4: **end for**
- 5: Refine $\hat{\mathbf{q}}$ on boundary subset $\mathcal{B}$ via LLM (Eq. 1, 2)
- 6: **Stage 1: Joint Training of Experts and Router**
- 7: Initialise $\{\Theta_\tau\}_{\tau \in \{sp, re, am\}}$ and router
- 8: **while** not converged **do**
- 9: **for** each query $(s, r, o^*, t_q) \in \mathcal{D}_{train}$ **do**
- 10: **for** $\tau \in \{sp, re, am\}$ **do**
- 11: Run Algorithm 1, obtain $\mathbf{p}_\tau^{str}$
- 12: Compute $\mathbf{p}_\tau^{mem}$ via memory pathway (Eq. 3–5)
- 13: Fuse to $\mathbf{p}_\tau$ via gate α_τ (Eq. 6)
- 14: **end for**
- 15: Compute routed mixture $\mathbf{p}_{final}$ (Eq. 7, 8)
- 16: Compute $\mathcal{L}$ (Eq. 9) and update parameters
- 17: **end for**
- 18: **end while**
- 19: **Stage 2: Inference on $\mathcal{D}_{test}$**
- 20: **for** each query $(s, r, ?, t_q) \in \mathcal{D}_{test}$ **do**
- 21: Compute $\mathbf{p}_\tau$ for each τ via Algorithm 1 and memory pathway
- 22: Predict $\mathbf{p}_{final}$ (no LLM, no soft label)
- 23: **end for**
- 24: **return** Θ^* and predictions

The sparse panel shows Naval (Vietnam) with no prior events on the subject, and the router places full mass on the sparse expert, which is the only one that ranks the gold tail *Naval Base (Japan)* first while the other two return generic country-level candidates. The recurrent panel shows Salva Kiir with several recent interactions on *Riek Machar* and related entities, and assigns 0.77 of the gate weight to the recurrent expert, which lifts the gold *Citizen (South Sudan)* from rank 3 under the sparse expert to rank 1. The ambiguous panel shows Iran with a dense and heterogeneous prefix covering diplomats, governments and domestic actors, and the router produces a split gate of 0.58 and 0.42 across the ambiguous and recurrent experts, which together recover *Hassan Rouhani* at rank 1 while the sparse expert collapses to rank 28. Across the three panels, the routed top-5 inherits the focal expert’s rank-1 entity and rearranges the lower ranks by borrowing from the secondary expert, mirroring the behaviour described in the main text.

C.2 LLM-based Soft Label Construction

This subsection completes the description of the soft labelling procedure introduced in §3.1. We specify (i) when the frozen LLM is invoked, (ii) what prompt it receives, and (iii) how its output

