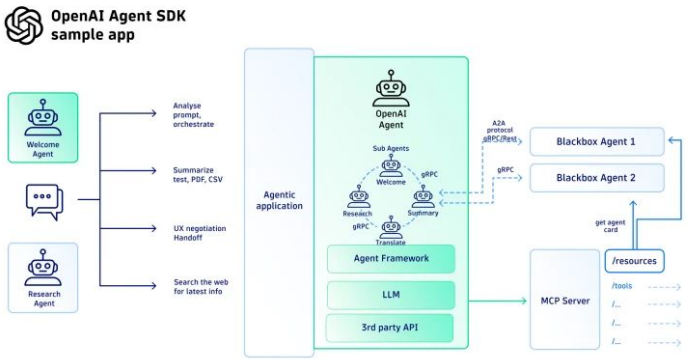
今使っている LLM について考える

2025 年 10 月 31 日
DS Team 山本 諒介

ともに挑む。ともに実る。
MIZUHO

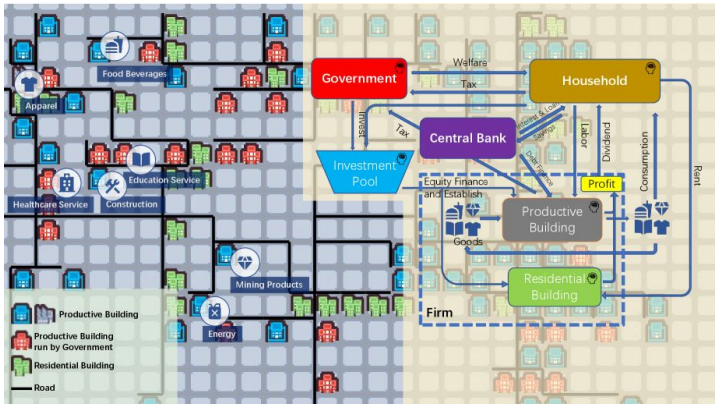
増加し続ける LLM のニーズ

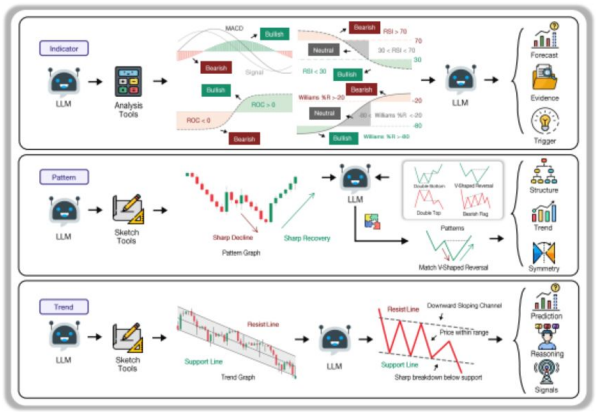
- 現在
  - LLM を組み合わせたプロダクト開発の増加
  - AI Agent, Agentic Workflow のような多くの推論過程を必要とするタスクの増加
- 今後
  - Multi Agent 化をはじめとする Agentic System の複雑化
  - 自動コーディングなどの長時間タスク
  - AI Agent を用いた高負荷な分析（人工市場シミュレーション、Agentic 投資戦略開発）



OpenAI Agents

SimCity





QunatAgent

問題提起

- 現在利用可能なモデルの選択肢
  - 言語モデル: gpt-4.1 (リリース日: 2025/4)
  - 埋め込みモデル: text-embedding-ada-002 (リリース日: 2022/12)

- 課題（今後想定される課題も含む）
  - 選択できるモデルの自由度がない（上記のモデルは古い）
  - リクエスト制限（RPM, RPD）により、並列利用の限界がある
  - トークン数上限（TPM, TPD）により、複雑なタスクを処理できない

- 解決案
  - デジ戦に他のモデルを使えるよう依頼
    - Pros: ただ？
    - Cons: 時間がかかる、手続きが面倒くさい、セキュリティ面の不安
  - Model Provider（Azure OpenAI Servie, Amazon Bedrock, Google Vertex AIなど）との契約
    - Pros: 自由度が高い
    - Cons: 時間がかかる、コストがかかる、手続きが面倒くさい、セキュリティ面の不安
  - ローカル LLM を自前でデプロイする
    - Pros: **自由度が高い、すぐできる、**（LLM への感度・知見が高まる）
    - Cons: ランニングコスト、マネジメントコストがかかる

