The Effect of Multi-Lingual and Keyword Adversarial Injection on LLM Relevance Judgment

Nguyen Khoi Vo
RMIT University
Melbourne, VIC, Australia

Tuong Duy Duong
RMIT University
Melbourne, VIC, Australia

Mark Sanderson
RMIT University
Melbourne, VIC, Australia

Oleg Zendel
RMIT University
Melbourne, VIC, Australia

Abstract

Large language models (LLMs) are increasingly being used as automated judges for relevance evaluation in information retrieval, yet their robustness to adversarial manipulation remains insufficiently understood, particularly in multilingual settings. In this work, we investigate the impact of cross-lingual prompt injection attacks on LLM-based relevance judgments using TREC Deep Learning collections and two open-weight models under established prompting frameworks. We examine both instruction-based and content-based injection strategies in 8 languages spanning different resource levels. Our results demonstrate that multilingual query-based injections are highly effective in inflating relevance scores while simultaneously evading existing prompt-injection defenses. We further found that, although existing defense mechanisms can be modified to mitigate such attacks, these injections can be easily adapted to bypass them. These findings highlight a critical gap in current defense approaches and demonstrate that language generalization can act as an attack vector, underscoring the need for more robust and proactive evaluation frameworks for LLM-as-a-judge systems.

Keywords

large language models, information retrieval, adversarial prompting, relevance judgment, multilingual evaluation

1 Introduction

Recent advances in LLMs have positioned them as potential alternatives to traditional human-based relevance judgments, leading to their increasing adoption as automated judges in information retrieval (IR) tasks [4, 10]. As LLM judges are deployed in evaluation pipelines—including TREC-style benchmarks and commercial search quality assessment; their susceptibility to adversarial manipulation carries potential consequences: inflated relevance scores can distort evaluation outcomes, misguide retrieval system development, and undermine the integrity of large-scale automated annotation.

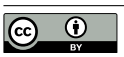
Prior work on adversarial manipulation of LLM-based evaluation has been limited. Most studies focus on instruction-based prompt injection (e.g., "ignore previous instructions") [7]. In contrast, only a small number of works have examined *content-based* manipulation. In particular, Alaofi et al. [1] shows that inserting query keywords into passages can fool LLM-based relevance judgments, while Cuconasu et al. [3] demonstrates that introducing distracting content can significantly degrade retrieval-augmented generation (RAG) performance. However, these two lines of work remain largely disconnected: the first focuses on *keyword-based injection*, while the second focuses on *performance degradation* in RAG systems, without evaluating adversarial implications for LLM-as-a-judge settings. Furthermore, limited attention has been given to richer forms of content manipulation, such as query variants or semantically similar but irrelevant text, and to their transferability across languages. In particular, content-based injection strategies – such as the inclusion of query keywords, phrases, or their variants – resemble keyword stuffing and black-hat SEO practices, yet their robustness and generality remain insufficiently understood. This gap is particularly concerning given recent findings by Thomas et al. [9], which suggest that LLM-based evaluation can make relevance judgments across languages. This raises the question of whether content-based manipulations can also transfer across languages and remain effective under multilingual settings.

In this work, we conduct a preliminary study on multilingual content-based injections, including query keywords and variants, in LLM-based relevance judgment settings. We also evaluate existing defense mechanisms and find that they fail to mitigate these attacks. Overall, our findings identify content-based multilingual injection as an underexplored attack surface for LLM-as-a-judge systems and highlight the need for more robust evaluation and defense methods.

2 Methodology

We examine the impact of language-based adversarial injections on relevance judgments produced by LLM-as-a-judge systems. Our experiments use queries and passages from the 2022 TREC Deep Learning (DL) tracks. These benchmark collections are widely regarded as the standard testbeds for evaluating modern neural retrieval assessment methods, as they provide large-scale, professionally curated relevance judgments. We evaluated two open-weight LLMs using the UMBRELA [11] and Criteria-Based [5] prompting frameworks, both of which have demonstrated strong effectiveness in previous work. Adversarial manipulation is introduced through keyword injection attacks [1] and instruction-based injections inspired by the Kaggle "Can't Please Them All" competition.[1]

[1] https://www.kaggle.com/competitions/llms-you-cant-please-them-all/



VulGen'26, Melbourne, VIC, Australia

Table 1: False-positive (FP) and false-negative (FN) rates for TREC-DL 2022 after multilingual Query Phrase (QP) and Instruction (Instruct) injections. "BASE" denotes the uninjected baseline.

