DateLogicQA: Benchmarking Temporal Biases in Large Language Models

Gagan Bhatia $^{\phi}$, **MingZe Tang** $^{\phi}$, **Cristina Mahanta** $^{\phi}$, **Madiha Kazi** $^{\phi}$
University of Aberdeen
`{g.bhatia.24,m.tang.24,c.mahanta.24,m.kazi.24}@abdn.ac.uk`

Abstract

This paper introduces DateLogicQA, a benchmark with 190 questions covering diverse date formats, temporal contexts, and reasoning types. We propose the Semantic Integrity Metric to assess tokenization quality and analyse two biases: Representation-Level Bias, affecting embeddings, and Logical-Level Bias, influencing reasoning outputs. Our findings provide a comprehensive evaluation of LLMs' capabilities and limitations in temporal reasoning, highlighting key challenges in handling temporal data accurately. The GitHub repository for our work is available at https://github.com/gagan3012/EAIS-Temporal-Bias [1]

1 Introduction

Accurate temporal reasoning is essential for real-world applications like event planning and historical questions. However, biases in Large Language Models (LLMs) can lead to misinterpretations or errors in date-related tasks. Understanding these biases is essential for precisely handling numerical structures and contextual meanings, making temporal reasoning ideal for identifying and analysing biases in tokenization, representation, and logical reasoning.

A significant source of these biases originates from the tokenization process. While tokenizers divide the text into subword units, inconsistencies in tokenizing dates can disrupt reasoning tasks. This can lead to two types of biases: Representation-Level Bias, caused by inconsistencies in embeddings affecting semantic structures of dates, and Logical-Level Bias, where correct tokens do not yield accurate outputs due to misaligned internal processing. Together, these biases highlight the challenges LLMs face in preserving the integrity and interpretability of temporal data across diverse formats and contexts.

[1] ϕ Denotes equal contribution

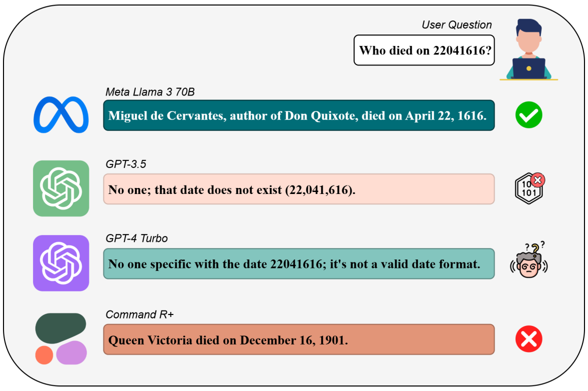


Figure 1: Examples of temporal biases in LLMs. **Incorrect** Response, **Faulty Date** but accurate reasoning indicating representation level temporal bias, **Faulty reasoning** but accurate date indicating logical level temporal bias, **Correct** response

This paper makes several contributions to understanding temporal biases in LLMs. **(1)** We introduce **DateLogicQA**, a dataset of 190 curated questions for evaluating temporal reasoning across various date formats, contexts (past, present, future), and reasoning types (commonsense, factual, conceptual, numerical). **(2)** We propose the Semantic Integrity Metric to assess tokenization quality, punishing unnecessary splits and excessive token counts. **(3)** We conduct human evaluations of model responses to analyse tokenization accuracy and reasoning quality, providing insights beyond automated metrics. **(4)** A thorough bias evaluation examines representation and logical-level biases using embeddings and outputs to investigate LLMs' treatment of temporal references.

We have organised the paper as follows: Section 2 reviews related works, summarising the impact of tokenization on LLM performance and past temporal reasoning approaches. Section 3 details the creation of the DateLogicQA dataset, including its design principles and examples. Section 4 outlines methods for evaluating tokenization, temporal reasoning, and biases. Section 5 presents experiment results, followed by a discussion of findings and bias mitigation in Section 6. Lastly, Section 7 summarises our contributions.

2 Related Works

Impact of Tokenization on Language Models Tokenization significantly affects the efficiency and reasoning abilities of large language models (LLMs). Research by Gu et al. (2024) and Goldman et al. (2024) highlights that tokenizers with higher compression rates enhance representation efficiency, particularly in smaller models. However, Schmidt et al. (2024) argue that effective tokenization also depends on pre-tokenization and vocabulary design. Studies like Ahia et al. (2023) show that poorly tokenized languages face performance and fairness issues. Furthermore, choices in tokenization impact reasoning; Zhang et al. (2024) and Singh and Strouse (2024) indicate that numerical tokenization can lead to errors in arithmetic and counting tasks. Rajaraman et al. (2024), Alberts et al. (2024), Minixhofer et al. (2024), and Gastaldi et al. (2024) show how well-designed tokenizers improve sequence pattern modelling and numerical reasoning through advanced embedding methods. Our study extends this work by examining tokenization's role in handling diverse date formats for temporal reasoning.

Temporal Reasoning in LLMs Temporal reasoning poses challenges for LLMs due to inherent biases. Zhu et al. (2024) discussed "nostalgia bias" (favouring outdated knowledge) and "neophilia bias" (speculative future predictions), while Tan et al. (2023b) observed inconsistent generalisation across different time periods. Structured approaches like temporal graphs (Xiong et al., 2024a) and synthetic datasets (Fatemi et al., 2024) enhance performance by explicitly encoding temporal relationships. Additionally, tokenization critically affects temporal reasoning; Zhao et al. (2024) found that temporal misalignment hampers accuracy, and Kishore and He (2024) identified inductive biases in models like GPT-3.5 and GPT-4. Su et al. (2024a) propose task-agnostic approaches to enhance temporal reasoning, while Gastaldi et al. (2024) and Rajaraman et al. (2024) link tokenization to reasoning performance. By analysing how tokenization strategies affect temporal reasoning, especially for date formats, our work fills a gap in understanding the interplay between tokenization and temporal task performance.

