arXiv:2611.32674v1 [stat.AP] 24 Nov 2026

LLM生成文本中水印比例的同时推断

Jingcheng He* Zinan Zhao† Wenguang Sun‡ Xintao Xia§

2026年11月25日

摘要

水印技术能够检测由大语言模型生成的文本，但实际审查还需要量化关注内容中保留的信号。读者可以根据内容选择段落，并在查看水印证据后调整选择。我们为选定词元中保留水印信号的比例构建同时置信下界。这些界对完整文档以及审查过程中选取的任意段落均联合成立。我们提出结构自适应闭合检验（SACT），利用序贯信号聚类的工作模型，为校准的符号检验构造得分。在条件零模型下，即使工作模型设定有误，这些界也具有有限样本同时有效性。精确参考族表示在支持高效重复查询的同时，可恢复完整闭合检验界。在一个构造的数据库恢复审查案例中，SACT 检测出文档范围内的水印信号，并为选定段落中的水印比例提供置信下界。将相关段落合并评估，可以得到更紧的合并信号计数下界。在涵盖两种水印机制以及一系列信号强度和密度的模拟中，SACT 的检验效能通常高于竞争方法。

关键词： 闭合检验；后选择推断；同时置信界；结构自适应推断；文本水印技术。

*浙江大学数学科学学院和数据科学中心。第一作者。

†密歇根大学统计学系

‡浙江大学数据科学中心§浙江大学数据科学中心；电子邮箱： xintaox@zju.edu.cn。通讯作者。

1 引言

借助大语言模型（大语言模型）起草的文档可能交错包含人类撰写的段落和生成文本段落（Lee 等人 2022年）。在科学和专业写作中（Noy & Zhang 2023年，Liang 等人 2025年），当读者评估某项主张或建议背后的推理时，这一区分可能十分重要。例如，一份数据库恢复审查可能会将已发表的故障报告摘录，与大语言模型对维护决策、备份故障和恢复流程生成的分析结合起来。审查中某处存在大语言模型生成的证据，并不能证明该模型参与了某一特定解释或建议的措辞。因此，读者可能会询问文档中有多少内容是由模型生成的，以及生成的措辞是否出现在支持某项决策的文本片段中。

文本水印技术通过在生成过程中嵌入验证器可使用密钥检测的统计信号，为这些判断提供证据（Kirchenbauer 等人（2023）、Kuditipudi 等人，2024、Dathathri 等人，2024、Li、Wen、Chen、Long、Jain、Joly、Lam、Song 与 Su 2026）。阳性的全局检验可以为水印提供证据，但无法量化文档中有多少内容包含该信号，也无法确定该信号出现在哪些文本片段中（Li, Ruan, Wang, Long & Su2025、2026年）。编辑可能会破坏生成文本中的信号（Li, Ruan, Wang, Long &Su2026）。因此，保留水印证据的词元比例为带水印模型的贡献值提供了一种保守衡量方式。近期研究估计了混合来源文档中的水印比例（Li、Wen、He、Wu、Long 与 Su 2025）、识别带水印片段（Li、Li 和 Zhang 2024、Zhao 等（2025）、Bonnerjee 等 2026），并选择具有水印技术证据的单个词元位置（Blanchet 等人，2026）。实际审查中的两个特点促成了我们的研究。

首先，审查过程中关注的段落可能会发生变化。读者在查看恢复演练建议时，可能会返回去阅读一段较早的说明，其中提到了备份失败，并

决定将这两段摘录放在一起评估。当水印证据影响这些选择时，就会出现后选择推断问题：针对预先指定文本片段校准的界，若应用于因证据有力而选出的文本片段，可能会高估水印比例。近期的水印技术通过数据分裂（Wei 等，2026）或对历史水印匹配设置同时界（Chen 2026）来处理选择问题。我们的目标是联合评估选出的段落，并给出一个置信下界，估计保留原始水印信号的词元比例，即使读者不断调整所选段落也能如此。

其次，利用词元顺序可以提高检测功效，并收紧关于水印比例的置信下界。水印信号往往会在生成文本段落中聚集，因此，来自相邻词元的证据有助于将带水印段落与孤立的大幅值得分区分开来。提高信号簇内各位置的权重，还有助于在编辑替换部分带水印词元或加入无水印文本时保持检验效能。挑战在于，如何利用这种序列结构，同时对读者选择的任意词元集合保持同时有效性。

我们针对水印比例的同时推断建立了一个框架。该框架提供了对所有非空词元位置集合同时成立的置信下界。为构造这些置信界，我们提出了结构自适应闭合检验 （SACT），利用工作模型从变换后的词元级统计量的幅值中学习序列结构，并保留其符号用于检验。在条件零模型下，即使工作模型设定错误，这些置信界仍然有效。我们还推导出了闭合检验界的一个精确参考族表示。在拟合工作模型并校准局部检验后，我们针对每个文档计算一次参考族。随后，当读者选择不同文本片段时，我们重复利用该参考族，以高效评估这些置信界。

在一次构造的数据库恢复审查中，SACT 检测文档级水印信号

并量化相关文本片段中的水印信号程度。这些界对所有选择同时成立，因此读者可以交互式地探索报告，并在问题出现时选择相应的文本片段。

1.1 贡献值

我们的贡献值如下。

1. 一种用于交互式水印推断的框架。 我们针对所有词元位置集合上的水印比例提供同时下置信界。即使读者根据文本和水印证据不断完善其选择，这些界仍然有效。该框架涵盖全局检测、文档级和文本片段级比例推断，以及词元级发现。

2. 采用精确计算捷径的结构自适应闭合检验。 我们开发了 SACT，通过结构自适应词元级评分来利用序贯信号聚类。我们推导出一种精确参考族表示，将文档级预处理与选定词元集合的界计算分离开来。在模型拟合和校准之后，该表示使得能够在 $O(n^2 + Mn)$ 时间内计算文档中 n 个词元的 M 个查询的完整闭合检验界。

3. 结构自适应推断的理论保证。 在假设1下，即使工作模型设定错误，我们仍建立了有限样本同时覆盖，从而使 SACT 能够利用序列结构，同时保持推断有效性。

1.2 相关工作

水印技术会在大语言模型生成的文本中嵌入可检测的信号。绿名单方法会修改词元概率，使其倾向于从词汇表的密钥指定子集中进行选择（Kirchenbauer 等人（2023）），可能会影响文本质量。无偏水印技术在对新的水印随机性取平均后，能够保留模型的下一个词元

分布（Hu 等人，2024、Kuditipudi 等人，2024）。我们的表述聚焦于无偏方案，包括 Gumbel–max 采样，以及 SynthID-Text 所采用的锦标赛采样（Dathathri 等人，2024）。

水印检测与比例推断。 近期研究基于词元级枢轴统计量开发了文档级水印检验，为统计高效的检测提供了框架（Li、Ruan、Wang、Long 与 Su 2025）。相关方法考虑了伪随机碰撞引起的依赖关系（Cai等人 2025），并使用截断拟合优度检验来提升对人工编辑的稳健性（Li、Ruan、Wang、Long 与 Su 2026年）。除了检测水印存在性之外，基于枢轴统计量的混合模型还能够估计混合文本中的水印比例（Li、Wen、He、Wu、Long 与 Su 2025）。将该框架应用于完整文档后，我们可以得到文档级保留信号比例的置信下界。正的置信下界意味着拒绝不存在水印信号的全局原假设。

混合来源文本中的定位。 确定文档中哪些部分保留了水印信号，对于实际应用十分重要，也促成了一系列定位方法的发展。窗口扫描和多尺度扫描能够识别水印证据集中的区域（Kirchenbauer 等人 2024、Pan 等人 2025、Zhao 等（2025）、Sander 等人 2026），而变点方法则估计片段边界（Li、Li 和 Zhang 2024、Li、Liu 和 Li 2025、Bonnerjee 等 2026）。词元级方法为选定位置提供渐近边际错误发现率控制（Blanchet 等人 2026）。FissionGuard 将文本片段的选择与验证分离，并控制所报告文本片段中的错误发现率（Wei 等，2026）。SACT 提供了不同的保证：对保留信号比例给出同时下界，即使读者使用相同的水印分数来选择文本片段并进一步细化选择，该下界仍然有效。

同时推断与结构自适应。 我们的方法借鉴探索性多重检验方法，构造在依赖数据的选择下仍然有效的同时置信界（Goeman 与 Solari（2011）、Goeman 等人 2019）。CountScan（Chen2026）对区间内的历史上下文–令牌匹配同时给出界，并通过汇总不相交区间的界，将这些界扩展到任意集合。SACT 的目标是保留其原始水印信号的词元，并使用结构自适应闭合检验，直接对选定集合中的此类词元比例给出界。闭合检验（Marcus 等人（1976）、Goeman 与 Solari（2011））将交集检验转换为同时真发现界。在 Goeman 等人 （2021年）的框架内，可容许的相干同时程序具有闭合检验表示。相关方法使用参考族（Blanchard 等（2020））或沿拒绝路径的 FDP包络（Hemerik 等人 2019、Katsevich 与 Ramdas（2020））；联合有效的参考族约束可以插值到任意集合。直接计算完全闭包需要指数级数量的交集检验。对于 Simes 检验，快捷方法能够高效地为任何选定集合给出界（Goeman 等人 2019）。FACT 利用局部检验的对称性和单调性，将计算复杂度降至多项式时间（Dobriban 2020）。我们的工作还与利用多重检验中结构信息的方法（Lei 和 Fithian 2018、Lei 等人 2021）以及保形推断（Zhao 和 Sun（2025））有关。在闭合检验中，局部检验的检验效能决定所得界所包含信息的丰富程度。相关构造包括 KJI（Li、Maathuis 和 Goeman 2024），它使用给定的抛硬币得分；以及 ACODE（Magnani 等人 2026），它学习利用参考观测进行校准的一致性得分。受这些进展启发，我们提出 SACT，将序列信号结构纳入任意词元集合上的同时推断，同时在工作模型设定错误下保持有效性。我们进一步推导出精确参考族表示，使完整闭合检验界能够高效计算。

本文其余部分的组织如下。第 2 节介绍水印证据和推断目标。第 3 节通过数据库审查应用发展 SACT，并总结相关数值研究。第 4节给出理论保证、参考族表示和精确捷径。第 5 节总结研究发现，并展望未来工作的方向。

2 Wa水印分析：问题与框架

第 2.1 节介绍全文评估和交互式内容审查中产生的问题。第 2.2 节定义保留信号比例，并阐述同时覆盖保证。第 2.3节概述 SACT 如何构造界。

2.1 应用背景与科学问题

由人类与大语言模型共同起草的文档可能同时包含人工撰写的段落和模型生成的文本段落（Lee 等人 2022年）。了解模型参与的位置十分重要，因为模型的建议可能影响最终文本所表达的观点（Jakesch 等人 2023年）。读者可以评估整篇文档，也可以在审查过程中选择段落进行更仔细的检查。这两种审查形式引出了以下问题：

1. 全文评估。 文档是否包含大语言模型生成的文本？如果包含，其比例是多少？

2. 交互式内容审查。 读者在审查过程中选择的段落中，有多少文本是由大语言模型生成的？

第一个问题既要求检测是否涉及大语言模型，也要求量化生成文本所占的比例。对于第二个问题，当审阅者根据内容选择文本片段，并依据证据不断完善选择时，结论必须依然有效。

另一个挑战是，单个词元可能只能提供很少关于大语言模型参与情况的证据。当生成文本出现在连续的文本片段中时，结合相邻词元的证据可以让信号更容易被检测到。

2.2 水印比例的同步推断

水印技术通过在词元采样过程中将统计信号嵌入生成文本，为第 2.1 节中的评估提供证据（Kirchenbauer 等人（2023）、Kuditipudi 等人，2024、Dathathri 等人，2024）。给定水印协议和密钥，验证器会在每个词元位置计算一个统计量，以量化该信号的证据（Li, Ruan,Wang, Long& Su 2025、2026年）。

令 $W_{1:n}$ =(W1、.....、 W_n) 表示观测到的、可能经过编辑的包含 n 个词元的文档，其中的词元可能代表单词或词语片段。对于 $j \in [n]$ ={1、.....、 n}，令 Y_j 表示位置 j 处的验证统计量，数值越大表示水印证据越强。遵循 Li、Ruan、Wang、Long 和 Su（2026）的编辑模型，令 $\xi_j \in$ {0、1} 表示位置 j 是否包含水印信号。若 W_j 未被改变，且其观测到的播种上下文与该步骤所使用的上下文相匹配，则设 $\xi_j = 1$ ；否则设 $\xi_j = 0$。因此，即使 W_j 未被改变，对前置上下文的编辑也可能破坏该信号。零假设和真实零假设位置的集合为

$$H_j : \xi_j = 0, \quad j \in [n], \qquad \mathcal{N} = \{j \in [n] : \xi_j = 0\}. \tag{1}$$

在假设1成立时， Y_j 在 $\xi_j = 0$时服从已知分布 F_0 。

令 $R \subseteq [n]$ 为选定进行审查的非空词元位置集合。真实发现数 $d(R)$ 和真实发现比例（TDP）TDP(R) 分别是 R 中保留水印信号的词元数量和比例：

$$d(R) = \sum_{j \in R} \xi_j = |R \setminus \mathcal{N}|, \qquad \mathrm{TDP}(R) = \frac{d(R)}{|R|}. \tag{2}$$

根据定义，$\mathrm{TDP}(R)$ 也是 R 中的水印比例。保留信号的词元必定来自带水印模型，但该模型生成的词元在编辑后可能失去信号。因此，$\mathrm{TDP}(R)$ 的置信下界也同时构成 R 中来自该模型的词元比例的下界。然而，如果 R 是在检查文本及其水印证据后才选定的，那么针对预先指定集合校准的置信界未必仍能达到其名义覆盖率。为解决这两个审查问题，我们构造 $\mathrm{TDP}(R)$ 的同时下置信界，并借鉴错误发现比例（FDP）的置信包络（Genovese 与 Wasserman（2006））以及用于探索性推断的闭合检验（Goeman 与 Solari（2011）、Goeman 等人（2021））。对于给定的$\alpha \in (0,\ 1)$，我们寻求满足以下条件的下界 $\mathrm{TDP}_\alpha(R)$：

$$\Pr\left\{\underline{\mathrm{TDP}}_\alpha(R) \le \mathrm{TDP}(R) \text{ for every nonempty } R \subseteq [n]\right\} \ge 1-\alpha. \tag{3}$$

$d(R)$对应的置信下界为 $d_\alpha(R) = |R|$真实发现比例$_\alpha(R)$，我们称之为认证计数。不同的 R 选择对应着两个审查问题。

1. 全文评估。 对于 $R = [n]$，该界给出了文档中保留水印信号的词元所占比例的下限。正的置信界也提供了水印存在性的证据，将全局检测（Li,Ruan, Wang,Long& Su 2025、2026年）与文档级比例推断（Li, Wen, He, Wu, Long & Su2025）联系起来。

2. 交互式内容审查。 读者可以利用词元级水印证据、待审查段落的内容，或两者兼用，来定义 R。

在词元层面，读者可以调整阈值 t 并查看选定位置在文本中的聚集位置 $R_t = \{j : Y_j \ge t\}$（Blanchet 等人）。对于每个非空的 R_t，$1-$ 真实发现比例$_\alpha$（R_t）都会给出实现的错误发现比例的上置信界，错误发现比例（R_t）$= 1-$ 真实发现比例（R_t）。对于期望的错误发现比例水平 γ，读者可以选择任意 t，使得 R_t 非空，且上界至多为 γ。

在内容层面， R 可能对应于单个文本片段，如段落级定位中所示（Li、Li 和 Zhang 2024、Zhao 等（2025）、Bonnerjee 等 2026），也可能包含来自多个文本片段的句子，这些句子共同支持同一主张或解释。该界适用于作为整体的选定文本；在查看证据后，读者进一步完善其选择时，该界仍然有效。

此外， R 还可以采用更灵活的形式，以契合特定任务的目标和约束，包括通过算法搜索得到的集合，如第 3.6 节所示。对于任何非空集合 R ，若该集合是在检查文本和水印证据后选定或修订的，公式（3）给出 $\Pr\{\text{真实发现比例}_{\alpha}(R) \leq \text{真实发现比例}(R)\} \geq 1-\alpha$ ，无需再针对选择进行调整。

2.3 交互式水印分析框架

闭合检验为公式（3）中的界提供了一种经典构造。SACT 将这一构造与结构自适应评分和精确的计算捷径相结合。图 1 展示了整体工作流程，下面概述具体步骤。

1. 结构自适应评分。 水印信号可能会在生成文本段落中聚集（Li、Li 和 Zhang 2024，Zhao 等（2025），Bonnerjee 等 2026）。我们对变换后的词元级统计量的幅值拟合一个工作模型，并将其符号保留用于检验。每个词元的拟合权重都会考虑相邻观测值。我们将变换后的统计量乘以该权重，得到用于检验的结构自适应得分。

2. 符号校准与闭合检验。 在条件零模型下，拟合得分满足符号翻转对称性（性质 1）。即使工作模型设定错误，这一性质也能对基于符号的局部检验进行校准（Barber 和 Candès 2015、Lei 和 Fithian（2018）、Zhao 和 Sun（2025））。闭合检验将这些检验结合起来

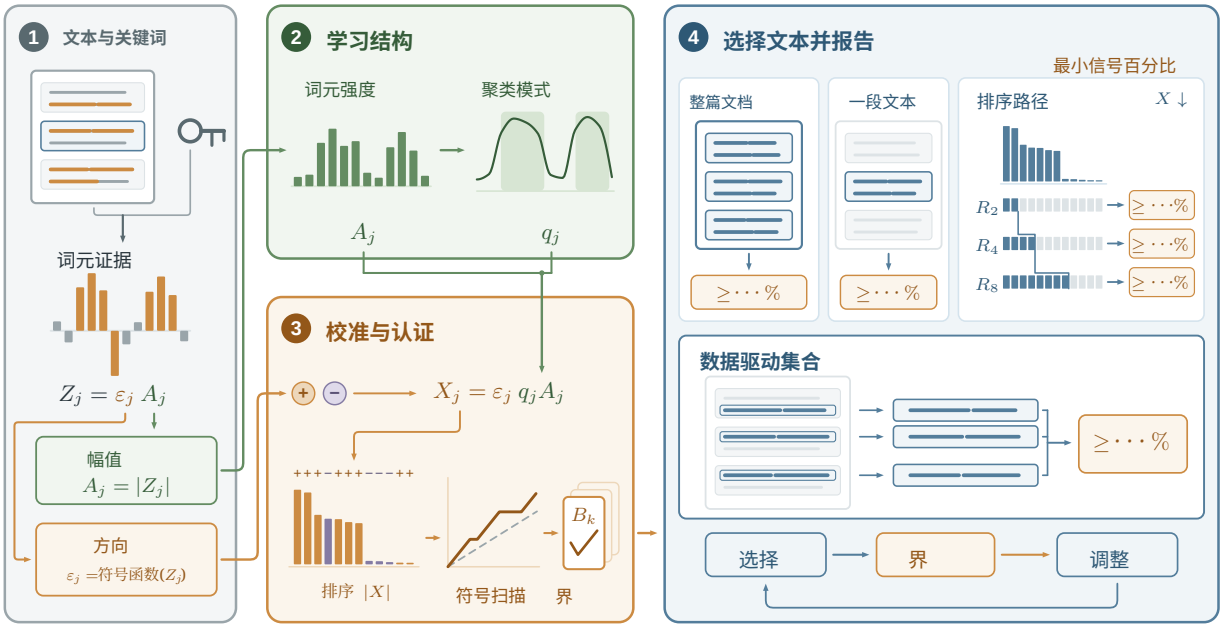

图 1：结构自适应闭合检验（SACT）概览。得分幅度指导模型拟合，符号为局部检验提供证据。闭合检验能够同时给出任意选定词元集合中信号比例的置信下界。

，从而为 d（R）和真实发现比例（R）针对每个非空$R \subseteq [n]$ （Marcus 等人（1976）、Goeman 与 Solari（2011））给出同时下置信界。

3. 高效评估选定集合。 基于参考族思想（Blanchard 等（2020）、Katsevich 与 Ramdas（2020）），我们推导出 SACT 界的精确表示，从而避免枚举所有交集检验。参考族针对该文档仅计算一次，随后可重复使用，以恢复每个新的选定集合对应的完整闭合检验界。

3 数据库恢复审查应用

我们将第 2.3 节中的框架应用于一次构建的数据库恢复审查，利用水印证据评估文档以及交互式审查期间所选择内容中的大语言模型参与程度。

3.1 文档构建与审阅问题

我们为一个假设的服务团队构建了一次数据库恢复审查，使用了归因于 GitLab 关于2017年1月故障的事后分析文章的摘录。[1] 我们围绕三个主题组织审查：破坏性维护操作的防护措施、备份恢复检查以及恢复演练。

针对每个主题，我们使用 Qwen3.5-9B（Qwen 团队，2026年）起草分析，并将其置于相应的原文摘录之后。维护、备份和恢复分析的目标长度分别为 512、768 和 1,024 个词元。我们采用 Gumbel–max 水印（Kuditipudi 等人，2024），使用温度为 1 的全词表采样，并根据前四个词元生成水印种子。为模拟人工编辑的影响，随后我们在禁用水印技术的情况下，使用同一个 Qwen3.5-9B 模型修改措辞，并将词元编辑限制在 3% 以内。最终文档包含 $5{,}454$ 个词元，其中$2{,}095$ 个词元位于生成后又经过编辑的部分。

