动手学大模型隐写

张玉龙

大模型在安全通信上的应用

上海交通大学 SHANGHAI JIAO TONG UNIVERSITY

- 隐写术（Steganography）是一种信息隐藏技术，其核心目的**是将信息嵌入到各种载体（如数字图像、音频、视频或文本）中，以实现隐蔽通信。**隐写术与加密技术不同，它**不仅隐藏信息的内容，还隐藏信息传输行为的存在性。**这种技术利用人类感知系统对某些信息不敏感的特性，**将秘密信息隐藏在数字载体的冗余信息中，使得信息在表面上看起来与普通载体无异，从而难以被攻击者察觉。**

- 隐写术的应用包括但不限于隐蔽传输、版权保护等。随着技术的发展，隐写术在军事、商业等领域变得越来越重要，同时也为恶意行为提供了便利，如间谍活动、恐怖袭击等。

隐写的意义

- 简单来说，隐写是与密码术不同的一种安全信息传输方法。
- 密码是“加密”明文，使得要传递的消息内容不被中间人获取到，但消息传递载体可能被中间人发现和获取从而篡改、拦截消息的传递。
- 隐写则是“隐藏”消息传递的事实，将要传递的消息内容隐藏在公开信道上，伪装成一般内容，目标是只有消息收受方可以发现隐藏的消息，不被中间人发现。

现实生活中的简单隐写案例：藏头诗、夹在书中的小纸条等等。

隐写的下游衍生技术：水印（可以理解为一种更重视鲁棒性，愿意被很多人发现并解读出信息的隐写）。

隐写的实例 与密码对比

- 隐写术（文本格式）：
  - **操作**：调整电子文档中字母间距（如0.1pt差异）、字体颜色（#000000 vs #010101），用二进制编码信息（如“间距大=1，正常=0”）1。
  - **目标**：**信息藏于公开文本中，肉眼不可见。**
- 密码学（维吉尼亚密码）：
  - **操作**：使用密钥词重复加密（如密钥"KEY"加密"HELLO"→"RIJVS"）。
  - **目标**：生成乱码密文，需密钥解密7。
  - 对比：
    - **隐写术：载体是正常文件（如合同），不引起怀疑。**
    - **密码学：密文本身暴露“有秘密”，但内容保密。**

隐写的实例

- **隐写术（图像LSB）**：
  - **操作**：修改图片像素最低有效位（LSB），嵌入秘密数据（如另一张图的二进制）。人眼无法察觉差异，但工具可提取29。
  - **目标**：信息**藏于普通照片（如旅游照）**，绕过审查。
- **密码学（AES-256）**：
  - 操作：用密钥将文件加密为乱码（如z.exe→G8x!gF2*...
  - 目标：**即使文件被截获，也无法破解内容。**
- **协同用例**：
  - 先用AES加密敏感数据；
  - 再将密文嵌入图片LSB中。
    → **双重保护**：既隐藏存在（像普通图片），又隐藏内容（需密钥解密）

隐写的实例

- 二维码中的隐写：

  原理： 设计一个看起来完全正常的二维码（指向一个无害网站），但在其纠错区域或通过精心设计码点图案，嵌入额外的隐藏信息（另一个URL、文本、小图片）。**普通扫码软件只能读出表面信息，需要定制软件才能读出隐藏层。**

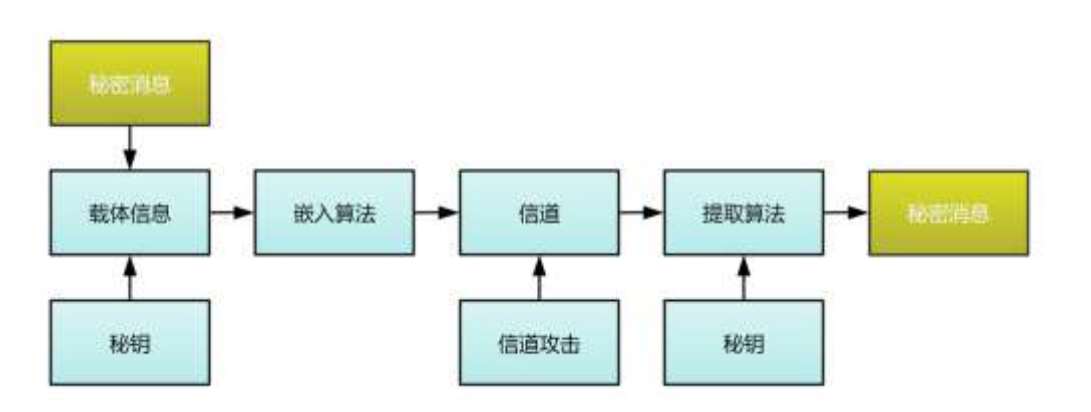
  示例： 餐厅菜单上的二维码，扫码显示菜品介绍（公开层），但用特定APP扫描能获取隐藏的当日优惠码或内部员工信息。海报上的二维码，普通扫码是活动介绍，隐藏层是VIP邀请函。

