Search Search Evolution

The next generation of search engines...

Alexander Reelsen
alr@spinscale.de | @spinscale

Today's goal

> Learn about the trends in search engines
>
> Understand that this is a highly volatile market in the coming years

Status quo

The power of search

- Speed (Search & Suggests)
- Scale (all the internet)
- **Relevance**
- Intent
- Personalization

Evolution of Use-Cases

- Text search
- Enterprise search
- Ecommerce search
- Log search
- Analytics
- Dashboards
- NLP
- Generative/Conversational Search

Relevancy

- SQL: Does row `r` match query `q`?
  - Answer: ✅/❌
- How well matches query `q` document `d`?
  - Answer: `[0..∞]`
- Scoring based on formula: `TF/IDF`, `BM25`
  - Dependent on corpus

Ranking

- Recency
- Rating
- Popularity
- Past (searches/purchases)
- Individualization

Trends

Going cloud native

- SaaS
- Splitting storage and compute
- Using blob storage, segment replication
- **Massive cost savings**

Learning to rank

- Scoring/Relevancy based on machine learning model
- Common: Reranking after first filtering
- Machine Learning models trained independently

Vector Search

- Vector search engines: translates content into vectors
- QDrant, Milvus, Weaviate, Pinecone, Deeplake, nucliadb
- Best model wins...
- Going hybrid: Will search engines add vector support or vector engines add search support?

Don't sleep on SQL engines!

- SQLite: vector extension, FTS3/4 extension
- Postgres: PostgresML - full model management and querying in Postgres!

Search on the edge

- Distributed search across regions
- Search on your browser
- Search on your phone
- Check out OramaSearch

ChatGPT

Generative/Conversational search

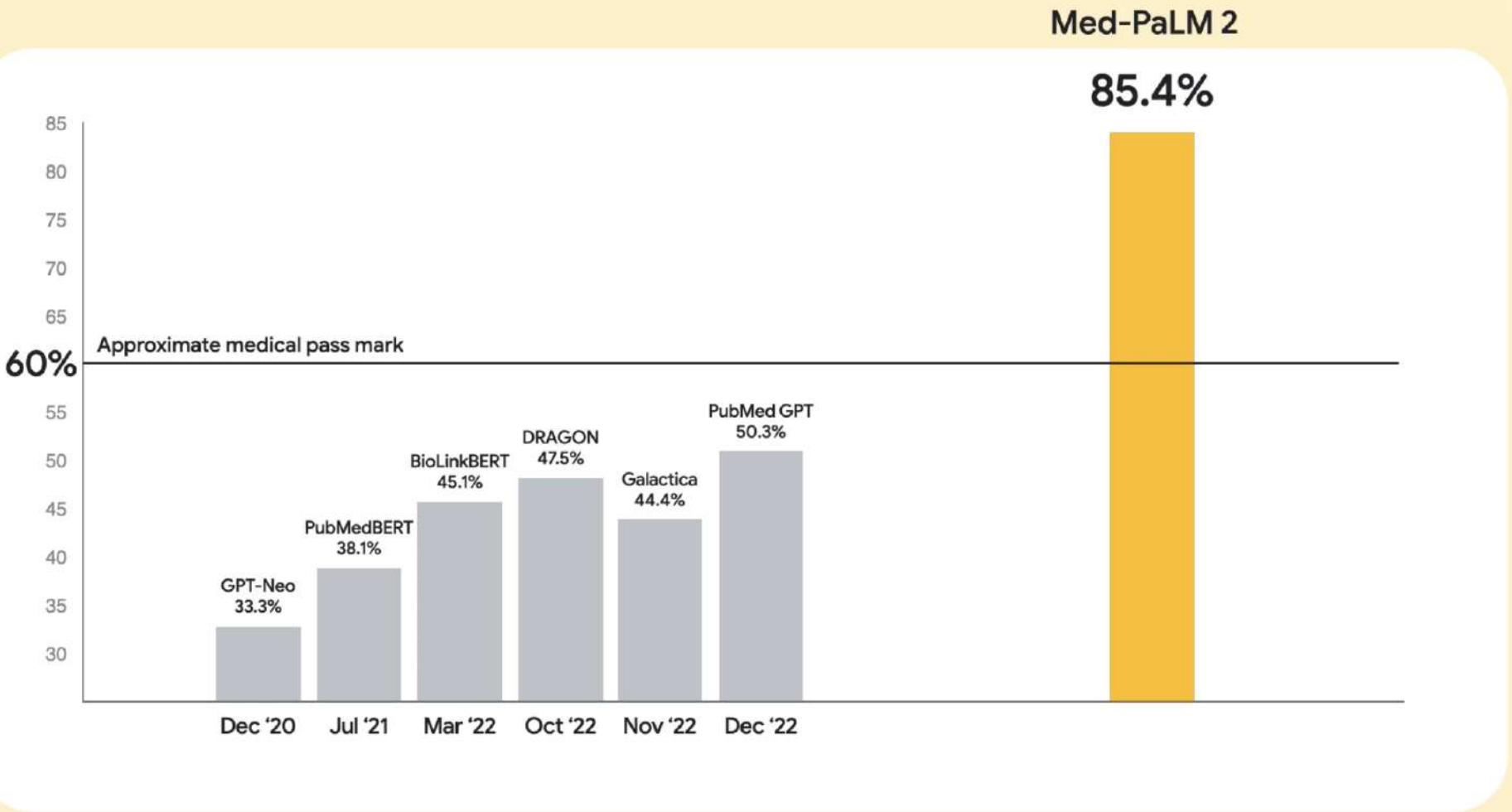
> blue dress with white stripes that has been shown on the last fashion week in milan

