Microsoft

Enablement Session GCR

听过这集，你就是 OpenAI 专家！

OpenAI will be everywhere!

OpenAI

它是什么，它不是什么？

With or without Azure

有什么区别，我们的价值如何体现？

GPT的限制

Be realistic

Best Practices

如何显得我们很专业！

各种花式应用

来自世界各地能说的，和不能说的应用案例！

Open AI 它是...

预训练模型家族
(俗称大模型)

学习方法

学习内容
(文字类素材)

派生

派生

GPT X	Codex	Embedding
生成文字	生成代码	生成向量

训练得到

Moderation (旧称Content Filter)

GPT架构

地表最强自然语音类深度学习架构，用来理解文字并做出相应的输出

0 1 0 1 0 1 0 1 0 1 0 1
1 0 1 0 1 0 1 0 1 0 1 0
0 1 0 1 0 1 0 1 0 1 0 1
1 0 1 0 1 0 1 0 1 0 1 0
0 1 0 1 0 1 0 1 0 1 0 1

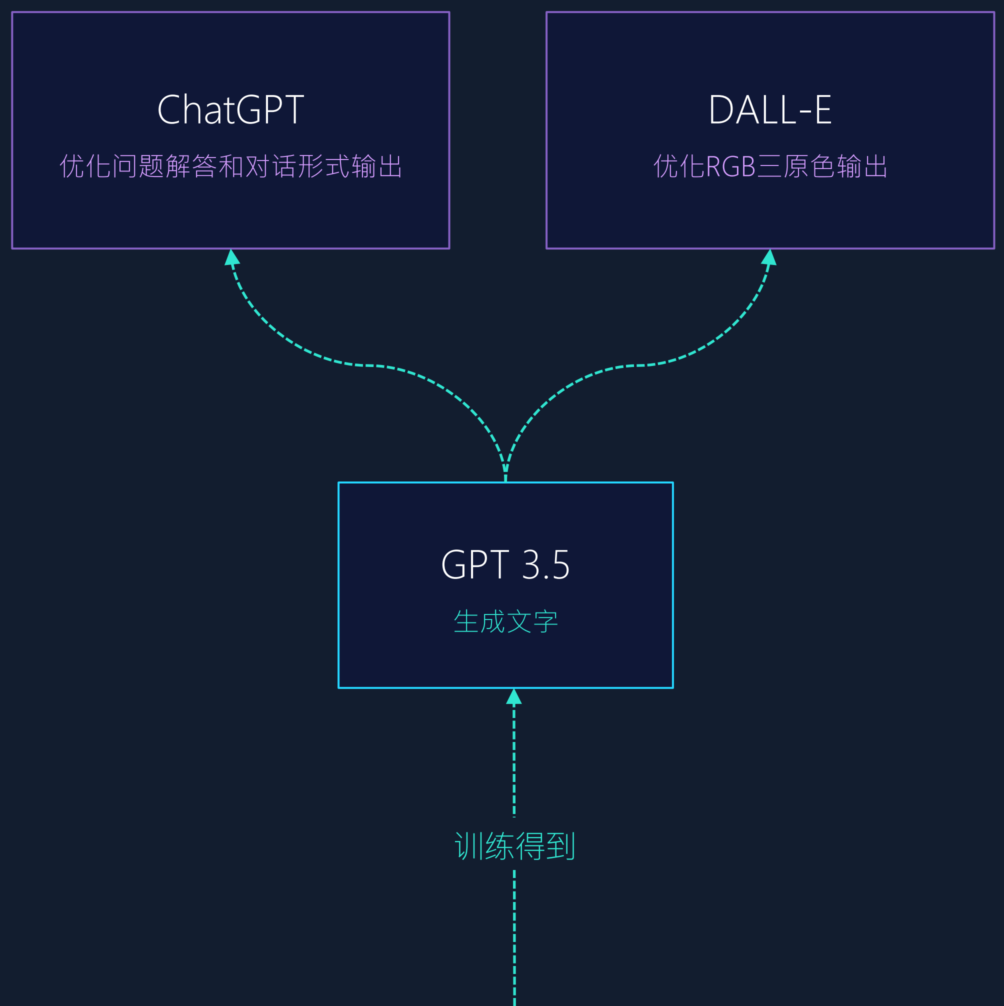
GPT模型具备一定的真实世界知识储备，基本常识和概念（世界观）

截止2021年6月

text-davinci-003

Open AI 它不是...

- Chat GPT 和 DALL-E 是以GPT3.5为基础衍生出来的应用
- 目的用来展示GPT可以做到什么程度
- 它不是模型服务，不能直接售卖
- 我们会有对应服务推出，暂定2月28日名为Enterprise ChatGPT



OpenAI能做什么？

GPT X	Codex	Embedding
生成文字	生成代码	生成向量

1.生成文章 (Completion)

a. Conversational AI （客服Bot）
b. 疑难解答 （QA） *
c. 翻译
d. 根据语言理解生成文字
e. 因果关系推理
f. 语句/语法/词汇改善

2. 归纳/总结 (Summarization)

3. 分类 (Classification)

a. 情緒分析 (Sentiment Analysis)
b. 打标签 / 分类

1.文字转程序

a. Auto testing
b. Debugging
c. Coding Quality Check
d. 代写代码*
e. 代码生成文档* (Reverse coding)

1.文字转矢量

a. 语义匹配* (Vectorized Semantic Search)
b. 疑难解答 (QA) *
c. 建立更强大的知识库 (KB)

* 表示可能需要多个能力组合实现

和Cognitive Service 怎样区别？

泛应用领域 (原子能力)

分类 | 生成文字 | 对话类 | 语言分析 | 归纳/总结

NLP / 语言类

翻译 | 视觉类 | 语音类

解决方案类 (Pre-built)

医学文档解析 | 视频分析/打标签 | 搜索

Pre-built OCR方案

- 发票/小票/收据识别
- 表单解析&识别
- 名片识别
- 名片识别

OpenAI

- 底层结构差异导致OpenAI各方面碾压传统DNN
- GPT本身丰富的语言理解能力和知识储备可实现零学习或Few-shot learning (举一反三)
- 未来Cognitive Services的底层会被GPT或其它Transfer架构取代 (比如 Florence)

认知服务

- OpenAI不具备的能力
- 商用成熟，包括定制化在内的业务支援丰富

认知服务

- 客戶寻求的解决方案级别的应用 (开箱即用)

不是对抗，而是联合！

- Cog Service本身的多元性 + 配合GPT强大的语言能力 = 提升现有能力或打造全新体验

Form Recognizer OCR发票，提取结构数据 → 整理出发票的性质，每个条目的类别及合规检查 → 更智能的业务自动化处理

STT实时会议转写 → 自动归纳会议主题，纪要，Action Points，重要时间线 +more → 更进阶的会议智能助手

游戏视频解析/打标 → 根据标签实施理解事件，生成情景描述 → 生产自然语音搭配虚拟人做直播介绍 → 全新的用户体验

语音标识一段需求（STT） → 理解需求，生成对应的代码 → 部署到Power Automate上！ → 把RPA玩到新高度！

那问题来了， Azure OpenAI是干嘛的？

打一个比喻...

钞能力

Microsoft

发行商

OpenAI

拍电影的团队

独家发行权

宣传
营销
公关
商流
法务
投产(aka 赚钱)

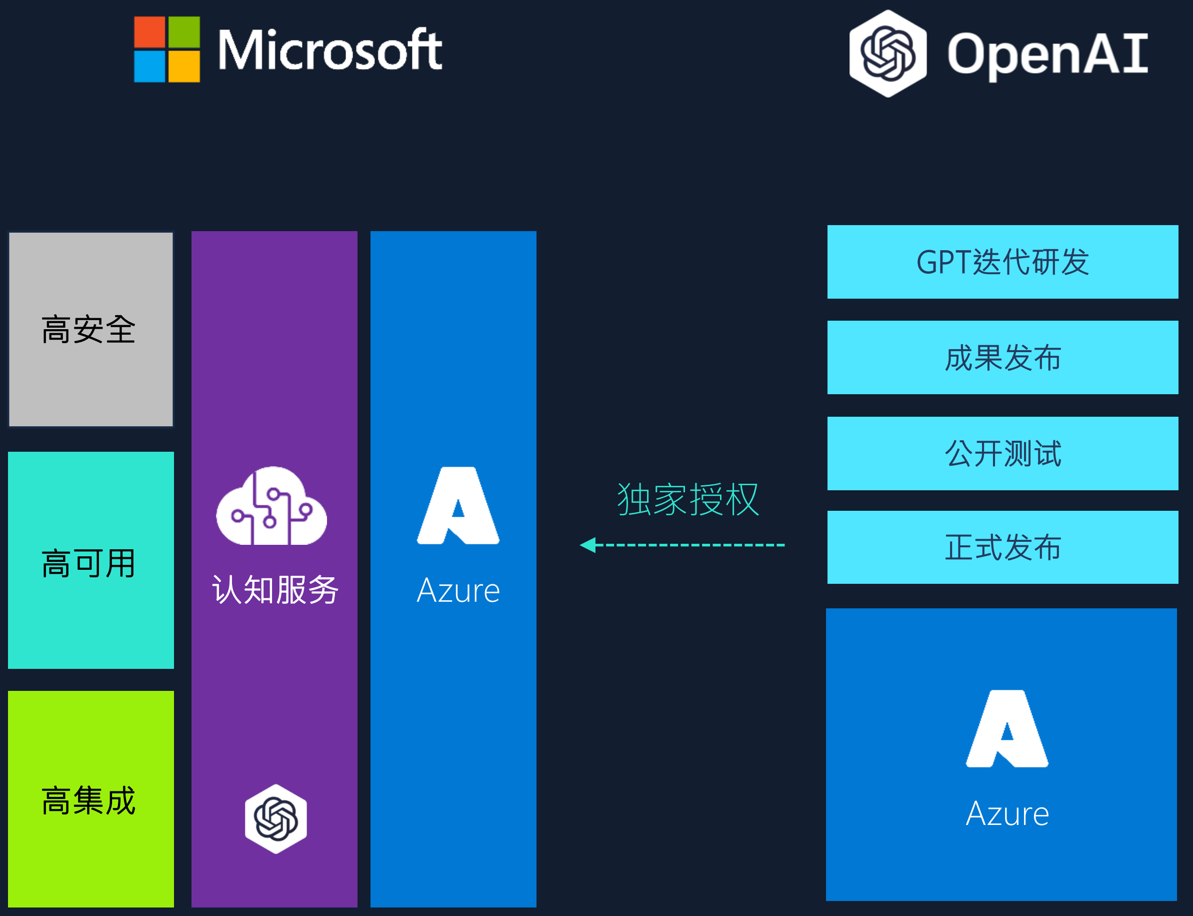
写剧本
雇人
拍片
导演
剪辑
后期

商务价值

- 私有网络（ vNet ）
- Private Endpoint/Link
- Custom Domain
- CMK
- RBAC/IAM 权限管理
- Azure AD集成
- 托管身份（Managed ID）
- 企业合规（SOC2， HIPAA， ISO，GDPR Data Privacy etc.）
- 企业伦理审查（RAI）
- 企业级有偿Support

- 多区域可用/容灾
- 保证SLA 99.9%
- 更高Rate Limit（TPM/QPM）
- 更高Token上限（Fine Tuning）
- Endpoint部署管理和监控
- 基于Azure的API扩充：
- Token监控
- 更好的Error & debug支持

- 集成现有Cognitive Service的服务架构
- 集成资源监控（Monitor / Log Analytics）
- 集成自动化（Automation Task， ARM部署）
- 集成Express Route
- 集成密匙管理（AKV）
- 集成数据加密（Azure Storage）



