APPENDIX CONTENTS

A THE USE OF LARGE LANGUAGE MODELS IN MMEVOKE

In this section, we elaborate on the precise role of large language models within MMEVOKE, as detailed below.

- **Usage 1: MMEVOKE's construction.** In Section 3.2, we specify that GPT-4o is employed for content summarization and QA generation, which aligns with current research practices.
- **Usage 2: MMEVOKE's evaluation.** In Section 4.2, we evaluate MMEVOKE using Gemini-2.0-Flash, Gemini-2.5-Pro, Perplexity AI, and GPT-4.1, following standard benchmarking practices.
- **Usage 3: General capability tests.** In Section 4.3, we employ MIA-Bench, MMDU, MathVista, and MathVision, whose evaluation requires large language models as judges—a practice consistent with current research standards.
- **Usage 4: Paper grammar polishing.** The paper is initially drafted by humans and subsequently polished for grammar using LMMs, a practice consistent with current research norms.

B MORE DETAILS ABOUT MMEVOKE

In this section, we further demonstrate the details of MMEVOKE, including benchmark presentation, complete subfields distribution, word cloud distribution, human study, fine-grained difficulty level results and release plan.

B.1 PRESENTATION OF MMEVOKE BENCHMARK

Figure 8 presents additional examples of MMEVOKE, encompassing four distinct subfields: Politics, Science, Video Game, and Songs. Each subfield showcases relevant Type, Knowledge Summary, Knowledge Image, Query, Query Image. Specifically, four examples are as follows:

Figure 8: **Examples of News/Entity Evolving Knowledge in MMEVOKE**, including Type, Knowledge Summary, Knowledge Image, Query, Query Image. Examples are taken from different clusters: **Politics** for News, **Science** for News, **Video Game** for Entity, and **Songs** for Entity.

- **Politics:** Describes the unsuccessful assassination attempt targeting former U.S. President Donald Trump at a campaign rally in Butler, Pennsylvania, on July 13, 2024. The query question asks for the identity of the individual depicted in the image.
- **Science:** Details the awarding of the 2024 Nobel Prize in Physics to John Hopfield and Geoffrey Hinton for their contributions. The query question inquires about the person who shared the Nobel Prize with the individual shown in the image.
- **Video Game:** Lists the video game Black Myth: Wukong, released on August 20, 2024. The query question focuses on the game's sales figures during its first month.

- **Songs:** Introduces the song Apt, performed by Russ and Bruno Mars. The query question concerns the drinking game that served as inspiration for the song.

These examples illustrate the diverse subfields of evolving knowledge captured within MMEVOKE, providing a more detailed demonstration.

B.2 WORD CLOUD DISTRIBUTION

(a) News Evolving Knowledge. (b) Entity Evolving Knowledge.

Figure 9: Word Cloud Distributions of MMEVOKE.

In Figure 9a, we show the word cloud distribution of News evolving knowledge. It can be found that Trump appears more often, which may be because MMEVOKE contains a large number of US political News data. Meanwhile, in Figure 9b, we present the word cloud distribution of entity names in the Entity evolving knowledge.

We have demonstrated the diversity of MMEVOKE benchmark through fine-grained subfields distribution, key statistics, word cloud distribution, and multiple perspectives. At the same time, our automated pipeline can continuously collect evolving knowledge and provide injection data for the knowledge injection field.

B.3 COMPLETE SUBFIELDS DISTRIBUTION

Figure 10: Fine-grained subfields distribution of News evolving knowledge.

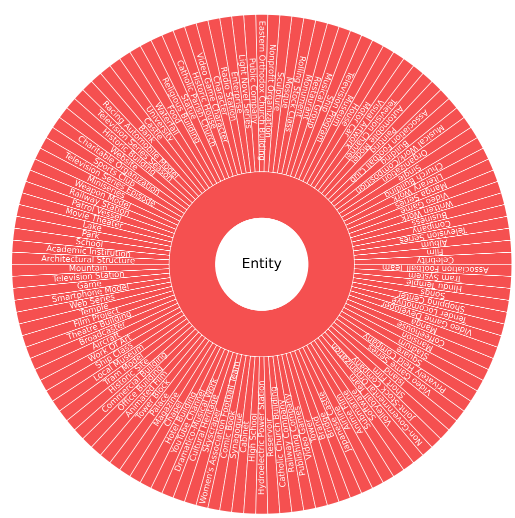

Figure 11: Fine-grained subfields distribution of Entity evolving knowledge.

In Figures 10 and 11, we comprehensively illustrate the fine-grained subfields distribution of the MMEVOKE benchmark, which includes 29 distinct subfields for News evolving knowledge and 130 subfields for Entity evolving knowledge, underscoring its exceptional diversity. This benchmark serves as a critical resource for the evolving knowledge injection domain, providing a robust foundation for advancing research and development in the field.

B.4 DENSITY DISTRIBUTION

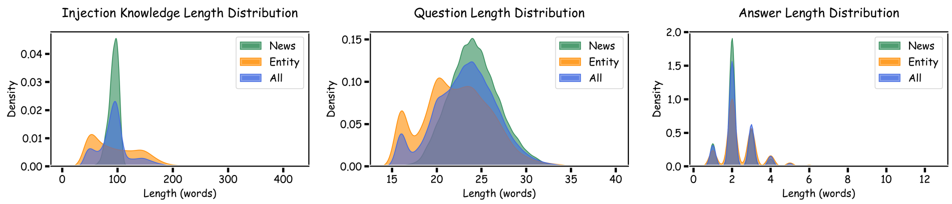

Figure 12: Density distribution based on evolving knowledge sources.

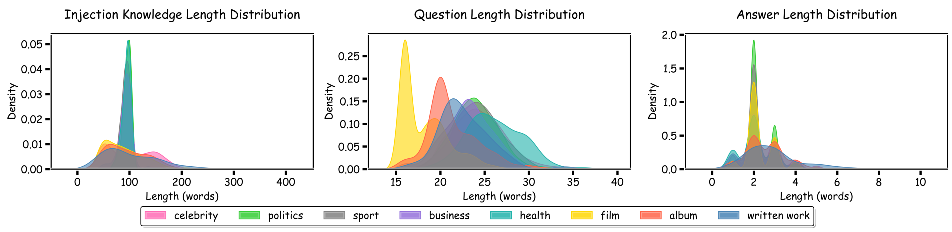

Figure 13: Density distribution of fine-grained subfields based on evolving knowledge.

B.5 HUMAN STUDY TOWARDS BENCHMARK QUALITY TEST

To verify the hallucination level of GPT-4o in data generation, We randomly selected 100 pieces of data from MMEVOKE during manual selection for human study. Specifically, four annotators scored the samples (1-5 scales, higher scores indicate greater purity) from the perspectives of content summarization, QA generation, and whether the summary contained information necessary to answer the question. According to the results in Table 4, MMEVOKE exhibits high quality, demonstrating minimal hallucination during the data construction process.

Table 4: Human Study Towards Benchmark Quality Test.

Dimension		ALL	News	Entity
MMEVOKE	Q&A	4.86 ($\pm$0.01)	4.87 ($\pm$0.01)	4.85($\pm$0.02)
	Summary	4.98 ($\pm$0.01)	4.97 ($\pm$0.01)	4.98 ($\pm$0.02)

B.6 FINE-GRAINED DIFFICULTY LEVEL OF MMEVOKE

Table 5: The performance of different difficulty levels on MMEVOKE.

Task	Method	ALL		News		Entity	
		CEM	**F1-Score**	**CEM**	**F1-Score**	**CEM**	**F1-Score**
SimpleVQA	Full-FT	16.55	14.82	17.43	14.12	15.53	15.61
	Sufficient Context	55.63	76.00	55.59	72.05	55.68	80.54
3-Hop	Full-FT	12.15	5.65	11.18	5.22	13.26	6.14
	Sufficient Context	40.49	52.58	38.16	51.49	43.18	53.82
Counterfactual Reasoning	Full-FT	70.42	70.42	74.01	74.01	66.29	66.29
	Sufficient Context	76.58	76.58	65.46	65.46	89.39	89.39

To further diversify MMEVOKE, we constructed 568 Counterfactual Reasoning and 3-Hop QA pairs using GPT-4o, and extracted their corresponding SimpleVQA data, yielding experimental results comparing fine-grained difficulty levels. ***The SimpleVQA here refers to the QA data of* MMEVOKE *itself.*** Table 5 shows the difficulty ranking: *Counterfactual Reasoning* < *SimpleVQA* < *3-Hop*, and 48.24% (avg) of cases have SimpleVQA failing while Counterfactual Reasoning succeeding, and 40.06% (avg) have SimpleVQA succeeding but 3-Hop failing.

C MORE RESULTS ABOUT MMEVOKE

C.1 MORE QUANTITATIVE EXPERIMENTAL RESULTS ABOUT RQ1

Table 6: **Performance of knowledge injection methods on MMEVOKE.** ALL, News.Avg, and Entity.Avg respectively show the performance of knowledge injection methods on entire MMEVOKE, News subset, and Entity subset. Orange value marks the best performance of methods on LLaVA-v1.5 and Qwen-VL-Chat, as well as the best performance of models in Web Search Engine and Sufficient Context **(vertical perspective)**. Red value indicates knowledge subfield with the best performance of the same method and model on different fine-grained subfields, while blue value indicates knowledge subfield with the worst performance **(horizontal perspective)**. PO: Politics; SP: Sports; BU: Business; HE: Health; CE: Celebrity; FI: Film; AL: Album; WR: Written Work.

