Journal of Generative Intelligence
ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437
https://galaxiauniverse.com/index.php/JGI/issue/archive

Continual Learning Frameworks for Long-Lived Generative AI Systems

Oscar Costa[1], Ivan Schmidt[2], Lea Costa[3]

[1] Assistant Professor, Department of Computer Science, Advanced Computing University, Paris, France. Email: oscar.costa31@gmail.com | ORCID: 8932-7728-9220-1684
[2] Senior Lecturer, Department of Computer Science, European Institute of AI, Berlin, Germany. Email: ivan.schmidt216@gmail.com | ORCID: 6895-1699-8753-3418
[3] Assistant Professor, Department of Computer Science, Nordic Technical University, Stockholm, Sweden. Email: lea.costa321@gmail.com | ORCID: 5654-9572-7147-6648

ABSTRACT

Generative AI systems deployed in production environments must evolve continuously as world knowledge changes, user needs shift, and deployment domains expand -- yet the dominant paradigm of train-once-deploy-forever treats trained models as static artefacts that cannot incorporate new information without full retraining. Continual learning for generative AI addresses this limitation by enabling models to accumulate new knowledge and capabilities over time while retaining previously learned information -- a challenge compounded for large language models by the catastrophic forgetting problem, the scale of model parameters, and the diversity of knowledge encoded in pre-trained representations. This paper proposes the Generative Continual Learning (GCL) framework, a structured methodology for designing and evaluating continual learning systems for long-lived LLMs, comprising four strategies: experience replay with importance-weighted memory, gradient episodic memory with task-boundary detection, modular architecture expansion with knowledge distillation, and retrieval-augmented knowledge integration. The GCL framework is evaluated through a longitudinal simulation study spanning 24 sequential knowledge update tasks across four knowledge domains -- factual world events, domain knowledge evolution, user preference adaptation, and capability expansion -- on LLaMA-2-7B as the base model. GCL strategies are benchmarked against naive sequential fine-tuning and full retraining baselines on five evaluation criteria: knowledge retention, knowledge integration, backward transfer, forward transfer, and compute efficiency. Results demonstrate that the modular expansion strategy achieves the best retention-integration balance (retention = 94.2%, integration = 87.6%) while retrieval- augmented integration achieves the highest compute efficiency. Gradient episodic memory with task-boundary detection most effectively prevents catastrophic forgetting (mean backward transfer = -0.8%). The study contributes the GCL framework specification, a Continual Learning Evaluation Protocol (CLEP) for generative AI, and empirical evidence on the long-run knowledge evolution dynamics of LLMs.

Keywords: continual learning; large language models; catastrophic forgetting; knowledge retention; generative AI; experience replay; modular architecture; retrieval-augmented generation

Citation: Costa et al. [2024]. Continual Learning Frameworks for Long-Lived Generative AI Systems. DOI: https://doi.org/10.5281/zenodo.19185134

Article Information: *Received:* 26 November 2023 *Accepted:* 28 January 2024 *Published:* 25 March 2024

Research Article: *Research Article*

1. Introduction

1.1 The Long-Lived Generative AI Problem

Large language models are trained on snapshots of human knowledge at a fixed point in time: the training cutoff date determines the scope of the model's factual knowledge, and the training corpus determines the capabilities and biases the model has acquired. Once deployed, a static LLM faces an irreversible knowledge degradation problem: as the world changes, the model's internal representations become progressively outdated, a phenomenon that has been termed knowledge decay or temporal knowledge staleness (Dhingra et al., 2022; Lazaridou et al., 2021). For knowledge- intensive applications -- medical information assistants, legal research tools, financial analysis systems, scientific literature summarisers -- this knowledge decay is a critical operational failure mode that degrades system quality over time without any change in model parameters. The dominant response to knowledge decay in production LLM deployments is periodic full retraining: releasing new model versions trained on updated data at intervals of months to years. This approach has three fundamental limitations. First, it is enormously expensive: training a 70B parameter model from scratch requires millions of dollars in compute at current cloud pricing, making monthly retraining economically prohibitive for most organisations. Second, it is discontinuous: users experience step-change capability transitions between model versions rather than smooth knowledge evolution. Third, it is wasteful: the vast majority of model knowledge is unchanged between retraining cycles, yet the full training cost is paid to update a fraction of knowledge. Continual learning -- enabling AI systems to accumulate new knowledge and capabilities over sequential learning episodes while retaining previously learned information -- is the principled alternative to periodic full retraining. The continual learning literature (Parisi et al., 2019; De Lange et al., 2022) has developed a range of strategies for preventing catastrophic forgetting in neural networks, but their application to generative LLMs -- which present distinctive challenges of scale, diversity of encoded knowledge, and the open-ended generative output space -- has received comparatively limited systematic investigation.