共有したいこと

- ローカル環境に LLM をホストしたい
- 今の埋め込みモデルはいい選択なのか
- 今のベクトルDBはいい選択なのか

ローカルLLMを使いたい

- 前提
  - AWS EC2 インスタンスにデプロイする
    - VRAM は 16GB は欲しいので g4dn-xlarge (T4, 84,989 円/月) を利用検討

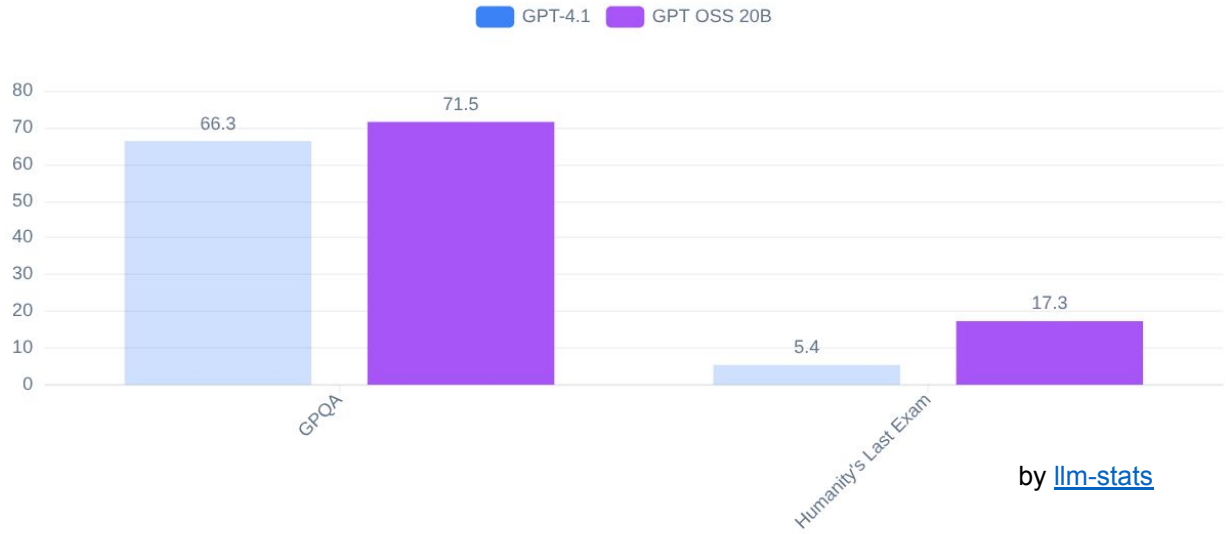
- どんなモデルが利用できるのか？

Model	Created by	Size	Context	Input
gpt-oss:20b	OpenAI	14GB	128K	Text
gemma3:4b	Google	3.3GB	128K	Text, Image
phi4:latest	Microsoft	9.1GB	16K	Text
llama3:8b	Meta	4.7GB	8K	Text
qwen2.5-coder:14b	Alibaba	9.0GB	32K	Text

