

From Proprietary to Public: Dataset Declassification for Question Answering

Stefano Campese¹ Alessandro Moschitti² Ivano Lauriola²

¹University of Trento ²Amazon AGI

stefano.campese@unitn.it, {amosch, lauivano}@amazon.com

Abstract

Training Question Answering (QA) systems on real customer interaction data is often infeasible to reproduce or share due to privacy and intellectual property constraints. We introduce dataset declassification, a framework that replaces proprietary QA datasets with public, semantically equivalent counterparts that preserve training and evaluation utility while satisfying typical legal and privacy requirements. The approach maps proprietary questions to public equivalents via large-scale retrieval and reconstructs answers from domain-matched corpora using LLM-based annotation. Experiments on four QA benchmarks show that models trained and evaluated on declassified datasets achieve performance within $\Delta \approx 0$ of the original data. The method targets datasets whose underlying information needs already exist in public knowledge sources (e.g., knowledge available on the web and used by public-facing services). We further provide explicit applicability conditions showing that declassification succeeds when sufficient semantic overlap exists between proprietary and public corpora, and degrades predictably when it does not.

1 Introduction

Modern machine learning research increasingly relies on large, high-quality Question Answering (QA) datasets derived from real user interactions. In practice, however, such datasets are almost always proprietary and sensitive, creating a persistent gap between industrial research and open scientific progress. Access to customer data inside organizations is constrained by strict governance policies, restricted execution environments, and limitations on data movement. Even when internal research is possible, the resulting work often cannot be reproduced or validated externally because the underlying datasets cannot be released.

In industrial settings, proprietary QA datasets fall into two regimes. The first (**Type A**) involves

enterprise-specific QA, where the underlying concepts themselves are private, such as internal procedures, proprietary products, or confidential entities with no public equivalent. The second (**Type B**) arises in public-facing systems such as search engines, voice assistants, and LLM-based assistants. Here, the underlying knowledge is already available on the web, while what remains proprietary is the *distribution of user interactions*: which questions are asked, how frequently they occur, and how they are paired with answers and labels. This work targets the Type B regime, where the information needs are public but the dataset structure is proprietary. This is a common industrial need: typical scenarios include publishing an internal benchmark to enable external research collaboration, sharing data with vendors for labeling under data governance constraints, and releasing evaluation data to make published industrial results externally reproducible. We provide an extended discussion of this distinction and its implications in Appendix F.

Existing data-release strategies do not resolve this tension. Traditional anonymization focuses on removing Personally Identifiable Information (PII), but privacy risks in natural language extend beyond explicit identifiers (Yermilov et al., 2023). Seemingly innocuous queries may still reveal sensitive information through contextual cues. For instance, a query such as “*where can I buy medication for my veterinary clinic near street X and landmark Y*” may uniquely identify a business in a small geographic area. As a result, anonymization is often either too weak to guarantee privacy or so aggressive that it destroys the semantic content required for QA tasks.

A natural alternative is to release only those examples that already appear in public sources. However, overlap between proprietary and public questions is primarily *semantic rather than lexical*: equivalent questions are typically expressed differently. Exact-match filtering therefore removes most

examples even when semantically equivalent public content exists, producing datasets that are both drastically smaller and distributionally distorted.

We propose *dataset declassification*, grounded in a simple principle: content already available in the public domain is generally not subject to privacy or confidentiality restrictions. Rather than modifying proprietary data, declassification replaces it entirely with semantically equivalent public content. Given a proprietary question-answer pair (q, a) , we seek a public pair (q', a') that induces similar training and evaluation behavior for downstream models.

We operationalize this idea through a practical pipeline with two steps. First, *question retrieval* maps proprietary questions to semantically equivalent public questions using large-scale retrieval and cross-encoder reranking. Second, *answer reconstruction* derives corresponding answers from domain-matched public corpora, preserving task structure and label distributions without generating new content. The result is a fully public dataset that is semantically aligned with the original while remaining content-disjoint from it.

Experiments on four QA benchmarks show that models trained and evaluated on declassified data achieve accuracy differences close to zero relative to the original datasets. We further establish explicit *applicability conditions*: declassification succeeds when sufficient semantic overlap exists between proprietary and public corpora and degrades predictably when this overlap is insufficient.

Finally, since reconstructed datasets are derived from public sources, they inherit the same potential exposure to large-scale LLM pretraining as other public benchmarks. This does not introduce new contamination risks beyond those already present in web-scale models. Instead, declassification enables a practical compromise: organizations can release a public *shadow dataset* for reproducible research while retaining the original proprietary benchmark internally for controlled evaluation.

2 Related Work

Releasing or comparing QA datasets derived from real user interactions is constrained by privacy, proprietary content, and benchmark contamination. While many approaches aim to protect sensitive data, none directly address the practical need to release or share QA datasets while preserving their training and evaluation behavior.

A common approach is **data sanitization** via anonymization or PII removal (Yermilov et al.,

2023). Although effective at removing explicit identifiers, PII redaction often degrades QA performance because entities frequently encode essential semantics (Pal et al., 2024). More fundamentally, privacy risks in natural language often arise from contextual combinations rather than explicit identifiers, making PII-based definitions incomplete in practice. Differential privacy offers formal guarantees (Dwork and Lei, 2009), and DP-based text sanitization rewrites private text under formal noise budgets (Yue et al., 2021; Chen et al., 2023). However, the noise required for text typically destroys semantic coherence, rendering the data unusable for QA. Our framework is not a formal privacy mechanism: rather than modifying the private text, it replaces the dataset entirely with pre-existing public content, using the private data only for semantic matching.

Synthetic data generation uses models to create artificial QA pairs that approximate real data distributions (Bai et al., 2024; Kotschenreuther, 2024). While useful for data augmentation, synthetic datasets often fail to preserve the linguistic diversity, reasoning complexity, and difficulty distribution of real QA benchmarks, and require extensive human validation. Crucially, they do not support scenarios where the distributional properties of specific real datasets must be preserved.

Paraphrasing and backtranslation methods (Liu et al., 2020; Team et al., 2022) modify surface form while preserving meaning. These techniques are effective for augmentation but insufficient for privacy: transformed content remains semantically equivalent to the original and can often be reconstructed or recognized.

Knowledge distillation transfers capabilities from models trained on private data to public students, but does not produce releasable datasets and offers weak privacy guarantees. Distilled models have been shown to retain significant information about training examples, remaining vulnerable to membership inference attacks (Cui et al., 2026).

Finally, **benchmark contamination**, where evaluation data appears in training corpora, undermines the validity of QA evaluation (Sainz et al., 2023; Balloccu et al., 2024). Existing mitigations, such as temporal splits or dynamic benchmarks, reduce risk but do not enable the release of realistic datasets.

Closest to our work are retrieval methods that reuse public content instead of generating new text. Prior work retrieves public sentences to augment private QA data (Arora et al., 2023) or rewrites

private documents into synthetic equivalents using LLMs (Bai et al., 2024; Yue et al., 2023). However, these approaches do not replace entire datasets with public counterparts, do not preserve question-level equivalence, and do not support releasing benchmarks with reliable evaluation utility. Our work differs from simple filtering or overlap detection in that it *reconstructs* a complete dataset from independent public sources and *validates* preservation of dataset-level training and evaluation properties.

3 Problem Formulation

We study the problem of dataset declassification for QA systems. A QA dataset is modeled as a collection of triplets $\mathcal{D} = \{(q_i, c_i, a_i)\}_{i=1}^N$, where q_i is a question, c_i is its associated context, and a_i is an answer label. The context captures task-specific information that determines the semantics or disambiguation of the question, such as previous dialog turns, supporting passages, or candidate answer sets. The answer a_i may represent a correct response (e.g., in extractive or generative QA) or a correctness label (e.g., in multiple-choice or verification tasks).

