

Elastic Weight Consolidation for Knowledge Graph Continual Learning: An Empirical Evaluation

Gaganpreet Jhajj^{1,*}, Fuhua Lin¹

¹School of Computing and Information Systems, Athabasca University, Canada

Abstract

Knowledge graphs (KGs) require continual updates as new information emerges, but neural embedding models suffer from catastrophic forgetting when learning new tasks sequentially. We evaluate Elastic Weight Consolidation (EWC), a regularization-based continual learning method, on KG link prediction using TransE embeddings on FB15k-237. Across multiple experiments with five random seeds, we find that EWC reduces catastrophic forgetting from 12.62% to 6.85%, a 45.7% reduction compared to naive sequential training. We observe that the task partitioning strategy affects the magnitude of forgetting: relation-based partitioning (grouping triples by relation type) exhibits 9.8 percentage points higher forgetting than randomly partitioned tasks (12.62% vs 2.81%), suggesting that task construction influences evaluation outcomes. While focused on a single embedding model and dataset, our results demonstrate that EWC effectively mitigates catastrophic forgetting in KG continual learning and highlight the importance of evaluation protocol design.

Keywords

Continual Learning, Knowledge Graphs, Elastic Weight Consolidation, Link Prediction

1. Introduction

Knowledge graphs (KGs) depict structured information as networks of entities and their relations [1], facilitating a variety of applications, including question answering [2] and recommendation systems [3] and educational systems [4, 5, 6, 7, 8, 9]. These real-world KGs evolve continuously as new information becomes available and existing knowledge is refined. Neural embedding models, such as TransE [10], generate vector representations of entities and relations for link prediction. However, adapting these models to accommodate new information while retaining prior knowledge presents a significant challenge.

Catastrophic forgetting occurs when neural networks, trained sequentially on multiple tasks, experience significant performance degradation on earlier tasks after learning new ones [11]. This phenomenon poses particular challenges for KG embeddings, where maintaining consistent representations across evolving information is essential. While continual learning methods have been developed for image classification and natural language processing, their effectiveness on KG link prediction remains underexplored.

We investigate how Elastic Weight Consolidation (EWC) [12], a regularization-based continual learning method, performs on KG link prediction. EWC protects important parameters learned in previous tasks by adding a quadratic penalty based on the Fisher Information Matrix, allowing networks to learn new tasks while preserving performance on old ones. We evaluate EWC on TransE embeddings using FB15k-237, a standard KG benchmark, and compare relation-based task partitioning (where triples are grouped by relation type) to random partitioning.

Our experiments reveal that EWC substantially reduces catastrophic forgetting. On relation-based partitioned tasks, naive sequential training results in 12.62% forgetting (measured as MRR degradation from post-task performance), while EWC with regularization strength $\lambda = 10$ reduces this to 6.85%, a 45.7% reduction. This demonstrates that regularization-based continual learning effectively preserves KG embeddings across sequential tasks.

NORA'25: 1st Workshop on Knowledge Graphs & Agentic Systems Interplay co-located with NeurIPS, Dec.1, 2025, Mexico City, Mexico

 gjhajj1@learn.athabascau.ca (G. Jhajj); oscarl@athabascau.ca (F. Lin)

 0000-0001-5817-0297 (G. Jhajj); 0000-0002-5876-093X (F. Lin)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

We also observe that the task partitioning strategy significantly affects the measured forgetting. Naive sequential training on relation-based tasks exhibits 12.62% forgetting, compared to only 2.81% on randomly partitioned tasks, a 9.8% point difference. This suggests that evaluation protocols, particularly how tasks are constructed from datasets, influence continual learning metrics and should be carefully explored when it comes time to design the study.

Our study focuses on TransE embeddings for the FB15k-237 dataset across four relation-based tasks. While this scope limits generalizability, it provides rigorous evidence that EWC reduces catastrophic forgetting in KG continual learning and raises essential questions about task construction in continual learning evaluation.

This work addresses a critical challenge for knowledge graph-based AI agents: maintaining knowledge representations as new information arrives. As agents increasingly rely on KG-based memory and reasoning, the ability to incorporate new knowledge while preserving existing facts becomes essential for long-term autonomous operation.

2. Related Work

KG Embeddings. TransE [10] represents relations as translations in embedding space, learning vectors such that $\mathbf{h} + \mathbf{r} \approx \mathbf{t}$ for true triples (h, r, t) . Extensions include TransH [13], RotatE [14], and ComplEx [15].

Continual Learning. Methods for mitigating catastrophic forgetting include regularization approaches like EWC [12] and Learning without Forgetting [16], replay-based methods that store and revisit previous examples [17], and architectural approaches that allocate separate parameters for different tasks [18]. EWC estimates parameter importance using the Fisher Information Matrix and adds regularization penalties to protect important weights during subsequent task training.

