TU Delft

Delft University of Technology

Temporal Blind Spots in Large Language Models

Wallat, Jonas; Jatowt, Adam; Anand, Avishek

DOI
10.1145/3616855.3635818

Publication date
2024

Published in
WSDM 2024 - Proceedings of the 17th ACM International Conference on Web Search and Data Mining

Citation (APA)
Wallat, J., Jatowt, A., & Anand, A. (2024). Temporal Blind Spots in Large Language Models. In *WSDM 2024 - Proceedings of the 17th ACM International Conference on Web Search and Data Mining* (pp. 683-692). ACM. https://doi.org/10.1145/3616855.3635818

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

Green Open Access added to TU Delft Institutional Repository

'You share, we take care!' - Taverne project

https://www.openaccess.nl/en/you-share-we-take-care

Otherwise as indicated in the copyright section: the publisher is the copyright holder of this work and the author uses the Dutch legislation to make this work public.

Temporal Blind Spots in Large Language Models

Jonas Wallat
L3S Research Center
Hannover, Germany
jonas.wallat@l3s.de

Adam Jatowt
Department of Computer Science
University of Innsbruck
Innsbruck, Austria
adam.jatowt@uibk.ac.at

Avishek Anand
Department of Software Technology
Delft University of Technology
Delft, The Netherlands
avishek.anand@tudelft.nl

ABSTRACT

Large language models (LLMs) have recently gained significant attention due to their unparalleled zero-shot performance on various natural language processing tasks. However, the pre-training data utilized in LLMs is often confined to a specific corpus, resulting in inherent freshness and temporal scope limitations. Consequently, this raises concerns regarding the effectiveness of LLMs for tasks involving temporal intents. In this study, we aim to investigate the underlying limitations of general-purpose LLMs when deployed for tasks that require a temporal understanding. We pay particular attention to handling factual temporal knowledge through three popular temporal QA datasets. Specifically, we observe low performance on detailed questions about the past and, surprisingly, for rather new information. In manual and automatic testing, we find multiple temporal errors and characterize the conditions under which QA performance deteriorates. Our analysis contributes to understanding LLM limitations and offers valuable insights into developing future models that can better cater to the demands of temporally-oriented tasks. The code is available[1].

CCS CONCEPTS

• **Information systems** → **Question answering**.

KEYWORDS

Temporal Information Retrieval, temporal query intents, question answering, large language models

ACM Reference Format:
Jonas Wallat, Adam Jatowt, and Avishek Anand. 2024. Temporal Blind Spots in Large Language Models. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM '24), March 4–8, 2024, Merida, Mexico.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3616855.3635818

1 INTRODUCTION

Autoregressive large language models (LLMs) like [8, 56] are versatile and a general-purpose solution to many natural language processing tasks. These models, benefiting from advanced natural language understanding capabilities, have demonstrated impressive zero-shot and few-shot performance. However, both pre-training and instruction-tuning data utilized in LLMs are often confined to a specific corpus and instruction demonstrations. However, there is a poor understanding of how much these large language models exhibit temporal knowledge and orientation in time. Consequently, this raises concerns regarding the effectiveness of LLMs for a range of tasks involving temporal intents like question answering and search over historical sources [60], QA over legal and personal temporal collections [23, 43], or fact checking [38].

`Which film won seven Oscars in 1994?`		
ChatGPT	Forrest Gump	✗ *Temporal shift:*
Correct	Schindler's List	F.G. won Oscars in **1995**
`Who lost the WBA boxing title, refusing to fight Tony Tucker in March 1995?`		
Alpaca	Mike Tyson	✗ *Time invariant:*
Correct	George Foreman	time disregarded
`Tom Brady played for which team in 2020?`		
Alpaca	New England Patriots	✗ *Temporal inertia*: Brady
Correct	Tampa Bay Buccaneers	joined Buccaneers in 2020

Table 1: Examples of temporal blindspots in LLMs.

Extensive research work towards understanding the capabilities and limitations of LLMs has focussed on knowledge probing [34, 42], adversarial examples [20], and risk analysis [16] among others. Probing methods have focussed on factual [34, 63], commonsense [55], social [67], numerical [37] and spatial [13] knowledge. However, most of the probing methods utilize bi-directional and non-instruction-tuned models. We do not focus on adversarial examples [20] but on natural questions that are incorrectly answered due to a lack of temporal understanding. Unlike the existing work, we analyze the temporal knowledge blind spots and temporal understanding of instruction-tuned LLMs for natural questions with a temporal intent. A few examples of temporally-induced errors on our datasets are shown in Table 1. To the best of our knowledge, our study is the first to inspect LLMs, their temporal blind spots, and their abilities in navigating over time.

1.1 Aim of this study

In this study, we aim to investigate the underlying limitations of general-purpose LLMs when deployed for tasks that require temporal knowledge and temporal understanding. Our research focuses on evaluating the abilities of LLMs in the context of temporal knowledge retrieval and processing and involves comprehensive testing

[1] https://github.com/jwallat/temporalblindspots

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
WSDM '24, March 4–8, 2024, Merida, Mexico
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0371-3/24/03…$15.00
https://doi.org/10.1145/3616855.3635818

over three temporal question-answering datasets: *TemporalQuestions* [58], *ArchivalQA* [60], and *TempLAMA* [18].

We pay particular attention to the handling of factual temporal knowledge and the processing of complex temporal information. Specifically, we experiment with different time-referencing schemes (absolute and relative), as well as experiment with the corruption of time references. Through this examination, we hope to shed light on the extent to which LLMs can effectively address the challenges associated with temporal knowledge management.

2 RELATED WORK

2.1 Temporal Aspects of Text

Several NLP and IR tasks have benefited from utilizing either of the two key temporal dimensions of texts (be it documents or queries), which are: *creation time* and *focus time* (aka. content time) [33]. The former refers to when a text was created (e.g., document timestamp or query issuing time), while the latter is the time mentioned or implicitly referred to in text, e.g., a document about WWII has the focus time of 1939 – 1945. Similarly, the focus time of a query "Winter Olympics 1988" would refer to when this sports event took place in Calgary (Feb 13 - Feb 28, 1988), and if this query were issued today, then its creation time would be the current day.

Temporal expressions embedded in a text may be relative, implicit, or underspecified, making their proper disambiguation challenging. In practice, texts may also contain sentences or paragraphs related to different time points (e.g., referring to multiple past events from different time periods). Hence, a document focus time may need to be represented as a set of time intervals [29]. Note that focus time is not always explicitly mentioned in the form of temporal expressions, which temporal taggers like SuTime [11] or HeidelTime [52] could extract (e.g., mentions of historical events from WWII without specifying any dates, or a query such as "Winter Olympics in Calgary"), in which case particular reasoning approaches may be required to anchor text over time dimension [29].