- ホスティング方法
  - Ollama を利用（とても簡単！）

gpt-4.1 vs gpt-oss:20b

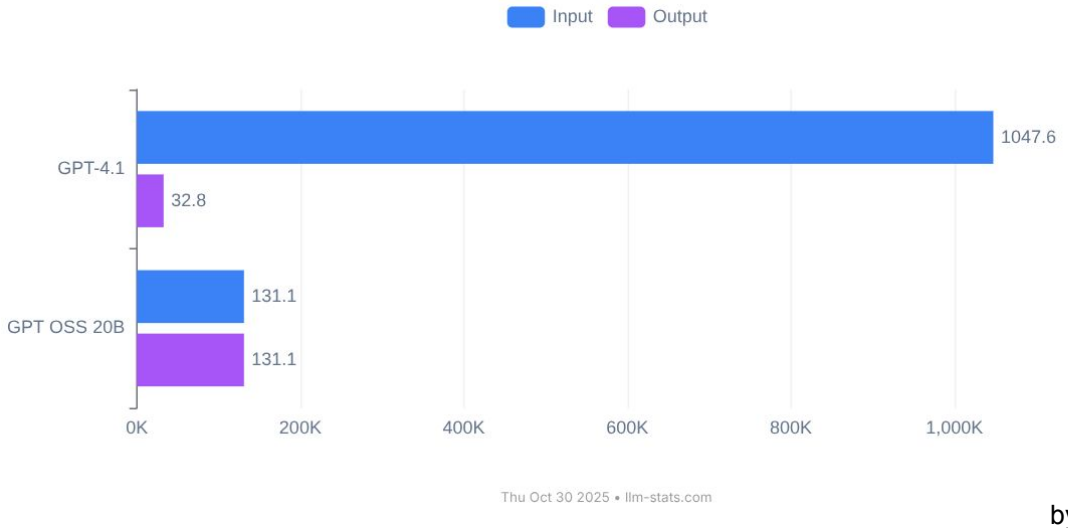
性能



by llm-stats

- GPQA
  - 48問の選択式問題からなる難易度の高いデータセット。
  - 生物学、物理学、化学の専門家によって作成された高品質な問題を収録。
  - 博士号取得者や博士課程の専門家でも正答率は65%、明らかなミスを除外しても74%にとどまる
- Humanity’s Last Exam
  - 人間の知識の最先端を反映した難問を集めたテスト
  - 100以上の科目（数学、科学、文学、芸術など）にわたる2,700問の挑戦的な問題

最大コンテキスト長



by llm-stats

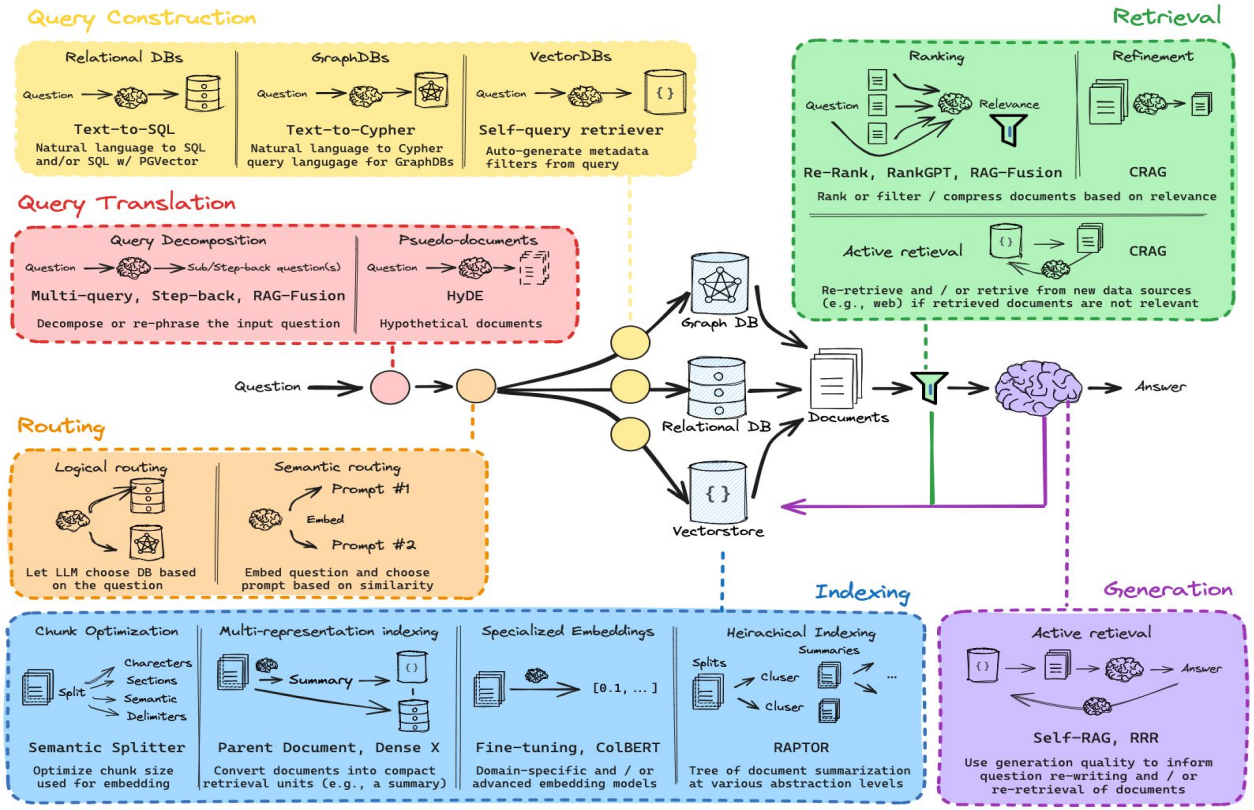
gpt-4.1 (2025-04-14)	米国中北部 スウェーデン中部	✅	入力: 128,000 出力: 16,384 トレーニング用コンテキストの例の長さ: 65,536	2024 年 5 月	テキストと視覚テキスト
gpt-4.1-mini (2025-04-14)	米国中北部 スウェーデン中部	✅	入力: 128,000 出力: 16,384 トレーニング用コンテキストの例の長さ: 65,536	2024 年 5 月	テキスト間
gpt-4.1-nano (2025-04-14)	米国中北部 スウェーデン中部	✅	入力: 128,000 出力: 16,384 トレーニング例のコンテキスト長: 32,768	2024 年 5 月	テキスト間