KG Continual Learning. Prior work has explored continual learning for KGs across various contexts. Wang et al. [19] studied multi-task feature learning for KG-enhanced recommendation. Daruna et al. [20] demonstrated continual learning methods for KG embeddings in robotic manipulation tasks, evaluating multiple architectures including TransE, DistMult, and ComplEx. Zhao et al. [21] introduced the PS-CKGE benchmark, focusing on pattern shifts and demonstrating that these shifts exacerbate catastrophic forgetting beyond what would be expected from simple data scaling. Recent work has also examined embedding adaptation [22] and temporal KGs. While these comprehensive benchmarking efforts establish the landscape of KG continual learning, a focused empirical analysis of classic regularization methods like EWC, with explicit investigation of task partitioning effects, remains limited, motivating our study.

3. Methodology

3.1. Problem Formulation

A KG is $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T})$ where \mathcal{E} is the set of entities, \mathcal{R} is the set of relations, and $\mathcal{T} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ is the set of true triples. TransE learns embeddings $\mathbf{h}, \mathbf{r}, \mathbf{t} \in \mathbb{R}^d$ by minimizing:

$$\mathcal{L} = \sum_{(h, r, t) \in \mathcal{T}} \sum_{(h', r, t') \in \mathcal{T}'} \max(0, \gamma + d(\mathbf{h} + \mathbf{r}, \mathbf{t}) - d(\mathbf{h}' + \mathbf{r}, \mathbf{t}')) \quad (1)$$

where \mathcal{T}' contains negative samples, $d(\cdot, \cdot)$ is L2 distance, and γ is the margin.

In continual learning, we partition \mathcal{G} into tasks $\mathcal{G}_1, \dots, \mathcal{G}_T$ and train sequentially. After training on task i , we measure performance M_i^j on task $j \leq i$. Forgetting for task j after learning task i is:

$$F_i^j = M_j^j - M_i^j \quad \text{for } i > j \quad (2)$$

We report average forgetting at the end of training:

$$\bar{F} = \frac{1}{T-1} \sum_{j=1}^{T-1} F_T^j \quad (3)$$

3.2. Elastic Weight Consolidation

EWC protects important parameters by adding a regularization term to the loss when training on task i :

$$\mathcal{L}_{\text{EWC}}^i = \mathcal{L}^i + \frac{\lambda}{2} \sum_k F_k (\theta_k - \theta_{k,i-1}^*)^2 \quad (4)$$

where $\theta_{k,i-1}^*$ are optimal parameters after task $i-1$, F_k is the Fisher Information diagonal approximation, and λ controls regularization strength. The Fisher Information Matrix diagonal is:

$$F_k = \mathbb{E}_{(h,r,t) \sim \mathcal{G}_{i-1}} \left[\left(\frac{\partial \log p(y|x; \theta)}{\partial \theta_k} \right)^2 \right] \quad (5)$$

We compute this empirically using all triples from the previous task, processed in mini-batches (see Appendix A for implementation details).

3.3. Task Partitioning

We evaluate two task partitioning strategies to assess how task construction affects catastrophic forgetting:

Relation-based partitioning: We partition FB15k-237 by grouping all triples that share the same relations. To create balanced task sizes, we sort the 237 relations by frequency (number of triples per relation) and assign them to four tasks in round-robin order: the most frequent relation goes to Task 1, the second-most to Task 2, the third-most to Task 3, the fourth-most to Task 4, the fifth-most back to Task 1, and so forth. This ensures each task receives approximately 59 relations with a mixture of common and rare relation types, while maintaining *relation-level coherence*—all triples with the same relation appear in the same task.

Random partitioning: We randomly shuffle all 272,115 training triples and divide them into four equal chunks of approximately 68,000 triples each. This distributes relation types across all tasks, creating *relation-level overlap*—most relations appear in multiple tasks. The key difference: relation-based partitioning creates distinct distribution shifts between tasks (each task focuses on different relation types). In contrast, random partitioning minimizes distribution shift (each task is a representative sample of all relations). This allows us to isolate the effect of task boundary definition on the difficulty of continual learning.

4. Experimental Setup

Dataset and Partitioning. We use FB15k-237 [23], which contains 14,505 entities, 237 relations, and 272,115 triples. Using relation-based partitioning (round-robin assignment by relation frequency), we create four tasks with approximately balanced sizes, each containing 59 relations and their associated triples.

Model Configuration. We use TransE with 50-dimensional embeddings, margin $\gamma = 1.0$, and L2 distance. Training uses the Adam optimizer [24] with a learning rate of 0.001, a batch size of 256, and 20 epochs per task. We found 20 epochs sufficient for convergence on FB15k-237 in preliminary experiments.

Methods Evaluated. We compare naive sequential training (no continual learning) with EWC at multiple regularization strengths ($\lambda \in \{0.1, 1.0, 10.0\}$), EWC combined with experience replay (500 examples per task), and replay-only baselines (random and wave-based sampling).