> summarize the quarterly earnings call, focus on numbers that differ strongly from the last three quarters

> Convert the following CDK snippet from Java to python

Generative search - context

- `blue dress with white stripes` requires image extraction
- `last fashion week in milan` requires **external** knowledge
- Your own dataset is not enough for a good search!


Med-PaLM 2
85.4%
85
80
75
70
65
60%
55
50
45
40
35
30
Approximate medical pass mark
GPT-Neo
33.3%
PubMedBERT
38.1%
BioLinkBERT
45.1%
DRAGON
47.5%
Galactica
44.4%
PubMed GPT
50.3%
Dec '20
Jul '21
Mar '22
Oct '22
Nov '22
Dec '22

inkeep

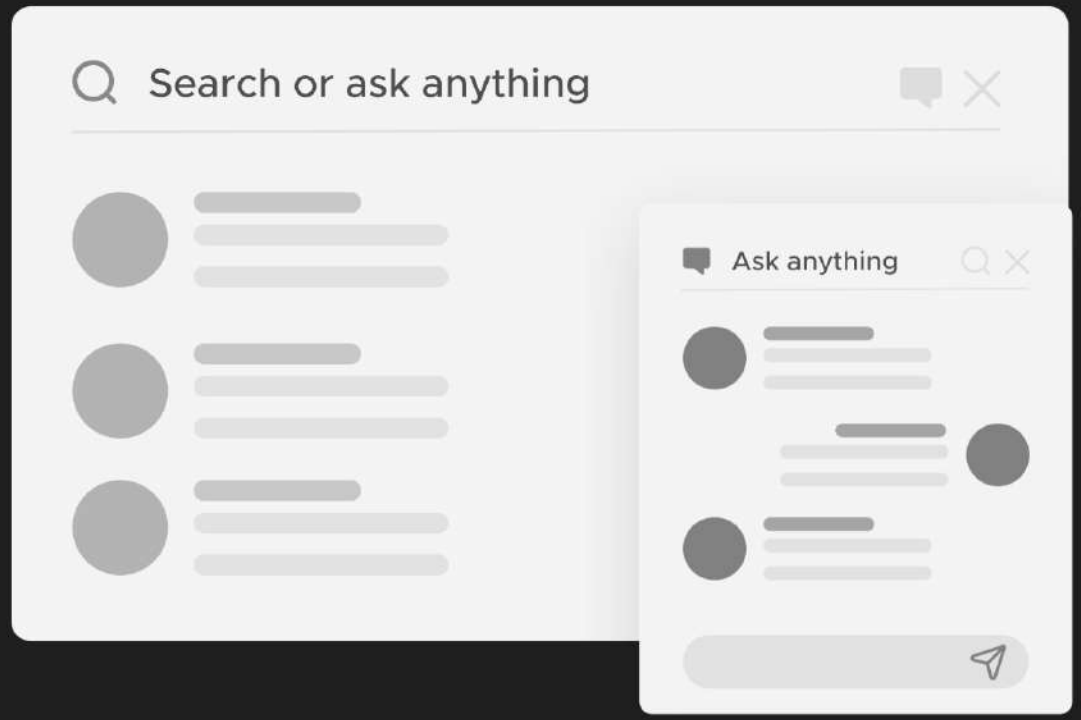
Conversational search for your developer product

Help developers do more with your product, faster with our AI-powered search and chat experiences.

- ✓ **Increase activation**
- ✓ **Support developers at scale**
- ✓ **Understand what users get stuck on**

Request early access

See a demo



Back to blog

StarCoder: A State-of-the-Art LLM for Code

Published May 4, 2023

Update on GitHub


lvwerra
Leandro von Werra


loubnabnl
Loubna Ben Allal

Introducing StarCoder

StarCoder and StarCoderBase are Large Language Models for Code (Code LLMs) trained on permissively licensed data from GitHub, including from 80+ programming languages, Git commits, GitHub issues, and Jupyter notebooks. Similar to LLaMA, we trained a ~15B parameter model for 1 trillion tokens. We fine-tuned StarCoderBase model for 35B Python tokens, resulting in a new model that we call StarCoder.

testgpt

3.0.3 • Public • Published 18 days ago

TestGPT

A command-line tool for generating unit tests for your files automatically using OpenAI GPT-3.5-turbo model (gpt-4 also supported for developers who have it).

