Measuring temporal effects of agent knowledge by date-controlled tool use

R. Patrick Xian[1], Qiming Cui[1,2], Stefan Bauer[3], Reza Abbasi-Asl[1]

[1]UC San Francisco [2]UC Berkeley [3]TU Munich & Helmholtz AI

✉: xrpatrick@gmail.com, qcui@berkeley.edu, st.bauer@tum.de, reza.abbasiasl@ucsf.edu

Abstract

Temporal progression is an integral part of knowledge accumulation and update. Web search is frequently adopted as the grounding for agent knowledge, yet an improper configuration affects the quality of the agent's responses. Here, we assess the agent behavior using distinct date-controlled tools (DCTs) as a stress test to measure the knowledge variability of large language model (LLM) agents. We demonstrate the temporal effects of an LLM agent as a writing assistant, which uses web search to complete scientific publication abstracts. We show that the temporality of search engines translates into tool-dependent agent performance but can be alleviated with base model choice and explicit reasoning instructions such as chain-of-thought prompting. Our results indicate that agent design and evaluations should take a dynamical view and implement effective measures to account for the temporal influence of external resources to improve agent reliability[1].

1 Introduction

AI agents based on LLMs and equipped with tools (Mialon et al., 2023; Wang et al., 2024) are well-suited for complex real-world tasks (Gao et al., 2024; Xu et al., 2024) because of their extended capabilities. Their potential to become virtual assistants, paraprofessionals, or "copilots" holds promise for improving the productivity and creativity of the scientific, medical workforces and beyond (Wachter and Brynjolfsson, 2024; Wornow et al., 2024; Bousetouane, 2025). The evaluation standards for AI agents are still in flux (Kapoor et al., 2024; Højmark et al., 2024) and they are urgently needed in specialized domains and realistic scenarios where the outcomes convey greater bearing on their adoption. Recent works demonstrated the feasibility of LLMs in predicting temporal events (Ye et al., 2024a) and carrying out time series forecasting (Tang et al., 2024), but their equivalents in agentic systems are not yet realized. Scientific knowledge and claims have a strong temporal dependence but they have so far been less studied in the context of generative language models (Zhao et al., 2024; Park et al., 2024). We devised a text completion task as a proxy to measure the agent's usability as a writing assistant with access to external sources (see Fig. 1a).

Web search is an essential tool for grounding agent knowledge in the current and bygone worlds (Pavlick, 2023) and it appears in many applications as a capability extender for models (Zhou et al., 2024a; Song et al., 2024). Nevertheless, web search is subject to the recency and primacy bias of the search engine (Lawrence and Giles, 1998) and the cognitive bias of the users who seek and collect information (Lau and Coiera, 2007). The term *search engine manipulation effect* (Epstein and Robertson, 2015) was coined to refer to the search results' influence on public opinions of societal issues. Independent of search engines, factual and scientific knowledge also experience constant but necessary updates over time (Arbesman, 2013).

While temporal generalization remains challenging for language models (Lazaridou et al., 2021; Wallat et al., 2024), explicit tuning of time-related tool parameters in LLM agents can offer an alternative way to reduce *temporal effects* of the base model. These effects are a source of performance reliability issues of agentic systems that warrant investigation (Ye et al., 2024b). Date control in reality can manifest *passively* because the tool-interfaced computer programs have an intrinsic time stamp or a versioned release over time (Zhang et al., 2009). Alternatively, date control can be imposed *actively* because of copyright, paywall, or local policy. Content access in the past can be controlled retroactively as policy changes (Aral and

[1]The code and datasets for the work are available at https://github.com/RealPolitiX/agent_oost.

Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025), pages 346–355

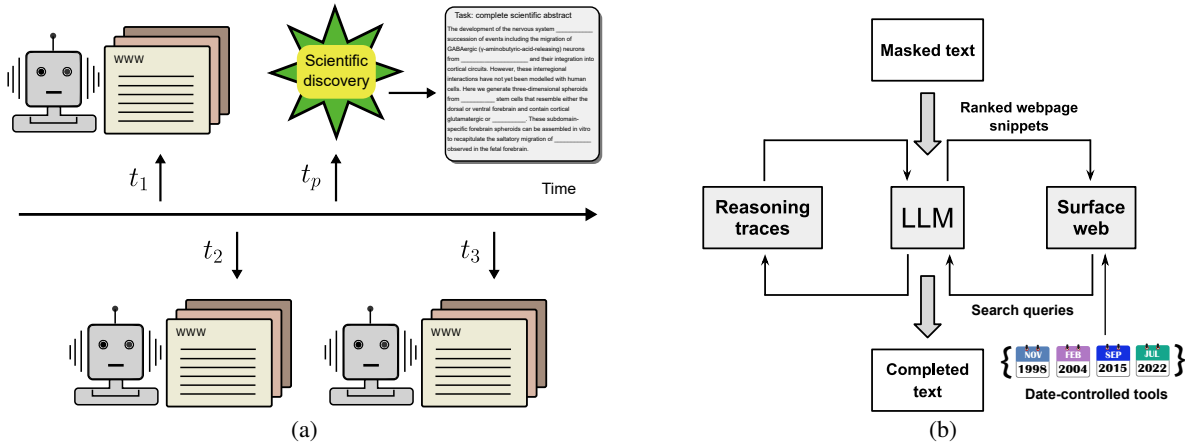


Figure 1: (a) Illustration of the stress testing framework for agent knowledge, t_p indicates the time of publication (b) Temporal tool selection in a ReAct-style agent that performs text completion task in (a) with a selected tool.

