Research Statement

Anirudh Phukan *Ph.D. Student in Computational and Data Sciences, IISc*

1 Importance of Continual Learning

I believe that the next set of breakthroughs in AI will require systems that can continue to acquire new knowledge and skills while in use, building on the capabilities they already possess. This ability, known as continual learning [1], is not merely desirable but necessary for any system aspiring toward *Artificial General Intelligence*. This is because real-world environments are too open-ended for us to enumerate all the tasks an intelligent agent may need to perform.

The need for continual learning is closely related to why machine learning succeeded where purely rule-based systems struggled. Rule-based systems require the designer to anticipate and specify behavior for each relevant situation in advance, which makes them brittle in complex environments. Machine learning moved beyond this limitation by allowing systems to learn patterns from data and generalize to new inputs similar to those seen during training. Continual learning is a step further in the same direction. Rather than learning only before deployment, it allows a system to keep improving after it is in use, as new needs arise.

2 Limitations with the current lens of Continual Learning

Continual learning has been studied for decades and since 2022 has its own dedicated conference, CoLLAs [2]. Yet it remains difficult because learning new information often requires modifying the same representations that support existing capabilities. New learning can therefore interfere with older knowledge, causing useful capabilities to degrade or be overwritten. The core challenge is to adapt to new tasks while preserving the general knowledge and skills that remain useful across tasks.

This tension is most commonly studied through catastrophic forgetting, the tendency of a model to lose previously acquired capabilities when it learns from new data [3]. While this emphasis is understandable, it has often dominated the discussion to the point that catastrophic forgetting is treated as almost synonymous with continual learning [4].

Recent work has broadened this focus from catastrophic forgetting alone to the larger problem of learning from data that changes over time [5, 6]. Under this view, a system receives new data over time from a distribution that differs from what it has seen before. The difficulty is not only that models may forget old information. Over time, they may also become less able to learn new information, or they may reuse old knowledge in ways that help some future tasks but hurt others [7, 8]. Continual learning is therefore fundamentally a problem of making learning cumulative, new experience should deepen and reorganize prior knowledge in ways that improve performance across both old and new tasks.

Part of the appeal of this framing is that it admits comparatively straightforward benchmarks. Researchers can compare a model that learns from data as it arrives with one trained on all the data at once [9]. However, this framing still assumes that the relevant training data will eventually be provided to the learner. In many real-world settings, a system first encounters a novel task or failure mode in deployment, after which the corresponding data must be gathered and made usable for learning [10]. In such cases, the challenge is not only to learn effectively from new data, but also to acquire the right data through exploration or interaction with the environment [11].

The analogy to chess is instructive. A player improves not only by learning techniques from a teacher, but also by exploring positions beyond the teacher's curriculum and discovering new strategies. That broader process is what enables performance that eventually surpasses the teacher [12]. For continual learning to meet its broader promise, the field must move beyond adaptation to changing data and address the end-to-end problem of acquiring, organizing, and learning from new experience. Figure 1 summarizes this shift.

3 The Benchmark Barrier to broader Continual Learning

A natural question, then, is why the field has not moved more decisively in this direction. I argue that the main obstacle is methodological. We lack end-to-end benchmarks based on tasks that humans regularly perform and handle successfully because they can continually learn. Such benchmarks would make the value of broader continual learning visible and encourage methods that address it directly.

A benchmark for broader continual learning should therefore satisfy the following desiderata:

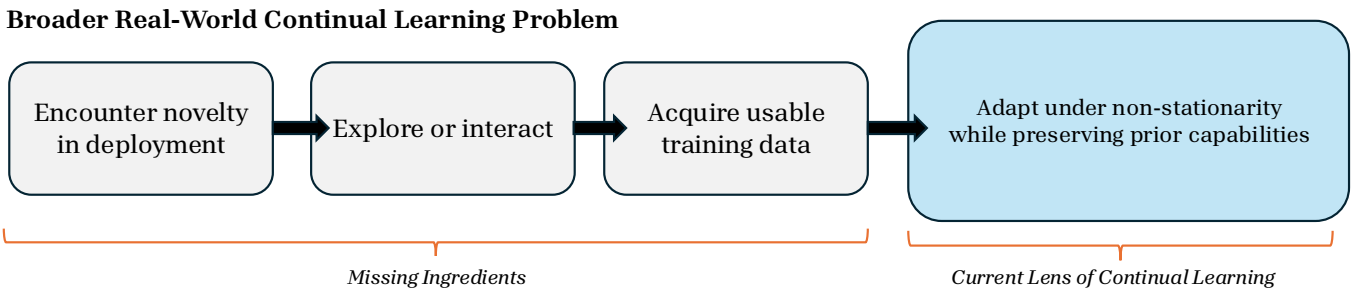

Figure 1: A broader view of continual learning includes discovering what needs to be learned, finding relevant experience, and learning from it. The current view focuses mainly on adapting to changing data.

