基于不变特征的鲁棒多比特自然语言水印

KiYoon Yoo[1] Wonhyuk Ahn[2] JihoJang[1] Nojun Kwak[1*]首尔国立大学 [2]
Webtoon AI {961230,geographic,nojunk}@snu.ac.kr
whahnize@gmail.com

arXiv:2305.01904v2 [cs.CL] 9 Jun 2023

摘要

近年来，订阅制媒体平台、网络小说平台以及大型语言模型的输出中，出现了大量有价值的原文自然语言内容。然而，若缺乏适当的安全措施，这些内容容易遭受非法盗版和潜在滥用。因此，需要一种安全的水印系统，通过泄漏追踪或所有权识别来保障版权保护。为了有效打击盗版并保护版权，多比特水印框架应能够嵌入足量的信息比特并在可能遭受破坏的情况下鲁棒地提取水印。在本研究中，我们借鉴图像水印中的一个著名命题，探索提升载荷和鲁棒性的方法，并找出自然语言中不受轻微破坏影响的不变特征。通过系统分析可能的错误来源，我们进一步提出了一种抗破坏填充模型。我们提出的完整方法在四个数据集、三种扰动类型和两种篡改比例下，平均鲁棒性较先前工作提升了 +16.8 个百分点。[1]

1 引言

近年来，原创且有价值的自然语言内容大量涌现，例如订阅制媒体平台（如金融时报、Medium）、网络小说平台（如Wattpad、Radish）——这一行业发展迅速，尤其是在东亚市场（HanSol，2022；Zeyi，2021年）——以及类人语言模型生成的文本（Ope-nAI，2022；Chiang 等人，2023年；Taori 等人，2023年）。然而，如果没有适当的安全措施，这些内容很容易遭受非法盗版和传播，从而给其创作者造成经济损失内容与市场产业。此外，近期 ChatGPT 等类人语言模型的出现，引发了人们对虚假信息大规模生成的担忧（Goldstein etal.，2023年）。因此，需要一种安全的水印系统，以确保版权保护或检测语言模型的滥用行为。

数字水印是一种能够以不易察觉的方式将信息嵌入多媒体（例如图像、视频和音频）中，同时不降低内容原有效用的技术。通过嵌入所有者/购买者 ID 等信息，其应用包括泄漏追踪、所有权识别、元数据绑定和防篡改。为了有效应对对手的蓄意规避或无意造成的数字退化，水印框架不仅应能够嵌入足够的比特信息，还应表现出抵抗潜在破坏的鲁棒性（Tao 等人，2014年；Zhu 等人，2018）。在深度学习出现之前，针对图像和视频内容的水印研究已经得到了广泛探索（Hsu 和 Wu，1999；Wolfgang 等人，1999；Wang 等人，2001年）。随着深度神经网络的兴起，深度水印作为一种新范式应运而生，改善了水印的三个关键方面：载荷（即嵌入的比特数量）、鲁棒性（即提取出的消息的准确率）以及嵌入媒体的质量。

自然语言水印通过以文本作为水印载体，对语义和/或句法特征进行不可察觉的修改。与改变视觉外观（Rizzo 等人，2019年）不同，这种修改使自然语言水印能够抵御基于人工转录的盗用。以往研究主要关注使用预定义规则和词典进行词汇替换，或进行结构变换等技术（Topkara 等人，2006a，b；Atallah 等人，2001年）。通过利用神经网络，

[1]智能与信息系，融合科学与技术研究生院。
https://github.com/bangawayoo/nlp-watermarking

近期研究要么用基于学习的方法替代预定义规则集（Ab- delnabi 和 Fritz，2021年，AWT），从而取消启发式方法；要么大幅提升词汇替换的质量（Yang 等人，2022，ContextLS）。然而，尽管近期研究优于传统方法，但仍存在局限：AWT 容易在消息提取过程中出错，尤其是在嵌入更多比特时；由于完全依赖神经网络，它偶尔会生成质量下降的水印样本；ContextLS 的载荷存在固定上限，更重要的是，它未考虑在遭受破坏时提取比特消息，因此鲁棒性较低。本研究致力于提升自然语言水印的载荷和鲁棒性。

为了构建有效且鲁棒的自然语言水印系统，我们从一项经典图像水印研究（Cox 等人，1997年）中的著名命题获得灵感：水印应“明确嵌入图像中感知上最重要的组成部分”。如果实现这一点，对手就必须破坏内容的基本结构才能摧毁水印。这会降低原文的效用，使盗版行为失去意义。

然而，由于图像本身具有感知冗余，只有图像才能将水印直接嵌入“感知上最重要的组成部分”。也就是说，相较于修改单个词语，修改单个像素要隐蔽得多。因此，尽管我们遵循该命题的核心思想，却不直接在最重要的组成部分中嵌入水印。相反，我们识别文本中在语义或句法上具有根本意义、因而不会因文本细微修改而改变的特征，并将它们用作锚点，以精确定位水印。提出自然语言鲁棒水印的通用框架后，我们通过实证研究，考察了从语义和句法成分中提取的各种潜在不变特征的有效性。通过逐步分析水印提取过程中可能出现的错误来源，我们进一步提出了一种抗破坏填充模型，该模型经过专门训练，能够抵御各种可能类型的破坏。

我们在涵盖四种不同写作风格的四个数据集上进行的实验结果，证明了以下方法的鲁棒性：（1）依靠不变特征进行水印嵌入；（2）使用经过鲁棒训练的填充模型。与先前工作相比，我们完整方法的鲁棒性平均绝对提升了 +16.8 个百分点，评估范围涵盖四个数据集、三种扰动类型和两种扰动比例。

2 预备知识

2.1 水印问题的形式化

在水印过程中，发送方将秘密消息 m 嵌入载体文本 X，生成带水印文本 $X_{\text{wm}} = \text{EMBED}(X, m)$。载体文本是需要保护的原始文档。消息例如可以是以比特表示的文档购买者或所有者的 ID。接收方[2] 尝试从 $\tilde{X}_{\text{wm}} = \text{CORRUPT}(X_{\text{wm}})$ 中提取嵌入的消息 $\hat{m} = \text{EXTRACT}(X_{\text{wm}})$；该文本可能受到对抗方的蓄意篡改，也可能在分发过程中发生自然退化（例如拼写错误）。我们关注盲水印，即无法访问原始载体文本的水印方式。发送方和接收方的主要目标是：（1）使 X_{wm} 在语义上尽可能与 X 相似，以免降低原始内容的效用；（2）设计嵌入 和提取 函数，使提取出的消息准确无误。

2.2 对 X_{wm} 的篡改

相反，对手会尝试通过破坏带水印文本来干扰消息提取阶段，同时维持文本原有的效用。例如，非法盗版方会希望避免水印被用于追踪泄漏点，同时仍希望保留文本以供非法传播。这限制了对手在数量和质量上对文本进行过度破坏。为此，我们借鉴对抗性攻击（Jin 等人，2020；Morris 等人，2020a）中的技术来修改文本并保持其原始语义。

我们考虑词语插入（Li 等人，2021年）、删除（Feng 等人，2018）和替换（Garg 和 Ramakrishnan，2020），破坏比例为每个句子中词语数量的 2.5% 至 5.0%，遵循 Abdelnabi 和 Fritz（2021年）。插入/替换/删除的词语数量等于 $\text{ROUND}(CR \times N)$，其中 CR 是破坏

[2]与发送方和接收方是不同角色的表述相反，双方也可能是同一方。

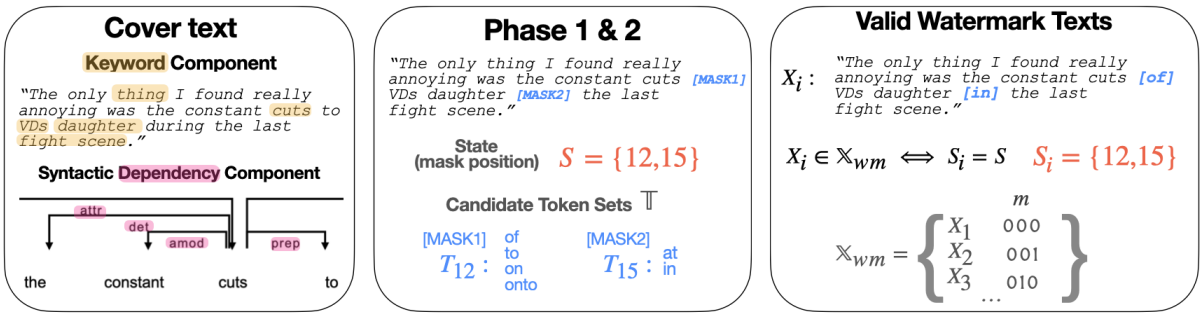


图1：最左侧展示了载体文本及其关键词和句法依存成分的示例（受篇幅限制，仅部分展示）；中间展示阶段 1 和阶段 2；最右侧展示了有效水印样本的示例。

比率以及 N 是句子中的词语数量。这可确保包含很少或没有破坏空间的较短句子不会遭到严重损坏。为进一步限制破坏后的文本偏离原文，我们使用预训练的句子转换器[3] *all-MiniLM-L6-v2*。该模型在由 10 亿对句子组成的多个数据集上训练，用于过滤与原文余弦相似度低于 0.98 的破坏文本。

2.3 填充模型

与 ContextLS（Yang 等人，2022）类似，我们使用预训练的填充模型来生成带水印数据集的候选项。给定一个掩码序列$X_{\setminus i} = \{x_1, \cdots, x_{i-1}$, 掩码 x_{i+1}, $\cdots$, $x_t\}$，填充语言模型可以预测适合填入掩码的词语。由 θ 参数化的填充模型会输出x_i 在词汇表（v）上的概率分布：

$$P(X_{\setminus i}|\theta) = p_i \in \mathbb{R}_+^{|v|}. \tag{1}$$

我们将填充模型输出的前 k 个候选词元集合记作

$$\{t_1^i, \cdots, t_k^i\} = \text{INFILL}(X_{\setminus i}; k). \tag{2}$$

3 鲁棒自然语言水印框架

我们用于自然语言水印的框架由两个阶段组成。阶段 1 是使用某个函数 g_1 从文本 X（或 X_{wm}）中获取状态 S，即 $\tilde{}$。S 可以被视为从文本中抽象出的特征，该特征包含足够的信息来确定嵌入过程。阶段 2 包含一个函数 g_2 该函数接收 X 和 S 作为输入，以生成有效的带水印文本。我们依赖掩码填充模型来生成带水印文本，这使得 S 能够确定掩码的位置。填充模型根据比特消息生成带水印文本 X_{wm}。总体概述如图1所示。