Dhillon, 2021). From a technical standpoint, invoking different DCTs is equivalent to changing the environment (here means the surface web, see Fig. 1a) of the agent, which requires the agent to adjust to in task execution.

Stress testing is the ultimate test for model behavior and trustworthiness. In the time domain, out-of-sample (OOS) testing is typically used for temporal prediction methods (Hansen and Timmermann, 2015). Analogous OOS assessments in the text domain include predicting future events (Ye et al., 2024a) or generating hypotheses (Zhou et al., 2024b) conditioned on existing (e.g. past) knowledge. We investigate the comparable problem from the tool use perspective, where the agent has access to changing internet-scale information. Because scientific breakthroughs often lead to significant knowledge updates, they are good markers for temporal knowledge progression[2]. In this work, we aim to investigate the following research questions:

RQ1: Can we manipulate agent knowledge by imposing date restrictions on the tools?

RQ2: Can agents determine the optimal date-controlled version of a tool to use for a task?

Our contributions along these directions are: (i) Formulation of a tool-based stress test for time-dependent knowledge for LLM agents; (ii) Introduction of the `SciBreak` dataset containing the publication records of public-endorsed scientific breakthroughs from 2000 to 2024. (iii) Investigation of the temporal effects of LLM agent performance and behavior. Besides, we also discuss the impact of temporal information on the agent capability and usability and its implications.

[2]Although the judgement on breakthroughs is ultimately subjective and can change with time, our motivation to use them is because of their noticeable footprints on the internet.

2 Related works

Temporal notion of LLMs Previous works have shown that the latent space of LLMs has a direction of time (Gurnee and Tegmark, 2024). Recent investigations show that model performance is affected by the lack of temporal grounding in the pre-training process (Zhao et al., 2024), which can hinder the elicitation of appropriate time-sensitive knowledge at task execution. Previous works have shown LLMs often struggle with tasks that require a consistent temporal grounding (Qiu et al., 2024). The limitation can be improved with techniques such as temporally-informed chain-of-thought (CoT) prompting (Xiong et al., 2024).

Out-of-sample testing Classical and learning-based time series forecasting commonly employ temporal OOS performance tests (Inoue and Kilian, 2005; Hansen and Timmermann, 2015; Cerqueira et al., 2020) to ensure model credibility and usability. It is also relevant from an online learning perspective where data are streamed in sequentially and are subject to distribution shifts. In deep learning, OOS testing is used to provide risk-based self-certification for neural networks (Pérez-Ortiz et al., 2021a,b). In generative models, it has been used for prompt selection (Perez et al., 2021) and controlled generation (Jie et al., 2024) in language models and for quality assessment of synthetic signal generators (Truong et al., 2019).

3 Tool-based stress testing

Definition 3.1 (Date-controlled tool (DCT))**.** A DCT $\mathcal{T}_t$ is a function interface $\mathcal{T}$ (base tool) with a

settable parameter t (upper terminal date) such that the effect of the tool at different times $t_1 \neq t_2$ are distinct, or $\mathcal{T}_{t_1} \neq \mathcal{T}_{t_2}$. The symbol $\mathcal{T}_{t_1}$ indicates that the tool was dated at t_1 or that it encompasses all that came before t_1, which is equivalent to $\mathcal{T}_{t \leqslant t_1}$. We use $\mathcal{T}_{t_1,t_2}$, or equivalently, $\mathcal{T}_{t_1 \leqslant t \leqslant t_2}$, to describe a tool assigned with a temporal window access in $t \in (t_1, t_2]$, where t_1 is the lower terminal date.

Definition 3.2 (LLM agent with tools)**.** An LLM agent $\mathcal{A}$ with the base model $\mathcal{M}$ equipped with an invocable tool $\mathcal{T}$ is $\mathcal{A} = \mathcal{M} \circ \mathcal{T}$. A single tool invocation by the agent given input X produces the trajectory $\tau_n = \{(O, R, G)_i\}_{i=1}^n$ involving the observation O, the reasoning trace R, and the action G. The output of the agent is described by the altered distribution

$$\Pr_{\mathcal{A}}(Y|X) \xrightarrow[\mathcal{T}]{\text{single use}} \Pr_{\mathcal{A}}(Y|X;\tau_1), \quad (1)$$

$$\mathcal{T} : S \to O \implies \mathcal{T}_t : S \to O_t. \quad (2)$$

The tool $\mathcal{T}$ converts the source information S from the environment into the observation O to support agent reasoning and action. In Fig. 1b, S refers to the surface web and O the ranked snippets.

