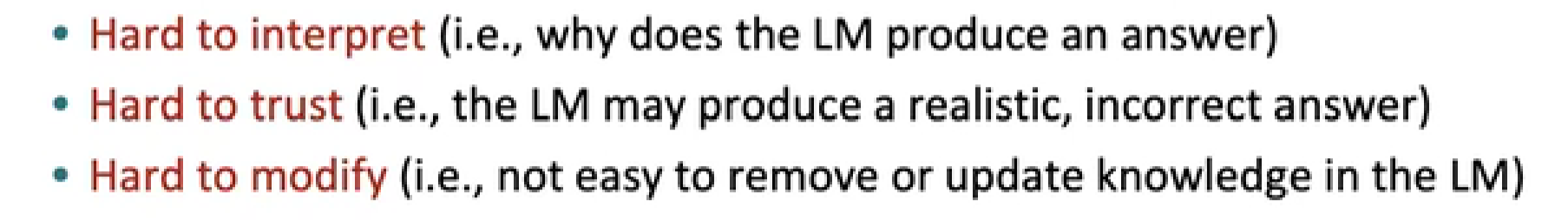
## 斯坦福机器学习课程

Lecture15

① 掩码语言模型（MLM）: MLM的训练方法是通过随机掩盖句子中的某些词语，然后训练模型根据上下文预测这些被掩盖的词语。

②Language Model可能可以扮演一个知识库的角色，这需要其拥有较高的知识质量。取代传统的为已经有的知识建立知识库，而是直接用自然语言对自然语言模型进行询问。并且这样建立的知识库更加灵活，面对更加多样的查询能做到更好的响应。语言模型作为知识库可能的缺点：



这也引出了一个问题——如何向模型中添加知识？

**方法一：加入预训练的实体嵌入;**

③实体链接：将文本中的实体（如人名、地名、组织名等）与知识库中对应的实体进行匹配和链接。将识别出的实体与知识库中的实体进行匹配，以确定它们具体指代的是什么，消除歧义。

④ERNIE（Enhanced Language Representation with Informative Entities）：知识图谱中的实体可以作为额外的知识来增强文本特征。 所以采用大规模文本语料库和知识图谱来训练增强型语言表示模型ERNIE。

底层文本编码器T-Encoder：

从输入的token中捕获基本词汇和句法信息。对每个Token进行Token嵌入，段落嵌入、位置嵌入并且拼接三者形成输入的嵌入。T-Encoder基于Transformer中的多头自注意力机制进行文本的语义特征提取，生成每个单词的向量表达。

上层知识编码器K-Encoder:

负责将额外的知识信息整合到底层的文本信息当中，以便将Token和实体的异质信息表示到统一的特征空间中。

**方法二：让模型接触到外部信息：**

⑤KGLM(Knowledge Graphs for Fact-Aware Language Modeling):结合了知识图谱和自然语言处理的深度学习方法，旨在提升模型的事实准确性，使其能够进行基于事实的语言建模。引入了一个新的实体/关系嵌入层，该层可以学习区分不同的实体和关系类型，从而允许模型学习知识图的结构。在这项工作中，我们展示了使用从知识图中提取的三元组进一步预训练带有该附加嵌入层的语言模型，然后进行标准微调阶段。当训练的知识图谱发生改变的时候，模型会根据不同的三元组给出不同的预测。

⑥KNN LM：结合了k近邻算法和语言模型的方法，利用预先训练好的语言模型构建数据存储，并在测试阶段执行k近邻搜索来插值基本语言模型的概率，从而提升模型的性能和泛化能力。

**方法三：修改模型训练数据：**

⑦WKLM：使用实体替换使得模型参数编码到更多的知识。在模型训练过程中，随机的将实体根据相似关系进行一定程度的替换，让模型判断哪些是经过替换的实体，进而产生一个loss，添加这样的loss有助于模型学习到更多知识。

