TIGER: Temporally Improved Graph Entity Linker
Supplementary Material

1. DATASET CONSTRUCTION

TempEL is a key benchmark dataset for temporal entity linking. To investigate if structured information present in the knowledge graph can help with this task, we combine TempEL[1] with Wikidata5M[2]. The dataset we've constructed is divided into five distinct parts. The text-based parts contain the **entity description** and the **mention context**. The graph-based parts include the **structure graph** representing entity relationships, the **feature graph** built from the embeddings of the entity descriptions, and the **feature matrix** derived from tokens within entity descriptions.

The construction process is shown in Figure S1 and we now walk through each step.

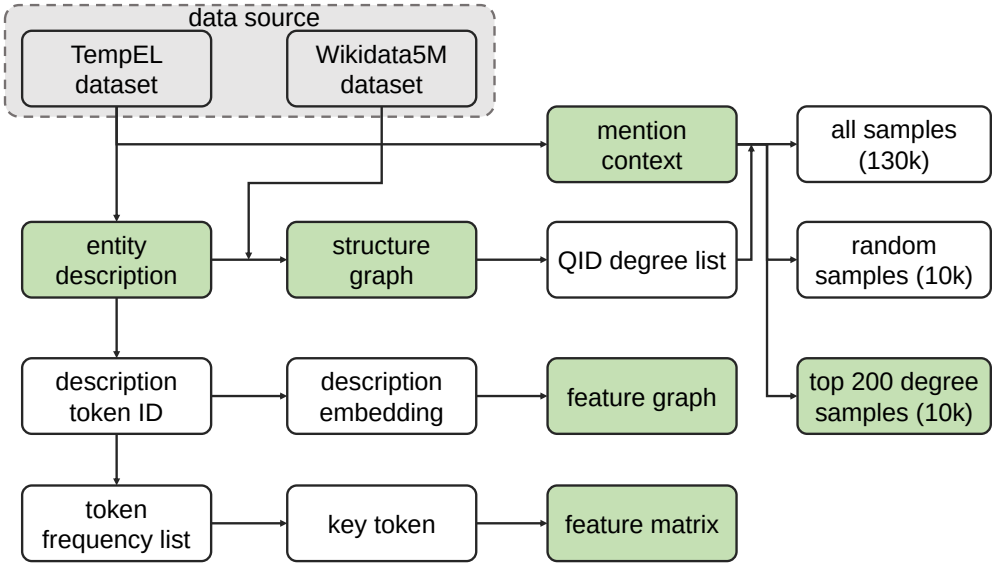


Fig. S1. Dataset construction process.

First, we categorized each year of data from the TempEL dataset into entity descriptions and mention context parts based on the year. The entity description comprises the title, text, document id, and importantly the unique id of the entity (its QID). The mention context consists of context left, context right, mention, label, QID, and category.

Second, we create a structure graph based on the relationship in Wikidata5M dataset and the entity IDs in TempEL dataset. There are numerous relationships among entities in the Wikidata5M dataset. To filter down the number of relationships, we matched the entity IDs (also QIDs) of these relationships with the QID in the entity descriptions from TempEL dataset. If both QIDs in a relationship in the Wikidata5M data are present in the existing entity description, we keep the relationship. The structure graph is an $n \times n$ adjacency matrix, where n represents the total number of entities in the dataset. Each row indicates whether a entity has a connection with other entity. The adjacency matrix is made up of 0s and 1s. If entity i and entity j are connected, the value in the i^{th} row and j^{th} column of the matrix is 1; otherwise, it is 0.

Note that, at this point, we may have introduced temporal leakage. This is because the cutoff for the Wikidata5M dataset is July 2019. If we construct a structure graph based on Wikidata5M,

[1]https://cloud.ilabt.imec.be/index.php/s/RinXy8NgqdW58RW
[2]https://deepgraphlearning.github.io/project/wikidata5m

temporal leakage might occur when dealing with data from years prior to 2019. For instance, if entity 1 and entity 2 had no relationship in 2013 but established one in 2019, our construction of the 2013 structure graph would inevitably link entity 1 and entity 2. However, the efficacy of our model can still be demonstrated. For example, when trained on 2013 data and tested in 2022, our model increased the accuracy from the baseline of 23.1% to 27.45%. During the training process, the input during training did not include entity relationships post-2019.

Third, we built the feature graph using the embeddings from entity descriptions. We employed the pre-trained bert-base-uncased model to embed the textual information associated with the "text" key in the entity description. By accessing the embedded information for each entity in the dataset, we established a kNN graph based on these entities, which we refer to as the feature graph. This graph highlights the connections between entities based on their entity descriptions. The feature graph is also an $n \times n$ adjacency matrix, where n represents the total number of entities in the dataset. Each row indicates whether an entity has a connection with other entities. If entity i and entity j are connected, the value in the i^{th} row and j^{th} column of the matrix is 1; otherwise, it is 0.

The difference between connections in the structure graph and those in the feature graph is that the connections in the structure graph are constructed based on actual existing links. For instance, on the Wikipedia page for "***Weightlifting (Q83462)***", the "***Summer Olympic Games (Q212434)***" is mentioned in the first two paragraphs. Hence a connection between these two entities exists in the structure graph. However, connections in the feature graph are generated based on the embeddings of entity descriptions, which might not exist in the KG but can still benefit the model. For example, while "***Arizona_Wildcats (Q4620330)***" and "***Summer Olympic Games (Q212434)***" aren't directly related, the embeddings of their descriptions have created a link between them. This makes "***Arizona_Wildcats (Q4620330)***" less likely to be confused with "***Arizona (Q816)***" or "***Wildcats (Q26665)***" a 1986 film by Michael Ritchie.

Four, we constructed a feature matrix representing each entity based on the tokens from entity descriptions in the dataset. After getting the token IDs for each entity using the pre-trained bert-base-uncased model, we filtered all token IDs based on their frequency of occurrence. We retained those token IDs that appeared between 46 and 200 times. We discarded highly frequent token IDs since these tokens, such as "is", "an", "the", and other common words, don't offer meaningful differentiation among entities. Also, the less frequent token IDs were removed due to the possibility of them being meaningless noise or random codes, and including an excess of these rare tokens would make the matrix too sparse, slowing down computation. The final feature matrix is an $n \times m$ dimensional matrix composed of 0s and 1s. Here, n represents the total number of entities in the dataset, while m is the number of retained token IDs. If the data in the i^{th} row and j^{th} column of the matrix is 1, it indicates that entity i contains the j^{th} token.

Finally, we generate distinct mention context subsets from all available mention context samples. Using the QID in each sample as the standard, different degrees of sampling were performed on the training set. Based on the annual entity structure graph, the degree of each QID was calculated, and the training set samples were sampled accordingly. The mention context was divided into the full training set (130k), random samples training set (10k), and top 200 degree samples training set (10k).

2. ADDITIONAL RESULTS - TOP 200 DEGREE SAMPLES TRAINING SET

Tables S1 through Table S7 present a comparison between the baseline model (BLINK) and our proposed model (TIGER). The training dataset, termed "top 200 degree samples training set," consists of approximately 10k samples, while the test dataset comprises around 1.5k previously unseen samples. Rows represent training datasets, while columns signify testing datasets. For instance, the entry at the intersection of the first row and tenth column in Table S1 is 0.2310, indicating the model's performance when trained on 2013 data and tested on 2022 data. Performance evaluations for both models were based on recall metrics, specifically @1, @2, @4, @8, @16, @32, and @64.

Table S1. Results between our model and the baseline model @1.