Exploiting the above-mentioned two kinds of temporal information has been gaining increased importance in temporal information retrieval and NLP, and their inter-relations have been utilized to develop various time-specific methods and applications [25, 26]. They have been applied to query and document matching in time-aware document ranking [22, 50] or temporal search in web archives [3]. More recently, documents' publication and content time have been harnessed in temporal question answering over longitudinal document collections [58] where questions could be time-scoped (contain explicit time expressions) or the time can be implicit. Other related works include search results diversification [7], clustering [54], summarization [6], document timestamping [59], and event ordering [27]. Additionally, Yamato et al. [65] explored changes in web facts' popularity over time, and Joho et al. [31] investigated temporal aspects in user search intents during web searches.

2.2 Time & Pre-trained Language Models

BERT [17] was one of the first breakthroughs contributing to the recent success of pre-trained language models. Its two pre-training tasks *masked language modeling* and *next sentence prediction* are time-agnostic, resulting in relatively poor temporal reasoning abilities of BERT and other subsequent models. Inspired by Salient Span Masking [24] and entity replacement incorporated into pretraining tasks [64], some researchers proposed to improve the performance of language models on various domain-specific tasks (e.g., entity-related tasks) by experimenting with the adaptation of pre-training tasks to specialize models for certain types of knowledge. Althammer et al. [2] proposed linguistically informed masking, demonstrating that paying attention to complex linguistic expressions benefits downstream tasks related to patent document processing.

Others have recently explored incorporating temporal knowledge into language models [18, 47, 48]. A simple modification to pre-training that parametrizes MLM objective with timestamp information using temporally-scoped knowledge has been proposed in [18] with the experiments conducted on a downstream task of question answering. Specifically, the authors focus on temporally scoped facts (e.g., "Cristiano Ronaldo played for Real Madrid in 2012", "Cristiano Ronaldo played for Juventus FC in 2019"), modifying the pretraining by transforming the input form with the prefixed temporal information (e.g., "year:2012 text: Cristiano Ronaldo plays for X" and "year:2019 text: Cristiano Ronaldo plays for X").

Rosin and Radinsky [48] and Rosin et al. [47] focus on the task of semantic change detection and characterization (i.e., detection of words that underwent semantic drift and calculation of the drift's extent), achieving a relatively modest improvement. The solution of Rosin and Radinsky [48] is to extend the self-attention mechanism by incorporating timestamp information to compute new attention scores, while the one of Rosin et al. [47] relies on training BERT by concatenating timestamp and text as input. The authors of [47] experiment also with the sentence time prediction task. Yet, their solution performs inferior to the fine-tuned BERT model.

More recently, Cole et al. [14] incorporated content time with the transformer encoder-decoder architecture (T5 model), masking the content time and experimenting with various temporal tasks. Lastly, Wang et al. [61] exploit both the creation and focus time during pre-training with transformer encoder-only architecture on a temporal news collection, experimenting on a range of tasks including temporal information retrieval, question answering over temporal collections, document timestamping, and event dating.

2.3 Temporal Knowledge Datasets

Quite a large number of question answering benchmarks have been introduced recently [5, 19]. The datasets for querying time-related factual knowledge are, however, relatively less common. NewsQA [57] is a machine reading comprehension dataset containing 119K text span answers created based on a collection of CNN news articles published over nine years (2007 - 2015). Questions in NewsQA require additional background knowledge from the original paragraphs, based on which they were generated, to be understood and correctly answered. Thus, they cannot be considered as forming a standalone open QA dataset.

ArchivalQA [60] is a large-scale open domain question answering dataset containing over half a million questions created from the New York Times archive [49] - a news article dataset that has been frequently used for temporal information retrieval [10, 33]. The *TemporalQuestions* dataset proposed in [58] covers the same time period. However, its questions concern relatively major and well-known events. For both ArchivalQA and TemporalQuestions,

Dataset	Example Question	Answer	#Qs	Scope	Type
TemporalQuestions	`Who did President Bush run against in 2004?`	John Kerry	1,000	1987-2007	major events
ArchivalQA	`What was Ankara's official aid bill for in 1997?`	Cyprus	60,000	1987-2007	detailed, news
TempLAMA	`Cristiano Ronaldo played for which team in 2020?`	Juventus FC	50,310	2010-2020	detailed, entities

Table 2: Overview of the temporal datasets used in this study.

half of the questions are time-scoped, while the other half do not contain any temporal expressions. Dhingra et al. [18] proposed *TempLAMA* - a dataset of temporally-scoped knowledge probes based on collecting subject-object relations from the 2020 Wikidata snapshot. The dataset contains over 50k data instances whose subjects and objects are both entities with their Wikipedia pages and which are in the form of manually written template cloze queries.

TempQuestions [30] contains 1,271 questions obtained by selecting time-related questions from other datasets[2] with additional curation and tagging of temporal cues. Another dataset, TimeSensitiveQA [12], provides about 40k questions of temporal nature obtained from wikidata after identifying time-evolving facts and considering four identifying common reasoning types ("in," "between," "before," "after"), requiring either event ordering or event locating information. Additionally, there is a growing number of datasets devoted to temporal commonsense reasoning [28, 62] rather than to asking about factual temporal knowledge; this task is, however, outside of the scope of our study. Finally, Temporal Information Retrieval datasets like [32] contain queries with their temporal search intents (e.g., past, future, temporal, or present).

2.4 Examining Abilities of LLMs

Existing literature on inspecting different types of knowledge contained in language models has been extensively researched in the last five years in the context of explainability and interpretability [4, 42]. Petroni et al. [42] first introduced the notion of *knowledge probing* that checked the ability of LMs to store factual knowledge about entities. Subsequently, many probing methods have been proposed to elicit factual [34, 63] or commonsense knowledge [55], including social [67], numerical [37] and spatial [13] knowledge. Jain et al. [28] analyzed temporal commonsense reasoning capabilities of LLMs on a range of datasets such as whether the models can correctly find typical time for certain actions, typical duration intervals, or common order. Different from existing work, we analyze the temporal knowledge, blind spots, and temporal understanding of LLMs. Other partially related lines of research suggest developing task-specific riskcards [16] for structured evaluation of LM risks in deployment scenarios. To the best of our knowledge, our study is the first to examine the temporal knowledge and abilities of LLMs. Existing research either takes an adversarial approach [41], focuses on specific risks [16], or does not center on LLMs [34, 42, 63].

3 STUDY DETAILS

3.1 Models

3.1.1 `alapaca-7B`. The `alapaca-7B` model[3] is an instruction-tuned derivative of the LLaMa foundation model [56]. The LLaMa model has been trained on the CommonCrawl[4], C4 [45], Github, August 2022 Wikipedia[5] dump, ArXiv, and StackExchange data. Then, the LLaMa model was trained to follow instructions using a set of 52k examples generated from OpenAI's `text-davinci-003` API. We use the smaller version with 7 billion parameters.

3.1.2 `text-davinci-003`. OpenAI's `text-davinci-003`[6] is a LLM from the GPT-3 [8] family. It was built from the InstructGPT model [39] and consists of 175B parameters. A mixture of training datasets has been used: a filtered version of CommonCrawl, an expanded version of WebText [44], two not further specified internet-based book corpora, and the English Wikipedia. The most recent information in `text-davinci-003`'s training data is from June 2021. To probe the model's parametric memory, we adapt the OpenAI playground's[7] default QA prompt to achieve shorter answers. For completeness, we add examples of our prompts with our code.