Azure AI Foundry より

Embedding はなぜ大事？

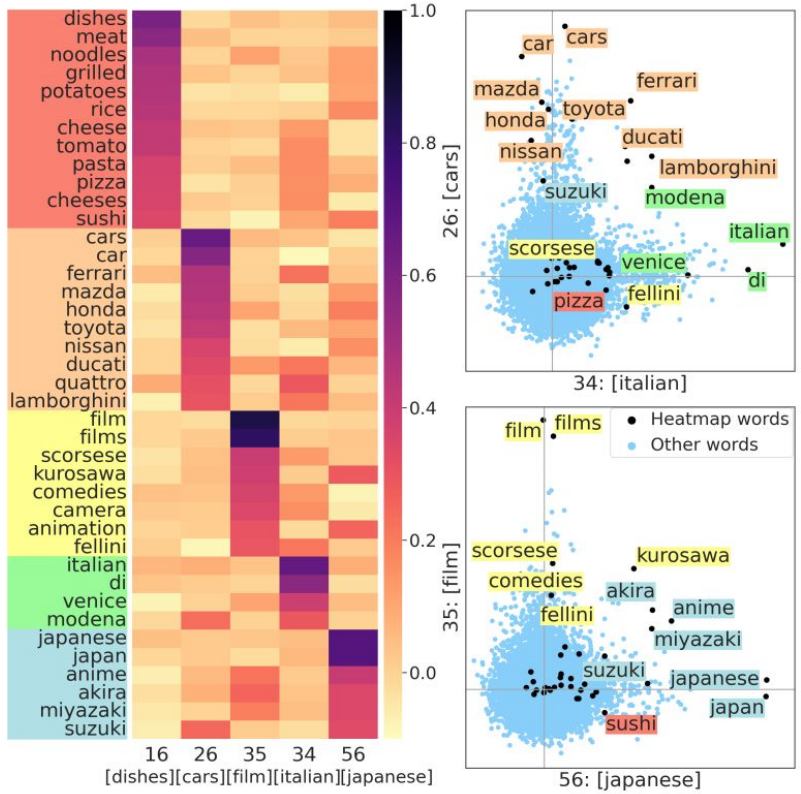
- Embedding (ここでは文章埋め込みの意味)は意味的類似性や構文的関係を測定できる潜在空間の表現であって、その幾何的特徴量が重要
- 最近では、単語埋め込みに普遍的な幾何的性質を利用して、解釈性の高い成分を抽出される手法が（少なくとも日本では）注目されている[Yamagiwa+2023]
- 実務では、クラスタリング、RAG 等、への応用があり、いずれも情報処理精度向上にあたっても Emebdding モデルの性能が鍵