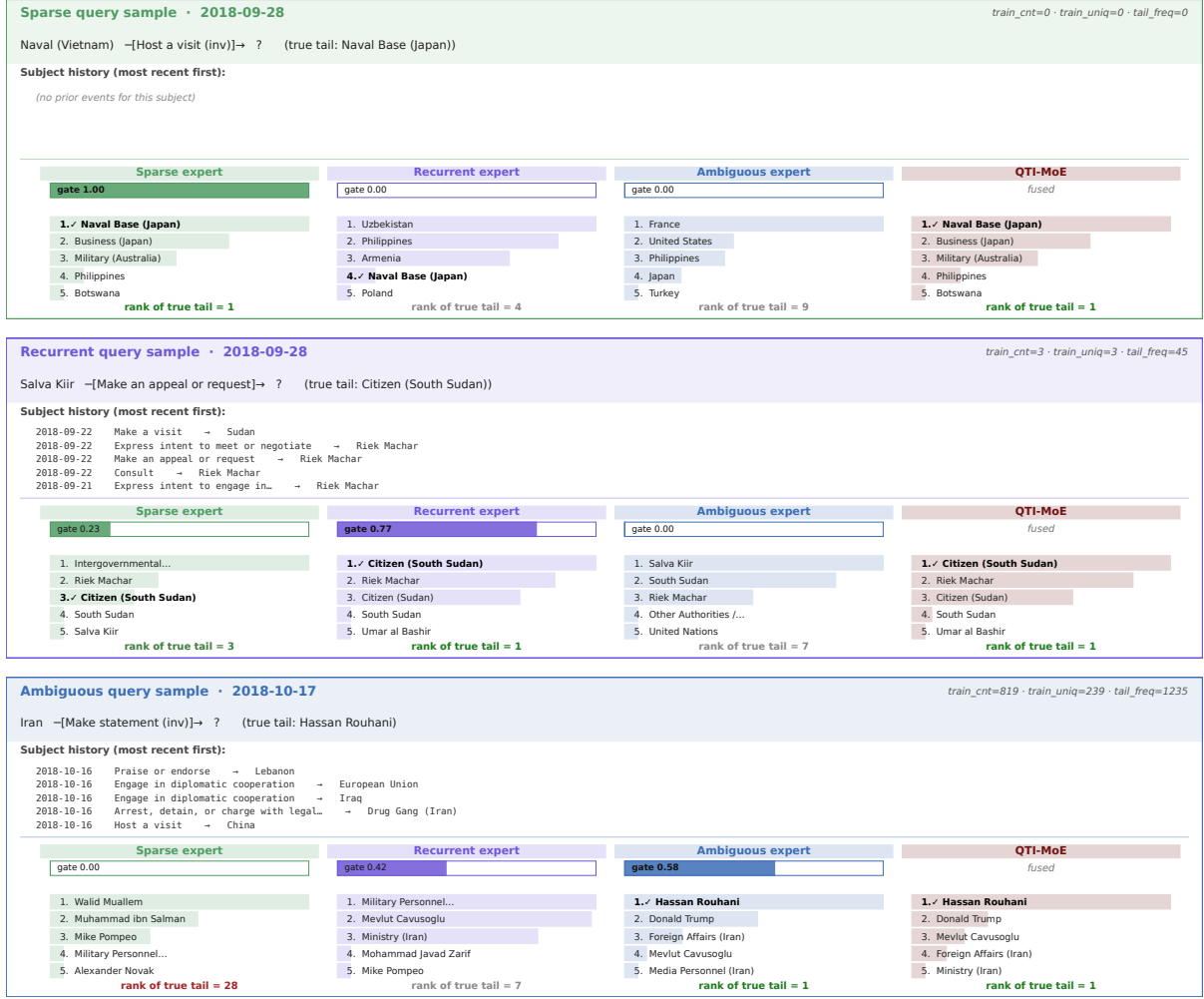


Figure 7: Extended case study on ICEWS18 with one query per regime, showing the historical events of the subject, the gate vector, the top-5 prediction of each expert, and the rank of the gold tail.

is converted into the three-way distribution $\hat{q}^{llm}$ that enters the fusion of Eq. (2). The LLM is used only during offline labelling. It is never called during training or inference of the QTI-MoE router, experts, or gate.

When the LLM is invoked. Following §3.1, the LLM is queried on a boundary subset $\mathcal{B}$ of training pairs. A pair (s, r, t) enters $\mathcal{B}$ when its rule label $\hat{q}_{rule}$ is uncertain. Two conditions trigger entry: the top two rule probabilities differ by less than a small margin, or the prefix statistics $(f_{sr}, d_{sr}, g_{sr}^*)$ lie close to a class-defining percentile threshold. Pairs outside $\mathcal{B}$ keep $\hat{q}_{rule}$ unchanged, so LLM compute is concentrated on the small fraction of queries where the rule is least reliable.

Prompt template. Every boundary pair is formatted with the same prompt template, which has four blocks: a task description, the three candidate labels with their intended meanings, the query con-

text together with strict prefix statistics, and a final instruction to answer with one letter. The block structure is summarised in Table 9, and a verbatim instantiation is provided in the released artefact.

Reading the LLM as a three-way distribution.