BLINK

Train \ Test	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.1848	0.2193	0.1979	0.2441	0.2166	0.1972	0.1897	0.2117	0.2000	0.2310
2014	0.1766	0.2172	0.1917	0.2062	0.1910	0.1862	0.1738	0.2083	0.1621	0.2407
2015	0.1883	0.2683	0.2352	0.2414	0.2179	0.2069	0.2090	0.2366	0.2248	0.2572
2016	0.2110	0.2724	0.2434	0.2414	0.2628	0.2159	0.2310	0.2483	0.2283	0.2662
2017	0.2117	0.2593	0.2400	0.2717	0.2414	0.2331	0.2145	0.2421	0.2097	0.2772
2018	0.2000	0.2131	0.2145	0.2559	0.2062	0.1959	0.1800	0.1979	0.1924	0.2586
2019	0.1834	0.2166	0.1945	0.2076	0.2021	0.1883	0.1772	0.1869	0.1938	0.2476
2020	0.2262	0.2531	0.2290	0.2731	0.2690	0.2303	0.2262	0.2372	0.2228	0.2586
2021	0.1952	0.2400	0.2179	0.2338	0.2497	0.2221	0.2248	0.2228	0.2014	0.2379
2022	0.1917	0.2228	0.2207	0.2297	0.2124	0.2034	0.2076	0.2166	0.2207	0.2372

TIGER

Train \ Test	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.2255	0.2628	0.2324	0.2731	0.2538	0.2221	0.2214	0.2317	0.2200	0.2745
2014	0.2131	0.2690	0.2276	0.2517	0.2579	0.2048	0.2124	0.2262	0.2069	0.2648
2015	0.1924	0.2386	0.2207	0.2366	0.2393	0.2014	0.1972	0.2200	0.1855	0.2366
2016	0.1717	0.2324	0.2000	0.2200	0.2248	0.1903	0.1910	0.2283	0.2034	0.2359
2017	0.2124	0.2669	0.2303	0.2614	0.2786	0.2228	0.2234	0.2421	0.2310	0.2669
2018	0.1986	0.2234	0.2055	0.2455	0.2234	0.1979	0.1993	0.2324	0.2117	0.2600
2019	0.1917	0.2393	0.2269	0.2214	0.2248	0.2028	0.1890	0.2097	0.2193	0.2455
2020	0.2028	0.2421	0.2062	0.2345	0.2228	0.1945	0.1959	0.2179	0.1993	0.2559
2021	0.1972	0.2469	0.2062	0.2400	0.2324	0.2048	0.2041	0.2228	0.2048	0.2434
2022	0.2186	0.2407	0.2297	0.2428	0.2255	0.2083	0.2083	0.2276	0.2117	0.2745

Table S2. Results between our model and the baseline model @2.

BLINK

Train \ Test	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.2738	0.3110	0.2793	0.3503	0.3124	0.2655	0.2600	0.2869	0.2862	0.3276
2014	0.2476	0.3048	0.2793	0.3110	0.2869	0.2641	0.2614	0.2848	0.2441	0.3186
2015	0.2897	0.3469	0.3352	0.3593	0.3366	0.2917	0.2931	0.3310	0.3131	0.3441
2016	0.2945	0.3607	0.3434	0.3683	0.3634	0.3021	0.3434	0.3628	0.3159	0.3483
2017	0.2966	0.3503	0.3345	0.3814	0.3483	0.3179	0.3172	0.3441	0.3028	0.3717
2018	0.2669	0.2966	0.2966	0.3531	0.3028	0.2669	0.2607	0.2903	0.2745	0.3400
2019	0.2641	0.2848	0.2807	0.3124	0.2972	0.2634	0.2538	0.2669	0.2607	0.3248
2020	0.3117	0.3586	0.3228	0.3786	0.3710	0.3007	0.3297	0.3297	0.3069	0.3545
2021	0.2903	0.3421	0.3145	0.3628	0.3607	0.3172	0.3193	0.3297	0.2876	0.3366
2022	0.2697	0.3000	0.3103	0.3434	0.3097	0.2779	0.2972	0.3152	0.2959	0.3352

TIGER

Train \ Test	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.3034	0.3490	0.3310	0.3897	0.3683	0.3159	0.3207	0.3455	0.3097	0.3607
2014	0.3069	0.3566	0.3200	0.3566	0.3476	0.2834	0.3152	0.3090	0.3048	0.3600
2015	0.2862	0.3152	0.3034	0.3428	0.3372	0.2703	0.2945	0.3179	0.2600	0.3331
2016	0.2538	0.3083	0.2821	0.3290	0.3283	0.2841	0.2924	0.3103	0.2807	0.3400
2017	0.2897	0.3517	0.3393	0.3779	0.3752	0.3207	0.3214	0.3407	0.3028	0.3641
2018	0.2779	0.3234	0.3014	0.3476	0.3234	0.2710	0.2779	0.3331	0.2855	0.3566
2019	0.2883	0.3241	0.3000	0.3234	0.3214	0.2786	0.2786	0.2938	0.2890	0.3317
2020	0.2890	0.3359	0.3131	0.3441	0.3159	0.2848	0.2966	0.2945	0.2855	0.3524
2021	0.2855	0.3393	0.3166	0.3524	0.3269	0.2772	0.2924	0.3179	0.2869	0.3372
2022	0.2924	0.3386	0.3186	0.3579	0.3269	0.2979	0.3048	0.3179	0.2993	0.3717

Table S3. Results between our model and the baseline model @4.

BLINK

Train \ Test	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.3441	0.3993	0.3717	0.4752	0.4345	0.3524	0.3531	0.3979	0.3759	0.4290
2014	0.3455	0.3993	0.3738	0.4193	0.4021	0.3490	0.3434	0.3821	0.3207	0.4090
2015	0.3910	0.4448	0.4248	0.4703	0.4531	0.3924	0.3986	0.4386	0.4145	0.4538
2016	0.3972	0.4724	0.4407	0.4766	0.4772	0.4097	0.4441	0.4710	0.4262	0.4655
2017	0.3972	0.4572	0.4269	0.4959	0.4648	0.4007	0.4090	0.4490	0.3903	0.4772
2018	0.3469	0.3890	0.3910	0.4510	0.3966	0.3593	0.3566	0.3841	0.3634	0.4324
2019	0.3366	0.3917	0.3614	0.4186	0.3828	0.3400	0.3483	0.3607	0.3359	0.4055
2020	0.3959	0.4531	0.4069	0.4766	0.4703	0.3966	0.4193	0.4434	0.4048	0.4634
2021	0.3752	0.4600	0.4069	0.4766	0.4586	0.4131	0.4269	0.4359	0.3862	0.4455
2022	0.3690	0.4028	0.3862	0.4559	0.4303	0.3662	0.3979	0.4103	0.4000	0.4497

TIGER

Train \ Test	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.3952	0.4469	0.4193	0.4862	0.4834	0.4207	0.4228	0.4448	0.3931	0.4648
2014	0.3931	0.4476	0.4159	0.4752	0.4510	0.3966	0.4179	0.4145	0.4179	0.4745
2015	0.3600	0.4131	0.4041	0.4497	0.4407	0.3593	0.3848	0.4041	0.3517	0.4352
2016	0.3455	0.4048	0.3786	0.4297	0.4276	0.3848	0.3952	0.4076	0.3738	0.4510
2017	0.3924	0.4538	0.4214	0.4848	0.4724	0.4241	0.4386	0.4517	0.4014	0.4703
2018	0.3669	0.4352	0.3759	0.4490	0.4366	0.3593	0.3959	0.4317	0.3800	0.4448
2019	0.3655	0.4145	0.3890	0.4372	0.4400	0.3690	0.3752	0.3828	0.3807	0.4372
2020	0.3752	0.4276	0.4103	0.4345	0.4234	0.3731	0.3834	0.3903	0.3828	0.4476
2021	0.3655	0.4283	0.4097	0.4731	0.4283	0.3703	0.4007	0.4221	0.4007	0.4497
2022	0.4021	0.4441	0.4200	0.4772	0.4524	0.3910	0.4090	0.4290	0.4055	0.4690

Table S4. Results between our model and the baseline model @8.

BLINK

Train \ Test	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.4455	0.5110	0.4834	0.5731	0.5497	0.4634	0.4641	0.5048	0.4855	0.5310
2014	0.4510	0.4986	0.4772	0.5324	0.5131	0.4545	0.4559	0.4869	0.4207	0.5179
2015	0.4931	0.5476	0.5324	0.5945	0.5510	0.5069	0.5131	0.5552	0.5338	0.5959
2016	0.5152	0.5759	0.5290	0.5938	0.5683	0.5324	0.5579	0.5628	0.5352	0.5855
2017	0.5062	0.5510	0.5345	0.6028	0.5676	0.5207	0.5290	0.5572	0.4959	0.5890
2018	0.4331	0.4903	0.4917	0.5469	0.4986	0.4462	0.4566	0.4703	0.4662	0.5200
2019	0.4248	0.4883	0.4579	0.5124	0.4766	0.4338	0.4545	0.4552	0.4228	0.5041
2020	0.4931	0.5607	0.5214	0.5814	0.5600	0.5248	0.5221	0.5407	0.5076	0.5717
2021	0.4745	0.5600	0.5048	0.5883	0.5517	0.5186	0.5386	0.5262	0.4986	0.5600
2022	0.4469	0.5228	0.4952	0.5703	0.5407	0.4890	0.5055	0.5048	0.5014	0.5393

TIGER

Train \ Test	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.4903	0.5600	0.5083	0.6069	0.5807	0.5248	0.5462	0.5476	0.5041	0.5710
2014	0.4772	0.5393	0.5228	0.5841	0.5428	0.5041	0.5172	0.5069	0.5248	0.5890
2015	0.4634	0.5076	0.5028	0.5752	0.5448	0.4752	0.4869	0.5179	0.4497	0.5441
2016	0.4441	0.5124	0.4676	0.5221	0.5393	0.4779	0.5014	0.4993	0.4821	0.5600
2017	0.5014	0.5579	0.5352	0.5952	0.5669	0.5228	0.5462	0.5614	0.4931	0.5662
2018	0.4628	0.5159	0.4800	0.5524	0.5262	0.4731	0.5021	0.5441	0.4903	0.5628
2019	0.4469	0.5090	0.4793	0.5538	0.5290	0.4793	0.4703	0.4834	0.4814	0.5345
2020	0.4662	0.5269	0.5110	0.5545	0.5310	0.4752	0.4828	0.4841	0.4917	0.5566
2021	0.4552	0.5310	0.5145	0.5724	0.5379	0.4924	0.5131	0.5131	0.5241	0.5655
2022	0.5014	0.5455	0.5234	0.5897	0.5441	0.5028	0.5172	0.5214	0.5159	0.5972

Table S5. Results between our model and the baseline model @16.