3 DateLogicQA

We introduce **DateLogicQA**, a dataset designed to explore how LLMs handle dates in various formats and contexts to tokenize, interpret, and reason with them. It consists of 190 questions divided into four categories: *commonsense*, *factual*, *conceptual*, and *numerical*. Each category features one of seven date formats across three temporal contexts: *past*, *present*, and *future*. This systematic variation allows for an in-depth analysis of LLMs' performance with temporal information.

Objective and Purpose The dataset aims to assess LLMs' tokenization and understanding of dates, as errors can lead to interpretative biases. By embedding dates within questions, we evaluate context-rich date interpretation, simulate real-world scenarios where dates carry contextual significance, and test LLMs' ability to extract and interpret date information accurately.

Concepts	Example
Numerical	What is the time 7 years and 9 months after 27101446?
Factual	Which of the people died on 23041616? A) Shah Jahan B) Miguel de Cervantes C) Princess Diana D) William Shakespeare
Conceptual	The first iPhone was released on 29062007. How many years has it been since its release?
Commonsense	John was born on 15-03-1985. He graduated from college on 01-05-2007. Was John older than 18 when he graduated?

Table 1: Dataset samples illustrating different temporal reasoning concepts.

Date Format	Example
DDMMYYYY	23041616
MMDDYYYY	04231616
DDMonYYYY	23April1616
DD-MM-YY	23-04-16
YYYY, Mon DD	1616, April 23
DD/YYYY (Julian calendar)	113/1616
YYYY/DD (Julian calendar)	1616/113

Table 2: Dataset samples illustrating different date formats used.

This approach comprehensively examines various temporal notations, including uncommon formats like Julian calendar representations.

Temporal Distribution DateLogicQA spans a broad temporal range, featuring dates from historical periods (e.g., the 1600s), modern contexts (e.g., the 2000s), and hypothetical futures (e.g., the 2100s). For clarity, we categorised dates into *past*, *present*, and *future*, with some questions covering multiple dates to assess LLMs' ability to manage temporal relationships across contexts.

Rationale for Design The dataset prioritises models' ability to interpret dates within broader narratives rather than as isolated data points. Its smaller size allows for careful curation of high-quality, linguistically diverse questions, focusing on specific nuances of temporal reasoning. This enables detailed analysis of model behaviour and understanding of temporal biases.

4 Methodology

The study proposes three interests to investigate temporal bias in models: tokenization process, temporal task capability, and internal computation across different LLMs.

4.1 Semantic Integrity

This experiment targets the tokenization process in different LLMs to identify how it influences the semantic interpretation of dates when presented in different formats. We specifically focus on the Semantic Integrity Metric, which measures the extent to which the original semantic meaning of a date is preserved after tokenization. Several key highlights are observed, such as how a single date input is presented after being tokenized, the extent of semantic preservation, and the ability to generalise across different date representations. These findings provide valuable insights into the tokenization process and its impact on temporal reasoning in LLMs.

Semantic integrity evaluates how well the tokenized date output maintains its original meaning and structure. The semantic integrity score ranges from 0 to 1. A higher score - closer to 1 - indicates that the date segmentation is nearly accurate, better preserving the intended structure and information. In contrast, a score closer to 0 indicates an inadequately tokenized date structure. The formula for calculating semantic integrity is as follows:

$$\mathrm{SI} = \max\left(0, \min\left(1, 1 - P - S - T - R\right)\right)$$

		Reasoning	
		Wrong	Correct
Date	Wrong	Incorrect (Hallucination)	Faulty Date, Accurate Reasoning (Representation-Level Temporal Bias)
	Correct	Accurate Date, Faulty Reasoning (Logical-Level Temporal Bias)	Correct

Figure 2: Human evaluation rubric

Unnecessary Splitting of Components (P) : If a date is tokenized into parts that do not correspond to the ideal format, a penalty of 0.1 is applied.

Preservation of Separators (S) : A penalty of 0.1 is applied when separators are lost during tokenization, reflecting incorrect date parsing (e.g., tokenizing `%Y-%m-%d` as `%Y%m%d`).

Token Count (T) This penalty design aims to penalise token outputs longer than necessary, such as `%Y-%m-%d`. If any tokenized output exceeds five splits, we apply a penalty for excessive fragmentation. A higher token count often indicates a loss of semantic information, hindering the model's ability to interpret temporal values.

Similarity with Baseline (R) This metric uses cosine similarity to measure how much a tokenized output deviates from a baseline reference of date tokens. Tokens with greater deviation incur higher penalties, while those resembling the baseline receive lower penalties.

4.2 Human-Led Temporal Bias Assessment

Understanding temporal contexts is crucial for analysing events over time. This includes grasping temporal references like *"How many years has it been since..."* (Past) and *"What will the contract's last day be..."* (Future), along with the maintenance of logical chronological order and adaptation to changes in context. For large language models, this capability is vital for tasks such as historical inquiries, time-sensitive query handling and predictions about future events. Assessing biases in temporal reasoning is essential for accuracy across various applications. We utilized the dataset referenced in Section 3.