3.1 Dataset Declassification

Given a private dataset $\mathcal{D}$, the goal of declassification is to construct a new dataset $\mathcal{D}'$ that: (i) contains no private or sensitive content, and (ii) preserves the utility of the original dataset for both training and evaluation. Formally, declassification is defined as a mapping $\Phi(\mathcal{D}) = \mathcal{D}'$, where $\mathcal{D}'$ consists exclusively of public content and is safe to release and share. While Φ may in principle operate on groups or structures of examples, in this work we focus on *pointwise declassification*, which enables scalable and controllable transformations.

3.2 Pointwise Declassification

We decompose the dataset-level mapping Φ into a pointwise function that operates on individual QA triplets: $\phi(q, c, a) = (q', c', a')$, which maps a private example to a declassified one. The dataset-level mapping is then obtained by aggregation: $\Phi(\mathcal{D}) = \{\phi(q_i, c_i, a_i)\}_{i=1}^N = \mathcal{D}'$.

This formulation acknowledges that more complex transformations, such as clustering or jointly declassifying groups of related questions, are possible. We deliberately restrict our scope to pointwise mappings for simplicity, robustness, and practical deployment. In our setting, ϕ is further decomposed into question, context, and answer mappings:

$\phi(q, c, a) = (\phi_Q(q), \phi_C(c), \phi_A(a))$, where each component is declassified independently while preserving semantic compatibility.

3.3 Utility Preservation

Utility preservation is defined with respect to a downstream model m and a performance metric $p_m(\cdot)$. We consider two complementary notions of utility.

Evaluation utility. Let $\mathcal{T}$ be an original test set and $\mathcal{T}' = \phi(\mathcal{T})$ its declassified counterpart. The evaluation utility loss is defined as $\Delta_{\text{test}} = p_m(\mathcal{T}) - p_m(\mathcal{T}')$. We say that declassification preserves evaluation utility if, for given $\delta > 0$ and $\epsilon > 0$, $\Pr(|\Delta_{\text{test}}| > \delta) < \epsilon$, where $\Pr$ indicates probability taken over draws of examples from the dataset.

Training utility. Let $\mathcal{T}_r$ be a training set and $\mathcal{T}'_r = \phi(\mathcal{T}_r)$ its declassified version. Let m be a model trained on $\mathcal{T}_r$ and m' a model trained on $\mathcal{T}'_r$. Both models are evaluated on the same test set $\mathcal{T}$. The training utility loss is defined as $\Delta_{\text{train}} = p_m(\mathcal{T}) - p_{m'}(\mathcal{T})$. Declassification preserves training utility if $\Pr(|\Delta_{\text{train}}| > \delta) < \epsilon$.

Together, these definitions formalize the requirement that declassified datasets behave equivalently to their private counterparts for both model evaluation and model training, up to a controlled and statistically bounded deviation.

4 The Declassification Framework

This section describes the concrete implementation of our declassification framework, built on two foundational design choices: retrieval over generation, and pointwise decomposition into question and answer mappings.

Retrieval over generation. A pointwise declassification function ϕ could in principle be implemented by prompting a large language model to generate a paraphrase that obfuscates sensitive content. However, we argue that unconstrained generation is unsuitable for declassification. Generated text is not guaranteed to remove all sensitive information, may hallucinate new private details, and relies on models whose training data and memorization behavior are opaque. Even rare failures are unacceptable in privacy-critical settings. Additionally, our experiments show that LLM-based approaches do not work effectively and are largely outperformed by retrieval-based methods.

For this reason, rather than generating new content, our declassification retrieves semantically

equivalent questions and answers from corpora of public data. This ensures that all declassified content is public by construction, eliminating privacy risk while enabling systematic reuse of existing resources.

Scope and assumptions. Our framework decomposes declassification into two components: *question declassification*, which replaces a private question with a semantically equivalent public one, and *answer declassification*, which reconstructs answers from public corpora while preserving task difficulty and label distributions. The pipeline schema is illustrated in Appendix G.

In this work, we assume that the semantics of each example is fully determined by the question alone, and that no additional private context (e.g., dialog history or user metadata) is required. This assumption holds for many industrial QA systems, including FAQ systems, knowledge-base search, open-domain search queries, and fact-seeking interactions in public assistants. Extending declassification to richer contextual settings (e.g., multi-turn dialog, document-grounded QA with private documents) is an important direction for future work discussed in Section 7.

4.1 Question Declassification

Let $\mathcal{Q}$ and $\mathcal{Q}'$ denote the distributions of private and public questions, respectively. Question declassification is defined as a mapping $\phi_Q : \mathcal{Q} \rightarrow \mathcal{Q}'$, which replaces a private question with a semantically equivalent public one.

We implement ϕ_Q using a retrieval-based pipeline. Given a private question, the system queries an index of public questions and selects the most semantically similar candidate. We adopt the question retrieval framework introduced by Campese et al. (2023), which consists of: (i) a large index of public question-answer pairs, (ii) a dense retriever to retrieve top- k candidates, and (iii) an optional cross-encoder reranker to refine the final selection. The index contains approximately 38M public QA pairs from WikiHow, Natural Questions, and PAQ (Lewis et al., 2021). Dense retrieval is performed using a MiniLM-based sentence encoder (Reimers and Gurevych, 2019), fine-tuned for question ranking, followed by a DeBERTa-v3-base cross-encoder trained for question reranking.

Because all retrieved questions originate from known public sources, question declassification guarantees that no proprietary content is leaked. The cross-encoder reranker improves mapping pre-

Algorithm 1 Answer Declassification. The LLM classifier at line 8 is Qwen-3 235B.

Input: Public corpus $\mathcal{C}$, mapped question q' , original answers $\{(a_j, y_j)\}$

Output: Declassified answers $\{(a'_j, y'_j)\}$

```

1: Retrieve top 500 passages from  $\mathcal{C}$  for  $q'$  using a dense
   passage retriever.
2: Segment passages into a sentence pool  $\mathcal{S} = \{s_1, s_2, \dots\}$ .
3: for each original answer  $(a_j, y_j)$  do
4:   Rank sentences  $s_i \in \mathcal{S}$  by semantic similarity to  $a_j$ .
5:   Let  $\mathcal{S}_K$  be the top  $K = 200$  ranked sentences.
6:    $found \leftarrow \text{FALSE}$ 
7:   for  $s_i \in \mathcal{S}_K$  in ranked order do
8:     Use LLM to classify  $s_i$  as correct or incorrect for
        $q'$ , yielding  $\hat{y}_i$ .
9:     if  $\hat{y}_i = y_j$  then
10:       $a'_j \leftarrow s_i, y'_j \leftarrow y_j, found \leftarrow \text{TRUE}$ 
11:     break
12:   end if
13: end for
14: if not  $found$  then
15:    $a'_j \leftarrow s_1, y'_j \leftarrow \hat{y}_1$   $\triangleright$  Fallback to most similar
16: end if
17: end for
```

cision: removing it and selecting the top-1 dense retrieval candidate increases the training utility delta from -0.5 to -0.8 on WikiQA and from -0.2 to -1.2 on TrecQA, while still outperforming the generative baselines.

4.2 Answer Declassification

Replacing the question alone is insufficient: answers associated with private questions may themselves contain sensitive information or become invalid once the question is replaced. Answer declassification reconstructs answers for the mapped question q' using public corpora, while preserving task structure and difficulty.

We consider datasets where answers are expressed as a single sentence. This simplification covers most traditional QA settings, including open-domain QA and Answer Sentence Selection (AS2) (Garg et al., 2020). AS2 is used to implement extractive QA by reranking retrieved candidate sentences and selecting the top-ranked sentence as the final answer.

