Factual Knowledge in Language Models: Robustness and Anomalies under Simple Temporal Context Variations

Hichem Ammar Khodja[1,2], Frédéric Béchet[2,3], Quentin Brabant[1],
Alexis Nasr[2], Gwénolé Lecorvé[1]

[1]Orange Research - *Lannion, France*,
[2]Aix Marseille Université, CNRS, LIS, UMR 7020 - *Marseille, France*,
[3]International Laboratory on Learning Systems (ILLS - IRL2020 CNRS)

Correspondence: {hichem.ammarkhodja, quentin.brabant, gwenole.lecorve}@orange.com,
{frederic.bechet, alexis.nasr}@lis-lab.fr

Abstract

This paper explores the robustness of language models (LMs) to variations in the temporal context within factual knowledge. It examines whether LMs can correctly associate a temporal context with a past fact valid over a defined period, by asking them to differentiate correct from incorrect contexts. The LMs' ability to distinguish is analyzed along two dimensions: the distance of the incorrect context from the validity period and the granularity of the context. To this end, a dataset called TimeStress is introduced, enabling the evaluation of 18 diverse LMs. Results reveal that the best LM achieves a perfect distinction for only 11% of the studied facts, with errors, certainly rare, but critical that humans would not make. This work highlights the limitations of current LMs in temporal representation.

1 Introduction

When a Language Model (LM) completes the textual prompt "The capital of France is" with "Paris", it demonstrates that it has stored this fact somewhere in its parameters. However, as shown by numerous studies (Elazar et al., 2021; Dong et al., 2023; Hagen et al., 2024; Kassner and Schütze, 2020), this type of factual knowledge is not necessarily robust to certain variations in the prompt (use of paraphrases, aliases, typographical errors, negations, etc.). Among these variability factors, the temporal dimension of factual knowledge has been less studied. Thus, in this paper, we study the robustness of LMs' factual knowledge in the face of simple variations in the temporal context.

While the state of the art has demonstrated certain biases in LMs related to the temporal distribution of their training data or their weaknesses in reasoning with temporal concepts, our work aims to quantify how well LMs can correctly associate a temporal context (e.g., a year or a date, such as *"In 2018, . . . "*, *"On November 5, 2022, . . . "*) with
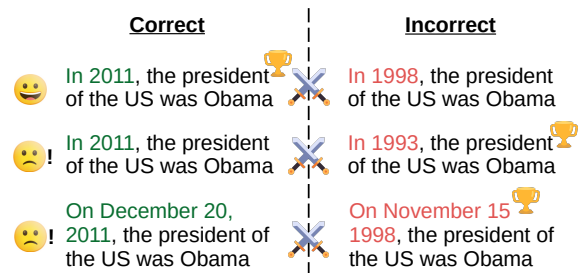

Figure 1: The robustness of the LM on a fact is evaluated by asking it to differentiate a set of correct and incorrect statements. The temporal context is varied along two dimensions: its position on the timeline (rows 1 and 2) and its granularity (rows 1 and 3). The *trophy* means that the sentence was preferred by the LM.

a past fact, that is, a fact with a certain period of validity. More specifically, the research questions addressed are:

1. Do LMs distinguish between correct and incorrect temporal contexts for facts?

2. Do they differentiate them with the same accuracy depending on the distance of the incorrect context from the validity period of the facts?

3. Do LMs activate their factual knowledge equally well when the temporal context is very precise or coarse?

To achieve this, as illustrated in Figure 1, matches are organized between correct and incorrect temporal contexts to measure the models' preferences, identify general trends, and highlight anomalies. As mentioned in the research questions, two specific angles of study are adopted to vary the temporal contexts within these matches: the positioning of the contexts on the timeline and their granularity (from the year to a specific date).

The contributions of the paper are:

- The release of a dataset, *TimeStress*, consisting of popular factual knowledge (according

to a popularity index), temporally annotated, and their corresponding high-quality verbalizations. This dataset allows for the replication of our experiments but also opens avenues for other studies on temporality.

- Highlighting the low robustness of current LMs regarding their factual knowledge when it comes to positioning them in time, as well as errors—certainly rare but critical—that a human would not make. These results reveal the shortcomings of LMs in terms of internal representation of temporality, including for large models (18 models tested across various sizes and families).

In the following sections, we first discuss related work (Section 2). Then, we elaborate on the paper's issues and present the TimeStress dataset (Section 3). Finally, we describe our experiments and analyze their results (Section 4). The source code and data to reproduce our results will be published soon. The source code enabling the reproduction of our experiments is published on GitHub[1] and TimeStress is distributed in Hugging Face[2].

2 Related work

This section presents related work to ours, focusing on the study of factual knowledge in LMs, the consideration of their temporal aspect, and their temporal reasoning abilities.

Robustness of factual knowledge in LMs. It has been demonstrated that LMs store a significant amount of factual knowledge (Petroni et al., 2019; Jiang et al., 2020; Sun et al., 2024). However, numerous studies indicate that this acquired knowledge often lacks consistency when faced with textual perturbations. For example, Kassner and Schütze (2020) highlighted the limitations of pretrained LMs in adapting to negations in questions, leading to contradictory answers. Robustness to paraphrasing and minor typographical errors has also been widely studied (Gan and Ng, 2019; von Geusau and Bloem, 2020; Matsuno and Tsuchiya, 2023; Mondal and Sancheti, 2024). Notably, Elazar et al. (2021) and Raj et al. (2022) found that LMs produce different answers for semantically equivalent factual queries. Similarly, Hagen et al. (2024)

discovered that recent LMs can be negatively impacted by minor typographical errors that preserve the original semantics.

Temporal alignment of knowledge in LMs. Since factual knowledge is constantly evolving, studies have been conducted to understand how to adapt LMs to this evolution. As expected, LMs have been shown to be incapable of predicting future facts (Lazaridou et al., 2021), highlighting the need to adapt them to maintain alignment with current knowledge. To address this issue, methods such as continual learning (Liska et al., 2022) and specific pretraining techniques have been proposed, including the joint modeling of text and its associated timestamp to facilitate the acquisition of new temporal knowledge (Dhingra et al., 2022); knowledge editing techniques (Meng et al., 2022; Hartvigsen et al., 2023; Yu et al., 2024; Zhang et al., 2023); or simply externalizing knowledge into an external database accessible by the LM through retrieval-augmented generation (Ram et al., 2023). In parallel, several datasets have been proposed to detect outdated facts in LMs (Zhao et al., 2024; Kim et al., 2024; Margatina et al., 2023; Kasai et al., 2023; Mousavi et al., 2024), and to update LMs' factual knowledge (Ammar Khodja et al., 2024; Yin et al., 2024a; Thede et al., 2025; Ge et al., 2024).

Temporal reasoning in LMs. Several studies have examined the temporal reasoning capabilities of LMs (Zhang and Choi, 2021; Chu et al., 2024; Wei et al., 2023; Fatemi et al., 2025; Dhingra et al., 2022; Xiong et al., 2024; Su et al., 2024). Notably, the works of Chen et al. (2021) and Tan et al. (2023) each proposed a dataset in which LMs are invited to answer questions involving the understanding of the temporality of facts. While these studies share similarities with ours in terms of data (temporally annotated facts), their objectives and methodologies differ. These studies test the mastery of certain temporal logic operators (date calculations, comparisons, etc.) and evaluate the average performance of LMs based on a one-test-per-fact principle. In contrast, we focus not on reasoning ability but on the robustness of knowledge, that is, the ability of an LM to recall the same fact across various temporal contexts.

---

3 Problem Statement and Dataset

The goal of this paper is to measure how robust a Language Model (LM) is to the temporal context associated with a fact. To achieve this, the proposed experimental protocol involves analyzing the LM's preferences when faced with correct or incorrect contexts for the same fact. This section first formalizes this problem and then presents the TimeStress dataset, which instantiates it.

3.1 Problem Statement

Facts and Temporal Contexts. Classically, we consider *facts* as RDF triplets (subject, relation, object), denoted as (s, r, o), where subjects and objects are entities or literals, and relations originate from an ontology (Petroni et al., 2019; Elsahar et al., 2018). When dealing with *temporal facts*, this representation is extended to include a validity period $[a, b]$, as done in other works (Yin et al., 2024b; Jain et al., 2020; Tan et al., 2023). For a quintuple (s, r, o, a, b), the subject s is connected to the object o via the relation r during the period from date a to date b. For example, (Barack Obama, president, USA, 20 January 2009, 20 January 2017) is a temporal fact.

We define the notion of a *temporal context* as a time interval over which we wish to test the validity of a temporal fact. To reduce the number of possibilities and frame our work, we limit these time intervals to either *entire years* (e.g., 1998, i.e., all days of the year 1998), an *entire month* of a given year (e.g., November 1998), or a *specific date* (e.g., November 15, 1998). Subsequently, these three distinct granularities will be denoted as Y for "Year," YM for "Year-Month," and YMD for "Year-Month-Day."

Considering a temporal fact $f = (s, r, o, a, b)$, a temporal context τ is said to be **correct** for f if τ is fully included in $[a, b]$ (i.e., $\tau \subseteq [a, b]$), **incorrect** if it is not included at all ($\tau \cap [a, b] = \varnothing$), or **transitional** otherwise ($\tau \cap [a, b] \neq \varnothing$ and $\tau \not\subseteq [a, b]$). For example, given the validity period $[2017, 2019]$, 2016 is incorrect, 2017 is transitional, and 2018 is correct.

To assess the ability of an LM to distinguish a correct context τ^+ from an incorrect context τ^- for a given temporal fact (s, r, o, a, b), two textual statements are constructed respectively. The form of the statements adopted in our work is that of a question about the fact (s, r, o) followed by its answer ("What is the r of s? o") and prefixed

by a verbalization of the temporal context τ^+ or τ^-. For the example about Barack Obama, two possible contexts are $\tau^+ = 2011$ and $\tau^- = 1998$, producing the statements "*In **2011**, who was the president of the USA? Barack Obama*" and "*In **1998**, who was the president of the USA? Barack Obama.*"

Finally, we say that an LM M distinguishes a correct context from an incorrect context when it assigns a higher probability to the answer o given the statement with τ^+ compared to conditioning on τ^-, i.e., $\Pr_M(o|s, r, \tau^+) > \Pr_M(o|s, r, \tau^-)$. The details of the computation of $\Pr_M$ can be found in Appendix D.

The overall estimation of this ability involves considering a large set of facts with varied entities, relations, and validity periods, and testing numerous pairs (τ^+, τ^-) for each fact. To make the results of these matches interpretable, we impose that the contexts of the same pair have the same granularity (Y, YM, YMD).