Method	ALL		News										Entity									
			Avg		PO		SP		BU		HE		Avg		CE		FI		AL		WR	
	CEM ↑	F1 ↑	CEM ↑	F1 ↑	CEM ↑	F1 ↑	CEM ↑	F1 ↑	CEM ↑	F1 ↑	CEM ↑	F1 ↑	CEM ↑	F1 ↑	CEM ↑	F1 ↑	CEM ↑	F1 ↑	CEM ↑	F1 ↑	CEM ↑	F1 ↑
LLaVA-v1.5																						
Vanilla	4.89	9.34	7.37	11.96	1.92	5.86	4.59	9.74	10.70	15.99	10.12	17.54	2.18	6.47	1.37	6.48	2.39	5.71	3.77	6.02	6.78	11.24
Full-FT	18.02	15.17	21.35	16.34	12.92	10.99	22.49	20.88	27.31	20.95	19.84	16.47	14.37	13.88	13.11	16.93	12.39	13.16	12.17	7.66	20.34	8.43
LoRA	15.23	18.31	17.72	19.42	10.54	12.96	19.11	21.50	20.66	24.03	17.81	23.76	12.51	17.09	12.20	21.19	12.39	15.82	10.72	8.72	20.34	12.94
MM-RAG$^{\text{Text-Only}}$	24.05	34.32	37.32	49.39	22.18	36.25	47.88	54.77	34.87	51.07	36.44	50.95	9.50	17.80	15.14	25.39	1.93	4.04	2.90	13.86	3.39	13.07
MM-RAG$^{\text{Image-Only}}$	25.25	37.11	19.28	26.76	9.35	16.96	33.37	39.19	19.56	29.46	18.22	28.60	31.80	48.45	26.37	43.01	39.09	47.58	40.29	58.14	28.81	53.68
MM-RAG$^{\text{UniIR}}$	40.68	57.51	40.12	53.21	21.81	35.08	56.23	65.94	39.85	57.08	35.22	50.93	41.30	62.23	41.01	63.94	48.86	58.98	41.45	63.02	35.59	60.09
Qwen-VL-Chat																						
Vanilla	5.84	10.99	7.75	12.72	3.21	7.69	4.47	10.37	10.52	14.92	10.93	19.32	3.74	9.10	1.78	8.06	8.18	13.10	4.35	6.93	8.47	16.81
Full-FT	10.16	16.61	13.35	18.22	6.42	11.80	12.70	17.11	16.42	22.27	17.00	25.42	6.65	14.83	5.39	14.68	11.59	17.95	5.22	10.83	15.25	21.69
LoRA	6.95	12.64	9.27	14.55	4.31	9.24	5.68	11.82	12.55	17.79	12.96	21.64	4.41	10.54	2.34	9.54	9.32	14.96	5.22	8.04	10.17	18.07
MM-RAG$^{\text{Text-Only}}$	21.79	31.28	31.51	41.14	20.71	29.81	30.71	40.75	32.29	43.38	33.20	47.56	11.13	20.47	13.36	24.27	8.41	14.02	6.67	15.27	11.86	19.60
MM-RAG$^{\text{Image-Only}}$	22.31	33.09	17.82	25.15	9.26	15.97	20.80	29.82	18.45	28.33	18.62	29.38	27.24	41.79	20.27	33.52	33.98	45.81	39.42	53.80	33.90	54.43
MM-RAG$^{\text{UniIR}}$	32.75	46.18	33.26	43.36	18.15	27.56	32.77	44.90	37.08	49.25	31.98	44.96	32.20	49.28	28.20	45.05	37.16	50.60	41.45	56.57	42.37	65.29
Commercial AI Web Search Engines																						
Gemini-2.0-Flash	18.21	26.52	21.23	27.75	10.91	16.87	21.64	27.45	22.88	30.03	17.41	28.32	14.91	25.16	10.11	20.35	28.64	37.47	14.49	23.87	16.95	28.77
Gemini-2.5-Pro	44.19	52.58	48.86	52.84	39.07	52.28	31.90	37.00	51.11	57.22	58.04	59.97	39.27	46.27	24.29	35.81	63.98	73.14	53.62	68.36	42.37	57.40
Perplexity AI	48.27	62.44	47.58	56.51	34.78	43.14	56.13	66.19	41.82	54.33	35.29	47.88	48.96	68.78	47.03	70.95	62.22	73.65	54.41	68.54	43.75	59.17
GPT-4.1	39.61	42.69	41.81	43.08	25.23	26.07	52.60	52.43	34.82	42.45	47.60	50.81	37.19	42.26	24.29	26.53	57.50	62.41	58.26	62.94	30.51	47.61
Sufficient Context																						
LLaVA-v1.5	56.13	75.77	56.78	72.37	38.77	58.44	75.09	84.69	54.61	74.33	48.58	67.01	55.43	79.50	52.08	78.83	75.91	89.71	57.39	78.80	49.15	69.96
Qwen-VL-Chat	48.96	66.02	49.98	63.42	35.20	50.29	52.00	68.90	50.55	67.25	48.18	62.02	47.84	68.87	43.29	66.15	62.05	75.92	58.55	75.41	47.46	67.79
Gemini-2.5-Pro	72.15	80.46	72.61	78.77	57.01	65.75	86.34	89.63	71.77	81.65	62.35	74.65	71.65	82.32	73.53	80.89	81.14	88.09	75.07	85.59	52.54	72.05
GPT-4.1	75.02	83.74	79.22	88.20	53.62	65.21	84.04	90.23	69.37	80.75	68.83	79.56	71.21	79.68	80.74	88.02	88.18	91.97	86.38	91.58	59.32	74.86

Table 6 presents the quantitative experimental results of RQ1, revealing that no method achieves robust injection performance, with significant performance variance observed across different fine-grained subfields knowledge. Specifically, We have obtained further observations:

- **Obs 1:** In Table 6, across nearly all evaluated methods, News knowledge injection performance consistently outperforms Entity knowledge. We attribute this gap to their fundamental differences in learning difficulty. Entity knowledge introduces entirely novel concepts to model, posing a substantial learning challenge. In contrast, News knowledge primarily establishes new and complex relationships among existing entities, which represents a comparatively lower learning barrier.
- **Obs 2:** The performance of knowledge in the same subfield varies depending on the method used. For example, in Full FT, LoRA, and MM-RAG$^{\text{Text-Only}}$, the performance of film knowledge is poor. In sharp contrast, it performs better when using MM-RAG$^{\text{Image-Only}}$, MM-RAG$^{\text{UniIR}}$, Sufficient Context, and Web Search.
- **Obs 3:** A significant performance variance among different strategies within same method. Notably, MM-RAG$^{\text{Text-Only}}$ is more effective for injecting News knowledge, while MM-RAG$^{\text{Image-Only}}$ is better suited for Entity knowledge. This discrepancy indicates that knowledge injection is optimized when the modality of the feature aligns with the nature of the knowledge source (textual features for News and visual features for Entity).
- **Obs 4:** The performance of the same subfield knowledge differs across models. For instance, Health and Written work perform better on Qwen-VL-Chat; Sport and Business perform better on LLaVA-v1.5. This is likely due to significant distributional differences in types of knowledge data encountered during pre-training of different models.
- **Obs 5:** Politics knowledge contains a wide range of professional terms and complex concepts that are difficult to learn, ranking lowest among almost all methods.

Observations

Observation 1: Current knowledge injection methods have significant domain specificity for different fine-grained subfield knowledge.

Table 7: **The performance of knowledge injection methods on Entity subset of MMEVOKE.** TEL: Television Series; COM: Company; VID: Video Game; CHU: Church Building; SIN: Single; OGR: Organization; PAI: Painting; MOT: Motor Car.

Method	TEL		COM		VID		CHU		SIN		ORG		PAI		MOT	
	CEM ↑	F1 ↑	CEM ↑	F1 ↑	CEM ↑	F1 ↑	CEM ↑	F1 ↑	CEM ↑	F1 ↑	CEM ↑	F1 ↑	CEM ↑	F1 ↑	CEM ↑	F1 ↑
LLaVA-v1.5																
Vanilla	6.15	9.77	1.12	5.69	0.00	3.16	0.00	6.39	4.55	9.51	2.70	6.31	0.00	11.90	0.00	4.76
Full-FT	13.97	10.29	29.21	14.15	10.34	7.32	26.53	22.67	15.91	8.55	27.03	15.52	17.86	13.83	7.14	6.21
LoRA	15.64	16.20	10.11	11.42	12.07	15.24	14.29	24.54	20.45	20.39	16.22	17.45	14.29	14.42	0.00	1.41
MM-RAG$^{\text{Text-Only}}$	3.35	6.15	4.49	14.31	5.17	21.81	8.16	18.10	2.27	20.72	2.70	13.69	14.29	21.31	7.14	27.55
MM-RAG$^{\text{Image-Only}}$	36.87	54.26	30.34	57.23	29.31	59.73	40.82	66.33	34.09	56.78	24.32	49.88	53.57	70.95	21.43	57.93
MM-RAG$^{\text{UniIR}}$	41.34	62.91	30.34	63.49	32.76	65.77	34.69	64.30	31.82	61.50	29.73	59.19	64.29	85.12	21.43	68.30
Qwen-VL-Chat																
Vanilla	7.82	11.33	1.12	7.32	1.72	2.59	0.00	10.20	6.82	11.33	0.00	2.88	7.14	13.10	0.00	10.37
Full-FT	8.94	16.49	1.12	11.05	3.45	15.54	2.04	16.91	6.82	15.75	5.41	8.61	10.71	12.93	7.14	15.48
LoRA	7.26	11.55	1.12	8.64	1.72	3.85	2.04	9.90	6.82	13.61	2.70	5.59	10.71	15.95	0.00	8.33
MM-RAG$^{\text{Text-Only}}$	7.26	13.22	7.87	23.37	8.62	25.35	4.08	12.90	13.64	31.20	13.51	19.91	14.29	23.45	14.29	30.36
MM-RAG$^{\text{Image-Only}}$	22.91	38.39	30.34	55.94	18.97	56.23	38.78	52.91	31.82	56.92	29.73	45.95	39.29	48.45	14.29	46.90
MM-RAG$^{\text{UniIR}}$	19.67	23.81	30.34	63.84	18.97	59.04	28.57	50.26	34.09	59.51	43.24	63.13	42.86	52.62	14.29	46.90
Commercial AI Web Search Engines																
Gemini-2.0-Flash	19.55	31.14	8.99	20.82	10.34	25.01	10.20	21.56	9.09	22.58	18.92	25.02	14.29	16.43	0.00	26.11
Gemini-2.5-Pro	58.10	74.71	41.57	66.09	46.55	65.25	20.41	33.07	43.18	66.37	43.24	59.98	46.43	38.27	7.14	35.48
Perplexity AI	43.90	54.59	30.00	52.08	33.33	48.41	62.50	75.83	50.00	70.00	33.33	54.07	85.71	83.67	33.33	13.33
GPT-4.1	50.28	62.08	52.81	57.02	53.45	65.23	22.45	29.31	38.64	47.03	45.95	52.43	17.86	20.53	0.00	15.99
Sufficient Context																
LLaVA-v1.5	56.42	81.18	41.57	78.05	34.48	68.72	44.90	72.48	45.45	68.79	45.95	79.70	75.00	90.12	35.71	73.15
Qwen-VL-Chat	51.96	72.08	39.33	73.62	25.86	63.28	34.69	62.88	36.36	62.62	43.24	65.69	42.86	55.60	42.86	73.47
Gemini-2.5-Pro	69.27	85.95	64.04	81.32	58.62	78.70	55.10	75.18	68.18	82.72	56.76	78.37	89.29	85.62	50.00	78.25
GPT-4.1	77.09	90.22	70.79	86.21	67.24	83.84	59.18	77.77	79.55	91.44	64.86	83.24	89.29	91.90	64.29	84.97

Table 8: **The performance of knowledge injection methods on News subset of MMEVOKE.** ENT: Entertainment; TEC: Tech; SCI: Science; TRA: Travel; FOO: Food; CLI: Climate; INV: Investing; STY: Style.

