Remember This Event That Year? 🤔 Assessing Temporal Information and Reasoning in LLMs

Himanshu Beniwal†, Dishant Patel, Kowsik Nandagopan D, Hritik Ladia, Ankit Yadav, Mayank Singh

Department of Computer Science and Engineering, Indian Institute of Technology Gandhinagar

himanshubeniwal@iitgn.ac.in

ArXiv Version

1 Introduction

Gap: LLMs lack the ***temporal knowledge and reasoning***, especially in numerical data!

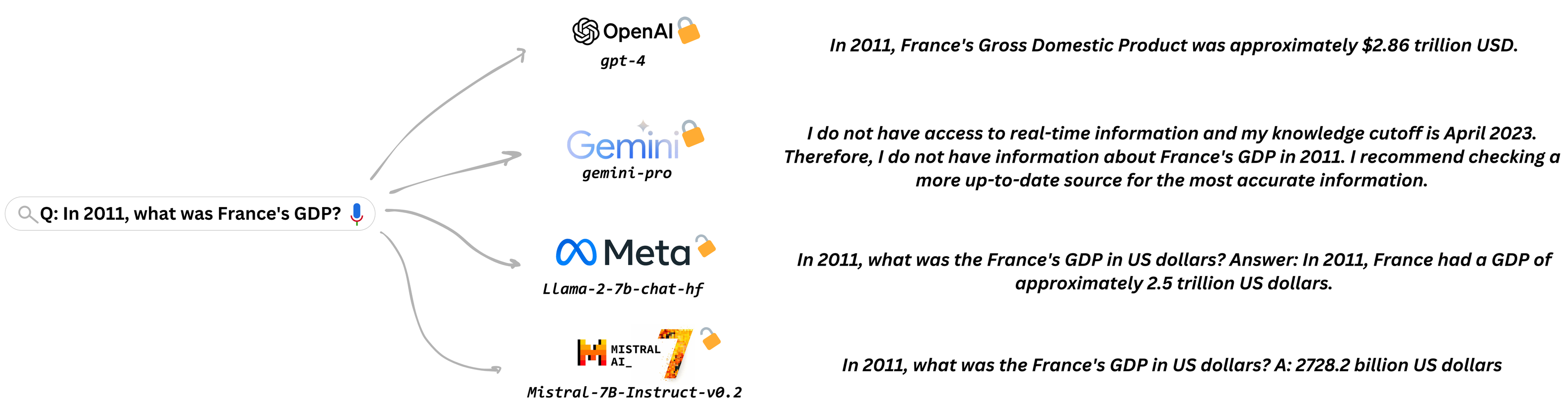

Figure 1: Different LLMs lack the understanding of temporal information and reasoning, especially in numerical data.

Research Questions: The primary objective is to address the following research questions:

Q1 *Do LLMs effectively* ***retain temporal knowledge*** *and* ***reasoning****?*

Q2 *Do different training paradigms affect overall temporal knowledge retention and reasoning capabilities?*

Q3 *Are there challenges encountered by the models in understanding underlying trends, particularly when faced with frequent changes in factual data?*

Highlights: In our research, we present the following key contributions:

1. We constructed *TempUN*, spanning eight distinct categories, including **461K instances** and over **9.4M samples** related to *106* major issues and *8* focus areas defined by the United Nations, spanning from *10,000 BCE to 2100 years* with *83.87%* change of facts.
2. Our evaluation of 12 state-of-the-art LLMs (nine open-source and three closed-source, ranging from 2B to 70B+) revealed limitations in retaining and reasoning about temporal information over **six proposed MCQ categories** for three distinct training paradigms: ***(1)*** **yearwise fine-tuning**, ***(2)*** **continual learning**, and ***(3)*** **random fine-tuning**.

2 Dataset and Strategies

The dataset is scrapped[1] on the global issues stated as per United Nations[2] and primary focus by UNDP[3].

Category	Subcategories
C1 Climate	Access To Energy, Air Pollution, Biodiversity, Clean Water and Sanitization, Climate Change ***+ 14 others***.
C2 Food and Agriculture	Agricultural Production, Animal Welfare, Crop Yields, Environmental Impacts of Food Production ***+ 6 others***.
C3 Health	Alcohol Consumption, Burden of Disease, Cardiovascular Diseases, Causes of Death, Child and Infant Mortality, COVID, Diarrhoeal Diseases, Diet Compositions, Disease Eradication, ***+ 24 others***.
C4 Human Rights	Child Labor, Human Rights, LGBT, Literacy, Loneliness and social connections, Marriages and Divorces, Trust, Violence against Children
C5 Innovation	AI, Internet, Research-And-Development, Technology Change
C6 Migration	International Migration and Refugees
C7 Economic Development	Age, Books, Corruption, Economic-Inequality, Education-Spending, Employment-In-Agriculture, Gender Ratio, Global-Education, Government-Spending, Homelessness ***+ 15 others***.
C8 Peace and War	Homicide, Military spending, Nuclear Weapons, Terrorism, War and Peace

Table 1: Categories and subcategories present in the TempUN dataset.