Metrics. We introduce two metrics. Given a fact f and a model M, we express the results using a *win rate* $\mathcal{W}(M, f) \in [0, 1]$ of M for f, which is the ratio of the number of times the model preferred a correct context over an incorrect context for the single fact f to the number of tests performed. Additionally, a *robustness* metric, denoted $\mathcal{R}(M, f)$, verifies that correct contexts consistently outperform incorrect ones, defined as: $\mathcal{R}(M, f) = \mathbb{1}[\mathcal{W}(M, f) = 1]$ where $\mathbb{1}[]$ is the indicator function. It is important to note that **transitional contexts are not used in any way for the calculation of these metrics**, as their validity is ambiguous. Given a set of facts, the average win rates and average robustness are denoted $\mathcal{V}(M)$ and $\mathcal{R}(M)$ respectively.

For segmentation purposes in the analyses, these global metrics can be restricted to tests conducted with temporal contexts of a specific granularity (Y, YM, or YMD).

Finally, to measure the distance of a context τ relative to the validity period $[a, b]$ of a fact, we calculate its *relative position*, denoted α, as the number of days between the midpoint of $[a, b]$ and the midpoint of τ, divided by the number of days in $[a, b]$. Thus, $|\alpha| < \frac{1}{2}$ for correct contexts, and $|\alpha| > \frac{1}{2}$ for incorrect contexts. For transitional contexts, the value $|\alpha|$ is explicitly set to $\frac{1}{2}$.

3.2 The TimeStress Dataset

We present the TimeStress dataset, which enables our study. This dataset contains over 521,000 statements (in the form of questions) generated from 2,003 temporal facts, covering 1,883 unique entities (1,385 unique subjects and 1,113 unique objects) and 86 relations. A brief sample is provided in Table 1.

On average, each fact is associated with 11 correct temporal contexts and 74 incorrect ones, distributed across the three granularities Y, YM, and YMD. **There are enough correct and incorrect contexts to make it nearly impossible for a random model to be robust on any fact by chance.**

In what follows, we briefly introduce how TimeStress was built, covering the quintuplet collection from Wikidata, their verbalization in natural language using GPT-4o, and how incorrect and correct contexts were sampled for each quintuplet in order to create statements.

A more detailed version of this section can be found in Appendix A.

3.2.1 Quintuplet Collection

The quintuplet collection process begins with a preprocessed version of Wikidata provided in Ammar Khodja et al. (2025). This source also provides a measure of each entity's popularity, defined as the median number of human visits per month to the Wikipedia article associated with the entity in 2020. This measure is used to define the popularity index of a quintuplet, calculated as the geometric mean of the popularity of its object and subject. Although the popularity of the subject and object does not imply the popularity of the fact, this index remains an interesting tool for finding facts "known" by LMs, as it is shown empirically in the experiments.

We collect and filter Wikidata facts following this procedure: (1) All quintuplets with a validity period (i.e., a start or end date mentioned) and whose objects are not literals, such as quantities and dates, are collected. (2) Quintuplets valid within two distinct periods are removed to simplify result analysis, as this allows all dates outside the validity period to be considered incorrect. (3) Quintuplets without a delimited validity period (i.e., a start AND end date mentioned) are removed. (4) Only quintuplets that were **valid prior to 2021** are retained, as this ensures that all these quintuplets are past facts for all studied LMs. (5) Only the quintuplets that are valid for **longer than three**

years are retained to ensure a minimal number of correct temporal contexts of Y granularity. (6) We keep only the most popular quintuplets using the popularity index. This results in a set of 2,098 quintuplets with a varied set of 86 relations.

3.2.2 Quintuplet Verbalization

The process of generating statements from quintuplets is carried out using GPT-4o. First, a prompt instructs GPT-4o to generate four linguistically diverse questions from a given tuple (subject, relation, object, year), with the following guidelines: the question must be in the past tense, begin with "In [YEAR],", be stated in a simple and concise manner without any detail that could give clues about the answer. It should be directly followed by the answer, which is the object. The quality of the generated questions was analyzed to identify and eliminate incorrect entries. Initially, out of the 2,098 facts intended for verbalization, 53 failed, and 64 questions mistakenly used the subject as the answer instead of the object. These erroneous cases were removed from the dataset, resulting in a total of 2,003 facts and $2003 \times 4 = 8012$ questions. A random sample of 50 questions was manually evaluated to ensure the overall quality of the generated questions. The evaluation revealed that only 1 out of 50 questions was incorrect, while the remaining questions were perfectly constructed (Wilson confidence interval at 95% = [0.85, 0.99]), which demonstrates the high quality of the questions in our dataset. Finally, the temporal context was removed and **each fact is randomly assigned one of its four associated questions**.

3.2.3 Context Sampling

For each fact, based on its validity interval $[a, b]$, centered on $m = \frac{a+b}{2}$ and of duration $d = b - a$[3], temporal contexts at the Y granularity are uniformly sampled over the wider interval $[m-5d, m+5d]$ with a step of $0.05 \times d$. From these Y-granularity contexts, YM-granularity contexts are generated by randomly selecting a month. Similarly, YMD-granularity contexts are determined by choosing a random day from each YM-granularity context[4]. This process creates a hierarchy among contexts derived from the same year for a given fact. Note that when a date d_2 is chosen from a higher-granularity date d_1, it is necessarily correct

---

[3]The median of dates (in day precision) is used to perform arithmetic operations between dates.

[4]This sampling does not produce erroneous dates such as February 29 for non-leap years, or April 31.

(or incorrect) if d_1 is. However, d_1 may be transitional while d_2 is correct or incorrect. In such cases, d_2 is excluded from the set of correct or incorrect dates. This **guarantees** that the number of correct and incorrect contexts does not vary by granularity, avoiding bias when comparing model robustness across granularities. The corresponding years of the produced contexts are mainly located in the contemporary period between 1800 and 2020 (Appendix E), because the popularity index used to select the facts in TimeStress draws more often recent facts. To produce the statements of each fact that will be used to compute the metrics, its corresponding statement is prefixed with the previously sampled temporal contexts associated with the fact.

4 Experimentation

This section details our experiments on the TimeStress dataset. As a reminder, our objectives are, in order, to measure the ability of models to distinguish correct and incorrect temporal contexts, analyze their robustness, and search for anomalies in this task when incorrect contexts are closer to or farther from the validity interval, and as the granularity of contexts becomes finer.

Numerous models from different families and sizes were tested: `Mistral-Nemo-Base-2407`, `Mistral-7B-v0.3` (Jiang et al., 2023); `OpenEML-{450M, 3B}` (Mehta et al., 2024); `gemma-2-{2b, 9b, 27b}` (Team et al., 2024); `Llama-3.1-{8B, 70B}` (Grattafiori et al., 2024). For each, both pretrained and instruction-tuned versions were considered, resulting in a total of 18 studied LMs. All models were sourced from *huggingface.co*.

In the first series of experiments, the statements were passed to the models as raw text rather than as instructions to enable the comparison between pretrained and instruction-tuned models. The use of an "instruction/message" format is explored in a second phase.

4.1 Overall Mastery of Temporal Contexts

Figure 2a shows the average win rate for the facts in TimeStress for the top 5 LMs and for each temporal granularity Y, YM, and YMD, as well as for their union. Results for other models are reported in Appendix E.

Overall, the results show that these top 5 LMs generally distinguish correct statements from incorrect ones well, with win rates ranging from 78% to 87%. Among our other findings, we observed that even smaller models (<500M parameters) perform better than chance, and the win rate logically improves with model size (Appendix E), with the best model being the largest, Llama-3.1-70B-Instruct.

Figure 3 provides a more detailed analysis by reporting the average $\log \Pr(o|f, \tau)$ as a function of the value α, which quantifies the relative distance of τ from the validity period of f (see Section 3.1). The average is calculated across all facts, for contexts at the year granularity, and across all 18 studied LMs. We observe that the highest probabilities correspond to contexts within the validity interval ($\alpha \in [-0.5, 0.5]$), while outside this interval, probabilities gradually decrease as $|\alpha|$ increases. Finally, we note that the probability assigned to transitional contexts (years that are neither fully correct nor fully incorrect) is significantly higher (based on the confidence intervals (CIs)) than that for incorrect contexts. We explain this phenomenon with the following hypothesis: in the training data of LMs, transitional years are more often associated with the considered fact than other years within the validity period, as they correspond to key events such as the beginning and end of the fact (e.g., the start or end year of a presidential term).

This strong alignment of LMs with the validity period of temporal facts leads us to conclude that LMs possess at least a basic representation of temporality.

4.2 Robustness and Anomalies

The temporal representation of LMs is not robust. Figure 2b shows the average robustness of the top 5 models across all facts in TimeStress. As a reminder, this metric is stricter and does not tolerate any error during matches for a given fact. Results for other models are reported in Appendix E.

The scores are generally low, indicating that win rates per fact rarely reach 100%. Interestingly, the most robust model is not the one with the highest win rate. The most robust model, gemma-2-27b-it, achieves an $\mathcal{R}$ value of only about 17% for the coarsest granularity Y. This score drops to 11% when all granularities are considered. Most other models do not exceed a global robustness score of 3%. Among our other results, we also observed that instruction-tuned models mostly outperform their pre-trained counterparts. A notable case is the Llama-3.1-70B-Instruct model; although it was fine-tuned on instructions, it is $3.6\times$ more robust than its pre-trained counterpart, Llama-3.1-70B.

Temporal fact	Temp. Cont.	Status	Statement
(Betty Ford, spouse, Gerald Ford, 1948-10-15, 2006-12-26)	1983-03-21	Correct	On March 21, 1983, who was the spouse of Betty Ford? Gerald Ford
(Beirut, country, Ottoman Empire, 1520, 1918)	1759-05	Correct	In May 1759, to which sovereign state did Beirut belong? Ottoman Empire
(Jimmy Butler, member of sports team, Chicago Bulls, 2011, 2017-06-22)	1989-06-17	Incorrect	On June 17, 1989, which basketball team did Jimmy Butler belong to? Chicago Bulls
(Samarkand, country, Soviet Union, 1922-12-30, 1991-08-31)	1789-03-31	Incorrect	On March 31, 1789, what was the sovereign state of Samarkand? Soviet Union
(United States of America, head of government, Andrew Johnson, 1865-04-15, 1869-03-04)	1865	Transitional	In 1865, who served as the head of government for the United States of America? Andrew Johnson
(Chris Evans, unmarried partner, Minka Kelly, 2007-05, 2014-10)	2014	Transitional	In 2014, who was Chris Evans romantically involved with? Minka Kelly

Table 1: Random sample of statements generated from various facts and temporal contexts in TimeStress.
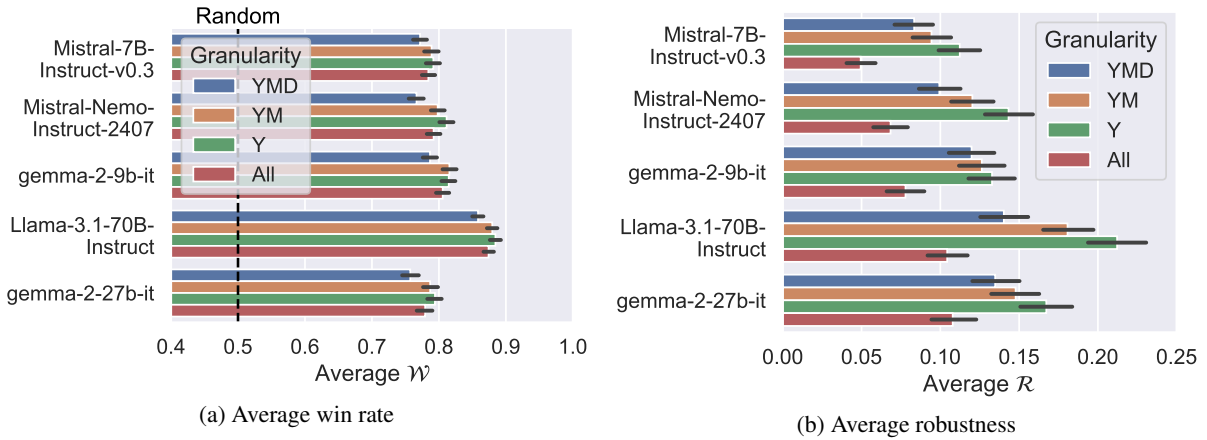

(a) Average win rate

(b) Average robustness

Figure 2: Average metrics on the TimeStress dataset for the 5 most robust models (95% CIs were determined using bootstrapping).