Evaluation Protocol. We use Mean Reciprocal Rank (MRR) for link prediction, computing filtered rankings that exclude known actual triples. For each task j , we record MRR immediately after training (M_j^j) and after each subsequent task (M_i^j for $i > j$). We run five random seeds (42, 123, 456, 789, 2024) and report means and standard deviations.

Hardware. Experiments ran on NVIDIA RTX 3070 Ti (8GB) with 20 hours total computation for 80 experiments (8 methods \times 5 seeds \times 2 partitioning strategies).

5. Results

5.1. Classical Methods on Relation-Based Partitioning

Table 1 shows forgetting results for classical continual learning methods on relation-based partitioned tasks.

Method	Forgetting (%)	Final MRR
Naive	12.62 ± 0.35	0.206 ± 0.006
EWC ($\lambda = 0.1$)	10.44 ± 0.26	0.229 ± 0.005
EWC ($\lambda = 1.0$)	7.51 ± 0.44	0.250 ± 0.006
EWC ($\lambda = 10$)	6.85 ± 0.33	0.242 ± 0.004
EWC + Wave Replay	9.91 ± 0.20	0.234 ± 0.005
Random Replay	13.78 ± 0.44	0.196 ± 0.006
Wave Replay	12.54 ± 0.14	0.216 ± 0.007

Table 1

Forgetting and final MRR on relation-based partitioned tasks (mean \pm std over 5 seeds).

EWC with $\lambda = 10$ achieves the best performance, reducing forgetting to 6.85% (std 0.33%), a 45.7% reduction compared to naive training. This demonstrates that regularization-based protection of important parameters effectively mitigates catastrophic forgetting in KG continual learning. Final MRR also improves from 0.206 to 0.242, indicating that EWC preserves not only previous task performance but also maintains overall embedding quality.

Interestingly, replay-based methods underperform. Random replay achieves 13.78% forgetting, worse than naive training, suggesting that simply revisiting old examples without principled parameter protection may interfere with learning. Wave-based replay performs similarly (12.54% forgetting). Combining EWC with wave replay (9.91% forgetting) improves over replay alone but underperforms pure EWC, indicating that regularization is the primary driver of performance.

Figure 1 visualizes these results, showing a clear separation between EWC and other methods.

5.2. Effect of Task Partitioning

We compare relation-based and random partitioning strategies to assess whether task construction affects measured forgetting. Table 2 and Figure 2 show results.

Partitioning	Naive Forgetting (%)	EWC Forgetting (%)
Relation-based	12.62 ± 0.35	6.85 ± 0.33
Random	2.81 ± 0.34	5.08 ± 0.22
Difference	9.81 pp	1.77 pp

Table 2

Forgetting comparison between relation-based and random task partitioning (mean \pm std over 5 seeds).

Relation-based partitioning results in substantially higher forgetting during naive training (12.62% vs 2.81%), a 9.8 percentage-point difference. We hypothesize this occurs because relation-based partitioning creates *task coherence*: each task focuses on a distinct subset of relation types (approximately 59 relations

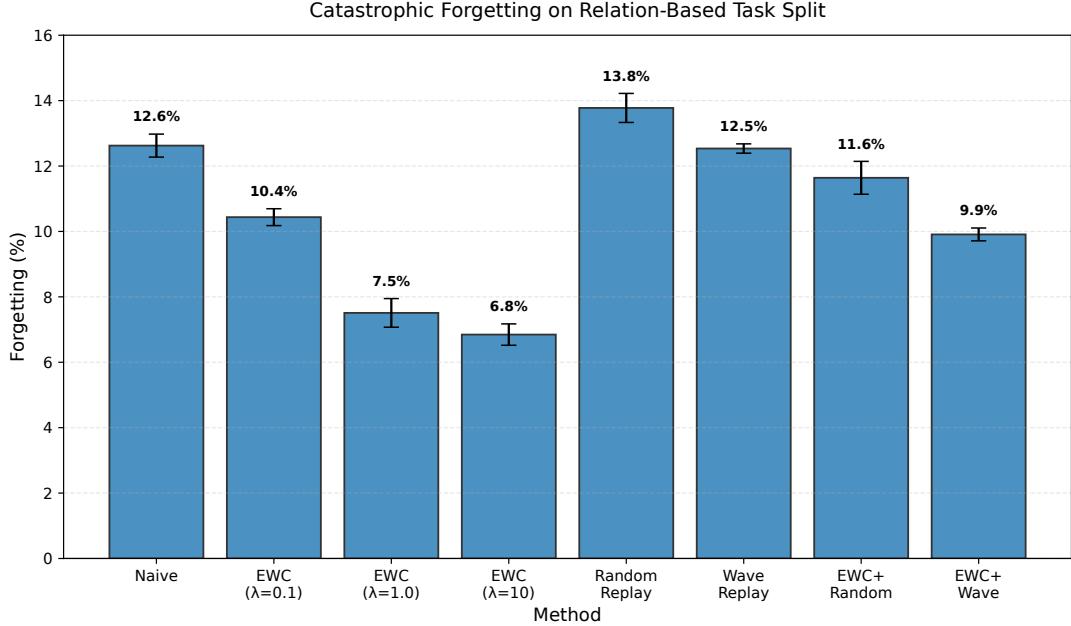


Figure 1: Catastrophic forgetting on relation-based partitioned tasks. EWC ($\lambda = 10$) substantially reduces forgetting compared to naive sequential training and replay-based methods.