Declassification process. Given a mapped question q' , answer declassification constructs a public replacement for each original answer sentence while preserving label distribution and answer difficulty. The procedure consists of three steps: passage retrieval, sentence segmentation, and answer replacement.

Passage Retrieval. Given the mapped question q' , we retrieve a set of candidate passages from a public corpus $\mathcal{C}$ using a dense passage retriever.

In our experiments, we retrieve the top 500 passages using BGE-base-v1.5 (Xiao et al., 2024). The choice of $\mathcal{C}$ depends on the dataset being declassified (e.g., CC-News¹ for TrecQA, Wikipedia for WikiQA²).

Sentence Segmentation. Retrieved passages are segmented into sentences via NLTK (Bird and Loper, 2004), yielding a candidate pool $\mathcal{S} = \{s_1, s_2, \dots\}$.

Answer Replacement. For each original answer a_j described by a sentence and a binary label $a_j = (s_j, y_j)$, we rank sentences in $\mathcal{S}$ by semantic similarity to a_j , using the MPNet model (Song et al., 2020). We then iterate through the top K candidates ($K = 200$) in ranked order: for each candidate $s_i \in \mathcal{S}$, we use an LLM-based classifier (Qwen-3 235B) to determine whether s_i correctly answers q' . If the predicted label $\hat{y}_i$ matches the original label y_j , we select s_i as the replacement and stop (early stopping). If no label-consistent sentence is found within the top K , we fall back to the most similar sentence s_1 and use the LLM’s predicted label $\hat{y}_1$.

This early stopping mechanism makes annotation efficient: we only call the LLM until we find a match, rather than annotating all candidates. The procedure approximates the original label distribution as well as the semantic and difficulty characteristics of the original answers, while ensuring that all reconstructed answers originate from public data. Algorithm 1 summarizes this process.

5 Experiments

We evaluate dataset declassification in a fully controlled setting using only public benchmarks. Since proprietary datasets cannot be shared, we simulate declassification by treating public QA datasets as if they were private and constructing declassified counterparts. This enables a direct and reproducible measurement of whether declassified data preserves evaluation behavior (test utility) and learning signal (training utility).

5.1 Setup

Tasks and datasets. We study two QA settings across four benchmarks.

Answer Sentence Selection (AS2): This is retrieval-based QA, which considers top- k candidate answers from a retrieval system, reranks them, and selects the top-one as the final response.

We use WikiQA (Yang et al., 2015) and TrecQA (Wang et al., 2007): both consist of open-domain questions paired with candidate answer sentences labeled as correct or incorrect. WikiQA contains 2,118/296/237 train/dev/test questions derived from Bing logs, with one correct sentence from Wikipedia and multiple incorrect sentences per question. TrecQA contains 94/65/68 train/dev/test factoid questions from TREC QA tracks, with news-domain answers.

Open-ended QA (parametric): We use OpenBookQA (Mihaylov et al., 2018) and SimpleQA (Wei et al., 2024). OpenBookQA contains 500 science questions; although originally multiple-choice, we evaluate it in an open-ended configuration. SimpleQA contains 4,326 deliberately obscure factual questions designed to challenge LLMs. It serves as a controlled stress test for the limits of retrieval-based declassification.

Models. For AS2, we use DeBERTa-v3-base 184M (He et al., 2023) as the main cross-encoder ranker, with MiniLM-L12 33M (Wang et al., 2020) reported as an additional model. The ranker scores each candidate sentence given the question as $[CLS] \text{ question } [SEP] \text{ sentence } [EOS]$ and ranks candidates by score. Training configurations and hyperparameters are in Appendix B. For open-ended QA we evaluate Phi-3 3.8B (Abdin et al., 2024a), Ministral 8B (Mistral AI, 2024), Phi-4 14.8B (Abdin et al., 2024b), and Qwen-3 30B (Qwen Team, 2025) in zero-shot settings.

Evaluation. Our main metric is answer accuracy: for AS2, labeled sentences are ranked and we use Precision@1 to measure success. For generative QA (OpenBookQA and SimpleQA), we use the same LLM-based judge and prompt from SimpleQA (Wei et al., 2024), fed with the reference answer of each target dataset, where unanswered or wrongly answered questions are counted as incorrect. Prompts are provided in Appendix A.

Declassification configurations. We compare: (i) Original data (Private), (ii) Generative declassification (Gen-PR, Gen-BT), and (iii) Retrieval-based declassification (Map).

Gen-PR paraphrases questions and answer sentences via Qwen-3 (80B), applied sentence-by-sentence. **Gen-BT** uses backtranslation (EN→DE→EN) via NLLB-200 (Team et al., 2022), also applied sentence-by-sentence. **Map** is our retrieval-based declassification (Section 4), which implements question declassification ϕ_Q and

¹<https://huggingface.co/datasets/stanford-oval/ccnews>

²<https://huggingface.co/datasets/olm/wikipedia>

answer declassification ϕ_A via retrieval from public corpora. End-to-end computational costs are reported in Appendix C.

Data	Model	Private	Gen-PR	Gen-BT	Map
Extraction-based QA (AS2)					
WikiQA	DeBERTa	81.4 $\pm$ 0.8	+5.1 $\pm$ 0.8	+4.6 $\pm$ 0.8	-0.8 $\pm$ 1.8
	MiniLM	70.6 $\pm$ 2.5	+3.6 $\pm$ 2.5	+2.6 $\pm$ 2.5	-2.0 $\pm$ 3.3
TrecQA	DeBERTa	85.8 $\pm$ 0.8	+1.4 $\pm$ 1.1	+0.5 $\pm$ 1.1	-0.5 $\pm$ 1.1
	MiniLM	72.5 $\pm$ 1.3	+0.5 $\pm$ 1.3	-1.5 $\pm$ 1.3	-0.6 $\pm$ 1.9
Generative-based QA					
OpenBookQA	Phi-3	53.5	-0.6	-0.9	0.0
	Ministral	49.2	-1.6	-5.0	-0.4
	Phi-4	59.8	-2.4	-3.0	0.0
	Qwen-3	60.8	-0.5	-0.2	-0.2
SimpleQA	Phi-3	4.0	-0.8	+0.8	+11.6
	Ministral	2.7	-0.4	-0.4	+8.4
	Phi-4	12.55	-0.9	+1.1	+12.0
	Qwen-3	18.5	-3.3	-4.1	+5.8

Table 1: **Evaluation utility** on declassified test sets. Private shows absolute accuracy; other columns show Δ relative to Private. $\pm$ denotes Std. Dev. across 3 runs (AS2 only).

5.2 Evaluation Utility (Declassify Test Only)

We first test whether declassified test sets preserve evaluation validity. Models are trained on the original training data and evaluated on (i) the original test set and (ii) the declassified test set.

Table 1 reports the results for each dataset and model. Column Private reports answer accuracy on the original data, while columns Gen-PR, Gen-BT, and Map report Δ , the difference between the accuracy on Private and on the three approaches, respectively.

We observe that: (i) for AS2, retrieval-based declassification (Map) yields very low Δ , outperforming generative methods; this indicates that declassified test sets preserve evaluation validity; (ii) Gen-BT performs comparably to Map on TrecQA, but is not applicable in practice as it carries a high risk of exposing private data; (iii) Map again outperforms Gen-PR and Gen-BT on the generative task OpenBookQA, with Δ close to 0; (iv) finally, SimpleQA behaves differently: Map produces positive deltas (+6 to +12 points), indicating a *difficulty shift* rather than model improvement. This behavior is expected and analyzed in Section 6.2.

5.3 Training Utility (Declassify Train Only)

We next measure whether declassified *training* data provides comparable learning signal. We train AS2 rankers on declassified training sets and evaluate them on the original (private) test set. Table 2 reports the performance delta. Since generative QA is zero-shot (no training), the table only shows AS2 results.