This suggests that the training data and possibly the training procedure play an important role in temporal robustness. Finally, early signs of failure in knowledge transfer between granularities are evident due to the substantial gap between individual robustness scores for granularities and the global score. This issue is explored in detail later in this section.

LMs are vulnerable to *easy* incorrect contexts. Table 4 investigates the impact of the relative positions of incorrect contexts of granularity Y, focusing on cases where incorrect contexts cause an LM to fail in a match for facts that seem "known" to the LM, as indicated by a very high win rate ($\mathcal{W} \geq 95\%$). For now, only the "Raw Text" column is of interest. The table reveals that these incorrect contexts are not entirely concentrated around the validity period, as might reasonably be expected. Instead, a significant proportion is located far from it. Specifically, LMs fail to achieve robustness due
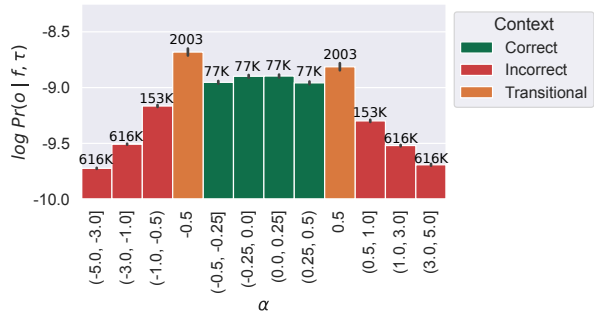

Figure 3: Evolution of $\log \Pr(o|f, \tau)$ with respect to the relative distance α, averaged across all facts in TimeStress and all LMs, for granularity Y (Bootstrap 95% CIs). The number of points used to compute each bar is indicated above it.

to contexts with a distance of $|\alpha| \geq 1$ in 19% of cases. This proportion decreases to 6% for $|\alpha| \geq 3$, which remains significant given the proximity of the win rate to 100% for the facts observed here. We conducted the same analysis using win rate

$|\alpha|$	Raw Text	Instruction
≥ 1	0.19 (±0.01)	0.25 (±0.01)
≥ 2	0.09 (±0.01)	0.13 (±0.01)
≥ 3	0.06 (±0.01)	0.08 (±0.01)
≥ 4	0.04 (±0.01)	0.05 (±0.01)

Figure 4: Proportion of incorrect dates favored over correct dates beyond a relative distance $|\alpha|$, when the win rate exceeds 95% (Wilson's 95% CIs).

thresholds higher than 95% (see Appendix B). As the threshold approaches 100%, vulnerability to "easy" incorrect dates gradually decreases but never completely disappears. Even when the win rate threshold is 99%, errors remain when $|\alpha| \geq 4$. We conclude that this vulnerability is inherent to current LMs. While the probabilistic nature of these models may provide a tangible explanation, this behavior is clearly undesirable, as these are typically errors that a human would not make when aware of a fact's validity period.

These conclusions hold for the instruction format. So far in our experiments, all models have been fed statements in Raw text rather than instructions. Since the performance of instruction-tuned LMs might have been underestimated, win rates and robustness scores were recalculated using an "instruction/message" format[5]. Figure 5 compares robustness scores calculated for the two formats. On average, robustness decreases with the use of the "instruction" format (notably for gemma-2 models), and global robustness scores remain low. However, no clear conclusions emerge regarding the positive or negative impact of this format, as the effect varies significantly across models. Next, the "Instruction" column of Table 4 complements our previous analysis on the impact of the relative position of incorrect contexts for high win-rate facts. This time, the "instruction" format degrades performance with more critical errors (i.e., far from the validity period). Based on the confidence intervals, these differences are statistically significant for all values of $|\alpha|$ studied. Examples of these critical errors are shown in Appendix E.

LMs fail to perfectly propagate their knowledge across granularities. We examine the ability of

---

[5]This involves constructing messages and injecting them into the chat template of each LM, as in the following example: {user: "In 2011, who was the president of the USA?", assistant: "Barack Obama"}.
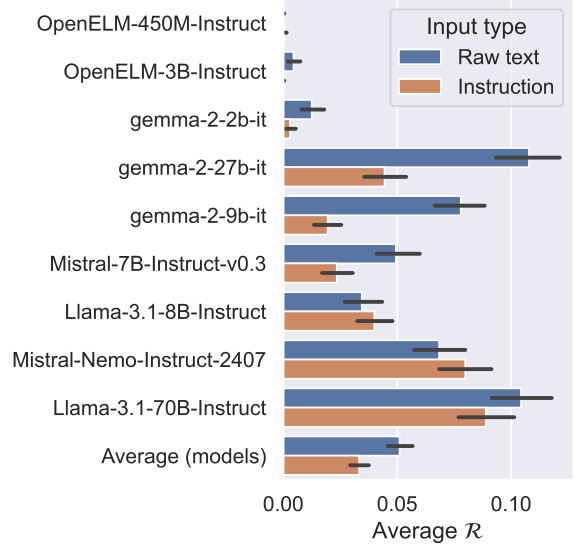

Figure 5: Average $\mathcal{R}$ across all granularities for facts in TimeStress based on the format of statements submitted to the models: raw text (blue) or instruction (orange). 95% CIs were determined using bootstrapping.
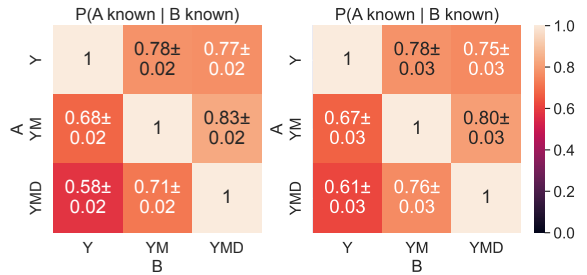

Figure 6: Average success rate of knowledge transfer between granularity pairs for the 5 most robust LMs with queries in **raw text (left)** or **instructions (right)**. Wilson confidence intervals at 95% are shown.

LMs to propagate knowledge of a fact across different temporal granularities. TimeStress allows comparisons between two granularities because the three studied granularities have the same number of correct and incorrect contexts for all temporal facts. The only difference between two granularities is the addition of a random month and/or day, which does not affect validity when transitioning from a lower granularity to a higher granularity (e.g., from Y to YM). For example, if a fact is incorrect for an entire year, it remains incorrect for any month or date within that year.

We consider a fact f to be "known" for a granularity by a model M if $\mathcal{R}(M, f) = 1$. This definition can apply to a given granularity. For example, a fact is "known" at the Y granularity if all matches with temporal contexts at the year granularity were won. For each of the 5 most robust LMs and for

each pair of granularities (A, B), we then calculate the proportion of facts that are "known" at granularity A, given that they are "known" at granularity B.

Figure 6 reports this transfer proportion from granularity B to A for the "raw text" format (left) and "instruction" format (right). On average, for the "raw text" format, LMs failed to generalize their knowledge to other granularities in 28% of cases (1 - average of all non-diagonal cells), which is surprisingly high given their perfect score on the starting granularity Y. Details for each model are available in Appendix C.2. Performance varies across LMs. For example, for the most robust model, gemma-2-27b-it, the transition from $B = Y$ to $A = YM$ is successful in 74±5% of cases, and the win rates for other transitions range between 68±6% and 88±5%. The general trend is that LMs fail more in transitions from coarse to fine granularities. No LM achieves perfect transitions for any pair of granularities. There are slight variations between the instruction (Figure 6, right) and raw formats, but the average success rate is nearly identical.

The possibility that poor knowledge propagation between granularities could be due to LMs' ignorance of the validity period boundaries[6]. This was confirmed in a similar analysis that takes context position into account (Appendix C.1). Indeed, consistency between granularities approaches perfect consistency as the context moves away from the validity period. However, perfect consistency is **never reached**; which reminds us of the vulnerability of LMs to *easy* incorrect contexts.

For exploratory purposes, we investigated whether including explanations about temporal concepts in the LMs' prompts could help them better transfer knowledge from one temporal granularity to another. To evaluate this, two prompts were prefixed to each TimeStress statement. The first explains the hierarchical nature of dates (i.e., a year consists of months, and a month consists of days), while the second is more direct and explains how knowledge of a temporal fact can be generalized from one granularity to another. Details of these prompts are provided in Appendix C.3. We recalculated the transfer proportions between granularities using the same 5 LMs as in Figure 6. The two explanatory prompts improved generalization in the "raw text" format from 73% to 76%. However, no substantial gain compared to not using an explanatory prompt was observed when using the "instruction" format.

Other observations. There is a positive correlation between the popularity of a fact and the robustness and win rate of LMs on it. Interestingly, LMs are robust on globally different facts. Indeed, a pair of LMs shares, on average, 11% of facts on which they are robust. This proportion reaches 31% when limited to the 5 most robust LMs. However, only 34 facts out of 384 (8.9%) are robust at the same time in these LMs. Furthermore, the longer a fact's validity period, the higher the win rate (on the 5 most robust LMs). This statistically significant correlation[7] is intriguing because it appears that the difficulty of situating a fact in time is the same whether it has a duration of 3 years or 30 years. One possible explanation is that facts with longer validity periods are more stable and unique (i.e., there are no alternative objects "o" for the same subject-relation pair "s,r"), so LMs can learn them without confusion or contradiction. However, this explanation is contradicted by another observation: when there are more alternative objects "o" for a given (s,r) pair, the win rate and robustness actually increase, not decrease. This contradiction raises the question of how to explain the observed phenomenon. Finally, the further a fact's validity period is from the present, the less robust the LMs are on it, with lower win rates as well. More details are in Appendix E.