The filled prompt is fed through a single forward call of the frozen LLM. We read the logits at the position that follows the literal answer cue and restrict attention to the three label letters $\{A, B, C\}$. Because the same letter may be tokenised with or without a leading space depending on the tokeniser, we take the maximum logit across these surface variants for each letter, then apply a softmax over the three resulting scores to obtain $\hat{q}^{llm}$ as defined in Eq. (1). Restricting the softmax to the three letter tokens has two consequences that match the role of the LLM in our pipeline. The LLM cannot escape the label set by emitting an unrelated token, and its uncertainty is fully expressed by the entropy of the

Block	Content
Task	The LLM is asked to classify the difficulty of a temporal KG query using only events strictly earlier than the query time.
Labels	Three candidate labels are listed in the prompt: A for <i>sparse</i> , B for <i>recurrent</i> , C for <i>ambiguous</i> . Each label is paired with a one-line description that mirrors the definitions in §3.1.
Query context	Surface forms of the subject and the relation, the query time, and the three prefix scalars f_{sr} , d_{sr} , g_{sr}^* introduced in §3.1. A short list of dataset-level reference scales is included so the LLM reasons about relative rather than absolute counts.
Decision and readout	Three short hints aligned with the rule (low evidence favours A, jointly high frequency and diversity favour C, intermediate evidence favours B), followed by the instruction to answer with one letter A, B, or C.

Table 9: Structure of the prompt template used by the frozen LLM on the boundary subset $\mathcal{B}$. All numeric fillers are computed from the strict temporal prefix of the pair, so no future event is exposed to the LLM. The privileged signal g_{sr}^* is included in the prompt for labelling only and is not accessible to the router or the experts.

three-way distribution rather than by free-form text that would require additional parsing.

Fusion with the rule. The fusion in Eq. (2) is governed by a single scalar $w(s, r, t)$. Outside $\mathcal{B}$ the weight is zero by construction. Inside $\mathcal{B}$ it is the product of two factors in $[0, 1]$: the confidence of $\hat{\mathbf{q}}^{llm}$, measured by its peak probability, and the disagreement between $\hat{\mathbf{q}}^{llm}$ and $\hat{\mathbf{q}}_{rule}$, measured as one minus their inner product. This factorisation has three properties that justify the claim in §3.1 that the rule remains dominant unless the LLM is both confident and points to a different type: (1) a low-confidence LLM vote contributes nothing, so noisy boundary cases are absorbed by the rule; (2) an LLM vote that agrees with the rule cannot sharpen the label further, so the LLM never amplifies an already-confident rule decision; and (3) the fusion weight is capped, so a single LLM call can only move a bounded fraction of the probability mass.

Privileged-signal containment. The gold occurrence count g_{sr}^* is the only privileged statistic that appears in the prompt. Following the discipline stated in §3.1, this slot is read by the labeller and the rule but is never exposed to the router, the experts, or the gate at any update step or at test time. Reproducing the soft labels therefore requires only the released LLM weights, the prompt template summarised in Table 9, and the strict prefix statistics.

Gold-free ablation. To verify that the containment above is not merely nominal, we rebuild the soft labels without g_{sr}^* , keeping only the frequency and diversity statistics, and retrain on ICEWS14 and ICEWS18 with all other settings unchanged.

Dataset	Soft label	MRR	H@1	H@3	H@10
ICEWS14	full	51.53	40.48	57.68	72.86
	w/o g_{sr}^*	51.00	39.83	57.08	72.41
ICEWS18	full	38.45	27.07	43.39	61.07
	w/o g_{sr}^*	38.16	26.62	43.28	61.00

Table 10: Soft labels constructed with and without the target occurrence count g_{sr}^* (%). All other settings are identical.

Table 10 reports the result. Removing the statistic costs less than 0.6 MRR on both datasets. The gold-free variant still exceeds the strongest baseline in Table 2 on both benchmarks, and remains far above the parameter-matched Ens-3 of Section 4.5. The target count therefore supplies useful supervision for label construction, but it is not the source of the gains reported in this paper.