We conduct a human evaluation to assess the temporal bias of LLMs as automated methods may exhibit inherent biases that affect results, ultimately undermining the evaluation's purpose. This methodology provides a more reliable analysis, identifying outliers that respond accurately without fully comprehending temporal aspects. Instead, it relies on contextual clues or learned patterns acquired during training or through retrieval-augmented generation.

Model responses are categorised based on colours in Figure 2, representing levels of temporal understanding. **Dark Orange** (■) denotes incorrect answers or temporal hallucinations from failure to tokenize dates or grasp context. **Light Orange** (■) reflects Representation-Level Temporal Bias, where the model tokenizes dates inaccurately but reaches the correct answer through logical reasoning. This suggests that some internal reasoning within the model compensates for misunderstanding the date format. **Light Teal** (■) signifies Logical-Level Temporal Bias, where the model tokenizes correctly but misapplies logic due to misattributing events or calculation errors. Finally, **Dark Teal** (■) denotes correct answers, indicating successful tokenization and logical reasoning. This illustrates a complete understanding of the question.

4.3 Understanding Temporal Bias

We investigate potential biases in the internal embedding space and softmax computations of large language models (LLMs) when processing texts with different temporal references, such as past, present, and future contexts. Temporal biases in LLMs fall into two main types: **Representation-Level Temporal Bias** indicates significant differences in internal embeddings across time references, revealing inconsistencies in encoding semantic information. In contrast, **Logical-Level Temporal Bias** occurs when output probabilities vary for semantically identical inputs due to changes in temporal references.

We established a controlled experimental framework to quantify these biases, analysing embeddings and softmax outputs across three temporal categories: *past*, *present*, and *future*. We measured representation-level biases using cosine similarity between averaged embeddings, with lower similarity indicating greater divergence. We assessed logical-level biases using KL divergence for softmax distributions, where higher divergence reflects substantial probability differences. Additionally, we examined sensitivity to seven date formats from Table 2.

We conducted experiments using *Llama 3.2 3B* (Dubey et al., 2024). For each prompt, we extracted two key outputs: internal embeddings, averaged from all the hidden layers of hidden states, and softmax probabilities, denoting the output distribution over the vocabulary. We analysed temporal biases by comparing these outputs across different references. Simultaneously, we assessed format sensitivity within each reference to determine how tokenization and variations in date formats affect model behaviour.

5 Results

5.1 Impact of tokenizers

Model	SI	TC	PC	PS
Baseline	1.00	4.30	✓	✓
OLMoE	0.77	5.08	≈	✓
OLMo	0.77	5.08	≈	✓
Davinci-003	0.75	5.17	×	✓
Llama 3	0.74	4.98	×	✓
GPT-3.5	0.74	4.98	×	✓
GPT-4	0.74	4.98	×	✓
GPT-4o	0.74	4.98	×	✓
Qwen	0.42	9.30	×	✓
Cohere	0.42	9.30	×	✓
Gemma	0.42	9.30	×	✓
DeepSeek	0.42	9.30	×	✓
Llama 2	0.37	10.30	×	✓
Mistral	0.37	10.30	×	✓
Phi 3.5	0.37	10.30	×	✓
Llama 1	0.37	10.30	×	✓

Table 3: Performance comparison of various models on semantic integrity, token count, and preservation of components and separators.

In Table 3, we present the Semantic Integrity scores with respect to token count **(TC)**, component preservation **(PC)** and separators **(PS)** for different tokenizers used by the tested models. Notably, newer released models tend to have a higher average semantic integrity score at ≈ 0.7 while token counts maintained close to the frequency of ≈ 5.

As described in Table 2, most of the tested date formats are composed of 5 components (e.g., DD-MM-YY), with some instances of a 3-component structure like the Julian date format (YYYY/DD). As a result, the baseline reference has an average token count of 4.30 rather than the ideal 5.

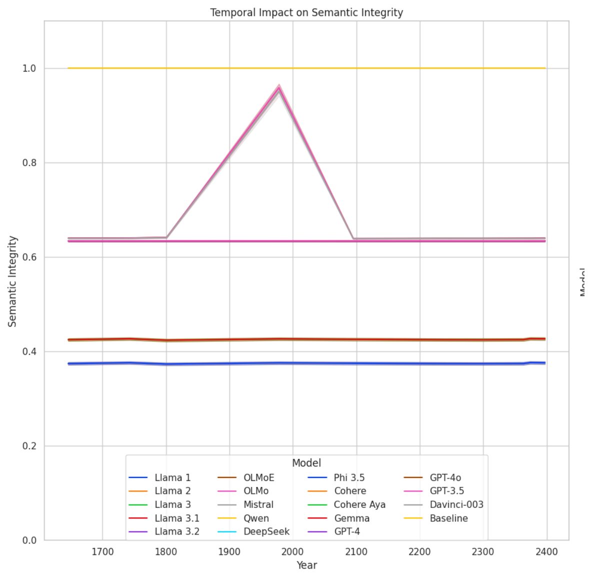


Figure 3: Temporal impact on semantic integrity

While many tokenizers struggled to consistently preserve the components, often producing tokenized outputs that deviated from the ideal, the tokenizers used by the OLMoE (Muennighoff et al., 2024) and OLMo (Groeneveld et al., 2024) models handled date inputs correctly on certain occasions.