5 Experimental Protocol: Motivations

There are seemingly more "natural" approaches for probing factual knowledge in language models, such as the evaluation protocols used in LAMA (Petroni et al., 2019), TriviaQA (Joshi et al., 2017), KAMEL (Kalo and Fichtel, 2022), and BEAR (Wiland et al., 2024). Instead of comparing probabilities across several temporal contexts, one could ask the LM to answer temporally contextual questions such as "In 2011, who was the president of the US?", and evaluate the LM based on the generated answers. However, our experimental protocol was preferred for several reasons.

First, our setup–where the LM must distinguish between statements with correct and incorrect temporal contexts by assigning probabilities–allows to target specific facts without ambiguity, even in the case of non-functional relations, such as "shares a

---

[6]In this case, robustness was achieved only by chance.

[7]The null hypothesis is the absence of correlation.

border with," where a subject-relation pair can have multiple valid objects. In generation-based settings, an LM may produce one or several correct answers, or even off-topic outputs, making evaluation less reliable and direct comparison across LMs more difficult. This is especially true given that classical generation-based metrics, such as ROUGE (Lin, 2004), can underestimate performance. Sometimes, the set of all correct answers is difficult to enumerate due to the vagueness of the relation (e.g., does asking for the borders of a country include continents and oceans?) and due to the sometimes large number of ways of expressing an answer.

Additionally, our evaluation protocol is efficient and scalable, as it does not require generation or answer validation.

Given the imperfections of other evaluation protocols, it would have been difficult to defend our claims–especially those involving sensitive metrics like robustness and the study of rare LM errors–if our results could be attributed to limitations of the evaluation method itself.

6 Conclusion

This study examined the robustness of LMs to simple temporal variations in factual knowledge. It assessed their ability to distinguish correct from incorrect temporal contexts based on two factors: the distance of contexts from the validity period of facts and their granularity. To facilitate this, the TimeStress dataset was introduced, featuring high-quality statements on popular temporal facts from Wikidata (according to a popularity index) and enabling the evaluation of 18 LMs of varying sizes and families. The results revealed that the best-performing LM was robust for only 11% of the studied facts, exhibiting errors, certainly rare, but critical that are uncommon to humans, which we frame as anomalies. These errors consist of a susceptibility to *easy* incorrect contexts and imperfect knowledge generalization across granularities. Notably, these findings held true regardless of whether the LM was pretrained or instruction-tuned, and whether the statements were presented in an instruction or raw format. This highlights the limits of current LMs in temporal representation. It is worth noting that since the studied temporal facts are relatively popular, these results likely represent an upper bound of LMs' performance on the general population of facts, given the strong link between knowledge popularity and its likelihood of

being learned by LMs (Kandpal et al., 2023; Kang and Choi, 2023).

Limitations

The study evaluates LMs using a probability-based approach to assess their understanding of temporal facts. While this method does not fully capture model performance in text generation scenarios, it is strongly related, as generated text is sampled from the LM's probability distribution. Additionally, prior research has shown that probability-based metrics correlate reasonably well with the generative performance of models in factual knowledge evaluation contexts, where the model is expected to generate specific entities (Dong et al., 2023; Lyu et al., 2024) as an answer, which is closely aligned with our experimental protocol. The advantage of our approach compared to generation metrics is that it allows for precise exploration of specific non-functional relations where multiple correct answers exist. This is more challenging with generation-based metrics, as LMs may produce another correct answer, unexpected responses, or off-topic outputs.

Second, the results of our study are limited to the format of the statements we chose, i.e., a temporal context followed by a question and an answer. It is possible that LMs would perform better in a different format. However, their current limitations on our data are already problematic.

Finally, the TimeStress dataset consists of statements in English, which may limit the applicability of our results to other languages due to potential linguistic differences that could affect temporal understanding. However, future research can easily expand the scope by adapting the GPT-4o prompt used to generate statements to target additional languages. As for entity labels, they are available in other languages in Wikidata.

References

Hichem Ammar Khodja, Abderrahmane Ait gueni ssaid, Frederic Bechet, Quentin Brabant, Alexis Nasr, and Gwénolé Lecorvé. 2025. Factual knowledge assessment of language models using distractors. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8043–8056, Abu Dhabi, UAE. Association for Computational Linguistics.

Hichem Ammar Khodja, Frédéric Béchet, Quentin Brabant, Alexis Nasr, and Gwénolé Lecorvé. 2024. WikiFactDiff: A large, realistic, and temporally adaptable

dataset for atomic factual knowledge update in causal language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17614–17624, Torino, Italia. ELRA and ICCL.

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. 2024. TimeBench: A comprehensive evaluation of temporal reasoning abilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1204–1228, Bangkok, Thailand. Association for Computational Linguistics.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. Time-aware language models as temporal knowledge bases. *Trans. Assoc. Comput. Linguistics*, 10:257–273.

Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Zhifang Sui, and Lei Li. 2023. Statistical knowledge assessment for large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard H. Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Trans. Assoc. Comput. Linguistics*, 9:1012–1031.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-REx: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Bahare Fatemi, Mehran Kazemi, Anton Tsitsulin, Karishma Malkan, Jinyeong Yim, John Palowitch, Sungyong Seo, Jonathan Halcrow, and Bryan Perozzi. 2025. Test of time: A benchmark for evaluating LLMs on temporal reasoning. In *The Thirteenth International Conference on Learning Representations*.

Wee Chung Gan and Hwee Tou Ng. 2019. Improving the robustness of question answering systems to question paraphrasing. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6065–6075. Association for Computational Linguistics.

Xiou Ge, Ali Mousavi, Edouard Grave, Armand Joulin, Kun Qian, Benjamin Han, Mostafa Arefiyan, and Yunyao Li. 2024. Time sensitive knowledge editing through efficient finetuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024 - Short Papers, Bangkok, Thailand, August 11-16, 2024*, pages 583–593. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, and Abhinav Jauhri et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Tim Hagen, Harrisen Scells, and Martin Potthast. 2024. Revisiting query variation robustness of transformer models. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 4283–4296. Association for Computational Linguistics.

Tom Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. 2023. Aging with GRACE: lifelong model editing with discrete key-value adaptors. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Prachi Jain, Sushant Rathi, Mausam, and Soumen Chakrabarti. 2020. Temporal Knowledge Base Completion: New Algorithms and Evaluation Protocols. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3733–3747, Online. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know. *Trans. Assoc. Comput. Linguistics*, 8:423–438.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Jan-Christoph Kalo and Leandra Fichtel. 2022. KAMEL: knowledge analysis with multitoken entities in language models. In *4th Conference on Automated Knowledge Base Construction, AKBC 2022, London, UK, November 3-5, 2022*.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 15696–15707. PMLR.

Cheongwoong Kang and Jaesik Choi. 2023. Impact of co-occurrence on factual knowledge of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7721–7735, Singapore. Association for Computational Linguistics.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. 2023. Realtime QA: what's the answer right now? In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Yujin Kim, Jaehong Yoon, Seonghyeon Ye, Sangmin Bae, Namgyu Ho, Sung Ju Hwang, and Se-Young Yun. 2024. Carpe diem: On the evaluation of world knowledge in lifelong language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 5401–5415. Association for Computational Linguistics.

Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Giménez, Cyprien de Masson d'Autume, Tomás Kociský, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. 2021. Mind the gap: Assessing temporal generalization in neural language models. In *Neural Information Processing Systems*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Adam Liska, Tomás Kociský, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien de Masson d'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, Ellen Gilsenan-McMahon, Sophia Austin, Phil Blunsom, and Angeliki Lazaridou. 2022. Streamingqa: A benchmark for adaptation to new knowledge over time in question answering models. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 13604–13622. PMLR.

Chenyang Lyu, Minghao Wu, and Alham Aji. 2024. Beyond probabilities: Unveiling the misalignment in evaluating large language models. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 109–131, Bangkok, Thailand. Association for Computational Linguistics.

Katerina Margatina, Shuai Wang, Yogarshi Vyas, Neha Anna John, Yassine Benajiba, and Miguel Ballesteros. 2023. Dynamic benchmarking of masked language models on temporal concept drift with multiple views. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 2873–2890. Association for Computational Linguistics.

Takumi Matsuno and Masatoshi Tsuchiya. 2023. Evaluating the robustness of question answering model against context variations. In *2023 10th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA)*, pages 1–6.

Sachin Mehta, Mohammad Sekhavat, Qingqing Cao, Max Horton, Yanzi Jin, Frank Sun, Iman Mirzadeh, Mahyar Najibikohnehshahri, Dmitry Belenko, Peter Zatloukal, and Mohammad Rastegari. 2024. Openelm: An efficient language model family with open training and inference framework. In *ICML Workshop*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Ishani Mondal and Abhilasha Sancheti. 2024. On the robustness of chatgpt under input perturbations for named entity recognition task. In *The Second Tiny Papers Track at ICLR 2024, Tiny Papers @ ICLR 2024, Vienna, Austria, May 11, 2024*. OpenReview.net.

Seyed Mahed Mousavi, Simone Alghisi, and Giuseppe Riccardi. 2024. Dyknow: Dynamically verifying time-sensitive factual knowledge in llms. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 8014–8029. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Harsh Raj, Domenic Rosati, and Subhabrata Majumdar. 2022. Measuring reliability of large language models through semantic consistency. In *NeurIPS ML Safety Workshop*.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Trans. Assoc. Comput. Linguistics*, 11:1316–1331.

Zhaochen Su, Juntao Li, Jun Zhang, Tong Zhu, Xiaoye Qu, Pan Zhou, Yan Bowen, Yu Cheng, and Min Zhang. 2024. Living in the moment: Can large language models grasp co-temporal reasoning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13014–13033, Bangkok, Thailand. Association for Computational Linguistics.

Kai Sun, Yifan Ethan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2024. Head-to-tail: How knowledgeable are large language models (llms)? A.K.A. will llms replace knowledge graphs? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 311–325. Association for Computational Linguistics.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023. Towards benchmarking and improving the temporal reasoning capability of large language models. In *Annual Meeting of the Association for Computational Linguistics*.

Gemma Team, Morgane Riviere, and Shreya Pathak et el. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Lukas Thede, Karsten Roth, Matthias Bethge, Zeynep Akata, and Tom Hartvigsen. 2025. Understanding the limits of lifelong knowledge editing in llms. *Preprint*, arXiv:2503.05683.

Paulo Alting von Geusau and Peter Bloem. 2020. Evaluating the robustness of question-answering models to paraphrased questions. In *Artificial Intelligence and Machine Learning - 32nd Benelux Conference, BNAIC/Benelearn 2020, Leiden, The Netherlands, November 19-20, 2020, Revised Selected Papers*, volume 1398 of *Communications in Computer and Information Science*, pages 1–14. Springer.