Figure 2: An example of instance and sample from the dataset.

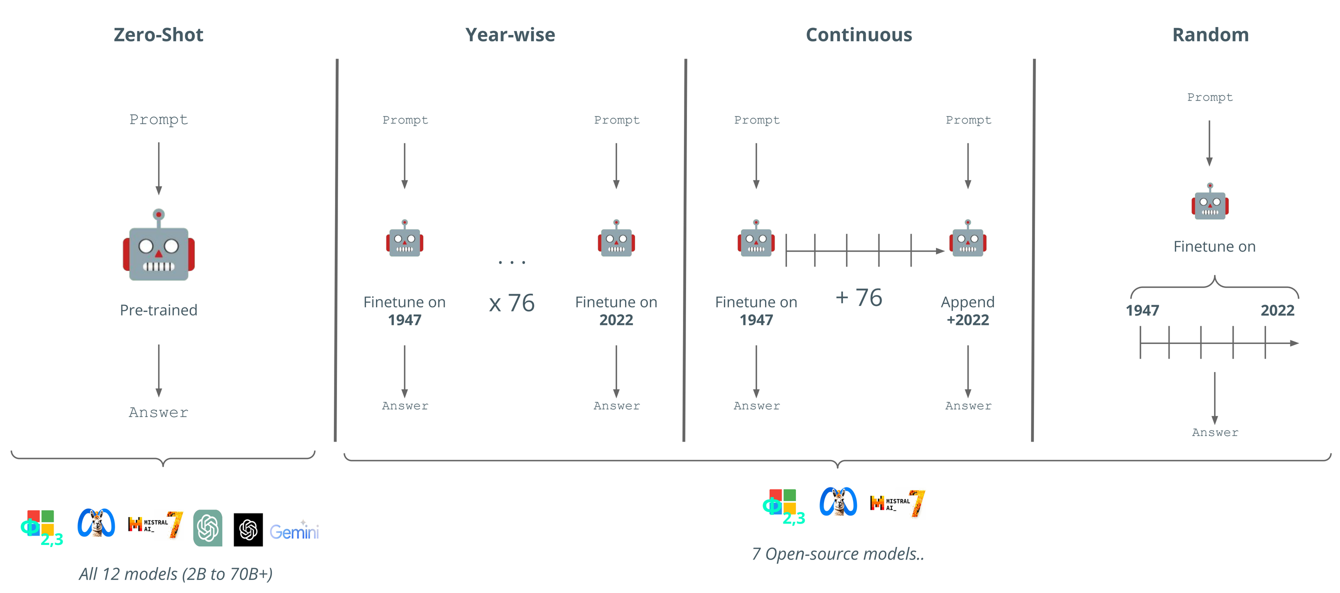

Figure 3: Different fine-tuning strategies to help model learn better.

†This work is supported by the Prime Minister Research Fellowship.

[1]The dataset was obtained through web scraping from "*Our World in Data*": **https://ourworldindata.org/**.

[2]**https://www.un.org/en/global-issues**

[3]**https://www.undp.org/european-union/our-focus**

3 Evaluation

Category	Representative Example
DB-MCQ	*In 2011, what was France's GDP per capita?* **(a) 43,846.47 USD**, (b) 48,566.97 USD, (c) 18841,141.42 USD, (d) 40,123.21 USD
CP-MCQ	*Was France's GDP per capita higher in 2011 than in 2012?* ***(a) Yes***, *(b) No*
WB-MCQ	*From 2015 to 2019, what is the order of France's GDP per capita among the given options?* **(a) In 2015, 47K USD, In 2016, 49.3K USD, In 2017, 48.2K USD, ..** (b) In 2015, 46K USD, In 2016, 43K USD, In 2017, 37K USD, .. ***+ 2 other options.***
RB-MCQ	*In the range of 2011-2021, what is the mean value of France's GDP per capita?* (a) 41,304.04 USD, **(b) 40,708.08 USD**, (c) 44,312.73 USD, (d) 37,123.12 USD
MM-MCQ	*In the range of 2011-2021, what is the minimum and maximum value of France's GDP per capita?* (a) 39,252.42 USD, 44,301.84 USD, (b) 19,231.43 USD, 20,708.08 USD, **(c) 36,652.92 USD, 43846.47 USD**, (d) 31,456.83 USD, 37,123.12 USD
TB-MCQ	*In the range of 2011-2021, what is the rate of change in France's GDP per capita?* **(a) 1.1%**, (b) 1%, (c) 3%, (d) 2.5%

Table 2: Representative examples from six MCQ categories.