Retrieval-based declassification preserves train-

Data	Model	Private	Gen-PR	Gen-BT	Map
WikiQA	DeBERTa	81.4 $\pm$ 0.8	-4.1 $\pm$ 2.1	-8.9 $\pm$ 1.8	-0.5 $\pm$ 1.8
	MiniLM	70.6 $\pm$ 2.5	-3.9 $\pm$ 2.7	-5.9 $\pm$ 2.1	-0.32 $\pm$ 2.2
TrecQA	DeBERTa	85.8 $\pm$ 0.8	-29.1 $\pm$ 6.4	-4.9 $\pm$ 2.1	-0.2 $\pm$ 0.7
	MiniLM	72.5 $\pm$ 1.3	-25.0 $\pm$ 16.4	-2.1 $\pm$ 1.4	-0.4 $\pm$ 1.4

Table 2: **Training utility** on AS2. Models are trained on declassified training sets and evaluated on the original test sets. Values are Δ relative to Private.

ing utility on both datasets ($|\Delta| \leq 2$), substantially outperforming generative declassification. Gen-PR is especially brittle on TrecQA ($\Delta = -29$ for DeBERTa), consistent with semantic drift or entity corruption that breaks alignment between questions and labeled answer sentences. These results support the claim that retrieval-based question mapping combined with domain-matched answer reconstruction best preserves the learning signal.

5.4 Benchmark Declassification (Declassify Train and Test)

Finally, we evaluate full benchmark declassification, where both training and test sets are declassified, and models are trained and evaluated entirely on declassified data. This setting corresponds to releasing a fully declassified benchmark intended to replace the original dataset.

Data	Model	Private	Gen-PR	Gen-BT	Map
WikiQA	DeBERTa	81.4 $\pm$ 0.8	-1.2 $\pm$ 0.8	-3.6 $\pm$ 0.9	-0.4 $\pm$ 1.8
	MiniLM	70.6 $\pm$ 2.5	-3.5 $\pm$ 2.5	-2.8 $\pm$ 2.5	-1.5 $\pm$ 3.3
TrecQA	DeBERTa	85.8 $\pm$ 0.8	-17.7 $\pm$ 0.8	-4.9 $\pm$ 0.8	-0.2 $\pm$ 1.2
	MiniLM	72.5 $\pm$ 1.3	-26.4 $\pm$ 1.3	-2.5 $\pm$ 1.4	-0.4 $\pm$ 1.9

Table 3: **Benchmark declassification** on AS2. Models are trained and evaluated on declassified data. Values are Δ relative to the private benchmark.

Table 3 shows that retrieval-based declassification yields $|\Delta| < 0.4$ for DeBERTa and < 1.5 for MiniLM in the full benchmark setting. This result enables the use of fully public “shadow benchmarks” that preserve both training and evaluation behavior. Generative baselines degrade substantially on TrecQA, confirming that small semantic and entity perturbations compound when both training and evaluation are declassified.

5.5 Proof of Concept on Proprietary Data

To validate declassification beyond simulated settings, we applied our framework to an internal dataset of 436 proprietary open-domain queries evaluated against a non-public commercial search engine. For each original and declassified query, we submit the query to the same system and ask expert annotators whether the returned document is relevant. Accuracy is the percentage of relevant results, and we report the absolute change $|\Delta|$

with respect to the original benchmark. Further details about the system and annotation procedure are omitted according to internal policy.

In addition to Map, Gen-BT, and random substitution, we consider two text-sanitization baselines (Yue et al., 2021; Chen et al., 2023): *PII removal* and *PII entity replacement*. Presidio³ detects PII in 82/436 queries (18.8%). PII removal discards them, leaving 354 queries. Entity replacement substitutes detected entities with alternatives of the same type and similar popularity using Qwen-3 235B (e.g., ‘john belushi cause of death’ → ‘chris farley cause of death’). Replacement succeeds on 80 PII-containing queries, yielding a mixed set of 434 queries together with the 354 unchanged ones. Additional examples are reported in Appendix E.

Method	N	Acc	$ \Delta $
Original	436	89.0	—
Map (retrieval + reranking)	436	88.5	0.5
Gen-BT (backtranslation)	436	84.6	4.4
Random substitution	436	84.6	4.4
PII removal	354	89.8	0.8
PII entity repl. (mixed)	434	87.3	1.7
PII entity repl. (PII only)	80	76.2	12.7

Table 4: Evaluation utility on proprietary queries. Map closely preserves the original benchmark, whereas entity-level sanitization degrades utility on transformed queries.

Table 4 confirms the trend observed on public benchmarks: Map nearly preserves evaluation utility ($|\Delta| = 0.5$), whereas Gen-BT and random substitution degrade accuracy by 4.4 points. PII replacement appears competitive on the mixed set ($|\Delta| = 1.7$), but 354/434 queries (81.6%) are unchanged. On the 80 actually transformed queries, $|\Delta|$ increases to 12.7, showing that replacing individual entities can alter the underlying information need and therefore the relevant documents.

In contrast, Map replaces the private query as a whole with a semantically aligned public query. The resulting public benchmark reproduces the proprietary evaluation within 0.5 accuracy points and can be shared with third parties, e.g., for external labeling. Despite its limited scale, this experiment provides direct evidence that the behavior observed in our simulated setting transfers to genuinely proprietary data.

6 Analysis and Discussion

We observed that, for WikiQA and TrecQA, retrieval-based declassification produces questions

that closely preserve the original information need. In contrast, generative approaches tend to introduce unnecessary creativity, leading to loss of precision, while Gen-BT may introduce translation artifacts. We report a number of examples in Appendix E. The remainder of this section analyzes why our approach consistently outperforms generative baselines, the central role of semantic similarity, and the conditions under which declassification succeeds or fails.

6.1 Semantic Similarity and Equivalence

To quantify these observations, we measure the average semantic similarity between original and mapped questions, as well as between original and reconstructed answer sentences, using a public STS model (MPNet-base⁴). We also manually assess question-level semantic equivalence on a sample of question pairs. Table 5 reports question similarity (Q sim.), answer similarity (S sim.) for private and declassified examples, and the percentage of manually judged semantically equivalent question pairs (Q equiv.). The manual evaluation was conducted on a subset of 600 total question pairs, where annotators judged whether two questions (private and declassified) have the same semantic information-seeking intent.

Data	Method	Q sim.	S sim.	Q equiv.
WikiQA	Gen-PR	76.5	63.3	90.0
	Gen-BT	84.7	74.2	92.0
	Map	84.6	72.3	94.0
TrecQA	Gen-PR	75.8	60.3	86.0
	Gen-BT	80.4	62.5	88.0
	Map	79.5	64.6	90.0
OpenBookQA	Gen-PR	78.5	62.1	90.0
	Gen-BT	91.0	75.4	94.0
	Map	91.2	67.6	94.0
SimpleQA	Gen-PR	84.4	77.3	92.0
	Gen-BT	88.4	80.0	94.0
	Map	51.1	55.7	16.0

Table 5: Average semantic similarity of questions (Q sim.) and answer sentences (S sim.), with manual assessment of question equivalence (Q equiv.).

Across WikiQA, TrecQA, and OpenBookQA, retrieval-based declassification achieves high semantic similarity and equivalence rates, consistent with the near-zero performance deltas observed in sections 5.2 and 5.4. The results support the hypothesis that high-quality semantic alignment between original and mapped questions is essential for precise and utility-preserving declassification.

³<https://github.com/microsoft/presidio>

⁴[sentence-transformers/paraphrase-mpnet-base-v2](#)

6.2 The Boundary of Applicability: SimpleQA

SimpleQA was deliberately constructed to stress-test LLMs with obscure, constraint-heavy questions unlikely to appear in standard public corpora. It represents a controlled probe for the limits of retrieval-based declassification rather than a typical deployment scenario.