3.2 水印生成与验证

在生成带水印文本时，密钥和当前位置之前的四个词元共同决定水印种子。解码器根据该种子，为每个候选词元生成一个伪随机值。Gumbel–max 规则将这些值与大语言模型提供的下一词元概率相结合，以选择下一个词元。

令 $W_{1:n}$ 表示最终报告，其包含 $n = 5,454$ 个词元。使用与生成时相同的分词器、密钥和播种规则，验证器根据观测到的上下文重构水印值。令 $U_{j,\,w}$ 表示词汇表词元 w 在位置 j处重构得到的值。对于观测词元 W_j，验证统计量及其 p值为

$$Y_j = U_{j,W_j}, \qquad p_j = 1 - Y_j.$$

[1]来源：GitLab，*1月31日*数据库故障事后分析，2017年2月10日。

更大的 Y_j，或等价地更小的 p_j，表明水印证据更强。验证无需获取生成过程中使用的下一词元概率。

重复的四词元上下文会重复使用水印值，因此我们会过滤掉每个上下文中除第一个词元位置以外的所有位置。这得到符合条件的位置集合 $\mathcal{E}$，其中有 $n_e = |\mathcal{E}| = 5{,}223$ 个位置。拟合时，我们将其余 231 个位置上的统计量视为缺失，但保留它们在序列中的位置。对于选定集合 R，只有符合条件的位置能够计入认证计数 $d_\alpha(R)$，而分母 $|R|$ 则包含所有选定词元。在符合条件的位置中，有 2,008 个仍保留水印信号。

3.3 构造结构自适应得分

为了优先审查某些词元，我们利用相邻位置的证据对水印分数加权。我们根据得分幅度学习权重，并将符号保留用于检验。

屏蔽与工作排放。 令 F_0 表示验证统计量 Y_j 的零分布函数。在假设 1 下，给定符合条件的位置、信号指示变量和非零统计量后，符合条件的原假设统计量是从 F_0 中独立抽取的。当 F_0 连续时，我们可以将每个符合条件的位置 $Y_j \quad j \in \mathcal{E}$ 上的统计量转换为正态分数，并将其符号与幅值分离：

$$Z_j = \Phi^{-1}\{F_0(Y_j)\} = \varepsilon_j A_j, \qquad \varepsilon_j = \operatorname{sign}(Z_j), \qquad A_j = |Z_j|,$$

其中 Φ 是标准正态分布函数。在相同条件下，原假设得分相互独立，且服从标准正态分布。进一步以其幅值为条件后，它们的符号函数仍然相互独立，并以相等概率取正或负。数据库审查使用 Gumbel–max 水印。在背景状态（$G_j = 0$）下，枢轴量 Y_j 服从均匀分布，因此 F_0（y） $= y$ 在 $[0, 1]$上。公式（10）给出了使用该水印生成的词元的枢轴量分布。这是贝塔分布的混合分布，其中一个

下一个词元概率为 $p > 0$ 的词元贡献一个 $\mathrm{Beta}(1/p,1)$ 分量，其权重为p。在 –对数 Y_j 尺度上，每个分量均为均值为 p的指数分布。验证器通常不知道这些概率。附录B.1.1 展示了如何通过对混合分布进行连续近似，得到活跃状态（$G_j = 1$）下的以下单参数工作模型，

$$-\log Y_j \mid G_j = 1 \sim \mathrm{Gamma}(\nu, 1), \qquad 0 < \nu \leq 1. \tag{4}$$

在 $\nu = 1$， –log Y_j服从均值为 1 的指数分布，因此 Y_j在背景状态下服从均匀分布。 ν 的值越小，枢轴量越倾向于集中在 1 附近。

背景状态下 A_j 对应的密度 g_0为标准半正态分布。要得到 g_1 在 $a > 0$处的活跃状态密度，我们将 Z_j 在 a 处的密度与 $-a$的密度相加。所得密度见公式（B.2）。对于 SynthID-Text，我们使用附录 B.3中定义的中心化二项式得分。在原假设下，给定得分幅值，该得分为正或为负的概率相同。

对信号聚类建模。我们用于信号聚类的主要工作模型是两状态隐马尔可夫模型（HMM），并在数据库审查中使用该模型。在模拟中，我们还考虑变点模型。两个模型都使用相同的密度 g_0 和 g_1 来描述符合条件的位置处的幅值，并将其余位置的观测视为缺失。

- 隐马尔可夫模型。 用于$G_{1:\,n}$ 通过相邻位置之间的信息共享发挥作用，而无需预先指定文本片段边界。即使单个观测只能提供微弱证据，活跃状态 $G_j = 1$也可以在一段文本中持续存在；编辑或播种上下文发生变化后，可能会出现这种情况。我们为转移参数设置连续先验，并对$\widehat{\ }$其后验分布上的前向–后向状态概率进行平均，以获得权重 $q_j = \Pr(G_j = 1 \mid A_{\mathcal{E}})$。附录 B.1.2

介绍了先验和模型拟合。

- 变点模型。受水印定位中变点（Li、Li 和 Zhang 2024）和流行病分段（Bonnerjee 等 2026方法的启发，我们的变点（CP）工作模型将活跃区域表示为 $\mathcal{I}$ 个 K 不相交区间的并集，其中 $G_j = \mathbf{1}\{j \in \mathcal{I}\}$。我们为 K 指定截断泊松先验，并在给定 K的条件下，为所有可接受的区间配置赋予相等的先验概率。我们使用动态规划对区间的数量和位置进行平均，从而得到权重 $q_j = \Pr(j \in \mathcal{I} \mid A_\mathcal{E})$。附录 B.1.3 给出了先验设定和递推公式。

在任一工作模型下，q_j 都是位置 j 属于活跃区域的拟合概率。使用后验概率与局部显著性指数（Sun 和 Cai 2009）有关。我们使用 q_j 对得分 Z_j加权，得到

$$X_j = q_j Z_j = \varepsilon_j q_j A_j, \qquad j \in \mathcal{E}. \tag{5}$$

幅值 $|X_j| = q_j A_j$ 优先考虑同时受到单个观测和相邻位置证据支持的位置。在第 3.7 节的模拟中，SACT 使用组合得分 $q_j Z_j$ 时通常比使用 Z_j 或 $\varepsilon_j q_j$时检验效能更高（图 C.5）。由于权重为正且仅取决于幅值，改变零假设符号不会改变拟合权重和得分幅度。在条件零模型下，即使聚类工作模型设定错误，加权得分仍保留符号翻转对称性。第 4.2 节确立了这种对称性（Barber 和 Candès 2015），我们利用该对称性校准第 3.4 节中的局部检验。

模型拟合与诊断。 对于数据库报告，隐马尔可夫模型的拟合结果为 $\hat{\nu} = 0.570$。平稳活跃状态概率和活跃运行时长的后验均值分别为 0.289 和 370 个词元。图 2 展示了拟合如何改变排序。在

符合条件的位置中，生成并编辑的分析部分的权重中位数为 0.995，周边材料中的权重中位数为 0.006。当选择集大小为文档词元的 10% 时，按 $X_j = q_j Z_j$ 降序排序可捕获已记录保留信号的 27.2%；而直接按 Z_j降序排序时，这一比例为 19.9%。

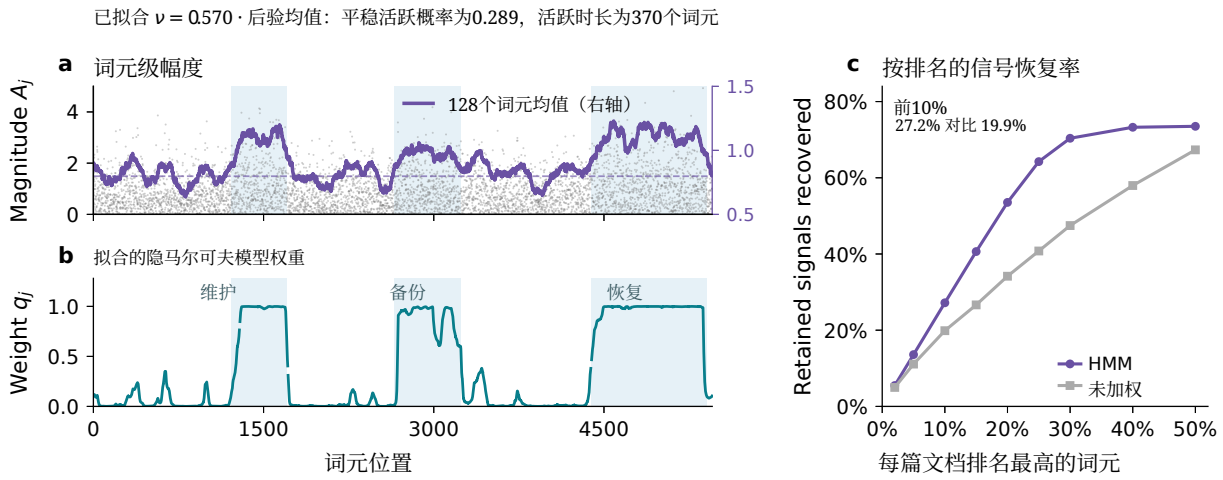

图2：数据库审查中的结构自适应评分。**a**，符合条件的词元幅度（左轴）及其在以中心对齐的128词元窗口上的移动均值（紫色，右轴），不包括不合格的位置。虚线表示右轴上的半正态零假设均值。**b**，根据幅值拟合的隐马尔可夫模型权重。**c**，由隐马尔可夫模型加权（递减 $X_j = q_j Z_j$）和未加权（递减 Z_j）排序捕获的已记录保留信号比例。阴影标示了大语言模型起草并随后修订的关于维护、备份和恢复的三个文本片段。

为了评估隐马尔可夫模型是否利用了序列聚类，我们在符合条件的位置之间随机置换观测幅度，并重新拟合模型。原始排序下，相较于全零假设模型，掩码对数似然的增益为 185.0；而在 64 种打乱后的排序下，增益为 114.1–121.5（中位数为 115.9）。因此，隐马尔可夫模型捕捉到了幅值排序中超出其边缘分布的信息。假设 1意味着零假设枢轴量服从均匀分布，并且相邻零位置上的正态分数不相关。本报告中，已知原假设枢轴量的经验分布与 Uniform(0

，*1*) 之间的柯尔莫哥洛夫–斯米尔诺夫距离为 0.012。相邻零位置上的未加权正态分数 Z_j 之间的样本相关性为 0.0045。这两项结果均与该假设一致。

3.4 校准检验并计算界

我们通过闭合检验构造同时下置信界，这要求对符合条件的位置 $\mathcal{E}$ 的每个子集进行信号缺失的局部检验。Donoho 和 Jin 2004 提出的高阶批判扫描思想启发了我们；我们采用符号扫描，以同时适应少数词元位置上的强水印信号，以及分布在更多位置上的较弱信号。

对于非空集合 $I \subseteq \mathcal{E}$，其大小为 m，令 H_I 表示假设：在 I 中没有任何位置保留水印信号。令 i_1、……、i_m 表示 I 中按得分幅度降序排列的位置，使得 $|X_{i_1}| \geq \cdots \geq |X_{i_m}|$；并列情况按照与符号函数无关的固定规则解决。这种幅度排序将符号保留用于检验，而图2中的带符号分数排序则用于确定优先审查的词元。具体而言，我们检验 H_I，使用

$$T_I = \max_{1 \leq k \leq m} \frac{\sum_{\ell=1}^{k} \varepsilon_{i_\ell}}{\sqrt{k}}, \qquad T_\varnothing = -\infty. \tag{6}$$

对于每个 k，分子表示在 k 个秩最高的位置中，正符号数超过负符号数的数量。根据 H_I，性质 1 表明，给定得分幅度后，有序符号相互独立，且取正号或负号的可能性相等。我们利用这一联合零分布，对所有 k 的最大值进行校准。局部检验$\phi_I^{\text{scan}} = \mathbf{1}\{T_I \geq c_{\alpha,\, m}\}$ 采用仅取决于 m 和 α 的截断值。在本研究中，我们使用 $\alpha = 0.05$ 以及精确截断点，使零假设拒绝概率至多为 α。附录 B.2.1介绍了它们的计算方法。

上述符号扫描检验对应于 $s = 2$ 公式（B.6）中的幂散度族成员，因此得名 SACT-2。我们将其作为默认检验，因为其统计量

具有基于有序符号累积和的简单形式。附录 C.3 和 C.4将其与其他 s 选择及备择局部检验进行比较。

将闭合检验程序（Marcus 等人（1976）、Goeman 与 Solari（2011）应用于这些局部检验，可得到任意选定集合中信号计数的下置信界R:

$$d_\alpha(R) = \min_{I \subseteq \mathcal{E}:\, \phi_I^{\text{scan}} = 0} |(R \cap \mathcal{E}) \setminus I|. \tag{7}$$

为说明为何最小值给出下界，考虑所有符合条件的零假设位置组成的集合。以至少 $1-\alpha$ 的概率，其局部检验不会拒绝原假设，因此该集合属于 I 公式（7）中的候选集合 I。对于这个特定的 I，$R\cap\mathcal{E}$）$\setminus I$ 中的每个位置都保留着一个信号。因此，所有候选集合中的最小值至多为 R中的真实信号计数 R。由于同一事件为每个 R都给出了该界，因此这些界可同时成立，正如定理2所述。对于非空的 R，比例下界为真实发现比例$_\alpha$（R） $= d_\alpha$（R）$/|R|$。

直接计算公式（7）需要遍历$\mathcal{E}$的所有子集。为避免这种搜索，我们按幅值对分数进行排序，并使用由此得到的嵌套集合来计算该界。第4.4节给出了精确参考族表示。

具体而言，首先按 $|X_j|$降序对位置进行一次排序，并令 S_k 包含前 k 个秩中的正得分位置，其中 $S_0 = \varnothing$。我们构造关于 S_k 中保留信号计数的下界 ℓ_k ，这些下界以至少 $1-\alpha$的概率联合成立。要将这些下界应用于一段文本 R，需要考虑 ℓ_k 所计数的信号可能位于何处。其中至多 $|S_k \setminus R|$个信号位于该文本之外。其余信号必定位于 R中，因此我们可以在所有 k中取这样的最大下界。同样的论证也适用于任何选定集合，包括段落并集，从而得到

$$d_\alpha(R) = \max_{0 \le k \le n_e} \big(\ell_k - |S_k \setminus R|\big)_+, \qquad (x)_+ = \max\{x, 0\}. \tag{8}$$

对于幅值不同的非零分数，该计算可恢复完整闭合检验

界。对于每个 R，定理 3 的证明构造了一个集合 I，其局部检验不拒绝，并且 $|(R \cap E) \setminus I|$ 等于公式（8）中的计数。对于包含 n 个词元的文档，在拟合和校准之后，构造参考集合 S_k 及其下界 ℓ_k 需要 $O(n^2)$ 时间，而评估每个选定集合的界需要 $O(n)$ 时间。因此，对于 M 个集合，总成本为 $O(n^2 + Mn)$。附录 A.1 中的算法 1 给出了预处理和查询步骤。

3.5 不同审查目标下的发现

我们首先评估完整报告中的水印存在性及其程度（$R = [n]$；问题 1），随后使用三类审查目标考察证据集中于何处：高排名词元集合、滑动窗口和句子对齐段落（问题 2）。为评估结构加权的贡献值，我们将 SACT-2 与其未加权版本进行比较（$X_j = Z_j$）。两种方法采用相同的局部检验和显著性水平，并在相同的窗口和选定段落上进行评估；词元级选择则分别依据各自的得分排名。

对于词元级选择，令 R_k 包含带符号得分 X_j最高的 k 个符合条件的位置。对于每个 FDP 目标 $\gamma \in \{0.1, 0.2, 0.5\}$，我们选择满足 $\mathrm{TDP}_\alpha(R_k) \geq 1 - \gamma$的最大 R_k；若没有符合条件的，则选择空集。对于段落级评估，我们评估长度为 128、256 和 512 个词元的滑动窗口，每次移动 16 个词元。为获得便于阅读的摘录，我们还在审查预算的限制下，对至少包含 64 个词元的连续完整句子组成的段落进行优化，以最大化 TDP 下界。对于每种方法，这些界在所有上述选择中同时成立。

全局检测和比例。 对于完整报告（$R = [n]$），SACT-2 认证了 5,454 个词元中至少有 869 个保留了水印信号，给出的置信下界为 15.9%（图3(c)）。无权重程序认证至少有 807 个信号，即

14.8%。两种程序都检测到了水印信号，其中加权方法给出的全局下界略高。作为比较，已知保留信号的 2,008 个符合条件的词元占完整报告的 36.8%。

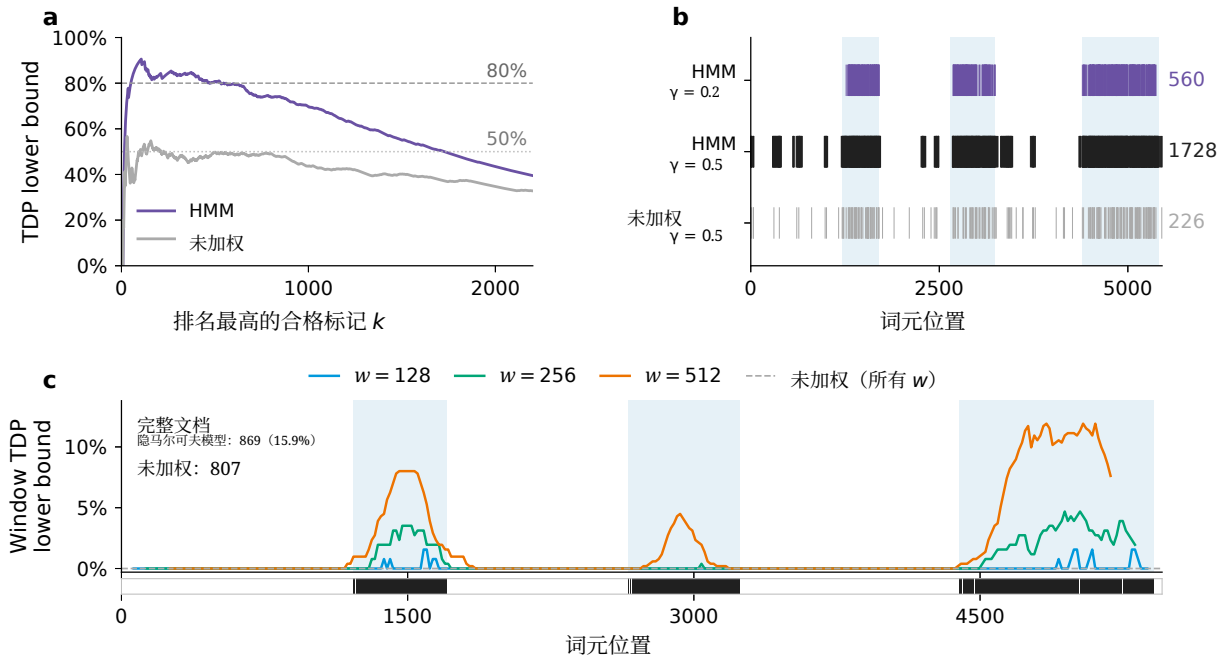

图3：词元级发现和窗口定位。**a**，排序后的前 k 个词元中信号比例的置信下界。**b**，在所示 FDP 目标下选择的词元位置，右侧显示计数。**c**，滑动窗口的比例下界。下方色带标记了记录的保留信号。

词元级发现。 词元级选择通过突出水印证据集中的位置，补充全文评估。在 FDP 目标 $\gamma = 0.2$下，HMM加权程序选择了 560 个位置，并证明其中至少有 448 个保留了水印信号（图 3（a）和（b））。在这一目标下，无权重程序的选择结果为空。生成记录显示，全部 560 个选定位置均保留了信号。将目标放宽至 $\gamma = 0.5$ 后，HMM加权选择扩大到 1,728 个位置，而无权评分下为 226 个。因此，结构自适应对词元级发现带来的增益，远大于其对文档级下界带来的增益。该

选定位置集中在维护、备份和恢复分析附近，引导审阅者更加仔细地检查报告的这些部分。

段落级定位。 接下来，我们直接使用滑动窗口定位水印证据。SACT-2 的界在维护、备份和恢复分析附近达到峰值（图 3（c））。信号比例的最大置信下界从 128 个词元窗口的 1.6% 上升到 256 个词元窗口的 4.7%，再升至 512 个词元窗口的 11.9%。较大的窗口可以积累证据，但这些证据难以在更细的分辨率下得到认证。无权重程序对所检查的每个窗口都给出零下界。

为了获得可读的摘录，我们还考虑由连续完整句子组成的文本片段，其长度受阅读预算限制。对于每个预算 B，我们选择最多包含 B 个词元且能使信号比例置信下界最大化的文本片段。在 256、512 和 1,024 个词元的预算下，SACT-2 分别认证了 251 个词元中至少 12 个保留信号（4.8%）、489 个词元中 54 个（11.0%）以及 994 个词元中 220 个（22.1%）（图 4（a）和（b））。这三段摘录均位于恢复分析中或其附近，涵盖恢复演练、依赖项以及关于恢复时间的假设。在这三段相同的摘录中，无权评分仅认证了 0、0 和 2 个保留信号。尽管完整报告的界相近（SACT-2为 15.9%，无权评分为 14.8%），SACT-2 在这三段恢复文本中每一段都认证出了显著更多的保留信号。

3.6 阅读预算下对文本片段的联合评估

前面的分析针对每种阅读预算选择了一个文本片段。评估拟议恢复演练的读者可能还希望查看此前有关维护和备份程序的文本片段。在固定的阅读预算下，读者可以从全部三项分析中选择相关文本片段。为说明这种审阅方式，我们使用一种算法来选择文本片段，并对选定文本中的水印证据进行联合评估。