BLINK

Train \ Test	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.5531	0.6055	0.6062	0.6607	0.6469	0.5766	0.5828	0.6028	0.5883	0.6359
2014	0.5483	0.5924	0.5945	0.6324	0.6090	0.5683	0.5662	0.6062	0.5297	0.6434
2015	0.5931	0.6455	0.6303	0.6952	0.6566	0.6159	0.6228	0.6497	0.6324	0.6931
2016	0.6152	0.6703	0.6503	0.6800	0.6759	0.6524	0.6538	0.6572	0.6269	0.6938
2017	0.6159	0.6524	0.6497	0.6710	0.6669	0.6400	0.6352	0.6490	0.6083	0.6855
2018	0.5103	0.5876	0.5959	0.6393	0.5862	0.5607	0.5524	0.5669	0.5600	0.6283
2019	0.5007	0.5752	0.5600	0.6117	0.5738	0.5331	0.5634	0.5297	0.5207	0.5841
2020	0.5717	0.6566	0.6172	0.6759	0.6586	0.6214	0.6186	0.6359	0.6110	0.6752
2021	0.5669	0.6483	0.6186	0.6655	0.6566	0.6310	0.6324	0.6283	0.6110	0.6634
2022	0.5393	0.6152	0.6076	0.6628	0.6448	0.5890	0.6159	0.6021	0.6041	0.6545

TIGER

Train \ Test	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.5786	0.6690	0.6159	0.6959	0.6607	0.6283	0.6503	0.6366	0.6124	0.6697
2014	0.5669	0.6407	0.6303	0.6703	0.6441	0.6124	0.6234	0.6103	0.6317	0.6828
2015	0.5614	0.6193	0.6221	0.6579	0.6469	0.5697	0.5897	0.6186	0.5655	0.6497
2016	0.5434	0.6097	0.5807	0.6159	0.6338	0.5821	0.6034	0.5910	0.5821	0.6586
2017	0.6048	0.6572	0.6248	0.6883	0.6510	0.6290	0.6345	0.6510	0.6062	0.6648
2018	0.5552	0.6021	0.5931	0.6503	0.6214	0.5855	0.5931	0.6338	0.5910	0.6703
2019	0.5276	0.6007	0.5959	0.6434	0.6290	0.5814	0.5752	0.5779	0.5841	0.6186
2020	0.5593	0.6097	0.6138	0.6503	0.6276	0.5807	0.5834	0.5724	0.5993	0.6359
2021	0.5641	0.6441	0.6138	0.6593	0.6262	0.5855	0.6166	0.6028	0.6090	0.6766
2022	0.5993	0.6338	0.6366	0.6841	0.6407	0.6131	0.6221	0.6090	0.6124	0.7117

Table S6. Results between our model and the baseline model @32.

BLINK

Train \ Test	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.6379	0.7138	0.6966	0.7352	0.7269	0.6924	0.6759	0.7138	0.6855	0.7503
2014	0.6476	0.6993	0.6931	0.7041	0.7131	0.6752	0.6669	0.6910	0.6338	0.7276
2015	0.6807	0.7393	0.7428	0.7586	0.7297	0.7255	0.7234	0.7372	0.7166	0.7683
2016	0.6959	0.7703	0.7497	0.7655	0.7559	0.7628	0.7428	0.7469	0.7083	0.7807
2017	0.6993	0.7462	0.7400	0.7359	0.7559	0.7531	0.7214	0.7303	0.6966	0.7669
2018	0.5979	0.6821	0.6938	0.7200	0.6669	0.6524	0.6538	0.6476	0.6524	0.7055
2019	0.5738	0.6621	0.6310	0.6993	0.6593	0.6441	0.6538	0.6124	0.6262	0.6752
2020	0.6545	0.7324	0.7297	0.7552	0.7366	0.7172	0.7069	0.7083	0.7234	0.7655
2021	0.6614	0.7455	0.7193	0.7531	0.7448	0.7366	0.7138	0.7152	0.7014	0.7669
2022	0.6290	0.7117	0.6993	0.7490	0.7297	0.7048	0.7131	0.6855	0.7028	0.7366

TIGER

Train \ Test	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.6593	0.7524	0.7069	0.7690	0.7434	0.7117	0.7276	0.7062	0.7090	0.7648
2014	0.6510	0.7276	0.7131	0.7552	0.7283	0.7172	0.7166	0.6986	0.7214	0.7641
2015	0.6455	0.7172	0.7000	0.7359	0.7317	0.6697	0.6834	0.7172	0.6669	0.7262
2016	0.6276	0.7076	0.6834	0.7062	0.7214	0.6807	0.7000	0.6814	0.6738	0.7407
2017	0.6897	0.7531	0.7248	0.7745	0.7359	0.7310	0.7117	0.7310	0.6828	0.7490
2018	0.6352	0.7007	0.6952	0.7393	0.7090	0.6752	0.6766	0.7076	0.6759	0.7455
2019	0.6166	0.6979	0.6986	0.7193	0.7214	0.6807	0.6628	0.6669	0.6786	0.6993
2020	0.6331	0.7048	0.6945	0.7103	0.7131	0.6834	0.6793	0.6648	0.6828	0.7352
2021	0.6490	0.7262	0.6938	0.7393	0.7117	0.6883	0.7055	0.6834	0.7152	0.7524
2022	0.6793	0.7276	0.7324	0.7600	0.7310	0.7131	0.7090	0.7055	0.6931	0.7752

Table S7. Results between our model and the baseline model @64.

BLINK

Train \ Test	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.7124	0.7993	0.7890	0.7945	0.8021	0.7834	0.7572	0.7869	0.7614	0.8214
2014	0.7324	0.7979	0.7586	0.7855	0.7848	0.7524	0.7503	0.7683	0.7124	0.8076
2015	0.7600	0.8131	0.8097	0.8283	0.8028	0.7952	0.7972	0.8062	0.7897	0.8421
2016	0.7800	0.8421	0.8193	0.8331	0.8228	0.8241	0.8124	0.8062	0.7821	0.8352
2017	0.7931	0.8276	0.8186	0.8124	0.8241	0.8193	0.7924	0.8014	0.7752	0.8414
2018	0.6834	0.7579	0.7690	0.7910	0.7655	0.7469	0.7269	0.7248	0.7310	0.7807
2019	0.6372	0.7497	0.7200	0.7779	0.7490	0.7324	0.7269	0.7062	0.7131	0.7469
2020	0.7345	0.8124	0.7924	0.8152	0.8138	0.7910	0.7752	0.7669	0.8166	0.8207
2021	0.7359	0.8124	0.7966	0.8166	0.8097	0.8269	0.7952	0.7979	0.7669	0.8421
2022	0.7228	0.7924	0.7800	0.8179	0.8014	0.7938	0.7931	0.7731	0.7924	0.8062

TIGER

Train \ Test	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.7393	0.8269	0.7959	0.8262	0.8083	0.7979	0.7972	0.7772	0.7862	0.8276
2014	0.7359	0.8166	0.8000	0.8145	0.8048	0.7931	0.7883	0.7772	0.8021	0.8352
2015	0.7255	0.7972	0.7738	0.8124	0.7952	0.7614	0.7931	0.7903	0.7566	0.8103
2016	0.7172	0.7952	0.7676	0.8021	0.7952	0.7703	0.7759	0.7703	0.7579	0.8152
2017	0.7814	0.8317	0.8090	0.8483	0.8041	0.8083	0.7828	0.8014	0.7703	0.8248
2018	0.7076	0.7897	0.7766	0.8214	0.7897	0.7662	0.7586	0.7807	0.7490	0.8290
2019	0.7028	0.7807	0.7772	0.7828	0.7807	0.7538	0.7352	0.7310	0.7607	0.7897
2020	0.7255	0.7848	0.7841	0.7903	0.7883	0.7710	0.7517	0.7366	0.7731	0.8090
2021	0.7269	0.8145	0.7931	0.8014	0.7848	0.7759	0.7897	0.7614	0.8041	0.8172
2022	0.7586	0.8186	0.8145	0.8331	0.8055	0.7979	0.7869	0.7779	0.7648	0.8400

3. ADDITIONAL RESULTS - RANDOM SAMPLES TRAINING SET

Tables S8 through Table S14 present a comparison between the baseline model (BLINK) and our proposed model (TIGER). The training dataset, termed "random samples training set," consists of approximately 10k samples, while the test dataset comprises around 1.5k previously unseen samples. Rows represent training datasets, while columns signify testing datasets. For instance, the entry at the intersection of the first row and tenth column in Table S8 is 0.4662, indicating the model's performance when trained on 2013 data and tested on 2022 data. Performance evaluations for both models were based on recall metrics, specifically @1, @2, @4, @8, @16, @32, and @64.

