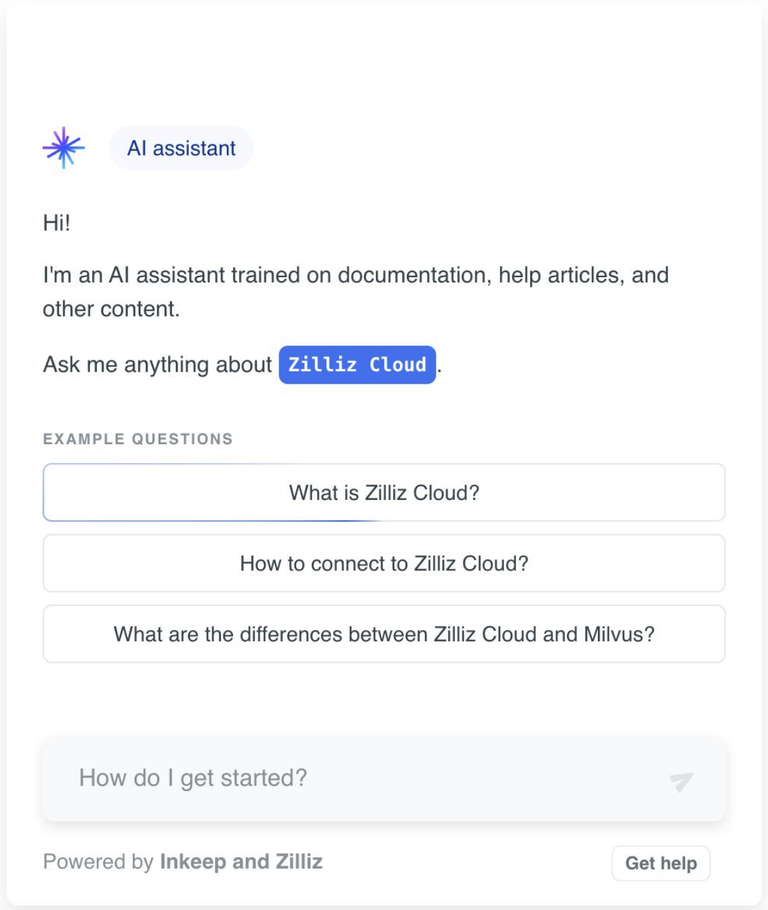
inkeep zilliz

How Inkeep & Zilliz built an AI Assistant

Robert Tran

Co-founder, CTO

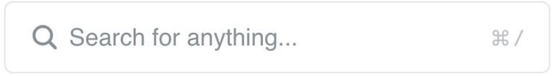
inkeep



Answer user questions, **instantly**.

As a **chat widget**

As a **search bar**



As a **Discord** or **Slack bot**

inkeep

AI support you can actually trust.

"Ask AI" for docs

Support team copilot

Community bots

POWERING AI SUPPORT FOR HIGH-GROWTH TEAMS

ANTHROP\C

POSTMAN

zilliz

clerk

scale

baseten

inkeep

What is the difference between Milvus and Zilliz Cloud?

How to get started with a cluster?

How to connect to Zilliz in my Python app?

How to scale the cluster?

How to query using REST API?

What do compute units mean?

Hybrid search?

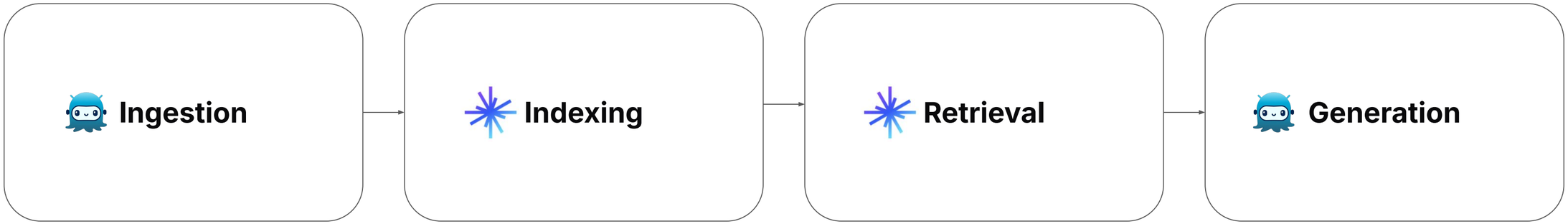
Should I use serverless?

Reciprocal Rank Fusion?

How to insert and upsert vectors?

How to include sparse vectors?

Retrieval Augmented Generation



Goals

- Ingest content from
  - Technical docs, support FAQs, GitHub repositories
- **Index and search with Zilliz**
  - **Metadata Filtering, Dense + Sparse Vectors, Hybrid Search**
- Integrate the AI Assistant into
  - The documentation at https://docs.zilliz.com
  - The support center at https://support.zilliz.com

Zilliz Vector Database

1. Ingestion

a. High quality metadata extraction and processing

2. Indexing

a. Traditional BM25 Sparse

b. Sparse Models

c. Dense Models

3. Search

a. Hybrid search with sparse + dense vectors

Version: User Guides (Cloud)

Search, Query & Get

This series of guides demonstrate similarity searches and scalar queries in a Zilliz Cloud collection.