A web-search agent has an implicit parameter $t = t_{\max}$ (i.e. the current date) for the tool $\mathcal{T}$, but it can be modified to an arbitrary value $t < t_{\max}$, which changes the observation in Eq. (2).

Definition 3.3 (Tool-based stress test)**.** A performance test that induces stress conditions by adjusting the tool parameters of an agentic system. A temporal version of the test alters the time information of tools and therefore measures the reliability of agent performance under such conditions.

4 Testing framework implementation

Dataset We constructed the `SciBreak` dataset, which has a clear time-delimited footprint on the internet—scientific breakthroughs. We extended the dataset collated in Wuestman et al. (2020) to the year of 2024. Each year contains up to ~ 20 publications, including multiple publications contributing to one breakthrough.

Agent configuration We integrated DCTs into the ReAct (Yao et al., 2023) agent which allows interleaved thinking and action. The agents were constructed from closed-source models, including OpenAI's GPT-3.5-turbo (`gpt-3.5-turbo-0125`), GPT-4-turbo (`gpt-4-turbo-2024-04-09`) (OpenAI, 2024a), and GPT-4o (`gpt-4o-2024-08-06`) (OpenAI, 2024b) as the base model. For the temporal tool selection task, we also included CoT into the agent pattern (ReAct+CoT) as a comparison. The temperature for model inference was chosen at 0.3 since lower values will significantly increase the failure rates, especially for agents based on less capable models.

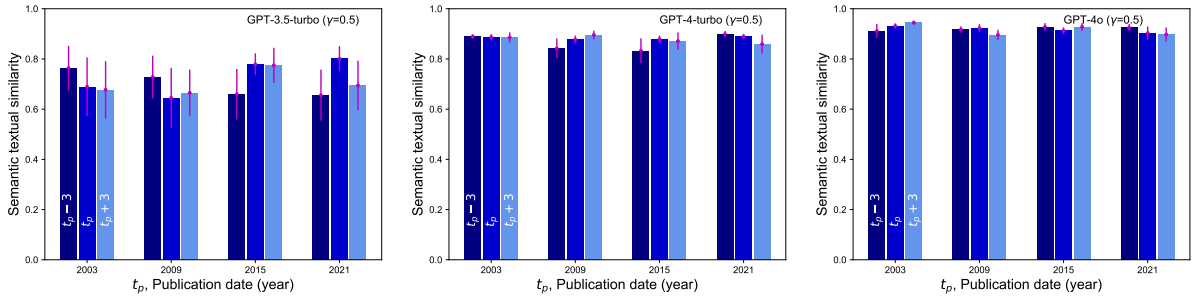
Task design and metrics The LLM agent acted as a writing assistant and was tasked to complete the abstract of scientific papers. All evaluations were in the form of cloze tests with random masking at the word level. The agent was allowed to seek relevant information through the Google Search API with specified dates to acquire information in the form of text snippets (Strzelecki and Rutecka, 2020) in the default ranking of the search engine. The agent then decides if the returned search results are relevant or it prefers to use its own knowledge otherwise. We modulated the information presented to the agent by changing the masking ratio, γ = #(masked words) / #(total words), which was compared for different runs at 0.5 and 0.75, respectively.

For **RQ1**, we evaluated how the text completion task is influenced by changing the upper terminal date of the web search ($\mathcal{T}_t$ with $t = t_p - 3, t_p, t_p + 3$ years) predating and postdating the time of the publication, t_p. We chose a separation of 3 years between the terminal dates from experience with the scientific publishing cycle (i.e., generally 1-2 or more years for major advances). For **RQ2**, we instructed the agent on temporal tool selection through CoT prompting (Wei et al., 2022; Chu et al., 2024). The agent was was presented with a set $\mathcal{T}_s$ of N differently date-controlled tools, $\mathcal{T}_s = \{\mathcal{T}_{t_i,t_i+1}\}_{i=1}^N$, each spanning the period of a year. The model needs to rely on the time parameter to make decisions. For simplicity, all API web searches were done in English.

We quantify the task performance by comparing the actual version of the scientific abstract using the text overlap metric Rouge-L (Lin, 2004) and the semantic text similarity (STS) computed with SentenceTransformer (Reimers and Gurevych, 2019). The STS is the primary performance metric, while Rouge-L is an indicator for verbatim completion.