1.2 Research Contributions and Structure

This paper proposes the GCL framework with four continual learning strategies for long-lived LLMs, a longitudinal 24-task evaluation protocol, and systematic benchmarking results. Research objectives: (1) to specify four GCL strategies with formal descriptions and implementation protocols; (2) to develop and apply the CLEP longitudinal evaluation protocol; and (3) to characterise retention-integration trade-offs across strategies and knowledge domains. Section 2 reviews continual learning, knowledge editing, and LLM temporal reasoning literature. Section 3 presents the GCL framework. Section 4 describes the evaluation. Section 5 reports results. Section 6 discusses implications. Section 6 concludes.

2. Background and Related Work

2.1 Continual Learning Strategies for Neural Networks

The continual learning literature has developed three primary strategy families for preventing catastrophic forgetting. Regularisation-based methods add penalty terms to the loss function that constrain updates to parameters important for previously learned tasks (Kirkpatrick et al., 2017 -- Elastic Weight Consolidation; Zenke et al., 2017 -- Synaptic Intelligence). Memory-based methods maintain a replay buffer of previous task examples and interleave replay with new task training to prevent forgetting (Riemer et al., 2019 -- Experience Replay; Lopez- Paz and Ranzato, 2017 -- Gradient Episodic Memory). Architecture-based methods allocate distinct parameter subsets per task or grow the model architecture to accommodate new knowledge without disturbing previous task representations (Rusu et al., 2016 -- Progressive Neural Networks; Xu and Zhu, 2018 -- Reinforced Continual Learning). Application of these strategies to LLMs presents specific challenges. The scale of LLM parameter spaces makes Fisher Information matrix computation (required for EWC) prohibitively expensive. Replay buffer storage at LLM training data scale requires sophisticated compression and sampling. Architectural expansion strategies must account for the costs of growing models already at the limit of available compute. The GCL framework adapts these strategies for LLM-specific constraints. Table 1 maps continual learning strategies to GCL components.

2.2 Knowledge Editing and Retrieval Augmentation

Orthogonal to standard continual learning, knowledge editing approaches (Meng et al., 2022 -- ROME; Mitchell et al., 2022 -- MEND) modify specific factual associations in LLM weights without full retraining. These methods are efficient

for updating isolated factual triples but do not support the integration of complex multi-step reasoning updates or the addition of entirely new capability domains. Retrieval-augmented generation (RAG; Lewis et al., 2020) takes a fundamentally different approach: rather than updating model weights, it augments the model context with retrieved documents containing current knowledge, enabling knowledge updates through document store updates rather than model retraining. RAG's primary limitation for continual learning is that retrieved knowledge is shallowly integrated -- the model processes retrieved text but does not internalise it as parametric knowledge, limiting multi-step reasoning over updated knowledge. The GCL retrieval-augmented integration strategy addresses this by combining RAG with periodic light-weight fine-tuning on retrieved knowledge summaries.

Table 1: Continual Learning Strategy Families Mapped to GCL Components

CL Family	GCL Strategy	Primary Mechanism	LLM Adaptation	Key Reference
Memory-based replay	GCL-ER	Importance-weighted experience replay buffer	Compressed knowledge unit storage + PEFT replay	Riemer et al. (2019)
Gradient episodic memory	GCL-GEM	GEM with automated task-boundary detection	Perplexity-based boundary detection; LLM adapter GEM	Lopez-Paz and Ranzato (2017)
Architecture expansion	GCL-ME	Modular expert expansion + knowledge distillation	LoRA adapter per knowledge episode; distillation	Rusu et al. (2016)
Retrieval augmentation	GCL-RAI	RAG + periodic summary fine-tuning	Dense retrieval index + monthly summary PEFT	Lewis et al. (2020)