3.1.3 Open-Source LLMs. We further use multiple openly available instruction-tuned LLMs. `open-llama-7B`[8] [21] was trained on a mixture of Falcon Refined-Web [40], Starcoder [36] and parts of the RedPajama dataset [15]. `red-pajama-3B`[9] and `red-pajama-7B`[10] were trained on 1.5T token of the RedPajama dataset [15], and `falcon-7B`[11] [1] was trained on Falcon Refined-Web as well as a mixture of curated corpora. Since the same instruction format was used to train `open-llama-7B` and `alapaca-7B`, we use the identical prompt for both models. The remaining models will be tested using the `text-davinci-003` prompt, as shown in the supplementary material available with our code.

3.2 Datasets

This study uses three datasets containing time-scoped questions: TemporalQuestions [58], ArchivalQA [60], and the TempLAMA dataset [18]. The overview of the three datasets, selected examples, and further information is given in Table 2.

3.2.1 TemporalQuestions. TemporalQuestions dataset [58] contains 1,000 human-generated questions about major events where half is explicitly and half implicitly time-scoped, meaning that half the questions contain temporal expressions while the remaining questions lack any temporal references. The questions were carefully selected from several history quiz websites, existing QA datasets such as SQUAD 1.1 [46] and TempQuestions [30], or manually created from Wikipedia's year pages[12].

[2] Free917, WebQuestions and ComplexQuestions datasets
[3] https://github.com/tatsu-lab/stanford_alpaca
[4] http://www.commoncrawl.org/
[5] https://en.wikipedia.org/
[6] https://platform.openai.com/docs/models/gpt-3-5
[7] https://platform.openai.com/playground
[8] https://huggingface.co/VMware/open-llama-7b-v2-open-instruct
[9] https://huggingface.co/togethercomputer/RedPajama-INCITE-Instruct-3B-v1
[10] https://huggingface.co/togethercomputer/RedPajama-INCITE-7B-Instruct
[11] https://huggingface.co/tiiuae/falcon-7b
[12] E.g., https://en.wikipedia.org/wiki/1989

3.2.2 ArchivalQA. The ArchivalQA train split of time-scoped questions, which we use in this analysis, was generated from the NYT News Corpus using T5-base model [45] fine-tuned on SQUAD 1.1 [46]. The questions were subsequently subject to multiple stages of systematic filtering, including syntactic filtering (e.g., dropping too short/long questions, questions with answers embedded in their content, ones with unresolved pronouns, and so on) as well as filtering questions that are not specific enough (overly general questions) and temporally ambiguous ones (i.e., questions with multiple possible answers over time) based on the application of several dedicated classifiers that were trained on specially prepared datasets.

ArchivalQA mostly contains detailed questions about the past, often on minor events, focussing on the period of 1987-2007, i.e., the time scope covered by the NYT corpus. Like TemporalQuestions, ArchivalQA contains a mixture of questions, including absolute time references and questions lacking any temporal expressions.

3.2.3 TempLAMA. The TempLAMA dataset is a set of KG triples for nine relations, such as "plays for" or "head of government." It covers more recent information from 2010 to 2020 than the other two datasets. While the first two datasets are distributed as question-answering datasets, TempLAMA is designed as a cloze task. Therefore, we reformulated the nine relations used in TempLAMA to actual question-answering pairs[13].

3.2.4 Evaluation. We evaluate the generated responses with the standard exact match (EM) and F1 score. Additionally, since the LLMs in our experiments tended to generate longer answers (leading to lower EM/F1 scores), we included a "contains" metric. This metric utilizes string matching to check if the answer is contained in the generated text (after removing punctuation and lowercasing). Lastly, we report the BERT-based answer equivalence metric (BEM) [9]. BEM utilizes a BERT model fine-tuned to detect equivalent answers by semantic (and not token) matching.

4 ANALYSIS

4.1 Temporal Knowledge of LLMs

First, we assess the general performance of the models on our temporal datasets. The results of several LLMs on the three datasets in this study can be seen in Table 3. Not surprisingly, we find the much bigger `text-davinci-003` to perform relatively better than the smaller models. We, however, generally observe relatively poor performance of our models, even on rather simple questions about major events from the past (TemporalQuestions). On the two other datasets, the performance of the LLMs is generally more lacking, and the numbers are quite low. As mentioned, ArchivalQA contains more fine-grained questions, i.e., ones about relatively minor or detailed events, than TemporalQuestions, while both cover the same time period. In the context of the smaller open-source models, we observe `red-pajama-7B` to perform relatively well among those models at answering temporally scoped questions.

Insight. The results suggest that LLMs exhibit limited ability in answering questions about the past and, overall, seem to lack knowledge regarding specific details of past events.

[13]Details in the supplementary material available with the code

4.2 Do LLMs Prioritize Recent Knowledge?

To further understand the performance of the LLMs on temporal datasets, we investigate whether the LLMs prioritize temporally newer over older knowledge. In other words, we wish to analyze the effect of the passage of time on the quality of answers. Other than existing related work (e.g., [35]), we do not investigate the model's behavior on new text produced after the model was trained, but how the model remembers historical information. To do so, we select those questions from the $60k$ questions of ArchivalQA that contain absolute year references. Next, we stratify the performance based on the years of the resulting set of around $27k$ QA pairs (Figure 1[14]). We see a slight trend of improvement over time, with the best results for the more recent years (2005-2007). This is somewhat expected and could be likely due to the forgetting effect (declining remembering of older years compared to the remembering of recent years), the trend that was also observed in large news article collections [66] and collections of tweets related to history [53].

Given that ArchivalQA does not cover the latest years, we next investigate the results from 2010 to 2020 repeating the same experiment on TempLAMA, keeping in mind that our models were trained on information up until late 2021 (`text-davinci-003`), mid 2022 (`alapaca-7B`), 2023 (others). The stratified per-year performance on the TempLAMA dataset can be seen in Figure 1b. Here, we also observe a trend of more recent information being better captured until the performance peaks in 2015 and 2017, respectively. Then, a decrease is seen for more recent information (2017-2020). We hypothesize that this recent decline could be due to older facts being more prevalent in the training data. At the same time, the newer information may not yet be sufficiently common in those data to be preferred by the models when answering questions on the latest events. Such *temporal inertia* can be expected as information within large document collections like the web is not being changed instantaneously at all relevant places but is rather gradually becoming updated. Furthermore, we believe that the models do not correctly model the temporality of the training data. If the models correctly recognize and adequately utilize the temporal signals (both publication and content time) in the documents used for training, they could prefer more recent information over older ones. In such a somewhat ideal case, even a few documents (or, in theory, even a single reliable document) should be enough for the models to update their parametric knowledge with more recent facts and thus correctly answer questions on dynamic topics.

Insight. LLMs seem to capture more recent information better than older information. However, such a pattern appears to occur only up to a certain point, as we observe in the case of TempLama (2017). Temporal inertia could be a possible explanation here, such that the most recent information is still not sufficiently prevalent in the training data of language models, while the models do not correctly distinguish between up-to-date and obsolete knowledge.