- **Human-grounded.** It should correspond to a real task that people plausibly perform, where success depends substantially on acquiring and using new knowledge. The task should also be accessible enough that model failure cannot be dismissed as merely a consequence of insufficient scale, compute, or engineering effort.
- **Dynamic.** It should refresh automatically and frequently, so that future models cannot simply absorb it into pre-training.
- **Automated Evaluation.** It should support automatic scoring to enable rapid iteration.
- **Non-saturated.** Frontier models should not already solve it reliably under standard prompting and tool use. Otherwise, the benchmark will not make a compelling case for continual learning.

A promising starting point is knowledge-based benchmarks, since real-world knowledge changes continuously and people regularly need to update what they know. Sources such as Wikipedia and Wikidata make this setting relatively easy to automate at scale, both for question generation and answer evaluation. Several recent benchmarks already explore this space [13, 14, 15, 16, 17, 18]. The difficulty is that many of these tasks reduce to simple or multi-hop question answering over latest information, which frontier language models can often handle with search or retrieval tools. In such cases, success reflects an ability to decompose the question into its constituent parts and generate effective retrieval queries for each of them, rather than a need for continual learning itself.

Therefore, the central challenge is to design a knowledge benchmark that remains dynamic and easy to evaluate, but cannot be solved by current models through retrieval alone. The benchmark should test whether a system internalizes newly acquired knowledge and applies it even when the question does not explicitly signal that the knowledge is needed.

4 Research Agenda

4.1 A Benchmark for Continual Knowledge Acquisition

My first research direction is to build a benchmark for continual knowledge acquisition based on what I call indirect probing questions. These are natural user queries whose correct answers depend on a recent world update, even though the query itself does not explicitly ask for that fact to be checked. The goal is to test whether a system can internalize new knowledge and apply it when its relevance is implicit.

Figure 2 shows the kind of failure I aim to study. The user asks for recommendations for potential PhD supervisors, and Gemini places Prof. Bhattacharyya at the top of its list, likely drawing on prior knowledge of his research contributions. However, Prof. Bhattacharyya passed away on 5 October 2025. The problem is not merely that the model lacks a recent fact, which it could retrieve. Rather, the user's query gives no explicit signal that the model should verify whether each recommended professor is still available to supervise students. A human aware of the update would naturally avoid such a recommendation.

To construct this benchmark, I extract updated facts from Wikidata after a chosen knowledge cutoff date, including events such as deaths, changes in official positions, and organizations being dissolved. For each update, I collect supporting context from sources such as the entity's Wikipedia page and use LLMs to generate natural user queries in which the update is necessary for a correct answer but is not directly mentioned. Evaluation can then check whether the model's answer violates the verified update, for example by recommending an entity that should now be excluded. In this way, indirect probing turns changing world knowledge into a dynamic and automated benchmark while avoiding the main weakness of direct latest-information QA, where the model is told in advance what it needs to look up.

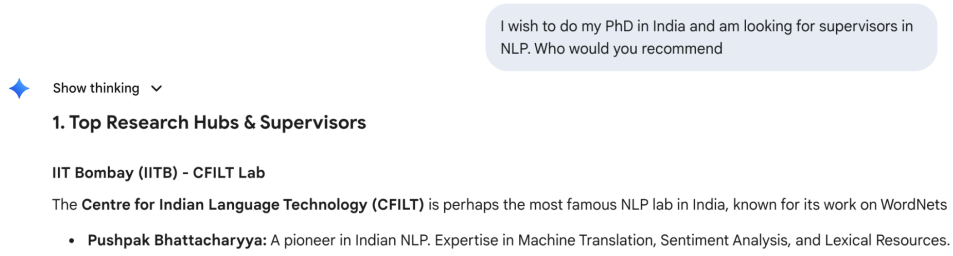

Figure 2: Indirect probing example. Gemini recommends Prof. Bhattacharyya as a potential PhD supervisor, although he passed away on 5 October 2025. The failure requires applying recent world knowledge that the user's query does not explicitly request.