Single-Vector Search

After your data is inserted, the next step is to send a `search` request to search for vectors that are similar to your query vector. A single-vector search compares your query vector against the existing vectors in your collection to find the most similar entities, returning their IDs and the distances between them. This process can optionally return the vector values and metadata of the results.

Hybrid Search [READ MORE]

Zilliz Cloud introduced multi-vector support and a hybrid search framework, which means users can bring in several vector fields into a single collection. These vectors in different columns represent diverse facets of data, originating from different embedding models or undergoing distinct processing methods. The results of hybrid searches are integrated using reranking strategies, such as Reciprocal Rank Fusion (RRF) and Weighted Scoring. To learn more about reranking strategies, refer to Reranking.

Get & Scalar Query

This guide demonstrates how to get entities by ID and conduct scalar filtering. A scalar filtering retrieves entities that match the specified filtering conditions.

Modify Cluster

Modify a dedicated cluster. Currently support upgrading dedicated clusters CU size.

POST `/v2/clusters/{CLUSTER_ID}/modify`

Base URL

The base URL for this API is in the following format:

https://api.cloud.zilliz.com

```
export BASE_URL="https://api.cloud.zilliz.com"
```

Parameters

Authorization `string` `header` `required`

The authentication token should be an API key with appropriate privileges.

Example Value: `Bearer {{TOKEN}}`

Accept `string` `header`

Use `application/json`.

```
export TOKEN="YOUR_API_KEY"
export CLUSTER_ID="inxx-xxxxxxxxxxxxxx"

curl --request POST \
--url "${BASE_URL}/v2/clusters/${CLUSTER_ID}/modify" \
--header "Authorization: Bearer ${TOKEN}" \
--header "Accept: application/json" \
--header "Content-Type: application/json" \
-d '{
    "cuSize": 2
}'
```

inkeep

Document Processing

- Source type
  - site, documentation, support
- Record type
  - text, code, programming language, version
- Hierarchical references
  - Parent, sibling, children records
- URLs, paths, tags
- Dates

inkeep

Vector Embeddings

- Traditional sparse BM25
- Sparse models
  - SPLADE
  - BGE-M3
- Dense models
  - MSMARCO
  - MPNET
  - BGE-M3

inkeep

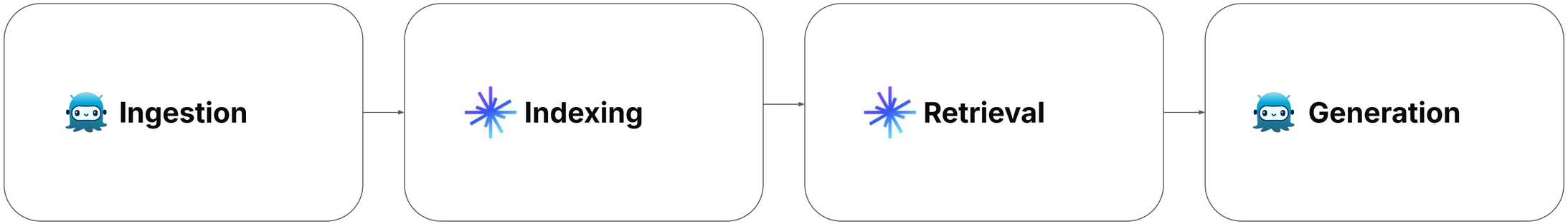
Vector Embeddings

- Traditional sparse BM25
- Sparse models
  - SPLADE
  - BGE-M3
- Dense models
  - MSMARCO
  - MPNET
  - BGE-M3

Hybrid Search

- Weighted Scoring
- Reciprocal Rank Fusion

Retrieval Augmented Generation



This means the connection attempt will time out if no response is received within 30 seconds after the request is sent.

- **IP Whitelist**: Ensure that your public IP address is added to the cluster's IP whitelist. Private IP addresses will not work for connection.
- **VPN/Proxy**: Turn off any VPN or proxy connections when attempting to connect to your cluster.

If you encounter connection issues, you can troubleshoot by:

- Verifying that the cluster status is RUNNING.
- Checking if your IP address is included in the IP whitelist.
- Testing port connectivity using the command:

```
telnet in01-(uuid).(region).vectordb.zillizcloud.com 19530
```

If you're still having trouble connecting after trying these steps, you can get more help at Support.