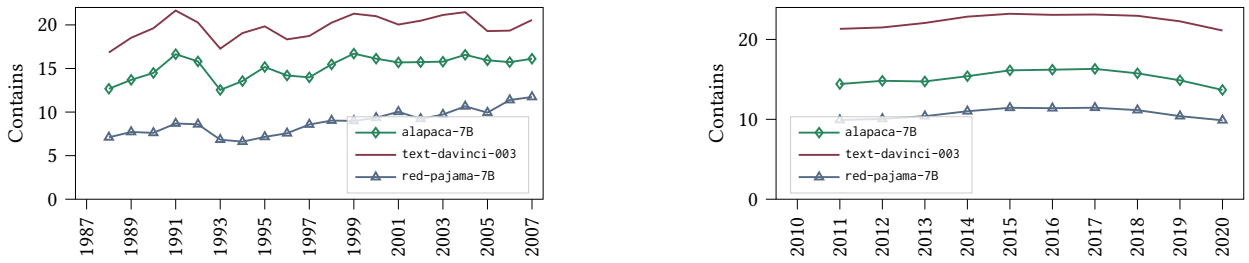
4.3 Relative vs. Absolute Time References

Next, we investigate to what extent the models are sensitive to the type of time information contained in questions. We start by

[14]Similar trends across other models and metrics are reported in the supplementary materials available at github.com/jwallat/temporalblindspots

Model	TemporalQuestions				ArchivalQA				TempLAMA			
	BEM	Cont.	F1	EM	BEM	Cont.	F1	EM	BEM	Cont.	F1	EM
text-davinci-003	**75.6**	**66.8**	**64.0**	**52.0**	**30.5**	**21.7**	**20.7**	**10.7**	**29.6**	**22.2**	**30.7**	**16.3**
alapaca-7B	57.1	46.2	37.2	26.6	30.0	16.5	11.0	4.4	28.2	15.1	16.4	3.8
open-llama-7B	28.7	23.1	22.9	16.5	14.7	9.4	7.8	3.3	12.0	7.0	12.7	3.5
falcon-7B	32.6	26.4	26.2	16.2	13.4	7.5	8.7	2.8	10.5	6.5	12.8	3.0
red-pajama-7B	42.6	35.1	40.2	33.7	14.5	9.3	11.1	6.4	15.7	10.6	19.0	11.0
red-pajama-3B	27.1	22.5	25.8	20.3	12.4	7.2	8.4	3.9	11.4	7.0	10.8	4.6

Table 3: Overall results of our models on the three temporal datasets - TemporalQuestions, the time-split of ArchivalQA, and the reformulated TempLAMA examples.



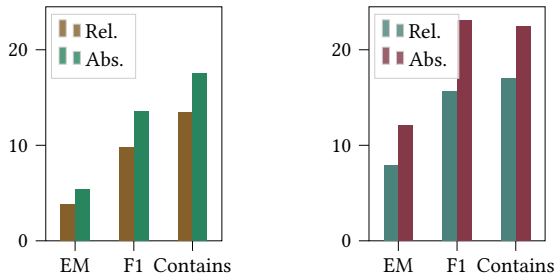
(a) ArchivalQA **(b) TempLAMA**

Figure 1: Stratification of the alapaca-7B, red-pajama-7B, and text-davinci-003 models on the ArchivalQA and TempLAMA datasets. Stratified by years, the trendline is the moving average with a window of 2. We do not show plots for the TemporalQuestions dataset since the dataset is not large enough for computing individual results per year.

comparing the usefulness of relative and absolute time references. We first experiment with absolute temporal expressions, but we later convert them into relative ones to see if there would be any change in results. As in the latter case, the temporal distances are now made to be relative, we expect the models to perform worse. This is because resolving back relative expressions to their absolute form requires certain rudimentary calculations, which we do not expect the models to always perform correctly (or to perform at all). On the other hand, if reasoning and knowledge retrieval done by the models were based purely on relative temporal expressions (without converting them to the absolute form), the models would not be effective either since the actual meaning of a relative temporal expression depends on the publication date of its containing document[15] - the type of information that does not seem explicitly utilized in the current mainstream LLMs.

Given a question such as "Who was the American president in 2018?" we then change the absolute reference of "2018" to a relative reference ("3 years ago"). Since both models responded 2021 when asked for which year it is now[16], we use 2021 as a reference year to compute all the relative references. The results on the subset of 27k ArchivalQA QA pairs that originally contained absolute year references can be seen in Figure 2.

We observe that the performance drops by 23-30% and 24-35% for alapaca-7B and text-davinci-003, respectively. This might be due to relative references requiring an additional reasoning



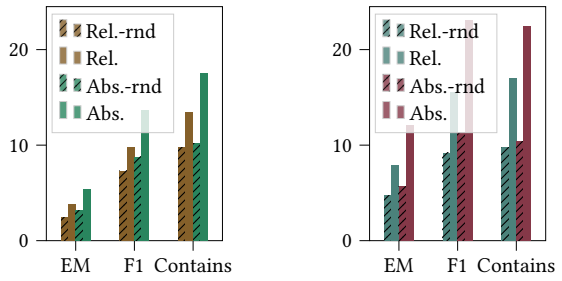
(a) alapaca-7B **(b) text-davinci-003**

Figure 2: Relative and absolute time referencing.

step. However, through qualitative tests, we have found the models (text-davinci-003, alapaca-7B) to be able to perform this reasoning when asked explicitly[17]. Yet, it seems that the models prefer absolute time references in arbitrary questions - potentially to match the same absolute reference tokens in the training data.

Given the studied language models have been trained without explicitly considering the publication time information, relative references in the training data are then less trustworthy (a mention of "2 years ago" is only useful when one knows when the document containing it was written), the models might learn to disregard these references and focus on matching absolute ones. Therefore, we find (relative) *time referencing* to be a blind spot in LLMs.

[15] Or, in some cases, on absolute time expressions stated earlier within the content of the document.
[16] Asking with text: "What year do we have?"
[17] We asked: "If today is 2021, what year was 4 years ago?"



(a) `alapaca-7B` (b) `text-davinci-003`

Figure 3: Effect of randomized relative and absolute time references. Textured bars show the randomized variants.

Insight. LLMs perform up to 35% worse when given relative time references (questions referring to time such as "4 years ago") than when absolute date expressions are used. Users are advised to use absolute temporal references when querying current LLMs.

4.4 How Important is Time Referencing?

Given that the performance of the models varies with the type of time expressions, we further investigate the importance of time references. To do so, we corrupt the time references by replacing a reference with a random year or by deliberately changing it to be off by a specific number of years.

4.4.1 Does Corruption of Time References has any Effect? First, Figure 3 shows the performance of `text-davinci-003` and the `alapaca-7B` models given absolute and relative time references with both the correct and random years. For these experiments, we sample 3k QA pairs from the 27k ArchivalQA questions containing an absolute year reference. For the random setting, we replace the years with random years between 1900 and 2021. After randomizing the time references, we observe the model performance to drop by an average of 26% and 44% (relative) as well as 40% and 53% (absolute) for `alapaca-7B` and `text-davinci-003`, respectively. Interestingly, randomizing the absolute time references has a similar impact on the model performance as when switching from absolute to relative references. However, randomizing the relative time references still results in decreased performance, indicating that relative references are not entirely useless or an active distraction.

