



生成式AI和信息安全

企业的实践尝试

孙琦

REEBUF

01 我们的一些AI知识储备

02 我们的一些AI实践和现状



人工智能发展的三个阶段

人工智能 (Artificial Intelligence) 是一门计算机学科，是一门人类尝试赋予机器模拟人类智能的一门学科。我们尝试让计算机系统能够模拟人类去执行各类任务，比如语音、图像的识别，处理自然语言以及做一些决策。

1980年 规则驱动的专家系统

2000年 特征+统计学习

2012年 大数据驱动的深度学习

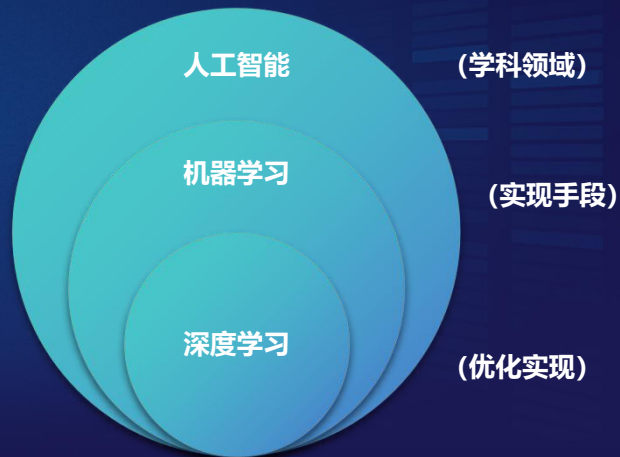




AI和机器学习的关系

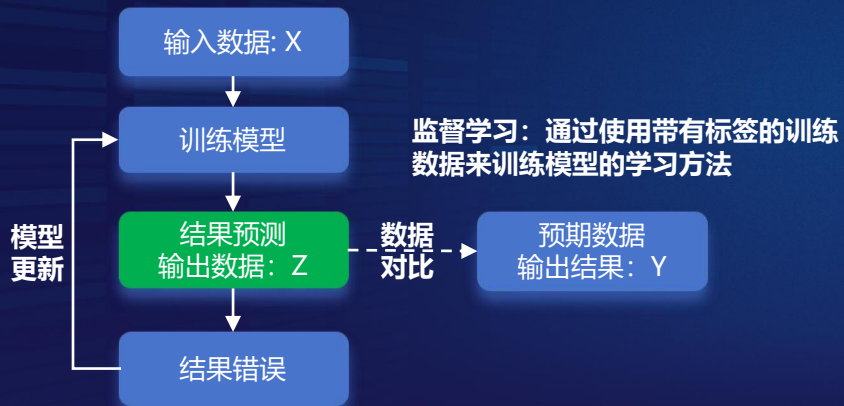
机器学习是人工智能学科领域的一个分支，通过数据和统计模型让机器自动学习和改进的一种方法。

机器学习的目标是设计和开发算法，让计算机系统能够在没有明确的编程的条件下从数据中学习，通过模型训练，机器学习能够做到识别模式、做出预测和决策等能力。



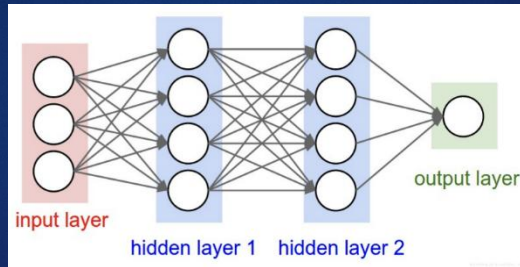


监督学习和无监督学习



深度学习

深度学习：是机器学习的一个特定领域，利用人工神经网络模型进行学习和训练，通过算法和模型让计算机系统从数据中学习的方法。深度学习模型由多个层次（称为神经网络的层）组成，每一层都会对输入数据进行变换和表示。这些网络层通过一系列的非线性转换将输入数据映射到输出结果。深度学习模型的核心是深度神经网络（Deep Neural Network, DNN），它可以通过大量的标记数据进行训练，从而实现高度准确的预测和分类任务。



