Perspectives on Cascading Pipelines for Sensitivity-Aware Search

Jack McKechnie
University of Glasgow
Glasgow, Scotland
j.mckechnie.1@research.gla.ac.uk

Abstract

Large document collections, such as email archives and meeting minutes, are produced by governments, institutions, and companies through their day-to-day activities. Such document collections contain information that would be useful to stakeholders across different sectors, were they to be made publicly available. Indeed, in the case of government document collections, under Open Government models, citizens must be able to access documents produced by their government in a timely manner. However, such document collections can contain *sensitive* information such as matters of national security, which prevent the collections from being made available to the public. *Sensitivity-Aware Search* (SAS) proposes a solution to making document collections potentially containing sensitive information publicly accessible by enabling the entire collection to be searched whilst protecting sensitive information from being exposed. In this paper, we argue that SAS should be addressed with a cascading retrieval pipeline, where documents are ranked in a staged manner by a sequence of different models. As such, results at each stage can be inspected, and each model can specialise in different aspects of retrieval to work in tandem, considering sensitivity at each stage. We present three arguments and provide perspectives from both a system and governance perspective for each argument. Specifically, we argue that the separation of concerns, efficiency, and inspectability that cascading pipelines offer make them particularly useful for deployment in open government scenarios. Further, we provide experimental evidence to support our thesis that SAS should be tackled as a cascading pipeline.

CCS Concepts

• **Information systems** → **Information retrieval**.

Keywords

Sensitivity-Aware Search, Open Government

ACM Reference Format:
Jack McKechnie. 2026. Perspectives on Cascading Pipelines for Sensitivity-Aware Search. In *Proceedings of The 1st AI & Open Government (AIOG) Workshop @ ICAIL 2026 (AI & OG @ ICAIL '26).* ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

AI & OG @ ICAIL '26, Singapore

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Governments, institutions, and companies produce large document collections as part of their operations. Such document collections contain useful information that would benefit multiple sectors if it were made more widely publicly available, e.g., by enabling cross-departmental collaboration and data sharing. Many governments operate under an *Open Government* model [4, 48], where citizens should be able to access the information that is produced by their government. For example, more than 64 governments are members of the Open Government Partnership [4], illustrating their commitment to making information accessible. However, interspersed with the useful information that is contained in government document collections, there is also *sensitive information*, such as matters of national security and personal information. Hence, such large-scale document collections typically cannot be made available to the public without expensive and time-consuming manual sensitivity review by experts. For example, during the 2014 US Presidential campaign, Hilary Clinton released ∼30k emails from her time working as a government official. The review process took a team of 25 people working full-time for nearly one year to complete [31].

Sensitivity-aware search (SAS) [19, 31, 32, 36, 43, 44] provides a complementary approach to manual sensitivity review for making document collections accessible to the public. In SAS, the entire document collection is indexed (both sensitive and non-sensitive documents), and a series of sensitivity-aware retrieval models is deployed to provide a set of ranked results. Typically, results are presented to an end user who wants to search the collection, but results can be passed to a sensitivity review expert for final validation, or to an agentic system to be used to answer questions, for example. Such retrieval models highly rank documents that are relevant to the query, but are not sensitive. Hence, sensitivity-aware search provides a promising avenue for quickly and inexpensively making government documents publicly available, in line with Freedom of Information legislation [9, 39] and the Treaty on the Functioning of the European Union (Article 15) [45]. Indeed, the Scottish government has identified that making government information available online is insufficient, and does not align with their Open Government Action Plan [1]. Users in Scotland must instead be able to filter and find noteworthy data, i.e., by using search engines.

Traditional search engines are often operationalised as *cascading ranking pipelines* [14, 47]. Given a collection of documents, C, a lightweight first-stage retriever is typically deployed to construct a set of $k \ll |C|$ candidate documents that is ordered by their estimated relevance to a query. Subsequently, a more computationally expensive and more effective reranker is applied on the set of candidate documents to improve the quality of the final ranked list presented to the user. In this position paper, we argue that each model that is deployed within a cascading ranking pipeline should be sensitivity-aware. Models that are placed in earlier stages of

the pipeline should aim to maximise recall, i.e. include as many relevant documents in the top k results as possible, whilst also filtering out sensitive document. Models in the later stages of the pipeline should aim to maximise precision at the top of the ranked list whilst demoting high-risk documents to low rank positions.