⑧在LMs中查询知识时使用更好的提示：模型可能对句子的短语结构敏感。有时稍微重写查询有助于模型回忆起它所学到的东西。

⑨知识驱动的下游任务：（需要微调）QA、实体类型、关系提取。

Lecture16

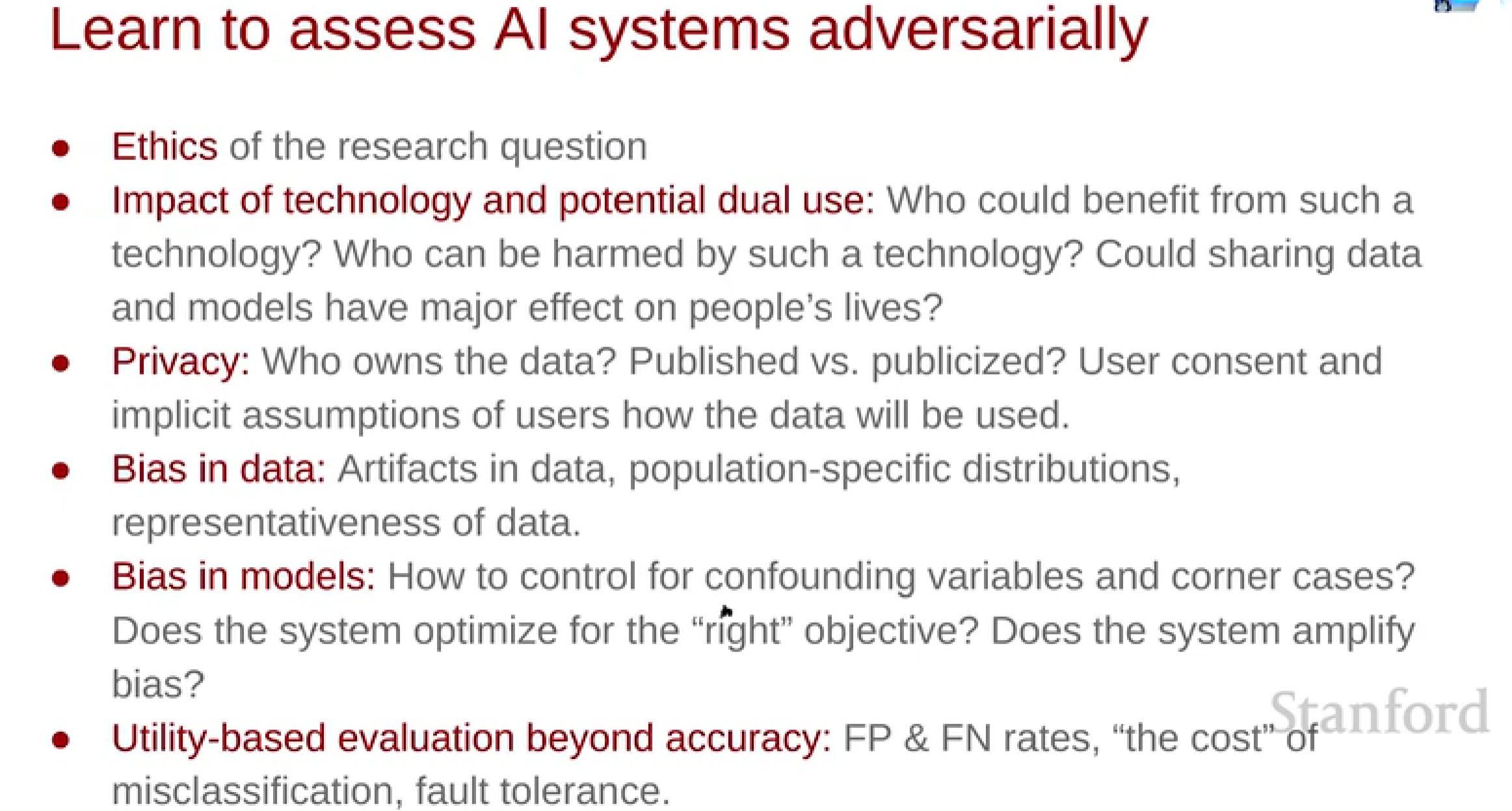
①Chicken dilemma：一个分类器的结果会导致鸡被杀掉活着分配去下蛋，这牵扯到的并不仅仅是一个分类问题，可能还有一定的道德伦理问题。