深度神经网络（Deep Neural Network, DNN）





生成式AI和深度学习的关系



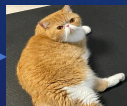
判别模型



是猫



生成模型



生成式AI是深度学习的分支



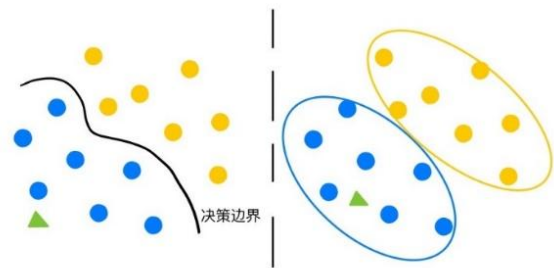


判别模型 和 生成模型

判别模型 (Discriminative Model) 和生成模型 (Generative Model) 是机器学习中两种不同类型的模型，它们的主要区别在于其对数据的建模方式和应用领域。

生成模型主要的优点是信息丰富、处理蛋类问题的灵活性比较强、增量学习以及对于缺失数据的合成；缺点也很明显，学习过程复杂，成本高昂；判别模型的优点是类间差异清晰、灵活的分界边界、学习简单和性能好；缺点主要表现在不能反应数据的特性、需要全量数据的学习

判别模型 vs 生成模型



判别模型 和 生成模型

判别模型是一种直接对条件概率进行建模的模型。它主要关注的是给定输入数据，预测输出类别或标签的概率分布。判别模型通过学习输入和输出之间的关系来建立决策边界，从而对新的输入数据进行分类。常见的判别模型包括逻辑回归、支持向量机和深度神经网络等。判别模型通常用于分类、回归和标注等任务。

生成模型是一种对联合概率分布进行建模的模型。它不仅学习输入和输出之间的关系，还学习了生成输入数据的过程。生成模型可以通过学习数据的分布和特征之间的关系来生成新的样本数据。常见的生成模型包括高斯混合模型（Gaussian Mixture Model, GMM）和生成对抗网络（Generative Adversarial Network, GAN）等。生成模型通常用于生成新的图像、语言模型和数据增强等任务。

判别模型和生成模型的选择取决于具体的问题和任务需求。判别模型更关注分类和预测的准确性，可以直接对输入和输出之间的关系进行建模。而生成模型更关注数据的生成过程，可以模拟数据的分布和生成新的样本。生成模型可以用于生成新的数据，但在分类和预测任务上可能不如判别模型准确。



