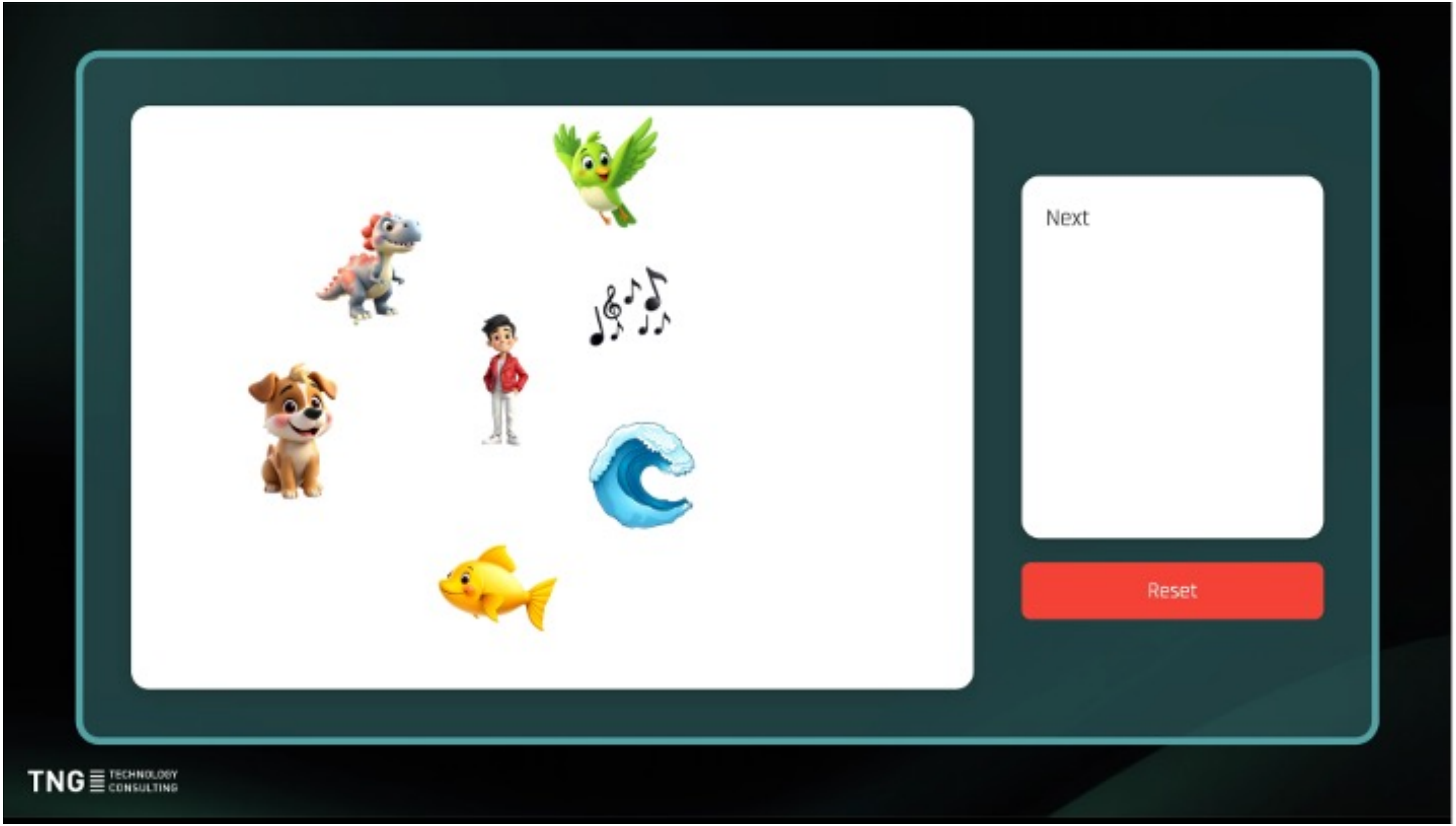
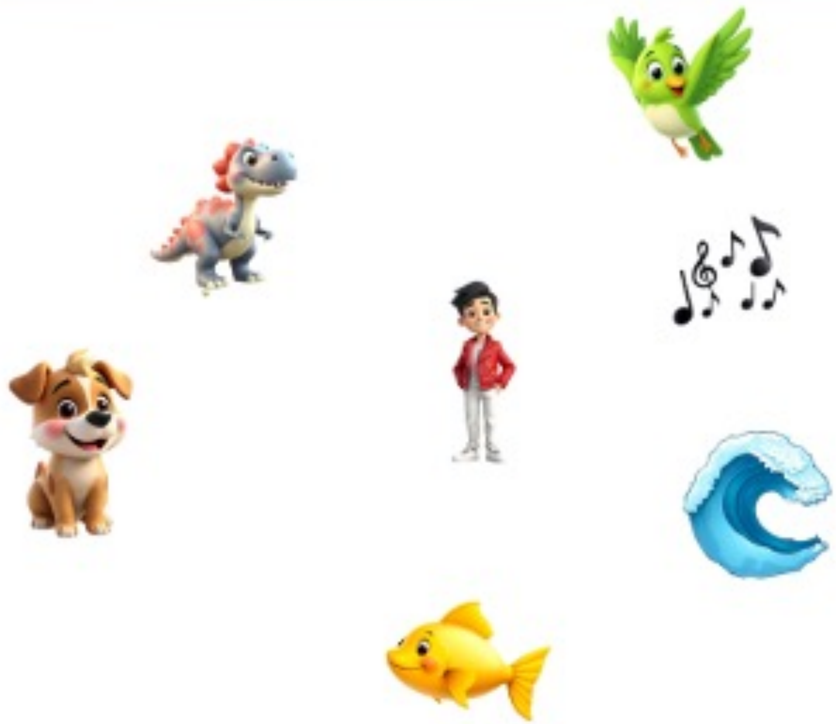
Finally Understand Embeddings

And You Will Never Have to Search for the Right Emoji Again

TNG TECHNOLOGY CONSULTING


Next
Reset
TNG TECHNOLOGY CONSULTING



Overview

- History
- Semantic Search
- Fine-Tuning Base Models
- Benchmarking Embeddings
- Embedding of Non-Text Data
- Semantic Search at Scale
- Outlook

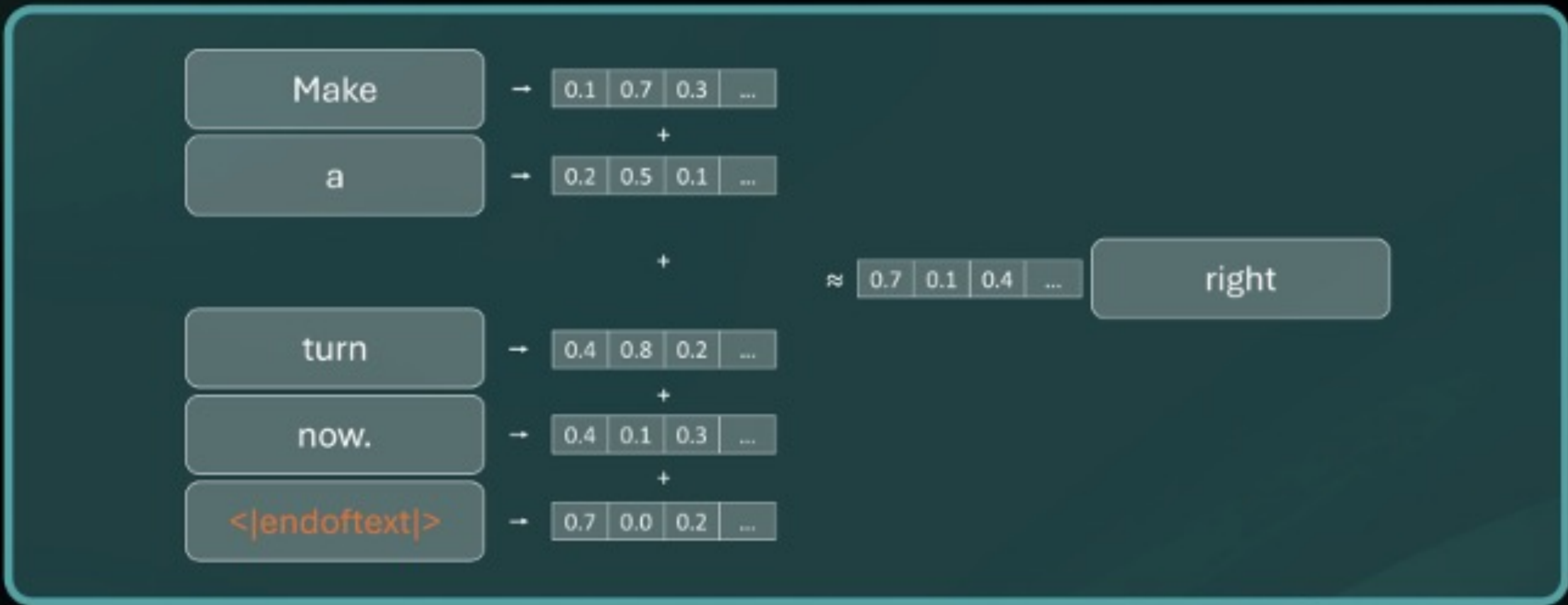
TNG TECHNOLOGY CONSULTING

The Idea of Embeddings

Word2Vec (2013)

The Idea of Embeddings

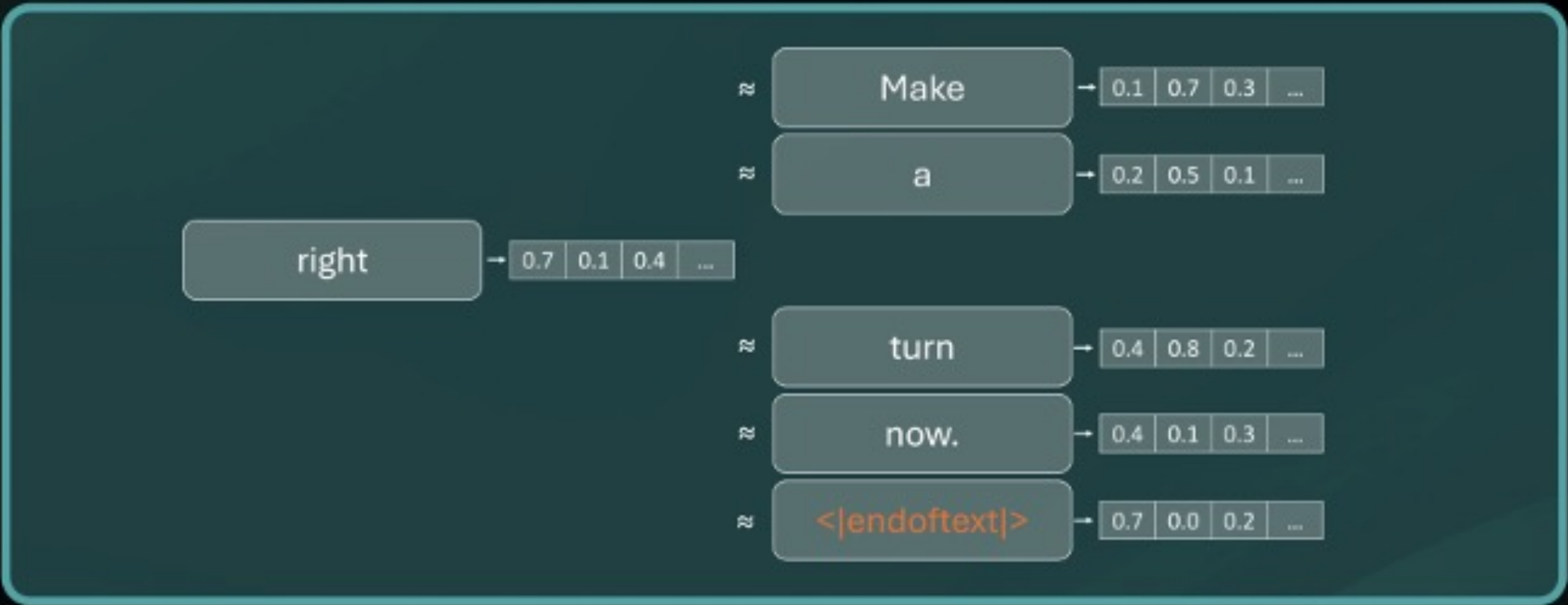
Word2Vec (2013)



TNG TECHNOLOGY CONSULTING

The Idea of Embeddings

Word2Vec (2013)



TNG TECHNOLOGY CONSULTING

Contextualized Word Embeddings

Transformers

TNG TECHNOLOGY CONSULTING

The Attention Mechanism

Make

a

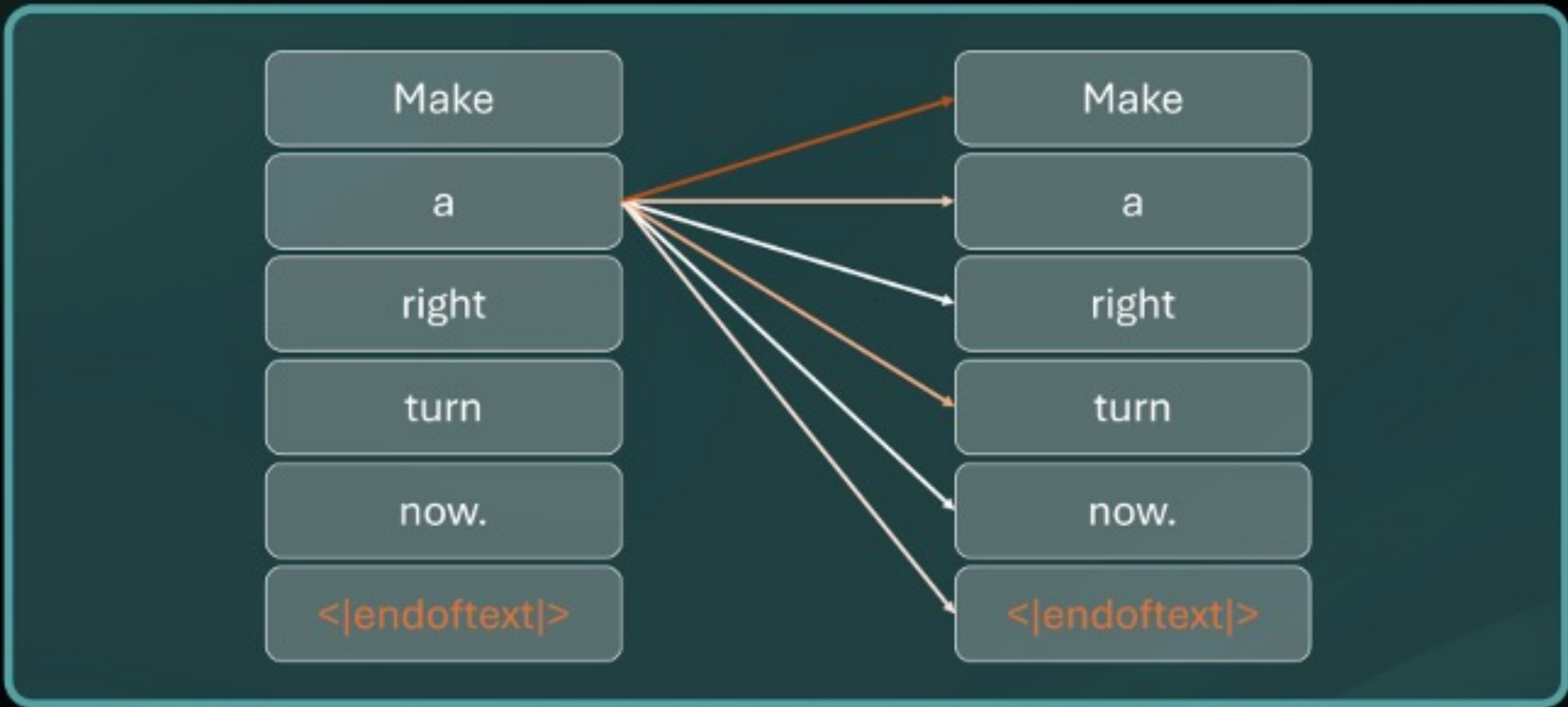
right

turn

now.