Yifan Wei, Yisong Su, Huanhuan Ma, Xiaoyan Yu, Fangyu Lei, Yuanzhe Zhang, Jun Zhao, and Kang Liu. 2023. Menatqa: A new dataset for testing the temporal comprehension and reasoning abilities of large language models. In *Conference on Empirical Methods in Natural Language Processing*.

Jacek Wiland, Max Ploner, and Alan Akbik. 2024. BEAR: A unified framework for evaluating relational knowledge in causal and masked language models.

In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 2393–2411. Association for Computational Linguistics.

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024. Large language models can learn temporal reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10452–10470, Bangkok, Thailand. Association for Computational Linguistics.

Xunjian Yin, Jin Jiang, Liming Yang, and Xiaojun Wan. 2024a. History matters: Temporal knowledge editing in large language model. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 19413–19421. AAAI Press.

Xunjian Yin, Jin Jiang, Liming Yang, and Xiaojun Wan. 2024b. History matters: Temporal knowledge editing in large language model. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 19413–19421. AAAI Press.

Lang Yu, Qin Chen, Jie Zhou, and Liang He. 2024. MELO: enhancing model editing with neuron-indexed dynamic lora. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 19449–19457. AAAI Press.

Michael Zhang and Eunsol Choi. 2021. SituatedQA: Incorporating extra-linguistic contexts into QA. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7371–7387, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zihan Zhang, Meng Fang, Ling Chen, Mohammad-Reza Namazi-Rad, and Jun Wang. 2023. How do large language models capture the ever-changing world knowledge? a review of recent advances. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8289–8311, Singapore. Association for Computational Linguistics.

Bowen Zhao, Zander Brumbaugh, Yizhong Wang, Hanna Hajishirzi, and Noah A. Smith. 2024. Set the clock: Temporal alignment of pretrained language models. In *Annual Meeting of the Association for Computational Linguistics*.

A   TimeStress: Details of the Construction Process

This section provides a detailed description of the construction process for the TimeStress dataset. Before discussing the collection process, we describe the main characteristics of TimeStress.

First, the dataset focuses on past facts valid strictly before 2021, ensuring that they are historical (not valid at the present) events for all recent LMs. TimeStress includes high-quality statements that are consistent with the facts and exhibit linguistic diversity to avoid biases stemming from a limited variety of questions. The statements are carefully selected to minimize typographical errors, verbs are systematically conjugated in the past tense, and future dates beyond 2020 are excluded to avoid absurd questions such as "*In 2052, who was the president of the USA?*". The dataset covers a diverse set of 86 relations to reduce biases associated with a restricted range. The targeted facts are popular, essential for evaluating the generalization of knowledge across different granularities—a task that becomes challenging if the LMs are unfamiliar with the facts. All facts are valid over a single validity period, ensuring that all contexts outside the validity period can be considered incorrect. Additionally, to ensure fairness, each granularity (Y, YM, YMD) has an equal number of correct and incorrect temporal contexts for all facts. Finally, the number of correct and incorrect contexts is sufficiently large to make it nearly impossible for a random model to be robust on any fact by chance.

The creation process for the TimeStress dataset was carefully designed to meet the properties described above, thereby effectively supporting the claims of this paper. This process consists of three main steps. First, an initial collection of 2,098 temporal facts is performed from Wikidata for inclusion in TimeStress. Second, questions are generated from these quintuplets using GPT-4o, accompanied by a quality evaluation to ensure high-quality questions. Finally, for each fact, correct and incorrect temporal contexts are identified and integrated into the questions to produce statements.

A.1   Quintuplet Collection Process

The process of collecting quintuplets begins with the post-processed version of Wikidata provided by (Ammar Khodja et al., 2025).

This source also provides a measure of an entity's popularity, defined as the median number of human visits to the Wikipedia article associated with that entity during the year 2020. This measure is used to define the popularity of a quintuplet, calculated as the geometric mean of the popularity of its object and subject. Figure 13 demonstrates the effectiveness of this popularity measure in identifying facts on which LMs are robust, illustrating that the likelihood of the robustness of LMs on a fact increases with its popularity.

Initially, all quintuplets with at least a start or end date and whose objects are not literals, such as quantities and dates, are collected, totaling over 2.1 million quintuplets. The quintuplets are then filtered to remove any (s, r, o, a, b) where another quintuplet (s, r, o, a', b') exists with a different validity period $[a', b']$, allowing us to assume that all dates outside $[a', b']$ are incorrect, which simplifies result analysis. This step eliminates a negligible amount of quintuplets (6.23%). Additionally, quintuplets without a start or end date are removed as their validity period is unbounded.

Only quintuplets with a popularity measure of at least 90,000[8] and a validity period strictly longer than three years are retained.

The final result is a dataset comprising the 2,098 most popular facts from Wikidata (according to the popularity index), with 1,910 unique entities, 1,435 unique subjects, 1,151 unique objects, and 86 relations, forming a well-diversified set of temporal facts.

A.2   Quintuplet Verbalization

The process of verbalizing quintuplets into natural language questions is carried out using GPT-4o. The prompt, adapted from Ammar Khodja et al. (2024) (Appendix B), was modified to generate questions instead of declarative sentences. The adapted *system* prompt instructs GPT-4o to take a tuple (subject, relation, object, timestamp) and generate four linguistically diverse questions. For example, for the input (`British India, capital, Kolkata, 1929`), a possible question could be: "*In 1929, what was the capital of British India? Kolkata*". The questions must adhere to specific guidelines: they must be in the past tense, begin with the year followed by a comma, and end with the answer. The questions should focus on the object, be simple and concise, and avoid any detail that could simplify the answer.

---

[8]This threshold was determined by gradually lowering the threshold from 150,000 in steps of 10,000 until the number of retrieved facts exceeded 2,000.

Here is the *system* prompt used:

```
You are an advanced knowledge verbalization system.
You take as input a knowledge quadruple (subject,
relation, object, time) and generate a list of 4
linguistically diverse questions on the quadruple.
For example, the input could be : (British India,
capital, Kolkata, 1929) and one of your questions may
be : "In 1929, what was the capital of British India?
Kolkata.".

All the questions you generate must be in past tense
because the facts are not valid anymore.
The questions must always start with the year, then a
comma, then the question itself, and then finally the
answer.
The questions must always be asked on the object.
The questions must be straightforward and concise.
The questions must not contain details that could make
them easier to answer.

Examples of questions:
- (Jimmy Butler, member of sports team, Chicago Bulls,
2014) -> "In 2014, which team did Jimmy Butler play
for? Chicago Bulls."
- (Philippines, head of state, Emilio Aguinaldo, 1900)
-> "In 1900, who was the head of state of Philippines?
Emilio Aguinaldo."
- (Coretta Scott King, spouse, Martin Luther King Jr.,
1960) -> "In 1960, who was Coretta Scott King married
to? Martin Luther King Jr."
- (European Union, currency, pound sterling, 2002)
-> "In 2002, what was one of the currencies of the
European Union? Pound sterling."
```

And here is the main prompt:

```
Here is the knowledge quadruple to verbalize:
([SUBJECT], [RELATION], [OBJECT], [YEAR]).

Due to the ambiguity that could arise from the provided
labels, here is their meaning:
- (subject) "[SUBJECT]" : "[SUBJECT_DESC]"
- (relation) "[RELATION]" : "[RELATION_DESC]"
- (object) "[OBJECT]" : "[OBJECT_DESC]"

Finally, here is an example where the relation
"[RELATION]" is employed :    ([EXAMPLE_SUBJECT],
[RELATION], [EXAMPLE_OBJECT]).
```

To use this main prompt, placeholders [SUBJECT], [RELATION], [OBJECT], [SUBJECT_DESC], [RELATION_DESC], and [OBJECT_DESC] are filled with the corresponding labels and descriptions from Wikidata. An example of the relation is also retrieved from Wikidata using the property *Wikidata property example (P1855)*. If no example is available, the last line of the main prompt is omitted. The year [YEAR] is selected as the midpoint of the quintuplet's validity period. GPT-4o then generates four questions and answers for each quintuplet. Next, the temporal context is removed from the question, and it is verified that the answer matches the object.

A.3 Quality of Generated Questions

The quality of the generated questions was analyzed to identify and eliminate incorrect entries. Initially, out of the 2,098 facts intended for ver-

balization, 53 failed, and 64 questions mistakenly used the subject as the answer instead of the object. These erroneous cases were removed from the dataset, resulting in a total of 2,003 facts and $2003 \times 4 = 8012$ questions.

A random sample of 50 questions was manually evaluated to ensure the overall quality of the generated questions. The evaluation revealed that only 1 out of 50 questions was incorrect, while the remaining questions were perfectly constructed (Wilson confidence interval at 95% = [0.85, 0.99])[9]. These results demonstrate the high quality of the questions in our dataset.

Finally, each fact is randomly assigned one of its four associated questions.

A.4 Test Generation

Arithmetic operations between temporal contexts are involved in this section. It is important to note that all operations between contexts are performed on the midpoint of the context (as the contexts studied are intervals). For example, when $a + b$ is calculated, the result is the midpoint of a added to the midpoint of b. The finest granularity a *midpoint* can have is the YMD granularity (i.e., Year-Month-Day). This approach bypasses the interval nature of dates.

For each quintuplet, the range of tested contexts is defined as $m \pm 5d$, where m is the midpoint of the validity period $(a + b)/2$, and d is the duration of the validity period $b - a$. To determine the dates of granularity Y (i.e., Year) to include in TimeStress, we perform an analysis starting from the midpoint and extending to the boundaries with a step size of $0.05 \times d$. This step size is chosen to limit the maximum number of correct and incorrect contexts to reasonable values of 21 and 180, respectively.

For each context of granularity Y, a context of granularity YM is chosen by randomly selecting a month within the year. Similarly, for each context of granularity YM, a context of granularity YMD is chosen by randomly selecting a day within the previously selected YM context[10]. This creates a hierarchical relationship between the different granularities (e.g., *2020, 2020-03, 2020-03-24*), enabling reasonable comparisons in terms of win rates and robustness, as they share the same year and/or month. All contexts are now classified as

---

[9]This confidence interval was calculated with a finite population correction.

[10]This sampling does not produce erroneous dates such as February 29 for non-leap years, or April 31.

correct, incorrect, or transitional (cf. Section 3.1).

Despite this setup, a fact may have a variable number of correct and incorrect contexts per granularity due to transitional contexts, which may be absent in finer granularities if the $0.05 \times d$ step skips over them. This difference could bias performance, particularly favoring granularity Y in the robustness metric, which is calculated on fewer tests. To address this issue, YM-granularity and YMD-granularity contexts associated with transitional Y-granularity contexts are removed from the correct and incorrect sets and assigned to a special class called **Discarded**.

Finally, the contexts are converted into text and prefixed to the questions to create statements for each context at each granularity for each fact.