- API 文档
- Fine-tuning
- API发布/管控
- Endpoint hosting

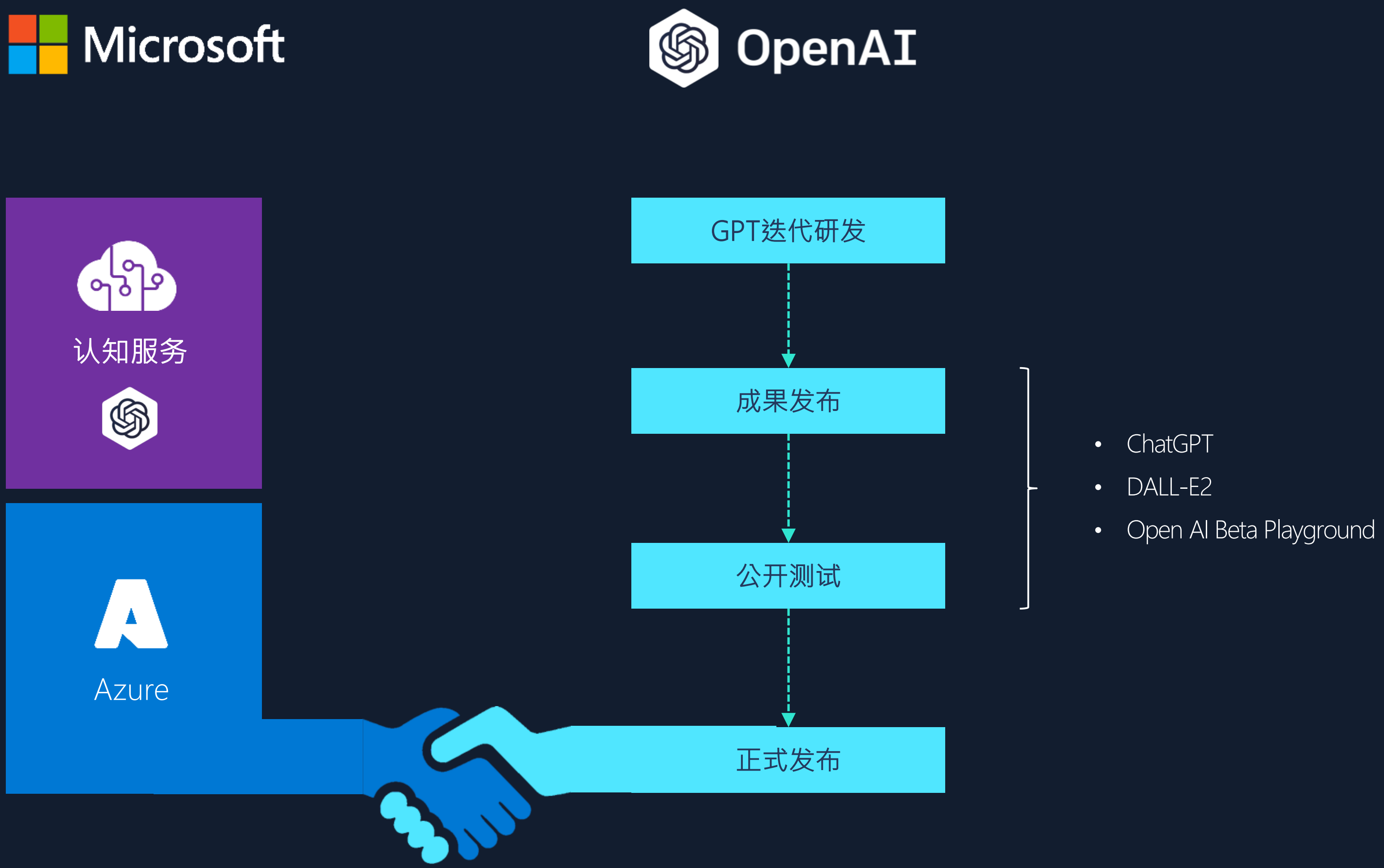
所有在 OpenAI.com 上产生的公共用量，全都是Azure的投资回报

模型服务

- AOAI PG有计划两边能力完全平等（最快3月开始）
- 未来正式模型，两者间会同步发布

但是

- 两边的gap不会完全消除
- 由于两边企业性质导致
- 客户用OpenAI作为early preview，然后再挪到Azure



解题思路❶ - 如何实现属于我的ChatGPT ？

因为前述GPT的学习特性，Customize的方法也和传统思路不太一样

以客服QA Bot为例

1. 使用Embedding模型把知识库(KM)转为矢量，并保存到Vector Database做匹配和搜索
2. 用戶对Bot输入一段自然语言
3. 对其做同样的矢量转换 (一样的Embedding模型)
4. 將输入的矢量和Vector Database做近似度计算（=语义匹配），抽取出关联度最高的条目/文章。這部分建议搭配Open Source (比如 Redis, FAISS)
5. 返回文章給GPT 作为prompt engineering的一部分，得到自然语言形式的回答
6. 返回结果給用戶

解题思路❷ - Embedding还能怎么玩？

Embedding不支持Fine-Tuning，但是它的应用最为广泛，手法也需要结合传统Machine Learning。 思路如下：

以舆情分析为例

1. 使用Embedding模型把评论数据转为矢量 ，并保存到Vector Database
2. 对每一个Review人工标注（Positive， Negative等）
3. 使用传统ML分类算法学习标注
4. 对新的评论，同样用Embedding转换为矢量
5. 用3学习到的模型进行预测，得到标签。

同样适用于其他类型的Machine Learning手法

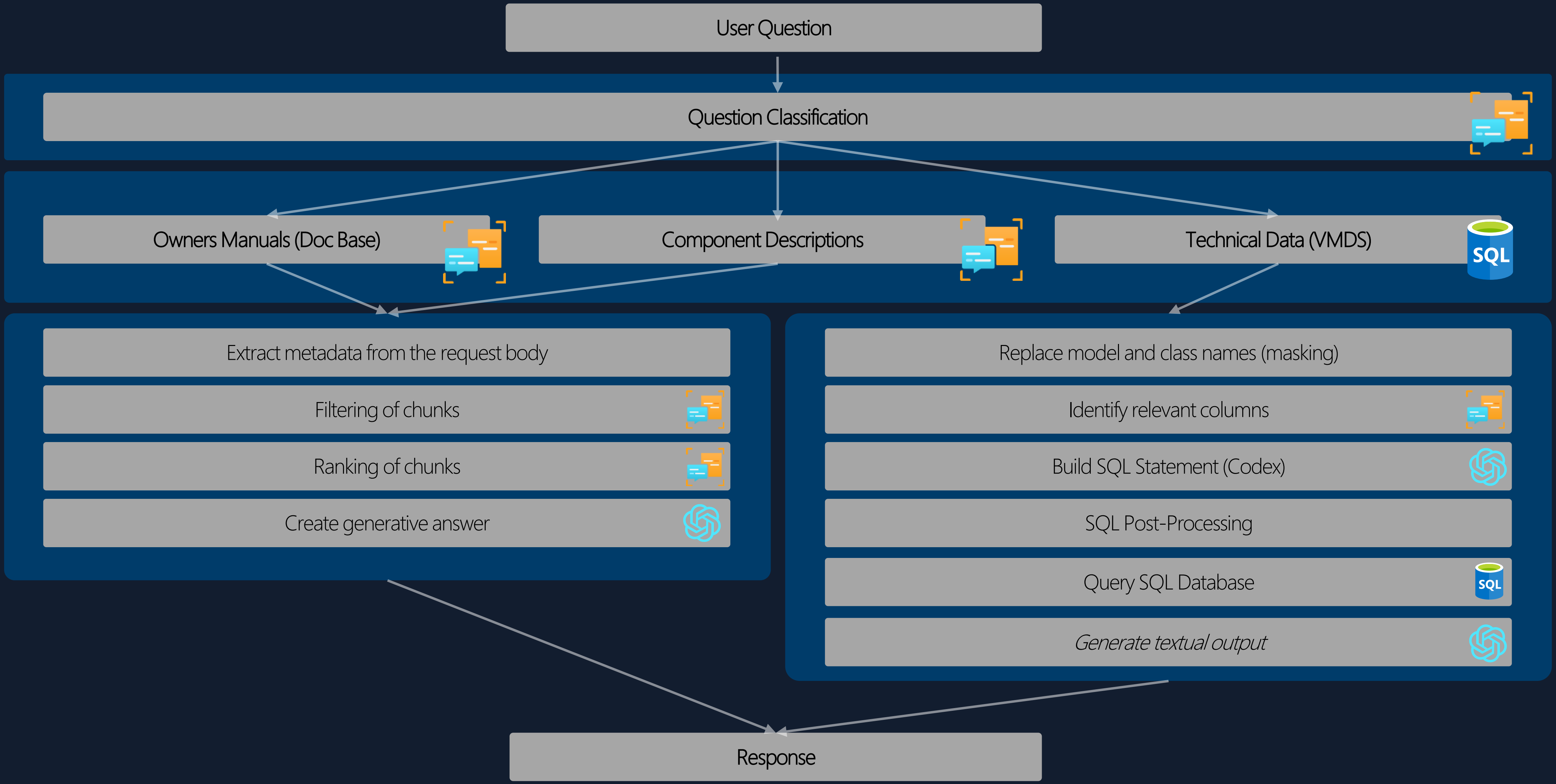
解题思路❸ - 怎么应用到你的领域？(Data Grounding)

OpenAI的解决方案从来都不应是单独的，它在实际应用中需要结合多个其它服务构成完整方案：

常见的经典构成

1. 分类 + (Codex) + GPT （根据不同场景，做不同处理）
2. Azur Bot + Embedding + Semantic Similarity + GPT （经典KB/QA, Virtual Agent）
3. LUIS/CLU 提取intent 转对应prompt + GPT (意图分类，限制不相干内容扩散)
4. GPT提取NER + Search Expansion + GPT (KG search的加强版应用)
5. Meta-prompt engineering 定型文做整体指引 (例如：总结对话，抽取大意, Call Center 合规检查)
6. 推荐场景新体验 (Search + 客户资料 作为 context + GPT)

某汽车客户QA GPT应用流程

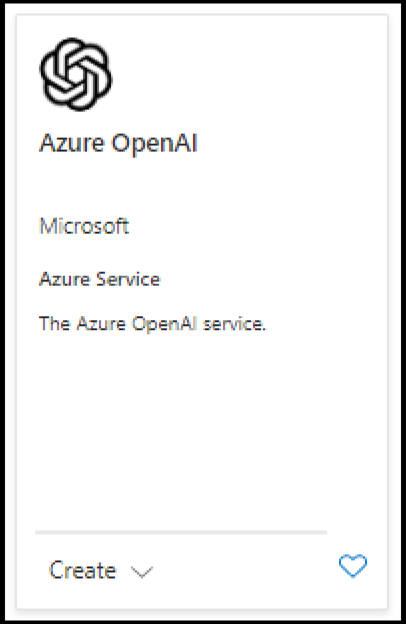


实战经验

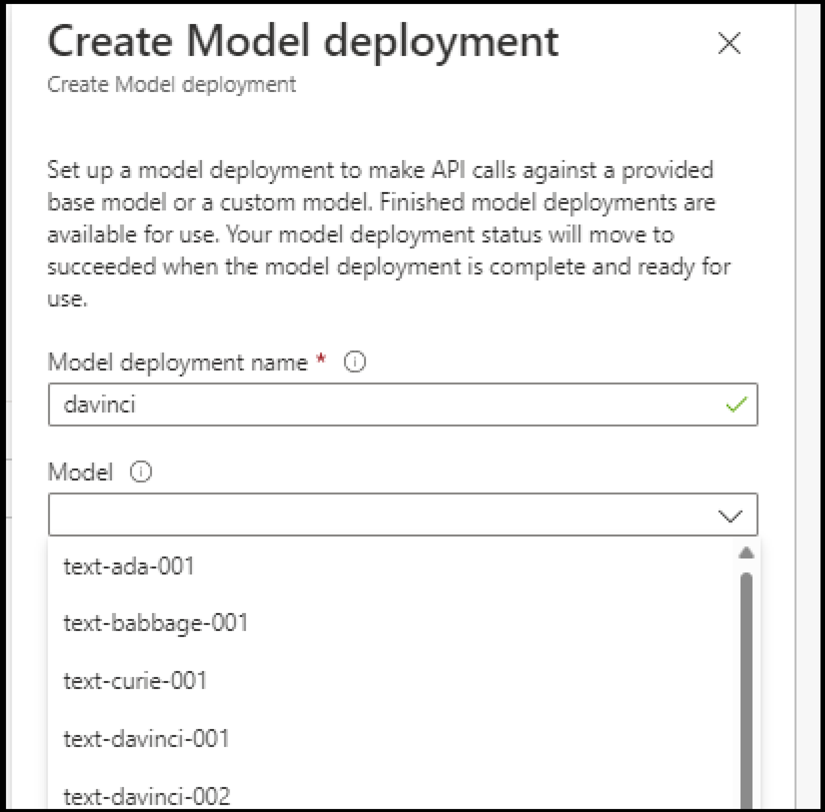
Best Practices

Azure OpenAI 测试体验： Azure Open AI Studio

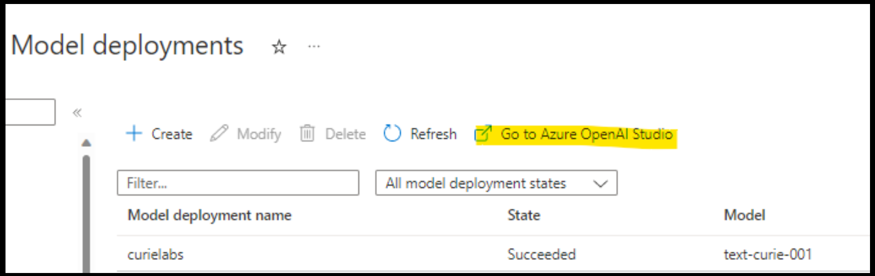
Azure Marketplace & create AOAI resource