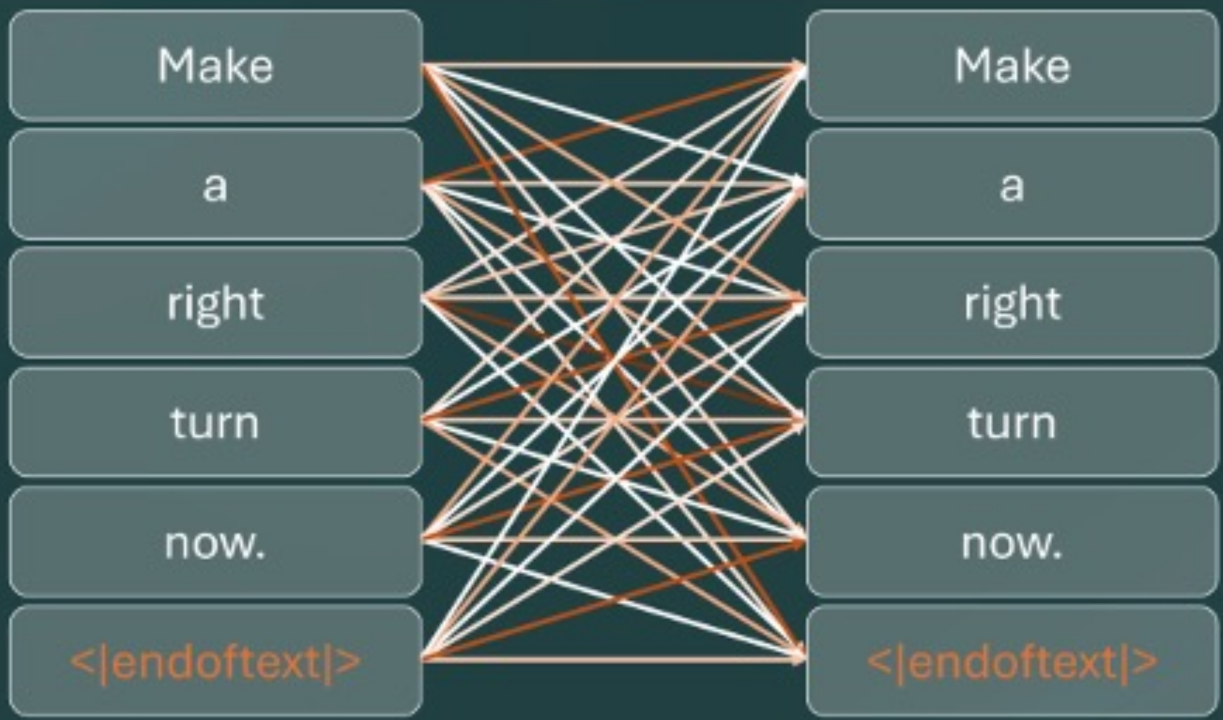
<|endoftext|>

The Attention Mechanism



TNG TECHNOLOGY CONSULTING

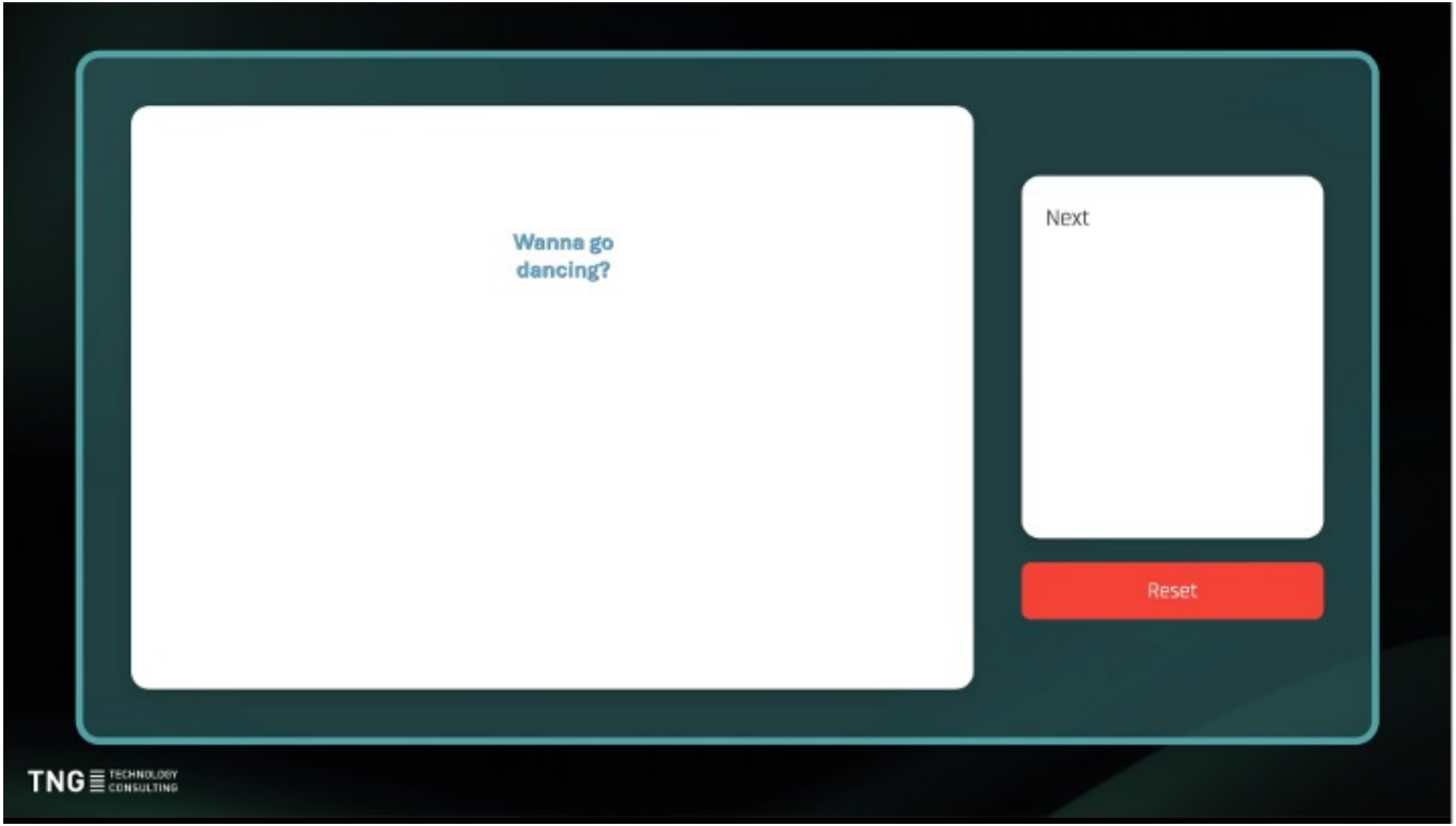
The Attention Mechanism

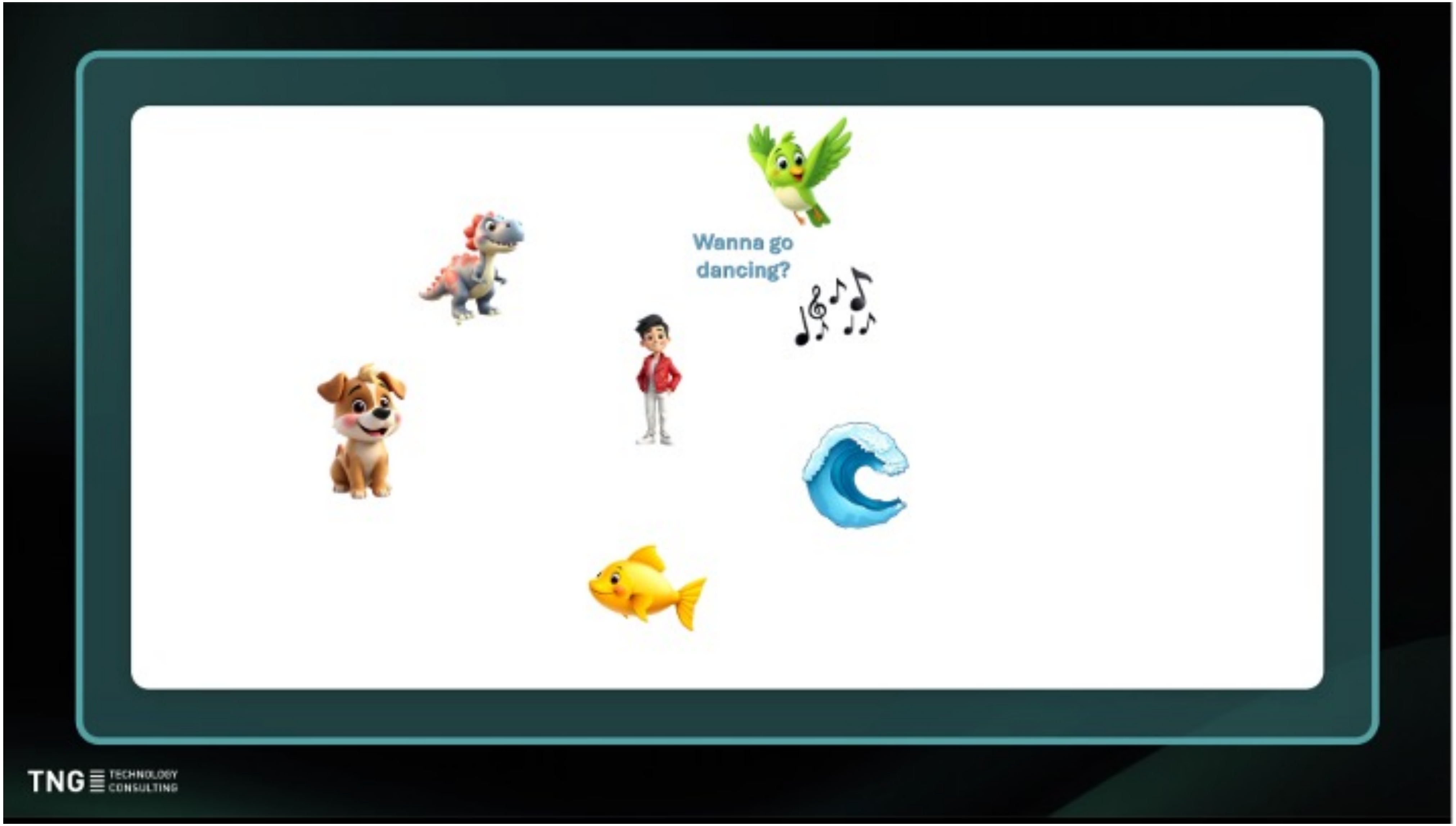


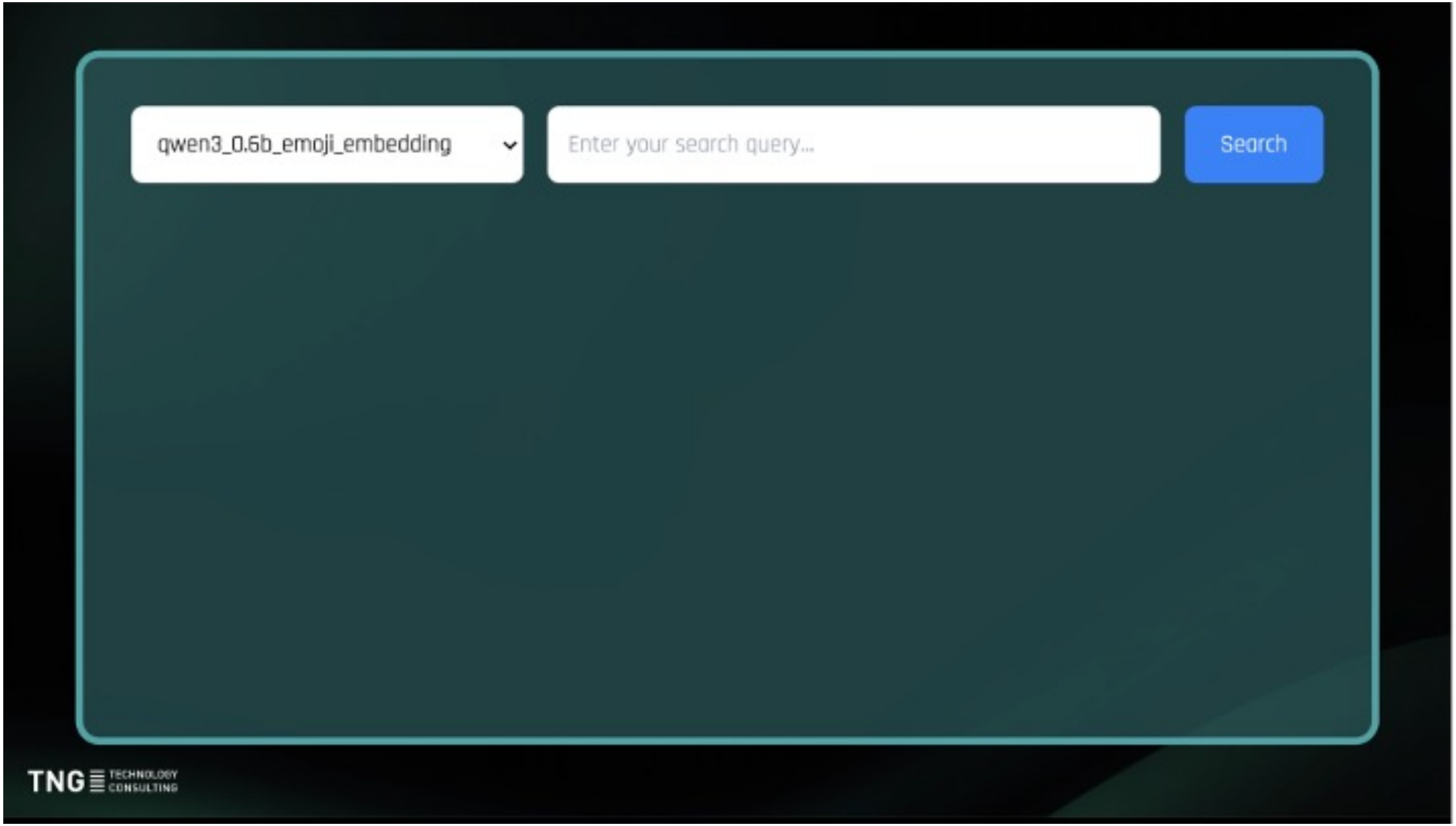
Overview

- History
- <u>Semantic Search</u>
- Fine-Tuning Base Models
- Benchmarking Embeddings
- Embedding of Non-Text Data
- Semantic Search at Scale
- Outlook

TNG TECHNOLOGY CONSULTING


Wanna go
dancing?
Next
Reset
TNG TECHNOLOGY CONSULTING


Wanna go dancing?
TNG TECHNOLOGY CONSULTING


qwen3_0.6b_emoji_embedding
Enter your search query...
Search
TNG TECHNOLOGY CONSULTING

Semantic Search

Model	Memory Usage (MB)	# Parameters	Embedding Dimensions
llama-embed-nemotron-8b	28629	7B	4096
gemini-embedding-001	Unknown	Unknown	3072
Qwen3-Embedding-8B	28866	7B	4096
Qwen3-Embedding-4B	15341	4B	2560
Qwen3-Embedding-0.6B	2272	595M	1024
gte-Qwen2-7B-instruct	29040	7B	3584
Linq-Embed-Mistral	13563	7B	4096

TNG TECHNOLOGY CONSULTING

Model Size

How many dimensions do I really need?

Google DeepMind

On the Theoretical Limitations of Embedding-Based Retrieval

Orion Weller*[1,2], Michael Boratko[1], Iftekhar Naim[1] and Jinhyuk Lee[1]
[1]Google DeepMind, [2]Johns Hopkins University

Vector embeddings have been tasked with an ever-increasing set of retrieval tasks over the years, with a nascent rise in using them for reasoning, instruction-following, coding, and more. These new benchmarks push embeddings to work for *any query* and *any notion of relevance* that could be given. While prior works have pointed out theoretical limitations of vector embeddings, there is a common assumption that these difficulties are exclusively due to unrealistic queries, and those that are not can be overcome with better training data and larger models. In this work, we demonstrate that we may encounter these theoretical limitations in realistic settings with extremely simple queries. We connect known results in learning theory, showing that the number of top-k subsets of documents capable of being returned as the result of some query is limited by the dimension of the embedding. We empirically show that this holds true even if we restrict to $k = 2$, and directly optimize on the test set with free parameterized embeddings. We then create a realistic dataset called LIMIT that stress tests models based on these theoretical results, and observe that even state-of-the-art models fail on this dataset despite the simple nature of the task. Our work shows the limits of embedding models under the existing single vector paradigm and calls for future research to develop methods that can resolve this fundamental limitation.

TNG TECHNOLOGY CONSULTING

Model Size

How many dimensions do I really need?

Why is this so relatable

Me when I see a dog wearing clothes

Selfie game strong today

Secret agent documenting the mission

Model Size

How many dimensions do I really need?

TNG TECHNOLOGY CONSULTING

Model Size

How many dimensions do I really need?

	😂	📸	👐	🤫	👀
Why is this so relatable	1	-1	-1	-1	-1
Me when I see a dog wearing clothes	-1	1	1	-1	1
Selfie game strong today	1	1	-1	-1	1
Secret agent documenting the mission	-1	-1	-1	1	1

Model Size

How many dimensions do I really need?

Model Size

How many dimensions do I really need?

	😂	📸	👐	🤫	👀
Why is this so relatable	1	-1	-1	-1	-1
Me when I see a dog wearing clothes	-1	1	1	-1	1
Selfie game strong today	1	1	-1	-1	1
Secret agent documenting the mission	-1	-1	-1	1	1

Model Size

How many dimensions do I really need?

	😂	📸	👐	🤫	👀
[0.2, -0.15, ..., -0.38, 0.67]	1	-1	-1	-1	-1
[-0.25, 0.33, ..., 0.44, -0.52]	-1	1	1	-1	1
0.15, -0.42, ..., -0.61, 0.35]	1	1	-1	-1	1
[-0.38, -0.22, ..., 0.19, -0.47]	-1	-1	-1	1	1

Model Size

How many dimensions do I really need?

	[0.2, ...]	[-0.2, ...]	[-0.9, ...]	[0.7, ...]	[-0.4, ...]
[0.2, -0.15, ..., -0.38, 0.67]	1	-1	-1	-1	-1
[-0.25, 0.33, ..., 0.44, -0.52]	-1	1	1	-1	1
0.15, -0.42, ..., -0.61, 0.35]	1	1	-1	-1	1
[-0.38, -0.22, ..., 0.19, -0.47]	-1	-1	-1	1	1

Model Size

How many dimensions do I really need?

