基于结构分布的专利交底书仿真数据构造方法研究

祝 苑[1,2]，杨沐昀[3*]，邓宇[4]

（1.浙江大学软件学院，浙江 宁波 315100；2.哈尔滨工业大学未来技术学院，黑龙江 哈尔滨 150001；3.哈尔滨工业大学计算学部，黑龙江 哈尔滨 150001；4.哈尔滨市阳光惠远知识产权代理有限公司 150001 黑龙江 哈尔滨）

摘要：技术交底书包含原始技术构思，但受保密性限制难以大规模获取，导致专利自动撰写研究面临高质量“交底书—说明书”配对数据稀缺的问题。为此，本文提出一种基于结构分布先验的伪交底书逆向构造方法。首先，基于真实文档对分析不同板块及位置的结构映射特征，并与现有替代数据进行比较；随后结合概率采样与大语言模型语义重写，从公开专利说明书逆向构造伪交底书。实验结果表明，该方法能够较好复现真实文档对的主要结构分布特征，伪交底书与真实交底书的平均语义一致性 F1 达到 0.8219；基线对比与消融实验进一步验证了细粒度结构先验对相似度分布一致性的改善作用。在此基础上，本文构建了 200 组“伪交底书—说明书”文档对，为专利自动撰写研究提供了可规模化的数据构造方案。

关键词：法律自然语言处理；专利文本生成；大语言模型

中图分类号：TP391　**文献标志码：**A

Synthetic Patent Disclosure Generation Based on Structural Distributions

ZHU Yuan[1,2]，YANG Muyun[3*]，DENG Yu[4]

（1.School of Software Technology, Zhejiang University, Ningbo 315100, China; 2. School of Future Technology, Harbin Institute of Technology, Harbin 150001, China; 3. Faculty of Computing, Harbin Institute of Technology, Harbin 150001, China; 4.Harbin Sunshine Huiyuan Intellectual Property Agency Co., Ltd., Harbin 150001, China）

Abstract: Technical disclosures contain original inventive concepts but are difficult to obtain at scale due to confidentiality, resulting in a shortage of high-quality disclosure–specification pairs for automated patent drafting. This paper proposes a structure-distribution-guided method for reverse-generating pseudo-disclosures. First, structural mapping patterns across sections and positions are analyzed using real document pairs and compared with existing alternative data. Then, pseudo-disclosures are generated from public patent specifications through probabilistic sampling and large language model-based semantic rewriting. Experiments show that the method effectively reproduces structural distribution patterns of real document pairs, achieving an average semantic F1 score of 0.8219 between pseudo-disclosures and real disclosures. Baseline comparisons and ablation studies further demonstrate that fine-grained structural priors improve the alignment of similarity distributions. Based on the validated method, a dataset of 200 pseudo-disclosure–specification pairs is constructed, providing a scalable data generation approach for automated patent drafting research.

Keywords: Legal natural language processing; Patent writing; Large language models

* 通信作者：yangmuyun@hit.edu.cn

专利撰写是将技术构思转化为规范化法律文本的重要环节，通常由专利代理师根据发明人提供的技术交底书进行扩充与规范化，形成专利说明书等申请材料。这一过程既涉及技术信息的传递，也包含从技术语言到法律语言的转化与增量加工。

随着大语言模型生成技术的发展，专利自动撰写等任务对高质量训练数据提出了更高要求。然而，技术交底书通常涉及未公开的技术信息，难以大规模获取，导致高质量“交底书—说明书”配对数据较为稀缺。现有研究因而尝试采用学术论文[1]、合成草稿[2]等作为替代数据，但其与真实技术交底书在结构映射特征上的差异仍缺乏系统分析。针对这一问题，本文首先基于真实“交底书—说明书”文档对量化二者的结构映射规律，并与现有替代数据进行比较；在此基础上，利用所得结构分布作为先验，从公开专利说明书逆向构造在结构分布和语义内容上接近真实交底书的伪交底书，为交底书稀缺条件下的数据扩充提供新的思路。