3.1 阶段1：掩码位置选择

为使水印系统能够抵御破坏，应选择 S，使其依赖于文本中相对不受破坏影响的属性。也就是说，S 应是文本不变特征的函数。更具体地说，理想的不变特征具有以下特征：

1. 必须修改文本中相当大的一部分，才能改变该特征。2. 因此，对于那些保留原文效用（例如语义和细微差别）的篡改，该特征是不变的。

根据构造，当 S 是理想不变特征的函数时，对于 X 和 X_{wm}，都能恢复相同的 $\tilde{}$ 状态 S，从而增强水印的鲁棒性。本质上，我们试图找出应该屏蔽哪些词语，才能使水印具有鲁棒性。

给定状态函数 $g_1(\cdot)$，令 $S = g_1(X)$，$\tilde{S} = g_1(X_{\text{wm}})$。那么，我们将 g_1 的**鲁棒性**定义如下：

$$\mathcal{R}_{g_1} := \mathbb{E}[\mathbb{1}(S = \tilde{S})]. \tag{3}$$

其中，$\mathbb{1}$ 表示指示函数，$\mathbb{E}$ 表示期望运算。

我们旨在发现自然语言中两个容易获取的领域——语义和句法成分中的不变特征。这些成分的示意图如图1 左侧。

[3] https://www.sbert.net/

鲁棒性	相关性类型	ContextLS (Yang 等人, 2022)	关键词	句法
$\mathcal{R}_{g_1}$	D	0.656	0.944	0.921
	I	0.608	0.955	0.959
	S	0.646	0.974	0.949

表 1：鲁棒性 g_1（$\mathcal{R}_{g_1}$），针对 IMDB 上 5% 扰动率下的三种扰动类型：删除（D）、插入（I）和替换（S），比较 ContextLS 和我们的方法（关键词、句法）的表现。完整结果见附录表 9 。

关键词组件在语义层面，我们首先确定为保持原文效用而应保留的关键词。我们的直觉是，关键词是句子中语义上至关重要的部分，因此即使受到破坏也会被保留且保持不变。这包括专有名词，因为在不改变语义的情况下，它们通常无法用同义词替换（例如电影、人物或地区的名称）；我们可以通过现成的命名实体识别模型提取这些专有名词。此外，我们使用一种名为 YAKE 的无监督方法（Campos 等人，2018），它能够输出语义上至关重要的词语。提取关键词后，我们将其用作锚点，并通过简单的启发式方法确定掩码的位置。例如，可以选择与关键词相邻的词作为掩码。

句法依存组件 在句法层面，我们使用现成的句法分析器构建依存句法树。依存句法分析器通过在中心词与其从属词之间构建有向边，来描述句子的句法结构。每个从属词都会根据其语法角色被标记为特定的依存关系类型。我们假设，句法分析树生成的整体语法结构对句子中的轻微破坏相对鲁棒。为选择应掩码的依存关系类型，我们构建了一个预定义顺序，以保持带水印句子的语义。具体而言，我们使用填充模型对每种依存关系类型进行掩码和替换，并在单独的留出数据集上比较其由自然语言推理模型（例如 RoBERTa-Large-NLI[4]）计算出的蕴含分数，从而构建该顺序，如算法 1 所示（更详细的流程和完整列表见附录 A.4）。利用生成的顺序，我们逐一掩码依存关系，直到达到目标掩码数量。对于这两种组件类型

[4]https://huggingface.co/roberta-large-mnli

算法 1： 句法依存排序
基于自然语言推理蕴含分数。

输入： 句子 X
输出： 排序列表 L
/* 查找 $x \in X$ 中每个词的依存关系
使用 spaCy */
x.依存句法 $\leftarrow$ spaCy(X, x)
/* 按依存类型初始化列表字典
依存类型 */
$D[x$.依存关系$]$: $[]$
$N \leftarrow$len(X)
/* 遍历词语并填充 */
对于 $i \leftarrow 0$ 至 N **执行**
  $X' \leftarrow$填充$(X_{\setminus i})$
  $s \leftarrow$ 自然语言推理(X', X)
  $D[x$.dep$]$.append(s)
对于v $\in D$.values() **执行**
  v $\leftarrow$v。均值()
$L =$
sorted([k for k,v in D.items()],
  key=lambda x:x[1])
return $L[:: -1]$

（语义和句法），我们确保关键词不会被掩码。

那么，前述组件抵御破坏的效果如何？表格中的结果1 支持了我们的假设：关键词和句法成分确实可以充当不变特征，因为根据掩码匹配样本的比例衡量，它们在三种不同类型的破坏下都表现出相当高的鲁棒性。相比之下，不依赖任何不变特征的 ContexLS（Yang 等人，2022）的 $\mathcal{R}_{g_1}$则低得多。这表明，破坏导致被掩码的词语发生变化，从而妨碍了水印提取过程。

3.2 阶段2：水印编码

在阶段 2 中，通过 $g_2(X, S)$ 生成一组有效的带水印文本，用于嵌入或提取消息。对于我们的方法，由于状态是掩码位置的集合，因此需要使用填充模型选择排名前 k的词语，并按字母顺序排序，以生成一组有效的水印。具体而言，使用 §2.3 中的符号， $g_2(X, S)$ 可分为以下步骤：

(1) $\mathcal{T}_i = \{t_1^i, \cdots, t_k^i\} = \textsc{infill}(X_{\setminus i}; k_1), \forall i \in S$

（2）过滤 $\mathcal{T}_i$ ，移除所有标点符号、子词和停用词。通过选择排名前 k_2 （$\leq k_1$）的词语并按字母顺序排序来更新 $\mathcal{T}_i$ 。

（3）对词元集合$\mathbb{T} = \mathcal{T}_{s_1} \times \cdots \times \mathcal{T}_{s_j}$取笛卡尔积，其中 $j = |S|$。令 $\mathbb{X}$ 为

将对应词元替换后的文本集合（$|\mathbb{X}| = |\mathbb{T}|$）。

(4) 生成有效的水印集合 $\mathbb{X}_{\text{wm}} = \{X_i \in \mathbb{X} | g_1(X_{wm}) = g_1(X_i)\} \subseteq \mathbb{X}$，并为集合中的每个元素分配一条比特消息$\mathbb{X}_{\text{wm}}$。

在（4）中，生成一组有效的水印意味着确保消息比特能够被无误提取。为此，只保留那些来自 $\mathbb{X}$ 且具有与 X相同状态的带水印文本（图1 中间和右侧）。在零损坏~（当 X_{wm}=X_{wm}时），阶段 2 将生成相同的 ~ 带水印文本集合，前提是 S 和 S 等价~ ~（即 $g_2(X, S) = g_2(X_{\text{wm}}, S)$）。因此，在没有任何破坏时，我们的方法能够无误地提取水印。

然而，带水印文本遭到破坏时会发生什么？即使恢复了确切的状态，也可能无法恢复相同的带水印文本集合，因为填充模型依赖局部上下文来填充掩码。考虑到这一点，我们还可以将 g_2 的**鲁棒性**定义为

$$\mathcal{R}_{g_2} := \mathbb{E}[\mathbb{1}(g_2(X, S) = g_2(\tilde{X}_{\text{wm}}, \tilde{S}))]. \quad (4)$$

图2 右侧展示了 $\mathcal{R}_{g_1}$ 以及 $\mathcal{R}_{g_1}$ 与 $\mathcal{R}_{g_2}$之间的差异。我们观察到，当我们选择原始预训练语言模型（如 BERT）作为填充模型时，$\mathcal{R}_{g_2}$ 显著低于 $\mathcal{R}_{g_1}$ 我们的方法。虽然不变特征的类型确实会影响 $\mathcal{R}_{g_2}$，但我们得到的关键结论是 $\mathcal{R}_{g_2}$ 在所有情况下都显著低于 $\mathcal{R}_{g_1}$ [5]。

有趣的是，对于 ContextLS，$\mathcal{R}_{g_1}$ 与 $\mathcal{R}_{g_2}$ 之间的差距几乎为零，这表明阶段 1 已经是实现鲁棒性的瓶颈。较小的差距可以归因于较小的 top-k_2(=2)以及增量式水印方案，该方案逐步增加要进行填充的序列长度。这可能降低破坏词影响填充模型的可能性。

3.3 鲁棒填充模型

为了彻底解决阶段 2 的脆弱性，我们通过微调 θ，构建一个能够抵御潜在篡改的鲁棒填充模型，使其在给定 $X_{\setminus i}$ 和 $X_{\setminus i}$（即 $X_{\setminus i}$ 的受损版本）的情况下，输出一致的词分布。这可以通过最小化.....的散度来实现。

[5]较大的 $\mathcal{R}_{g_2}$ 并不一定意味着更低的比特错误率，因为该指标并未衡量 $g_2(X, S)$ 与 $g_2(\tilde{X}_{\text{wm}}, \tilde{S})$ 之间差异的程度。

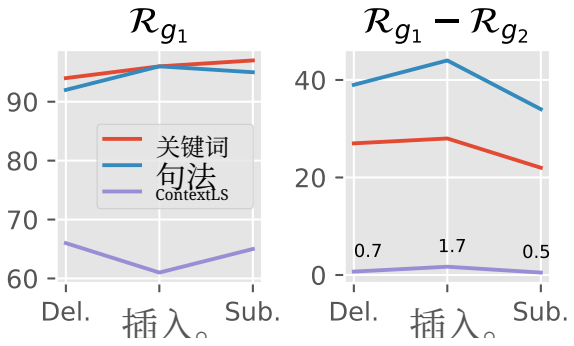


图 2：在 IMDB 数据集上 5% 扰动率下，g_1 的鲁棒性以及 g_1 与 g_2 之间鲁棒性的差异。

数据集	$\Delta\mathcal{R}_{g_1}$	$\Delta\mathcal{R}_{g_2}$
D1	.005±.004	.113±.013
D2	.009±.007	.070±.024
D3	.0±.002	.142±.051
D4	.0±.002	.151±.048

表2：应用鲁棒填充模型对阶段1和2的鲁棒性的影响（应用与未应用之差），在三种扰动类型上的平均值，四舍五入至小数点后三位。四个数据集（D1-D4）分别为 IMDB、Wikitext-2、德古拉和《呼啸山庄》。有关数据集的更多详细信息，请参见 §4。

这两个分布 p_i 和 $\tilde{p}_i$ 其中 $\tilde{p}_i$ 指的是~受扰序列的词分布，$X_{\setminus i}$。我们不再使用原始词分布作为目标分布（它在 $>$ 30,000 个词元（对于 BERT-base）上分布密集），而是通过将其余词元置零，并在前 k_1 个词元上进行归一化，构建一个稀疏的目标分布。这是因为在我们的水印框架中仅使用前 k_1个词元（见 §3.2）。

