Qdrant

— VECTOR SPACE MEETUP · 2026

Retrieval in the Age of Agents

June 11, 2026 · Merantix AI Campus, Berlin

Evgeniya Sukhodolskaya
SENIOR DEVELOPER ADVOCATE · QDRANT

— VECTOR SPACE MEETUP BERLIN

Tonight.

18:20	**Keynote** · Vector Search in the Age of LLMs Andrey Vasnetsov, CTO @ Qdrant
18:55	**Panel** · Retrieval in the Age of Agents driven by YOUR questions
19:50	Networking with pizza
21:00	Wrap

- **Tonight is filmed.** Panel video lands on Qdrant's YouTube.

— WHO WE ARE

A vector search engine. Not a database.

Berlin-based · Open source · Written in Rust · github.com/qdrant/qdrant

We're on a Vector Space Meetup → Space is full of stars → Add one to this GitHub →

★ github.com/qdrant/qdrant

01

Vectors aren't data.

An embedding is a **transformation** of data you already have. It creates no new facts, only a new way to reach the ones you've got.

02

There's no universal store.

Every workload accepts different constraints, and **each constraint you give up unlocks a different axis of optimization.**

03

So we built an engine.

Purpose-built for vectors, **standalone**, and tuned end-to-end for retrieval.

— THE THESIS

Search is a primitive.

Search can be tuned.

01

Dense

Meaning-based similarity.

02

Sparse

Keyword precision.

03

Filter

Structured constraints on metadata.

04

Multivector

Multi-aspect – ColBERT, late interaction.

05

Score

Ranking control and custom scoring.

— COMPOSE IT · CONTROL IT · CONFIGURE IT

Three layers of control, one engine.

Composability

which operations, one query

- dense ANN
- sparse / lexical
- metadata filters
- multi-vector
- custom scoring
- prefetch
- fusion (RRF / weighted)
- rerank

Controllability

how it scores, at query time

- fusion weights
- score formulas
- oversampling + rescoring
- search-time ef
- accuracy ⇄ latency

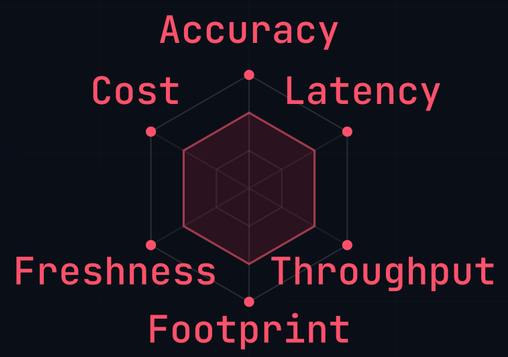
Configurability

how the engine runs each one

- quantization (scalar / binary / product)
- RAM ⇄ mmap ⇄ SSD tiering
- HNSW m / ef_construct
- payload indexes

— RETRIEVAL · ONE WORD, MANY SHAPES

Every workload wants something different.



Each spoke is a demand. **Further out = more demanding** on that axis. No two workloads share a shape.

RAG & assistants

Product search

Recommendations

Agent memory

Multimodal similarity

Large-scale archive

— SEARCH IN PRODUCTION

One engine, behind every workload.

MULTIMODAL
Tripadvisor

2–3× more revenue from AI-engaged users · 1B+ items

RECOMMENDATIONS + RAG
HubSpot

Breeze AI — chosen after it outperformed every alternative on retrieval

SEARCH AT SCALE
Bazaarvoice

2.7B vectors · ~100× smaller footprint · <100 ms @ ~98% recall

HEALTHCARE
Roche

SEMANTIC FILTERING
OpenTable

AI Concierge — sparse embeddings filter 60K+ restaurants

REAL-TIME MULTIMODAL
Flipkart

9 h → <1 min fraud-detection turnaround on live retrieval

AGENTS · KNOWLEDGE
Dust

5,000+ data sources · multi-second → sub-second · ¼ the RAM

AGENTS
Lyzr

>90% latency cut · 300–500 ms → 20–50 ms

PLATFORM SCALE
Canva

Standardizing vector search across AI features · millions of users

MANUFACTURING
Bosch

The workloads from earlier — **one engine behind all of them.**

— UP NEXT · KEYNOTE

Retrieval in the Age of Agents: Our Plans

You did the whole event for us: Your questions made not only the panel but the keynote.