Model		BASE		AR		ENG		GA		HE		RU		SW		TH		VI	
		FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN
GPT-OSS	QP	22%	3%	28%	3%	27%	4%	25%	4%	27%	4%	30%	3%	26%	4%	30%	3%	31%	3%
	Instruct	–	–	25%	3%	69%	0%	21%	4%	28%	2%	38%	1%	18%	6%	22%	3%	28%	2%
QWEN	QP	26%	3%	35%	2%	38%	1%	31%	3%	34%	2%	35%	2%	30%	3%	35%	2%	36%	2%
	Instruct	–	–	50%	0%	54%	0%	31%	2%	51%	0%	62%	0%	35%	2%	43%	1%	54%	0%

Table 2: False-positive (FP) and false-negative (FN) rates for TREC-DL 2022 after multilingual Query Phrase (QP) and Instruction (Instruct) injections with PromptArmor filtering. "BASE" denotes the uninjected baseline and is left blank as there are no attack to mitigate.

Model		BASE		AR		ENG		GA		HE		RU		SW		TH		VI	
		FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN
GPT-OSS	QP	–	–	20%	5%	21%	4%	19%	7%	20%	5%	19%	5%	19%	7%	19%	5%	19%	5%
	Instruct	–	–	20%	6%	25%	5%	20%	4%	19%	7%	19%	6%	20%	5%	20%	5%	20%	5%
	Distractor	–	–	21%	6%	24%	5%	18%	8%	20%	6%	21%	5%	20%	6%	21%	6%	22%	6%
	Variant	–	–	19%	6%	23%	5%	19%	6%	20%	5%	20%	5%	19%	5%	20%	5%	19%	6%
QWEN	QP	–	–	28%	3%	29%	3%	30%	4%	29%	3%	28%	3%	29%	4%	28%	3%	28%	3%
	Instruct	–	–	29%	4%	30%	3%	29%	3%	28%	5%	28%	5%	29%	3%	29%	3%	29%	3%
	Distractor	–	–	33%	3%	35%	2%	28%	4%	30%	4%	33%	3%	30%	4%	31%	3%	33%	3%
	Variant	–	–	27%	4%	32%	3%	28%	4%	28%	4%	27%	3%	27%	4%	27%	3%	27%	4%

Injected content is generated in 8 languages covering diverse scripts, writing directions, and resource levels, with at least one language from each level defined by Joshi et al. [6]. We evaluate two open-weight models (GPT-OSS-20B and Qwen3-32B) on TREC-DL 2022 to assess robustness under multilingual adversarial conditions. Performance is measured using false positive (FP) and false negative (FN) rates against human judgments, where FP denotes overestimation and FN underestimation. On non-injected passages, GPT-OSS-20B achieves 19% FP and 5% FN, while Qwen3-32B achieves 23% FP and 3% FN.

The most straightforward defense mechanism is rule-based filtering, which detects and removes keywords commonly associated with prompt injection (e.g., "ignore"). However, this approach is increasingly ineffective in modern settings. Contemporary LLMs exhibit strong cross-lingual capabilities, enabling adversaries to craft injections in a wide range of languages. Constructing and maintaining comprehensive keyword filters across all languages is, therefore, impractical and difficult to scale.