Sources

FAQ: Cluster
Docs

How do I get started?

zilliz benchmarks

Ask AI

⌘ ↵

Guides (10) | BYOC (8) | Reference (0) | Support (1) | Customers (1) | Events (2) | Learn (9) | GitHub (1)

- Performance Benchmarking with...
  Docs
- Single-Vector Search
  Docs
- FAQ: Get Started
  Docs
- Best Practices
  Docs
- FAQ: Pipelines
  Docs

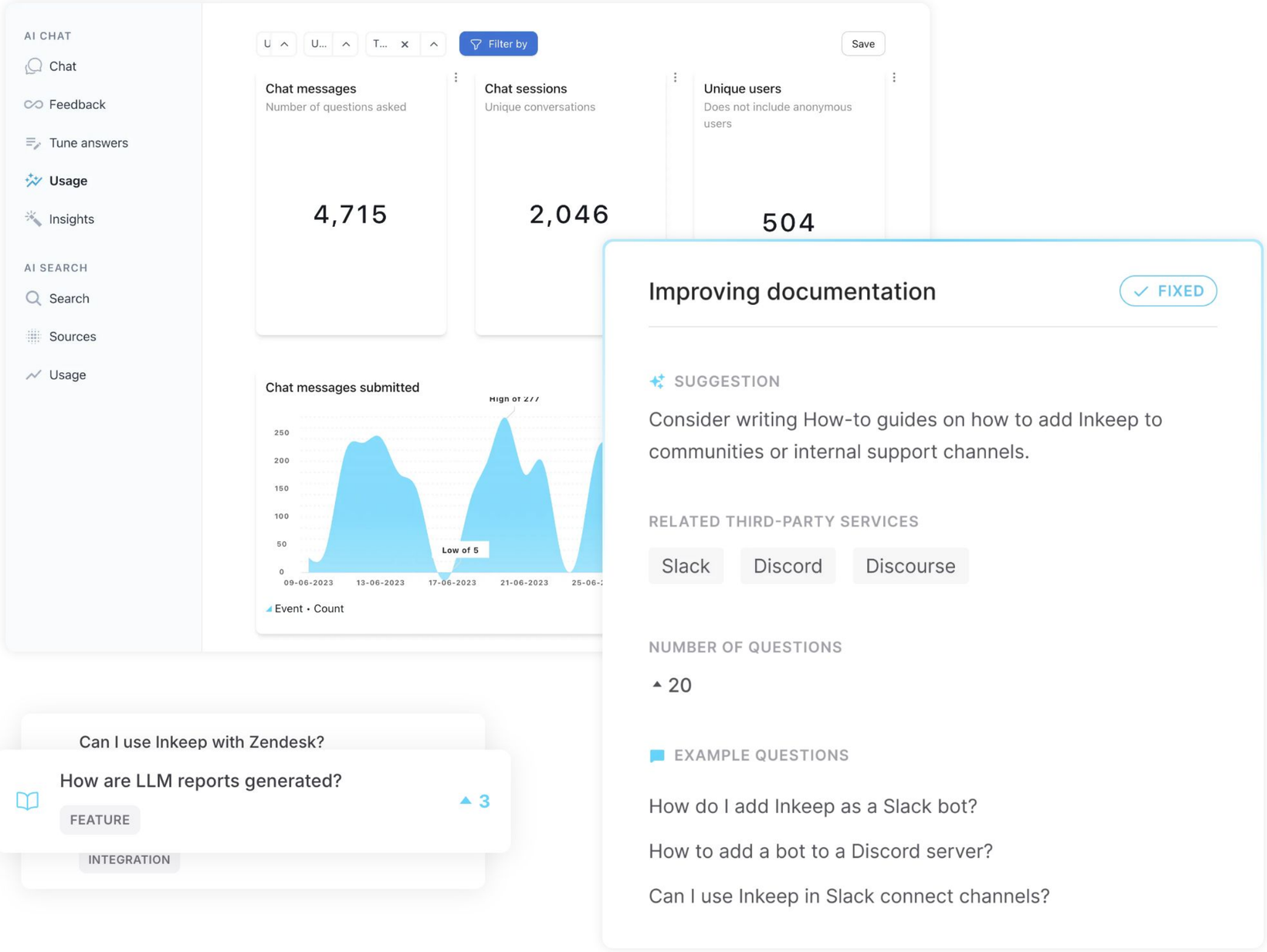
By inkeep & zilliz

Performance Benchmarking with VectorDBBench

Docs

... the performance test results of **Zilliz** Cloud. Overview & VectorDBBench ...

ON THIS PAGE


AI CHAT
Chat
Feedback
Tune answers
Usage
Insights
AI SEARCH
Search
Sources
Usage
Filter by
Save
Chat messages
Number of questions asked
4,715
Chat sessions
Unique conversations
2,046
Unique users
Does not include anonymous users
504
Chat messages submitted
250
200
150
100
50
0
Low of 5
09-06-2023
13-06-2023
17-06-2023
21-06-2023
Event • Count
Improving documentation
FIXED
SUGGESTION
Consider writing How-to guides on how to add Inkeep to communities or internal support channels.
RELATED THIRD-PARTY SERVICES
Slack
Discord
Discourse
NUMBER OF QUESTIONS
20
EXAMPLE QUESTIONS
How do I add Inkeep as a Slack bot?
How to add a bot to a Discord server?
Can I use Inkeep in Slack connect channels?
Can I use Inkeep with Zendesk?
How are LLM reports generated?
FEATURE
3
INTEGRATION

inkeep

Thank you!

follow.inkeep.com/linkedin

follow.inkeep.com/twitter