Table S8. Results between our model and the baseline model @1.

BLINK

Train \ Test	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.4903	0.4931	0.4490	0.5000	0.4655	0.4738	0.4248	0.4917	0.3828	0.4662
2014	0.4586	0.4945	0.4641	0.4828	0.4731	0.4710	0.4476	0.4634	0.3869	0.4662
2015	0.4090	0.4703	0.4497	0.4890	0.4524	0.4297	0.4186	0.4779	0.3924	0.4372
2016	0.4359	0.4848	0.4172	0.4572	0.4545	0.4186	0.4110	0.4131	0.3545	0.4421
2017	0.3834	0.4421	0.3841	0.4545	0.4117	0.4331	0.4117	0.4379	0.3138	0.3834
2018	0.4503	0.4614	0.4593	0.4786	0.4828	0.5152	0.4462	0.4759	0.3538	0.4676
2019	0.3655	0.4214	0.3517	0.4228	0.4097	0.4179	0.3828	0.4159	0.3807	0.4207
2020	0.4269	0.4766	0.4172	0.4600	0.4407	0.4269	0.4172	0.4538	0.3766	0.4124
2021	0.4207	0.4524	0.4283	0.4510	0.4193	0.4131	0.4090	0.4497	0.3634	0.4317
2022	0.4290	0.4690	0.3903	0.4434	0.4214	0.4269	0.3834	0.4552	0.3407	0.4483

TIGER

Train \ Test	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.4172	0.4793	0.4297	0.4710	0.4110	0.4186	0.3745	0.4497	0.3634	0.4476
2014	0.4959	0.5021	0.4731	0.5007	0.4890	0.4924	0.4407	0.4931	0.4048	0.4669
2015	0.4524	0.4441	0.3931	0.4772	0.4159	0.4283	0.4062	0.4572	0.3959	0.4290
2016	0.4531	0.4566	0.4434	0.4607	0.4366	0.4297	0.3972	0.4462	0.3386	0.4228
2017	0.4166	0.4600	0.4207	0.4848	0.4366	0.4634	0.3869	0.4297	0.3345	0.4131
2018	0.4503	0.4634	0.4214	0.4793	0.4159	0.4903	0.4317	0.4738	0.3345	0.4552
2019	0.3566	0.4200	0.3655	0.4545	0.3966	0.4110	0.3662	0.4048	0.3338	0.3883
2020	0.4131	0.4683	0.4207	0.4986	0.4345	0.3952	0.3448	0.4283	0.2931	0.4083
2021	0.3841	0.4455	0.3814	0.3959	0.3979	0.3938	0.3890	0.4255	0.3510	0.4090
2022	0.4366	0.4807	0.4214	0.5117	0.4607	0.4772	0.4255	0.4910	0.3841	0.4421

Table S9. Results between our model and the baseline model @2.

BLINK

Train \ Test	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.6524	0.6234	0.5897	0.6490	0.6131	0.6372	0.5883	0.6628	0.5110	0.6262
2014	0.6379	0.6138	0.6290	0.6248	0.6069	0.6255	0.6041	0.6214	0.4917	0.6441
2015	0.5821	0.5910	0.5924	0.6379	0.6097	0.5855	0.5876	0.6352	0.5083	0.5938
2016	0.6221	0.6124	0.5634	0.6186	0.6283	0.5793	0.5414	0.5931	0.4607	0.5966
2017	0.5455	0.5731	0.5338	0.5986	0.5745	0.5676	0.5710	0.5903	0.4552	0.5607
2018	0.6172	0.5793	0.6048	0.6462	0.6372	0.6428	0.6131	0.6262	0.4931	0.6248
2019	0.5193	0.5379	0.4972	0.5731	0.5572	0.5669	0.5297	0.5752	0.4814	0.5662
2020	0.5931	0.5910	0.5683	0.6214	0.6069	0.5779	0.5993	0.6366	0.4938	0.6000
2021	0.5945	0.5897	0.5772	0.6076	0.5972	0.5766	0.5676	0.6124	0.4800	0.6014
2022	0.6034	0.6124	0.5469	0.6014	0.5634	0.5903	0.5738	0.6145	0.4586	0.5890

TIGER

Train \ Test	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.5876	0.6055	0.5841	0.6214	0.5690	0.5710	0.5372	0.6221	0.4959	0.5959
2014	0.6510	0.6324	0.6221	0.6331	0.6290	0.6276	0.5993	0.6510	0.5200	0.6248
2015	0.6152	0.5862	0.5634	0.6234	0.5952	0.5724	0.5655	0.6345	0.5255	0.5993
2016	0.6297	0.6090	0.5807	0.6166	0.6262	0.5952	0.5641	0.6069	0.4531	0.5910
2017	0.6062	0.5841	0.5476	0.6352	0.5959	0.5945	0.5448	0.5862	0.4655	0.5759
2018	0.6062	0.5986	0.5738	0.6290	0.6090	0.6083	0.5772	0.6034	0.4669	0.6172
2019	0.5283	0.5503	0.5117	0.6076	0.5517	0.5669	0.5207	0.5703	0.4428	0.5366
2020	0.5628	0.5890	0.5683	0.6462	0.6021	0.5317	0.4731	0.5690	0.3952	0.5517
2021	0.5662	0.5697	0.5262	0.5393	0.5683	0.5400	0.5490	0.5834	0.4621	0.5683
2022	0.5979	0.6083	0.5703	0.6490	0.6297	0.5959	0.6021	0.6241	0.4993	0.6021

Table S10. Results between our model and the baseline model @4.

BLINK

Train \ Test	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.7814	0.7338	0.7097	0.7462	0.7462	0.7517	0.7290	0.7683	0.6103	0.7379
2014	0.7710	0.7200	0.7352	0.7234	0.7297	0.7352	0.7276	0.7441	0.6041	0.7552
2015	0.7441	0.7055	0.7303	0.7510	0.7455	0.7083	0.7193	0.7331	0.6234	0.7434
2016	0.7593	0.7117	0.6917	0.7234	0.7552	0.7110	0.7159	0.7283	0.5593	0.7297
2017	0.7034	0.7014	0.6814	0.7276	0.7193	0.7110	0.7076	0.7214	0.5600	0.6807
2018	0.7772	0.6959	0.7310	0.7352	0.7683	0.7593	0.7248	0.7379	0.6041	0.7497
2019	0.6903	0.6448	0.6469	0.6903	0.6924	0.6890	0.6800	0.6924	0.5938	0.6959
2020	0.7359	0.7083	0.6779	0.7097	0.7448	0.7200	0.7234	0.7455	0.6055	0.7062
2021	0.7428	0.7152	0.7131	0.7179	0.7193	0.7034	0.7152	0.7303	0.5972	0.7303
2022	0.7469	0.7241	0.6814	0.7152	0.6938	0.7241	0.7048	0.7248	0.5855	0.7007

TIGER

Train \ Test	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.7441	0.7055	0.7021	0.7372	0.7028	0.7145	0.6759	0.7317	0.5993	0.7200
2014	0.7800	0.7393	0.7379	0.7421	0.7531	0.7393	0.7434	0.7283	0.6248	0.7338
2015	0.7600	0.7069	0.6897	0.7379	0.7345	0.7076	0.7103	0.7441	0.6228	0.7103
2016	0.7731	0.7159	0.6972	0.7152	0.7545	0.7179	0.7048	0.7428	0.5703	0.7214
2017	0.7483	0.7076	0.6855	0.7400	0.7214	0.7124	0.6834	0.7055	0.5641	0.7055
2018	0.7517	0.7062	0.6724	0.7379	0.7503	0.7234	0.7138	0.7276	0.5724	0.7297
2019	0.6931	0.6614	0.6428	0.7000	0.6917	0.6883	0.6641	0.6855	0.5586	0.6600
2020	0.7110	0.6959	0.7021	0.7359	0.7262	0.6641	0.6469	0.7034	0.5214	0.6897
2021	0.7359	0.6862	0.6483	0.6600	0.6931	0.6614	0.7021	0.7179	0.5924	0.7117
2022	0.7428	0.7117	0.7055	0.7510	0.7510	0.7179	0.7490	0.7297	0.6034	0.7214

Table S11. Results between our model and the baseline model @8.

