Dense retrieval methods have shown great promise over sparse retrieval methods in a range of NLP problems. Among them, dense phrase retrieval-the most fine-grained retrieval unit-is appealing because phrases can be directly used as the output for question answering and slot filling tasks. In this work, we follow the intuition that retrieving phrases naturally entails retrieving larger text blocks and study whether phrase retrieval can serve as the basis for coarse-level retrieval including passages and documents. We first observe that a dense phrase-retrieval system, without any retraining, already achieves better passage retrieval accuracy (+3-5% in top-5 accuracy) compared to passage retrievers, which also helps achieve superior end-to-end QA performance with fewer passages. Then, we provide an interpretation for why phrase-level supervision helps learn better fine-grained entailment compared to passage-level supervision, and also show that phrase retrieval can be improved to achieve competitive performance in document-retrieval tasks such as entity linking and knowledge-grounded dialogue. Finally, we demonstrate how phrase filtering and vector quantization can reduce the size of our index by 4-10x, making dense phrase retrieval a practical and versatile solution in multi-granularity retrieval.Hybrid retrievers can take advantage of both sparse and dense retrievers. Previous hybrid retrievers leverage indexing-heavy dense retrievers. In this work, we study "Is it possible to reduce the indexing memory of hybrid retrievers without sacrificing performance"? Driven by this question, we leverage an indexing-efficient dense retriever (i.e. DrBoost) and introduce a LITE retriever that further reduces the memory of DrBoost. LITE is jointly trained on contrastive learning and knowledge distillation from DrBoost. Then, we integrate BM25, a sparse retriever, with either LITE or DrBoost to form light hybrid retrievers. Our Hybrid-LITE retriever saves 13X memory while maintaining 98.0% performance of the hybrid retriever of BM25 and DPR. In addition, we study the generalization capacity of our light hybrid retrievers on out-of-domain dataset and a set of adversarial attacks datasets. Experiments showcase that light hybrid retrievers achieve better generalization performance than individual sparse and dense retrievers. Nevertheless, our analysis shows that there is a large room to improve the robustness of retrievers, suggesting a new research direction.With the rapid development of large-scale language models, Retrieval-Augmented Generation (RAG) has been widely adopted. However, existing RAG

paradigms are inevitably influenced by erroneous retrieval information, thereby reducing the reliability and correctness of generated results. Therefore, to improve the relevance of retrieval information, this study proposes a method that replaces traditional retrievers with GPT-3.5, leveraging its vast corpus knowledge to generate retrieval information. We also propose a web retrieval based method to implement fine-grained knowledge retrieval, Utilizing the powerful reasoning capability of GPT-3.5 to realize semantic partitioning of problem.In order to mitigate the illusion of GPT retrieval and reduce noise in Web retrieval,we proposes a multi-source retrieval framework, named MSRAG, which combines GPT retrieval with web retrieval. Experiments on multiple knowledge-intensive QA datasets demonstrate that the proposed framework in this study performs better than existing RAG framework in enhancing the overall efficiency and accuracy of QA systems.Despite the recent progress in long-context language models, it remains elusive how transformer-based models exhibit the capability to retrieve relevant information from arbitrary locations within the long context. This paper aims to address this question. Our systematic investigation across a wide spectrum of models reveals that a special type of attention heads are largely responsible for retrieving information, which we dub retrieval heads. We identify intriguing properties of retrieval heads:(1) universal: all the explored models with long-context capability have a set of retrieval heads; (2) sparse: only a small portion (less than 5\%) of the attention heads are retrieval. (3) intrinsic: retrieval heads already exist in models pretrained with short context. When extending the context length by continual pretraining, it is still the same set of heads that perform information retrieval. (4) dynamically activated: take Llama-2 7B for example, 12 retrieval heads always attend to the required information no matter how the context is changed. The rest of the retrieval heads are activated in different contexts. (5) causal: completely pruning retrieval heads leads to failure in retrieving relevant information and results in hallucination, while pruning random non-retrieval heads does not affect the model's retrieval ability. We further show that retrieval heads strongly influence chain-of-thought (CoT) reasoning, where the model needs to frequently refer back the question and previously-generated context. Conversely, tasks where the model directly generates the answer using its intrinsic knowledge are less impacted by masking out retrieval

heads. These observations collectively explain which internal part of the model seeks information from the input tokens. We believe our insights will foster future research on reducing hallucination, improving reasoning, and compressing the KV cache.New retrieval tasks have always been emerging, thus urging the development of new retrieval models. However, instantiating a retrieval model for each new retrieval task is resource-intensive and time-consuming, especially for a retrieval model that employs a large-scale pre-trained language model. To address this issue, we shift to a novel retrieval paradigm called modular retrieval, which aims to solve new retrieval tasks by instead composing multiple existing retrieval modules. Built upon the paradigm, we propose a retrieval model with modular prompt tuning named REMOP. It constructs retrieval modules subject to task attributes with deep prompt tuning, and yields retrieval models subject to tasks with module composition. We validate that, REMOP inherently with modularity not only has appealing generalizability and interpretability in preliminary explorations, but also achieves comparable performance to state-of-the-art retrieval models on a zero-shot retrieval benchmark.\footnote{Our code is available at \url{https://github.com/FreedomIntelligence/REMOP}}Large language models (LLMs) have the remarkable ability to solve new tasks with just a few examples, but they need access to the right tools. Retrieval Augmented Generation (RAG) addresses this problem by retrieving a list of relevant tools for a given task. However, RAG's tool retrieval step requires all the required information to be explicitly present in the query. This is a limitation, as semantic search, the widely adopted tool retrieval method, can fail when the query is incomplete or lacks context. To address this limitation, we propose Context Tuning for RAG, which employs a smart context retrieval system to fetch relevant information that improves both tool retrieval and plan generation. Our lightweight context retrieval model uses numerical, categorical, and habitual usage signals to retrieve and rank context items. Our empirical results demonstrate that context tuning significantly enhances semantic search, achieving a 3.5-fold and 1.5-fold improvement in Recall@K for context retrieval and tool retrieval tasks respectively, and resulting in an 11.6% increase in LLM-based planner accuracy. Additionally, we show that our proposed lightweight model using Reciprocal Rank Fusion (RRF) with LambdaMART outperforms GPT-4 based retrieval. Moreover, we observe context