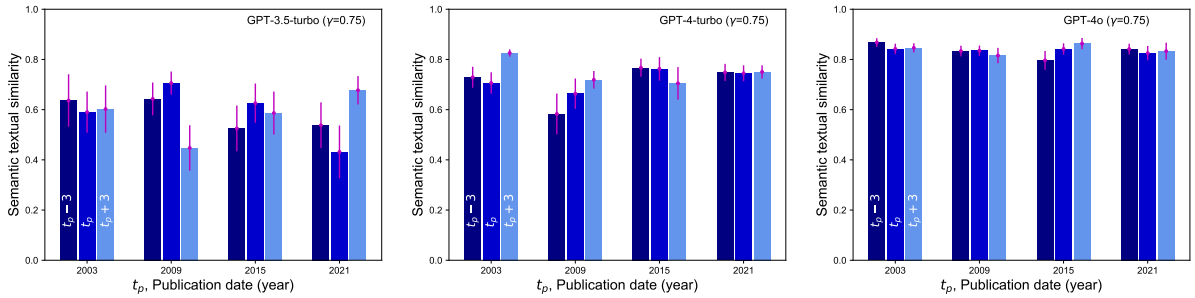
5 Results

Reasoning about time In our evaluation experiments, the agent's reasoning behavior related to its awareness of time (e.g. does the tool have time-



(a) GPT-3.5-turbo (b) GPT-4-turbo (c) GPT-4o

Figure 2: Temporal effects of the search engine on agent performance in scientific abstract completion ($\gamma = 0.5$).



(a) GPT-3.5-turbo (b) GPT-4-turbo (c) GPT-4o

Figure 3: Temporal effects of the search engine on agent performance in scientific abstract completion ($\gamma = 0.75$).

appropriate utility?) is triggered in two scenarios: (i) When the web search returns nothing or little relevant information to assist task completion. The agent then proceeds to complete the task with the internal knowledge of the base LLM. (ii) In temporal tool selection, when the agent is given an explicit CoT stepwise instruction (Chu et al., 2024) to direct its reasoning towards considering the relevance of information to the topic.

Table 1: Performance of LLM agents on text completion ($\gamma = 0.5$) with temporal tool selection.

Agent model	Agent pattern	Pub. in 2003		Pub. in 2015	
		Rouge-L	STS	Rouge-L	STS
GPT-3.5 -turbo	ReAct	0.447	**0.732**	0.447	**0.794**
	ReAct + CoT	0.487	0.640	0.518	0.607
GPT-4 -turbo	ReAct	0.635	0.888	0.588	0.783
	ReAct + CoT	0.649	**0.897**	0.644	**0.859**

Table 1 shows that the STS increases by including CoT prompting (ReAct + CoT) than with ReAct only, when the agent by default selects $\mathcal{T}_{t_+-1,t_+}$ as the tool. Here, $t \in [t_-, t_+]$ being the date range of the tools. The agent then explores more date choices driven by its internal understanding of the scientific concepts present in the input paragraph. These behavioral characteristics allow the agent to handle non-existent and underspecified contexts in the stress test setting. The performance boost of ReAct + CoT agent pattern requires a model with sufficient reasoning capability such as GPT-4 (OpenAI, 2024a), while for GPT-3.5, it is more prone to failure and the performance gain is reversed.

Table 2: Performance of LLM agents on text completion ($\gamma = 0.75$) with temporal tool selection.

Agent model	Agent pattern	Pub. in 2003		Pub. in 2015	
		Rouge-L	STS	Rouge-L	STS
GPT-3.5 -turbo	ReAct	0.297	**0.604**	0.438	**0.779**
	ReAct + CoT	0.212	0.416	0.546	0.774
GPT-4 -turbo	ReAct	0.343	0.756	0.311	0.666
	ReAct + CoT	0.447	**0.796**	0.304	**0.693**

Temporal effects across models and masking High-capacity models with greater reasoning capabilities are capable of making more sensible choices on the tool dates in the temporal tool selection task. This task evaluated the model capability with two different levels of text masking determined by the masking ratio γ. The results in Figs. 2-3 and Tables 1-2 contains two major trends: (i) The more advanced models can recover more of the missing semantic content in the masked input, as indicated by the significant increase of STS from

LLM agents based on GPT-3.5 to GPT-4o (OpenAI, 2024b). (ii) There is noticeable variability of agent performance between knowledge generated more recently than before 2010. (iii) For the same model, varying the masking ratio of input largely preserves the date sensitivity in the agent performance. Similar time-dependent performance change has also been described in a different context for LLMs (Zhao et al., 2024). Overall, the temporal effects are less severe in more capable models.

6 Discussion

Agent vulnerability and tool-based control Our work shows that agents with access to external tools are subject to manipulation by corrupted tools (Ye et al., 2024b) to compromise their generated information for knowledge-intensive domains, extending the previous example on misinformation in LLMs (Han et al., 2024). We provide evidence that agentic reasoning and model capabilities can counter the limited information quality of search engines. In agentic search, carefully designed controls will allow filtering of unreliable information and improve agent performance. Imposing a date restriction on search is similar to reranking and partial deletion of the search results. Therefore, agent designs with verification of content freshness and temporality will ensure more reliable use.

