Knowledge Retrieval Over Public and Private Data

Simran Arora,[1] **Patrick Lewis,** [2] **Angela Fan,** [2] **Jacob Kahn**[*], [2] **Christopher Ré**[*1]

[1] Stanford University
[2] Facebook AI Research
simran@cs.stanford.edu

Abstract

Users and organizations are generating ever-increasing amounts of private data from a wide range of sources. Incorporating private context is important to personalize open-domain tasks such as question-answering, fact-checking, and personal assistants. State-of-the-art systems for these tasks explicitly retrieve information that is relevant to an input question from a background corpus before producing an answer. While today's retrieval systems assume relevant corpora are fully (e.g., publicly) accessible, users are often unable or unwilling to expose their private data to entities hosting public data. We define the PUBLIC-PRIVATE AUTOREGRESSIVE INFORMATION RETRIEVAL (PAIR) problem involving retrieval over multiple privacy scopes. We introduce a foundational benchmark with which to study PAIR, as no existing benchmark includes data from a private distribution. Our dataset, CONCURRENTQA, includes data from distinct public and private distributions and is the first textual QA benchmark requiring concurrent retrieval over multiple distributions. Finally, we show that existing retrieval approaches face significant performance degradations when applied to our proposed retrieval setting and investigate approaches with which these tradeoffs can be mitigated. We release the QA system and new benchmark: https://github.com/facebookresearch/concurrentqa

1 Introduction

The world's information is split between publicly and privately accessible scopes and the ability to simultaneously reason over both scopes is useful to support personalized tasks. However, retrieval-based machine learning (ML) systems, which first retrieve relevant information to a user input from background knowledge sources before providing an output, do not consider retrieving from private data that organizations and individuals aggregate locally. Neural retrieval systems are achieving impressive performance across applications such as language-modeling (Borgeaud et al. 2021), question-answering (Chen et al. 2017), and dialogue (Dinan et al. 2019), and we focus on the underexplored question of how to personalize these systems while preserving privacy.

Consider the following examples that require retrieving information from both public and private scopes. Individuals could ask *"With my GPA and SAT score, which universities should I apply to?"* or *"Is my blood pressure in the normal range for someone 55+?"*. In an organization, an ML engineer could ask: *"How do I fine-tune a language model, based on public StackOverflow and our internal company documentation?"*, or a doctor could ask *"How are COVID-19 vaccinations affecting patients with type-1 diabetes based on our private hospital records and public PubMed reports?"*. To answer such questions, users manually cross-reference public and private information sources. We initiate the study of a retrieval setting that enables using public (global) data to enhance our understanding of private (local) data.

Modern retrieval systems typically collect documents that are most-similar to a user's question from a massive corpus, and provide the resulting documents to a separate model, which reasons over the information to output an answer (Chen et al. 2017). Multi-hop reasoning (Welbl, Stenetorp, and Riedel 2018) can be used to answer complex queries over information distributed across multiple documents, e.g. news articles and Wikipedia. For such queries, we observe that using multiple rounds of retrieval (i.e., combining the original query with retrieved documents at round i for use in retrieval at round $i+1$) provides over 75% gains in performance vs. using one round of retrieval (Section 5). Iterative retrieval is now common in retrieval (Miller et al. 2016; Feldman and El-Yaniv 2019; Asai et al. 2020; Xiong et al. 2021; Qi et al. 2021; Khattab, Potts, and Zaharia 2021, inter alia.).

Existing multi-hop systems perform retrieval over a single privacy scope. However, users and organizations often cannot expose data to public entities. Maintaining terabyte-scale and dynamic data is difficult for many private entities, warranting retrieval from *multiple* distributed corpora.

To understand why distributed multi-hop retrieval implicates privacy concerns, consider two illustrative questions an employee may ask. First, to answer "*Of the products our competitors released this month, which are similar to our unreleased upcoming products?*", an existing multi-hop sys-

[*]These authors contributed equally.

Accepted to Workshop on Knowledge Augmented Methods for Natural Language Processing, in conjunction with AAAI 2023.

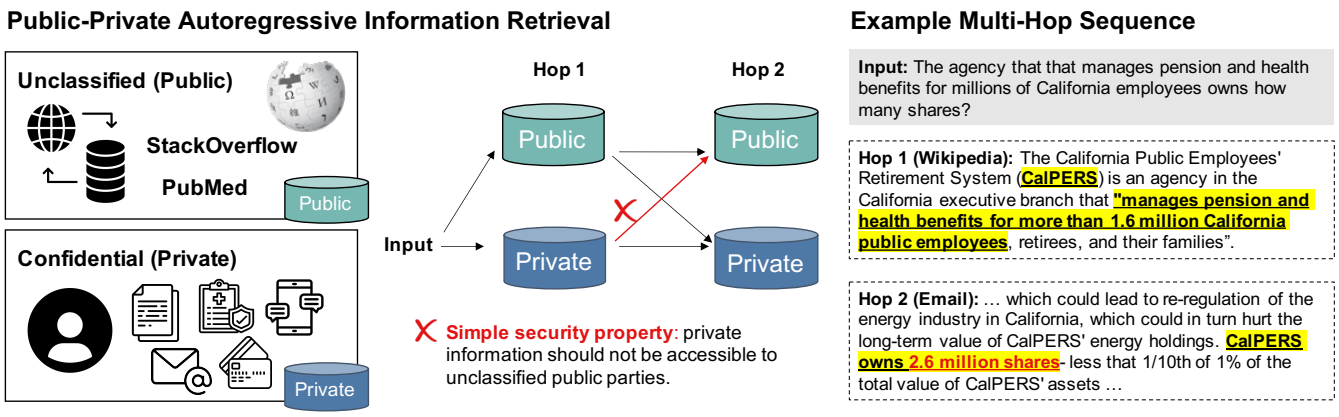

Figure 1: Multi-hop retrieval systems use beam search to collect information from a massive corpus: retrieval in hop_{i+1} is conditioned on the top documents retrieved in hop_i. The setting of retrieving from corpora distributed across multiple privacy scopes is unexplored. Here, the content of a private document retrieved in hop_i is revealed to the entity hosting public data if used to retrieve public documents in hop_{i+1}.

tem likely (1) retrieves public documents (e.g., news articles) about competitors, and (2) uses these to find private documents (e.g., company emails) about internal products, leaking no private information. Meanwhile, "*Have any companies ever released similar products to the one we are designing?*" entails (1) retrieving private documents detailing the upcoming product, and (2) performing similarity search for public products *using information from the confidential documents*. The latter reveals private data to an untrusted entity hosting a public corpus. An effective privacy model will minimize leakage.

We introduce the PUBLIC-PRIVATE AUTOREGRESSIVE INFORMATION RETRIEVAL (PAIR) problem. Public and private document distributions usually differ and our first observation is that *all* existing textual benchmarks require retrieving from one data-distribution. To appropriately evaluate PAIR, we create the first textual multi-distribution benchmark, CONCURRENTQA, which spans Wikipedia in the public domain and emails in the private domain, enabling the study of two novel real-world retrieval setups: (1) multi-distribution and (2) privacy-preserving retrieval:

- **Multi-distribution retrieval** The ability for a model to effectively retrieve over multiple distributions, even in the absence of privacy constraints, is a precursor to effective PAIR systems since it is unlikely for all private distributions to be reflected at train time. However, the typical retrieval setup requires retrieving over a single document distribution with a single query distribution (Thakur et al. 2021). We initiate the study of the real-world multi-distribution setting. We find that the SoTA multi-hop QA model trained on 90.4k Wikipedia data *underperforms* the same model trained on the 15.2k CONCURRENTQA (Wikipedia and Email) examples by 20.8 F1 points on questions based on Email passages. Further, we find the performance of the model trained on Wikipedia improves by 4.3% if we retrieve the top $\frac{k}{2}$ passages from each distribution vs. retrieving the overall top k passages, which is the standard protocol.
- **Privacy-preserving retrieval** We then propose a framework for reasoning about the privacy tradeoffs required for SoTA models to achieve as good performance on public-private QA as is achieved in public-QA. We evaluate performance when *no* private information is revealed, and models trained only on public data (e.g. Wikipedia) are utilized. Under this privacy standard, models sacrifice upwards of 19% performance under PAIR constraints to protect document privacy and 57% to protect query privacy when compared to a baseline system with standard, non-privacy aware retrieval mechanics. We then study how to manage the privacy-performance tradeoff using selective-prediction, a popular approach for improving the reliability of QA systems (Kamath, Jia, and Liang 2020; Lewis et al. 2021; Varshney, Mishra, and Baral 2022).

In summary: (1) We are the first to report on problems with applying existing neural retrieval-systems to the public and private retrieval setting, (2) We create CONCURRENTQA, the first textual multi-distribution benchmark to study the problems, and (3) We provide extensive evaluations of existing retrieval approaches under the proposed real-world retrieval settings. We hope this work encourages further research on private retrieval.

2 Background

Retrieval-Based Systems Open-domain applications, such as question-answering and personal assistants, must support user inputs across a broad range of topics. *Implicit-memory* approaches for these tasks focus on memorizing the knowledge required to answer questions within model parameters (Roberts, Raffel, and Shazeer 2020). Instead of memorizing massive amounts of knowledge in model parameters, *retrieval-based systems* introduce a step to retrieve information that is relevant to a user input from a massive corpus of documents (e.g., Wikipedia), and then provide this to a separate task model that produces the

output. Retrieval-free approaches have not been shown to work convincingly in multi-hop settings (Xiong et al. 2021).