一般的なRAGのフレームワーク



https://github.com/langchain-ai/rag-from-scratch

ICA (独立成分分析)による意味成分抽出



https://doi.org/10.18653/v1/2023.emnlp-main.283

今使っている埋め込みモデルは良くない？

- MTEB（Embedding Model のタスク性能を評価するリーダーボード）では、OpenAI の埋め込みモデルの性能は決して高くはない
- Qwen3-Embedding が性能とコンテキスト長の両面で優位。またコードドメイン特化ではSOTA

Rank (Borda)	Model	Zero-shot	Memory Usage (MB)	Number of Parameters	Embedding Dimensions	Max Tokens	Mean (Task)	Mean (TaskType)
1	llama-embed-nemotron-8b	99%	28629	7B	4096	32768	69.46	61.09
2	gemini-embedding-001	99%	Unknown	Unknown	3072	2048	68.37	59.59
3	Qwen3-Embedding-8B	99%	28866	7B	4096	32768	70.58	61.69
4	Qwen3-Embedding-4B	99%	15341	4B	2560	32768	69.45	60.86
5	Qwen3-Embedding-0.6B	99%	2272	595M	1024	32768	64.34	56.01
9	embeddinggemma-300m	99%	578	307M	768	2048	61.15	54.31
18	text-embedding-3-large	⚠️ NA	Unknown	Unknown	3072	8191	58.93	51.41
35	text-embedding-3-small	⚠️ NA	Unknown	Unknown	1536	8191	54	47.18
47	granite-embedding-278m-multilingual	96%	530	278M	768	512	53.74	47.22
173	text-embedding-ada-002	⚠️ NA	Unknown	Unknown	1536	8191	⚠️ NA	⚠️ NA

by MTAB

複数の embedding を使い分ける

- RAG においてHybrid 検索は検索精度（Context Recall）を高める上で重要な手法であり、表現が異なる複数のモデルを組み合わせたほうがいい（スパースな埋め込みでも勿論◎）

日本語特化

Model	Avg.	Retrieval
sbintuitions/sarashina-embedding-v1-1b	**75.50**	**77.61**
OpenAI/text-embedding-3-large	74.05	74.48
jinaai/jina-embeddings-v3	73.44	75.22
cl-nagoya/ruri-large	73.31	73.02
pkshatech/GLuCoSE-base-ja-v2	72.23	73.36
pkshatech/RoSEtta-base-ja	72.04	73.21
cl-nagoya/ruri-base	71.91	69.82
cl-nagoya/ruri-small	71.53	69.41
intfloat/multilingual-e5-large	70.90	70.98
OpenAI/text-embedding-3-small	69.18	66.39
intfloat/multilingual-e5-base	68.61	68.21
intfloat/multilingual-e5-small	67.71	67.27
pkshatech/GLuCoSE-base-ja	67.29	59.02
OpenAI/text-embedding-ada-002	67.21	64.38