Cascading ranking pipelines are well studied in traditional search tasks [14, 47] and are deployed in industry for example in web and product search [11]. Cascading ranking pipelines offer advantages over single-stage retrieval: **(1) Efficiency:** A small number of candidate documents that have been identified to be potentially relevant by a first-stage retriever can be passed to expensive rerankers, minimising unnecessary scoring of non-relevant documents. Efficiency is important when deploying cascading ranking pipelines for SAS as highly-skilled reviewers can be part of the process, and their time is expensive; **(2) Specialisation:** models at different stages of the pipeline can perform different functions that suit their position. Recall of non-sensitive documents can be prioritised early in the pipeline and relevance of the top documents later; **(3) Resilience:** in a single-stage ranking pipeline, the order in which documents are ranked by the first stage retriever is final; there are no further stages that can be deployed to refine the ranking. However, with a cascading ranking pipeline, ranking errors that are made by the first-stage retriever can be corrected at later stages.

In this paper, we posit that SAS should be addressed with a cascading ranking pipeline designed for the task. We make three arguments in favour of cascading ranking pipelines for SAS and present our arguments from two distinct perspectives. Our first perspective is a governance perspective, where we consider the impact of cascading ranking pipelines on legislation, compliance, and accountability. Our second is a system perspective, where we consider technical and implementation details to make our argument. We present experimental evidence in favour of cascading retrieval pipelines. Using the OHSUMED dataset, we deploy cascading ranking pipelines with a sensitivity-aware learned sparse retrieval [13] model as a first-stage retriever, a T5-based [41] classifier filter stage, and a cross-encoder reranker trained with sensitivity-aware negative sampling [32]. We show that sensitivity-aware cascading retrieval pipelines outperform pipelines where sensitivity is either not considered at all or is considered at only certain stages, with up to a 64% decrease in the retrieval of sensitive documents and 31% improvements in sensitivity vs. retrieval effectiveness metrics.

In the following, Section 2 presents related literature; Section 3 argues that the propagation of errors throughout the pipeline can be mitigated by considering SAS at each stage; Section 4 outlines some advantages of cascading pipelines for SAS; Section 5 discusses efficiency advantages; Section 6 presents experimental evidence. Section 7 provides concluding remarks.

2 Related Work

This section discusses relevant literature on cascading retrieval pipelines, SAS, and information retrieval for Open Government.

Cascading Retrieval Pipelines. In traditional search tasks, it is typically impractical to deploy the most effective ranking models to score each document in the corpus with respect to each incoming query, as such models are computationally expensive and corpora are large. Cascading retrieval pipelines offer a practical solution to this problem by applying increasingly expensive models to a decreasing number of documents in a staged manner. Inexpensive first-stage retrieval models are applied over the full document collection to retrieve a set of candidate documents. The candidate subset of documents is subsequently passed to one or more stages of reranking by a more effective, but more inefficient, reranker. As such, a set of ranked search results that are relevant to a user-issued query is efficiently produced [14, 47]. Typically, in traditional search tasks, the early stages of cascading retrieval pipelines focus on attaining high recall [47]. That is, relevant documents should be included in the candidate set, no matter where they are in the ranking [28]. Later stages subsequently focus on the precision of the results, i.e., ensuring that the relevant documents contained in the candidate set are pushed up to top ranks [16]. However, cascading retrieval pipelines remain under-explored in the SAS literature, with previous works focusing on reranking for the task. Consequently, in this paper, we argue for the use of cascading ranking pipelines in the SAS task.

Sensitivity-Aware Search. Traditionally, government document collections have undergone manual human review to identify and remove sensitive information when being made publicly available, in compliance with Freedom of Information Act (FOIA) legislation. Gollins et al. [15] outlined the challenges of reviewing collections of born-digital documents and identified that such issues could potentially be alleviated with automatic sensitivity classifiers. Subsequently, a body of literature emerged on automatic sensitivity classification. For example, McDonald et al. [29] deployed SVM [10] classifiers with specific sensitivity features, Branting et al. [8] trained a BERT [12] model as a classifier to reduce sensitivity review time, and Baron et al. [5] prompted Large Language Models to classify FOIA sensitivities. Most recently, Larooij [22] proposed that a sensitivity-aware intent-clarification agent could be deployed to mediate user interactions with collections potentially containing sensitive information.