Multi-hop Question Answering We focus on open-domain QA (ODQA), a classic application for retrieval-based systems. ODQA entails providing an answer a to a question q, expressed in natural language and without explicitly provided context from which to find the answer (Voorhees 1999). A *retriever* collects relevant documents to the question from a corpus, then a *reader* model extracts an answer from selected documents.

Our setting is concerned with complex queries where supporting evidence for the answer is distributed across multiple (public and private) documents, termed multi-hop reasoning (Welbl, Stenetorp, and Riedel 2018). To collect the distributed evidence, systems use multiple *hops* of retrieval: representations of the top passages retrieved in $\mathrm{hop_i}$ are used to retrieve passages in $\mathrm{hop_{i+1}}$ (Miller et al. 2016; Feldman and El-Yaniv 2019; Asai et al. 2020; Wolfson et al. 2020; Xiong et al. 2021; Qi et al. 2021; Khattab, Potts, and Zaharia 2021). Finally, we discuss the applicability of existing multi-hop benchmarks to our problem setting in Section 4.

Privacy Framework Private retrieval for open-domain applications is underexplored (Si and Yang 2014). Prior works securely retrieve sparse encodings of the passages and do not extend to neural-retrievers (Schoppmann et al. 2020; Lai et al. 2018). To enable private retrieval, Lai et al. (2018) represents each passage by a set of pre-defined attributes, while Marujo et al. (2014) requires re-running the protocol for each new set of queries. Modern retrieval systems use dense passage encodings and seek to ingest new passages and queries over time.

The computational overhead of existing cryptographic methods for private retrieval, such as obfuscating queries by interleaving real and fake queries or performing secure approximate nearest neighbor search, remain prohibitive (Gervais et al. 2014; Servan-Schreiber 2021) more so for queries requiring *multiple* hops. For instance, applying multi-party computation, Murugesan et al. (2010) requires 4k minutes to retrieve over a corpus of 1.5k documents with 1.5k queries for a single-hop, again using sparse encodings. Applications such as search require low latency. Cao et al. (2019) proposes fully on-device search, but scaling the amount of public data that can be hosted locally, not to mention updating at the of rate public data updates, remains challenging.

Finally, federated learning (FL) with differential privacy (DP) is a popular strategy for training models without exposing training data (McMahan et al. 2016; Dwork et al. 2006). This can help produce a retriever that generalizes to public and private distributions, granted there is training data for private distributions. Numerous attacks have been demonstrated showing FL leaks private information and DP can degrade quality (Shokri et al. 2017; Nasr, Shokri, and Houmansadr 2019). Furthermore, once the retriever is trained, it is used by public and private entities to encode the *raw documents*; FL and DP *do not protect* at this stage.

We initiate the study of private neural (and iterative) retrieval and we present a baseline that provides perfect-privacy (Shannon 1949).

3 Problem Definition

Objective Given a multi-hop input q, a set of private documents $p \in D_P$, and public documents $d \in D_G$, the objective is to provide the user with the correct answer a, which is contained in the documents. Figure 1 (Right) provides an example. Overall, the PUBLIC-PRIVATE AUTOREGRESSIVE INFORMATION RETRIEVAL problem entails maximizing quality, while protecting query and document privacy.

Standard, Non-Privacy Aware QA Standard non-private multi-hop ODQA involves answering q with the help of passages $d \in D_G$, using beam search. In the first iteration of retrieval, the k passages from the corpus, $d_1, ..., d_k$, that are most relevant to q are retrieved. The text of a retrieved passage is combined with q using function f (e.g., concatenating the query and passages sequences) to produce $q_i = f(q, d_i)$, for $i \in [1..k]$. Each q_i (which contains d_i) is used to retrieve k more passages in the following iteration.

We now introduce the PAIR retrieval problem. The user inputs to the QA system are the private corpus D_P and questions q. There are two key properties of the problem setting.

Property 1: Data is likely stored in multiple enclaves and personal documents $p \in \mathbf{D_P}$ can not leave the user's enclave. Users and organizations own private data, and untrustworthy (e.g., cloud) services own public data. First, we assume users likely do not want to publicly expose their data to create a single public corpus nor blindly write personal data to a public location. Next, we also assume it is challenging to store global data locally in many cases. This is because not only are there terabytes of public data and user-searches follow a long tail (Bernstein et al. 2012) (i.e. it is challenging to anticipate all a user's information needs), but public data is also constantly being updated (Zhang and Choi 2021). Thus, D_P and D_G are hosted as two separate corpora.

Now given q, the system must perform one retrieval over D_G and one over D_P rank the results such that the top-k passages will include k_P private and k_G public passages, and use these for the following iteration of retrieval. If the retrieval-system stops after a **single-hop**, there is no privacy risk since no $p \in D_P$ is publicly exposed. However for **multi-hop** questions, if $k_P > 0$ for an initial round of retrieval, meaning there exists some $p_i \in D_P$ which was in the top-k passages, it would sacrifice privacy if $f(q, p_i)$ were to be used to perform the next round of retrieval from D_G. Thus, for the strongest privacy guarantee, public retrievals should precede private document retrievals. For less privacy-sensitive use cases, this strict ordering can be weakened.

Property 2: Inputs that entirely rely on private information should not be revealed publicly. Given the multiple indices, D_P and D_G, q may be entirely answerable using multiple hops over the D_P index, in which case, q would never need to leave the user's device. For example, the query from an employee standpoint, *Does the search team use any infrastructure tools that our personal assistant team does not use?*, is fully answerable with company information. Prior work demonstrates that queries are very revealing of user interests, intents, and backgrounds (Xu et al. 2007; Gervais et al. 2014). There is an observable difference in the

search behavior of users with privacy concerns (Zimmerman et al. 2019) and an effective system will protect queries.

4 CONCURRENTQA Benchmark

Here we develop a testbed for studying public-private retrieval. We require questions spanning two corpora, D_P and D_G. First, we consider using existing benchmarks and describe the limitations we encounter, motivating the creation of our new benchmark, CONCURRENTQA. Then we describe the dataset collection process and its contents.

4.1 Adapting Existing Benchmarks

We first adapt the widely used benchmark, HotpotQA (Yang et al. 2018), to study our problem. HotpotQA contains multi-hop questions, which are each answered using two Wikipedia passages. We create HotpotQA-PAIR by splitting the Wikipedia corpus into D_G and D_P. This results in questions entirely reliant on $p \in D_P$, entirely on $d \in D_G$, or reliant on a mix of one private and one public document, allowing us to evaluate performance under PAIR constraints.

Ultimately however, D_P and D_G come from a single Wikipedia distribution in HotpotQA-PAIR. Private and public data will often reflect different linguistic styles, structures, and topics. We observe all existing textual multi-hop benchmarks require retrieving from a single distribution. We cannot combine two existing benchmarks over two different corpora because this will not yield any questions requiring one document from each corpus. To evaluate with a realistically private set of information and PAIR set up, we create a new benchmark: CONCURRENTQA. We quantitatively demonstrate the limitations of using HotpotQA-PAIR and CONCURRENTQA in the experiments and analyses.

4.2 CONCURRENTQA Overview

We create and release a new multi-hop QA dataset, CONCURRENTQA, which is designed to more closely resemble a practical use case for PAIR. CONCURRENTQA contains questions spanning Wikipedia documents as D_G and Enron employee emails (Klimt and Yang 2004) as D_P. We propose two unique evaluation settings for CONCURRENTQA: performance (1) conditioned on the sub-domains in which the question evidence can be found (Section 5), and (2) conditioned on the degree of privacy protection (Section 6).

Each benchmark example includes the *question* that requires reasoning over multiple documents, *answer* which is a span of text from the supporting documents, and the specific *supporting sentences* in the documents which are used to arrive at the answer and can serve as supervision signals.

Benchmark Collection We used Amazon Mechanical Turk for collection. In question generation, crowdworkers were shown sets of documents and asked to submit a question that requires reasoning over all the documents in the set. To select workers, we first used an onboarding exam to grant access to the main task. We then reviewed initial submissions from each candidate and allowed workers providing high-quality submissions to generate questions. We manually reviewed over 2.5k examples and prioritized including these in the final test and dev splits. Through reviewing, we identified key failure modes and used these insights to develop a second *validation* task. The validation task contained a multiple choice questionnaire about the previously generated QA pairs to filter low-quality submissions.

4.3 Benchmark Analysis

The corpora contain 47k emails (D_P) and 5.2M Wikipedia passages (D_G), and the benchmark contains 18,439 examples (Table 2). Questions require three main reasoning patterns: (1) *bridge questions* require identifying an entity or fact in Hop_1 on which the second retrieval is dependent, (2) *attribute questions* require identifying the entity that satisfies all attributes in the question, where attributes are distributed across passages, and (3) *comparison questions* require comparing two similar entities, each appearing in a separate passage. We estimate the benchmark is 80% bridge, 12% attribute, and 8% comparison questions. We focus on factoid QA; Figure 6 shows the distribution of answers' NER tags.