	[0.2, ...]	[-0.2, ...]	[-0.9, ...]	[0.7, ...]	[-0.4, ...]
[0.2, -0.15, ..., -0.38, 0.67]	+	-	-	-	-
[-0.25, 0.33, ..., 0.44, -0.52]	-	+	+	-	+
0.15, -0.42, ..., -0.61, 0.35]	+	+	-	-	+
[-0.38, -0.22, ..., 0.19, -0.47]	-	-	-	+	+

Model Size

How many dimensions do I really need?

$$d \geq rank_{\pm} Q_{rel} - 1$$

Model Size

How many dimensions do I really need?

$$d \geq rank_{\pm} Q_{rel} - 1$$

$$\# 😩 \leq -10.5322 + 4.0309d + 0.0520d^2 + 0.0037d^3$$

TNG TECHNOLOGY CONSULTING

Model Size

How many dimensions do I really need?

$$d \geq rank_{\pm} Q_{rel} - 1$$

$$\#😩 \leq -10.5322 + 4.0309d + 0.0520d^2 + 0.0037d^3$$

# 😩	500k	1.7m	4m	107m	250m
d	512	768	1024	3072	4096

Semantic Search

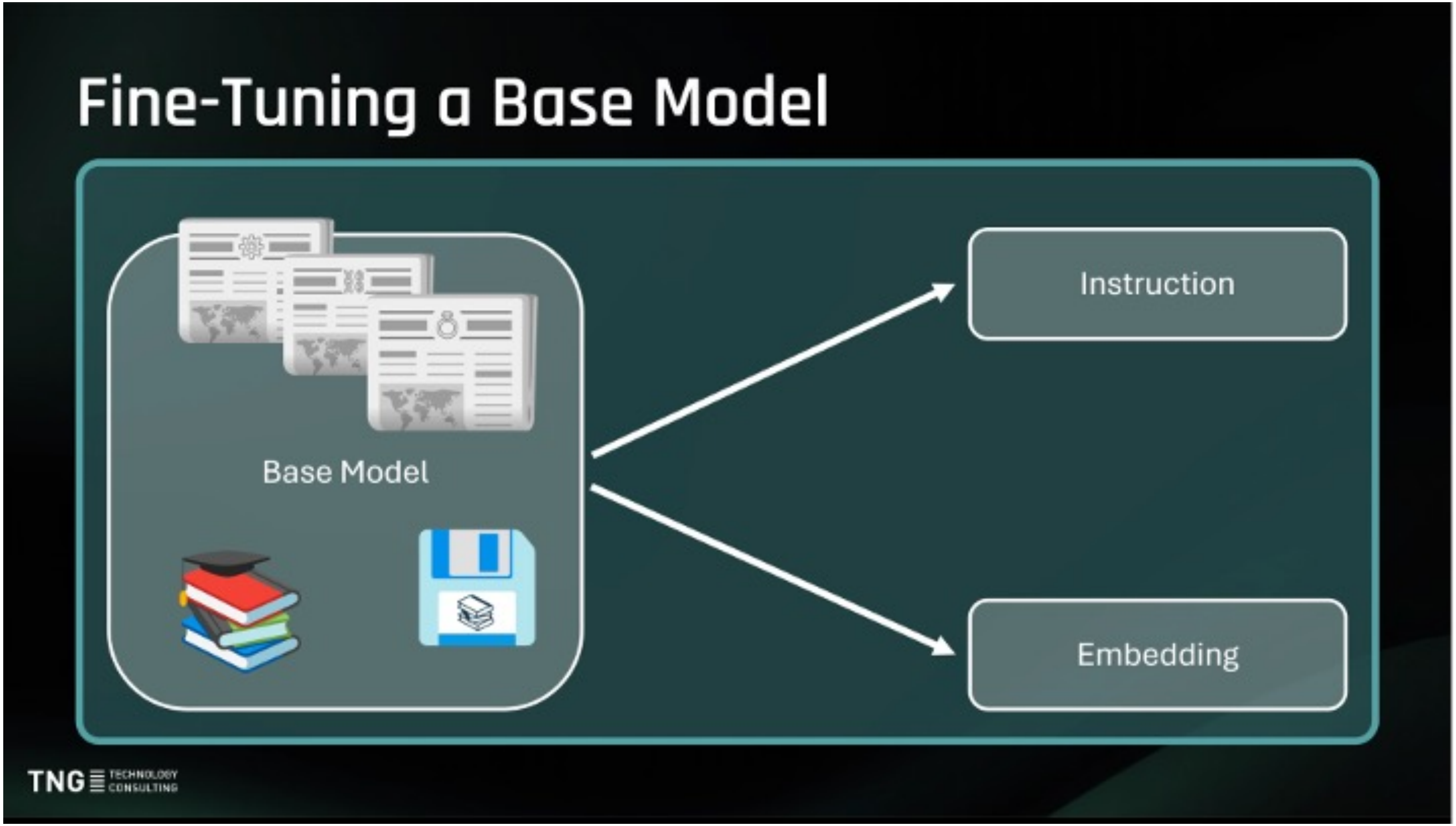
Model	Memory Usage (MB)	# Parameters	Embedding Dimensions
llama-embed-nemotron-8b	28629	7B	4096
gemini-embedding-001	Unknown	Unknown	3072
Qwen3-Embedding-8B	28866	7B	4096
Qwen3-Embedding-4B	15341	4B	2560
Qwen3-Embedding-0.6B	2272	595M	1024
gte-Qwen2-7B-instruct	29040	7B	3584
Linq-Embed-Mistral	13563	7B	4096

TNG TECHNOLOGY CONSULTING

Overview

- History
- Semantic Search
- <u>Fine-Tuning Base Models</u>
- Benchmarking Embeddings
- Embedding of Non-Text Data
- Semantic Search at Scale
- Outlook

TNG TECHNOLOGY CONSULTING


Fine-Tuning a Base Model
Base Model
Instruction
Embedding
TNG TECHNOLOGY CONSULTING

Base and Instruction Model Behavior

Qwen 3

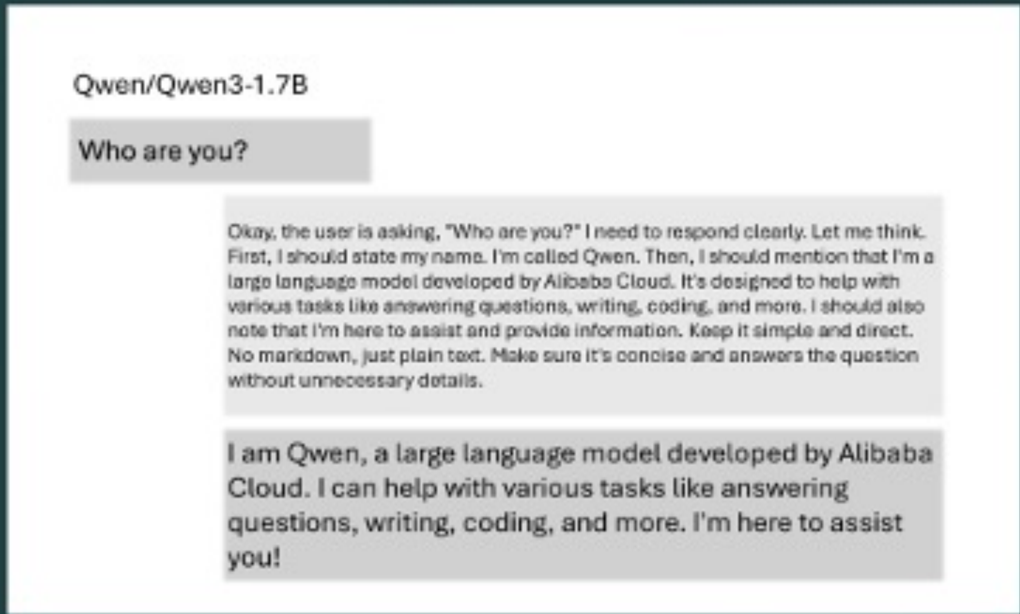
Qwen/Qwen-1.7B-Base

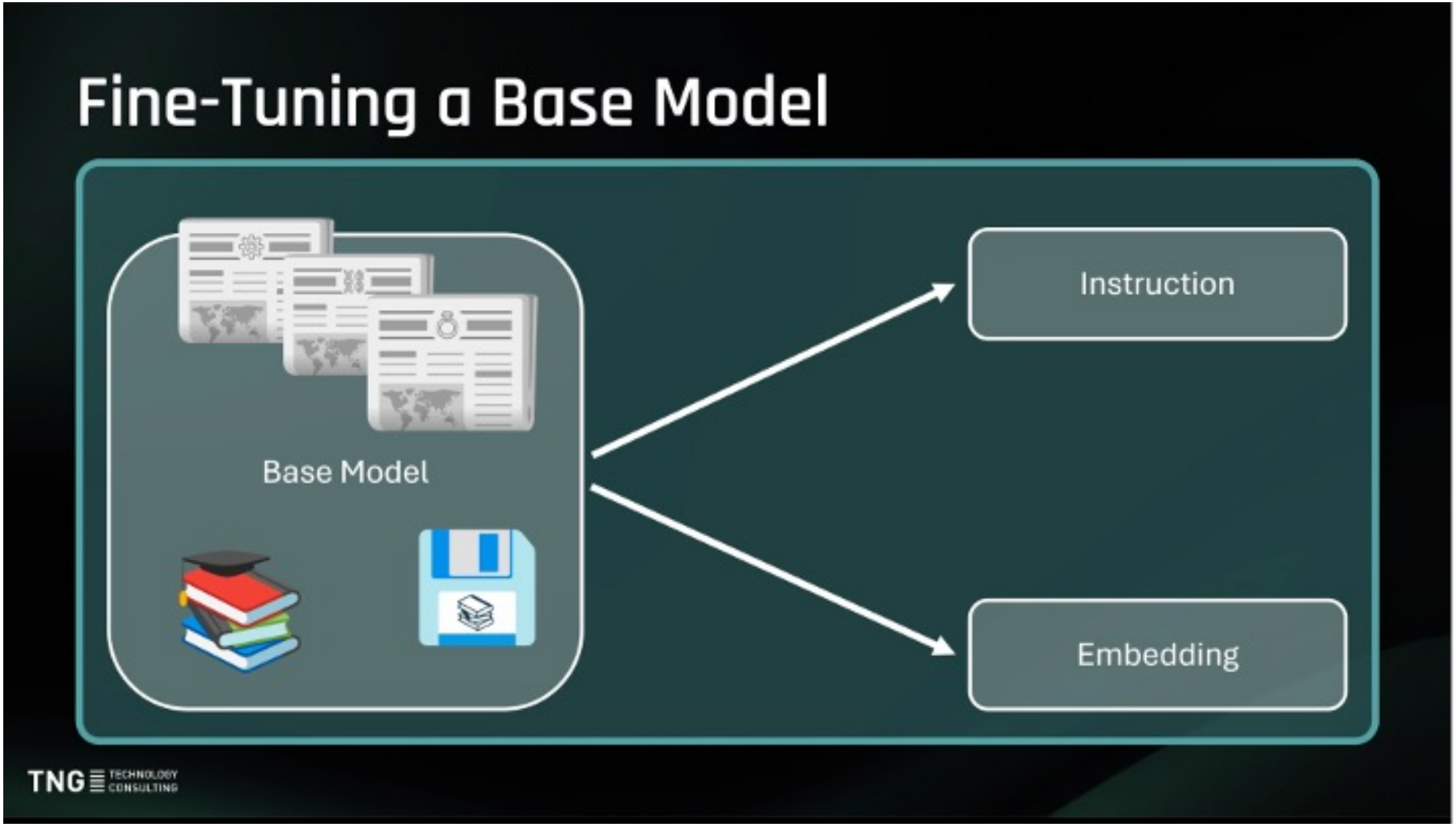
Who are you?

I am an artificial intelligence language model, trained to assist with a wide range of tasks and provide information on various topics. I am designed to understand and respond to user queries in a helpful and informative manner.

Base and Instruction Model Behavior

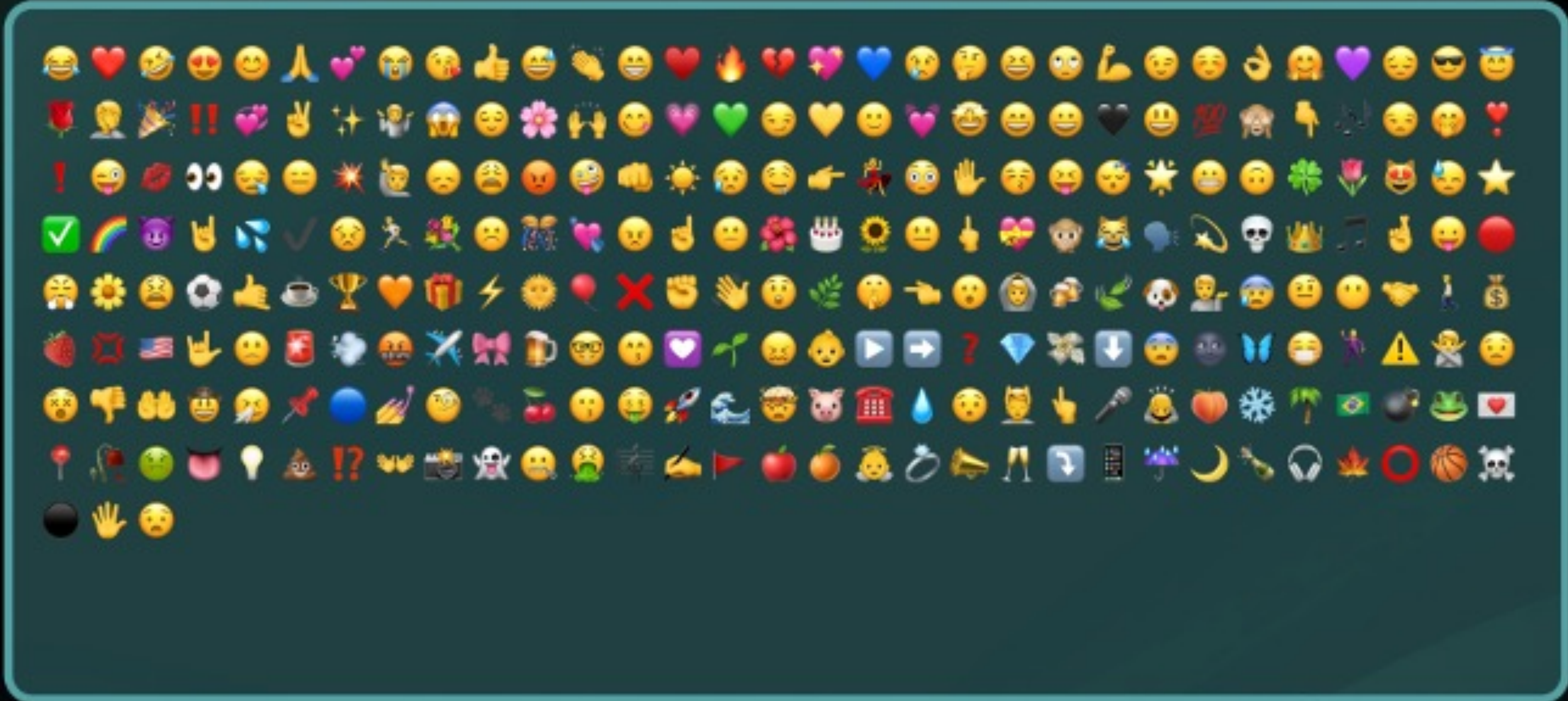
Qwen 3





Fine-Tuning a Base Model
Base Model
Instruction
Embedding
TNG TECHNOLOGY CONSULTING

