Milestone 3: Research Proposal

Stealthy Backdoors in RAG-Based LLMs

Michael Hudson

July 2025

1 Introduction & Motivation

Retrieval-Augmented Generation (RAG) integrates a Large Language Model (LLM) with an external retrieval module, enabling it to incorporate relevant information from a large corpus of passages during generation. A typical RAG pipeline consists of three components: a knowledge base of passages; a retriever that embeds and ranks passages by semantic similarity to the input query; and an LLM that conditions its output on both the original query and the retrieved context.

This modular design brings practical benefits. Since the knowledge base can be updated independently of the LLM, RAG enables grounded responses with current or domain-specific context without retraining. It has gained traction across industries, including: health, see Wang et al. (2024); finance, see Loukas et al. (2023); and legal, see Mahari (2021). Furthermore, it is often cheaper and faster to deploy than fine-tuning an LLM. However, this modularity also introduces new vulnerabilities. Since the LLM often treats retrieved context as trustworthy, even minor manipulations of the corpus can yield significant behavioural changes. Moreover, corpus poisoning does not require LLM access, only the ability to insert or modify passages, thereby significantly lowering the barrier to entry.

2 Literature Review

Retrieval-Augmented Generation (RAG) pipelines have emerged as a prime target for adversarial attacks, largely due to their modular design and reliance on external corpora. Recent work has sought to exploit this vulnerability through various attack vectors. Zou et al. (2024) propose PoisonedRAG, which demonstrates that a single, adversarially crafted passage — generated using HotFlip-style gradient-based learning — can be inserted into various corpora to reliably hijack retrieval for a fixed query. However, the attack's impact is limited by its specificity to particular queries and the detectability of the resulting adversarial documents, which often exhibit unnatural phrasing.

Xue et al. (2024) put forward a more general approach in BadRAG. Contrastive Optimisation on a Passage (COP) leverages gradient-based learning to craft passages of a given length that are retrieved only in the presence of a specific trigger word in the query. This represents a broader attack with an additional degree of stealth, but still relies on a fixed, user-selected trigger that is held constant during the optimisation process.

In contrast, Cheng et al. (2024) introduce a model-level attack in TrojanRAG, where the retriever is fine-tuned on query-passage pairs containing certain trigger tokens. This inserts persistent backdoors into the retriever's embedding space, allowing the attack to generalise across multiple triggers and queries. TrojanRAG achieves strong results across different retriever architectures and LLMs, but assumes that the attacker can modify the retriever's parameters.

Together, these works expose several attack surfaces in RAG pipelines but stop short of exploring the inverse problem: trigger optimisation for a fixed passage.

In AgentPoison, Chen et al. (2024) present an attack that targets RAG-based LLM agents that follow multi-step reasoning protocols. It jointly optimises both a trigger and the associated adversarial passages using HotFlip-style gradient-based learning. Whilst the attack is effective in influencing agent behaviour, it is designed for RAG-based LLM agents rather than standard RAG-based LLMs. Additionally, its emphasis on multi-step action control leaves open the question of whether joint trigger-passage optimisation can be adapted for retrieval-time attacks in standard RAG-based LLMs.

Other notable work includes that of Chen et al. (2025), which influences the opinions expressed by RAG-based LLMs on sensitive topics by modifying retrieval rankings through surrogate model training. Zhong et al. (2023) demonstrate that inserting a small number of adversarial passages into a corpus can cause retrievers to consistently rank them highly. Wallace et al. (2019) show that fixed triggers can be appended to inputs to cause consistent, targeted failures across various NLP models.

Several techniques facilitate optimisation of discrete language inputs. Ebrahimi et al. (2018)'s HotFlip uses gradient-based character substitutions to craft adversarial passages. Attacks such as PoisonedRAG and AgentPoison adopt similar discrete optimisation techniques at the token level. Jang et al. (2017)'s Gumbel-Softmax provides a continuous approximation of categorical sampling, making it possible to apply gradient-based optimisation to discrete variables.

3 Proposed Contributions

(C1) **Ablation of Adversarial Passage Length**. Conduct a detailed ablation study of adversarial passage length, which is fixed to 50 tokens in the original BadRAG framework. By systematically varying passage length and evaluating its effect on retrieval success and stealth, uncover key trade-offs that inform the design of more efficient and covert attacks.