Emails and Wiki passages differ in several ways. **Format:** Wiki passages for entities of the same type tend to be similarly structured. Emails can contain portions of forwarded emails, lists of articles, or spam advertisements. **Noise:** Unlike Wiki passages, emails contain typos, URLs, and inconsistent capitalization. **Entities:** While a Wiki passage focuses on one entity, emails can cover multiple (possibly unrelated) topics. Enron entities are mentioned in multiple emails while public entities correspond to one Wiki passage. In the trainset gold supporting passages, Enron entities occur 9 times and Wiki entities occur 4 times on average. **Length:** emails are 3x longer than Wiki passages on average.

Limitations As in HotpotQA, workers see the gold supporting passages when writing questions, which can result in lexical overlap between the questions and passages. We mitigate these effects through validation task filtering and by limiting the allowed lexical overlap via the Turk interface. Next, our questions are not organic user searches, however existing search and dialogue logs do not contain questions over public and private data to our knowledge. Finally, Enron was a major public corporation; data encountered during model pretraining could impact the distinction between public vs. private data. We investigate this in Section 5.

Ethics Statement The Enron Dataset is already widely-used in NLP research (Heller 2017). That said, we acknowledge the origin of this data as collected and made public by the U.S. FERC during their investigation of Enron. We note that many of the individuals whose emails appear in the dataset were not involved in wrongdoing. We defer to using inboxes that are frequently used in prior work.

In the next sections, we evaluate CONCURRENTQA in the PAIR setting. We first ask how a range of SoTA retrievers perform in the mixed-domain retrieval setting in Section

The Enron Corpus includes emails generated by 158 employees of Enron Corporation and are in the public domain.

Since information density is generally lower in emails vs. Wiki passages, this helps crowdworkers generate meaningful questions. Lengths chosen within model context window.

Question	Hop 1 and Hop 2 Gold Passages
What was the estimated 2016 population of the city that generates power at the Hetch Hetchy hydroelectric dams?	*Hop 1* An email mentions that San Francisco generates power at the Hetch Hetchy dams. *Hop 2* The Wikipedia passage about San Francisco reports the 2016 census-estimated population.
Which firm invested in both the 5th round of funding for Extraprise and first round of funding for JobsOnline.com?	*Hop 1* An email lists 5th round Exraprise investors. *Hop 2* An email lists round-1 investors for JobsOnline.com.

Table 1: Example CONCURRENTQA queries based on Wikipedia passages (D_G) and emails (D_P).

Split	Total	EE	EW	WE	WW
Train	15,239	3762	4002	3431	4044
Dev	1,600	400	400	400	400
Test	1,600	400	400	400	400

Table 2: Size statistics. The CONCURRENTQA evaluation splits are balanced between questions with gold passages as emails (E) vs. Wikipedia (W) passages for Hop_1 and Hop_2.

5, then introduce baselines for CONCURRENTQA under the perfect-privacy privacy model in Section 6.

5 Evaluating Multi-domain Retrieval

Here we study the SoTA multi-hop model performance on CONCURRENTQA in the novel multi-distribution setting. The ability for models trained on public data to generalize to private distributions, with little or no labeled data, is a precursor to solutions for PAIR. In the commonly studied zero-shot retrieval setting (Guoa et al. 2021; Thakur et al. 2021) the top k of k passages will be from a single distribution, however users often have diverse questions and documents.

We first evaluate multi-hop retrievers. Then we apply strong single-hop retrievers to the setting, to understand the degree to which iterative retrieval is required.

5.1 Benchmarking Multi-Hop Retrievers

Retrievers We evaluate the multi-hop dense retrieval model (MDR) (Xiong et al. 2021), which achieves SoTA on multi-hop QA and multi-hop implementation of BM25, a classical bag-of-words method, as prior work indicates its strength in OOD retrieval (Thakur et al. 2021).

MDR is a bi-encoder model consisting of a query encoder and passage encoder. Passage embeddings are stored in an index designed for efficient retrieval (Johnson, Douze, and Jégou 2017). In Hop_1, the embedding for query q is used to retrieve the k passages $d_1, ..., d_k$ with the highest *retrieval score* by the maximum inner product between question and passage encodings. Next, the retrieved passages are each appended to q and encoded, and each of the k resulting embeddings are used to collect k more passages in Hop_2, yielding k^2 passages. The top-k of the passages after the final hop are inputs to the reader, ELECTRA-Large (Clark et al. 2020). The reader selects a candidate answer in each passage. The candidate with the highest *reader score* is outputted.

Baselines We evaluate using four retrieval baselines: **(1) CONCURRENTQA-MDR**, a dense retriever trained on the CONCURRENTQA train set (15.2k examples), to understand the value of in-domain training data for the task; **(2) HotpotQA-MDR**, trained on HotpotQA (90.4K examples), to understand how well a publicly trained model performs on the mixed distribution; **(3) Subsampled HotpotQA-MDR**, trained on subsampled HotpotQA data of the same size as the CONCURRENTQA train set, to investigate the effect of dataset size; and **(4) BM25** sparse retrieval. Results are in Table 3. Experimental details are in Appendix 8.1.

Training Data Size *Strong dense retrieval performance requires a large amount of training data.* Comparing CONCURRENTQA-MDR and Subsampled Hotpot-QA MDR, the former outperforms by 12.6 F1 points as it is evaluated in-domain. However, the HotpotQA-MDR baseline, trained on the full HotpotQA training set, performs nearly equal to CONCURRENTQA-MDR. Figure 2 shows the performance as training dataset size varies. Next we observe the sparse method matches the zero-shot performance of the Subsampled HotpotQA model on CONCURRENTQA. For larger dataset sizes (HotpotQA-MDR) and in-domain training data (CONCURRENTQA-MDR), dense outperforms sparse retrieval. Notably, it may be difficult to obtain training data for all private or temporally arising distributions.

Domain Specific Performance *Each retriever excels in a different subdomain of the benchmark*. Table 3 shows the retrieval performance of each method based on whether the gold supporting passages for Hop_1 and Hop_2 are email (E) or Wikipedia (W) passages (EW is Email-Wiki for Hop_1-Hop_2). HotpotQA-MDR performance on WW questions is far better than on questions involving emails. The sparse retriever performs worse than the dense models on questions involving W, but better on questions with E in Hop_2. When training on CONCURRENTQA, performance on ques-

Xiong et al. (2021) compares ELECTRA and other readers such as FiD (Izacard and Grave1 2021), finding similar performance. We follow their approach and use ELECTRA.

We check for dataset leakage stemming from the "public" models potentially viewing "private" email information in pretraining. Using the MDR and ELECTRA models fine-tuned on HotpotQA, we evaluate on CONCURRENTQA using a corpus of only Wiki passages. Test scores are 72.0 and 3.3 EM for questions based on two Wiki and two email passages respectively, suggesting explicit access to emails is important.

Retrieval Method	Overall		Domain-Conditioned			
	EM	*F1*	*EE*	*EW*	*WE*	*WW*
ConcurrentQA-MDR	**48.9**	**56.5**	**49.5**	**66.4**	**41.8**	68.3
HotpotQA-MDR	45.0	53.0	28.7	61.7	41.1	**81.3**
Subsampled HotpotQA-MDR	37.2	43.9	23.8	51.1	28.6	72.1
BM25	33.2	40.8	44.2	30.7	50.2	30.5
Oracle	74.1	83.4	66.5	87.5	89.4	90.4

Table 3: ConcurrentQA results using four retrieval approaches, and oracle retrieval. On the right, we show performance (F1 scores) by the domains of the Hop_1 and Hop_2 gold passages for each question (email is "E", Wikipedia is "W", and "EW" indicates the gold passages are email for Hop_1 and Wikipedia for Hop_2).

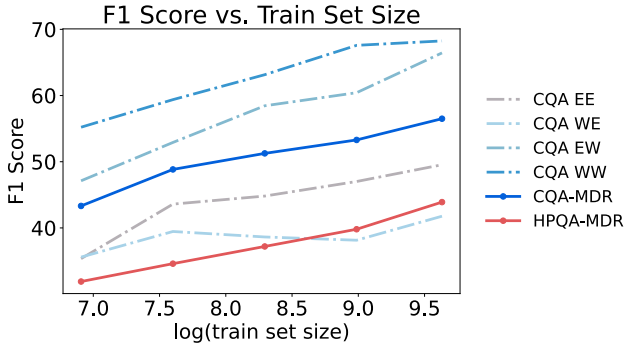

Figure 2: F1 score vs training data size, training MDR on subsampled HotpotQA (HPQA) and subsampled ConcurrentQA (CQA) training data. We also show trends by the question domain for CQA (dotted lines).

tions involving E improves significantly, but remains low on W-based questions. Finally, we explicitly provide the gold supporting passages to the reader model (Oracle). EE oracle performance also remains low, indicating room to improve the reader.

How well does the retriever trained on public data perform in the PAIR setting? We observe the HotpotQA-MDR model is biased towards retrieving Wikipedia passages. On examples where the gold Hop_1 passage is an email, 15% of the time, no emails appear in the top-k Hop_1 results; meanwhile, this only occurs 4% of the time when Hop_1 is Wikipedia. On the slice of EE examples, 64% of Hop_2 passages are E, while on the slice of WW examples, 99.9% of Hop_2 passages are W. If we simply *force* equal retrieval ($\frac{k}{2}$) from each domain on each hop, we observe 2.3 F1 points (4.3%) improvement in ConcurrentQAperformance, compared to retrieving the overall top-k. *Optimally* selecting the allocation for each domain is an exciting question for future work.

