# Evaluation Plan for Dense‑X Replicability

## Objective

We want to verify the performance gains reported by the **Dense X Retrieval** paper when moving from passage‑level indexes to proposition‑level indexes. The paper shows that indexing Wikipedia at the level of “propositions” (atomic, self‑contained facts) improves dense retrieval and downstream QA performance. Specifically, the authors report that **the average improvement over passage‑based retrieval of recall@20 is +10.1 points for unsupervised dense retrievers and +2.2 points for supervised retrievers**[[1]](https://arxiv.org/html/2312.06648v2#:~:text=Wikipedia%20is%20indexed%20by%20passage%2C,provide%20a%20higher%20density%20of). Our goal is to design a small, reproducible evaluation that can confirm these improvements on our own data and models.

## Evaluation Dataset

* **Number of Q/A pairs:** 30–50. A moderate‑sized set is large enough to detect meaningful differences while being small enough for manual annotation.
* **Domain:** neutral and general‑purpose (e.g., Wikipedia or well‑known facts) to avoid domain bias. Questions should cover a range of topics and difficulty levels.
* **Ground‑truth annotations:**
* **Relevant documents:** For each query, manually identify all documents (passages or propositions) that contain the answer.
* **Citation spans:** For each relevant document, record the character start and end positions of the exact answer span. These spans will be used to compute citation‑span accuracy.

## Metrics

The evaluation uses both order‑unaware and order‑aware retrieval metrics as well as a citation‑quality metric.

| Metric | Definition | Source |
| --- | --- | --- |
| **Recall@K** | Measures how many relevant items are returned in the top‑K results. If a query has  relevant documents and the system returns relevant documents among the top  retrieved items, recall@K =  . Higher values indicate more complete retrieval. Pinecone’s guide describes recall@K as one of the most interpretable offline metrics[[2]](https://www.pinecone.io/learn/offline-evaluation/#:~:text=Recall%40K). | Pinecone evaluation metrics overview[[2]](https://www.pinecone.io/learn/offline-evaluation/#:~:text=Recall%40K) |
| **nDCG@K (normalized discounted cumulative gain)** | An order‑aware metric that measures ranking quality. It rewards placing highly relevant items near the top of the ranking and normalises the score so that 1.0 represents a perfect ranking. nDCG is computed by dividing the discounted cumulative gain (DCG) by the ideal DCG; Pinecone’s tutorial explains how DCG and nDCG@K are derived[[3]](https://www.pinecone.io/learn/offline-evaluation/#:~:text=Normalized%20Discounted%20Cumulative%20Gain%20). | Pinecone evaluation metrics overview[[3]](https://www.pinecone.io/learn/offline-evaluation/#:~:text=Normalized%20Discounted%20Cumulative%20Gain%20) |
| **Citation‑span accuracy** | Measures whether the retrieved citation spans match the ground‑truth spans. For each predicted span, compute the overlap with the true span as an F1‑score (precision: proportion of predicted span that overlaps with the true span; recall: proportion of the true span captured by the prediction). Average the best matching F1 over all predicted spans. No widely used standard exists for this metric, so it should be treated as an internal measure [Unverified]. | Internal definition [Unverified] |

## Experimental Conditions

We will compare two retrieval pipelines on the same dataset:

1. **Passage‑chunk baseline:** The corpus is split into fixed‑length passages (e.g., 100 words) and indexed with a dense retriever. For each query, retrieve the top‑ passages and compute recall@K, nDCG@K and citation‑span accuracy.
2. **Proposition‑level index:** The corpus is segmented into atomic propositions (as described in the Dense X paper) and indexed with the same dense retriever. Retrieval and evaluation follow the same protocol as the baseline.

Both pipelines should use identical model architectures and training data; the only difference is the granularity of the indexed units. Following the paper’s findings, we expect to observe higher recall@20 on the proposition index (potentially around **+10 points** on unsupervised models and **+2 points** on supervised models[[1]](https://arxiv.org/html/2312.06648v2#:~:text=Wikipedia%20is%20indexed%20by%20passage%2C,provide%20a%20higher%20density%20of)). The citation‑span accuracy is also hypothesised to be higher, since propositions provide more focused text with less extraneous content.

### Parameter Settings

* **Top‑K values:** Evaluate at to measure early‑retrieval and full‑retrieval effectiveness.
* **Embedding dimension:** Use the same vector dimension for both indexes (e.g., 384‑dimensional BGE‑small or 1024‑dimensional BGE‑m3). This ensures that any difference arises from retrieval unit size rather than embedding capacity.
* **Evaluation frequency:** Run the evaluation on a single snapshot of the index. For reproducibility, record the model checkpoints and random seeds.

## Workflow

1. **Data preparation:** Curate 30–50 Q/A pairs and annotate the relevant passages/propositions and citation spans. Save as a tabular file (CSV/JSON).
2. **Index construction:** Build two separate indexes (passage and proposition) using the same embedding model. Record any differences in indexing time or storage requirements.
3. **Retrieval:** For each query, retrieve the top‑ items from each index and record both the retrieved IDs and the model‑generated citation spans.
4. **Metric computation:** Use the provided Python harness (see eval‑harness.zip) to compute recall@K, nDCG@K and citation‑span accuracy for each query. Aggregate results by taking the mean across all queries.
5. **Analysis:** Compare the metrics between passage and proposition indexes. Highlight absolute improvements in recall@20 and nDCG@K. Determine whether the observed gains replicate the paper’s reported **+10.1/+2.2** improvements[[1]](https://arxiv.org/html/2312.06648v2#:~:text=Wikipedia%20is%20indexed%20by%20passage%2C,provide%20a%20higher%20density%20of).

## Confidence and Limitations

* **Confidence level:** Medium. The metrics definitions are derived from standard information retrieval literature[[2]](https://www.pinecone.io/learn/offline-evaluation/#:~:text=Recall%40K)[[3]](https://www.pinecone.io/learn/offline-evaluation/#:~:text=Normalized%20Discounted%20Cumulative%20Gain%20). However, citation‑span accuracy is an internal metric without an established baseline [Unverified].
* **Potential biases:** With only 30–50 queries, variance may be high. The results may not generalise beyond the chosen domain. Annotation quality directly affects citation‑span accuracy.
* **Reproducibility:** By publishing the dataset, index construction scripts, and evaluation harness, others can replicate and extend the experiments.

[[1]](https://arxiv.org/html/2312.06648v2#:~:text=Wikipedia%20is%20indexed%20by%20passage%2C,provide%20a%20higher%20density%20of) Dense Retrieval: What Retrieval Granularity Should We Use?

<https://arxiv.org/html/2312.06648v2>

[[2]](https://www.pinecone.io/learn/offline-evaluation/#:~:text=Recall%40K) [[3]](https://www.pinecone.io/learn/offline-evaluation/#:~:text=Normalized%20Discounted%20Cumulative%20Gain%20) Evaluation Measures in Information Retrieval | Pinecone

<https://www.pinecone.io/learn/offline-evaluation/>