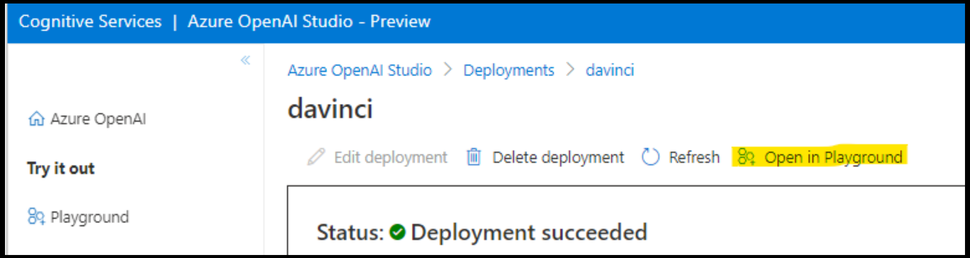
Create your base model



Azure Open AI Studio

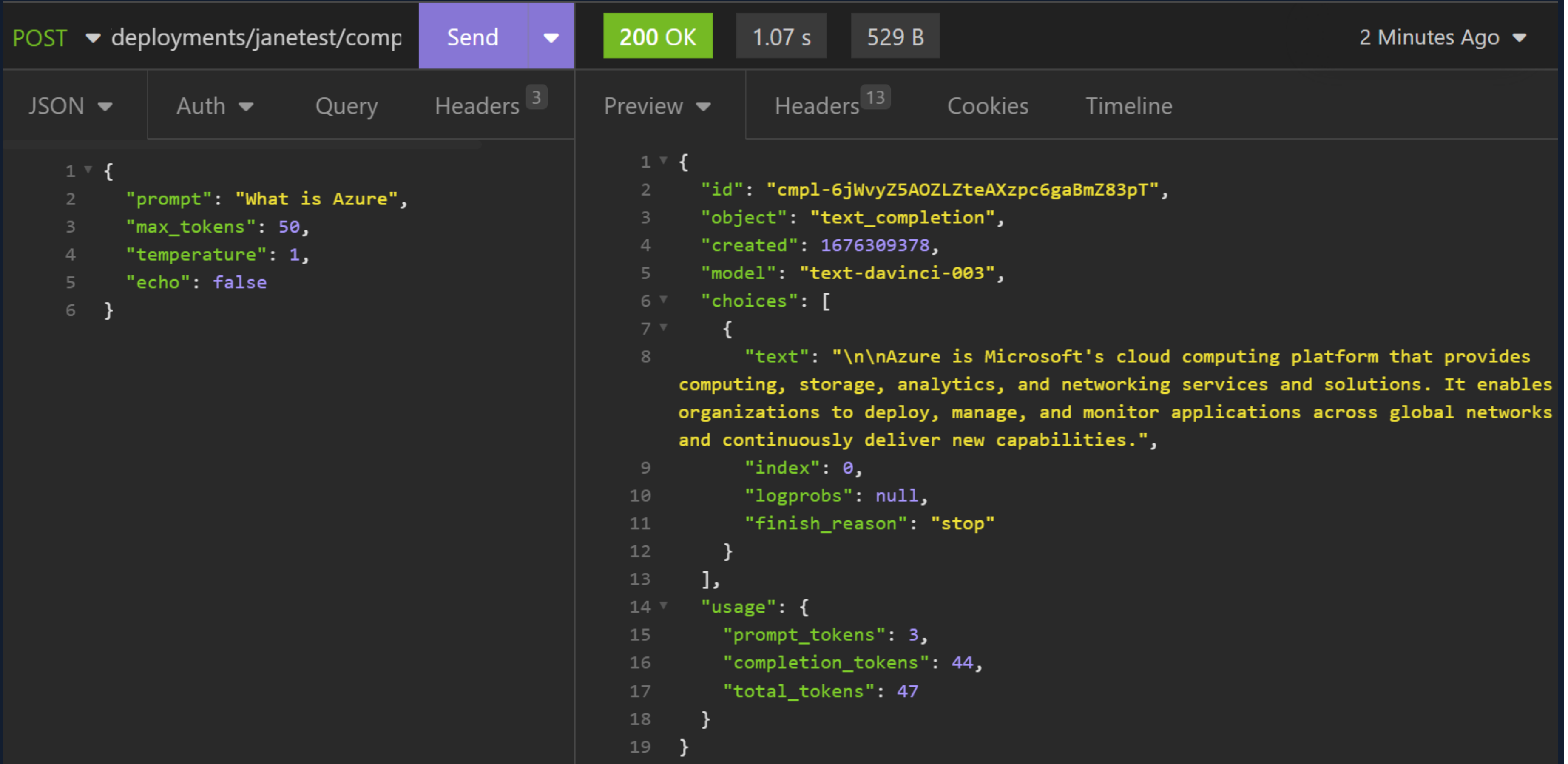


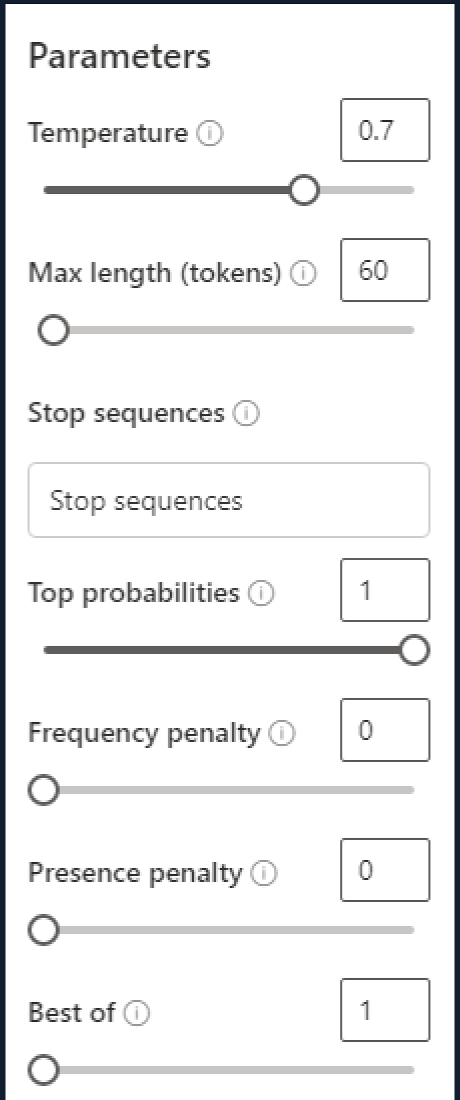
Open playground in your deployments & model to check this out



重点理解这些最常用的参数

- Parameters 参数
  - Temperature – Randomness / Creativity slider
  - Max tokens – After Max Tokens are reached, stop the completion.
  - Freq penalty – Don’ t repeat words in your completions
  - Presence penalty – Don’ t use language that has already appeared in the prompt



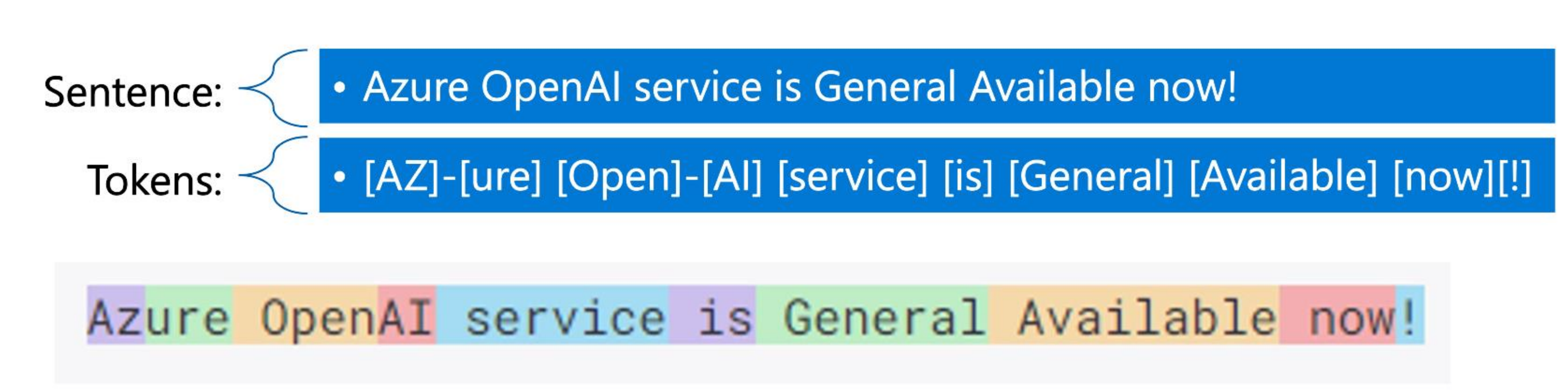


* 我们的Completion调用计价方式和OpenAI一样

如何评估Completion调用费用

- 英文750~800 words = 1000 token （1.33倍率）, input+output 两侧token都要计费
- 中文以及其它全角字符token倍率差异比较大，难在纸面上预测。 越是生僻复杂结构的字，token消耗越多。
- 最好方法是测试一些接近真实用例的情景，衡量token消耗情况。

估算方式*：OpenAI API (官方) or Token estimator (gpttools.com)



In-context 调用可能可以满足部分定制化需求

Zero-shot - Predicting with no sample provided

Zero-shot
The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
Translate English to French:     <------ task description
Cheese =>                        <------ prompt
```

One-shot - Predicting with one sample provided

One-shot
In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
Translate English to French:     <------ task description
Sea otter => loutre de mer       <------ example
Cheese =>                        <------ prompt
```

Few-shot – Predicting with a few samples provided

Few-shot
In addition to the task description, the model sees a few examples of the task. No gradient updates are performed

```
Translate English to French:     <------ task description
Sea otter => loutre de mer       <------ examples
Peppermint => menthe poivre      <------
Plush giraffe => girafe peluche  <------
Cheese =>                        <------ prompt
```

正确理解Fine Tuning

如果有客户抱怨怎么我越Tune越差？？

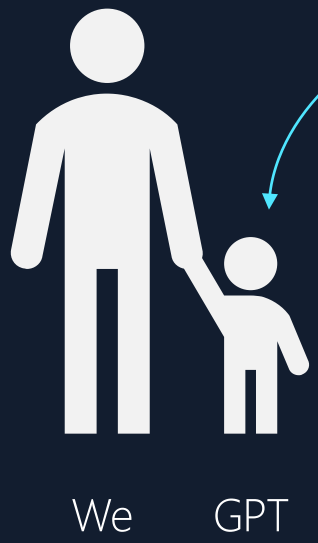
Here's why

正确理解Fine Tuning

所以我们作为大人，也要用榜样/例子來引导孩子模仿，而非直接教他对或错，因为这样会破坏它的现有认知 (Supervised Learning)

