Vector Retrieval Metrics — Ground Truth First

Vector Retrieval Metrics — Ground Truth First

Author: Sayed Amir Karim (NextX AG)
Organization: NextX AG, Hauptstrasse 20, 6418 Rothenthurm, Switzerland
Contact: compression@aqea.ai

Primary reference (DOI backlink):

```
Karim, Sayed Amir (NextX AG).
"AQEA: Technical Report - Domain-Adaptive Semantic Compression of Embeddings."
Version 0.1, 2026-01-03. DOI: 10.5281/zenodo.18138436 (https://doi.org/10.5281/zenodo.18138436).
Reproducibility source: aqea-compress commit f093b2d.
```

One-sentence rule

If a metric name does not state the ground truth, it is **operationally ambiguous** and **not decision-grade**.

Why this matters

In vector retrieval, teams commonly use "Recall@k" to mean two different things:

1) **Baseline overlap** ("How close are we to a baseline model's top-k?")

2) **Truth-based retrieval** ("How good are we vs human/business truth?")

Mixing these is as unprofessional as mixing physical units (e.g., power vs energy): you can't optimize what you can't define.

Why the industry ended up here (real systems)

- **Term drift**: IR terms were reused in ML without carrying the explicit relevance definition.
- **GT is missing**: labels are expensive, so teams substitute baseline rankings or biased logs.
- **Incentives**: ambiguous metrics look comparable on dashboards, even when they are not.

Ground truth types (GT)

- **GT-H**: Human labels/scores
- **GT-W**: Weighted scores (document the weights)
- **GT-P**: Pairwise preferences
- **GT-L**: Log-based truth (biased; debias assumptions required)
  - **Minimum bar (5 lines):**
    - **Position bias**: clicks depend on rank/exposure, not only relevance.
    - **Selection bias**: observed interactions are conditional on what was shown.
    - **Feedback loops**: the serving policy changes the data you collect.

    - ∗ **Counterfactual evaluation**: state whether you use IPS / doubly-robust / randomized buckets.
    - ∗ **Assumptions**: explicitly list the debias assumptions you rely on.
- **baseline=B**: Baseline model reference (not truth)
  - Recommended: `baseline=original-float32` when comparing against uncompressed embeddings

Canonical metric names

Customer-relevant (GT-based)

- `nDCG@k[GT-H]` / `nDCG@k[GT-W]` (graded or weighted truth)
- `SetOverlap@k[GT-*]` (set overlap with GT-defined top-k; alias: `HitRate@k[GT-*]`)
- `Precision@k[GT-*]`, `Recall@k[GT-*]` (only for binary truth)

Diagnostic only (baseline-based)

- `BO@k[baseline=B]` = **Baseline Overlap @k** (alias: `BRecall@k[baseline=B]`)
  Use for regression checks, not for product-quality claims.

Legacy → canonical mapping

Legacy phrase	What it usually meant	Use instead
"Recall@10"	baseline overlap	`BO@10[baseline=B]`
"Hits@k"	often baseline overlap (ANN / embedding eval)	`BO@k[baseline=B]` (diagnostic) or `Recall@k[GT-*]` (if GT is explicit)
"Accuracy@k" (in retrieval)	either "contains a relevant item" or baseline overlap	`SetOverlap@k[GT-*]` (GT) or `BO@k[baseline=B]` (baseline)
"Quality retention"	Spearman vs baseline similarities	`Spearman[baseline=B]`
"Retrieval@10"	unclear	`nDCG@10[GT-*]` or `SetOverlap@10[GT-*]`

Minimal example

If baseline top-10 disagrees with humans, then "Recall@10 vs baseline" can increase while real user utility decreases. Therefore: always report GT-based metrics for decisions; baseline-based metrics only for diagnostics.

Toy mini-example (why BO@k GT utility)

Baseline top-2: `{A, B}`
GT top-2 (human/weighted): `{A, C}`
System top-2: `{A, C}`

- `BO@2[baseline=B]` = overlap(system, baseline) = 1/2 (looks "worse")
- `SetOverlap@2[GT-*]` = overlap(system, GT) = 2/2 (is actually "better")

Same system change, opposite conclusions — depending on the ground truth.

Visual intuition (3-set overlap)

Figure: 3-set overlap (Baseline vs GT vs System)