(C2) **Trigger Optimisation for Fixed Passages**. Reverse the BadRAG problem by fixing the passage and optimising the trigger instead. This formulation introduces a complementary attack surface and facilitates the exploration of how trigger length and position influence retrieval.

(C3) **Joint Trigger–Passage Optimisation**. Build on (C2) by jointly optimising both the trigger and the adversarial passage, treating them as a coupled pair. This formulation will necessitate the use of alternative optimisation techniques and provide insight into how coordinated query and passage manipulations can enhance the effectiveness of retrieval-time attacks.

(C4) **Improving Fluency of Adversarial Passages**. Investigate methods to improve the fluency of adversarial passages generated in (C3), aiming to reduce detectability. Specifically, explore incorporating fluency constraints during optimisation and partially constraining edits to preserve LLM-generated content, balancing attack effectiveness with linguistic plausibility.

4 Potential Impact

This project introduces a novel perspective on corpus poisoning attacks in RAG-based LLMs by systematically exploring the roles of passage length and trigger structure. In addition, the use of gradient-based trigger learning and differentiable optimisation over discrete text expands the methodological toolkit for studying retrieval-time adversarial behaviour.

Each contribution addresses a practically relevant threat model. (C1) deepens understanding of how passage length affects retrieval, informing how attackers might balance control with stealth. (C2) is critical in settings where attackers cannot modify the knowledge base directly but can manipulate user inputs — a common real-world constraint. (C3) generalises prior work by jointly learning query and passage modifications, yielding optimal backdoors that may pose greater risks in multi-step or agentic systems, as highlighted in AgentPoison. (C4) moves the attack beyond proof-of-concept by improving fluency and allowing content to be adversarially controlled — a key step for practical exploitation.

The findings will lay groundwork for future adversarial research, including better understanding of how and why retrieval-time attacks succeed, and how to optimise them for stealth, generalisation, and domain transfer. Even more importantly, this work will guide defensive strategies. By identifying which attack configurations are most effective, it supports the development of targeted mitigations — such as input sanitisation, retrieval auditing, or adversarial training — ultimately contributing to more secure and trustworthy RAG pipelines.

5 Feasibility & Timeline

This project builds on well-established codebases: both PoisonedRAG and BadRAG have open-source implementations available on OpenReview. Experiments will be run on the Natural Questions (NQ) dataset, Kwiatkowski et al. (2019), which is widely used in RAG research. For all experiments, I will work with a 10,000-document subset of NQ, which facilitates efficient experimentation whilst preserving statistical robustness and retrieval diversity. This will be particularly beneficial for the ablation study. I also have access to a GPU server at Imperial, enabling efficient gradient-based optimisation and retrieval evaluation. Given this foundation, the project is technically feasible within the proposed timeline.

w/c	Research Activities	Potential Challenges
7 Jul	(C1) Conduct ablation study on adversarial passage length. Vary the number of tokens in adversarial passages and evaluate performance using rank and similarity metrics. This isolates the role of passage length in retrieval.	Longer passages may dominate retrieval; investigate quantifying the optimal balance between control and stealth.
14 Jul	(C2) Implement trigger optimisation using HotFlip-style updates. Systematically vary trigger length and insertion position (start, end, random). This reveals structural factors that influence trigger efficacy.	Semantically odd triggers may emerge; enforce constraints on tokens used.
21 Jul	(C3) Develop joint optimisation of trigger–passage pairs. Start with HotFlip-style updates, then move to Gumbel-Softmax sampling for smoother convergence. Compare to independent (disjoint) optimisation baselines.	Joint optimisation may be unstable; tune learning rate and sampling temperature.
28 Jul	(C4 part 1) Incorporate fluency constraints into adversarial passage generation. Add a language model perplexity penalty to the loss function.	Fluency constraints may reduce attack effectiveness; tune the weight of the perplexity term.
4 Aug	(C4 part 2) Test constrained optimisation where only the first and/or last tokens of an LLM-generated adversarial passage are editable. Evaluate both fluency and attack effectiveness.	As above.
11 Aug	Finalise all experiments and produce a first draft of the report.	Will need to manage GPU usage across the multiple experiments.
18 Aug	Refine results and discussion sections of the report, ensuring that plots, tables and visualisations are clear.	This may take some time; have planned such that this can spill into w/c 25 Aug.
25 Aug	Final code and report clean-up. Check reproducibility and documentation. Archive datasets and model checkpoints.	