augmentation at plan generation, even after tool retrieval, reduces hallucination.These lecture notes focus on the recent advancements in neural information retrieval, with particular emphasis on the systems and models exploiting transformer networks. These networks, originally proposed by Google in 2017, have seen a large success in many natural language processing and information retrieval tasks. While there are many fantastic textbook on information retrieval and natural language processing as well as specialised books for a more advanced audience, these lecture notes target people aiming at developing a basic understanding of the main information retrieval techniques and approaches based on deep learning. These notes have been prepared for a IR graduate course of the MSc program in Artificial Intelligence and Data Engineering at the University of Pisa, Italy.Generative retrieval generates identifiers of relevant documents in an end-to-end manner using a sequence-to-sequence architecture for a given query. The relation between generative retrieval and other retrieval methods, especially those based on matching within dense retrieval models, is not yet fully comprehended. Prior work has demonstrated that generative retrieval with atomic identifiers is equivalent to single-vector dense retrieval. Accordingly, generative retrieval exhibits behavior analogous to hierarchical search within a tree index in dense retrieval when using hierarchical semantic identifiers. However, prior work focuses solely on the retrieval stage without considering the deep interactions within the decoder of generative retrieval. In this paper, we fill this gap by demonstrating that generative retrieval and multi-vector dense retrieval share the same framework for measuring the relevance to a query of a document. Specifically, we examine the attention layer and prediction head of generative retrieval, revealing that generative retrieval can be understood as a special case of multi-vector dense retrieval. Both methods compute relevance as a sum of products of query and document vectors and an alignment matrix. We then explore how generative retrieval applies this framework, employing distinct strategies for computing document token vectors and the alignment matrix. We have conducted experiments to verify our conclusions and show that both paradigms exhibit commonalities of term matching in their alignment matrix.Generative retrieval is a promising new neural retrieval paradigm that aims to optimize the retrieval pipeline by performing both indexing and retrieval with a single transformer model.

However, this new paradigm faces challenges with updating the index and scaling to large collections. In this paper, we analyze two prominent variants of generative retrieval and show that they can be conceptually viewed as bi-encoders for dense retrieval. Specifically, we analytically demonstrate that the generative retrieval process can be decomposed into dot products between query and document vectors, similar to dense retrieval. This analysis leads us to propose a new variant of generative retrieval, called Tied-Atomic, which addresses the updating and scaling issues by incorporating techniques from dense retrieval. In experiments on two datasets, NQ320k and the full MSMARCO, we confirm that this approach does not reduce retrieval effectiveness while enabling the model to scale to large collections.This article presents a summary graph to show the relationships between Information Retrieval (IR) and other related disciplines. The figure tells the key differences between them and the conditions under which one would transition into another.We study retrieving a set of documents that covers various perspectives on a complex and contentious question (e.g., will ChatGPT do more harm than good?). We curate a Benchmark for Retrieval Diversity for Subjective questions (BERDS), where each example consists of a question and diverse perspectives associated with the question, sourced from survey questions and debate websites. On this data, retrievers paired with a corpus are evaluated to surface a document set that contains diverse perspectives. Our framing diverges from most retrieval tasks in that document relevancy cannot be decided by simple string matches to references. Instead, we build a language model based automatic evaluator that decides whether each retrieved document contains a perspective. This allows us to evaluate the performance of three different types of corpus (Wikipedia, web snapshot, and corpus constructed on the fly with retrieved pages from the search engine) paired with retrievers. Retrieving diverse documents remains challenging, with the outputs from existing retrievers covering all perspectives on only 33.74% of the examples. We further study the impact of query expansion and diversity-focused reranking approaches and analyze retriever sycophancy. Together, we lay the foundation for future studies in retrieval diversity handling complex queries.Despite the advantages of their low-resource settings, traditional sparse retrievers depend on exact matching approaches between high-dimensional bag-of-words (BoW) representations of