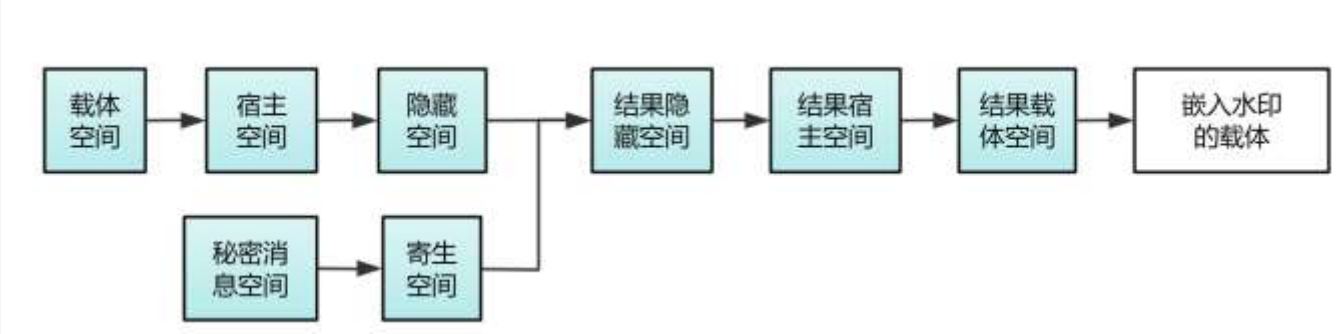
  生活场景： **营销活动（寻宝、解锁优惠），内部信息传递（员工公告），版权保护（在公开二维码中嵌入所有者信息）。**

  要点： 需要理解二维码结构（尤其是纠错码的冗余性）、需要生成双层的特殊二维码工具。

隐写的模型

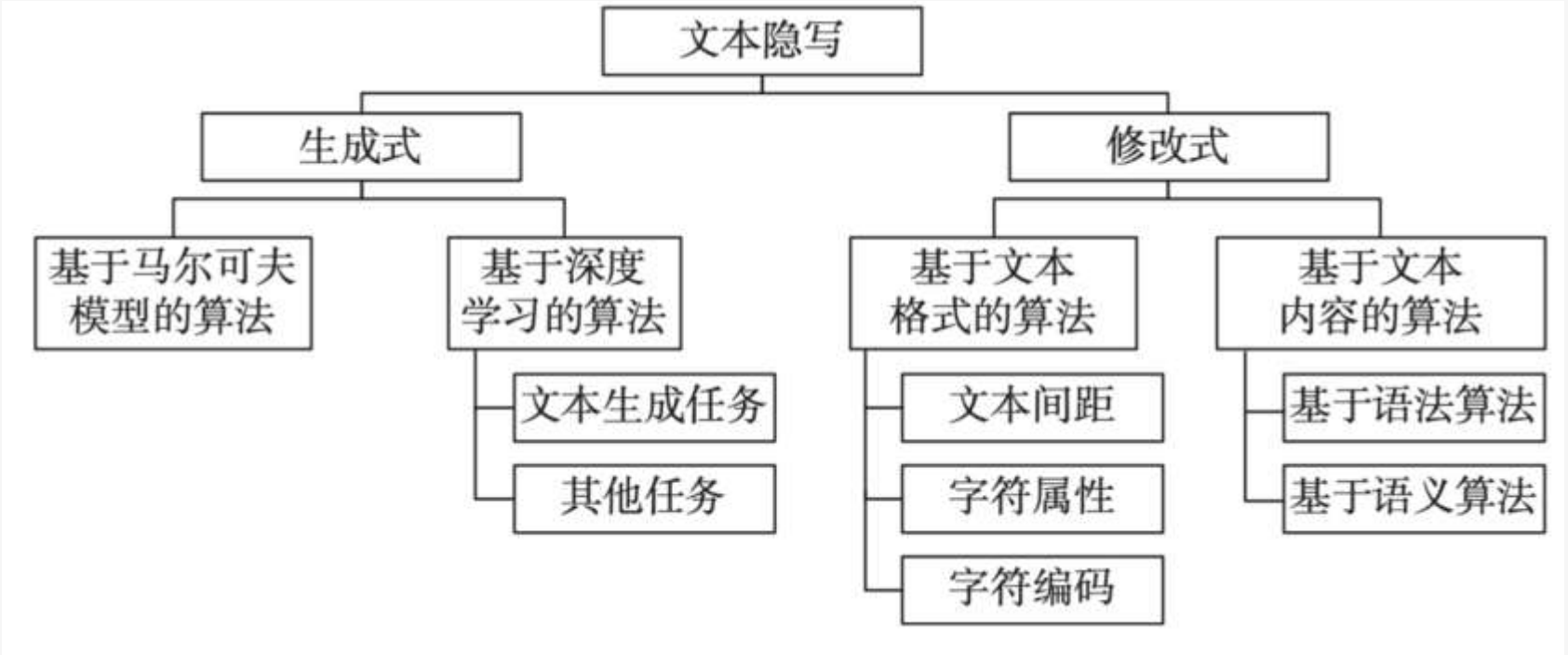
- 目前在数字媒体载体的选择上，**隐写一般选择图像或视频，因为有大量的冗余空间可以插入信息。**
- 而**文本作为隐写载体的话，冗余度低，但其也具有灵活、内容量大等特点，可以比较容易地嵌入其他内容中。**这次我们用大语言模型来介绍大模型隐写。
- **隐写（左）与水印（右）各自的过程模型如下图所示。**