Method	ENT		TEC		SCI		TRA		FOO		CLI		INV		STY	
	CEM ↑	F1 ↑	CEM ↑	F1 ↑	CEM ↑	F1 ↑	CEM ↑	F1 ↑	CEM ↑	F1 ↑	CEM ↑	F1 ↑	CEM ↑	F1 ↑	CEM ↑	F1 ↑
LLaVA-v1.5																
Vanilla	6.79	9.35	6.79	9.35	6.79	9.35	11.90	18.57	10.26	17.83	8.11	13.87	18.28	23.71	13.93	16.20
Full-FT	18.67	11.47	28.29	17.02	15.79	12.56	28.57	24.16	35.90	24.54	27.03	13.02	44.09	25.06	31.15	19.17
LoRA	16.98	15.70	27.63	25.96	8.77	18.73	23.81	29.91	20.51	18.83	16.22	18.02	34.41	28.13	19.67	19.45
MM-RAG$^{\text{Text-Only}}$	39.81	48.79	46.05	55.21	36.84	55.71	38.10	54.50	33.33	50.85	37.84	53.51	37.63	47.06	68.85	78.51
MM-RAG$^{\text{Image-Only}}$	21.76	28.07	23.03	28.02	22.81	38.42	21.43	30.09	23.08	36.32	18.92	26.04	25.81	31.61	22.13	25.67
MM-RAG$^{\text{UniIR}}$	52.16	63.67	42.11	51.77	33.33	52.89	47.62	62.83	41.03	57.78	35.14	53.06	38.71	48.23	59.84	67.32
Qwen-VL-Chat																
Vanilla	6.79	9.90	14.47	16.10	8.77	14.95	9.52	16.59	10.26	16.24	10.81	12.07	23.66	29.27	13.11	16.19
Full-FT	11.27	14.64	17.11	18.79	8.77	13.78	14.29	23.89	17.95	27.35	18.92	21.42	35.48	38.34	16.39	19.18
LoRA	7.41	11.01	16.45	18.76	8.77	13.93	7.14	15.00	7.69	17.52	13.51	14.77	24.73	30.44	15.57	17.72
MM-RAG$^{\text{Text-Only}}$	31.48	38.00	46.71	51.27	42.11	48.99	38.10	50.56	20.51	39.66	35.14	46.65	43.01	52.75	60.66	66.14
MM-RAG$^{\text{Image-Only}}$	20.06	24.82	22.37	27.06	33.33	42.59	21.43	31.67	20.51	27.35	24.32	31.40	30.11	36.37	19.67	23.81
MM-RAG$^{\text{UniIR}}$	42.75	50.25	41.45	45.18	47.37	55.69	40.48	50.46	28.21	44.36	32.43	44.34	43.01	52.93	51.64	56.70
Commercial AI Web Search Engines																
Gemini-2.0-Flash	24.69	29.98	38.82	46.00	15.79	22.97	16.67	30.40	23.08	30.52	10.81	19.28	38.71	45.72	30.33	32.60
Gemini-2.5-Pro	59.72	61.28	63.82	60.26	31.58	37.64	52.38	63.00	48.72	56.44	48.65	44.35	52.69	51.29	69.67	68.13
Perplexity AI	59.85	64.15	47.06	55.20	45.45	49.13	50.00	70.05	33.33	40.74	37.50	64.58	33.33	40.12	71.88	74.36
GPT-4.1	46.30	43.64	57.24	59.50	22.81	35.29	50.00	50.29	66.67	56.89	40.54	35.21	55.91	55.73	50.82	50.84
Sufficient Context																
LLaVA-v1.5	65.12	78.31	63.82	77.61	47.37	66.30	57.14	72.37	51.28	76.58	51.35	63.07	60.22	72.83	75.41	85.18
Qwen-VL-Chat	61.42	68.99	62.50	72.69	43.86	63.14	45.24	58.56	51.28	64.66	48.65	56.68	53.76	65.04	68.03	75.70
Gemini-2.5-Pro	81.17	83.08	75.00	82.33	61.40	66.34	73.81	82.47	66.67	81.28	70.27	74.10	75.27	77.29	82.79	83.34
GPT-4.1	78.70	83.73	82.89	85.12	61.40	72.69	69.05	80.41	69.23	78.69	62.16	67.85	68.82	77.61	89.34	91.33

Tables 7 and 8 present richer experimental results of fine-grained subfields, further verifying the significant domain specificity of existing knowledge injection methods and their inability to robustly implement knowledge injection.

C.2 SEQUENTIAL FINE-TUNING

C.2.1 SEQUENTIAL FINE-TUNING BASED ON TASKS

Sequential Fine-Tuning refers to the process of incrementally training models on new tasks and data. Specifically, model weights obtained from previous tasks and data are used to initialize model parameters (Chen et al., 2025). In this section, ***we explore whether Sequential Fine-Tuning is more***

effective than One-Time Injection? We employed MMEVOKE for knowledge injection, randomly dividing the data into subsets of 4, 8, and 12 tasks. We consider each subset as a task and use these subsets to Sequential Fine-Tuning the model.

Sequential Fine-Tuning impede the effective injection of multimodal evolving knowledge. As illustrated in Figure 14, the performance of LMMs exhibits a declining trend with progressive Sequential Fine-Tuning based on tasks. This degradation primarily stems from the disruption of previously fine-tuning parameters during each subsequent fine-tuning iteration. Consequently, the overall performance of LMMs progressively deteriorates. Furthermore, our investigation into the impact of Sequential Fine-Tuning steps revealed a negative correlation between the number of steps g and LMMs performance, as evidenced by the values corresponding to the terminal points in each line graph. These findings underscore the importance of minimizing Sequential Fine-Tuning in practical applications to preserve model efficacy.

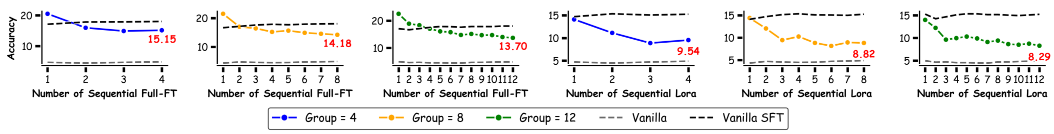

Figure 14: **The results of LLaVA-v1.5 on Sequential Fine-Tuning based on Tasks.** The data $\mathcal{D}_{\mathcal{K}}$ and $\mathcal{D}_{\mathcal{Q}}$ are evenly divided into $g \in \{4, 8, 12\}$ parts, namely $\mathcal{D}_{\mathcal{K}} = \left\{d_k^1, d_k^2, \ldots, d_k^n\right\}_{n=1}^{g}$ and $\mathcal{D}_{\mathcal{Q}} = \left\{d_q^1, d_q^2, \ldots, d_q^n\right\}_{n=1}^{g}$. Sequential Fine-Tuning based on tasks refer to the situation where if the current m-th Sequential Fine-Tuning has ended, it indicates that the model is being trained on $d_k^1, d_k^2, \ldots, d_k^m$ in sequence; and evaluated on $\left\{d_q^1 \cup d_q^2 \cup \cdots \cup d_q^m\right\}$.

C.2.2 SEQUENTIAL FINE-TUNING BASED ON SUBSETS

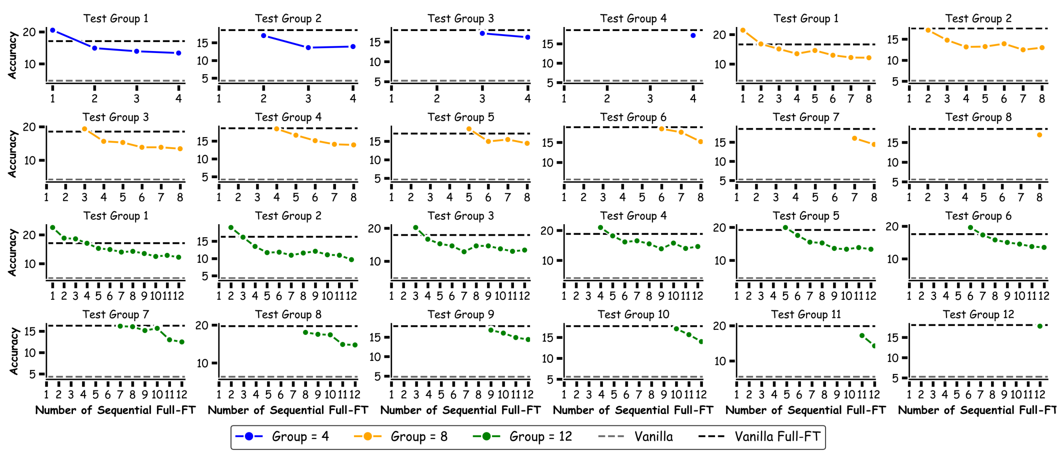

Figure 15: **The results of LLaVA-v1.5 on Sequential Full-FT based on Subsets.** Sequential Full-FT based on subset refer to the situation where if the current m-th Sequential Full-FT has ended, it indicates that the model is being trained on $d_k^1, d_k^2, \ldots, d_k^m$ in sequence; and evaluate sequentially on **one of** $d_q^1, d_q^2, \ldots, d_q^m$.

