G5 Project Proposal: Can Multimodal Models Discern Episodic Events Across Memory Cues? – An Episodic Memory Benchmark

Cassandra TAN Hui Ming & Divya MITTAL & CHUA Joon Kiat & SIM Wei Xuan Samuel
School of Computing and Information System
Singapore Management University
Singapore
{hm.tan.2023,divyam.2023}@phdcs.smu.edu.sg,
{jk.chua.2023,samuel.sim.2024}@msc.smu.edu.sg

1 Introduction and Motivation

The ability to discern experiences and events across space and time in humans is linked to episodic memory. Episodic memory allows us to encode and retrieve events from the past and the contextual content associated with the event (Mayes & Roberts, 2001; Tulving, 2002; Cox et al., 2018). This key memory hallmark enables the retrieval of autobiographical life experiences and specific events across time and space (Tulving, 2002; Mayes & Roberts, 2001), such as, recalling a birthday party that happened prior to a family holiday and the people involved. Beyond that, episodic memory plays several crucial roles in human cognition, such as mental time travel for predicting and planning ahead (Szpunar et al., 2013), and improving learning and problem solving efficacy across different age groups (Vandermorris et al., 2013; Peters et al., 2019). Collectively, we cannot understate episodic memory's importance in enabling key cognitive functions in humans.

In Artificial Intelligence, the capacity for episodic-like memory capabilities has been proposed as vital for Large Language Models (LLMs)-based agents to operate and reason across long contexts for task planning and handling (Pink et al., 2025; Fountas et al., 2024; Wu et al., 2025). In robotics, episodic memory representations has led to improved decision-making and action-planning (Bärmann et al., 2021). As such, recent research has focused on embedding human-inspired episodic memory frameworks into LLMs and Multimodal Models (MLLMs), to improve long-context understanding and efficient knowledge update (Das et al., 2024; Wang et al., 2025; Lando et al., 2025).

Despite the increasing interest in episodic memory capabilities, existing benchmarks are not fully-suited to assess episodic event retrieval capabilities in MLLMs. The reasons are multifold. Most existing benchmarks focus on action-object centric spatial-temporal reasoning over videos (Li et al., 2024; Plizzari et al., 2025). While important, they primarily test the model's ability to interpret and reason about visual context, and lean towards information retrieval rather reconstructing episodic events (Huet et al., 2025). Additionally, episodic benchmarks such as Ego4D (Grauman et al., 2022) and EMQA (Bärmann & Waibel, 2022) are focused on egocentric (first-person) videos, and questions are phrased as *"Where was the object last seen"* or *"What happens after a given action"* in a continuous video. These tasks do not assess the ability to track entity states or relationships across different episodic events (Huet et al., 2025). For instance, they do not ask questions related to the temporal order of episodic events across shared cues (e.g., *"What is the order of events that occurred in the same space with the same people?"*). Collectively, this suggests a need for more complex episodic memory tasks in MLLMs beyond egocentric perspectives.

To close this gap, we propose a new taxonomy of episodic memory tasks that culminates into a new **Episodic Memory Tasks and Benchmarks (EMTB)**. EMTB is designed to probe episodic retrieval capabilities of MLLMs from non-egocentric perspectives. In doing so, we aim to answer the following research questions - (1) Can MLLMs Discern and Track Episodic Events Across Different Memory Cues? (2) How Good are MLLMs Episodic Memory Capabilities beyond Egocentric Videos? (3) Can MLLMs Reason Beyond Short or Long Context Videos? (4) Do multiple-choice options act as cues that inflate model's accuracy? With this, we aim not only to develop a series of tasks to evaluate MLLMs more robustly but also to push the understanding of MLLMs' episodic-memory capabilities and their viability as real-world planning agents.