Performance on WE questions is notably worse than on EW questions. We hypothesize this is because several emails discuss each Wikipedia-entity, which may increase the noise in Hop_2 (i.e., WE is a one-to-many hop, while for EW, W typically contains one valid entity-specific passage). The latter is intuitively because individuals refer to a narrow set of public entities in private discourse.

5.2 Benchmarking Single-Hop Retrieval

In Section 3, we identify that iterative retrieval implicates document privacy. Therefore, an important preliminary question is to what degree multiple-hops are actually required? We investigate this question using both HotpotQA and ConcurrentQA. We evaluate MDR using just the first-hop results and Contriever (Izacard et al. 2021), the SoTA single-hop dense retrieval model.

Results Using the off-the-shelf HotpotQA fine-tuned MDR model and retrieving for just 1-hop results in 41% worse Recall@10 compared to 2-hop MDR on HotpotQA. Next, using the strong single-hop retriever, Contriever, we find Recall@10 remains 32% and 17% worse compared to 2-hop MDR on HotpotQA, when using the off-the-shelf pretrained and MS-MARCO (Nguyen et al. 2016) fine-tuned variants of this single-hop model respectively. Strong single-hop models trained over more diverse publicly available data may help address the PAIR problem.

However, on ConcurrentQA, we find that Contriever MS-MARCO Recall@10 is 34% worse than 2-hop MDR. By sub-domain, the Contriever model finds the gold first-hop passage for 85% of questions where both gold passages are from Wikipedia, but less than 39% when at least one gold passage (hop_1 and/or hop_2) is an email.

Analysis We observe the MS-MARCO fine-tuned Contriever model performs well with a single-hop over HotpotQA, but leaves a large gap on ConcurrentQA. In error analysis, we find the model retrieves the first-hop passage 49% of the time and second-hop passage 25% of the time. By sub-domain, the Contriever model finds the gold first-hop passage for 85% of questions where both gold passages are from Wikipedia, but less than 39% when at least one gold passage (hop_1 or hop_2) is from Enron, suggesting a quality gap between the distributions. Finally, Contriever is challenging to fine-tune on ConcurrentQA and we hypothesize this is because Contriever used a manual and careful negative sampling procedure. Overall, iterative retrieval provides large improvements over current single-hop retrievers.

6 Evaluation under Privacy Constraints

In this section, we provide baselines for ConcurrentQA under privacy constraints. There are many possible privacy constraints as users find different information to be sensitive (), so we demonstrate how to apply a classical privacy

Privacy Level	Sample Questions Answered under Each Privacy Level
Answered with **No Privacy**, but *not* under Document Privacy	*Q1* In which region is the **site of a meeting** between Dabhol manager Wade Cline and Ministry of Power Secretary A. K. Basu located? *Q2* What year was the state-owned regulation **board that was in conflict** with Dabhol Power over the DPC project formed?
Answered with **Document Privacy**	*Q1* The U.S. Representative from New York who served from 1983 to 2013 requested a summary of what **order concerning a price cap complaint**? *Q2* **How much of the company** known as DirecTV Group does General Motors own?
Answered with **Query Privacy**	*Q1* Which CarrierPoint backer has a partner on SupplySolution's board? *Q2* At the end of what year did Enron India's managing director responsible for managing operations for Dabhol Power believe it would go online? ***All evidence is in private emails and not in Wikipedia.**

Table 4: Examples of queries answered under different privacy restrictions. **Bold** indicates private information.

Model	HotpotQA-PAIR		ConcurrentQA	
	EM	*F1*	*EM*	*F1*
No Privacy Baseline	62.3	75.3	45.0	53.0
No Privacy Multi-Index	62.3	75.3	45.0	53.0
Document Privacy	56.8	68.8	36.1	43.0
Query Privacy	34.3	43.3	19.1	23.8

Table 5: Multi-hop QA datasets using the dense retrieval baseline (MDR) under each privacy setting.

model, perfect-privacy, to the PAIR retrieval setting.

The perfect-privacy guarantee is that as users interact with the system, the probability that adversaries learn private information does not increase (Shannon 1949; Miklau and Suciu 2004). Perfect-privacy based access-control frameworks are actively used in practice for highly sensitive settings such as government and medical data (Bell and LaPadula 1976; Hu, Ferraiolo, and Kuhn 2006), motivating our study of the framework.

Setup We use models trained on Wikipedia data, to evaluate performance under privacy restrictions both in the in-distribution multi-hop HotpotQA-PAIR (an adaptation of the HotpotQA benchmark to the PAIR setting (Yang et al. 2018)) and mixed-distribution ConcurrentQA (Wikipedia and Enron based) settings. Motivating the latter, sufficient training data is seldom available for all private distributions. We use the multi-hop SoTA model, MDR, which is representative of the iterative retrieval procedure that is used across multi-hop solutions (Miller et al. 2016; Feldman and El-Yaniv 2019; Xiong et al. 2021, inter alia.).

We construct Hotpot-PAIR by randomly assigning passages to the private (D_P) and public (D_G) corpora. To enable a clear comparison, we ensure that the sizes of D_P and D_G, and the proportions of questions for which the gold documents are public and private in Hop_1 and Hop_2 match those in ConcurrentQA.

6.1 Applying Perfect-Privacy to PAIR

We evaluate performance when no private information (neither queries nor documents) is revealed whatsoever. We compare four baselines, shown in Table 5. **(1) No Privacy Baseline**: We combine all public and private passages in one corpus, ignoring privacy concerns. **(2) No Privacy Multi-Index**: We create two corpora and retrieve the top k from each index in each hop, and retain the top-k of these $2k$ documents for the next hop, without applying any privacy restriction. Note performance should match single-index performance. **(3) Document Privacy**: We use the process in (2), but cannot use a private passage retrieved in Hop_1 to subsequently retrieve from public D_G. **(4) Query Privacy**: The perfect-privacy baseline to keep q private is to only retrieve from D_P.

Overall, we can answer many complex questions *while maintaining perfect-privacy* (see Table 4). However, in maintaining document privacy, the end-to-end QA performance degrades by 9% HotpotQA and 19% for ConcurrentQA compared to the quality of the non-private system; degradation is worse under query privacy. Perfect-privacy is a natural starting point and we hope future work studies alternate privacy models using the resources we provide.

Setup Selective prediction aims to provide the user with an answer only when the model is confident. The goal is to answer as many questions as possible (*high coverage*) with as high performance as possible (*low risk*). Given query q, and a model which outputs $(\hat{a}, c)$, where $\hat{a}$ is the predicted answer and $c \in R$ represents the model's confidence in $\hat{a}$, we output $\hat{a}$ if $c \geq \gamma$ for some threshold $\gamma \in R$, and abstain otherwise. As γ increases, risk and coverage both tend to decrease. The QA model outputs an answer and score for each of the top-k retrieved passages — we compute the softmax over the top-k scores and use the top softmax score as c (Hendrycks and Gimpel 2017; Varshney, Mishra, and Baral

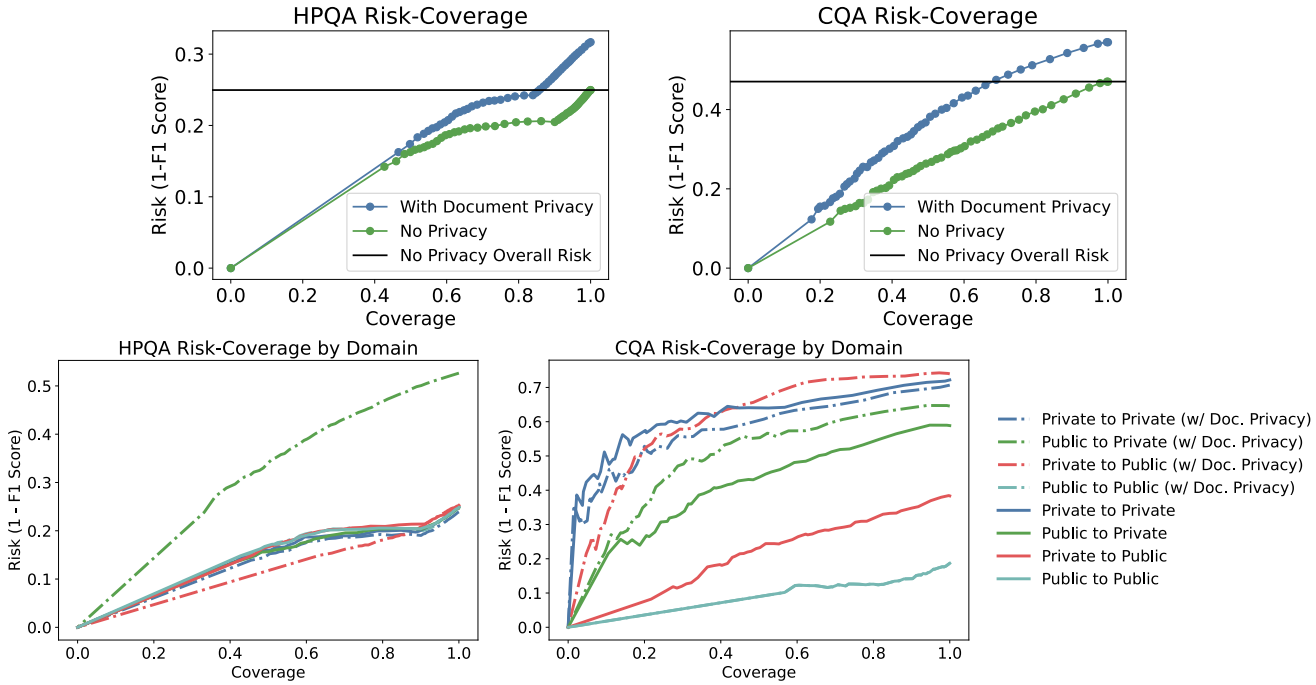

Figure 3: Risk-coverage curves using the model trained on Wikipedia data for HotpotQA-PAIR and mixed-distribution CONCURRENTQA retrieval, both under No Privacy and perfect Document Privacy (which restricts $Private$ to $Public$ retrieval). The left shows the overall test results, and the right is split by the the domains of the gold supporting passages for the question at hand, for Hop_1 to Hop_2.