Ethical Considerations. This project studies vulnerabilities in RAG-based LLMs with the goal of improving their robustness. No personal data will be used, and all experiments will be conducted on public datasets. Results will be defence-oriented, in line with established norms in adversarial research.

References

Chen, Z., Liu, J., Gong, Y., Chen, M., Liu, H., Cheng, Q., Zhang, F., Lu, W., Liu, X. and Wang, X. (2025), 'FlippedRAG: Black-Box Opinion Manipulation Adversarial Attacks to Retrieval-Augmented Generation Models'.
URL: *https://arxiv.org/abs/2501.02968*

Chen, Z., Xiang, Z., Xiao, C., Song, D. and Li, B. (2024), AgentPoison: Red-teaming LLM Agents via Poisoning Memory or Knowledge Bases, *in* A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak and C. Zhang, eds, 'Advances in Neural Information Processing Systems', Vol. 37, Curran Associates, Inc., pp. 130185–130213.
URL: *https://arxiv.org/abs/2407.12784*

Cheng, P., Ding, Y., Ju, T., Wu, Z., Du, W., Yi, P., Zhang, Z. and Liu, G. (2024), 'TrojanRAG: Retrieval-Augmented Generation Can Be Backdoor Driver in Large Language Models'.
URL: *https://arxiv.org/abs/2405.13401*

Ebrahimi, J., Rao, A., Lowd, D. and Dou, D. (2018), HotFlip: White-Box Adversarial Examples for Text Classification, *in* I. Gurevych and Y. Miyao, eds, 'Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)', Association for Computational Linguistics, Melbourne, Australia, pp. 31–36.
URL: *https://aclanthology.org/P18-2006/*

Jang, E., Gu, S. and Poole, B. (2017), 'Categorical Reparameterization with Gumbel-Softmax'.
URL: *https://arxiv.org/abs/1611.01144*

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q. and Petrov, S. (2019), 'Natural Questions: A Benchmark for Question Answering Research', *Transactions of the Association for Computational Linguistics* **7**, 452–466.
URL: *https://aclanthology.org/Q19-1026/*

Loukas, L., Stogiannidis, I., Diamantopoulos, O., Malakasiotis, P. and Vassos, S. (2023), Making LLMs Worth Every Penny: Resource-Limited Text Classification in Banking, *in* 'Proceedings of the Fourth ACM International Conference on AI in Finance', ICAIF '23, Association for Computing Machinery, New York, NY, USA, p. 392–400.
URL: *https://doi.org/10.1145/3604237.3626891*

Mahari, R. Z. (2021), 'AutoLAW: Augmented Legal Reasoning through Legal Precedent Prediction'.
URL: *https://arxiv.org/abs/2106.16034*

Wallace, E., Feng, S., Kandpal, N., Gardner, M. and Singh, S. (2019), Universal Adversarial Triggers for Attacking and Analyzing NLP, *in* K. Inui, J. Jiang, V. Ng and X. Wan, eds, 'Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)', Association for Computational Linguistics, Hong Kong, China, pp. 2153–2162.
URL: *https://aclanthology.org/D19-1221/*

Wang, C., Ong, J., Wang, C., Ong, H., Cheng, R. and Ong, D. (2024), 'Potential for GPT Technology to Optimize Future Clinical Decision-Making Using Retrieval-Augmented Generation', *Annals of Biomedical Engineering* **52**(5), 1115–1118.
URL: *https://doi.org/10.1007/s10439-023-03327-6*

Xue, J., Zheng, M., Hu, Y., Liu, F., Chen, X. and Lou, Q. (2024), 'BadRAG: Identifying Vulnerabilities in Retrieval Augmented Generation of Large Language Models'.
URL: *https://arxiv.org/abs/2406.00083*

Zhong, Z., Huang, Z., Wettig, A. and Chen, D. (2023), Poisoning Retrieval Corpora by Injecting Adversarial Passages, *in* H. Bouamor, J. Pino and K. Bali, eds, 'Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Singapore, pp. 13764–13775.
URL: *https://aclanthology.org/2023.emnlp-main.849/*

Zou, W., Geng, R., Wang, B. and Jia, J. (2024), 'PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models'.
URL: *https://arxiv.org/abs/2402.07867*