此外，为了改善训练动态，我们遵循 §3.1 中提出的掩码策略来选择要掩码的词语，而不是采用原始预训练阶段使用的随机掩码策略。这样可以使训练时和测试时被掩码词语的分布保持一致，从而在相同计算时间下获得更好的性能（鲁棒性）。与此相反，原始掩码策略会随机选择一定比例的词语进行掩码，这会为填充模型提供较弱的学习信号。

我们使用Kullback–Leibler（KL）散度作为评估指标。更具体地说，我们将“反向KL”用作损失项，其中预测分

布（而非目标分布）用于对数分布差异进行加权，这与变分贝叶斯中的做法一致（Kingma 和 Welling，2014年）。这有助于防止模型输出“零强制”预测分布。两个分布之间的一致性损失定义为

$$\mathcal{L}_{con} = \sum_{i \in S} \mathrm{KL}(\tilde{p}_i | p_i), \tag{5}$$

$$\text{where} \quad \tilde{p}_i = P(\tilde{X}_{\setminus i} | \theta), \tag{6}$$

$$p_i = P(X_{\setminus i} | \textsc{freeze}(\theta)) \tag{7}$$

所有 i 掩码词元。图的输出p 会被分离，用于训练一个模型，使其在接收到受破坏的输入时输出一致的结果。正如我们所预期的，将鲁棒填充模型用于句法组件，能显著提升 $\mathcal{R}_{g_2}$，而对 $\mathcal{R}_{g_1}$ 的提升则可以忽略不计（表2）。

受破坏的输入按照与§2.2 中相同的策略生成，使用的是一个单独的训练数据集。我们在§5.3中对设计选择进行了消融实验。

总而言之，所提出的框架

1. 在不存在破坏时，能够无误地嵌入和提取水印。2. 能够将不变特征用于水印嵌入，从而在存在破坏时实现鲁棒性。3. 在阶段 2 中进一步利用鲁棒填充模型增强鲁棒性。

		IMDB			
指标		方法 ContextLS	关键词	句法	+RI
每词比特数 (↑)		0.100	0.116	0.125	**0.144**
比特错误率(↓) @压缩率=0.025	D	0.219	0.127	0.100	**0.074**
	I	0.303	0.153	0.153	**0.106**
	S	0.273	0.142	0.133	**0.110**
比特错误率(↓) @压缩率=0.05	D	0.392	0.252	0.277	**0.200**
	I	0.355	0.201	0.242	**0.163**
	S	0.343	0.218	0.220	**0.177**
		Wikitext-2			
指标		方法 AWT	ContextLS	关键词句法	+RI
每词比特数 (↑)		0.100	0.083	0.092	0.090	**0.136**
比特错误率(↓)@压缩率=0		0.264	0.0	0.	0.	0.
比特错误率(↓) @CR=0.025	D	0.273	0.224	0.202	0.162	**0.136**
	I	0.272	0.289	0.222	0.216	**0.205**
	S	0.279	0.266	0.176	0.155	**0.157**
比特错误率(↓) @CR=0.05	D	0.284	0.410	0.326	0.321	**0.282**
	I	0.272	0.338	0.246	0.235	**0.201**
	S	0.289	0.342	0.256	0.228	**0.201**
		德古拉			
每词比特数 (↑)		0.100	0.089	0.126	0.117	**0.146**
比特错误率(↓)@压缩率=0		0.111	0.	0.	0.	0.
BER(↓) @压缩率=0.025	D	0.236	0.201	0.116	0.076	**0.030**
	I	0.218	0.299	0.181	0.133	**0.063**
	S	0.231	0.272	0.140	0.130	**0.081**
BER(↓) @压缩率=0.05	D	0.286	0.373	0.255	0.248	**0.177**
	I	0.264	0.375	0.228	0.279	**0.155**
	S	0.281	0.337	0.207	0.229	**0.164**
		《呼啸山庄》			
每词比特数（↑）		0.100	0.076	0.088	0.097	**0.114**
比特错误率(↓)@压缩率=0		0.100	0.	0.	0.	0.
比特错误率(↓) @压缩率=0.025	D	0.224	0.194	0.102	0.088	**0.063**
	I	0.212	0.284	0.144	0.132	**0.068**
	S	0.224	0.271	0.161	0.143	**0.096**
比特错误率(↓) @压缩率=0.05	D	0.283	0.379	0.253	0.240	**0.169**
	I	0.258	0.363	0.224	0.268	**0.133**
	S	0.276	0.363	0.231	0.245	**0.161**

表 3：四个数据集上的载荷与鲁棒性对比。 +RI 表示在我们的句法组件中加入鲁棒填充模型。**Top-1** 数值以粗体显示。

4 实验

数据集 为评估所提出方法的有效性，我们使用了四种风格各异的数据集。IMDB（Maas 等人，2011）是一个电影评论数据集，语言风格更加口语化。WikiText-2（Merity 等人，2016）由维基百科文章组成，风格更具信息性。我们还使用了两部小说进行实验：《德古拉》和《呼啸山庄》（WH）。这两部作品的风格与现代英语明显不同，且均可从古腾堡计划获取（Bram，1897；Emily，1847）。

指标 对于载荷，我们计算每词比特数（BPW）。对于鲁棒性，我们计算提取出的消息的比特错误率（BER）。我们还通过将带水印文本与原始封面文本进行比较，来衡量其质量。遵循Yang 等人 (2022);Abdelnabi 和 Fritz (2021年)的方法，我们使用自然语言推理模型（RoBERTa-Large-NLI）计算蕴含分数（ES），并通过比较预训练模型输出表示的余弦相似度来计算语义相似度（SS），该预训练模型为 sentencetransformer（stsb-RoBERTa-base-v2）。我们还开展了人工评估，以评估语义质量。

实现细节 对于我们的方法和 Con-textLS（Yang 等人，2022），二者都针对单个句子进行操作，因此我们使用最小的现成模型（*en-core-web-sm*）来自 spaCy （ Honnibal 和 Montani，2017 ）来拆分句子。同一个 spaCy 模型也用于 NER（命名实体识别器）以及为我们的方法构建依存关系句法分析器。两种方法都使用 BERT-base 作为填充模型，并选择排名前 32 的（k_1）词元。我们通过控制每个句子中的掩码数量（$|S|$）和排名前k_2 个词元（§3.2），将我们的载荷设置为与对比方法相近的程度；每个数据集的这些配置见附录表12。对于每个数据集，我们为前 5,000 个句子添加水印，并使用 TextAttack（Morris

等人，2020b）创建损坏样本。为了实现鲁棒填充，我们在各个数据集上对 BERT 进行 100 个训练轮次的微调。更多详情请参阅附录。

对比方法 我们在实验中将我们的方法与基于深度学习的方法（Abdelnabi 和 Fritz，2021，AWT）（Yang 等人，2022，ContextLS）进行比较，因为作为深度学习出现之前的方法（Top- kara 等人，2006b；Hao 等人，2018）的纯规则方法具有较低的载荷和/或较低的语义质量（稍后将在表 4 中展示）。有关所比较方法的更多详细信息，请参见 §6。

4.1 主要实验

表 3 展示了在全部四个数据集上的水印结果。附录 A.2详细介绍了我们在训练 AWT 时遇到的一些挑战以及克服这些挑战的方法。由于 AWT 在 IDMB 上的损失未能收敛（详见附录 A.3），因此我们省略了相关结果。

我们在 2.5% 和 5% 的破坏率（CR）下测试了各方法的鲁棒性。对于我们的方法，我们对句法依存组件应用鲁棒填充，最终一列中的 +RI 表示这一点。AWT 受较高破坏率的影响较小，有时在不使用 RI 时优于我们的方法。然而，零破坏率下的 BER 不可忽略，而这对可靠的水印系统至关重要。此外，我们通过定性观察发现，AWT 经常在水印集合中重复词语或替换代词，这似乎为提取消息提供了信号——但代价是质量严重下降。附录 A.7 和表 17-19 中展示了一些示例。

我们的最终模型在所有数据集和扰动率下都大幅优于 ContextLS。此外，在所有数据集中，即使不使用鲁棒填充，语义成分和句法成分也都比 ContextLS 鲁棒得多。在 CR=2.5% 下，相对于 ContextLS，使用句法成分在各种扰动类型下的比特错误率绝对改善幅度，在四个数据集上分别为 13.6、8.2、14.4 和 12.9 个百分点；在 CR=5% 下，分别为 10.0、10.2、11.0 和 11.7 个百分点。

4.2 水印的语义分数

表4 展示了语义指标的结果。虽然我们的方法落后于 ContextLS，但其语义得分优于所有其他方法，同时实现了鲁棒性。ContextLS 通过明确使用自然语言推理模型筛除候选词元，从而能够保持较高的语义相似度。然而，如前一节所示，在存在破坏的情况下，提取出的消息的准确率会严重下降。使用按照蕴含分数排序的有序依存关系，相较于使用随机排序的依存关系，能够显著提高语义指标，后者记为“– 自然语言推理排序”。结果见附录表15。

		[1]	[2]	AWT	ContextLS	Ours
IMDB	ES	0.843	0.867	0.958	0.985	0.975
	SS	0.916	0.943	0.973	0.982	0.981
Wikitext-2	ES	0.888	0.907	0.935	0.986	0.966
	SS	0.941	0.945	0.991	0.989	0.993
德古拉	嵌入成功率	0.869	0.915	0.869	0.985	0.963
	SS	0.910	0.889	0.855	0.986	0.971
WH	ES	0.882	0.893	0.947	0.984	0.964
	语义相似度	0.929	0.934	0.968	0.989	0.975

表 4： [1]： Topkara 等人 (2006b)， [2]： Hao 等人(2018)。水印集合相对于原始载体文本的语义分数（ES：蕴含分数，SS：语义相似度）。除我们的方法外，所有数值均引自Yang 等人 (2022)

指标	AWT	ContextLS	Ours
流畅度Δ(↓)	1.32±0.7	0.25±0.4	0.26±0.4
语义相似度(↑)	2.97±0.8	4.22±0.5	3.90±0.8

表 5：李克特量表上的人工评估结果（20 个样本和 5 名标注者）。

我们还进行了人工评估，在表 Δ5 中比较了带水印文本与载体文本的流畅度（流畅度），以及与原始封面文本相比语义的保持程度（语义相似度；SS）。实验详情见附录 A.6。这与我们在自动指标中的发现一致，但显示出我们的方法与 AWT 之间存在明显差距。值得注意的是，与原始封面文本相比，我们的方法和 ContextLS 的流畅度变化程度几乎相同。