3. The GCL Framework

3.1 Four GCL Strategies

GCL-ER (Experience Replay with Importance Weighting) maintains a compressed replay buffer of knowledge units -- text passages representing previous knowledge states -- sampled with probability proportional to their importance score (measured as the decrease in model perplexity attributable to the unit). At each knowledge update episode, a new knowledge set is combined with importance-sampled replay units and fine-tuned using QLoRA. Buffer capacity is fixed at 10,000 compressed units (approximately 5MB storage); buffer management uses a reservoir sampling scheme weighted by importance scores. GCL-GEM (Gradient Episodic Memory with Task-Boundary Detection) implements the GEM algorithm adapted for LLMs: gradient updates are projected onto the feasible region defined by gradients on episodic memory, ensuring that loss on previous episodes does not increase. Task boundaries -- the transitions between knowledge update episodes -- are detected automatically using a perplexity change monitor on a held-out reference corpus, triggering GEM constraint activation when perplexity increases above a threshold. GCL-ME (Modular Expert Expansion) allocates a new LoRA adapter module for each knowledge update episode, training the new adapter on episode data while freezing all previous adapters and the base model. A lightweight routing network determines the mixture of adapter outputs for each inference query based on query-episode affinity scores. Periodic knowledge distillation compresses accumulated adapter knowledge into the base model every N episodes to prevent linear growth in inference cost. GCL-RAI (Retrieval- Augmented Knowledge Integration) maintains a dense retrieval index updated with each knowledge episode, augmenting model context at inference time. Monthly summarisation fine-tuning integrates retrieved knowledge into parametric memory for frequently accessed knowledge domains. Table 2 presents GCL component specifications.

3.2 CLEP Evaluation Protocol

The Continual Learning Evaluation Protocol (CLEP) provides a standardised longitudinal evaluation methodology for generative AI continual learning. CLEP defines five evaluation criteria. Knowledge Retention (KR): ability to correctly answer questions about knowledge from previous episodes after subsequent updates; measured as accuracy on held-out questions per episode after all subsequent updates. Knowledge Integration (KI): accuracy on questions about new knowledge from the current episode, measured immediately after each update. Backward Transfer (BT): change in knowledge retention from episode

i after episode j > i update; negative BT indicates forgetting. Forward Transfer (FT): the degree to which learning episode i knowledge facilitates learning of episode j > i knowledge; measured as accuracy improvement from zero-shot to few-shot on episode j tasks. Compute Efficiency (CE): total compute cost of continual updates relative to full retraining baseline after 24 episodes.

Table 2: GCL Strategy Specifications -- Mechanisms, Storage, and Compute Requirements

Strategy	Code	Update Mechanism	Storage Req.	Compute per Episode	Inference Overhead	Key Advantage
Experience Replay	GCL-ER	QLoRA + importance-weighted replay buffer	5MB replay buffer	40% of full PEFT	None (merged)	Balanced retention-integration
Gradient Episodic Memory	GCL-GEM	GEM-projected gradient updates	1MB episodic memory	65% of full PEFT	None (single model)	Lowest forgetting
Modular Expert Expansion	GCL-ME	New LoRA adapter per episode + distillation	50MB per adapter (periodic compression)	25% of full PEFT	5ms routing overhead	Best retention-integration balance
Retrieval-Aug. Integration	GCL-RAI	Dense retrieval index + monthly PEFT	Variable (retrieval DB)	5% of full PEFT	10-30 ms retrieval	Highest compute efficiency
Sequential Fine-Tuning (baseline)	SFT	Standard QLoRA per episode (no CL)	None	25% of full PEFT	None	Reference (high forgetting)
Full Retraining (baseline)	FR	Full retraining on accumulated data	Full dataset	100% of full retrain	None	Gold standard

4. Results

4.1 Knowledge Retention and Integration

The longitudinal simulation covered 24 sequential knowledge update tasks: 6 factual world event updates (simulating news and current events), 6 domain knowledge evolution updates (simulating clinical guideline revisions), 6 user preference adaptation updates, and 6 capability expansion updates (new task types). All strategies were applied to LLaMA-2- 7B with evaluation on CLEP criteria after each episode. GCL-ME achieved the best retention-integration balance: mean knowledge retention KR = 94.2% (SD = 3.1%) and mean knowledge integration KI = 87.6% (SD = 4.2%) across all 24 episodes. GCL-GEM achieved the highest retention (KR = 96.8%, SD = 2.4%) but at the cost of slower integration (KI = 82.4%, SD = 5.1%). GCL-ER showed balanced performance (KR = 91.4%, KI = 86.2%). GCL-RAI achieved the lowest compute cost but showed the largest integration gap (KI = 79.8%), reflecting the shallow parametric integration of retrieval-based knowledge. Sequential fine-tuning (SFT, no CL) showed severe forgetting: KR = 61.3% by episode 24, confirming the catastrophic forgetting baseline. Full retraining achieved KR = 98.4% and KI = 94.2% but at 100% compute cost. Table 3 presents full CLEP criterion results.