Token Count and Semantic Integrity From Table 3, we identify an indirect relationship between the token count and the Semantic Integrity score, specifically at lower Semantic Integrity often corresponds to more token splits, suggesting that excessive and inefficient token splitting could badly impact the interpretability of the original input. Figure 8 further establishes this observation by showing that the Semantic Integrity scores tend to fall near the lower end in the area of higher token count.

Shared Characteristics Between Tokenizers Results produced by some tokenizers have been identical, as observed in Table 3. For example, the tokenizers that Llama 1 (Touvron et al., 2023a) and Mistral (Jiang et al., 2023) use produce similar Semantic Integrity scores, a trend also observed in other models. Therefore, we investigated this further by evaluating the output from the tokenizers used by all models, presenting the results in Table 4. This finding seems accurate, as we identified similar tokenized date outputs from the tokenizers used in certain models. For instance, (Macijauskas, 2024) has also concluded a similar finding for the tokenizer used by GPT-4 (OpenAI et al., 2023) and Llama 3 (Dubey et al., 2024), which both produced similar outputs during his experiment. From our observation, this helps justify the idea of shared tokenizers between several models.

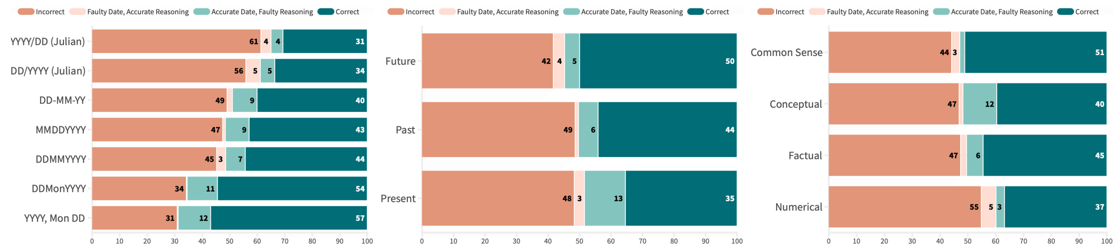
Performance in Different Temporal References In addition, Figure 3 reveals that certain tokenizers exhibit a temporal bias, resulting in varying scores produced in different timelines. Some tokenizers yield better Semantic Integrity scores for dates closer to the present (1900s-2100s). Although this is not consistently true across all tokenizers, this pattern highlights a potential bias in a tokenizer, with some favouring more recent dates over others. Slight variations in Semantic Integrity scores are particularly noticeable between the 1700s and 1900s for the tokenizers from Gemma (Team et al., 2024) and Phi 3.5 (Abdin et al., 2024) model. In contrast, the tokenizers used by OLMoE (Muennighoff et al., 2024) and OLMo (Groeneveld et al., 2024) exhibited more significant score fluctuations in more recent years.

5.2 Temporal Reasoning Analysis

Temporal reasoning, including processing and drawing inferences from historical and future dates, is one of the most challenging tasks for large language models. The current study investigates whether there are any differences in LLM performance when reasoning with historical dates, such as "July 20, 1969", and future dates, such as "January 1, 2050". To this end, we present the testing of 12 state-of-the-art LLMs using a question-answer dataset encompassing different date formats and various temporal contexts. This paper examines their skills in tokenization, comprehension, and inference on dates. We classify the answers into four categories based on their accuracy and treatment of the dates and logical structure involved, thereby providing a systematic evaluation framework.

In order to ensure that the assessment is robust, four human annotators, each with at least four years of experience in computer science, evaluated the responses across the four categories. The labelling achieved a high inter-annotator agreement with a Cohen's kappa (K) score of 0.80, confirming the reliability of the evaluation framework. These results evidence two critical areas where LLMs shine and their struggles, giving further information about their strengths and limitations concerning temporal reasoning.

Performance of Selected LLMs The evaluation of 12 language models, accessed through Hugging



(a) Date Format visualisation (b) Time period visualisation (c) Question Type visualisation

Figure 4: Results Visualisations

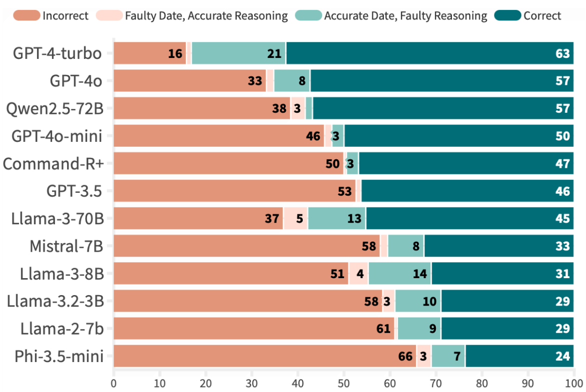


Figure 5: Each bar is segmented into four colors representing the quality of responses: **Incorrect** Response, **Faulty Date** but accurate reasoning indicating representation level temporal bias, **Faulty reasoning** but accurate date indicating logical level temporal bias, **Correct** response