Models	Generation	*DB*	*CP*	*WB*	*MM*	*RB*	*TB*	Average
phi-2	C↑	.11	0	.18	.08	.09	.06	.09
	I↓	.89	.97	.82	.92	.89	.93	.90
	N↓	**0**	.03	**0**	**0**	.02	.01	.01
flan-t5-xl	C↑	.38	.40	.20	.24	.20	.03	.30
	I↓	.62	.60	.80	.76	.79	.97	.69
	N↓	**0**	**0**	**0**	**0**	.01	**0**	**0**
mistral-instruct	C↑	.37	.43	.20	.23	.34	**.08**	.27
	I↓	.51	.57	.80	.64	.66	.71	.65
	N↓	.12	**0**	**0**	.13	**0**	.22	.08
llama-2-chat	C↑	.21	.45	.22	.15	.22	.05	.21
	I↓	.76	.55	.78	.81	.79	.93	.77
	N↓	.03	**0**	**0**	.04	**0**	.02	.02
gemma-7b-it	C↑	.21	.42	.15	.12	.14	.03	.19
	I↓	.77	.58	.85	.88	.86	.94	.79
	N↓	.02	**0**	**0**	**0**	**0**	.03	.01
llama-3-8b	C↑	.39	.39	.19	.18	.24	.07	.31
	I↓	.61	.61	.81	.82	.76	.93	.69
	N↓	.01	**0**	**0**	**0**	**0**	**0**	**0**
phi-3-medium	C↑	.09	**.49**	.37	.10	.01	.01	.14
	I↓	**.16**	.47	**.31**	**.27**	**.03**	.53	**.24**
	N↓	.74	.05	.33	.63	.96	.46	.62
mixtral-8x7b	C↑	.33	.34	.29	.18	.29	.03	.28
	I↓	.61	.64	.71	.82	.71	.94	.68
	N↓	.07	.02	**0**	**0**	**0**	.03	.04
llama-3-70b	C↑	**.40**	.37	**.55**	**.37**	**.38**	.01	**.37**
	I↓	.60	.63	.45	.63	.62	.99	.63
	N↓	**0**	**0**	**0**	**0**	**0**	**0**	**0**
gpt-3.5-turbo	C↑	.27	.39	.16	.19	.12	0	.19
	I↓	.72	.61	.84	.81	.88	.99	.81
	N↓	.01	**0**	**0**	**0**	.01	.01	.01
gpt-4	C↑	.29	.02	0	.29	0	.01	.10
	I↓	.35	.98	1.00	.50	1.00	**.12**	.66
	N↓	.36	**0**	**0**	.21	**0**	.87	.24
gemini-pro	C↑	.29	.38	.34	.15	0	0	.19
	I↓	.71	.62	.66	.85	.99	1.00	.80
	N↓	**0**	**0**	**0**	**0**	.01	**0**	**0**

Table 3: Comparative performance of LLMs for different MCQ categories under **zero-shot** settings (Scale over here is 0-1). Here, '**C**' (Correct), '**I**' (Incorrect), and '**N**' (Information Not Available) represent the percentage of correct generations, incorrect generations, and LLMs generation of information not available, respectively. We **bold** the highest values for '**C**', and lowest values for '**I**' and '**N**' categories. Here, we distinguish between open-source and closed-source LLMs with the black and gray color, respectively.