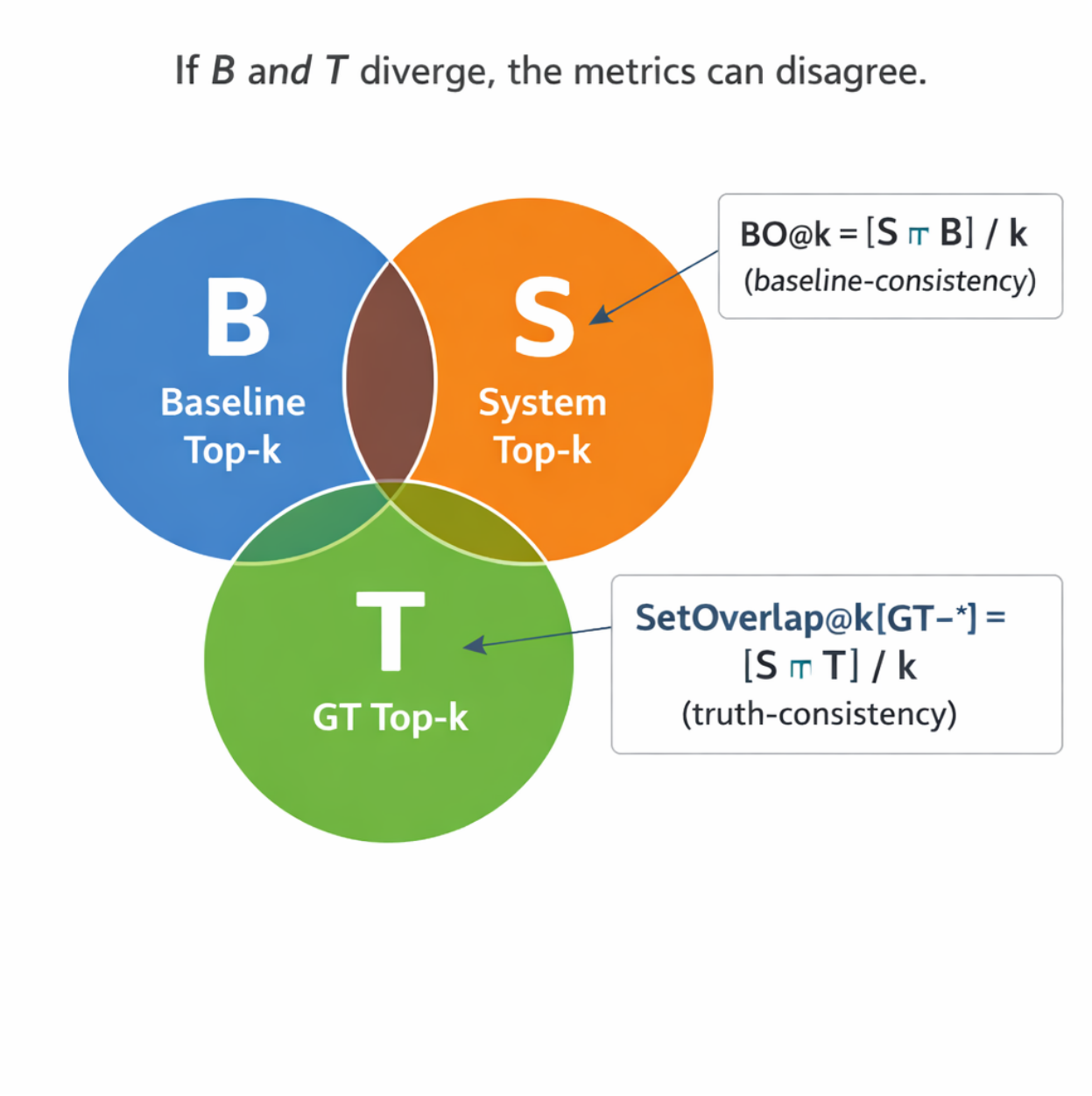

Figure 1: 3-set overlap: Baseline vs GT vs System

Key: - `BO@k[baseline=B]` measures overlap(S, B) (**baseline-consistency**) - `SetOverlap@k[GT-*]` measures overlap(S, T) (**truth-consistency**)

Note: Canonical asset is stored at `aqea-technical-report/assets/venn_3set_overlap.png` (this bundle includes a copy for PDF builds).

Observed patterns (domains)

Legal search

Baseline semantic similarity often ignores legal structure (authority, jurisdiction, procedural posture). Tuning toward **GT-W** can improve `nDCG@k[GT-W]` while decreasing `BO@k[baseline=B]`.

E-commerce search / recommendations

Business relevance depends on constraints (stock, price bands, intent ambiguity, policy filters). Improving `nDCG@k[GT-W]` can require intentionally departing from baseline semantics, so `BO@k[baseline=B]` may drop.

Machine-readable metric header (mini schema)

To make this linter-friendly in real systems, attach a small structured header (YAML/JSON) to every reported metric:

```yaml
task_type: retrieval            # retrieval | ranking | correlation | reconstruction
gt_type: GT-W                   # GT-H | GT-W | GT-P | GT-L | baseline
baseline_id: original-float32   # required if gt_type=baseline
candidate_pool_id: pool_v1      # stable ID
candidate_pool_desc: "tenant=acme, locale=en, category=shoes, hard-negatives"
metric: SetOverlap
k: 10
ci_method: bootstrap
ci_level: 0.95
```

Audit checklist (minimum)

Any reported metric must state: - task (retrieval / correlation / reconstruction) - ground truth (`GT-*` or `baseline=B`) - candidate pool definition

If one is missing, the number is ambiguous and should be treated as **diagnostic-only**.

Call to action

Adopt ground-truth-aware labels (`[GT-*]` / `[baseline=*]`) in: - dashboards, - docs, - papers, - and PR review checklists.

Reference (backlink)

Primary reference for AQEA's evaluation discipline and auditable claims:

- **DOI**: `10.5281/zenodo.18138436` (`https://doi.org/10.5281/zenodo.18138436`)
- Source: `aqea-technical-report/TECHNICAL_REPORT.md`