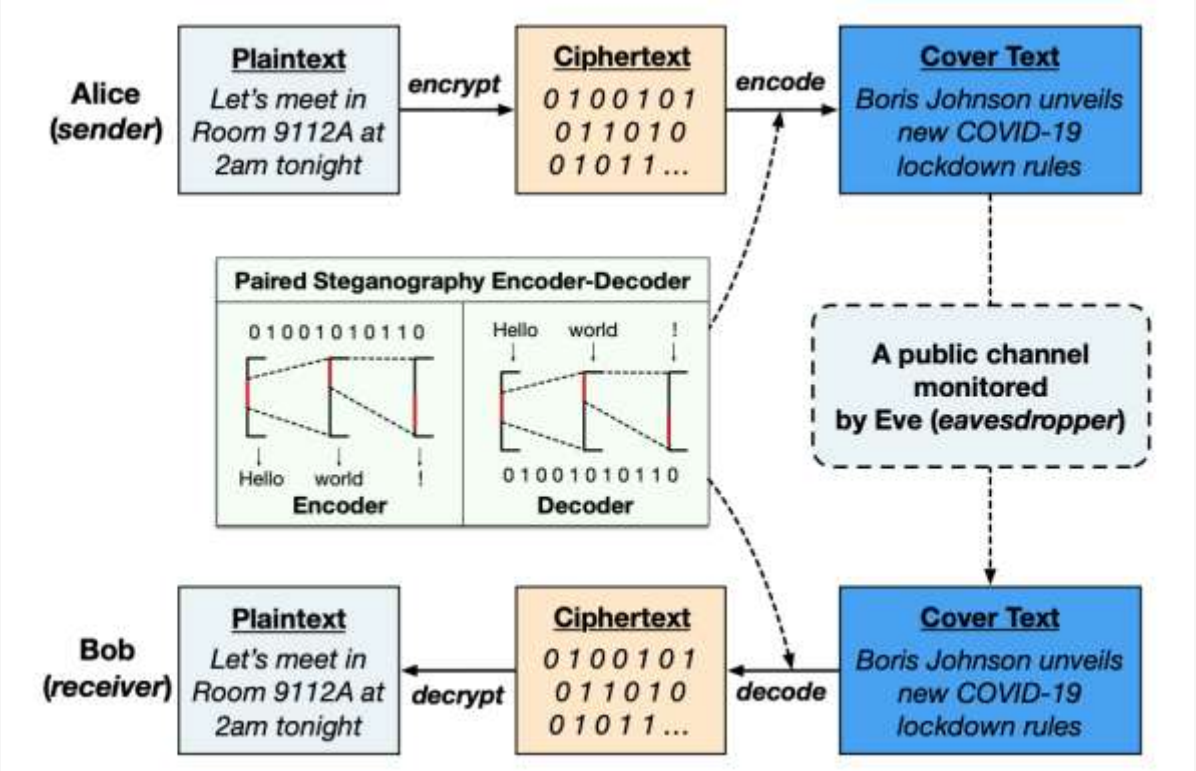
文本隐写的分类

- 文本隐写主要包含两类：修改式与生成式。
- 近年来随着深度学习特别是大模型的发展，**生成式文本隐写已经成为主流。**



大模型文本隐写[1-3]