本文开展了一系列实证分析与方法设计，主要贡献如下：**1）构建了细粒度的专利文本结构映射分析框架**，基于真实“交底书—说明书”文档对，利用双重相似度指标量化句级映射关系，并与现有替代数据进行比较，揭示二者在结构分布特征上的差异，为后续数据构造提供实证依据； **2）提出并验证了一种基于结构分布的伪交底书逆向构造方法**，基于真实文档对的结构分布规律，设计“双指标—双阈值”分级机制，结合概率采样与语义重写实现伪交底书构造，并通过结构与语义评价、基线对比及消融实验验证方法的有效性； **3）构建了规模化伪交底书数据集**，基于所提出并验证的构造方法，从公开专利说明书逆向生成 200 篇伪交底书，为交底书稀缺条件下的数据扩充及后续专利自动撰写模型的训练提供数据支持。

1 相关工作

随着预训练语言模型的发展，专利领域的自动化任务逐渐由文本分类拓展至生成式任务[3]。早期研究主要关注专利文档内部不同板块间的映射生成。Lee[4]探索了利用 GPT-2 模型捕捉专利标题、摘要与权利要求之间的生成关系；Christofidellis 等[5]进一步在 GPT-2 和 Patent Transformer 架构上验证了板块间对应生成的可行性。近年来，研究逐渐拓展至更具挑战性的长文本生成任务。Jiang 等[6]研究了从具体实施方式生成权利要求的方法；Wang 等[7]利用权利要求书与附图文本生成说明书内容；Yang 等[8]进一步引入多模态信息，利用权利要求与专利附图共同指导说明书生成。总体而言，现有研究主要利用专利文档内部板块之间的对应关系开展生成，而真实专利撰写通常以发明人提供的技术交底书为基础。由于技术交底书难以公开获取，现有研究对“交底书—说明书”之间的结构映射关系仍缺乏系统分析。

高质量训练语料是专利生成模型的重要基础。当前公开专利数据集主要包含专利摘要[4,9,10]、权利要求书[4,5,8,11-13]、公开说明书[14-18]或申请初始文件[10,13]，较少涉及真实技术交底书。为缓解源数据不足的问题，研究者尝试采用其他数据源或合成数据作为替代。Knappich 等[1]提出利用学术论文作为输入源生成专利文本，为构建大规模专利生成数据提供了新的思路。然而，学术论文与技术交底书在写作目的、内容组织和表达方式等方面存在差异，其作

为技术交底书替代数据的适用性仍有待进一步分析。Wang 等[2]提出 AutoPatent 框架，通过多智能体协作生成专利文档，并利用大语言模型问答构造专利草稿。上述研究为技术交底书难以获取条件下的专利生成提供了不同的数据构造思路，但现有替代数据与真实技术交底书之间的差异仍缺乏基于真实配对数据的系统比较。因此，如何利用真实数据分析技术交底书的文本特征，并进一步构建更接近真实数据的替代语料，仍是值得研究的问题。

2 问题定义与数据准备

为了对后续结构分析与逆向构造过程进行统一描述，本文首先定义研究问题及相关符号。设$\mathcal{S}$表示公开的专利说明书集合，$\mathcal{D}$表示对应的真实技术交底书集合。对于任意一对非对称文档(S, D)，其中$S \in \mathcal{S}$且$D \in \mathcal{D}$，本文旨在构建一种逆向生成方法Φ，使生成的伪交底书 $\widehat{D} = \Phi(S)$在结构分布和语义内容上接近真实交底书D。为进行细粒度的结构分析，本文将文档S和D划分为$K = 5$个标准板块，记为$\mathcal{K} = \{TF, BG, SU, DR, DE\}$，分别对应技术领域、背景技术、发明内容、附图说明和具体实施方式。其他符号定义如表 1 所示。