"What does the search engine of the future look like? Will we still need stand-alone search engines and vector databases?"

Founder

"Does traditional search have a future, or is it destined to die out completely?"

Data/ML Engineer

"What is the next big thing in vector databases*?"

Senior AI Architect

**or vector search engines :)*

KEYNOTE SPEAKER

Andrey Vasnetsov

CTO @ Qdrant

The person who decides what the future of Qdrant search will look like.

Vector Search in the age of LLMs

And what's Qdrant vision about it

Who am I

- Name is Andrey Vasnetsov
- CTO @ Qdrant
- Working with search since 2014

Question Asking Guide

wrong: Is `grep` all I need?

better: Do I need indexing?

top: What is the slowest and most expensive part which might benefit from indexing?

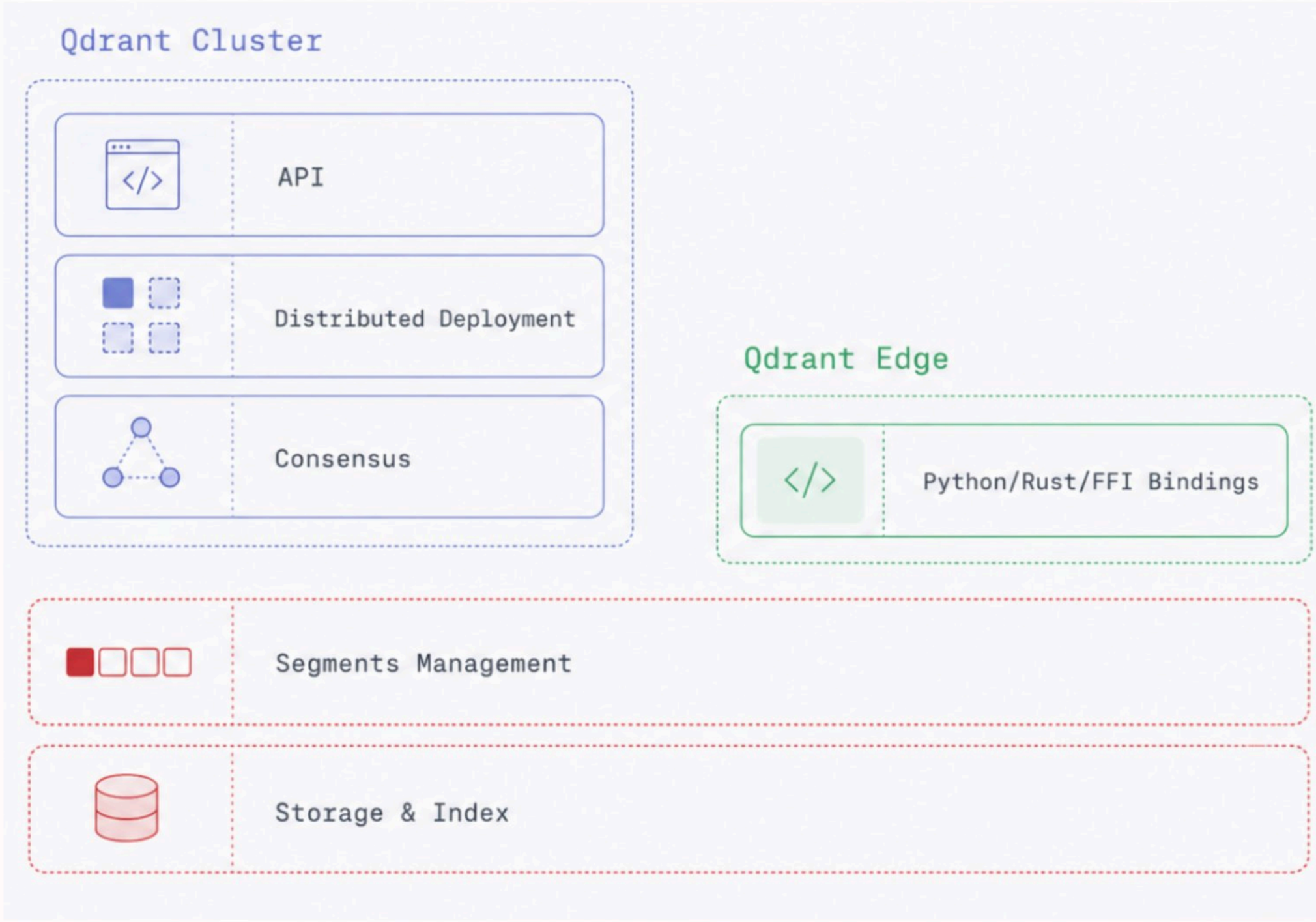
wrong: Should I put my vectors into relational database?