5 讨论

5.1 与 ContextLS 的比较

我们与 ContextLS 不同的一项设计选择是 top-k_2 $>$ 2，它决定了每个掩码对应的候选词元数量。通过选择更高的 k_2，我们可以根据需求增加载荷。然而，对于 ContextLS，增加 k_2 却会违反直觉地导致载荷 <escape></*escape*>更低<*escape*></escape>。这是因为 ContextLS 会确定有效的水印集合（即能够在不发生错误的情况下提取消息的集合），而这些集合.....

	top-k_2	2	3	4
BPW	ContextLS	0.100	0.033	0.021
	Ours	0.100	0.161	0.211
前向传播	ContextLS	1994	2386	2801
	Ours	94	94	94

表6：top-k_2 对载荷、填充模型的前向传播次数以及 Con-textLS 和我们的方法在 IMDB 上的墙钟时间的影响。我们将关键词比例固定为 0.11。

协调

科幻电影/电视节目通常资金不足，也不受重视**并且**[也不] 被误解。（嵌入成功率=0，语义相似度=0）

我觉得主要反派塑造得相当不错**而且**[不过] 演技也相当出色。（嵌入成功率=0.994，语义相似度=0.994）

命名实体

这部电影之所以没有被评为 1 分（糟糕），唯一的原因是**艾达**[艾达] 卢皮诺和**罗伯特**[罗伯特] 瑞安的演技十分出色。（嵌入成功率=0.993，语义相似度=0.961）

我没看过**"犯罪**[罪犯] 医生"系列的其他电影，所以无法进行任何比较。（嵌入成功率=0.994，语义相似度=0.990）

表7：载体文本与带水印文本之间的蕴含分数。此处展示了**原文**[带水印] 词语。

ror），并采用更强的约束（详情见Yang 等人 (2022）论文中的公式 5、6、7）。这还需要使用逐步增大的窗口对整个句子进行穷举搜索，由于神经网络需要多次前向传播，嵌入/提取时间因此大幅延长。例如，在 IMDB 的 1000 个句子中进行嵌入，耗时是 ContextLS 上的 20 多倍（81 分钟，而后者为 4 分钟）。更多结果汇总于表 6。将我们的鲁棒填充模型应用于 ContextLS 的结果见附录 A.4。

5.2 自动语义指标的陷阱

尽管自动语义指标确实提供了有意义的信号，有助于保留原始语义，但它们并未呈现全貌。首先，替换并列依存关系（例如 and、or、nor、but、yet）时，这些分数无法准确反映语义变化。如表 7 所示，蕴含分数和语义相似度分数都忽略了一些人类很容易察觉的语义变化。这也体现在我们在 §3.1 中构建的依存关系列表排序上——填充并列依存关系后，平均自然语言推理分数为 0.974，排名第二。一个简单的解决办法是将其放在最末位，或直接舍弃它。我们可以通过

		Ran. 掩码（FKL）	运行。掩码（RKL）	Ours
每词比特数 (↑)		0.121	0.129	0.144
比特错误率(↓) @压缩率=0.025	D	0.106	0.101	0.074
	I	0.141	0.139	0.106
	S	0.138	0.137	0.110

表 8：掩码设计选择的消融实验（FKL：正向 KL，RKL：反向 KL）。我们的方法是主要实验中使用的最终版本（我们的掩码策略 +RKL）。

将并列依存关系放在最后一位来解决，或干脆将其舍弃。附录表11 显示，这样做也能获得相当的每词比特数和鲁棒性。

我们观察到自然语言推理模型的另一种异常情况是，当人物或地区等命名实体被掩码时。表7 展示了 ContextLS 中的一个示例，以及嵌入成功率为何异常偏高。如果角色名称被修改，这类水印可能会显著损害小说的效用。我们的方法通过排除命名实体（使用 NER 检测）作为可能的掩码候选项，避免了这个问题。

5.3 消融实验及其他结果

消融实验 本节对若干设计选择进行消融分析。首先，我们在表8中比较了掩码策略的设计选择（随机策略与我们的方法）以及损失项（正向 KL 与反向 KL）。与随机掩码词语相比，我们的掩码策略提升了每词比特数和鲁棒性。尽管初步实验表明，RKL 在载荷较高时更有效且鲁棒性更强，但进一步实验显示，采用我们的掩码策略时，KL 的类型对最终鲁棒性影响不显著。我们还展示了基于字符的破坏下的结果，并在附录 A.4中比较了面对不同扰动类型时的鲁棒性。

句法成分的压力测试 我们测试了所提出的句法成分在更高破坏率下的表现。结果见附录图 3。在压缩率 =0.1 时，面对插入和替换，鲁棒性仍高于 0.9；但面对删除时，鲁棒性会迅速下降。这表明我们的句法成分最容易受到删除的影响。

6 相关工作

自然语言水印通过操纵语义或句法特征来嵌入信息，而不是改变词语、行和文档的视觉外观

（Rizzo 等人，2019年）。这使自然语言水印能够抵御文件重新格式化或文本人工转录（Topkara 等人，2005年）。早期的自然语言水印研究依赖同义词替换（Topkara 等人，2006b）、句法结构重组（Atallah 等人，2001年）或释义改写（Atallah 等人，2003）。依赖预定义规则集往往会导致比特容量较低，而嵌入过程中缺乏对上下文的考量，则可能降低带水印文本的效用，使其听起来不自然或怪异。

随着神经网络的出现，一些研究不再像以往的工作那样依赖预定义的规则集。对抗式水印变换器（Abdelnabi 和 Fritz，2021年，AWT）提出了一种编码器-解码器变换器架构，该架构能够学习从解码后的带水印文本中提取消息。为了保持带水印文本的质量，他们利用句子变换器和语言模型提供的信号。然而，由于消息嵌入和提取完全依赖神经网络，即使不存在破坏，提取出的消息也容易出现错误，尤其是在载荷较高时，并且部分样本中还会出现明显的伪影，例如重复词元。Yang 等人（2022）采用算法方法进行消息嵌入和提取，从而实现无错误处理。此外，与自然语言推理模型结合使用神经填充模型，已被证明在词汇替换方面比更传统的方法（例如 WordNet）具有更好的质量。然而，该方法并未考虑文本遭受破坏时的鲁棒性。

图像水印 显式考虑破坏以增强鲁棒性，并利用多媒体的不同域，这些都与盲图像水印密切相关；盲图像水印已得到广泛研究（Mun 等人，2019年；Zhu 等人，2018；Zhong 等人，2020；Luo 等人，2020）。与我们的鲁棒填充训练类似，Zhu 等人；Luo 等人显式考虑可能发生的图像破坏，以提高鲁棒性。与此同时，使用离散余弦变换等变换方法将像素域转换到各种频域，已被证明既有效又更具鲁棒性（Potdar 等人，2005年）。在我们的工作中，使用关键词和依存关系确定嵌入位置，可以类似地视为分别将原始文本转换到语义域和句法域。

其他研究方向 隐写术是类似的一类研究，旨在将秘密数据隐藏到载体媒体中，侧重于隐蔽性而非鲁棒性。自然语言领域已研究了多种方法（Tina Fang 等人，2017；Yang 等人，2018；Ziegler 等人，2019年；Yang 等人，2020；Ueoka 等人，2021年）。这类研究与水印的不同之处在于，载体文本可以任意生成以隐藏秘密消息，因此无需严格维持原始语义。

最近，He 等人（2022a）提出对语言模型的输出添加水印，以防止模型窃取和提取。尽管这些工作的主要目标（He 等人，2022a，b）与我们的方法不同，但其方法可以调整为直接对文本添加水印。然而，这些方法仅限于零比特水印（例如，判断文本是否来自语言模型），而我们的方法允许嵌入任意多比特信息。同样，Kirchenbauer 等人（2023年）提出在解码时以零比特方式对语言模型的输出添加水印，以区分机器生成文本和人工撰写文本。

7 结论

我们提出利用自然语言的不变特征来嵌入能够抵抗破坏的鲁棒水印。我们通过实证验证了两个可由现成模型轻易发现的潜在组成部分。与近期基于神经网络的水印方法相比，所提出的方法在鲁棒性和载荷方面表现更优，同时具有相当的语义质量。我们并不声称本文研究的不变特征是最优方案。相反，我们为未来的研究探索遵循该框架的其他有效领域和解决方案铺平了道路。

局限性

尽管具有鲁棒性，但与最新研究相比，我们的方法在自动语义指标上的结果较差。这可能是可感知性与鲁棒性之间权衡的自然结果（Tao 等人，2014年；De Vleeschouwer 等人，2002）：更强的水印往往会干扰原文内容。尽管如此，通过采用一些技术技巧（例如神经填充模型、按自然语言推理排序），我们的方法仍然能够优于其他所有方法，包括两种传统方法和一种基于神经网络的方法。

我们采用了对抗性攻击技术来模拟工作中可能出现的破坏。然而，这些自动攻击并不总是会对原文进行不可察觉的修改（Morris 等人，2020a）。因此，我们工作中使用的破坏可能只是对真实对手为规避水印而采取的行动的粗略估计。此外，我们的方法尚未针对改写进行测试，而改写可能会大幅改变文本的句法成分。阻碍我们尝试基于改写的攻击的一个现实原因是，与其他能够对被破坏词语的数量进行细粒度控制的攻击相比，改写缺乏可控性。同样，对于重视细微差别的小说等文本资源而言，上述特性可能会令对手不愿使用改写来破坏水印。

致谢

本研究获得韩国政府通过 IITP 项目 2022-0-00320、2021-0-01343，NRF 项目 2021R1A2C3006659，以及 2022 年 NAVER WEBTOON 的 Webtoon AI 资助。

参考文献

Sahar Abdelnabi 和 Mario Fritz。2021年。对抗性水印 Transformer：通过数据隐藏追踪文本来源。载于 *2021*年 *IEEE* 安全与隐私研讨会（*SP*），第 121–140 页。IEEE。

Mikhail J Atallah、Victor Raskin、Michael Croga n、Christian Hempelmann、Florian Kerschbau m、Dina Mohamed 和 Sanket Naik。2001年。自然语言水印：设计、分析与概念验证实现。发表于国际信息隐藏研讨会，第 185–200 页。施普林格。

Mikhail J Atallah、Victor Raskin、Christian F Hempel- mann、Mercan Karahan、Radu Sion、Umut Topkara 和 Katrina E Triezenberg。2003 年。《自然语言水印与防篡改》。载于国际信息隐藏研讨会，第 196–212 页。施普林格。

Stoker Bram。1897年。《呼啸山庄》。

Ricardo Campos、Vítor Mangaravite、Arian Pasquali、Alípio Mário Jorge、Célia Nunes 和 Adam Jatowt。2018。Yake！与语料集无关的自动关键词提取器。载于欧洲信息检索会议，第 806–810 页。施普林格。