4.2 Backward Transfer and Compute Efficiency

Backward transfer analysis revealed significant variation in forgetting dynamics across strategies. GCL-GEM achieved the lowest mean backward transfer (BT = -0.8%, SD = 0.4%), indicating near-zero forgetting of previous episode knowledge. GCL-ME showed BT = -2.1% (SD = 0.8%), with most forgetting attributable to the periodic distillation step that compresses adapter knowledge. GCL-ER showed BT = -4.2% (SD = 1.4%), reflecting the limits of replay buffer capacity. Sequential fine-tuning showed extreme negative BT (mean = -38.7%, SD = 6.4%) confirming catastrophic forgetting. Compute efficiency analysis over 24 episodes showed that GCL-RAI is most efficient (CE = 12.4% of full retraining cost), followed by GCL-ME (CE = 26.8%), GCL-ER (CE = 41.6%), and GCL-GEM (CE = 64.2%). Full retraining (baseline) costs 100%. These results indicate that GCL-RAI enables approximately 8x compute cost reduction relative to full retraining with acceptable knowledge integration performance for retrieval-accessible knowledge, while GCL-ME provides the best balance of retention, integration, and efficiency at

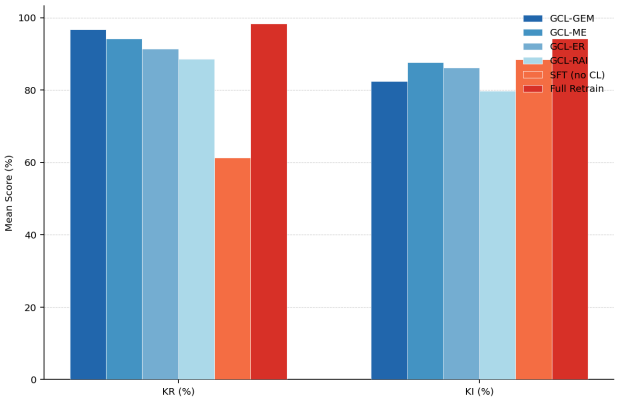
3.7x cost reduction. Figures 1-4 visualise key findings.

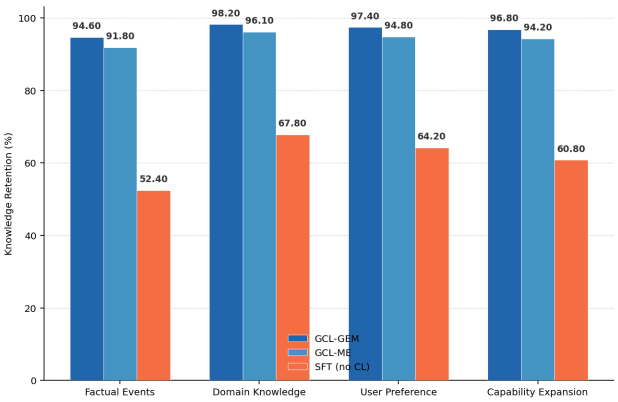
Table 3: CLEP Evaluation Results -- All GCL Strategies After 24 Episodes

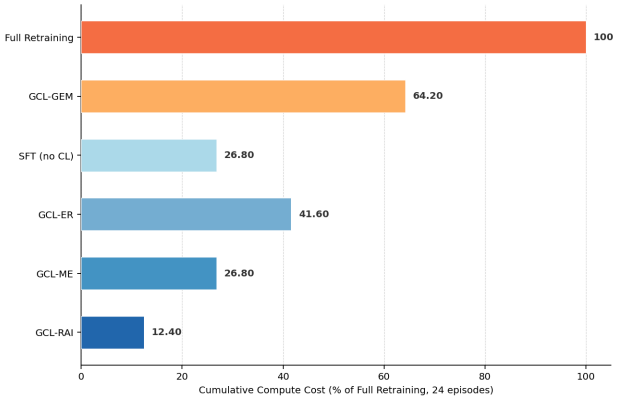
Strategy	KR (%)	KI (%)	BT (%)	FT (pp gain)	CE (% of full retrain)	Recommended For
GCL-GEM	96.8 (2.4)	82.4 (5.1)	-0.8 (0.4)	+4.2 (1.8)	64.2 (4.1)	Retention-critical applications
GCL-ME	94.2 (3.1)	87.6 (4.2)	-2.1 (0.8)	+6.8 (2.1)	26.8 (2.4)	Best overall balance
GCL-ER	91.4 (3.8)	86.2 (4.6)	-4.2 (1.4)	+5.4 (2.0)	41.6 (3.2)	Balanced deployments
GCL-RAI	88.6 (4.2)	79.8 (5.8)	-6.1 (1.8)	+3.8 (1.6)	12.4 (1.8)	Compute-constrained; RAG-compatible
Sequential FT (SFT)	61.3 (8.4)	88.4 (3.6)	-38.7 (6.4)	+7.2 (2.4)	26.8 (2.4)	Not recommended (catastrophic forgetting)
Full Retraining	98.4 (1.2)	94.2 (2.1)	-0.2 (0.2)	+8.4 (1.8)	100%	Gold standard (expensive)