BLINK

Train \ Test	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.8566	0.8262	0.8241	0.8234	0.8283	0.8283	0.8234	0.8407	0.7083	0.8193
2014	0.8628	0.8186	0.8131	0.8200	0.8345	0.8234	0.8200	0.8262	0.7007	0.8152
2015	0.8607	0.8138	0.8359	0.8248	0.8428	0.8200	0.8007	0.8379	0.7117	0.8159
2016	0.8538	0.7959	0.8186	0.8007	0.8400	0.8152	0.8097	0.8166	0.6628	0.8228
2017	0.8400	0.7897	0.7883	0.8069	0.8352	0.8062	0.8007	0.8117	0.6683	0.7800
2018	0.8676	0.8028	0.8317	0.8200	0.8517	0.8517	0.8124	0.8248	0.6876	0.8166
2019	0.8200	0.7524	0.7662	0.7772	0.8000	0.7959	0.7710	0.7890	0.6938	0.7883
2020	0.8552	0.8014	0.7966	0.7903	0.8359	0.8241	0.8207	0.8228	0.6986	0.7890
2021	0.8593	0.8248	0.8248	0.7972	0.8317	0.8110	0.8138	0.8255	0.7028	0.8172
2022	0.8545	0.8076	0.8103	0.7890	0.7903	0.8145	0.8007	0.8145	0.6786	0.7779

TIGER

Train \ Test	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.8483	0.8055	0.8152	0.8076	0.8124	0.8179	0.7931	0.8193	0.7048	0.8062
2014	0.8752	0.8297	0.8379	0.8297	0.8441	0.8234	0.8372	0.8124	0.7159	0.8055
2015	0.8648	0.8214	0.8069	0.8255	0.8462	0.8214	0.7972	0.8248	0.7055	0.7993
2016	0.8662	0.8090	0.8131	0.7966	0.8400	0.8048	0.8145	0.8234	0.6697	0.8103
2017	0.8531	0.8069	0.8255	0.8103	0.8407	0.8110	0.7890	0.8048	0.6648	0.7986
2018	0.8683	0.8000	0.7897	0.8172	0.8297	0.8248	0.7938	0.8255	0.6772	0.8186
2019	0.8034	0.7669	0.7745	0.7779	0.8028	0.7945	0.7710	0.7828	0.6621	0.7566
2020	0.8310	0.8090	0.8131	0.8124	0.8200	0.7772	0.7545	0.8014	0.6303	0.7807
2021	0.8372	0.7848	0.7717	0.7586	0.8097	0.7786	0.7986	0.8214	0.7103	0.8000
2022	0.8434	0.8069	0.8276	0.8159	0.8469	0.8310	0.8317	0.8152	0.6945	0.8083

Table S12. Results between our model and the baseline model @16.

BLINK

Train \ Test	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.9276	0.8959	0.8938	0.8855	0.8897	0.8800	0.8931	0.9055	0.7897	0.8855
2014	0.9152	0.8890	0.9007	0.8855	0.8959	0.8890	0.8786	0.8910	0.7821	0.8641
2015	0.9214	0.8993	0.8917	0.8876	0.9131	0.8731	0.8510	0.8945	0.7862	0.8669
2016	0.9207	0.8766	0.8876	0.8759	0.9083	0.8814	0.8655	0.8759	0.7317	0.8710
2017	0.9124	0.8655	0.8834	0.8703	0.9055	0.8862	0.8676	0.8683	0.7517	0.8434
2018	0.9159	0.8814	0.9021	0.8759	0.9048	0.9028	0.8710	0.8828	0.7545	0.8607
2019	0.9014	0.8510	0.8559	0.8517	0.8814	0.8655	0.8490	0.8676	0.7807	0.8448
2020	0.9152	0.8731	0.8800	0.8497	0.9131	0.8766	0.8724	0.8855	0.7848	0.8441
2021	0.9138	0.8855	0.8979	0.8772	0.9069	0.8731	0.8800	0.8897	0.8000	0.8807
2022	0.9179	0.8786	0.8828	0.8600	0.8828	0.8738	0.8703	0.8800	0.7807	0.8510

TIGER

Train \ Test	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.9228	0.8848	0.8855	0.8724	0.9014	0.8731	0.8559	0.8779	0.7869	0.8641
2014	0.9262	0.8952	0.8938	0.8959	0.9179	0.8807	0.8966	0.8821	0.7897	0.8552
2015	0.9221	0.8897	0.8910	0.8862	0.9103	0.8759	0.8579	0.8876	0.7834	0.8655
2016	0.9228	0.8793	0.8876	0.8731	0.9097	0.8731	0.8772	0.8862	0.7455	0.8676
2017	0.9172	0.8752	0.8841	0.8855	0.9131	0.8876	0.8634	0.8724	0.7517	0.8621
2018	0.9241	0.8828	0.8752	0.8717	0.8952	0.8834	0.8579	0.8869	0.7779	0.8621
2019	0.8959	0.8566	0.8648	0.8531	0.8910	0.8552	0.8476	0.8524	0.7607	0.8131
2020	0.9062	0.8717	0.8924	0.8772	0.8972	0.8538	0.8366	0.8717	0.7359	0.8510
2021	0.9090	0.8662	0.8703	0.8421	0.8966	0.8538	0.8717	0.8821	0.8041	0.8607
2022	0.9083	0.8724	0.8945	0.8807	0.9034	0.8848	0.8834	0.8897	0.7834	0.8821

Table S13. Results between our model and the baseline model @32.

BLINK

Train \ Test	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.9559	0.9338	0.9283	0.9269	0.9338	0.9228	0.9310	0.9317	0.8593	0.9366
2014	0.9510	0.9366	0.9393	0.9234	0.9372	0.9297	0.9193	0.9290	0.8455	0.9097
2015	0.9579	0.9359	0.9352	0.9276	0.9510	0.9138	0.9110	0.9324	0.8497	0.9138
2016	0.9545	0.9303	0.9303	0.9283	0.9428	0.9241	0.9186	0.9269	0.8159	0.9166
2017	0.9510	0.9069	0.9255	0.9186	0.9469	0.9200	0.9062	0.9200	0.8317	0.8952
2018	0.9566	0.9221	0.9324	0.9166	0.9379	0.9324	0.9152	0.9262	0.8345	0.8966
2019	0.9366	0.9041	0.9166	0.9055	0.9331	0.9076	0.9138	0.9234	0.8421	0.8966
2020	0.9517	0.9214	0.9241	0.9014	0.9421	0.9124	0.9103	0.9248	0.8497	0.8993
2021	0.9552	0.9228	0.9386	0.9317	0.9434	0.9172	0.9166	0.9248	0.8655	0.9193
2022	0.9559	0.9241	0.9345	0.9069	0.9297	0.9166	0.9138	0.9262	0.8414	0.9007

TIGER

Train \ Test	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.9552	0.9276	0.9331	0.9110	0.9393	0.9110	0.8986	0.9159	0.8572	0.9159
2014	0.9566	0.9352	0.9366	0.9303	0.9531	0.9228	0.9262	0.9303	0.8476	0.8979
2015	0.9510	0.9303	0.9324	0.9345	0.9476	0.9234	0.9193	0.9221	0.8428	0.9131
2016	0.9579	0.9255	0.9276	0.9200	0.9483	0.9110	0.9207	0.9283	0.8200	0.9055
2017	0.9476	0.9214	0.9193	0.9234	0.9434	0.9241	0.9076	0.9172	0.8352	0.9083
2018	0.9538	0.9297	0.9228	0.9179	0.9441	0.9221	0.9097	0.9317	0.8462	0.8993
2019	0.9359	0.9076	0.9152	0.9034	0.9393	0.8966	0.9028	0.9117	0.8297	0.8779
2020	0.9483	0.9172	0.9338	0.9179	0.9393	0.8972	0.8917	0.9186	0.8124	0.9007
2021	0.9531	0.9221	0.9179	0.8993	0.9324	0.9014	0.9117	0.9214	0.8738	0.9124
2022	0.9428	0.9145	0.9297	0.9207	0.9407	0.9200	0.9200	0.9262	0.8524	0.9221

Table S14. Results between our model and the baseline model @64.