Robustness and reproducibility of agentic systems Agentic systems for scientific problems should adapt to different levels of prior knowledge available to the domain (Vinuesa et al., 2024). From the robustness viewpoint, temporal shifts can be counteracted through the use of external resources. Task-oriented requirements specification (Xian et al., 2025) is useful for improving the usability and avoiding unnecessary artifacts from model imperfections and the reliability of external tooling and information sources. From the reproducibility viewpoint, agentic tool use should always incorporate essential information of the key parameters. Our work indicates that more research is needed in principled maintenance of agentic frameworks under constant updates of external resources to facilitate reliable agent design (Kapoor et al., 2024).

Limitations

Our work is focused on models with tool-calling and reasoning capabilities, yet the phenomenon demonstrated here has equivalents in less capable models not investigated here. The test examples we chose simulate a realistic application setting of an agentic writing assistant, yet such an effect could already manifest in more ordinary tasks such as knowledge-related question answering or in malicious settings where bad actors are trying to pollute the information system (e.g. internet or proprietary databases) through more elaborate search engine manipulation. We also didn't investigate the scenario where the LLM agent has possession of a proprietary tool (e.g. for fact-checking) independent of web search, which could be an alternative way to improve performance.

Ethics statement

The present work illustrates the importance of temporal factors when working with LLM agents that have access to the internet. Our results provide an initial assessment of the factors that can influence an agentic writing assistant's ability to properly utilize time-bounded search results in its reasoning process. We acknowledge that reliance on time-bounded search results presents ethical considerations related to misinformation, data freshness, and accuracy. Agents may misinterpret outdated or contextually misaligned information, leading to erroneous conclusions. Furthermore, temporal biases in search tools, such as the prioritization of newer content over historically relevant sources, can skew results, potentially reinforcing recency bias or omitting crucial context.

References

Sinan Aral and Paramveer S. Dhillon. 2021. Digital Paywall Design: Implications for Content Demand and Subscriptions. *Management Science*, 67(4):2381–2402. Publisher: INFORMS.

Samuel Arbesman. 2013. *The Half-Life of Facts: Why Everything We Know Has an Expiration Date*, reprint edition. Penguin Publishing Group.

Fouad Bousetouane. 2025. Agentic Systems: A Guide to Transforming Industries with Vertical AI Agents. *arXiv preprint*. ArXiv:2501.00881 [cs].

Vitor Cerqueira, Luis Torgo, and Igor Mozetič. 2020. Evaluating time series forecasting models: an empirical study on performance estimation methods. *Machine Learning*, 109(11):1997–2028.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2024. Navigate through Enigmatic Labyrinth A Survey of Chain of Thought

Reasoning: Advances, Frontiers and Future. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1173–1203, Bangkok, Thailand. Association for Computational Linguistics.

Robert Epstein and Ronald E. Robertson. 2015. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences*, 112(33):E4512–E4521. Publisher: National Academy of Sciences.

Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and Marinka Zitnik. 2024. Empowering biomedical discovery with AI agents. *Cell*, 187(22):6125–6151.

Wes Gurnee and Max Tegmark. 2024. Language Models Represent Space and Time. In *The Twelfth International Conference on Learning Representations*.

Tianyu Han, Sven Nebelung, Firas Khader, Tianci Wang, Gustav Müller-Franzes, Christiane Kuhl, Sebastian Försch, Jens Kleesiek, Christoph Haarburger, Keno K. Bressem, Jakob Nikolas Kather, and Daniel Truhn. 2024. Medical large language models are susceptible to targeted misinformation attacks. *npj Digital Medicine*, 7(1):1–9. Publisher: Nature Publishing Group.

Peter Reinhard Hansen and Allan Timmermann. 2015. Equivalence Between Out-of-Sample Forecast Comparisons and Wald Statistics. *Econometrica*, 83(6):2485–2505.

Axel Højmark, Govind Pimpale, Arjun Panickssery, Marius Hobbhahn, and Jérémy Scheurer. 2024. Analyzing Probabilistic Methods for Evaluating Agent Capabilities. In *Workshop on Socially Responsible Language Modelling Research*.

Atsushi Inoue and Lutz Kilian. 2005. In-Sample or Out-of-Sample Tests of Predictability: Which One Should We Use? *Econometric Reviews*, 23(4):371–402. Publisher: Taylor & Francis.

Renlong Jie, Xiaojun Meng, Lifeng Shang, Xin Jiang, and Qun Liu. 2024. Prompt-Based Length Controlled Generation with Multiple Control Types. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1067–1085, Bangkok, Thailand. Association for Computational Linguistics.

Sayash Kapoor, Benedikt Stroebl, Zachary S. Siegel, Nitya Nadgir, and Arvind Narayanan. 2024. AI Agents That Matter. *arXiv preprint*. ArXiv:2407.01502 [cs].