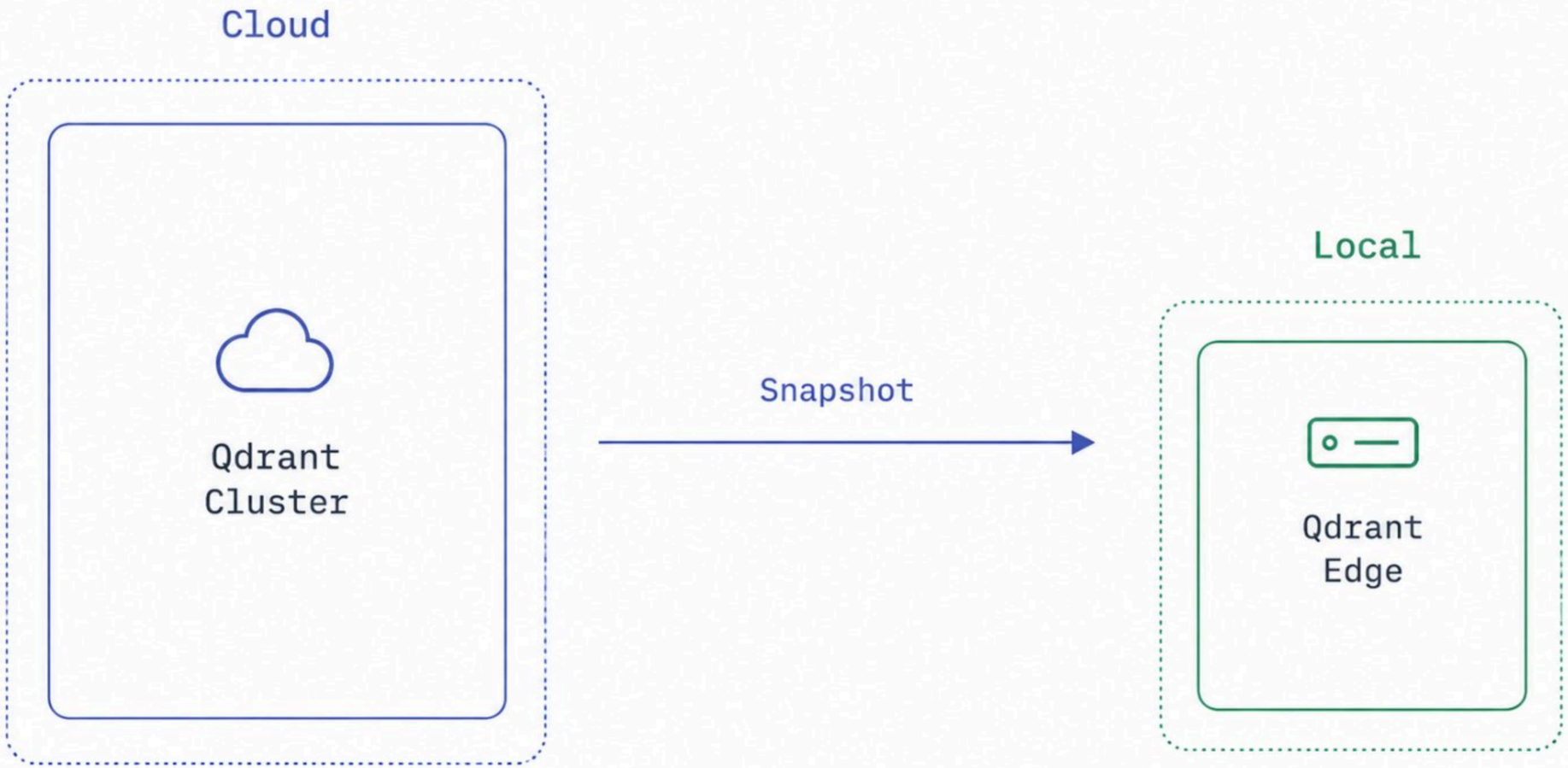
better: What guarantees do I need from vector store?