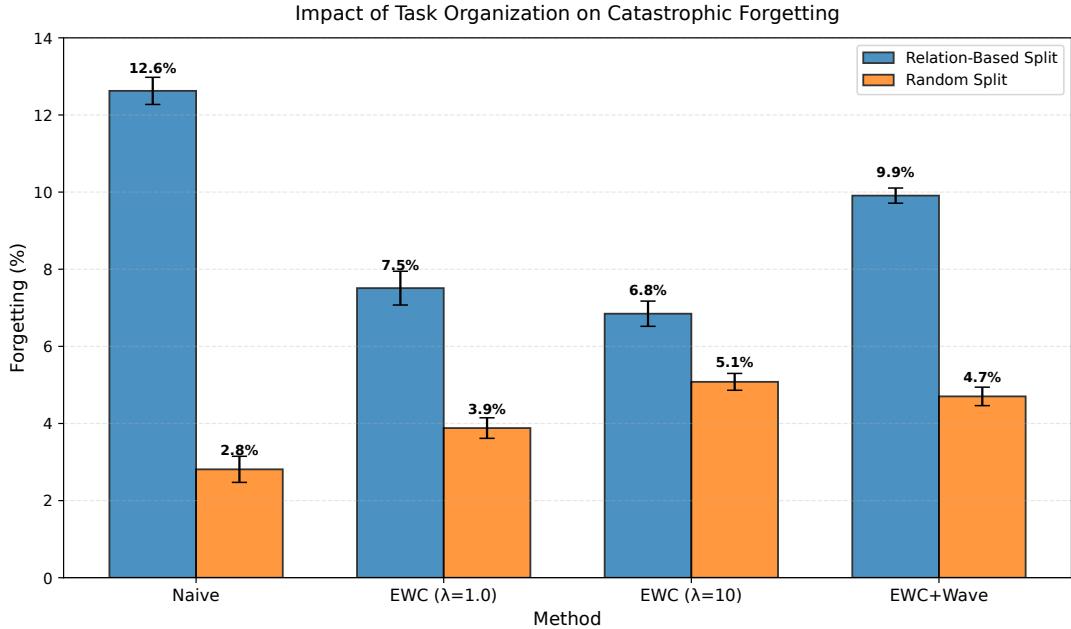


Figure 2: Effect of task partitioning on forgetting. Relation-based partitioning creates more challenging continual learning scenarios, with 9.8 percentage points of higher forgetting in naive training compared to random partitioning.

per task), inducing larger distribution shifts when transitioning between tasks. When the model moves from Task 1’s relations to Task 2’s completely different relation set, the parameter updates required are more disruptive to previously learned representations. In contrast, random partitioning creates *relation-level overlap*: each task contains a representative sample of most relation types, so the distribution shift between tasks is minimal. This naturally regularizes learning, as the model continually encounters all relation types across tasks. This observation suggests that task construction significantly influences the difficulty of continual learning and that evaluation protocols should explicitly consider and report partitioning strategies.

Notably, EWC reduces this gap: the difference between relation-based and random partitioning under EWC is only 1.77 percentage points (6.85% vs 5.08%), suggesting that effective continual learning methods can generalize across different task construction approaches.

6. Discussion

KG embeddings have structured parameter spaces where specific dimensions encode semantic properties. The Fisher Information Matrix identifies parameters critical for encoding relation types and entity characteristics learned in previous tasks. By protecting these parameters, EWC enables new-task learning while preserving the semantic structure of the embedding space.

The superior performance of EWC compared to replay methods suggests that principled parameter protection is more effective than simply revisiting old examples when working memory (replay buffer) is limited. This aligns with neuroscience findings that synaptic consolidation, rather than replay alone, enables long-term memory retention.

Our EWC forgetting rate (6.85%) on relation-based partitioned tasks demonstrates effective continual learning for KG link prediction. While direct comparison with prior work is challenging due to differences in datasets and evaluation protocols, our results are consistent with EWC’s performance on image classification tasks [12] and suggest that regularization-based continual learning generalizes to structured knowledge representations.

The task partitioning effect we observe (a 9.8 percentage-point difference) highlights an important consideration for continual learning evaluation: reported forgetting rates depend on how tasks are constructed. This suggests that future work should explicitly consider task construction methodology when designing experiments and reporting results.

7. Limitations and Future Work

Our study has several limitations that define the scope and future directions.