However, comparatively few works have investigated safely searching collections that potentially contain sensitive information. Oard et al. [36] proposed the task of sensitivity-aware search and outlined how the task could be applied in large archival and government collections. Further, Oard et al. [36] introduced an evaluation measure for SAS based on the nDCG [20] metric. Their proposed metric, named nCSDCG, rewards the system for retrieving relevant documents whilst also penalising the system for returning sensitive documents. We use nCSDCG in Section 6 to measure the tradeoff between relevance and sensitivity in our experiment with cascading retrieval pipelines for SAS. Iqbal et al. [19] and Sayed et al. [43] investigated how test collections can be designed to consider different perspectives on what is considered sensitive. Subsequently, Sayed et al. [44] proposed a learning-to-rank model that leveraged sensitivity features and optimised for the nCSDCG metric. However, there have been considerable improvements in ad-hoc search over learning-to-rank approaches by deploying neural Information Retrieval approaches. Recent literature has proposed using such models for the task of SAS [30]. Mckechnie et al. [32] defined sensitivity-aware negative sampling approaches that have been used to train cross-encoder rerankers for SAS. We leverage sensitivity-aware training approaches in our experiments shown in Section 6. However, SAS literature has so far focused on reranking retrieval results from a relevance-only first-stage retriever. Hence,

in this paper, we discuss the need to investigate sensitivity at each stage of a cascading retrieval pipeline.

The Intersection of Information Retrieval and Open Government. Open Government principles, such as open data, transparency, and accountability [48], can be implemented by deploying Information Retrieval techniques [6]. As such, several sub-domains of IR have emerged at the intersection of Open Government and IR research. For example, e-discovery [35, 37] aims to keep companies and governments accountable during legal trials by effectively identifying and retrieving all information from digital documents that is pertinent to ongoing litigation. Technology-Assisted Review has also seen much research from IR practitioners, with the task aiming to prioritise documents for human assessment in order to reduce human review effort while maintaining high recall. Further, High-Recall Retrieval studies retrieval scenarios in which failing to retrieve relevant documents carries a significant risk, such as in legal proceedings. More broadly, Legal Information Retrieval addresses search tasks using legal corpora [7, 46] and emphasises the importance of transparency and explainability. Such areas are adjacent to sensitivity-aware search but show how IR practices can be successfully deployed in Open Government scenarios to improve access, transparency, and accountability. In the next section, we give our first argument for cascading ranking pipelines in sensitivity-aware search.

3 Error Propagation in Cascading Pipelines

In this section, we argue for the use of cascading ranking pipelines for sensitivity-aware search by investigating how errors are propagated and exacerbated during ranking when sensitivity is not considered at each stage. We first discuss error propagation from a governance perspective, and secondly from a systems perspective.

Propagation of errors from a governance perspective. The European Union's Digital Services Act (DSA) mandates that platforms that deploy algorithmic systems must be able to audit how those systems operate and, critically, where failures occur. Specifically, the legislation states that the providers of such systems should be able to "diligently identify, analyse and assess any systemic risks in the Union stemming from the design or functioning of their service and its related systems, including algorithmic systems". It is typically difficult to identify exactly *why* a state-of-the-art dense retrieval model ranks documents as it does [3]. For example, the dense retrieval models that often achieve the best performance on retrieval benchmarks are not explainable, as queries and documents are encoded as high-dimensional vectors that cannot be reasoned about. Consequently, when a sensitive document is highly ranked and exposed to the user, single-stage retrieval systems offer no way of determining where, how, or why this failure happened. Cascading retrieval pipelines, on the other hand, offer natural audit points. When a sensitive document is erroneously returned to a user, the staged structure of the pipeline allows us to identify where the failure occurred by tracing the rank position of the document throughout the pipeline. Auditing at each stage of the pipeline can be done manually by human reviewers, or with the assistance of a separate explanation model that has been developed for the task. By considering sensitivity during retrieval at every stage of the pipeline, system designers can not only reduce the likelihood of sensitive documents being exposed to users but also ensure that when failures do occur, there is a clear and auditable record of how the system behaved, as is increasingly required by legislation such as the DSA.