by JMTEB

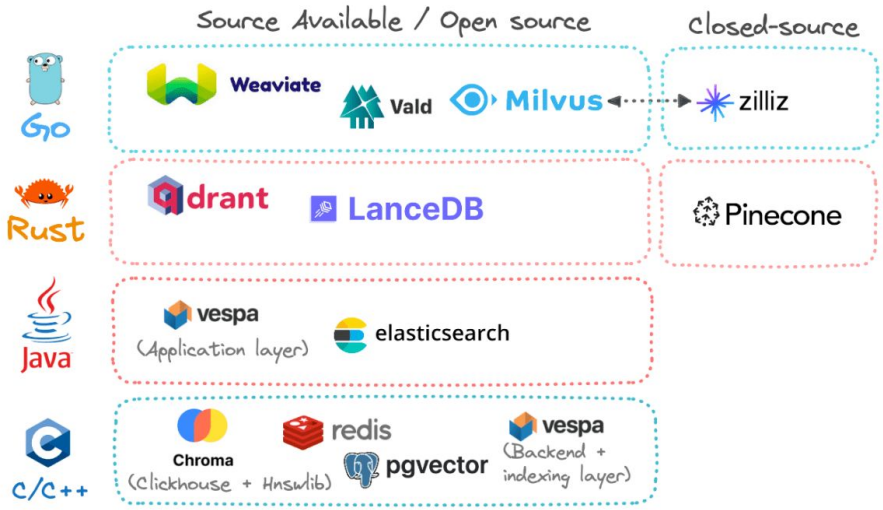
金融ドメイン特化

Model	Overall (EN)
Fin-e5	0.692
voyage-3-large	0.692
text-embedding-3-large	0.675
e5-mistral-7b-instruct	0.654
NV-Embed v2	0.653
bge-large-en-v1.5	0.647
gte-Qwen1.5-7B-instruct	0.641
bge-m3	0.638
text-embedding-3-small	0.633
bge-en-icl	0.619
all-MiniLM-L12-v2	0.613

