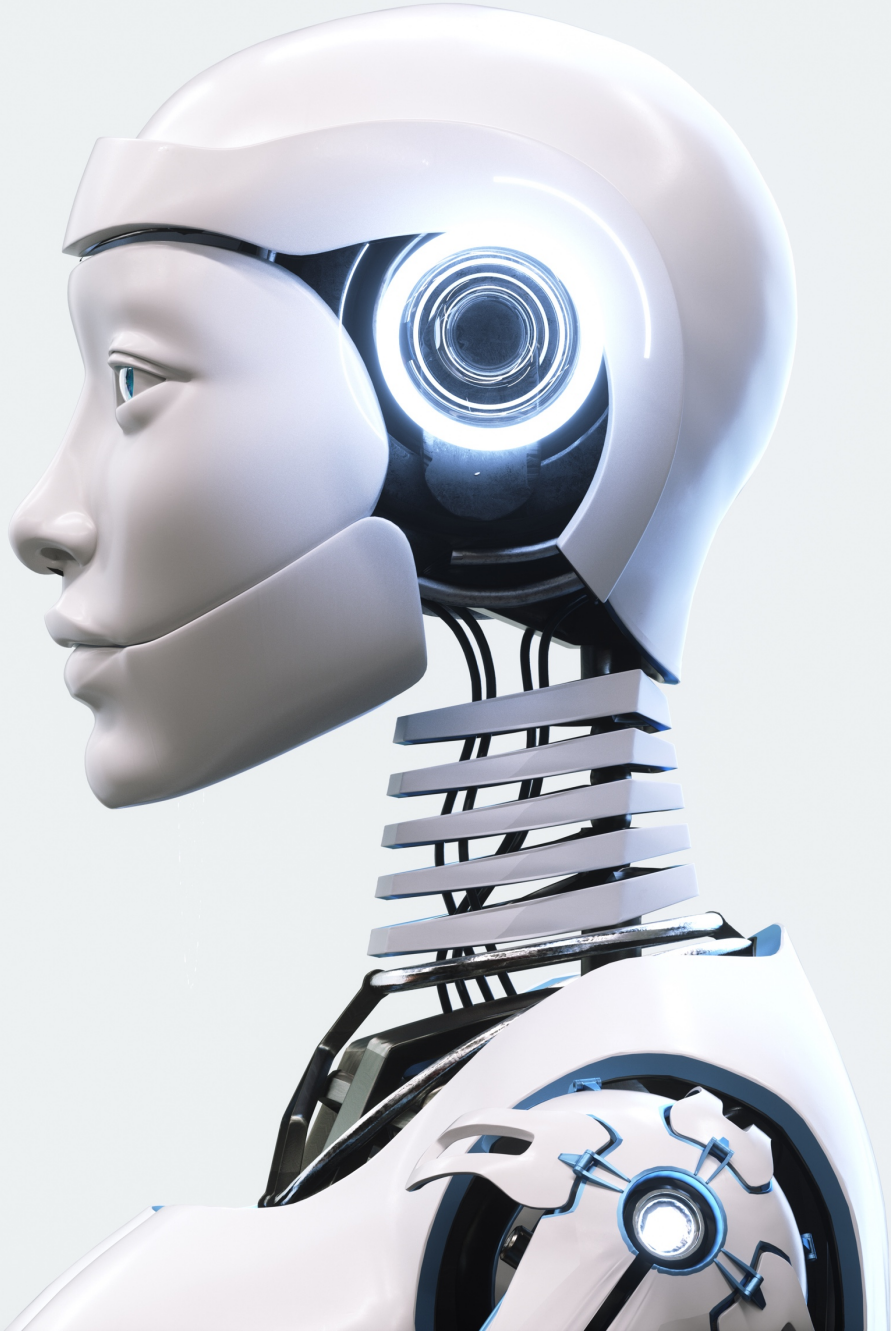
OpenAI | Microsoft

Azure OpenAI Service

for .NET Developers



Microsoft® Most Valuable Professional

Mert Yeter

Cloud Solutions Architect

360 Dotnet

Certified Traefik Ambassador

Cloud Community Champion (2020)