表 1　符号说明

Tab. 1　Notation and Definitions

符号	定义
$\mathcal{S}^{(k)}$	说明书第k板块的句子集合
$\mathcal{D}^{(k)}$	交底书第k板块的句子集合
$s_i^{(k)}$	说明书中第k板块的第i个句子
$d_j^{(k)}$	交底书中第k板块的第j个句子
$Sim(s_i^{(k)}, d_j^{(k)})_{BLEU}$	$s_i^{(k)}$与$d_j^{(k)}$之间的 BLEU 相似度分数
$Sim(s_i^{(k)}, d_j^{(k)})_{Jac}$	$s_i^{(k)}$与$d_j^{(k)}$之间的 Jaccard 相似度分数
$\sigma(s_i^{(k)})_{BLEU}$	$s_i^{(k)}$的最大 BLEU 相似度
$\sigma(s_i^{(k)})_{Jac}$	$s_i^{(k)}$的最大 Jaccard 相似度
$\tau_{BLEU}^{high}, \tau_{Jac}^{high}$	BLEU、Jaccard 相似度上界
$\tau_{BLEU}^{low}, \tau_{Jac}^{low}$	BLEU、Jaccard 相似度下界
$\mathscr{L}$	$\mathscr{L} = \{KEEP, DISTILL, DELETE\}$
$y_i^{(k)}$	句子$s_i^{(k)}$被分配的操作标签

本文使用两类语料开展分析。核心研究数据包含 50 组非公开的发明专利技术交底书及其对应的公开初始提交版本说明书，均由合作专利代理机构授权提供。样本涉及电学、机械和计算机等类别，发明内容均与人工智能技术相关，用于分析真实“交底书—说明书”文档对的结构与语义对应关系。此外，本文通过 Google Patents 数据库，以“人工智能”为关键词采集 100 篇已授权的中文发明专利说明书作为辅助语料。该部分数据仅用于后续阈值设定阶段构造随机专利文本负样本，以辅助确定相似度阈值下界。

3 专利文本结构映射与分布特征分析

为量化现有专利替代数据与真实“交底书—说明书”文档对在结构演化特征上的差异，本文首先对真实“交底书—说明书”非对称文档对以及 Knappich 等[1]构建的“学术论文—说明书”语料进行平行的分布特征分析。通过比较两类数据在不同板块及位置上的相似度分布规律，分析现有替代数据与真实数据之间的差异，并为后续伪交底书构造提供实证依据。

3.1 句子相似度度量与分布统计

3.1.1 相似度度量指标

BLEU[19]是自然语言生成领域常用的自动化评估指标之一，其核心思想是基于 N-gram 精确匹配率衡量生成文本与参考文本在字词及局部短语层面的重合程度。对于给定的句子对，BLEU 分数定义如下：

$$BLEU = BP \times exp\left(\sum_{n=1}^{N} w_n log p_n\right) \tag{1}$$

其中，p_n为 N-gram 精确匹配率，BP为长度惩罚项，用于避免生成文本过短。

本文利用 BLEU 衡量句子间局部词汇及短语结构的重合程度。

Jaccard 相似度是一种基于集合的相似性度量，用于衡量两个集合之间的重叠程度。将句子表示为词集合，对于给定的句子对$(s_i^{(k)} d_j^{(k)})$，其定义为交集大小与并集大小之比：

$$Jaccard\left(s_i^{(k)}, d_j^{(k)}\right) = \frac{\left|Set\left(s_i^{(k)}\right) \cap Set\left(d_j^{(k)}\right)\right|}{\left|Set\left(s_i^{(k)}\right) \cup Set\left(d_j^{(k)}\right)\right|} \tag{2}$$

其中，$Set(\cdot)$表示句子分词去重后的词集合。

与 BLEU 相比，Jaccard 对词序变化不敏感，可用于衡量句子间整体词汇的重合程度。因此，本文结合 BLEU 与 Jaccard，从局部短语匹配和整体词汇重合两个角度刻画说明书与交底书之间的句子对应关系。

3.1.2 阈值设定

为减少人工经验设定阈值带来的主观性，本文采用基于正负样本统计分布的经验阈值设定方法。通过构建复述关系的正样本集与非复述关系的负样本集，分别统计两类样本的相似度特征，并据此确定高、低相似度句子的经验判定阈值。具体过程如下：

本文从真实语料中随机抽取 2 篇技术交底书，选取 132 个句子用于构建正样本集。利用 DeepSeek v3.2，在保留核心语义的前提下对样本句进行复述生成。为进一步筛选生成结果，将复述句与原句成对输入模型，判断二者是否构成复述关系。最终 132 个句子对均被判定为复述关系，并作为正样本集$\mathcal{D}_{pos}$，用于表征复述句的相似度分布特征。