数据和标签之间的关系



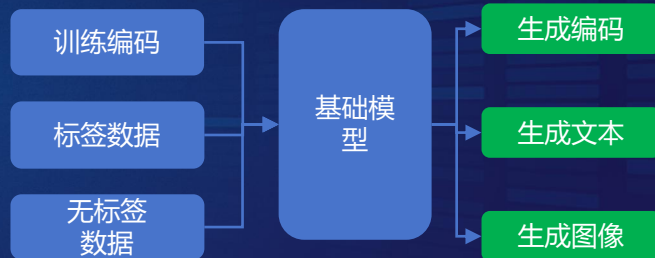
数据的分布和特征之间的关系



生成式AI的监督、半监督和无监督学习



传统的监督、无监督学习，将训练数据和标注数据喂给模型，可以作出预测、分类和聚类。



生成式 AI 的监督、半监督、无监督学习，将训练数据、打标数据和未打标数据给基础模型，然后生成新的内容，最终实现文本、代码和图片的生成。





生成式AI和传统编程、神经网络的区别

定义 狗

类型：动物

腿：4

耳朵：2

毛皮：有

传统的编程方式，需要硬编码来描述猫的一些特征。

请问这是一只猫吗？



是



不是

神经网络算法可以通过学习是不是猫的样本，然后你给出一张图片它可以判断是否为一个猫。



LaMDA、PaLM、GPT 等生成式模型在喂了大量内容后，可以直接问猫是什么，它讲给出它所知道的答案





生成式AI的定义

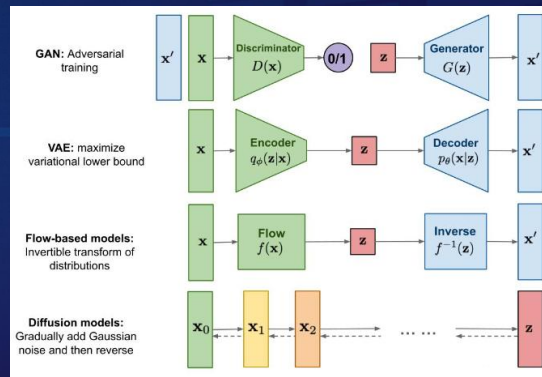
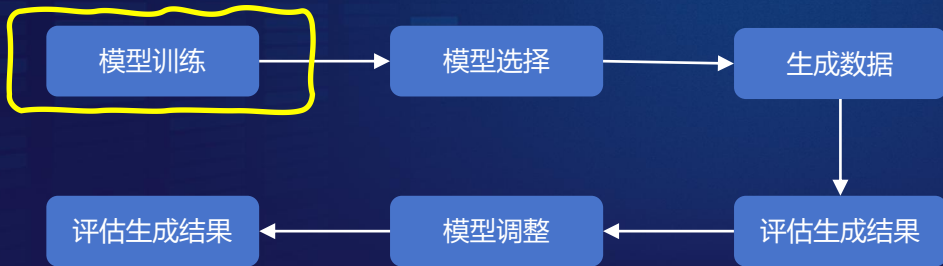
生成式 AI 是什么？

- 生成式 AI 是人工智能的一个分支，可以根据已经学习的内容生成新的内容。
- 从现有的内容中学习的过程叫做训练，训练的结果是创建一个统计模型。
- 当用户给出提示词，生成式 AI 将会使用统计模型去预测答案，生成新的文本来回答问题。



生成式AI的工作流

生成式AI的基本原理是使用概率模型或神经网络模型，将已有数据的结构和规律学习到模型中，并基于这些结构和规律生成新的数据。



生成模型的数学原理



生成式模型的分类

生成式语言模型

Generative language models

生成式语言模型 是基于自然语言处理的技术，通过学习语言的规律和模式来生成新的文本。它可以根据之前的上下文和语义理解生成连贯的句子或段落。生成式语言模型的训练基于大规模的文本数据，例如新闻文章、小说或网页内容。通过学习文本中的单词、短语和句子之间的关系，生成式语言模型可以自动生成新的、具有逻辑和语法正确性的文本，如文章、对话和诗歌等。

生成式图片模型

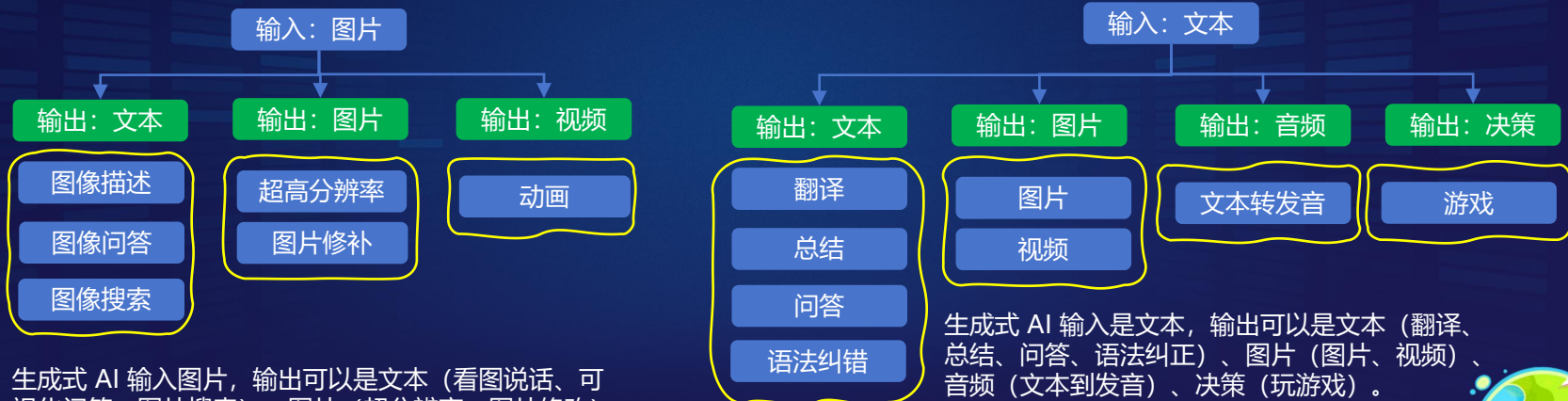
Generative image models

生成式图片模型是基于计算机视觉的技术，通过学习图像的特征和结构来生成新的图像。它可以从之前的训练数据中学习到图像的特征表示和统计规律，然后使用这些知识生成新的图像。生成式图片模型的训练通常基于大规模的图像数据集，例如自然图像或艺术作品。通过学习图像的纹理、颜色、形状和物体之间的关系，生成式图片模型可以生成具有视觉真实感或艺术风格的新图像，如自然风景、人像或抽象艺术作品等。