C.3 Full Type-Stratified Expert Results

Figures 8, 9, 10, and 11 present the complete type-stratified sweep on the four benchmarks under MRR and Hits@1, 3, 10, extending the H@1 view in Figure 3. The two patterns identified in Section 4.6 hold consistently across datasets and metrics. In every column the focal expert leads the other two, and its lead widens as the subset narrows toward the extreme end of the regime. The QTI-MoE curve tracks the focal expert at all thresholds and rises above the strongest single expert in the mid-range of each regime, where the router has room to combine signals from neighbouring types. ICEWS05-15 and GDELT, which differ from the two main-text datasets in size and temporal granularity, produce the same shape of curves, indicating that the three-way correspondence between regime and expert is not tied to a particular benchmark

Dataset	Type	n	Single expert			QTI-MoE
			sp	re	am	
ICEWS14	sp	423	14.43	10.92	7.61	12.88
	re	13,729	42.81	50.45	46.10	53.16
	am	590	24.48	37.43	38.23	41.24
ICEWS18	sp	1,471	12.69	10.45	8.51	11.66
	re	93,655	31.95	36.61	35.39	39.37
	am	3,964	19.36	23.24	25.17	27.09
ICEWS05-15	sp	1,909	22.61	21.37	22.01	25.56
	re	86,716	53.14	57.93	57.11	59.74
	am	3,693	47.08	57.13	56.82	58.19
GDELT	sp	284	8.34	2.53	4.56	8.23
	re	585,779	21.01	23.13	22.05	27.86
	am	24,419	14.51	15.62	15.81	20.65

Table 11: MRR (%) of each expert in isolation and of the routed QTI-MoE on the three query types of the test set. Each expert is evaluated with the router bypassed. Bold marks the strongest single expert in a row; it coincides with the in-type expert in 11 of the 12 rows.

scale.

C.4 Necessity of the Three Query Types

Table 11 complements the threshold sweep of Section 4.6 with a fixed partition of the test set. We tag every test query with the type attaining the largest rule-based score of Section 3.1, computed from its temporal prefix; the tags group queries for analysis and never enter the model. Each test fact yields an object query and an inverse-relation query, so the counts n are twice the test sizes in Table 1. Each expert is then evaluated with the router bypassed.

The in-type expert is the strongest single expert in 11 of the 12 rows, the exception being a near-tie on ICEWS05-15 ambiguous, so no expert is redundant. The routed model exceeds the best single expert on every recurrent and ambiguous row, indicating that queries in these types benefit from combining neighbouring biases rather than from committing to one. On the sparse rows, the rarest of the three, it stays close to the sparse expert and trails it slightly on two datasets, which marks the setting in which routing is hardest. These results also bear on the discretisation noted in the Limitations section: the routing weights of Equation 7 and the soft label of Equation 2 are continuous, and Equation 8 always mixes all three experts, so the three types act as anchors of a continuous mixing space rather than as disjoint classes. We do not claim the anchors to be complete or optimal, only that each of the three is necessary.

Dataset	Regime	n	Selected expert (%)		
			sp	re	am
ICEWS14	sp	423	63.1	8.5	28.4
	re	13,729	29.0	45.7	25.3
	am	590	25.8	35.3	39.0
ICEWS18	sp	1,471	60.8	0.3	39.0
	re	93,655	25.8	42.4	31.8
	am	3,964	19.5	31.0	49.5
ICEWS05-15	sp	1,909	97.9	0.1	2.0
	re	86,716	37.6	41.4	21.0
	am	3,693	23.5	31.0	45.5
GDELT	sp	284	62.3	2.8	34.9
	re	585,779	25.8	43.8	30.4
	am	24,419	20.7	22.2	57.0

Table 12: Share of test queries in each regime whose largest routing weight goes to each expert (%). Rows may not sum to 100 because of rounding. Bold marks the in-type expert, which is the most frequently selected one in all 12 regimes. The counts n match those of Table 11.

C.5 Router Behaviour across Query Types

The router reads only $\mathbf{f}_{sr} = [\mathbf{h}_s; \mathbf{e}_r]$, which leaves open whether it responds to historical shape or to the identity of the (s, r) pair. Two properties of the design bear on this. First, $\mathbf{h}_s$ is the recurrent state at query time (Equations 11 and 12) rather than a static lookup, so the same subject queried at two different times gives two different router inputs. Second, the router is supervised by $\mathcal{L}_{distill}$, whose targets are computed from prefix statistics.