BLINK

Train \ Test	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.9786	0.9559	0.9593	0.9566	0.9614	0.9517	0.9531	0.9552	0.9083	0.9628
2014	0.9800	0.9600	0.9634	0.9441	0.9676	0.9476	0.9414	0.9648	0.8993	0.9483
2015	0.9793	0.9641	0.9600	0.9552	0.9703	0.9490	0.9428	0.9559	0.9062	0.9469
2016	0.9745	0.9566	0.9579	0.9497	0.9669	0.9531	0.9497	0.9593	0.8897	0.9503
2017	0.9731	0.9372	0.9510	0.9455	0.9683	0.9455	0.9400	0.9559	0.8952	0.9359
2018	0.9779	0.9510	0.9531	0.9352	0.9655	0.9545	0.9428	0.9579	0.8931	0.9338
2019	0.9752	0.9372	0.9476	0.9359	0.9621	0.9359	0.9310	0.9476	0.8924	0.9255
2020	0.9717	0.9503	0.9497	0.9345	0.9662	0.9448	0.9352	0.9579	0.9034	0.9448
2021	0.9731	0.9531	0.9676	0.9490	0.9697	0.9462	0.9434	0.9579	0.9255	0.9517
2022	0.9793	0.9462	0.9579	0.9331	0.9593	0.9379	0.9462	0.9545	0.8952	0.9441

TIGER

Train \ Test	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.9834	0.9545	0.9607	0.9428	0.9676	0.9352	0.9317	0.9483	0.9041	0.9545
2014	0.9800	0.9648	0.9690	0.9545	0.9745	0.9462	0.9517	0.9572	0.8979	0.9428
2015	0.9793	0.9628	0.9648	0.9662	0.9766	0.9531	0.9476	0.9566	0.8966	0.9497
2016	0.9772	0.9538	0.9559	0.9476	0.9710	0.9428	0.9490	0.9572	0.8917	0.9455
2017	0.9766	0.9517	0.9428	0.9483	0.9662	0.9448	0.9345	0.9524	0.8966	0.9414
2018	0.9759	0.9531	0.9455	0.9386	0.9690	0.9469	0.9386	0.9621	0.9090	0.9386
2019	0.9676	0.9386	0.9428	0.9310	0.9641	0.9352	0.9283	0.9441	0.8897	0.9276
2020	0.9759	0.9469	0.9634	0.9407	0.9641	0.9262	0.9400	0.9545	0.8848	0.9455
2021	0.9766	0.9524	0.9531	0.9290	0.9614	0.9262	0.9455	0.9531	0.9255	0.9510
2022	0.9697	0.9434	0.9559	0.9448	0.9628	0.9421	0.9441	0.9566	0.8979	0.9517

4. ADDITIONAL RESULTS - TEMPORAL INTERVAL (YEARS)

Table S15 and Table S16 illuminates the effectiveness of our model in mitigating temporal degradation using results derived from the expanded TempEL dataset. Here, we use the Top 200 degree samples (10k) as the training set because we aim to explore the impact of entity relationship information on the results. Each column in the table represents the years' gap between the training and testing datasets, as denoted by the digits from 0 to 9. For instance, 0 implies that training and testing datasets come from the same year, while 9 indicates that model was trained in 2013 and tested in 2022. The rows are divided based on various metrics: @1 and @8. The complete table includes @1, @2, @4, @8, @16, @32 and @64 can be found in the supplementary material.

"Forward and Backward" means the results include two scenarios: training in the past and testing in the future, and training in the future and testing in the past. "Only Forward" means just the scenario where the model is trained in the past and tested in the future. For instance, when @1 and the gap is 9, our model's result under "Forward and Backward" is 0.2466, representing the average of two situations: training in 2013 and testing in 2022, and training in 2022 and testing in 2013. Under "Only Forward", our model's result is 0.2745, which only considers the situation of training in 2013 and testing in 2022. It's worth noting that when the gap is 0, it means the training and testing datasets come from the same year. Thus, the "Forward and Backward" and "Only Forward" values are identical.

Table S15. Results between our model and the baseline model Forward and Backward.

		Forward and Backward									
		0	1	2	3	4	5	6	7	8	9
@1	BLINK	0.2169	0.2222	0.2208	0.2247	0.2228	0.2205	0.2216	0.2197	0.2147	0.2114
	TIGER	0.2298	0.2220	0.2248	0.2232	0.2261	0.2234	0.2190	0.2258	0.2307	0.2466
	Boost	5.95%	-0.09%	1.81%	-0.67%	1.48%	1.32%	-1.17%	2.78%	7.45%	16.65%
@2	BLINK	0.3104	0.3135	0.3139	0.3182	0.3139	0.3093	0.3109	0.3065	0.2988	0.2987
	TIGER	0.3170	0.3141	0.3189	0.3133	0.3165	0.3188	0.3161	0.3217	0.3235	0.3266
	Boost	2.13%	0.19%	1.59%	-1.54%	0.83%	3.07%	1.67%	4.96%	8.27%	9.34%
@4	BLINK	0.4097	0.4115	0.4128	0.4198	0.4104	0.4090	0.4085	0.4024	0.3907	0.3990
	TIGER	0.4144	0.4145	0.4185	0.4126	0.4158	0.4204	0.4150	0.4202	0.4193	0.4335
	Boost	1.15%	0.73%	1.38%	-1.72%	1.32%	2.79%	1.59%	4.42%	7.32%	8.65%
@8	BLINK	0.5117	0.5146	0.5180	0.5238	0.5144	0.5171	0.5164	0.5116	0.5002	0.4890
	TIGER	0.5170	0.5182	0.5216	0.5206	0.5182	0.5208	0.5176	0.5229	0.5235	0.5362
	Boost	1.04%	0.70%	0.69%	-0.61%	0.74%	0.72%	0.23%	2.21%	4.66%	9.65%
@16	BLINK	0.6148	0.6141	0.6205	0.6224	0.6181	0.6118	0.6192	0.6089	0.6035	0.5876
	TIGER	0.6162	0.6180	0.6178	0.6176	0.6186	0.6187	0.6150	0.6263	0.6233	0.6345
	Boost	0.23%	0.64%	-0.44%	-0.77%	0.08%	1.13%	-0.68%	2.86%	3.28%	7.98%
@32	BLINK	0.7054	0.7077	0.7088	0.7140	0.7076	0.7044	0.7048	0.7025	0.6966	0.6897
	TIGER	0.7022	0.7058	0.7088	0.7056	0.7065	0.7066	0.7011	0.7076	0.7124	0.7221
	Boost	-0.45%	-0.27%	0.00%	-1.18%	-0.16%	0.31%	-0.52%	0.73%	2.27%	4.70%
@64	BLINK	0.7791	0.7867	0.7858	0.7896	0.7836	0.7807	0.7768	0.7781	0.7743	0.7721
	TIGER	0.7818	0.7830	0.7862	0.7834	0.7905	0.7839	0.7825	0.7907	0.7917	0.7931
	Boost	0.35%	-0.47%	0.05%	-0.79%	0.88%	0.41%	0.73%	1.62%	2.25%	2.72%

Table S16. Results between our model and the baseline model Only Forward.

