

Performance of FileMaker with and without AI

AI in FileMaker - Performance Perspective

Performance Optimization Story: Parker

FileMaker 2025 Performance News

HOnza Koudelka, 24U Software

HOnza Koudelka

- Co-founder and co-owner of 24U
- FileMaker developer since 1991
- Mad Optimizer
- Achiever of the Impossible...
- Speaking at FileMaker events since 2011
- For more about me see honza.guru





AI in FileMaker

Performance Perspective



New FileMaker Built-In AI Features

- Vector Embeddings & Semantic Search
- Perform Find or SQL Query by Natural Language
- Generate Response From Model
- Retrieval Augmented Generation
- Model Fine-Tuning
- Built-In Model Server

Vector Embeddings & Semantic Search

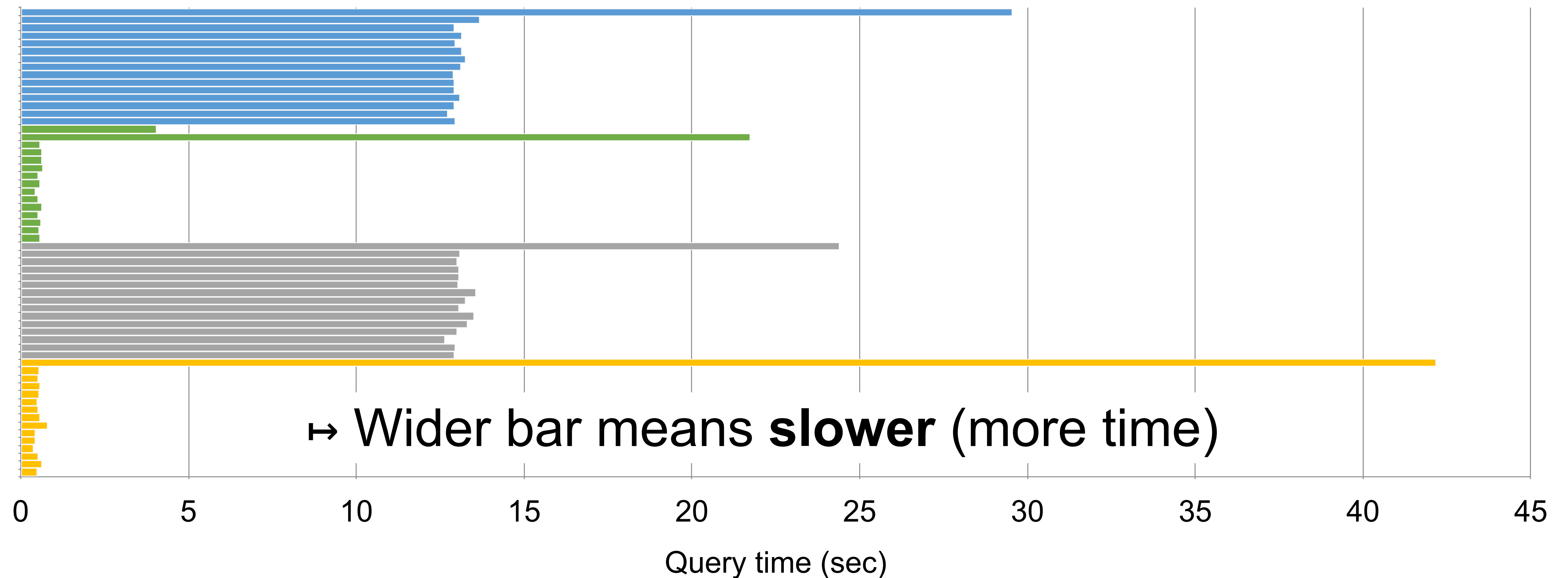


Filmoteka (honza.24u.cz)	
Filmotéka	Cute christmas comedy suitable for children
Elf	0,5153   
Santa Klaus	0,4831   
Bad Santa	0,4734   
The Santa Clause 3: The Escape Clause	0,4703   
Bob the Builder: A Christmas to Remember	0,4690   
The Santa Clause 2	0,4508   
Christmas in Boston	0,4408   

Semantic Search by Container Storage

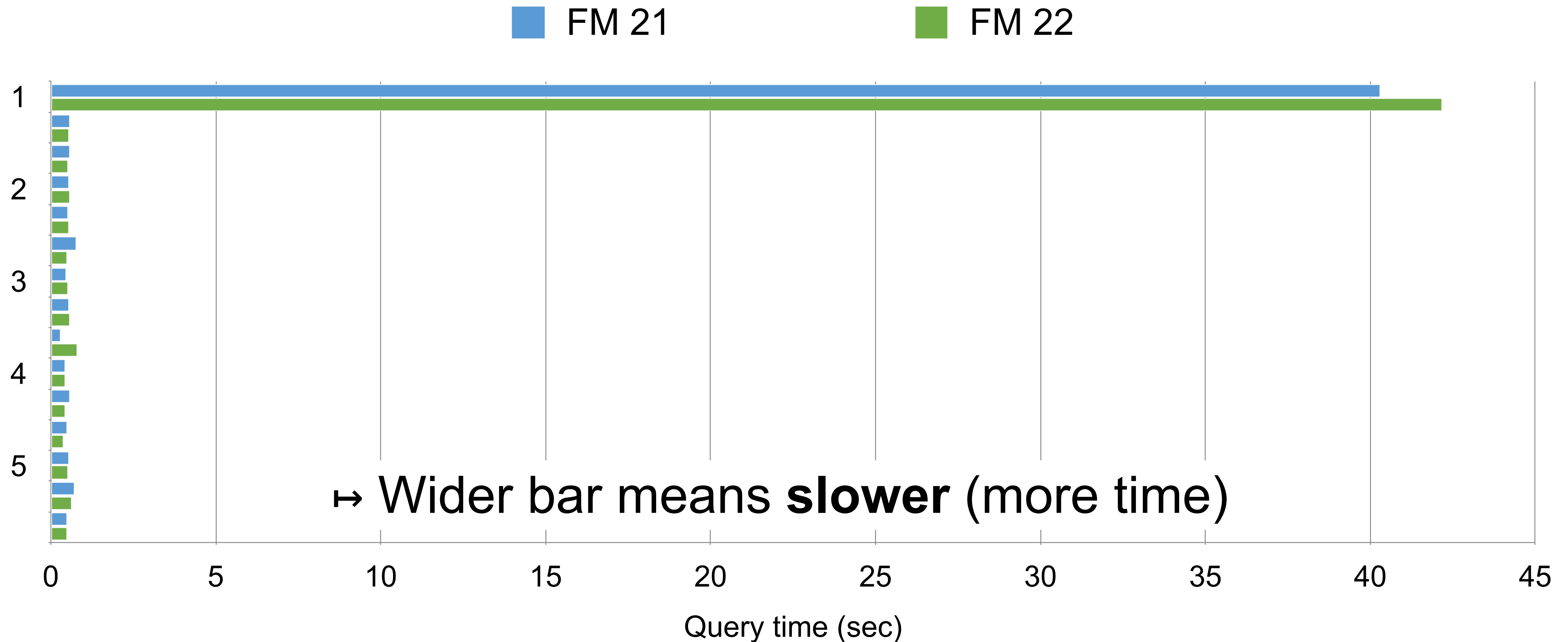
5 different queries, each repeated 3 times, 1:1 related field