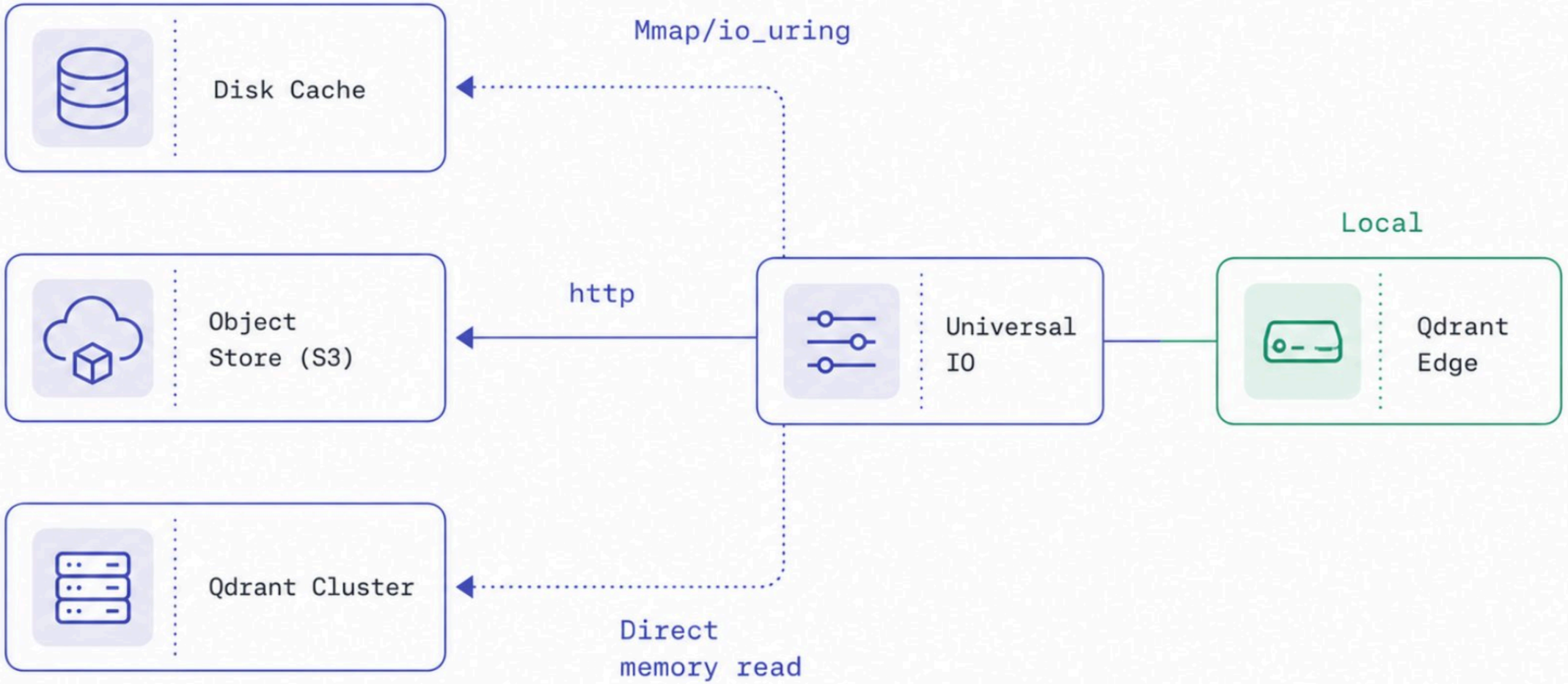
top: What trade-offs am I ready to make for better performance?