具体而言，令 $\mathcal{C}_{B,m}$ 为由至多 m 个不相交段落取并集构成的非空集合的集合。每个文本片段由完整句子组成，且至少包含 64 个词元；合并后的长度不超过 B 个词元。我们要在 $R \in \mathcal{C}_{B,m}$中最大化 $\mathrm{TDP}_{\alpha}(R)$，其中 $B \in \{256, 512, 1024, 1536, 2048\}$，且 $m \in \{1, 2, 3\}$。对于 $m = 1$，我们枚举所有可行的文本片段。对于 $m > 1$，我们使用附录 B.4 中的启发式方法组合候选段落，并调整其句子边界，以提高联合下界。选定的文本片段会随阅读预算和段落上限而变化（图 4(a) 和 (b)）。在 $B = 1536$ 和 $m = 3$条件下，选定的摘录涵盖全部三个主题，并联合证明，在 1,522 个词元中至少有 522 个保留信号（34.3%）。在 $B = 2048$ 和 $m = 3$条件下，选择三个文本片段将待审阅文本从 5,454 个词元缩减至 1,868 个词元。SACT 证明这些文本片段中至少有 674 个保留信号（36.1%），而完整报告中有 869 个。根据生成记录，选定文本片段中的词元有 97.5% 保留了水印信号。聚焦这些文本片段，可大幅减少待审阅文本量，同时捕获报告中大部分保留的水印信号。

为了解联合评估带来的提升，我们分别评估相同的三个文本片段（图 4<style id='5''>（c））。使用相同的拟合得分和校准检验，单个段落的下界分别为 41、23 和 209。三者之和证实了 273 个保留信号，占合并文本的 14.6%；而直接评估并集时，则为 674 个信号（36.1%）。汇总三个文本片段的证据，可以对其总体水印比例给出更严格的下界；将它们放在一起阅读，还能看出大语言模型如何为报告中对失败的解释及其建议作出贡献。

在阅读预算和段落限制内，搜索旨在获得较高的联合界。审阅者也可以根据内容选择段落，并在查看证据后修改选择。同时覆盖适用于任一选择规则（Goeman 与 Solari（2011））。

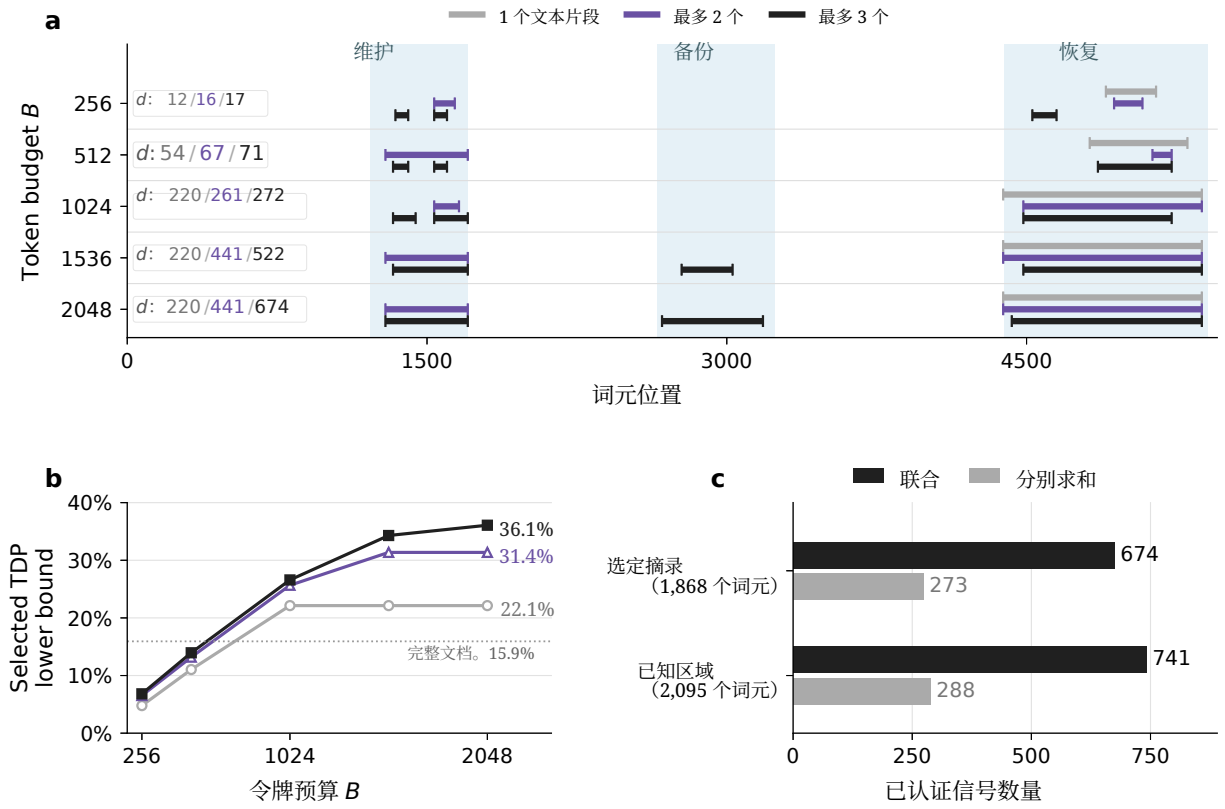

图4：在阅读预算 B 和段落限制 m下，对句子对齐段落进行联合评估。**a**，选定段落位置和联合认证计数，计数按照图例顺序列出。**b**，选定文本中保留信号比例的置信下界。虚线表示全文档界。**c**，联合认证计数与对同一段落分别进行认证所得计数之和的比较，展示选定段落（ $B = 2048$、 $m = 3$ ）以及三个边界已知的完整分析结果。三个阴影段落是大语言模型起草后又经修订的段落。标签标示其主题。

3.7 支持性数值研究

为进一步评估 SACT 的性能，我们在不同信号比例和强度下，将其与其他同时推断程序进行比较。我们从 Colossal Clean Crawled Corpus (C4) 中抽取背景文档。该语料库由清理后的英语网页文本组成，

来自 Common Crawl 的网页文本（Raffel 等，2020）。在每份文档中，我们随机选取两个彼此分隔的区间，并将其替换为等长的文本片段。这些片段由 Qwen3.5-9B（ L Qwen Team 2026）在采用四词元局部上下文种子的 Gumbel–max 水印技术下生成。

我们改变文档长度 n、段落长度 L和采样温度 t （每种设置重复实验 100 次）。各方法共享验证统计量、符合条件的位置、窗口和排序路径。有关具有代表性的设置，请参见表 1 ，详细信息请参见附录 B 和 C 。

方法	HMM/CP 加权	全局 检验效能	认证 计数	窗口 计数	$\lvert R_Y \rvert$	$\lvert R_X \rvert$ HMM	$\lvert R_X \rvert$ CP	中央处理器（秒）
SACT-1	✓	0.950 [0.887, 0.984]	66.6 (3.4)	26.1 (1.5)	25.2 (3.8)	112.0 (7.2)	104.7 (7.3)	0.033
SACT-1.5	✓	**0.960** [0.901, 0.989]	67.8 (3.4)	26.4 (1.5)	22.0 (3.8)	114.2 (7.3)	105.6 (7.5)	0.034
SACT-2	✓	**0.960** [0.901, 0.989]	68.4 (3.4)	26.7 (1.5)	18.8 (3.7)	**114.7** (7.4)	106.5 (7.5)	0.033
SACT-2-CP	✓	**0.960** [0.901, 0.989]	**69.2** (3.4)	27.2 (1.5)	21.1 (3.9)	109.5 (7.8)	**119.2** (7.5)	0.033
SACT-BC	✓	0.950 [0.887, 0.984]	33.4 (2.0)	17.7 (1.2)	27.2 (3.6)	62.8 (4.1)	59.9 (4.3)	0.032
SACT-KJI-B	✓	0.950 [0.887, 0.984]	48.7 (2.7)	22.9 (1.3)	**29.5** (3.8)	87.2 (5.7)	82.6 (5.8)	0.018
SACT-KJI-D	✓	0.950 [0.887, 0.984]	46.9 (2.7)	21.4 (1.3)	**29.5** (3.8)	82.2 (5.6)	77.6 (5.6)	0.014
SACT-q	✓	**0.960** [0.901, 0.989]	68.2 (3.4)	25.7 (1.6)	0.0 (0.0)	25.8 (6.5)	23.7 (6.5)	0.032
SACT-Z	×	0.940 [0.874, 0.978]	50.5 (3.3)	1.4 (0.3)	14.4 (3.3)	1.5 (0.8)	1.5 (0.8)	0.032
GoF-1	×	0.940 [0.874, 0.978]	52.9 (5.0)	2.0 (0.2)	14.0 (2.2)	4.4 (0.8)	4.6 (0.9)	0.520
GoF-1.5	×	0.930 [0.861, 0.971]	39.2 (3.1)	2.1 (0.2)	11.6 (1.7)	5.1 (0.8)	5.3 (0.8)	1.134
GoF-2	×	0.880 [0.800, 0.936]	21.8 (2.0)	1.7 (0.2)	8.4 (1.3)	4.4 (0.6)	4.5 (0.7)	1.067
Simes	×	0.620 [0.517, 0.715]	1.3 (0.2)	1.0 (0.1)	2.6 (0.3)	2.5 (0.3)	2.5 (0.3)	1.117
CountScan	×	0.950 [0.887, 0.984]	40.7 (2.1)	**28.2** (1.2)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.037

表 1：在 $(n,\ L,\ t) = (1024,\ 128,\ 1)$ Y、隐马尔可夫模型和 CP 排序方法下的模拟比较，基于 100 次重复实验。计数对应完整文档和最佳的 256 词元窗口。选定大小使用共同的 Y、隐马尔可夫模型和 CP 排序方法，FDP 目标为 0。5。CPU 时间包括预处理以及针对 $n = 2048$的 1,000 次子集查询，详情见附录 C.1.4。方括号表示 95% 二项区间；圆括号表示蒙特卡罗标准误。粗体标示列最大值。

在 Gumbel–max 下，SACT-2及其 CP 变体具有有竞争力的检测功效，且通常比无权评分、基于 p 值的程序和 CountScan 认证更多信号（图 C.3）。向固定的加水印文本片段附加背景文本后，它们的计数几乎保持稳定。对枢轴量、资格指标和信号标签进行联合置换并重新拟合模型后，两种计数均随长度增加而下降。此时，SACT-2 给出的计数与 GoF-1 相近（图 C.10）。

来自 s 的增益　结构加权对局部推断的影响更为显著。CountScan

在较短窗口中往往能认证更多信号，而 SACT 变体在较长窗口中给出的界更大（图 C.6）。在 FDP 目标为 0。5 时，基于共同隐马尔可夫模型排名的 SACT-2通常比无权评分或基于 q 的得分选择更多词元。BC 和 KJI–B 在更严格的目标下可能会选择更多词元（图 C.4 和 C.5）。

在 SynthID-Text 水印技术（Dathathri 等人，2024）下，两种变体在 20 个未置换信号设置中的平均检测功效和认证计数方面均优于 CountScan。其最佳 256 个词元窗口的平均比例界大于采用无权评分或 GoF-1 时的结果（图 C.7 和C.8）。附录 B.3 介绍了实现细节。

4 方法论与理论基础

我们为 SACT 建立了两个理论结果。首先，在给定的条件零假设下，即使信号聚类的工作模型存在误设，这些界对于所有词元集合仍具有有限样本同时覆盖。其次，我们推导出一种精确参考族表示，为在重复查询中评估完整的 SACT-2 闭合检验界提供了高效捷径。

4.1 水印机制与条件零模型

我们引入 Gumbel–max 验证统计量（Li, Ruan, Wang, Long & Su 2025，2026年）。令 $P_j = (P_{j,w})_{w\in\mathcal{V}}$ 表示给定提示词及位置 j 之前词元时的下一词元分布。在理想化随机性假设下，给定 P_j，$\zeta_j = (U_{j,w})_{w\in\mathcal{V}}$ 的各项是相互独立的 Uniform(0，1) 随机变量。Gumbel–max 解码器选择

$$W_j = \arg\max_{w:P_{j,w}>0} \{\log P_{j,w} - \log(-\log U_{j,w})\} = \arg\min_{w:P_{j,w}>0} \frac{-\log U_{j,w}}{P_{j,w}}. \tag{9}$$

对水印随机性取平均后，该规则以概率 w 选择每个词元$P_{j,\ w}$ 因此保留了下一词元分布（Li, Ruan, Wang, Long & Su 2025）。枢轴量 $Y_j = U_{j,\ W_j}$ 是 ζ_j 中由观测词元 W_j索引的分量。如果观测词元 W_j 与验证向量 ζ_j相互独立，那么 Y_j 服从均匀分布，其零分布函数为 $F_0(y) = y$，定义在 $[0,1]$上。对于通过方程（9）中的水印解码器选取的词元，给定 P_j 时的条件分布函数为

$$F_{1,P_j}(y) = \sum_{w:P_{j,w}>0} P_{j,w} y^{1/P_{j,w}}, \qquad 0 \le y \le 1. \tag{10}$$

例如，如果 P_j 在 s 个词元上均匀分布，那么 $Y_j \sim$ 服从 Beta(s,1)。如果 P_j 将全部概率赋予一个词元，那么即使在水印生成下，枢轴量仍服从均匀零分布。信号强度的这种变化促使我们在第 3.3 节引入工作模型。

对不同位置进行校准需要一个关于零假设枢轴量的联合模型。当位置经过筛选后，我们按照文档顺序为符合条件的位置编索引，并在整个过程中以这些位置的所在位置为条件；这些位置用于拟合工作模型。以下假设涉及这些位置处的枢轴量。记作 $Y_I = (Y_j)_{j\in I}$ and $\mathcal{N}^c = [n] \setminus \mathcal{N}$。

假设1（条件零假设枢轴量）。 给定信号配置 $\xi_{1:n}$:

（i）零假设枢轴量相互独立且服从共同分布 F_0：$\mathcal{L}(Y_{\mathcal{N}} \mid \xi_{1:n}) = F^{\otimes|\mathcal{N}_0|}$。（ii）零假设与非零假设枢轴向量相互独立：$Y_{\mathcal{N}} \perp\!\!\!\perp Y_{\mathcal{N}^c} \mid \xi_{1:n}$。

独立同分布零假设模型是水印检测中的一种常见工作理想化（Li, Ruan,Wang, Long& Su 2025、2026年）。假设1(i) 将这一理想化应用于在给定信号配置条件下的零位置，而第 (ii) 部分要求，在进一步对非零证据进行条件化后，它们的联合分布保持不变。这一表述与 Li、Ruan、Wang、Long 与 Su(2026年) 的潜在源模型一致，该模型将普通词元采样与水印随机性区分开来。对于预先指定的信号配置，这也符合 Cai等人所讨论的固定下一词元分布理想化

(2025)，在该理想化下，标记级枢轴被视为相互独立。直观地说，只需重建得到的伪随机向量 ζ_j 在零位置处彼此独立，且独立于完整文档及其生成随机性。例如，假设一位无法访问水印密钥的作者在带水印文本后附加一段独立撰写的文本片段。假设附加的文本片段使用了与带水印文本不同的播种上下文。在各上下文随机性相互独立的理想化假设下，该附加文本片段因此具有相互独立的均匀枢轴Y_j，且这些枢轴独立于带水印文本中的枢轴。

4.2 符号翻转对称性与校准

零假设符号翻转不会改变拟合权重，因为它们只取决于得分幅度。在假设 1 下，可得以下对称性。

性质 1（符号翻转对称性）。 假设假设1 成立，且 F_0 连续。给定任意固定的信号指标 $\xi_{1:n}$，方程（5）中的分数满足

$$(X_1, \ldots, X_n) \stackrel{d}{=} (\eta_1 X_1, \ldots, \eta_n X_n) \tag{11}$$

对于每个固定的 $\eta \in \{-1,\ \mathbf{1}\}^n$，其中 $\eta_j = 1$ 对于 $j \notin \mathcal{N}$。等价地，在给定 $\xi_{1:n}$、所有分数幅度（$|X_j|$）$_{j\in[n]}$以及非零得分（X_j）$_{j\notin\mathcal{N}}$的条件下，零符号（符号函数(X_j)）$_{j\in\mathcal{N}}$相互独立，且以等概率取值 $+1$ 和 -1 。

为说明这一点，在 $\xi_{1:n}$、非零枢轴量以及用于拟合的幅值 $A_{1:n}$ 给定的条件下进行分析。这些量决定拟合权重，而零假设符号仍相互独立，且取正或负的概率相等。由于权重严格为正，加权不会改变符号。对任意一组零假设符号取反，都不会改变条件得分分布。对 $A_{1:n}$ 和非零枢轴量取平均，得到公式（11）。同样的校准论证也适用于其他得分在条件下具有符号对称性的机制，包括附录B.3中定义的 SynthID-Text 中心化二项式得分。

公式（6）中的扫描与Arias-Castro& Wang (2017) 的 CUSUM 符号统计量形式相同。性质1 给出了结构自适应分数的零假设参考分布。设 $e_1, \ldots, e_m$ 为相互独立的 Rademacher 变量，每个变量以 $+1$ 和 -1为取值，且两者等概率。对于 $m \geq 1$ 和 $0 < \alpha < 1/2$，定义

$$T_m^0 = \max_{1 \leq k \leq m} \frac{\sum_{\ell=1}^{k} e_\ell}{\sqrt{k}}, \qquad c_{\alpha,m} = \min\left\{c \in \mathrm{supp}(T_m^0) : \Pr(T_m^0 \geq c) \leq \alpha\right\}, \tag{12}$$

若显示的集合为空，则为 $c_{\alpha,m} = +\infty$ 。记 $c_m = c_{\alpha,m}$ ，并令 $c_0 = +\infty$。对于非空的 I，局部检验为 $\phi_I^{\mathrm{scan}} = \mathbf{1}\{T_I \geq c_{\alpha,|I|}\}$；令 $\phi_\varnothing^{\mathrm{scan}} = 0$。

在 H_I 下， I 中的每个位置都是原假设。给定 $\xi_{1:n}$ 和得分幅值后，排序固定，且 I 中的有序符号是相互独立的 Rademacher 变量。因此， T_I 的条件分布与 $T_{|I|}^0$ 相同，局部检验以不超过 α 的条件概率拒绝。

4.3 同时有效性

记 $d_\alpha^{\mathrm{CT}}(R)$ 和 真实发现比例$_\alpha^{\mathrm{CT}}(R)$ 分别表示第 3.4 节中的闭合检验计数界和比例界，其中 $[n]$表示可检验位置。以下定理证明它们同时有效。

定理 2 (SACT 的同时有效性)。 假设分数具有性质 *1* 中所述的符号翻转对称性，并令 $0 < \alpha < 1/2$。按照方程 (5) 构造分数，所用权重严格为正且仅取决于 $A_{1:n}$；得分幅值相同时，使用与符号无关的固定规则确定排序。对于第 *4.2* 节中的符号扫描局部检验，使用方程 (12)中的精确临界值进行校准后，闭合检验界满足

$$\Pr\left\{\underline{\mathrm{TDP}}_\alpha^{\mathrm{CT}}(R) \leq \mathrm{TDP}(R) \text{ for every nonempty } R \subseteq [n]\right\} \geq 1 - \alpha.$$

该保证在有限样本中成立，且不要求工作模型设定正确。

论证遵循 Goeman 与 Solari (2011年）。以至少 $1-\alpha$ 的概率，局部检验不会拒绝真实零集 $\mathcal{N}$。在此事件下，$\mathcal{N}$ 是方程（7）中最小化问题的一个候选项，因此 $d_\alpha^{\mathrm{CT}}(R) \le |R \setminus \mathcal{N}| = d(R)$ 对每个 R 都同时成立。

4.4 精确参考族表示

参考族通过插值联合有效的信号计数下界，为任意集合得出比例界（Blanchard 等（2020））。令 $\mathfrak{R} = \{(S_k,\ \ell_k) :\ k \in \mathcal{K}\}$ 由集合$S_k \subseteq [n]$ 及满足条件的整数下界 $0 \le \ell_k \le |S_k|$ 组成

$$\Pr\left\{d(S_k) \ge \ell_k \text{ for all } k \in \mathcal{K}\right\} \ge 1-\alpha. \tag{13}$$

在此事件下，S_k 中认证的信号至多有 $|S_k \setminus R|$ 个位于 R 之外。对于非空的 R，取剩余界中最强者可得

$$\underline{\mathrm{TDP}}_\alpha^{\mathrm{RF}}(R) = \frac{1}{|R|}\max_{k\in\mathcal{K}}\Big(\ell_k - |S_k \setminus R|\Big)_+, \qquad d_\alpha^{\mathrm{RF}}(R) = |R|\underline{\mathrm{TDP}}_\alpha^{\mathrm{RF}}(R).$$

这些界对所有 R 同时成立。现在我们构造一个 SACT-2 参考族，其插值可恢复完整闭合检验界。

设 $0 < \alpha < 1/2$，且拟合得分均非零，其绝对值两两不同。使用公式（12）中的精确临界值，并令 π 按$|X_j|$降序排列各位置。对于 $k \in \mathsf{K} = \{0, \dots, n\}$，定义 $S_k = \{\pi(\ell) :\ 1 \le \ell \le k, \varepsilon_{\pi(\ell)} = 1\}$、$P_k = |S_k|$和 $Q_k = k - P_k$。因此，S_k 包含前 k 秩中的正得分位置，而 Q_k 则统计这些秩中的负分数。集合 S_k 是嵌套的。