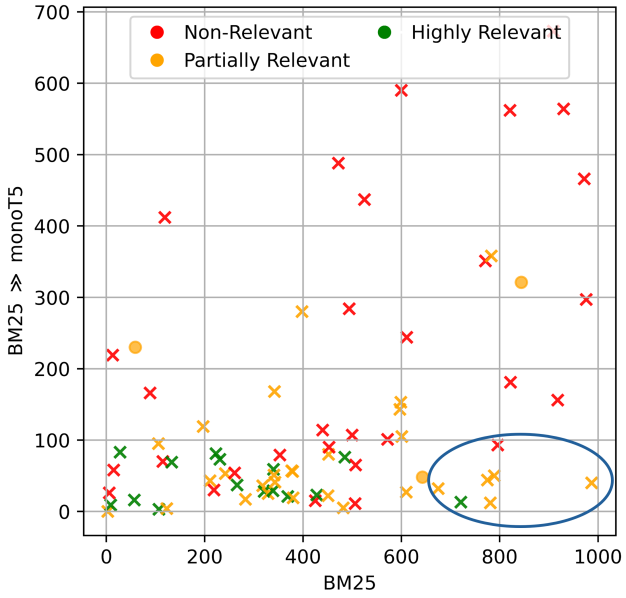

Figure 1: Risk of misclassification when sensitivity is not considered at each stage of the ranking pipeline. Sensitive documents are marked with × and non-sensitive documents are marked with •. Highlighted with a blue circle are misclassified sensitive documents that are highly ranked during reranking, exposing sensitive information.

Propagation of errors from a systems perspective. One simple approach for developing a sensitivity-aware search pipeline is to deploy a trained sensitivity classifier to filter out predicted sensitive documents from the ranking, before passing on the filtered results to a relevance-only reranker. However, we argue that not considering sensitivity at all stages of the pipeline can lead to relevant but sensitive documents being exposed to the user. We make our argument using an example, given in Figure 1. We use a commonly deployed ranking pipeline, where BM25 [42] is used as a first-stage retriever and the top 1000 documents retrieved by BM25 are reranked by the monoT5 [34] reranker. We show retrieval results for an example query from the OHSUMED [17] collection. Each point in Figure 1 is a retrieved document, with the position on the x-axis indicating the rank of that document during the BM25 first-stage retrieval and the position on the y-axis indicating the rank of that document after being reranked with monoT5. Therefore, documents in the bottom right-hand side have been upranked during reranking, and documents in the top-left-hand side have been downranked. Sensitive documents are marked with ×, and non-sensitive documents are marked with •. We deploy a T5-based classifier (F1: 0.7997, Precision: 0.8229, Recall: 0.7777, and BAC: 0.877) on the retrieved documents to illustrate how misclassification errors can be propagated when sensitivity is not considered at each stage of a cascading ranking pipeline. Figure 1 shows documents that are misclassified by the classifier. We can observe in Figure 1 that the documents that are highlighted in the blue circle are ranked low in

the ranking by the BM25 first-stage retrieval, but are brought up to the top of the ranking by the relevance-only monoT5 reranker. As these sensitive documents are misclassified as non-sensitive, the relevance-only reranker exposes sensitive information to the user. However, sensitivity-aware rerankers provide an opportunity for such misclassification errors to be mitigated and recovered from, as sensitive documents are downranked in the later stages of the pipeline. Consequently, Sensitivity should be considered at each stage of the ranking pipeline.

4 Separation of Concerns in Cascading Pipelines

In this section, we argue for the use of cascading ranking pipelines for sensitivity-aware search by outlining how different components of the system can specialise in different aspects of the task. We first make this argument from a governance perspective and secondly from a systems perspective.

Separation of concerns from a governance perspective The General Data Protection Regulation (GDPR) legislation of the European Union requires that personal data not be processed further than is required. The separation of concerns that cascading ranking pipelines for sensitivity-aware search offer allows sensitive personal information not to be unnecessarily processed. First-stage sensitivity-aware retrieval aims to ensure that only relevant non-sensitive documents are retrieved. As such, sensitive documents that contain personal information are not subject to processing by later stages of reranking as they have been filtered out of the ranking by prior stages. Cascading ranking pipelines help SAS systems comply with GDPR regulations, as is required by each EU country. Further, the separation of concerns that cascading ranking pipelines offer allows for quick adaptation of sensitivity-aware retrieval systems to changing legislation. Each model in the pipeline can focus on one aspect that makes information sensitive, or on identifying one specific legal definition of sensitivity that is likely to be present in the collection. Situations change over time; what information is sensitive under the UK Freedom of Information Act [39] changes depending on what is publicly available, for example. As public information and legislation change, models in a SAS pipeline can be swapped in and out to handle this. As such, the whole system does not need to be retrained to factor in one small change. In this way, the separation of concerns that is inherent to cascading pipelines saves time, research effort, and model training.