Local External Local Internal Related External Related Internal

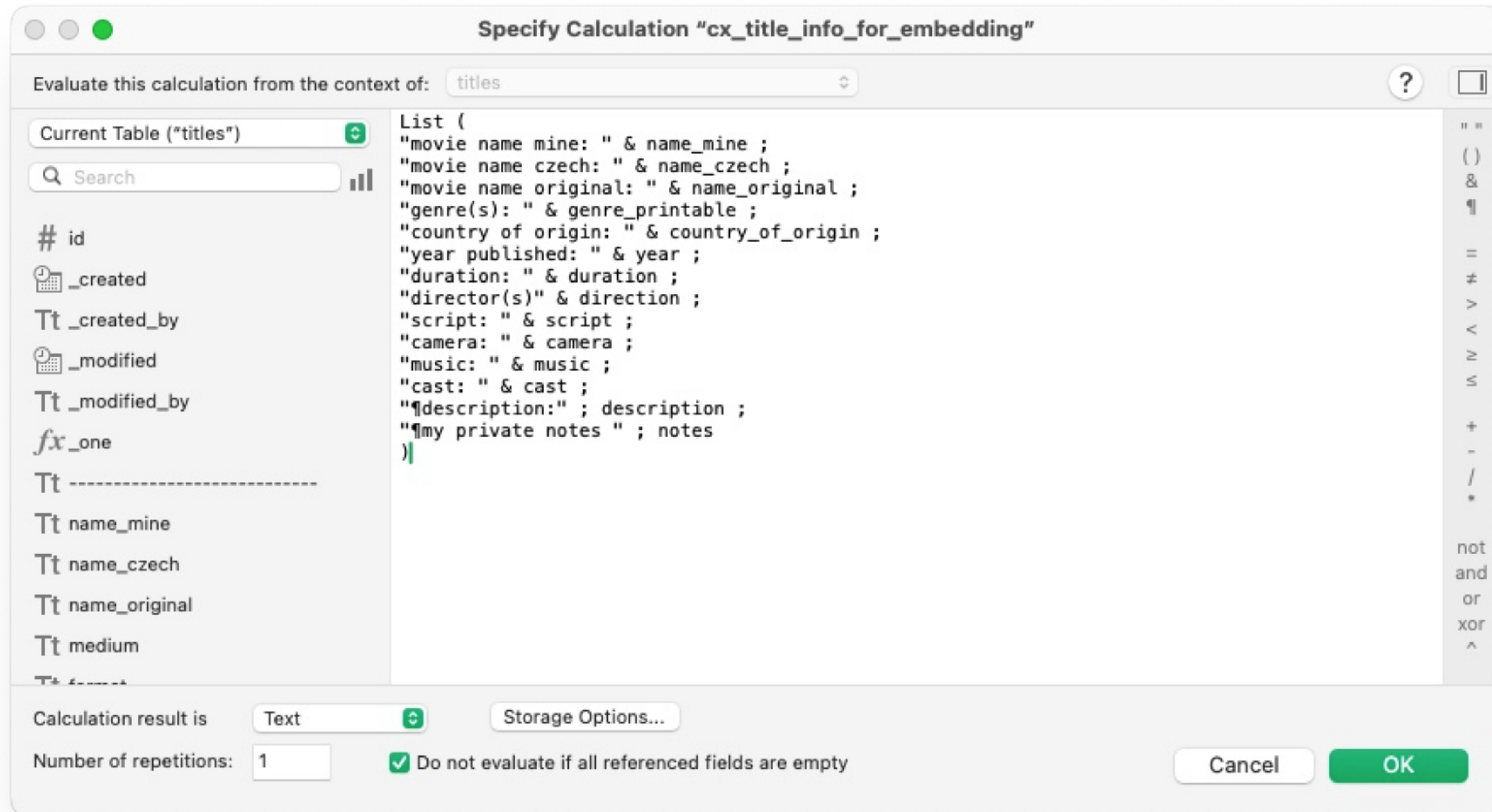


Semantic Search FM 21 vs FM 22 over WAN

5 different queries, each repeated 3 times, 1:1 related field



Single combined value for embedding



Asked ChatGPT to recommend models

What embedding model available for download from HuggingFace is the closest match to OpenAI's text-embedding-3-large in terms of accuracy embedding English and Czech text?

Thought for 58s

Short answer: `intfloat/multilingual-e5-large` (or its instruction-tuned variant `multilingual-e5-large-instruct`) is the closest downloadable HuggingFace match to OpenAI's `text-embedding-3-large` for high-accuracy embeddings in **English and Czech**.

Why this pick

- **State-of-the-art multilingual accuracy:** The MMTEB paper finds the best *publicly available* model to be `multilingual-e5-large-instruct` (≈ 560 M params) across massive multilingual tasks. Hugging Face
- **Designed for 100 languages (via XLM-RoBERTa-large):** The model card explains it's initialized from `XLM-R-large` and inherits its 100-language coverage (which includes Czech). Hugging Face +1
- **Well-documented & easy to use:** It's shipped on HuggingFace with examples; use E5's instruction prefixes (e.g. `query:` vs `passage:`) for best retrieval performance. Hugging Face

Models I tried

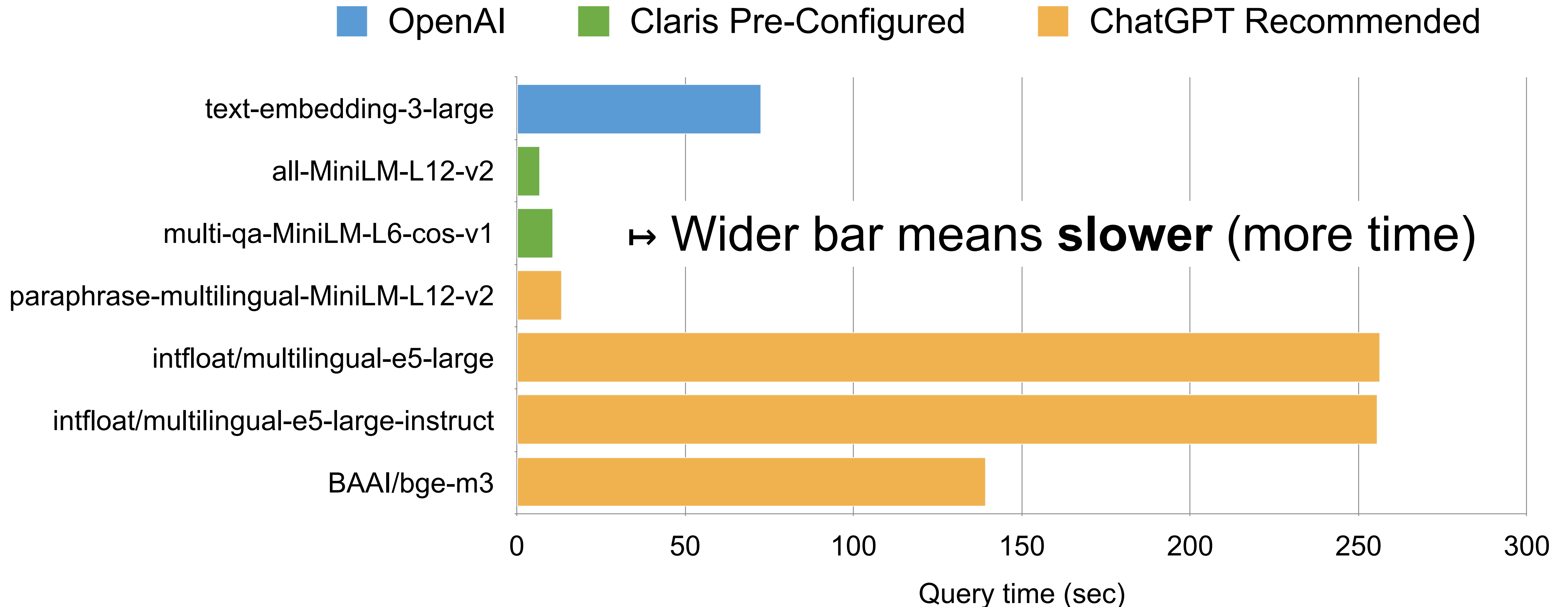
- OpenAI
 - text-embedding-3-large
- Claris pre-configured
 - sentence-transformers/all-MiniLM-L12-v2
 - sentence-transformers/multi-qa-MiniLM-L6-cos-v1
- ChatGPT recommended as multi-lingual alternative to all-MiniLM-L12-v2
 - sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2
- ChatGPT recommended as closest offline match for OpenAI's text-embedding-3-large
 - intfloat/multilingual-e5-large
 - intfloat/multilingual-e5-large-instruct
 - BAAI/bge-m3