Retrieval-based declassification exhibits substantially lower similarity and equivalence rates on SimpleQA (Q sim. 51.1, Q equiv. 16%). This explains the large positive deltas in Table 1: the mapped questions are systematically easier than the originals. The examples in Table 9 illustrate this: retrieval fails to recover highly specific information such as the identifier “gal56” or the precise details of a particular soccer match, leading to mapped questions that omit critical constraints. Additional evidence comes from absolute model performance: across LLMs, accuracy on the declassified SimpleQA increases by an average of +9.5 points, with relative improvements ranging from approximately 132% to 400% relative to the original dataset, confirming that gains reflect a difficulty shift rather than improved evaluation fidelity. Crucially, this limitation does not undermine the applicability of our approach in its intended setting. The goal of declassification is to enable sharing and reuse of datasets from public-facing QA systems (Type B), which capture the behavior of broad user populations over publicly available knowledge. Extremely specific, tail questions, such as those deliberately curated in SimpleQA, are not representative of typical customer interaction data. While Gen-PR and Gen-BT achieve smaller deltas on SimpleQA by preserving semantic specificity, this same property is precisely what raises privacy concerns in realistic settings.

6.3 Applicability Guidelines

To help practitioners determine when declassification can be reliably applied, we provide the following diagnostic framework based on our experimental analysis.

Similarity-based diagnostics. We evaluate model performance on subsets of mapped SimpleQA questions filtered by similarity thresholds. For each threshold, only mapped questions with similarity above the given value are retained, and accuracy is compared to that on the corresponding original questions.

Table 6 shows that when similarity is high

Model	SimpleQA				
Phi-3	0.0	0.0	0.0	-2.5	-2.2
Ministral	0.0	0.0	-2.8	-2.5	-2.2
Phi-4	0.0	3.7	8.3	7.5	7.0
Qwen-3	0.0	0.0	0.0	-2.5	-1.8
Sim. Threshold	95.0	90.0	85.0	80.0	75.0
# Questions	16	27	36	40	46

Table 6: Accuracy Δ on SimpleQA subsets filtered by similarity threshold. When similarity is ≥ 0.90 , $\Delta \approx 0$ for most models.

(≥ 0.90), accuracy deltas are close to zero across models. Phi-3 and Qwen-3 achieve $\Delta = 0$ at both the 0.95 and 0.90 thresholds, while Ministral remains near zero. As similarity decreases to 0.80 or 0.75, larger deviations emerge, with negative deltas for most models and positive shifts for Phi-4, reflecting model-specific sensitivity to difficulty changes.

Practical recommendations. These results suggest the following guidelines for deploying dataset declassification:

Pre-deployment diagnostic: Before declassifying a dataset, organizations should compute the distribution of semantic similarity scores between original and mapped questions. If the majority of examples have similarity ≥ 0.85 , high-fidelity declassification is expected.

Partial declassification: When a fraction of examples falls below the similarity threshold, organizations can declassify only the high-similarity subset. Even partial declassification is strictly more informative than complete inaccessibility: releasing 70% of a dataset with validated fidelity is a significant advance over releasing nothing.

Boundary detection: The similarity distribution itself is diagnostic. A concentration of low-similarity mappings signals that the dataset contains tail or domain-specific questions outside the coverage of available public corpora. Such cases may require complementary strategies (e.g., controlled generation, domain-specific corpora) or may indicate a Type A regime where full declassification is not applicable.

6.4 On the Benchmark Contamination

Benchmark contamination, where evaluation examples appear in model pretraining data, is a general concern for NLP systems trained on web-scale corpora. This issue is not specific to dataset declassification: many public QA benchmarks may already overlap with the data used to train modern language models.

Declassification therefore operates within the same contamination regime as standard public evaluations. Moreover, for AS2 (WikiQA, TrecQA) we train models from scratch on task-specific data, reducing the risk of web-scale memorization. For LLM evaluation (OpenBookQA), both original and mapped questions are public, so any contamination is symmetric across conditions. The reconstructed datasets are derived exclusively from public sources and thus inherit the same potential exposure as any other public benchmark. Mapping proprietary questions to semantically equivalent public ones does not introduce new contamination risks beyond those already present in web-scale pretraining.

In practice, contamination is controlled through standard evaluation procedures. Researchers select training and test splits to avoid overlap with known training data or previously used benchmarks. The same procedure can be applied to declassified datasets by selecting evaluation subsets that do not intersect with the training data used for downstream models. As a result, the framework preserves the same experimental safeguards used in conventional QA evaluation.

7 Conclusion

We introduced dataset declassification, a framework for releasing QA datasets that preserve training and evaluation utility while containing only public content. By combining semantic question retrieval with domain-matched answer reconstruction, declassified datasets achieve performance within $\Delta \approx 0$ of their proprietary counterparts on standard benchmarks, and enable faithful benchmark replacement.

Our analysis shows that declassification succeeds when semantic alignment between original and mapped examples is high, a condition that holds for public-domain, distribution-sensitive QA data typical of public-facing systems (Type B). Domain-matched answer reconstruction preserves training signal, while cross-encoder reranking is critical for difficulty preservation. On tail distributions lacking public equivalents (Type A regime or adversarial benchmarks like SimpleQA), utility degrades gracefully and measurably, clarifying the intended scope. At a minimum, declassification provides a concrete and verifiable guarantee: it identifies portions of an internal dataset for which sufficiently similar public content exists, and enables their release in full. Even partial declassification,

i.e., releasing the high-overlap subset with validated fidelity, replaces organizational uncertainty with an explicit, auditable notion of public overlap, allowing companies to safely release data that would otherwise remain entirely inaccessible.

Future Work This work opens several directions for future research. First, extending declassification beyond question-determined semantics to *context-dependent* settings, such as multi-turn dialog, document-grounded QA with private passages, or user-log-based interactions. Such settings require declassifying not just questions and answers but also supporting context (e.g., conversation history, retrieved passages), while preserving contextual coherence and grounding.

Second, while our retrieval-based approach is effective for Type B data, handling tail and highly specific questions may require *hybrid retrieval-generation pipelines* that complement retrieval when no sufficiently similar public content exists, using controlled generative methods with formal privacy guarantees. Finally, important next steps include validation on proprietary industrial datasets and integration with complementary privacy techniques such as differential privacy.

Limitations

Our framework has several limitations. First, effectiveness depends on semantic overlap between proprietary questions and public retrieval corpora; adversarial or tail-heavy benchmarks like SimpleQA exhibit difficulty shifts when no equivalent questions exist, though this boundary is measurable via similarity diagnostics. Second, the answer reconstruction pipeline requires domain-appropriate corpora, limiting applicability to domains with suitable public text (e.g., Wikipedia for encyclopedic, CC-News for news, PubMed for medical). Third, our current formulation assumes question-determined semantics; extending to context-dependent QA remains open. Fourth, our evaluation focuses on English; multilingual declassification would require both multilingual question databases and appropriate corpora in target languages. Fifth, computational costs scale linearly with dataset size, with LLM annotation dominating processing time (approximately 12 seconds per example).

Furthermore, our conclusions are strongest in the regime where private information needs have close public semantic counterparts. The proof of

concept on 436 proprietary queries (Section 5.5) supports this setting but remains limited in scale, and its annotation process cannot be fully disclosed. Strengthening the claim that declassification generally applies to industrial QA datasets will require larger and more precisely characterized proprietary evaluations, including similarity and coverage distributions on internal queries, inter-annotator agreement on expert judgments, analyses of the cases where the mapping fails or abstains, and explicit contamination checks for mapped public questions.