both the queries and the collection. As a result, retrieval performance is restricted by semantic discrepancies and vocabulary gaps. On the other hand, transformer-based dense retrievers introduce significant improvements in information retrieval tasks by exploiting low-dimensional contextualized representations of the corpus. While dense retrievers are known for their relative effectiveness, they suffer from lower efficiency and lack of generalization issues, when compared to sparse retrievers. For a lightweight retrieval task, high computational resources and time consumption are major barriers encouraging the renunciation of dense models despite potential gains. In this work, we propose boosting the performance of sparse retrievers by expanding both the queries and the documents with linked entities in two formats for the entity names: 1) explicit and 2) hashed. We employ a zero-shot end-to-end dense entity linking system for entity recognition and disambiguation to augment the corpus. By leveraging the advanced entity linking methods, we believe that the effectiveness gap between sparse and dense retrievers can be narrowed. We conduct our experiments on the MS MARCO passage dataset. Since we are concerned with the early stage retrieval in cascaded ranking architectures of large information retrieval systems, we evaluate our results using recall@1000. Our approach is also capable of retrieving documents for query subsets judged to be particularly difficult in prior work. We further demonstrate that the non-expanded and the expanded runs with both explicit and hashed entities retrieve complementary results. Consequently, we adopt a run fusion approach to maximize the benefits of entity linking.We propose a framework for discriminative Information Retrieval (IR) atop linguistic features, trained to improve the recall of tasks such as answer candidate passage retrieval, the initial step in text-based Question Answering (QA). We formalize this as an instance of linear feature-based IR (Metzler and Croft, 2007), illustrating how a variety of knowledge discovery tasks are captured under this approach, leading to a 44% improvement in recall for candidate triage for QA.Despite the advantages of their low-resource settings, traditional sparse retrievers depend on exact matching approaches between high-dimensional bag-of-words (BoW) representations of both the queries and the collection. As a result, retrieval performance is restricted by semantic discrepancies and vocabulary gaps. On the other hand, transformer-based dense retrievers introduce significant

improvements in information retrieval tasks by exploiting low-dimensional contextualized representations of the corpus. While dense retrievers are known for their relative effectiveness, they suffer from lower efficiency and lack of generalization issues, when compared to sparse retrievers. For a lightweight retrieval task, high computational resources and time consumption are major barriers encouraging the renunciation of dense models despite potential gains. In this work, I propose boosting the performance of sparse retrievers by expanding both the queries and the documents with linked entities in two formats for the entity names: 1) explicit and 2) hashed. A zero-shot end-to-end dense entity linking system is employed for entity recognition and disambiguation to augment the corpus. By leveraging the advanced entity linking methods, I believe that the effectiveness gap between sparse and dense retrievers can be narrowed. Experiments are conducted on the MS MARCO passage dataset using the original qrel set, the re-ranked qrels favoured by MonoT5 and the latter set further re-ranked by DuoT5. Since I am concerned with the early stage retrieval in cascaded ranking architectures of large information retrieval systems, the results are evaluated using recall@1000. The suggested approach is also capable of retrieving documents for query subsets judged to be particularly difficult in prior work.In this paper, we consider the phase retrieval problem in which one aims to recover a signal from the magnitudes of affine measurements. Let $\{{\mathbf a}_j\}_{j=1}^m \subset {\mathbb H}^d$ and ${\mathbf b}=(b_1, \ldots, b_m)^\top\in{\mathbb H}^m$, where ${\mathbb H}={\mathbb R}$ or ${\mathbb C}$. We say $\{{\mathbf a}_j\}_{j=1}^m$ and $\mathbf b$ are affine phase retrievable for ${\mathbb H}^d$ if any ${\mathbf x}\in{\mathbb H}^d$ can be recovered from the magnitudes of the affine measurements $\{|<{\mathbf a}_j,{\mathbf x}>+b_j|,\, 1\leq j\leq m\}$. We develop general framework for affine phase retrieval and prove necessary and sufficient conditions for $\{{\mathbf a}_j\}_{j=1}^m$ and $\mathbf b$ to be affine phase retrievable. We establish results on minimal measurements and generic measurements for affine phase retrieval as well as on sparse affine phase retrieval. In particular, we also highlight some notable differences between affine phase retrieval and the standard phase retrieval in which one aims to recover a signal $\mathbf x$ from the magnitudes of its linear measurements. In standard phase retrieval, one can only recover $\mathbf x$ up to a

unimodular constant, while affine phase retrieval removes this ambiguity. We prove that unlike standard phase retrieval, the affine phase retrievable measurements $\{{\mathbf a}_j\}_{j=1}^m$ and $\mathbf b$ do not form an open set in ${\mathbb H}^{m\times d}\times {\mathbb H}^m$. Also in the complex setting, the standard phase retrieval requires $4d-O(\log_2d)$ measurements, while the affine phase retrieval only needs $m=3d$ measurements.Multi-vector retrieval models such as ColBERT [Khattab and Zaharia, 2020] allow token-level interactions between queries and documents, and hence achieve state of the art on many information retrieval benchmarks. However, their non-linear scoring function cannot be scaled to millions of documents, necessitating a three-stage process for inference: retrieving initial candidates via token retrieval, accessing all token vectors, and scoring the initial candidate documents. The non-linear scoring function is applied over all token vectors of each candidate document, making the inference process complicated and slow. In this paper, we aim to simplify the multi-vector retrieval by rethinking the role of token retrieval. We present XTR, ConteXtualized Token Retriever, which introduces a simple, yet novel, objective function that encourages the model to retrieve the most important document tokens first. The improvement to token retrieval allows XTR to rank candidates only using the retrieved tokens rather than all tokens in the document, and enables a newly designed scoring stage that is two-to-three orders of magnitude cheaper than that of ColBERT. On the popular BEIR benchmark, XTR advances the state-of-the-art by 2.8 nDCG@10 without any distillation. Detailed analysis confirms our decision to revisit the token retrieval stage, as XTR demonstrates much better recall of the token retrieval stage compared to ColBERT.Developing a universal model that can efficiently and effectively respond to a wide range of information access requests -- from retrieval to recommendation to question answering -- has been a long-lasting goal in the information retrieval community. This paper argues that the flexibility, efficiency, and effectiveness brought by the recent development in dense retrieval and approximate nearest neighbor search have smoothed the path towards achieving this goal. We develop a generic and extensible dense retrieval framework, called \framework, that can handle a wide range of (personalized) information access requests, such as keyword search, query by example, and complementary item recommendation. Our proposed