Separation of concerns from a system perspective. In traditional search tasks, where the relevance of retrieval results is the only concern, each stage of a cascading ranking pipeline is designed to focus on one goal. Typically, first-stage retrievers focus on achieving high recall; as many relevant documents as possible should be included in the top k retrieval results, no matter their rank position. Later stages in the pipeline subsequently focus on precision, ensuring that the relevant documents are brought to the top of the ranked list. In the sensitivity-aware search task, we can see that first-stage retrievers should not focus on retrieving as many relevant documents as possible, but on retrieving as many relevant non-sensitive documents as possible. In this way, later stages of the pipeline can focus on ensuring that such relevant non-sensitive documents are ranked highly. Further, as each sensitivity is highly specific to the use case, errors can happen, which result in sensitive information being processed by later pipeline stages unnecessarily. Deploying multiple models in sequence allows errors in early stages to be recovered from in later stages to limit unnecessary processing of sensitive information later in the pipeline. Having multiple ranking models that are deployed in sequence allows each model to focus on what is necessary at that stage of the pipeline, and consequently return results to the user that are both relevant to their query and do not contain sensitive information.

5 Efficiency of Cascading Pipelines

In this section, we argue for the use of cascading ranking pipelines for sensitivity-aware search by considering efficiency. We first discuss efficiency from a governance perspective, and secondly, we discuss efficiency from a system perspective.

Efficiency from a governance perspective. Recital 83 of the United Kingdom General Data Protection Regulation [2] (UK GDPR) asserts that the cost of implementation of systems that handle private information should be considered by data controllers and processors. Typically, the most effective ranking models are LLM-based rerankers [24, 40], which require many calls to external LLM APIs or a large amount of local computation on GPU. Consequently, such effective models are prohibitively expensive to deploy in real-world scenarios. The use of a cascading ranking pipeline results in fewer documents needing to be scored by expensive models, as only documents that are likely to be relevant to the user are included in the candidate set. Cascading ranking allows for LLM-based SAS models to be practically deployed in real open government scenarios.

Further, in the SAS task, different queries require different levels of attention. SAS retrieval models are unlikely to return sensitive documents when queries on innocuous topics or information that is publicly available are issued to the model. Consequently, when such queries are processed by the retrieval system, it is not necessary to deploy the most effective SAS model available; a simpler and more efficient model would suffice. Cascading ranking pipelines allow for flexibility when processing different queries, as approaches such as *early exit* strategies can be deployed, where documents are not passed through the entire pipeline; rather, the ranking process is terminated once the results are satisfactory. Consequently, government compute budgets are saved by not passing results through unnecessary stages of the pipeline. Additionally, given a graded access scenario, where different types of users have different levels of access and trust, early exit strategies can be designed to trade off access level with retrieval latency. Models can then be reused by deploying the appropriate exit strategy for the level of access of the user, saving research and development budget. Efficient early exit strategies that cascading ranking pipelines allow for enable SAS to be compliant with GDPR and to save governments time and money.

Efficiency from a system perspective. Information Retrieval (IR) systems typically demonstrate an effectiveness versus efficiency tradeoff. That is, effective IR systems leverage complex ranking functions and are hence less computationally efficient than simpler systems. Document collections are large, and users can issue many queries to the IR system. As such, scoring every document for every query using the most effective (but inefficient) retrieval model is not feasible. High latency frustrates users and leads to a negative perception of the IR system [33, 49]. Cascading retrieval pipelines allow the system designer to strike a balance between the effectiveness and the efficiency of the system. An efficient first-stage retrieval

Table 1: Experimental results showing SAS leveraging different first-stage retrievers. Best performance for each category of pipeline; without reranking (w/o Rerank), reranking with a relevance-only reranker (Rel. Rerank), and reranking with a SAS reranker (SAS Rerank) are shown in bold. Significant improvements with the SPLADE RELEVANCE and SPLADE SANS pipelines in each category are denoted with * and †, respectively (paired t-test, $p < 0.05$, Bonferroni Correction).