Table 12 reports which expert receives the largest routing weight, grouped by the type tags of Section C.4. The in-type expert is the most frequently selected one in all 12 rows, against the 33.3% that uniform routing would give. The share is highest on the sparse rows, where the recurrent expert is almost never chosen, consistent with prefixes that offer no repetition to exploit, and lower on the recurrent rows, where the remaining mass falls on the neighbouring types and matches the mixing behaviour that Section C.4 shows to be beneficial. Pair identity does not account for this alignment: sparse queries make up 0.05% of the GDELT test set, far too few to support memorisation, yet a clear majority of them are still routed to the sparse expert.

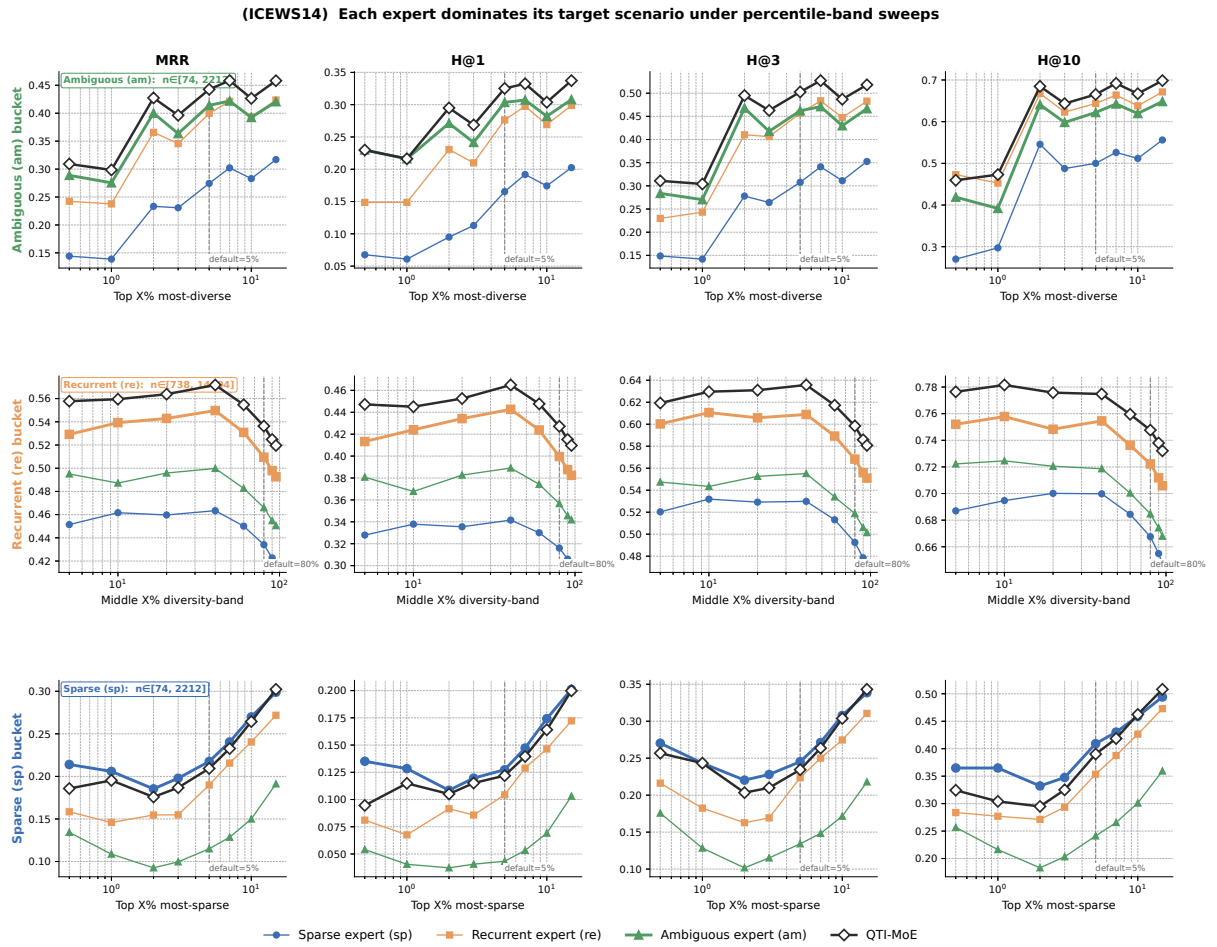


Figure 8: Full-type-stratified sweep on ICEWS14 under MRR and Hits@{1, 3, 10}.