Performance by Model

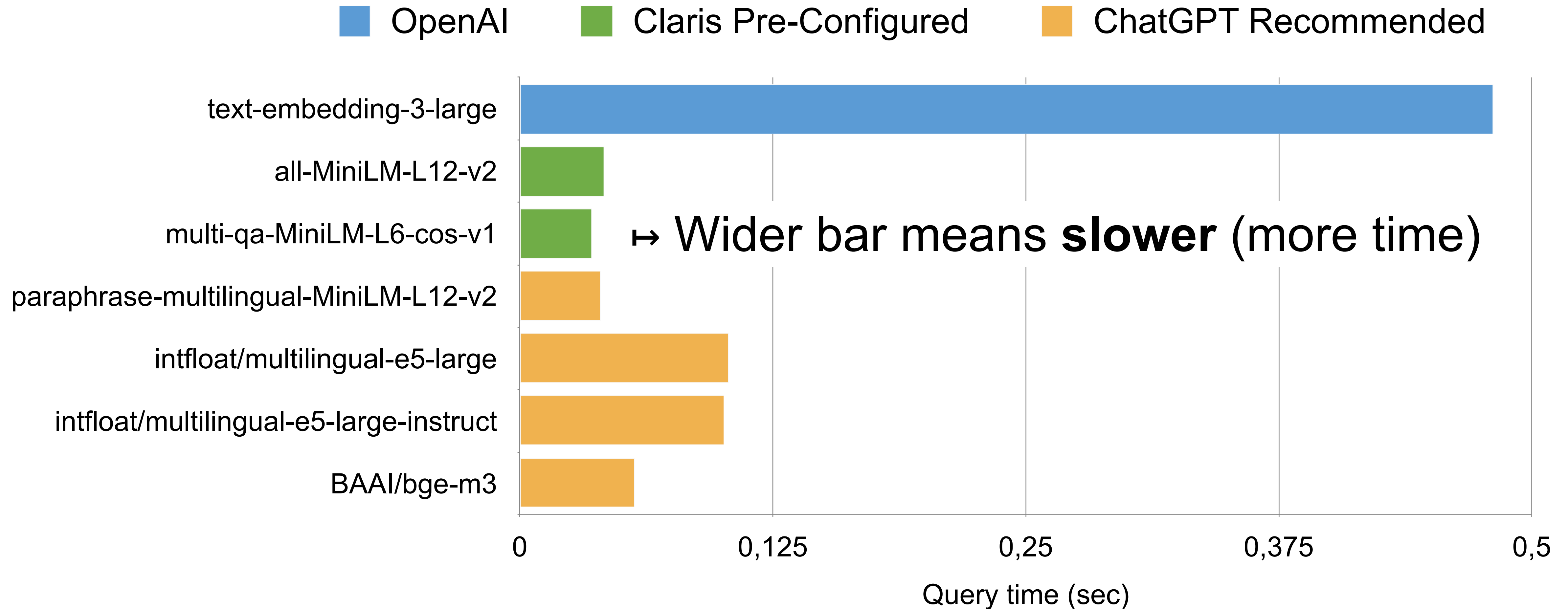
1073 movies, searching for top 100 matching query, time in seconds

Model	Size	Embedding	Semantic Search
OpenAI/text-embedding-3-large	N/A	72,676338	0,480949
sentence-transformers/all-MiniLM-L12-v2	134.2 MB	6,873841	0,041680
sentence-transformers/multi-qa-MiniLM-L6-cos-v1	91.6 MB	10,674411	0,035676
sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2	479.7 MB	13,238713	0,040026
intfloat/multilingual-e5-large	2.26 GB	256,310571	0,103274
intfloat/multilingual-e5-large-instruct	1.14 GB	255,668089	0,101083
BAAI/bge-m3	4.56 GB	139,393981	0,056764

Generating Vector Embeddings for 1073 movies



Semantic Search for Top 100 out of 1073 Movies



OpenAI text-embedding-3-large - not too bad, but...

Filmoteka (honza.24u.cz)

Filmotéka   Cute christmas comedy suitable for children 

Elf				0,5151			
Santa Klaus				0,4829			
Bad Santa				0,4734			
Bob the Builder: A Christmas to Remember				0,4708			
The Santa Clause 3: The Escape Clause				0,4700			
The Santa Clause 2				0,4503			
Christmas in Boston				0,4415			

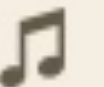
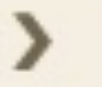
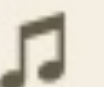
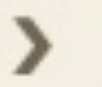
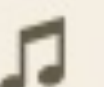
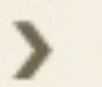
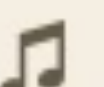
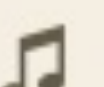
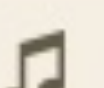
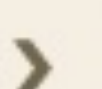
all-MiniLM-L12-v2 - slightly worse

Filmoteka (honza.24u.cz)							
Filmotéka				Cute christmas comedy suitable for children			
Last Christmas				0,5310			
Bad Santa				0,4851			
The Santa Clause 2				0,4811			
Santa Klaus				0,4385			
Christmas in Boston				0,4162			
The Santa Clause 3: The Escape Clause				0,4104			
Bob the Builder: A Christmas to Remember				0,4046			

multi-qa-MiniLM-L6-cos-v1 - absolutely worst

Filmoteka (honza.24u.cz)

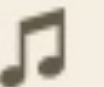
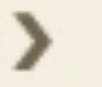
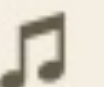
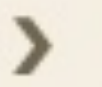
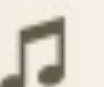
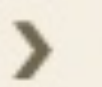
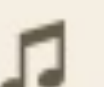
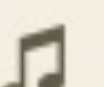
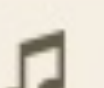
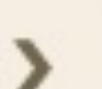
Filmotéka   Cute christmas comedy suitable for children 

The Santa Clause 2				0,3428			
Christmas in Boston				0,3316			
Santa Klaus				0,3271			
Last Christmas				0,3239			
Bad Santa				0,3052			
Happy Gilmore				0,2717			
The Santa Clause 3: The Escape Clause				0,2680			

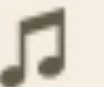
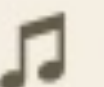
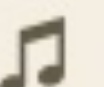
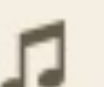
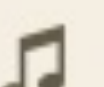
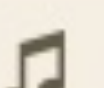
paraphrase-multilingual-MiniLM-L12-v2 - close to OpenAI

Filmoteka (honza.24u.cz)

Filmotéka   Cute christmas comedy suitable for children 

Bad Santa				0,4392			
The Santa Clause 2				0,4311			
The Santa Clause 3: The Escape Clause				0,4162			
Christmas in Boston				0,4080			
Bob the Builder: A Christmas to Remember				0,4030			
Last Christmas				0,3966			
Santa Klaus				0,3681			

intfloat/multilingual-e5-large - higher cosine similarity

Filmoteka (honza.24u.cz)						
Filmotéka				Cute christmas comedy suitable for children		
Bad Santa				0,7971		 
Christmas in Boston				0,7846		 
Santa Klaus				0,7845		 
Malí poníci a jejich přátelé				0,7832		 
Joyeuses Pâques				0,7823		 
Bob the Builder: A Christmas to Remember				0,7821		 
Elias: The Little Rescue Boat				0,7789		 

intfloat/multilingual-e5-large-instruct - winner

Filmoteka (honza.24u.cz)