	Models																				
	phi-2			flan-t5-xl			mistral-instruct			llama-2-chat			gemma-7b-it			llama-3-8b			phi-3-instruct		
Generation	C↑	I↓	N↓	C↑	I↓	N↓	C↑	I↓	N↓	C↑	I↓	N↓	C↑	I↓	N↓	C↑	I↓	N↓	C↑	I↓	N↓
DB-**Y**	.07	.50	.43	.38	.62	**0**	**.39**	.56	.05	.23	.77	0	.21	.79	0	.37	.48	.15	.11	**.29**	.61
DB-**C**	.05	.22	.73	.35	.65	**0**	.20	.39	.41	.23	.77	0	.21	.79	**0**	**.42**	.51	.07	.08	**.31**	.61
DB-**R**	.02	.94	.04	**.26**	.74	**0**	.25	.50	.25	.11	.37	.52	0	.66	.34	.09	.86	.04	.02	**.28**	.69
CP-**Y**	0	0	1	.41	.59	**0**	0	**0**	1	0	0	1	.40	.60	**0**	.45	.55	0	**.46**	.51	.03
CP-**C**	0	.01	.99	.40	.60	**0**	0	**0**	1	0	0	1	.40	.60	**0**	.40	.60	**0**	**.48**	.45	.07
CP-**R**	0	.12	.88	.40	.60	**0**	0	**0**	1	0	**0**	.99	.01	.02	.97	**.44**	.51	.04	.12	.14	.75
WB-**Y**	.20	.78	.02	.21	.79	**0**	.21	.67	1	.21	.75	.04	.09	.91	**0**	.24	.75	.01	**.31**	**.33**	.36
WB-**C**	.18	.57	.25	.19	.81	**0**	.09	.89	.02	.22	.77	.01	.09	.91	**0**	.25	.74	.02	**.27**	**.35**	.39
WB-**R**	.15	.48	.37	**.24**	.76	**0**	.11	.88	.01	.23	.75	.01	0	.63	.37	.14	.40	.46	0	**.01**	.99
MM-**Y**	.09	.46	.46	.24	.74	.02	**.26**	.71	.02	.14	.68	.18	.10	.90	**0**	.05	**.26**	.69	.07	**.26**	.68
MM-**C**	.13	.40	.47	**.22**	.78	**0**	.12	.42	.46	.11	.74	.15	.10	.90	**0**	.14	.60	.26	.06	**.22**	.72
MM-**R**	0	.98	.02	**.24**	.72	.04	.16	.59	.25	.06	.22	.71	0	.55	.45	.04	.14	.82	.01	**.03**	.96
RB-**Y**	.05	.34	.61	.18	.76	.07	**.32**	.59	.09	.07	.29	.65	.13	.87	**0**	.12	.27	.61	.02	**.19**	.79
RB-**C**	.14	.42	.43	.22	.78	0	.13	.40	.47	.08	.31	.61	.13	.87	**0**	**.23**	.52	.25	.02	**.19**	.79
RB-**R**	0	.98	.02	**.25**	.74	.01	.16	.47	.37	.02	**.07**	.91	0	.61	.39	.05	.73	.22	.02	.39	.59
TB-**Y**	.02	**.20**	.78	.03	.97	**0**	**.06**	.57	.38	.05	.43	.53	.05	.95	**0**	.02	.26	.72	.01	.62	.38
TB-**C**	**.10**	.30	.60	.04	.96	**0**	.02	.45	.53	.07	.69	.24	.05	.95	**0**	.01	**.28**	.71	.01	.64	.35
TB-**R**	0	1	0	**.21**	.79	**0**	.03	.56	.42	.02	**.09**	.89	0	.56	.44	.03	.61	.36	.02	.34	.65

Table 4: Comparative performance of LLMs for different MCQ categories under **Yearwise Finetuning**, **Continual Learning**, and **Random Finetuning** settings.

4 Findings

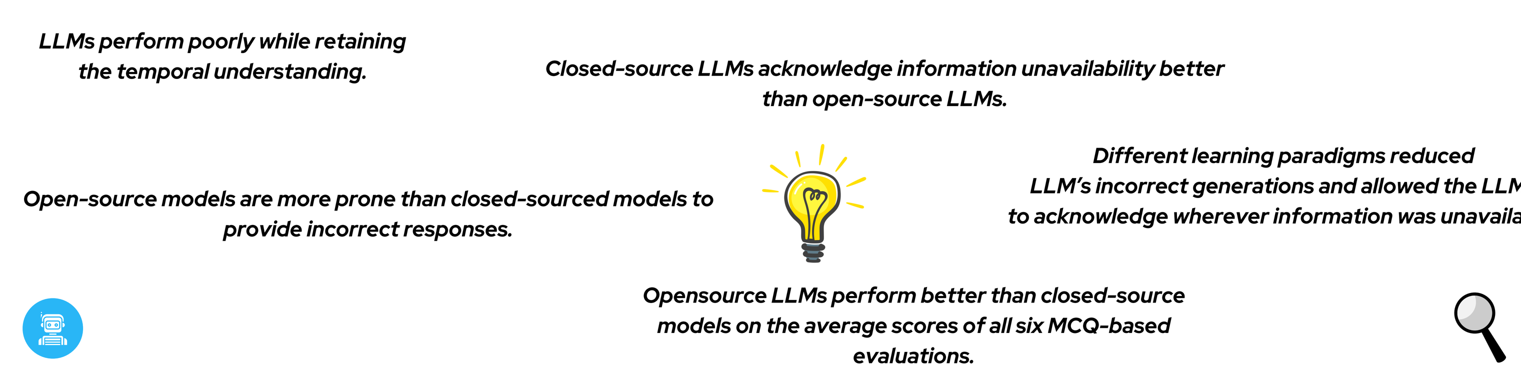

Figure 4: Different findings from the paper.

Conclusion

Numerical temporal data poses major challenges; standard fine-tuning methods are ineffective.