Fine Tuning就是用例子來引导GPT的过程

而怎么设计Prompt是个很深的学问。 微软可以帮助客户设计Prompt Design



GPT像个小孩, 它阅遍了全世界的书籍

它是通过主观例子来学习世界。例子没有对错 ，只有相对好坏 (Reinforcement Learning)

Fine Tuning, not “Fine” Tuning

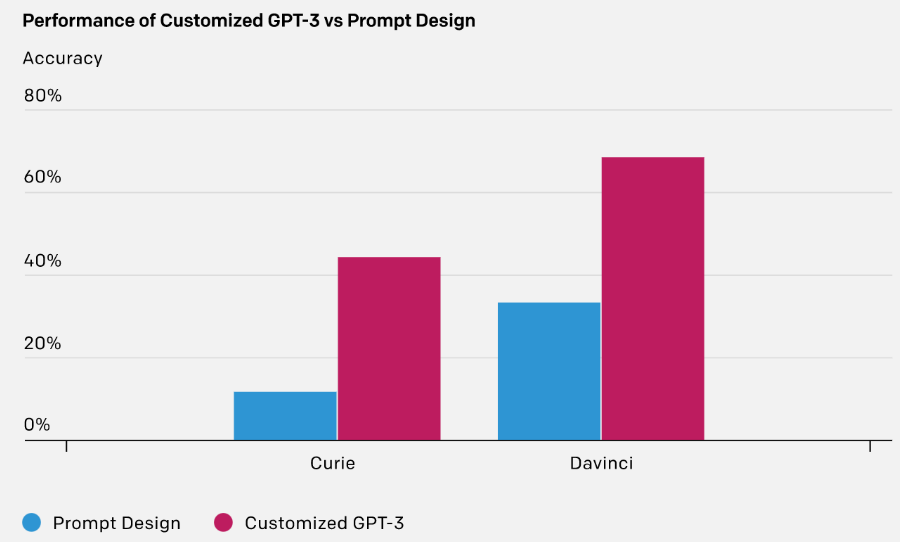
避免撞墙的小技巧

- 不要过度重复利用同一个数据（技术语言：正确设置参数)
- 尽量接近真实用户用例和习惯（技术语言：保持学习和验证数据格式高度一致)
- 不同的case要有对应的Prompt Design，不要模板化
- Meta prompt engineering 明确指示弱输出范例，有效避免 “不懂装懂”
- 示例要足够明确性和具体性。配合逻辑解释会帮助获得更好Fine Tune表现，尤其数据量不大时。
- Fine Tune可以循序渐进，避免一次喂太多数据。 从几百开始，逐渐加大数据量。

更多具体Best Practices可以联系 GBB / Specialized CSA 以及PG来协助客户 ！

什么时候用Fine Tuning ？

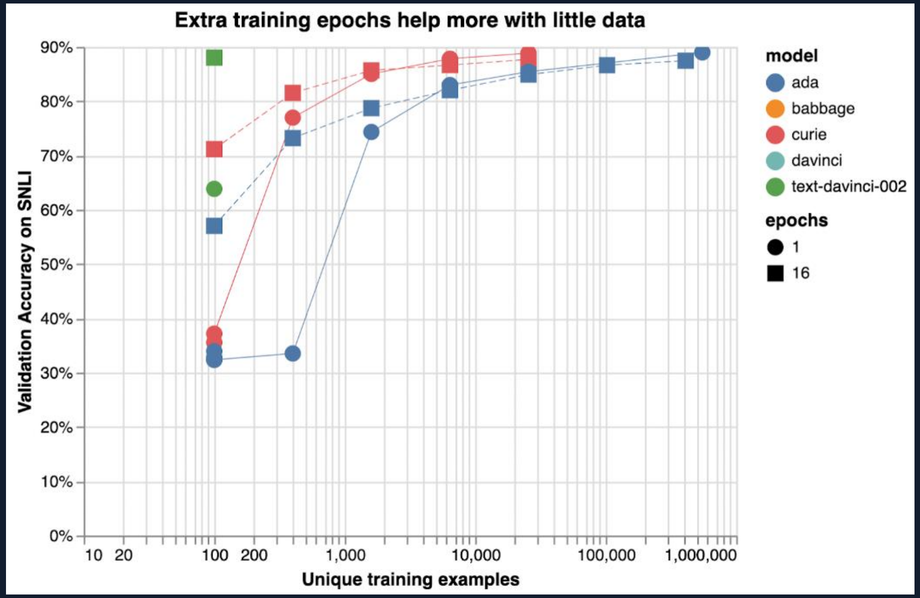
- Fine-tuning 主要是用来改善长期输出形式。相对的Prompt Engineering是短期/一次性改变输出的方法。
- Fine-tuning的长期表现会随着样本数量的翻倍呈现类线性特性的增长。
- 建议首先做prompt engineering来验证提升潜力。 然后转为Fine-tuning 做长线改善以帮助节省token，降低TCO。
- Fine-tuning可以用来限定范围，提高domain相关度
  - 根据某一类特定产品标签生成描述
  - 根据某一篇博文内容风格写推广广告
  - 企业级（知识库）客服Bot

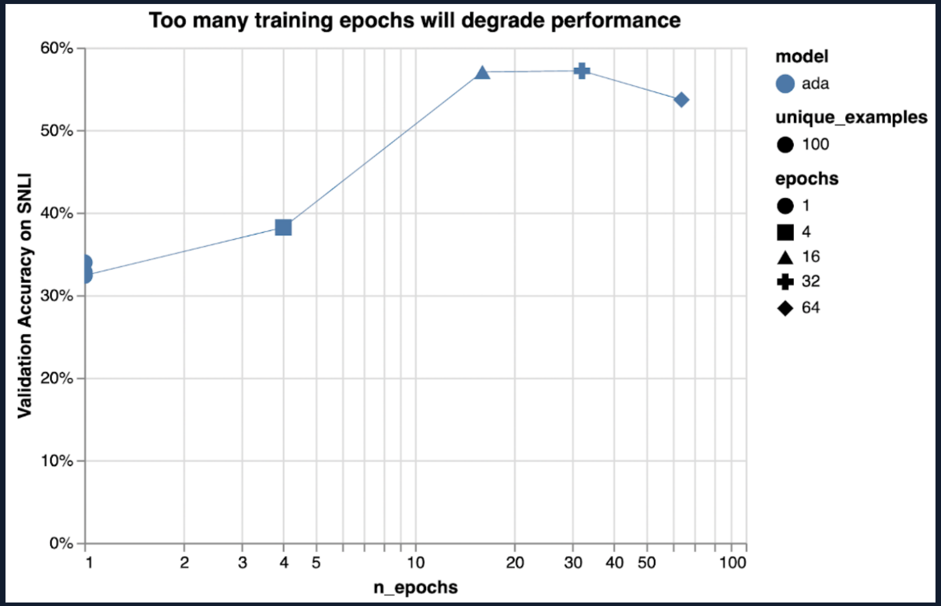


Two sizes of GPT-3 models, Curie and Davinci, were fine-tuned on 8,000 examples from one of our most challenging research datasets, Grade School Math problems. We compare the models' ability to solve problems when 10 completions are created.

如何正确选择Fine Tune超参数！

同一组数据（SNLI）下不同epoch参数的Fine Tune表现：

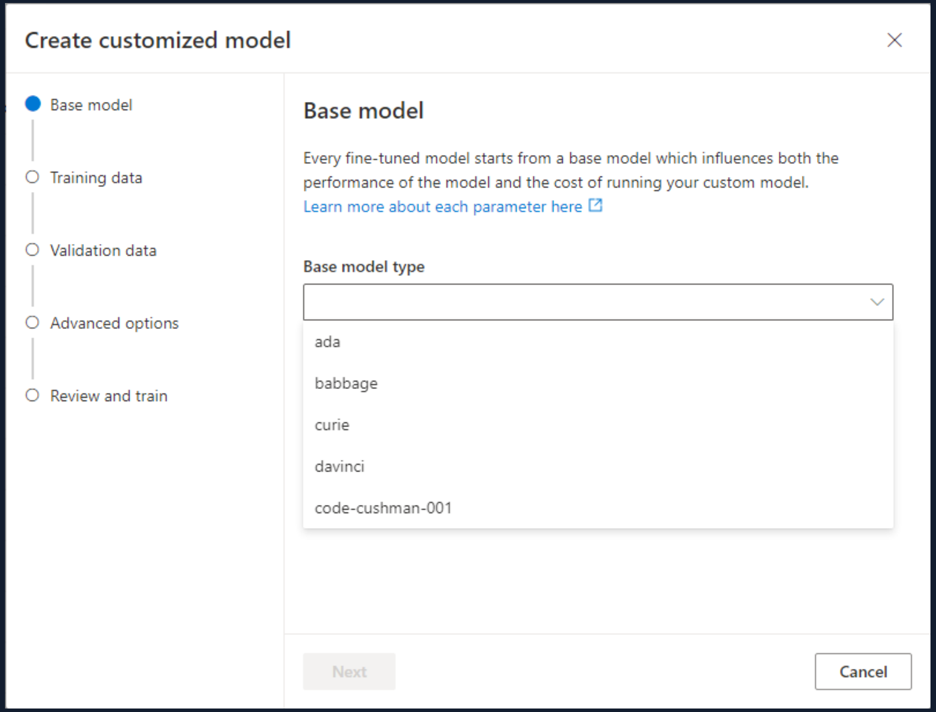


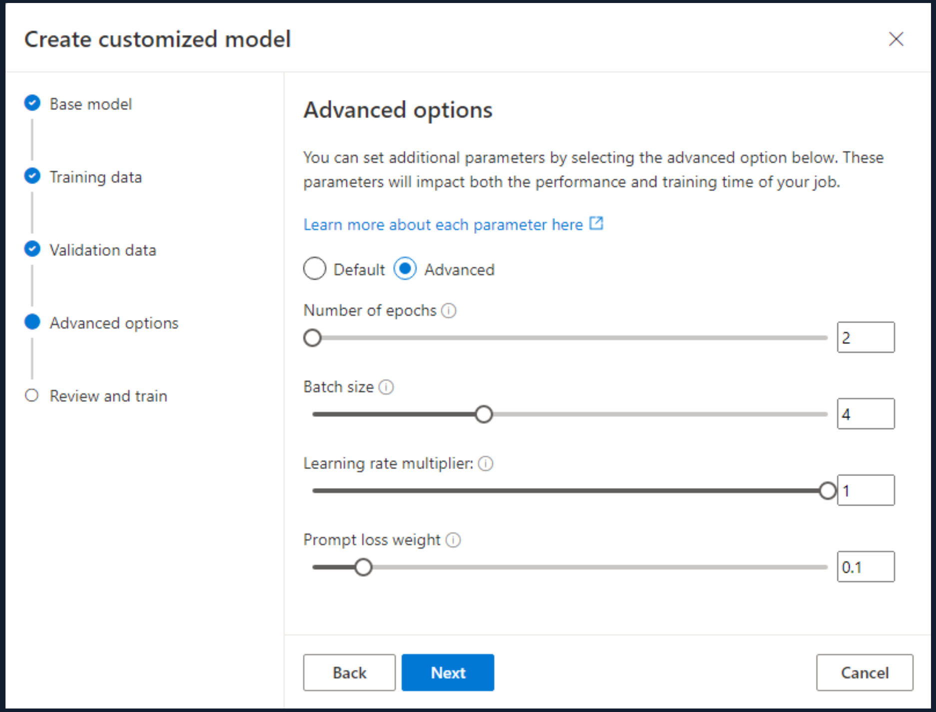


General建议

- 如果样本量不大或质量不保证（如未经校准的数据），控制epoch 8~16 有助于有效提升同精度表现
- 如果样本量足够，建议控制epoch 1~4，可以有效避免过拟合
- 模型越复杂epoch可以越小，实现更低的训练成本