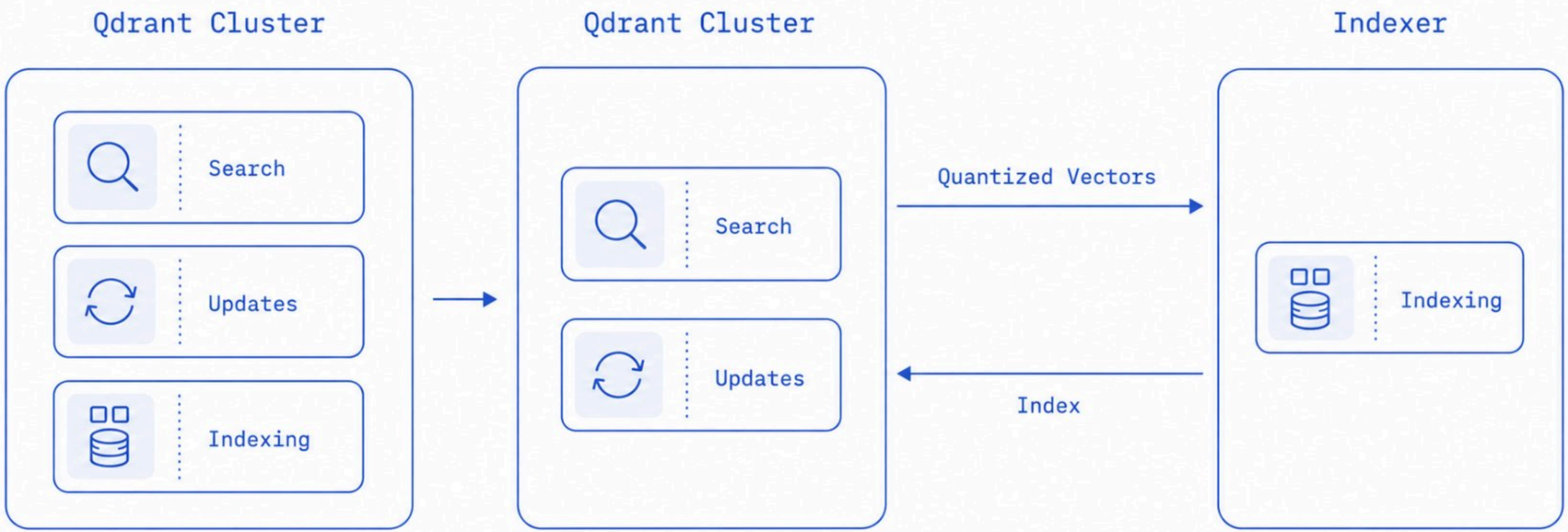
Qdrant in 2026

- Qdrant Edge
- Storage and compute separation
- Distributed indexing


Qdrant Cluster
API
Distributed Deployment
Consensus
Qdrant Edge
Python/Rust/FFI Bindings
Segments Management
Storage & Index


Cloud
Qdrant
Cluster
Snapshot
Local
Qdrant
Edge


Disk Cache
Mmap/io_uring
Object
Store (S3)
http
Universal
IO
Local
Qdrant
Edge
Qdrant Cluster
Direct
memory read


Qdrant Cluster
Search
Updates
Indexing
Qdrant Cluster
Search
Updates
Quantized Vectors
Index
Indexer
Indexing

— KEYNOTE · ANDREY VASNETSOV

Thank you. Questions?

Next in the schedule → Panel · Retrieval in the Age of Agents

— THE PANEL

You drove this panel.