令 $m_0 = |\mathcal{N}|$ 表示未知的零位置数量。对于候选零假设集合大小 r及有限临界值 c_r，SACT-2 扫描给出的信号数量下界为

$$\ell_k(r) = \left[P_k - \min\left\{r, \left\lceil\frac{r + c_r\sqrt{r}}{2}\right\rceil - 1, \left\lceil Q_k + \frac{c_r^2}{2} + c_r\sqrt{2Q_k + \frac{c_r^2}{4}}\right\rceil - 1\right\}\right]_+. \tag{14}$$

直观而言，扫描的不拒绝限制了 S_k 中可由大小为 r的零假设集合解释的正分数数量。P_k 超过这一上限的部分，为保留信号计数提供了一个下界。令 $\ell_k(0) = P_k$，并令 $\ell_k(r) = (P_k - r)_+$，当 $c_r = +\infty$ 时。

如果扫描未拒绝真实零集，那么对于每个 k，$d(S_k) \geq \ell_k(m_0)$。由于总信号计数为 $n - m_0$，这一事件也意味着 $m_0 + \ell_k(m_0) \leq n$。因此，我们取与所有秩相一致的最大候选规模，

$$\widehat{m} = \max\left\{r \in \{0, \ldots, n\} : r + \ell_k(r) \leq n \text{ for all } 0 \leq k \leq n\right\}, \tag{15}$$

并存储相应的参考界

$$\ell_k = \ell_k(\widehat{m}), \qquad 0 \leq k \leq n. \tag{16}$$

根据引理2，$\ell_k(r)$关于 r单调不增。由于 $m_0 \leq \widehat{m}$ 在不拒绝事件上，$d(S_k) \geq \ell_k(m_0) \geq \ell_k$ 对所有 k均成立，由此得到方程（13）中的联合保证。下述定理表明，对于任意选定集合，通过该参考族进行插值所得的下界与完整闭合检验程序所得的下界完全相同，且不会引入额外的保守性。

定理 3 (SACT下界的参考族表示)。 假设 $0 < \alpha < 1/2$且拟合得分均非零，绝对值两两不同。使用方程 (12)中的完整符号扫描和精确临界值。那么，下界由方程 (16) 给出的参考族满足 $d_\alpha^{\mathrm{RF}}(R) = d_\alpha^{\mathrm{CT}}(R)$ 对每个 $R \subseteq [n]$均成立。

定理 3 中的参考族表示无需枚举所有交集假设，即可恢复完整闭合检验界。在交互式审查过程中，读者可能会多次查询同一份文档，以比较文本片段或修改选择。每次查询都会复用针对该文档一次性计算出的参考集合和下界。具体而言，在完成拟合和校准后，为 n 词元构建参考族需要$O(n^2)$ 时间，而对每次查询评估公式（8）需要 $O(n)$ 时间。因此，

M 集合的总成本为 $O(n^2 + Mn)$。当底层闭合检验程序具有可容许性时，精确等价也会保留可容许性（Goeman 等，2021）。附录 A给出了证明和算法，包括在其校准和单调性条件下对功率散度和 BC 扫描的扩展。

5 结论

我们针对任意非空词元位置集合中保留水印信号的词元比例，构建了同时下置信界。SACT 利用了序贯信号聚类。读者可以检查文本和水印证据，并在保持相同覆盖保证的同时完善其选择。在所述原假设条件下，即使工作模型设定错误，该保证在有限样本中仍然成立。每份文档只需计算一次参考族，审阅者即可针对每种选择高效地恢复完整闭合检验界。在 Gumbel–max 和 SynthID-Text 水印技术下进行的模拟中，结构加权通常改善了段落级界。数据库恢复审查表明，将相关文本片段放在一起评估，可以收紧其保留信号总数的下界。

一个重要的未来方向是扩展原假设模型，以纳入重复播种上下文和编辑所引起的依赖关系。这包括刻画条件符号对称性何时得以保持，开发针对相关原假设得分的校准方法，以及量化偏离对称性时的覆盖误差。进一步的理论研究可以将信号强度、聚类长度和文档长度与检测功效及同时界的紧致性联系起来，从而阐明结构自适应在何种情况下最为有效。对替换、插入、删除和释义改写开展系统研究，将通过评估 SACT 多大程度上能够整合编辑后仍存留的水印证据，为这一分析提供补充。

参考文献

Arias-Castro, E. & Wang, M. (2017), “稀疏异质混合的分布无关检验”, *TEST***26**(1), 71–94。

Barber, R. F. & Candès, E. J. (2015), ‘通过 knockoffs 控制错误发现率’, 《统计年鉴》**43**(5), 2055–2085.

Berk, R. H. & Jones, D. H. (1979), ‘优于柯尔莫哥洛夫统计量的拟合优度检验统计量’,概率论及相关领域杂志**47**, 47–59.

Blanchard, G., Neuvial, P. & Roquain, E. (2020), ‘使用参考族对假阳性进行事后置信界估计’,《统计年鉴》**48**(3), 1281–1303.

Blanchet, J. H., Cai, T. T., Li, X., Liu, H., Long, Q.& Su, W. J. (2026),‘混合来源大语言模型文本中的最优水印定位’, arXiv 预印本 arXiv:2608.14906.

Bonnerjee, S.、Roy, S. 与 Karmakar, S.（2026年），“通过流行病式变点框架对大语言模型生成的带水印文本进行快速分割”，arXiv 预印本，arXiv:2509.21160v2。

Cai, T. T.、Li, X.、Long, Q.、Su, W. J. 与 Wen, G. G.（2025年），“采用伪随机碰撞对语言水印进行最优检测”，arXiv 预印本，arXiv:2510.22007。

Chen, G.（2026年），“CountScan：对所有文本区间中的历史水印匹配给出同时下界”，arXiv 预印本，arXiv:2611.07888v1。

Dathathri, S.、See, A.、Ghaisas, S.、Huang, P.-S.、McAdam, R.、Welbl, J.、Bachani, V.、Kaskasoli, A.、Stanforth, R.、Matejovicova, T.、Hayes, J.、Vyas, N.、Al Merey, M.、Brown-Cohen, J.、Bunel, R.、Balle, B.、Cemgil, T.、Ahmed, Z.、Stacpoole, K.、Shumailov, I.、Baetu, C.、Gowal, S.、Hassabis, D. 和 Kohli, P.（2024），《用于识别大型语言模型输出的可扩展水印技术》，《自然》 **634**，818–823。Dobriban, E.（2020），《针对可交换局部检验的快速闭合检验》，《*Biometrika*》**107**（3），761–768。Donoho, D. 和 Jin, J.（2004），《用于检测稀疏异质混合分布的高阶批判》，《统计年鉴》**32**（3），962–994。Genovese, C.R. 和 Wasserman, L.（2006），《错误发现比例的超越控制》，《美国统计学会杂志》**101**（476），1408–1417。

Goeman, J. J.、Hemerik, J. 和 Solari, A.（2021），《只有闭合检验程序才适用于控制错误发现比例》，《统计年鉴》 **49**（2），1218–1238。

Goeman, J. J., Meijer, R. J., Krebs, T. J. P. & Solari, A.（2019年），《大规模多重假设检验中所有错误发现比例的同时控制》，*Biometrika* **106**(4), 841–856。

Goeman, J. J. & Solari, A.（2011年），《探索性研究中的多重检验》，《统计科学》**26**(4), 584– 597。

Hemerik, J., Solari, A. & Goeman, J. J.（2019年），《基于置换的错误发现比例同时置信界》，*Biometrika* **106**(3), 635–649。

Hu, Z., Chen, L., Wu, X., Wu, Y., Zhang, H. & Huang, H.（2024年），《大语言模型的无偏水印》，载于 《第十二届国际学习表征会议》，第45408–45436页。

Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L. & Naaman, M. (2023年), 与具有观点的语言模型协同写作会影响用户的观点，载于《2023年计算机系统中的人为因素大会论文集》，ACM，第 1–15 页。

Katsevich, E. & Ramdas, A. (2020年), 《结构化、回归和在线情境下错误发现比例的同时高概率界》，《统计年鉴》**48**(6), 3465–3487。

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers,I. & Goldstein, T.(2023年), 大语言模型的水印, 载于《第 40 届国际机器学习会议论文集》，第 202 卷，《机器学习研究论文集》，第 17061–17084 页。

Kirchenbauer，J.、Geiping，J.、Wen，Y.、Shu，M.、Saifullah，K.、Kong，K.、Fernando，K.、Saha，A.、Goldblum，M. 和 Goldstein，T.（2024年），《大语言模型水印的可靠性》，载于《第十二届国际学习表征会议》，第 49660–49704 页。

Kuditipudi，R.、Thickstun，J.、Hashimoto，T. 和 Liang，P.（2024年），《语言模型的稳健无失真水印》，《机器学习研究汇刊》。

Lee, M., Liang, P. & Yang, Q. (2022)，CoAuthor：设计人类—人工智能协同写作数据集以探索语言模型能力，见《2022年计算系统人为因素会议论文集》，ACM，第1–19页。

Lei, L. & Fithian, W. (2018)，“AdaPT：利用侧信息进行多重检验的交互式程序”，英国皇家统计学会会刊：*B*辑（统计方法学）**80**(4)，649–679。

Lei, L.、Ramdas, A. 与 Fithian, W.（2021年），“结构约束下错误发现率控制的通用交互式框架”，*Biometrika* **108**(2)，253–267。

Li, J.、Maathuis, M. H. 与 Goeman, J. J.（2024年），“通过 knockoffs 和闭合检验获得同时错误发现比例界”，英国皇家统计学会杂志*B*辑：统计方法学**86**(4)，966–986。

Li, X.、Li, G. 与 Zhang, X.（2024年），从语言模型中分割带水印文本，载于《神经信息处理系统进展》，第37卷，第14634–14665页。

Li, X.、Liu, X. 与 Li, G.（2025年），“用于分割语言模型生成的带水印文本的自适应检验”，*Stat***14**(4)，e70118。

Li, X.、Ruan, F.、Wang, H.、Long, Q. 与 Su, W. J.（2025年），“大语言模型水印的统计框架：枢轴量、检测效率与最优规则”，《统计年鉴》**53**(1)，322–351。

Li, X., Ruan, F., Wang, H., Long, Q. & Su, W. J.（2026年），《人工编辑下大语言模型水印的稳健检测》，英国皇家统计学会杂志*B*辑：统计方法学**88**(2)，491–515。

Li, X., Wen, G., Chen, X., Long, Q., Jain, A., Joly, F., Lam, M., Song, Q. & Su, W.（2026年），textGrain：经熵校准的语言模型文本水印技术，技术报告，OpenAI。

Li, X., Wen, G., He, W., Wu, J., Long, Q. & Su, W. J.（2025年），《人工智能–人类混合文本中水印比例的最优估计》，arXiv 预印本 arXiv:2506.22343。

Liang, W., Zhang, Y., Wu, Z., Lepp, H., Ji, W., Zhao, X., Cao, H., Liu, S., He, S., Huang, Z., Yang, D., Potts, C., Manning, C.D. & Zou, J.（2025年），《量化大语言模型在科学论文中的使用情况》，*Nature Human Behaviour* **9**(12)，2599–2609。

Magnani, C. G., Sesia, M. & Solari, A.（2026年），《采用保形化闭合检验进行集体异常值检测与枚举》，美国统计学会杂志，第 1–12 页。doi:10.1080/01621459.2026.2699463。

Marcus, R., Peritz, E. & Gabriel, K. R.（1976），《关于闭合检验程序，特别是有序方差分析》，*Biometrika* **63**(3), 655–660。

Noy, S. & Zhang, W.（2023年），‘生成式人工智能生产率效应的实验证据’，*Science* **381**（6654），187–192。

Pan, L., Liu, A., Lu, Y., Gao, Z., Di, Y., Wen, L., King, I. & Yu, P. S.（2025年），WaterSeeker：率先实现大型文档中带水印片段的高效检测，载于‘计算语言学协会：NAACL 2025论文集’，第2866–2882页。

Qwen 团队（2026年），‘Qwen3.5：迈向原生多模态智能体’，Qwen技术博客和Qwen3.5-9B模型卡。

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. & Liu, P. J.（2020年），‘探索统一文本到文本转换器的迁移学习极限’，机器学习研究杂志**21**（140），1–67。

Sander, T., Chang, H., Souček, T., Tran, T., Lacatusu, V., Rebuffi, S.-A., Mourachko, A., Parimi, S., Ropers, C., Moritz, R., Stark, V., Elsahar, H. & Fernandez, P. (2026年), ‘TextSeal：一种用于溯源与蒸馏保护的局部化大语言模型水印’，arXiv 预印本 arXiv:2605.12456。

Sun, W. & Cai, T. T. (2009), ‘依赖关系下的大规模多重检验’，英国皇家统计学会杂志*B*辑：统计方法学 **71**(2), 393–424。

Wei, Z., Bao, Y., Ren, H. & Chen, Z. (2026年), ‘FissionGuard：采用错误发现率控制检验局部水印比例’，arXiv 预印本 arXiv:2610.11625。

Zhao, X., Liao, C., Wang, Y.-X. & Li, L. (2025), 在混合来源文本中高效识别带水印片段，载于‘计算语言学协会第63届年会论文集（第1卷：长篇论文）’，第 6304–6316 页。

Zhao, Z. & Sun, W. (2025), ‘结构化多重检验的错误发现率控制：非对称规则与 conformal Q-值’,《美国统计学会杂志》**120**(550), 805–817。

LLM生成文本中水印比例的同时推断的补充材料

附录 A 阐述了所提出的算法恢复完整闭合检验界的条件。附录B 介绍了案例研究和模拟中所用方法的实现。附录C展示了关于同时覆盖、检测功效以及信号聚类对界的影响的模拟设计与结果。

A 证明

我们首先证明，SACT-2 的参考族界等于完整闭合检验界，并将这一论证推广至其他符号扫描。随后，我们给出一种用于对称单调局部检验的独立算法，并用它计算拟合优度和 Simes 界。

A.1 SACT-2 的参考族表示

对于非空查询集 R，满足方程（13）中联合保证的参考族给出比例和计数下界

$$\underline{\mathrm{TDP}}_{\alpha}^{\mathrm{RF}}(R)=\frac{1}{|R|}\max_{k\in\mathcal{K}}\Big(\ell_k-|S_k\setminus R|\Big)_+,\qquad d_{\alpha}^{\mathrm{RF}}(R)=|R|\underline{\mathrm{TDP}}_{\alpha}^{\mathrm{RF}}(R). \tag{A.1}$$

设 $d_{\alpha}^{\mathrm{RF}}(\varnothing)=0$。在下文的证明和算法中，$[n]$ 表示符合条件的位置。对于原始文档中的选择集 R，将算法应用于其中符合条件的位置$R_e=R\cap\mathcal{E}$。如果 R_e 为空，则认证计数为零。否则，对于 R_e，比例界 L 给出计数界 $|R_e|L$，适用于 R。将此计数除以完整的选择集大小 $|R|$，即可得到任意非空 R的比例界。

构造信号数量下界。 我们采用第 4.4 节的记号，并按照分数绝对值递减的顺序重新标记位置，使得 $S_k=\{i\le k:\ \ \varepsilon_i=1\}$。记

$P = P_n$ 且 $H = Q_n$。固定一个秩 k，令 $q = Q_k$ 表示前 k 个位置中的负分数数量。假设一个大小为 r 的候选零假设集合包含这些位置中的 v 个正分数。该候选集合还剩下 $r - v$ 个位置，因此至多可以包含 u_r 个 $(v,\ q) = \min\{q,\ r - v\}$ 负分数。对于给定的 v，纳入这么多负分数会得到最小的标准化符号和。由于 $v + u_r\,(v,\ q) = \min\{r,\ v + q\}$，满足该约束的最大 v 为

$$G_r(q) = \max\left\{v \in \{0, \ldots, r\} : v < \frac{\min\{r, v+q\} + c_r\sqrt{\min\{r, v+q\}}}{2}\right\}. \tag{A.2}$$

按照约定，令 $v = 0$ $G_0(q) = 0$，并在 $c_r = +\infty$ 时取 $G_r(q) = r$。当$v + u_r(v,\ q) > 0$ 时，同一约束可写为 $(v - u_r(v,\ q))/\sqrt{v + u_r(v,\ q)} < c_r$。从观测到的正得~~分数~~ ~~$P_k$~~ 中减去最大的可行计数 $G_r(Q_k)$，得到$\ell_k(r) = (P_k - G_r(Q_k))_+$。因此，存储的下界为 $\ell_k = (P_k - G_{\widehat{m}}(Q_k))_+$。每个有限的 c_r 都为正，因为当第一个符号函数为正时，$T_r^0 \geq 1$，而该事件的概率为 $1/2 > \alpha$。

对于有限的 $c > 0$，记 $L_c(q) = \lceil q + c^2/2 + c\sqrt{2q + c^2}/4\rceil - 1$。则

$$G_r(q) = \min\left\{r, L_{c_r}(q), \left\lceil\frac{r + c_r\sqrt{r}}{2}\right\rceil - 1\right\}. \tag{A.3}$$

第二项利用了只有 q 个负分数可用这一事实。第三项利用了候选者只有 r 个位置这一事实。对于整数 v，严格界 $v < a$ 等价于 $v \leq \lceil a\rceil - 1$。下面的引理表明，可行的正计数构成一个初始区间。我们将在下文利用这一性质构造一个达到闭合检验界的未拒绝集合。

引理1 (边界计算和初始区间)。对于有限的 $c_r > 0$，方程 (A.3)和 (A.2) 给出相同的 $G_r(q)$值。对于每个 r，q，方程 v 中的可行整数 (A.2)构成一个从零开始的初始区间。

证明。对于 $v+u>0$，比值 $(v-u)/\sqrt{v+u}$ 随 v 增加而增加，随 u 增加而减少。随着 v 增加，$u_r(v,\ q)$ 不会增加。一旦 v 的某个取值违反约束，所有更大的取值也都会违反约束。这就证明了初始区间性质，并按照约定将空状态包括在内。

为得到显式边界，先允许所有 q 负得分位置。对于 $v>q$，求解 $v-q<c\sqrt{v+q}$ 可得 $v<q+c^2/2+c\sqrt{2q+c^2/4}$；对于 $v\le q$，除空状态外，不等式自动成立。因此，最大的可行整数是 $L_c(q)$。对于 $r>0$ 和 $v+q>0$，施加第二个限制，即最多允许 $r-v$ 个负得分位置，可得

$$\frac{v-\min(q,r-v)}{\sqrt{v+\min(q,r-v)}}=\max\left\{\frac{v-q}{\sqrt{v+q}},\frac{2v-r}{\sqrt{r}}\right\}.$$

因此，当且仅当右侧两项都小于 c_r 时，该约束成立；等价地，

$$v\le L_{c_r}(q),\qquad v<\frac{r+c_r\sqrt{r}}{2},$$

连同 $v\le r$，即可证明等式（A.3）。空状态约定与这些非负界一致。

□

引理 2（信号计数下界的单调性）。采用等式（12）中的 $\ell_k(r)\ge\ell_k(r+1)$，对所有 $0\le k\le n$ 和 $0\le r<n$ 均成立。

引理 2 的证明。由于 $\ell_k(r)=(P_k-G_r(Q_k))_+$，只需证明对每个 q，$G_r(q)\le G_{r+1}(q)$。固定一个规模为 $r\ge 1$ 的可行 v，并令 $u=\min(q,r-v)$。空状态仍然可行。否则，令 $s=(v-u)/\sqrt{v+u}$。若 $s\le 0$，增大 r 至少能容纳同样多的负得分位置，且不会增大比值，因此，由每个有限截断值均为正可知，可行性仍然成立。

对于 $s>0$，考虑一个长度为 r 的序列：先是 u 个负得分位置，接着是 v 个正得分位置，最后是 $r-u-v$ 个负得分位置。其扫描最大值为 s：正值块将标准化和增加到

s，最后的负得分位置会使其减小。再添加一个负得分位置后，这个最大值仍然不变。因此， s 对 T_r^0 和 T_{r+1}^0 都是可取值。

由于 $s < c_r$ 是可达到的， c_r 的规范定义意味着 $\Pr(T_r^0 \geq s) > \alpha$，即使在 $c_r = +\infty$时也是如此。通过前 r 个符号将两个零假设随机游走耦合起来，可得 $T_{r+1}^0 \geq T_r^0$，因此

$$s < c_r \implies \Pr(T_{r+1}^0 \geq s) \geq \Pr(T_r^0 \geq s) > \alpha \implies s < c_{r+1}.$$

最后一个蕴含关系来自截断值的定义，因为长度为 $r+1$的随机游走可以达到 s。在规模为 $r+1$ 时， v 的比率不大于 s，因此 v 仍然可行。于是$G_r(q) \leq G_{r+1}(q)$，且信号计数下界不递增； $r = 0$ 的情形是显然的。

□

与完整闭合检验的等价性。

定理3的证明。 我们首先证明，每个局部未拒绝集合的补集都满足存储的信号计数下界，从而得到这两个程序之间的一个不等式。随后，我们构造一个未拒绝集合，使其对每个查询都达到参考族界，从而证明反向不等式。