We evaluate only TransE on FB15k-237 across four tasks, a deliberate choice given consumer-grade GPU constraints (NVIDIA RTX 3070 Ti, 8GB) that enabled rigorous multi-seed experimentation within feasible time scales. Results may not generalize to: (1) complex embedding methods (RotatE, ComplEx, TuckER); (2) other datasets (WN18RR, YAGO, Wikidata); (3) more extended task sequences (10+ tasks). We compare EWC with basic replay baselines, but do not exhaustively benchmark all continual learning methods. Our relation-based partitioning uses round-robin frequency assignment; alternative strategies (entity-based, domain-based) warrant investigation.

Experiments were constrained to academic resources without HPC clusters or large-scale GPU infrastructure. While this motivated our focused design, a comprehensive multi-model evaluation would require substantially more resources. We view this as proof-of-concept evidence for future large-scale studies.

Future work should evaluate EWC across multiple embedding methods and datasets to assess generalizability. Scaling studies with 10+ tasks would reveal the long-term dynamics of continual learning. Systematic investigation of task construction strategies could formalize the relationship between task partitioning and forgetting. Additionally, combining EWC with more sophisticated replay strategies or architectural approaches may yield further improvements.

A particularly promising direction is to evaluate continual learning methods on educational KGs, where knowledge evolves as curriculum content updates and student learning data accumulates. Recent work has highlighted the potential of open-source LLMs to transform educational contexts through adaptive and personalized learning [25]. However, the challenge of maintaining such systems as they continually learn from new educational content remains underexplored. Building on recent work in curriculum modeling and adaptive learning [6], we plan to investigate how EWC performs when domain-specific educational KGs are continually updated with new learning resources, pedagogical relations, and student interaction patterns. The integration of open-source LLMs (e.g., Llama [26],

Mistral [27], Qwen [28]) with educational KGs for personalized learning support presents unique continual learning challenges: both the symbolic KG structure and LLM parameters must adapt to new content while preserving existing pedagogical knowledge. Our findings on task partitioning effects may inform how educational content updates should be structured to minimize interference with previously learned material.

Neuromorphic approaches using spiking neural networks with spike-timing-dependent plasticity offer promising directions for extending this work. Building on prior work demonstrating SNN-based relational inference and knowledge representation for knowledge graphs [29], we plan to investigate whether the biological learning mechanisms inherent in STDP can provide natural solutions to catastrophic forgetting. The structured, relational nature of KGs may align particularly well with neuromorphic computation, where synaptic plasticity mechanisms could enable task-consolidation without explicit regularization.

While our preliminary experiments on consumer-grade GPUs were inconclusive, a comprehensive evaluation using specialized neuromorphic hardware (Intel Loihi, IBM TrueNorth, SpiNNaker) could reveal whether biological learning mechanisms provide advantages for continual learning in knowledge graphs. The structured, relational nature of KGs may align well with neuromorphic computation, particularly for consolidating knowledge across tasks. Given EWC’s success through parameter-wise importance weighting, bio-inspired plasticity mechanisms that selectively strengthen or weaken synaptic connections based on usage patterns warrant thorough exploration with appropriate hardware infrastructure.

8. Conclusion

We evaluated EWC for KG continual learning using TransE embeddings on FB15k-237. Across multiple experiments with five random seeds, we found that EWC reduces catastrophic forgetting from 12.62% to 6.85%, a 45.7% reduction compared to naive sequential training. This demonstrates that regularization-based continual learning effectively preserves KG embeddings across sequential tasks.

We also observed that the task partitioning strategy significantly affects measured forgetting: relation-based partitioning exhibits 9.8 percentage points higher forgetting than randomly partitioned tasks for naive training. This suggests that the evaluation protocol’s design, particularly the task-construction methodology, influences continual learning measurements and should be carefully considered in experimental design.

While our study focuses on a single embedding model and dataset, it provides rigorous evidence for EWC’s effectiveness and raises essential questions about evaluation methodology in KG continual learning.

Acknowledgments

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), Alberta Innovates, Alberta Advanced Education, and Athabasca University, Canada. We would also like to thank the reviewers for their suggestions on how to improve this work.

Declaration on Generative AI

During the preparation of this work, the author(s) used Grammarly and Claude (Anthropic) for Grammar and spelling checks.