> If you have access to GPT-4 API, you can now pass `--model` / `-m` option, see below for an example.

> Now there is a VScode extension the process even faster, check the extension here (You have to install the latest version of testgpt for it to work)

> NOTE: From version 3.0.0 and upwards, `testgpt.config.json` was replaced with `testgpt.config.yaml`, and a new custom property (`examples`) is added.

Install

```
> npm i testgpt
```

Repository
github.com/fayez-nazzal/testgpt

Homepage
github.com/fayez-nazzal/testgpt#readme

Weekly Downloads
118

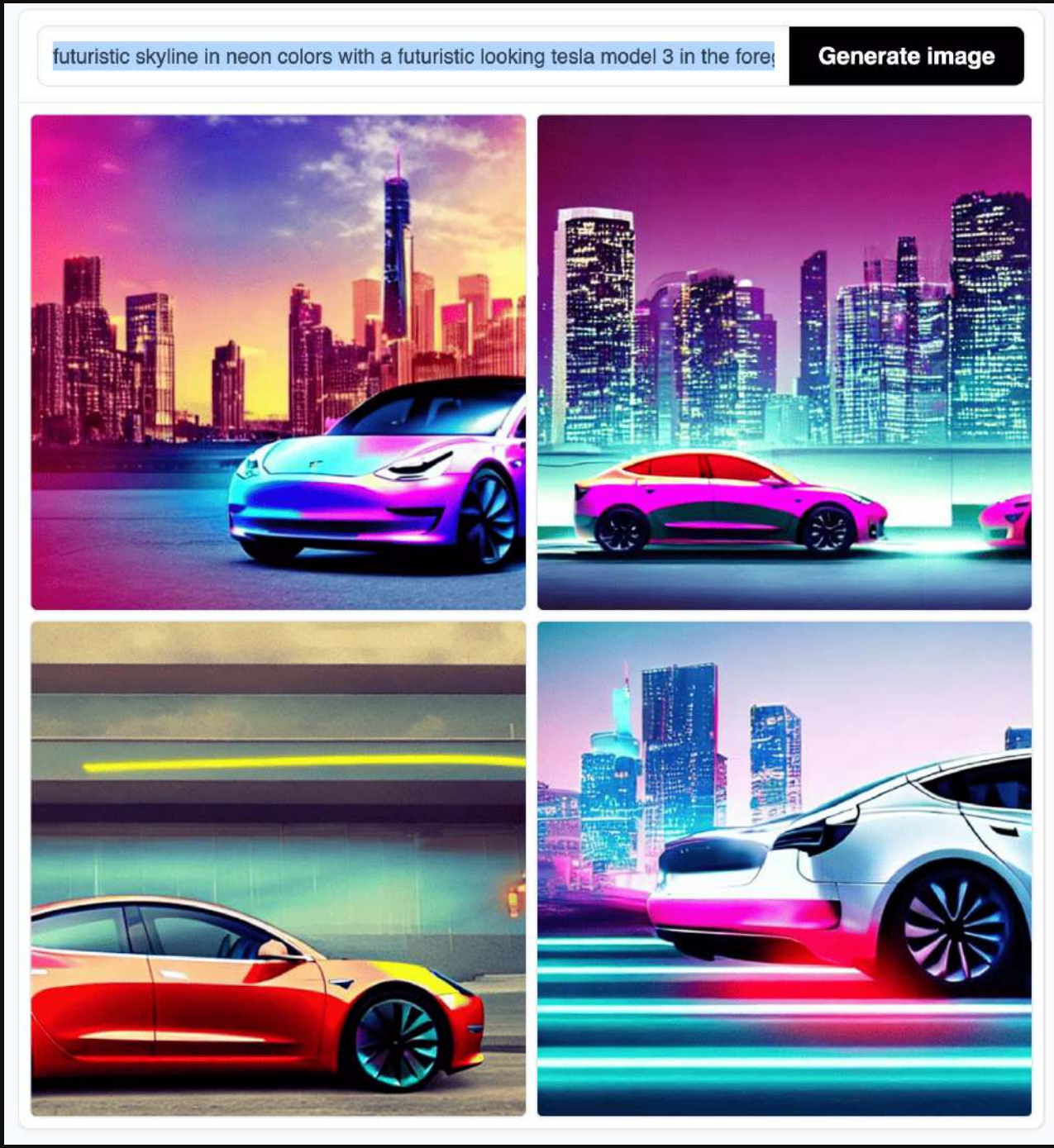
Version
3.0.3

License
MIT

Prompt to any

Stable diffusion

> futuristic skyline in neon colors with a futuristic looking tesla model 3 in the foreground