The results of Sequential Fine-Tuning based on subsets are shown in Figure 15 and 16. Each subgraph displays the performance changes of the LMMs on the same subset as the Sequential Fine-Tuning process progresses. It can be observed that whether using Full-FT or LoRA as training strategies, as the number g of Sequential Fine-Tuning increases, the performance of the model on the same subset shows a downward trend. This discovery further indicates that Sequential Fine-Tuning is not conducive to injecting up-to-date knowledge into the LMMs.

Observations

Observation 2: Both sequential task and subset fine-tuning impede the efficacy of knowledge injection, with performance degradation correlating with an increased number of tasks or subsets.

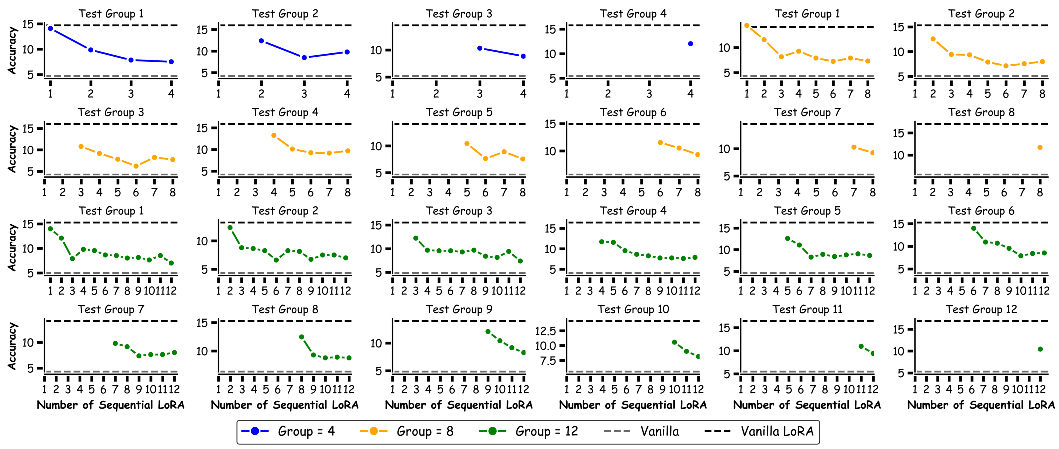

Figure 16: **The results of LLaVA-v1.5 on Sequential LoRA based on Subsets.** Sequential LoRA based on subset refer to the situation where if the current m-th Sequential LoRA has ended, it indicates that the model is being trained on $d_k^1, d_k^2, \ldots, d_k^m$ in sequence; and evaluate sequentially on **one of** $d_q^1, d_q^2, \ldots, d_q^m$.

C.3 Ablation experiments in MM-RAG

Retrieval strategy, ***Example Number***, and ***Pool Size*** are critical factors influencing the performance of MM-RAG, as demonstrated by the experimental results presented in Figure 17 and 18.

- ***Effect of Retrieval Strategy in MM-RAG.*** An interesting observation appears in the "News" subgraph, where the Text-Only approach significantly outperforms the Image-Only strategy. The reason for this difference is that textual information is more important for news understanding than visual information, as valuable data cannot be retrieved solely through images. On the contrary, for Entity knowledge, visual information is more valuable than textual information.
- ***Effect of Example Number in MM-RAG.*** We compared $K \in \{1, \ldots, 5\}$, and in the first row of Figure 17, the direct correlation between the performance of model and Example Number is shown. Our experiment revealed a convincing trend that the model performs using a monotonically increasing function of Example Number K for three retrieval strategies. This observation indicates that an increase in the example number brings more diverse reference information, which has a positive effect on the model's understanding and utilization of evolving knowledge.
- ***Effect of Retrieval Pool Size in MM-RAG.*** Regarding the ablation experiment of pool size, our setup is to randomly select 20% of the corresponding data from $\mathcal{D}_\mathcal{Q}$ and $\mathcal{D}_\mathcal{K}$ as ${\mathcal{D}_\mathcal{Q}}^{20\%}$ and ${\mathcal{D}_\mathcal{K}}^{20\%}$; For instance, when Pool Size = 20%, Retrieve Pool = ${\mathcal{D}_\mathcal{Q}}^{20\%}$; When Pool Size = 60%, Retrieve Pool = ${\mathcal{D}_\mathcal{K}}^{20\%} + \mathcal{D}_\mathcal{J}$,where $\mathcal{D}_\mathcal{J}$ is a randomly selected 40% data from the $\mathcal{D}_\mathcal{K} \setminus {D_K}^{20\%}$. The evaluation data is always ${\mathcal{D}_\mathcal{Q}}^{20\%}$. The experimental results, presented in the second row of Figure 18, demonstrate an inverse correlation between MM-RAG's performance and Pool Size. This suggests that larger pool sizes hinder the retriever's ability to identify relevant information, a critical consideration for practical MM-RAG applications.

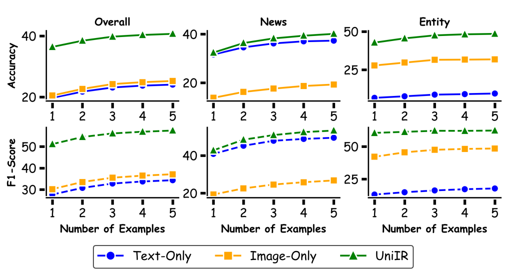

Figure 17: The results of LLaVA-v1.5's ablation study on MM-RAG about **Retrieval Strategy and Example Number** analysis.

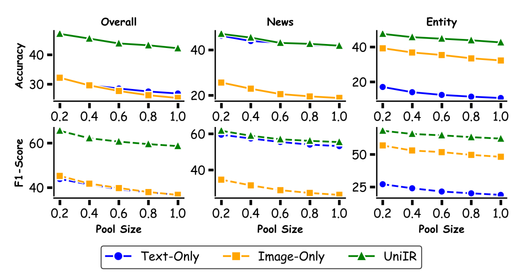

Figure 18: The results of LLaVA-v1.5's ablation study on MM-RAG about **Retrieval Strategy and Pool Size** analysis.

Observations

Observation 3: Cross-modal retrieval strategies, a larger number of examples, and a smaller retrieval pool size all contribute to strengthening knowledge injection performance.

C.4 More qualitative results about MMEvoke

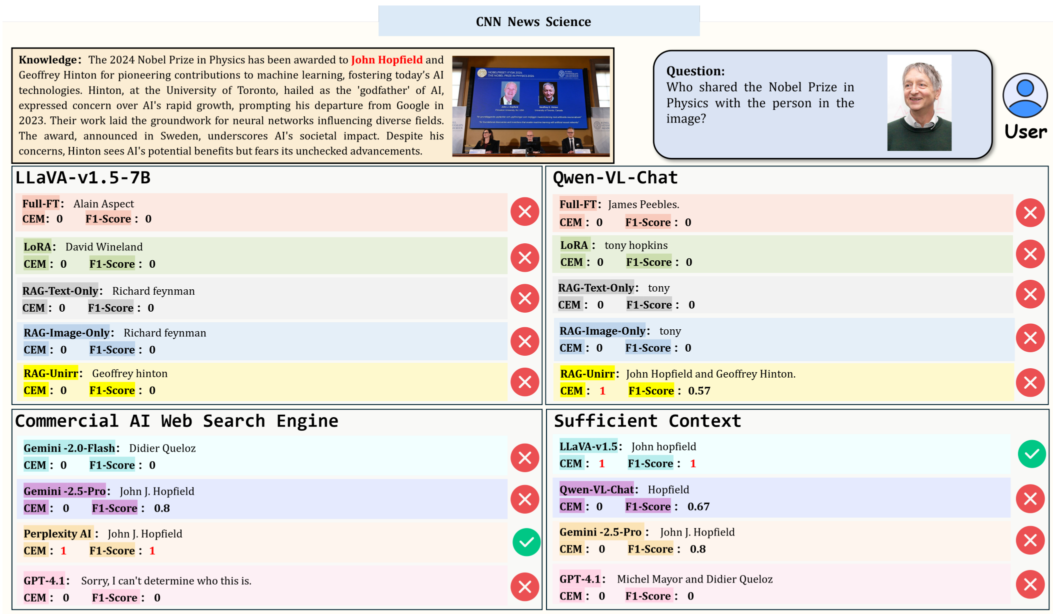

Figure 19: Qualitative example of CNN News science knowledge.

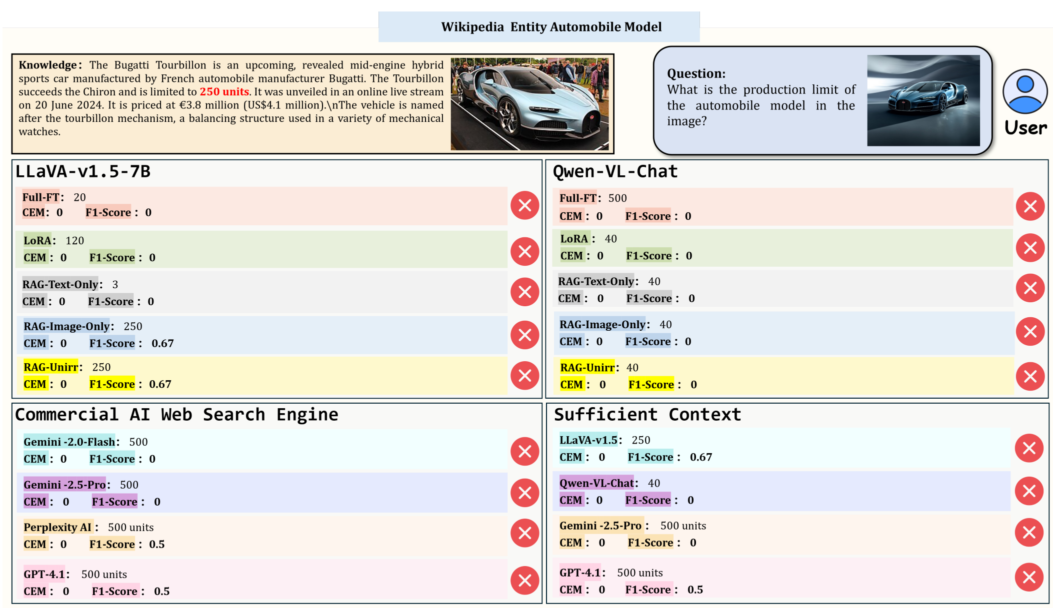

Figure 20: Qualitative example of Wikipedia Entity automobile model knowledge.