Finally, while all released content is public by construction, the mapping itself is derived from private data and constitutes a residual attack surface: the selection of public questions reflects the distribution of private queries and may indirectly reveal distributional properties of the underlying information needs. We note that the private-to-public correspondence is never released, that reverse-engineering individual mappings from a pool of 38M candidates is infeasible without access to the private data, and that this concern applies to any dataset release strategy, including synthetic generation. A formal analysis of mapping leakage remains an important direction for future work.

Ethics Statement

Dataset declassification is intended to enable beneficial data sharing while protecting privacy. However, the same techniques could potentially be misused to circumvent legitimate data access restrictions. We emphasize that declassification should complement, not replace, proper data governance practices. Organizations should verify that declassification satisfies their specific legal and ethical obligations before using it for regulatory compliance.

References

- Marah Abdin and 1 others. 2024a. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.
- Marah Abdin and 1 others. 2024b. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.
- Simran Arora, Patrick Lewis, Angela Fan, Jacob Kahn, and Christopher Ré. 2023. [Reasoning over public and private data in retrieval-based systems](#). *Transactions of the Association for Computational Linguistics*, 11:902–921.
- Fan Bai, Keith Harrigan, Joel Stremmel, Hamid Hasanzadeh, Ardavan Saeedi, and Mark Dredze. 2024. [Give me some hard questions: Synthetic data generation for clinical qa](#). *Preprint*, arXiv:2412.04573.
- Simone Balloccu, Patrícia Schmidová, Mateusz Lango, and Ondrej Dusek. 2024. [Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian’s, Malta. Association for Computational Linguistics.
- Steven Bird and Edward Loper. 2004. [NLTK: The natural language toolkit](#). In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.
- Stefano Campese, Ivano Lauriola, and Alessandro Moschitti. 2023. [QUADRo: Dataset and models for QUESTION-answer database retrieval](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15573–15587, Singapore. Association for Computational Linguistics.
- Sai Chen, Fengran Mo, Yanhao Wang, Cen Chen, Jian-Yun Nie, Chengyu Wang, and Jamie Cui. 2023. [A customized text sanitization mechanism with differential privacy](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5747–5758, Toronto, Canada. Association for Computational Linguistics.
- Ziyao Cui, Minxing Zhang, and Jian Pei. 2026. [On membership inference attacks in knowledge distillation](#). *SIGKDD Explor. Newsl.*, 28(1):32–40.
- Cynthia Dwork and Jing Lei. 2009. [Differential privacy and robust statistics](#). In *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*, STOC ’09, page 371–380, New York, NY, USA. Association for Computing Machinery.
- Siddhant Garg, Thuy Vu, and Alessandro Moschitti. 2020. [Tanda: Transfer and adapt pre-trained transformer models for answer sentence selection](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7780–7788.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. [DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing](#). *Preprint*, arXiv:2111.09543.
- Konstantin Kotschenreuther. 2024. [EHR-DS-QA: A Synthetic QA Dataset Derived from Medical Discharge Summaries for Enhanced Medical Information Retrieval Systems](#). *PhysioNet*. Version 1.0.0.
- Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenertorp, and Sebastian Riedel. 2021. [PAQ: 65 million probably-asked questions and what you can do with them](#). *Transactions of the Association for Computational Linguistics*, 9:1098–1115.

- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. [Multilingual denoising pre-training for neural machine translation](#). *Transactions of the Association for Computational Linguistics*, 8:726–742.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. [Can a suit of armor conduct electricity? a new dataset for open book question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.
- Mistral AI. 2024. Mistral 8b. *Mistral AI Blog*.
- Anwesan Pal, Radhika Bhargava, Kyle Hinsz, Jacques Esterhuizen, and Sudipta Bhattacharya. 2024. [The empirical impact of data sanitization on language models](#). *Preprint*, arXiv:2411.05978.
- Qwen Team. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. [NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.
- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MpNet: masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20*, Red Hook, NY, USA. Curran Associates Inc.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. [No language left behind: Scaling human-centered machine translation](#). *Preprint*, arXiv:2207.04672.
- Mengqiu Wang, Noah A. Smith, and Teruko Mitamura. 2007. [What is the Jeopardy model? a quasi-synchronous grammar for QA](#). In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 22–32, Prague, Czech Republic. Association for Computational Linguistics.
- Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20*, Red Hook, NY, USA. Curran Associates Inc.
- Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. [Measuring short-form factuality in large language models](#). *Preprint*, arXiv:2411.04368.
- Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muenighoff, Defu Lian, and Jian-Yun Nie. 2024. [C-pack: Packed resources for general chinese embeddings](#). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24*, page 641–649, New York, NY, USA. Association for Computing Machinery.
- Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. [WikiQA: A challenge dataset for open-domain question answering](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.
- Oleksandr Yermilov, Vipul Raheja, and Artem Chernodub. 2023. [Privacy- and utility-preserving NLP with anonymized data: A case study of pseudonymization](#). In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 232–241, Toronto, Canada. Association for Computational Linguistics.
- Xiang Yue, Minxin Du, Tianhao Wang, Yaliang Li, Huan Sun, and Sherman S. M. Chow. 2021. [Differential privacy for text analytics via natural text sanitization](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3853–3866, Online. Association for Computational Linguistics.
- Xiang Yue, Huseyin Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Hoda Shajari, Huan Sun, David Levitan, and Robert Sim. 2023. [Synthetic text generation with differential privacy: A simple and practical recipe](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1321–1342, Toronto, Canada. Association for Computational Linguistics.

A Annotation Prompts

A.1 Answer Correctness Annotation

For answer candidate annotations we use the prompt in Listing 1.

Listing 1: Answer Correctness Annotation

You are evaluating whether a candidate sentence correctly answers a question.

```

Question: {question}

Reference answer (known correct): {reference}

Candidate: {candidate}

Does the candidate correctly answer the question?
Respond with only "correct" or "incorrect".

```

A.2 SimpleQA Evaluation

We use the original SimpleQA grading prompt from Wei et al. (2024), which classifies responses as “correct”, “incorrect”, or “not attempted”. The system prompt is shown in Listing 2 and the user prompt in Listing 3.

Listing 2: SimpleQA evaluation system prompt

```

Your job is to look at a question, a gold target,
and a predicted answer, and then assign a
grade of either ["CORRECT", "INCORRECT",
"NOT_ATTEMPTED"].

First, I will give examples of each grade, and then
you will grade a new example.

The following are examples of CORRECT predicted
answers.

    Question: What are the names of Barack Obama’s
    children?
    Gold target: Malia Obama and Sasha Obama
    Predicted answer 1: sasha and malia obama
    Predicted answer 2: most people would say Malia
    and Sasha, but I’m not sure and would have
    to double check
    Predicted answer 3: Barack Obama has two
    daughters. Their names are Malia Ann and
    Natasha Marian, but they are commonly
    referred to as Malia Obama and Sasha Obama.

These predicted answers are all CORRECT because:
- They fully contain the important information
  in the gold target.
- They do not contain any information that
  contradicts the gold target.
- Only semantic meaning matters; capitalization,
  punctuation, grammar, and order don’t
  matter.
- Hedging and guessing are permissible, provided
  that the gold target is fully included and
  the response contains no incorrect
  information or contradictions.

The following are examples of INCORRECT predicted
answers.

    Question: What are the names of Barack Obama’s
    children?
    Gold target: Malia and Sasha
    Predicted answer 1: Malia.
    Predicted answer 2: Malia, Sasha, and Susan.
    Predicted answer 3: Barack Obama does not have
    any children.
    Predicted answer 4: I think it’s either Malia
    and Sasha. Or it could be Malia and Jackie.
    Or it could be Joey and Malia.
    Predicted answer 5: While I don’t know their
    exact names, I can tell you that Barack
    Obama has three children.

These predicted answers are all INCORRECT because:
- A factual statement in the answer contradicts
  the gold target. Incorrect statements that
  have some hedging (e.g., "it is possible
  that", "although i’m not sure, i think")
  are also considered incorrect.

The following are examples of NOT_ATTEMPTED
predicted answers.