Filmotéka



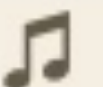
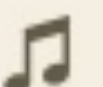
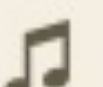
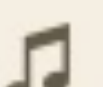
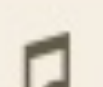
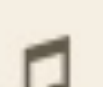


Cute christmas comedy suitable for children

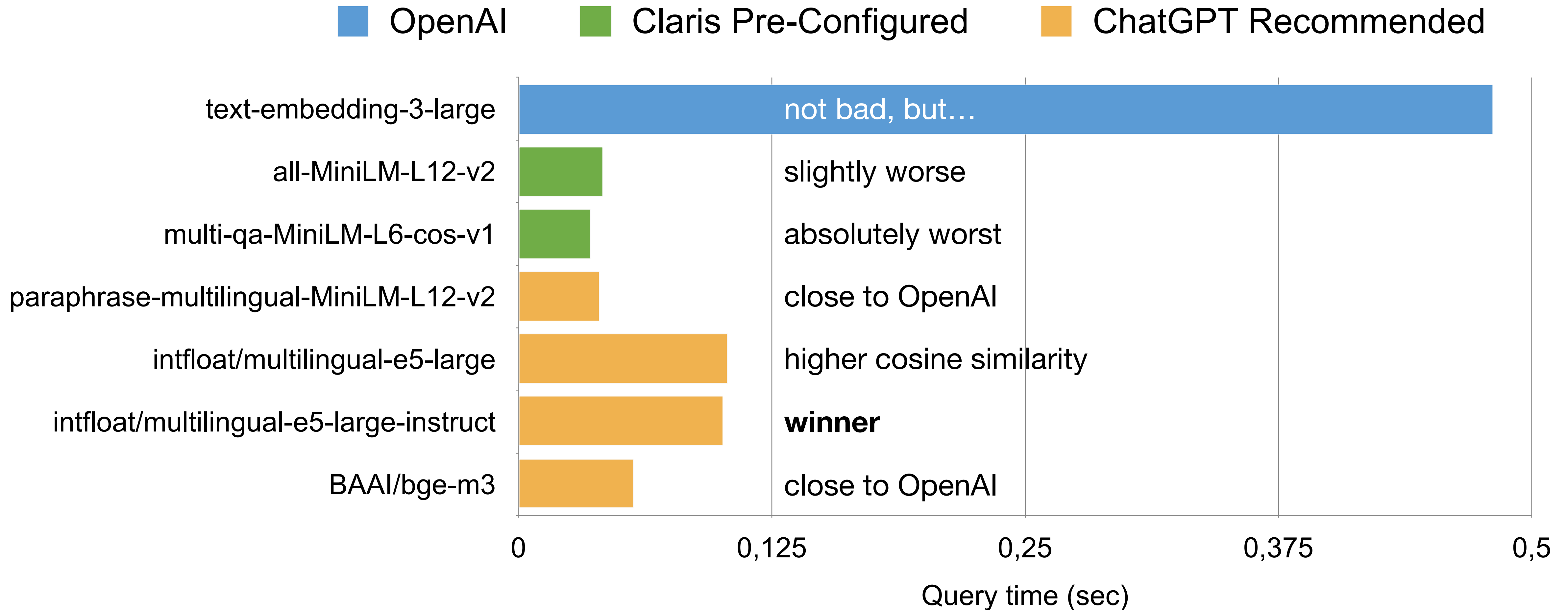


Santa Klaus				0,8056			
Thomas & Friends: Calling All Engines!				0,8020			
The Santa Clause 3: The Escape Clause				0,8019			
Bad Santa				0,8018			
The Santa Clause 2				0,8009			
Bob the Builder: A Christmas to Remember				0,7986			
Elf				0,7971			

BAAI/bge-m3 - close to OpenAI

Filmoteka (honza.24u.cz)									
Filmotéka				Cute christmas comedy suitable for children					
Bad Santa				0,5354					
Last Christmas				0,5316					
Elf				0,5197					
The Santa Clause 2				0,5125					
Santa Klaus				0,5101					
Christmas in Boston				0,5100					
Bob the Builder: A Christmas to Remember				0,5058					

Let's put it into context...



Claris AI Model Server

AI Services

AI services manage downloading open source models and loading them into memory, perform inference using loaded models, perform retrieval-augmented generation (RAG), fine-tune base models with your own training data, and manage API keys to control access to the Model Server.

- Model Server
- Keys
- Models
- Fine-Tuned Models

Model Server

The AI Model Server performs embedding generation, text and query generation, retrieval-augmented generation (RAG), and fine-tuning without sending your data to third-party model providers.

Status	Running	
Endpoint	https://honza.24u.cz/llm/v1/	Copy
Hugging Face Token	hf_...iri	Change Revoke