2 Related Works

2.1 Spatial-Temporal Benchmarks in Multimodal Models

Several benchamrks have emered to evaluate the spaital-temporal reasoning capabilities of MLLMs (Li et al., 2024; Liu et al., 2024; Plizzari et al., 2025; Fu et al., 2025). For instance, MVBench (Li et al., 2024) emphasizes dynamic video tasks to systematically test a range of temporal skills, demonstrating that MLLMs have far from satisfactory understanding. Similarly, TempCompass (Liu et al., 2024) evaluates MLLMs on fine-grained temporal aspects (e.g, action, speed, direction, attribute change, and event order). More recently, benchmarks like Video-MME (Fu et al., 2025) and Omnia de EgoTempo (Plizzari et al., 2025) evaluate MLLMs on video understanding across diverse video types and temporal conditions. These benchmarks advance spatial–temporal evaluation, but they tend to emphasize perceptual reasoning, retrieval of a single information based on a given cue (e.g., *what*) Cheng et al. (2025); Wang et al. (2025), rather than reconstructing or retrieving the complex trace of episodic events and context (e.g., *who-where*) (Huet et al., 2025). Therefore, there is a need to develop episodic memory relevant tasks as the existing spatial-temporal tasks do not adequately assess episodic memory capabilities. Thus, we extend this by developing a new taxonomy of tasks that focuses on episodic events retrieval and reasoning.

2.2 Existing Episodic Memory Tasks and Benchmarks

Another line of work focuses more directly on episodic memory event retrieval. In the context of LLMs, Huet et al. (2025) proposed an episodic memories evaluation benchmark, which focused on tracking entities across time and space in the context of book chapters. Results show that SOTA LLMs struggle with episodic memory tasks, especially when dealing with multiple related events across time and space. In the context of videos, egocentric benchmarks like Ego4D (Grauman et al., 2022) and EMQA (Bärmann & Waibel, 2022) focus on Question-Answer and retrieval tasks from an egocentric perspective. These benchmarks test aspects of episodic recall in retrieving information from the correct temporal window, but lack the necessary complexity of ordering and disentanglement of event relationships across time and space. As such, the availability of more complex episodic memory tasks are still relatively unexplored in MLLMs. Also, such tasks are focused on retrieving from a single video context. Therefore, little is known about how MLLMs can perform when retrieving and reasoning about episodic events across multiple videos. We build on this by developing more complex episodic memory tasks in non-egocentric videos.

3 Methodology

3.1 Episodic Memory Representation

We extend the concept of *episodic memory cues* as proposed by Huet et al. (2025) to encode the key components of our episodic memory benchmark using a structured tuple:

$$(\text{Time } t,\ \text{Location } \phi,\ \text{Subjects } \sigma,\ \text{Events } \epsilon,\ \text{Label } \alpha).$$

These represent the fundamental dimensions of episodic memory, where any permutation of the components can serve as retrieval cues (Huet et al., 2025). Time (t) denotes when the event occurs, represented either by explicit timestamps or a temporal index starting from zero. Location (ϕ) specifies the spatial context or setting of the scene. Subjects (σ) identify the participants involved in the event. Events (ϵ) capture the actions or situations depicted within the clip. Finally, Label (α) encodes the sequential order of the clip (e.g., episode 1 – segment 1; episode 2 – segment 3). The context dimension indicates whether related events occur within the same episode (e.g., episode 1 – segments 1 and 2) or across different episodes (e.g., episode 1 – segment 3 and episode 2 – segment 1), allowing the representation of linked events both within and between episodic memories. In our benchmark, each memory is constructed from *script-based event extraction*, where the video script or subtitles are processed to identify key episodic attributes — time, location, subjects, events, and label. These attributes are used to form structured memory tuples representing individual episodic events. Furthermore, our proposed tasks evaluates models on holistic question-answer pairs constructed from various permutations of these cues. This enables assessment of how well models reason across various episodic memory cues.