250 Emojis



Training Data

Send help… or don't, this is entertaining

🤪

Accidentally invented a new dance move

My brain is officially on vacation

Choosing the Right Training Mode

Let's dance the night away! <|endoftext|>

💃

Choosing the Right Training Mode

Instruction query + Let's dance the night away! <|endoftext|>

Answer query + 💃

Choosing the Right Training Mode

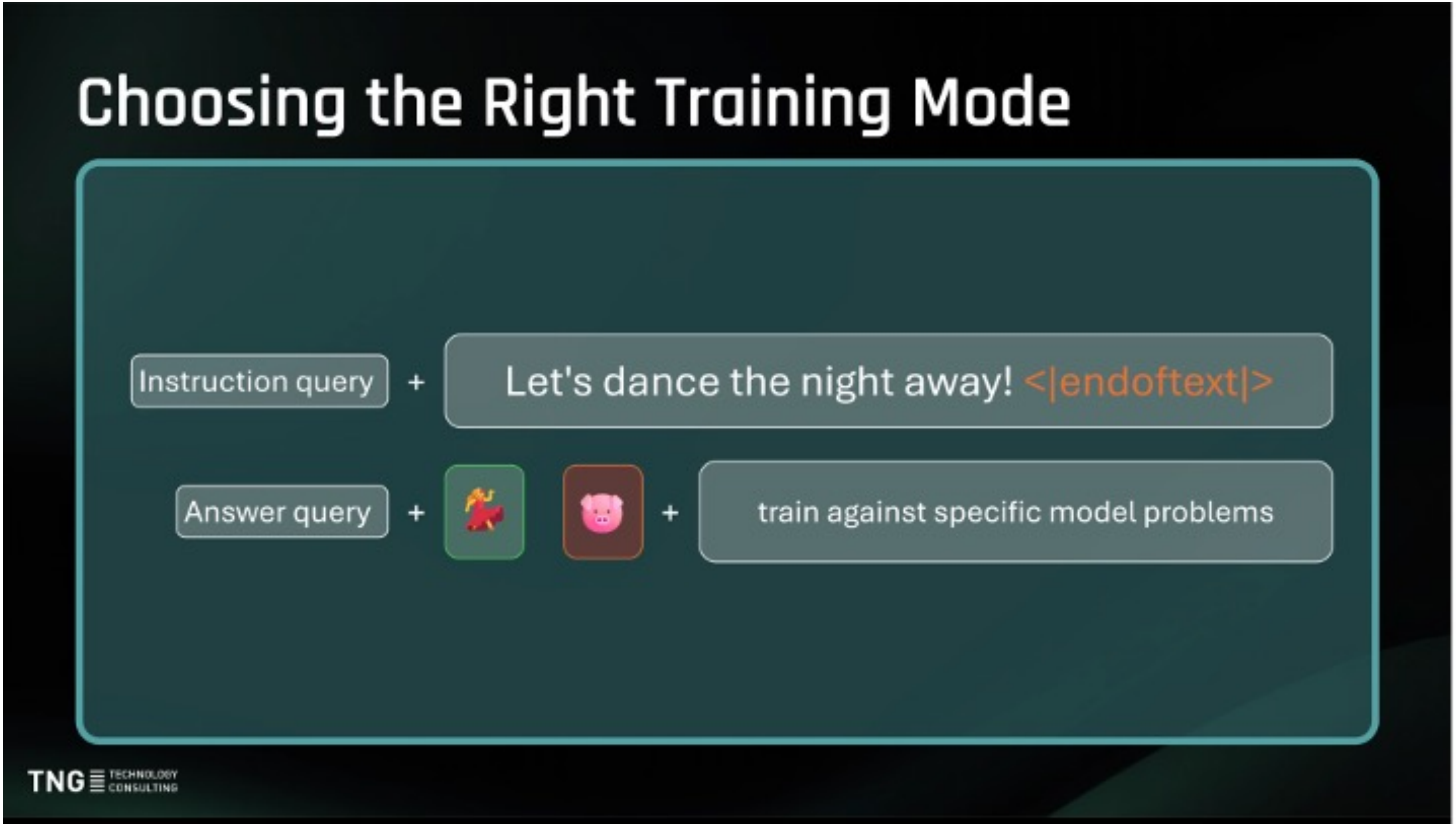
Instruction query + Let's dance the night away! <|endoftext|>

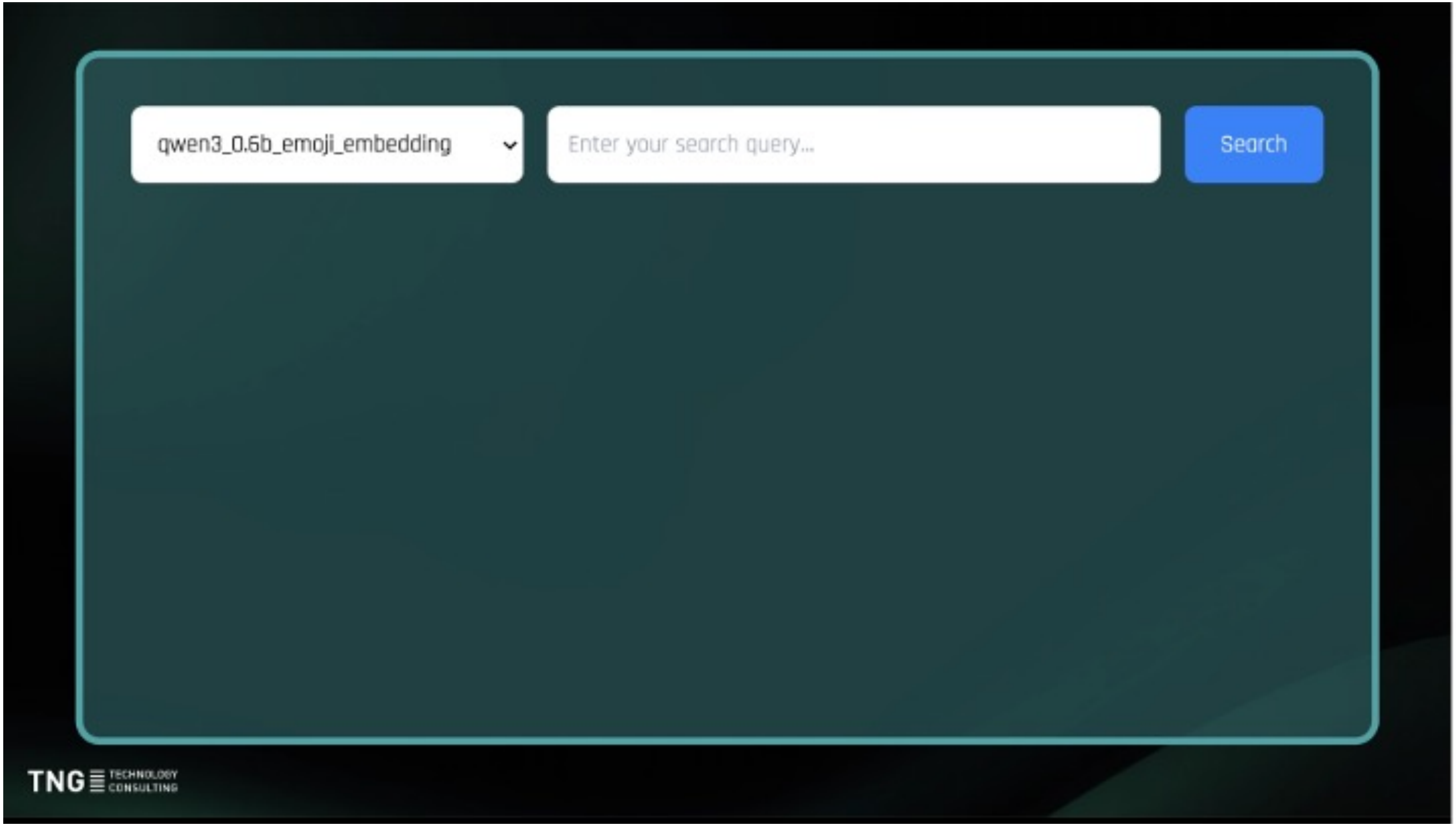
Answer query + 💃 😴

Choosing the Right Training Mode

Instruction query + Let's dance the night away! <|endoftext|>

Answer query + 💃 🐷

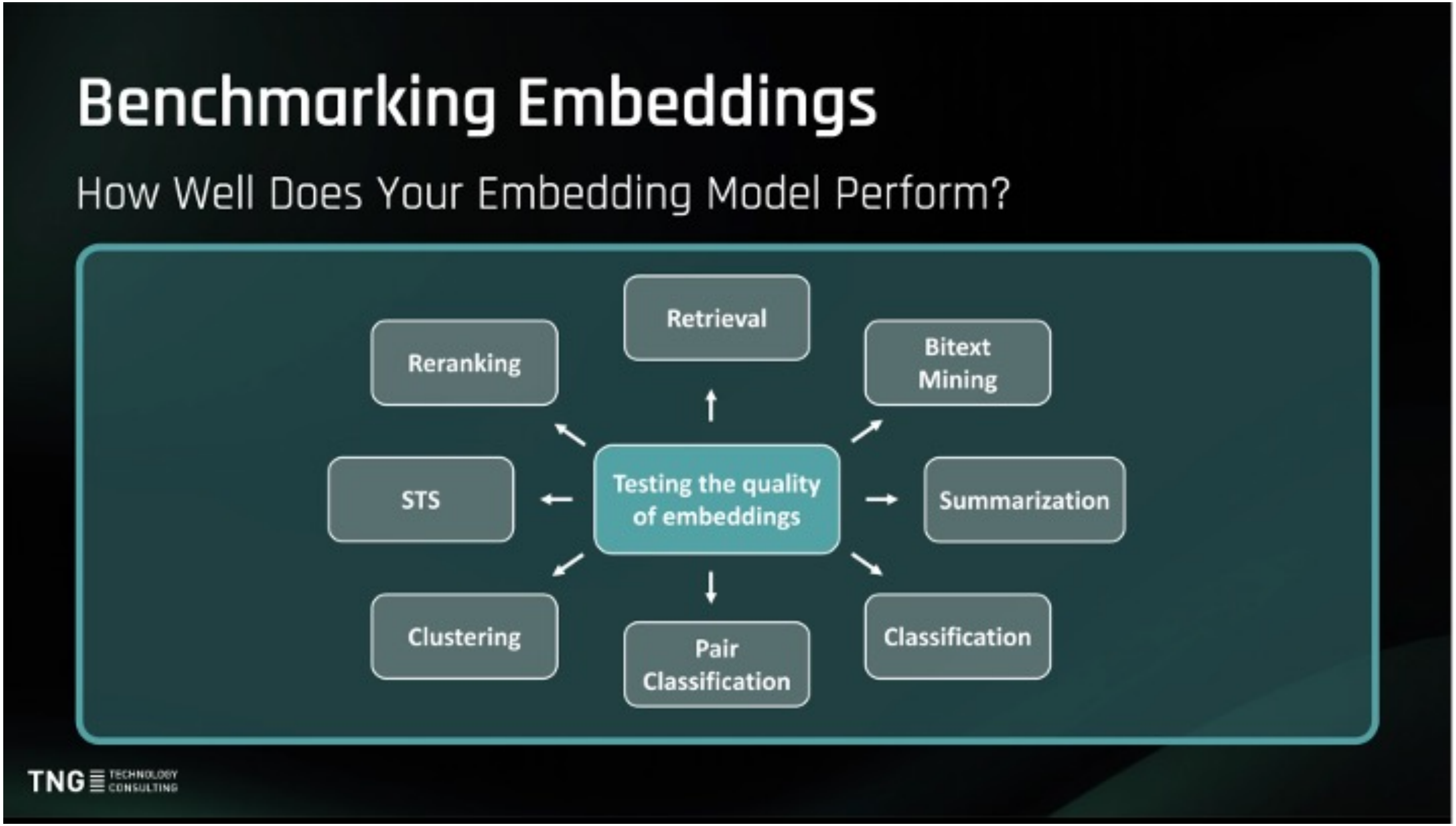

Choosing the Right Training Mode
Instruction query
+
Let's dance the night away! <|endoftext|>
Answer query
+
+
train against specific model problems
TNG TECHNOLOGY CONSULTING

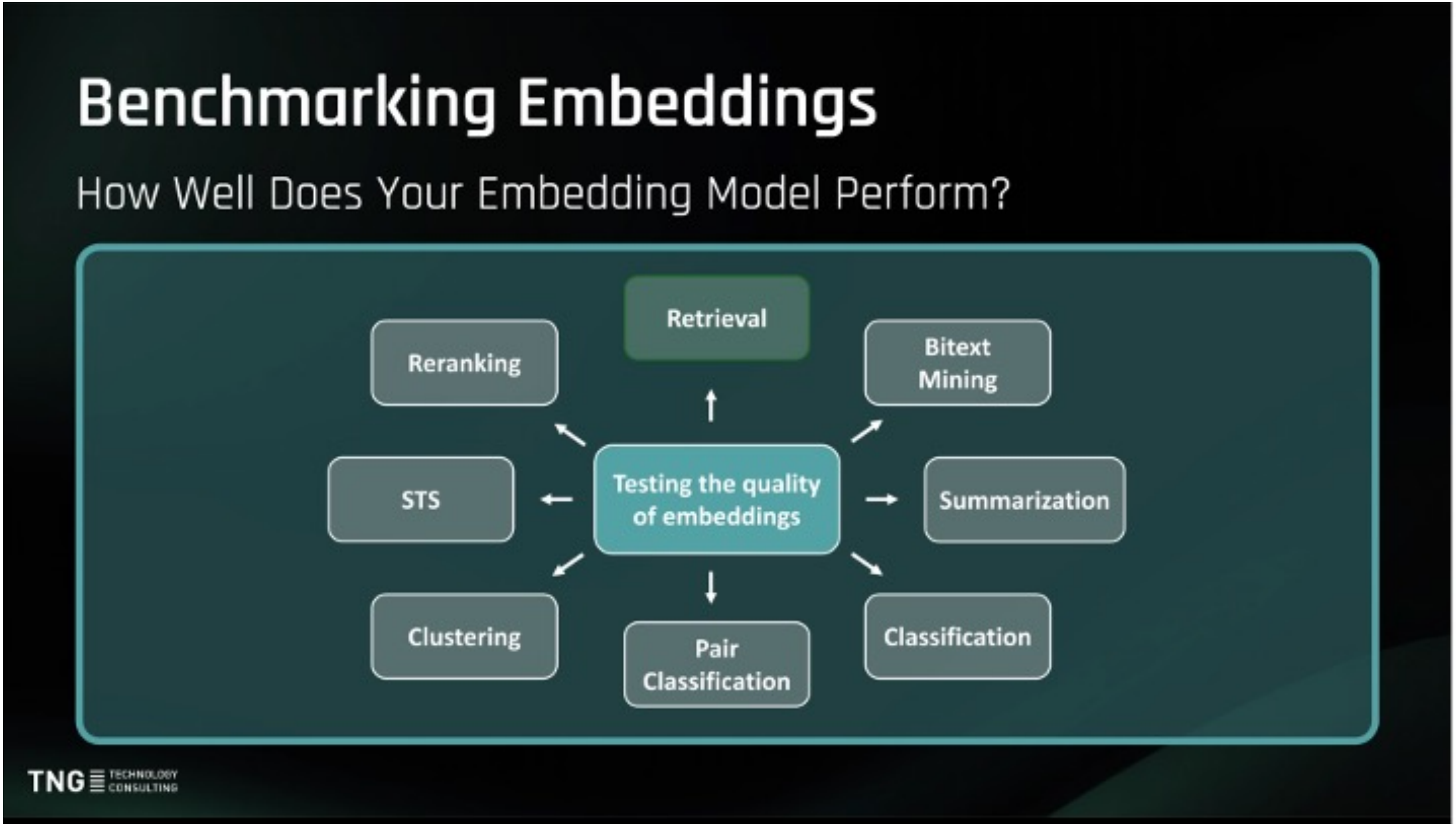

qwen3_0.6b_emoji_embedding
Enter your search query...
Search
TNG TECHNOLOGY CONSULTING

Overview

- History
- Semantic Search
- Fine-Tuning Base Models
- Benchmarking Embeddings
- Embedding of Non-Text Data
- Semantic Search at Scale
- Outlook

TNG TECHNOLOGY CONSULTING


Benchmarking Embeddings
How Well Does Your Embedding Model Perform?
Retrieval
Reranking
Bitext Mining
STS
Testing the quality of embeddings
Summarization
Clustering
Pair Classification
Classification
TNG TECHNOLOGY CONSULTING


Benchmarking Embeddings
How Well Does Your Embedding Model Perform?
Retrieval
Reranking
Bitext Mining
STS
Testing the quality of embeddings
Summarization
Clustering
Pair Classification
Classification
TNG TECHNOLOGY CONSULTING

Top-3 benchmark

When the autocorrect ruins your heartfelt text

Top-3 benchmark

When the autocorrect ruins your heartfelt text

Top-3 benchmark

When the autocorrect ruins your heartfelt text

😢

😡 😢 👻

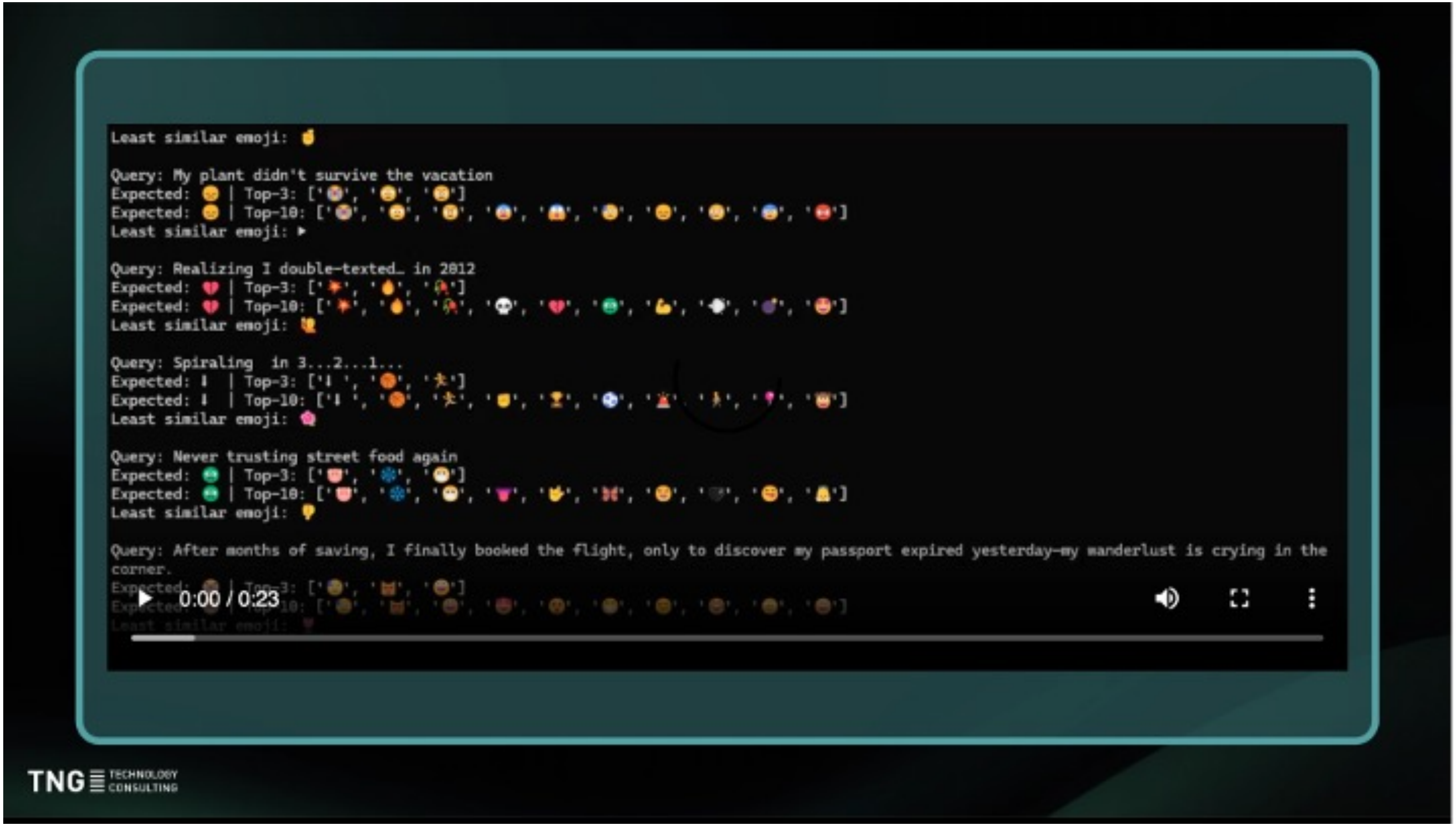
Top-3 benchmark

I thought I'd impress my boss by volunteering for that big presentation...