Settings

Services

Embedding	Enabled	
Generation	Enabled	
RAG	Enabled	
Fine-Tuning	Enabled	

Stress-testing the AI Model Server

Keep AI Server Busy

Keep AI Server Busy

Text Embedding

Image Embedding

Text Processing

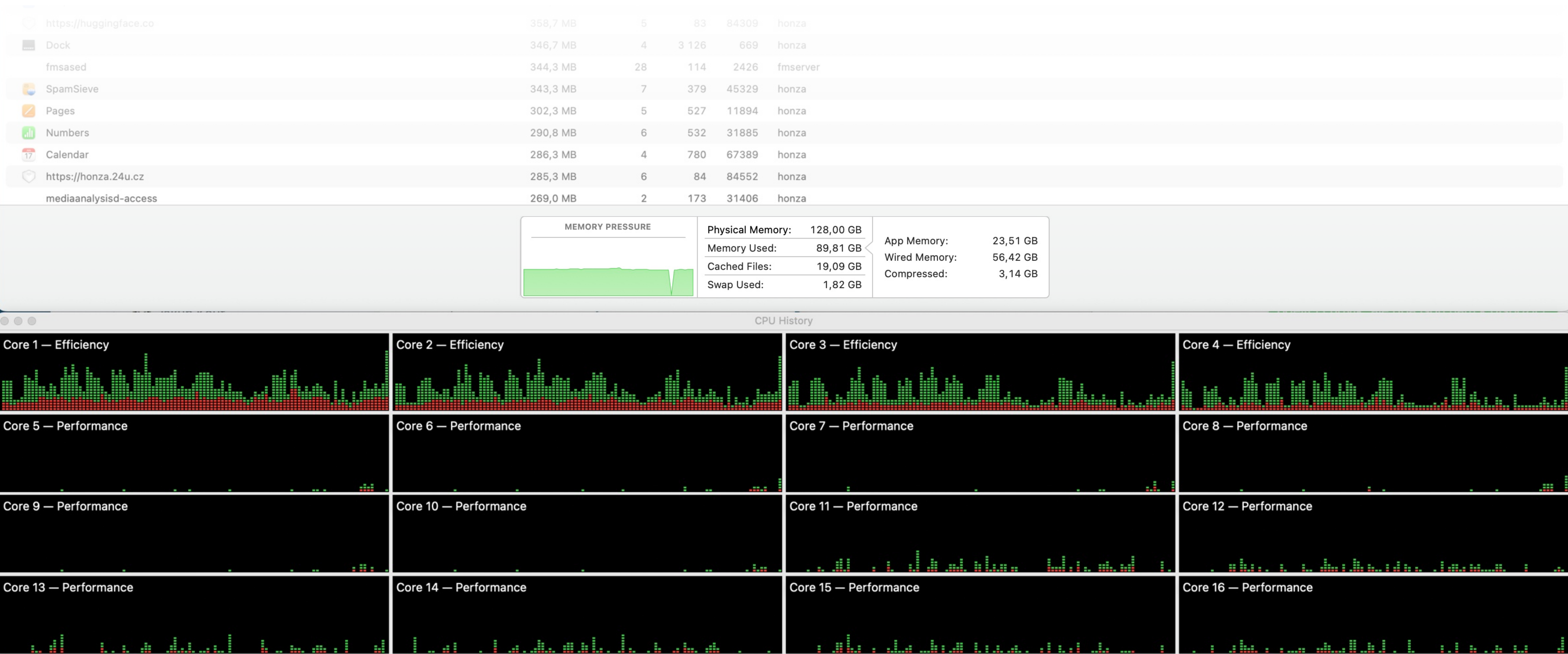


A young woman stands in a field, looking determined, suggesting a story about resilience or overcoming challenges.

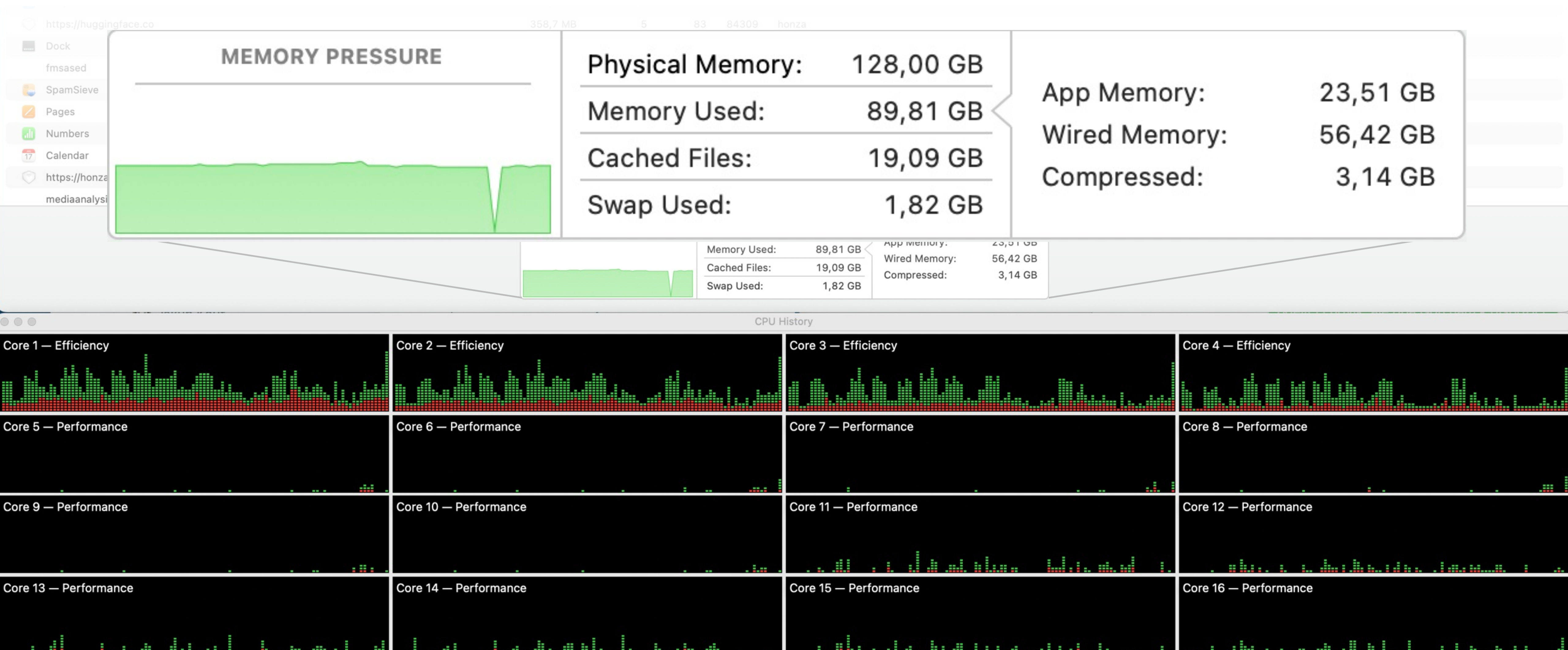
A transgender woman's road trip to pick up her estranged son from prison unexpectedly becomes a journey of self-discovery for both of them.