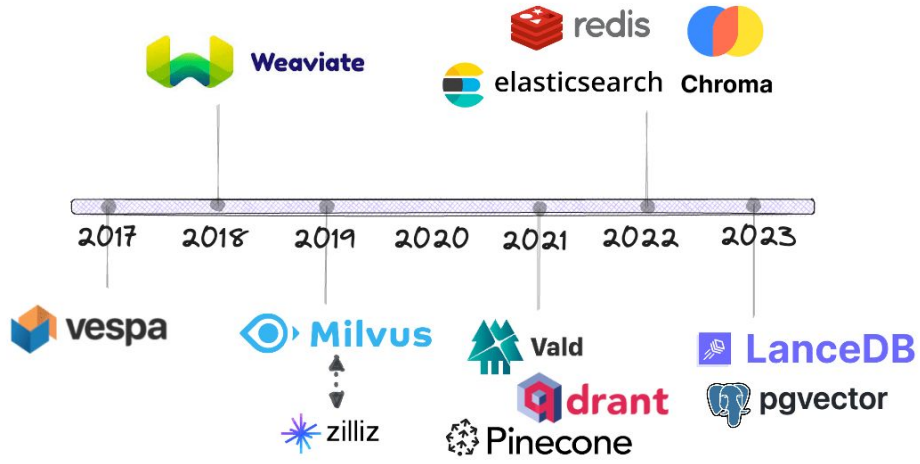
by FinMTEB

ついでに Vector Store

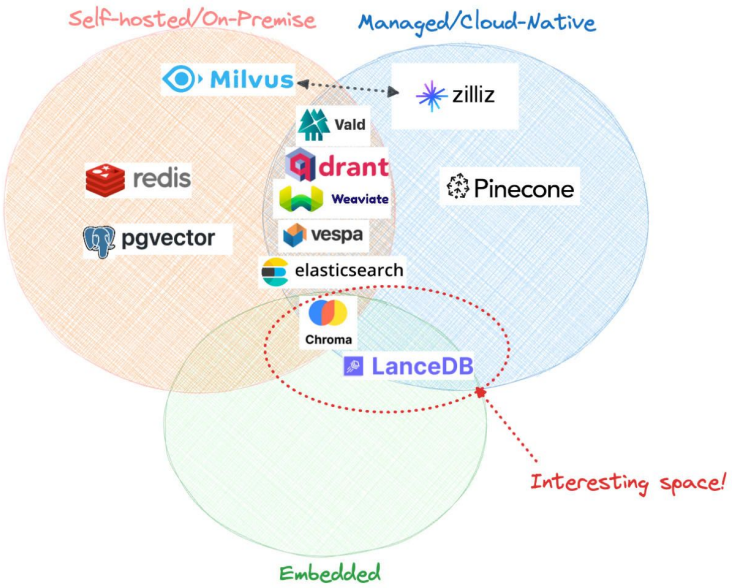
開発言語



リリース日



ホスティング方法

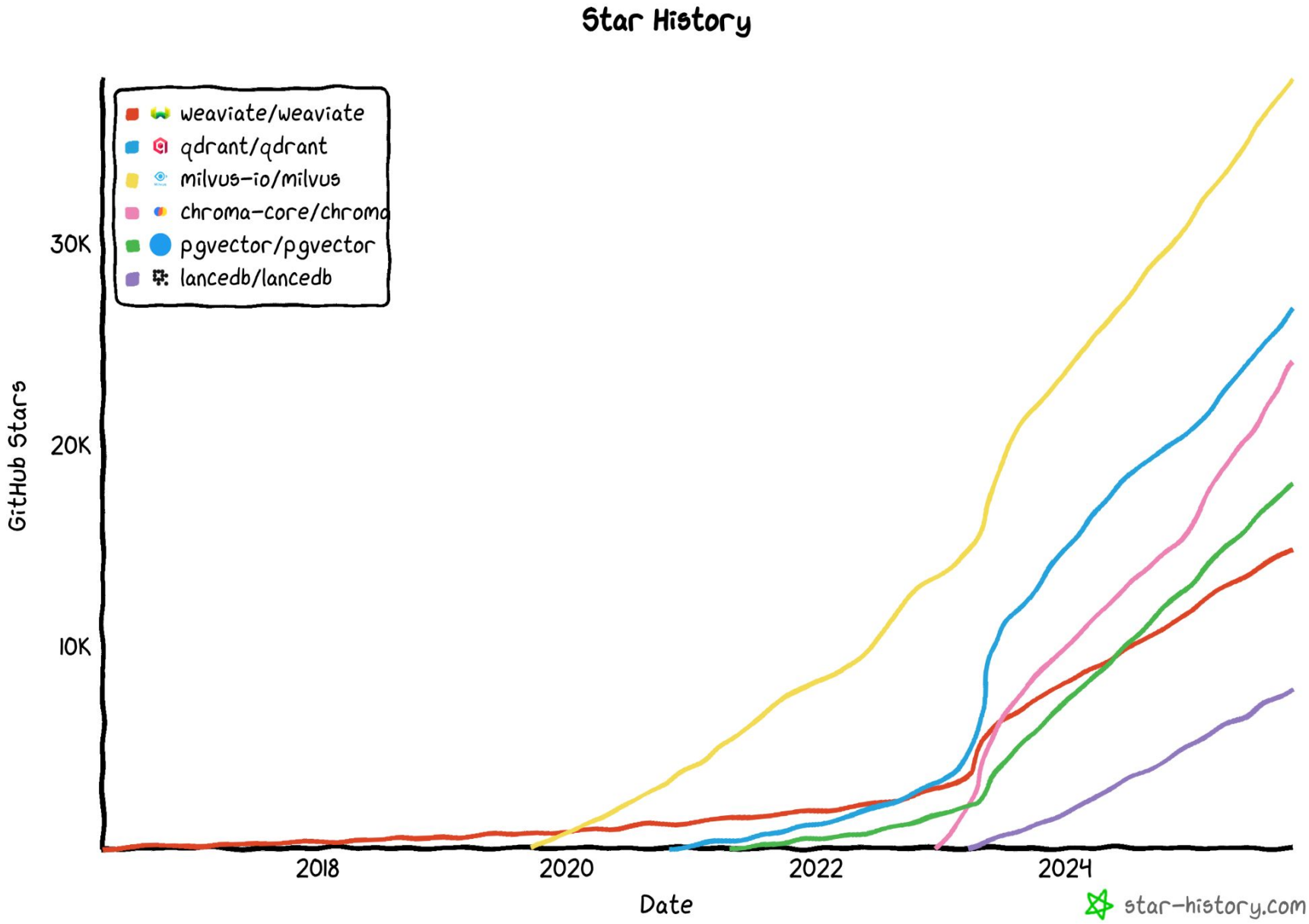


インデクシング方法

- Pinecone ..... Proprietary composite index
- milvus / zilliz ..... Flat, Annoy, IVF, HNSW/RHNSW (Flat/PQ), DiskANN
- Weaviate ..... Customized HNSW, HNSW (PQ), DiskANN (in progress...)
- qdrant ..... Customized HNSW
- chroma ..... HNSW
- LanceDB ..... IVF (PQ), DiskANN (in progress...)
- vespa ..... HNSW + BM25 hybrid
- Vald ..... NGT
- elasticsearch ..... Flat (brute force), HNSW
- redis ..... Flat (brute force), HNSW
- pgvector ..... IVF (Flat), IVF (PQ) in progress...

https://thedataquarry.com/blog/vector-db-1/

ついでに Vector Store



ともに挑む。ともに実る。

MIZUHO