2022). Models are still trained in the *public domain*, i.e. on HotpotQA.

Results Risk-coverage curves for HotpotQA and CONCURRENTQA are in Figure 3. Under Document Privacy, the "No Privacy" score of 75.3 F1 for HotpotQA and 53.0 F1 for CONCURRENTQA are achieved at 85.7% and 67.8% coverage respectively.

First, we observe in the top plots that in the absence of privacy concerns, the risk-coverage trends are worse for CONCURRENTQA vs. HotpotQA (i.e. the quality degrades more quickly as the coverage increases). Out-of-distribution selective prediction is actively studied (Kamath, Jia, and Liang 2020). However, this setting differs from the standard setup. The bottom plots show on CONCURRENTQA that that the risk-coverage trends differ widely based on the sub-domains of the questions; the standard retrieval setup typically has a single distribution (Thakur et al. 2021).

Second, privacy restrictions correlate with degredations in the risk-coverage curves on both CONCURRENTQA and HotpotQA. Critically, HotpotQA is in-distribution for the retriever. Strategies beyond selective prediction via max-prob, the prevailing approach in NLP (Varshney, Mishra, and Baral 2022), may be useful for the PAIR setting.

7 Conclusion

We ask how to personalize neural retrieval-systems in a privacy-preserving way and report on how arbitrary retrieval over public and private data poses a privacy concern. We define the PAIR retrieval problem, present the first textual multi-distribution benchmark to study the novel setting, and empirically characterize the privacy-quality tradeoffs faced by neural retrieval systems.

In Section 4, we motivated the creation of CONCURRENTQA, rather than simply repurposing existing benchmarks such as HotpotQA, by noting CONCURRENTQA is multi-distributional. In summary, we qualitatitvely analyzed how the public Wikipedia and private emails in Section 4.3, and demonstrated the unique retrieval challenges of applying models trained on one distribution (e.g. public) to the mixed-distribution (e.g. public and private) setting in Sections 5 and 6. Private-public retrieval is intuitively often a mixed-distribution problem, warranting the new benchmark.

Private neural retrieval is underexplored and we hope the benchmark-resource and evaluations we provide inspire further research, for instance under alternate privacy models.

References

Asai, A.; Hashimoto, K.; Hajishirzi, H.; Socher, R.; and Xiong, C. 2020. Learning to Retrieve Reasoning Paths over Wikipedia Graph for Question Answering. In *International Conference on Learning Representations (ICLR)*.

Bell, D. E.; and LaPadula, L. J. 1976. SECURE COMPUTER SYSTEM: UNIFIED EXPOSITION AND MULTICS INTERPRETATION. *The MITRE Corporation*.

Bernstein, M. S.; Teevan, J.; Dumais, S.; Liebling, D.; ; and

Horvitz, E. 2012. Direct answers for search queries in the long tail. *SIGCHI*.

Borgeaud, S.; Mensch, A.; Hoffmann, J.; Cai, T.; Rutherford, E.; Millican, K.; van den Driessche, G.; Lespiau, J.-B.; Damoc, B.; Clark, A.; de Las Casas, D.; Guy, A.; Menick, J.; Ring, R.; Hennigan, T.; Huang, S.; Maggiore, L.; Jones, C.; Cassirer, A.; Brock, A.; Paganini, M.; Irving, G.; Vinyals, O.; Osindero, S.; Simonyan, K.; Rae, J. W.; Elsen, E.; and Sifre, L. 2021. Improving language models by retrieving from trillions of tokens. In *arXiv:2112.04426v2*.

Cao, Q.; Weber, N.; Balasubramanian, N.; and Balasubramanian, A. 2019. DeQA: On-Device Question Answering. In *The 17th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys)*.

Chen, D.; Fisch, A.; Weston, J.; and Bordes, A. 2017. Reading Wikipedia to answer open-domain questions. In *Association for Computational Linguistics (ACL)*.

Clark, K.; Luong, M.-T.; Le, Q. V.; and Manning, C. D. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations (ICLR)*.

Dinan, E.; Roller, S.; Shuster, K.; Fan, A.; Auli, M.; and Weston, J. 2019. Wizard of Wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations (ICLR)*.

Dwork, C.; McSherry, F.; Nissim, K.; and Smith, A. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. *Theory of Cryptography Conference (TCC)*.

Feldman, Y.; and El-Yaniv, R. 2019. Multi-Hop Paragraph Retrieval for Open-Domain Question Answering. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Gervais, A.; Shokri, R.; Singla, A.; Capkun, S.; and Lenders, V. 2014. Quantifying Web-Search Privacy. In *ACM Conference on Computer and Communications Security (SIGSAC)*.

Guoa, M.; Yanga, Y.; Cera, D.; Shenb, Q.; and Constant, N. 2021. MultiReQA: A Cross-Domain Evaluation for Retrieval Question Answering Models. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*.

Heller, N. 2017. What the Enron E-mails Say About Us.

Hendrycks, D.; and Gimpel, K. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*.

Hu, V. C.; Ferraiolo, D. F.; and Kuhn, D. R. 2006. Assessment of Access Control Systems. National Institute of Standards and Technology (NIST).

Izacard, G.; Caron, M.; Hosseini, L.; Riedel, S.; Bojanowski, P.; Joulin, A.; and Grave, E. 2021. Unsupervised Dense Information Retrieval with Contrastive Learning.

Izacard, G.; and Grave1, E. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Johnson, J.; Douze, M.; and Jégou, H. 2017. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*.

Kamath, A.; Jia, R.; and Liang, P. 2020. Selective Question Answering under Domain Shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; and tau Yih, W. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Khattab, O.; Potts, C.; and Zaharia, M. 2021. Baleen: Robust Multi-Hop Reasoning at Scale via Condensed Retrieval. *arXiv:2101.00436v2*.

Klimt, B.; and Yang, Y. 2004. Introducing the enron corpus. In *Proceedings of the 1st Conference on Email and Anti-Spam (CEAS)*.

Lai, J.; Mua, Y.; Guoa, F.; Jiang, P.; and Susilo, W. 2018. Privacy-enhanced attribute-based private information retrieval. *Information Sciences*.

Lewis, P.; Wu, Y.; Liu, L.; Minervini, P.; Küttler, H.; Piktus, A.; Stenetorp, P.; and Riedel, S. 2021. PAQ: 65 Million Probably-Asked Questions and What You Can Do With Them. In *Transactions of the Association for Computational Linguistics (TACL)*.

Marujo, L.; Portêlo, J.; de Matos, D. M.; Neto, J. P.; Gershman, A.; Carbonell, J.; Trancoso, I.; and Ra, B. 2014. Privacy-Preserving Important Passage Retrieval. *PIR'14,Privacy-Preserving IR: When Information Retrieval Meets Privacy and Security, SIGIR 2014 Workshop*.

McMahan, H. B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2016. Communication-Efficient Learning of Deep Networks from Decentralized Data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Miklau, G.; and Suciu, D. 2004. A Formal Analysis of Information Disclosure in Data Exchange. *SIGMOD*.

Miller, A. H.; Fisch, A.; Dodge, J.; Karimi, A.-H.; Bordes, A.; and Weston, J. 2016. Key-Value Memory Networks for Directly Reading Documents. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Murugesan, M.; Jiang, W.; Clifton, C.; Si, L.; and Vaidya, J. 2010. Efficient privacy-preserving similar document detection. *The VLDB Journal*.

Nasr, M.; Shokri, R.; and Houmansadr, A. 2019. Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning. *IEEE Symposium on Security and Privacy*.

Nguyen, T.; Rosenberg, M.; Song, X.; Gao, J.; Tiwary, S.; Majumder, R.; and Deng, L. 2016. MS Marco: A human generated machine reading comprehension dataset. *CoCo@NIPS*.

Qi, P.; Lee, H.; Sido, O. T.; and Manning, C. D. 2021. Retrieve, Read, Rerank, then Iterate: Answering Open-Domain Questions of Varying Reasoning Steps from Text. arXiv:2010.12527. Version 3.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Roberts, A.; Raffel, C.; and Shazeer, N. 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Schoppmann, P.; Vogelsang, L.; Gascón, A.; and Balle, B. 2020. Secure and Scalable Document Similarity on Distributed Databases: Differential Privacy to the Rescue. *Proceedings on Privacy Enhancing Technologies*.