②IQ classifier：指一个分类器，用于根据人物的外表评估一个人的智力水平。但在科学和伦理上，用外表来推断智力是没有依据的。智力是一个复杂的特质，受到多种因素的影响。此外，这种方法涉及到的伦理问题，可能侵犯个人隐私和尊严。基于外表的评估可能会导致不公平的待遇和机会的丧失。

③AI Gaydar：指的是用人工智能技术来预测或推断一个人的性取向。这种技术尝试通过分析人的外貌、行为、社交媒体活动等数据，得出某个人的性取向。AI模型的预测准确性通常依赖于其训练数据和算法的设计。如果模型的训练数据存在偏差，或者算法不够完善，预测结果可能不准确。即使预测准确，这种技术的使用也可能引发严重的伦理问题。可能会被用于歧视、骚扰,强化刻板印象。

④对抗性地评估AI系统：通过模拟恶意攻击或输入来发现模型的弱点和潜在的安全漏洞。这对于确保AI系统的稳健性和安全性十分重要，特别是在高风险的应用领域。

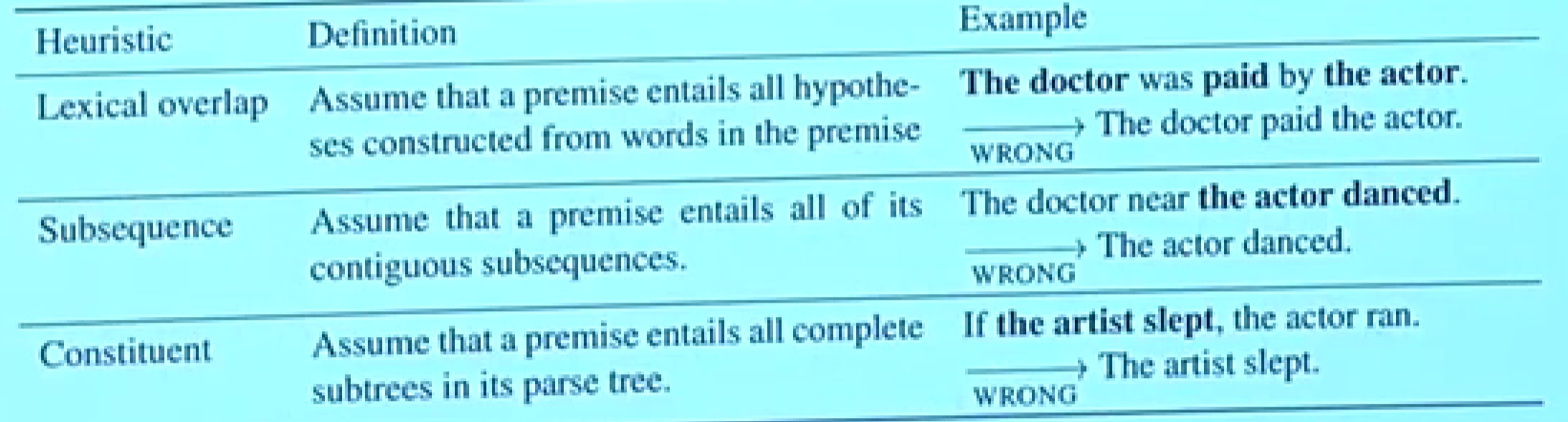
⑤在进行模型配置前要考虑的问题：



⑥data bias: 数据偏差是指在构建NLP模型的过程中，由于数据集的不平衡或代表性不足，导致模型的输出结果具有系统性的偏差。这种偏差可能导致模型在特定人群、话题或场景下的表现不佳，引发歧视或不公正。

Lecture17

①HANS测试：用于评估自然语言推理（NLI）模型的一种基准测试，检测模型对一些简单启发式的依赖。帮助了解模型是否真的在进行语言理解和推理，还是仅仅依赖于数据中的一些表面特征或简单规则。一些hans例子：



②知识评估（Knowledge Evaluation） ：指评估模型在特定知识领域的理解和应用能力。这通常涉及测试模型是否能够正确回答与特定领域相关的问题，或者是否能够生成符合领域知识的文本。