```

```

Question: What are the names of Barack Obama’s
children?
Gold target: Malia and Sasha
Predicted answer 1: I don’t know.
Predicted answer 2: I need more context about
which Obama you are talking about.
Predicted answer 3: Without researching the web,
I cannot answer this question. However, I
can tell you that Barack Obama has two
children.
Predicted answer 4: Barack Obama has two
children. I know that one of them is Malia,
but I’m not sure about the other one.

```

```

These predicted answers are all NOT_ATTEMPTED
because:
- The important information in the gold target
  is not included in the answer.
- No statements in the answer contradict the
  gold target.

Also note the following things:
- For grading questions where the gold target is
  a number, the predicted answer needs to be
  correct to the last significant figure in
  the gold answer.
- The gold target may contain more information
  than the question. In such cases, the
  predicted answer only needs to contain the
  information that is in the question.
- Do not punish predicted answers if they omit
  information that would be clearly inferred
  from the question.
- Do not punish for typos in people’s name if
  it’s clearly the same name.

```

Listing 3: SimpleQA evaluation user prompt

```

Here is a new example. Simply reply with either
CORRECT, INCORRECT, NOT ATTEMPTED. Don’t
apologize or correct yourself if there was a
mistake; we are just trying to grade the
answer.
'''
Question: {question}
Gold target: {target}
Predicted answer: {predicted_answer}
'''

Grade the predicted answer of this new question as
one of:
A: CORRECT
B: INCORRECT
C: NOT_ATTEMPTED
Just return the letters "A", "B", or "C", with no
text around it.

```

B Training Details

For AS2 experiments, we perform hyperparameter optimization for each of the 3 runs to ensure fair comparison across configurations. The hyperparameter search space includes learning rate $\in \{5 \times 10^{-6}, 7 \times 10^{-6}, 1 \times 10^{-5}, 2 \times 10^{-5}\}$, batch size $\in \{8, 16, 32, 64\}$, warmup of 10% of training steps, maximum sequence length of 256 tokens, and training for up to 15 epochs with early stopping (patience of 3 epochs). We optimize on Precision@1 using the development set. All experiments were conducted on a machine with 8 NVIDIA L40S GPUs (48GB each). Training a single AS2 model takes approximately 30–50 minutes depending on dataset size and hyperparameters.

C Computational Resources

Table 7 reports the computational time required by our declassification framework on WikiQA. The time consumption is measured on a machine equipped with 8 NVIDIA L40S GPUs (48 GB each), 192 Intel Xeon CPU cores, and 1 TB of RAM. The total runtime is approximately 7 hours (~ 12 seconds per example), with LLM annotation dominating processing time. Cost scales linearly with dataset size and can be reduced via batching and retrieval pruning.

Stage	Time
Question retrieval	1–2 min
Question reranking	8 min
Corpus retrieval	45 min
LLM annotation	~ 6 hrs
Total	~ 7 hrs

Table 7: Computational costs for WikiQA declassification (2,118 examples, $8 \times$ L40S GPUs).

D Additional Results

We tested our approach on complete benchmark declassification with additional metrics. Table 8 shows that the delta remains low across Precision@1, MAP, and MRR.

Config	P@1	MAP	MRR
Private	81.43 ± 0.84	87.98 ± 0.73	89.09 ± 0.74
Map	0.50 ± 1.8	-0.05 ± 0.74	-0.11 ± 0.75

Table 8: Full AS2 results (P@1/MAP/MRR) for WikiQA with DeBERTa.

E Additional Examples

This appendix presents additional illustrative examples of declassified questions. Tables 10 and 11 report representative cases drawn from three benchmarks (WikiQA, TrecQA, and SimpleQA), including both successful and failed mappings produced by different systems. While some reformulations preserve the original intent and constraints, others exhibit semantic drift, loss of specificity, or unintended changes in meaning. Table 12 reports examples of PII entity replacement from the proof of concept on proprietary data (Section 5.5), drawn from our public index rather than from the private queries.

F Scope and Applicability Discussion

This appendix provides a detailed discussion of the scope, assumptions, and applicability boundaries

Type	Question
WikiQA	
Original	how are pointe shoes made?
Gen-PR	what is the process of creating ballet slippers for dancing?
Gen-BT	How are high-heeled shoes made ?
Map	how pointe shoes are made?
TrecQA	
Original	What is the name of the promising anticancer compound derived from the pacific yew tree?
Gen-PR	what is the name of the anticancer drug isolated from the bark of a pacific species
Gen-BT	What's the name of the promising anti-cancer compound that comes from the Pacific tea forest?
Map	what is the name of the anticancer drug isolated from the bark of a pacific yew tree
SimpleQA	
Original	What year did the Lego part with ID gal56 first release?
Gen-PR	In which year was the Lego component identified by the code gal56 initially made available?"
Gen-BT	In what year was the Lego part with ID gal56 first released?
Map	when was the first lego set released
Original	How many corners did Barcelona take in the Champions League semi-final match between Barcelona and Milan on April 27, 2006?
Map	what year was barcelona eliminated from the champions league?
Original	Who taught Cab Calloway how to sing in the scat style?
Map	who taught cab calloway to sing in the scat style

Table 9: Examples of declassified test questions across three benchmarks.

of dataset declassification, elaborating on points raised during peer review.

Type A vs. Type B data regimes. Proprietary QA datasets differ fundamentally in what makes them proprietary. In the *Type A regime*, the underlying concepts are themselves private: internal procedures, proprietary product specifications, or confidential entities that have no public equivalent. Declassification cannot apply to such data because no semantically equivalent public content exists by definition. In the *Type B regime*, the knowledge queried by users is already publicly available (e.g., facts accessible via search engines or encyclopedias), while the proprietary component is the *distribution* of user interactions: which questions are asked, how frequently, and how they are labeled. Our framework targets exclusively the Type B regime. This distinction is critical: the near-zero performance deltas we observe hold precisely because sufficient semantic overlap exists between the proprietary questions and publicly available content.

Type	Question
WikiQA	
Original	how are pointe shoes made
Map	how pointe shoes are made?
Gen-PR	what is the process of creating ballet slippers for dancing?
Gen-BT	How are high-heeled shoes made ?
Original	how do pineapples grow
Map	How do pineapples grow?
Gen-PR	what is the process of pineapple cultivation?
Gen-BT	how golden pineapples grow
TrecQA	
Original	What is the name of the promising anticancer compound derived from the pacific yew tree?
Map	what is the name of the anticancer drug isolated from the bark of a pacific yew tree
Gen-PR	what is the name of the anticancer drug isolated from the bark of a pacific species
Gen-BT	What's the name of the promising anti-cancer compound that comes from the Pacific tea forest?
Original	Who is the author of the book, "The Iron Lady: A Biography of Margaret Thatcher"?
Map	who is the author of the book the book iron lady
Gen-PR	Can you identify the writer of the biography titled "The Iron Lady: Margaret Thatcher's Life Story"?
Gen-BT	Who is the author of "The Iron Lady: A Biography of Margaret Thatcher"?

Table 10: Additional examples of declassified test questions from WikiQA and TrecQA.

Concrete deployment scenarios. The Type B regime corresponds to a common industrial setting: companies operating search engines, voice assistants, or LLM-based chatbots collect massive QA interaction logs that cannot be released or reused despite the underlying knowledge being public. Concrete scenarios include: (i) an organization wants to publish an internal benchmark to enable external research collaboration; (ii) a company needs to share data with a vendor for labeling under data governance constraints; (iii) researchers inside a company want to release evaluation data to enable reproducibility of published results and let the community work on realistic datasets. In all cases, the content is public-domain, but the dataset as a unit cannot be released due to governance policies.

