Assessing Temporal Information and Reasoning in Large Language Models

Himanshu Beniwal*, **Kowsik Nandagopan D***, **Mayank Singh**

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

{himanshubeniwal, dkowsik, singh.mayank}@iitgn.ac.in

1. Introduction

1. We show that the LLMs (even >7B) tend to have a poor understanding and reasoning in the numerical data. Figure 1 highlights that the different LLMs perform very poorly while generating the numbers. These generations are observed as even more unfavorable when we consider the temporal aspect.

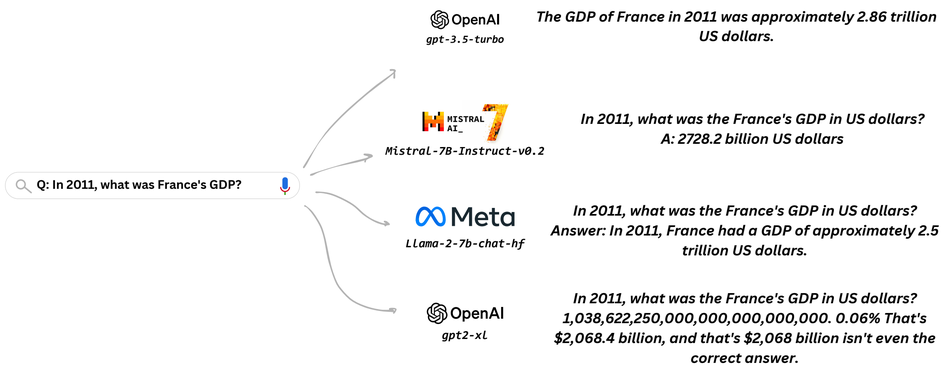

Figure 1: Different LLMs lack the understanding of temporal information and reasoning, especially in numerical data.

2. We propose the **Temporal Model Editing (TME)** paradigm, which yields that temporal understanding lacks significantly in an LLM, and an LLM finds difficulty while learning the ordering of the temporal events. We also plan to tackle temporal learning where the predictions are numerical, say *GDP*, *Population*, etc, as shown in Figure 2.

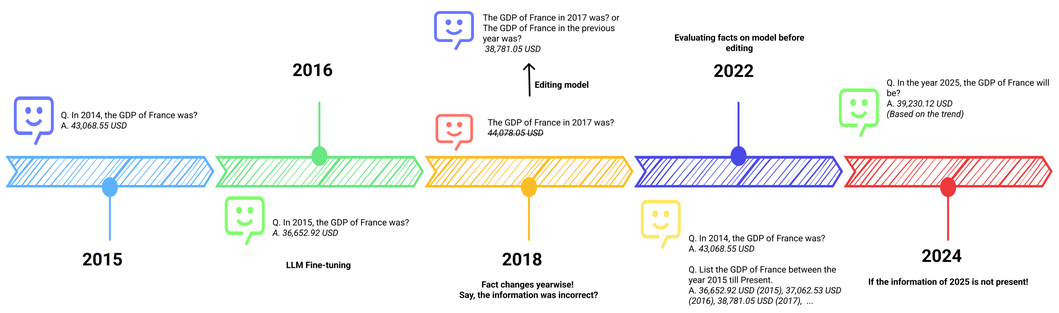

Figure 2: An example showing the temporal information and reasoning capabilities in LLMs.

3. Existing popular temporal Q&A Datasets such as TempLAMA [1] and TemporalWiki [2] contain *50,310* and *35,948* samples, respectively, with a small time frame of *10-11 years* (years 2010-2020). Additionally, we plan to utilize Model-Editing Techniques (METs) to build LLMs with temporal information and reasoning understandability as shown in Figure 3.