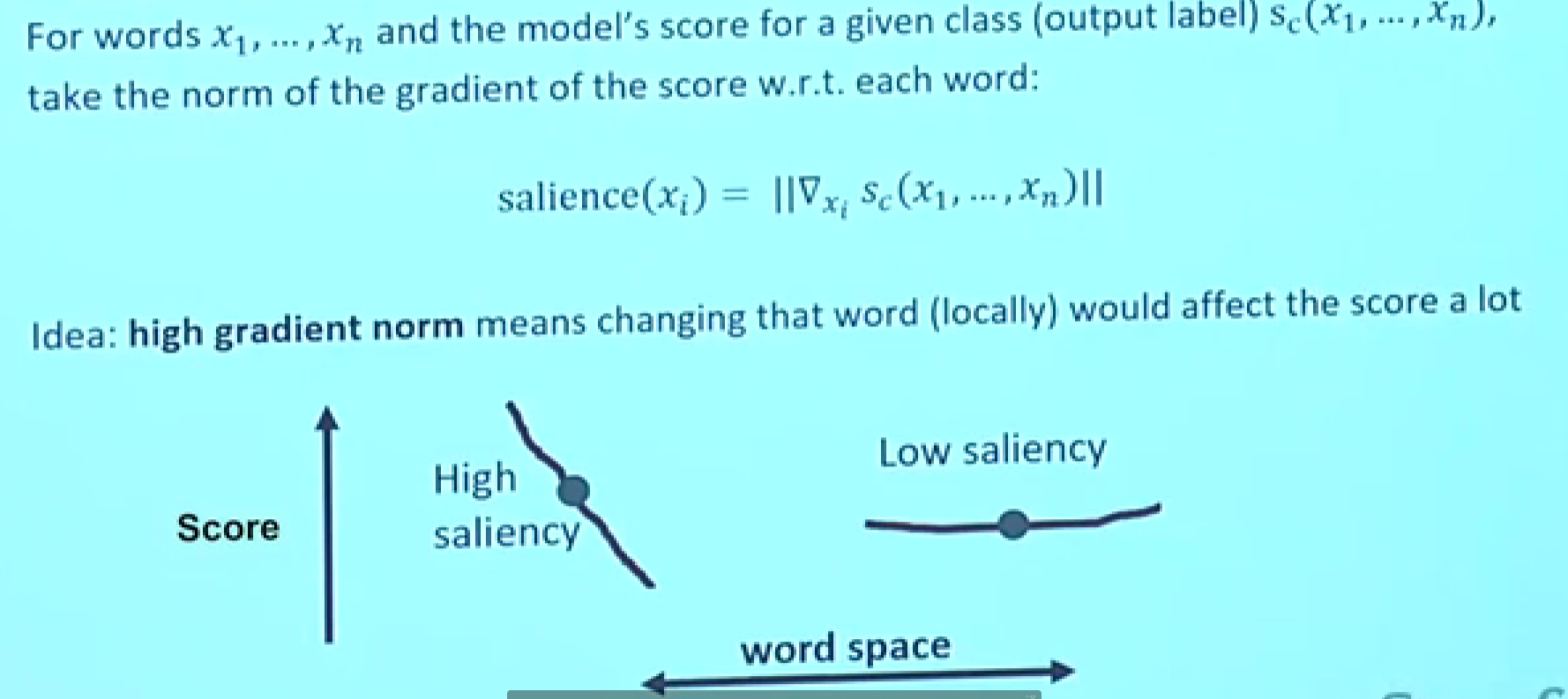
模型分析（Model Analysis） ：指对模型的性能和行为进行详细分析，以理解模型的优点和缺点，并识别改进的机会。模型分析通常包括性能评估、错误分析和特征重要性分析等。其分析范围更加广泛。

③影响研究（Influence Study） ：指分析和理解训练数据中的每个样本对模型性能的影响。通过影响研究，可以识别哪些样本对模型的预测结果有显著影响，从而帮助改进模型的鲁棒性和性能。

④对抗样本（Adversarial Examples） ：指通过微小的扰动生成的输入样本，这些样本能够误导模型，导致其做出错误的预测。对抗样本研究旨在理解和提高模型的鲁棒性，防止模型在面对对抗攻击时失效。