@m3rtyeter

/in/mertyeter

/mertyeter

/azureishlive
/yenimshowto
/TraefikLabsTurkey

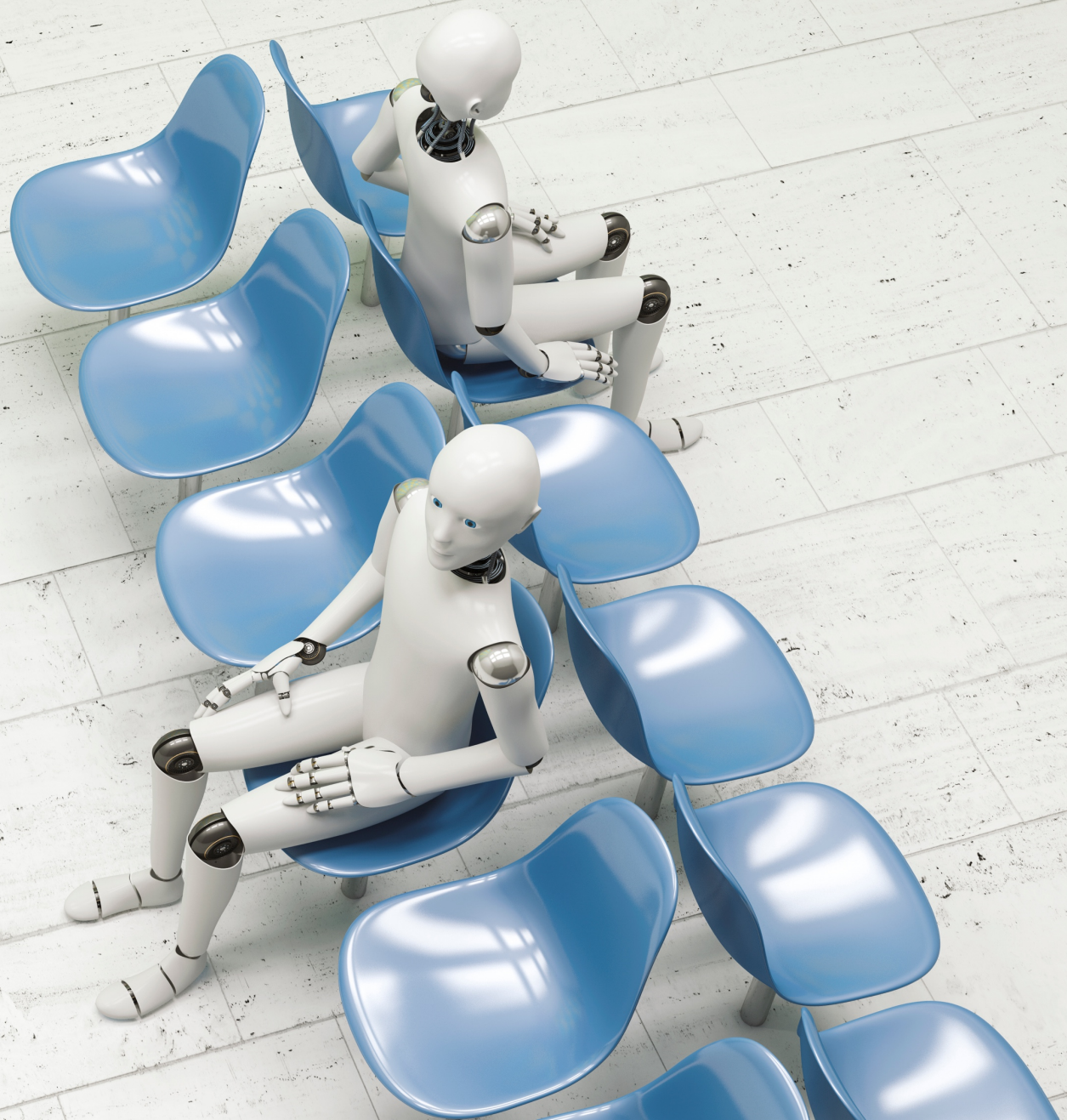
/azureishlive
/mshowto
/traefikistanbul

AZUREISH Live!