3.2 Benchmark Dataset

EMTB builds upon **TVQA-Long**, introduced in *Goldfish: Vision-Language Understanding of Arbitrarily Long Videos* (Ataallah et al., 2024). TVQA-Long extends the original **TVQA** dataset; comprising 152,545 QA pairs across 21,793 video clips from six popular TV shows (*Friends*, *The Big Bang Theory*, *How I Met Your Mother*, *House M.D.*, *Grey's Anatomy*, *Castle*). This sums up to over 460 hours of video. To evaluate long-form comprehension, the authors aggregated 5, 10, and 20 consecutive clips into longer continuous segments averaging 6, 12, and 24 minutes, preserving the same question–answer pairs across full episodes. TVQA-Long uses the same train/test splits as TVQA, with an additional 10% validation subset introduced for development. Each question has five multiple-choice answers (random baseline = 20%), and evaluation metrics include QA accuracy and retrieval precision for locating evidence segments. We use this readily available dataset to construct our benchmark.

3.2.1 Dataset Annotation Strategy

This section outlines our proposed benchmark dataset creation pipeline for long video understanding. In the *memory formation* stage, long videos are divided into event clips using **script-based event extraction**. A few events are manually annotated with key attributes — *time*, *location*, *subject*, *event*, and *label* — to provide reference examples. These examples are then used to perform **in-context learning (ICL)** on a LLM, which automatically extracts similar attributes from the remaining clips. This approach converts unstructured video content into a structured memory representation efficiently and with minimal manual effort.

3.3 Task Taxonomy and Question-Answer Pairs Construction

To evaluate MLLMs' performance on episodic event awareness, our proposed **EMTB** introduces new episodic memory tasks, which are constructed from event tuples extracted from non-egocentric video contexts annotated with $\{$*Time, Location, Subject, Event, Label*$\}$. The objective is to examine whether MLLMs can demonstrate episodic-memory-linked spatial and temporal reasoning, by accurately recalling, ordering, or linking events across time and space.

Question answer (QA) pairs are generated from structured event tuples. As illustrated in Appendix A, a template-based generation procedure maps elements of the event tuple to natural language questions. Each question is instantiated with corresponding contextual values, producing both open-ended (OE) and closed-form multiple-choices question (MCQ) versions using a unified question format. Closed-form MCQ options are generated by sampling alternative candidates for the target attribute while maintaining a consistent format. In short, each task follows a cue and target formulation with unified question templates for both OE and MCQ formats (see Appendix B).

3.3.1 Task Dimensions

Our proposed tasks encompasses three major dimensions aligned with episodic memory representation. The first task is **Temporal Ordering**, where we test MLLMs' ability to reconstruct event chronology across overlapping contexts that require tracking the relationship in subjects, locations and events across time and space. Next, we evaluate **Episodic Recall**, where we provide single-target and dual-cue recalls. Tasks here emphasize understanding temporal transitions across episode based on the given cues. Lastly, we evaluate **inter-video task performance** beyond single continuous contexts, to assess whether MLLMs can integrate episodic information across multiple temporally or thematically linked videos, simulating human-like episodic coherence across contexts. An elaboration of the above tasks with examples can be found in the task taxonomy in Appendix B.

3.4 Experiment Setup and Protocol

The evaluation protocol follows two paradigms. (1) **Vanilla MLLM Evaluation**: Models are tested under their native context window constraint. We will use standardize the context window based on the smallest model. Discontinuous event clips, confined within the maximum token capacity, are input into the model alongside QA pairs that reference only the provided clips. This setup isolates the model's intrinsic ability to encode and integrate temporal, spatial, entity, and event-level cues within a limited context. (2) **Planned models and baselines:** We plan to compare across the following vanilla open-weight models - Qwen2.5-VL-7B, Qwen2.5-VL-32B, Qwen3-VL, LLaVa-Video-Llama-3.1-8B. For proprietary models, we plan to use Gemini and OpenAI's ChatGPT (if possible). (3) **Episodic Memory Framework Evaluation**: Subject to time constraints, we will also utilize existing episodic memory frameworks (Wang et al., 2025) to retrieve relevant event clips from the entire corpus based on each question. The retrieved content is then used for QA evaluation. This paradigm examines the frameworks' effectiveness in long-range episodic retrieval and their use of multi-cue memory access mechanisms.