	Approach	nDCG@10	sens_docs@10	nCSDCG@10	nDCG@20	sens_docs@20	nCSDCG@20
w/o Rerank	BM25	0.3917	1.3396	0.7077	0.3760	2.6132	0.6259
	SPLADE RELEVANCE	0.4375	1.2170	0.7276	0.4192	2.5094	0.6354
	SPLADE SANS	**0.4557**	0.8302*	0.7848*	**0.4374**	1.8302*	0.7509*
	SPLADE DISTIL	0.2841	**0.2736***†	**0.7983***	0.2771	**0.5189***†	**0.8263***†
Rel. Rerank	BM25 ≫ MONOT5 RELEVANCE	0.4808	1.4245	0.7146	0.4565	2.7547	0.6099
	SPLADE RELEVANCE ≫ MONOT5 RELEVANCE	0.4904	1.3396	0.7385	0.4649	2.6509	0.6380
	SPLADE SANS ≫ MONOT5 RELEVANCE	**0.4936**	1.1981	0.7440	**0.4701**	2.3868	0.6913
	SPLADE DISTIL ≫ MONOT5 RELEVANCE	0.4563	**0.6415***†	**0.8047***†	0.4165*†	**1.0755***†	**0.8181***†
SAS Rerank	BM25 ≫ MONOT5 SANS	0.3766	0.7925	0.7684	0.3865	1.6132	0.7339
	SPLADE RELEVANCE ≫ MONOT5 SANS	**0.4273**	0.7547	0.7919	0.4234	1.6415	0.7507
	SPLADE SANS ≫ MONOT5 SANS	0.4205	0.7547	0.7858	**0.4219**	1.5943	0.7748
	SPLADE DISTIL ≫ MONOT5 SANS	0.3982	**0.4340***†	**0.8127**†	0.3738*†	**0.7925***†	**0.8312***†

model, such as BM25 [42] or SPLADE [13], is deployed to select a set of documents that is likely to include relevant documents. As such, only a small subset of documents are passed to the expensive reranker, saving time and resources by not scoring documents that are unlikely to be relevant, as they are not included in the candidate set. Cascading ranking pipelines allow effective retrieval models to be efficiently deployed for effective search results.

6 Experimental Evidence

In this section, we provide experimental evidence that shows how cascading ranking pipelines that consider sensitivity at all stages of the pipeline are effective for sensitivity-aware search. We first describe the experimental setup used for this experiment, and secondly, we discuss the results of this experiment.

Experimental Setup We experiment with the OHSUMED [17] dataset, which comprises 14,430 documents and 106 queries. Following prior literature [32, 44], we take documents that are associated with Medical Subject Headings [23] C12 (Male Urogenital Diseases) and C13 (Female Urogenital Diseases and Pregnancy Complications) as our sensitive documents (~12% of the collection). We use the PyTerrier [26] Information Retrieval platform to perform our experiments, using a PISA [27] index with the PyTerrier PISA [25]. We use the nDCG@k [20] metric to measure the relevance of our results. To measure the propensity of sensitive documents in the search results, we use the sens_docs@k measure, i.e., the average number of sensitive documents in the top k results across all queries. Finally, to measure the relevance versus sensitivity tradeoff, we use the nCSDCG@k [36] measure, which is an nDCG-based measure that rewards the system for returning relevant documents and penalises the system for returning sensitive documents.

We experiment with cascading retrieval pipelines in one of two forms; (1) R_1 % 100; (2) R_1 % 100 ≫ R_2, where % k indicates truncation at rank cutoff k and $a \gg b$ denotes passing the results from stage a to stage b. R_1 is one of BM25 (a traditional, relevance-only, statistical retriever), SPLADE RELEVANCE (a learned sparse retrieval model trained for relevance only), SPLADE SANS (a learned sparse retrieval model trained for SAS using sensitivity-aware negative sampling (SANS) [32] using CrossEntropy loss and the FLOPs regulariser [38]), or SPLADE DISTIL (an LSR model trained for SAS using knowledge distillation with the KL divergence loss [18, 21]). R_2 is a monoT5 [34] reranker trained either for relevance only (MONOT5 RELEVANCE) or for trained for SAS using SANS (MONOT5 SANS).

Discussion Table 1 shows the results of our experiments with cascading retrieval pipelines for SAS. Each section shows a different type of pipeline: with no reranking, with reranking for relevance only, and with reranking for SAS. We can observe that considering sensitivity during first-stage retrieval is important, as both SPLADE SANS and SPLADE DISTIL outperform relevance-only baselines in terms of the relevance versus sensitivity tradeoff (0.7077 nCSDCG@10 with BM25 → 0.7848 with SPLADE SANS and 0.7983 with SPLADE DISTIL). Reranking using a relevance-only reranker improves the relevance of the search results (e.g. 0.4557 nDCG@10 for SPLADE SANS → 0.4936 for SPLADE DISTIL. However, relevance-only reranking hurts sensitivity performance (e.g. 0.8302 sens_docs@10 → 1.1981). This degradation is to be expected, as some queries specifically ask for sensitive information, and if sensitivity is not considered in the reranking stage, then any relevant sensitive documents should be upranked during reranking. On the other hand, if sensitivity is considered during reranking and first stage retrieval, we see the best performance in terms of the relevance vs. sensitivity tradeoff (e.g. 0.7983 w/o reranking → 0.8047 w/ relevance reranking → 0.8127 w/ SAS reranking). By deploying a cascading retrieval pipeline that considers sensitivity at all stages, we are able to provide the user with results that are both relevant to their query and that do not contain sensitive information.