为构建具有不同文本距离的负样本，本文针对每个原句设计了 5 种扰动采样策略：1）近邻前句：该句在文档中的前一句；2）近邻后句：该句在文档中的后一句；3）篇章内随机前文：文档前文中的任意一句；4）篇章内随机后文：文档后文中的任意一句；5）跨文档长度匹配负例：从辅助语料中随机抽取字数相近的句子。

通过上述策略，本文共获得 652 个负样本对，并采用与正样本相同的方式进行校验，所

有样本对均被判定为非复述关系，最终构成负样本集$\mathcal{D}_{neg}$。

基于上述$\mathcal{D}_{pos}$和$\mathcal{D}_{neg}$，分别计算其中所有句子对的 BLEU 和 Jaccard 分数，并取其算术平均值作为高、低相似度的经验判定阈值。定义四个阈值参数：1）$\tau_{BLEU}^{high}, \tau_{Jac}^{high}$：$\mathcal{D}_{pos}$中 BLEU 和 Jaccard 的均值，作为高相似度判定的下界；2）$\tau_{BLEU}^{low}, \tau_{Jac}^{low}$：$\mathcal{D}_{neg}$中 BLEU 和 Jaccard 的均值，作为低相似度判定的上界。

根据表 2，BLEU 的低、高相似度阈值分别为 0.0300 和 0.4381，Jaccard 的低、高相似度阈值分别为 0.1383 和 0.6339。上述双指标阈值用于后续句子级相似度分类与分布特征统计。利用上述阈值，本文将说明书句子$s_i^{(k)}$与交底书句子集合$\mathcal{D}$的匹配关系划分为三个层级：

判定 1：高相似度句子　当说明书句子s_i与最佳匹配交底书句子$d_{best,BLEU}^{(k)}$和$d_{best,Jac}^{(k)}$（即使得$s_i^{(k)}$达到$\sigma(s_i^{(k)})_{BLEU}$或$\sigma(s_i^{(k)})_{Jac}$的匹配句）的 BLEU 和 Jaccard 相似度均不低于对应的高相似度阈值时，将其标记为高相似度句子。判定公式如下：

$$Condition_{high}: \left(Sim\left(s_i^{(k)}, d_{best,BLEU}^{(k)}\right)_{BLEU} \geq \tau_{BLEU}^{high}\right) \wedge \left(Sim\left(s_i^{(k)}, d_{best,Jac}^{(k)}\right)_{Jac} \geq \tau_{Jac}^{high}\right) \tag{3}$$

判定 2：低相似度句子　当两个指标均低于对应的低相似度阈值时，将其标记为低相似度句子。判定公式如下：

$$Condition_{low}: \left(Sim\left(s_i^{(k)}, d_{best,BLEU}^{(k)}\right)_{BLEU} < \tau_{BLEU}^{low}\right) \wedge \left(Sim\left(s_i^{(k)}, d_{best,Jac}^{(k)}\right)_{Jac} < \tau_{Jac}^{low}\right) \tag{4}$$

判定 3：中相似度句子　不满足上述两类条件的句子标记为中相似度句子。

表 2　相似度经验阈值

Tab. 2　Empirical Similarity Thresholds

样本对	Jaccard 相似度	BLEU
正样本对	0.6339	0.4381
负样本对	0.1383	0.0300

通过上述阈值设定，本文将说明书与交底书之间的句子匹配关系划分为高、中、低三个相似度层级，为后续结构分布分析与伪交底书构造提供量化基础。

3.1.3 相似度分布统计

本文分别从全文档和板块内部两个层面统计句子的相似度分布。考虑到不同专利说明书在篇幅上存在差异，为在统一尺度下衡量句子在板块内部的位置分布，首先定义句子的归一化相对位置。对于说明书第k个板块中的第i个句子$s_i^{(k)}$，其相对位置$pos(s_i^{(k)})$定义为：

$$pos\left(s_i^{(k)}\right) = \frac{i}{N_k - 1} \tag{5}$$

其中，N_k为该板块的总句数，$i \in [0, N_k - 1]$。

基于归一化相对位置，本文将每个板块划分为三个区间：前段（$Front, 0-30\%$）、中段（$Middle, 30-70\%$）和后段（$Rear, 70-100\%$）。为进一步统计高、中、低相似度句子在说明书不同位置的分布特征，本文定义局部发生概率指标，用于衡量特定板块及区间内某类句子占该区间总句数的比例：

$$P(Label \mid Region) = \frac{Count(Sentence\ in\ Region\ with\ Label)}{Count(All\ Sentence\ in\ Region)} \quad (6)$$