生成式模型的分类



生成式AI的工作原理

make a presentation about AI and cybersecurity

✓ Searching for: AI and cybersecurity presentation FCIS 2023
✓ Generating answers for you...

Great! Here's a brief outline of the presentation:

Introduction

- Definition of AI and cybersecurity
- Importance of AI in cybersecurity

AI and Cybersecurity

- AI in threat detection
- AI in malware detection
- AI in phishing detection
- AI in network security

AI and Cybersecurity Challenges

- AI in adversarial attacks
- AI in data poisoning
- AI in model stealing

生成式语言模型学习训练数据中的语言模式，然后给出一些文本，它们将会预测后面的内容是什么。将用户的输入进入 Transformer 模型的编码器和解码器进行处理，然后在生成式预训练模型中进行处理，最终将结果输出给用户。

输入：我今天心情不太好

Transformer

编码组件

解码组件

预训练：

- 海量数据
- 数十亿参数
- 无监督学习

生成式预训练Transformer模型

输出：

巧了，我今天心情也不太好，要不我们一起去找老板聊聊天吧



Transformer

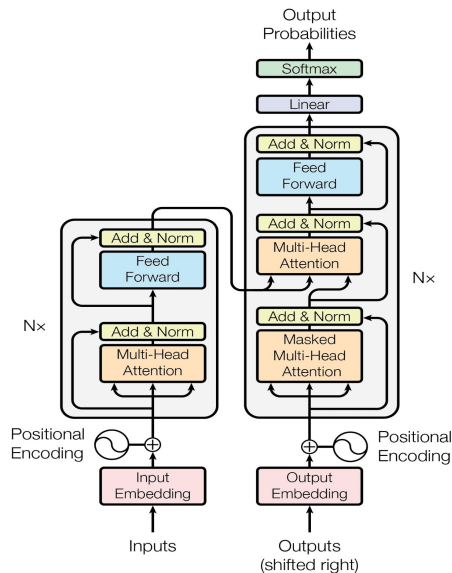
Transformer模型是由谷歌公司提出的一种基于自注意力机制的神经网络模型，用于处理序列数据，在自然语言处理领域中得到了广泛应用。

作为一种基于自注意力机制的神经网络模型，Transformer模型能够对序列中的每个元素进行全局建模，并在各个元素之间建立联系。**Transformer模型中包含了多层encoder和decoder**，每一层都由多个注意力机制模块和前馈神经网络模块组成。encoder用于将输入序列编码成一个高维特征向量表示，decoder则用于将该向量表示解码成目标序列。在Transformer模型中，还使用了残差连接和层归一化等技术来加速模型收敛和提高模型性能。**Transformer模型的核心是自注意力机制（Self-Attention Mechanism）**，其作用是为每个输入序列中的每个位置分配一个权重，然后将这些加权的位置向量作为输出。

自注意力机制的计算过程包括三个步骤：

1. 计算注意力权重：计算每个位置与其他位置之间的注意力权重，即每个位置对其他位置的重要性。
2. 计算加权和：将每个位置向量与注意力权重相乘，然后将它们相加，得到加权和向量。
3. 线性变换：对加权和向量进行线性变换，得到最终的输出向量。

通过不断堆叠多个自注意力层和前馈神经网络层，可以构建出Transformer模型。对于Transformer模型的训练，通常采用无监督的方式进行预训练，然后再进行有监督的微调。在预训练过程中，通常采用自编码器或者掩码语言模型等方式进行训练，目标是学习输入序列的表示。在微调过程中，通常采用有监督的方式进行训练，例如在机器翻译任务中，使用平行语料进行训练，目标是学习将输入序列映射到目标序列的映射关系。