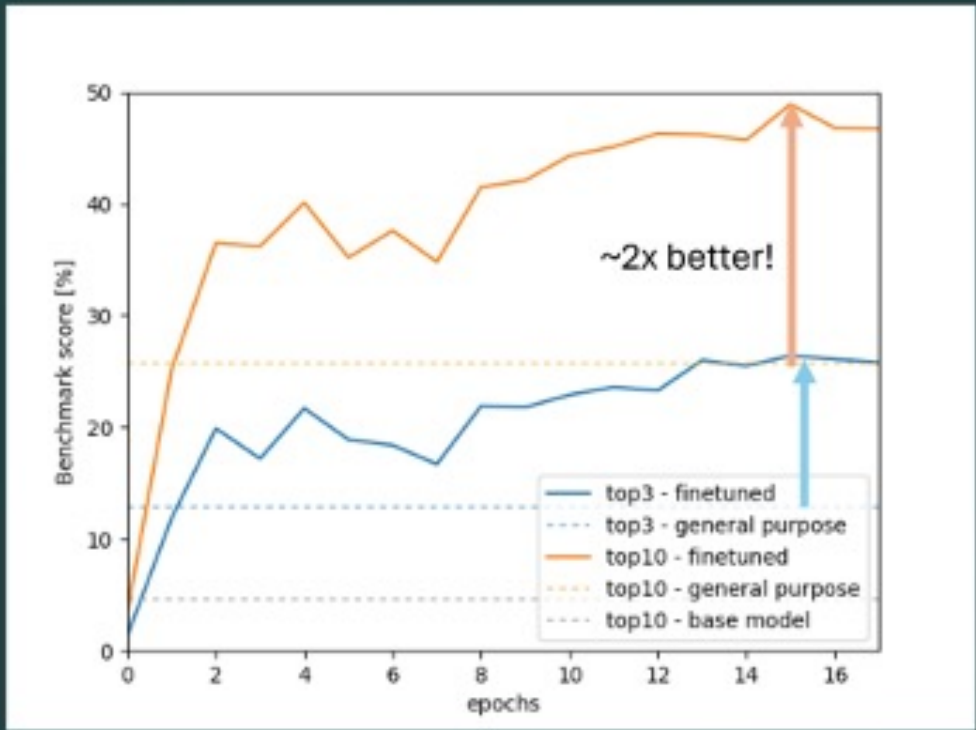
Top-3 benchmark

I thought I'd impress my boss by volunteering for that big presentation...

Top-3 benchmark

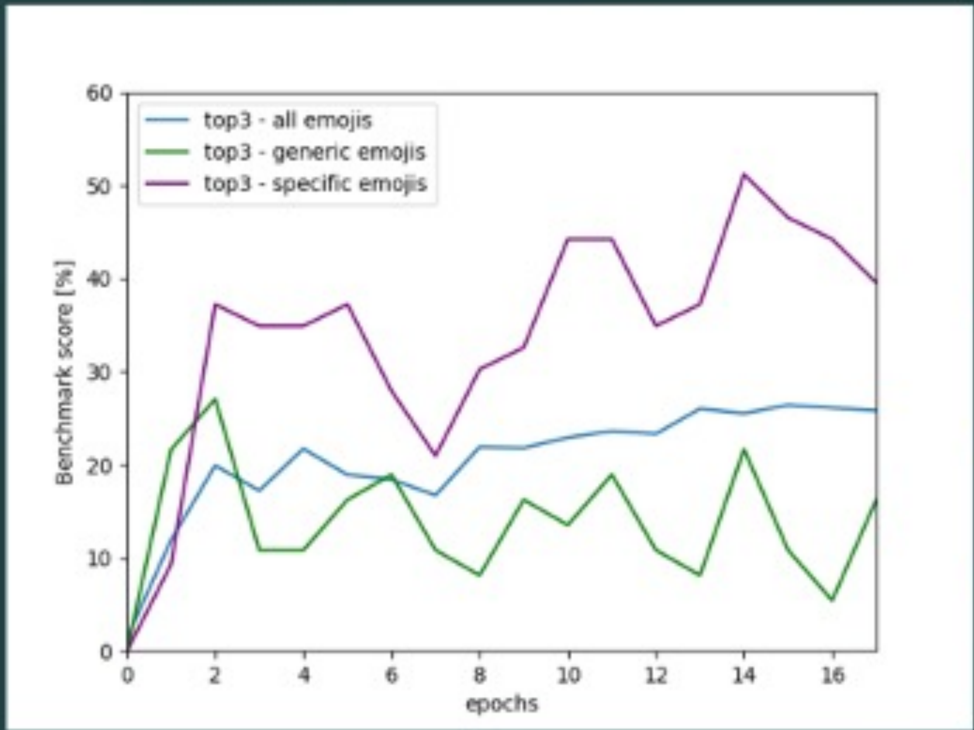

Least similar emoji:
Query: My plant didn't survive the vacation
Expected: | Top-3:
Expected: | Top-10:
Least similar emoji:
Query: Realizing I double-texted… in 2012
Expected: | Top-3:
Expected: | Top-10:
Least similar emoji:
Query: Spiraling in 3...2...1...
Expected: | Top-3:
Expected: | Top-10:
Least similar emoji:
Query: Never trusting street food again
Expected: | Top-3:
Expected: | Top-10:
Least similar emoji:
Query: After months of saving, I finally booked the flight, only to discover my passport expired yesterday—my wanderlust is crying in the corner.
0:00 / 0:23
TNG TECHNOLOGY CONSULTING

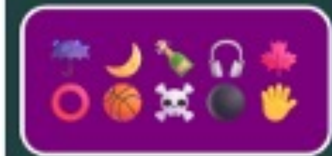
Re-fining fine-tuning



TNG TECHNOLOGY CONSULTING

Re-fining fine-tuning

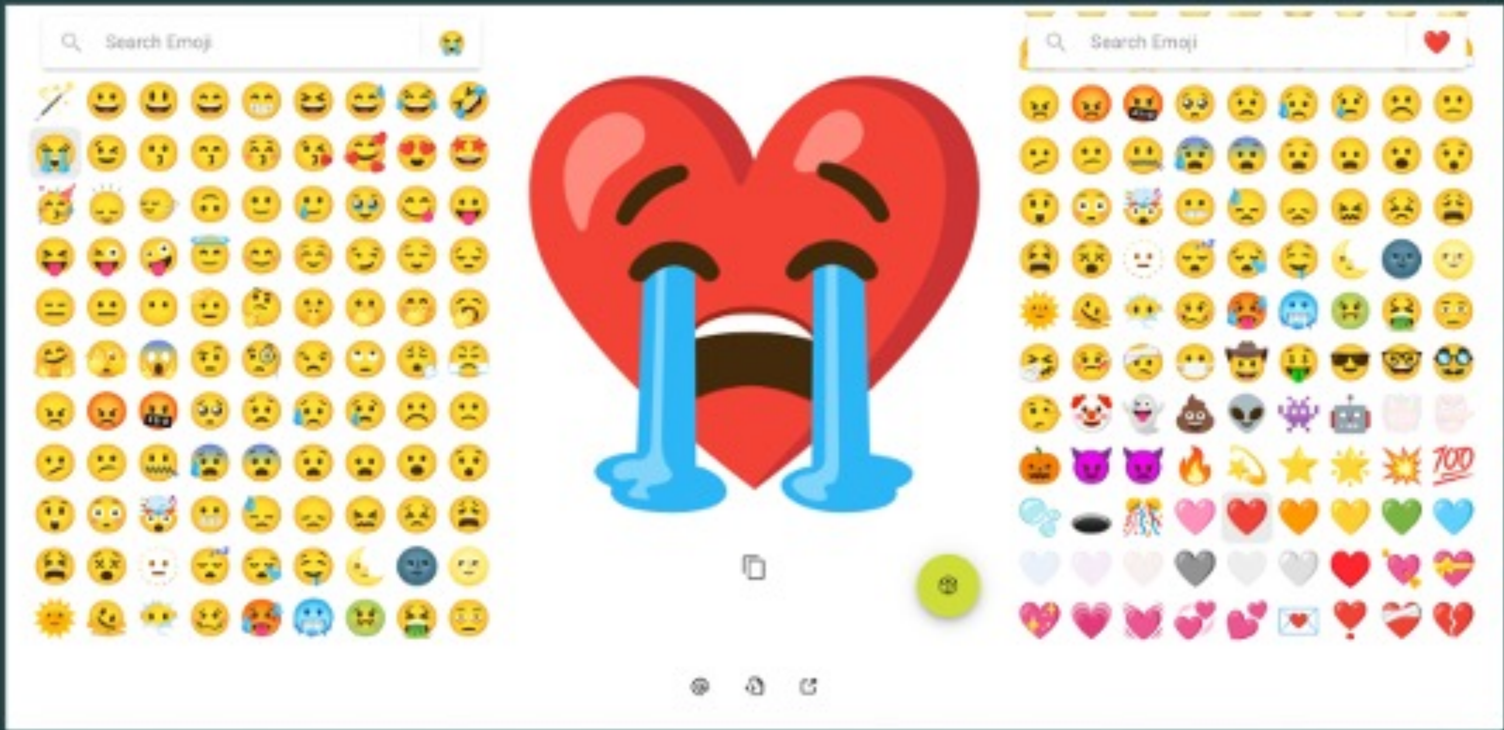




Overview

- History
- Semantic Search
- Fine-Tuning Base Models
- Benchmarking Embeddings
- Embedding of Non-Text Data
- Semantic Search at Scale
- Outlook

Contrastive Language-Image Pre-Training

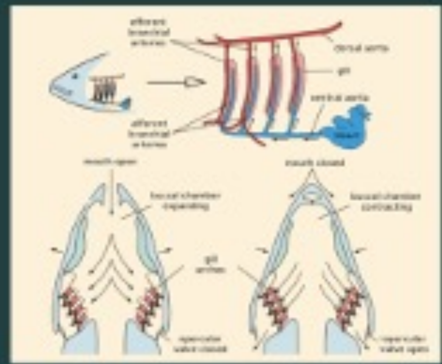


TNG TECHNOLOGY CONSULTING

Contrastive Language-Image Pre-Training

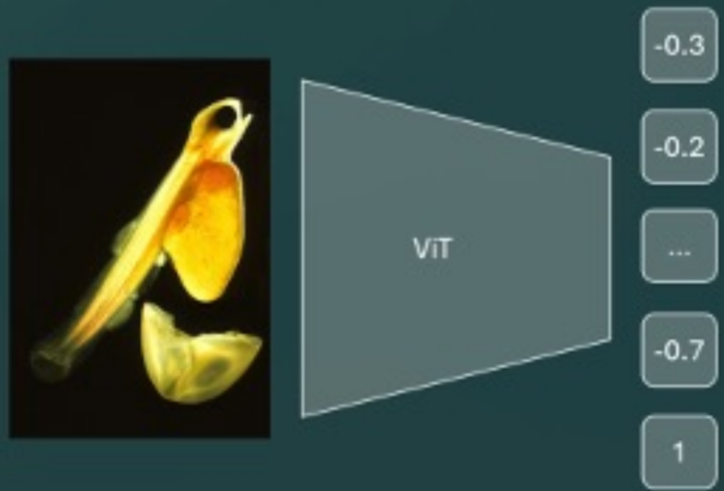
A fish

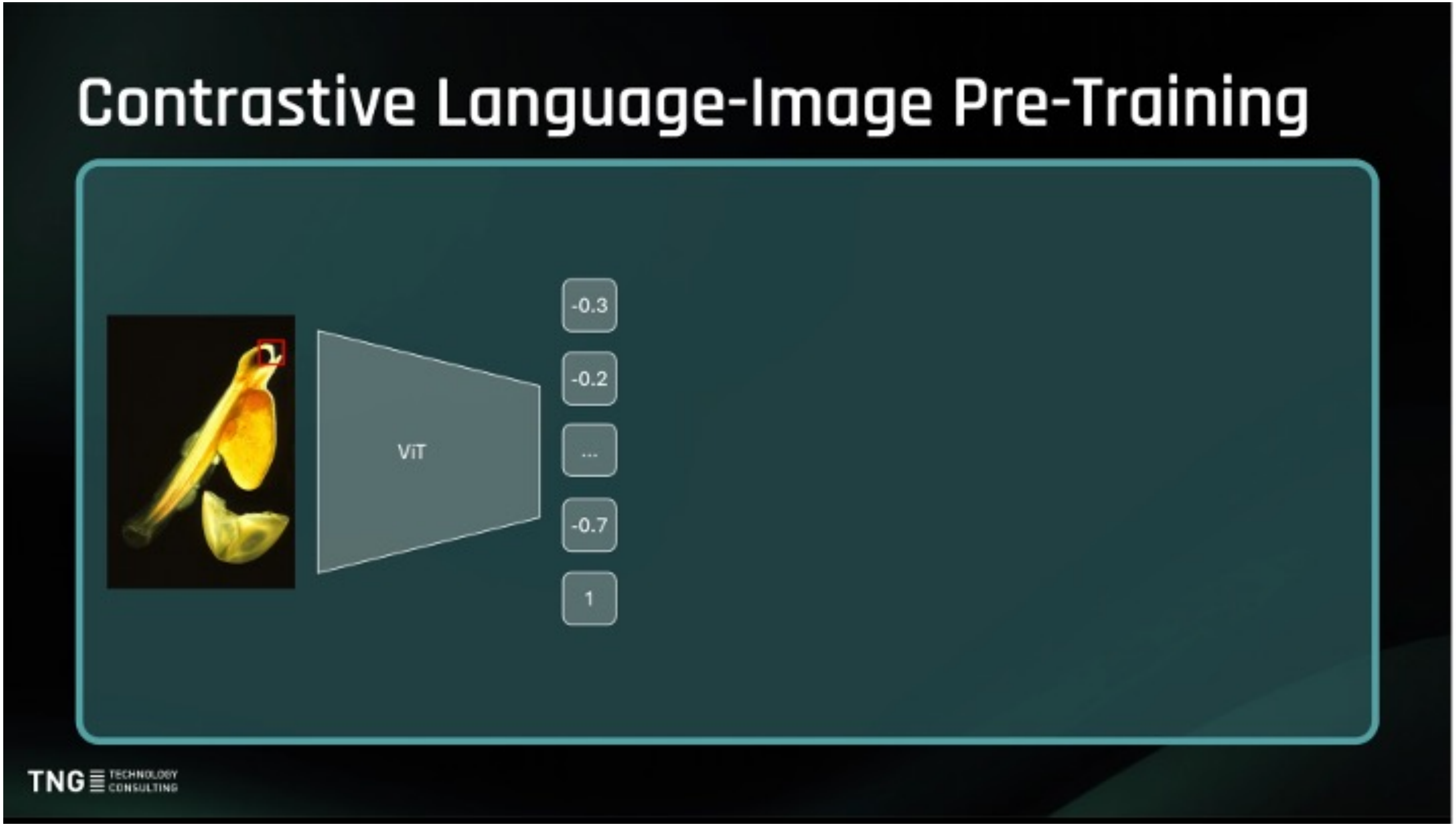
Salmon fry hatching
from the egg,
keeping its yolk sac