Wei-Lin Chiang、Zhuohan Li、Zi Lin、Ying Sheng、Zhanghao Wu、Hao Zhang、Lianmin Zheng、Siyuan Zhuang、Yonghao Zhuang、Joseph E. Gonzalez、Ion Stoica 和 Eric P. Xing。2023年。Vicuna：一款开源聊天机器人，表现媲美 GPT-4，ChatGPT 质量达 90%*。

Ingemar J Cox、Joe Kilian、F Thomson Leighton 和 Talal Shamoon。1997 年。用于多媒体的安全扩频水印。*IEEE* 图像处理汇刊，6(12):1673–1687。

Christophe De Vleeschouwer、J-F Delaigle 和 Benoit Macq。2002。感知水印中的不可见性与应用功能概述。*IEEE* 论文集，90(1):64–77。

勃朗特·艾米莉。1847年。《呼啸山庄》。

Shi Feng、Eric Wallace、II Alvin Grissom、Pedro Rodriguez、Mohit Iyyer 和 Jordan Boyd-Graber。2018。神经模型的病态特征使得解释变得困难。发表于自然语言处理实证方法会议。

Siddhant Garg 和 Goutham Ramakrishnan。2020。Bae：用于文本分类的基于 BERT 的对抗样本。载于*2020* 年自然语言处理实证方法会议（*EMNLP*）论文集，第 6174–6181 页。

Josh A Goldstein、Girish Sastry、Micah Musser、Renee DiResta、Matthew Gentzel 和 Katerina Sedova。2023年。生成式语言模型与自动化影响行动：新兴威胁与潜在缓解措施。*arXiv* 预印本 *arXiv:2301.04246*。

Park HanSol。2022。网络小说乘势走红，成为网络漫画和电视剧改编的素材来源。《韩国时报》。

Wei Hao、Lingyun Xiang、Yan Li、Peng Yang 和 Xi- aobo Shen。2018。使用同义词替换和算术编码的可逆自然语言水印。

Xuanli He、Qiongkai Xu、Lingjuan Lyu、Fangzhao Wu 和 Chenguang Wang。2022a。使用词汇水印保护语言生成 API 的知识产权。载于 *AAAI* 人工智能会议论文集，第 36 卷，第 10758–10766 页。

Xuanli He、Qiongkai Xu、Yi Zeng、Lingjuan Lyu、Fangzhao Wu、Jiwei Li 和 Ruoxi Jia。2022b。Cater：通过条件水印保护文本生成 API 的知识产权。载于 神经信息处理系统进展。

Matthew Honnibal 和 Ines Montani。2017年。spaCy 2：使用 Bloom 嵌入、卷积神经网络和增量解析实现自然语言理解。即将发表。

Chiou-Ting Hsu 和 Ja-Ling Wu。1999。图像中的隐蔽数字水印。*IEEE* 图像处理汇刊，8(1)：58–68。

Di Jin、Zhijing Jin、Joey Tianyi Zhou 和 Peter Szolovits。2020。BERT 真的鲁棒吗？文本分类和蕴含任务中的自然语言攻击基线。载于*AAAI* 人工智能会议论文集，第 34 卷，第 8018–8025 页。

Diederik P Kingma 和 Max Welling。2014年。自动编码变分贝叶斯。载于国际学习表征会议。

John Kirchenbauer、Jonas Geiping、Yuxin Wen、Jonathan Katz、Ian Miers 和 TomGoldstein。2023年。大型语言模型的水印。*arXiv* 预印本 *arXiv:2301.10226*。

Dianqi Li、Yizhe Zhang、Hao Peng、Liqun Chen、Chris Brockett、Ming-Ting Sun 和 William B Dolan。2021年。用于文本对抗攻击的上下文化扰动。载于 *2021*年北美计算语言学协会分会年会：人类语言技术论文集，第 5053–5069 页。Xiyang Luo、Ruohan Zhan、Huiwen Chang、Feng Yang 和 Peyman Milanfar。2020。失真无关深度水印。载于 *IEEE/CVF*计算机视觉与模式识别会议论文集，第 13548–13557 页。Andrew L. Maas、Raymond E. Daly、Peter T. Pham、Dan Huang、Andrew Y. Ng 和 Christopher Potts。2011。 用于情感分析的词向量学习。载于 计算语言学协会人类语言技术第 *49* 届年会论文集，第 142–150 页，美国俄勒冈州波特兰。计算语言学协会。Stephen Merity、Caiming Xiong、James Bradbury 和 Richard Socher。2016。指针哨兵混合模型。 *arXiv* 预印本 *arXiv:1609.07843*。John Morris、Eli Lifland、Jack Lanchantin、Yangfeng Ji 和 Yanjun Qi。2020a。重新评估自然语言中的对抗样本。载于 *2020*年自然语言处理实证方法会议论文集：研究成果，第 3829–3839 页。

John X Morris、Eli Lifland、Jin Yong Yoo 和 Yanjun Qi。2020b。Textattack：一个用于自然语言处理对抗攻击的框架。*2020*年*EMNLP*会议论文集，*Arvix*。

Seung-Min Mun、Seung-Hun Nam、Haneol Jang、Dongkyu Kim 和 Heung-Kyu Lee。2019年。从攻击中寻找鲁棒域：盲水印的学习框架。*Neurocomputing*，337：191–202。

OpenAI。2022。 介绍 ChatGPT。

Vidyasagar M Potdar、Song Han 和 Elizabeth Chang。2005年。数字图像水印技术综述。载于 *INDIN'05*。*2005*年第*3*届*IEEE*国际工业信息学会议，*2005*年。，第709–716页。IEEE。

Stefano Giovanni Rizzo、Flavio Bertini 和 Danilo Montesi。2019年。用于知识产权保护的细粒度水印。*EURASIP*信息安全期刊，2019年(1)：1–20。

Hai Tao、Li Chongmin、Jasni Mohamad Zain 和 Ahmed N Abdalla。2014年。鲁棒图像水印理论与技术：综述。应用研究与技术期刊，12(1)：122–138。

Rohan Taori、Ishaan Gulrajani、Tianyi Zhang、Yann Dubois、Xuechen Li、Carlos Guestrin、Percy Liang 和 Tatsunori B. Hashimoto。2023年。Stanford al- paca：一个指令跟随型Llama模型。 https://github.com/tatsu-lab/stanford_alpaca。Tina Tina Fang、Martin Jaggi 和 Katerina Argyraki。2017。使用长短期记忆网络生成隐写文本。载于第*55*届计算语言学协会年会论文集-学生研究研讨会，CONF，第100–106页。Mercan Topkara、Giuseppe Riccardi、Dilek Hakkani- Tür 和 Mikhail J Atallah。2006a。自然语言水印：构建实用系统的挑战。载于多媒体内容安全、隐写与水印 *VIII*，第6072卷，第106–117页。SPIE。Mercan Topkara、Cuneyt M Taskiran 和 Edward J Delp III。2005年。自然语言水印。载于多媒体内容安全、隐写与水印 *VII*，第5681卷，第441–452页。SPIE。Umut Topkara、Mercan Topkara 和 Mikhail J Atallah。2006b。歧义的隐藏优势：通过同义词替换实现具有可量化韧性的自然语言文本水印。载于第*8*届多媒体与安全研讨会论文集，第164–174 页。Honai Ueoka、Yugo Murawaki 和 Sadao Kuro- hashi。2021年。使用掩码语言模型实现令人惊讶地简单的基于编辑的语言隐写。载于*2021*年北方会议论文集

计算语言学协会美国分会：人类语言技术，第5486–5492页。

Ran-Zan Wang、Chi-Fang Lin 和 Ja-Chen Lin。2001年。通过最优最低有效位替换和遗传算法进行图像隐写。模式识别，34(3)：671–683。

Raymond B Wolfgang、Christine I Podilchuk 和 Edward J Delp。1999年。用于数字图像和视频的感知水印。*IEEE*论文集，87(7)：1108–1126。

Xi Yang、Jie Zhang、Kejiang Chen、Weiming Zhang、Ze-hua Ma、Feng Wang 和 Nenghai Yu。2022年。通过上下文感知的词汇替换追踪文本来源。在*AAAI*人工智能大会论文集，第36卷，第11613–11621页。

Zhong-Liang Yang、Xiao-Qing Guo、Zi-Ming Chen、Yong-Feng Huang 和 Yu-Jin Zhang。2018年。Rnn-stega：基于循环神经网络的语言隐写。*IEEE*信息取证与安全汇刊，14(5)：1280–1295。

Zhong-Liang Yang、Si-Yu Zhang、Yu-Ting Hu、Zhi-Wen Hu 和 Yong-Feng Huang。2020。Vae-stega：基于变分自编码器的语言隐写术。*IEEE*信息取证与安全汇刊，16:880–895。

Yang Zeyi。2021年 中国正在重塑世界的阅读方式。*Protocol*。

Xin Zhong、Pei-Chi Huang、Spyridon Mastorakis 和 Frank Y Shih。2020。一种基于深度神经网络的自动化鲁棒图像水印方案。*IEEE*多媒体汇刊，23:1951–1961。

Jiren Zhu、Russell Kaplan、Justin Johnson 和 Li Fei-Fei。2018。Hidden：使用深度网络隐藏数据。载于欧洲计算机视觉会议论文集（*ECCV*），第657–672页。

Zachary Ziegler、Yuntian Deng 和 Alexander M Rush。2019年。神经语言隐写。载于*2019*年自然语言处理实证方法会议和第九届自然语言处理国际联合会议（*EMNLP-IJCNLP*）论文集，第1210–1215页。

鲁棒性	相关性类型	ContextLS (Yang 等人, 2022)	关键词	句法
$\mathcal{R}_{g_1}$	D	0.656	0.944	0.921
	I	0.608	0.955	0.959
	S	0.646	0.974	0.949
$\mathcal{R}_{g_2}$	D	0.649	0.679	0.535
	I	0.591	0.679	0.517
	S	0.641	0.756	0.612

表 9：在 IMDB 数据集上，5% 扰动率下，g_1 和 g_2 的三个组件针对三种扰动类型（删除（D）、插入（I）和替换（S））的鲁棒性。

$\mathcal{R}_{g_1}$	相关性类型	关键词	句法
Wikitext-2	D	0.878	0.871
	I	0.909	0.939
	S	0.935	0.963
德古拉	D	0.947	0.940
	I	0.953	0.972
	S	0.987	0.963
WH	D	0.945	0.934
	I	0.963	0.965
	S	0.977	0.936