如何正确选择Fine Tune超参数！

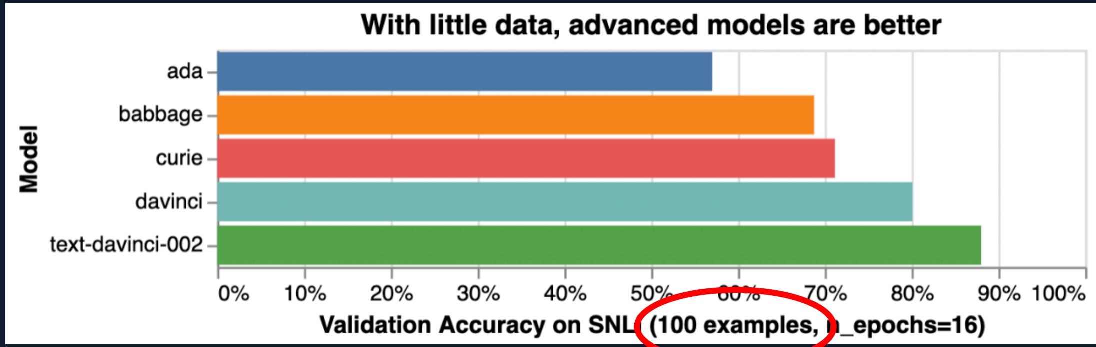


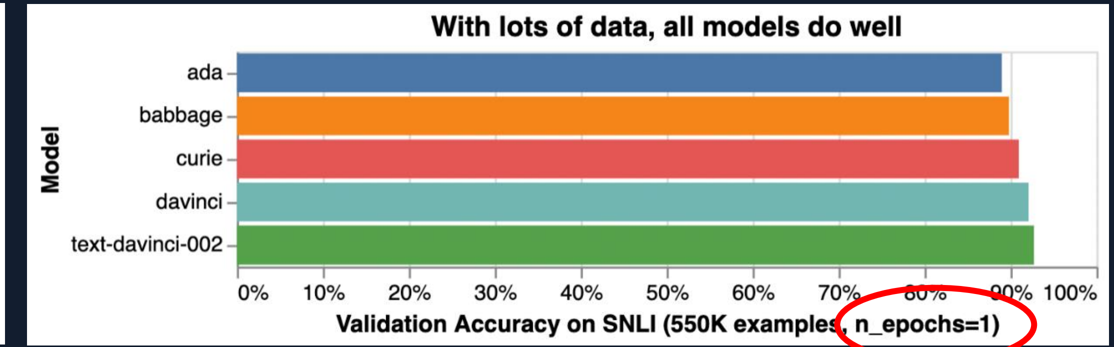


Parameter name	Description
Number of epochs	The number of epochs to train the model for. An epoch refers to one full cycle through the training dataset.
Batch size	The batch size to use for training. The batch size is the number of training examples used to train a single forward and backward pass.
Learning rate multiplier	The learning rate multiplier to use for training. The fine-tuning learning rate is the original learning rate used for pre-training, multiplied by this value.
Prompt loss weight	The weight to use for loss on the prompt tokens. This value controls how much the model tries to learn to generate the prompt (as compared to the completion, which always has a weight of 1.0.) Increasing this value can add a stabilizing effect to training when completions are short.

模型并非越新，越大就越好！

同一组数据（SNLI）下不同模型Fine-tuning的结果





Model	Illustrative accuracy*	Training cost (Open AI)	Inference cost
ada	89%	$0.0004 / 1K tokens (~3,000 pages per dollar)	$0.0016 / 1K tokens (~800 pages per dollar) 3
babbage	90%	$0.0006 / 1K tokens (~2,000 pages per dollar)	$0.0024 / 1K tokens (~500 pages per dollar)
curie	91%	$0.003 / 1K tokens (~400 pages per dollar)	$0.012 / 1K tokens (~100 pages per dollar)
davinci	92%	$0.03 / 1K tokens (~40 pages per dollar)	$0.12 / 1K tokens (~10 pages per dollar)

模型并非越新，越大就越好！

假设有两种选择

Which one do you choose ?

D 模型

- 10倍价格
- 3倍延迟
- +1%精度

VS

A 模型

- 1/10价格
- 速度快2倍
- -5%精度

- 模型精度是有代价的。 选哪个更合适是个Business Question而非Technical Question
- 每个use case都有一个独特的成本收效平衡点

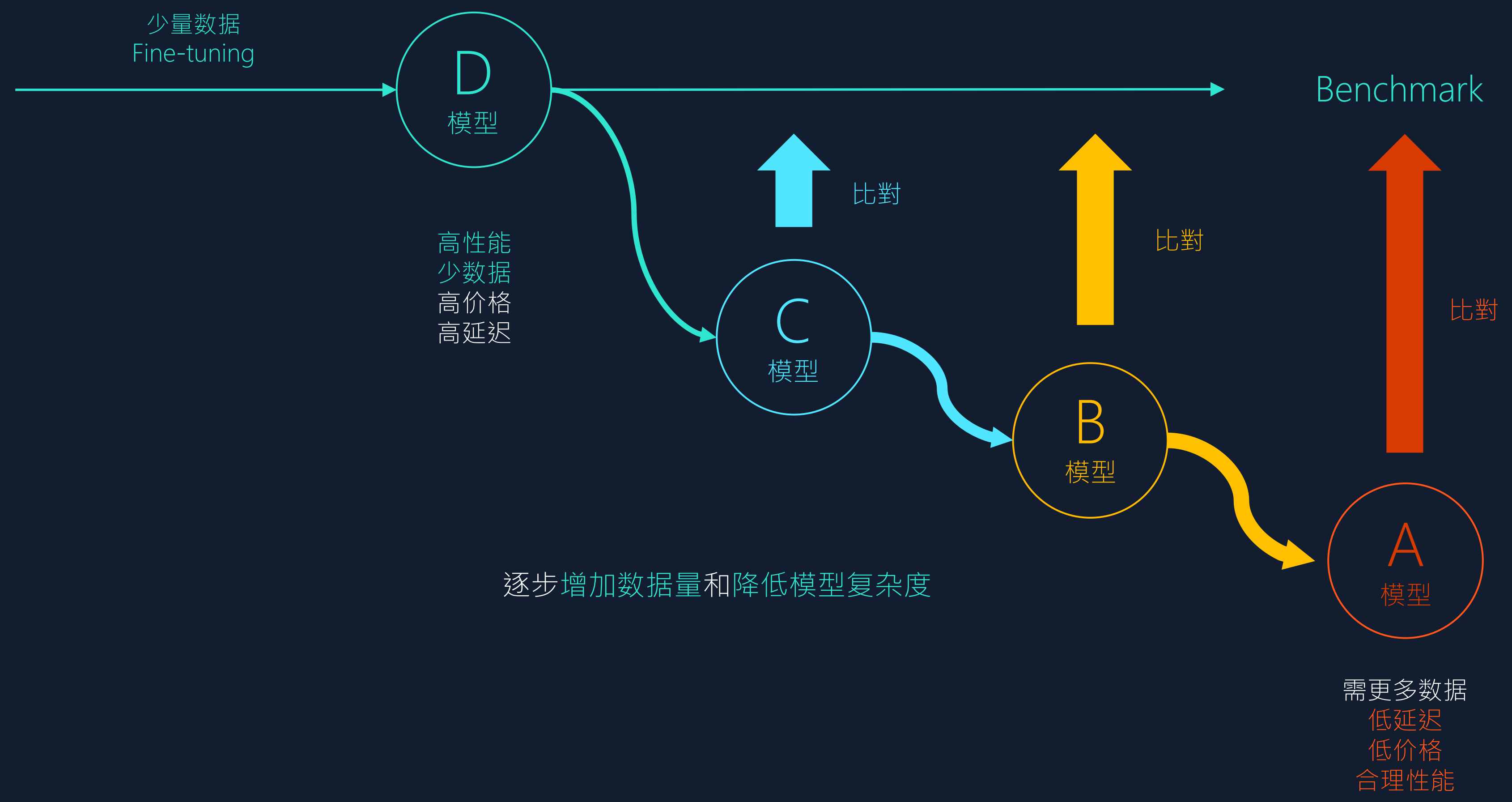
模型并非越新, 越大就越好！

General建议

- 有大量优质（指人工审核/达标的）数据，建议低复杂度模型，精度或许能媲美Davinci，而ROI更高。
- 如果是分类/归纳这类相对简单的任务，建议A模型
- 如果数据量不大（几百）或质量不保证（如未经校准的数据），则建议D模型

General建议

- 步骤建议先从Davinci开始建立一个benchmark，随后逐一降低模型复杂度验证Fine-Tuning能达到的结果
- 选出一个客户认为最均衡的模型



* 我们的finetune计价方式和OpenAI不一样， 无法互换

如何评估预算

- 英文750~800 words = 1000 token （1.33倍率）
- 中文以及其它全角字符token倍率差异比较大，难在纸面上预测。 越是生僻复杂结构的字，token消耗越多。
- 最好方法是测试一些接近真实用例的情景，衡量token消耗情况。

估算方式*

(月活跃用户数 (MAU) X 用户平均消耗字数（双向合计） X 1.33 Token倍率) + 每月Fine-tune时长 + 模型Hosting时长 = TCO

单价 1 (per 1,000 tokens)

单价 2

单价 3

(月活跃用户数 (MAU) X 用户平均消耗字数（双向合计） X 1.33 Token倍率) + 每月Fine-tune时长 + 模型Hosting时长

案例

- 某知名电商2C客服APP有 10万月均活跃客戶，每人每月平均10次互动，每次总计250字数 (英文)
- 模型每周训练一次20小时，每月24x7供应

以Ada为例：

(100,000 * 250 * 1.33 / 1000) * $0.0004 + (20 * 4 * $20) + (730 * $0.05) = **$1649.8** /每月

* 我们的finetune计价方式和OpenAI不一样， 无法互换

知识产权归客户所有， 同样法律责任也是！

详见Azure open AI FAQ

https://learn.microsoft.com/en-us/legal/cognitive-services/openai/data-privacy?context=%2Fazure%2Fcognitive-services%2Fopenai%2Fcontext%2Fcontext

知识产权和数据隐私保护声明

承诺不擅用客户数据，数据内部处理流程等细节

https://learn.microsoft.com/en-us/legal/cognitive-services/openai/data-privacy?context=%2Fazure%2Fcognitive-services%2Fopenai%2Fcontext%2Fcontext

GPT / OpenAI 限制

无法在离线跑 。 现在不会，未来也不见得会。

GPT不擅长数字和推理类处理。

GPT缺乏可靠性 (甚至会胡扯)。 其商用稳定性需要大量依赖精良的Fine Tuning！

只有GPT X和Codex能Fine-Tune, Embedding不支持*

* Embedding有特殊的训练方法

Use Cases

Get Inspired

Media Example

News Analyses & Article Creations

News Broadcast
(Global Warming)

Azure
Speech
Transcription

Azure
OpenAI
Text Completion

Search
SEO
Virtual Agent
Analytics
Reporting
Knowledge Mining

Content Analysis

Topic Classification

Global warming, Deforestation,
Carbon footprint

Entity Extraction

Organizations: IPCC, UNFCCC, Green Peace
Geography: Canada, USA

Key Word Extraction

Human activities, fossil fuels,
earth atmosphere

Question and Answer

What is the Intergovernmental Panel on Climate Change (IPCC)?
The IPCC is an international organization that studies climate change and the effectiveness …

Content Creation

Ideation
Productivity
Personalization
Accessibility
A/B Testing

Video summarization

The article discusses about global warming and its effects on the Earth's atmosphere, wildlife, and human communities. It states that the primary cause of global warming is ….