approach extends the capabilities of dense retrieval models for ad-hoc retrieval tasks by incorporating user-specific preferences through the development of a personalized attentive network. This allows for a more tailored and accurate personalized information access experience. Our experiments on real-world e-commerce data suggest the feasibility of developing universal information access models by demonstrating significant improvements even compared to competitive baselines specifically developed for each of these individual information access tasks. This work opens up a number of fundamental research directions for future exploration.Cross-modal retrieval aims to search for data with similar semantic meanings across different content modalities. However, cross-modal retrieval requires huge amounts of storage and retrieval time since it needs to process data in multiple modalities. Existing works focused on learning single-source compact features such as binary hash codes that preserve similarities between different modalities. In this work, we propose a jointly learned deep hashing and quantization network (HQ) for cross-modal retrieval. We simultaneously learn binary hash codes and quantization codes to preserve semantic information in multiple modalities by an end-to-end deep learning architecture. At the retrieval step, binary hashing is used to retrieve a subset of items from the search space, then quantization is used to re-rank the retrieved items. We theoretically and empirically show that this two-stage retrieval approach provides faster retrieval results while preserving accuracy. Experimental results on the NUS-WIDE, MIR-Flickr, and Amazon datasets demonstrate that HQ achieves boosts of more than 7% in precision compared to supervised neural network-based compact coding models.Large language models (LLMs) inevitably exhibit hallucinations since the accuracy of generated texts cannot be secured solely by the parametric knowledge they encapsulate. Although retrieval-augmented generation (RAG) is a practicable complement to LLMs, it relies heavily on the relevance of retrieved documents, raising concerns about how the model behaves if retrieval goes wrong. To this end, we propose the Corrective Retrieval Augmented Generation (CRAG) to improve the robustness of generation. Specifically, a lightweight retrieval evaluator is designed to assess the overall quality of retrieved documents for a query, returning a confidence degree based on which different knowledge retrieval actions can be triggered. Since retrieval from static and limited corpora

can only return sub-optimal documents, large-scale web searches are utilized as an extension for augmenting the retrieval results. Besides, a decompose-then-recompose algorithm is designed for retrieved documents to selectively focus on key information and filter out irrelevant information in them. CRAG is plug-and-play and can be seamlessly coupled with various RAG-based approaches. Experiments on four datasets covering short- and long-form generation tasks show that CRAG can significantly improve the performance of RAG-based approaches.Retrieval approaches that score documents based on learned dense vectors (i.e., dense retrieval) rather than lexical signals (i.e., conventional retrieval) are increasingly popular. Their ability to identify related documents that do not necessarily contain the same terms as those appearing in the user's query (thereby improving recall) is one of their key advantages. However, to actually achieve these gains, dense retrieval approaches typically require an exhaustive search over the document collection, making them considerably more expensive at query-time than conventional lexical approaches. Several techniques aim to reduce this computational overhead by approximating the results of a full dense retriever. Although these approaches reasonably approximate the top results, they suffer in terms of recall -- one of the key advantages of dense retrieval. We introduce 'LADR' (Lexically-Accelerated Dense Retrieval), a simple-yet-effective approach that improves the efficiency of existing dense retrieval models without compromising on retrieval effectiveness. LADR uses lexical retrieval techniques to seed a dense retrieval exploration that uses a document proximity graph. We explore two variants of LADR: a proactive approach that expands the search space to the neighbors of all seed documents, and an adaptive approach that selectively searches the documents with the highest estimated relevance in an iterative fashion. Through extensive experiments across a variety of dense retrieval models, we find that LADR establishes a new dense retrieval effectiveness-efficiency Pareto frontier among approximate k nearest neighbor techniques. Further, we find that when tuned to take around 8ms per query in retrieval latency on our hardware, LADR consistently achieves both precision and recall that are on par with an exhaustive search on standard benchmarks.Legal case retrieval is a special Information Retrieval~(IR) task focusing on legal case documents. Depending on the downstream tasks of the retrieved case documents, users'

information needs in legal case retrieval could be significantly different from those in Web search and traditional ad-hoc retrieval tasks. While there are several studies that retrieve legal cases based on text similarity, the underlying search intents of legal retrieval users, as shown in this paper, are more complicated than that yet mostly unexplored. To this end, we present a novel hierarchical intent taxonomy of legal case retrieval. It consists of five intent types categorized by three criteria, i.e., search for Particular Case(s), Characterization, Penalty, Procedure, and Interest. The taxonomy was constructed transparently and evaluated extensively through interviews, editorial user studies, and query log analysis. Through a laboratory user study, we reveal significant differences in user behavior and satisfaction under different search intents in legal case retrieval. Furthermore, we apply the proposed taxonomy to various downstream legal retrieval tasks, e.g., result ranking and satisfaction prediction, and demonstrate its effectiveness. Our work provides important insights into the understanding of user intents in legal case retrieval and potentially leads to better retrieval techniques in the legal domain, such as intent-aware ranking strategies and evaluation methodologies.Evaluation of information retrieval systems (IRS) is a prominent topic among information retrieval researchers--mainly directed at a general population. Children require unique IRS and by extension different ways to evaluate these systems, but as a large population that use IRS have largely been ignored on the evaluation front. In this position paper, we explore many perspectives that must be considered when evaluating IRS; we specially discuss problems faced by researchers who work with children IRS, including lack of evaluation frameworks, limitations of data, and lack of user judgment understanding.In this work, we address multi-modal information needs that contain text questions and images by focusing on passage retrieval for outside-knowledge visual question answering. This task requires access to outside knowledge, which in our case we define to be a large unstructured passage collection. We first conduct sparse retrieval with BM25 and study expanding the question with object names and image captions. We verify that visual clues play an important role and captions tend to be more informative than object names in sparse retrieval. We then construct a dual-encoder dense retriever, with the query encoder being LXMERT, a multi-modal pre-trained transformer. We further show that dense retrieval significantly outperforms