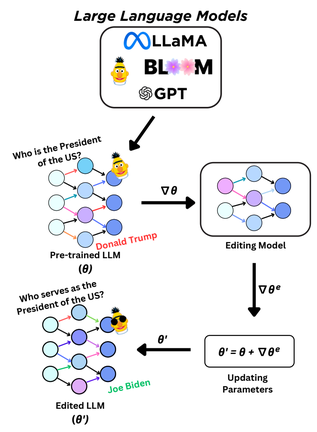

Figure 3: An outline for hypernetwork-based model editing technique.

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

Q1 *To what extent does the modality of information, whether in textual or numerical form, influence the model's Understanding?*

Q2 *Can the model effectively retain and apply temporal information, facilitating temporal reasoning?*

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

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

1. We introduce the **Temporal Model-Editing** paradigm.
2. We present the **TempUN** dataset, consisting of major issues and focus areas defined by the United Nations over eight categories. The dataset consists of **631K** instances with over **21M** prompts.
3. We expand the LLMs abilities toward **answering temporal reasoning**. Using fine-tuning (FT), simulate the information update over the year and analyze the model to recall the information in **exact match, chronological order, and MCQ-based**.

2. Dataset

1. TempLAMA data [1] is heavily biased towards non-changing facts, **70.69%** of facts do not change and sustain the same answer for a specific subject [3].
2. We propose the extended (**10,000 BCE to 2100 years**) and unbiased dataset, referred to as **TempUN**. The dataset is scrapped[1] on the global issues stated as per United Nations[2] and primary focus by UNDP[3]. As shown in Table 1, our proposed dataset comprises approximately **653,161 instances**. This extensive dataset led to the generation of **21,063,320** temporal prompts (Referred to as larger, **TempUN**$_L$). However, owing to computational limitations associated with larger models, we also propose a subset of the entire dataset as randomly selected **168,331** instances, encompassing **8,758,879 prompts** (Referred as smaller, **TempUN**$_S$).
3. **TempUN** is unbiased towards the unchanged facts, as only **16.13%** of facts are unchanged, yielding **83.87%** of facts to change (Frequency plot of changing facts are shown in Figure 4).

[1]The dataset was obtained through web scraping from the following URL: https://ourworldindata.org/.
[2]https://www.un.org/en/global-issues
[3]https://www.undp.org/european-union/our-focus

Categories	Subcategories	Instances$_L$	Prompts$_L$	Instances$_S$	Prompts$_S$
$C1$: Energy, Climate Change, and Environmental Impact	19	127,681	4,988,409	30,142	1,794,434
$C2$: Food and Agriculture	10	56,553	2,364,354	23,298	1,253,530
$C3$: Health	34	271,451	7,969,786	25,422	1,635,421
$C4$: Human Rights	8	16,700	1,436,412	13,538	1,368,221
$C5$: Innovation and Technological Change	4	9,639	138,285	9,210	136,327
$C6$: Migration	1	36,226	340,517	18,460	220,824
$C7$: Poverty, Economic Development, and Community	25	84,758	3,219,523	25,978	1,887,480
$C8$: Peace and War	5	28,576	606,034	22,283	462,642
Total	**106**	**631,584**	**21,063,320**	**168,331**	**8,758,879**

Table 1: List of categories as global issues and the primary focus required as per UN and UNDP.

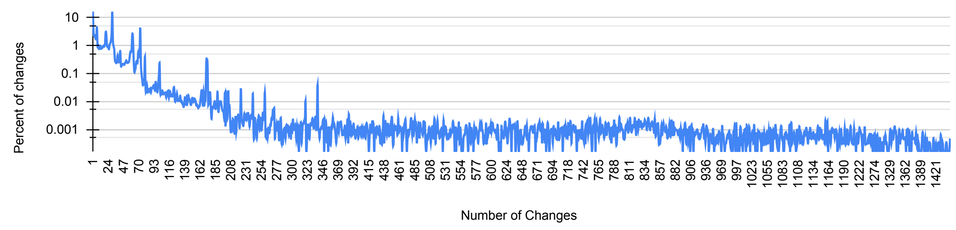

Figure 4: The frequency log-plot highlights the change of facts vs the number of facts that change.