C.5 Error Analysis

Observing the qualitative examples in Figures 19, 20, and 21, we find that, as demonstrated by the results in Table 6, existing knowledge injection methods perform poorly on MMEvoke, with even sufficient context failing to achieve perfect performance. Here, we conduct a detailed analysis of sufficient context.

Even when provided with sufficient context, the model still generates hallucinations. For instance, in Figure 19, the response given by GPT-4.1 is entirely unrelated to the question and does not appear in the sufficient context, representing a severe hallucination phenomenon. A similar hallucination issue persists in Figure 20. These concrete results indicate that merely improving the sufficiency of context is far from adequate—the model's inherent reasoning and ability to utilize contextual information are equally critical. Hallucination remains an urgent problem to be addressed.

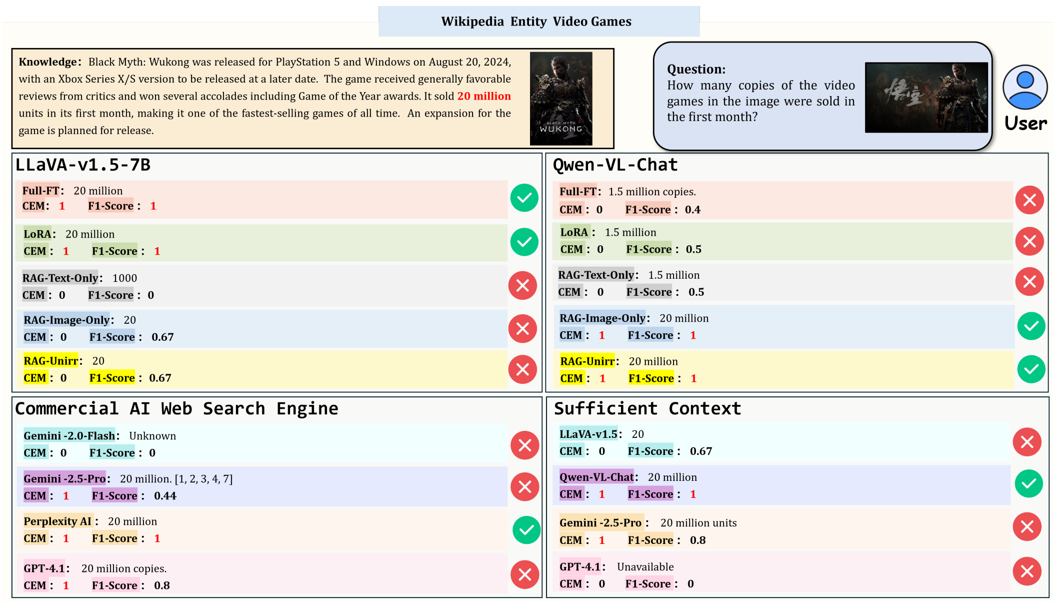

Figure 21: Qualitative example of Wikipedia Entity video games knowledge.

Observations

Observation 4: Despite being provided with sufficient context, the model still exhibits severe hallucinations.

D MORE DETAILS ON CAPABILITY DEGRADATION

D.1 CAPABILITY DEGRADATION RANKING

Table 9: **The degree of general capability degradation results.** The displayed values are obtained by calculating the mean based on the results in Table 3.

Method	Comprehensive		OCR		Multidisciplinary		Instruction		Multi-Round		Mathematical		Hallucination	
	Loss ↓	Rank ↓	Loss ↓	Rank ↓	Loss ↓	Rank ↓	Loss ↓	Rank ↓	Loss ↓	Rank ↓	Loss ↓	Rank ↓	Loss ↓	Rank ↓
Full-FT	↓33.40%	4	↓13.85%	3	↓9.63%	2	↓61.93%	7	↓50.59%	6	↓6.20%	**1**	↓35.98%	5
LoRA	↓25.24%	4	↓19.32%	3	↓15.20%	2	↓55.28%	7	↓48.05%	6	↓5.76%	**1**	↓37.25%	5
Knowledge Augmentation for Text														
Knowledge Agnostic	↓16.60%	3	↓15.51%	2	↓11.87%	**1**	↓65.48%	7	↓59.76%	6	↓25.16%	4	↓34.21%	5
Knowledge Aware (+3)	↓14.62%	3	↓5.36%	2	↓3.78%	**1**	↓64.36%	7	↓60.03%	6	↓17.48%	4	↓20.89%	5
Knowledge Augmentation for Images														
Knowledge Agnostic	↓16.95%	**1**	↓19.58%	3	↓17.44%	2	↓67.41%	7	↓59.46%	6	↓22.60%	4	↓38.07%	5
Knowledge Aware (+3)	↓24.58%	4	↓12.75%	2	↓4.88%	**1**	↓72.85%	7	↓59.73%	6	↓28.91%	5	↓24.06%	3
Knowledge Retention Methods														
$\text{Replay}^{\text{Full-FT}}_{+10\%}$	↓10.02%	4	↓3.69%	3	↑0.09%	**1**	↓22.81%	6	↓31.40%	7	↓1.06%	2	↓13.09%	5
$\text{Replay}^{\text{LoRA}}_{+10\%}$	↓8.95%	5	↓4.14%	3	↓0.93%	2	↓6.03%	4	↓26.77%	7	↓0.70%	**1**	↓9.69%	6
EWC	↓24.65%	4	↓14.96%	3	↓8.89%	2	↓55.09%	7	↓49.34%	6	↓5.83%	**1**	↓31.38%	5
LwF	↓18.94%	4	↓17.16%	3	↓16.58%	2	↓45.44%	6	↓48.12%	7	↓6.41%	**1**	↓33.42%	5
MoELoRA	↓4.56%	4	↓18.34%	6	↓0.97%	**1**	↓2.05%	3	↓29.24%	7	↓1.16%	2	↓9.18%	5

Based on Table 3, we calculate the mean degradation levels for each capability dimension. Table 9 reveals that both Full-FT and LoRA exhibit a consistent ranking of capability degradation: Instruction Following → Multi-Round QA → Hallucination → Comprehensive Evaluation → OCR → Multidisciplinary → Mathematical Reasoning. The identical ranking is also maintained in knowledge retention. Only $\text{Replay}^{\textbf{LoRA}}_{\textbf{+10\%}}$ and MoELoRA show significantly alleviated degradation rankings in instruction-following, rising to 4th and 3rd place respectively.

D.2 FINE-GRAINED DIMENSIONAL RESULTS ON GENERAL CAPABILITY TESTS

To effectively evaluate the specific capability degradation caused by knowledge injection in LMMs, we utilized 12 benchmarks across 7 task categories:

1. **MME** (Fu et al., 2023) is a comprehensive evaluation benchmark designed to assess the performance of LMMs across 14 distinct tasks, encompassing both perception and cognition abilities. To ensure fair and accurate comparisons, MME provides concise, manually designed instruction-answer pairs, eliminating the need for extensive prompt engineering.
2. **MMBench** (Liu et al., 2024b) is a bilingual benchmark designed to evaluate the comprehensive capabilities of LMMs across multiple modalities. It offers a meticulously curated dataset with over 3,000 multiple-choice questions covering 20 distinct ability dimensions, such as object localization and social reasoning. Additionally, MMBench provides questions in both English and Chinese, enabling comparative evaluations of LMM performance across these languages.
3. **SEEDBench2_Plus** (Li et al., 2024a) comprehensively evaluates LMMs' understanding of text-rich visuals (charts, maps, web pages). Comprising 2,300 multiple-choice questions across these categories, it assesses reasoning capabilities in real-world scenarios where text and visuals intertwine—addressing gap for applications like document analysis and web content understanding.
4. **OCRBench** (Liu et al., 2023b) is a comprehensive evaluation benchmark designed to assess the OCR)capabilities of LMMs. It encompasses 29 datasets across five key tasks: Text Recognition, Scene Text-Centric VQA, Document-Oriented VQA, Key Information Extraction (KIE), and Handwritten Mathematical Expression Recognition (HMER). The benchmark aims to provide a thorough assessment of LMMs' performance in various text-related visual tasks, highlighting their strengths and weaknesses, particularly in handling multilingual text, handwritten text, non-semantic text, and mathematical expressions.
5. **MMMU** (Yue et al., 2024) is a comprehensive benchmark designed to evaluate LMMs on tasks that require college-level subject knowledge and deliberate reasoning. It comprises 11,500 meticulously curated multimodal questions sourced from college exams, quizzes, and textbooks, spanning six core disciplines: Art & Design, Business, Science, Health & Medicine, Humanities & Social Science, and Technology & Engineering. These questions cover 30 subjects and 183 subfields, featuring 30 diverse image types such as charts, music sheets, and chemical structures.

6. **MIA-Bench** (Qian et al., 2024) is a benchmark designed to evaluate the ability of LMMs to adhere strictly to complex instructions. It comprises a diverse set of 400 image-prompt pairs, each crafted to challenge models' compliance with layered instructions, requiring accurate and contextually.
7. **MMDU** (Liu et al., 2025) is a comprehensive evaluation framework designed to assess the capabilities of LMMs in handling multi-turn, multi-image dialog scenarios. It focuses on understanding complex interactions involving multiple images and sequential dialog turns, which are critical for real-world applications like visual storytelling, medical diagnosis, and interactive AI systems. The benchmark includes a diverse dataset with rich annotations, enabling models to be fine-tuned and evaluated on tasks requiring contextual reasoning, image-text alignment, and temporal coherence.
8. **MathVista** (Lu et al., 2024) evaluates foundation models' mathematical reasoning in visual contexts. It comprises 6,141 examples from 28 existing multimodal datasets, augmented with three new datasets (IQTest, FunctionQA, PaperQA), requiring fine-grained visual understanding and compositional reasoning.
9. **MathVision** (Wang et al., 2025) is a meticulously curated dataset comprising 3,040 high-quality mathematical problems, each embedded within a visual context and sourced from real mathematics competitions. This benchmark spans 16 distinct mathematical disciplines and is organized across five levels of difficulty, offering a comprehensive platform to evaluate the mathematical reasoning abilities of LMMs.
10. **HallusionBench** (Guan et al., 2024) is a comprehensive benchmark designed to evaluate LMMs on their ability to accurately interpret and reason about visual data, specifically addressing issues of language hallucination and visual illusion. It comprises 346 images paired with 1,129 questions among visual dependent and visual supplement. The benchmark introduces a novel structure for visual questions, enabling quantitative analysis of models' response tendencies, logical consistency, and various failure modes.
11. **POPE** (Li et al., 2023b) is a benchmark designed to systematically assess object hallucination in LMMs. Object hallucination refers to the tendency of these models to generate descriptions containing objects not present in the corresponding images. POPE addresses this issue by implementing a polling-based query method that evaluates models' accuracy in identifying the existence of specific objects within images. This approach provides a more stable and flexible evaluation of object hallucination, revealing that current LMMs often generate objects inconsistent with the target images.