Insight. Randomizing the time reference leads to a performance decrease of 40% (`alapaca-7B`) and 53% (`text-davinci-003`).

4.4.2 How Robust are LLMs Under Wrong Time References? Lastly, we more closely analyze the effect of errors in temporal references. In real life, users may misremember or confuse years; hence, such cases are not completely unrealistic. To do so, we corrupt the time references of the 27k ArchivalQA questions containing years to be wrong by X-years and observe the impact on model performance. That is, we change the time references by X years (e.g., 2014 to 2011). The results are shown in Figure 5.

Slight imprecisions in the time references only have minor effects. When asking the model about the past, being off by three years reduces the model's performance by 3% (relative reference) and 10% (absolute reference). More significant deviations lead to performance converging towards the performance seen with randomized references. Corrupting the time expression to be off by 20 years reduces `alapaca-7B`'s performance by 33-35% for relative and absolute expressions. We note that the performance degrades with the degree of the introduced error, suggesting that it is harder for the models to recover from larger errors. When comparing the effect of corrupting the references to removing the entire time reference, we observe that relative references distract the model.

Insight. The amount of error in the time references embedded in questions appears to correlate with the answer quality. The larger the introduced error, the worse the results. Not using time references at all seems to be preferential over relative references.

5 TEMPORAL ERRORS

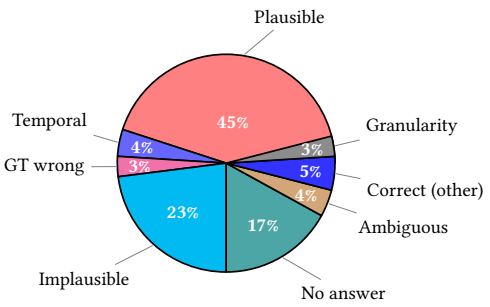
5.1 Error Analysis

Given the rather low performance of our models on the temporal datasets, we now investigate the possible reasons for failures. In the first open coding step [51], we use 100 wrong `alapaca-7B` predictions to identify failure cases. Next, we manually annotate 100 random, wrong answers of `alapaca-7B` and `text-davinci-003` on the ArchivalQA dataset concerning their failure types (see Figure 4).

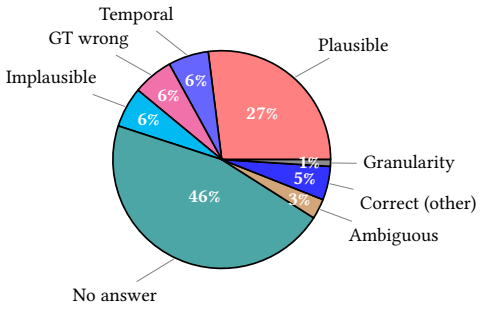
For nearly half of `alapaca-7B`'s wrong answers, we observe factually wrong yet plausible answers: These are usually questions where the model gives the **correct answer type but factually wrong entity** (such as a person or a location). As part of the plausible answers, we also find a set of *temporally shifted* entities as answers, i.e., successor or predecessor of the correct entity. Next, we observe a set of **answers that represent uncertainty in the model** ("unknown", "I do not know", etc.), followed by **technically correct answers with a different granularity than the ground truth** (e.g., answering with a country instead of a city when asked where a person was born). We also found a set of **implausible answers** where the model did not produce the right type of answer but repeated the question or produced incoherent and wrong text. Lastly, we also observed a few cases of **wrong ground truth answers**, which stem from the ArchivalQA dataset being automatically generated and cleaned. When inspecting the manual analysis results for `text-davinci-003`, we observe many questions that were not answered - suggesting that `text-davinci-003` is better calibrated not to answer in case of uncertainty. There are also fewer implausible answers and a similar amount of temporal errors.

Insight. Temporal hallucinations occur when `alapaca-7B` and `text-davinci-003` answer temporally-scoped questions.

5.1.1 Effect of Question Words. Next, we analyze the performance of our model on different types of questions. To do so, we stratify the performance on ArchivalQA on the question's leading words (e.g., who, where, what) and present the results in Figure 6. It is apparent that both models' performance on the question words "how" and "when" is much worse than for the other question words (i.e., "which," "where," "who," and "what"). We conjecture that "when" (very time-focused) and "how" (procedural or ambiguous) are more difficult than the other questions. Especially "when" requires the



(a) `alapaca-7B`



(b) `text-davinci-003`

Figure 4: Manual analysis of 100 of `alapaca-7B`'s and `text-davinci-003`'s wrong answers from the ArchivalQA dataset.

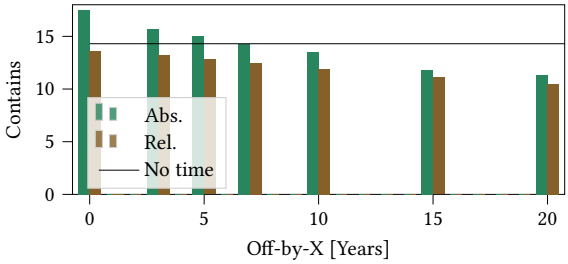


Figure 5: The effect of corrupted time of the `alapaca-7B` model. The relative time reference is calculated via a reference year (2021). Off-by-X means that the year is X years apart from the correct year. No time refers to the effect of entirely removing the time reference from the questions.

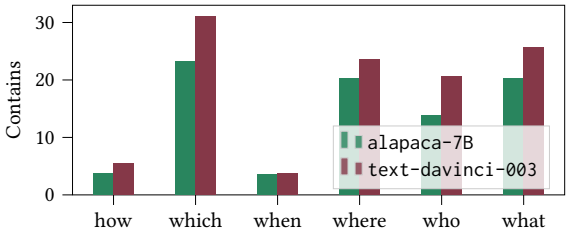


Figure 6: Performance of `alapaca-7B` and `text-davinci-003` on ArchivalQA stratified by the leading words of questions.

model to be aware of the time so as not to ignore the time references (both in questions and in training documents) compared to other question types for which it might be less critical, as we saw in our time reference randomization experiments (Figure 3). Interestingly, we find `text-davinci-003` to perform similarly to the much smaller `alapaca-7B` on the temporal "when" questions.

Insight. `alapaca-7B` and `text-davinci-003` perform poorly on questions requiring explicit temporal knowledge ("when").

5.2 Types of Temporal Errors

Next, we use our experiments and the manual error analysis to characterize the types of temporal errors exhibited by the models. Our analysis revealed four potential categories of temporal errors: *temporal shifts, time invariance, temporal inertia, and referencing errors*. Table 4 contains examples of these temporal errors.

5.2.1 Temporal Shifts. Temporal shifts occur when the model incorrectly disambiguates the actual time of concern. For instance, when asked about the winners of the Oscars in 1994, the model might mistakenly predict the winners of 1995 instead. This error highlights the model's difficulty in accurately determining and aligning the specific temporal context with the question's intent.