表 10：在 5% 扰动率下，g_1 对我们提出的组件抵御三种扰动类型的鲁棒性：删除（D）、插入（I）和替换（S）。

A 附录

A.1 实现细节

数据集划分 按照 ContextLS 的做法，我们对前 5000 个句子进行子采样，并在所有方法中使用相同的子集。初步实验表明，对其他样本进行子采样只会带来轻微波动：3 次试验中每词比特数（BPW）均值的标准误为 0.002。为避免任何混杂因素，我们在所有实验中都使用同一子集。对于具有随机性的鲁棒性实验，插入和替换情况下比特错误率（BER）的标准误也都很小（均为 0.004），与性能差距相比微不足道。

为微调我们的鲁棒填充模型，我们需要使用一个不同于待加水印测试集的训练集。对于 IMDB 和 Wikitext-2，我们使用原始训练划分。对于小说数据集，我们将文本前 40% 作为训练集，其余部分作为测试集。AWT 的训练也采用相同的划分。

破坏 为测试鲁棒性，我们对其进行破坏

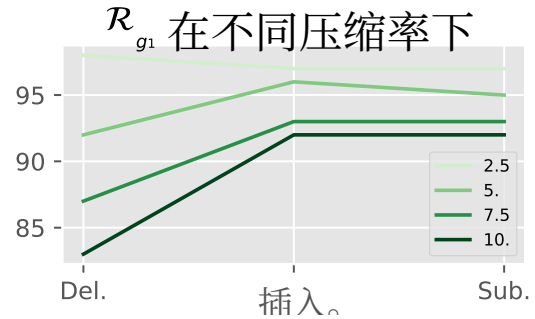


图 3：在更高破坏率下 g_1 的鲁棒性。

测试集中的前 1000 个句子（共 5000 个）。由于对于给定消息，我们的方法和 ContextLS 的水印嵌入过程都是确定性的，因此我们使用固定的随机种子运行一次嵌入实验。由于 TextAttack 的实现方式，某些破坏模块可能具有非确定性，从而导致比特错误率不确定。我们发现，所使用的删除模块是确定性的，因此只运行一次鲁棒性实验。另一方面，对于插入和替换，我们为每个样本生成五个破坏后的样本，并报告我们的方法和 ContextLS 的平均值。

计算时间 实际的水印过程不需要计算梯度。整个流程中最大的瓶颈是填充模型的前向传播。实际耗时和前向传播次数详见§5.1。训练填充模型需要耗费最多的计算时间。我们在单 GPU 环境中使用 Titan RTX 或 RTX 3090 对所有模型进行微调。在各个数据集中，Wikitext-2 的微调时间最长，100 个训练轮次约需 22 个 GPU 小时。

填充模型的训练细节 我们使用 AdamW，学习率为 5e-5，并在总训练步数的 0.1 处进行线性预热。所有模型均训练 100 个训练轮次，并使用最后一个检查点。对于随机掩码，我们采用整词掩码策略，随机掩盖 15% 的词语。

A.2 AWT 实现细节

我们使用官方实现，并且除非另有说明，大体上遵循 AWT 所采用的超参数。原文中的实验仅在 Wikitext-2 数据集上针对较低载荷（每词比特数为 =0.05）进行，因此，对于较高载荷（每词比特数为 =0.1），其实现细节或】【。需要调整其他数据集。

首先，我们用 GPT-2 替换了 AWD-LSTM 语言模型，从而获得更强的语言建模能力。其次，当载荷增加到 BPW=0.1 时，我们在 AWT 的第二个训练阶段将重建损失的权重项（见第 IV-D 节）加倍，以使模型收敛。第三，由于数据不足，无法针对每个数据集分别训练模型，因此我们将《德古拉》和《呼啸山庄》的数据合并为一个数据集，用于训练和评估 AWT 模型。

为了在鲁棒性实验中进行公平比较，我们将水印片段拼接后拆分成句子，并以句子为单位施加破坏。最后，我们使用破坏后的片段报告针对各种攻击的比特错误率。此外，AWT 会在嵌入水印之前，使用语料库构建词元词典。这可能会导致插入和替换时出现未知词元，在这种情况下，我们会将这些词元排除。

A.3 IMDB 数据集上的 AWT

IMDB 数据集的文本重建损失未能收敛。这导致带水印句子的质量严重下降，如下表13所示。尽管如此，我们仍测试了其在扰动下的鲁棒性。三种扰动类型的 BER@CR=0.05 分别为 0.283、0.278 和 0.299。

A.4 更多结果

NLI 排序与舍弃并列结构为了定义句法依存关系的排序，我们对训练集中每个依存关系分别进行掩码，然后填充被掩码的依存关系。将填充后的句子与原始句子进行比较。算法 1 展示了一个样本的 Python 风格算法。 1。我们对 IMDB 的 500 个样本执行了此操作。所得排序如表14所示。

如 §5.2所述，替换并列依存关系（CC）通常会导致一种自动指标无法检测到的语义漂移。我们还在表11中给出了从 NLI 排序列表中舍弃 CC 后的 BPW 和鲁棒性结果。

基于字符的破坏我们还尝试了基于字符的破坏，这种情况可能会在人工转录过程中无意发生。我们通过随机将一个字符与相邻字符交换来模拟这种类型的破坏。

使用 TextAttack 对字符进行攻击。与我们的主要实验类似，我们在 CR={2.5%, 5%} 条件下进行测试。在 IMDB 数据集上，我们的句法依存关系组件模型的比特错误率分别为 0.079 和 0.167。尽管我们的 RI 模型并未专门针对这类错误进行训练，但其鲁棒性仍有所提升，错误率分别降至 0.063 和 0.142。

ContextLS +RobustInfill 使用微调后的填充模型使我们的方法在所有数据集上的鲁棒性都得到了显著提升。该模型对 ContextLS 同样有效吗？使用通过随机掩码训练的填充模型并不总能提升 ContextLS 的鲁棒性，与我们的方法相比，其改进幅度也很有限（附录表 16）。鉴于我们在 §3.1 中的分析，这一点是意料之中的：阶段 1 是 ContextLS 的主要瓶颈。不过，我们认为，如果在微调填充模型时采用 ContextLS 中使用的特定掩码策略，还可以进一步提升其性能。

A.5 更多讨论

计算比特错误率对于我们的方法和 ContextLS，每个句子所包含的比特数各不相同。这导致在计算比特错误率时出现一个问题：预测消息的比特数可能少于或多于真实消息。为准确评估比特错误率，我们假设在提取过程中不知道真实的比特数。当提取出的比特数少于真实值时，我们将所有未预测的比特视为错误。相反，当提取出的比特数更多时，我们会将其截断，并将所有多提取出的比特视为错误。

A.6 人工评估

我们通过 ClickWorker 收集了关于带水印文本的人工标注，并告知参与者，其回答可能会用于研究目的。参与者来自美国、英国和爱尔兰，年龄在 20 至 99 岁之间，且自认为以英语为母语。问卷设计为大约需要 40–60 分钟完成，报酬为 20 欧元，高于这三个国家的最低工资标准。我们仅采用了那些对于未经过修改、与原始载体文本保持一致的带水印文本，给出较高比例“语义得到了完全保留”回答的问卷结果，以确保参与者遵循了问卷说明。当以 0.5 为该比例的阈值时，7 份回答中有 2 份被剔除。问卷截图见图4的最后一页。该问卷由 10 个随机样本组成——分别从德古拉和《呼啸山庄》中选取样本。由于 AWT 将实体名称预处理为未知词元，这可能导致标注者评估的流畅度大幅下降，因此我们排除了 Wikitext-2。由于 AWT 的文本重构损失未能收敛，导致生成的句子难以理解，因此我们排除了 IMDB。第 1 部分要求对每个句子的流畅度进行评分，包括原始封面文本。流畅度 Δ 通过从原文的流畅度中减去水印样本的流畅度计算得出。第 2 部分要求根据参考句（载体文本）对语义保留程度进行评分。

A.7 水印示例

水印文本示例见表 17-20。水印词语以颜色标记。对于我们的方法和 ContextLS，如果原文包含在有效水印集合中，部分文本可能与载体文本保持不变。对于 AWT，只有在片段的其他部分嵌入了水印时才可能出现这种情况，因为它通常以多个句子（40 个词语）作为输入。因此，我们仅展示经过修改的示例用于定性分析。（相反，在人工评估中，我们会随机抽取句子。）对于包含大量实体的 Wikitext-2，AWT 输出中的许多实体都被标记为未知词元。为了便于展示，我们手动替换了这些词元。

指标		使用 CC	舍弃 CC
IMDB			
每词比特数（↑）		0.130	0.151
比特错误率(↓) @压缩率=0.025	D	0.072	0.085
	I	0.113	0.123
	S	0.111	0.125
比特错误率(↓) @CR=0.05	D	0.195	0.224
	I	0.161	0.194
	S	0.187	0.200
嵌入成功率 (↑)		0.970	0.963
语义相似度（↑）		0.974	0.978
Wikitext-2			
BPW（↑）		0.099	0.115
比特错误率(↓) @压缩率=0.025	D	0.137	0.132
	I	0.197	0.180
	S	0.142	0.140
比特错误率(↓) @压缩率=0.05	D	0.274	0.231
	I	0.195	0.172
	S	0.194	0.179
嵌入成功率（↑）		0.966	0.961
语义相似度（↑）		0.993	0.993
德古拉			
每词比特数（↑）		0.146	0.135
比特错误率(↓) @压缩率=0.025	D	0.030	0.062
	I	0.063	0.093
	S	0.081	0.099
比特错误率(↓) @CR=0.05	D	0.177	0.193
	I	0.155	0.234
	S	0.164	0.179
嵌入成功率（↑）		0.963	0.944
语义相似度（↑）		0.971	0.965
《呼啸山庄》			
每词比特数（↑）		0.114	0.113
比特错误率（↓） @压缩率=0.025	D	0.063	0.075
	I	0.068	0.114
	S	0.096	0.117
比特错误率(↓) @压缩率=0.05	D	0.169	0.204
	I	0.133	0.200
	S	0.161	0.190
嵌入成功率 (↑)		0.964	0.942
语义相似度 (↑)		0.975	0.969

表 11：在 IMDB 上舍弃并列依存关系后水印嵌入和提取的结果。

	超参数。	关键词	句法
IMDB	KR	0.06	0.05
	k_2	4	4
Wikitext-2	KR	0.06	0.07
	k_2	4	4*
德古拉	KR	0.07	0.03
	k_2	4	3
WH	KR	0.05	0.03
	k_2	4	4

表 12：每个数据集采用的配置，以确保载荷约为每词比特数=0.1。KR 表示关键词数量与句子中词语数量的比率。我们确保每个句子至少选取一个关键词。

原文与水印文本
“预算限制、时间限制，拍摄剧本，然后删减、删减、再删减.....这个剧组是一群优秀的年轻电影人；
政治/战略 放映时间 *非常 拍摄剧本，然后删减、删减、再删减.....这个剧组是一群优秀的年轻吉尔伯特