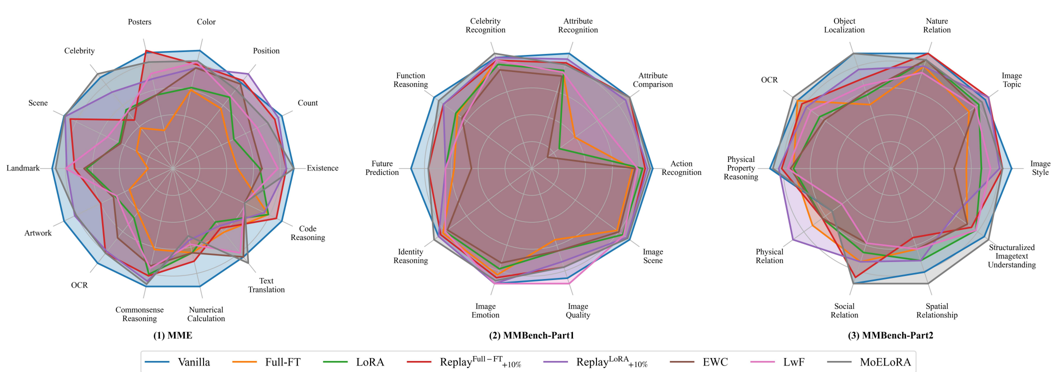

Figure 22: Fine-grained dimensional results on MME and MMBench.

According to Figures 22, 23, 24, 25, and 26, we conduct result analysis for each benchmark.

1. **MME:** Results on the MME benchmark indicate that both Full-FT and LoRA significantly degrade LLaVA's perception and cognition capabilities, with perception exhibiting a more pronounced decline. We attribute this primarily to MMEVOKE's focus on cognition tasks and its lack of substantial perception content. While the replay method effectively mitigates forgetting in perception abilities (e.g., outperforming Vanilla in Position tasks), it shows limited efficacy for cognition (e.g., poor performance in ***Numerical Calculation*** and ***Text Translation***). This disparity likely stems from LLaVA's original training data heavily emphasizing perception. Overall, EWC and LwF are less effective at mitigating forgetting than MoELoRA, though all three methods perform relatively well on the ***Text Translation*** task.
2. **MMBench:** Experimental results show that both Full-FT and LoRA significantly degrade LLaVA's performance in the perceptually demanding Attribute Comparison task, while enabling superior

performance in the Physical Relationship task due to MMEVOKE's relational data. For capability degradation mitigation, Replay and MoELoRA remain most effective. Notably, the EWC method underperforms even Full-FT and LoRA across 16 tasks (including ***Attribute Comparison***, ***Attribute Recognition***, ***Celebrity Recognition***, and ***Function Reasoning***), directly indicating the instability of this parameter-regularization approach.

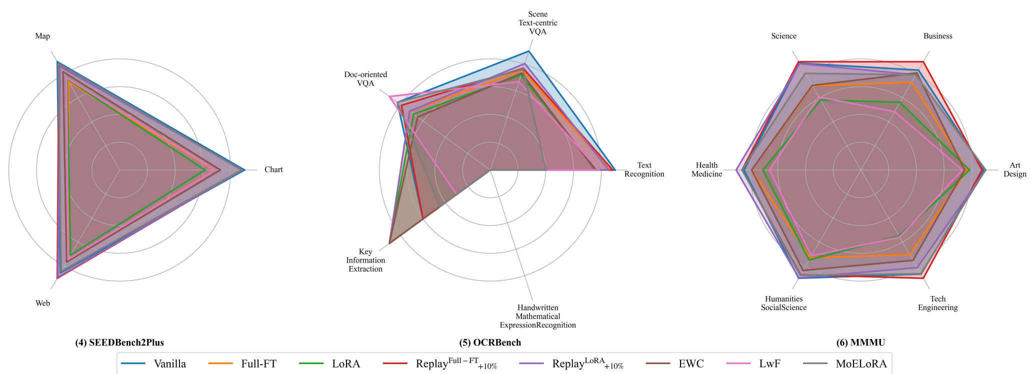

Figure 23: Fine-grained dimensional results on SEEDBench2_Plus, OCRBench and MMMU.

3. **SEEDBench2_Plus:** Both Full-FT and LoRA reduce LLaVA's performance on SEEDBench2_Plus, with LoRA underperforming compared to Full-FT. Among knowledge retention methods, only Replay outperforms the Vanilla approach in ***Web*** tasks.
4. **OCRBench:** Experimental result shows Full-FT and LoRA exhibit relatively less degradation in OCR tasks, potentially due to their text-information focus, while outperforming Vanilla in Key Information Extraction. However, LwF and MoELoRA demonstrate unstable degradation mitigation—underperforming Full-FT/LoRA in ***Text Recognition*** and ***Scene Text Centric VQA***, yet showing opposite trends to all other methods (Full-FT, LoRA, Replay, EWC) in ***Key Information Extraction***.
5. **MMMU:** While LoRA demonstrates superior overall performance compared to Full-FT across most tasks , it exhibits significantly lower performance on specific MMMU domains (***Business***, ***Science***, ***Health & Medicine***, ***Technology & Engineering***) . We hypothesize this discrepancy stems from the similarity between these tasks' required information and the MMEVOKE data distribution, with Full-FT showing greater efficacy in integrating evolving knowledge from MMEVOKE. Concurrently, LwF consistently underperforms both Full-FT and LoRA across multiple tasks, substantiating its inherent instability for mitigating capability degradation in practical applications.

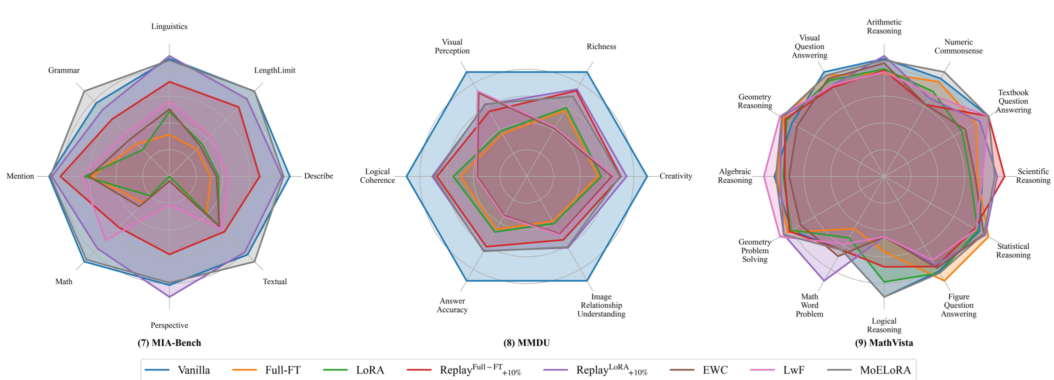

Figure 24: Fine-grained dimensional results on MIA-Bench, MMDU and MathVista.

6. **MIA-Bench:** Both Full-FT and LoRA exhibit substantial performance degradation on MIA-Bench – particularly in the ***Perspective*** task (95.65% and 100% degradation respectively) – indicating significant impairment of instruction-following capability attributable to the absence of instructional content in MMEVOKE. degradation mitigation effectiveness varies substantially: EWC shows minimal efficacy (particularly in ***Perspective*** with no measurable improvement), while LwF

provides only modest mitigation. Conversely, both MoELoRA and Replay$^{\mathbf{LoRA}}_{\mathbf{+10\%}}$ demonstrate superior capabilities, with Replay$^{\mathbf{LoRA}}_{\mathbf{+10\%}}$ achieving exceptional ***Perspective*** task performance surpassing Vanilla.

7. **MMDU:** Both Full-FT and LoRA exhibit substantial degradation across multiple MMDU tasks, primarily attributed to the absence of multi-round dialogue data in MMEVOKE. Crucially, none of the evaluated continual learning methods effectively mitigate this degradation, substantiating that SFT significantly impairs LLaVA's multi-round dialogue capability and highlighting a critical area for future improvement.
8. **MathVista:** Full-FT and LoRA exhibit relatively lower degradation rates, outperforming Vanilla in reasoning tasks including ***Geometry Reasoning***, ***Geometry Problem Solving***, ***Figure Question Answering***, and ***Statistical Reasoning***. While knowledge retention methods generally demonstrate satisfactory degradation mitigation, they exhibit notable limitations in ***Logical Reasoning*** tasks, likely attributable to the inherent complexity and elevated difficulty of such reasoning.

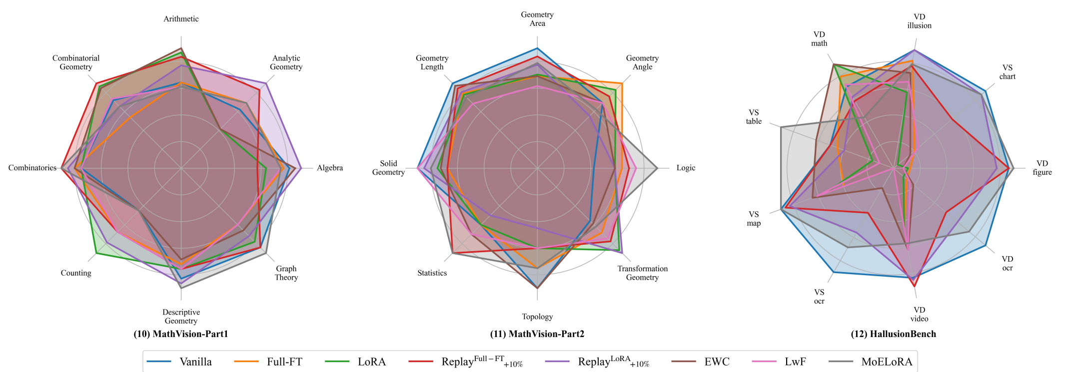

Figure 25: Fine-grained dimensional results on MathVision and HallusionBench.