MusicLM: Generating Music From Text

| paper | dataset |

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, Christian Frank

Google Research

Abstract We introduce MusicLM, a model generating high-fidelity music from text descriptions such as *"a calming violin melody backed by a distorted guitar riff"*. MusicLM casts the process of conditional music generation as a hierarchical sequence-to-sequence modeling task, and it generates music at 24 kHz that remains consistent over several minutes. Our experiments show that MusicLM outperforms previous systems both in audio quality and adherence to the text description. Moreover, we demonstrate that MusicLM can be conditioned on both text and a melody in that it can transform whistled and hummed melodies according to the style described in a text caption. To support future research, we publicly release MusicCaps, a dataset composed of 5.5k music-text pairs, with rich text descriptions provided by human experts.

LLMs

- Large size, trained on massive datasets
- Open Source: Langchain
- Prompt engineering
- Classification, Question Answering, Summarization, Fill-mask, Translation
- Hallucination & Model bias
- Conversational memory
- Learning from queries (dangerous?)
- Agents for LLMs (execute a calculator, SQL query, use mechanical turk)

Voice based search

- Cars
- Mobile

Summary

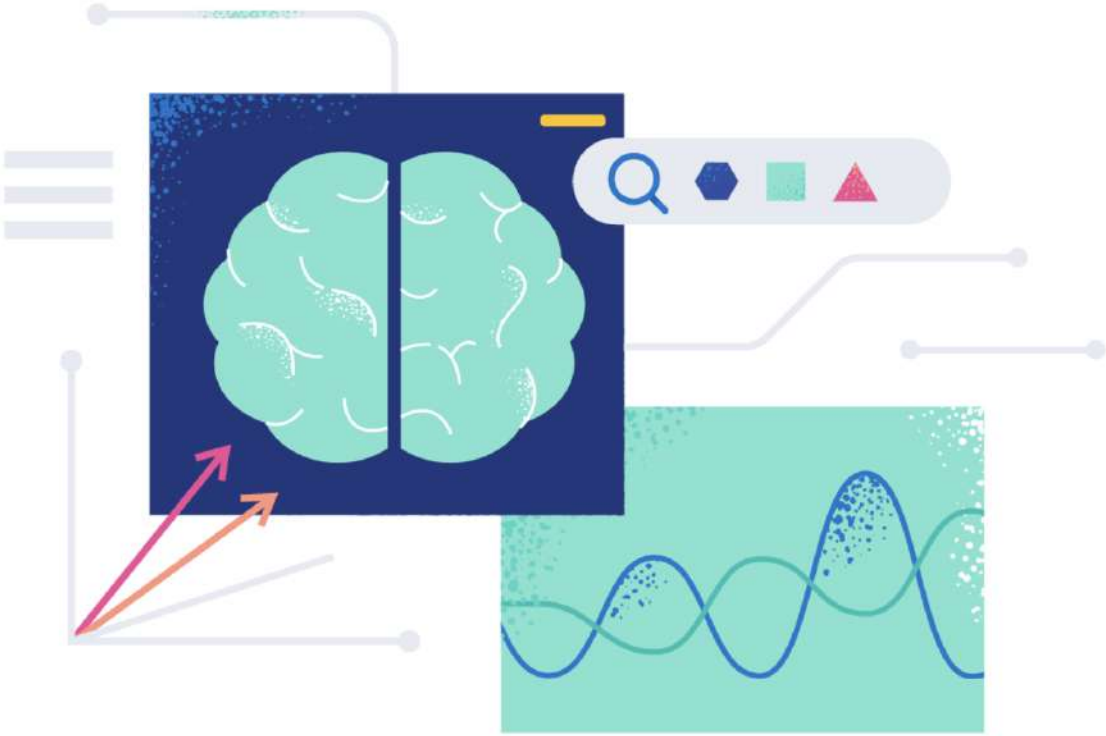
Summary

- Search becomes hybrid: Will the existing search engines adapt?
- Search customization is expensive - A brief history of code search at GitHub
- Search engine becomes the commodity
- Rent your industry specific LLM!
- Privacy LLMs might be a thing
- Expect a lot of movement, lots of "AI integrations" and even more hot air...

ESRE Elasticsearch Relevance Engine™

Powering the generative AI era

The Elasticsearch Relevance Engine™ (ESRE) is designed to power artificial intelligence-based search applications. Use ESRE to apply semantic search with superior relevance out of the box (without domain adaptation), integrate with external large language models (LLMs), implement hybrid search, and use third-party or your own transformer models.



Thank you!

Q & A

Alexander Reelsen
alr@spinscale.de | @spinscale