Servan-Schreiber, S. 2021. Private Nearest Neighbor Search with Sublinear Communication and Malicious Security.

Shannon, C. E. 1949. Communication Theory of Secrecy Systems.

Shokri, R.; Stronati, M.; Song, C.; and Shmatikov, V. 2017. Membership Inference Attacks against Machine Learning Models. *In the proceedings of the IEEE Symposium on Security and Privacy*.

Si, L.; and Yang, H. 2014. Privacy-Preserving IR: When Information Retrieval Meets Privacy and Security. *SIGIR'14*.

Thakur, N.; Reimers, N.; Ruckle, A.; Srivastav, A.; and Gurevych, I. 2021. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS)*.

Varshney, N.; Mishra, S.; and Baral, C. 2022. Investigating Selective Prediction Approaches Across Several Tasks in IID, OOD, and Adversarial Settings. *Findings of the Association for Computational Linguistics: ACL 2022*.

Voorhees, E. M. 1999. The TREC-8 question answering track report. In *TREC*.

Wadden, D.; Lin, S.; Lo, K.; Wang, L. L.; van Zuylen, M.; Cohan, A.; and Hajishirzi, H. 2020. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Welbl, J.; Stenetorp, P.; and Riedel, S. 2018. Constructing Datasets for Multi-hop Reading Comprehension Across Documents. In *Transactions of the Association for Computational Linguistics (TACL)*.

Wolfson, T.; Geva, M.; Gupta, A.; Gardner, M.; Goldberg, Y.; Deutch, D.; and Beran, J. 2020. BREAK It Down: A Question Understanding Benchmark. In *Transactions of the Association for Computational Linguistics (TACL)*.

Xiong, W.; Li, X. L.; Iyer, S.; Du, J.; Lewis, P.; Wang, W.; Mehdad, Y.; tau Yih, W.; Riedel, S.; Kiela, D.; and Oguz, B. 2021. Answering Complex Open-Domain Questions with Multi-Hop Dense Retrieval. In *International Conference on Learning Representations (ICLR)*.

Xu, Y.; Zhang, B.; Chen, Z.; and Wang, K. 2007. Privacy-Enhancing Personalized Web Search. In *Proceedings of the 16th international conference on World Wide Web (WWW)*.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; andRuslan Salakhutdinov, W. W. C.; and Manning, C. D. 2018. HOTPOTQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2369–2380.

Zhang, M. J.; and Choi, E. 2021. SituatedQA: Incorporating Extra-Linguistic Contexts into QA. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Zimmerman, S.; Thorpe, A.; Fox, C.; and Kruschwitz, U. 2019. Investigating the Interplay Between Searchers' Privacy Concerns and Their Search Behavior. In *Proceedings of the 42nd International ACM Conference on Research and Development in Information Retrieval (SIGIR)*.

8 Appendix

8.1 Experimental Details

Dense Retrieval We use the model implementations for the MDR (dense retriever) provided by Xiong et al. (2021). For the non-private experiments, we use the base retrieval algorithm; we extend the base implementation for the private-retrieval modes described in Section 3 and release our implementation. We construct the dense passage corpus using FAISS (Johnson, Douze, and Jégou 2017), and use exact inner product search as in the original implementation. For the HotpotQA-MDR experiments, we directly use the provided question encoder and passage encoder checkpoints in the code base. For training the CONCURRENTQA-MDR and Subsampled HotpotQA-MDR experiments (Section 5), we train the MDR model from scratch, finding the hyperparameters in Table 8 work best.

The MDR retriever is trained with a contrastive loss as in Karpukhin et al. (2020), where each query is paired with a (gold annotated) positive passage and m negative passages to approximate the softmax over all passages. We consider two methods of collecting negative passages: first, we use random passages from the corpus that do not contain the answer (random), and second, we use one top-ranking passage from BM25 that does not contain the answer as a hard-negative paired with remaining random negatives. We do not observe much difference between the two approaches for CONCURRENTQA-results (also observed in Xiong et al. (2021)), and thus use random negatives for all experiments. We hope to experiment with additional methods of selecting negatives for CONCURRENTQA in future work.

The number of retrieved passages per retrieval, k, is an important hyperparameter as increasing k tends to increase recall, but sacrifice precision. Using larger values of k is also less efficient at inference time. We use $k = 100$ for all experiments in the paper and Table 6 shows the effect of using different values of k on performance.

https://github.com/facebookresearch/multihop_dense_retrieval

k	Avg-PR	F1
$k = 1$	41.4	33.5
$k = 10$	55.9	44.7
$k = 25$	63.3	48.0
$k = 50$	68.4	50.4
$k = 100$	73.8	53.0

Table 6: Retrieval performance (Average Passage-Recall@k, F1) for $k \in \{1, 10, 25, 50, 100\}$ retrieved passages per hop using the retriever trained on HotpotQA for OOD CONCURRENTQA test data.

k	F1
$k = 1$	22.0
$k = 10$	34.6
$k = 25$	37.8
$k = 50$	39.3
$k = 100$	40.8

Table 7: F1 score on the CONCURRENTQA test data for $k \in \{1, 10, 25, 50, 100\}$ retrieved passages per hop using BM25 sparse retrieval.

Learning Rate	5e-5
Batch Size	150
Maximum passage length	300
Maximum query length at initial hop	70
Maximum query length at 2nd hop	350
Warmup ratio	0.1
Gradient clipping norm	2.0
Training epoch	64
Weight decay	0

Table 8: Retrieval hyperparameters for MDR training on CONCURRENTQA and Subsampled-HotpotQA.

Finally, in our ablations with Contriever (Izacard et al. 2021), we also use the released checkpoints.

Sparse Retrieval For the sparse retrieval baseline, we use the Pyserini BM25 implementation using default parameters. We consider different values of $k \in \{1, 10, 25, 100\}$ per retrieval and report the retrieval performance in Table 7. We generate the second hop query by concatenating the text of the initial query and first hop passages.

QA Model We use the provided ELECTRA-Large reader model checkpoint from Xiong et al. (2021) for all experiments. The model was trained on HotpotQA training data. Using the same reader is useful to understand how retrieval quality affects performance, in the absence of reader modifications.

https://github.com/facebookresearch/contriever
https://github.com/castorini/pyserini

8.2 Additional Experimental Results

Figures In Table 9, we provide QA results for the CONCURRENTQA Dev split under the PAIR restrictions. The main paper includes Test results.

In Table 4, we provide examples of CONCURRENTQA that are successfully answered under each PAIR privacy restriction. Excitingly, we are able to answer many questions spanning public and private data, without sacrificing privacy. This ability has not previously been demonstrated with existing retrievers.

Next, Figure 4 shows that there is a clear separation between the relevance score distributions from the email vs. Wikipedia corpus for questions based on Wikipedia (public) passages, but this is is not the case for questions based on email passages. The relevance score distributions are not-necessarily well-aligned in the mixed-distribution retrieval setting, contributing to the difficulty and difference vs. zero-shot retrieval. We observe that for questions requiring private (email, red) documents, there are still several public (Wikipedia, blue) passages being selected.

Error Analysis of Retrieval Methods on CONCURRENTQA Here we include a qualitative discussion of representative errors observed for each retrieval method, corresponding to the results in Section 5.

Dense Retrievers First, HotpotQA-MDR appears biased towards Wikipedia passages. On examples where the gold Hop_1 passage is an email, 15% of the time, no emails appear in the top-k Hop_1 results; meanwhile, this only occurs 4% of the time for Hop_1 Wikipedia. On the slice of EE examples, 64% of Hop_2 passages are E, while on the slice of WW examples, 99.9% of Hop_2 passages are W. If we simply *force* equal retrieval from each domain on each hop, we observe up to 2.3 F1 points improvement on overall CONCURRENTQA. However, this is a heuristic choice and should be explored further in future work.

Performance on WE questions is notably worse than on EW questions. We hypothesize two reasons: (1) Wikipedia passages generally follow consistent structures, so it may be easier to retrieve Wikipedia passages in Hop_2 after retrieving Wikipedia in Hop_1, and (2) several emails discuss each Wikipedia-entity, which may increase the noise in Hop_2 (i.e., WE is a one-to-many hop, while for EW, W typically contains one valid entity-specific passage). The latter is intuitively because individuals owning private data truly care about a narrow set of public entities.

Sparse Retrievers We observe the sparse model often "cheats" by retrieving the Hop_2 passage, without the Hop_1 passage. For questions where BM25 retrieves the gold Hop_2 passage in the first hop, the score is 64.2 F1, and when this is not the case, the score is 18.3 F1. Next, given how CONCURRENTQA is constructed, i.e., crowdworkers see passages before writing questions, it may underevalue the skills dense models provide (e.g., fuzzy semantic matching) and overvalue direct matching, a strength of sparse methods. We observe several other benchmarks reported in (Thakur et al. 2021) on which BM25 outperforms dense retrieval, use similar annotation pipelines during question generation (e.g.,

Benchmark	Model	EM	F1
ConcurrentQA	No Privacy Baseline	49.3	55.8
	Multi-Index Baseline	49.3	55.8
	Document Privacy Baseline	38.6	45.0
	Query Privacy Baseline	19.1	23.9

Table 9: Multi-hop QA datasets using MDR under each privacy setting. Here we include results for the ConcurrentQA Dev split.