News article generation (or blogs, social media)

Global warming is the gradual increase in the overall temperature of the Earth's atmosphere, primarily caused …

Script Generation

Act 1: The show opens with a shot of a beautiful coastal town
Act 2: As the town struggles to cope with the crisis …

Personalized Content generation (or Advertising)

Simon, as someone passionate about global warming, you are aware of the urgent threat it poses to our planet …

Sports & Entertainment Example

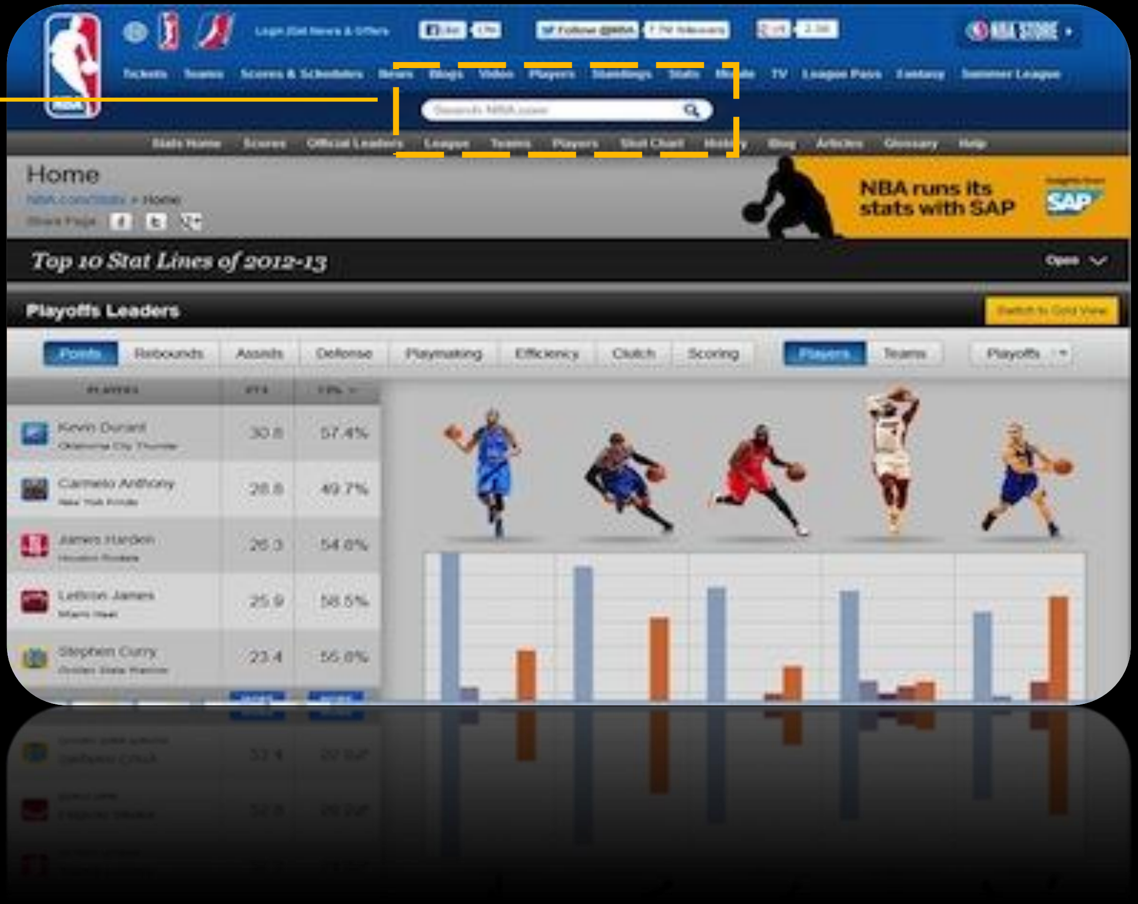
Natural language to SQL to surface stats data (Fan Engagement)

Sports Stats Database

Azure OpenAI Codex

"How many points did Lebron James score in 2021?"

```sql
SELECT SUM(points) FROM basketball_stats WHERE player = 'LeBron James' AND game_date BETWEEN '2021-01-01' AND '2021-12-31';
```

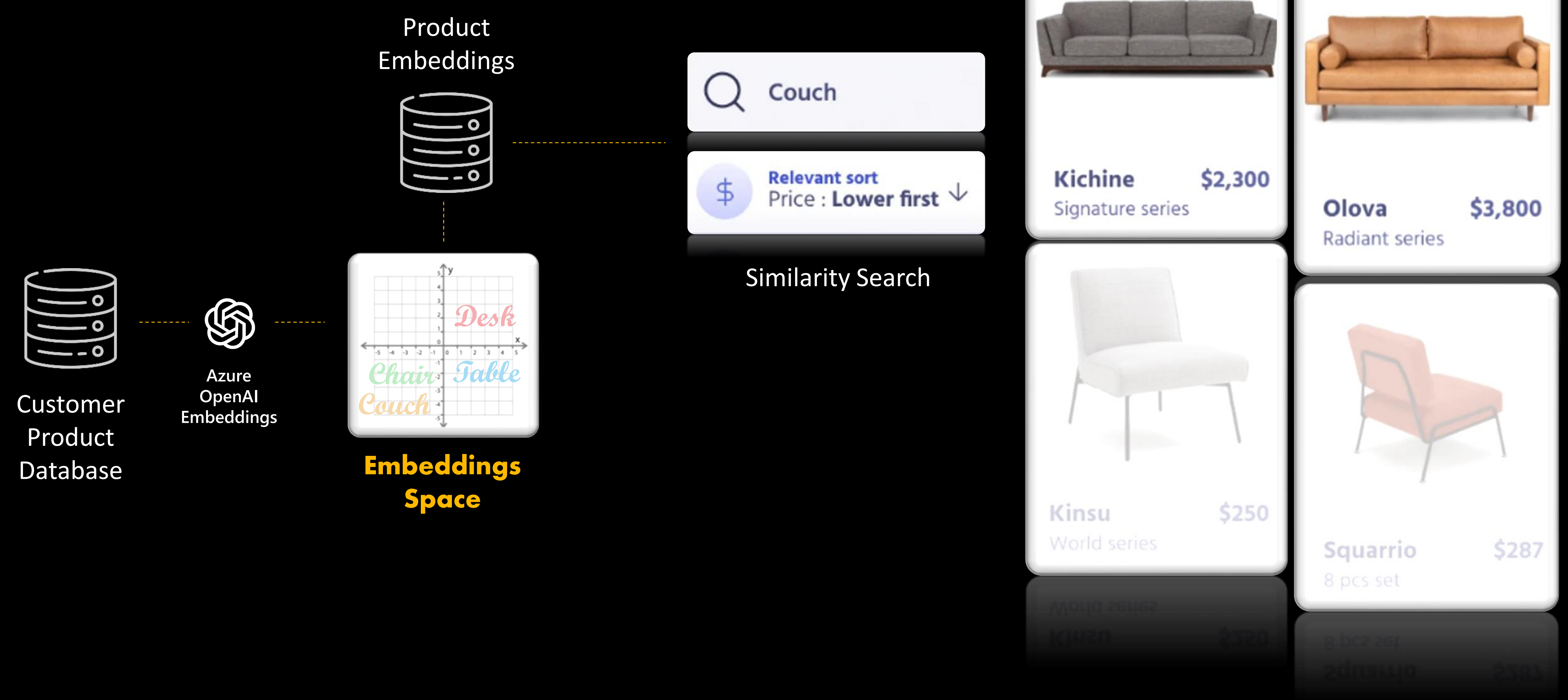


basketball_stats (player, team, points, rebounds, assists, steals, blocks, turnovers, games_played)
("LeBron James", "Los Angeles Lakers", 25, 7, 10, 1, 0, 3, 82), ("Kevin Durant", "Brooklyn Nets", 30, 7, 4, 1, 2, 3, 72)

basketball_games (home_team, away_team, home_score, away_score, game_date)
("Los Angeles Lakers", "Brooklyn Nets", 110, 105, '2021-12-01'),
("Milwaukee Bucks", "Golden State Warriors", 120, 115, '2021-12-02')

Retail Example

Similarity Search



Marketing Example

Digital Creative Assistant – Dalle-2

Synthetic Brand Ambassador (GDPR-safe)

Input

Output

Generate a white female

Generate

Generate images, concepts and ideas

Input

Output

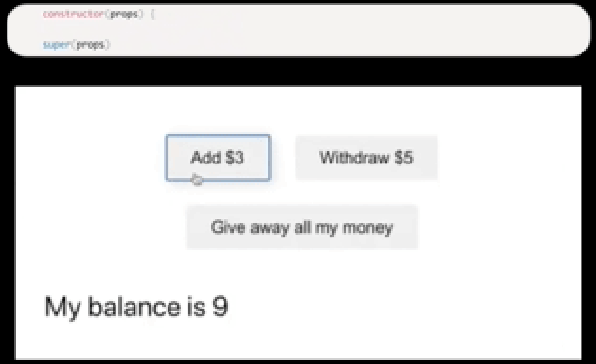
a green c

No code Web and App Development

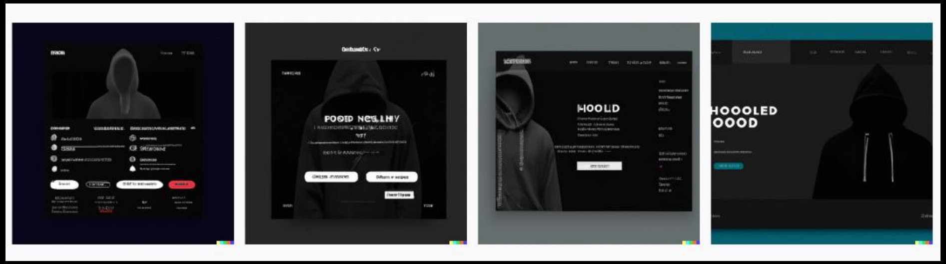
Input

Output

a button that says "add $3" and a
button that says "withdraw $5",
and a button that says
"Give away all my money",
then show me my balance

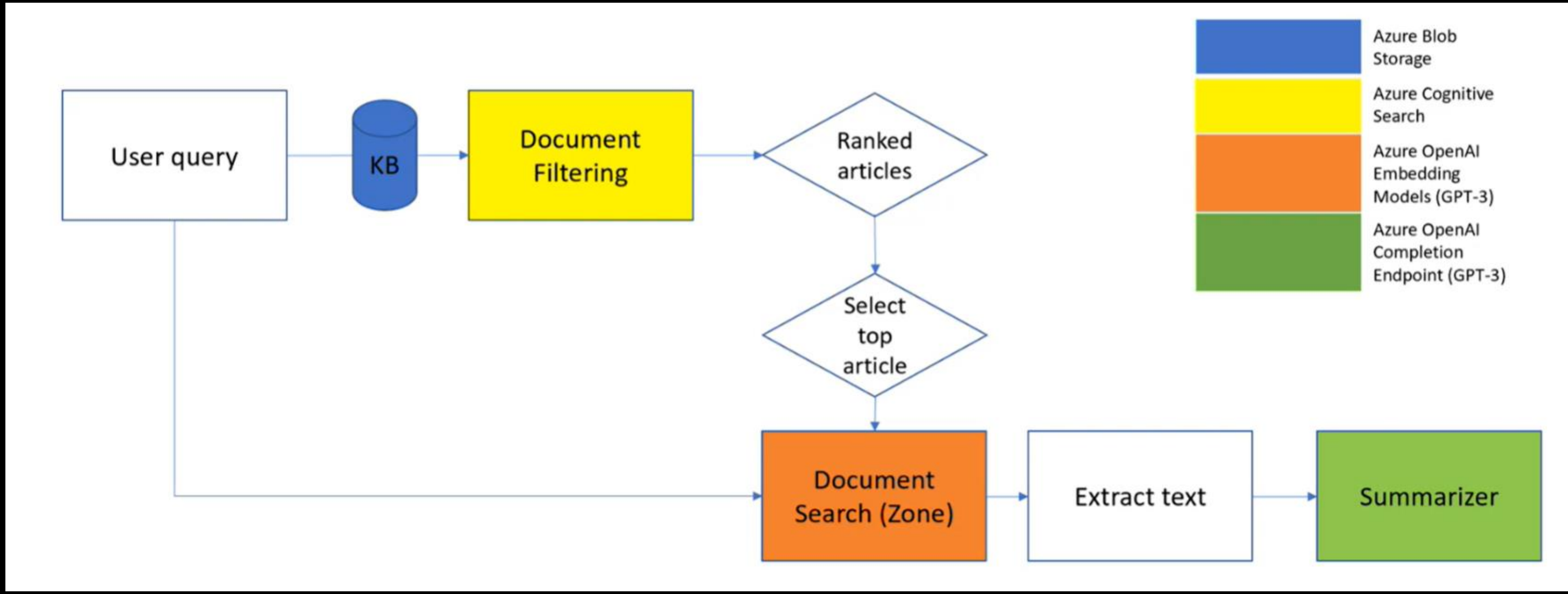


Produce rough Layouts



In-house Analyst Example

Query-based Summarization

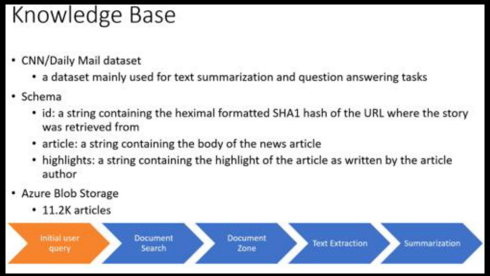


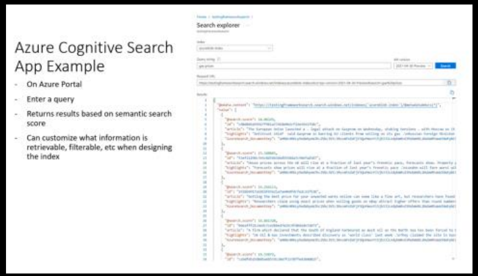
Use case overview:

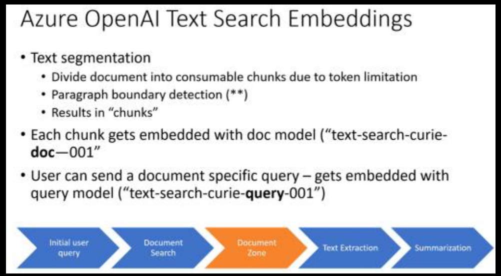
- Document search
- Document Zone search
- Text summarization

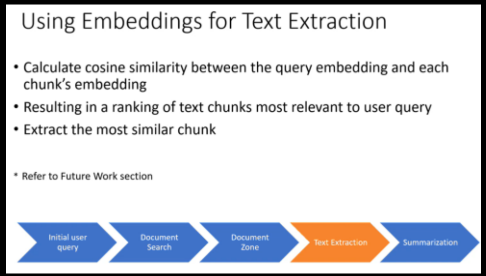
Advantages of query-based summarization:

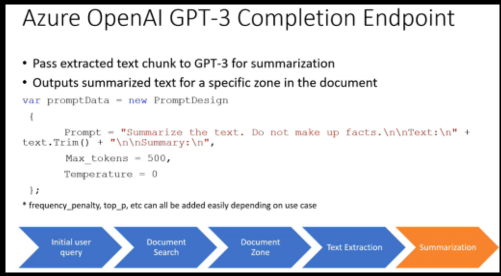
1. Shortens reading time
2. Improves the effectiveness of searching for information
3. Removes bias from human summarization techniques
4. Increases bandwidth for humans to focus on more in-depth analysis