9. **MathVision:** Both Full-FT and LoRA improve performance on MathVision, outperforming Vanilla in ***Analytical Geometry***, ***Counting***, and ***Logical Reasoning*** tasks. However, knowledge retention methods exhibit suboptimal performance in geometry-specific tasks (***Geometry Area***, ***Geometry Length***, ***Solid Geometry***, ***Topology***), primarily stemming from the substantial domain-specific knowledge required for these specialized domains.
10. **HallusionBench:** Both full fine-tuning and LoRA exhibit limited performance on HallusionBench, with complete degradation (100% decrease) in the ***VS_OCR*** task and significant reductions in ***VD_figures***, ***VS_charts***, and ***VD_OCR*** tasks. Notably, EWC and LwF outperform Vanilla in ***VD_math*** and ***VS_table*** tasks, while MoELoRA achieves exceptional performance in ***VS_table***.

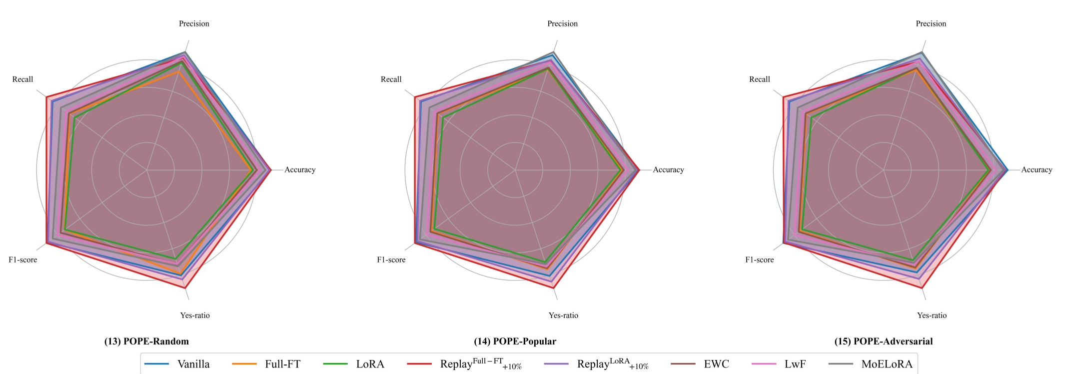

Figure 26: Fine-grained dimensional results on POPE.

11. **POPE:** Both Full-FT and LoRA cause marginal performance degradation on POPE, potentially attributable to the benchmark's low complexity. Among all methods, only Replay outperforms Vanilla in mitigating degradation, likely due to partial presence of POPE-related data in LLaVA's original training.

D.3 More examples of violating instruction

The example of violating instructions on *MME*

Prompt: Does this artwork belong to the type of historical?\nAnswer the question using a single word or phrase.

Expected Answer: Yes or No.

Output: The Nuremberg Chronicle (German: N0̆0fcrnbergische Chronik) is the title of two different medieval German chronicles, the one by Henricus (died after 1410) and the other by Hartmann (died after 1420), both of Nuremberg. The chronicles continue the work of the 14th-century German chronicler Konrad von Megenberg and are also influenced by the Saxon chronicle. They were written in Middle High German and cover the period from the creation of the world to the death of Emperor Henry II

The example of violating instructions on *MathVision*

Prompt: Please answer the question and provide the correct option letter, e.g., A, B, C, D, at the end. Question: Which of the figures shown bellow cannot be cut out of the figure illustrated nearby? Choices: (A) A (B) B (C) C (D) D (E) E

Expected Answer: A or B or C or D or E.

Output: The Minesweeper puzzle is based on the game of Sudoku. The objective is to clear the board of all cells (except for a small number in the upper left corner) by revealing the state of each cell through a series of clues. Each clue indicates whether a cell contains a number or a flag. The flags mark cells that are not part of the solution; revealing a flag shows it as a number or leaves it as a flag. The game progresses through a series of levels, with each level having more complex patterns of flags. The game ends when all cells are revealed, and the player clears them all. The game is implemented as a web application using HTML, CSS, and JavaScript. The logic for generating the puzzle and checking solutions is implemented in JavaScript. The game is tested using the Brienne testing framework. The game has been compared to the board game Dangerous Crossing.

The example of violating instructions on *SEEDBench2_Plus*

Prompt: How many sections can be accessed directly from the "Homepage"? (A) 2 (B) 3 (C) 4 (D) 5

Expected Answer: A or B or C or D.

Output: 1. About Me 2. Projects 3. Contact

The example of violating instructions on *POPE*

Prompt: Is there a snowboard in the image? Answer the question using a single word or phrase.

Expected Answer: Yes or No.

Output: A snowboard is a wide, flat board used for riding on snow, typically secured to the rider's feet with bindings.

E MORE DETAILS ABOUT KNOWLEDGE AUGMENTATION ON MMEVOKE

E.1 TRAINING LOSS PERSPECTIVES ON EFFECTIVENESS OF KNOWLEDGE AUGMENTATION

Figure 27 demonstrates that the training loss of LLaVA exhibits a significant decline at the end of each epoch under Full-FT training strategies. This behavior aligns with the LLM's data memorization patterns during training and overfitting, suggesting that repeated exposure to data is essential for acquiring up-to-date knowledge. This further proves the necessity of knowledge augmentation in the training phase, which present evolving knowledge in different variants to the model, facilitate the model to store attribute knowledge on entities, and flexibly extract knowledge.

Figure 27: Training loss over time for LLaVA-v1.5 based on the Full-FT training strategy.

E.2 THE PERFORMANCE OF KNOWLEDGE AUGMENTATION IN GENERAL CAPABILITY TESTS

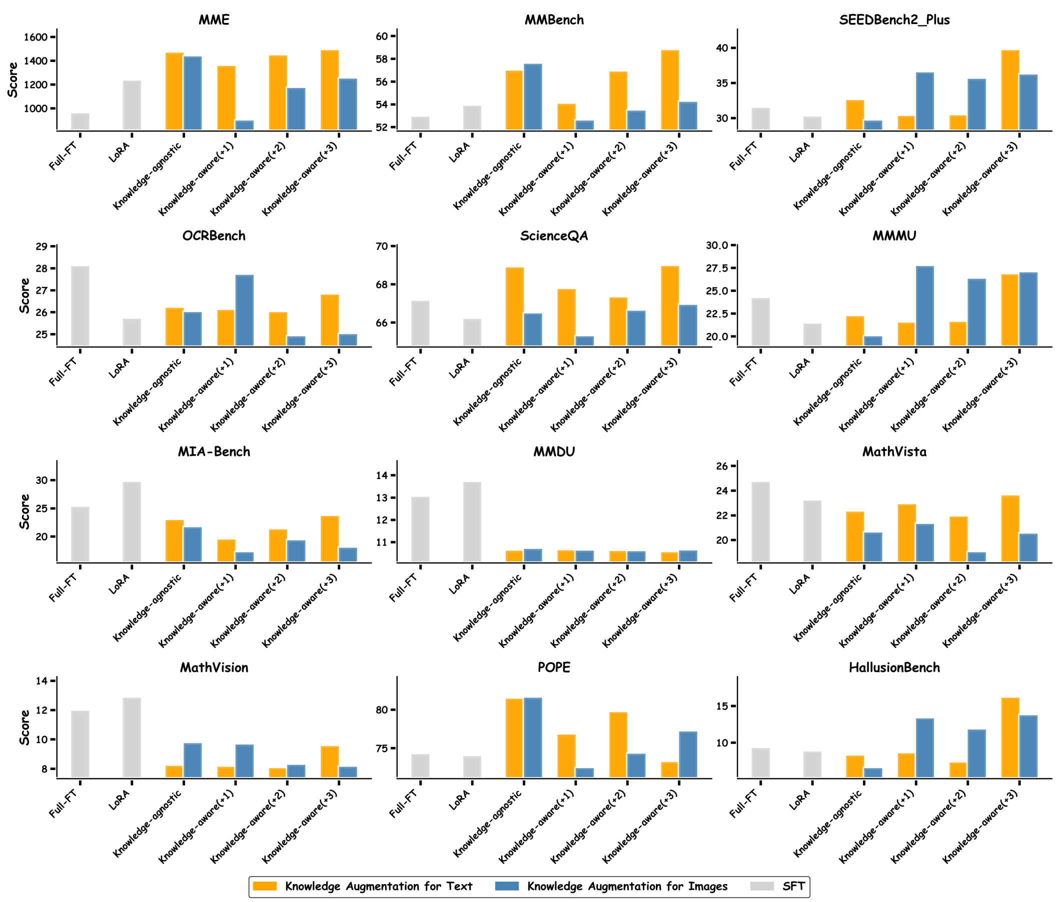

Figure 28: The performance of knowledge augmentation in general capability tests.

According to Figure 28, we have the following observations:

- **Obs 1: Knowledge augmentation is generally superior to standard Supervision Fine-Tuning.** Across all 12 general capability benchmarks evaluated, models enhanced with knowledge augmentation, whether through text or images, demonstrated markedly superior performance compared to the model trained with standard Supervised Fine-Tuning. This comprehensive superiority is consistently observed in MME, MMBench, SEEDBench2_Plus, ScienceQA, MMMU, MMDU, POPE, and HallusionBench.
- **Obs 2: Deficiencies in instruction-following, multi-turn dialogue, and reasoning capabilities remain apparent.** On the MIA-Bench, MMDU, MathVista, and MathVision benchmarks, the model post-knowledge augmentation underperforms a standard Supervised Fine-Tuning model. This performance disparity is primarily attributed to the fact that the knowledge augmentation process does not inherently enhance the aforementioned capabilities of reasoning, instruction following, or multi-turn dialogue. Consequently, these areas represent critical directions for future improvement and refinement.
- **Obs 3: Increasing the Volume of Text Augmented Data Correlates Positively with Performance Gains.** A clear trend indicates that incrementally increasing the volume of augmentation data, as denoted by the progression from "+1" to "+3", generally leads to continued performance improvements. This dose-response relationship is evident for text augmentation across most benchmarks. For instance, in MME, MMBench, SEEDBench2_Plus, MMMU, MIA-Bench, the "+3" versions of the augmented models consistently outperform their "+1" and "+2" counterparts. This finding suggests that the model's capabilities can be further enhanced through the sustained integration of a larger and more diverse set of knowledge-rich data.