Face and OpenAI APIs, provided a comprehensive overview of their performance on temporal reasoning tasks. Small models like Llama-3.2-3B (Dubey et al., 2024) and Phi-3.5-mini (Abdin et al., 2024) gave bad performances, with 58% and 66% incorrect answers, respectively. Due to their restricted processing and resources, these models performed poorly in tokenization and reasoning. Mid-sized models, including Mistral-7B (Jiang et al., 2023), Llama-3-8B (Dubey et al., 2024), and Llama-2-7B (Touvron et al., 2023b), demonstrated a more moderate improvement. They had trouble with complex reasoning problems, although they were able to improve their tokenization accuracy. Larger models, including Llama-3-70B (Dubey et al., 2024), Qwen2.5-72B (Yang et al., 2024), and Command R+ (Cohere, 2024), were more robust in their performance, especially in date interpretation and logical reasoning. However, there were inconsistencies in specific formats. Proprietary models, including GPT-3.5 (Brown et al., 2020), GPT-4-turbo (OpenAI et al., 2023), GPT-4o, and GPT-4o-mini (OpenAI et al., 2024) outperformed all the rest, with GPT-4-turbo leading on correct responses with 63% and the lowest rate of incorrect answers at 16%. These results emphasise that model size, architecture, and diversity of pretraining data all bear on performance related to temporal reasoning tasks.

Performance Based on Date Formats The format of the date had a significant impact on model performance. Models performed best for formats that included clear separators and natural language cues, such as "YYYY, Mon DD" with 57% correct and "DDMonYYYY" with 54% correct. The poorest performance was from formats like "YYYY/DD (Julian)" and "DD/YYYY (Julian)", with only 31% and 34% correct, respectively, since the representation is less common and more complex in tokenization. This trend indicates format standardisation's apparent relevance in improving date processing efficiency in LLMs.

Performance Across Temporal Contexts Temporal context also mattered a lot. Models were better with future dates, 50% correct, compared to historical dates, 44%, and present dates, 35%. This runs contrary to the expectations and may point to the fact that future-oriented reasoning tasks tap into the generative and predictive capabilities of the models. Historical and present contexts, which often require exact recall or conformity to training data, proved more difficult due to inconsistencies in the coverage of pretraining corpora.

Performance by Question Type Question type further modified results, with commonsense reasoning questions reaching the highest percentage of correctness: 51%. These questions depended less on explicit tokenization and more on logical inference, which LLMs did comparatively well. Factual questions were at 45%, while conceptual questions reached slightly lower performances of 40%. Numerical reasoning questions were the hardest; only

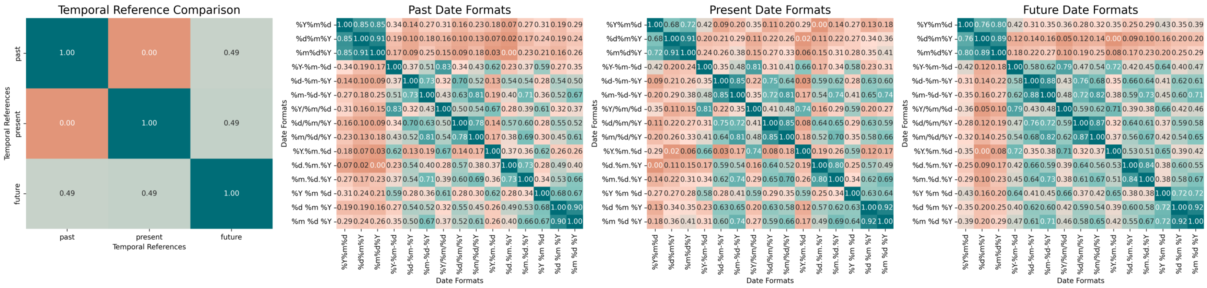


Figure 6: Representation level Temporal Bias Analysis using LLama 3.2 3B

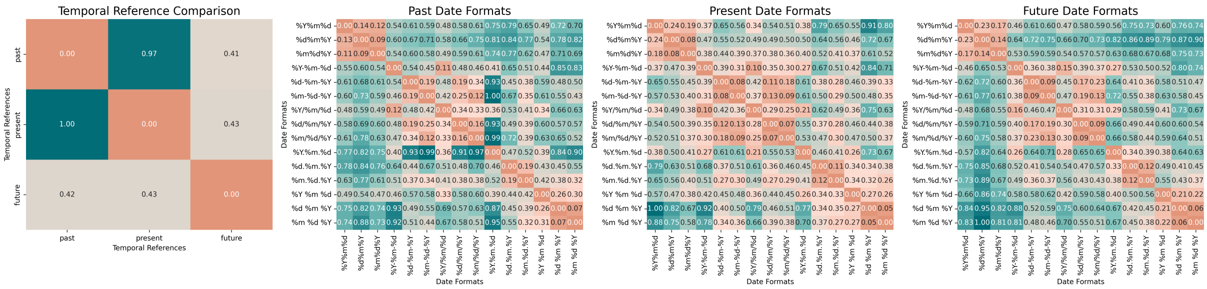


Figure 7: Logical level Temporal Bias Analysis using LLama 3.2 3B

37% were correct since these often included some calculation or logical deduction that exposed the weaknesses in the models' reasoning capability.

5.3 Temporal Sensitivity Analysis

We analyse temporal biases and format sensitivity by examining the embeddings and softmax outputs of the model for prompts across three temporal categories — past, present, and future — and multiple date formats, as shown in Table 2. We organise the findings into representation-level bias, logical-level bias, and format sensitivity, and we show the results in Figure 6 and Figure 7.

Representation-Level Bias We evaluated representation-level bias by calculating the cosine similarity between the averaged embeddings for prompts across the three temporal references. The leftmost heatmap in Figure 6 illustrates these similarities.