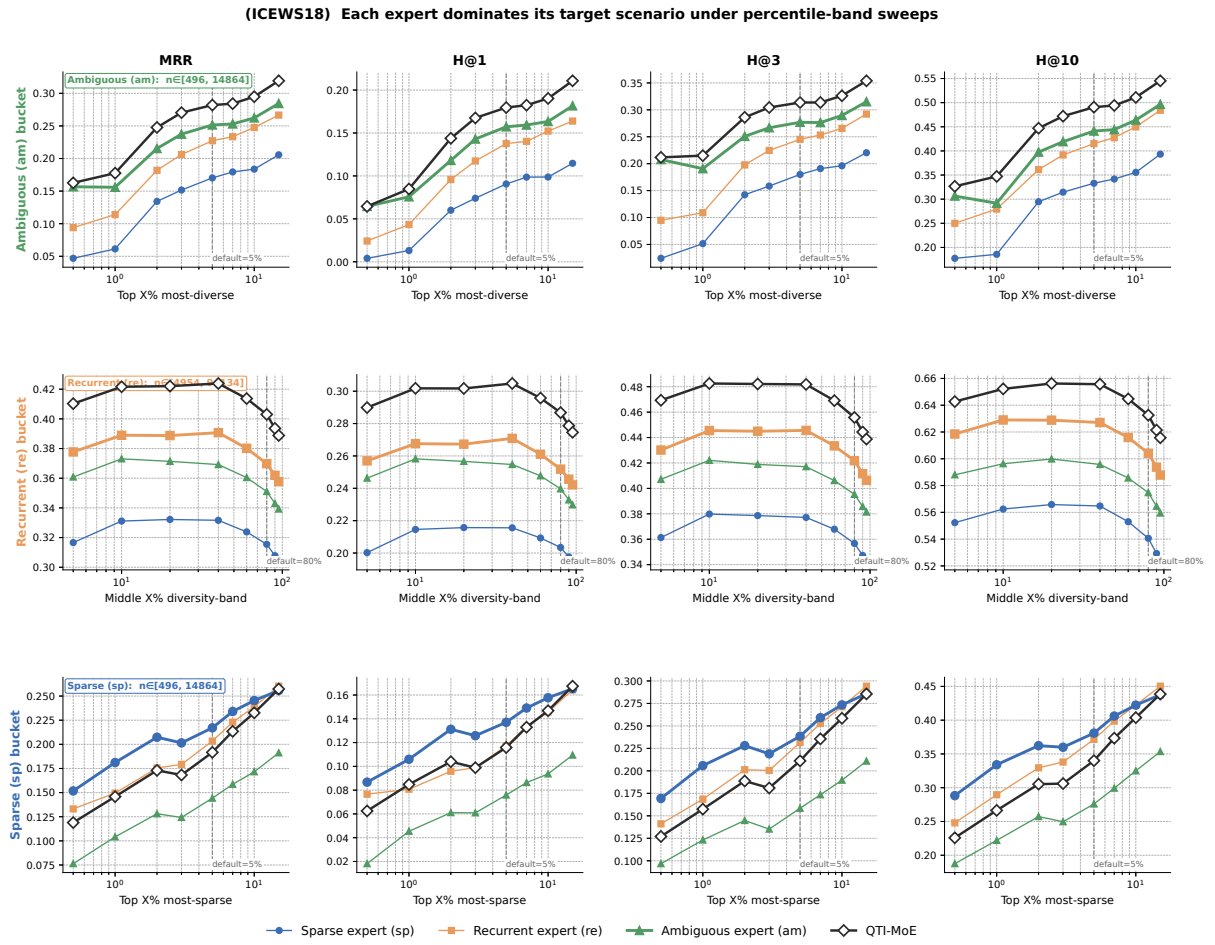


Figure 9: Full type-stratified sweep on ICEWS18 under MRR and Hits@{1, 3, 10}.