步骤 *1*：每个未拒绝候选都满足参考约束。令 $A = \{I \subseteq [n] : T_I < c_{|I|}\}$。固定 $I \in A$ 中大小为 r的集合。在前 k 个排序位置中，令

$$v = |I \cap S_k|, \qquad u = |\{i \in I : i \leq k,\ \varepsilon_i = -1\}|.$$

在 I 的排序中，这些是前 $v+u$ 个分数。如果 $v+u > 0$，它们的标准化符号和为 $(v-u)/\sqrt{v+u}$。由于 I 未被拒绝，它低于 c_r。将 u 替换为 $\min(Q_k, r-v)$ 后，总和仍低于 c_r，因此 $v \leq G_r(Q_k)$。当 $v = u = 0$ 时，该界也成立。因此，

$$\ell_k(r) = (P_k - G_r(Q_k))_+ \leq P_k - v = |S_k \setminus I| \leq n - r.$$

因此 $r + \ell_k(r) \le n$ 对每个 k均成立，因此 $r \le \widehat{m}$。引理 2 给出 $|S_k \setminus I| \ge \ell_k(r) \ge \ell_k$。对于每个查询 R 以及每个 k,

$$|R \setminus I| \ge |S_k \setminus I| - |S_k \setminus R| \ge \ell_k - |S_k \setminus R|.$$

对未被拒绝的 I 取最小值即可证明

$$d_\alpha^{\mathrm{CT}}(R) \ge d_\alpha^{\mathrm{RF}}(R). \tag{A.4}$$

步骤 *2*：在最大可行规模下，确定要选择多少个正分数。设$r_* = \widehat{m}$ 和 $p_* = r_* - H$。所有负得分位置组成的集合未被局部拒绝，因此步骤 **1** 给出 $r_* \ge H$ 和 $p_* \ge 0$。令 $S_+ = \{i : \varepsilon_i = 1\}$。对于 $A \subseteq S_+$，定义

$$K(A) = \min\left\{|A|, \min_{0 \le k \le n}\Big(|A \setminus S_k| + P_k - \ell_k\Big)\right\}. \tag{A.5}$$

r_* 的可行性给出

$$p_* \le P - \ell_k \quad \text{for every } k,$$

因此 $K(S_+) \ge p_*$。若 $r_* < n$ 且 $K(S_+) > p_*$，则由整数性可得，对于每个 k, $P - \ell_k \ge p_* + 1$。由单调性可知

$$r_* + 1 + \ell_k(r_* + 1) \le r_* + 1 + \ell_k \le n,$$

因此 $r_* + 1$ 可行，这就产生矛盾。若 $r_* = n$，则 $K(S_+) \le P = p_*$。无论哪种情况，$K(S_+) = p_*$。K 的公式关于包含关系是单调的，因此对于每个 $A \subseteq S_+$，$0 \le K(A) \le p_*$。

步骤 *3*：为每个查询构造一个未拒绝集合。固定 R 并设置 $A = R \cap S_+$ 和$K = K(A)$。包含所有具有负分数的 H 位置。在 A中的位置里，选择得分幅度最小的 K。用仍可用的最小正分数填补剩余位置，直到集合包含具有正分数的 p_* 个位置。这是可行的，因为 $K \le p_* \le P$。令 I_R 表示由此得到的、大小为 r_*的集合。

在排名靠前的 k 个位置中，为 I_R 选择的正分数数量为

$$|I_R \cap S_k| = \max\left\{[K - |A \setminus S_k|]_+,\ [p_* - (P - P_k)]_+\right\}. \tag{A.6}$$

在秩 k 之后，共有 $|A \setminus S_k|$ 个位置来自 A，以及 $P - P_k$ 个正分数。因而，要从 A 中选择 K 个位置，我们至少需要秩 1 至k中的第一项。要总共选择 p_* 个正分数，则至少需要这些秩中的第二项。优先选择最小的分数，恰好得到这两个计数中较大的一个。

方程（A.6）中的第一项根据方程（ $P_k - \ell_k$ 至多为A.5）；第二项则因 r_* 的可行性而满足相同的不等式。因此 $v = |I_R \cap S_k| \le P_k - \ell_k =$ 最小值$\{P_k,\ G_{r_*}(Q_k)\}$。由于 $v \le p_*$,

$$r_* - v \ge H \ge Q_k, \qquad u_{r_*}(v, Q_k) = Q_k.$$

在前 k 个位置中， I_R 包含 v 个正分数和全部 Q_k 个负得分位置。若$v + Q_k > 0$，则在秩 $v + Q_k$ 处对 I_R 的扫描值为（$v - Q_k$）$/\sqrt{v + Q_k}$。根据引理1，这小于 ~~c_{r_*}~~，因为 $v \le G_{r_*}$（ Q_k ）。由于 k 从 0 运行到 n，每当我们到达 I_R 中的一个位置时， $v + Q_k$ 就加一。这涵盖了扫描中的每个秩，因此 $I_R \in$A。若 $r_* = 0$，则按照约定，空集属于 A。

步骤*4*：证明所构造的集合达到了界。 集合 I_R 包含 R 中的每个负得分位置，以及其至少 K 个正得分位置。因此

$$d_\alpha^{\mathrm{CT}}(R) \le |R| - |I_R \cap R| \le |A| - K.$$

根据公式（A.5）和 $A = R \cap S_+$， $|A| - K =$ 最大值$_k$（$\ell_k - |S_k \setminus R|$）$_+ = d_\alpha^{\mathrm{RF}}$（$R$）。将其与公式（A.4）结合，即可证明其与完全闭包等价，并表明 I_R 达到了界。根据步骤 1，每个未拒绝集合至多包含 r_* 个元素。步骤 3 中构造的集合恰好包含 r_*，因此 $r_* =$ 最大值$_{I \in \mathcal{A}} |I|$。最后，单调性给出 最小值$_{r \le \widehat{m}}\ell_k$（$r$） $= \ell_k$（$\widehat{m}$）。

□

扫描值等于截断值时会导致拒绝，这一点体现在方程（A.2）中的严格不等式中。如果 $c_r = a/\sqrt{b} > 0$ 且 $v > u$，则该秩处的扫描值低于

当且仅当 $b(v-u)^2 < a^2(v+u)$ 时，截断值成立。可以使用整数运算或经认证运算进行比较，以避免截断值处的舍入误差。复杂度陈述按算术运算次数计算。

预处理与重复查询。 算法 1实现精确表示。

它只需计算一次排序和参考族下界，然后将其复用于每个选定令牌集。

算法 1 SACT-2：预处理与重复查询

输入： 具有不同幅值的非零分数；精确截断点 $c_0, \ldots, c_n$。

过程 预处理$(X_{1:n},\ c_{0:n})$

按递减绝对值对分数进行排序，并计算累积计数 P_k 和 Q_k 的正分数和负分数。

设定 $\widehat{m} \leftarrow 0$。

对于 $r = 1, \ldots, n$ **do**

计算 $\ell_k(r)$ 对所有 k 根据公式（14）及其约定。

如果 $r + \ell_k(r) \le n$ 对于每个 $k = 0, \ldots, n$ **那么**

设置 $\widehat{m} \leftarrow r$。

结束 if

结束 for

为 $k = 0$、……、n 设置 $\ell_k \leftarrow \ell_k(\widehat{m})$。

返回 $(\pi、\ \varepsilon、\ \ell)$，其中 $\varepsilon_j =$ 符号函数(X_j)。

结束过程

过程 查询$(R, \pi, \varepsilon, \ell)$

为 R 构造成员掩码；设置 $a \leftarrow |R|$、$C \leftarrow 0$ 和 $d \leftarrow 0$。

对于 $k = 1, \ldots, n$ **do**

如果 $\varepsilon_{\pi(k)} = 1$ 且 $\pi(k) \notin R$ **则**

设置 $C \leftarrow C + 1$。

结束条件判断

设置 $d \leftarrow$ 最大值$\{d, \ell_k - C\}$。

结束循环

返回$(d, d/(a \vee 1))$。

结束过程

A.2 对其他符号扫描的扩展

我们将参考族构造扩展到功率散度扫描和 Barber–Candès 扫描。

A.2.1 幂散度扫描

对于公式（B.6）中的功率散度扫描，我们对每种扫描采用其自身的信号计数下界，并应用第 4.4 节中的构造。对于大小为 r 的候选集合，考虑幅度排序中的前 k 个位置。令 $u = \min\{Q_k, r - v$，并使用二分搜索找到最大的正计数 $0 \le v \le r$，使其贡献值（$v + u$）$d_s\{v/(v + u), 1/2\}$ 低于检验大小为 r 的集合时的截断值；当 $v \le u$ 时，将贡献值视为零。该搜索成立，是因为贡献值关于 v 单调不减。当 $u = Q_k$ 时，增加一个正值会同时提高总计数和正值比例。一旦 $u = r - v$，总计数便保持为 r，而正值比例继续增加。如果最大的可行计数为 $v_{\max}$，则信号数量下界为 $\ell_k(r) = (P_k - v_{\max})_+$。如果对于每个秩 k，$\ell_k(r)$ 关于 r 单调不增，则附录 A.1 中的论证表明，参考族给出的界与完整闭合检验相同。

A.2.2 Barber–Candès程序

令 c_r^{BC}为长度为r 的公平符号BC扫描支持集中的最小点，其包含式上尾概率至多为 α；若不存在这样的点，则令其为 $+\infty$。设定 $c_0^{\mathrm{BC}} = +\infty$。定义

$$G_r^{\mathrm{BC}}(q) = \max\left\{v \in \{0, \dots, r\} : \frac{v}{1 + \min(q, r - v)} < c_r^{\mathrm{BC}}\right\},$$

其中 $G_0^{\mathrm{BC}}(q) = 0$ 且对于无限截断值，$G_r^{\mathrm{BC}}(q) = r$。对于有限截断值 $c = c_r^{\mathrm{BC}}$，

$$G_r^{\mathrm{BC}}(q) = \min\left\{r, \lceil c(1 + q)\rceil - 1, \left\lceil \frac{c(r + 1)}{1 + c}\right\rceil - 1\right\}. \tag{A.7}$$

命题*1*（*BC*表示与复杂度）。对于幅值不同的非零分数以及 $0 < \alpha < 1/2$，在公式（*15*）–（*16*）中用 $\ell_k^{\mathrm{BC}}(r) = (P_k - G_r^{\mathrm{BC}}(Q_k))_+$ 替代 $\ell_k(r)$，可得到一个参考族，其插值等于采用精确*BC*局部检验时的完全闭包。这些信号计数下界关于 r 单调不增。构造该参考族需要 $O(n^2)$ 时间，而计算其针对一个查询集的界需要 $O(n)$ 时间。

证明。 可行的正计数构成一个初始区间，因为 $v/(1+u)$ 随 v 增加而随 u 减少。恒等式

$$\frac{v}{1+\min(q, r-v)} = \max\left\{\frac{v}{1+q}, \frac{v}{1+r-v}\right\}$$

给出了约束条件 $v < c\,(1+q)$ 和 $v < c\,(r+1)\,/\,(1+c)$，从而证明了公式（A.7）。

对于可行的 $v > 0$，令 $u = \min(q, r-v)$。由 u 负得分位置、v 正得分位置和 $r-u-v$ 负得分位置组成的符号序列具有 BC 最大值 $s = v/\,(1+u)$，追加一个负得分位置不会改变该值。引理 2 中的校准论证给出 $s < c^{\mathrm{BC}}_{r+1}$。在规模 $r+1$ 下，我们至少可以纳入同样多的负得分位置，因此 v 仍然可行。按照约定，计数 $v = 0$ 是可行的。

这些论证给出了定理 3证明中使用的两个性质：可行计数构成初始区间，并且信号计数下界随候选规模增加而非递增。因此，我们可以使用 G_r^{BC}代替 G_r来构造相同的未拒绝集合。这证明了其与完全闭包相等。最后，公式（A.7）在常数时间内计算每个下界，因此预处理和查询成本保持不变。

□

A.3 对称坐标逐点不减局部检验的闭合检验界

对于第 3.7 节中的拟合优度和 Simes 检验，我们使用算法 2 计算每个查询的闭合检验界，而无需构造参考族。这两种局部检验在证据分数 $X_j = -p_j$ 上均为置换不变且坐标逐点不减，因为降低 p-值不会影响任何拒绝。以下给出的条件下，该算法也适用于基于符号函数的检验。对于此处考虑的检验，每个查询所需时间为 $O(n^2)$。对于使用 SACT-2 的重复查询，算法 1 更高效，因为它只需花费 $O(n^2)$ 的时间构造一次参考族，然后用 $O(n)$ 的时间评估每个查询。

算法 2 查找查询 R 与交集之间的最大重叠，该交集

算法2 精确的检验界 输入：固定证据得分 $X_{1:n}$；非空 R；局部检验 $\phi_I = h_{|I|}(X_I)$，其具有对称的、坐标单调的 h_m。输出：闭合检验下置信界 $\mathbf{TDP}_\alpha^{\mathrm{CT}}(R)$。设置 $r \leftarrow |R|$, $m \leftarrow n$ 和 $b \leftarrow 0$。将 $[n]$ 排序为 $(w_1, \ldots, w_n)$，使得 $X_{w_1} \leq \cdots \leq X_{w_n}$。令 $(u_1, \ldots, u_r)$ 为属于 R 的 $(w_1, \ldots, w_n)$ 的子序列。当 $b < \min(r, m)$ 时，设置 $s \leftarrow b+1$ 和 $L_{m,s} \leftarrow \{u_1, \ldots, u_s\}$。将 $m-s$ 个得分最小的位置从 $[n] \setminus L_m$ 加入 $L_{m,s}$。如果 $\phi_{L_{m,s}} = 1$，则 $m \leftarrow m-1$ 否则 $b \leftarrow |L_{m,s} \cap R|$ 返回 $1 - b/r$

未被局部拒绝。为组织这一搜索，固定一个非空的 R，满足 $r = |R|$，并定义

$$\mathcal{C}_{m,s}(R) = \{I \subseteq [n] : |I| = m,\ |I \cap R| \geq s\}, \qquad 1 \leq s \leq \min(m, r).$$

构造 $L_{m,\ s}$ 首先取 s 中幅值最小的证据分数 R，然后在所有剩余位置中加入 $m-s$ 分数最小的位置。其中一些新增位置也可能属于 R，因此最终集合可能与 R 重叠超过 s 个位置。在每个秩上，$L_{m\text{中的分数}s}$ 在该族集合中是可能的最小值。因而检验 $L_{m,\ s}$ 即可确定该族是否包含未拒绝集合。这是 FACT（Dobriban 2020）所依据的对称性与单调性原理。

引理3（对称性与单调性）。对于固定大小 m 和截断值，基于公式（6）、（B.6）和（B.7）的*sign*扫描、功率散度和*BC*规则，在幅值两两不同的非零得分向量上均具有置换不变性和坐标逐点不减性。

证明。固定规模 m 及其截断值。不同幅值确定唯一的有序符号序列，与坐标标签无关，因此具有置换不变性。

现在增加一个 带符号得分，同时保持其他项不变。如果它为负，其幅值

会减小，并在排序中后移；如果为正，则会前移。与得分符号相同的项交换，不会改变符号序列。符号相反的项交换时，$(-1,\ +1)$会替换为$(+1,\ -1)$，使两个秩中较前者的累积正号计数增加，而其他所有秩上的计数保持不变。如果得分越过零点，它会在最后一个秩处由负得分位置变为正。由此，对于每个 k，前 k 个秩中的正号数量都不会减少。

在秩 k 处，扫描统计量的贡献随前 k 个得分中正号的数量增加而不减。对于标准化符号和与 BC 比率，这一结论直接成立；对于高于阈值 $k/2$ 的幂散度贡献也成立，而低于该阈值时贡献为零。对 k 取最大值并应用固定阈值，仍能保持单调性。

□

定理 4 (单个查询的精确计算)。 固定证据分数 $X_{1:n}$ 和局部规则$\phi_I = h_{|I|}(X_I)$，它们具有置换不变性且坐标逐点不减。算法 2返回对应于方程 $d_\alpha^{\mathrm{CT}}(R)/|R|$ 的比例界 $d_\alpha^{\mathrm{CT}}(R)/|R|$，即方程 (7)，至多需要 n次局部检验评估。对于符号检验、功率散度检验、*BC* 检验、拟合优度检验和 *Simes* 检验，拟合与校准后，每个查询的计算成本为$O(n^2)$。

证明。我们首先证明，L_m 的排序得分在集合 C$_{m,s}$(R) 中是 $_s$ 坐标逐点最小的。对于阈值 t，记 $a(t) = |\{j \in R:\ X_j \le t\}|$，且 $c(t) = |\{j \notin R : X_j \le t\}|$。C$_{m,s}$($R$) 中的元素至多包含 $a(t)$ 个不大于 t 的 $+\,c(t)$ 得分。此外，它必须选取至少 $[s - a(t)$ 个得分大于 t 的查询位置，因此不大于 t 的位置至多有 $m - [s - a(t)]_+$ 个。对于每个 t，$L_{m,s}$ 的构造都达到了这两个上限中的较小值，因此每个 $I \in$C$_{m,s}$(R) 都满足

$|\{j \in I:\ X_j \le t\}| \le$ 最小$\{a(t) + c(t)\ \ m - [s - a(t)]_+\} = |\{j \in L_{m,\ s}:\ X_j \le t\}|$.因此，

e $L_{m,\ s}$ 的排序得分按坐标均不大于 I的排序得分。

对称性和单调性意味着 $\phi_{L_{m,\ s}} \leq \phi_I$。由于 $L_{m,\ s}$ 本身属于该族，

$$\exists I \in \mathcal{C}_{m,s}(R) : \phi_I = 0 \quad \Longleftrightarrow \quad \phi_{L_{m,s}} = 0. \tag{A.8}$$

接下来，令 $b^* = \max_{\phi_I=0} |I \cap R|$ 表示算法要寻找的重叠。在每次迭代中，b 是未拒绝集合已达到的重叠。任何重叠更大的未拒绝集合，其大小都必须至多为 m。初始时 $m = n$，且空集达到 $b = 0$。由于查询本身只有 r 个位置，因此可能达到的最大重叠满足

$$b \leq b^* \leq \max\{b, \min(m, r)\}.$$

如果局部检验拒绝 $L_{m,\ b+1}$，式（A.8）便排除了所有大小为 m 且重叠大于 b 的集合，因此我们可以减小 m。否则，我们将 b 更新为该集合的重叠，$|L_{m,\ b+1} \cap R| \geq b + 1$。这两种更新都会保持 b^* 的这些界。算法终止时，要么$b = r$，要么 $m \leq b$；没有剩余集合能够提高已达到的重叠，而给出的界表明 $b = b^*$。因此，返回 $1 - b/r$ 即可得到闭合检验比例界。最后，每次检验都会使非负整数 $m - b$ 至少减少 1；其初始值为 n，因此最多需要进行 n 次评估。对证据分数排序，以及对于基于符号函数的检验对其幅值排序，需要花费 $O(n \log n)$。给定这些排序后，对于上述检验，构造一个代表集合并评估其局部检验，各自需要 $O(n)$ 时间。因此，总成本为 $O(n \log n + n^2) = O(n^2)$，不包括拟合和校准。

□

B 技术细节

我们介绍如何拟合 Gumbel–max 分数的工作模型并校准局部检验。随后，我们给出 SynthID-Text 对应的分数和工作模型，并说明如何在数据库恢复审查中，在阅读预算范围内搜索文本片段。

B.1 工作模型的设定与估计

对于 Gumbel–max 水印，隐马尔可夫模型和变点模型对分数绝对值采用相同的发射分布族。我们推导出这一分布族，并说明每个模型如何估计公式（5）中的权重。

B.1.1 发射分布

对于 Gumbel–max 水印技术，$U_j = F_0(Y_j) = Y_j$ 在原假设下服从均匀分布。在备择假设下，对于固定的下一词元分布 P_j，对方程（10求导可得枢轴密度

$$h_{P_j}(u) = \sum_{w:P_{j,w}>0} u^{1/P_{j,w}-1} = \int_{[1,\infty)} u^{x-1}\, N_{P_j}(dx), \qquad N_{P_j} = \sum_{w:P_{j,w}>0} \delta_{1/P_{j,w}}. \tag{B.1}$$

这里 δ_x 表示在 x 处的单位点质量。每一项都是 $\mathrm{Beta}(1/P_{j,w},\ 1)$ 密度乘以其混合权重 $P_{j,w}$。因此，h_{P_j} 是概率密度，尽管 N_{P_j} 是计数测度。验证器通常不知道下一词元概率，而且这些概率可能随位置而变化。我们将方程（B.1）中的计数测度替换为连续混合分布，并对核函数进行归一化。取极限后即可得到一个仅含单个信号强度参数的工作族。

对于 G 在 $[1,\infty)$ 上的概率律，替换计数测度并进行归一化后可得

$$h_G(u) = \frac{\int u^{x-1} G(dx)}{\int x^{-1} G(dx)} = \frac{\mathbb{E}_G(u^{X-1})}{\mathbb{E}_G(X^{-1})}.$$

对于 $X = 1 + S$wi th $S \sim$ Gamma(k, 率 $= r$)，$\mathbb{E}(u^S) = \{r/(r -$ 对数 $u)\}^k$。约去 r^k 得到

$$h_{k,r}(u) = \frac{(r - \log u)^{-k}}{\int_0^1 (r - \log v)^{-k}\, dv}, \qquad 0 < u < 1.$$

对于 $0 < k < 1$，代入 $t =$ –log u 可得

$$\int_0^1 (r - \log u)^{-k}\, du \longrightarrow \int_0^\infty t^{-k} e^{-t}\, dt = \Gamma(1-k) \qquad \text{as } r \downarrow 0.$$

设定 $\nu = 1 - k$ 并纳入均匀端点 $\nu = 1$ 可得到方程（4）中的工作族。验证器无需知道下一词元概率即可拟合该工作族。拟合不佳可能会影响检验效能。定理 2 保证在条件零模型下，即使工作族拟合不佳，也能实现同时覆盖。