The embeddings for past and present references exhibit no measurable similarity (0.00), emphasising that the model encodes historical and contemporary contexts with distinct semantic structures. However, the moderate similarity between future and present suggests some shared contextual features between these categories while maintaining semantic differentiation. The moderate similarity between past and future indicates that these categories share overlapping contextual features while remaining semantically distinct. This further implies that the model is somewhat confused regarding futuristic references, which may frequently misattribute to a different temporal category, likely reflecting the training data distribution.

Logical-Level Bias We assessed logical-level bias by measuring the Kullback-Leibler (KL) divergence between softmax outputs for prompts across temporal references. The leftmost heatmap in Figure 7 illustrates these divergences—prompts referencing the present exhibit the lowest divergence, indicating stable and consistent output probabilities. However, significant divergence is observed between past-present and future-present comparisons, highlighting the model's reliance on different priors when predicting tokens for noncontemporary contexts.

The moderate divergence between past and future outputs suggests that the model differentiates between these temporal categories while leveraging some shared contextual grounding. The distinct KL divergences for non-present prompts indicate a logical-level bias, where the model's probabilistic outputs are sensitive to the temporal context, even when the semantic content of the prompts remains equivalent.

5.4 Format Sensitivity Analysis

Figures 6 and 7 (second to fourth columns) show the model's sensitivity to variations in date formats for each temporal reference. Both embeddings and softmax outputs reveal notable patterns of variability across formats.

Representation-Level Bias The cosine similarity heatmaps in Figure 6 indicate that date formats with standard separators (e.g., `%Y-%m-%d`) yield higher consistency, particularly for present references. Non-standard formats (e.g., `%Y%m%d`, `%d%m%Y`) result in lower similarity, especially for past and future prompts. The future category exhibits the highest variability in embeddings across formats, suggesting that futuristic contexts rely more on consistent input structures. In contrast, embeddings for present references remain robust across formats, likely due to the dominance of contemporary contexts in the training data.

Logical-Level Bias The KL divergence heatmaps in Figure 7 reflect similar trends. Standardised data formats (e.g., `%Y-%m-%d`) produce more stable predictions, while non-standard formats (e.g., `%d%m%Y`) introduce higher variability. This sensitivity manifests most prominently in future references, where the KL divergence values are consistently higher, indicating that the model's predictions have increased uncertainty. In contrast, present references remain relatively stable, reinforcing the model's preference for standardised inputs and contemporary contexts.

The results highlight two fundamental temporal biases in the model. First, representation-level biases reveal that the model encodes temporal contexts with distinct semantic structures, likely shaped by training data distribution. Second, logical-level biases indicate inconsistencies in output probabilities across temporal references, underscoring the challenges of achieving temporal generalisation. Furthermore, the heightened sensitivity to non-standard date formats underscores the importance of input standardisation for ensuring consistent model behaviour in temporal reasoning tasks.

6 Discussion

This study highlights the need for targeted strategies to address temporal biases in large language models (LLMs). A key step is to enhance pre-training datasets to ensure temporal diversity, incorporating historical, contemporary, and futuristic contexts. While resources like Redpajama (Weber et al., 2024) and Dolma (Soldaini et al., 2024) are open source, researchers should develop data focused on temporal reasoning with varied formats and cultural contexts.

Post-training methods, such as Direct Preference Optimization (DPO) (Rafailov et al., 2024), offer a promising avenue for fine-tuning models using curated datasets specifically designed to improve their logical temporal reasoning capabilities (Su et al., 2024b; Tan et al., 2023a). These approaches can help align the models' outputs with human-preferred logical reasoning patterns, addressing specific shortcomings in temporal tasks. Additionally, Retrieval-Augmented Generation (RAG) (Liu et al., 2024) enhances LLMs by integrating external knowledge dynamically during inference, allowing the models to access up-to-date or context-specific temporal information beyond their static training data. Moreover, prompting techniques such as Chain of Thought (CoT) prompting (Wei et al., 2023) enable models to break down complex temporal reasoning tasks into incremental steps, improving interpretability and logical coherence (Liu et al., 2024; Xiong et al., 2024b).

However, while these post-training methods significantly mitigate biases in temporal reasoning and improve model performance, they are not sufficient to completely eliminate inherent biases. Factors such as the limitations of pre-trained embeddings, the static nature of foundational knowledge, and the variability in task-specific datasets mean that biases are likely to persist at some level. Thus, post-training approaches should be viewed as an important step toward reducing biases.

7 Conclusion

Our paper addresses the challenges of temporal biases in large language models (LLMs) and proposes a structured approach to analyse their performance with temporal data. We introduced the DateLogicQA dataset and the Semantic Integrity Metric to evaluate the impact of diverse date formats and contexts on tokenization and reasoning. Our findings highlighted representation-level biases, where temporal contexts are inconsistently encoded, and logical-level biases, evident in varying outputs for similar prompts. We suggest mitigation strategies, such as temporally balanced pretraining datasets, post training and prompting methods.

Limitations

Future Scalability The manual human evaluation approach for temporal reasoning performance analysis was time-consuming and challenging for future scalability. Furthermore, the evaluation technique requires high consensus among evaluators, especially when team size expands. Maintaining the evaluation quality in a larger team is also particularly difficult, and it might require more effort to cross-validate the results.

Ethical Considerations

AI usage. It's pertinent to acknowledge the role of AI tools such as ChatGPT in our project. Specifically, Grammarly was utilized minimally and primarily for grammar corrections in our documents. This use was strictly confined to enhancing linguistic accuracy and improving the readability of our written materials. It's important to clarify that the core research, analysis, and development were conducted independently by our team.

