Impact Statement

Daily Oracle serves as an up-to-date, continuous evaluation benchmark for assessing AI models' forecasting accuracy and temporal generalization. These abilities have broad applications in areas such as finance, healthcare, and policy-making. The performance drop we observe highlights the risk of outdated knowledge and the importance of continuous model updates to keep AI systems reliable. In the future, by assessing continuous model update strategies in Daily Oracle, the broader ML community can gain valuable insights into how to maintain AI systems that are relevant and well-informed on recent and upcoming events.

Acknowledgment

This work was supported in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) with a grant funded by the Ministry of Science and ICT (MSIT) of the Republic of Korea in connection with the Global AI Frontier Lab International Collaborative Research. (No. RS-2024-00469482 & RS-2024-00509279) We also thank the Microsoft Accelerating Foundation Models Research program for providing Azure cloud compute credits for the LLM APIs. The compute was also supported by the NYU High Performance Computing resources, services, and staff expertise.

References

Agarwal, O. and Nenkova, A. Temporal effects on pre-trained models for language processing tasks. *Transactions of the Association for Computational Linguistics*, 2022.

Anderson, J. R. and Schooler, L. J. Reflections of the environment in memory. *Psychological Science*, 2(6):396–408, 1991.

Anthropic. The claude 3 model family: Opus, sonnet, haiku. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf, 2024.

Chen, W., Wang, X., Wang, W. Y., and Wang, W. Y. A dataset for answering time-sensitive questions. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Dempsey, W. H., Moreno, A., Scott, C. K., Dennis, M. L., Gustafson, D. H., Murphy, S. A., and Rehg, J. M. iSurvive: An interpretable, event-time prediction model for mHealth. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996.

Fleiss, J. L. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., and Leahy, C. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Gillingham, K., Nordhaus, W., Anthoff, D., Blanford, G., Bosetti, V., Christensen, P., McJeon, H., and Reilly, J. Modeling uncertainty in integrated assessment of climate change: A multimodel comparison. *Journal of the Association of Environmental and Resource Economists*, 5(4): 791–826, 2018.

Gokaslan, A. and Cohen, V. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus, 2019.

Halawi, D., Zhang, F., Yueh-Han, C., and Steinhardt, J. Approaching human-level forecasting with language models. In *Advances in Neural Information Processing Systems*, 2024.

Hamborg, F., Meuschke, N., Breitinger, C., and Gipp, B. news-please: A generic news crawler and extractor. In *Proceedings of the 15th International Symposium of Information Science*, 2017.

Jang, J., Ye, S., Yang, S., Shin, J., Han, J., Kim, G., Choi, S. J., and Seo, M. Towards continual knowledge learning of language models. In *International Conference on Learning Representations*, 2022.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., Casas, D. d. l., Hanna, E. B., Bressand, F., et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Jin, W., Khanna, R., Kim, S., Lee, D.-H., Morstatter, F., Galstyan, A., and Ren, X. ForecastQA: A question answering challenge for event forecasting with temporal text data. In *Proceedings of the 59th Annual Meeting of*

the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021.

Jin, X., Zhang, D., Zhu, H., Xiao, W., Li, S.-W., Wei, X., Arnold, A., and Ren, X. Lifelong pretraining: Continually adapting language models to emerging corpora. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022.

Karger, E., Bastani, H., Yueh-Han, C., Jacobs, Z., Halawi, D., Zhang, F., and Tetlock, P. E. Forecastbench: A dynamic benchmark of ai forecasting capabilities. In *International Conference on Learning Representations*, 2025.

Kasai, J., Sakaguchi, K., Le Bras, R., Asai, A., Yu, X., Radev, D., Smith, N. A., Choi, Y., Inui, K., et al. Realtime qa: what's the answer right now? In *Advances in Neural Information Processing Systems*, 2024.

Ke, Z., Shao, Y., Lin, H., Konishi, T., Kim, G., and Liu, B. Continual pre-training of language models. In *International Conference on Learning Representations*, 2022a.

Ke, Z., Shao, Y., Lin, H., Xu, H., Shu, L., and Liu, B. Adapting a language model while preserving its general knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022b.

Kobayashi, S. Homemade bookcorpus. https://github.com/soskek/bookcorpus, 2018.

Lazaridou, A., Kuncoro, A., Gribovskaya, E., Agrawal, D., Liska, A., Terzi, T., Gimenez, M., de Masson d'Autume, C., Kocisky, T., Ruder, S., Yogatama, D., Cao, K., Young, S., and Blunsom, P. Mind the gap: Assessing temporal generalization in neural language models. In *Advances in Neural Information Processing Systems*, 2021.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, 2020.