7 Conclusions

This work has argued for the use of cascading ranking pipelines for sensitivity-aware search. We have made three arguments: separation of concerns, efficiency of cascading pipelines, and error propagation. Further, we have made each of our arguments from two standpoints, one from a system perspective and one from a governance perspective. We have provided experimental evidence in favour of cascading ranking pipelines for sensitivity-aware search, with improvements of ~12% over relevance-only baseline models. Sensitivity-aware search shows promise in helping governments uphold open government principles by enabling users to *safely* search collections that may contain sensitive information, thereby enhancing transparency and accountability.

References

[1] 2018. Scotland's Open Government Action Plan 2018–2020. https://assets.publishing.service.gov.uk/media/605f24ccd3bf7f7188ffaea6/Scottish_NAP_2018-20_V2.pdf Accessed section at p. 14.

[2] 2018. UK General Data Protection Regulation (UK GDPR). https://www.legislation.gov.uk/eur/2016/679/contents.

[3] Avishek Anand, Lijun Lyu, Maximilian Idahl, Yumeng Wang, Jonas Wallat, and Zijian Zhang. 2022. Explainable information retrieval: A survey. *arXiv preprint arXiv:2211.02405* (2022).

[4] Judie Attard, Fabrizio Orlandi, Simon Scerri, and Sören Auer. 2015. A systematic review of open government data initiatives. *Gov. Inf. Q.* 32, 4 (2015).

[5] Jason R Baron, Nathaniel W Rollings, and Douglas W Oard. 2023. Using ChatGPT for the FOIA Exemption 5 Deliberative Process Privilege.. In *LegalAIIA @ ICAIL*.

[6] Floris Bos, Marc van Opijnen, and Maarten Marx. 2025. Linking References to Documents in Parliamentary Debates. In *TPDL*.

[7] Floris Bos, Marc van Opijnen, and Maarten Marx. 2025. Linking References to Documents in Parliamentary Debates. In *TPDL*.

[8] Karl Branting, Bradford Brown, Chris Giannella, James Van Guilder, Jeff Harrold, Sarah Howell, and Jason R Baron. 2025. Decision support for detecting sensitive text in government records. *AI & Law* 33, 1 (2025).

[9] United States Congress. 1966. Freedom of Information Act. https://www.foia.gov/foia-statute.html

[10] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *ML* 20, 3 (1995).

[11] Sunhao Dai, Jiakai Tang, Jiahua Wu, Kun Wang, Yuxuan Zhu, Bingjun Chen, Bangyang Hong, Yu Zhao, Cong Fu, Kangle Wu, et al. 2025. Onepiece: Bringing context engineering and reasoning to industrial cascade ranking system. *arXiv preprint arXiv:2509.18091* (2025).

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*.

[13] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade: Sparse lexical and expansion model for first stage ranking. In *Proc. of SIGIR*.

[14] Luke Gallagher, Ruey-Cheng Chen, Roi Blanco, and J Shane Culpepper. 2019. Joint optimization of cascade ranking models. In *Proc. of WSDM*.

[15] Timothy Gollins, Graham McDonald, Craig Macdonald, and Iadh Ounis. 2014. On Using Information Retrieval for the Selection and Sensitivity Review of Digital Public Records.. In *PIR @ SIGIR*.

[16] Donna Harman. 1992. Evaluation Issues in Information Retrieval. *Inf. Process. Manag.* 28, 4 (1992).

[17] William Hersh, Chris Buckley, TJ Leone, and David Hickam. 1994. OHSUMED: An interactive retrieval evaluation and new large test collection for research. In *Proc. of SIGIR*.

[18] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

[19] Modassir Iqbal, Katie Shilton, Mahmoud F Sayed, Douglas Oard, Jonah Lynn Rivera, and William Cox. 2021. Search with discretion: value sensitive design of training data for information retrieval. *PACMHCI* 5, CSCW1 (2021).

[20] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *TOIS* 20, 4 (2002).