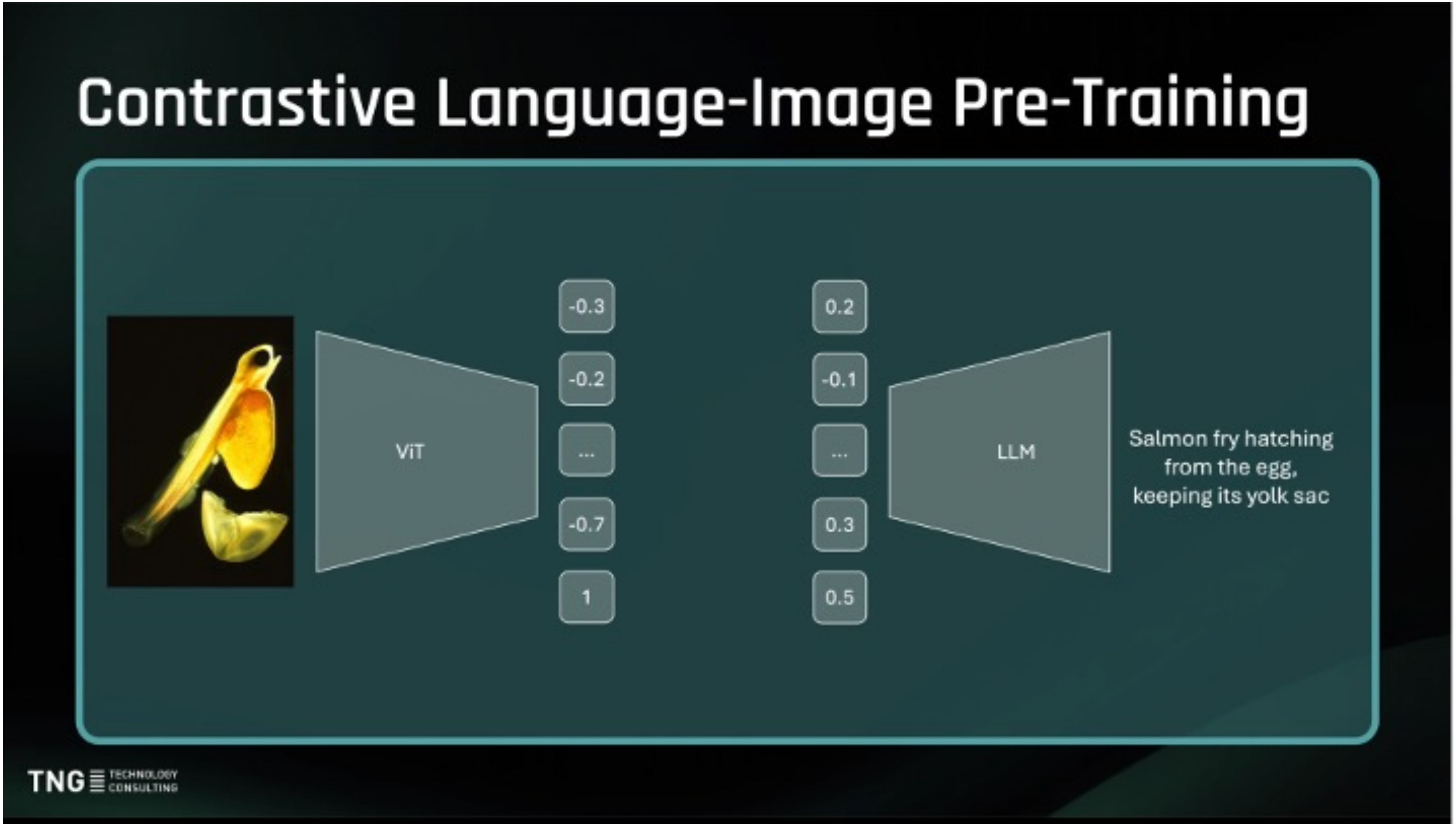


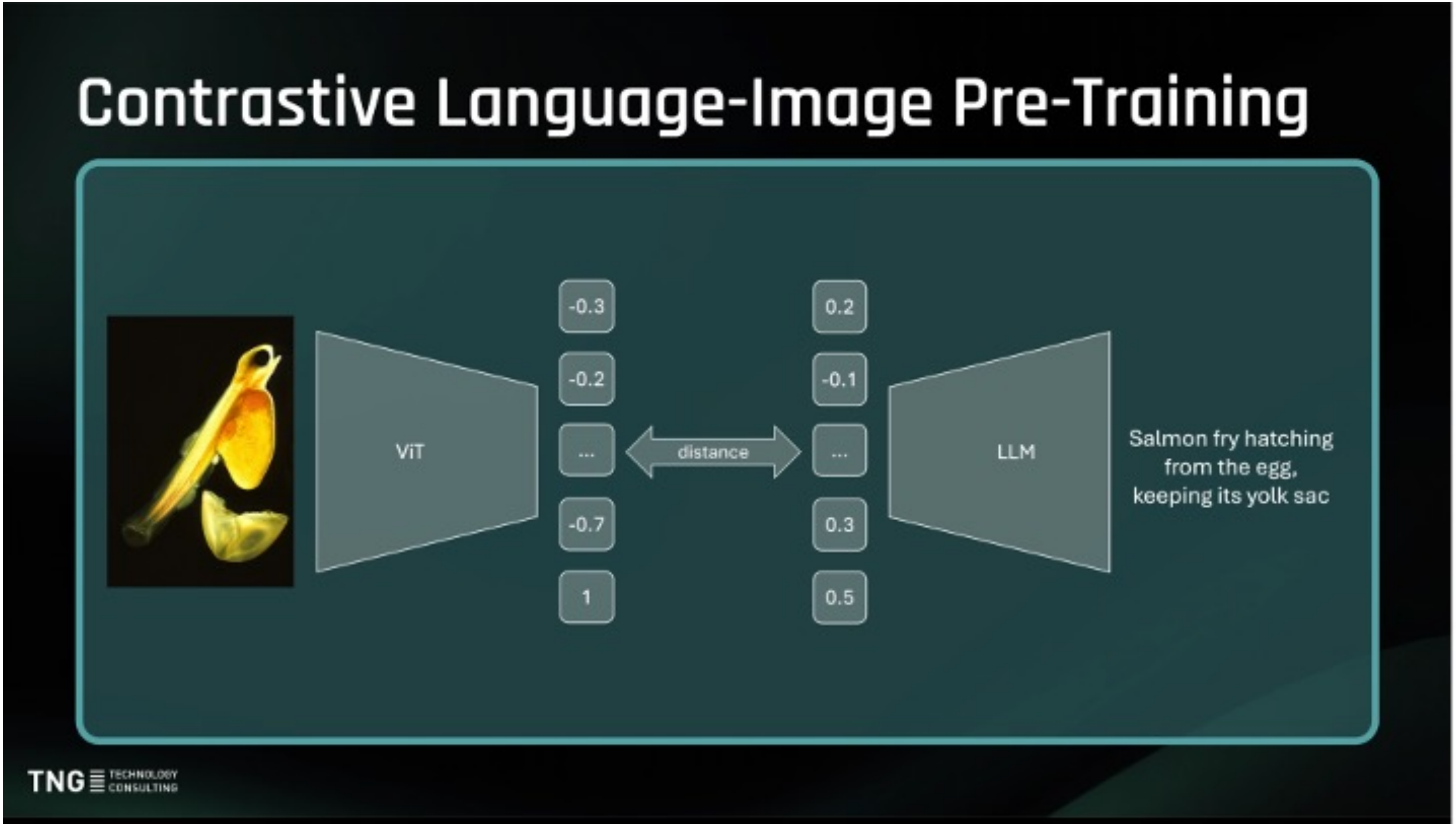
The fish heart pumps
blood to the gills,
where it picks up
oxygen. The blood
then flows without
further pumping to
the body, from where
it returns to the
heart.

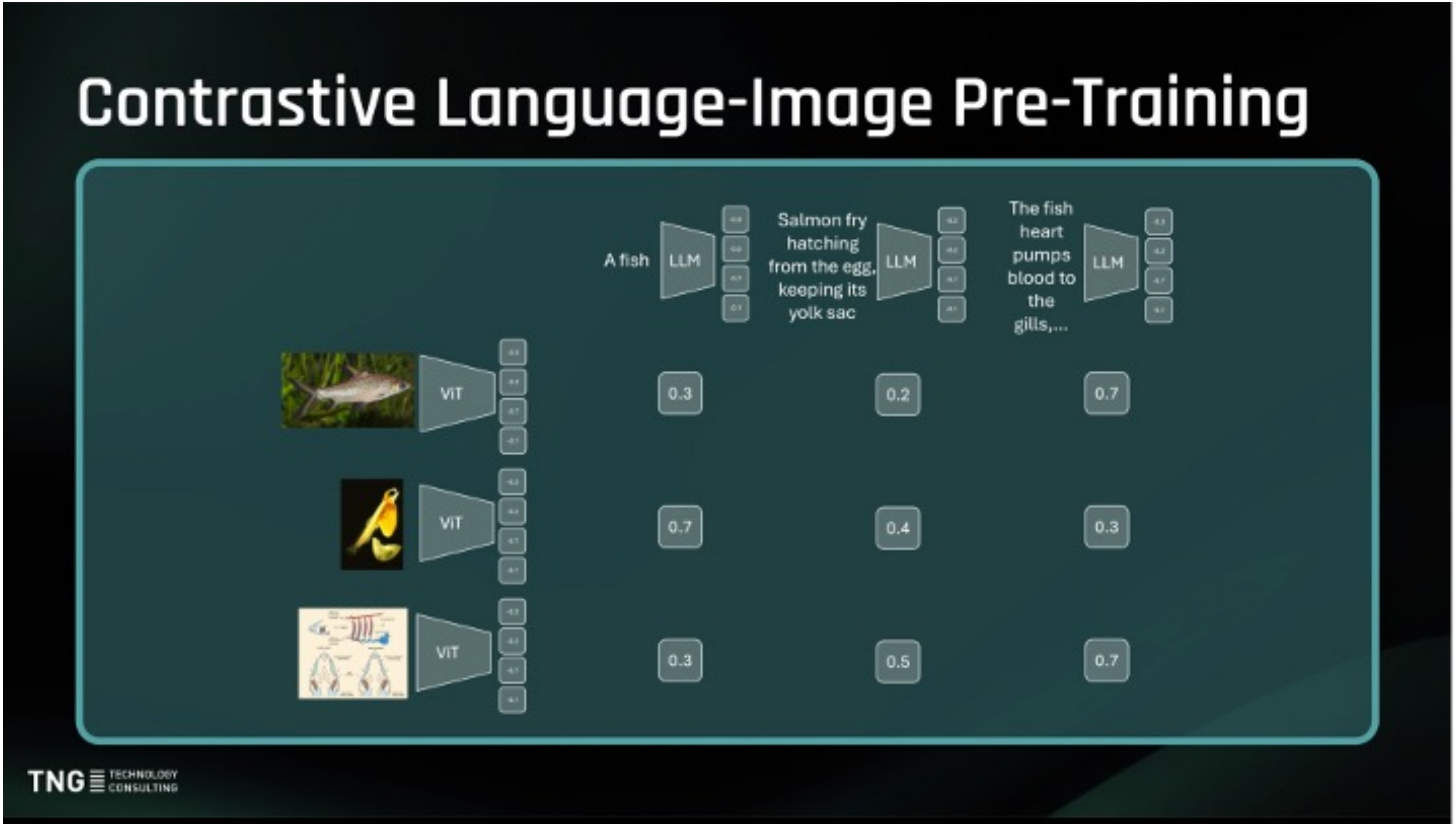
Contrastive Language-Image Pre-Training




Contrastive Language-Image Pre-Training
ViT
-0.3
-0.2
...
-0.7
1
TNG TECHNOLOGY CONSULTING

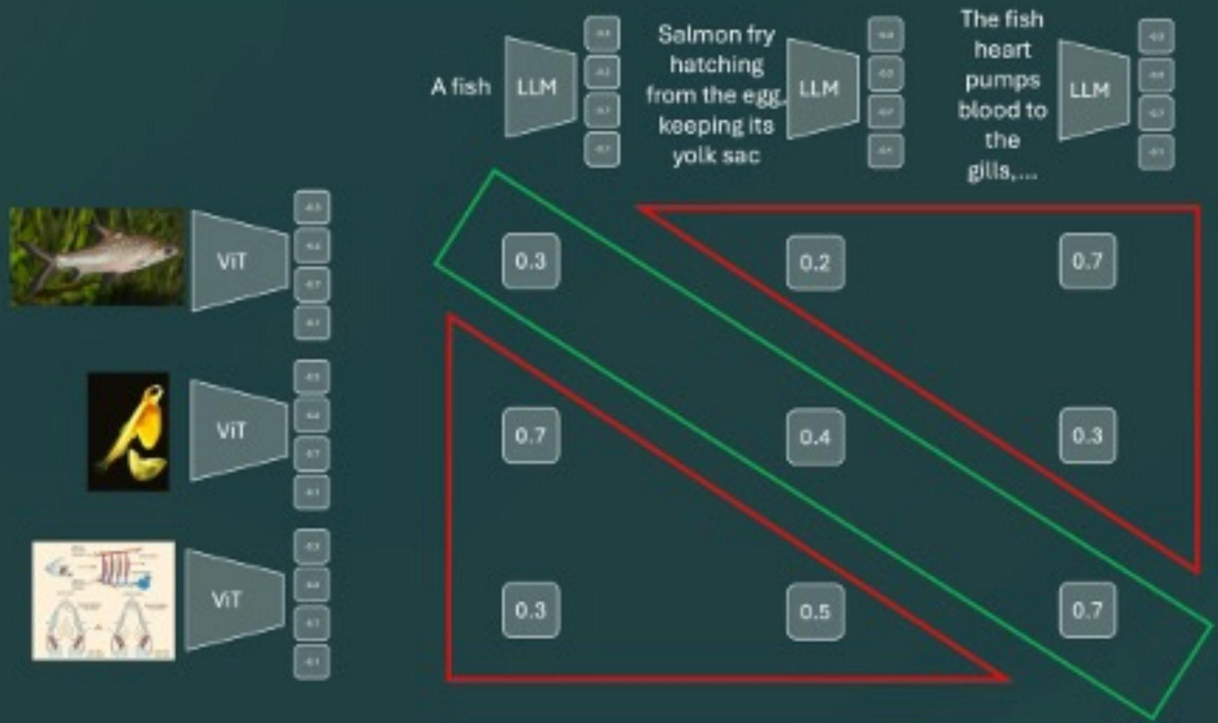

Contrastive Language-Image Pre-Training
ViT
-0.3
-0.2
...
-0.7
1
0.2
-0.1
...
0.3
0.5
LLM
Salmon fry hatching
from the egg,
keeping its yolk sac
TNG TECHNOLOGY CONSULTING


Contrastive Language-Image Pre-Training
ViT
-0.3
-0.2
...
-0.7
1
distance
0.2
-0.1
...
0.3
0.5
LLM
Salmon fry hatching
from the egg,
keeping its yolk sac
TNG TECHNOLOGY CONSULTING



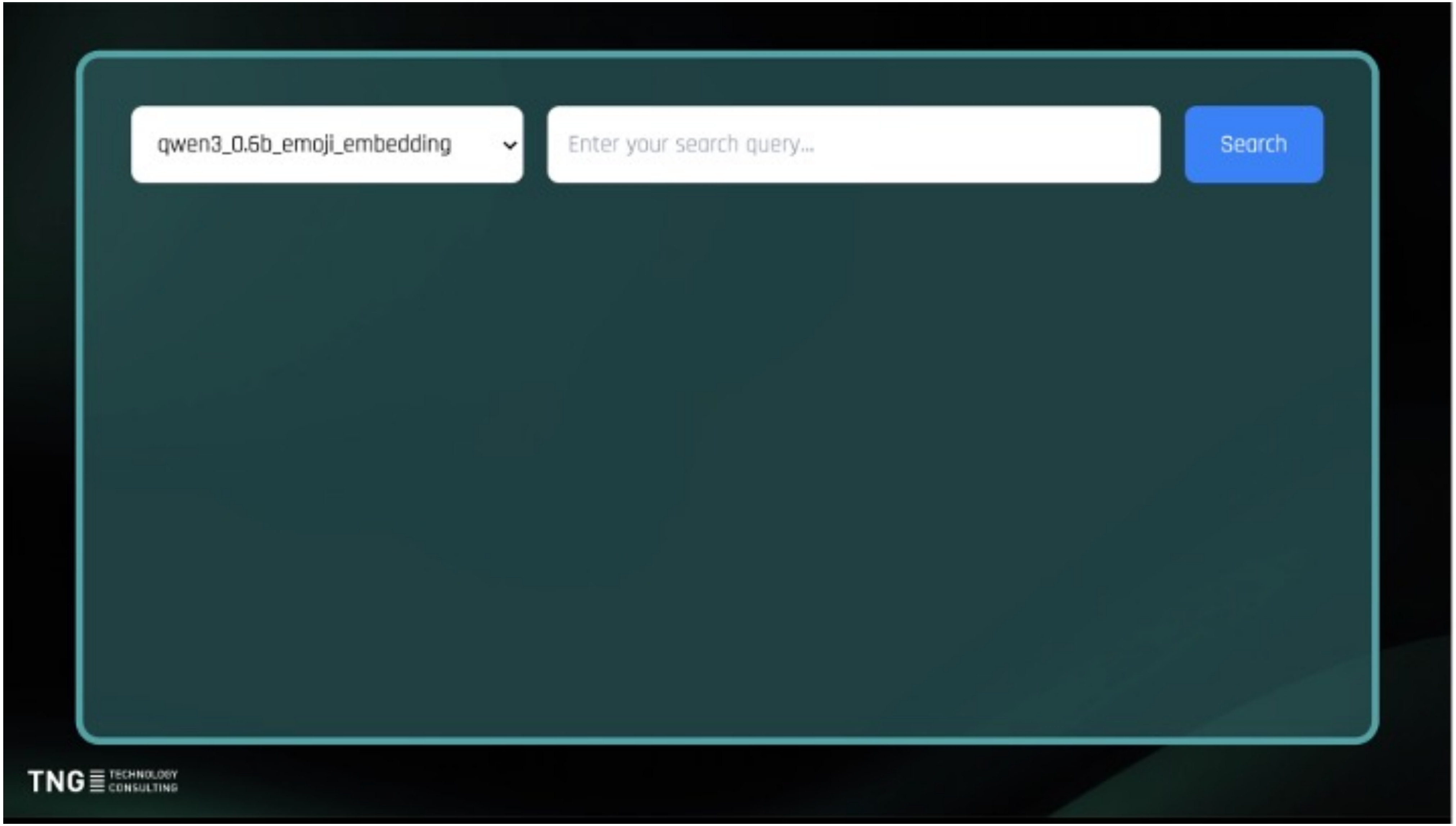
Contrastive Language-Image Pre-Training
A fish
LLM
Salmon fry hatching from the egg, keeping its yolk sac
LLM
The fish heart pumps blood to the gills,...
LLM
ViT
0.3
0.2
0.7
ViT
0.7
0.4
0.3
ViT
0.3
0.5
0.7
TNG TECHNOLOGY CONSULTING