The resulting dataset, named **TimeStress**, includes 521,000 statements generated from 2,003 temporal facts. On average, it contains 11 correct dates and 74 incorrect dates, encompassing 1,883 unique entities, 1,385 unique subjects, 1,113 unique objects, and 86 relations. A random sample of TimeStress is presented in Table 2.

B Vulnerability to *Easy* Incorrect Contexts: Analysis of Results at Different Win Rate Thresholds

In Section 4.2, we demonstrated that LMs, even when they are almost robust on a fact (i.e., a high win rate but inferior to 100%), often fail to achieve robustness due to their vulnerability to easy contexts that are far outside the validity period (Table 4). In this section, we extend this analysis by experimenting with different win rate thresholds to observe how the distribution of incorrect contexts favored over correct contexts evolves as the threshold approaches 100%.

The results in Figure 7 indicate that even as the threshold approaches 1, LMs remain vulnerable to *easy* incorrect contexts that are significantly distant from the validity period. We would expect LMs to definitively exclude highly distant contexts once they have acquired sufficient information about the validity period. However, this is not the case here, as even when the win rate is very close to 1, LMs continue to fail on these contexts. These results suggest that language models may never achieve true robustness, as the proportion of incorrect contexts converges toward zero but never fully reaches it. This implies that there will always be a possibility for an LM to fail on a distant incorrect context.

This last point suggests that the already low percentage of robust facts could be even lower if we increased the number of incorrect and correct contexts used to calculate robustness.

C Generalization of Knowledge Across Granularities

This section provides additional details and results regarding the generalization of knowledge across granularities.

C.1 Consistency Across Granularities Based on Relative Distance

In this section, we examine the consistency of LM predictions across different granularities (Y, YM, YMD) as the distance between the tested context and the validity period increases.

To evaluate this, and solely for this section, we introduce a metric called local robustness. Local robustness for a fact, a LM, and a given incorrect context is equal to 1 if all correct contexts are preferred over this incorrect context, and 0 otherwise.

We group all statements in TimeStress according to the relative distance α from their temporal context, and restricting ourselves to the 5 most robust MLs and to the "known" facts[11] at least on one granularity by these LMs. These statements are categorized according to the interval of which their relative distance α is part. The chosen intervals are $]s, s + \frac{1}{2}]$, where s can take values from $\{-5, -4.5, \ldots, 4.5\}$. For each interval, the contexts are aligned by fact and by granularity hierarchically (e.g., 2020, 2020-04, 2020-04-23), which is guaranteed to be possible due to the properties of TimeStress (cf. Section 3.2). Local robustness is then calculated for each incorrect context, and the accuracy[12] between these measures is computed for all granularity pairs (i.e., Y-YM, Y-YMD, and YM-YMD). These coefficients are averaged across all granularity pairs, all facts, and the 5 most robust LMs, with the results presented in Figure 8.

The results indicate that the inconsistency between granularities is mainly caused by incorrect contexts located at the boundaries of the validity

---

[11]We recall that "known" in the context of this article means that the ML in question has a robustness equal to 1 on the fact in question, i.e., all correct contexts are preferred to incorrect contexts by the ML.

[12]Accuracy measures the proportion of identical elements between two vectors, that is, the number of positions where the values are equal, divided by the total number of elements compared.
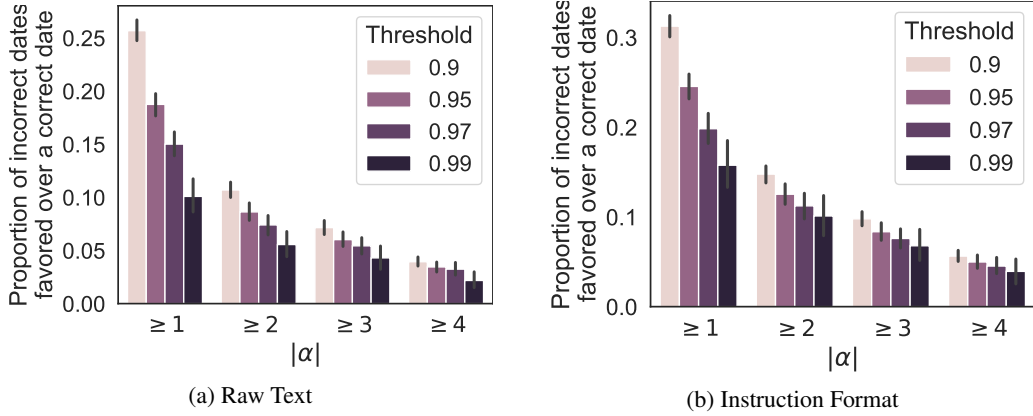
(a) Raw Text

(b) Instruction Format

Figure 7: Proportion of incorrect contexts favored over correct contexts that are beyond a relative distance α from the validity period, when the win rate exceeds the threshold, for the 5 most robust LMs. Experiments were conducted with granularity Y. 95% confidence intervals were calculated using *bootstrapping*.

period. As the context moves away from the validity period, the consistency approaches a perfect score of 1 but never reaches it regardless of the ML, the statement type and the α interval used.

C.2 Generalization Matrices for Each LM

In Section 4.2, we explored the ability of language models to generalize their temporal knowledge from one granularity to another. We provided two matrices (one for instruction-based questions and one for raw text questions) containing the generalization rate between each granularity pair averaged over the 5 most robust LMs. Complementing these average performances, the generalization rate matrices for individual models are presented in Figure 9.

C.3 Explanatory Prompts

In section 4.2, we investigated whether including explanations of temporal concepts in the prompt could help LMs better generalize their knowledge across granularities. Two prompts prefixed to each instruction in TimeStress were used:

Prompt 1 : Hierarchical natures of dates

```
A  date  is  a  specific  point  in  time,
expressed  through  a  year,  a  month,  and  a
day.  A  year  is  divided  into  months,  and
a  month  is  divided  into  days.  Answer  the
following  question.
```

Prompt 2 : Knowledge transfer between granularities

```
A date is a specific point in time. If a
fact is valid for a specific year, it holds
true for all dates within that year.  If
a fact is valid for a specific month of a
specific year, it holds true for all dates
within that month.  Answer the following
question.
```

The first explains the hierarchical nature of dates, while the second is more straightforward and explains how knowledge of a temporal fact can be generalized across granularities.

In addition to the average performance in the 4.2 section, figure 10 shows the average generalization matrices across the same 5 models as in figure 6, using raw text and an instruction format.

D Conditional Probability Calculations in LMs

Since our experiments rely entirely on the calculation (by the LM) of the conditional probability of one text given another, it is crucial that these calculations are rigorously implemented.

Given that different tokenizers split a text differently, we require a universal algorithm to best calculate the probability of generating a text given a prompt, even when the end of the prompt might be in the middle of a token.

Below are the general steps we used to compute $P(A \mid B)$ where A and B are strings:

1. Tokenize $A + B$ into a sequence of tokens $s = (t_1, t_2, \ldots, t_n)$[13].

2. Find the smallest token sequence $(t_k, \ldots, t_n)$ in s that contains B, starting from the end.

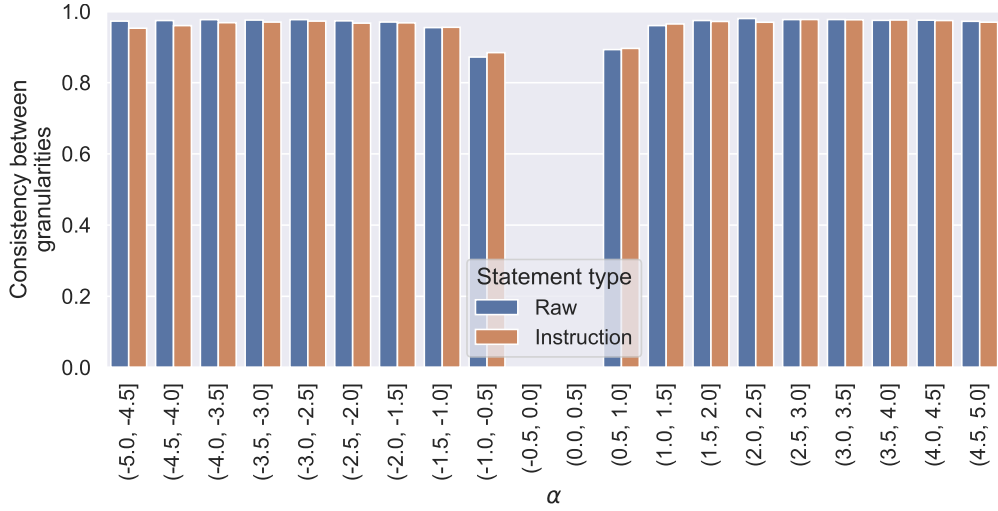[13]+ represents the string concatenation operation.

Figure 8: For each α segment, the average local robustness correlation across all granularity pairs is calculated over all facts and the 5 most robust LMs.
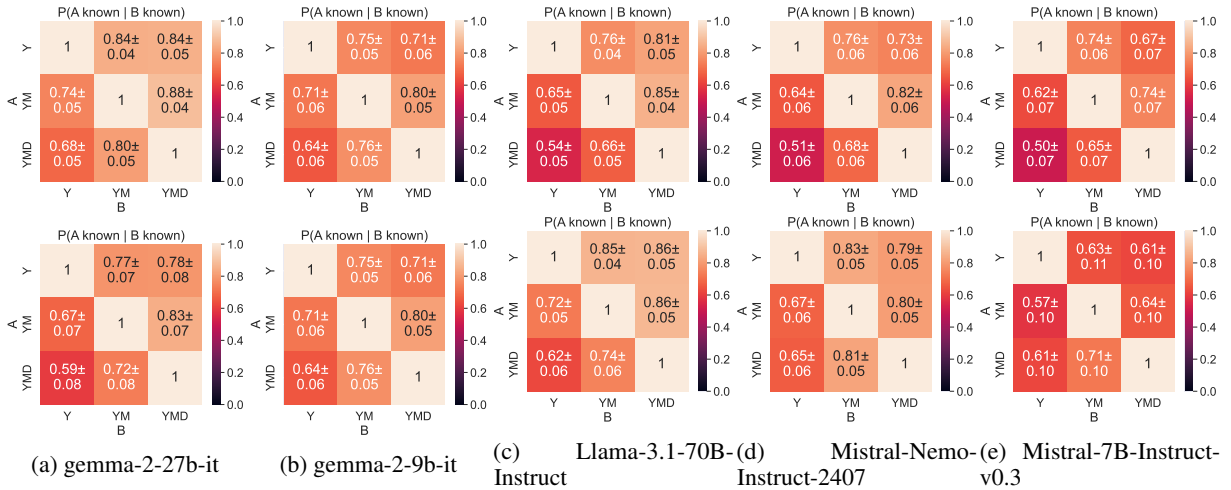

Figure 9: Generalization matrics between pairs of granularities on the 5 most robust LMs. In the **first row**, the statements are presented in a **raw** format, and in the **second row**, they are presented in a **instruction** format.