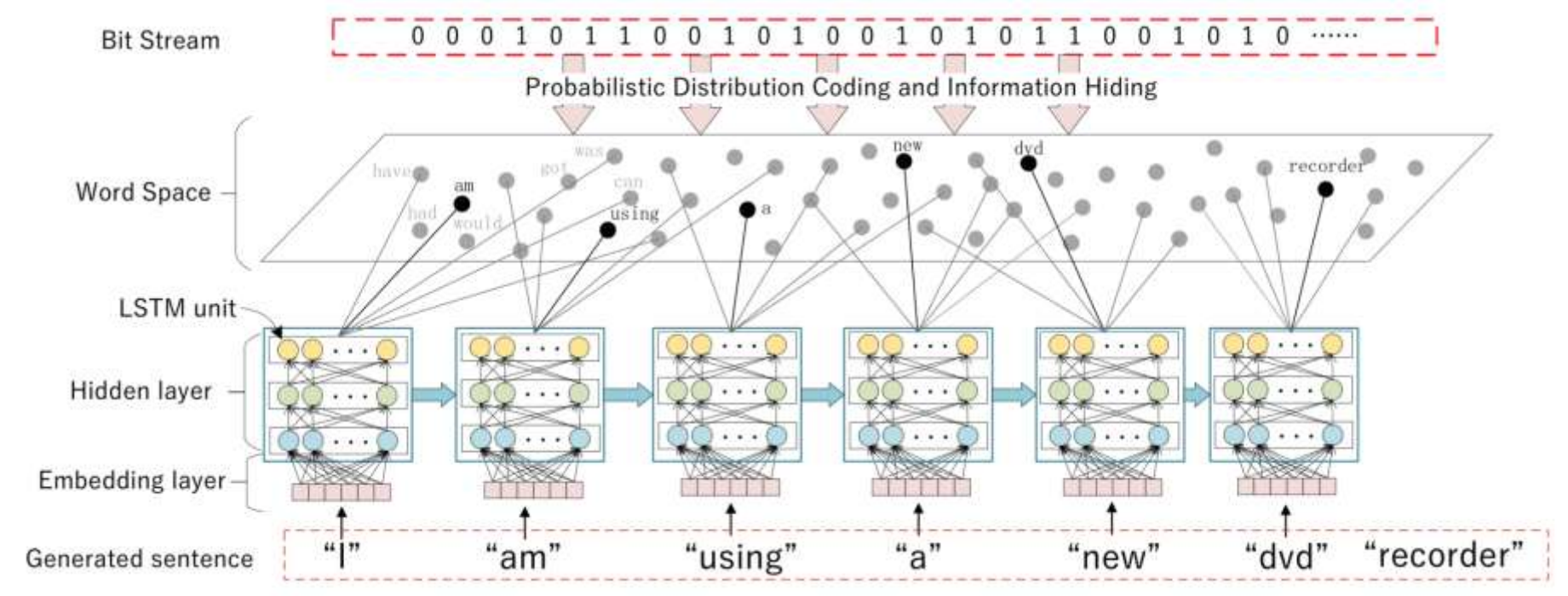
- 文本生成模型都是“序列化”的数据。
- 当输入上文和各参数后，模型会返回一个对“next token”的预测logits表，表内是从最高概率到最低概率的token ids。
- 正常情况下，模型会根据参数来随机选择n个token中的一个来作为下一个token，然后继续循环生成。
- 而生成式隐写则是在模型推理过程中，根据设置好的规则干预模型对next token的选择，从而将想要嵌入的信息在模型推理过程中嵌入到生成文本里。



[1] Yang, Zhong-Liang, et al. "RNN-stega: Linguistic steganography based on recurrent neural networks." IEEE Transactions on Information Forensics and Security 14.5 (2018): 1280-1295.
[2] Tang, Yifan, et al. "Linguistic Steganalysis via LLMs: Two Modes for Efficient Detection of Strongly Concealed Stego." IEEE Signal Processing Letters (2024).
[3] Yang, Zhong-Liang, et al. "VAE-Stega: linguistic steganography based on variational auto-encoder." IEEE Transactions on Information Forensics and Security 16 (2020): 880-895.