References

- [1] A. Sheth, S. Padhee, A. Gyrard, Knowledge graphs and knowledge networks: the story in brief, *IEEE Internet Computing* 23 (2019) 67–75.
- [2] G. Jhajj, Y. Nomura, Jack and the beansTALK: Towards question answering in plant biology, in: Eighth Widening NLP Workshop (WiNLP 2024) Phase II, 2024. URL: <https://openreview.net/forum?id=0DlJEPHHK>.
- [3] Q. Guo, F. Zhuang, C. Qin, H. Zhu, X. Xie, H. Xiong, Q. He, A survey on knowledge graph-based recommender systems, 2020. URL: <https://arxiv.org/abs/2003.00911>. arXiv: 2003.00911.
- [4] G. Jhajj, X. Zhang, J. R. Gustafson, F. Lin, M. P.-C. Lin, Educational knowledge graph creation and augmentation via LLMs, in: International Conference on Intelligent Tutoring Systems, Springer, 2024, pp. 292–304.
- [5] M. R. Kabir, F. Lin, An LLM-Powered Adaptive Practicing System., in: LLM@ AIED, 2023, pp. 43–52.
- [6] F. Lin, R. Morland, Curriculum Modeling for Adaptive Learning, Springer Nature Switzerland, 2025, p. 57–72. URL: http://dx.doi.org/10.1007/978-3-031-92970-0_5. doi:10.1007/978-3-031-92970-0_5.
- [7] G. Jhajj, F. Lin, Augmenting japanese language acquisition via llms and asr, in: IEEE Smart World Congress 2025 (IEEE SWC'25), Calgary, Canada, 2025, p. 3.84.
- [8] J. R. D. Gustafson, X. Zhang, G. Jhajj, F. Lin, Representing and tracing students' cognitive processes in Project-Based learning through the Function-Behavior-Structure framework and knowledge graphs, in: IEEE Smart World Congress 2025 (IEEE SWC'25), Calgary, Canada, 2025, p. 6.56.
- [9] J. R. D. Gustafson, G. Jhajj, X. Zhang, F. O. Lin, Enhancing Project-Based Learning With a GenAI Tool Based on Retrieval: Augmented Generation and Knowledge Graphs, IGI Global, 2024, p. 161–194. URL: <http://dx.doi.org/10.4018/979-8-3693-5443-8.ch006>. doi:10.4018/979-8-3693-5443-8.ch006.
- [10] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, *Advances in neural information processing systems* 26 (2013).
- [11] M. McCloskey, N. J. Cohen, Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem, in: *Psychology of Learning and Motivation*, volume 24, Elsevier, 1989, pp. 109–165. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0079742108605368>. doi:10.1016/S0079-7421(08)60536-8.
- [12] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, R. Hadsell, Overcoming catastrophic forgetting in neural networks, *Proceedings of the National Academy of Sciences* 114 (2017) 3521–3526. URL: <https://pnas.org/doi/full/10.1073/pnas.1611835114>. doi:10.1073/pnas.1611835114.
- [13] Z. Wang, J. Zhang, J. Feng, Z. Chen, Knowledge Graph Embedding by Translating on Hyperplanes, *Proceedings of the AAAI Conference on Artificial Intelligence* 28 (2014). URL: <https://ojs.aaai.org/index.php/AAAI/article/view/8870>. doi:10.1609/aaai.v28i1.8870.
- [14] Z. Sun, Z.-H. Deng, J.-Y. Nie, J. Tang, RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space, in: International Conference on Learning Representations, 2019. URL: <https://openreview.net/forum?id=HkgEQnRqYQ>.
- [15] T. Trouillon, J. Welbl, S. Riedel, E. Gaussier, G. Bouchard, Complex Embeddings for Simple Link Prediction, in: M. F. Balcan, K. Q. Weinberger (Eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, PMLR, New York, New York, USA, 2016, pp. 2071–2080. URL: <https://proceedings.mlr.press/v48/trouillon16.html>.
- [16] Z. Li, D. Hoiem, Learning without forgetting, *IEEE transactions on pattern analysis and machine intelligence* 40 (2017) 2935–2947.
- [17] D. Rolnick, A. Ahuja, J. Schwarz, T. Lillicrap, G. Wayne, Experience Replay for Continual Learning, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d. Alché-Buc, E. Fox,

R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 32, Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/fa7cdfad1a5aaf8370ebeda47a1ff1c3-Paper.pdf.