Summarize image and text in loop

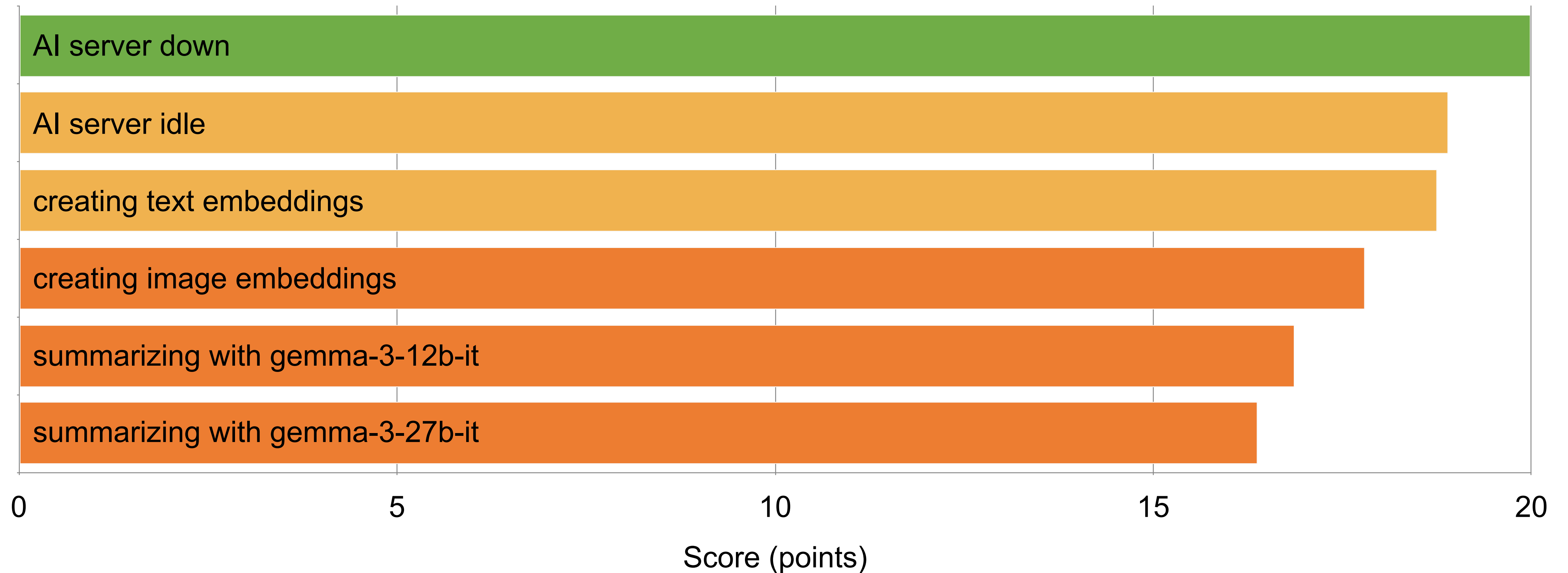
Stress-testing the AI Model Server



Stress-testing the AI Model Server

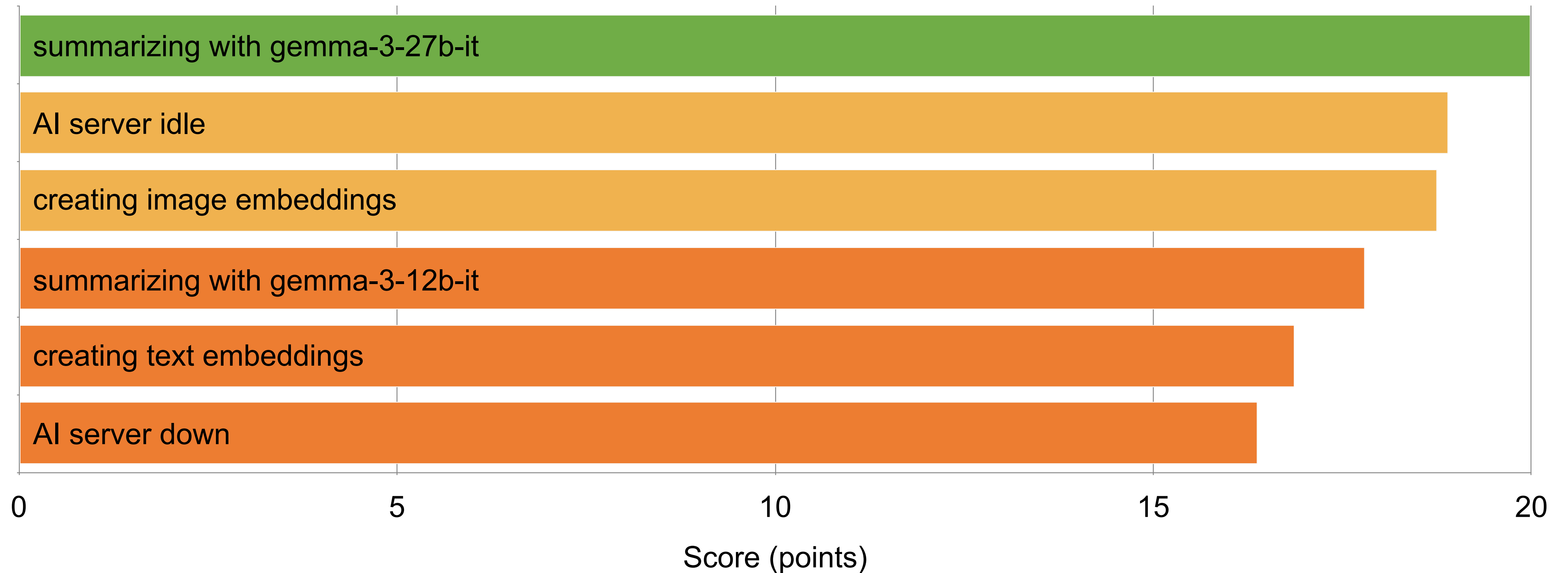


BenchTest - Standard Test (expected)



⇒ Wider bar means **faster**

BenchTest - Standard Test (actual)



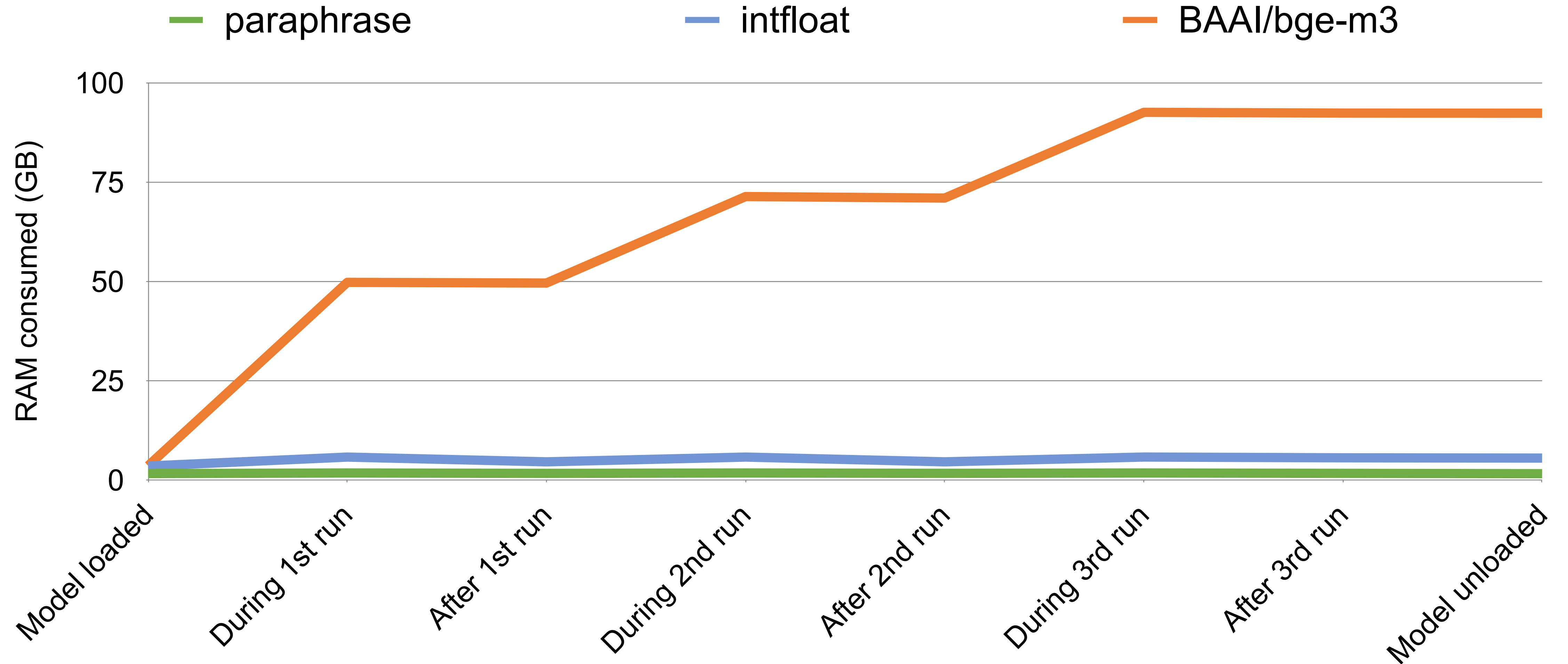
⇒ Wider bar means **faster**

Test Setup

- Configuration
 - MacBook Pro M4 Max 128GB RAM
 - FileMaker Pro 22.0.2.202, FileMaker Server 22.0.1.66
- Models
 - Text embedding: all-MiniLM-L12-v2
 - Image embedding: clip-ViT-B-32
 - Text Generation: google/gemma-3-12b-it & google/gemma-3-27b-it

Embedding RAM Consumption

Generating vector embeddings for found set of 1073 movies



Best Practices

- Claris White Paper: Running your own AI Server - <https://24usw.com/cais>
 - Use dedicated machine for AI Model Server, disable access to databases
 - Disable AI Model Server in Admin Console for database server
- Ask the right people
 - Ian Jempson
 - Cris Ippolite
 - Wim Decorte
 - Ernest Koe & Todd Geist
 - Chris Moyer & Heidi Porter
 - Michael Wallace



Image Summarization by google/gemma-3-27b-it



Image Summarization by google/gemma-3-27b-it

A determined woman stands in a futuristic cityscape, suggesting a science fiction action or thriller movie.



A group of people are running through a forest, suggesting an action or thriller movie.

A young woman with fiery red hair stands defiant against a backdrop of swirling chaos, suggesting a story of rebellion and strength in a fantastical world.