5.2.2 Time Invariance. Biases due to popularity arise from a strong semantic association between the answer entity and an entity mentioned in the question. This association often leads the model to disregard the temporal constraints specified in the question. In some cases, this bias can even override entity type constraints. For example, our analysis showed that the model incorrectly predicted Mike Tyson as the answer due to the strong association between boxing and Tyson, disregarding the temporal context.

5.2.3 Temporal Inertia. Furthermore, we hypothesized that statistical support for certain past entity-entity relationships could cause LLMs to answer questions about the recent past incorrectly. We attributed these potential statistical errors to temporal inertia. This suggests that past statistical patterns might influence the model's predictions, resulting in inaccurate responses when the more recent past deviates from those patterns.

5.2.4 Referencing Errors. Referencing errors occur due to an incorrect understanding of the current time. Table 4 shows an example of incorrect referencing of the past time. Also, future time, defined as temporal intents after the model's training time, is beyond the scope of the model's parametric memory. Hence, the model is expected not to predict future events accurately.

5.2.5 Measuring Temporal Errors. We use the TempLAMA dataset to measure temporal errors since it contains ground truth for neighboring years, allowing us to test for temporal shifts. Specifically, we estimate temporal shifts by looking at examples where the object of the relation changes (e.g., a player changes his team) and whether

Model	Question	Temporal error	Explanation
alapaca-7B Answer	Tom Brady played for which team in 2020? New England Patriots Tampa Bay Buccaneers	Temporal inertia	Brady joined Tampa (2020) after 20 years with Patriots
red-pajama-7B Answer	Cristiano Ronaldo played for which team in 2019? Real Madrid Juventus FC	Time invariance	The model predicts Real Madrid for all years
alapaca-7B Answer	In the 2003 Wimbledon Men's Singles Tennis Championship, who beat Tim Henman? Andy Roddick Sebastien Grosjean	Temporal shift	Roddick beat Henman in the following year (2004)
text-davinci-003 Answer	Who painted a portrait of George Washington for Martha in 1789? Gilbert Stuart John Ramage	Temporal shift	Stuart did paint Washington but in 1795 and 1796
text-davinci-003 Answer	What state did Tuscany join 162 years ago? Unknown. Italy	Temporal referencing	correct answer of Italy with absolute reference (1859)

Table 4: Examples of temporal blind spots.

the model predicts the next or previous object. Similarly, we investigate the extent of temporal inertia via the rate at which models fail to adapt to the last relation change of a subject. The amount of time-invariant relations is estimated by the ratio of relations for which the model always predicts the same output, no matter the year specified in the question. Lastly, we report the performance degradation of changing to relative references on the ArchivalQA dataset. The results are presented in Table 5.

Temp. Error	alapaca-7B	text-davinci-003	red-pajama-7B
Shift	5.0%	6.3%	**3.9%**
Inertia	12.7%	18.8%	**9.2%**
Invariance	**21.4%**	62.5%	66.9%
Referencing	30%	35%	**11%**

Table 5: Estimation of temporal error occurrences.

We observe that all categories of temporal errors frequently occur in our LLMs - posing new opportunities to improve the temporal abilities of LLMs. While temporal shifts seem rare, a relevant portion of facts is answered wrongly because the model did not incorporate new knowledge in its parametric memory (inertia). The most pronounced seems to be an invariance toward the time reference for many relations for all three models and the negative effects of using the relative time referencing. The latter is causing the performance to degrade by 11-35%. While we offer a novel framework for temporal errors of LLMs and estimate their prevalence, a more fine-grained analysis of the causes will be necessary to alleviate these and to build capable and trustworthy language models. This, however, is out of the scope of this work.

6 DISCUSSION AND CONCLUSIONS

In this study, we aimed to identify the degree of temporal understanding in Language Models (LLMs) by evaluating their performance on temporal QA tasks. Our findings shed light on several vital aspects of LLMs' temporal understanding abilities, offering insights into their limitations and blind spots. First, we observed that LLMs exhibit lower effectiveness in answering questions about the past, particularly when detailed information is required. This indicates a relative weakness in their ability to recall accurately and reason about temporal aspects of historical events (cf. Table 3). Furthermore, our investigation revealed that LLMs tend to capture more recent information better than older information up to a certain point. However, we hypothesize that the statistical support of older information can sometimes overshadow newer information, leading to a potential decline in performance (Figure 1). To address this drawback, we propose that modeling the training data's creation and focus time could improve LLMs' temporal comprehension capabilities. We also find that relative time references proved more challenging for LLMs than absolute references. Task performance dropped by up to 35% when relative time references were used, as demonstrated in Figure 2. Moreover, randomizing the absolute time reference had a similar detrimental effect on model performance, resulting in a decrease of up to 55%. By corrupting the time references, we showed that using relative, or absolute references that are wrong by more than a few years, results in worse model performance than using no time cues. In an extensive manual analysis, we find multiple temporal errors such as temporal shifts or the inability to update the parametric memory - and device automatic tests to detect them. Our research underscores the importance of further improving LLMs' temporal understanding abilities.

ACKNOWLEDGMENTS

This research was partially funded by the Federal Ministry of Education and Research (BMBF), Germany under the project LeibnizKILabor with grant No. 01DD20003 and Cubra with grant No. 13N16052. It is also partially supported by the EU – Horizon 2020 Program, Grant Agreement n.871042, "SoBigData++".

REFERENCES

[1] E. Almazrouei, H. Alobeidli, A. Alshamsi, A. Cappelli, R. Cojocaru, M. Debbah, E. Goffinet, D. Heslow, J. Launay, Q. Malartic, B. Noune, B. Pannier, and G. Penedo. Falcon-40B: an open large language model with state-of-the-art performance. 2023.

[2] S. Althammer, M. Buckley, S. Hofstätter, and A. Hanbury. Linguistically informed masking for representation learning in the patent domain. *CoRR*, abs/2106.05768, 2021.

[3] A. Anand, S. J. Bedathur, K. Berberich, and R. Schenkel. Index maintenance for time-travel text search. In W. R. Hersh, J. Callan, Y. Maarek, and M. Sanderson, editors, *The 35th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '12, Portland, OR, USA, August 12-16, 2012*, pages 235–244. ACM, 2012.

[4] A. Anand, L. Lyu, M. Idahl, Y. Wang, J. Wallat, and Z. Zhang. Explainable information retrieval: A survey. *CoRR*, abs/2211.02405, 2022.

[5] R. Baradaran, R. Ghiasi, and H. Amirkhani. A survey on machine reading comprehension systems. *Nat. Lang. Eng.*, 28(6):683–732, 2022.

[6] C. Barros, E. Lloret, E. Saquete, and B. Navarro-Colorado. NATSUM: narrative abstractive summarization through cross-document timeline generation. *Inf. Process. Manag.*, 56(5):1775–1793, 2019.

[7] K. Berberich and S. Bedathur. Temporal diversification of search results. In *Proceedings of SIGIR 2013 workshop on time-aware information access*, 2013.

