Overview of the SISAP 2025 Indexing Challenge



Eric S. Tellez, Edgar Chavez, Martin Aumüller, Vladimir Mic

Task 1: Resource-Limited Indexing

How do you search through 23.9 million scientific abstracts when you only have 16 GB RAM, 8 CPUs, and a 12 hours time limit to build an index structure?

Task 1 challenges participants to build memoryefficient approximate nearest neighbor (ANN) indexes under strict resource constraints.

Goal: Achieve ≥70% average recall of 30 NN queries while maximizing search throughput on unseen query set.

Dataset: 23.9M Sentence-BERT embeddings (384) dimensions) from the **PUBMED** corpus— 2x too large to fit in memory!

Twist: Queries come from titles, not abstracts, introducing a distribution shift.

Evaluation: Throughput beyond recall threshold, tested on unseen queries.

Task 2: k-NN Graph Constructing

How do you build a k-nearest neighbor graph (k=15) for millions of vectors under tight resource limits?

Task 2 challenges participants to design memoryefficient graph construction algorithms using 16 GB RAM, 8 CPUs, and 12 hours.

Goal: Achieve ≥80% average recall (of graph neighborhood) while minimizing total computation.

Dataset: Google Q&A corpus with 3M Sentence-BERT embeddings (384 dimensions), totaling 7.4 GB.

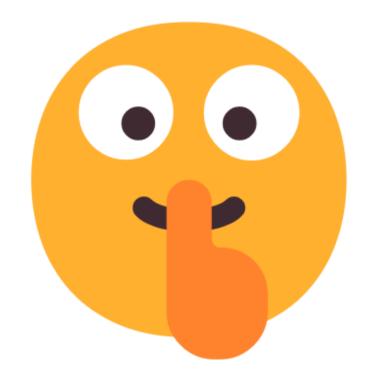
Twist: Accurate index might be too time-consuming to build!

Evaluation: Based on end-to-end runtime (preprocessing → graph computation → postprocessing) and quality metrics.

Participating Teams

| TEAM | MEMBERS | TASK |
|---------------------|--|------|
| BrownCICESE | Foster, Magdaleno-Gatica, Kimia | 1, 2 |
| cm-lll | Lou, Ma, Luo, Ruan, Wu, Lu, Mao | 1 |
| Crusty Coders | Dearle, Connor, Claydon, McKeogh | 1, 2 |
| DCC-UChile | Bustos, Chen | 2 |
| hforest | Imamura | 1, 2 |
| JLapeyra | Lapeyra | 1, 2 |
| TeamDoubleFiltering | Higuchi, Imamura, Shinohara, Hirata, Kuboyama | 1 |

Final Ranking



Join the session on Thursday starting at 14:30!