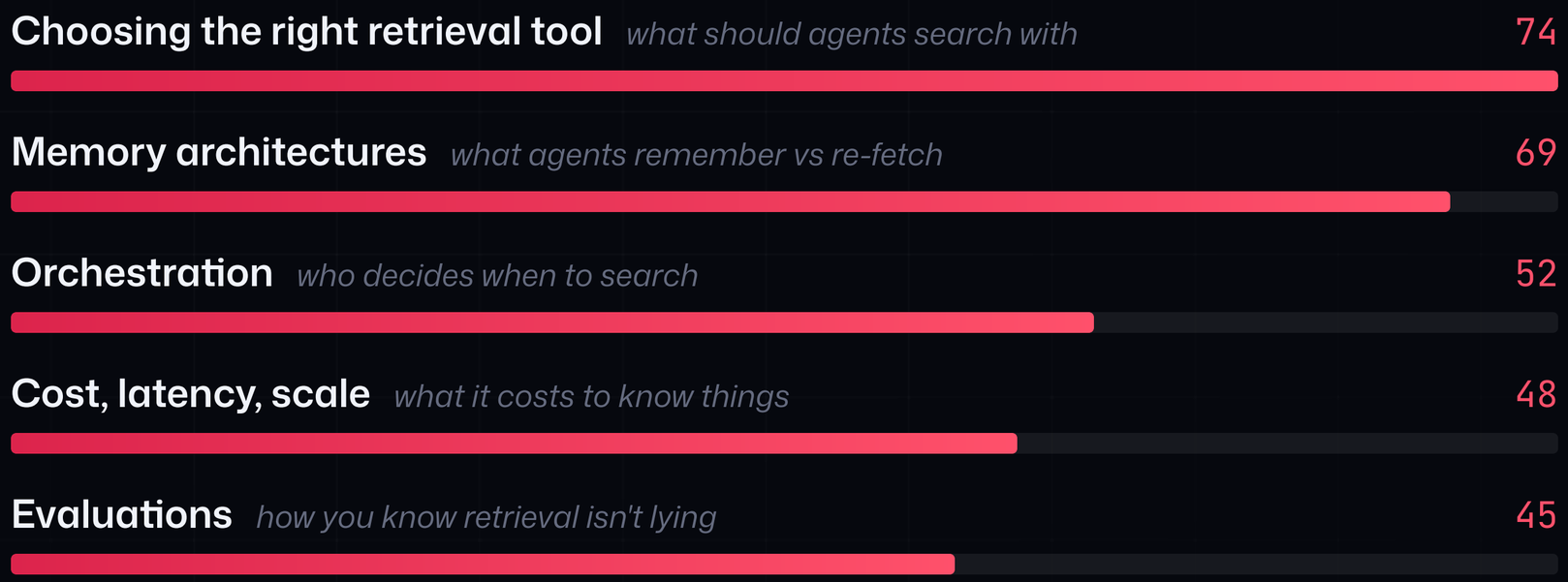
136

questions submitted

4

certified jokes:)

"WHAT ARE YOU CURIOUS ABOUT"



Every question got a home: **panel tonight**, networking whiteboards, company staff.

QD Qdrant team · R RAG in production · A Agentic retrieval in production · N New to this · T Tonight's panelist

THE PANEL

Retrieval in the Age of Agents

01 The Shift

02 Starting Right

03 What to Pick

04 Context Engineering

05 Evals & Debugging

06 Sustainability & Cost

07 Hot Take

PANELISTS

Andrey (Qdrant)

Vasilije (Cognee)

Bilge (deepset/Haystack)

Clelia (LlamaIndex)

Marcel (n8n)