Human Annotation. The human annotators involved in this project are professionals with expertise in computer science. No sensitive or personally identifiable data was used in the annotation process, adhering to ethical guidelines and data privacy standards. The human annotators are co authors on this paper.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David R. Mortensen, Noah A. Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? tokenization in the era of commercial language models.

Marvin Alberts, Gianmarco Gabrieli, and Irina Espejo Morales. 2024. Interleaving text and number embeddings to solve mathemathics problems.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Cohere. 2024. Command r+ model documentation. Comprehensive documentation and model details for the Command R+ model.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan

Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield,

Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models.

Bahare Fatemi, Mehran Kazemi, Anton Tsitsulin, Karishma Malkan, Jinyeong Yim, John Palowitch, Sungyong Seo, Jonathan Halcrow, and Bryan Perozzi. 2024. Test of time: A benchmark for evaluating llms on temporal reasoning.

Juan Luis Gastaldi, John Terilla, Luca Malagutti, Brian DuSell, Tim Vieira, and Ryan Cotterell. 2024. The foundations of tokenization: Statistical and computational concerns.

Omer Goldman, Avi Caciularu, Matan Eyal, Kris Cao, Idan Szpektor, and Reut Tsarfaty. 2024. Unpacking tokenization: Evaluating text compression and its correlation with model performance.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. 2024. Olmo: Accelerating the science of language models.

Shuhao Gu, Mengdi Zhao, Bowen Zhang, Liangdong Wang, Jijie Li, and Guang Liu. 2024. Retok: Replacing tokenizer to enhance representation efficiency in large language model.

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.

Sindhu Kishore and Hangfeng He. 2024. Unveiling divergent inductive biases of llms on temporal data.

Jingyu Liu, Jiaen Lin, and Yong Liu. 2024. How much can rag help the reasoning of llm?

Augustas Macijauskas. 2024. Are llama 3 and gpt-4 tokenizers the same?

Benjamin Minixhofer, Edoardo Maria Ponti, and Ivan Vulić. 2024. Zero-shot tokenizer transfer.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, Yuling Gu, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, Ali Farhadi, Noah A. Smith, Pang Wei Koh, Amanpreet Singh, and Hannaneh Hajishirzi. 2024. Olmoe: Open mixture-of-experts language models.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. Gpt-4o system card.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie

Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model.

Nived Rajaraman, Jiantao Jiao, and Kannan Ramchandran. 2024. Toward a theory of tokenization in llms.

Craig W. Schmidt, Varshini Reddy, Haoran Zhang, Alec Alameddine, Omri Uzan, Yuval Pinter, and Chris Tanner. 2024. Tokenization is more than compression.

Aaditya K. Singh and DJ Strouse. 2024. Tokenization counts: the impact of tokenization on arithmetic in frontier llms.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research.

Zhaochen Su, Jun Zhang, Tong Zhu, Xiaoye Qu, Juntao Li, Min Zhang, and Yu Cheng. 2024a. Timo: Towards better temporal reasoning for language models.

Zhaochen Su, Jun Zhang, Tong Zhu, Xiaoye Qu, Juntao Li, Min Zhang, and Yu Cheng. 2024b. Timo: Towards better temporal reasoning for language models.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023a. Towards benchmarking and improving the temporal reasoning capability of large language models.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023b. Towards robust temporal reasoning of large language models via a multi-hop qa dataset and pseudo-instruction tuning.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Maurice Weber, Daniel Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. 2024. Redpajama: an open dataset for training large language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024a. Large language models can learn temporal reasoning.

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024b. Large language models can learn temporal reasoning.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report.

Xiang Zhang, Juntai Cao, and Chenyu You. 2024. Counting ability of large language models and impact of tokenization.

Bowen Zhao, Zander Brumbaugh, Yizhong Wang, Hannaneh Hajishirzi, and Noah A. Smith. 2024. Set the clock: Temporal alignment of pretrained language models.

Chenghao Zhu, Nuo Chen, Yufei Gao, Yunyi Zhang, Prayag Tiwari, and Benyou Wang. 2024. Is your llm outdated? evaluating llms at temporal generalization.