msHOWTO ÇÖZÜM SANATTIR

AI Landscape

- Artificial Intelligence
- Machine Learning
- Deep Learning
- Generative AI



Capabilities of OpenAI AI models

Generating Natural Language

Text completion: Generate and edit text

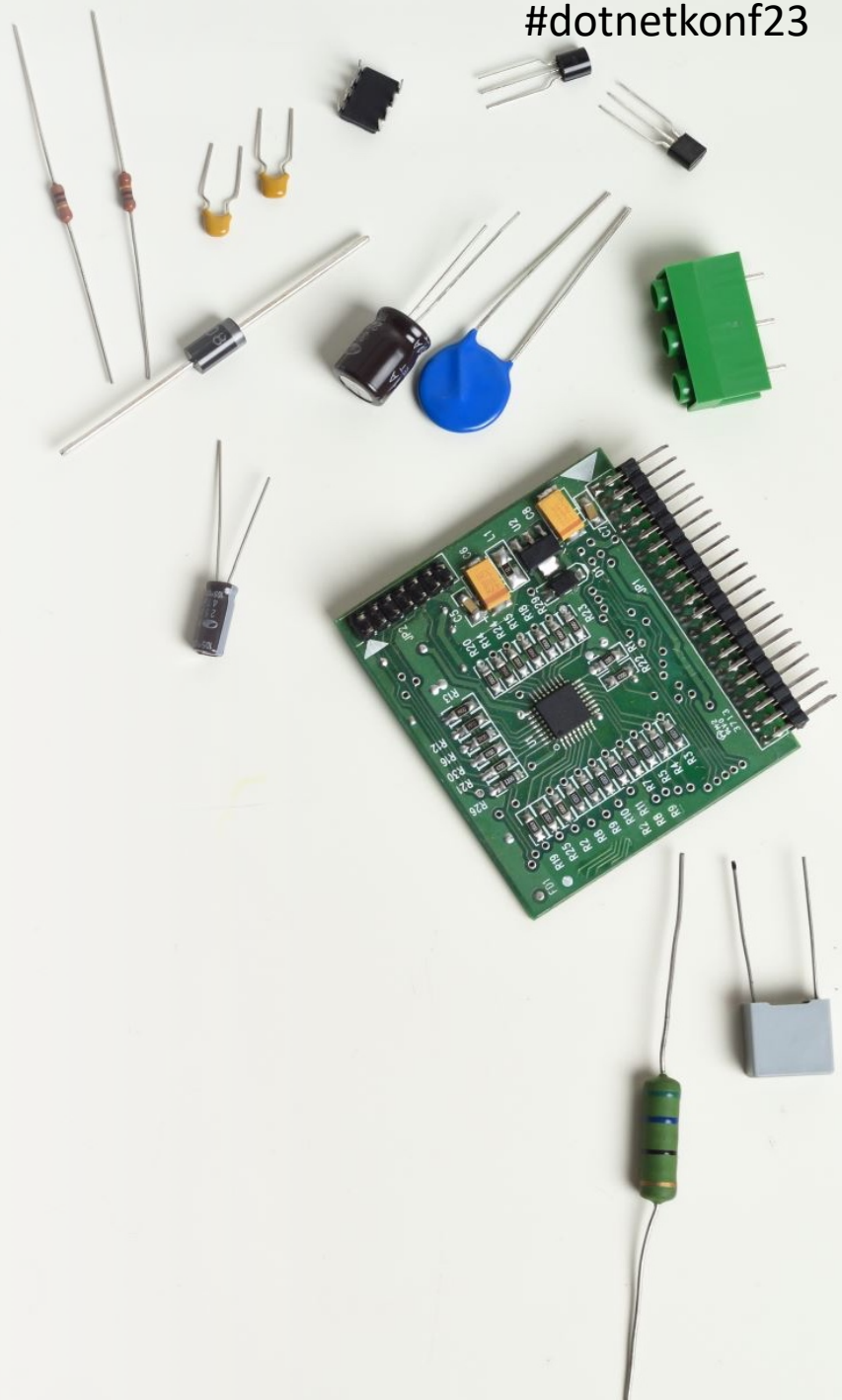
Embeddings: Search, classify and compare text

Generating Code

Generating Images

Azure OpenAI components

- Pre-trained generative AI models
- Customization capabilities; the ability to fine-tune AI models with your own data
- Built-in tools to detect and mitigate harmful use cases so users can implement AI responsibly
- Enterprise-grade security with role-based access control (RBAC) and private networks