- [18] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, R. Hadsell, Progressive Neural Networks, 2022. URL: <https://arxiv.org/abs/1606.04671>.
- [19] H. Wang, F. Zhang, M. Zhao, W. Li, X. Xie, M. Guo, Multi-task feature learning for knowledge graph enhanced recommendation, in: The world wide web conference, 2019, pp. 2000–2010.
- [20] A. Daruna, M. Gupta, M. Sridharan, S. Chernova, Continual learning of knowledge graph embeddings, IEEE Robotics and Automation Letters 6 (2021) 1128–1135. doi:10.1109/LRA.2021.3056071.
- [21] T. Zhao, J. Chen, Y. Ru, Q. Lin, Y. Geng, H. Zhu, Y. Pan, J. Liu, Rethinking continual knowledge graph embedding: Benchmarks and analysis, in: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 138–147. URL: <https://doi.org/10.1145/3726302.3730073>. doi:10.1145/3726302.3730073.
- [22] M. Delange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, T. Tuytelaars, A continual learning survey: Defying forgetting in classification tasks, IEEE Transactions on Pattern Analysis and Machine Intelligence (2021) 1–1. URL: <https://ieeexplore.ieee.org/document/9349197>. doi:10.1109/TPAMI.2021.3057446.
- [23] K. Toutanova, D. Chen, Observed versus latent features for knowledge base and text inference, in: A. Allauzen, E. Grefenstette, K. M. Hermann, H. Larochelle, S. W.-t. Yih (Eds.), Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality, Association for Computational Linguistics, Beijing, China, 2015, pp. 57–66. URL: <https://aclanthology.org/W15-4007/>. doi:10.18653/v1/W15-4007.
- [24] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2017. URL: <https://arxiv.org/abs/1412.6980>. arXiv:1412.6980.
- [25] M. P.-C. Lin, D. Chang, S. Hall, G. Jhajj, Preliminary Systematic Review of Open-Source Large Language Models in Education, Springer Nature Switzerland, 2024, p. 68–77. URL: http://dx.doi.org/10.1007/978-3-031-63028-6_6. doi:10.1007/978-3-031-63028-6_6.
- [26] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. URL: <https://arxiv.org/abs/2302.13971>. arXiv:2302.13971.
- [27] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, 2023. URL: <https://arxiv.org/abs/2310.06825>. arXiv:2310.06825.
- [28] Qwen, ., A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, Z. Qiu, Qwen2.5 technical report, 2025. URL: <https://arxiv.org/abs/2412.15115>. arXiv:2412.15115.
- [29] G. Jhajj, J. R. D. Gustafson, R. Morland, C. E. Gutierrez, M. P.-C. Lin, M. A. A. Dewan, F. Lin, Neuromorphic Knowledge Representation: SNN-Based Relational Inference and Explainability in Knowledge Graphs, in: International Conference on Intelligent Tutoring Systems, Springer, 2025, pp. 159–165.

A. EWC Implementation Details

We compute the Fisher Information Matrix diagonal after training each task using all triples from that task, processed in mini-batches of 256. For each batch b , we compute the empirical Fisher approximation:

$$F_k \approx \frac{1}{B} \sum_{b=1}^B \left(\frac{\partial \mathcal{L}_b}{\partial \theta_k} \right)^2 \quad (6)$$

where B is the total number of batches in the task and \mathcal{L}_b is the loss on batch b . This accumulates squared gradients across all training examples from the previous task.

During training on task i , we apply the EWC penalty for all previous tasks:

$$\mathcal{L}_{\text{total}}^i = \mathcal{L}^i + \sum_{j=1}^{i-1} \frac{\lambda}{2} \sum_k F_k^j (\theta_k - \theta_{k,j}^*)^2 \quad (7)$$

B. Random Partitioning Results

Table 3 shows full results for random partitioning.

Method	Forgetting (%)	Final MRR
Naive	2.81 ± 0.34	0.277 ± 0.008
EWC ($\lambda = 0.1$)	2.88 ± 0.34	0.270 ± 0.005
EWC ($\lambda = 1.0$)	3.88 ± 0.27	0.244 ± 0.005
EWC ($\lambda = 10$)	5.08 ± 0.22	0.222 ± 0.008

Table 3

Full results on random partitioning.

Interestingly, stronger EWC regularization increases forgetting on random partitioning (5.08% for $\lambda = 10$ vs 2.81% naive). We hypothesize that random partitioning naturally distributes relation types across tasks, reducing interference. Strong regularization may over-constrain parameters, preventing necessary adaptation. This suggests that optimal regularization strength depends on task construction.

C. Hyperparameter Sensitivity

Table 4 shows EWC performance across λ values we tested.

λ	Relation-based (%)	Random (%)	Final MRR (Relation-based)
0.1	10.44 ± 0.26	2.88 ± 0.34	0.229 ± 0.005
1.0	7.51 ± 0.44	3.88 ± 0.27	0.250 ± 0.006
10.0	6.85 ± 0.33	5.08 ± 0.22	0.242 ± 0.004

Table 4

EWC performance across tested regularization strengths.

For relation-based partitioning, $\lambda = 10$ achieves the lowest forgetting. For random partitioning, weaker regularization ($\lambda = 0.1$) performs best. This suggests that optimal regularization strength depends on task construction: relation-grouped tasks require stronger protection of essential parameters, while randomly distributed tasks benefit from more flexibility.

D. Performance-Forgetting Trade-off

Figure 3 visualizes the relationship between final MRR and forgetting for all methods.