4 Evaluation

The evaluation stage is designed to systematically assess MLLMs' performance on our proposed **EMTB**. Following the benchmark design, evaluation focuses on testing both open-ended (OE) and closed-form multiple-choices question (MCQ) outputs under different episodic-memory conditions. For every task, performance is computed using a combination of quantitative metrics that capture correctness, recall, generative quality, and ordering accuracy. The evaluation serves two primary objectives - (1) To provide a grounded and reproducible framework for measuring spatial-temporal-event awareness in MLLMs. (2) To benchmark performance differences between baseline models and those augmented with episodic-memory representations or retrieval mechanisms.

4.1 Evaluation Metrics

We propose the following metrics to comprehensively capture multiple aspects of model performance on our episodic memory tasks, including retrieval accuracy, precision–recall balance, generation quality, and temporal ordering consistency. The first set of metrics comprise of **Accuracy**, **Precision**, **Recall**, and **F1 Score**. We use these to assess the correctness of retrieved or generated answers relative to ground truth event labels. Next, we have **String-Matching Metrics** where we use Exact Match (EM) to assess the strict correctness of text-based responses in structured QA. We also use **BLEU** and **ROUGE** to measure n-gram overlaps between model-generated text and gold references, reflecting the quality outputs in non-closed form answers. Lastly, we use **Kendall's** τ, to evaluate rank correlation between predicted and true event sequences, to assess chronological coherence in temporal ordering tasks. These metrics collectively enable both fine-grained evaluation (at the individual QA level) and holistic assessment (aggregated across episodes or task types). This aligns with the benchmark's goal of understanding how well MLLMs perform on **EMTB's** episodic memory tasks. A complete table of metrics with examples are provided in Appendix C.

5 Risks

This section highlights the risks based on the scope and duration given for project completion. (i) The first risk is the time required to annotate the large dataset across multiple videos. To mitigate this, we will use SOTA models and automated scripts where possible. (ii) The second risk is the computational requirements needed to run some of the MLLMs. We will start with smaller baseline models, and try to run larger models if possible. (iii) Lastly, to work within the time constraint, we may ultimately annotate a smaller subset of the dataset first to run our experiments.

6 Team Execution Plan

Our Team Leader for the project is Cassandra Tan. We have broken down our project plan into four phases - (i) Research and Exploration, (ii) Dataset and Labels Annotation, (iii) Experiment Setup and Run, (iv) Final Report and Compilation. All members are expected to contribute to all sub-tasks within each phase, but a lead will be assigned to each phase for planning and strategizing. Details on the team execution plan broken down into phases and sub-tasks by week is found in Appendix D.

References

Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Mingchen Zhuge, Jian Ding, Deyao Zhu, Jürgen Schmidhuber, and Mohamed Elhoseiny. Goldfish: Vision-language understanding of arbitrarily long videos. In *European Conference on Computer Vision*, pp. 251–267. Springer, 2024.

Leonard Bärmann, Fabian Peller-Konrad, Stefan Constantin, Tamim Asfour, and Alex Waibel. Deep episodic memory for verbalization of robot experience. *IEEE Robotics and Automation Letters*, 6(3):5808–5815, 2021.

Leonard Bärmann and Alex Waibel. Where did i leave my keys? — episodic-memory-based question answering on egocentric videos. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1559–1567, 2022. doi: 10.1109/CVPRW56347.2022.00162.

Zixu Cheng, Jian Hu, Ziquan Liu, Chenyang Si, Wei Li, and Shaogang Gong. V-star: Benchmarking video-llms on video spatio-temporal reasoning. *arXiv preprint arXiv:2503.11495*, 2025.

Gregory E Cox, Pernille Hemmer, William R Aue, and Amy H Criss. Information and processes underlying semantic and episodic memory across tasks, items, and individuals. *Journal of Experimental Psychology: General*, 147(4):545, 2018.