表 13：AWT 在 IMDB 上无法重构载体文本的示例。

依存关系的类型		
1. 解释	6. 助动词	11. 前置限定词
2. 并列连词	7. 介词	12. 格标记
3. 被动助动词	8. 限定词	13. 从句主语
4. 施事者	9. 小品词	14. 访问控制列表
5. 标记	10. 并列关系	15. 状语从句

表 14：按自然语言推理蕴含分数排序的依存关系列表（前 15 项）。有关各项依存关系的详细信息，请参阅斯坦福依存关系手册。

数据集	指标	关键词	句法	+RI	-NLI 序位
D1	ES	0.932	0.975	0.975	0.854
	SS	0.967	0.982	0.981	0.946
D2	ES	0.895	0.966	0.966	0.696
	SS	0.979	0.993	0.993	0.953
D3	ES	0.920	0.960	0.963	0.835
	SS	0.964	0.974	0.971	0.939
D4	ES	0.910	0.964	0.964	0.790
	SS	0.967	0.976	0.975	0.941

表 15：我们方法各变体的水印集合语义评分（ES：蕴含分数，SS：语义相似度）。

指标		ContextLS	Δ	Ours	Δ
每词比特数 (↑)		0.100	+0.0	0.130	+1.3%
比特错误率(↓) @压缩率=0.025	D	0.219	+2.0%	0.100	+2.8%
	I	0.303	-0.5%	0.153	+4.0%
	S	0.273	+1.6%	0.133	+2.2%
比特错误率(↓) @压缩率=0.05	D	0.392	+1.4%	0.279	+9.4%
	I	0.362	+2.0%	0.236	+7.9%
	S	0.343	0.0%	0.224	+4.5%

表16：在 IMDB 的前 1,000 个句子上，对 ContextLS 使用鲁棒填充（RI）模型的效果。正数 表示比特错误率有所改善。作为参考，我们展示了我们的方法的改进幅度。

德古拉

原文	我们的方法	我们的方法（丢弃 *CC*）	上下文-*LSAWT*
我担心浓重的气味会让处于虚弱状态的亲爱的孩子难以忍受，于是把它们全都拿走，并打开了一点窗户，让新鲜空气透进来。			
我担心浓重气味会让虚弱状态下的亲爱的孩子难以承受，于是把它们全都拿走了**但**打开了一点窗户，让新鲜空气进来。			
我担心**如果**浓重气味会让她处于虚弱状态的亲爱的孩子难以承受，所以我把它们都拿走了，并稍稍打开窗户，让一点新鲜空气进来。			
我担心浓重气味对她虚弱状态下的 **孩子** 来说会太过 **强烈** ，所以 **我** 把所有东西都拿走了，并打开了一点窗户，以便**让**一些新鲜空气进来。			
<eos> <eos> 考虑到浓重气味对虚弱状态下的亲爱的孩子来说可能太过强烈，于是我把它们全都拿走，并打开了**呵呵呵**窗户，让一些新鲜空气进来。			
他在大厅里打开餐厅门，我们走了进去，他小心翼翼地关上了门。 在他身后 .			
在大厅里，他打开了餐厅门，**当**我们走进去时，他小心地关上了门 在他身后 .			
他在大厅里打开餐厅门，我们走了进去，他小心翼翼地关上了门。 在他身后 .			
他在大厅里打开餐厅门，我们走了进去，他小心地关上了门。 在他身后 .			
在大厅里**我**打开餐厅门，我们走了进去，**在**小心地关上门时 在他身后 .			
显然，他已经读过了，正托着额头坐在那里，沉思着。			
显然，他已经读过这段文字，此刻正用手托着额头，若有所思地琢磨着。			
他显然读过了，正在反复思量**，这时**他坐在那里，手扶着额头。			
<escape>他显然已经读过了，此刻正坐在那里，一只手托着**头**，沉思着。</escape>			
他显然读过了，正坐在那里沉思，**手**放在**裙子**上。			
我已经尽了自己的一份力，现在接下来的任务就是保持体力。			
我已经尽了自己的职责，**但**现在我的首要任务是保持体力。			
我**已经** 完成了自己的职责，现在我的下一项任务就是保持体力。			
我已经**尽了**自己的职责，现在我**新**的任务是保持体力。			
我已经尽了自己的本分，现在我的下一项任务是**保持** 保持体力。			
我本来不打算打架，所以就等着食物，像狼、狮子和老虎那样用我的脑袋行事。我本来不打算打架，所以就等着食物，**或者**像狼、狮子和老虎那样用我的脑袋行事。我本来不打算打架，所以就等着食物，像狼、狮子和老虎那样用我的脑袋行事。我本来不打算打架，所以就等着食物，像狼、狮子和老虎那样用我的脑袋行事。**<句末>** 没打算**被追赶**打架，所以 **<句末>** 等着食物，像狼、狮子和老虎那样用我的脑袋行事。			

表17：带水印文本样例。原始封面文本显示在第一行。

《呼啸山庄》

原文	我们的方法	我们的方法（丢弃 *CC*）	上下文-*LSAWT*

“总的来说，我承认确实如此，艾伦，”她接着说，“但除非我也参与其中，否则希斯克利夫遭受的苦难又怎能让我满足？“总的来说，我承认确实如此，艾伦，”她接着说；“**而且**希斯克利夫遭受的苦难又怎能让我满足，除非我也参与其中？“总的来说，我承认确实如此，艾伦，”她接着说，“但除非我也参与其中，否则希斯克利夫遭受的苦难又怎能让我满足？“总的来说，我承认确实如此，艾伦，”她接着说，“但除非我也参与其中，否则希斯克利夫遭受的苦难又怎能让我满足？**那**“总的来说，我承认确实如此，艾伦，”她接着说，“但除非我也参与其中，否则希斯克利夫遭受的苦难又怎能让我满足？他亲自负责她的教育，并把这当成一种乐趣。他亲自负责她的教育，**但** 把这当成~~一种乐趣。他亲自负责她的教育**为** 自己，并把这当成一种乐趣。他把她的**学业** 全都~~亲自包办，并把这当成一种乐趣。他亲自负责她的教育，并把这当成一种乐趣。

我相信，目睹这个恶魔生命的终结，你也会和我一样感到欣慰；除非你能胜过他，否则他会要了你的命，而他也会毁了我。

我相信，亲眼见证这个恶魔生命的终结，你也会和我一样感到痛快；除非你能胜过他，否则他迟早会要了你的命，而他也会毁了我。

我相信，亲眼见证这个恶魔走向生命终点，你也会像我一样感到无比愉悦；如果**你胜过他；他也会成为我的毁灭**。

我敢肯定，你也会和**我**一样乐于见证恶魔的**现身**落幕；除非你先发制人，否则他会要了你的命；而他也会毁了我。

我相信你也会像**一样**乐于见证结局**如同**恶魔的
存在；**作为**除非你对他要求过分，否则他会成为你的死因；而他会是**多语者，**毁灭。

令我欣慰的是，他给出这番明智的劝告后便离开了我们，而欣德利则伸展开身子躺在了地上。
炉边地面 .
令我欣喜的是，他给出这番明智的忠告后便离开了我们，**而**欣德利则在
炉边地面 .
令我欣喜的是，他给出这番明智的忠告后便离开了我们，而欣德利则在
炉边地面 .
令我欣喜的是，他*说完*这番明智的忠告后便离开了我们，而欣德利则躺下休息
炉边地面上 .
令我欣喜的是，**在**离开我们后，欣德利听从了这番明智的劝告，舒展了身子 **<终止>**
炉边地面 .

我听见主人上楼的声音——冷汗从我的额头上流下：我吓坏了。
我听见主人上楼的声音——冷汗从**我的**额头上流下：我吓坏了。
我听见主人上楼的声音——冷汗从**我的**额头上流下：我吓坏了。
我听见主人上楼梯的声音——冷汗从额头上流下来：我吓坏了。
的听见主人上楼梯的声音——冷汗从额头上流下来：我吓坏了。

表18：带水印文本示例。原始封面文本显示在第一行。

Wikitext-2

原文 我们的方法 我们的方法（丢弃 *CC*） 上下文-*LSAWT*

他被好友兼前同事严武接替，严武获任成都总督。他被好友兼前同事严武接替，严武获任成都总督。他被好友兼前同事严武接替，严武获任成都总督。

他被严武接替，严武是他的朋友兼**前**同事，**获任**成都总督。

他被接替，**一位**严武，他是他的朋友兼前同事，获任成都总督。Keiser认为，这种情况使得有必要将这支分裂的师分成两支特种部队加以控制和指挥。Keiser认为，这种~~情况使得有必要将这支分裂的师分成两支特种部队加以控制和指挥。Keiser决定**因为**~~这种情况使得有必要将分裂的师分成两支特种部队加以控制和指挥。Keiser认为，这种情况使得有必要将分裂的部队分成两支特种部队加以控制和指挥。Keiser认为，这种情况使得有必要将分裂的**师作为他的**两支特种部队。他最大的抱负是为国家效力，成为一名成功的公务员，但他未能作出必要的妥协。

他最大的抱负是成为一名成功的公务员，为国家效力，**尽管**事实证明，他无法作出必要的妥协。

他最大的志向是为祖国服务，成为一名**出色的**公务员，但他无法作出必要的妥协。

他最大的抱负是服务于他的**国家**，成为一名成功的公务员，但他无法做出必要的妥协。
他最大的抱负是服务于他的国家**成为**一名成功的公务员，但他无法做出必要的妥协。