Annie Y.S. Lau and Enrico W. Coiera. 2007. Do People Experience Cognitive Biases while Searching for Information? *Journal of the American Medical Informatics Association*, 14(5):599–608.

Steve Lawrence and C. Lee Giles. 1998. Searching the World Wide Web. *Science*, 280(5360):98–100. Publisher: American Association for the Advancement of Science.

Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomáš Kočiský, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. 2021. Mind the Gap: Assessing Temporal Generalization in Neural Language Models. In *Advances in Neural Information Processing Systems*.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Grégoire Mialon, Roberto Dessi, Maria Lomeli, Christoforos Nalmpantis, Ramakanth Pasunuru, Roberta Raileanu, Baptiste Roziere, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. 2023. Augmented Language Models: a Survey. *Transactions on Machine Learning Research*.

OpenAI. 2024a. GPT-4 Technical Report. *arXiv preprint*. ArXiv:2303.08774 [cs].

OpenAI. 2024b. GPT-4o System Card. *arXiv preprint*. ArXiv:2410.21276 [cs].

Yein Park, Chanwoong Yoon, Jungwoo Park, Donghyeon Lee, Minbyul Jeong, and Jaewoo Kang. 2024. ChroKnowledge: Unveiling Chronological Knowledge of Language Models in Multiple Domains. *arXiv preprint*. ArXiv:2410.09870 [cs].

Ellie Pavlick. 2023. Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 381(2251).

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True Few-Shot Learning with Language Models. In *Advances in Neural Information Processing Systems*.

Maria Pérez-Ortiz, Omar Rivasplata, Emilio Parrado-Hernandez, Benjamin Guedj, and John Shawe-Taylor. 2021a. Progress in Self-Certified Neural Networks. *arXiv preprint*. ArXiv:2111.07737 [cs].

María Pérez-Ortiz, Omar Rivasplata, John Shawe-Taylor, and Csaba Szepesvári. 2021b. Tighter risk certificates for neural networks. *J. Mach. Learn. Res.*, 22(1):227:10326–227:10365.

Yifu Qiu, Zheng Zhao, Yftah Ziser, Anna Korhonen, Edoardo Ponti, and Shay Cohen. 2024. Are Large Language Model Temporally Grounded? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7064–7083, Mexico City, Mexico. Association for Computational Linguistics.

David Reich, Richard E. Green, Martin Kircher, Johannes Krause, Nick Patterson, Eric Y. Durand, Bence Viola, Adrian W. Briggs, Udo Stenzel, Philip L. F. Johnson, Tomislav Maricic, Jeffrey M. Good, Tomas Marques-Bonet, Can Alkan, Qiaomei Fu, Swapan Mallick, Heng Li, Matthias Meyer, Evan E. Eichler, Mark Stoneking, Michael Richards, Sahra Talamo, Michael V. Shunkov, Anatoli P. Derevianko, Jean-Jacques Hublin, Janet Kelso, Montgomery Slatkin, and Svante Pääbo. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, 468(7327):1053–1060. Publisher: Nature Publishing Group.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Yueqi Song, Frank Xu, Shuyan Zhou, and Graham Neubig. 2024. Beyond Browsing: API-Based Web Agents. *arXiv preprint*. ArXiv:2410.16464 [cs].

Artur Strzelecki and Paulina Rutecka. 2020. Direct Answers in Google Search Results. *IEEE Access*, 8:103642–103654.

Hua Tang, Chong Zhang, Mingyu Jin, Qinkai Yu, Zhenting Wang, Xiaobo Jin, Yongfeng Zhang, and Mengnan Du. 2024. Time Series Forecasting with LLMs: Understanding and Enhancing Model Capabilities. *arXiv preprint*. ArXiv:2402.10835 [cs].

Nhan Duy Truong, Levin Kuhlmann, Mohammad Reza Bonyadi, Damien Querlioz, Luping Zhou, and Omid Kavehei. 2019. Epileptic Seizure Forecasting With Generative Adversarial Networks. *IEEE Access*, 7:143999–144009.

Ricardo Vinuesa, Jean Rabault, Hossein Azizpour, Stefan Bauer, Bingni W. Brunton, Arne Elofsson, Elias Jarlebring, Hedvig Kjellstrom, Stefano Markidis, David Marlevi, Paola Cinnella, and Steven L. Brunton. 2024. Opportunities for machine learning in scientific discovery. *arXiv preprint*. ArXiv:2405.04161 [cs].

Robert M. Wachter and Erik Brynjolfsson. 2024. Will Generative Artificial Intelligence Deliver on Its Promise in Health Care? *JAMA*, 331(1):65–69.

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, pages 683–692, New York, NY, USA. Association for Computing Machinery.

Zhiruo Wang, Zhoujun Cheng, Hao Zhu, Daniel Fried, and Graham Neubig. 2024. What Are Tools Anyway? A Survey from the Language Model Perspective. In *First Conference on Language Modeling*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*.