sparse retrieval that uses object expansion. Moreover, dense retrieval matches the performance of sparse retrieval that leverages human-generated captions.This paper introduces the first two pixel retrieval benchmarks. Pixel retrieval is segmented instance retrieval. Like semantic segmentation extends classification to the pixel level, pixel retrieval is an extension of image retrieval and offers information about which pixels are related to the query object. In addition to retrieving images for the given query, it helps users quickly identify the query object in true positive images and exclude false positive images by denoting the correlated pixels. Our user study results show pixel-level annotation can significantly improve the user experience. Compared with semantic and instance segmentation, pixel retrieval requires a fine-grained recognition capability for variable-granularity targets. To this end, we propose pixel retrieval benchmarks named PROxford and PRParis, which are based on the widely used image retrieval datasets, ROxford and RParis. Three professional annotators label 5,942 images with two rounds of double-checking and refinement. Furthermore, we conduct extensive experiments and analysis on the SOTA methods in image search, image matching, detection, segmentation, and dense matching using our pixel retrieval benchmarks. Results show that the pixel retrieval task is challenging to these approaches and distinctive from existing problems, suggesting that further research can advance the content-based pixel-retrieval and thus user search experience. The datasets can be downloaded from \href{https://github.com/anguoyuan/Pixel_retrieval-Segmented_instance_retrieval}{this link}.Retrieving paragraphs to populate a Wikipedia article is a challenging task. The new TREC Complex Answer Retrieval (TREC CAR) track introduces a comprehensive dataset that targets this retrieval scenario. We present early results from a variety of approaches -- from standard information retrieval methods (e.g., tf-idf) to complex systems that using query expansion using knowledge bases and deep neural networks. The goal is to offer future participants of this track an overview of some promising approaches to tackle this problem.Large Language Models (LLMs) excel in various language tasks but they often generate incorrect information, a phenomenon known as "hallucinations". Retrieval-Augmented Generation (RAG) aims to mitigate this by using document retrieval for accurate responses. However, RAG still faces hallucinations due to vague queries. This

study aims to improve RAG by optimizing query generation with a query-document alignment score, refining queries using LLMs for better precision and efficiency of document retrieval. Experiments have shown that our approach improves document retrieval, resulting in an average accuracy gain of 1.6%.Document retrieval systems have experienced a revitalized interest with the advent of retrieval-augmented generation (RAG). RAG architecture offers a lower hallucination rate than LLM-only applications. However, the accuracy of the retrieval mechanism is known to be a bottleneck in the efficiency of these applications. A particular case of subpar retrieval performance is observed in situations where multiple documents from several different but related topics are in the corpus. We have devised a new vectorization method that takes into account the topic information of the document. The paper introduces this new method for text vectorization and evaluates it in the context of RAG. Furthermore, we discuss the challenge of evaluating RAG systems, which pertains to the case at hand.In multitask retrieval, a single retriever is trained to retrieve relevant contexts for multiple tasks. Despite its practical appeal, naive multitask retrieval lags behind task-specific retrieval in which a separate retriever is trained for each task. We show that it is possible to train a multitask retriever that outperforms task-specific retrievers by promoting task specialization. The main ingredients are: (1) a better choice of pretrained model (one that is explicitly optimized for multitasking) along with compatible prompting, and (2) a novel adaptive learning method that encourages each parameter to specialize in a particular task. The resulting multitask retriever is highly performant on the KILT benchmark. Upon analysis, we find that the model indeed learns parameters that are more task-specialized compared to naive multitasking without prompting or adaptive learning.Though pre-training vision-language models have demonstrated significant benefits in boosting video-text retrieval performance from large-scale web videos, fine-tuning still plays a critical role with manually annotated clips with start and end times, which requires considerable human effort. To address this issue, we explore an alternative cheaper source of annotations, single timestamps, for video-text retrieval. We initialise clips from timestamps in a heuristic way to warm up a retrieval model. Then a video clip editing method is proposed to refine the initial rough boundaries to improve retrieval performance. A student-teacher network is