Image Summarization by google/gemma-3-27b-it



A determined woman stands in a futuristic cityscape, suggesting a science fiction action or thriller movie.



A group of people are running through a forest, suggesting an action or thriller movie.



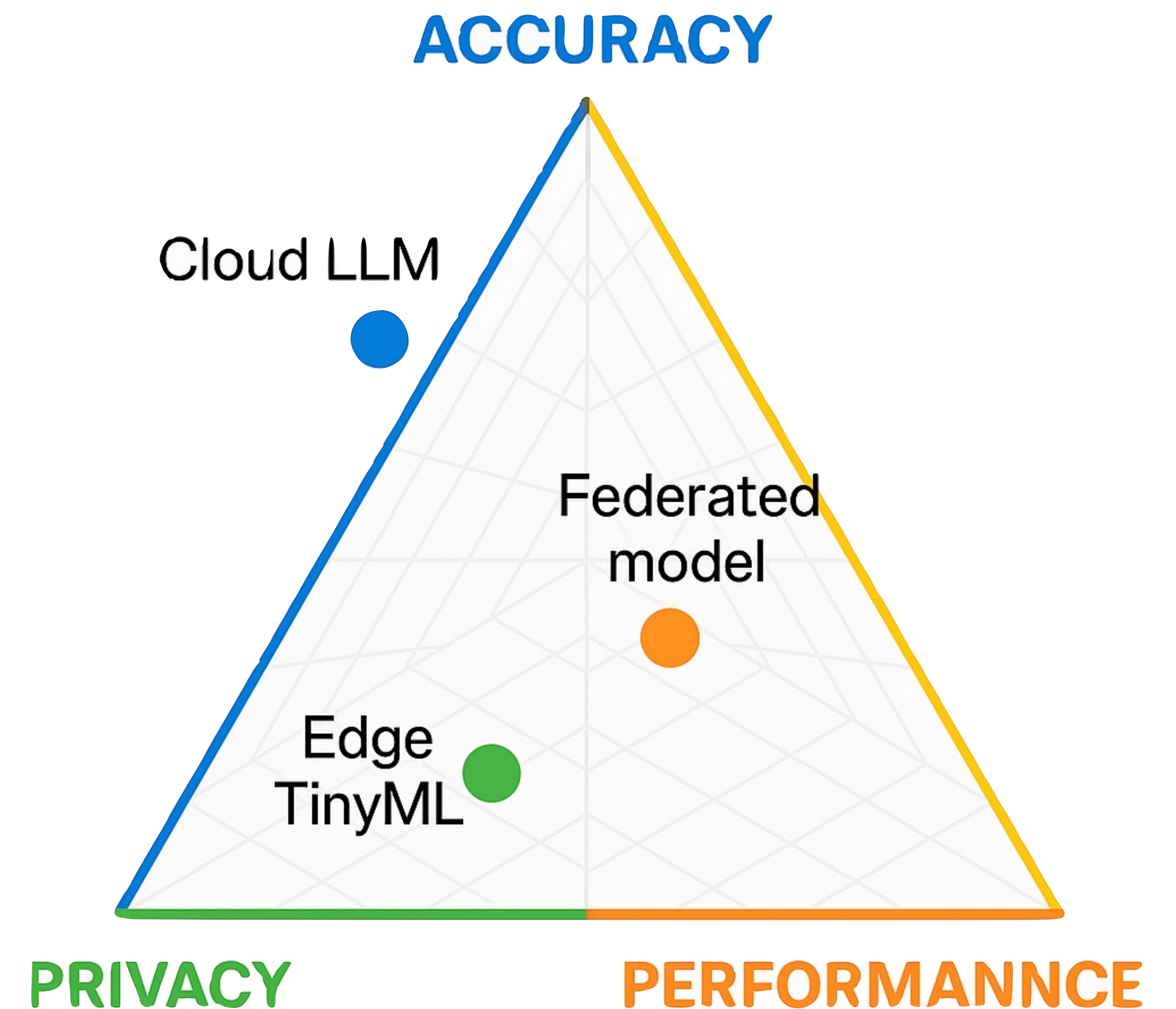
A young woman with fiery red hair stands defiant against a backdrop of swirling chaos, suggesting a story of rebellion and strength in a fantastical world.

New Way of Thinking (not only) about Performance

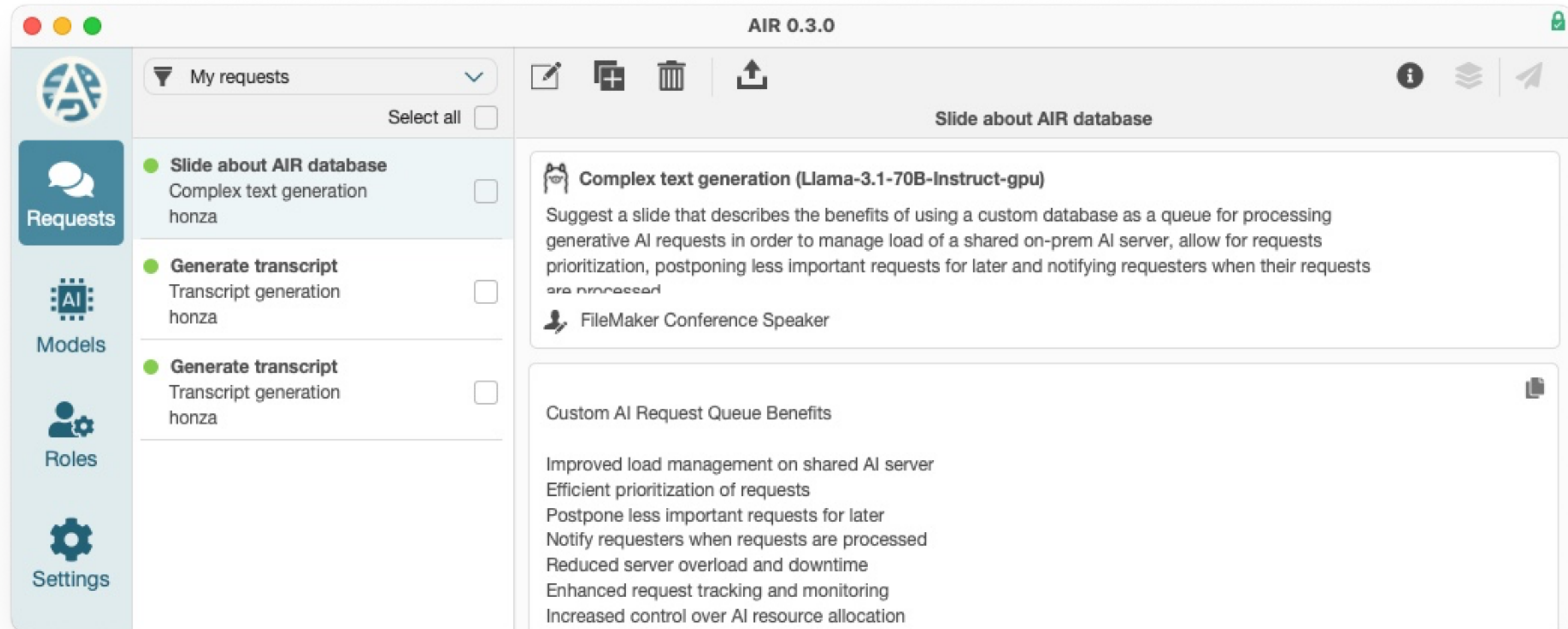
Traditional Features	AI-powered Features
Deterministic: What is a faster way to achieve the exact same result?	Probabilistic: What is the fastest way to achieve sufficiently accurate result?
Simple metrics: Amount of data to process per time	Complex metrics: Number of tokens to process, number of tokens to generate, number of iterations to achieve required accuracy, etc.
Fewer factors: Network latency, network bandwidth, storage bandwidth, processing power	More factors: CPU vs GPU vs MLX, RAM vs VRAM, LLM parameters count vs vector size vs training data size vs training data quality, dedicated vs shared resources