伽马工作族的枢轴密度为 $h_\nu(u) = (-\log u)^{\nu-1}/\Gamma(\nu)$，其中 $0 < u < 1$。要获得 $A_j = |Z_j|$ 在 $a > 0$处的密度，我们将 Z_j 在 a 处和 $-a$的密度相加。在工作备择下，带符号得分的密度为 $\phi(z)h_\nu\{\Phi(z)\}$。两个幅值密度为

$$g_0(a) = 2\phi(a), \qquad g_1(a;\nu) = \phi(a)\Big[h_\nu\{\Phi(a)\} + h_\nu\{\Phi(-a)\}\Big], \qquad a \geq 0. \tag{B.2}$$

这里 ϕ 表示标准正态密度。为拟合任一结构模型，我们使用发射对数似然比

$$\ell_j(\nu) = \log \frac{g_1(A_j;\nu)}{g_0(A_j)} = \log\left[\frac{h_\nu\{\Phi(A_j)\} + h_\nu\{\Phi(-A_j)\}}{2}\right].$$

在不合格位置，我们将该对数似然比设为零，因此缺失观测不会偏向任一工作状态。这些位置仍保留在序列和查询分母中，但不纳入符号扫描。

B.1.2 隐马尔可夫模型

我们从其平稳分布 $(1 - \pi, \pi)$ 开始马尔可夫链。这里 π 是活跃状态的平稳概率，而 d 是活跃区段的平均长度。其转移矩阵为

$$\mathbf{P}(\pi, d) = \begin{pmatrix} 1 - \dfrac{\pi}{d(1-\pi)} & \dfrac{\pi}{d(1-\pi)} \\[2ex] \dfrac{1}{d} & 1 - \dfrac{1}{d} \end{pmatrix}, \tag{B.3}$$

其中 $d > 1$ 和 $0 < \pi < d/(d+1)$，因此所有转移概率都严格介于零和一之间。幅度发射见公式（B.2）。

我们通过以下方式对转移参数进行正则化

$$\operatorname{logit}(\pi) \sim N\{\operatorname{logit}(0.25), 1\}, \qquad \log d \sim N\{\log 32, (\log 4)^2\}, \tag{B.4}$$

在截断到有效区域之前独立进行。我们使用张量积高斯–厄米特求积规则对这一连续转移先验进行积分。

我们通过最大化观测幅度的边际似然来估计 $\nu \in [0.001, 1]$。我们使用前向–后向递推对状态路径求和，并在每个坐标上使用七阶求积对两个转移参数进行积分。我们舍弃参数空间之外的求积节点，并对剩余的先验权重重新归一化。我们从多个初始值出发，在 $\log \nu$ 上运行 L-BFGS-B，并保留似然最高的拟合结果。为计算梯度，我们使用后验状态概率对发射导数进行加权。在进行 B 次似然评估并保留 Q 个张量积节点的情况下，拟合成本为 $O(BQn)$。

令 $(\pi_b、\ d_b)、\ b = 1、\ldots\ldots、\ Q$为求积节点，并令 $\widehat{\omega}_b$为其拟合后的后验权重。对这些节点取平均，可得到公式（5）中使用的活跃状态概率，

$$q_j = \sum_{b=1}^{Q} \widehat{\omega}_b \Pr_{\widehat{\nu},\pi_b,d_b}(G_j = 1 \mid A_{1:n}). \tag{B.5}$$

转移节点位于内部，且发射密度为正，因此 $q_j \in (0,\ 1)$。根据连续枢轴分布，零假设得分幅度 $q_j A_j$ 因此几乎必然为正。

B.1.3 变点模型

令 $K(G)$统计长度为 n的二进制路径中的非空活动段。我们采用截断至 $0 \le K \le K_{\max}$的泊松(λ)先验，其中 $K_{\max} = 32 \wedge \lfloor (n+1)/2 \rfloor$，并为具有相同 K的所有路径赋予等概率。此类路径共有 $\binom{n+1}{2K}$ 条，其中包括接触

文档边界的区段。因此

$$\Pr(G=g)=\frac{p_{K(g)}}{\binom{n+1}{2K(g)}},\qquad p_k=\frac{\lambda^k/k!}{\sum_{h=0}^{K_{\max}}\lambda^h/h!}.$$

全原假设路径具有 $K=0$。我们在主要 CP 分析中使用 $\lambda=2$，并使用 $\lambda=0.5$ 重复分析，以评估先验的敏感性。

对于发射对数比 $\ell_j(\nu)$，令 $F_j(s,\ k)$ 对以状态s 结束且具有 k 个活跃区段的路径的似然比求和。将 $F_0(0,\ 0)\ =1$ 初始化，并将所有其他条目设为零。递推关系为

$$F_j(0,k)=F_{j-1}(0,k)+F_{j-1}(1,k),$$

$$F_j(1,k)=e^{\ell_j(\nu)}\{F_{j-1}(1,k)+F_{j-1}(0,k-1)\}.$$

索引无效的项均为零。边际似然比为

$$L_{\mathrm{CP}}(\nu)=\sum_{k=0}^{K_{\max}}\frac{p_k}{\binom{n+1}{2k}}\{F_n(0,k)+F_n(1,k)\}.$$

随后通过反向递推得到区间归属的后验概率以及 K的后验分布。我们在对数尺度上计算递推，并使用与隐马尔可夫模型相同的参数边界和优化方法，对 ν 进行最大化。每次目标函数和梯度计算的成本为 $O(nK_{\max})$。两个模型都为方程（ q_j 5）中的得分提供。

B.2 局部检验与校准

我们将第3.7节中使用的程序分为基于符号的检验、基于 p值的检验，以及 CountScan 的基于区间的构造。这些程序均使用相同的验证统计量和候选查询集。对于经过检验的集合 I，记 $m=|I|$。

B.2.1 基于符号的检验

利用公式（5）中的拟合得分，我们采用第 4.2 节中的条件符号翻转分布校准局部检验。在 I 内，按递减绝对值对得分排序。令 $P_{I,k}$ 表示前 k 个得分中正号的数量。

幂散度扫描。 在每个秩 k 处，我们将正号的比例 $P_{I,k}/k$ 与 $1/2$ 进行比较。我们使用幂散度函数

$$d_s(u,v) = \frac{u^s v^{1-s} + (1-u)^s (1-v)^{1-s} - 1}{s(s-1)},$$

其在 $s = 1$ 处具有连续极限。我们取正号占多数的秩中的最大值，得到

$$T_{I,s} = \max_{k: P_{I,k} > k/2} k\, d_s(P_{I,k}/k, 1/2), \qquad s > 0. \tag{B.6}$$

空最大值为零。 $s = 1$ 情形给出了对排序符号进行的 Berk–Jones扫描（Berk 和 Jones 1979）。在 $s = 2$ 时，统计量等于 $\{\max(T_I, 0)\}^2/2$，从而得到 SACT-2。我们还评估 $s = 0.5$、1、1.5、4，而 $s = 3$ 与 $s = 2$一致。SACT-Z 和 SACT-q保留 $s = 2$ 检验，但分别使用 Z_j 和 $\mathrm{sign}(Z_j)q_j$。

Barber–Candès程序。

基于 Barber–Candès程序（Barber 和 Candès 2015）的扫描为

$$T_{I,\mathrm{BC}} = \max_{1 \le k \le m} \frac{P_{I,k}}{1 + k - P_{I,k}}. \tag{B.7}$$

对于每个交集大小 m，我们使用可达到的最小截断值，使原假设上尾概率至多为 0。05；如果不存在这样的截断值，则使用 $+\infty$。我们通过递归计算该概率，在依次加入 m 个独立公平符号时，对正号进行计数。所得闭合检验界使用附录A.1 和A.2中的参考族算法。并列情况按位置解决，与符号无关。在计算时，我们赋予与该顺序一致的不同幅度。

KJI。 KJI 在预先指定的检查点检验相同的正号累积计数 $P_{I,\ k}$

(Li、Maathuis 和 Goeman 2024年)，从而得到

$$\phi_I^{\mathrm{KJI}} = \mathbf{1}\{\exists v:\ P_{I,\min(m,k_v+v-1)} \ge k_v\},$$

各交集使用固定的阈值。网格 A 使用 $v_i = i$。网格 B 从 $v_1 = 1$开始，并使用 $v_i = \lfloor i^2/2 \rfloor$ 作为 $i \geq 2$。网格 C 使用从 (1,2)开始的斐波那契数列，而网格 D 使用 $v_i = 2^{i-1}$。在每个网格中，我们仅保留满足 $v_i \leq \lfloor n/3 \rfloor$的检查点。我们使用 Li、Maathuis 和 Goeman（Li、Maathuis 和 Goeman (2024年）的算法 2，模拟 200,000 次公平符号游走，目标穿越概率为 0。045。随后，我们精确计算穿越概率并调整阈值，以确保显著性水平为 0。05。与 SACT 不同，KJI 使用预先指定的检查点，并在不同交集大小之间共享阈值。

为便于计算，令 A_v 包含按绝对值递减顺序排列的第 v 个负得分位置之前的正得分位置；如果负得分位置少于 v 个，则包含所有正得分位置。预先计算这些集合可得到 $d_\alpha^{\mathrm{KJI}}(R) = \max\{0, \max_v(|R \cap A_v| - k_v + 1)\}$，这是 Li 、 Maathuis 和 Goeman(2024 年) 的定理 4.1 所给出的完全闭包的精确表示。

B.2.2 基于 p 值的检验

使用上文定义的幂散度函数 d_s ，拟合优度评估经过排序的 p 值，从I 到

$$G_{I,s} = m \max_{k:k/m \geq p_{(k)}} d_s(k/m, p_{(k)}), \qquad s \in \{1, 1.5, 2\},$$

包括 $k = m$。如果 $p_{(m)} = 1$，则该最后秩贡献 d_s（1，1） $= 0$。当观测统计量超过第 r 大者（在 $M = 200,000$ 个零假设校准统计量中）时，拟合优度拒绝，其中 $r = \lfloor 0.05(M + 1) \rfloor$。校准样本彼此独立，且独立于实验文档。对随机校准样本取平均后，零假设拒绝概率至多为0.05 。当 $\min_k mp_{(k)}/k \leq \alpha$ 时，Simes 拒绝。对于这两种方法，完整闭合检验界由算法2计算，并使用证据分数$-p_j$。

B.2.3 CountScan

我们使用截断为 2.5、对冲权重为 0.5 的 CountScan C1 证据函数、共同的对数证据阈值 $\log(n/\alpha)$，以及通过最大化每个查询集中所含不相交区间上原始界之和而得到的界（Chen2026）。此处 n 是过滤前的完整文档长度。对于一般集合，我们按其保留的文档顺序，将符合条件的位置拆分为连续段，用动态规划计算每段内的最优区间聚合，再将结果求和。排序决定哪些位置进入每个集合。在集合内部，扫描使用文档顺序。所有方法使用相同的枢轴、符合条件的位置和排序。每个比例的分母包括选定集合中的全部词元。比较针对的是保留信号标签，而非 CountScan 的全键历史匹配目标。对于 SynthID-Text 和高斯实验，同一 C1 构造在实验的条件零模型下使用保守的二项分布与均匀原假设 p 值。

B.3 SynthID-Text 水印的扩展

对于 SynthID-Text，我们将 Gumbel–max 枢轴和 Gamma 发射替换为离散验证分数和二项工作分布。

在Dathathri 等人（2024年）的二元锦标赛中，密钥会针对每个播种上下文和层，为候选词元分配相互独立且公平的比特$g_{j,\,\ell}(w)$。同一候选词元的重复出现共享同一个比特。每一层将候选词元两两配对，优先选择比特为 1 的候选词元，若平局则均匀随机打破平局。从 $p_j^{(0)}{}_{,\,w} = P_{j,\,w}$ 开始，获胜者分布满足

$$
\begin{aligned}
r_{j,\ell} &= \sum_w p_{j,w}^{(\ell-1)} g_{j,\ell}(w), \\
p_{j,w}^{(\ell)} &= p_{j,w}^{(\ell-1)} \{1 + g_{j,\ell}(w) - r_{j,\ell}\}.
\end{aligned}
\tag{B.8}
$$

在我们的实验中，我们使用 $D = 31$ 层，并利用前四个

词元为水印设定种子。我们从 $p_j^{(D)}$ 中采样每个词元。

验证计算 $S_j = \sum_{\ell=1}^{D} g_{j,\ell}(W_j)$。在词元–比特独立性下，$S_j$ ∼服从二项分布$(D, 1/2)$。定义 $Z_j = (2S_j - D)/\sqrt{D}$、$A_j = |Z_j|$，以及–$B_j = \max(S_j, D - S_j)$。由于 D 为奇数，零假设得分非零，且在给定其幅值的条件下，其符号函数是公平的。在假设1的条件零模型下，这些符号函数还相互独立，从而具备校准所需的符号翻转性质。我们采用工作备择二项分布(D, ϑ)，$\vartheta \geq 1/2$，其掩码质量为

$$g_0(b) = 2^{1-D}\binom{D}{b}, \qquad g_1(b;\vartheta) = \binom{D}{b}\{\vartheta^b(1-\vartheta)^{D-b} + \vartheta^{D-b}(1-\vartheta)^b\}. \tag{B.9}$$

利用这些掩码质量，我们在附录B.1.2 和B.1.3中拟合隐马尔可夫模型和CP模型，方法是对 $\mathrm{logit}\ \vartheta \in [0, 8]$ 上的仅幅值边际似然进行最大化，从而得到 q_j 和$X_j = q_j Z_j$。对于 Gumbel–max，工作模型的模型设定错误可能影响检验效能，但不会影响条件零模型下的有效性。校准使用相同的公平符号分布；出现并列时，则根据位置独立于符号函数进行处理。所述 p值方法采用 $p_j = \Pr\{\text{二项分布}(D, 1/2) \geq S_j\}$。

B.4 段落选择启发式方法

第 3.6 节中的搜索使用句子边界、拟合得分和公式（8）中的精确界。对于 $m = 1$，我们在每个预算范围内枚举至少包含 64 个词元的所有连续句子段落，并保留比例下界最大的段落。对于 $m > 1$，我们构建候选段落、比较它们的并集，并按如下方式细化选定的边界。

1. 构建段落库。 按 $\sum_{j \in J \cap}$ 对候选段落 J 排序ε $X_j/\sqrt{|J|}$。将段落长度按上限划分为以下区间：64、96、128、192、256、384、512、768、1,024、1,536 和 2,048 个词元。在每个区间内，按得分从高到低最多保留四个段落；若某个候选段落与该区间内已

保留段落的交并比重叠度超过 0.5，则跳过该候选段落。为纳入较弱簇，还需使用 $q_j \geq 0.5$ 划分区域，合并间隔不超过 32 个词元的区域，并保留至少包含 64 个高权重位置的区域。对于每个此类区域和长度区间，添加词元至少有 80% 位于该区域内且得分最高的段落。最后，纳入每种预算下最优的单段落选择，并删除重复项。

2. 评估段落组合。 从该库中枚举二元组和三元组，仅保留彼此分隔且总长度不超过预算的文本片段。根据公式（8）联合评估每个并集，并为每个 (B, m)保留比例下界最大的集合，同时将单段落选择也作为候选项。

3. 细化选定集合。 对选定的文本片段最多进行两轮处理。每次处理一个文本片段，将其端点沿任一方向移动最多两个句子边界，并保留最佳可行集合。在段落上限允许的情况下，还要考虑删除一个文本片段，或从库中添加一个文本片段。每个候选项都必须满足长度、分隔和预算约束。

对于每个预算和段落上限，我们还会考虑在其他上限下评估的集合，只要它们满足当前约束。该候选池包括在 $B = 128$ 下评估的集合。我们报告 $B \geq 256$ 的结果。若出现并列，我们依次优先选择认证计数较大的集合、总长度较短的集合、段落较少的集合，最后选择边界更早的集合。搜索仅使用观测到的文本和水印证据。对于启发式方法返回的每个集合，我们都会计算精确的闭合检验界。即使搜索未找到最优集合，同时覆盖仍然成立。

C 补充结果

我们介绍模拟设计，并在比较 Gumbel–max 和 SynthID-Text 水印技术下的检测功效、词元级发现和段落定位之前，报告经验错误率。最后的实验考察信号聚类、添加背景文本，以及工作模型和局部检验的选择如何影响界。

C.1 实验设计

在 $\alpha = 0.05$，我们比较使用不同局部检验、得分或工作模型的 SACT 变体。我们还在相同的文档、窗口以及排序后词元集合上评估 CountScan。附录 B.2 详细说明了 SACT 局部检验和比较流程。

C.1.1 文档生成与符号标签

我们使用来自 C4 realnewslike 子集的英文网络文本（Raffel 等，2020）作为背景，并使用大语言模型 Qwen3.5-9B（Qwen 团队，2026）生成的段落构造混合文档，同时禁用其思考模式。在每个 n词元的文档中，我们将两个间隔至少四个词元的 L 连续词元位置区间替换为 Gumbel–max 加水印段落。我们按照文档顺序生成这些段落，并使用更新后的前置文本作为第二个段落的上下文。播种使用紧邻前方的四个词元。提示词在段落两侧各包含最多 128 个词元。我们从紧邻段落前方的 32 个词元开始生成。这样，第一个生成的词元便拥有与其所替换位置相同的四词元播种上下文。在生成 L 个词元之前，我们会抑制特殊词元和提前终止。如果某个水印播种上下文已经出现，则下一个词元将从 P_j 中采样，且不使用水印技术。生成的段落不会经过编辑。

我们改变文档长度 $n \in \{1024、2048\}$、段落长度 $L \in \{64、96、128、192、256\}$，

以及大语言模型采样温度 $t \in \{0。8、1。0\}$。我们使用 $t = 1$ 进行主要比较，并针对每种设置运行 100 次重复实验。对于每种文档长度，我们排除重复文本和之前使用过的文本，然后对 L 和 t的所有取值使用相同的背景文档。较高的温度会使下一词元分布变得更加平坦，并通常增强 Gumbel–max 信号。验证会重建 Y_j、 $p_j = 1 - Y_j$和 $Z_j = \Phi^{-1}$（Y_j）。只有当某个位置保留有效的带水印词元及匹配的生成上下文时，该位置才会获得 $\xi_j = 1$。不合格位置仍保留在查询分母中，但会从测试中排除。

SynthID-Text 实验采用相同的 C4 设计、查询窗口、FDP 目标，以及置换和稀释比较；每种长度均使用 100 个完全原假设文档。我们在附录B.3中介绍 SynthID-Text 的生成与验证。

C.1.2 比较程序

默认的 SACT 程序使用 Gamma-HMM 得分 $q_j Z_j$ 进行 Gumbel–max 水印。SACT-s改变散度指标，而 SACT-2-CP 将隐马尔可夫模型替换为 CP 工作模型。SACT-Z 和 SACT-q 分别使用 Z_j 和 $\varepsilon_j q_j$，并采用 $s = 2$ 检验。SACT–BC 和 SACT–KJI 在保留拟合得分的同时改变局部检验。GoF-s 和 Simes 使用枢轴量 p值，而 CountScan 使用其 C1 校准器和区间聚合扩展。附录 C.3 比较了两种水印机制下的性能，附录 C.4考察了结构自适应的影响。

C.1.3 结构实验

为考察新增背景的影响，我们从包含一个 512 个词元带水印区间的 1,024 个词元文档开始。我们将 C4 来源中文档后的文本追加到每个文档中，在保留初始枢轴和信号标签的同时，将长度增加至 1,280、1,536、1,792、2,048、2,560、3,072、3,584 和 4,096。对于置换比较，我们将枢轴、资格指标和信号标签一同置换，然后重新拟合模型。我们

以 16 个词元为步长移动 128、256 和 512 个词元的窗口。为评估已知生成模型时的得分估计，我们还模拟了一个高斯隐马尔可夫模型，其 $n = 1024$、平稳信号概率为 0.25、信号运行平均长度为 128、方差为单位方差、原假设均值为零，备择均值为 0.75、1.25、2。我们将拟合得分与使用生成参数计算的预言机分数进行比较。

C.1.4 计算评估

表 1 中的计时结果涵盖 100 份文档，参数为 $(n, L, t) = (2048, 128, 1)$。对于每份文档，我们分别计时预处理步骤，以及对随机选取的 1,000 组、每组含 256 个词元的查询集进行评估所需的时间。所有方法使用相同的查询集。SACT、拟合优度、Simes 和 KJI 程序使用 AMD EPYC 9654 处理器，并配备 64 个单线程工作进程。符号函数和 BC 扫描使用算法 1 中的参考族快捷方法；拟合优度和 Simes 使用算法 2；KJI 则使用附录 B.2.1 中描述的参考族表示。对于 CountScan，我们使用其精确的区间聚合实现，并配备 128 个单线程工作进程。我们不计入加载数据、生成查询集、拟合模型、校准检验和验证结果所花费的时间。除全零表外，区间均为跨文档逐点计算的 95% 蒙特卡洛区间；全零表中的区间则是对各键取平均得到的。

C.2 经验误差控制

C.2.1 Gumbel–max 和高斯实验

图 C.1 报告了三种错误率。第一种是同时覆盖失败的重复实验所占的比例。第二种是至少有一个窗口中的认证计数超过真实信号计数的重复实验所占的比例。对于第三种，我们针对排序与目标的每种组合，估计实现的 FDP 超过其目标的频率，并报告这些频率中的最大值。所有方法都使用相同的窗口和排序路径进行这些