Gaming

Minecraft Mod Creation

Mount&Blade II BANNERLORD

情景

- 根据玩家交流，动态生成NPC互动
- 互动内容会参考NPC所处位置，职业，状态，游戏进程及当时世界观。
- 非常优秀的Fine-Tuning案例

机会

结合微软TTS，Custom Neural Voice和Cross Lingual Adaptation实现更强大的NPC表现力及游玩体验

AI DUNGEON

情景

- 文字类冒险RPG
- 玩家自行编辑剧情，游戏自动生成图文并茂的开放式故事线，供玩家互动

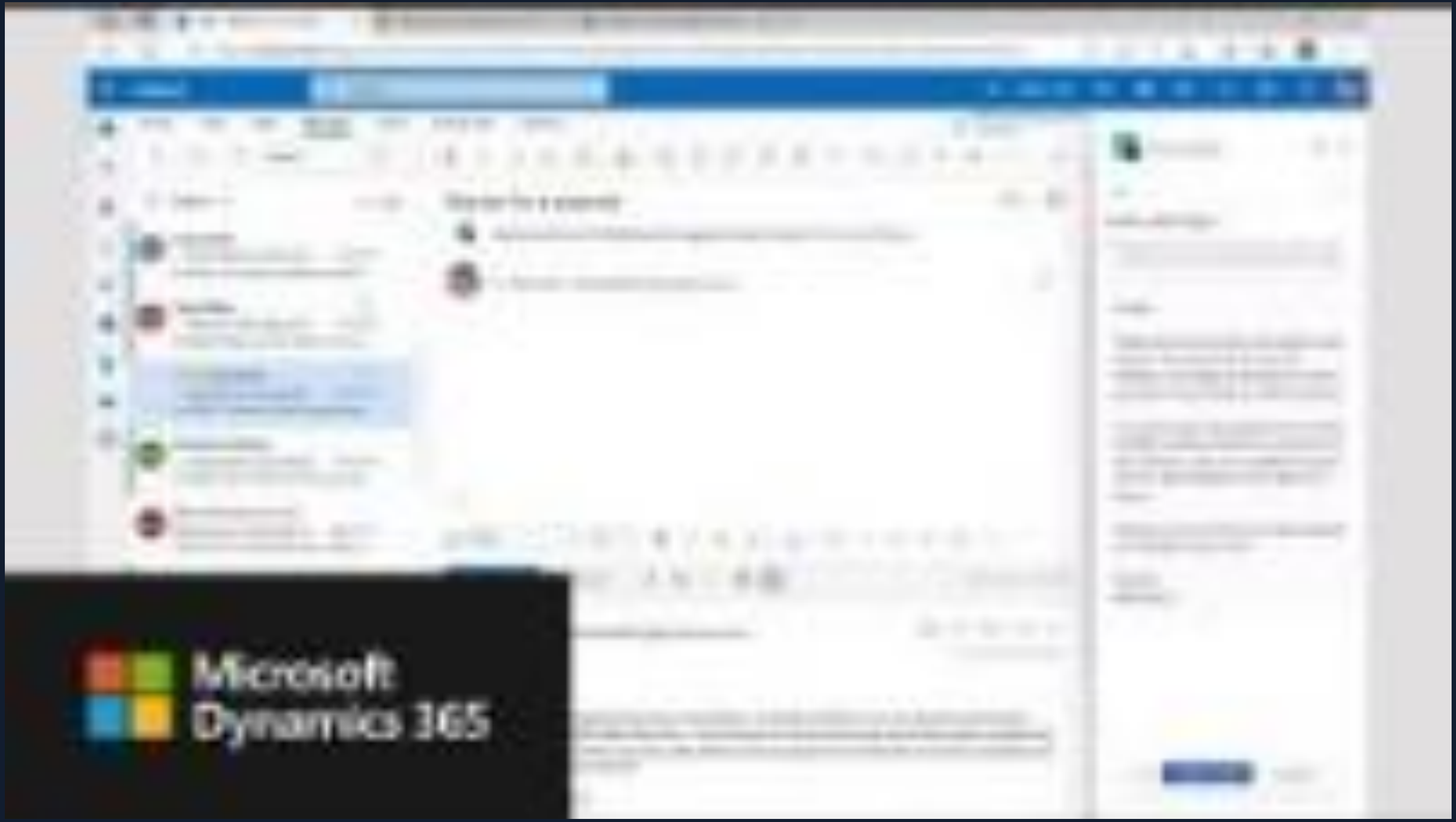
机会

结合DALL-E2，Unity Engine自动生成沉浸式游玩体验。 配合定制化的Content Moderation加固生成内容的安全性/合规性

Viva Sales

情景

根据上下文邮件信息，结合CRM最新资料自动撰写邮件

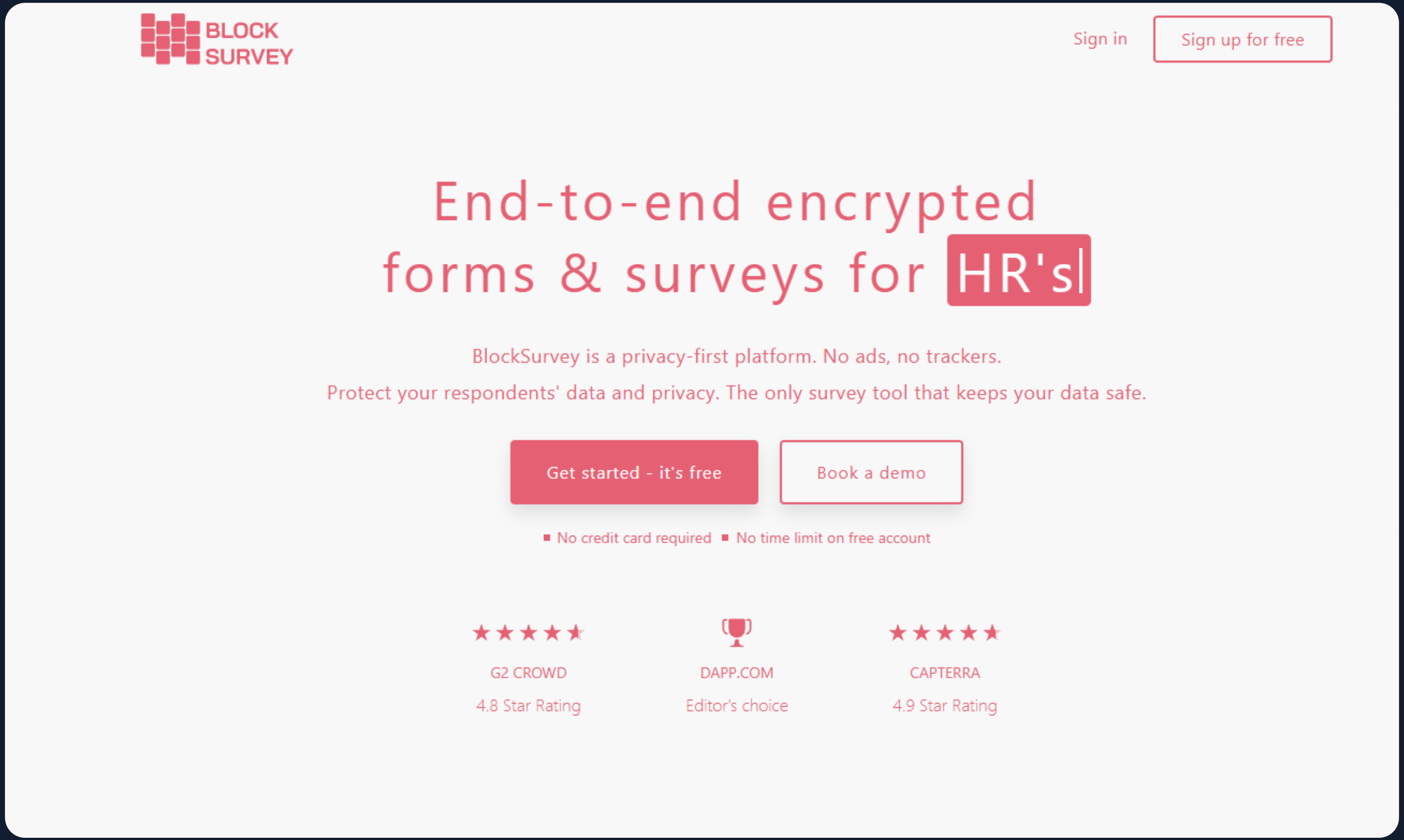


Block Survey

情景

动态生成问卷调查，总结并提示改善点

- 根据提示，自动编写调查问卷（生成问题）
- 解析用户反馈（文字类）
- 总结
- 提出汇总改善点

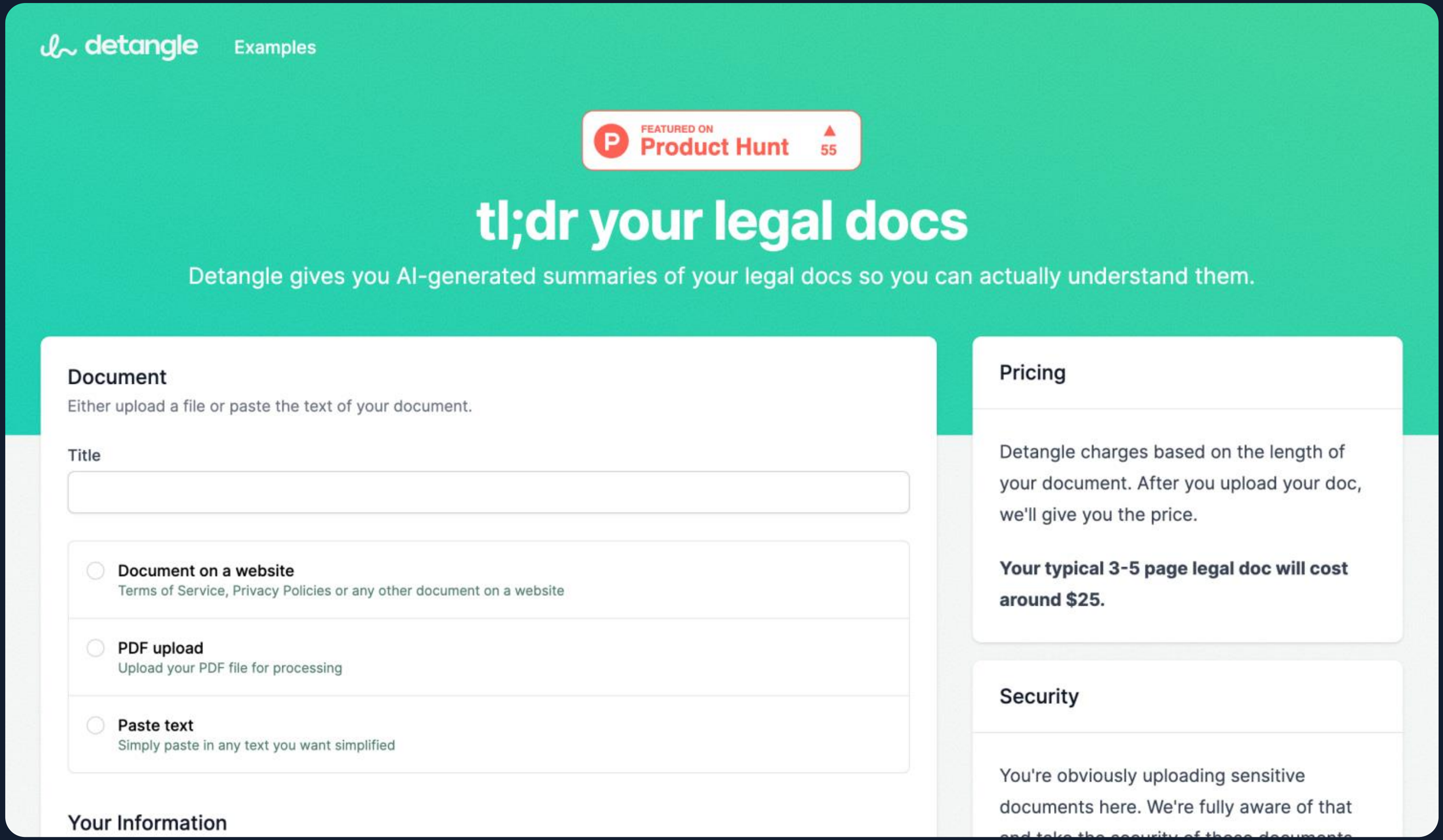


Detangle AI

情景

用户上传法律文献/文件，自动解析并归纳成重点汇总

- 解析文章
- 总结/归纳

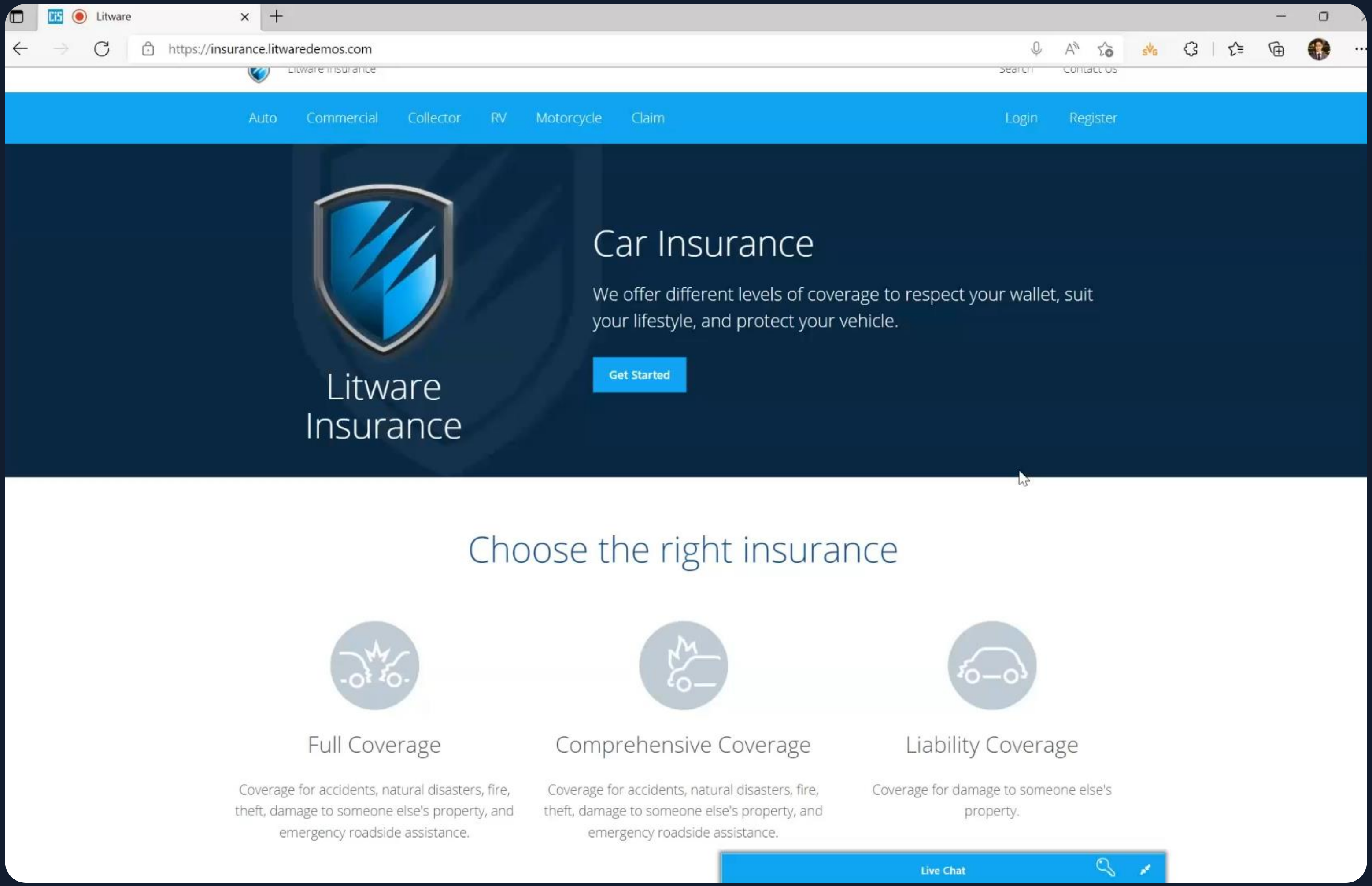


更好的客服AI体验

情景

更好地协助完成自动交易，更自然的交互和客户体验

- 语义理解
- 按情景动态生成自然对话
- 多轮对话（ 上下文继承 ）
- 多语种对话
- 图像识别和解析
- 声纹辨识
- Text-to-speech

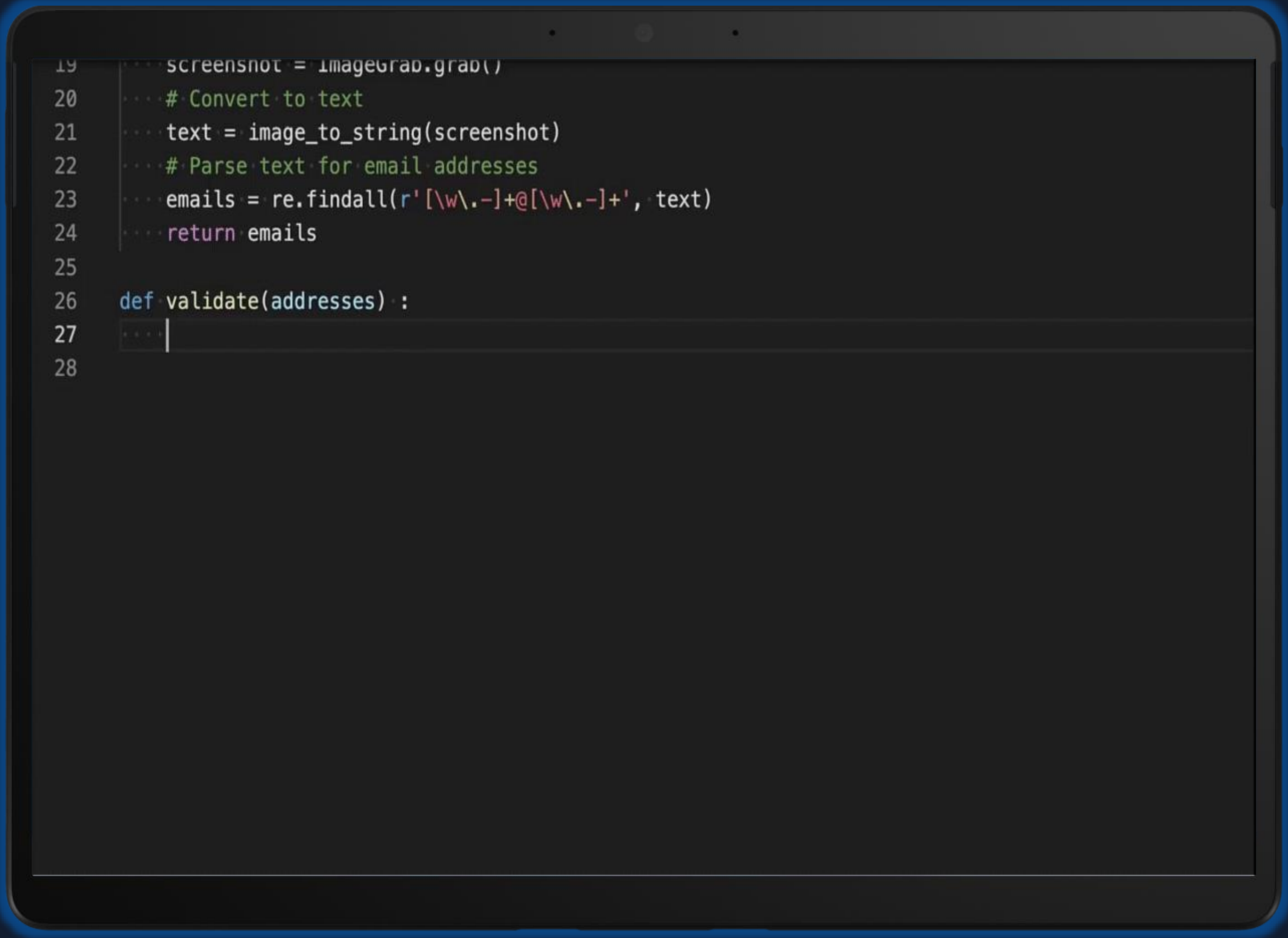


代码生成，品控，优化和Debug

情景

根据用户指示和需求，自动生成满足需求的代码（可切换语言）。 实现代码自动品控，测试，优化和除错等情景。

- CODEX代码生成
- GPT Embedding 进行语义理解和匹配



数字虚拟人游戏直播

- 游戏情节语音解析（ STT ）
- 视频标签和图像分析 （解析场景，关键物品，行为）
- 玩家对话内容解析
- 事件驱动的场景描述
- 仿真语音合成（ TTS ）
- 声纹模仿

OpenAI

语言服务

Microsoft

未完待续...