表 19：带水印文本示例。原始封面文本显示在第一行。

IMDB

原文	我们的方法	我们的方法（丢弃 *CC*）	上下文-*LS*
摄影师加里（大卫·哈塞尔霍夫）未经许可，在岛屿度假村为琳达（凯瑟琳·希克兰饰演，她的声音和举止像《小熊维尼》动画片中的屹耳）拍照。琳达是一位正在学习巫术的处女。	摄影师加里（大卫·哈塞尔霍夫）正在为琳达（凯瑟琳·希克兰饰演，她的声音**和**举止像《小熊维尼》动画片中的屹耳）拍照。琳达是一位正在岛屿度假村学习巫术的处女，未经许可。	摄影师加里（大卫·哈塞尔霍夫）正在拍照，**和**琳达（凯瑟琳·希克兰饰演，她的声音和举止像《小熊维尼》动画片中的屹耳）一起。琳达是一位正在岛屿度假村学习巫术的处女，未经许可。	摄影师加里（大卫·哈塞尔霍夫）正在**拍摄**照片，为琳达（凯瑟琳·希克兰饰演，她的声音和举止像《小熊维尼》动画片中的屹耳）拍摄。琳达是一位正在岛屿度假村学习巫术的处女，未经许可。
各方面都很业余。各方面都很业余。	各方面都很业余，各方面都很业余，各方面都很业余**透顶**。		
一部很有潜力的电影，却可能受预算所限。一部很有潜力的电影，却可能受预算所限。	一部很有潜力的电影，**可能** 受预算所限。一部很有潜力的电影，却可能受预算所限。		
一群人在马萨诸塞州的一座岛屿度假村聚集，却遭到邪恶女巫施展的黑魔法袭击。女巫以残酷、折磨人的手段逐一杀害众人。	一群人在马萨诸塞州的一座岛屿度假村聚集，**却遭到** 邪恶女巫施展的黑魔法袭击，女巫以残酷、折磨人的手段逐一杀害众人。	一群人在马萨诸塞州的一座岛屿度假村聚集，**却** 遭到邪恶女巫施展的黑魔法袭击，女巫以残酷、折磨人的手段逐一杀害众人。	一群人在马萨诸塞州的一座岛屿度假村聚集，却遭到邪恶女巫施展的黑魔法袭击。女巫以残酷、折磨人的手段逐一杀害众人。
我没看过《犯罪医生系列》的其他电影，所以无法进行比较。	我没看过“犯罪医生系列”的其他电影，**所以**无法进行任何比较。	我没看过“犯罪医生系列”的其他电影，所以无法进行任何比较。	我没看过“**犯罪** 医生系列”的其他电影，所以无法进行任何比较。

表20：带水印文本示例。原始封面文本见第一行。

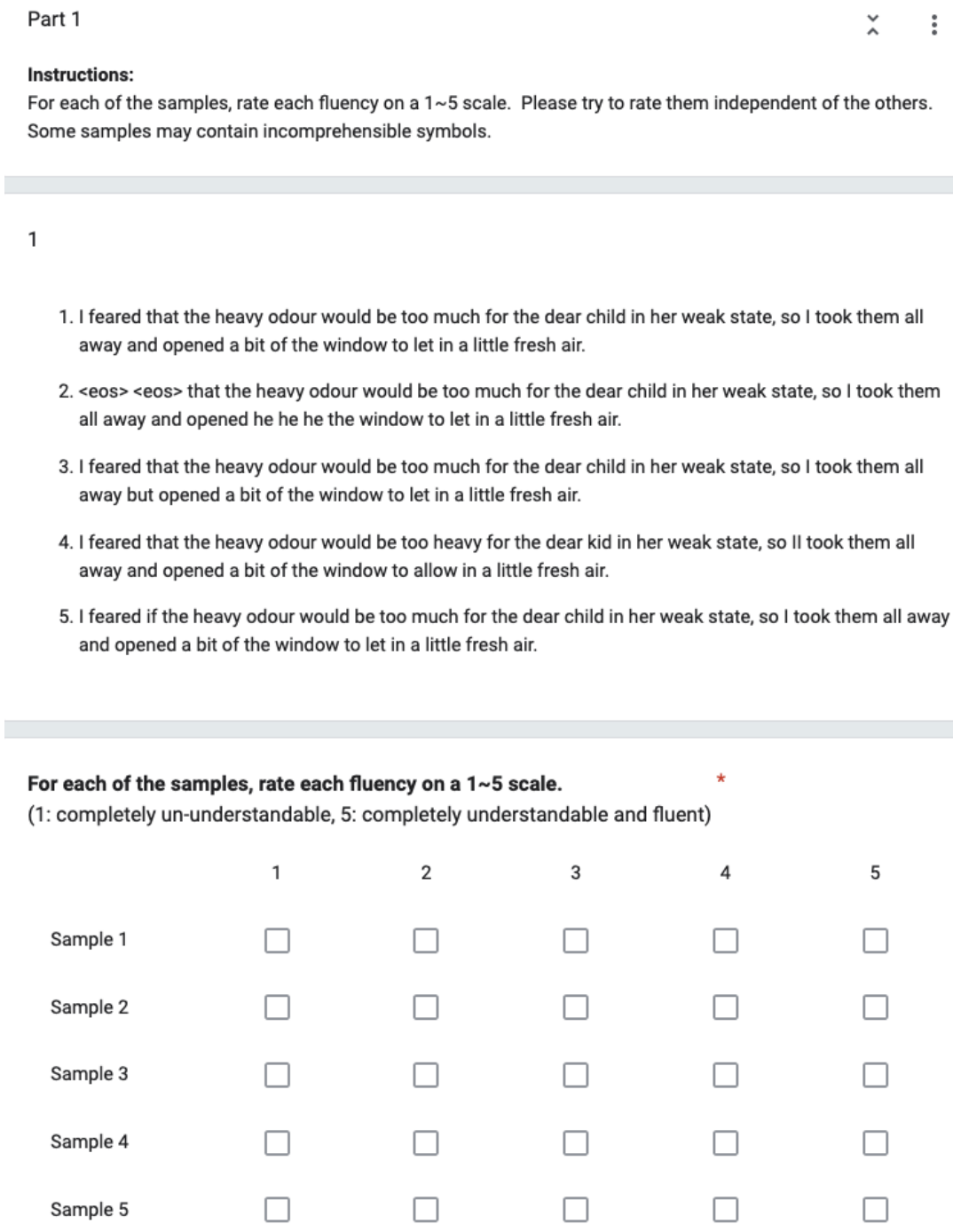
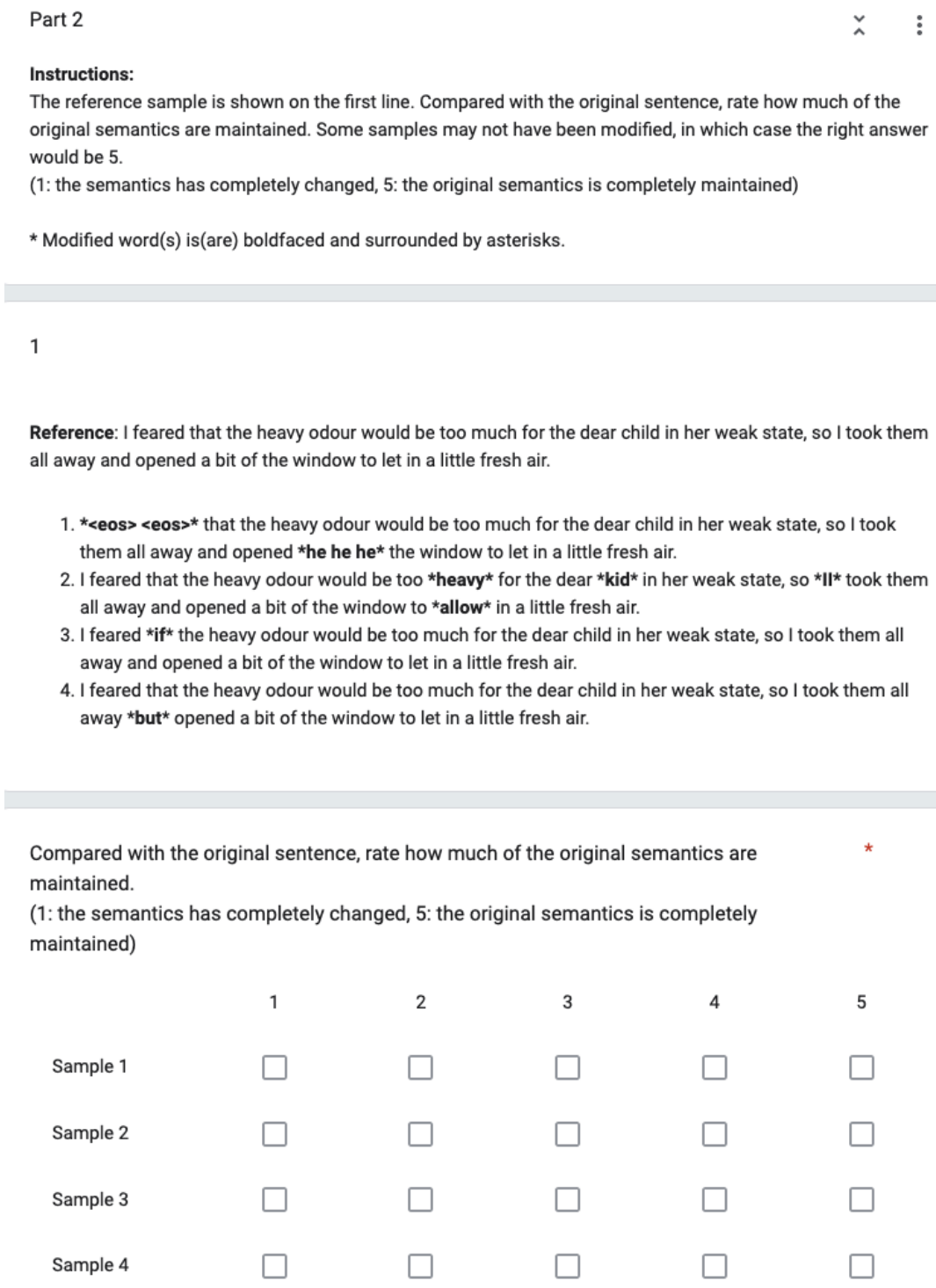


图 4：评估流畅度的人工评估问卷截图。


Part 2

Instructions:
The reference sample is shown on the first line. Compared with the original sentence, rate how much of the original semantics are maintained. Some samples may not have been modified, in which case the right answer would be 5.
(1: the semantics has completely changed, 5: the original semantics is completely maintained)

* Modified word(s) is(are) boldfaced and surrounded by asterisks.

1

Reference: I feared that the heavy odour would be too much for the dear child in her weak state, so I took them all away and opened a bit of the window to let in a little fresh air.

1. ***<eos> <eos>*** that the heavy odour would be too much for the dear child in her weak state, so I took them all away and opened ***he he he*** the window to let in a little fresh air.
2. I feared that the heavy odour would be too ***heavy*** for the dear ***kid*** in her weak state, so ***ll*** took them all away and opened a bit of the window to ***allow*** in a little fresh air.
3. I feared ***if*** the heavy odour would be too much for the dear child in her weak state, so I took them all away and opened a bit of the window to let in a little fresh air.
4. I feared that the heavy odour would be too much for the dear child in her weak state, so I took them all away ***but*** opened a bit of the window to let in a little fresh air.

Compared with the original sentence, rate how much of the original semantics are maintained. *
(1: the semantics has completely changed, 5: the original semantics is completely maintained)

	1	2	3	4	5
Sample 1	☐	☐	☐	☐	☐
Sample 2	☐	☐	☐	☐	☐
Sample 3	☐	☐	☐	☐	☐
Sample 4	☐	☐	☐	☐	☐

图 5：人工评估问卷的截图，用于评估与原始封面文本相比的语义。