introduced for video clip editing. The teacher model is employed to edit the clips in the training set whereas the student model trains on the edited clips. The teacher weights are updated from the student's after the student's performance increases. Our method is model agnostic and applicable to any retrieval models. We conduct experiments based on three state-of-the-art retrieval models, COOT, VideoCLIP and CLIP4Clip. Experiments conducted on three video retrieval datasets, YouCook2, DiDeMo and ActivityNet-Captions show that our edited clips consistently improve retrieval performance over initial clips across all the three retrieval models.The landscape of information retrieval has broadened from search services to a critical component in various advanced applications, where indexing efficiency, cost-effectiveness, and freshness are increasingly important yet remain less explored. To address these demands, we introduce Semi-parametric Vocabulary Disentangled Retrieval (SVDR). SVDR is a novel semi-parametric retrieval framework that supports two types of indexes: an embedding-based index for high effectiveness, akin to existing neural retrieval methods; and a binary token index that allows for quick and cost-effective setup, resembling traditional term-based retrieval. In our evaluation on three open-domain question answering benchmarks with the entire Wikipedia as the retrieval corpus, SVDR consistently demonstrates superiority. It achieves a 3% higher top-1 retrieval accuracy compared to the dense retriever DPR when using an embedding-based index and an 9% higher top-1 accuracy compared to BM25 when using a binary token index. Specifically, the adoption of a binary token index reduces index preparation time from 30 GPU hours to just 2 CPU hours and storage size from 31 GB to 2 GB, achieving a 90% reduction compared to an embedding-based index.Users and organizations are generating ever-increasing amounts of private data from a wide range of sources. Incorporating private data is important to personalize open-domain applications such as question-answering, fact-checking, and personal assistants. State-of-the-art systems for these tasks explicitly retrieve relevant information to a user question from a background corpus before producing an answer. While today's retrieval systems assume the corpus is fully accessible, users are often unable or unwilling to expose their private data to entities hosting public data. We first define the PUBLIC-PRIVATE AUTOREGRESSIVE INFORMATION RETRIEVAL (PAIR) privacy framework for

the novel retrieval setting over multiple privacy scopes. We then argue that an adequate benchmark is missing to study PAIR since existing textual benchmarks require retrieving from a single data distribution. However, public and private data intuitively reflect different distributions, motivating us to create ConcurrentQA, the first textual QA benchmark to require concurrent retrieval over multiple data-distributions. Finally, we show that existing systems face large privacy vs. performance tradeoffs when applied to our proposed retrieval setting and investigate how to mitigate these tradeoffs.Recent advances in large language models have enabled the development of viable generative retrieval systems. Instead of a traditional document ranking, generative retrieval systems often directly return a grounded generated text as a response to a query. Quantifying the utility of the textual responses is essential for appropriately evaluating such generative ad hoc retrieval. Yet, the established evaluation methodology for ranking-based ad hoc retrieval is not suited for the reliable and reproducible evaluation of generated responses. To lay a foundation for developing new evaluation methods for generative retrieval systems, we survey the relevant literature from the fields of information retrieval and natural language processing, identify search tasks and system architectures in generative retrieval, develop a new user model, and study its operationalization.Retrieving relevant plots from the book for a query is a critical task, which can improve the reading experience and efficiency of readers. Readers usually only give an abstract and vague description as the query based on their own understanding, summaries, or speculations of the plot, which requires the retrieval model to have a strong ability to estimate the abstract semantic associations between the query and candidate plots. However, existing information retrieval (IR) datasets cannot reflect this ability well. In this paper, we propose Plot Retrieval, a labeled dataset to train and evaluate the performance of IR models on the novel task Plot Retrieval. Text pairs in Plot Retrieval have less word overlap and more abstract semantic association, which can reflect the ability of the IR models to estimate the abstract semantic association, rather than just traditional lexical or semantic matching. Extensive experiments across various lexical retrieval, sparse retrieval, dense retrieval, and cross-encoder methods compared with human studies on Plot Retrieval show current IR models still struggle in capturing abstract semantic association between

texts. Plot Retrieval can be the benchmark for further research on the semantic association modeling ability of IR models.Building dense retrievers requires a series of standard procedures, including training and validating neural models and creating indexes for efficient search. However, these procedures are often misaligned in that training objectives do not exactly reflect the retrieval scenario at inference time. In this paper, we explore how the gap between training and inference in dense retrieval can be reduced, focusing on dense phrase retrieval (Lee et al., 2021) where billions of representations are indexed at inference. Since validating every dense retriever with a large-scale index is practically infeasible, we propose an efficient way of validating dense retrievers using a small subset of the entire corpus. This allows us to validate various training strategies including unifying contrastive loss terms and using hard negatives for phrase retrieval, which largely reduces the training-inference discrepancy. As a result, we improve top-1 phrase retrieval accuracy by 2~3 points and top-20 passage retrieval accuracy by 2~4 points for open-domain question answering. Our work urges modeling dense retrievers with careful consideration of training and inference via efficient validation while advancing phrase retrieval as a general solution for dense retrieval.Recently people started to understand that applications of the mathematical formalism of quantum theory are not reduced to physics. Nowadays, this formalism is widely used outside of quantum physics, in particular, in cognition, psychology, decision making, information processing, especially information retrieval. The latter is very promising. The aim of this brief introductory review is to stimulate research in this exciting area of information science. This paper is not aimed to present a complete review on the state of art in quantum information retrieval.Recent research has shown that transformer networks can be used as differentiable search indexes by representing each document as a sequences of document ID tokens. These generative retrieval models cast the retrieval problem to a document ID generation problem for each given query. Despite their elegant design, existing generative retrieval models only perform well on artificially-constructed and small-scale collections. This has led to serious skepticism in the research community on their real-world impact. This paper represents an important milestone in generative retrieval research by showing, for the first time, that generative retrieval models can be trained to perform effectively on