Michael Wornow, Avanika Narayan, Krista Opsahl-Ong, Quinn McIntyre, Nigam Shah, and Christopher Ré. 2024. Automating the Enterprise with Foundation Models. *Proc. VLDB Endow.*, 17(11):2805–2812.

Mignon Wuestman, Jarno Hoekman, and Koen Frenken. 2020. A typology of scientific breakthroughs. *Quantitative Science Studies*, 1(3):1203–1222.

R. Patrick Xian, Noah R. Baker, Tom David, Qiming Cui, A. Jay Holmgren, Stefan Bauer, Madhumita Sushil, and Reza Abbasi-Asl. 2025. Robustness tests for biomedical foundation models should tailor to specification. *arXiv preprint*. ArXiv:2502.10374 [cs].

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024. Large Language Models Can Learn Temporal Reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10452–10470, Bangkok, Thailand. Association for Computational Linguistics.

Frank F. Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Z. Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, Mingyang Yang, Hao Yang Lu, Amaad Martin, Zhe Su, Leander Maben, Raj Mehta, Wayne Chi, Lawrence Jang, Yiqing Xie, Shuyan Zhou, and Graham Neubig. 2024. TheAgentCompany: Benchmarking LLM Agents on Consequential Real World Tasks. *arXiv preprint*. ArXiv:2412.14161 [cs].

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *The Eleventh International Conference on Learning Representations*.

Chenchen Ye, Ziniu Hu, Yihe Deng, Zijie Huang, Mingyu Derek Ma, Yanqiao Zhu, and Wei Wang. 2024a. MIRAI: Evaluating LLM Agents for Event Forecasting. *arXiv preprint*. ArXiv:2407.01231 [cs].

Junjie Ye, Sixian Li, Guanyu Li, Caishuang Huang, Songyang Gao, Yilong Wu, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024b. ToolSword: Unveiling Safety Issues of Large Language Models in Tool Learning Across Three Stages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2181–2211, Bangkok, Thailand. Association for Computational Linguistics.

Ruiqiang Zhang, Yi Chang, Zhaohui Zheng, Donald Metzler, and Jian-yun Nie. 2009. Search Engine Adaptation by Feedback Control Adjustment for Time-sensitive Query. In *Proceedings of Human Language Technologies: The 2009 Annual Conference*

of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, pages 165–168, Boulder, Colorado. Association for Computational Linguistics.

Bowen Zhao, Zander Brumbaugh, Yizhong Wang, Hannaneh Hajishirzi, and Noah Smith. 2024. Set the Clock: Temporal Alignment of Pretrained Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15015–15040, Bangkok, Thailand. Association for Computational Linguistics.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2024a. WebArena: A Realistic Web Environment for Building Autonomous Agents. In *The Twelfth International Conference on Learning Representations*.

Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. 2024b. Hypothesis Generation with Large Language Models. In *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pages 117–139, Miami, FL, USA. Association for Computational Linguistics.

A Agentic task

The web-search agent is configured according to the ReAct architecture using a helpful assistant system prompt ("You are a helpful writing assistant."). The instruction prompt is as follows.

> Instruction: Browse the internet using keywords or phrases in the following paragraph with masked text. Make use of the search results to fill in each [UNK] with a word or punctuation. Output your final results after Final Answer: to indicate the beginning of your completed text. Use your own judgement to decide what information from the search results is useful. If nothing is useful, then try to complete the task with your own knowledge.
> Requirements: Use the given parameters in the tools to solve the following problem and don't reset them. Don't change the number of arguments supplied to the tool you use.
> Masked text: ... `[the masked text]` ...

In our experiments, `the masked text` is replaced with the masked scientific abstracts. The instruction prompt contains a task description and suppresses undesired agent behaviors that can cause errors in execution. In our empirical investigation, we also found that the reasoning process of tool-calling agents tend to reset the date parameter. For the experiments, we added an instruction to specifically forbid that behavior.

B Dataset preprocessing

The `SciBreak` dataset is partly based on peer-reviewed publications collated and categorized in Wuestman et al. (2020), including the annual top-ten-ranked scientific breakthroughs from mid-1990s till 2012 collated by the journal Science at the end of each year. The publications are drawn from various journals in the physical, biomedical, and engineering sciences, which constitute the scope of the ranking. We chose records from year 2000 to 2012 and extended to year 2024 by self-curating the extra years of ranked publications from the published tally in each year. The links to the yearly breakdown is provided as follows: 2012 (15), 2015 (12), 2018 (17), 2021 (14), 2024 (11), etc[3]. The number in the parenthesis indicates the number of publications featured in the top-ten ranking of the corresponding year.

We collected the abstracts of the associated publications through web scraping from the public databases PubMed[4] and SAO/NASA ADS Abstract Service[5] using the Digital Object Identifiers of the publications, which are also provided in the dataset.