⑤显著图（saliency maps）：突出显示图像中最重要或“显著”的部分。在语言模型中，这些图可以帮助我们理解一个模型在做决策时关注的词汇，或者哪些部分对模型的输出最有影响。

如何衡量某个词对输出的影响呢？可以采用简单梯度法通过计算在某个单词上关于模型分数的梯度：



⑥Probing Model：探针模型通常是简单的机器学习模型，如线性模型、逻辑回归模型等，用于预测特定的语言特征或属性，从而评估预训练模型的表示能力。例子：

1.加载预训练的 BERT 模型和分词器：使用transformers库加载预训练的 BERT 模型和分词器。

2.提取 BERT 的表示：将输入句子转换为 BERT 的输入格式，并提取最后一层的隐藏状态作为特征。

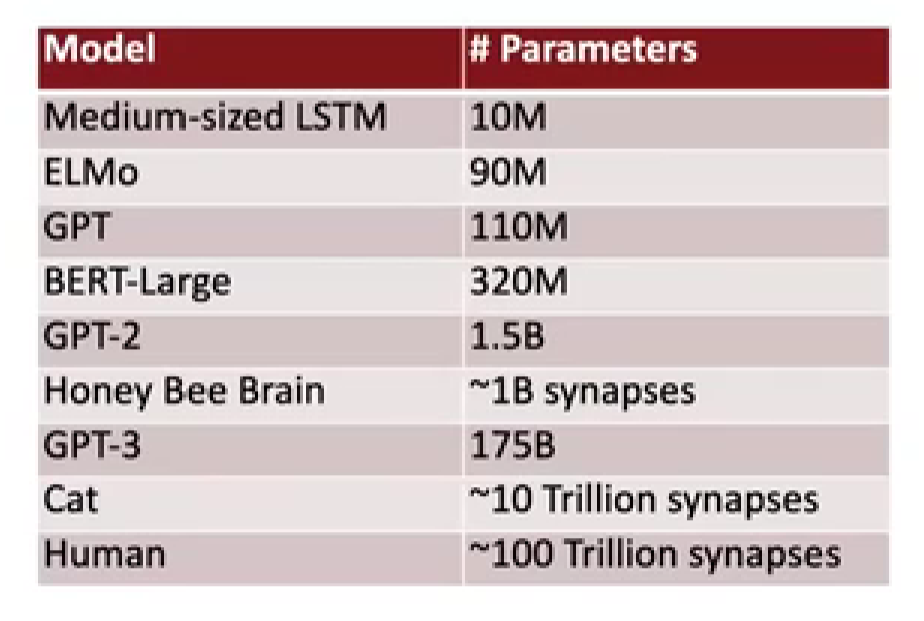
3.训练一个简单的线性探针模型：使用线性模型来预测词性标注。

评估探针模型的性能：使用准确率来评估探针模型的性能。

通过这种方式可以评估 BERT 的表示能力，理解其在处理输入数据时所学到的特征。

Lecture18

①常见的模型的参数大小：



②GPT3的in-context learning

③系统性与组合性：组合性指的是能够将简单的语言单元组合成复杂结构的能力。系统性强调的是这些组合遵循一致的规则，并且模型能够理解和应用这些规则，甚至在新的上下文中正确地生成或理解语言。

组合性在指的是一个系统通过将较小的、独立的部分组合在一起，从而生成更大、复杂结构。在语言中，这通常意味着词语通过语法规则组合成短语和句子。如在自注意力机制允许模型在生成输出时注重输入序列中的不同部分。这种机制使得模型能够理解和生成由多个单元组合而成的复杂结构。

验证模型的组合性，要验证将简单的语言元素组合成复杂的结构，并在新的上下文中正确应用这些组合规则。

④动态基准：用来评估模型性能的方法它与传统静态基准不同，能够随着时间的推移或模型的改进而动态变化。这种方法可以更好地捕捉模型的进步，并避免模型过度拟合固定的测试集。

⑤在不同的规模上塑造出不同的语言模型，其中多模态是一个重要方向（指的是将语言与视觉、听觉、触觉等相联系）：