A Appendix

B Contributions

Format	Model	Date	Year	Time Period	Century	TC	Tokenized Output	SI	SC	PS
MMDDYYYY	Baseline	10271606	1606	Historical (Pre-2000)	17th Century	3	10 27 1606	1.00	false	true
MMDDYYYY	OLMoE	10271606	1606	Historical (Pre-2000)	17th Century	4	10 27 16 06	0.66	true	true
MMDDYYYY	OLMo	10271606	1606	Historical (Pre-2000)	17th Century	4	10 27 16 06	0.66	true	true
MMDDYYYY	Llama 3	10271606	1606	Historical (Pre-2000)	17th Century	3	102 716 06	0.60	true	true
MMDDYYYY	Llama 3.1	10271606	1606	Historical (Pre-2000)	17th Century	3	102 716 06	0.60	true	true
MMDDYYYY	Llama 3.2	10271606	1606	Historical (Pre-2000)	17th Century	3	102 716 06	0.60	true	true
MMDDYYYY	Davinci-003	10271606	1606	Historical (Pre-2000)	17th Century	3	1027 16 06	0.60	true	true
MMDDYYYY	GPT-3.5	10271606	1606	Historical (Pre-2000)	17th Century	3	102 716 06	0.60	true	true
MMDDYYYY	GPT-4o	10271606	1606	Historical (Pre-2000)	17th Century	3	102 716 06	0.60	true	true
MMDDYYYY	GPT-4	10271606	1606	Historical (Pre-2000)	17th Century	3	102 716 06	0.60	true	true
MMDDYYYY	Cohere Aya	10271606	1606	Historical (Pre-2000)	17th Century	8	1 0 2 7 1 6 0 6	0.45	true	true
MMDDYYYY	Gemma	10271606	1606	Historical (Pre-2000)	17th Century	8	1 0 2 7 1 6 0 6	0.45	true	true
MMDDYYYY	DeepSeek	10271606	1606	Historical (Pre-2000)	17th Century	8	1 0 2 7 1 6 0 6	0.45	true	true
MMDDYYYY	Cohere	10271606	1606	Historical (Pre-2000)	17th Century	8	1 0 2 7 1 6 0 6	0.45	true	true
MMDDYYYY	Qwen	10271606	1606	Historical (Pre-2000)	17th Century	8	1 0 2 7 1 6 0 6	0.45	true	true
MMDDYYYY	Phi 3.5	10271606	1606	Historical (Pre-2000)	17th Century	9	_ 1 0 2 7 1 6 0 6	0.40	true	true
MMDDYYYY	Llama 2	10271606	1606	Historical (Pre-2000)	17th Century	9	_ 1 0 2 7 1 6 0 6	0.40	true	true
MMDDYYYY	Mistral	10271606	1606	Historical (Pre-2000)	17th Century	9	_ 1 0 2 7 1 6 0 6	0.40	true	true
MMDDYYYY	Llama 1	10271606	1606	Historical (Pre-2000)	17th Century	9	_ 1 0 2 7 1 6 0 6	0.40	true	true

Table 4: Generated by Spread-LaTeX

Type of Q	Question	Answer	Date Format	Llama3-70B	Qwen2.5-72B	GPT-4o	GPT-4-turbo	CMDR+
Factual	Which of the following famous people died on 23041616? Answer only with A, B, C, or D: A) Shah Jahan B) Miguel de Cervantes C) Princess Diana D) William Shakespeare	William Shakespeare **OR** Miguel de Cervantes	DDMMYYYY	B) Miguel de Cervantes and D) William Shakespeare	B, D	B) Miguel de Cervantes	D)William Shakespeare	A
Conceptual	The Eiffel Tower was built between 028/1887 to 090/1889. How long did it take to build the Eiffel Tower?	2 years, 2 months, and 3 days	DD/YYYY (Julian)	The Eiffel Tower took 2 years and 1 month to build.	2 years, 10 months, 2 days.	2 years and 1 month.	It took 2 years and 2 months to build.	Construction of the Eiffel Tower took about two years and two months.
Common Sense	A new electric car model is set to be released on 273/2040. Pre-orders open 6 months earlier. On what date can customers start preordering?	On March 30, 2040	DD/YYYY (Julian)	273/2040 - 6 months = 93/2039. Customers can start preordering on this date.	Pre-orders open on 273/2039.	Customers can start pre-ordering on March 273/2039.	Customers can start preordering on March 273, 2040.	The pre-order date is July 273, 2039.
Numerical	What is the time 7 years and 9 months after 10271446?	July 27, 1454	MMDDYYYY	October 27, 2040.	10271446 + 7 years 9 months = 10353406	October 2023 plus 7 years and 9 months is July 2031.	Time: 10429846 (Unix timestamp format).	10279141

Table 5: Model Performance on DateLogicQA

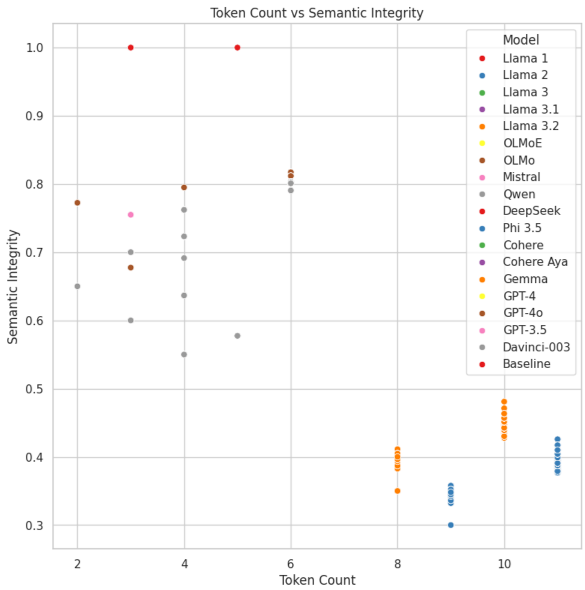


Figure 8: Correlation plot between semantic integrity score against token count

Sections	Contributors
1 - Introduction	All
2 - Related Works	All
3 - DateLogicQA	Cristina & Madiha
4.1 - Semantic Integrity	Tang
4.2 - Human-Led Temporal Bias Assessment	Madiha
4.3 - Understanding Temporal Bias	Gagan
5.1 - Impact of Tokenizers	Tang
5.2 - Temporal Reasoning Analysis	Cristina
5.3 - Temporal Sensitivity Analysis	Gagan
6 - Discussion	All
7 - Conclusion	All
Editing Document	All

Figure 9: Contributions Chart