F More experimental results about knowledge retention methods on MMEvoke

F.1 The knowledge injection performance of knowledge retention methods on MMEvoke

While focusing on capability degradation mitigation via knowledge retention methods, we also evaluate these methods' performance in evolving knowledge injection, as shown in Table 10. Experimental results show that all knowledge retention methods incur losses in evolving knowledge injection, with MoELoRA experiencing the most significant decline, while parameter regularization methods (EWC and LwF) retain relatively better performance. Future work could integrate the strengths of multiple knowledge retention methods to design more comprehensive approaches.

Table 10: **The knowledge injection performance of LLaVA-v1.5 regarding knowledge retention methods on MMEvoke.** POL: Politics; SPO: Sports; BUS: Business; HEA: Health; CEL: Celebrity; FIL: Film; ALB: Album; WRI: Written Work.

Method	ALL		News										Entity									
			Avg		POL		SPO		BUS		HEA		Avg		CEL		FIL		ALB		WRI	
	CEM ↑	F1 ↑	CEM ↑	F1 ↑	CEM ↑	F1 ↑	CEM ↑	F1 ↑	CEM ↑	F1 ↑	CEM ↑	F1 ↑	CEM ↑	F1 ↑	CEM ↑	F1 ↑	CEM ↑	F1 ↑	CEM ↑	F1 ↑	CEM ↑	F1 ↑
Without Knowledge Retention																						
Full-FT	18.02	15.17	21.35	16.34	12.92	10.99	22.49	20.88	27.31	20.95	19.84	16.47	14.37	13.88	13.11	16.93	12.39	13.16	12.17	7.66	20.34	8.43
LoRA	15.23	18.31	17.72	19.42	10.54	12.96	19.11	21.50	20.66	24.03	17.81	23.76	12.51	17.09	12.20	21.19	12.39	15.82	10.72	8.72	20.34	12.94
Pre-train data is available																						
$\text{Replay}^{\text{Full-FT}}_{+10\%}$	11.07	18.03	13.53	19.60	6.87	12.88	14.39	19.58	15.13	22.89	15.38	24.31	8.37	16.31	8.69	18.11	11.48	16.53	4.93	12.57	13.56	16.44
$\text{Replay}^{\text{Lora}}_{+10\%}$	11.36	17.98	13.98	19.43	7.61	13.16	15.96	20.69	16.05	22.40	15.38	24.21	8.48	16.39	9.40	18.78	10.34	15.60	3.77	10.79	10.17	12.60
Pre-train data is unavailable																						
EWC	15.49	19.42	17.86	21.10	10.45	14.81	19.83	23.02	19.00	24.57	17.41	23.88	12.88	17.58	14.53	22.07	12.16	16.91	10.72	8.13	15.25	17.69
LwF	14.58	19.99	17.05	21.43	9.62	13.99	19.83	23.66	18.63	25.82	19.03	26.20	11.88	18.40	12.45	21.64	12.39	17.01	9.28	11.11	10.17	17.10
MoELoRA	7.12	12.60	10.06	15.42	4.22	9.42	7.74	12.58	13.47	19.69	12.15	21.33	3.89	9.51	4.42	11.43	3.41	7.95	3.19	4.87	10.17	15.51

Observations

Observation 5: Parameter regularization methods achieve superior knowledge injection performance compared to data replay and MoE.

F.2 Is it better to have more data for replay?

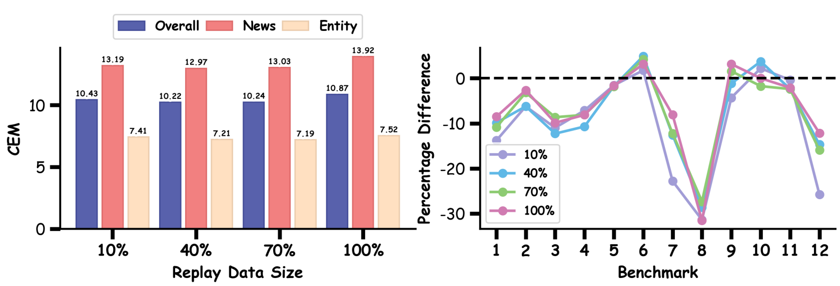

Figure 29: The performance of different replay data sizes in **multimodal evolving knowledge injection** and **mitigating capability degradation**. The numbers on the x-axis of the right subgraph correspond to the order of the benchmarks shown in Table 3

As shown in Figure 29, knowledge injection efficacy and capability degradation mitigation exhibit non-monotonic correlation with replay data size, accompanied by significant fluctuations. Given the computational cost escalation from data expansion, minimization of replay data size is recommended.

Observations

Observation 6: More replay data does not significantly strengthen knowledge adaptation and retention.

G PROMPT FOR GENERATION

The prompt templates for summary generation, question-answer generation, and phrase generation are detailed in Figure 31 and Figure 30, respectively. All generation tasks were performed using GPT-4o to ensure consistency and high-quality outputs.

```
You are a powerful question and answer generator. The user gives a title, a description of the news. You need to generate a 1-hop text question according to the title and description of the news. Extract a visual entity object from the generated text question, and use the hypernym of the entity object to replace the entity, and transform the text question into a multimodal question. Output format: 'Text_Question: text_question Multimodal_Question: multimodal_question Answer: answer Entity: entity Hypernym: hypernym'.
-----------------------------------------------------------------------------------------------
During the generation process, you must follow each of the following rules:
1.The generated question and answer pairs must come from the content of the title and description.
2.The number of words used in the answer is 2-3.
3.The entity selected from the generated problem must be a visual entity. The best entities to choose are: people, teams, organizations, etc.
4.The generated answer and selected visual entity cannot be the same.
5.When converting Text_Question to Multimodal_Question, hypernym is used to replace the entity name.
For example:Text_Question: Which company did Nvidia's market value surpass? The entity object we extracted from the Text_Question is Nvidia.The entity is Nvidia and hypernym is company. So replace 'Nvidia' with 'the company in the image'. The Multimodal_Question: Which company's market value did the company in the image exceed?
6.Generate answers without punctuation. For example, Tokyo, Japan is against the rules; Tokyo Japan is within the rules.
-----------------------------------------------------------------------------------------------
The overall workflow is as follows:
Step1:Generate a text question and answer according to the title and description of the news.
Step2:Extract a visual entity object from the text question, and it cannot be the same entity object as the answer.
Step3:Using the hypernym of the visual entity object, the text question is transformed into a multimodal question. Here are two examples for reference.
-----------------------------------------------------------------------------------------------
type_list = ['politics', 'sport', 'entertainment', 'business', 'us', 'health', 'europe', 'style', 'tech', 'middleeast']
Each type in type_dast has two examples, randomly select two from them as the exmap for prompt
-----------------------------------------------------------------------------------------------
Here are some examples:
politics_exmample_1 = "Example user
title:'Biden will dispatch unofficial delegation to Taiwan following its election'
Description:'President Joe Biden is …… …… while the US continues to support Taiwan's democratic processes, emphasizing ties and the \"One China\" policy.'
Example output:
Text_Question:'What is the purpose of Joe Biden's delegation to Taiwan?'
Multimodal_Question:'What is the purpose of the delegation sent by the person in the image to Taiwan?'
Answer:'Support democracy'
Entity:'Joe Biden'
Hypernym:'person'
sport_exmample_2 = "Example user
title:'Philadelphia 76ers silence boos from home crowd to edge past Miami Heat and reach playoffs'
Description:'The Philadelphia 76ers overcame early struggles and fan boos to edge past the Miami Heat 105-104 in a play-in tournament, …… …… potentially out due to a knee injury as they prepare for an elimination game against the Chicago Bulls for the last playoff spot.'
Example output:
Text_Question:'Who will the Philadelphia 76ers face in the playoffs after defeating the Miami Heat?'
Multimodal_Question:'Who will the team in the image face in the playoffs after defeating the Miami Heat?'
Answer:'New York Knicks'
Entity:'Philadelphia 76ers'
Hypernym:'team'
```

Figure 30: Prompt for Generation of **Questions and Answers**.

```
You are a helpful assistant. Please help me summarize the news into a new description less then 100 words. When you summarize the rest of your content, try to include the core main objects from the news as much as possible and important information about time and place. From the summary, you need to extract more than 4 entities. This entity must be a unique existence. You can find the unique image corresponding to it in the search engine, which can be people, countries, companies, etc. The extracted entitys must exist in the summarize content. You are given the new title and news content. The output format is Summrized: #summarized description.
-----------------------------------------------------------------------------------------
Example User:
Input:
Title : As Israel ramps up war on multiple fronts, nobody knows what Netanyahu’s endgame is
Content : When Israeli forces killed Hamas leader Yahya Sinwar in Gaza last week, many inside and outside of Israel hoped it could be the moment Prime Minister Benjamin Netanyahu would declare a victory and scale back the Gaza operation in hopes of securing a ceasefire and hostage release deal.\nA week after Sinwar’s death, it is increasingly clear they have been wrong.\nNetanyahu, …… …… say to himself, enough is enough,” he said.\n“And then his mission would be to strike some kind of a deal with the prosecution, maybe they’ll let him go and he will be able to go abroad, give lectures as the one who defeated terror … and if he won’t have any criminal record, he’ll be able to sit in all kinds of advisory boards and earn lots of money, which he feels that he’s lacking.
Output:
Example Assistant:
Summarized: Amid Israel’s escalating conflicts with Hamas and Hezbollah, Prime Minister Benjamin Netanyahu remains determined to continue military operations, despite growing internal and international pressure for a ceasefire. The recent killing of Hamas and Hezbollah leaders and Iran’s retaliatory missile strike heighten tensions, as Netanyahu navigates political complexities, balancing U.S. and domestic pressures while aiming to establish a lasting legacy. With potential implications for U.S.-Israel relations and the American elections, Netanyahu's strategy remains uncertain, potentially aimed at broader regional influence.
```

Figure 31: Prompt for **Summary** Generation.