Microsoft AI Portfolio

Business Users
& Citizen Developers

Applications

Microsoft 365 | Microsoft Dynamics 365 | Microsoft Edge | Microsoft Bing | Windows | XBOX

Power Platform


Power BI


Power Apps


Power Automate


Power Virtual Agents

Developers &
Data Scientists

Azure AI

Applied AI Services


Bot Service


Cognitive Search


Form Recognizer


Video Indexer


Metrics Advisor


Immersive Reader

Cognitive Services


Vision


Speech

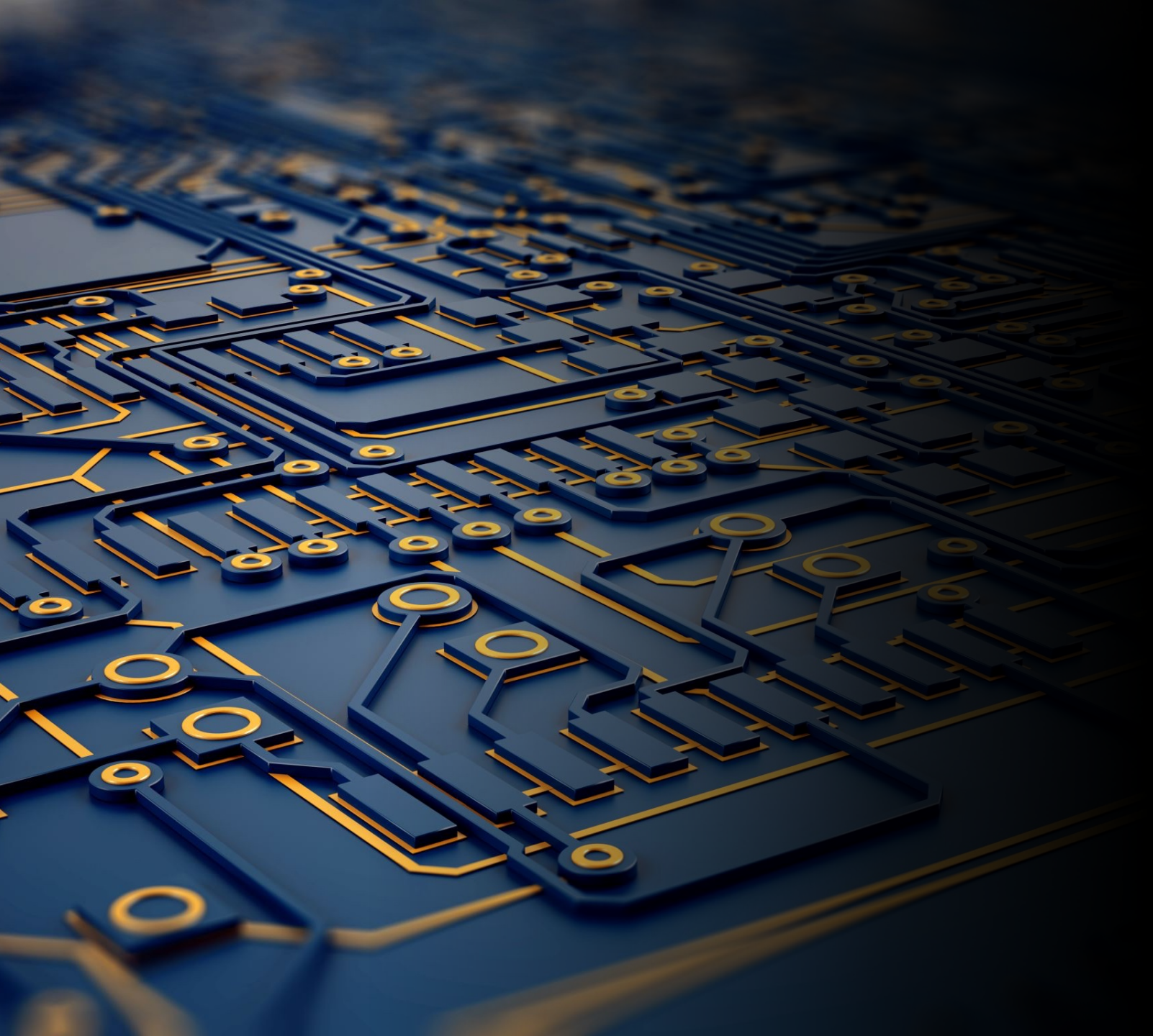

Language


Decision

OpenAI Service

ML Platform


Azure Machine Learning



Models

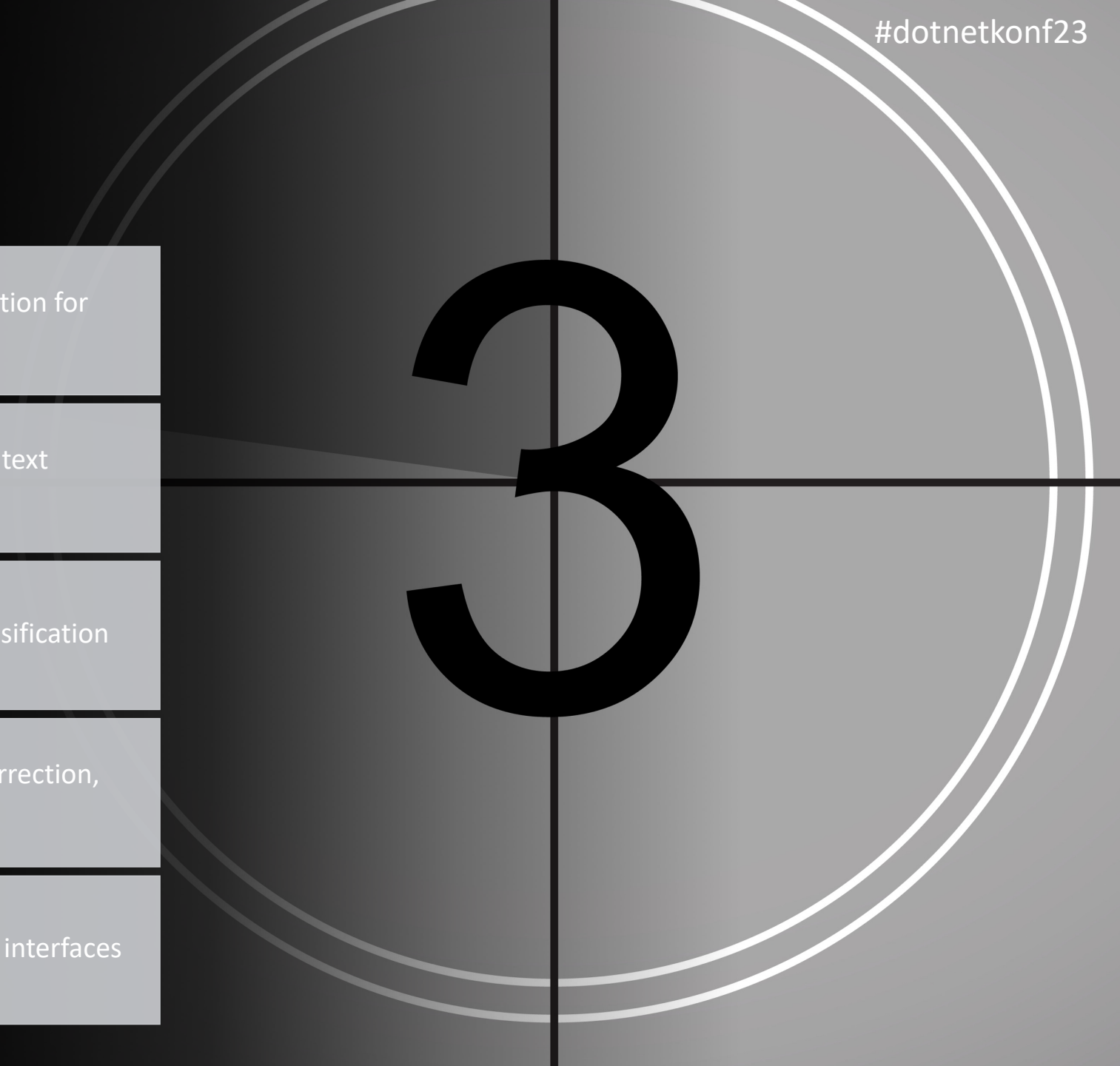
- GPT-4
- GPT-3
- Codex
- Embeddings

GPT-4 Models

- **Gpt-4**
  - supports 8192 max input tokens
- **Gpt-4-32k**
  - supports up to 32,768 tokens