3. Compute $P(t_k, \ldots, t_n \mid t_1, \ldots, t_{k-1})$, which can be done using the *logits* produced by the LM.

   Other considerations, such as the automatic addition of special tokens by the tokenizer, must also be accounted for. A detailed implementation of this method (the function `LanguageModel.credibility_text`) that handles these details is available in the source code.

E Supplementary Results

- The average robustness score and win rate across the 18 studied LMs are presented in Figure 12.

- The relationship between the number of parameters in LMs and their performance is shown in Figure 11.

- Figure 15 illustrates the evolution of $logP(o \mid f, \tau)$ with respect to the relative distance of the date from the validity period α, which is equivalent to Figure 3 but with more details.

- Figure 16 displays the relations that were most robustly known on average by the studied LMs ("raw text" format statements).

- Figure 17 shows examples where LMs were vulnerable to easy incorrect contexts.

- Figure 14 shows the year distribution of temporal contexts across the entire TimeStress dataset.

- Figure 18 shows the influence of fact distance

17

(a) Prompt 1, raw text   (b) Prompt 1, instruction format   (c) Prompt 2, raw text   (d) Prompt 2, instruction format

(e)   No   explanation prompt, raw text
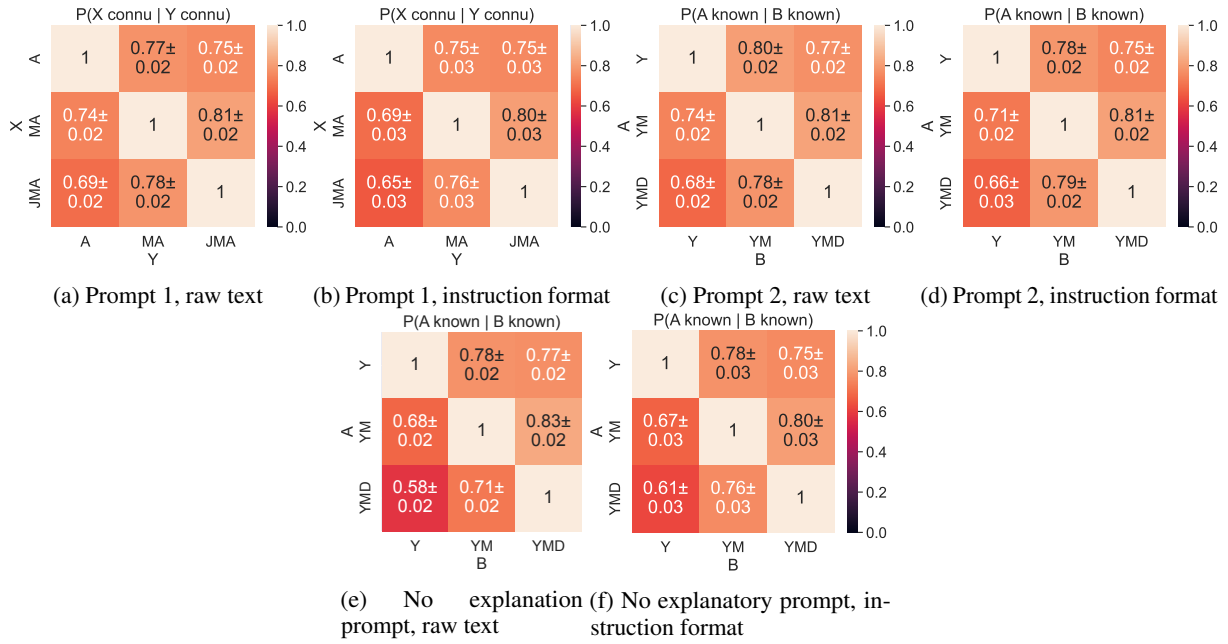
(f) No explanatory prompt, instruction format

Figure 10: Effect of adding explanations on temporal concepts through an explanatory prompt

from the present (here, the year 2021), as well as their durations, on the robustness and win rate of the 5 most robust MLs. The time unit used for both metrics is the **year**.
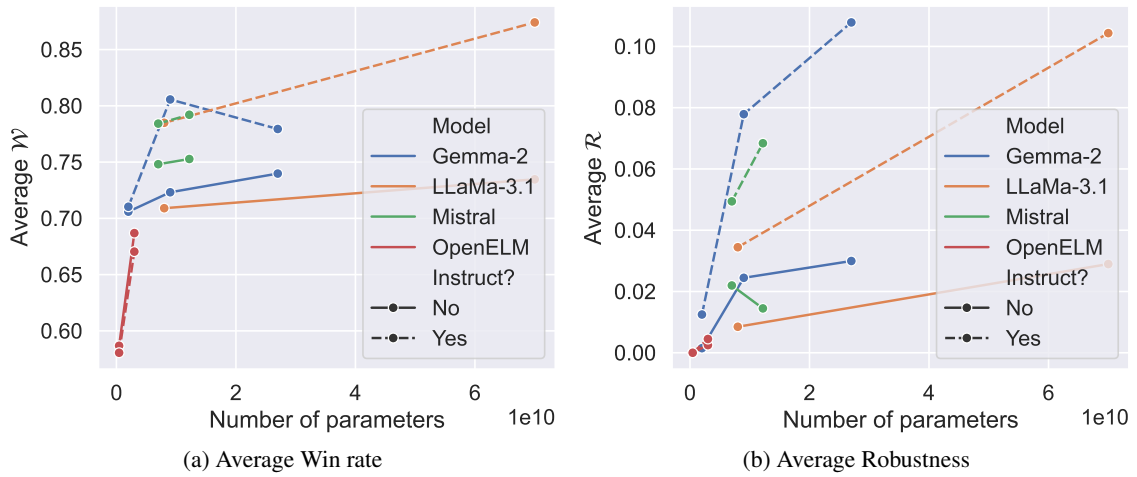
(a) Average Win rate

(b) Average Robustness

Figure 11: Relationship between the number of parameters in an LM and the metric used (across all granularities Y, YM and YMD). Pretrained models are represented by straight lines, while models finetuned on instructions are represented by dotted lines.

(a) Average win rate

(b) Average robustness
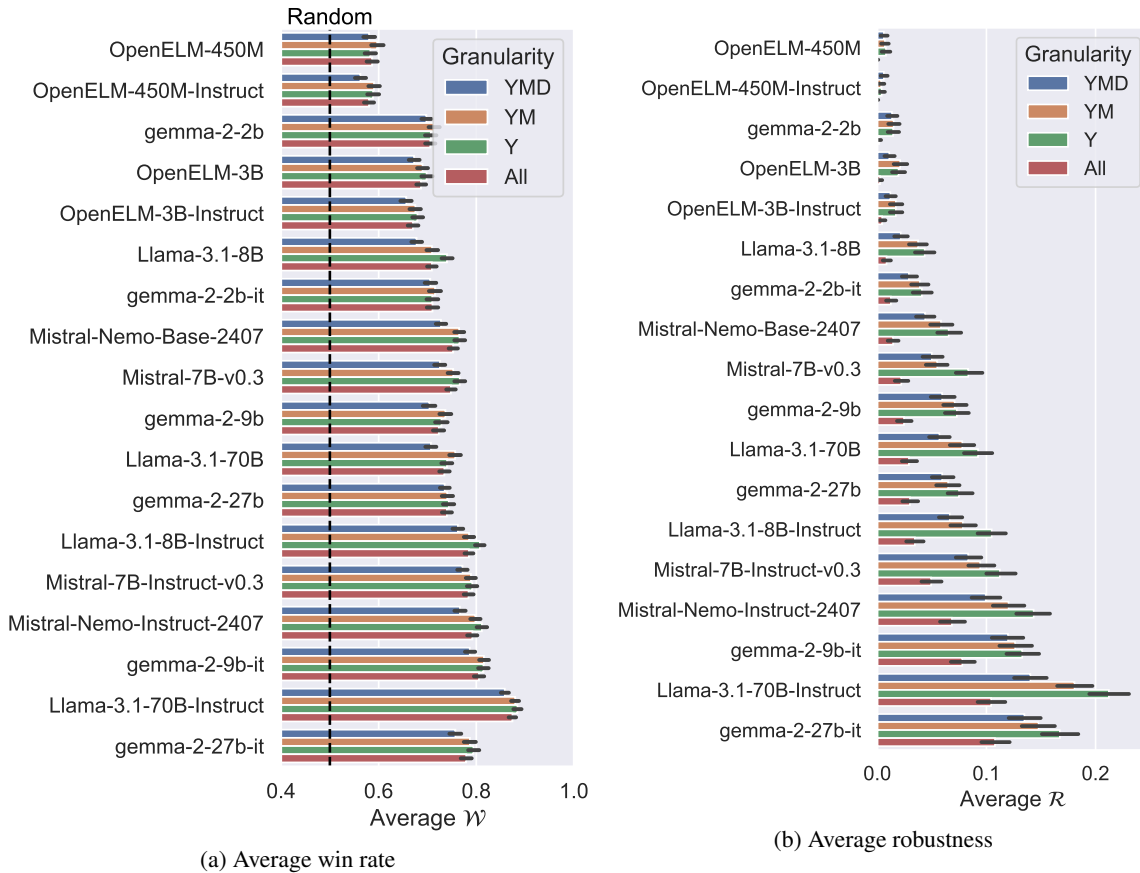
Figure 12: Average metrics across all facts in TimeStress for the 18 studied LMs with 95% confidence intervals (determined using bootstrapping).