C Extended results

Table 3: Performance of ReAct-style LLM agents on text completion ($\gamma = 0.5$, see Fig. 2) following web search with DCTs. For the row of Input, the metrics are computed between the input and the ground truth.

Agent model	$\mathcal{T}_t$ cut-off (t years)	Pub. in 2003		Pub. in 2015	
		Rouge-L	STS	Rouge-L	STS
Input	—	0.486	0.657	0.488	0.658
GPT-3.5 -turbo	$t_p - 3$	0.481	**0.764**	0.623	0.659
	t_p	0.462	0.689	0.438	**0.779**
	$t_p + 3$	0.427	0.677	0.546	0.774
GPT-4 -turbo	$t_p - 3$	0.627	**0.889**	0.601	0.833
	t_p	0.641	0.887	0.629	**0.879**
	$t_p + 3$	0.614	0.886	0.624	0.871
GPT-4o	$t_p - 3$	0.735	0.911	0.727	**0.928**
	t_p	0.741	0.931	0.744	0.912
	$t_p + 3$	0.733	**0.944**	0.701	**0.928**

Table 4: Performance of ReAct-style LLM agents on text completion ($\gamma = 0.75$, see Fig. 3) following web search with DCTs. For the row of Input, the metrics are computed between the input and the ground truth.

Agent model	$\mathcal{T}_t$ cut-off (t years)	Pub. in 2003		Pub. in 2015	
		Rouge-L	STS	Rouge-L	STS
Input	—	0.486	0.657	0.488	0.658
GPT-3.5 -turbo	$t_p - 3$	0.418	**0.636**	0.286	0.525
	t_p	0.279	0.590	0.274	**0.626**
	$t_p + 3$	0.285	0.602	0.309	0.586
GPT-4 -turbo	$t_p - 3$	0.324	0.729	0.325	**0.767**
	t_p	0.341	0.706	0.376	0.762
	$t_p + 3$	0.390	**0.826**	0.319	0.705
GPT-4o	$t_p - 3$	0.478	**0.867**	0.405	0.796
	t_p	0.440	0.842	0.448	0.842
	$t_p + 3$	0.446	0.846	0.458	**0.863**

Extended results for **RQ1** include Table 3, which includes a portion of the results in Fig. 2, and Table 4, which is similarly related to Fig. 3. These tables also contain the corresponding values of the Rouge-L score (Lin, 2004). The performance metrics in the tables with a higher masking ratio ($\gamma = 0.75$) trend lower than the conditions in tables when $\gamma = 0.5$, similar characteristics hold for the time dependence of the STS for publications appeared more recently (since the 2010s) and those

[3] A complete list is found at the online repository
[4] `https://pubmed.ncbi.nlm.nih.gov/`
[5] `https://ui.adsabs.harvard.edu/`

before. In the temporal tool selection task, the ReAct + CoT pattern provides a performance gain over ReAct-only agents.

Observation: None

Thought: The search results provide detailed information about the discovery and genetic analysis of Denisovans. I can use this information to fill in the blanks in the masked text. The text seems to describe the discovery of Denisovan remains in Denisova Cave, Siberia, and their genetic relationship to Neanderthals and modern humans.

(a) No date restriction on tool.

Observation: The search did not return any results. This could be because the discovery of Denisovans and the relevant genetic analysis occurred after 2004. Therefore, I need to adjust my approach to fill in the blanks based on general knowledge available up to 2004 about ancient hominins, their discoveries, and genetic analysis techniques.

Thought: Since the Denisovan discovery itself was not known before 2004, I will use general knowledge about ancient DNA studies and hominin discoveries in Siberia to fill in the blanks.

(b) Tool restricted to date before event.

Figure 4: Example reasoning paths (emphasized by underlines) from the LLM agent before and after imposing a date restriction on the tool. The example here uses the discovery of Denisovan hominins. Important parts of the verbalized reasoning are underlined.

D Examples of temporal awareness

The example in Fig. 4 includes the typical reasoning trace of the LLM agent put under testing to the breakthrough discovery of Denisovan hominins (an ancestor of modern humans) around 2008, which became widely reported in the English media a couple of years later thanks to the major scientific publication (Reich et al., 2010) and contributed significantly to Svante Pääbo's Nobel Prize in 2022.

The Denisova cave in Siberia has existed as a geographical name for much longer on the internet, but primarily in the Russian language, so largely inaccessible through English language search before 2008. Moreover, Denisova is used as a surname, which appears upon search in English. However, neither of these facts informs the model about potential content in the masked text about the scientific discovery that was consolidated by genomic sequencing. When the clock of the search engine is set to before 2008, the LLM agent attempted to confront the absence of results and reasoned that the work was not known before the cut-off date of the search and instead switched to using its parametric knowledge to complete the text. If the search clock was unset, then the information is readily available, as compared in Fig. 4.