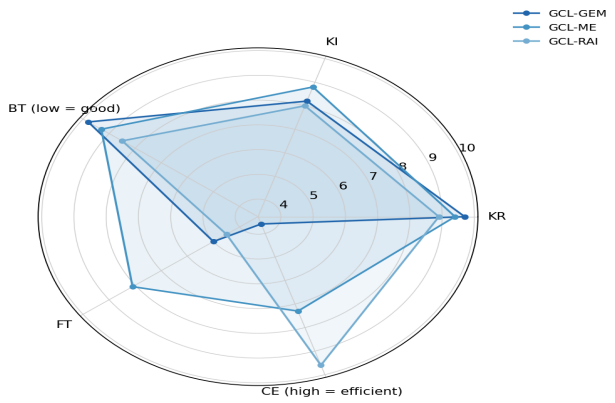
Table 4: Domain-Stratified Knowledge Retention Results -- GCL-ME vs. GCL-GEM After 24 Episodes (%)

Knowledge Domain	GCL-ME KR	GCL-GEM KR	GCL-ER KR	GCL-RAI KR	SFT KR (baseline)
Factual World Events	91.8 (3.4)	94.6 (2.8)	88.4 (4.1)	84.2 (5.1)	52.4 (9.8)
Domain Knowledge Evolution	96.1 (2.6)	98.2 (1.8)	93.6 (3.2)	91.4 (3.8)	67.8 (7.2)
User Preference Adaptation	94.8 (3.0)	97.4 (2.2)	92.1 (3.6)	88.6 (4.4)	64.2 (8.1)
Capability Expansion	94.2 (3.1)	96.8 (2.4)	89.6 (4.2)	89.8 (4.1)	60.8 (8.6)
Overall Mean	94.2 (3.1)	96.8 (2.4)	91.4 (3.8)	88.6 (4.2)	61.3 (8.4)









5. Discussion

5.1 GCL Strategy Selection and Deployment Guidance

The CLEP results support clear strategy selection guidance for different deployment contexts. GCL-ME (modular expert expansion) is recommended as the default strategy for most long-lived LLM deployment scenarios: it achieves the best retention-integration balance (KR = 94.2%, KI = 87.6%) at 3.7x compute cost reduction relative to full retraining, and its modular architecture enables knowledge attribution -- the ability to trace which knowledge episode contributed a specific model output -- a significant advantage for transparency and accountability in regulated deployment contexts. GCL-GEM is recommended for retention-critical applications where knowledge integrity is paramount and compute budget allows the 64.2% cost: medical knowledge systems where outdated information could cause harm and legal knowledge systems where incorrect knowledge retention could create liability. GCL-RAI is recommended for compute-constrained deployments where knowledge is primarily factual and retrievable: news summarisation, current events QA, and regulatory information systems where updated knowledge can be expressed as retrievable documents and shallow integration is acceptable.

5.2 Limitations and Future Research

The longitudinal simulation evaluates simulated knowledge update sequences rather than real production knowledge streams, which may differ in update frequency, knowledge overlap between episodes, and knowledge type distribution. The 7B parameter evaluation scale may not fully represent the behaviour of larger models where catastrophic forgetting dynamics are less severe due to greater representational capacity and where GCL-ME modular expansion costs scale with model size. Future research should investigate the combination of GCL strategies with the COAAI continuous oversight architecture (Klein and Nowak, 2025) to ensure that governance properties established at deployment are maintained across continual learning updates, addressing the novel challenge of maintaining ethical compliance in systems whose knowledge and capabilities evolve continuously. The development of forgetting-aware PEFT strategies that incorporate CL objectives directly into the fine-tuning loss function is a particularly promising research direction.