19
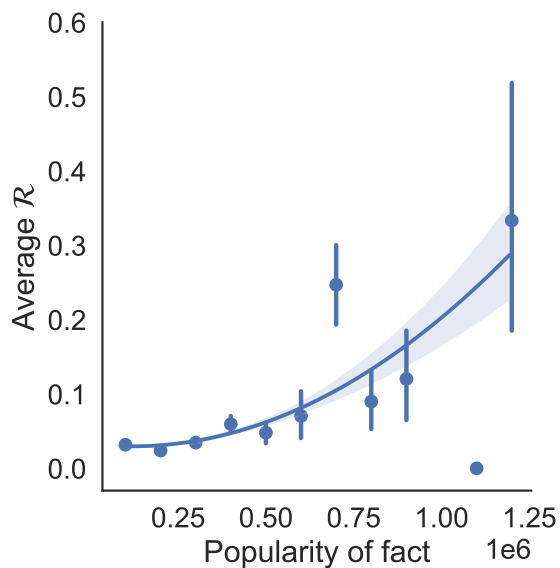
Figure 13: Relationship between fact popularity and robustness metric calculated across granularities Y, YM, YMD. The Pearson coefficient is equal $+0.065$ (p-value $< 10^{-51}$).
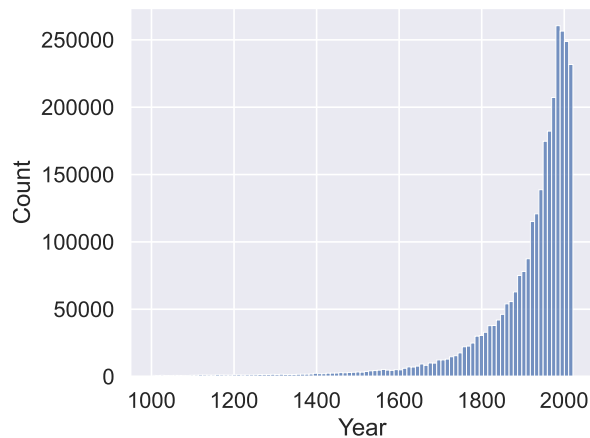
Figure 14: Distribution of the the years of all the temporal contexts in TimeStress.
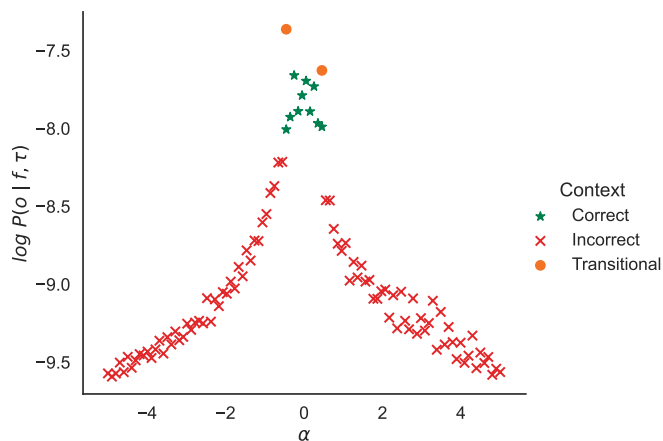
Figure 15: The evolution of $logP(o|f,\tau)$ with respect to the relative distance of the context from the validity period α. Each point is an average over many data points.
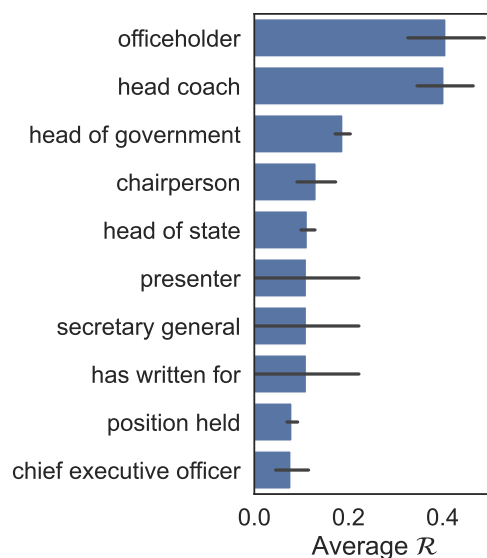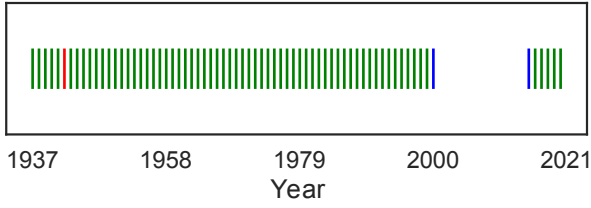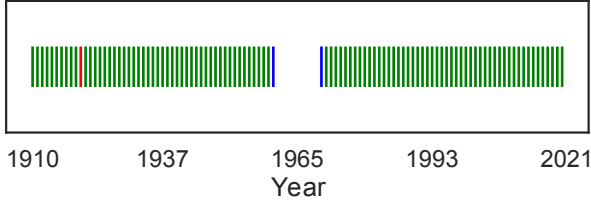

Figure 16: The 10 most known relationships (across all granularities) in TimeStress on average by the studied LMs.

Question : In [YEAR], who led the government of Texas? Rick Perry
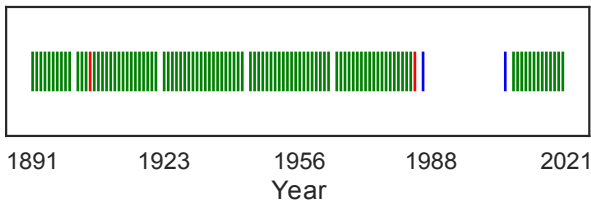
Type : Instruction
Model : Mistral-Nemo-Instruct-2407

Question : In [YEAR], of which band was Paul McCartney a member? The Beatles
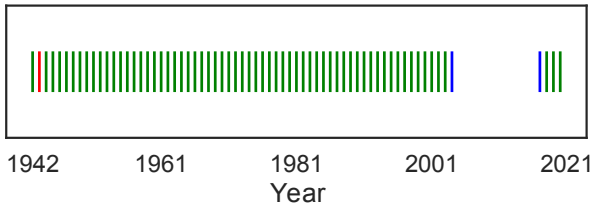
Type : Instruction
Model : Llama-3.1-8B-Instruct

Question : In [YEAR], who was the owner of Pixar? Steve Jobs

Type : Raw
Model : gemma-2-9b-it

Question : In [YEAR], which football club was Wayne Rooney associated with? Manchester United F.C.

Type : Instruction
Model : gemma-2-27b-it

Figure 17: Examples of vulnerability to *easy* incorrect contexts for different LMs. The color blue represents the boundaries of the validity period, the color green represents incorrect contexts that are never preferred to correct contexts, and the color red, on the contrary, represents incorrect contexts that were preferred to one or more correct contexts.
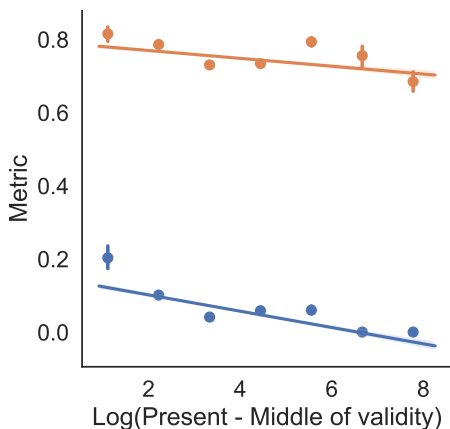
(a) Logarithm of the distance of the fact with respect to the present.
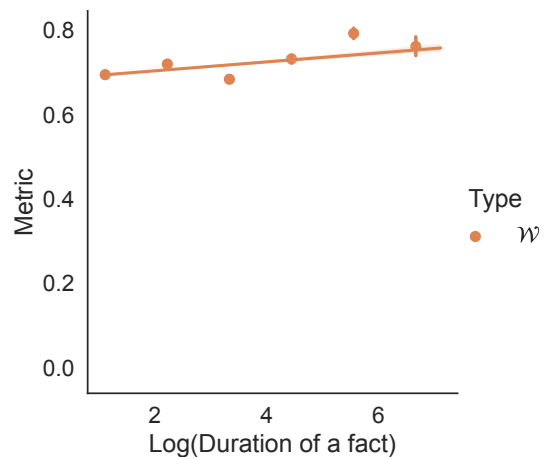
(b) Logarithm of the duration of the fact.

Figure 18: The influence of two factors on the robustness and win rate of the 5 most robust LMs. All correlations are statistically significant where the null hypothesis is the absence of linear correlation. Robustness is missing from Figure b because its analysis is not relevant as the duration of a fact is confounded with another variable: the number of matches of a fact. Indeed, the longer a fact is, the more matches it has, and the lower is the robustness.

Temporal Fact	Statement	Status
(Alexander Graham Bell, country of citizenship, United States of America, 1882, 1922)	In July 1734, what was Alexander Graham Bell's country of citizenship? United States of America	Incorrect
(Lauren Bacall, spouse, Jason Robards, 1961-07-04, 1969-09-10)	In July 1984, who was the spouse of Lauren Bacall? Jason Robards	Incorrect
(Vatican City, head of state, John Paul II, 1978-10-16, 2005-04-02)	In July 2006, who held the highest authority in Vatican City? John Paul II	Incorrect
(Gareth Barry, member of sports team, Manchester City F.C., 2009, 2014)	In July 2020, which football team included Gareth Barry as a player? Manchester City F.C.	Incorrect
(Pierce Brosnan, spouse, Cassandra Harris, 1980, 1991)	In 1954, who did Pierce Brosnan have as his wife? Cassandra Harris	Incorrect
(Metallica, has part, Jason Newsted, 1987, 2001-01-17)	In 1971, who was included in Metallica's lineup? Jason Newsted	Incorrect
(Eliza Dushku, unmarried partner, Rick Fox, 2009, 2014)	In 2003, who was Eliza Dushku in a relationship with? Rick Fox	Incorrect
(United Kingdom, head of state, George VI, 1936-12-11, 1952-02-06)	On July 1, 1892, who served as the king of the United Kingdom? George VI	Incorrect
(Linda Lee Cadwell, spouse, Bruce Lee, 1964, 1973-07-20)	In 1929, who was the spouse of Linda Lee Cadwell? Bruce Lee	Incorrect
(George Harrison, part of, The Beatles, 1960, 1970)	On July 2, 1971, what was the name of the band that George Harrison was associated with? The Beatles	Incorrect
(Philippines, head of state, Corazon Aquino, 1986-02-25, 1992-06-30)	On July 2, 1969, who served as the leader of the Philippines? Corazon Aquino	Incorrect
(Jawaharlal Nehru, position held, Prime minister of India, 1947-08-15, 1964-05-27)	In 1985, what position did Jawaharlal Nehru hold? Prime Minister of India	Incorrect
(Vienna, country, Austria-Hungary, 1867-03-30, 1918-11-11)	In July 1769, which country did Vienna belong to? Austria-Hungary	Incorrect
(Mileva Marić, spouse, Albert Einstein, 1903, 1919)	In July 1907, who was Mileva Marić married to? Albert Einstein	Correct
(Mayte Garcia, spouse, Prince, 1996, 2000)	In July 1979, who was the spouse of Mayte Garcia? Prince	Incorrect
(Abkhazia, country, Soviet Union, 1921, 1991)	In July 1956, which country did Abkhazia belong to? Soviet Union	Correct
(Georgia, member of, Commonwealth of Independent States, 1993-12-03, 2009-08-18)	In 1930, what group included Georgia as a member? Commonwealth of Independent States	Incorrect
(Abraham Lincoln, member of political party, Whig Party, 1834, 1854)	In 1808, which political party was Abraham Lincoln a member of? Whig Party	Incorrect
(Wales, located in the administrative territorial entity, Kingdom of England, 1284, 1707-04-30)	On July 1, 1072, which territorial entity included Wales? Kingdom of England	Incorrect
(Frédéric Chopin, residence, Paris, 1831, 1849)	On July 2, 1847, which city was home to Frédéric Chopin? Paris	Correct

Table 2: Random sample from TimeStress.