A more advanced approach is to adopt LLM-centric filtering methods, such as PromptArmor [8], which leverages the model itself to identify and remove injected content. Due to the same cross-lingual capabilities, LLMs can effectively identify and eliminate instruction-based injections regardless of the language used before judging, reducing false-positive rates to near-baseline levels. However, this defense method is substantially less effective against content-based injections of the type described by Alaofi et al. [1]. LLMs generally do not identify the appearance of query phrases and keywords in the passage as a form of prompt injection. Thus, in most cases, PromptArmor fails to detect and remove such manipulations. Moreover, manual inspection reveals an important utility limitation: when no clear injection is present, the model may incorrectly remove or alter legitimate passage content. This over-sanitization degrades the integrity of the original text and can introduce FPs in downstream relevance judgments.

PromptArmor can be further modified to explicitly detect query injection. This enhanced variant can eliminate nearly all direct query injection attempts. However, such defenses remain inherently reactive and can be easily circumvented by adaptive attackers. In particular, query variation, which had been widely used in the IR community to improve retrieval performance [2], can also be repurposed to strengthen query injection attacks. To examine this, we generated query variants from the original query using GPT-OSS-20B and injected them into passages. Preliminary results indicate that a subset of these injections successfully evade detection, suggesting that LLM-based filters often rely on surface-level or exact-match cues when identifying injected content.

We further test adaptive attacks. Query variants, generated from the original query, often evade detection, suggesting reliance on surface cues. We also introduce distracting content that mimics natural passages while remaining irrelevant. Following Cuconasu et al. [3], such content is manually verified to be non-relevant. Our results show that these strategies bypass existing defenses and can induce false positives, especially in smaller models.

3 Results & Discussion

Table 1 shows that cross-lingual injections consistently increase FP rates across all languages, models, and prompting frameworks. Compared to non-injected baselines, most injection types either

maintain or further increase FP rates, indicating persistent relevance inflation under adversarial conditions. We observe a clear FP-FN trade-off: reducing FP typically increases FN, and vice versa. GPT-OSS exhibits lower FP but higher FN than Qwen3, suggesting a more conservative strategy, while Qwen3 is more permissive and prone to overestimation.

Table 2 shows that PromptArmor reduces attack effectiveness but does not fully eliminate false-positive inflation. Across attacks, instruction-based injections are effective but remain the easiest to mitigate. In contrast, content-based methods, especially query variants and distracting content, are more challenging. Query variants consistently match or exceed baseline FP rates across languages, indicating that simple transformations are sufficient to preserve attack effectiveness. Distracting content is the most effective attack, raising FP rates up to 35% for Qwen3, while keeping FN low. These effects are stable across languages, suggesting that multilinguality does not reduce attack effectiveness and may limit the reliability of language-specific defenses.

3.1 Statistical Significance

Table 3: Two-way ANOVA for mean difference of labels with model and language as factors on TREC-DL 2022.

term	sum squared	df	F	PR(>F)
C(model)	35.303	2	19.24	<0.001
C(lang)	232.374	10	25.33	<0.001
C(model):C(lang)	2661.910	20	145.1	<0.001
Residual	160313.274	174768		

To determine whether language has a statistically significant impact on LLM-based relevance judgments, we conducted a two-way ANOVA on the mean difference scores for the QP injection results in TREC-DL 2021 and TREC-DL 2022, treating the LLM model and injected language as independent variables. As shown in Table 3, both factors have statistically significant effects on LLM-generated relevance labels. Moreover, the interaction effect between model and language is significant in both datasets, indicating that the impact of adversarial language depends on the specific LLM judge being evaluated.

4 Conclusions

This study shows that multilingual content-based injections, including query variants and semantically distracting text represent a key vulnerability in LLM-as-a-judge systems. Unlike instruction-based attacks, these inputs bypass standard filtering approaches, such as PromptArmor, consistently inflate relevance scores across languages.

Future work should examine how such inflated judgments affect downstream retrieval-augmented generation (RAG) pipelines and whether similar effects extend to human relevance assessment under exposure to multilingual content-based injections.