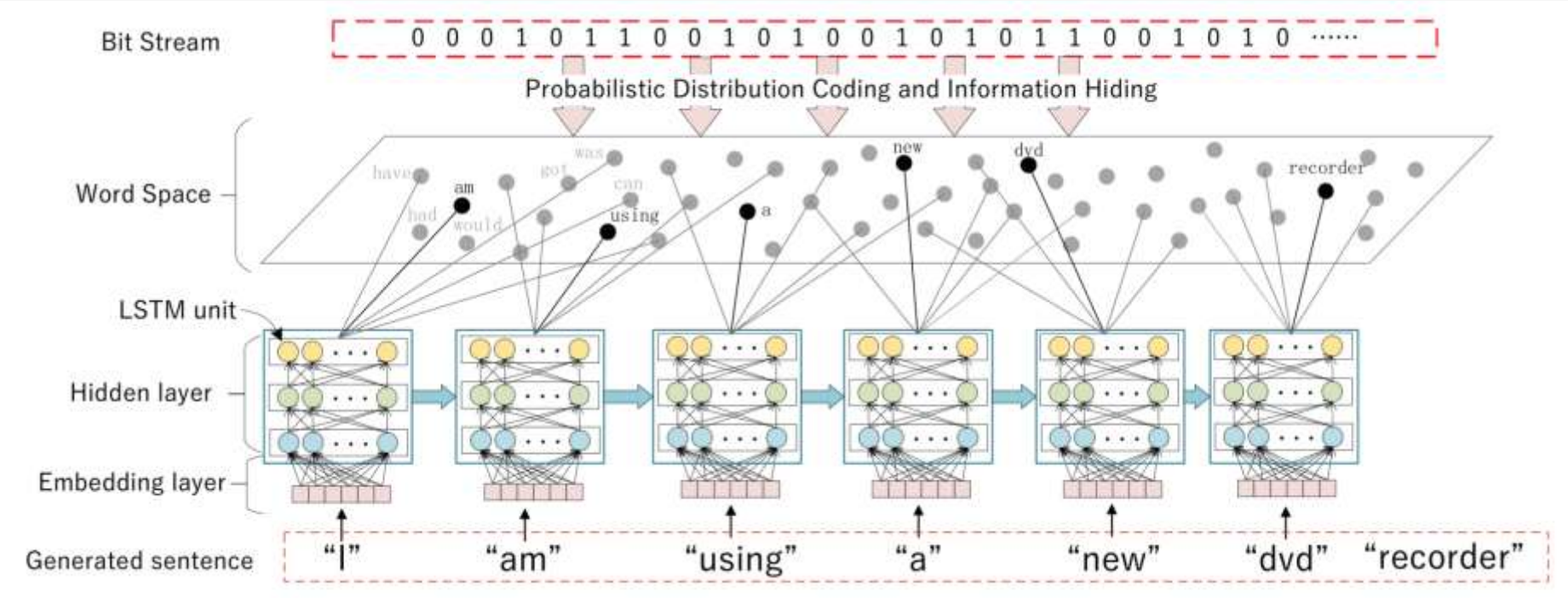
大模型文本隐写

- 如下图为例，假设我要嵌入被转换为的二进制的秘密信息，我可以**将每次模型返回的**next token**的前n个编码**，如当我输入“I”之后，模型返回的前8个token按概率大小顺序排列**分别是**have(000)、am(001)、was(010)、can(011)....**每个token都能对应一个3位二进制的值，当我要隐写嵌入“001”时，我选择“am”作为**next token**，后面循环继续。**