Why not simply filter or release overlapping content? Even for Type B data, direct release of overlapping records is typically infeasible for three reasons. First, governance and legal policies treat proprietary datasets as indivisible units; organizations generally cannot release individual records regardless of their content. Second, overlap between proprietary and public questions is predominantly semantic rather than lexical: equivalent questions are expressed differently, so exact-match filtering discards most examples. Third, releasing only fil-

Type	Question
SimpleQA	
Original	What year did the Lego part with ID gal56 first release?
Map	when was the first lego set released
Gen-PR	In which year was the Lego component identified by the code gal56 initially made available?"
Gen-BT	In what year was the Lego part with ID gal56 first released?
Original	How many corners did Barcelona take in the Champions League semi-final match between Barcelona and Milan on April 27, 2006?
Map	what year was barcelona eliminated from the champions league?
Gen-PR	Can you tell me the number of times Barcelona scored in the Champions League semi-final game against Milan on the 27th of April, 2006?
Gen-BT	How many turns did Barcelona make in the Champions League semi-final between Barcelona and Milan on April 27, 2006?
Original	In the lore of Dungeons and Dragons, what is the name of the fortress in the Astral Plane used as a lair by the red great wyrm Ashardalon?
Map	what is the name of the fortress in dungeons and dragons
Gen-PR	Within the mythos of Dungeons and Dragons, can you identify the name of the celestial stronghold in the Astral Plane that serves as the dwelling of the crimson colossal dragon known as Ashardalon?
Gen-BT	In the legend of Dungeons and Dragons, what is the name of the fortress in the astral plane that was used as a hiding place by the great red worm Ashardalon?
Original	Who taught Cab Calloway how to sing in the scat style?
Map	who taught cab calloway to sing in the scat style
Gen-PR	Which instructor guided Cab Calloway in mastering scat singing?
Gen-BT	Who taught Cab Calloway how to sing scat-style?
Original	What was the name of male impersonator Hetty King's half-sister?
Map	what was the name of the half sister of music hall artist hetty king
Gen-PR	Can you tell me the name of Hetty King's half-sibling who was also a male impersonator?
Gen-BT	What was the name of the half-sister of the male impersonator, Hetty King?
Original	Which was the second largest republic in the Soviet Union?
Map	What is the second largest republic of the soviet union?
Gen-PR	Identify the republic with the second highest population within the former Soviet Union.
Gen-BT	What was the second largest republic of the Soviet Union?

Table 11: Additional examples of declassified test questions from SimpleQA. Note how retrieval-based declassification preserves semantics for well-represented questions (bottom rows) but loses specificity for tail questions (top rows).

tered subsets does not preserve the statistical and difficulty properties of the original dataset. Declassification addresses all three issues by reconstructing an independently sourced public dataset and empirically validating that it preserves training and evaluation utility.

Original query	Entity-replaced
what was daisy jane the song before daisy jane	what was britney spears the song before britney spears
what university is in monterey california	what university is in san diego california
who testified in the case of margaret garner	who testified in the case of harriet tubman
what is the elo rating of chess grand master levon aronian	what is the elo rating of chess grand master magnus carlsen
what remains were found in the pyramid of senusret iii	what remains were found in the pyramid of tutankhamun

Table 12: Examples of PII entity replacement (Section 5.5). Finding entities of comparable popularity is challenging, and popularity-matched substitutions still alter which documents are relevant. Examples are drawn from our public index, as the original queries are private.

Evaluation settings and contamination. A potential concern is that mapping proprietary questions to public equivalents may introduce contamination asymmetry: if mapped questions appear in LLM pretraining data more frequently than the originals, performance comparisons could be biased. We address this by noting the distinct evaluation settings in our experiments. For Answer Sentence Selection (WikiQA, TrecQA), we use fine-tuned models (DeBERTa-v3-base, MiniLM) trained from scratch on task-specific data; these models have no mechanism for memorizing web-scale content, so contamination does not apply. For LLM zero-shot evaluation (OpenBookQA), both original and mapped questions are public, making any contamination effect symmetric. For SimpleQA, asymmetry is expected by construction (originals are deliberately obscure, mapped questions are more common), and the observed large positive deltas (+6 to +12 points) are consistent with this, confirming that the framework surfaces rather than masks contamination effects.

When declassification fails. Our analysis on SimpleQA (Section 6.3) demonstrates that declassification degrades predictably when the proprietary dataset contains questions with no adequate public equivalent. The similarity-based diagnostics in Section 6.4 provide a practical tool for organizations to assess applicability before deployment: when the majority of mapped questions have similarity ≥ 0.85 , high-fidelity declassification is expected; below this threshold, partial declassification or complementary strategies may be needed.

G Declassification Framework Architecture

Figure 1 provides an overview of the proposed declassification framework.

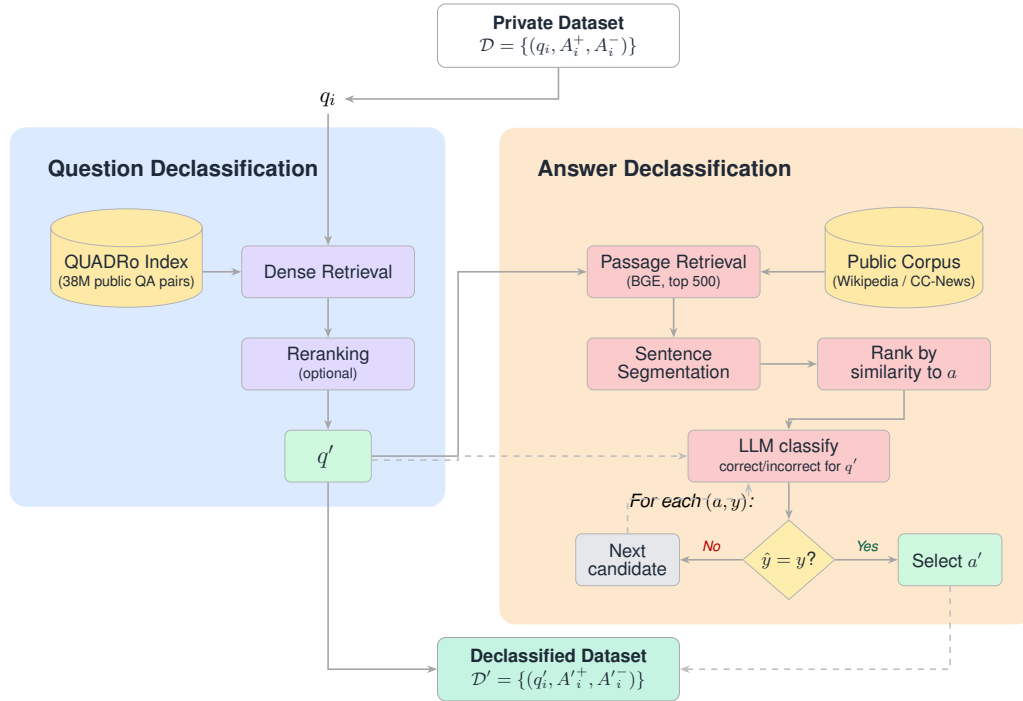


Figure 1: The declassification pipeline, here in the AS2 instantiation, where answers are sets of labeled candidate sentences. Given a proprietary example $(q, \mathcal{A}^+, \mathcal{A}^-)$, question declassification finds a similar public question q' via QUADRo retrieval and reranking. Answer declassification then reconstructs answers from a domain-appropriate corpus: for each original answer a with label y , we retrieve passages for q' , segment them into sentences, rank sentences by similarity to a , and use an LLM to find a replacement a' with matching label (with early stopping for efficiency).