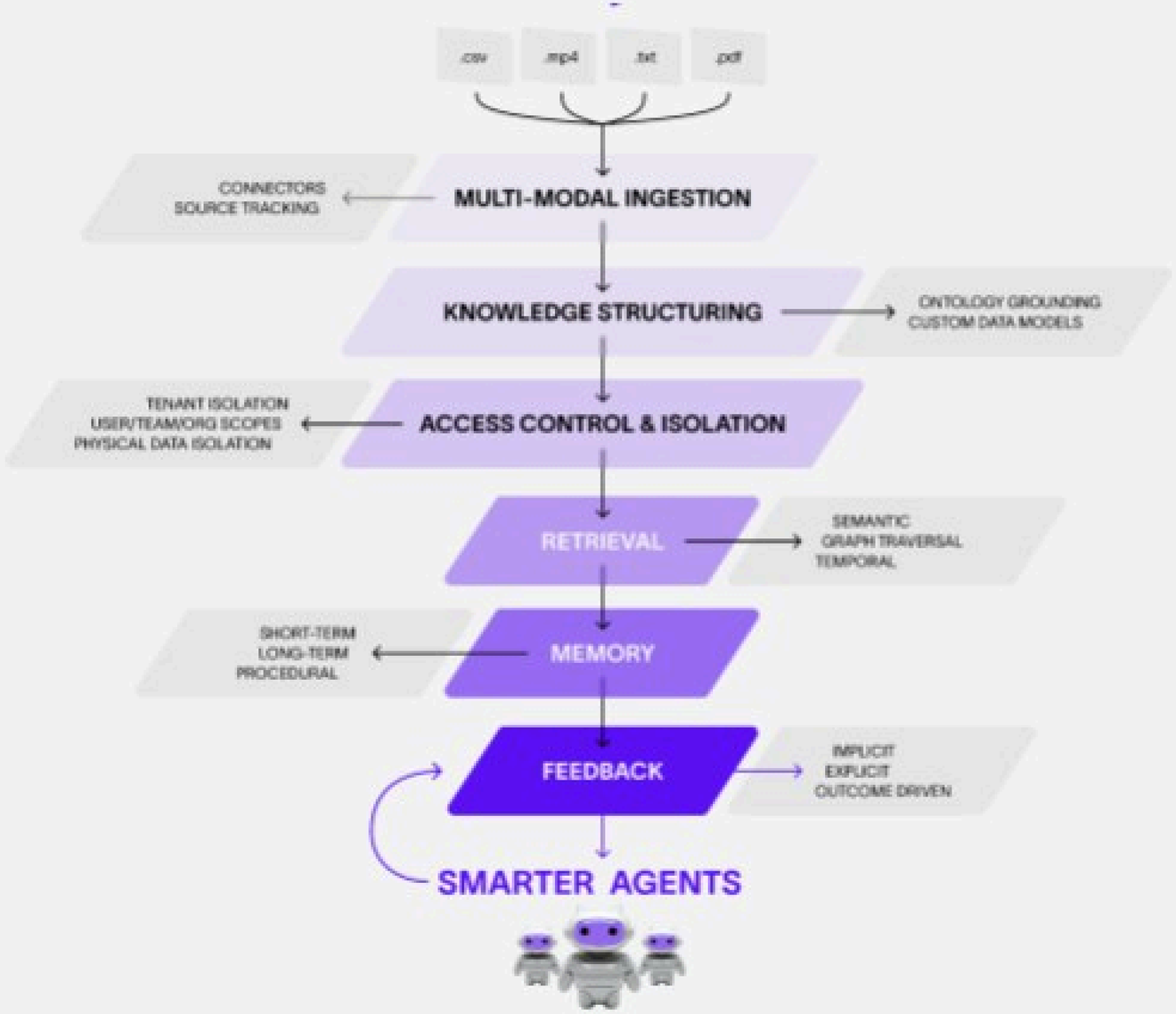
cognee

Most enterprise AI journeys start with a **static RAG**, or a **chatbot**.

They are easy to try, but bring **limited disruption**.

They break in edge cases, **repeat the same mistakes**, and require extensive input.

Cognee provides the **persistence** and **adaptability** for AI systems to work in critical workflows.



n8n
Webhook
AI Agent
When chat message received
webhook?
true
false
Respond to Webhook
respond to chat
alculator
n8n Ambassador
Marcel Claus-Ahrens

LlamaIndex: The Developer Platform for Building Document Agents

LlamaIndex

Agent Templates

- Equity Analyst Agent
- Underwriting Agent
- Invoice Processing Agent
- Indexing & Retrieval

LlamaAgents

Workflows And LlamaIndex

- Composable Steps
- Stateful Orchestration
- Observability
- HITL
- Agent Tools and Memory

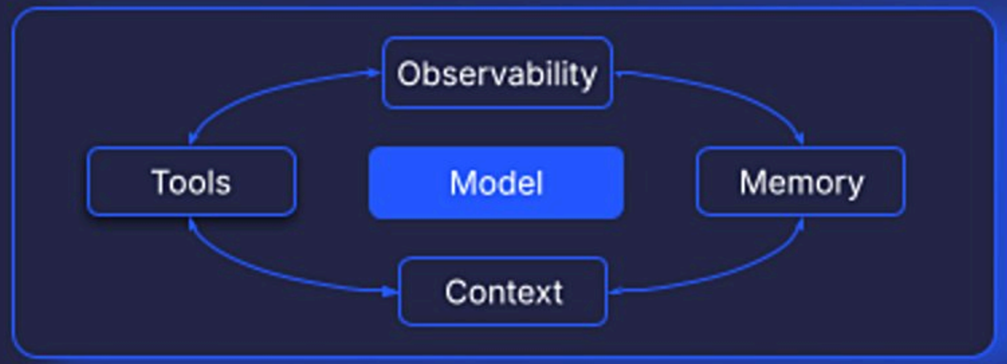
LlamaParse Agentic Document Processing

- Data Connectors
- Classification
- Parsing
- Structured Extraction
- Indexing & Retrieval

deepset x Haystack

AI Orchestration

Pipelines: Advanced RAG, Agents, Multimodal, IDP, Text2SQL, Search



Agents: Autonomous, conversational, reliable

deepset | haystack Open Source | haystack Enterprise Platform

Foundational Models

OpenAI · cohere · ANTHROPIC · Google DeepMind · MISTRAL AI_ · deepseek

Infrastructure & Hardware

groq · NVIDIA · Google · aws · Microsoft · ORACLE · STACKIT · HPE · IONOS

— QUESTION 1 OF 7 · THE SHIFT

“Classic RAG is relatively deterministic and auditable. Agentic retrieval is dynamic and emergent. Are we trading reliability for capability, and is that a good deal?”

“Many RAG prototypes work because retrieval is a fixed step. In agentic systems, the agent decides whether, when, and how to retrieve. … keep this flexible without making the system unreliable or too slow?”