4.2 Consequence-Aware Continual Learning

The benchmark above exposes a gap between factual recall and knowledge that changes behavior. My next objective is to develop methods that close this gap by making a model ask, whenever it learns that the world has changed, how that change should affect future behavior.

A natural starting point is continual pre-training (CPT), the predominant approach for updating LLMs with new knowledge [19]. In CPT, a model continues training on raw text by predicting the next likely word [20]. This can improve performance on direct factual questions, where the query closely matches the new information [21]. However, it is less reliable when the new knowledge must be used compositionally, recognized under paraphrase, or applied in situations where its relevance is only indirect [21].

I believe this limitation is partly a data problem. Raw text is often not the right unit of experience for continual learning. A news article, Wikipedia edit, or Wikidata update may contain the relevant fact, but not present it in the forms in which the model will need to use it later. Prior work showing that paraphrased or question-answer-style training data can improve factual knowledge injection supports the broader point that how update data is organized matters [22, 23]. If the goal is to change future behavior, then training data should include not only the new fact, but also examples of when that fact should matter.

The central hypothesis of my methods agenda is that new knowledge becomes more useful when it is organized around its consequences before the model is updated. Rather than training only on raw text, I will treat each verified change in the world as an opportunity for the model to reason from its own prior knowledge about what is now different, what remains true, and which future questions or decisions should be affected. This moves part of the work from the moment a user asks a question to the moment the system encounters new information. For instance, if an athlete changes teams, the model should infer that this affects questions about current rosters and upcoming matches, but not historical facts about earlier seasons. These inferred consequences can then be turned into training examples, including direct factual questions, indirect probing questions, before-and-after comparisons, preservation examples for still-valid knowledge, and cases that teach the model when retrieval is needed. This gives me a concrete first research question. Do existing update mechanisms become more effective when the model is trained on consequence-aware examples rather than raw sources alone?

As follow-up work, I will study how consequence-aware learning should interact with retrieval-augmented systems. Retrieval is often sufficient for answering a direct information-seeking query [24]. If a user asks about a recent event, the system can search for current evidence and use it in the response. But from the perspective of continual learning, this interaction should not end when the answer is produced. The same evidence can also become a candidate for learning. Once verified, it can be organized into consequence-aware training data, not to replace retrieval, but to make useful changes persist beyond the original interaction. This view treats RAG as a mechanism for both immediate access and longer-term learning. Over time, a system that learns from retrieval episodes should become better at recognizing when its prior knowledge may be stale, deciding what needs to be checked, and applying newly verified information in later queries that do not explicitly ask for it.

5 Broader Impact and Relevance to Google

Knowledge acquisition provides a concrete setting for the broader view of continual learning I advocate. Deployed AI systems must identify useful changes in the world, convert raw updates from sources such as news, web pages, and knowledge bases into usable knowledge, and apply that knowledge later without degrading existing capabilities. The indirect probing benchmark makes this problem measurable, while consequence-aware continual learning makes it actionable by converting verified updates into training signals that shape future behavior.

This agenda is closely aligned with Google's mission to organize the world's information and make it universally accessible and useful [25]. Google already organizes and retrieves information at global scale. My research advances Google's mission by helping users benefit from the world's changing information, even when they do not know which updates are relevant to their needs.

References

[1] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural networks*, vol. 113, pp. 54–71, 2019.

[2] CoLLAs, "Fifth conference on lifelong learning agents," Online, 2026. [Online]. Available: https://lifelong-ml.cc/

[3] M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars, "A continual learning survey: Defying forgetting in classification tasks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 7, pp. 3366–3385, 2021.

[4] R. Hadsell, D. Rao, A. A. Rusu, and R. Pascanu, "Embracing change: Continual learning in deep neural networks," *Trends in Cognitive Sciences*, vol. 24, no. 12, pp. 1028–1040, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364661320302199

[5] P. Buzzega, M. Boschini, A. Porrello, D. Abati, and S. Calderara, "Dark experience for general continual learning: a strong, simple baseline," *Advances in neural information processing systems*, vol. 33, pp. 15 920–15 930, 2020.