幻觉&提示词

幻觉

模型通过学习大量的文本数据，尝试预测下一个单词或短语。然而，有时候模型会生成一些不符合语法规则或意义不明的词语或短语，这被称为"幻觉 (hallucinations)"。幻觉可以视为模型在生成过程中的错误或缺陷，可能由于训练的数据量不够、模型的训练数据质量差、没有给模型足够的上下文、没有给模型足够的约束导致的。

提示词

提示词是作为大语言模型输入的一段文本，它可以以各种方式用来控制模型的输出。提示词设计是创建提示的过程，从而从大型语言模型中生成期望的输出。正如我们之前提到的，生成 AI 在很大程度上取决于你输入的训练数据。它分析输入数据的模式和结构，生成内容。因此输入的质量决定了输出的质量。





生成式模型的分类

文本到文本生成模型旨在接收一个文本输入，并生成一个相关的文本输出。这种模型可用于机器翻译、文本摘要、对话生成、故事生成等任务。生成模型可以学。

文本到图像生成模型接收一个文本描述作为输入，并生成对应的图像输出。这种模型可以将自然语言描述转化为视觉内容，用于图像生成、图像标注、图像编辑等任务。通过学习文本描述和图像之间的语义关联，模型可以生成与文本描述相匹配的图像。常见应用场景：图像生成、图像标注、图像编辑修复。

文本到视频或三维生成模型接收一个文本输入，并生成相应的视频或三维模型输出。这些模型可以用于视频生成、场景合成、三维模型生成等任务。模型可以学习从文本描述到视频序列或三维模型的转换过程，生成与文本描述相符的动态视频或立体模型。常见应用场景：视频生成、场景合成、三维模型生成。

文本到任务生成模型旨在根据文本输入执行特定任务。这些模型可以接收自然语言指令或问题，并生成相应的任务执行结果。例如，问答生成模型可以接收问题，并生成相应的答案；代码生成模型可以接收自然语言描述，并生成相应的代码实现。这种模型能够将文本指令转化为任务执行的具体操作。常见应用场景：问答生成：根据问题生成相应的答案或解决方案。代码生成：将自然语言描述转化为代码实现。指令执行：根据自然语言指令执行特定的任务，如图像处理、数据操作等。





Model Garden

Model Garden 是 Google 的一个开源项目，旨在为研究者和开发者提供预训练的机器学习模型和相关的训练和优化技术。这些模型覆盖了许多不同的机器学习任务，例如图像分类、物体检测和自然语言处理等。

Model Garden 中的模型分为两大类：语言模型和视觉模型。

语言模型：这类模型可以进行一些特定的语言处理任务，如：提取（Extraction）：这类任务包括语法分析（Syntax Analysis），它可以理解文本的语法结构。分类（Classification）：这类任务包括实体分析（Entity Analysis，识别文本中的特定实体，如人名、地名等）、内容分类（Content Classification，根据内容的主题进行分类）、情感分析（Sentiment Analysis，评估文本的情感倾向，如积极、消极等）、实体情感分析（Entity Sentiment Analysis，评估特定实体在文本中的情感倾向）。

视觉模型：这类模型可以进行一些特定的视觉任务，如：分类（Classification）：这类任务包括对象检测（Object Detector，识别图片中的特定对象）。检测（Detection）：这类任务包括人流量分析（Occupancy Analytics，分析特定区域内的人流量）、人/车辆检测（Person/Vehicle Detector，识别图片中的人或车辆）、个人防护设备检测（PPE Detector，识别图片中是否有人穿戴个人防护设备）、人物模糊（Person Blur，将图片中的人物进行模糊处理）。这些模型都是为特定的任务进行训练和优化的，可以用来解决一些具体的实际问题。