Senior Software Engineer

“How do we know agents retrieve data properly , can they mistake there ?”

Platform architect

“How could I transition best from a classic RAG to agentic retrieval … using Qdrant?”

Software Engineer

— QUESTION 2 OF 7 · STARTING RIGHT

“For startups building on (agentic) retrieval right now: which technical decisions are hardest to reverse later, so they're worth getting right early?”

“What is the most common mistake … integrating agentic retrieval?”

“For agentic retrieval in real products, what is the hardest part to get right: choosing the retrieval method, designing memory, orchestrating tools, or evaluating quality? …”

Founder / CEO

“How can we best minimize brittleness of agentic pipelines?”

Senior FDE

— QUESTION 3 OF 7 · WHAT TO PICK

Retrieval stack decisions look very different depending on resources and scale. What would you recommend to a solo developer, a PM building on your tools, or an enterprise team with strict requirements?

> “For solo entrepreneurs or small teams building their first AI agent … how do you choose between vector search, graph retrieval, and lexical/grep? … What's the right retrieval stack when your entire company knowledge base fits on one VPS … and there's no ML engineer on staff?”
>
> *AI Marketing Engineer*

> “How can PMs brainstorm around all the agentic search stuff?”
>
> *Product Manager*

> “Which retrieval architectures work best for highly structured enterprise documents … where both semantic similarity and exact positional logic are important?”
>
> *Founder*

— QUESTION 4 OF 7 · CONTEXT ENGINEERING

Default belief: more context, better agent. When does more retrieved context make things worse, and what actually earns a place in the window?

the most popular topic in your submissions

> “... what if most agents are hindered by information overload (defining technology as doing more with less)?”
>
> *Digital Solutions Intern*

> “How do you deal with the inevitable context rot and discarding of non-relevant chunks?”
>
> *Founder*

> “As agents become increasingly stateful, what do you believe should be the source of truth for memory: retrieval systems, structured memory stores, or model-internal representations? ...”
>
> *Senior R&D Engineer*

— QUESTION 5 OF 7 · EVALS & DEBUGGING

How do you measure and maintain retrieval quality: from knowing a single result is good to knowing the whole agent loop landed right?

"... what signals actually indicate retrieval quality beyond chunk similarity scores?"

Research Associate

"How do you evaluate agentic/iterative retrieval?"

Co-Founder

"When an agent does multiple retrieval steps, the failure modes compound: query drift, context window pollution, and retrieval of stale or contradicting chunks. ... how do you debug it when the agent's reasoning chain is opaque?"

Senior AI Engineer

"What are the biggest telltale signs that a RAG system's poor performance is caused by bad chunking strategies versus bad data?"

Targeting Lead

— QUESTION 6 OF 7 · SUSTAINABILITY & COST

“Looks like nowadays the bottleneck became the cost/money burn for LLMs. How do we reduce it?”

“With agent costs rising rapidly, what guardrails, technical or organizational, do you recommend to prevent over-retrieval, over-agentization, and blind trust in agent outputs ...?”

MLOps Engineer

“How we can reduce the use of tokens and still have fast responsiveness”

AI Engineer

“More often than not, carefully crafted AI pipelines meant to remain in production over years are obsoleted within months ... What's your take on sustainable development and operation of agentic solutions?”

Team Lead Data Science

“What are your suggestions to decrease latency while keeping the same level of performance?”

NLP Engineer

— QUESTION 7 OF 7 · HOT TAKES

“Which current trends do you think are overhyped relative to their real-world adoption or usefulness, and why?”

— AUDIENCE Q&A

Your turn. The fastest hand and the most controversial question.

3-2-1 GO! :)

AND DON’T FEEL LEFT OUT:

Guided networking is **THE second panel** tonight.

The second but more important one!:)

— GUIDED NETWORKING

The second panel: all of you.

Whiteboard 01
(We're trying to) Make It Into Production

Whiteboard 02
Retrieval Tool Choice

Whiteboard 03
In the Context Window

Qdrant Discord → #ask-for-help

Your lu.ma questions are already printed on the boards.
Grab a marker and add more!

Thank you for coming! We hope you enjoy tonight's event!

QD	R	A	N	T
Qdrant team	RAG in production	Agentic retrieval	New to this	Tonight's panelist

Qdrant qdrant.tech github.com/qdrant