(ICEWS05-15) Each expert dominates its target scenario under percentile-band sweeps

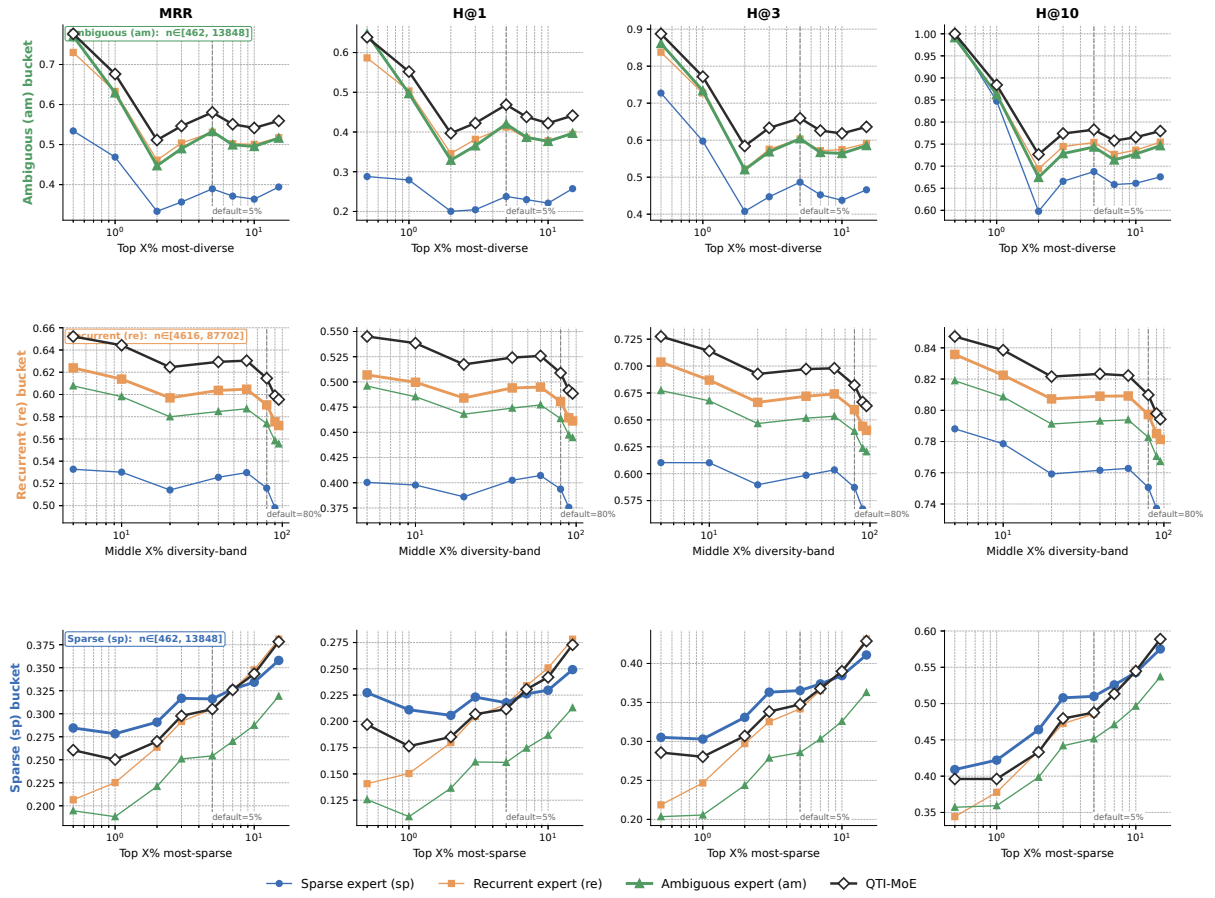


Figure 10: Full type-stratified sweep on ICEWS05-15 under MRR and Hits@{1, 3, 10}.