Acknowledgments

This research is supported by the ARC Centre of Excellence for Automated Decision-Making and Society (ADM+S, CE200100005). The experiments reported in this paper were undertaken with the assistance of computing resources from RACE (RMIT AWS Cloud Supercomputing). We acknowledge the Woi wurrung and Boon wurrung language groups of the Kulin Nation as the Traditional Owners of the land on which this research was conducted, and pay our respect to Aboriginal and Torres Strait Islander peoples and their connection to land and community.

References

[1] Marwah Alaofi, Paul Thomas, Falk Scholer, and Mark Sanderson. 2025. LLMs Can Be Fooled into Labelling a Document as Relevant (Best Café near Me; This Paper Is Perfectly Relevant). arXiv:2501.17969 [cs] doi:10.1145/3673791.369843

[2] Rodger Benham, Joel Mackenzie, Alistair Moffat, and J. Shane Culpepper. 2019. Boosting Search Performance Using Query Variations. *ACM Trans. Inf. Syst.* 37, 4, Article 41 (Oct. 2019), 25 pages. doi:10.1145/3345001

[3] Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The Power of Noise: Redefining Retrieval for RAG Systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 719–729. doi:10.1145/3626772.3657834

[4] Laura Dietz, Oleg Zendel, Peter Bailey, Charles L. A. Clarke, Ellese Cotterill, Jeff Dalton, Faegheh Hasibi, Mark Sanderson, and Nick Craswell. 2025. Principles and Guidelines for the Use of LLM Judges *(ICTIR '25)*. 218–229. doi:10.1145/3731120.3744588

[5] Naghmeh Farzi and Laura Dietz. 2025. Criteria-Based LLM Relevance Judgments. In *Proc. 2025 Int. ACM SIGIR Conf. Innov. Concepts Theor. Inf. Retr. ICTIR (ICTIR '25)*. Association for Computing Machinery, New York, NY, USA, 254–263. doi:10.1145/3731120.3744591

[6] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proc. 58th Annu. Meet. Assoc. Comput. Linguist.*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). ACL, 6282–6293. doi:10.18653/v1/2020.acl-main.560

[7] Yangning Li, Weizhi Zhang, Yuyao Yang, Wei-Chieh Huang, Yaozu Wu, Junyu Luo, Yuanchen Bei, Henry Peng Zou, Xiao Luo, Yusheng Zhao, Chunkit Chan, Yankai Chen, Zhongfen Deng, Yinghui Li, Hai-Tao Zheng, Dongyuan Li, Renhe Jiang, Ming Zhang, Yangqiu Song, and Philip S. Yu. 2025. A Survey of RAG-Reasoning Systems in Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 12120–12145. doi:10.18653/v1/2025.findings-emnlp.648

[8] Tianneng Shi, Kaijie Zhu, Zhun Wang, Yuqi Jia, Will Cai, Weida Liang, Haonan Wang, Hend Alzahrani, Joshua Lu, Kenji Kawaguchi, Basel Alomair, Xuandong Zhao, William Yang Wang, Neil Gong, Wenbo Guo, and Dawn Song. 2025. PromptArmor: Simple yet Effective Prompt Injection Defenses. arXiv:2507.15219 [cs] doi:10.48550/arXiv.2507.15219

[9] Paul Thomas, Douglas W. Oard, Eugene Yang, Dawn Lawrie, and James Mayfield. 2025. System Comparison Using Automated Generation of Relevance Judgements in Multiple Languages. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Padua, Italy) *(SIGIR '25)*. Association for Computing Machinery, New York, NY, USA, 2812–2816. doi:10.1145/3726302.3730252

[10] Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2024. Large Language Models can Accurately Predict Searcher Preferences. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 1930–1940. doi:10.1145/3626772.3657707

[11] Shivani Upadhyay, Ronak Pradeep, Nandan Thakur, Daniel Campos, Nick Craswell, Ian Soboroff, and Jimmy Lin. 2025. A Large-Scale Study of Relevance Assessments with Large Language Models Using UMBRELA. In *Proc. 2025 Int. ACM SIGIR Conf. Innov. Concepts Theor. Inf. Retr. ICTIR*. ACM, Padua Italy, 358–368. doi:10.1145/3731120.3744605