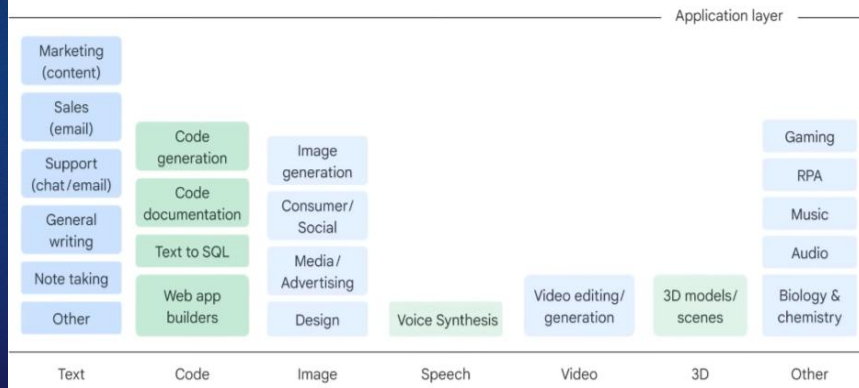




生成式AI应用

生成式 AI 在文本（生成写作、AI 笔记、销售文案、聊天机器人、邮件编写等）、代码（代码生成、代码文档、文本转SQL、Web 应用构建等）、图片、发音、视频、3D 等领域都有大量的市场。

The generative AI Application Landscape



生成式AI的商业应用

金融

AI是普惠金融的核心驱动力之一，AI可以赋能金融企业节省大量人力成本提高效率，从而改善用户体验和减少信息不对称，助力金融客户实现智能化升级。

智能客服 身份识别 信息审核
智能双录 智能营销

新零售

阿里AI技术渗透新零售领域，构建数据打通、场景贯通、深度触达的“AI+零售”体系，利用人工智能算法技术将人与货、人与场实时结合、真正打穿，全面提升运营效率、提升消费者体验，助力零售业数字化升级。

智能客服 商品识别分析
消费者识别分析 精准营销

教育

随着AI技术的引入，教育行业正在脱离单一教育辅助的角色，为受教育者提供科技赋能、内容完善、效果优良的课程。结合海量优质资源覆盖终身学习场景，实现高质量教育的可持续发展目标。

智慧课堂 口语评测 智能助教
试卷分析 学情监测

政务

以阿里云AI技术为基石，把人工智能技术属性和社会属性高度融合，辅助政府在经济、治理、民生等领域的管理变得更加精细化、智慧化，整合并高效利用政务资源，助力政务数智化转型。

电子政务 客服机器人 智能办公
访客预约

交通

AI智慧赋能交通行业，可助力交通信息广泛应用与服务，提升交通系统运行效率和管理水平，打造实时、准确、高效的城市交通智能体。

电子收费 便捷出行 智慧停车
突发事件处理 电子警察

司法

阿里AI正在利用大数据和人工智能推进着一场数字化、智能化革命升级，集中AI能力服务于中国司法行业，能有效提高司法效率、保证司法公开公正、提升司法公信力等作用和价值，为行业带来更多值得期待的创新。

智慧辅助审理系统 庭审记录
智能速记员 断案辅助功能



内容引用：百度AI开放平台



ChatGPT

ChatGPT (chat Generative Pre-trained Transformer) 是由总部位于旧金山的初创公司 OpenAI 开发的人工智能聊天机器人。OpenAI 于 2015 年由 Elon Musk 和 Sam Altman 共同创立，并得到了知名投资者的支持——最著名的是微软。它是生成式人工智能的几个例子之一。这些工具允许用户输入书面提示并接收人工智能生成的新的类人文本或图像和视频。