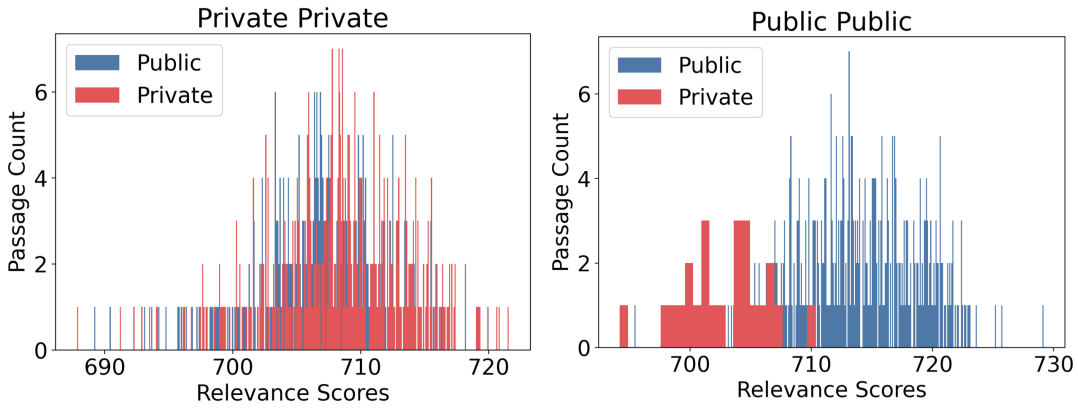

Figure 4: Number of passages retrieved in Hop_1 by relevance score, for each type of ConcurrentQA question, based on the gold supporting passage types.

Avg. words per question	28
Avg. words per Email passage	149
Avg. words per Wiki passage	44
Avg. words per answer	2

Table 10: Length statistics for ConcurrentQA.

Wadden et al. (2020); Yang et al. (2018)).

Single-Hop Baselines We observe the MS-MARCO fine-tuned Contriever model (Izacard et al. 2021), performs well with a single-hop over HotpotQA (Thakur et al. 2021), but leaves a large gap on ConcurrentQA. In error analysis, we find the model retrieves the first-hop passage 49% of the time and second-hop passage 25% of the time. By sub-domain, the Contriever model finds the gold first-hop passage for 85% of questions where both gold passages are from Wikipedia, but less than 39% when at least one gold passage (hop_1 or hop_2) is from Enron, suggesting a quality gap between the distributions. The gold passage is typically retrieved when the query consists of specific entities or phrases that occur simultaneously in very few passages.

8.3 Additional ConcurrentQA Analysis

Here we provide additional insight into the contents of ConcurrentQA and failure modes incurred by baseline retrieval methods on the benchmark.

Dataset	Size	Domain
WebQuestions	6.6K	Freebase
WebQSP	4.7K	Freebase
WebQComplex	34K	Freebase
MuSiQue	25K	Wiki
DROP	96K	Wiki
HotpotQA	112K	Wiki
2Wiki2MultiHopQA	193K	Wiki
Natural-QA	300K	Wiki
ConcurrentQA	18.4K	Email & Wiki

Table 11: Existing textual multi-hop benchmarks are designed over a single-domain.

First, to augment discussion of our motivation for creating ConcurrentQA, namely all existing textual multi-hop benchmarks require retrieving from a single domain, we include Table 11.

Size Statistics Table 2 gives ConcurrentQA size statistics. We provide statistics for the number of ConcurrentQA questions that require gold supporting passages from each set of privacy scopes. Note that the evaluation data is balanced in questions requiring two supporting emails, two Wikipedia passages, and one of each corpus.

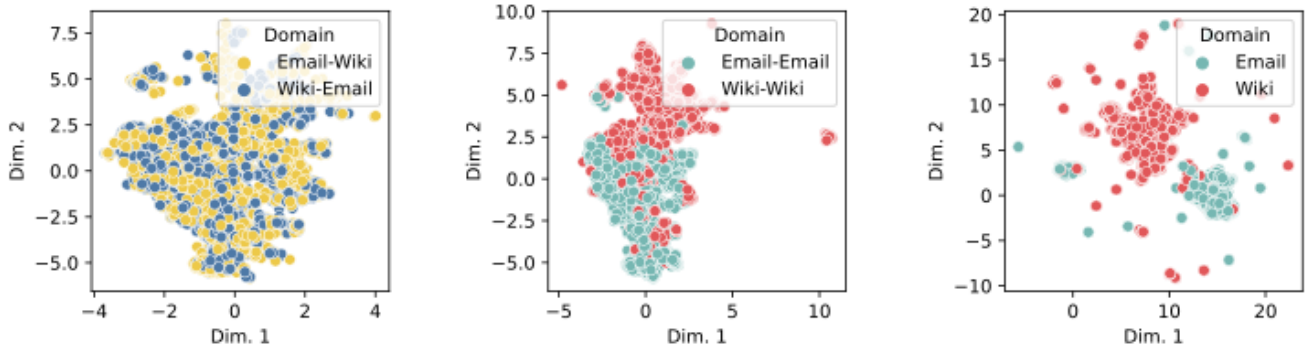

Figure 5: UMAP of BERT-base embeddings, using Reimers and Gurevych (2019), of CONCURRENTQA **questions** based on the domains of the gold passage chain to answer the question (left and middle). I.e., questions that require an Email passage for hop 1 and Wikipedia passage for hop 2 are shown as "Wiki-Email". Embeddings for all gold **passages** are also shown, split by domain (right).

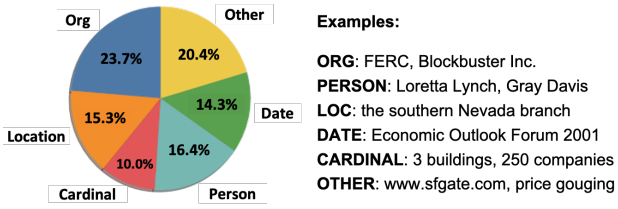

Figure 6: NER-types of CONCURRENTQA answers.

Analysis by Domain Next we further illustrate the question, passage, and answer types involved in CONCURRENTQA. First, in Table 12 we provide additional examples to illustrate examples of key linguistic features and question types (recall the three main reasoning types are bridge, attribute, and comparison questions). Next, Figure 5 (Left, Middle) shows the UMAP plots of CONCURRENTQA questions using BERT-base representations, split by whether the gold hop passages are both from the same domain (e.g., two Wikipedia or two email passages) or require one passage from each domain. We observe a separation in the clusters for Wiki-based vs. email-based questions and passages. Next, Table 10 shows the average length for each component of CONCURRENTQA. Note that Wiki passages tend to be longer than email passages. Finally, we focus on factoid QA, in which answers are short spans of text in the retrieved passages, for example containing common nouns, entities, or properties. Figure 6 shows the Named Entity Recognition (NER) type distribution for answers in CONCURRENTQA.

8.4 CONCURRENTQA Details

CONCURRENTQA is collected by showing crowdworkers pairs of passages and asking them to submit a question that requires reasoning over all the passages. Passages should be related to yield meaningful questions. We use the insight that we can obtain meaningful passage-pairs by showing workers passages that mention similar or overlapping entities. Entity tags are readily available in Wikipedia, however for private emails we 1) collect candidate entities with the SpaCy NER tagger, 2) split the full set into candidate public and candidate private entities by identifying Wikipedia linked entities amongst the spans tagged by the NER model, using the open-source SpaCy entity-linker, and 3) post-process the entity lists. For bridge questions, we present pairs of passages that mention the same entity.

For comparison questions, we present pairs of passages each mentioning a different entity of the same *type*. Wikidata types are readily available for public entities and we heuristically assign types to private entities. We release our code for preprocessing, annotating, filtering, and deduplication along with the benchmark.

https://spacy.io/

Example 1: shows how a question can be answered by an alternate retrieval path than the gold path. The *Alternate Hop 1* passage also depicts typos which are more prevalent in Enron compared to Wikipedia passages.

Multi-hop Question Reliant Energy is based in a city located in which Texas county?

Gold Hop 1 (Email) **Reliant Energy of Houston**, another company that resisted demands for business records, on Wednesday signed a confidentiality agreement with Dunn's committee and will begin bringing 250,000 documents to a depository in Sacramento, said Reliant spokesman Marty Wilson. Dunn said other companies have begun to deliver documents to Sacramento, but not all are fully complying with subpoenas. ...

Alternate Hop 1 (Email) ... "Independent power generators have come under increasing scrutiny and are being investigated by the state's Attorney General Bill Lockyer's office fo r gaming the market." Generators are being investigated as to whether they hav e shut plants for maintenance in order to spike prices during peak periods an d periods when the California Independent System Operator declares alerts whe n power reserves drop below certain levels in the state. **Reliant Energy is based in Houston, Texas**....

Example 2: shows how the same email passage can cover multiple topics. In contrast to Wikipedia, where passages are about a single entity, other types of documents including emails can cover many topics in the same passage. Thus, the single dense embedding generated per passage in retrieval methods such as DPR may not be as effective. This is a *bridge* question.