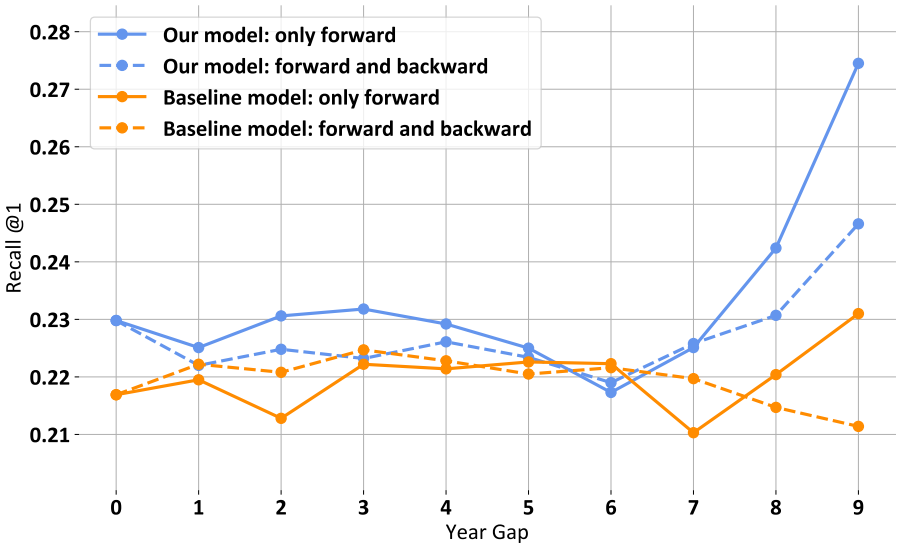
		Only Forward									
		0	1	2	3	4	5	6	7	8	9
@1	BLINK	0.2169	0.2195	0.2128	0.2222	0.2214	0.2226	0.2223	0.2103	0.2204	0.2310
	TIGER	0.2298	0.2251	0.2306	0.2318	0.2292	0.2250	0.2173	0.2251	0.2424	0.2745
	Boost	5.95%	2.55%	8.36%	4.32%	3.52%	1.08%	-2.25%	7.04%	9.98%	18.83%
@2	BLINK	0.3104	0.3113	0.3065	0.3165	0.3125	0.3091	0.3016	0.2917	0.3024	0.3276
	TIGER	0.3170	0.3172	0.3256	0.3226	0.3193	0.3188	0.3074	0.3278	0.3349	0.3607
	Boost	2.13%	1.90%	6.23%	1.93%	2.18%	3.14%	1.92%	12.38%	10.75%	10.10%
@4	BLINK	0.4097	0.4099	0.4058	0.4188	0.4126	0.4076	0.4038	0.3908	0.3925	0.4290
	TIGER	0.4144	0.4195	0.4273	0.4229	0.4198	0.4174	0.4100	0.4326	0.4338	0.4648
	Boost	1.15%	2.34%	5.30%	0.98%	1.75%	2.40%	1.54%	10.70%	10.52%	8.34%
@8	BLINK	0.5117	0.5168	0.5116	0.5255	0.5160	0.5197	0.5176	0.5071	0.5017	0.5310
	TIGER	0.5170	0.5292	0.5304	0.5304	0.5212	0.5216	0.5157	0.5388	0.5466	0.5710
	Boost	1.04%	2.40%	3.67%	0.93%	1.01%	0.37%	-0.37%	6.25%	8.95%	7.53%
@16	BLINK	0.6148	0.6186	0.6182	0.6189	0.6220	0.6210	0.6288	0.6085	0.6159	0.6359
	TIGER	0.6162	0.6297	0.6254	0.6248	0.6217	0.6234	0.6212	0.6393	0.6476	0.6697
	Boost	0.23%	1.79%	1.16%	0.95%	-0.05%	0.39%	-1.21%	5.06%	5.15%	5.32%
@32	BLINK	0.7054	0.7146	0.7067	0.7106	0.7124	0.7143	0.7161	0.7053	0.7066	0.7503
	TIGER	0.7022	0.7147	0.7135	0.7105	0.7090	0.7137	0.7085	0.7179	0.7366	0.7648
	Boost	-0.45%	0.01%	0.96%	-0.01%	-0.48%	-0.08%	-1.06%	1.79%	4.25%	1.93%
@64	BLINK	0.7791	0.7911	0.7816	0.7809	0.7856	0.7927	0.7876	0.7805	0.7845	0.8214
	TIGER	0.7818	0.7914	0.7886	0.7869	0.7940	0.7918	0.7866	0.7965	0.8107	0.8276
	Boost	0.35%	0.04%	0.90%	0.77%	1.07%	-0.11%	-0.13%	2.05%	3.34%	0.75%

5. PERFORMANCE ANALYSIS

Figure S2 to Figure S5 displays recall@N results from the expanded TempEL dataset. We assessed our proposed model against the baseline bi-encoder using recall metrics. The x-axis indicates the year gap between training and testing sets, while the y-axis represents the recall rate. Overall, our model consistently outperforms the baseline.

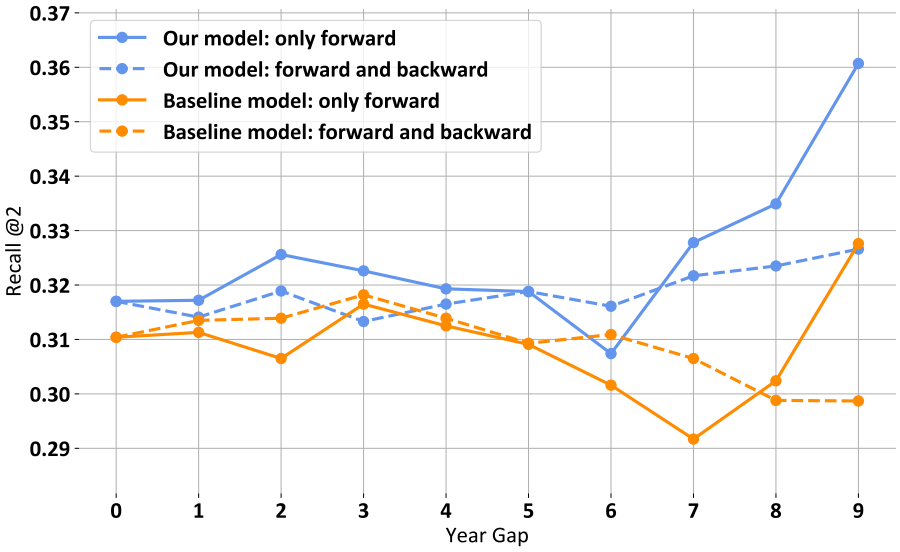
Two testing scenarios are considered: "Forward and Backward" (training on past data and testing on future data, and vice versa) and "Only Forward" (training on past data and testing on future data). As an example, for a gap of 9 years at @1, our model achieves a recall of 0.2466 in "Forward and Backward", averaging results from both 2013 to 2022 and 2022 to 2013. In the "Only Forward" scenario, the recall is 0.2745, solely accounting for 2013 to 2022. Notably, a gap of 0 indicates identical training and testing years, making "Forward and Backward" and "Only Forward" values the same.

At all recall@N levels, our model's "Only Forward" setting consistently outperforms its "Forward and Backward" counterpart. This indicates that our proposed model exhibits superior generalization capabilities, particularly when predicting the future based on past data. Also, a trend can be observed where the recall rates for both models tend to increase as we move from recall@1 to recall@8. This is expected as a higher recall@N allows for more chances to correctly identify the relevant entities.

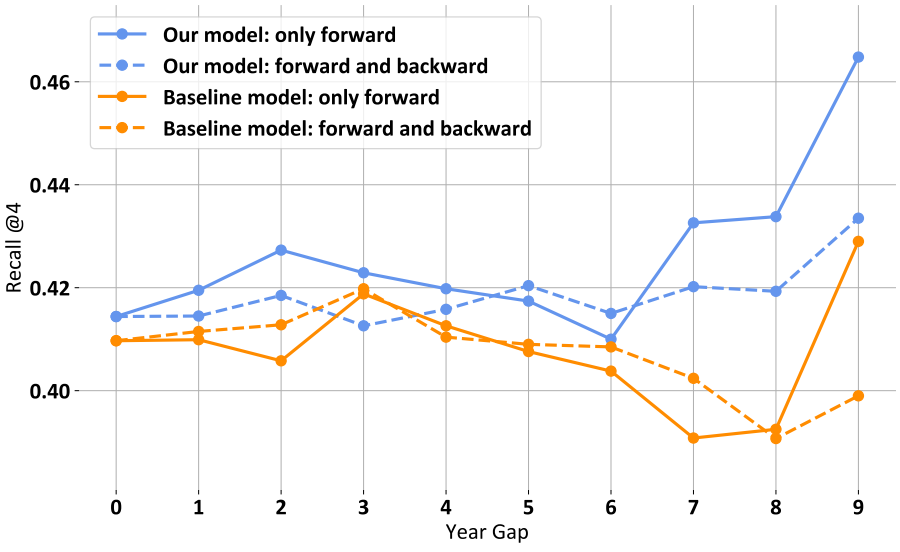


(a) @1

Fig. S2. The improvement in the metrics @1.