Payel Das, Subhajit Chaudhury, Elliot Nelson, Igor Melnyk, Sarathkrishna Swaminathan, Sihui Dai, Aurélie Lozano, Georgios Kollias, Vijil Chenthamarakshan, Jiří Navrátil, Soham Dan, and Pin-Yu Chen. Larimar: large language models with episodic memory control. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Zafeirios Fountas, Martin A Benfeghoul, Adnan Oomerjee, Fenia Christopoulou, Gerasimos Lampouras, Haitham Bou-Ammar, and Jun Wang. Human-like episodic memory for infinite context llms. *arXiv preprint arXiv:2407.09450*, 2024.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24108–24118, 2025.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Alexis Huet, Zied Ben Houidi, and Dario Rossi. Episodic memories generation and evaluation benchmark for large language models, 2025.

Giuseppe Lando, Rosario Forte, Giovanni Maria Farinella, and Antonino Furnari. How far can off-the-shelf multimodal large language models go in online episodic memory question answering?, 2025.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos?, 2024.

Andrew R Mayes and Neil Roberts. Theories of episodic memory. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 356(1413):1395–1408, 2001.

Sarah L Peters, Carina L Fan, and Signy Sheldon. Episodic memory contributions to autobiographical memory and open-ended problem-solving specificity in younger and older adults. *Memory & cognition*, 47(8):1592–1605, 2019.

M Pink, Q Wu, VA Vo, et al. Position: episodic memory is the missing piece for long-term llm agents (2025), 2025.

Chiara Plizzari, Alessio Tonioni, Yongqin Xian, Achin Kulshrestha, and Federico Tombari. Omnia de egotempo: Benchmarking temporal understanding of multi-modal llms in egocentric videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24129–24138, 2025.

Karl K Szpunar, Donna Rose Addis, Victoria C McLelland, and Daniel L Schacter. Memories of the future: New insights into the adaptive value of episodic memory. *Frontiers in behavioral neuroscience*, 7:47, 2013.

Endel Tulving. Episodic memory: From mind to brain. *Annual review of psychology*, 53(1):1–25, 2002.

Susan Vandermorris, Signy Sheldon, Gordon Winocur, and Morris Moscovitch. Differential contributions of executive and episodic memory functions to problem solving in younger and older adults. *Journal of the International Neuropsychological Society*, 19(10):1087–1096, 2013.

Yun Wang, Long Zhang, Jingren Liu, Jiaqi Yan, Zhanjie Zhang, Jiahao Zheng, Xun Yang, Dapeng Wu, Xiangyu Chen, and Xuelong Li. Episodic memory representation for long-form video understanding, 2025.

Yaxiong Wu, Sheng Liang, Chen Zhang, Yichao Wang, Yongyue Zhang, Huifeng Guo, Ruiming Tang, and Yong Liu. From human memory to ai memory: A survey on memory mechanisms in the era of llms. *arXiv preprint arXiv:2504.15965*, 2025.

A Details of QA Pair Generation

Question answer (QA) pairs are generated from structured event tuples. These event tuples are derived from our episodic memory representation as explained in Section 3.1. As illustrated in Figure 1 below, a template-based generation procedure maps elements of the event tuple to natural language questions. Each question is instantiated with corresponding contextual values, producing both open-ended (OE) and closed-form multiple-choices question (MCQ) versions using a unified question format.

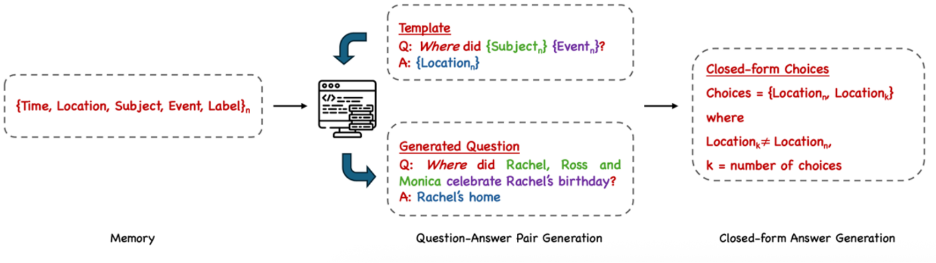
{Time, Location, Subject, Event, Label}$_n$
Template
Q: *Where* did {Subject$_n$} {Event$_n$}?
A: {Location$_n$}
Generated Question
Q: *Where* did Rachel, Ross and Monica celebrate Rachel's birthday?
A: Rachel's home
Closed-form Choices
Choices = {Location$_n$, Location$_k$}
where
Location$_k$ ≠ Location$_n$,
k = number of choices
Memory
Question-Answer Pair Generation
Closed-form Answer Generation