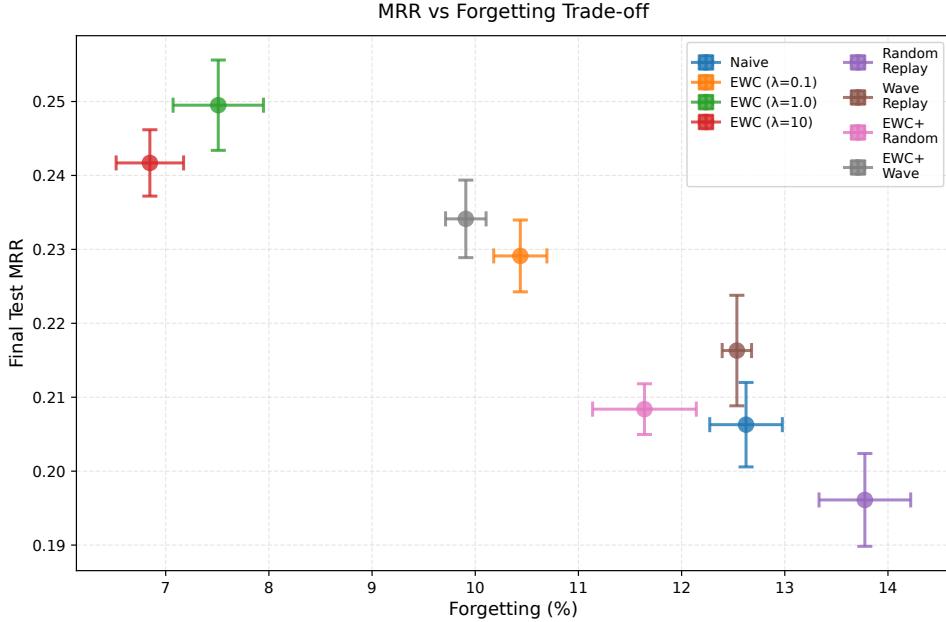


Figure 3: Performance-forgetting trade-off. EWC ($\lambda = 10$) achieves the best balance, with low forgetting (6.85%) and competitive final MRR (0.242).

EWC with $\lambda = 10$ occupies the optimal region (low forgetting, competitive performance). Replay methods cluster in the high-forgetting, low-performance region. This visualization confirms that EWC achieves superior trade-offs compared to alternative approaches.

E. Implementation Details

Code Structure. Experiments implemented in PyTorch 1.13.

Negative Sampling. We use uniform random tail corruption with a 1:1 ratio of positive to negative samples. For each positive triple (h, r, t) , we generate one negative sample (h, r, t') by randomly replacing the tail entity t' sampled uniformly from \mathcal{E} . We use unfiltered negative sampling during training (negative samples may coincidentally be true triples), but employ filtered evaluation for link prediction metrics.

Hardware. Experiments ran on NVIDIA RTX 3070 Ti (8GB) with approximately 18-20 hours total computation for all experimental runs (8 methods \times 5 seeds \times 2 partitioning strategies).

F. Task Retention Visualization

Figure 4 shows how performance on each task degrades over subsequent training.

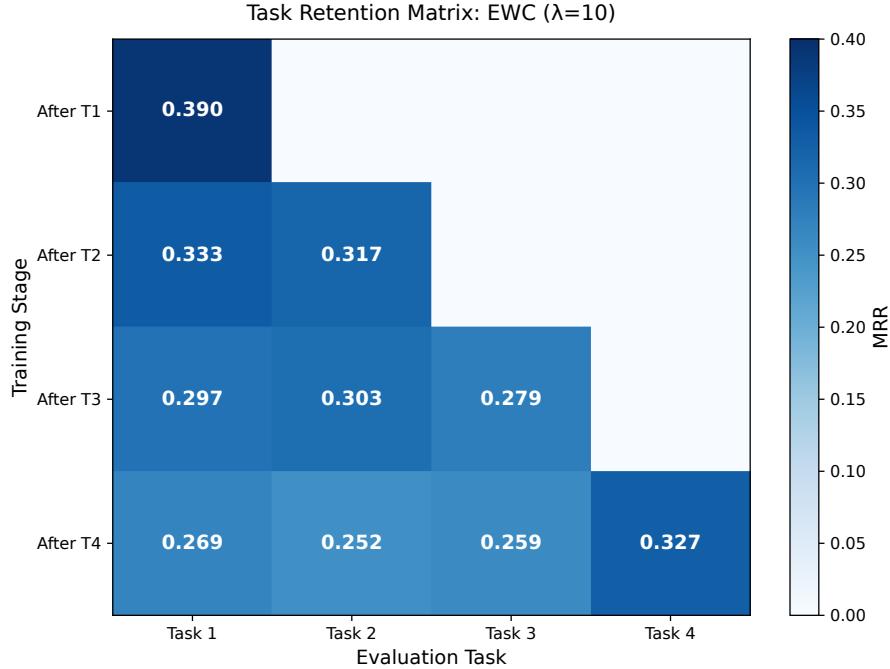


Figure 4: Task retention matrix for EWC ($\lambda = 10$). The diagonal shows performance immediately after learning each task; the line below the diagonal shows retention after subsequent tasks. Minimal degradation indicates effective continual learning.

The heatmap shows that EWC maintains relatively stable performance across tasks, with limited degradation on earlier tasks as new tasks are learned. This visualization confirms that EWC effectively protects previous task performance.