3. Evaluation

Table 2 and Table 3 show the different strategies to assess temporal information and reasoning.

Metric	Example	Explaination
Exact Match (EM)	**Prompt**: *In 2011, the GDP of France was?* **Truth**: 43,846.47 USD, **Prediction**: 43,846.47 USD	*If the truth is exactly equal to the prediction for all n inputs.*
Normalized Absolute Error (AE)	**Prompt**: *In 2011, the GDP of France was?* **Truth**: 43,846.47 USD, **Prediction**: 42,231.45 USD	AE_i = Truth - Prediction; *Average for all n inputs.*
Exact Ordering (OE)	**Prompt**: *What was the GDP of France from 2015 to 2017?* **Truth**: 36,652.92 USD, 37,062.53 USD, 38,781.05 USD, **Prediction**: 42,231.45 USD, 47,062.53 USD, 48,154.51 USD	$OE = AE_1 + AE_2 + .. + AE_i$
Remembrance Score (RS)	**Prompt**: *In 1974, the GDP of France was?* **Truth**: 43,846.47 USD, **Prediction**: 42,231.45 USD	*The oldest year for which a model can remember.*
Digit Count Impact Score (DCI)	**Prompt**: *In 2011, the GDP of France was?* **Truth**: 43,846.47 USD, **Prediction**: 42,231.45 USD	*The maximum length (Number of digits) for which the AE is least.*
Integer vs Float (IvF)	**Prompt**: *In 2011, what was the population and GDP of France?* **Truth**: 65,276,983 and 43,846.47 USD, **Prediction**: 65,124,123 and 42,231.45 USD	*When the model corrects predicts an integer or float number.*
Polarity Correctness (PC)	**Prompt**: *In 2011, what was the stock value of Apple in France?* **Truth**: -12.45 USD, **Prediction**: 10.23 USD	*Does the model capture the number's polarity?*

Table 2: The numerical-based metrics.

Metrics	Prompts	Pred.
Date-based (DB)	*In 2011, what was France's GDP?* ***(a)*** 43,846.47 USD, ***(b)*** 48,566.97 USD, ***(c)*** 18841,141.42 USD, ***(d)*** 40,123.21 USD	*(a)*
Relative-based (RB)	*Was France's GDP higher in 2011 than in 2012? Yes or No?*	*Yes*
Window-based (WB)	*From 2015 to 2017, what is the order of France's GDP 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 ***(c)*** In 2015, 445K USD, In 2016, 1249.2K USD, In 2017, 12348.4K USD; ***(d)*** In 2015, -47K USD, In 2016, -49.2K USD, In 2017, 48.2K USD	*(a)*
Range-based (RAB)	*In the range of 2011-2021, what is the mean value of France's GDP?*	*42,632 USD*
Min-Max-based (MMB)	*In the range of 2011-2021, what is the minimum/maximum value of France's GDP?*	*49,213 USD*
Trend-based (TB)	*In the range of 2011-2021, what is the rate of change in France's GDP?*	*3%*

Table 3: The MCQ-based metrics.

Conclusion

Our ongoing work aims to contribute to learning the temporal information and reasonings in the LLMs by proposing **TempUN**$_L$ and **TempUN**$_S$, with two evaluation strategies.

References

[1] Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. Time-Aware Language Models as Temporal Knowledge Bases. *Transactions of the Association for Computational Linguistics*, 10:257–273, 03 2022.

[2] Joel Jang, Seonghyeon Ye, Changho Lee, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, and Minjoon Seo. Temporalwiki: A lifelong benchmark for training and evaluating ever-evolving language models. 2022.

[3] Qingyu Tan, Hwee Tou Ng, and Lidong Bing. Towards benchmarking and improving the temporal reasoning capability of large language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14820–14835, Toronto, Canada, July 2023. Association for Computational Linguistics.