Figure 1: Overview of QA pair generation. Event tuples {Time, Location, Subject, Event, Label} are used to generate both open-ended and closed-form MCQ question answer, by passing a templates to either a SOTA model or an automated script.

B Task Taxonomy

The taxonomy encompasses three major dimensions aligned with the core of an episodic memory:

- **Temporal Ordering:** Tests MLLMs' ability to reconstruct event chronology across overlapping contexts where models are required to track the relationship in subjects, locations and events across time and space.
- **Episodic Event Recall:** Evaluates recall when provided with single-target and dual-cue recalls. Tasks here emphasize understanding temporal transitions across episode based on the given cues.
- **Inter-Video Recall:** Extends beyond single continuous contexts to assess whether MLLMs can integrate episodic information across multiple temporally or thematically linked videos, simulating human-like episodic coherence across contexts.

Category	Subcategory	Cues	Target	Template of Questions
Single Target Recall	Single Event Look-up	T+L+S	E	Given {T,L,S}, what happened?
	Single Subject Look-up	T+L+E	S	Who performed {E} at {T} in {L}?
	Single Location Look-up	T+S+E	L	Where did {S} do {E} at {T}?
	Single Temporal Look-up	L+S+E	T	When did {S} do {E} in {L}?
Dual-Cue Integration	Spatio-TemporalSubject	T+L	S	At {T} in {L}, who was involved?
	Spatio-TemporalEvent	T+L	E	At {T} in {L}, what happened?
	Temporal+SubjectLocation	T+S	L	Where was {S} at {T}?
	Temporal+EventLocation	T+E	L	Where did {E} occur at {T}?
	Location+SubjectTime	L+S	T	When was {S} at {L}?
	Location+EventTime/Subject	L+E	T or S	At {L}, when (or who) for event {E}?
Temporal Reasoning	Temporal Ordering	L+S	$\{E_n\}$	List what happened at {L} when {S} was/were there in chronological order.
	Latest Event Retrieval	L+S	E	What happened at {L} when {S} was/were there most recently?

Table 1: **Cue→Target Taxonomy for Event QA** with unified questions. Events are tuples {Time (T), Location (L), Subject (S), Event (E)}. Each row uses the *same question* for both open-ended (OE) and multiple-choice (MCQ).