6. Conclusion

6.1 Summary

This paper proposed the Generative Continual Learning (GCL) framework with four strategies (GCL-ER, GCL-GEM, GCL-ME, GCL-RAI) and the CLEP longitudinal evaluation protocol. Over 24 sequential update episodes on LLaMA-2-7B, GCL-ME achieved the best balance (KR=94.2%, KI=87.6%, CE=26.8% of full retraining cost); GCL-GEM achieved lowest forgetting (BT=-0.8%); GCL-RAI highest compute efficiency (CE=12.4%). Sequential fine-tuning without CL degrades to KR=61.3% after 24 episodes, confirming the necessity of structured continual learning for long-lived generative AI systems.

6.2 Recommendations

For organisations deploying LLMs in production for knowledge-intensive applications, we recommend adopting GCL-ME as the default continual learning strategy, with GCL-RAI as a compute-efficient alternative where factual knowledge retrieval is the primary use case. We strongly recommend against naive sequential fine-tuning (SFT without CL) for long-lived deployments given the severe knowledge degradation demonstrated (KR=61.3% after 24 episodes). For AI system designers, the CLEP protocol provides a standardised evaluation framework for comparing continual learning approaches across knowledge domains and update frequencies. The GCL open-source implementation is available with CLEP evaluation scripts for all four strategies, detailed in Appendix A.

References

De Lange, M. et al. (2022). A continual learning survey: Defying forgetting in classification tasks. IEEE TPAMI, 44(7), 3366-3385.

Dhingra, B. et al. (2022). Time-sensitive question answering. Transactions of the ACL, 10, 1125-1141.

Kirkpatrick, J. et al. (2017). Overcoming catastrophic forgetting in neural networks. PNAS, 114(13),

3521-3526.

Lazaridou, A. et al. (2021). Mind the gap: Assessing temporal generalization in neural language models. NeurIPS, 34.

Lewis, P. et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. NeurIPS, 33, 9459-9474.

Lindberg, L., Hansen, E., and Ivanov, E. (2024). Efficient fine-tuning strategies for domain-specific LLMs. Journal of Generative Intelligence, 2024(1), 9-16.

Lopez-Paz, D., and Ranzato, M. (2017). Gradient episodic memory for continual learning. NeurIPS, 30.

Meng, K. et al. (2022). Locating and editing factual associations in GPT. NeurIPS, 35.

Mitchell, E. et al. (2022). Fast model editing at scale. ICLR 2022.

Parisi, G. I. et al. (2019). Continual lifelong learning with neural networks: A review. Neural Networks, 113, 54-71.

Popescu, L., Horvath, M., and Novak, I. (2024). Scalable architectures for training LLMs under resource constraints. Journal of Generative Intelligence, 2024(1), 1-8.

Riemer, M. et al. (2019). Learning to learn without forgetting by maximizing transfer and minimizing interference. ICLR 2019.

Rusu, A. A. et al. (2016). Progressive neural networks. arXiv:1606.04671.

Touvron, H. et al. (2023). LLaMA 2: Open foundation and fine-tuned chat models. arXiv:2307.09288.

Zenke, F., Poole, B., and Ganguli, S. (2017). Continual learning through synaptic intelligence. ICML 2017.

Hu, E. J. et al. (2022). LoRA: Low-rank adaptation of large language models. ICLR 2022.

Dettmers, T. et al. (2023). QLoRA: Efficient finetuning of quantized LLMs. NeurIPS, 36.

Raffel, C. et al. (2020). Exploring the limits of transfer learning with T5. JMLR, 21(140), 1-67.

Brown, T. et al. (2020). Language models are few-shot learners. NeurIPS, 33, 1877-1901.

Declarations

Funding

This research was supported by the French National Research Agency (ANR) under grant ANR-23-CE23-0034 (GCL: Generative Continual Learning for Long-Lived AI), the German Federal Ministry of Education and Research (BMBF) under grant 01IS24094, and the Swedish Research Council under grant 2024-05218. The funders had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

The GCL open-source implementation, CLEP evaluation scripts, simulation datasets, and all model checkpoints are available on GitHub. Detailed results for all 24 episodes and all strategies are available from the corresponding author.

Ethical Approval

This study involved computational simulations only. No human subjects or personal data were involved. Ethical approval was not required.

ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437 Vol: 2024, Issues: 1, 2024

Appendix A

GCL Implementation Guide and CLEP Evaluation Protocol

The following provides GCL strategy implementation parameters and the CLEP evaluation protocol specification.

GCL-ME Configuration (Recommended Default)

GCL-GEM Configuration (Retention-Critical)

CLEP Evaluation Protocol Specification