[8] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[9] J. Bulian, C. Buck, W. Gajewski, B. Börschinger, and T. Schuster. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 291–305. Association for Computational Linguistics, 2022.

[10] R. Campos, G. Dias, A. M. Jorge, and A. Jatowt. Survey of temporal information retrieval and related applications. *ACM Comput. Surv.*, 47(2):15:1–15:41, 2014.

[11] A. X. Chang and C. D. Manning. Sutime: A library for recognizing and normalizing time expressions. In N. Calzolari, K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 3735–3740. European Language Resources Association (ELRA), 2012.

[12] W. Chen, X. Wang, and W. Y. Wang. A dataset for answering time-sensitive questions. *CoRR*, abs/2108.06314, 2021.

[13] A. G. Cohn and J. Hernández-Orallo. Dialectical language model evaluation: An initial appraisal of the commonsense spatial reasoning abilities of llms. *CoRR*, abs/2304.11164, 2023.

[14] J. R. Cole, A. Chaudhary, B. Dhingra, and P. Talukdar. Salient span masking for temporal understanding. pages 3044–3052, 2023.

[15] T. Computer. Redpajama: An open source recipe to reproduce llama training dataset, 2023.

[16] L. Derczynski, H. R. Kirk, V. Balachandran, S. Kumar, Y. Tsvetkov, M. R. Leiser, and S. Mohammad. Assessing language model deployment with risk cards. *CoRR*, abs/2303.18190, 2023.

[17] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[18] B. Dhingra, J. R. Cole, J. M. Eisenschlos, D. Gillick, J. Eisenstein, and W. W. Cohen. Time-aware language models as temporal knowledge bases. *Trans. Assoc. Comput. Linguistics*, 10:257–273, 2022.

[19] D. Dzendzik, J. Foster, and C. Vogel. English machine reading comprehension datasets: A survey. pages 8784–8804, 2021.

[20] D. Ganguli, L. Lovitt, J. Kernion, A. Askell, Y. Bai, S. Kadavath, B. Mann, E. Perez, N. Schiefer, K. Ndousse, A. Jones, S. Bowman, A. Chen, T. Conerly, N. DasSarma, D. Drain, N. Elhage, S. E. Showk, S. Fort, Z. Hatfield-Dodds, T. Henighan, D. Hernandez, T. Hume, J. Jacobson, S. Johnston, S. Kravec, C. Olsson, S. Ringer, E. Tran-Johnson, D. Amodei, T. Brown, N. Joseph, S. McCandlish, C. Olah, J. Kaplan, and J. Clark. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *CoRR*, abs/2209.07858, 2022.

[21] X. Geng and H. Liu. Openllama: An open reproduction of llama, May 2023.

[22] D. Gupta and K. Berberich. Identifying time intervals of interest to queries. In J. Li, X. S. Wang, M. N. Garofalakis, I. Soboroff, T. Suel, and M. Wang, editors, *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*, pages 1835–1838. ACM, 2014.

[23] J. P. Gupta, Z. Qin, M. Bendersky, and D. Metzler. Personalized online spell correction for personal search. In L. Liu, R. W. White, A. Mantrach, F. Silvestri, J. J. McAuley, R. Baeza-Yates, and L. Zia, editors, *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 2785–2791. ACM, 2019.

[24] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR, 2020.

[25] H. Holzmann and A. Anand. Tempas: Temporal archive search based on tags. In J. Bourdeau, J. Hendler, R. Nkambou, I. Horrocks, and B. Y. Zhao, editors, *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11-15, 2016, Companion Volume*, pages 207–210. ACM, 2016.

[26] H. Holzmann, W. Nejdl, and A. Anand. On the applicability of delicious for temporal search on web archives. In R. Perego, F. Sebastiani, J. A. Aslam, I. Ruthven, and J. Zobel, editors, *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, pages 929–932. ACM, 2016.

[27] O. Honovich, L. T. Hennigen, O. Abend, and S. B. Cohen. Machine reading of historical events. In D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7486–7497. Association for Computational Linguistics, 2020.

[28] R. Jain, D. Sojitra, A. Acharya, S. Saha, A. Jatowt, and S. Dandapat. Do language models have a common sense regarding time? revisiting temporal commonsense reasoning in the era of large language models. In *Proceedings of 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1–25, 2023.

[29] A. Jatowt, C. A. Yeung, and K. Tanaka. Estimating document focus time. In Q. He, A. Iyengar, W. Nejdl, J. Pei, and R. Rastogi, editors, *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013*, pages 2273–2278. ACM, 2013.

[30] Z. Jia, A. Abujabal, R. S. Roy, J. Strötgen, and G. Weikum. Tempquestions: A benchmark for temporal question answering. In P. Champin, F. Gandon, M. Lalmas, and P. G. Ipeirotis, editors, *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018*, pages 1057–1062. ACM, 2018.

[31] H. Joho, A. Jatowt, and R. Blanco. A survey of temporal web search experience. In L. Carr, A. H. F. Laender, B. F. Lóscio, I. King, M. Fontoura, D. Vrandecic, L. Aroyo, J. P. M. de Oliveira, F. Lima, and E. Wilde, editors, *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013, Companion Volume*, pages 1101–1108. International World Wide Web Conferences Steering Committee / ACM, 2013.

[32] H. Joho, A. Jatowt, and R. Blanco. NTCIR temporalia: a test collection for temporal information access research. In C. Chung, A. Z. Broder, K. Shim, and T. Suel, editors, *23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014, Companion Volume*, pages 845–850. ACM, 2014.

[33] N. Kanhabua and A. Anand. Temporal information retrieval. In R. Perego, F. Sebastiani, J. A. Aslam, I. Ruthven, and J. Zobel, editors, *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, pages 1235–1238. ACM, 2016.

[34] N. Kassner, P. Dufter, and H. Schütze. Multilingual LAMA: investigating knowledge in multilingual pretrained language models. In P. Merlo, J. Tiedemann, and R. Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 3250–3258. Association for Computational Linguistics, 2021.

[35] A. Lazaridou, A. Kuncoro, E. Gribovskaya, D. Agrawal, A. Liska, T. Terzi, M. Gimenez, C. de Masson d'Autume, T. Kociský, S. Ruder, D. Yogatama, K. Cao, S. Young, and P. Blunsom. Mind the gap: Assessing temporal generalization in neural language models. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 29348–29363, 2021.

[36] R. Li, L. B. Allal, Y. Zi, N. Muennighoff, D. Kocetkov, C. Mou, M. Marone, C. Akiki, J. Li, J. Chim, Q. Liu, E. Zheltonozhskii, T. Y. Zhuo, T. Wang, O. Dehaene, M. Davaadorj, J. Lamy-Poirier, J. Monteiro, O. Shliazhko, N. Gontier, N. Meade, A. Zebaze, M. Yee, L. K. Umapathi, J. Zhu, B. Lipkin, M. Oblokulov, Z. Wang, R. M. V, J. Stillerman, S. S. Patel, D. Abulkhanov, M. Zocca, M. Dey, Z. Zhang, N. Moustafa-Fahmy, U. Bhattacharyya, W. Yu, S. Singh, S. Luccioni, P. Villegas, M. Kunakov, F. Zhdanov, M. Romero, T. Lee, N. Timor, J. Ding, C. Schlesinger, H. Schoelkopf, J. Ebert, T. Dao, M. Mishra, A. Gu, J. Robinson, C. J. Anderson, B. Dolan-Gavitt, D. Contractor, S. Reddy, D. Fried, D. Bahdanau, Y. Jernite, C. M. Ferrandis, S. Hughes, T. Wolf, A. Guha, L. von Werra, and H. de Vries. Starcoder:

may the source be with you! *CoRR*, abs/2305.06161, 2023.