large-scale standard retrieval benchmarks. For doing so, we propose RIPOR- an optimization framework for generative retrieval that can be adopted by any encoder-decoder architecture. RIPOR is designed based on two often-overlooked fundamental design considerations in generative retrieval. First, given the sequential decoding nature of document ID generation, assigning accurate relevance scores to documents based on the whole document ID sequence is not sufficient. To address this issue, RIPOR introduces a novel prefix-oriented ranking optimization algorithm. Second, initial document IDs should be constructed based on relevance associations between queries and documents, instead of the syntactic and semantic information in the documents. RIPOR addresses this issue using a relevance-based document ID construction approach that quantizes relevance-based representations learned for documents. Evaluation on MSMARCO and TREC Deep Learning Track reveals that RIPOR surpasses state-of-the-art generative retrieval models by a large margin (e.g., 30.5% MRR improvements on MS MARCO Dev Set), and perform better on par with popular dense retrieval models.Private information retrieval (PIR) protocols make it possible to retrieve a file from a database without disclosing any information about the identity of the file being retrieved. These protocols have been rigorously explored from an information-theoretic perspective in recent years. While existing protocols strictly impose that no information is leaked on the file's identity, this work initiates the study of the tradeoffs that can be achieved by relaxing the requirement of perfect privacy. In case the user is willing to leak some information on the identity of the retrieved file, we study how the PIR rate, as well as the upload cost and access complexity, can be improved. For the particular case of replicated servers, we propose two weakly-private information retrieval schemes based on two recent PIR protocols and a family of schemes based on partitioning. Lastly, we compare the performance of the proposed schemes.We consider algorithm selection in the context of ad-hoc information retrieval. Given a query and a pair of retrieval methods, we propose a meta-learner that predicts how to combine the methods' relevance scores into an overall relevance score. Inspired by neural models' different properties with regard to IR axioms, these predictions are based on features that quantify axiom-related properties of the query and its top ranked documents. We conduct an evaluation on TREC Web Track data and find that the

meta-learner often significantly improves over the individual methods. Finally, we conduct feature and query weight analyses to investigate the meta-learner's behavior.Two key assumptions shape the usual view of ranked retrieval: (1) that the searcher can choose words for their query that might appear in the documents that they wish to see, and (2) that ranking retrieved documents will suffice because the searcher will be able to recognize those which they wished to find. When the documents to be searched are in a language not known by the searcher, neither assumption is true. In such cases, Cross-Language Information Retrieval (CLIR) is needed. This chapter reviews the state of the art for CLIR and outlines some open research questions.Pre-trained and fine-tuned transformer models like BERT and T5 have improved the state of the art in ad-hoc retrieval and question-answering, but not as yet in high-recall information retrieval, where the objective is to retrieve substantially all relevant documents. We investigate whether the use of transformer-based models for reranking and/or featurization can improve the Baseline Model Implementation of the TREC Total Recall Track, which represents the current state of the art for high-recall information retrieval. We also introduce CALBERT, a model that can be used to continuously fine-tune a BERT-based model based on relevance feedback.Information retrieval (IR) evaluation measures are cornerstones for determining the suitability and task performance efficiency of retrieval systems. Their metric and scale properties enable to compare one system against another to establish differences or similarities. Based on the representational theory of measurement, this paper determines these properties by exploiting the information contained in a retrieval measure itself. It establishes the intrinsic framework of a retrieval measure, which is the common scenario when the domain set is not explicitly specified. A method to determine the metric and scale properties of any retrieval measure is provided, requiring knowledge of only some of its attained values. The method establishes three main categories of retrieval measures according to their intrinsic properties. Some common user-oriented and system-oriented evaluation measures are classified according to the presented taxonomy.Image-based retrieval in large Earth observation archives is challenging because one needs to navigate across thousands of candidate matches only with the query image as a guide. By using text as information supporting the visual query, the retrieval system gains in

usability, but at the same time faces difficulties due to the diversity of visual signals that cannot be summarized by a short caption only. For this reason, as a matching-based task, cross-modal text-image retrieval often suffers from information asymmetry between texts and images. To address this challenge, we propose a Knowledge-aware Text-Image Retrieval (KTIR) method for remote sensing images. By mining relevant information from an external knowledge graph, KTIR enriches the text scope available in the search query and alleviates the information gaps between texts and images for better matching. Moreover, by integrating domain-specific knowledge, KTIR also enhances the adaptation of pre-trained vision-language models to remote sensing applications. Experimental results on three commonly used remote sensing text-image retrieval benchmarks show that the proposed knowledge-aware method leads to varied and consistent retrievals, outperforming state-of-the-art retrieval methods.Recent advances in dense retrieval techniques have offered the promise of being able not just to re-rank documents using contextualised language models such as BERT, but also to use such models to identify documents from the collection in the first place. However, when using dense retrieval approaches that use multiple embedded representations for each query, a large number of documents can be retrieved for each query, hindering the efficiency of the method. Hence, this work is the first to consider efficiency improvements in the context of a dense retrieval approach (namely ColBERT), by pruning query term embeddings that are estimated not to be useful for retrieving relevant documents. Our proposed query embeddings pruning reduces the cost of the dense retrieval operation, as well as reducing the number of documents that are retrieved and hence require to be fully scored. Experiments conducted on the MSMARCO passage ranking corpus demonstrate that, when reducing the number of query embeddings used from 32 to 3 based on the collection frequency of the corresponding tokens, query embedding pruning results in no statistically significant differences in effectiveness, while reducing the number of documents retrieved by 70%. In terms of mean response time for the end-to-end to end system, this results in a 2.65x speedup.Cross-modal contrastive learning has led the recent advances in multimodal retrieval with its simplicity and effectiveness. In this work, however, we reveal that cross-modal contrastive learning suffers from

