(19.0%), Llama3-8b (23.9%), and Llama3-8b-SFT (21.6%). These results indicate that the **unstructured format outperforms the structured format for LLMs in script event prediction** (unsure for temporal knowledge graph completion).

B.6 Prompt Templates

The prompt templates for six tasks are shown in Table 10. The prompt templates for dataset construction are shown in Table 11 and Table 12.

Type		Definition
Misssing	Semantic	The predictions and labels refer to the same event but differ in textual representation.
	Commensense	The predictions cannot be verified through existing evidences but can be easily inferred from them through commonsense reasoning. For example, "the escalating tension between the two countries" can be easily inferred from the "increased military deployment and confrontation in the border region", even if there is no relevant reports.
	Ongoing	The predictions are ongoing events of given background but are not mentioned in labels and subsequent reports. For example, given event background including assistance, there will be ongoing material and medical assistance until the rescue operation is finished, even if there is no relevant content in subsequent report.
overestimation	Semantic	The prediction and label describe different events but are erroneously matched.
	Pessimistic	The model makes a pessimistic forecasting but actually not. For example, in testing sample about "2023 Chitral cross-border attacks", the prediction is "The Pakistani military will launch a major offensive against TTP insurgents in the border regions, resulting in a significant increase in violence and displacement". In contrast, the real development is "Pakistani ambassador Asif Durrani led a visit of delegation to Afghanistan and Afghan authorities took actions to mitigate the conflict".
	plausible	The prediction is plausible, but not happen according to the event background and relevant reports.

Table 6: The definitions of the error types in open-ended evaluation methods.

Task	**2023 Chitral cross-border attacks**
Extracted Timeline	(1) On 7 September 2023, the TTP attacked five military checkposts from Afghan territory, killing 4 Pakistani soldiers. (2) On 7 September 2023, the Pakistani military thwarted the attack and killed 12 TTP insurgents. (3) Following the attack, Pakistan's primary Torkham border crossing with Afghanistan remained closed for a second consecutive day on 7 September 2023, resulting in a backlog of trucks carrying goods. (4) The Pakistani army initiated a combing operation in response to the attacks, targeting the neighboring Afghan provinces of Kunar and Nuristan. (5) On 8 September 2023, Pakistan summoned the head of the Afghan diplomatic mission in Islamabad to lodge a robust protest concerning the attacks. (6) On 8 September 2023, the Foreign Office (FO) issued a stern demarche to the Afghan interim government, emphasizing that such attacks and the utilization of Afghan territory by the TTP would pose a significant threat to peace and security. (7) The Caretaker Foreign Minister, Jalil Abbas Jilani, described the recent terrorist attack on two military posts in the Lower Chitral district of Khyber Pakhtunkhwa as an ïsolated incident.̈
Prediction on 8 September	(1) The TTP will continue to carry out cross-border attacks from Afghanistan into Pakistan. (2) The Pakistani military will respond with retaliatory strikes against TTP hideouts in Afghanistan. (3) The Pakistani government will request the Afghan government to take action against TTP insurgents operating from its territory. (4) The Afghan government will deny the presence of TTP insurgents on its soil and refuse to take action. (5) The Pakistani government will seek international support to pressure the Afghan government to take action against TTP insurgents. (6) The TTP will intensify its attacks on Pakistani military and civilian targets, leading to increased tensions between Pakistan and Afghanistan. (7) The Pakistani military will launch a major offensive against TTP insurgents in the border regions, resulting in a significant increase in violence and displacement. (8) The international community will call for a ceasefire and a negotiated settlement to the conflict. (9) The Pakistani government and the TTP will enter into peace talks, but the talks will break down due to irreconcilable differences. (10) The conflict will continue to simmer, with occasional outbreaks of violence, for the foreseeable future.

Table 7: An example of Mixtral-8x7b on STF task. The model makes pessimistic forecasting while actually not.

Model	STF						LTF					
	LRAE w/o RA			LRAE			LRAE w/o RA			LRAE		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Llama2-7b	9.6	51.7	15.7	47.5	83.7	57.9	11.7	42.3	17.0	39.7	64.3	46.3
Llama2-13b	11.9	57.4	18.9	47.0	81.4	56.6	11.0	34.0	15.2	33.3	54.6	38.5
Llama2-70b	11.1	61.0	18.3	42.7	81.7	57.2	10.9	45.7	16.4	41.4	69.5	48.2
Vicuna-13b	7.3	35.9	11.5	30.2	59.0	36.7	6.5	18.3	9.0	25.5	42.7	29.6
Wizardlm-13b	11.3	35.0	15.9	40.3	57.6	44.5	6.7	23.6	9.5	26.6	42.6	29.7
Falcon-40b	11.0	19.7	13.1	30.3	35.1	31.6	9.6	20.6	11.2	31.9	41.4	32.7
Mixtral-8x7b	10.6	49.2	16.4	42.8	74.2	51.3	8.0	33.2	11.9	35.4	54.1	39.8
Llama3-8b	19.9	50.0	27.0	58.7	69.5	60.3	13.6	34.2	17.9	38.0	55.4	43.0
SFT	24.0	17.9	19.7	39.6	25.0	29.4	10.1	14.1	10.5	12.9	16.2	13.2

Table 8: Evaluation results on STF and LTF tasks using LRAE evaluation. *RA* denotes the retrieval augmentation module.

Model	STF						LTF					
	BEM			BEM+RA			BEM			BEM+RA		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Llama2-7b	2.5	12.0	4.0	11.8	32.4	16.6	4.6	19.2	6.9	14.5	33.6	18.3
Llama2-13b	1.6	8.4	2.7	11.5	25.3	14.6	5.2	20.4	7.7	13.9	30.0	17.3
Llama2-70b	1.7	8.7	2.8	11.4	28.9	15.3	8.7	36.8	13.0	13.6	43.8	19.1
Vicuna-13b	1.6	7.7	2.3	11.6	21.4	14.0	2.2	6.8	3.2	15.7	22.2	16.7
Wizardlm-13b	2.0	5.8	2.9	14.9	21.9	16.5	4.0	7.9	4.8	15.6	22.4	17.4
Falcon-40b	2.0	3.7	2.3	10.2	12.8	10.5	3.0	6.5	3.6	17.4	21.1	17.5
Mixtral-8x7b	2.3	8.3	3.5	18.6	31.8	21.9	1.5	6.1	2.3	10.8	21.5	13.5
Llama3-8b	6.0	6.9	6.3	21.9	18.7	19.1	1.9	2.8	2.2	11.9	13.9	12.3
SFT	13.5	13.9	13.4	13.6	13.9	13.4	36.5	43.9	35.0	38.0	44.9	36.0

Table 9: Evaluation results on STF and LTF tasks using BEM evaluation. *RA* denotes the retrieval augmentation module.