局限性

- 答案可能是错误的
- 可能会产生有害的指引或有偏见的内容
- 因为数据样本仅限于2021年之前，所以对于世界的认知并不全面





信息安全场景下的生成式AI应用探索实践

场景：日志分析

痛点：日志太多，跨度太大，即使有siem也存在海量日志被忽略或者威胁遗漏等问题

需求：通过ML对现有的日志进行处理，并通过生成式AI提供完整的关联分析和统计

现状：数据的准确性是一个巨大的问题，可能是团队的技术沉淀不够也可能是数据的问题

场景：智能安全小助手

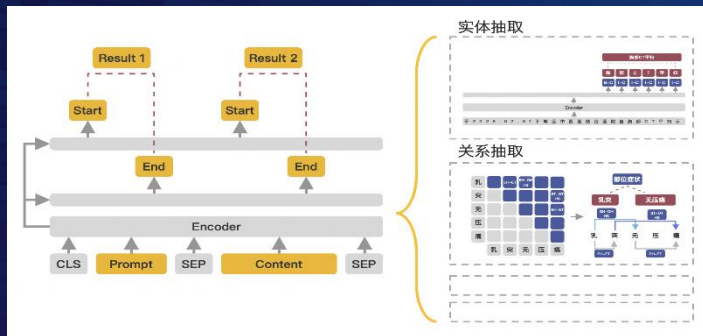
痛点：安全人员工作量太重

需求：通过在app内集成小助手，实现交互式信息获取及指令下达

现状：小助手太幼稚，指令下达有问题，企业内部AI的应用真的有较高的门槛



信息安全场景下的生成式AI应用探索实践



背景：抽取医疗病历中的信息，为医生科研，药企制药用

（右肝肿瘤）肝细胞性肝癌（II-III级，实索型和假腺管型），肿瘤包膜不完整，紧邻肝被膜，侵及周围肝组织，未见脉管内癌栓（MVI分级：M0级）及卫星灶形成。（肿物1个，大小4.2×4.0×2.8cm）。

→ **肿瘤级别** II-III级

→ **脉管内癌栓分级** M0级

→ **肿瘤的个数** 1个

→ **肿瘤的大小** 4.2×4.0×2.8cm

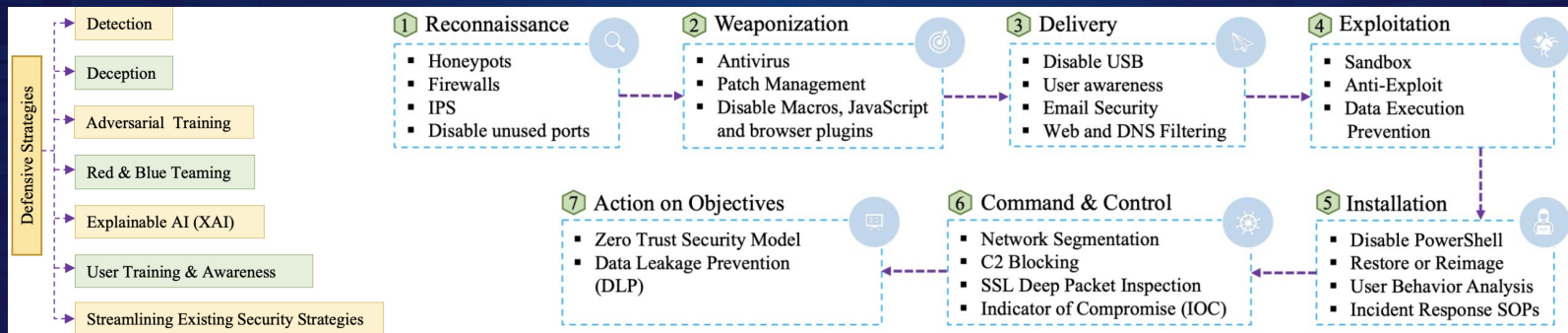
	金融		医疗		互联网	
	0-shot	5-shot	0-shot	5-shot	0-shot	5-shot
uie-base (12L768H)	46.43	70.92	71.83	85.72	78.33	81.86
uie-medium (6L768H)	41.11	64.53	65.40	75.72	78.32	79.68
uie-mini (6L384H)	37.04	64.65	60.50	78.36	72.09	76.38
uie-micro (4L384H)	37.53	62.11	57.04	75.92	66.00	70.22
uie-nano (4L312H)	38.94	66.83	48.29	76.74	62.86	72.35
uie-m-large (24L1024H)	49.35	74.55	70.50	92.66	78.49	83.02
uie-m-base (12L768H)	38.46	74.31	63.37	87.32	76.27	80.13

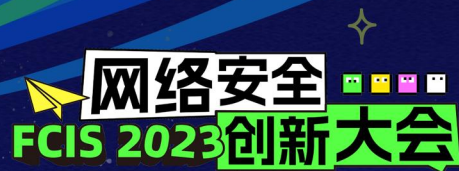
UIE(Universal Information Extraction): Yaojie Lu等人在ACL-2022中提出了通用信息抽取统一框架UIE。该框架实现了实体抽取、关系抽取、事件抽取、情感分析等任务的统一建模，并使得不同任务间具备良好的迁移和泛化能力。

数据引用：Unified Structure Generation for Universal Information Extraction <https://arxiv.org/pdf/2203.12277.pdf>



信息安全场景下的生成式AI应用探索实践





THANKS





THANKS





THANKS