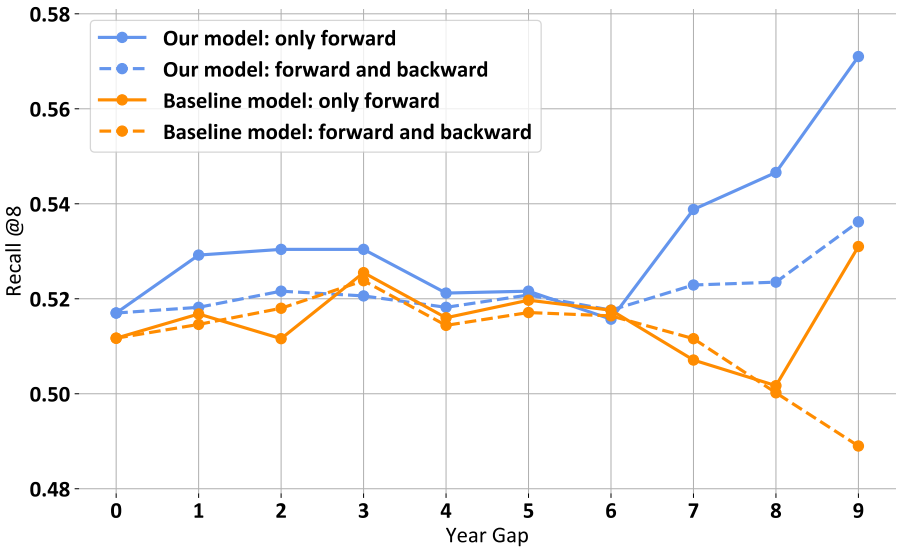


(a) @2

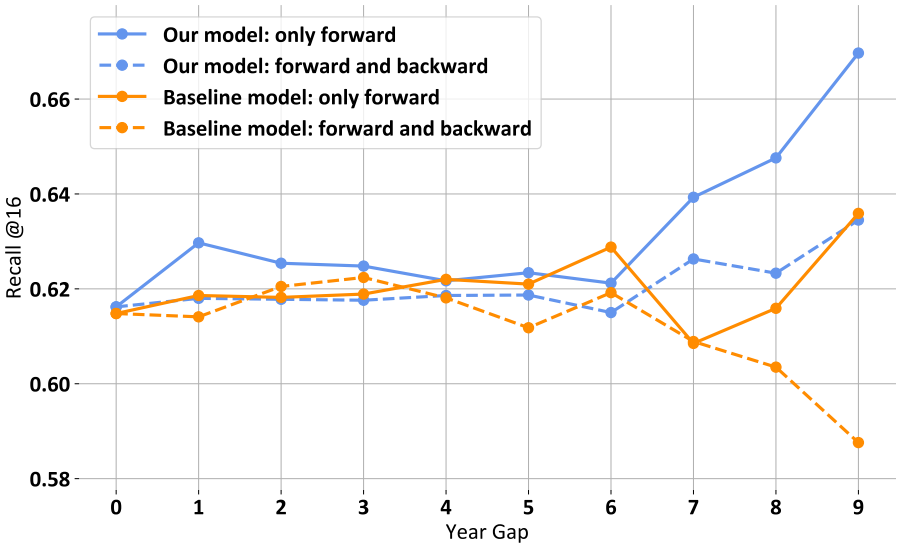


(b) @4

Fig. S3. The improvement in the metrics @2 and @4.

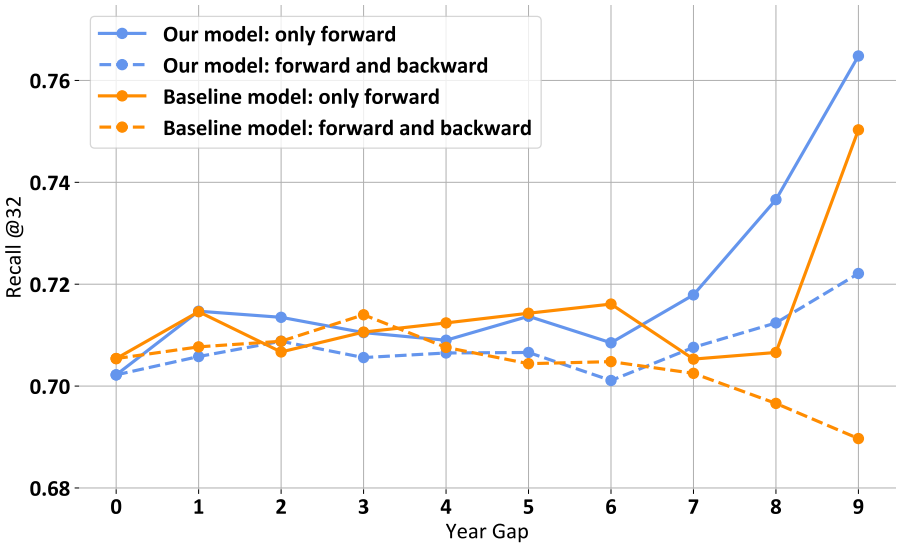


(a) @8

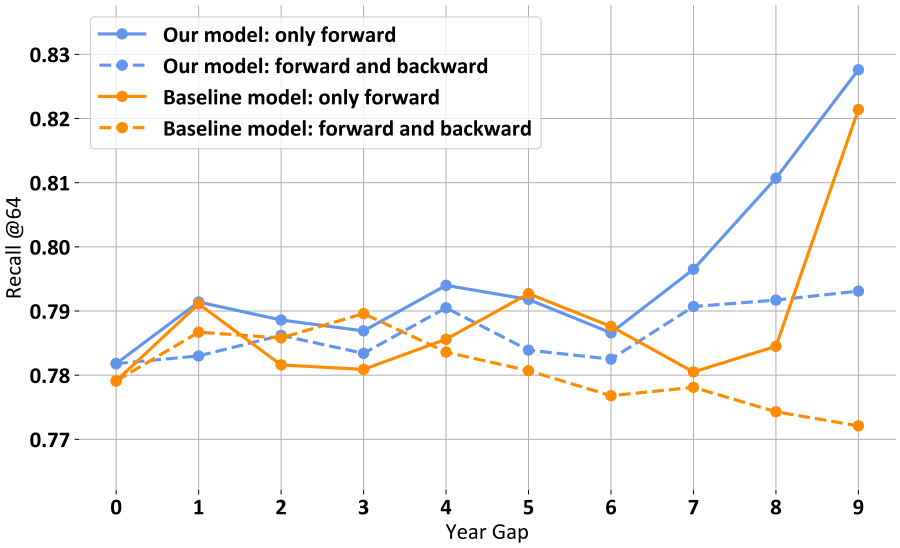


(b) @16

Fig. S4. The improvement in the metrics @8 and @16.



(a) @32

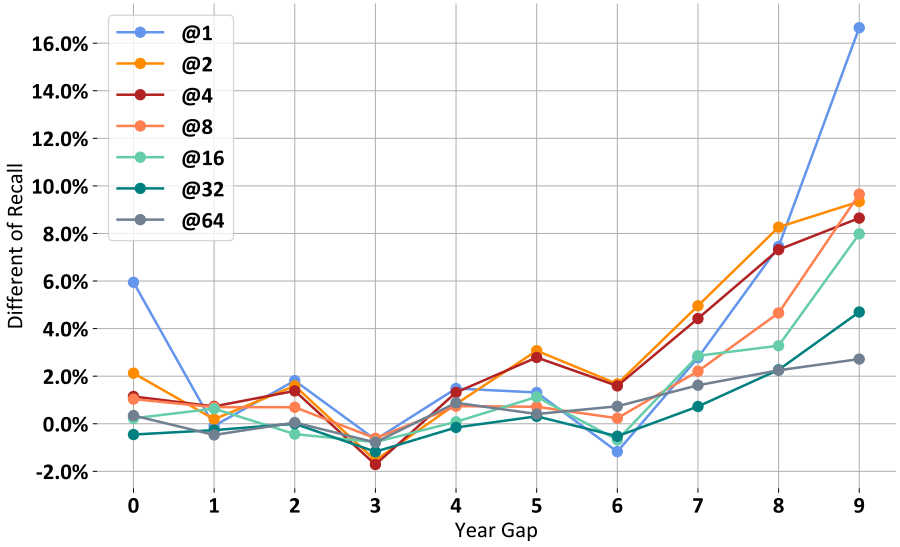


(b) @64

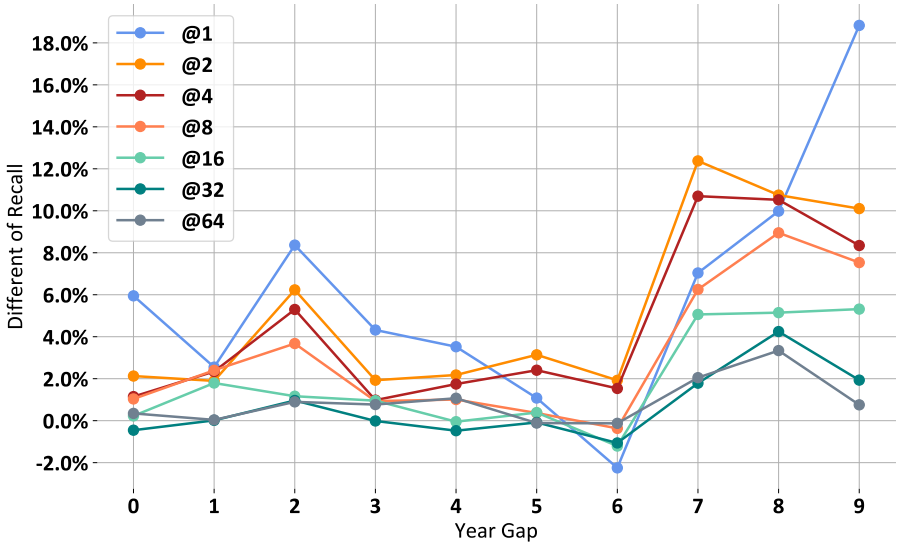
Fig. S5. The improvement in the metrics @32 and @64.

6. DIMINISHING PERFORMANCE GAINS

It is also worth noting that the improvement effect of our model diminishes gradually as the metric threshold becomes more lenient, as shown in Figure S6. A plausible explanation for this observation is when using @64 threshold, the model only needs to correctly predict one out of the top 64 answers, allowing for a higher tolerance of errors. Consequently, the relative performance improvement of our model becomes less evident at these higher thresholds. This indicates a trade-off between prediction accuracy and the employed scale, suggesting the need for careful balance in practical applications.



(a) Forward and Backward



(b) Only Forward

Fig. S6. Different of Recall changes as the metric changes.