[6] L. Wang, X. Zhang, Q. Li, M. Zhang, H. Su, J. Zhu, and Y. Zhong, "Incorporating neuro-inspired adaptability for continual learning in artificial intelligence," *Nature Machine Intelligence*, vol. 5, no. 12, pp. 1356–1368, 2023.

[7] S. Dohare, J. F. Hernandez-Garcia, P. Rahman, A. R. Mahmood, and R. S. Sutton, "Maintaining plasticity in deep continual learning," *arXiv preprint arXiv:2306.13812*, 2023.

[8] D. Lopez-Paz and M. Ranzato, "Gradient episodic memory for continual learning," *Advances in neural information processing systems*, vol. 30, 2017.

[9] G. M. Van de Ven and A. S. Tolias, "Three scenarios for continual learning," *arXiv preprint arXiv:1904.07734*, 2019.

[10] G. Kim, C. Xiao, T. Konishi, Z. Ke, and B. Liu, "Open-world continual learning: Unifying novelty detection and continual learning," *Artificial Intelligence*, vol. 338, p. 104237, 2025.

[11] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell, "Curiosity-driven exploration by self-supervised prediction," in *International conference on machine learning*. PMLR, 2017, pp. 2778–2787.

[12] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel *et al.*, "Mastering chess and shogi by self-play with a general reinforcement learning algorithm," *arXiv preprint arXiv:1712.01815*, 2017.

[13] H. Zhou, A. Yu, Y. Fan, J. Shi, L. Kang, H. Geng, Y. Zhang, Y. Fan, Y. Wu, T. He *et al.*, "Livesearchbench: An automatically constructed benchmark for retrieval and reasoning over dynamic knowledge," *arXiv preprint arXiv:2511.01409*, 2025.

[14] J. Ouyang, T. Pan, M. Cheng, R. Yan, Y. Luo, J. Lin, and Q. Liu, "Hoh: A dynamic benchmark for evaluating the impact of outdated information on retrieval-augmented generation," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 6036–6063.

[15] A. Özer and Ç. Yıldız, "Question answering under temporal conflict: Evaluating and organizing evolving knowledge with llms," *arXiv preprint arXiv:2506.07270*, 2025.

[16] Y. Kim, J. Yoon, S. Ye, S. Bae, N. Ho, S. J. Hwang, and S.-Y. Yun, "Carpe diem: On the evaluation of world knowledge in lifelong language models," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2024, pp. 5401–5415.

[17] Q. Lin, J. Li, and H. T. Ng, "Dynaquest: A dynamic question answering dataset reflecting real-world knowledge updates," in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025, pp. 26 918–26 936.

[18] Y. Li, T. Xu, K. Tang, K. Livescu, D. McAllester, and J. Zhou, "Okbench: Democratizing llm evaluation with fully automated, on-demand, open knowledge benchmarking," *arXiv preprint arXiv:2511.08598*, 2025.

[19] T. Wu, L. Luo, Y.-F. Li, S. Pan, T.-T. Vu, and G. Haffari, "Continual learning for large language models: A survey," *arXiv preprint arXiv:2402.01364*, 2024.

[20] K. Gupta, B. Thérien, A. Ibrahim, M. L. Richter, Q. Anthony, E. Belilovsky, I. Rish, and T. Lesort, "Continual pre-training of large language models: How to (re) warm your model?" *arXiv preprint arXiv:2308.04014*, 2023.

[21] A. O. Li and T. Goyal, "Memorization vs. reasoning: Updating LLMs with new knowledge," in *Findings of the Association for Computational Linguistics: ACL 2025*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 25 853–25 874. [Online]. Available: https://aclanthology.org/2025.findings-acl.1326/

[22] Z. Jiang, Z. Sun, W. Shi, P. Rodriguez, C. Zhou, G. Neubig, X. Lin, W.-t. Yih, and S. Iyer, "Instruction-tuned language models are better knowledge learners," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 5421–5434.

[23] M. Kang, S. J. Hwang, G. Lee, and J. Cho, "Latent paraphrasing: perturbation on layers improves knowledge injection in language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 119 689–119 716, 2024.

[24] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.

[25] Google, "Our approach to search," Online, 2026. [Online]. Available: https://www.google.com/intl/en_us/search/howsearchworks/our-approach/