Finding the Right Balance

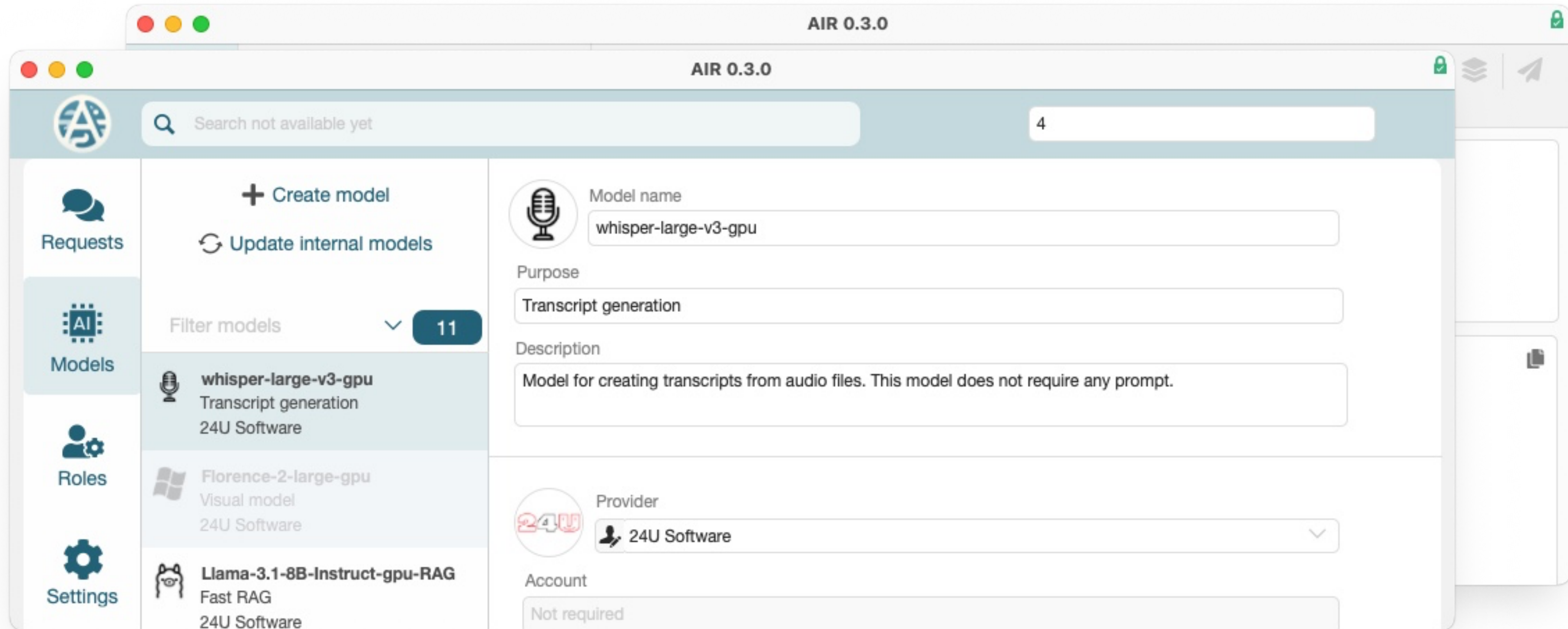
- Accuracy costs performance
- Cloud models offer better accuracy at same performance at the cost of privacy
- Performance can be increased by
 - Simpler model
 - Quantization
 - More resources (and money)



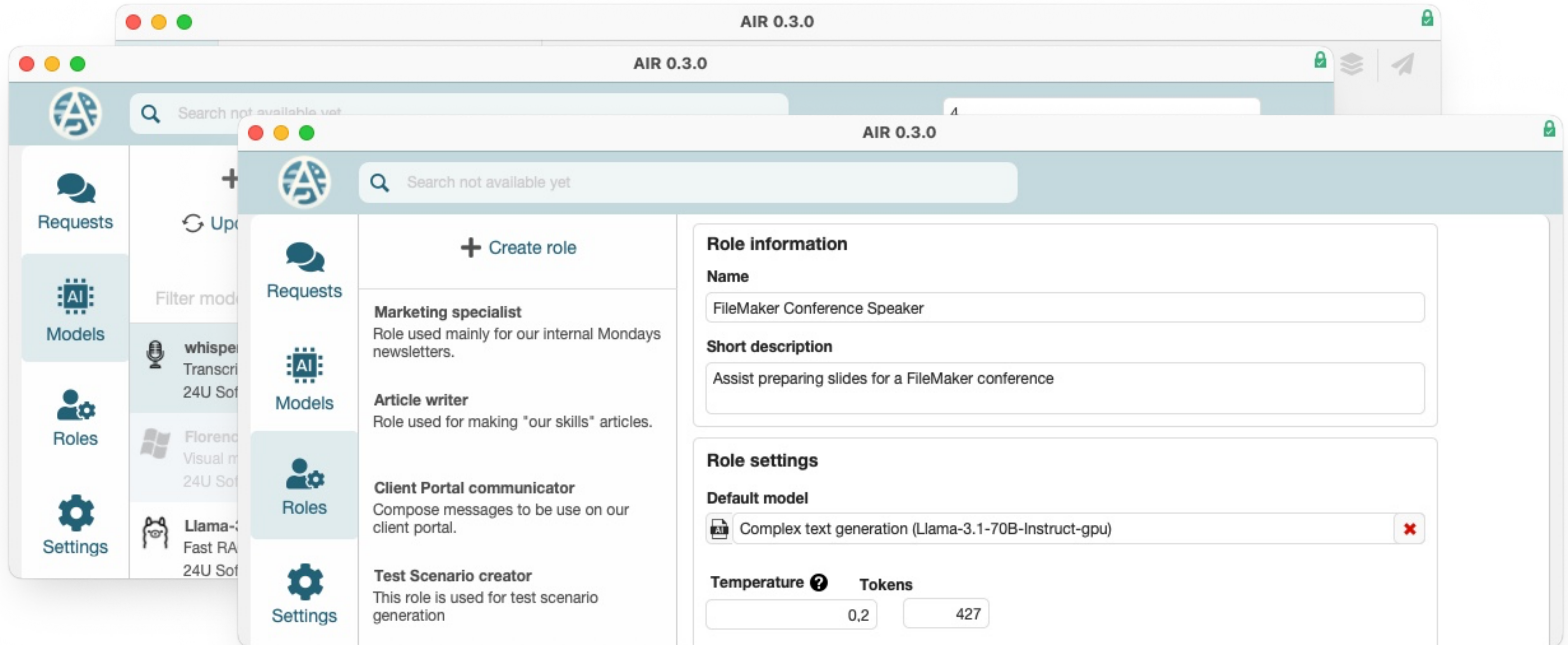
AI Request Dispatcher



AI Request Dispatcher



AI Request Dispatcher



Combining Probabilistic and Deterministic

- Separate embedding fields for title, description, and people
- **Perform Find by Natural Language** to narrow initial found set by exact matches
- **Perform Semantic Find** on each field separately, returning results as JSON
- Fuse results using calculation by calculating combined score for each record:

$$\text{score}(d) = \sum_{\ell \in \{\text{plot}, \text{title}, \text{people}\}} w_{\ell} \cdot \frac{1}{k + \text{rank}_{\ell}(d)}$$

- **Go To List Of Records** to reveal the results
- On larger data sets can result in both higher **accuracy** and better **performance**