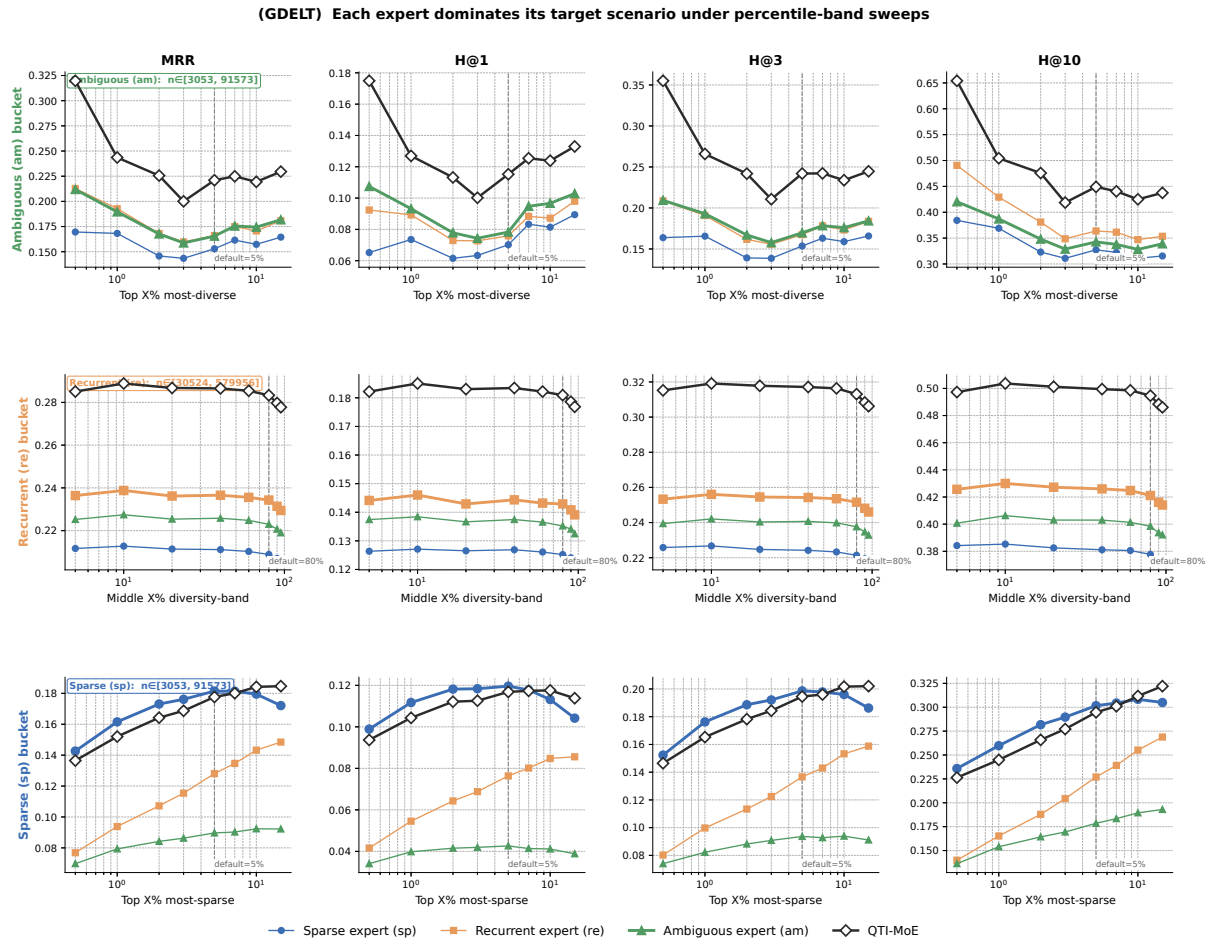


Figure 11: Full type-stratified sweep on GDELT under MRR and Hits@{1, 3, 10}.