[21] Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *Ann. Math. Stat.* 22, 1 (1951).

[22] Maik Larooij. 2026. Sensitivity-Aware Retrieval-Augmented Intent Clarification. In *CoSCIN @ ECIR2026*.

[23] Carolyn E Lipscomb. 2000. Medical subject headings (MeSH). *Bull. Med. Libr. Assoc.* 88, 3 (2000).

[24] Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2024. Fine-tuning llama for multi-stage text retrieval. In *Proc. of SIGIR*.

[25] Sean MacAvaney and Craig Macdonald. 2022. A Python Interface to PISA!. In *Proc. of SIGIR*.

[26] Craig Macdonald and Nicola Tonellotto. 2020. Declarative experimentation in information retrieval using PyTerrier. In *Proc. of ICTIR*.

[27] Antonio Mallia, Michal Siedlaczek, Joel Mackenzie, and Torsten Suel. 2019. PISA: Performant indexes and search for academia. *Proc. of the Open-Source IR Replicability Challenge* (2019).

[28] Christopher D Manning. 2008. *Introduction to information retrieval*. Syngress Publishing,.

[29] Graham McDonald, Craig Macdonald, Iadh Ounis, and Timothy Gollins. 2014. Towards a classifier for digital sensitivity review. In *Proc. of ECIR*.

[30] Jack McKechnie. 2024. Cascading Ranking Pipelines for Sensitivity-Aware Search. In *Proc. of ECIR*.

[31] Jack McKechnie and Graham McDonald. 2024. SARA: A Collection of Sensitivity-Aware Relevance Assessments. *arXiv preprint arXiv:2401.05144* (2024).

[32] Jack McKechnie, Graham McDonald, and Craig Macdonald. 2024. Bi-Objective Negative Sampling for Sensitivity-Aware Search. In *Proc. of SIGIR*.

[33] Fiona Fui-Hoon Nah. 2004. A study on tolerable waiting time: how long are web users willing to wait? *BIT* 23, 3 (2004).

[34] Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. In *EMNLP*.

[35] Douglas W Oard, Jason R Baron, Bruce Hedin, David D Lewis, and Stephen Tomlinson. 2010. Evaluation of information retrieval for E-discovery. *AI & Law* 18, 4 (2010).

[36] Douglas W Oard, Katie Shilton, and Jimmy Lin. 2016. Evaluating Search Among Secrets.. In *EVIA @ NTCIR*.

[37] Douglas W Oard and William Webber. 2013. Information retrieval for e-discovery. *FnTIR* 7, 2–3 (2013).

[38] Biswajit Paria, Chih-Kuan Yeh, Ian EH Yen, Ning Xu, Pradeep Ravikumar, and Barnabás Póczos. 2020. Minimizing flops to learn efficient sparse representations. *arXiv preprint arXiv:2004.05665* (2020).

[39] UK Parliament. 2000. Freedom of Information Act 2000. https://www.legislation.gov.uk/ukpga/2000/36/contents

[40] Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023. Rankvicuna: Zero-shot listwise document reranking with open-source large language models. *arXiv preprint arXiv:2309.15088* (2023).

[41] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR* 21, 140 (2020).

[42] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proc. of TREC*.

[43] Mahmoud F Sayed, William Cox, Jonah Lynn Rivera, Caitlin Christian-Lamb, Modassir Iqbal, Douglas W Oard, and Katie Shilton. 2020. A test collection for relevance and sensitivity. In *Proc. of SIGIR*.

[44] Mahmoud F Sayed and Douglas W Oard. 2019. Jointly modeling relevance and sensitivity for search among sensitive content. In *Proc. of SIGIR*.

[45] European Union. 2012. Consolidated version of the Treaty on the Functioning of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012E/TXT

[46] Ruben van Heusden, Maik Larooij, Jaap Kamps, and Maarten Marx. 2025. A collection of FAIR Dutch Freedom of Information Act documents. *Sci. Data* 12, 1 (2025).

[47] Lidan Wang, Jimmy Lin, and Donald Metzler. 2011. A cascade ranking model for efficient ranked retrieval. In *Proc. of SIGIR*.

[48] Bernd W Wirtz and Steven Birkmeyer. 2015. Open government: Origin, development, and conceptual perspectives. *IJPA* 38, 5 (2015).

[49] Iris Xie and Colleen Cool. 2009. Understanding help seeking within the context of searching digital libraries. *J. Assoc. Inf. Sci. Technol.* 60, 3 (2009).