Contrastive Language-Image Pre-Training

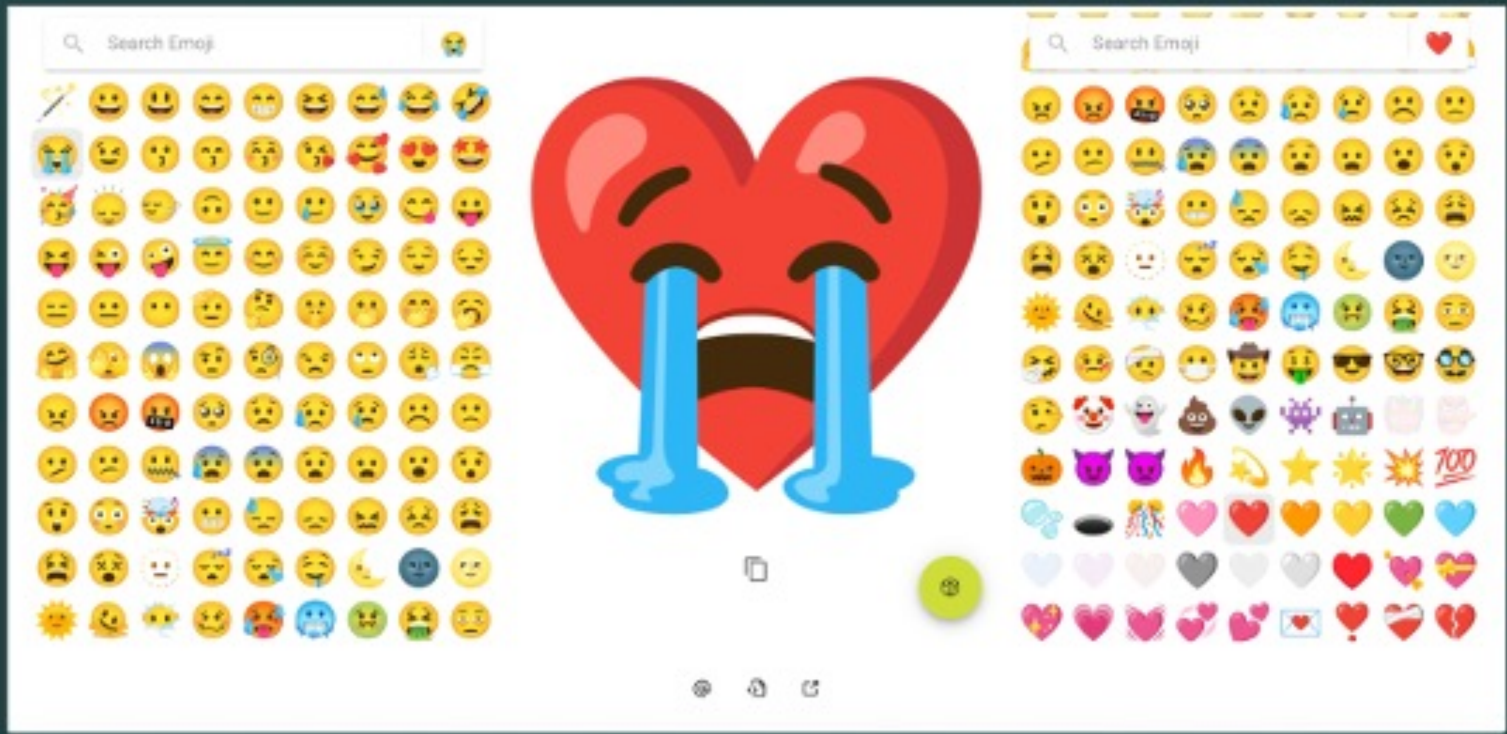


Contrastive Language-Image Pre-Training

CLIP (OpenAI)	→	CLIP (Meta)	→	CLIP 2 (Meta)
English only		English only		300+ Languages
2021		2023		2025
400M Pairs		2.5B Pairs		44% English Pairs


qwen3_0.6b_emoji_embedding
Enter your search query...
Search
TNG TECHNOLOGY CONSULTING

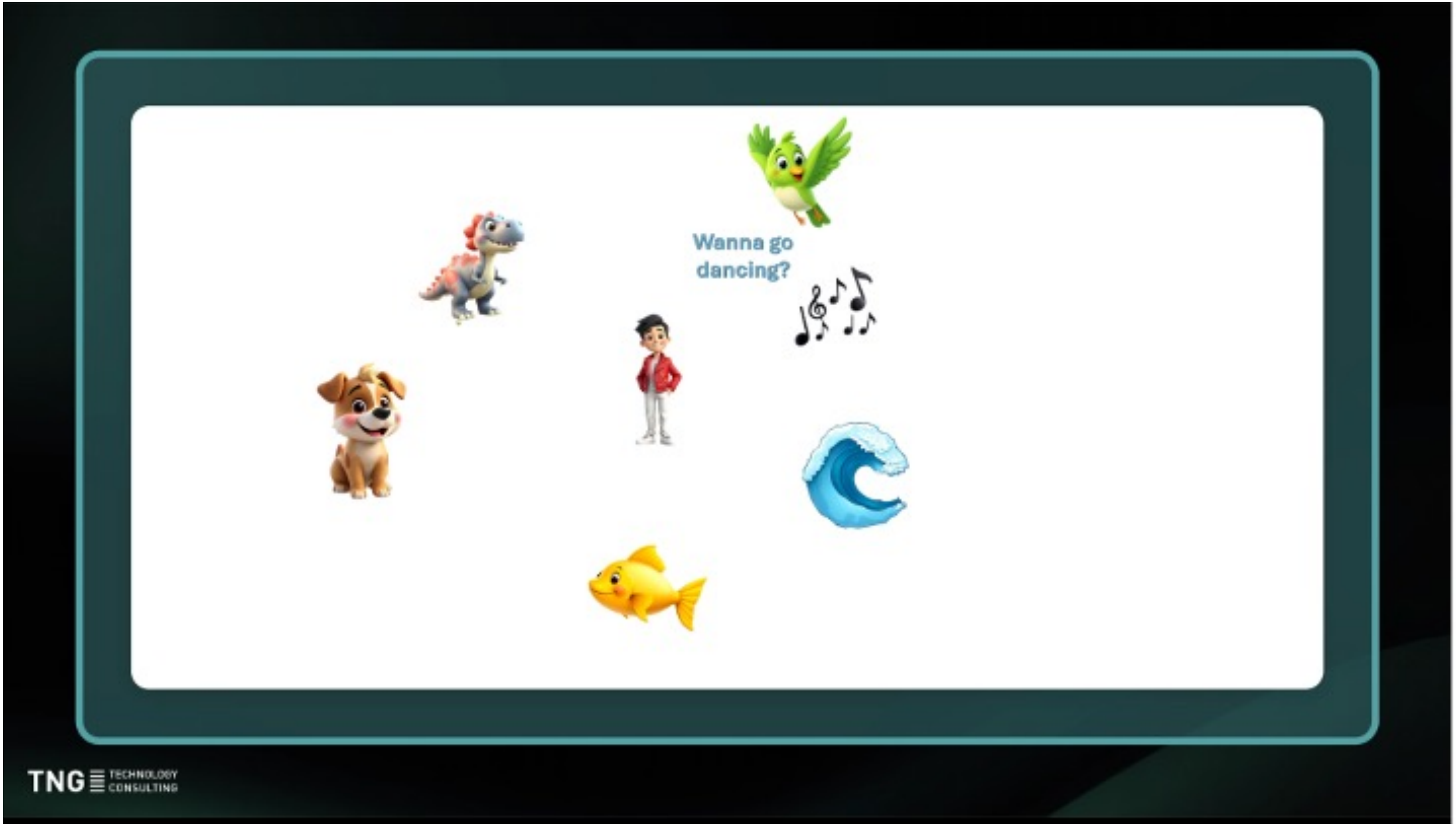
Contrastive Language-Image Pre-Training



TNG TECHNOLOGY CONSULTING

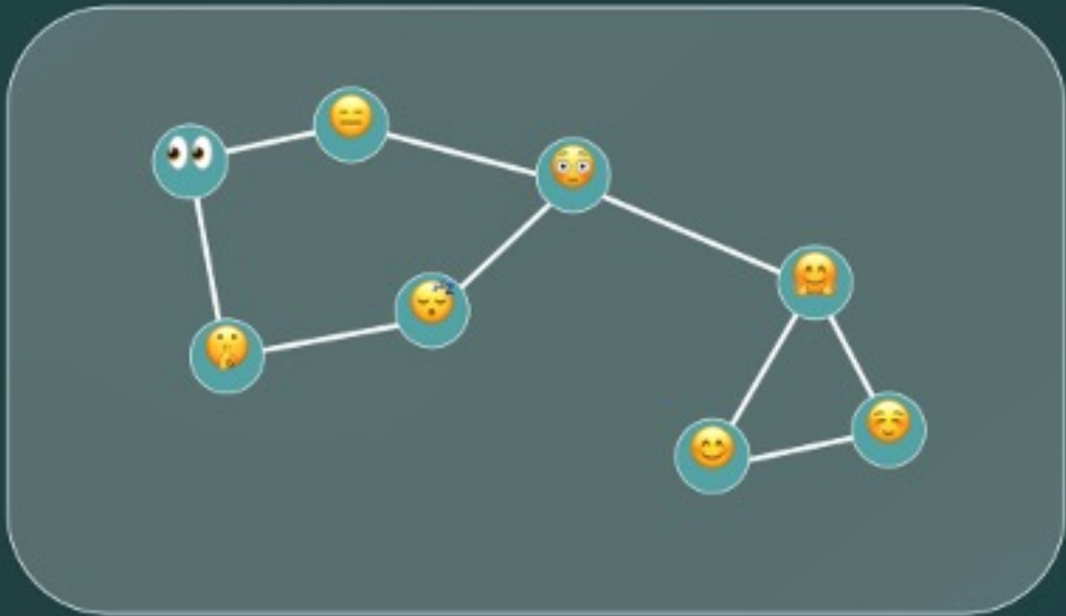
Overview

- History
- Semantic Search
- Fine-Tuning Base Models
- Benchmarking Embeddings
- Embedding of Non-Text Data
- Semantic Search at Scale
- Outlook