Li, C. and Flanigan, J. Task contamination: Language models may not be few-shot anymore. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Liska, A., Kocisky, T., Gribovskaya, E., Terzi, T., Sezener, E., Agrawal, D., Cyprien De Masson, D., Scholtes, T., Zaheer, M., Young, S., et al. Streamingqa: A benchmark for adaptation to new knowledge over time in question answering models. In *International Conference on Machine Learning*, 2022.

Longpre, S., Mahari, R., Lee, A., Lund, C., Oderinwale, H., Brannon, W., Saxena, N., Obeng-Marnu, N., South, T., Hunter, C., et al. Consent in crisis: The rapid decline of the ai data commons. In *Advances in Neural Information Processing Systems*, 2024.

Lopez-Lira, A. and Tang, Y. Can chatgpt forecast stock price movements? return predictability and large language models. *arXiv preprint arXiv:2304.07619*, 2023.

McIntosh, T. R., Susnjak, T., Arachchilage, N., Liu, T., Xu, D., Watters, P., and Halgamuge, M. N. Inadequacies of large language model benchmarks in the era of generative artificial intelligence. *IEEE Transactions on Artificial Intelligence*, 2025.

Nagel, S. Common crawl news dataset. https://data.commoncrawl.org/crawl-data/CC-NEWS/index.html, 2016.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

OpenAI. New embedding models and api updates. https://openai.com/index/new-embedding-models-and-api-updates/, 2024a.

OpenAI. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024b.

Rae, J. W., Potapenko, A., Jayakumar, S. M., Hillier, C., and Lillicrap, T. P. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations*, 2020.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21: 1–67, 2020.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., Gatford, M., et al. Okapi at trec-3. *Nist Special Publication Sp*, 109:109, 1995.

Röttger, P. and Pierrehumbert, J. Temporal adaptation of BERT and performance on downstream document classification: Insights from social media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021.

Sainz, O., Campos, J., García-Ferrero, I., Etxaniz, J., de Lacalle, O. L., and Agirre, E. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

Saxton, D., Grefenstette, E., Hill, F., and Kohli, P. Analysing mathematical reasoning abilities of neural models. In *International Conference on Learning Representations*, 2019.

Team, G., Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Tetlock, P. E. and Gardner, D. *Superforecasting: The art and science of prediction*. Random House, 2016.

Tiedemann, J. Finding alternative translations in a large corpus of movie subtitle. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016.

Vu, T., Iyyer, M., Wang, X., Constant, N., Wei, J., Wei, J., Tar, C., Sung, Y.-H., Zhou, D., Le, Q., and Luong, T. FreshLLMs: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics ACL 2024*, 2024.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in neural information processing systems*, 2022.

Xu, C., Guan, S., Greene, D., and Kechadi, M.-T. Benchmark data contamination of large language models: A survey. *arXiv preprint arXiv:2406.04244*, 2024.

Yan, Q., Seraj, R., He, J., Meng, L., and Sylvain, T. Autocast++: Enhancing world event prediction with zero-shot ranking-based context retrieval. In *The Twelfth International Conference on Learning Representations*, 2024.

Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Ye, C., Hu, Z., Deng, Y., Huang, Z., Ma, M. D., Zhu, Y., and Wang, W. Mirai: Evaluating llm agents for event forecasting. *arXiv preprint arXiv:2407.01231*, 2024.

Yıldız, Ç., Ravichandran, N. K., Punia, P., Bethge, M., and Ermis, B. Investigating continual pretraining in large language models: Insights and implications. *arXiv preprint arXiv:2402.17400*, 2024.

Zhang, M. and Choi, E. SituatedQA: Incorporating extralinguistic contexts into QA. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

Zhang, Z., Cao, Y., Ye, C., Ma, Y., Liao, L., and Chua, T.-S. Analyzing temporal complex events with large language models? a benchmark towards temporal, long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024.

Zhu, C., Chen, N., Gao, Y., Zhang, Y., Tiwari, P., and Wang, B. Is your LLM outdated? a deep look at temporal generalization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2025.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, 2015.

Zou, A., Xiao, T., Jia, R., Kwon, J., Mazeika, M., Li, R., Song, D., Steinhardt, J., Evans, O., and Hendrycks, D. Forecasting future world events with neural networks. In *Advances in Neural Information Processing Systems*, 2022.