Multi-hop Question How much power can the company reported on October 1 2001 to be in talks to acquire an Indian Enron stake generate?

Gold Hop 1 (Email) World Watch The Wall Street Journal, 10 01 01 INDIA: Panel suggests Indian govt pay in Enron row-paper. Reuters English News Service, 10 01 01 INDIA: **Tata Power said in talks to buy India Enron stake**. Reuters English News Service, 10 01 01 Greece Awards 4 Electricity Production, 8 Supply Permits ... Portland Oregonian, 09 29 01 Firms Push Edison Near Bankruptcy Energy...

Gold Hop 2 (Wiki) The Tata Power Company Limited is an Indian electric utility company based in Mumbai, Maharashtra, India and is part of the Tata Group. The core business of the company is to generate, transmit and distribute electricity. With an installed **electricity generation capacity of 10,577MW**, it is India's largest integrated power company. At the end of August 2013, its market capitalisation was $2.74 billion.

Example 3: shows an example requiring *list-based reasoning*. This occurs in several benchmark questions. This is a *bridge* question.

Multi-hop Question The seven economic commentators at Economic Outlook Forum 2001 were Ben Hermalin's co-chair, Severin Borenstein, Jerry Engel, Rich Lyons, Ken Rosen, Janet Yellen, and a professor who was born in what year?

Gold Hop 1 (Email) Dear Haas Evening MBA Students, On Friday afternoon November 9, 2001, some of the School's most distinguished economists and I will participate in a "teach-in" about the US economy. "Economic Outlook Forum 2001" will examine ... Professor Hermalin will moderate the panel presentations and following discussion. In addition to myself, our **economic commentators will be Professors Severin Borenstein, Jerry Engel, Rich Lyons, Ken Rosen, Hal Varian, and Janet Yellen**...

Gold Hop 2 (Wiki) Hal Ronald Varian (**born March 18, 1947** in Wooster, Ohio) is an economist specializing in microeconomics and information economics. He is the chief economist at Google and he holds the title of emeritus professor at the University of California, Berkeley where he was founding dean of the School of Information. He has written ...

Table 12: Illustrative examples of properties of CONCURRENTQA.

Example 4: shows an example of an *attribute* style question, in which both passages provide an attribute about the same entity (i.e., "Idealab!").

Multi-hop Question Funding Metiom filed for Chapter 11 after investors backed out of which company founded by Bill Gross in 1996?

Gold Hop 2 (Email) ... DigiPlex Raises $48 Million Equity, $35 Million Debt STSN Gets $66.5M of Series D Debt and Equity Tribune Media Services Takes Majority Stake in TVData Viator Closing $5M to $10M Series C Round in Next Two Weeks bad news WorkingWoman.com Lays Off 63%; Looking for Buyers, Funding **Metiom Files for Chapter 11 after Investors Back Out Idealab!**

Gold Hop 2 (Wiki) **Idealab was founded by Bill Gross (not the same Bill Gross of PIMCO) in March 1996.** Prior to Idealab, he founded GNP Loudspeakers (now GNP Audio Video), an audio equipment manufacturer; GNP Development Inc., acquired by Lotus Software; and Knowledge Adventure, an educational software company, later acquired by Cendant...

Example 5: shows an example of a *yes-no* style question. For these questions (a subset of the comparison questions), the answer is not a span in the passages.

Multi-passage Question Did the company who appointed Carol S. Schmitt as vice president secure all of its expected first round of funding?

Passage 1 ... Fabless Semiconductor Firm Secures $8.2 Million in Round One AGOURA HILLS, Calif. – Internet Machines, a fabless semiconductor company that develops software and services for data communications markets, said it secured $8.2 million in its first round of funding. ... Management App Firm Gets $5 Million of $8 Million Round One CAMBRIDGE, Mass. – **Bluesocket, which develops management software for Bluetooth-enabled networks, said it secured $5 million of its expected $8 million first round of funding** from St. Paul Venture Capital and Osborn Capital.

Passage 2 ... **Bluesocket, which develops security and management products for wireless local area networks, said it appointed Carol S. Schmitt as vice president** of business development. Prior to joining the company, Ms. Schmitt was a business and market development consultant in Los Gatos, Calif. Bluesocket is backed by Osborn Capital and ...

Example 6: shows an example of a non yes-no *comparison* style question. For these questions, the answer is the one of the two entities being compared, where one entity appears in each passage.

Multi-passage Question Which company out of Regency Capital and StellaService started its business operations first?

Passage 1 ... NEW YORK (VENTUREWIRE) – Privacy Protection, which does business as Eprivex.com and is a developer of electronic privacy technology and persona l privacy protection services, said it must cease operations unless it can complete its seed round of $1.5 million, wholly or incrementally, from individual or private investors. **The company, which was founded in March 2000, has received prior financing from individual investors including Roge r Dietch, founder of Regency Capital,** as well as from Jesse L. Martin, Jerry Orbach, and Sam Waterston, all of whom are actors on the NBC television sho w Law and Order.

Passage 2 StellaService Inc. is a privately held American information and measurement company with headquarters in New York City (USA). The company measures and rates the customer service performance of online companies in a process audited by global accounting and auditing firm KPMG. **Founded in 2009,** it produces both Stella Metrics (a mystery shopping platform) and Stella Connect (a customer feedback system).

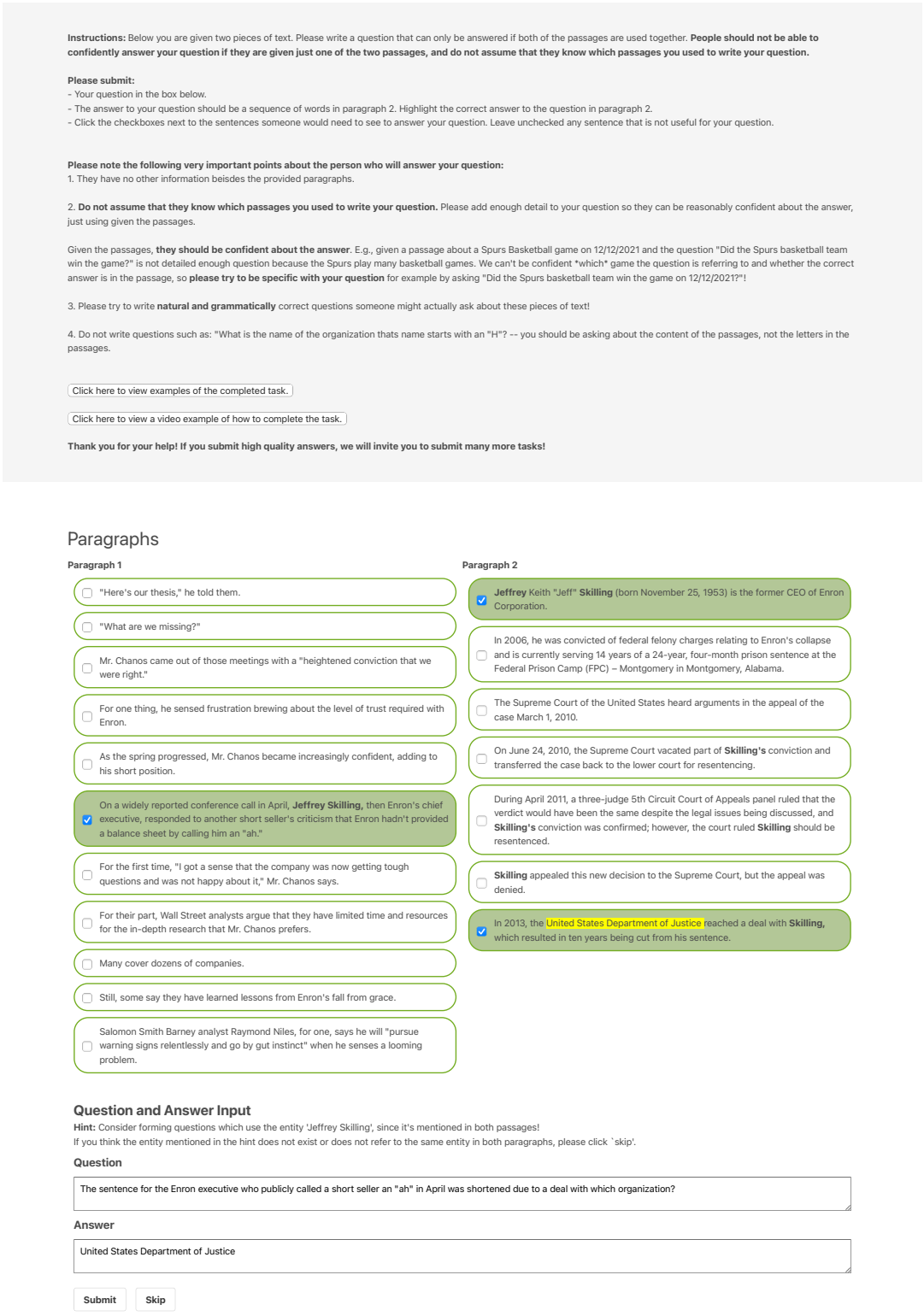

Figure 7: Mechanical Turk interface for CONCURRENTQA data collection. Crowdworkers select checkboxes for supporting passages, highlight the answer span, and write the question in the text box.