比较。我们保留所有重复实验，即使每个下界都为零。对于 CountScan，我们将每个区间界与该区间内的保留信号计数进行比较。如果每个区间界都正确，那么对不相交区间求和也同样正确。在图 C.1所示的 33 种设置中，三种诊断指标中任意一种观测到的最大错误率为 7%（100 次重复实验中有 7 次）。对于均值，我们使用普通或配对 St udent-t 区间。对于概率，我们使用 Clopper–Pearson 区间。

表C.1 显示，全零假设拒绝率接近名义 5% 水平。我们使用 50 个独立验证密钥评估每个队列。对于每个密钥，我们重新计算枢轴量并重新拟合模型。由于这 50 个密钥均基于相同文本进行评估，因此我们计算每个密钥的平均拒绝率，并使用这 50 个拒绝率构造置信区间。

零假设队列	n	每个密钥的文本数	隐马尔可夫模型 (%)	CP，$\lambda = 2$ （%）
主要	1024	1,000	4.98 [4.77、5.18]	5.03 [4.81, 5.24]
主要	2048	1,000	4.98 [4.78, 5.18]	5.00 [4.78，5.22]
独立队列	1024	2,000	4.97 [4.82，5.11]	5.02 [4.87，5.17]

表 C.1：在 50 个独立验证密钥上取平均的全零假设拒绝率。使用 SACT-2 并以 5% 的水平，对 Gamma-HMM 和 Gamma-CP 进行评估。每个密钥下均使用相同的文本进行评估。括号内为逐点 95% Student-t 区间，该区间根据 50 个密钥级别的平均率计算，并以这些文本队列为条件。

C.2.2 SynthID-Text 实验

SACT-2 和 SACT-2-CP 在未置换信号的设置中，经验同时覆盖率为 99.25%（图 C.2）。在 $n = 1024$ 时，它们的全零假设拒绝率分别为 4% 和 3%；在 $n = 2048$ 时，分别为 7% 和 6%。四个逐点 95% 二项区间均包含 5% 的名义水平。

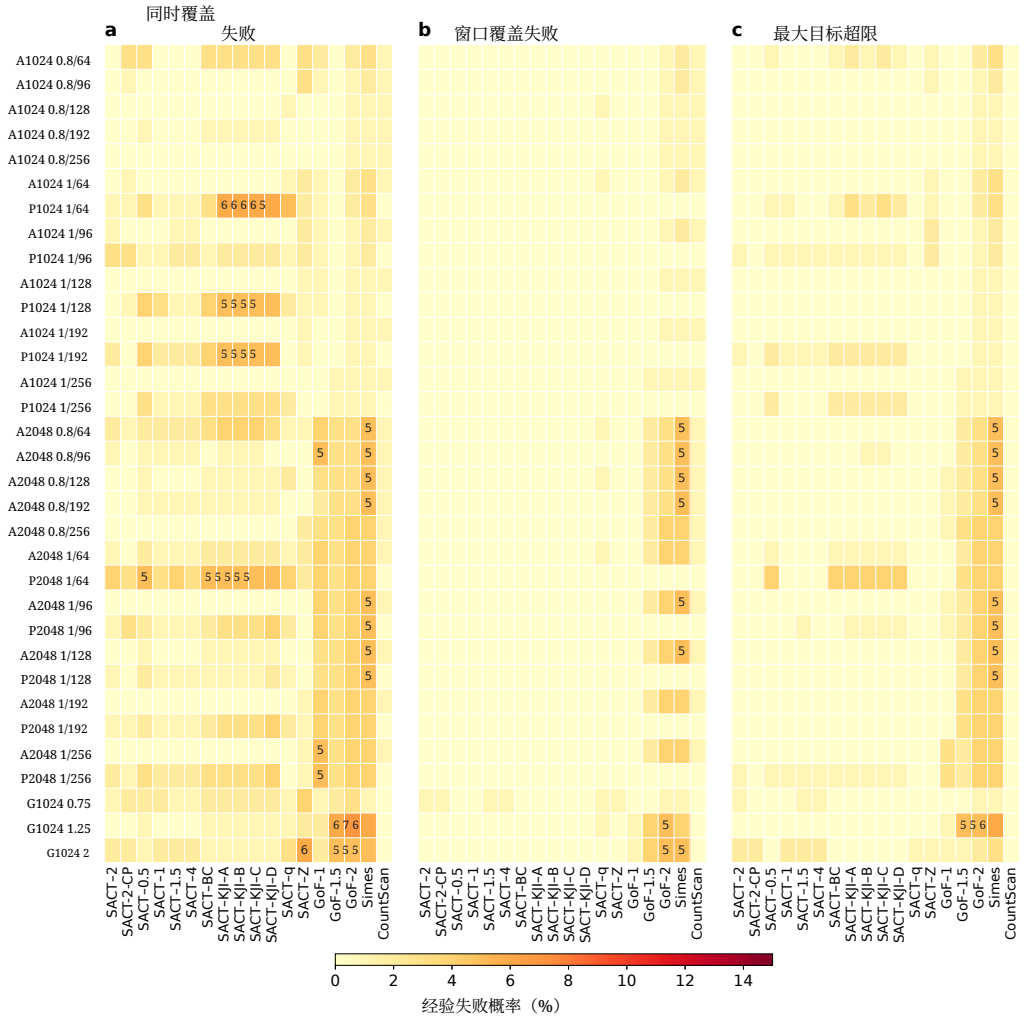

图 C.1：包含信号的 33 种设置中所有方法的经验错误率（每种设置重复实验 100 次）。面板 (a) 显示同时覆盖失败的频率，即，对于真实零假设位置集合$\mathcal{N}$，$d_\alpha(\mathcal{N}) > 0$。面板 (b) 显示至少一个窗口中的界超过真实计数的频率。面板 (c) 给出分别针对两条排序路径和三个目标计算的六个 FDP 超限率中的最大值。对于 CountScan，面板 (a) 检查所有区间上的等价失败事件，面板 (c) 则使用其在相同路径上的一般集合界。在 Gumbel–max 各行中，A 表示原始词元顺序，P 表示置换顺序。标签依次给出文档长度 n、温度 t 和段落长度 L，对应 t/L。高斯各行 (G) 给出 n 和备择均值。单元格中的数字表示不低于 5% 的比率。

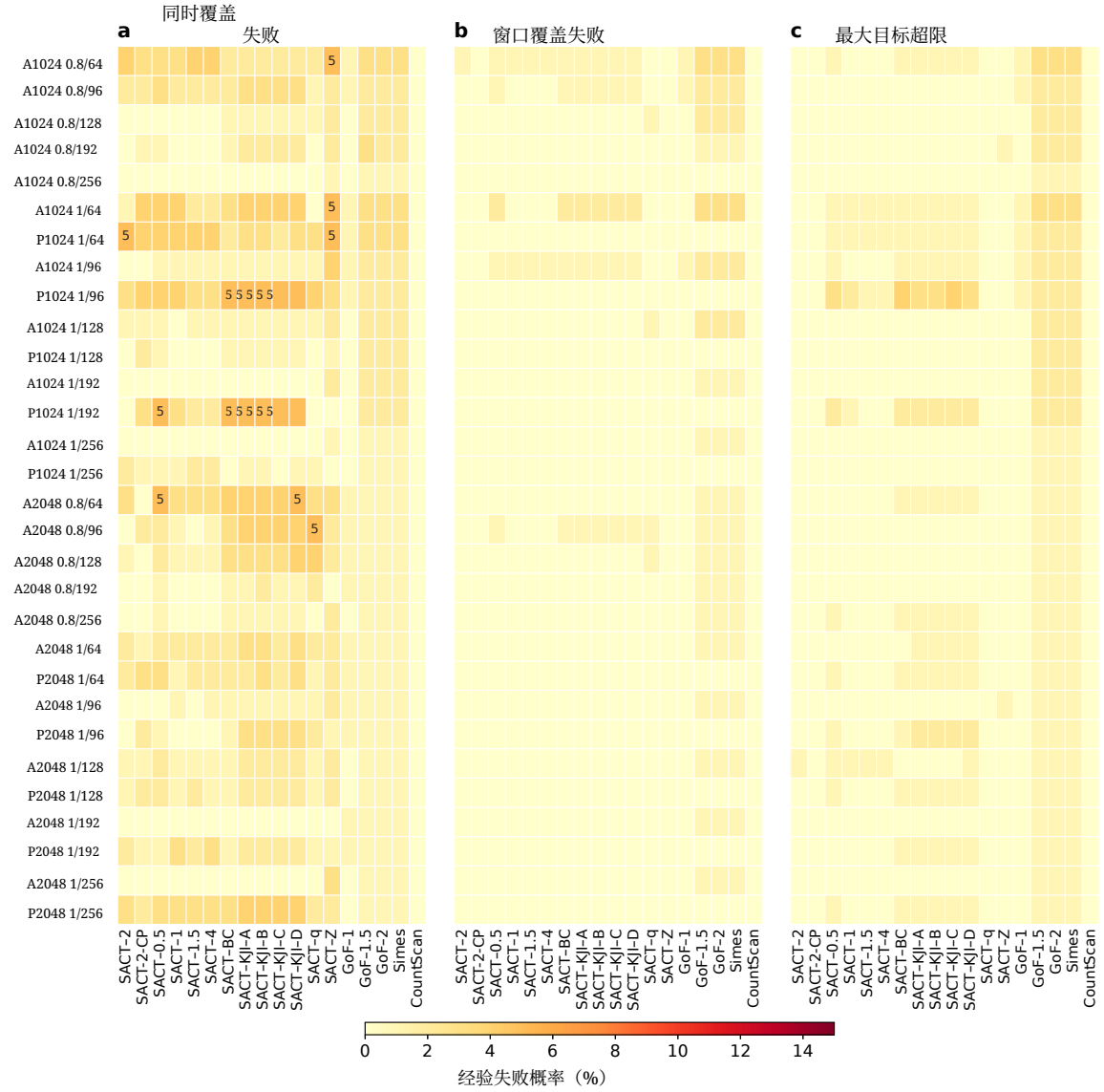

图 C.2：SynthID-Text：在 30 种完整或置换信号设置和 18 种方法下，使用每种设置 100 份文档进行的错误诊断。各面板分别报告同时覆盖失败、至少一个窗口中的过度陈述，以及两个排序和三个目标对应的六个 FDP 超限率中的最大值。对于 CountScan，第一个面板检查等价的全区间失败事件，第三个面板则使用其一般集合界评估相同的排序路径。行标签给出原始（A）或置换（P）词元顺序、文档长度 n，以及温度和段落长度，具体为 t/L。达到至少 5% 的比率以注释标出。

C.3 性能比较

我们采用附录 C.1 中的设计，比较两种水印机制下的文档级推断、词元级发现和段落定位。

C.3.1 Gumbel–max 水印

全局检测和比例下界。 图C.3显示，SACT 变体和 GoF-1 具有相似且通常较高的检测功效，但 SACT-2和 SACT-2-CP 的认证计数多于 GoF-1 和 CountScan。我们通过对 GoF-1 的局部检验应用闭合检验，获得其计数界。

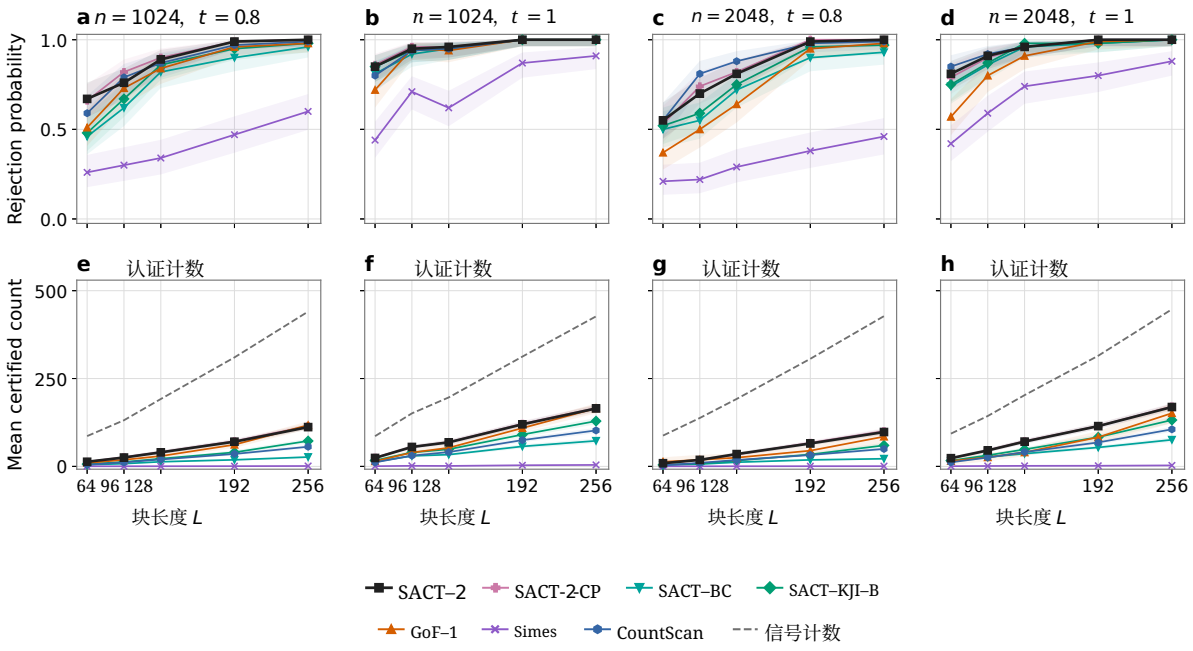

图 C.3：C4 文本中包含两个采用 Gumbel–max 水印的段落时，全局拒绝概率（上）和全文档认证计数（下）。各列展示了块长度 L 下的 $(n,\ t)$ =(1024，0.8)，(1024，1)，(2048，0.8)，(2048，1)。底行中的虚线表示真实信号计数。每个点为 100 次重复实验的均值。色带表示 95% 的蒙特卡洛区间。

词元级发现。 所有方法在三个排序中的每个排序内查询相同的前k 个集合：原始 Y_j、基于 HMM 的 X_j和基于 CP 的 X_j。与第3.5节一样，我们选择满足每个 FDP 目标 $\gamma \in \{0.1、0.2、0.5\}$的最大集合，从而分离排序和局部检验的影响。

按任一工作模型进行排序，比仅按 Y 排序能将更多信号词元置于列表顶部（图C.4）。BC 与 KJI–B 在更严格的 FDP 目标下给出更大的选择，而在 $\gamma = 0.5$ 时，SACT-2与 SACT-2-CP 在各自基于模型的路径上给出最大的选择。CountScan 仅对较大的 top-k 集合给出正界。即便对于这些较大的集合，CountScan 的界也过小，无法满足三个 FDP 目标中的任何一个。

图 C.5 对 $n = 1024$路径比较作了补充，并给出所有变体在 $n = 2048$处的选定大小和真实发现。

段落级定位。 对于每个窗口宽度，每种方法都会从公共网格中选择认证计数最大的窗口。结构自适应方法的计数高于 GoF-1 和 Simes，而对于较短的窗口，CountScan 通常会给出更大的认证计数（图 C.6）。

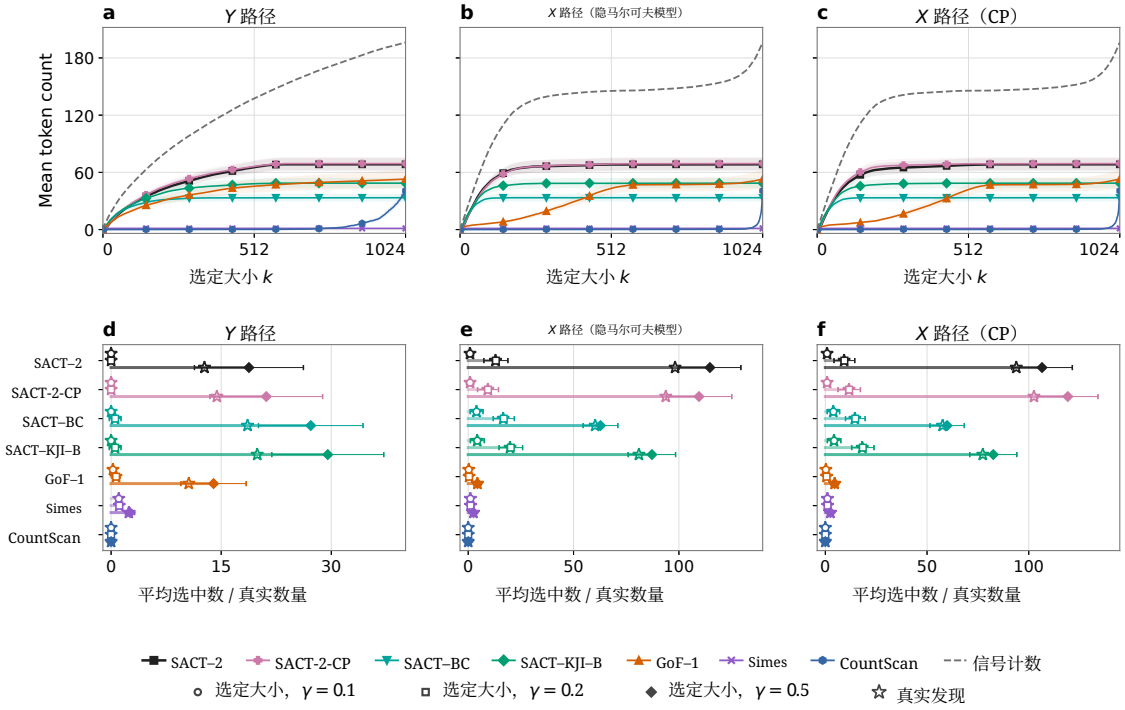

图 C.4：在 $(n, L, t) = (1024, 128, 1)$ 条件下，两个带水印段落中的词元级发现。

各列分别使用 Y、基于 HMM 的 X 和基于 CP 的 X 排序。上排显示沿每条路径的认证计数，虚线曲线表示信号计数。下排显示选定词元的数量，以及在 FDP 目标$\gamma = 0.1$、0.2、0.5 下的真实发现数量（圆形、方形和菱形）；空心星号表示真实发现。每种方法在每条路径内查询相同的集合。图中数值为 100 次重复实验的均值。带状区域和误差条表示 95% 区间。

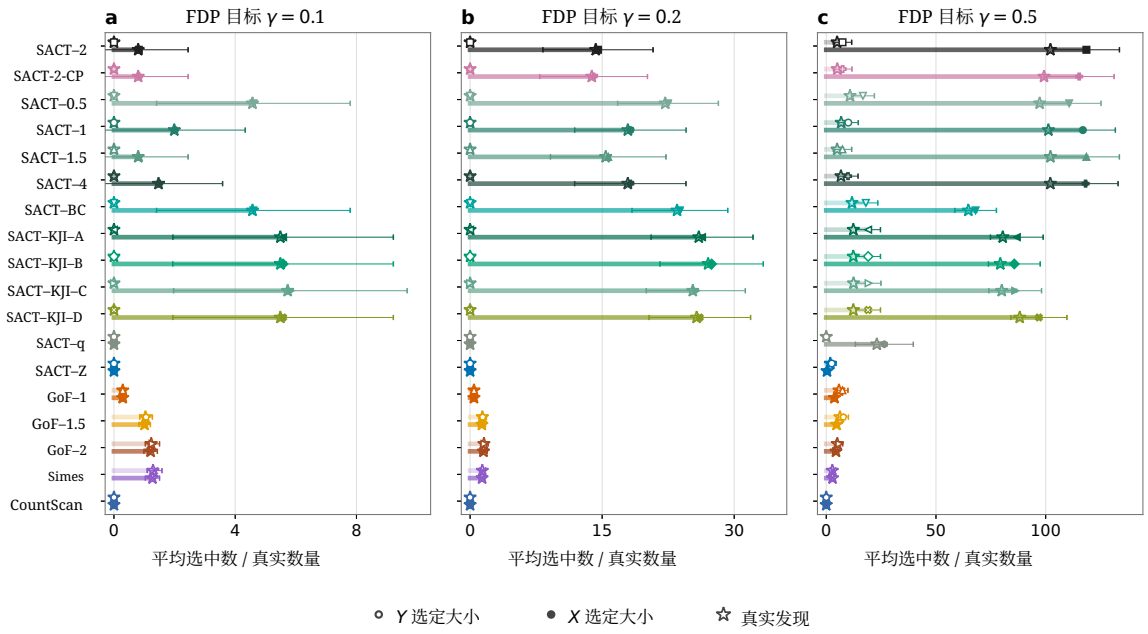

图 C.5：在 $(n,\ L,\ t) = (2048,\ 128,\ 1)$ 条件下使用 Gumbel–max 水印进行词元选择。各列显示 FDP 目标 0。1、0。2 和 0。5。空心圆和实心圆分别表示使用 Y 和基于 HMM 的 X 排序所选择词元数量的平均值。空心星号表示真实发现。柱状图表示 100 次重复实验的 95% 区间。