incorrect normalization of the sum retrieval probabilities of each text or video instance. Specifically, we show that many test instances are either over- or under-represented during retrieval, significantly hurting the retrieval performance. To address this problem, we propose Normalized Contrastive Learning (NCL) which utilizes the Sinkhorn-Knopp algorithm to compute the instance-wise biases that properly normalize the sum retrieval probabilities of each instance so that every text and video instance is fairly represented during cross-modal retrieval. Empirical study shows that NCL brings consistent and significant gains in text-video retrieval on different model architectures, with new state-of-the-art multimodal retrieval metrics on the ActivityNet, MSVD, and MSR-VTT datasets without any architecture engineering.Generative retrieval is a promising new paradigm in text retrieval that generates identifier strings of relevant passages as the retrieval target. This paradigm leverages powerful generative language models, distinct from traditional sparse or dense retrieval methods. In this work, we identify a viable direction to further enhance generative retrieval via distillation and propose a feasible framework, named DGR. DGR utilizes sophisticated ranking models, such as the cross-encoder, in a teacher role to supply a passage rank list, which captures the varying relevance degrees of passages instead of binary hard labels; subsequently, DGR employs a specially designed distilled RankNet loss to optimize the generative retrieval model, considering the passage rank order provided by the teacher model as labels. This framework only requires an additional distillation step to enhance current generative retrieval systems and does not add any burden to the inference stage. We conduct experiments on four public datasets, and the results indicate that DGR achieves state-of-the-art performance among the generative retrieval methods. Additionally, DGR demonstrates exceptional robustness and generalizability with various teacher models and distillation losses.Dense retrieval has shown great success in passage ranking in English. However, its effectiveness in document retrieval for non-English languages remains unexplored due to the limitation in training resources. In this work, we explore different transfer techniques for document ranking from English annotations to multiple non-English languages. Our experiments on the test collections in six languages (Chinese, Arabic, French, Hindi, Bengali, Spanish) from diverse language families reveal that zero-shot model-based transfer using mBERT

improves the search quality in non-English mono-lingual retrieval. Also, we find that weakly-supervised target language transfer yields competitive performances against the generation-based target language transfer that requires external translators and query generators.Large-scale datasets are essential for the success of deep learning in image retrieval. However, manual assessment errors and semi-supervised annotation techniques can lead to label noise even in popular datasets. As previous works primarily studied annotation quality in image classification tasks, it is still unclear how label noise affects deep learning approaches to image retrieval. In this work, we show that image retrieval methods are less robust to label noise than image classification ones. Furthermore, we, for the first time, investigate different types of label noise specific to image retrieval tasks and study their effect on model performance.Synonymous keyword retrieval has become an important problem for sponsored search ever since major search engines relax the exact match product's matching requirement to a synonymous level. Since the synonymous relations between queries and keywords are quite scarce, the traditional information retrieval framework is inefficient in this scenario. In this paper, we propose a novel quotient space-based retrieval framework to address this problem. Considering the synonymy among keywords as a mathematical equivalence relation, we can compress the synonymous keywords into one representative, and the corresponding quotient space would greatly reduce the size of the keyword repository. Then an embedding-based retrieval is directly conducted between queries and the keyword representatives. To mitigate the semantic gap of the quotient space-based retrieval, a single semantic siamese model is utilized to detect both the keyword--keyword and query-keyword synonymous relations. The experiments show that with our quotient space-based retrieval method, the synonymous keyword retrieving performance can be greatly improved in terms of memory cost and recall efficiency. This method has been successfully implemented in Baidu's online sponsored search system and has yielded a significant improvement in revenue.We investigate knowledge retrieval with multi-modal queries, i.e. queries containing information split across image and text inputs, a challenging task that differs from previous work on cross-modal retrieval. We curate a new dataset called ReMuQ for benchmarking progress on this task. ReMuQ requires a system to retrieve

knowledge from a large corpus by integrating contents from both text and image queries. We introduce a retriever model ``ReViz'' that can directly process input text and images to retrieve relevant knowledge in an end-to-end fashion without being dependent on intermediate modules such as object detectors or caption generators. We introduce a new pretraining task that is effective for learning knowledge retrieval with multimodal queries and also improves performance on downstream tasks. We demonstrate superior performance in retrieval on two datasets (ReMuQ and OK-VQA) under zero-shot settings as well as further improvements when finetuned on these datasets.Creating an intelligent search and retrieval system for artwork images, particularly paintings, is crucial for documenting cultural heritage, fostering wider public engagement, and advancing artistic analysis and interpretation. Visual-Semantic Embedding (VSE) networks are deep learning models used for information retrieval, which learn joint representations of textual and visual data, enabling 1) cross-modal search and retrieval tasks, such as image-to-text and text-to-image retrieval; and 2) relation-focused retrieval to capture entity relationships and provide more contextually relevant search results. Although VSE networks have played a significant role in cross-modal information retrieval, their application to painting datasets, such as ArtUK, remains unexplored. This paper introduces BoonArt, a VSE-based cross-modal search engine that allows users to search for images using textual queries, and to obtain textual descriptions along with the corresponding images when using image queries. The performance of BoonArt was evaluated using the ArtUK dataset. Experimental evaluations revealed that BoonArt achieved 97% Recall@10 for image-to-text retrieval, and 97.4% Recall@10 for text-to-image Retrieval. By bridging the gap between textual and visual modalities, BoonArt provides a much-improved search performance compared to traditional search engines, such as the one provided by the ArtUK website. BoonArt can be utilised to work with other artwork datasets.