GPT-3 Models

text-davinci-003	Complex intent, cause and effect, summarization for audience
text-curie-001	Language translation, complex classification, text sentiment, summarization
text-babbage-001	Moderate classification, semantic search classification
text-ada-001	Parsing text, simple classification, address correction, keywords
gpt-35-turbo	Language model designed for conversational interfaces



Codex Models

code-davinci-002

Stronger when it comes to analyzing complicated tasks

code-cushman-001

Capable model for many code generation tasks. Cushman typically runs faster and cheaper than Davinci

Embeddings Models

Clustering, regression, anomaly detection, visualization

text-similarity-ada-001
text-similarity-babbage-001
text-similarity-curie-001
text-similarity-davinci-001

Search, context relevance, information retrieval

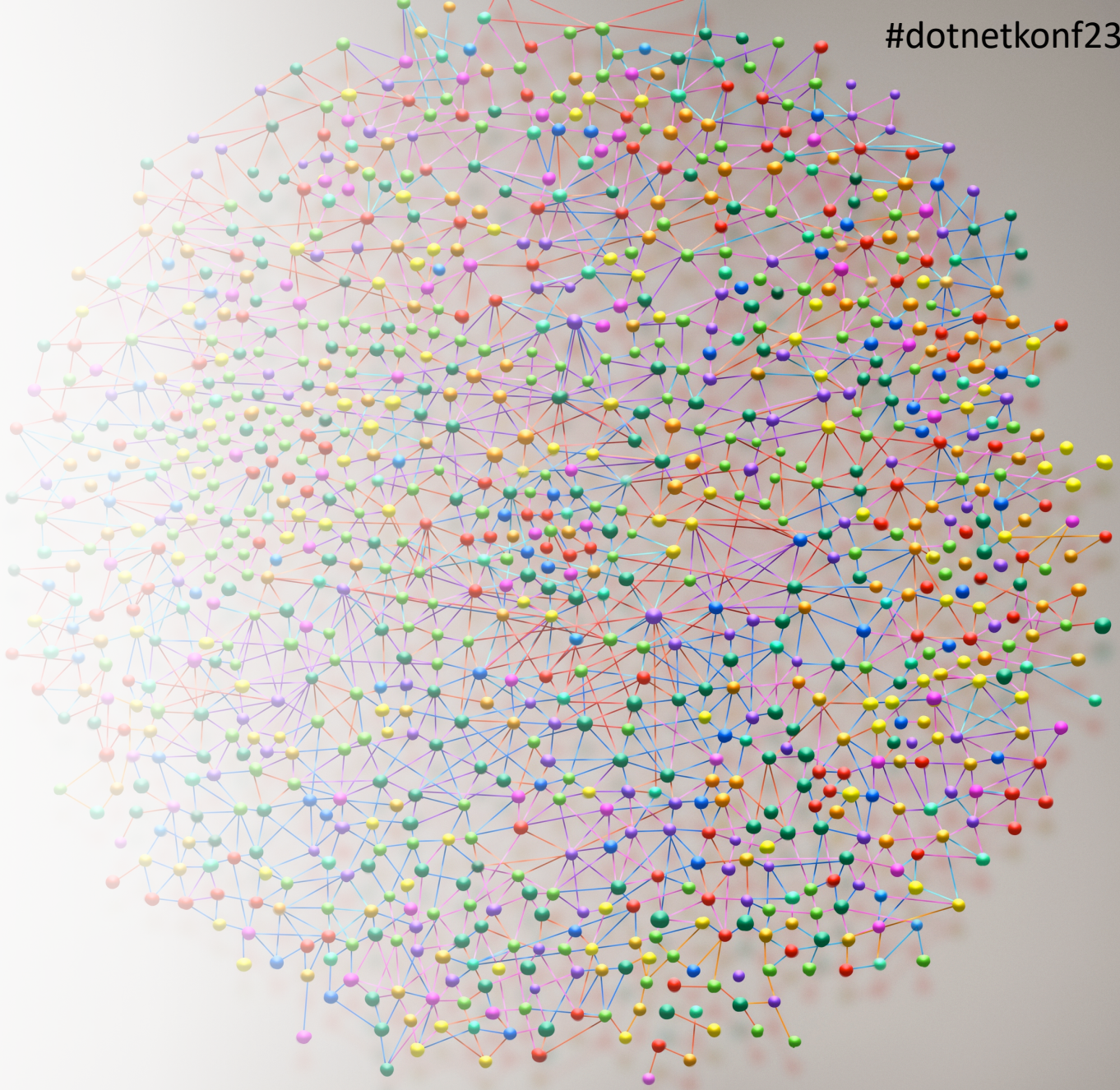
text-search-ada-doc-001
text-search-ada-query-001
text-search-babbage-doc-001
text-search-babbage-query-001
text-search-curie-doc-001
text-search-curie-query-001
text-search-davinci-doc-001
text-search-davinci-query-001