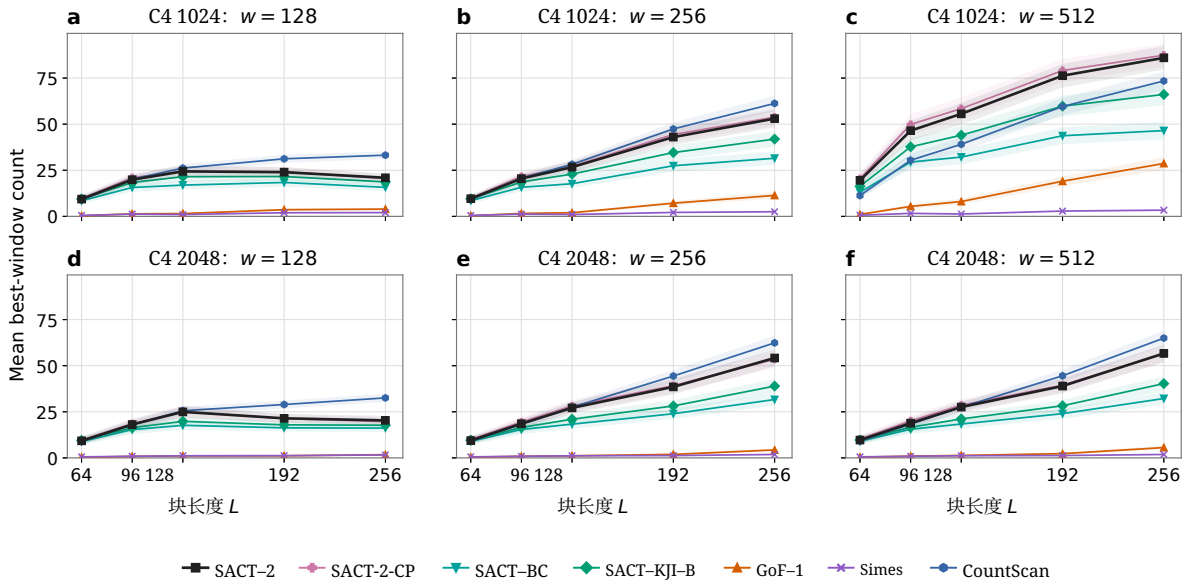

图 C.6：在 $t = 1$ 处滑动窗口上的最大认证计数。行对应文档长度，列对应窗口宽度。窗口每次以 16 个词元为步长向前移动。每种方法仅根据界选择窗口。每个数据点是 100 次重复实验的均值。色带表示 95% 蒙特卡洛区间。

C.3.2 SynthID-Text 水印

使用附录 B.3 中的验证统计量和工作模型，我们在 SynthID-Text 水印下重复附录 C.1 中的比较。附录 C.2 报告了相应的错误诊断。

全局检测和段落定位。 在信号保持原始顺序的 20 个设置中，SACT-2 和 SACT-2-CP 的平均检测功效分别为 79.70% 和 82.55%，平均认证计数分别为 52.03 和 54.27（图 C.7）。CountScan 在 55.8% 的文档中检测到信号，平均认证 10.84 个信号。对于每个文档，我们还取所有 256 个词元窗口中的最大比例界。在相同的 20 个设置上取平均后，对于 SACT-2，该界为 4.90%；对于 SACT-2-CP，为 5.54%；对于 CountScan，为 3.23%；对于 SACT-Z，为 0.29%；对于 GoF-1，为 0.10%。图 C.8 显示了 $t = 1$ 在三种窗口宽度下对应的窗口计数。

词元级发现。 包括 CountScan 在内的所有方法，都会在原始计数和基于 HMM 的排序下评估相同的前 k 个集合。在 $(n,\ L,\ t) =$（1024，128，1）时，CountScan 在三个 FDP 目标下均未选择任何词元（图 C.9）。

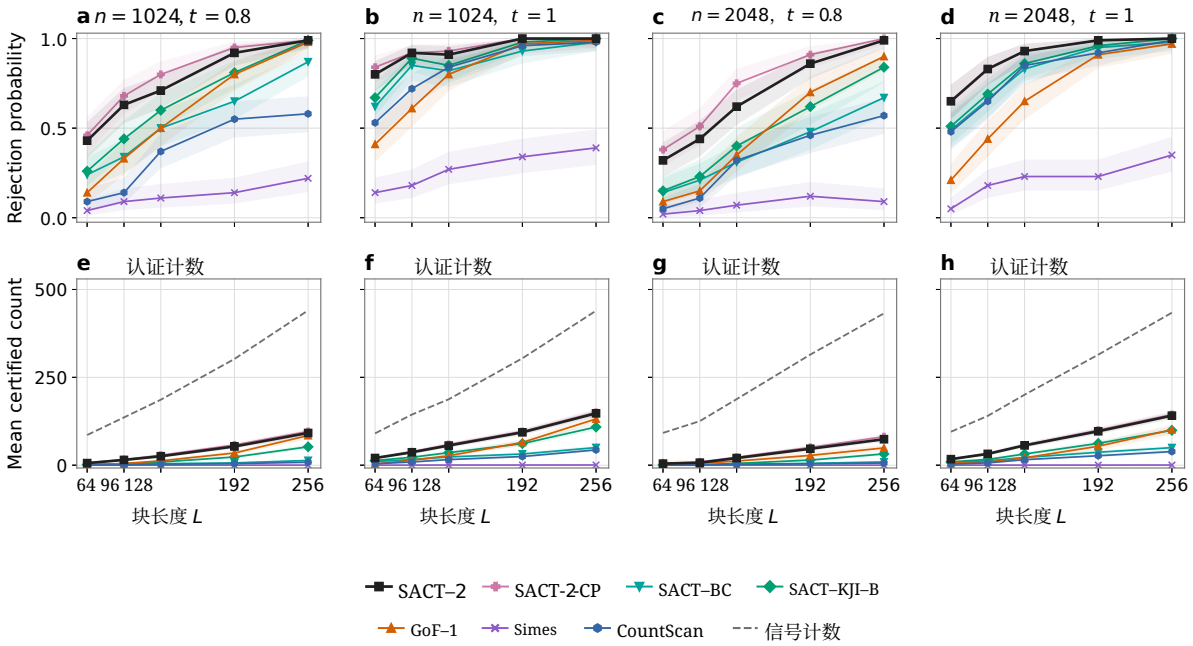

图 C.7：SynthID-Text 水印下的全局检测功效（上）和全文档认证计数（下）。每列展示一种文档长度与采样温度的组合。每个点表示 100 篇文档的均值。虚线表示信号计数，色带表示逐点 95% 蒙特卡洛区间。SACT-2-CP 使用 CP先验和 $\lambda = 2$；其他结构方法使用 HMM权重。

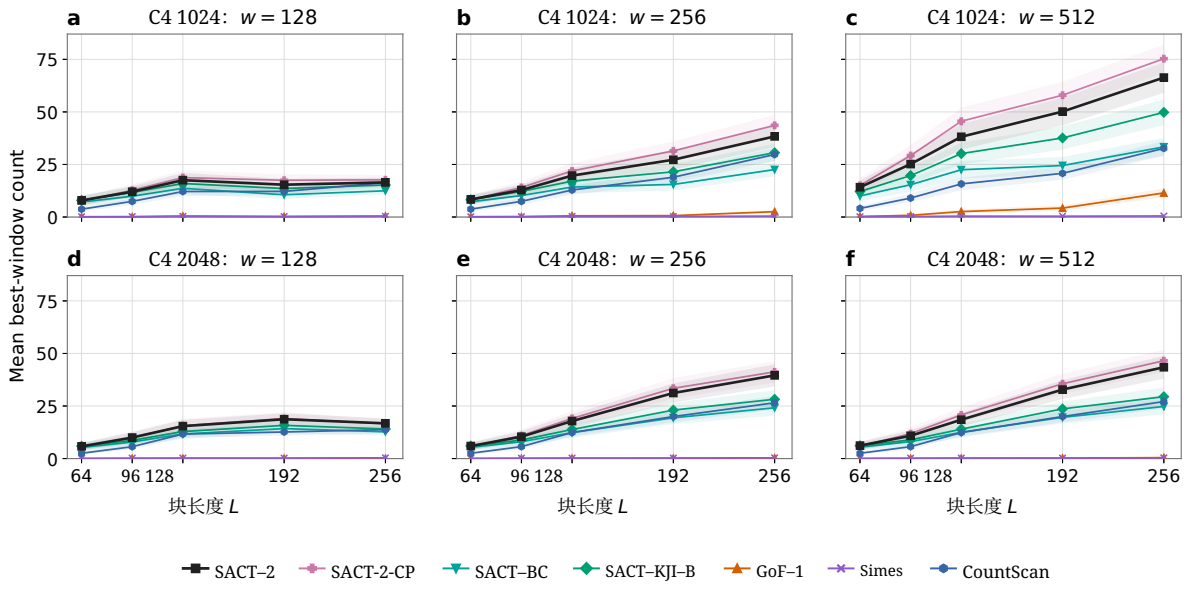

图 C.8：在 $t = 1$ 条件下使用 SynthID-Text 水印进行段落定位。对于每种宽度（128、256 或 512 个词元），每种方法都会报告间隔 16 个词元的窗口中最大的认证计数。均值和 95% 区间基于每种设置下的 100 份文档计算。

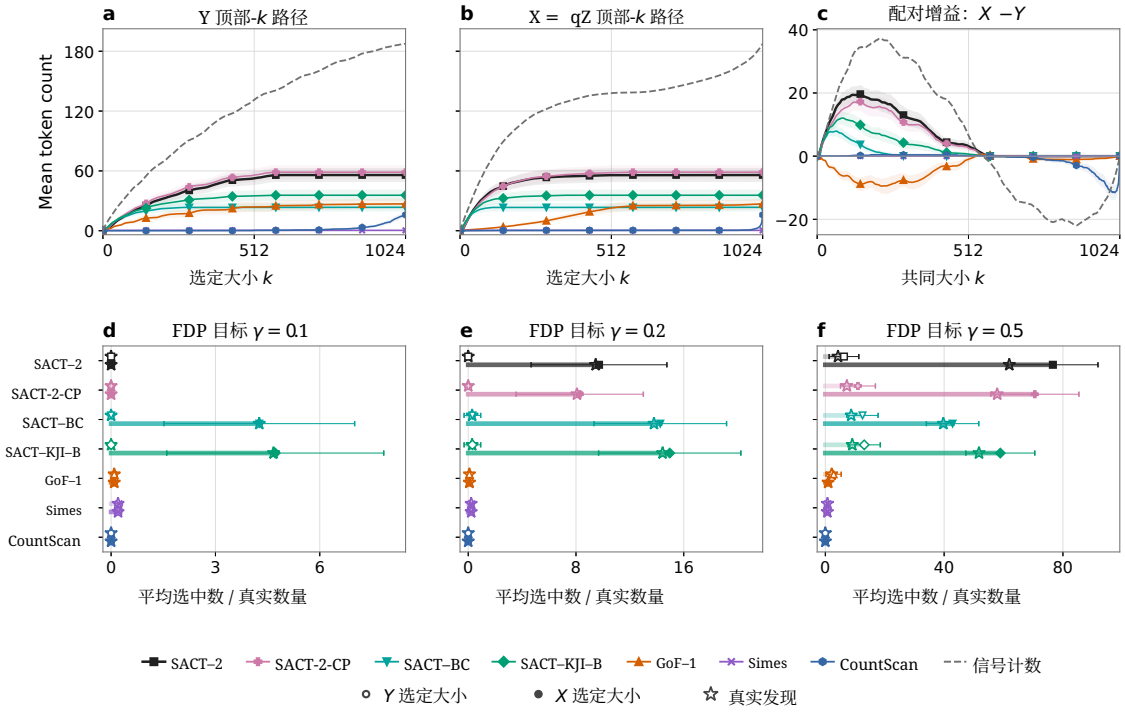

图 C.9：SynthID-Text：在 $(n,\ L,\ t) = (1024,\ 128,\ 1)$ 下的词元级发现。面板 (a) 和 (b) 展示了在原始计数（Y）和基于 HMM（X）排序下，前 k 个词元中的认证计数均值。面板 (c) 展示了每个 k 处的配对差值均值（X 减去 Y）。虚线曲线分别表示 (a) 和 (b) 中对应的真实信号计数，以及 (c) 中的差值。在底行，空心圆和实心圆分别表示在每个 FDP 目标下，依据 Y 和 X 排序所选择的数量均值。空心星号表示真实发现的数量均值。所有方法均使用相同的两种排序。区间基于 100 份文档计算。

C.4 结构自适应的影响

我们首先考察两种水印机制下的信号聚类与稀释现象，然后比较工作模型和局部检验，其中包括一项使用预言机分数的高斯实验。

C.4.1 信号聚类与背景稀释

Gumbel–max 水印。 在固定的 512 个词元带水印区间中，随着背景文本的增加，结构自适应 SACT 变体和 CountScan 的认证计数几乎保持稳定，而拟合优度计数则有所下降（图 C.10(a)）。经过联合置换后，随着文档长度增加，SACT 的计数会减少，而 SACT-2 的计数与拟合优度计数相近（图 C.10(b)）。

SynthID-Text 水印。 对于两种 SACT-2 变体而言，当加水印段落保持完整时，添加背景文本对认证计数的影响小于词元被置换时的影响（图C.11）。

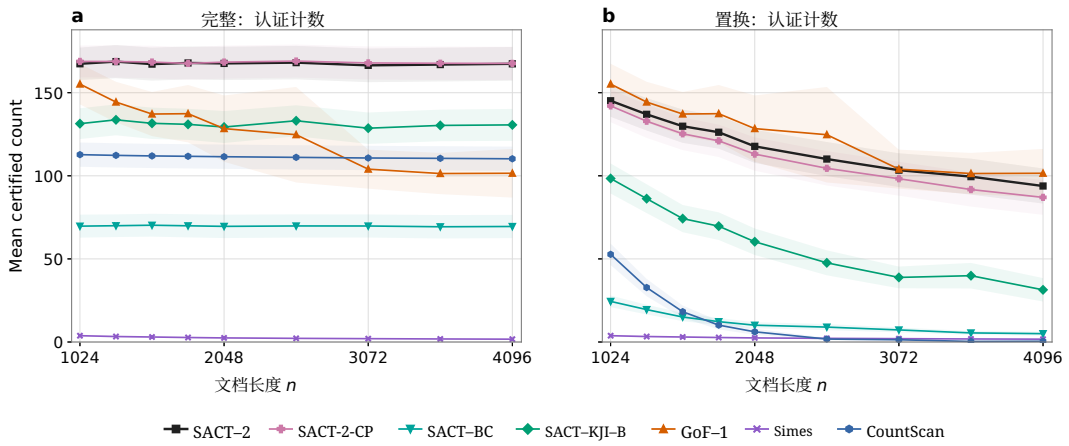

图 C.10：在 $t = 1$ 处进行固定信号稀释时，联合置换枢轴、资格指标和信号标签前后的全文档认证计数。每个文档最初包含 1,024 个词元，其中包括一个 512 个词元的带水印区间。我们通过追加背景文本来延长每个文档，同时保持带水印区间不变。带状区域表示 100 个配对文档中的 95% 区间。

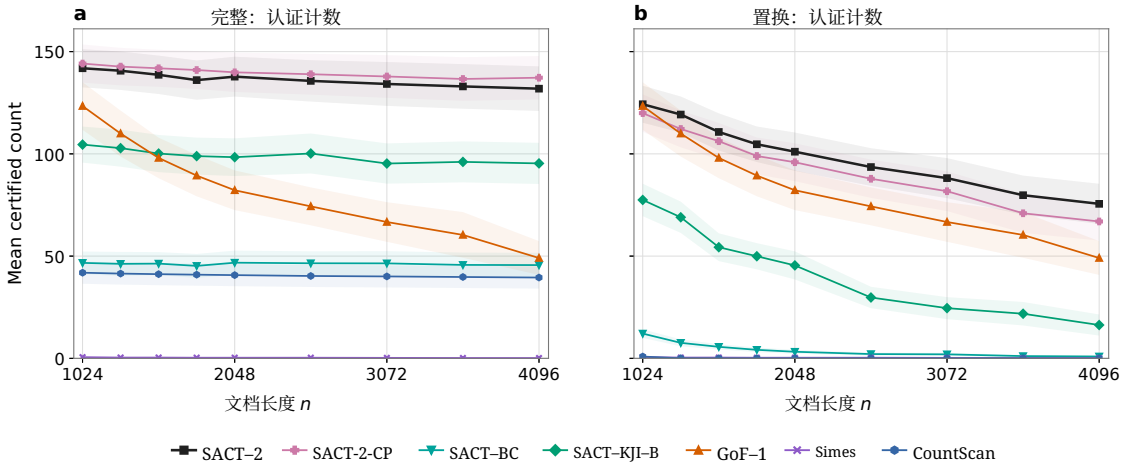

图 C.11：向包含一个位于 $t = 1$ 处的 512 个词元 SynthID-Text 加水印段落的文档中添加背景文本时的全文档认证计数。面板 (a) 保持原始词元顺序。面板 (b) 对枢轴、资格指标和信号标签进行联合置换。带状区域表示 100 个配对文档中的 95% 区间。

C.4.2 工作模型和局部检验

我们使用相同的 Gamma发射族和 SACT- 2 检验比较 HMM 与 CP 权重。我们首先在每个背景文档内对配对差值取平均，然后再对 200 份文档取平均。对于采用 $\lambda = 2$ 的 CP，相较于 HMM，平均增益为：0.82 个认证信号（95% 区间 $[0.42, 1.23]$）、256 个词元窗口上的最大比例下界提高 0.28 个百分点（$[0.19, 0.38]$），以及检验效能提高 0.45 个百分点（$[-0.38, 1.28]$）。$\lambda = 2$ 与 $\lambda = 0.5$ 的比较在检测功效（90.25% 对 89.45%）、平均认证计数（67.75 对 67.86）和平均窗口界（8.69% 对 8.28%）方面给出了相似结果。

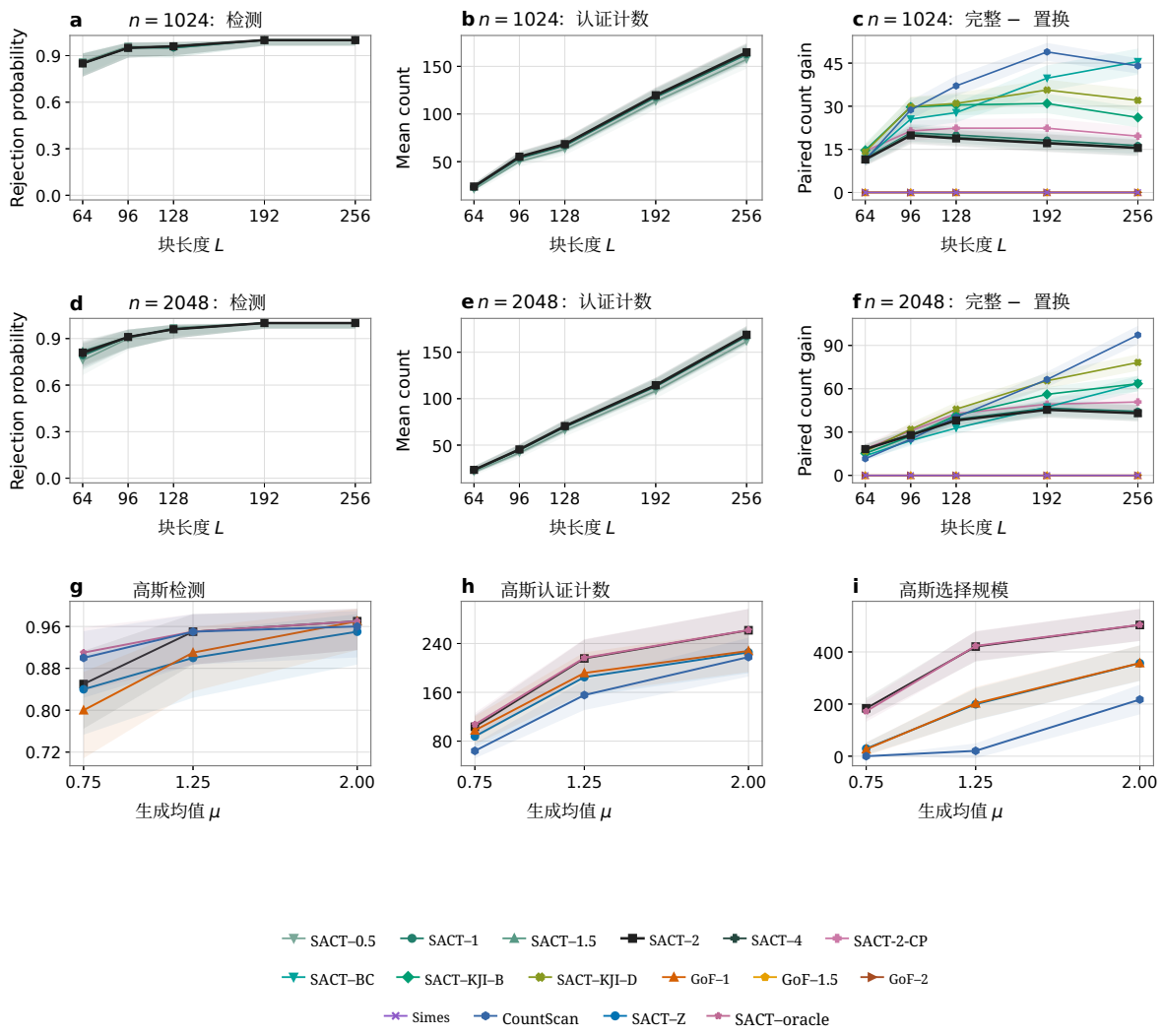

图 C.12：局部检验与结构比较。前两行分别展示 Gumbel–max 在 $t = 1$ 下对 $n = 1024$ 和 2048 的结果。各列分别展示检测功效、全文档认证计数，以及认证计数中的配对差值均值（原始顺序减去置换顺序）。底行展示高斯实验中的检测功效、认证计数，以及 FDP 目标为 0。5 时的选定位置数量。CountScan 包含在配对计数对比和全部三个高斯结果中。预言机使用生成参数计算用于推断的分数。所评估的集合仍遵循根据拟合后的 X 分数得到的排序。带状区域表示普通比较或配对比较中的 95% 二项区间或 95%t 区间。