C Evaluation Metrics

Metric	Definition	Formulation	Range
Accuracy	Proportion of correctly predicted answers.	$\text{Acc} = \frac{\text{Correct}}{\text{Total}}$	$[0, 1]$
Example:			
Question:	*Who said "We were on a break"?*		
Ground truth:	*Ross Geller*		
Model's answer:	*Ross (correct)*		
Computation:	*If 8 of 10 predictions are correct, Accuracy* $= \frac{8}{10} = 0.80$.		
Precision	Fraction of retrieved items that are relevant (penalizes false positives).	$P = \frac{TP}{TP + FP}$	$[0, 1]$
Example:			
Question:	*Retrieve episodes where Monica cooks Thanksgiving dinner.*		
Ground truth:	*10 relevant episodes.*		
Model's answer:	*Returns 10 episodes; 7 are truly Thanksgiving-with-Monica.*		
Computation:	$TP = 7,\ FP = 3 \Rightarrow P = \frac{7}{7+3} = 0.70$.		
Recall	Fraction of relevant items that are retrieved (penalizes false negatives).	$R = \frac{TP}{TP + FN}$	$[0, 1]$
Example:			
Question:	*Retrieve episodes where Monica cooks Thanksgiving dinner.*		
Ground truth:	*10 relevant episodes.*		
Model's answer:	*Returns 7 correct episodes total.*		
Computation:	$TP = 7,\ FN = 3 \Rightarrow R = \frac{7}{7+3} = 0.70$.		
F1 Score	Harmonic mean of precision and recall (balances over/under-retrieval).	$F1 = 2\,\frac{P \cdot R}{P + R}$	$[0, 1]$
Example:			
Question:	*Same as above.*		
Ground truth:	$P = 0.70,\ R = 0.70$.		
Model's answer:	—		
Computation:	$F1 = 2 \cdot \frac{0.7 \cdot 0.7}{0.7 + 0.7} = 0.70$.		
Exact Match (EM)	Strict correctness: 1 if prediction equals gold string, else 0.	$\text{EM} = \nVdash[\text{pred} = \text{gold}]$	$\{0, 1\}$
Example:			
Question:	*Where do Joey and Chandler live?*		
Ground truth:	*# Apartment 19*		
Model's answer:	*"Apartment 19" (exact string)*		
Computation:	*EM* $=$ *1 for this question; averaged over a set, EM is the mean of 0/1 outcomes.*		
BLEU / ROUGE	N-gram overlap metrics for generated text vs. reference.	BLEU: $\exp\left(\sum_n w_n \log p_n\right)$; ROUGE (Recall-style): $\frac{\text{Overlap}}{\text{Reference}}$	$[0, 1]$
Example:			
Question:	*Why didn't Rachel marry Barry? (short explanation)*		
Ground truth:	*"Rachel left Barry at the altar."*		
Model's answer:	*"Rachel abandoned Barry on the wedding day."*		
Computation:	*High ROUGE and moderate BLEU due to shared n-grams/semantics; EM*$=0$ *since strings differ.*		
Kendall's τ	Rank correlation between predicted and gold event orderings.	$\tau = \frac{C - D}{\frac{n(n-1)}{2}}$	$[-1, 1]$
Example:			
Question:	*Order events chronologically: (A) Ross's London wedding; (B) Rachel moves in with Monica; (C) Joey gets* Days of Our Lives*.*		
Ground truth:	*B* $\rightarrow$ *A* $\rightarrow$ *C (illustrative).*		
Model's answer:	*B* $\rightarrow$ *C* $\rightarrow$ *A.*		
Computation:	*Count concordant/discordant pairs; if 2 of 3 pairs agree,* $\tau = \frac{2-1}{3} = 0.33$.		

Table 2: An overview of the Evaluation Metrics used to measure the model's performance across our tasks. Metrics cover retrieval, generation, ranking, and judgment-based evaluation.

D TEAM EXECUTION PLAN

This appendix section describes the overall team execution plan, broken down into four main phases. Each phase contains several sub-tasks necessary to complete each phase. A team member is assigned as the respective lead for each phase, and they will take charge of planning and strategizing the sub-tasks. Nevertheless, all members are expected to work and contribute meaningfully to each sub-task.

Project Timeline	**W3-W6**	**W7**	**W8**	**W9**	**W10**	**W11**	**W12**	**W13**	**W14**
Phase 1: Research and Exploration - Cassandra									
Literature Review	✓								
Technical Research	✓								
Dataset Selection	✓								
Proposal Writing	✓	✓	✓						
Phase 2: Dataset and Labels Annotation - Divya									
Labels Generation			✓	✓	✓				
Video Slicing			✓	✓	✓				
Label Validation			✓	✓	✓				
Phase 3: Experiment Setup and Run - Joon Kiat									
QA Generation				✓	✓	✓			
Evaluation					✓	✓	✓	✓	
Phase 4: Final Report - Samuel Sim									
Compile Code								✓	✓
Compile Results								✓	✓
Report Writing							✓	✓	✓

Table 3: An overview of our execution plan broken down into phases and sub-tasks for each of the corresponding week (W). The lead for each phase of the project are stated next to the phase title. Although each task will have a lead, all members are expected to contribute meaningfully to each sub-task