Code search and relevance

code-search-ada-code-001
code-search-ada-text-001
code-search-babbage-code-001
code-search-babbage-text-001

The length of the numerical vector returned by the service, based on model capability

- Ada: 1024 dimensions
- Babbage: 2048 dimensions
- Curie: 4096 dimensions
- Davinci: 12288 dimensions



Azure OpenAI Service for Developers?

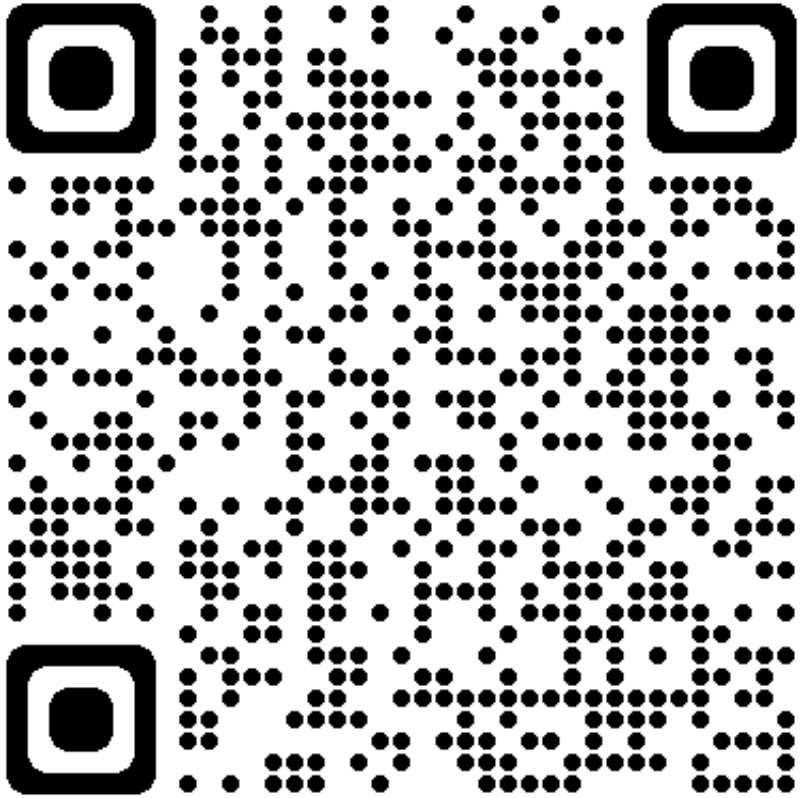
- REST API
- .NET
- Java
- JavaScript


DEMO

Quotas and limits

Limit Name	Limit Value
OpenAI resources per region per Azure subscription	3
Requests per minute per model*	Davinci-models (002 and later): 120 ChatGPT model: 300 GPT-4 models: 18 All other models: 300
Tokens per minute per model*	Davinci-models (002 and later): 40,000 ChatGPT model: 120,000 GPT-4 8k model: 10,000 GPT-4 32k model: 32,000 All other models: 120,000
Max fine-tuned model deployments*	2
Ability to deploy same model to multiple deployments	Not allowed
Total number of training jobs per resource	100
Max simultaneous running training jobs per resource	1
Max training jobs queued	20
Max Files per resource	50
Total size of all files per resource	1 GB
Max training job time (job will fail if exceeded)	720 hours
Max training job size (tokens in training file) x (# of epochs)	2 Billion

Resources



Azure OpenAI Service - Documentation