[37] B. Y. Lin, S. Lee, R. Khanna, and X. Ren. Birds have four legs?! numersense: Probing numerical commonsense knowledge of pre-trained language models. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6862–6868. Association for Computational Linguistics, 2020.

[38] P. Nakov, D. P. A. Corney, M. Hasanain, F. Alam, T. Elsayed, A. Barrón-Cedeño, P. Papotti, S. Shaar, and G. D. S. Martino. Automated fact-checking for assisting human fact-checkers. In Z. Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4551–4558. ijcai.org, 2021.

[39] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.

[40] G. Penedo, Q. Malartic, D. Hesslow, R. Cojocaru, A. Cappelli, H. Alobeidli, B. Pannier, E. Almazrouei, and J. Launay. The refinedweb dataset for falcon LLM: outperforming curated corpora with web data, and web data only. *CoRR*, abs/2306.01116, 2023.

[41] E. Perez, S. Huang, H. F. Song, T. Cai, R. Ring, J. Aslanides, A. Glaese, N. McAleese, and G. Irving. Red teaming language models with language models. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3419–3448. Association for Computational Linguistics, 2022.

[42] F. Petroni, T. Rocktäschel, S. Riedel, P. S. H. Lewis, A. Bakhtin, Y. Wu, and A. H. Miller. Language models as knowledge bases? In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2463–2473. Association for Computational Linguistics, 2019.

[43] Z. Qin, Z. Li, M. Bendersky, and D. Metzler. Matching cross network for learning to rank in personal search. In Y. Huang, I. King, T. Liu, and M. van Steen, editors, *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 2835–2841. ACM / IW3C2, 2020.

[44] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. 2019.

[45] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020.

[46] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. Squad: 100, 000+ questions for machine comprehension of text. pages 2383–2392, 2016.

[47] G. D. Rosin, I. Guy, and K. Radinsky. Time masking for temporal language models. In K. S. Candan, H. Liu, L. Akoglu, X. L. Dong, and J. Tang, editors, *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*, pages 833–841. ACM, 2022.

[48] G. D. Rosin and K. Radinsky. Temporal attention for language models. pages 1498–1508, 2022.

[49] E. Sandhaus. The new york times annotated corpus. ldc2008t19. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752, 2008.

[50] J. Singh, W. Nejdl, and A. Anand. History by diversity: Helping historians search news archives. *CoRR*, abs/1810.10251, 2018.

[51] A. Strauss and J. M. Corbin. Basics of qualitative research : techniques and procedures for developing grounded theory. 1998.

[52] J. Strötgen and M. Gertz. Heideltime: High quality rule-based extraction and normalization of temporal expressions. In K. Erk and C. Strapparava, editors, *Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval@ACL 2010, Uppsala University, Uppsala, Sweden, July 15-16, 2010*, pages 321–324. The Association for Computer Linguistics, 2010.

[53] Y. Sumikawa and A. Jatowt. Analyzing history-related posts in twitter. *Int. J. Digit. Libr.*, 22(1):105–134, 2021.

[54] K. M. Svore, J. Teevan, S. T. Dumais, and A. Kulkarni. Creating temporally dynamic web search snippets. In W. R. Hersh, J. Callan, Y. Maarek, and M. Sanderson, editors, *The 35th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '12, Portland, OR, USA, August 12-16, 2012*, pages 1045–1046. ACM, 2012.

[55] A. Talmor, J. Herzig, N. Lourie, and J. Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4149–4158. Association for Computational Linguistics, 2019.

[56] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023.

[57] A. Trischler, T. Wang, X. Yuan, J. Harris, A. Sordoni, P. Bachman, and K. Suleman. Newsqa: A machine comprehension dataset. In P. Blunsom, A. Bordes, K. Cho, S. B. Cohen, C. Dyer, E. Grefenstette, K. M. Hermann, L. Rimell, J. Weston, and S. Yih, editors, *Proceedings of the 2nd Workshop on Representation Learning for NLP, Rep4NLP@ACL 2017, Vancouver, Canada, August 3, 2017*, pages 191–200. Association for Computational Linguistics, 2017.

[58] J. Wang, A. Jatowt, M. Färber, and M. Yoshikawa. Improving question answering for event-focused questions in temporal collections of news articles. *Inf. Retr. J.*, 24(1):29–54, 2021.

[59] J. Wang, A. Jatowt, and M. Yoshikawa. Event occurrence date estimation based on multivariate time series analysis over temporal document collections. In F. Diaz, C. Shah, T. Suel, P. Castells, R. Jones, and T. Sakai, editors, *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 398–407. ACM, 2021.

[60] J. Wang, A. Jatowt, and M. Yoshikawa. Archivalqa: A large-scale benchmark dataset for open-domain question answering over historical news collections. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 3025–3035. ACM, 2022.

[61] J. Wang, A. Jatowt, M. Yoshikawa, and Y. Cai. Bitimebert: Extending pre-trained language representations with bi-temporal information. In H. Chen, W. E. Duh, H. Huang, M. P. Kato, J. Mothe, and B. Poblete, editors, *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 812–821. ACM, 2023.

[62] G. Wenzel and A. Jatowt. An overview of temporal commonsense reasoning and acquisition. *CoRR*, abs/2308.00002, 2023.

[63] P. West, C. Quirk, M. Galley, and Y. Choi. Probing factually grounded content transfer with factual ablation. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3732–3746. Association for Computational Linguistics, 2022.

[64] W. Xiong, J. Du, W. Y. Wang, and V. Stoyanov. Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[65] Y. Yamamoto, T. Tezuka, A. Jatowt, and K. Tanaka. Supporting judgment of fact trustworthiness considering temporal and sentimental aspects. In J. Bailey, D. Maier, K. Schewe, B. Thalheim, and X. S. Wang, editors, *Web Information Systems Engineering - WISE 2008, 9th International Conference, Auckland, New Zealand, September 1-3, 2008. Proceedings*, volume 5175 of *Lecture Notes in Computer Science*, pages 206–220. Springer, 2008.

[66] C. A. Yeung and A. Jatowt. Studying how the past is remembered: towards computational history through large scale text mining. In C. Macdonald, I. Ounis, and I. Ruthven, editors, *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011*, pages 1231–1240. ACM, 2011.

[67] D. Yin, H. Bansal, M. Monajatipoor, L. H. Li, and K. Chang. Geomlama: Geo-diverse commonsense probing on multilingual pre-trained language models. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2039–2055. Association for Computational Linguistics, 2022.