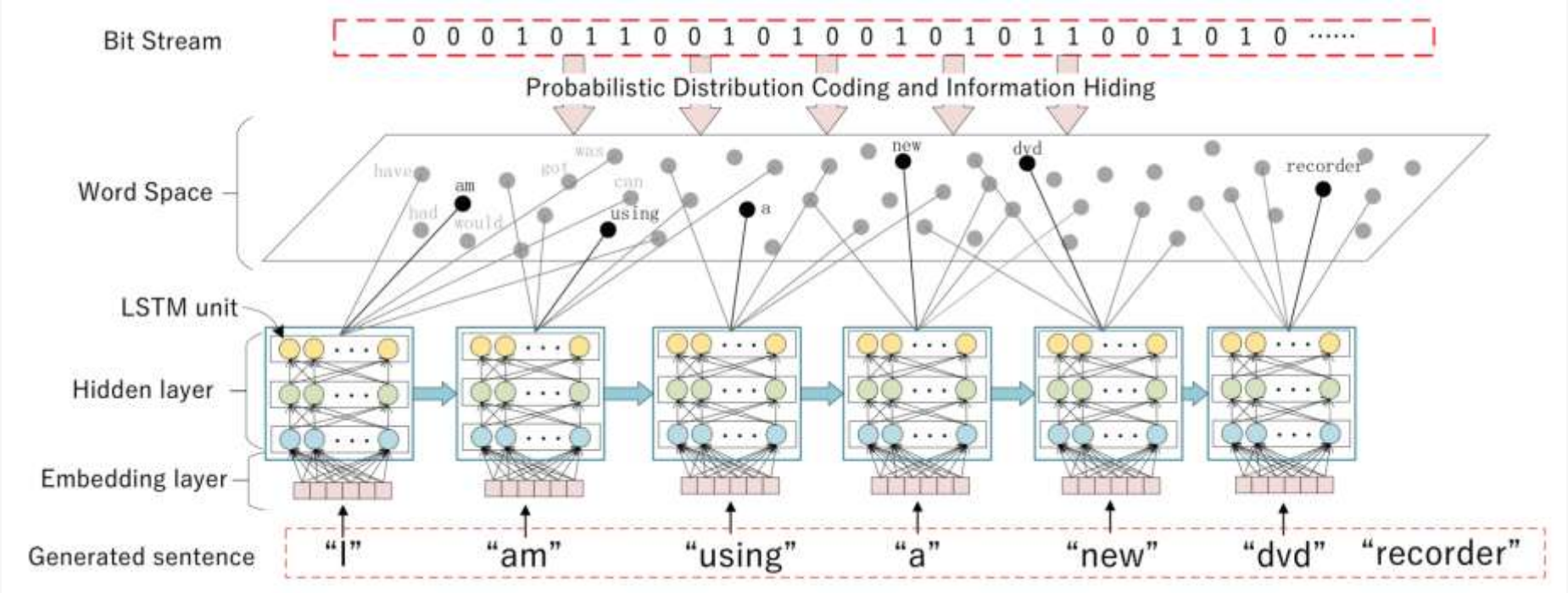
大模型文本隐写

- 这里的本质上方法就是对token候选词的前n个进行编码，构造一个隐式的隐写空间，而当我想要解码信息的时候，**使用相同的模型、上文、参数，那么可以保证每次模型会返回一个跟生成时完全相同的编码空间，从而反向从token来获取到原来的二进制信息流**，比如从“am”反向解码得到001。



大模型文本隐写

- 当然前述方法使用了非常固定的编码方法（直接按顺序对前n个token编码），**下图实际上是使用霍夫曼编码方式，从而确保尽可能嵌入更多信息，提升嵌入率。**



大模型隐写的意义和未来

- 大模型隐写：使用大模型进行隐写的优势是能利用大模型的性能，输出更自然、更难被发现的隐写内容。
- 编码空间：**隐写的不同方法很依赖编码方式，如果有新的构造编码空间的方法，就可以创造出一门新的隐写术。**
- 交叉扩展：**大模型生成式文本隐写的本质是"序列化数据的循环编码"**，这种技术也可以迁移到其他媒介，比如DNA也是序列化数据，可以使用仪器配合大模型输出对DNA编码进行隐写，这样可以实现军事上"把隐藏信息藏在一瓶水、一些细胞、一些毛发上"，从而实现极高水平的隐藏信息传递。
- 大模型水印：现有**大模型水印技术也可以与隐写互相结合和推进**。比如著名的KGW水印也是在推理过程中对next token的选择进行约束。
- 内容安全：隐写的反制技术"隐写分析（Steganalysis）"可以用于对网络媒体内容进行内容评估和筛选检测。**不触发关键词的"阴阳怪气"、"网暴骂人"本质上也是一种"隐写"，因此也有被人工智能使用隐写检测的方法自动侦测到的可能性**，这将有利于未来人类网络空间安全的维护。