Wanna go
dancing?
TNG TECHNOLOGY CONSULTING

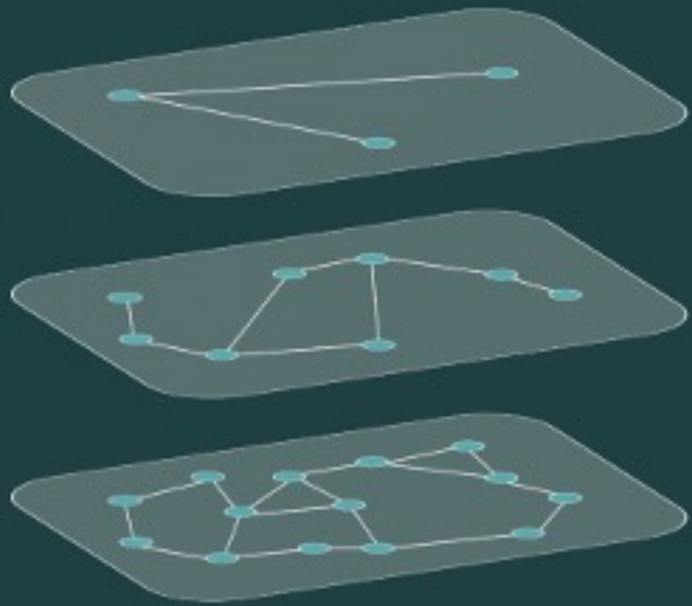
Scale Your Search to Large Datasets

Hierarchical Navigable Small World



Scale Your Search to Large Datasets

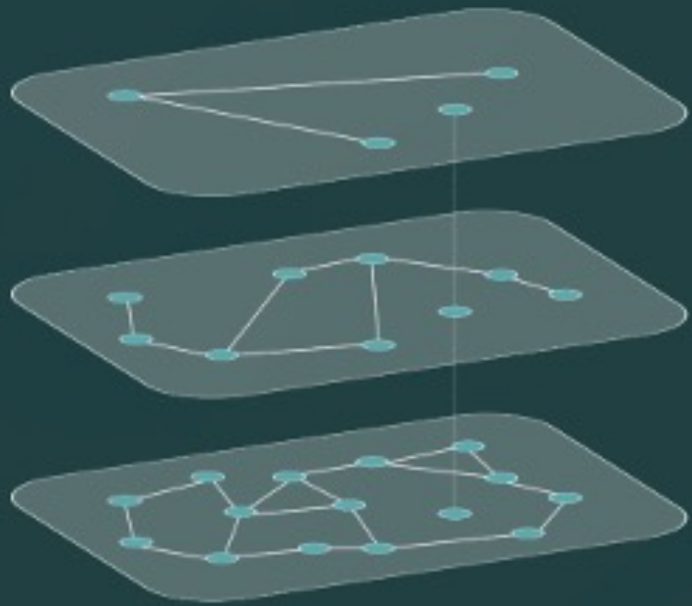
Hierarchical Navigable Small World



TNG TECHNOLOGY CONSULTING

Scale Your Search to Large Datasets

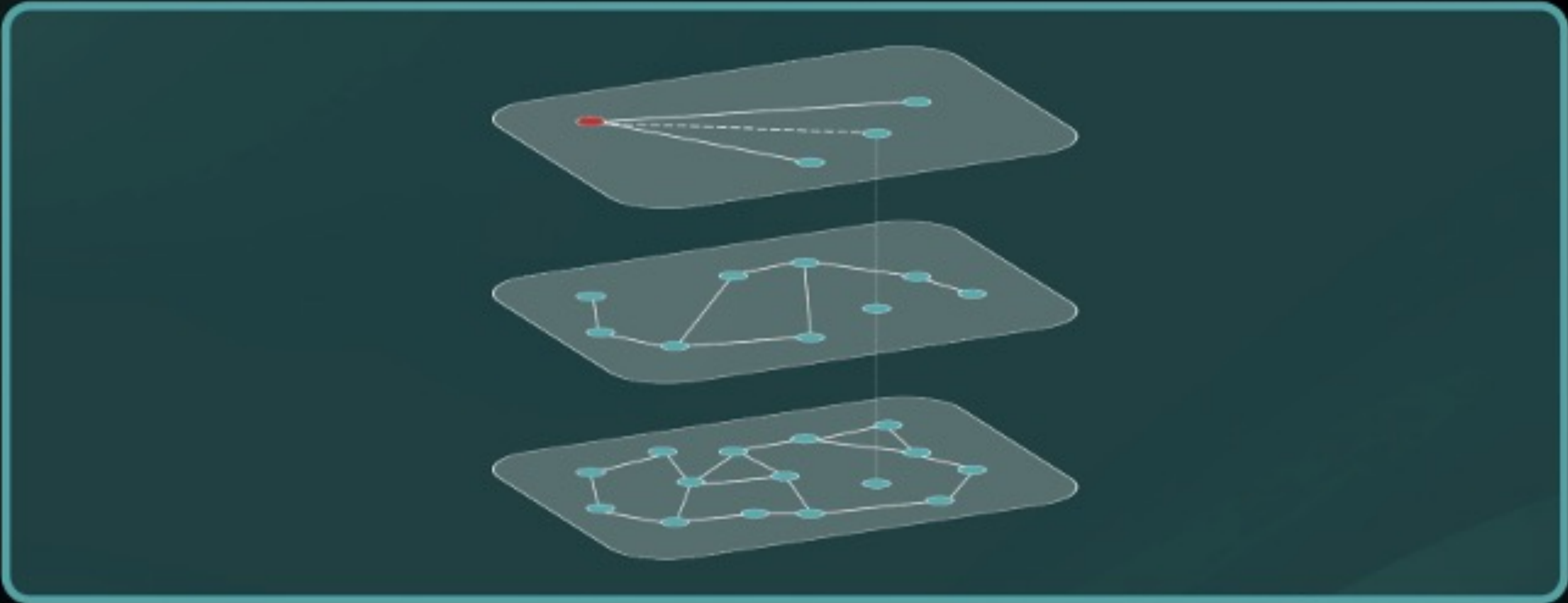
Hierarchical Navigable Small World



TNG TECHNOLOGY CONSULTING

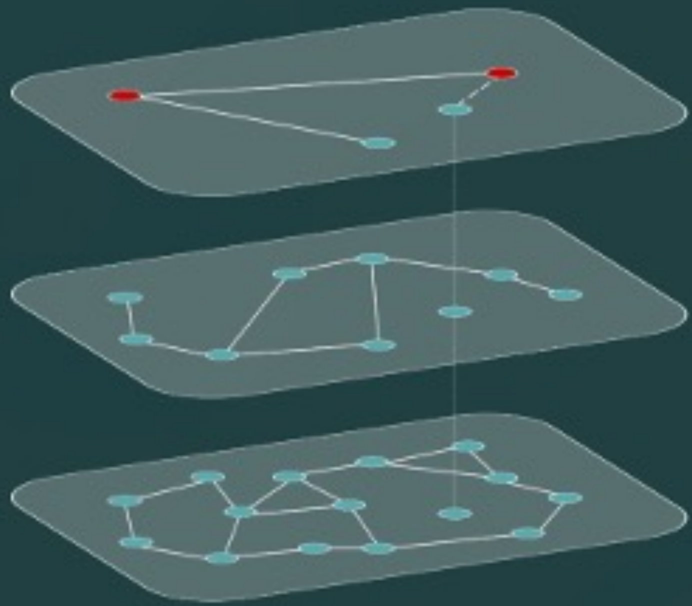
Scale Your Search to Large Datasets

Hierarchical Navigable Small World



Scale Your Search to Large Datasets

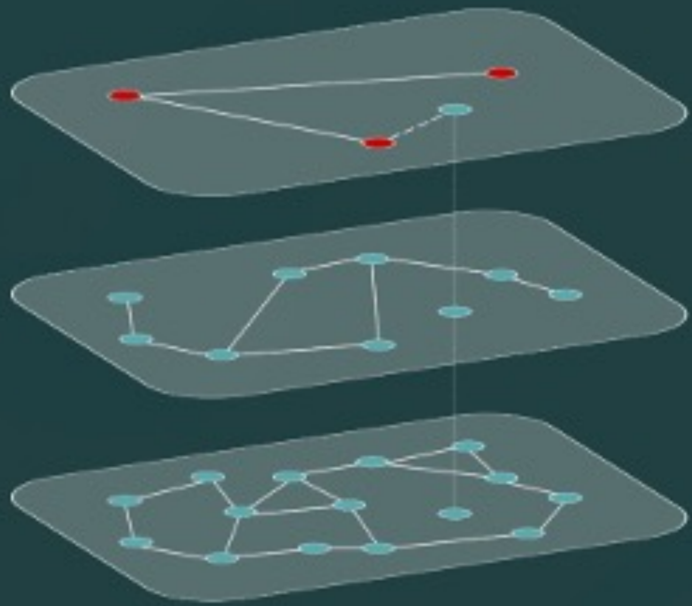
Hierarchical Navigable Small World



TNG TECHNOLOGY CONSULTING

Scale Your Search to Large Datasets

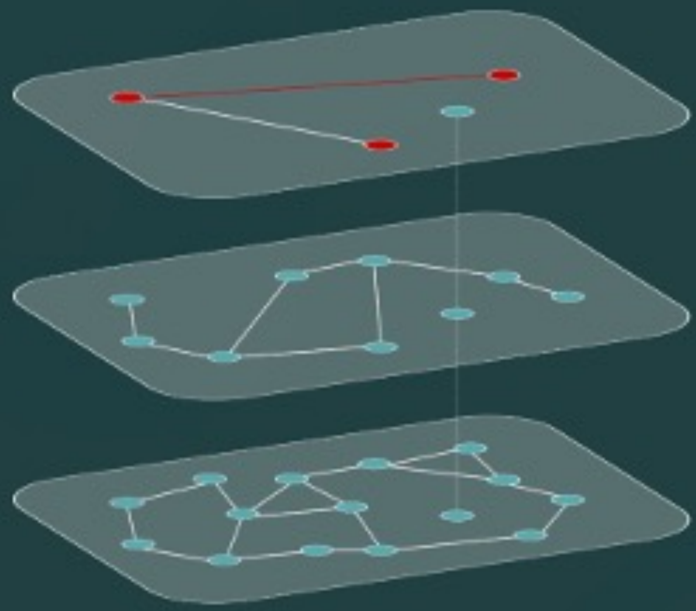
Hierarchical Navigable Small World



TNG TECHNOLOGY CONSULTING

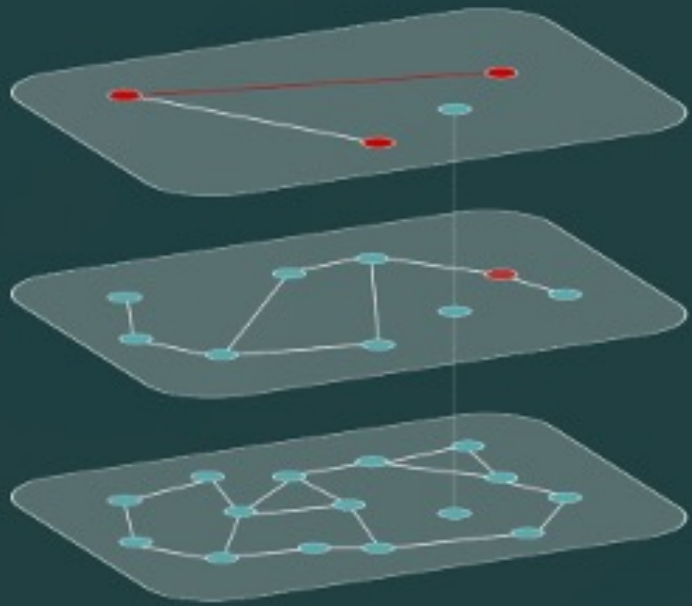
Scale Your Search to Large Datasets

Hierarchical Navigable Small World



Scale Your Search to Large Datasets

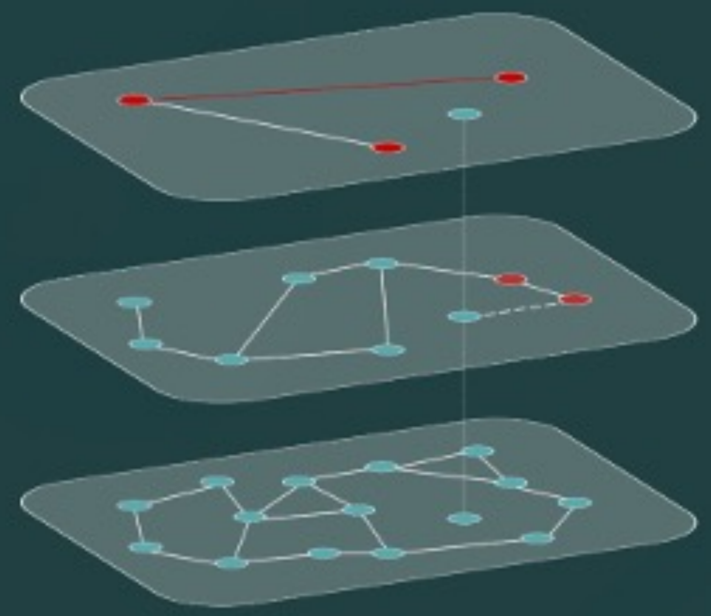
Hierarchical Navigable Small World



TNG TECHNOLOGY CONSULTING

Scale Your Search to Large Datasets

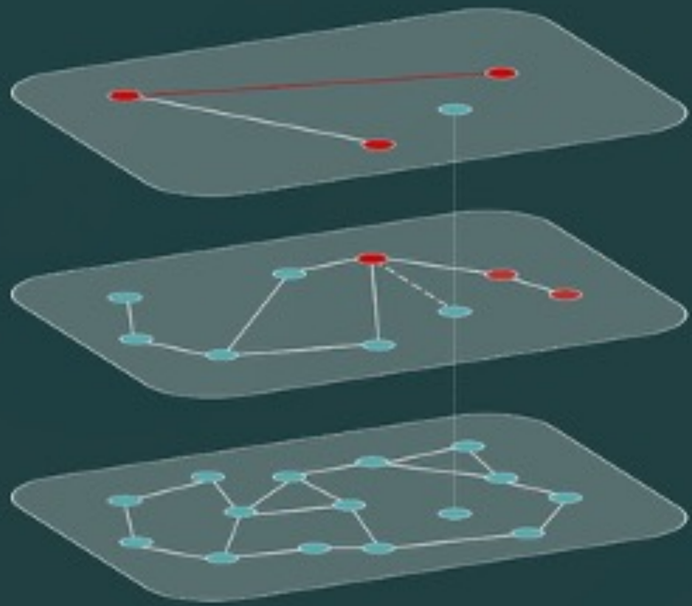
Hierarchical Navigable Small World



TNG TECHNOLOGY CONSULTING

Scale Your Search to Large Datasets

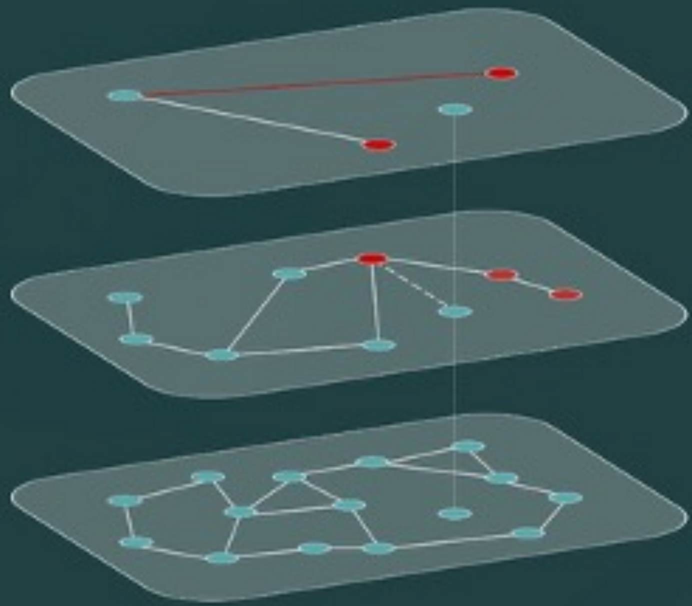
Hierarchical Navigable Small World



TNG TECHNOLOGY CONSULTING

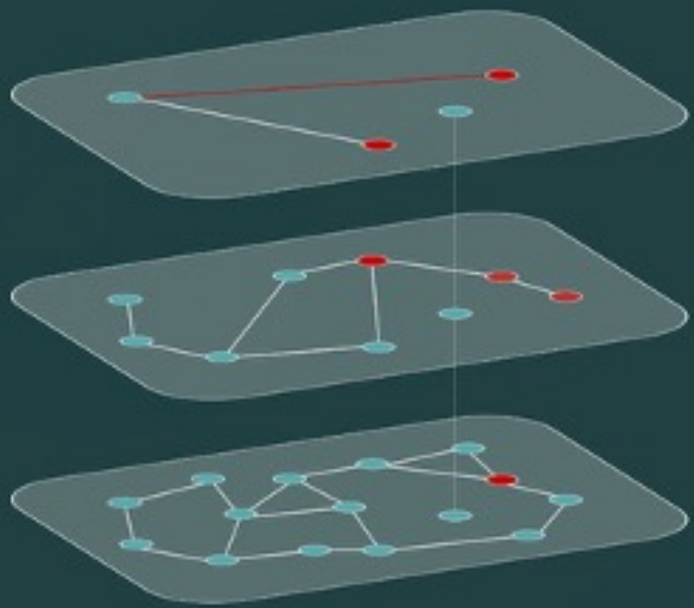
Scale Your Search to Large Datasets

Hierarchical Navigable Small World



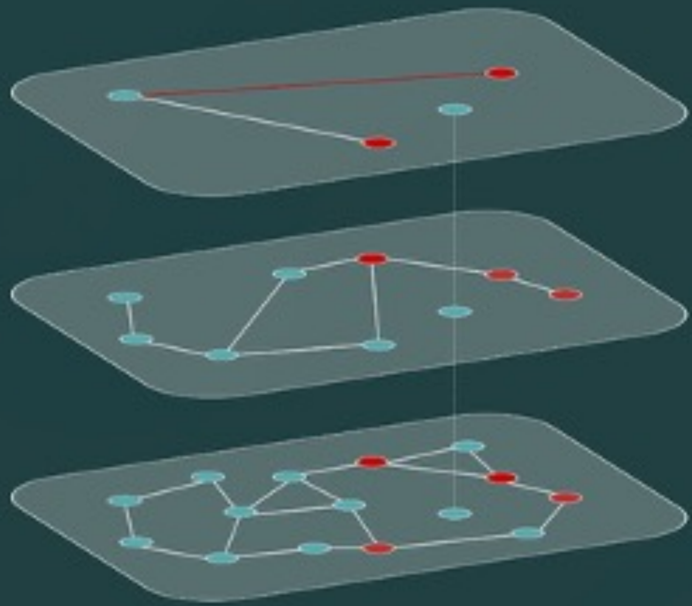
Scale Your Search to Large Datasets

Hierarchical Navigable Small World



Scale Your Search to Large Datasets

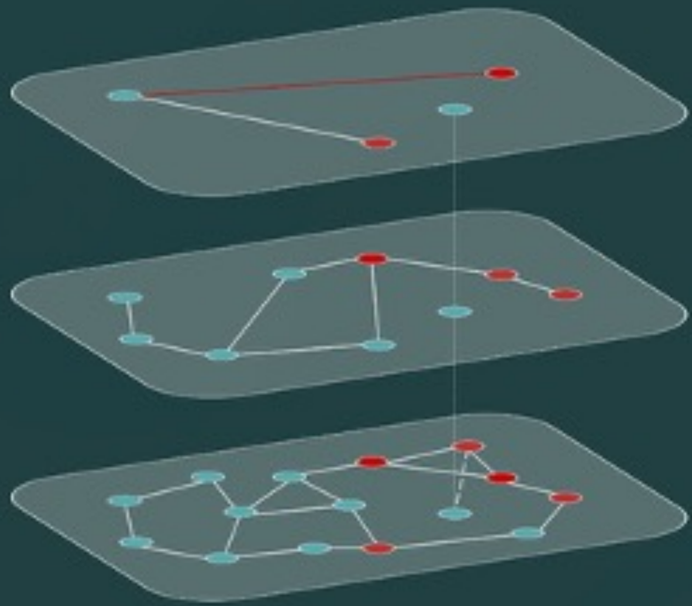
Hierarchical Navigable Small World



TNG TECHNOLOGY CONSULTING

Scale Your Search to Large Datasets

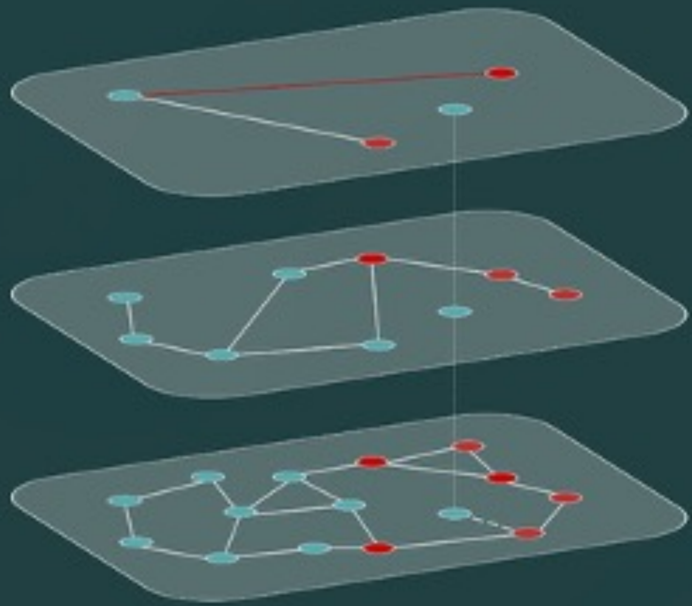
Hierarchical Navigable Small World



TNG TECHNOLOGY CONSULTING

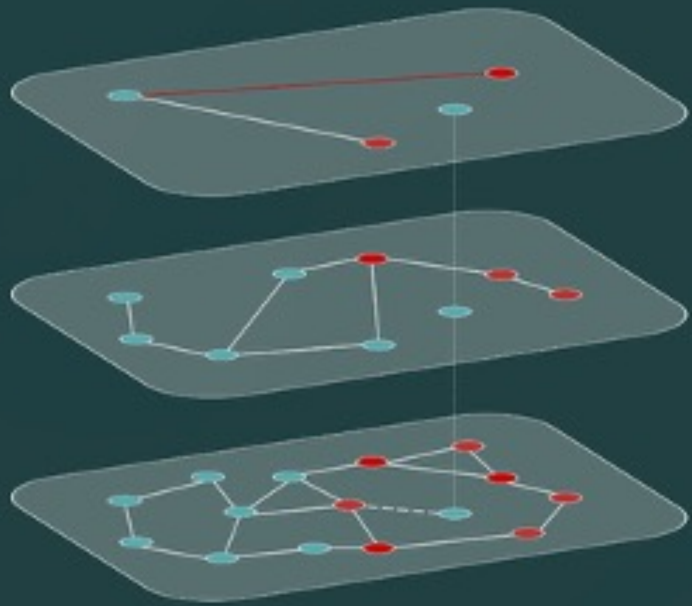
Scale Your Search to Large Datasets

Hierarchical Navigable Small World



Scale Your Search to Large Datasets

Hierarchical Navigable Small World



Overview

- History
- Semantic Search
- Fine-Tuning Base Models
- Benchmarking Embeddings
- Embedding of Non-Text Data
- Semantic Search at Scale
- Outlook

TNG TECHNOLOGY CONSULTING

Outlook

Recommender Systems

Outlook

Fraud Detection

Uber ebay

Thank you!

Elias Schecke

Senior Consultant

Dennis Schulz

Senior Consultant