该指标用于刻画不同相似度句子随板块位置变化的分布规律，并作为后续伪交底书构造过程中概率采样的分布先验。同时，由于技术领域板块通常仅包含单个句子，不具有明显的板块内部位置分布，因此不参与该指标的统计。

3.2 真实交底书与说明书的映射分布

本文首先进行全局语义匹配，即将说明书中的每个句子作为查询项，在对应交底书全文范围内检索最佳匹配句。结果显示，说明书中高相似度句子占比约为 42.9%，表明说明书中较大比例的文本与交底书具有较高的文本重合度；低相似度句子占比约为 3.8%，说明完全处于低相似度区间的句子相对较少，其余文本主要分布于中相似度区间。

为进一步分析不同板块的映射差异，本文实施板块内约束匹配，即仅计算交底书与说明书同名板块内部的句子相似度，各板块的分布统计如表 3 所示。背景技术板块的高相似度句子占比最高，达到 81.20%，表明该板块在两类文档之间具有较高的文本重合程度。具体实施方式板块的高相似度句子占比为 23.05%，同时低相似度句子占比达到 19.19%，反映出该板块既包含与交底书高度重合的内容，也包含较多文本差异较大的内容。

基于 3.1.3 节定义的局部发生概率指标$P(Label \mid Region)$，本文进一步分析各板块不同区间内高、低相似度句子的分布情况，结果如表 4 所示。该指标刻画了不同相似度句子在板块内部的位置分布，并作为后续伪交底书构造过程中概率采样的经验参数。

表 3　真实交底书—说明书分板块匹配结果

Tab. 3　Section-wise Matching Results between Real Disclosures and Patent Specifications

板块	字数占比（%）	高于上限	中间范围	低于下限
技术领域	0.4474	0.1122	0.8469	0.0408
背景技术	8.4562	0.8120	0.1640	0.0240
发明内容	22.0862	0.2175	0.7588	0.0237
附图说明	2.0100	0.1019	0.7690	0.1292
具体实施方式	65.9710	0.2305	0.5775	0.1919

从高相似度句子的分布来看，背景技术板块的高相似度占比随位置推移逐渐下降，由前段的 91.83%下降至后段的 70.25%，但各区间整体仍保持较高水平，说明该板块与交底书具有较为稳定的文本重合关系。发明内容板块则呈现逐渐上升的趋势，高相似度句子占比由前段的 16.51%增加至后段的 27.92%，表明该板块后部与交底书的文本对应程度相对更高。具体实施方式板块的高相似度句子主要集中于中、后段，其中中段占比最高，为 33.16%，明显高于前段的 8.01%，表明该板块内部的文本对应关系具有较明显的位置差异。

从低相似度句子的分布来看，背景技术板块在各区间的占比均约为 2.3%，整体较低且分布稳定；发明内容板块的低相似度占比也保持在较低水平，但由前段的 1.14%逐渐增加至后段的 3.65%。具体实施方式板块各区间的低相似度占比分别为 19.23%、15.91%和 23.43%，整体明显高于背景技术和发明内容，说明该板块包含较多与交底书文本重合程度较低的内容。附图说明板块的低相似度占比则随位置推移逐渐下降，由前段的 19.62%下降至后段的

5.30%，表明其后部文本与交底书之间的对应程度有所提高。

表 4　真实交底书—说明书局部发生概率

Tab. 4　Position-dependent Operation Probabilities by Section

板块	高相似度句子分布			低相似度句子分布		
	0~30%	30~70%	70~100%	0~30%	30~70%	70~100%
背景技术	0.9183	0.8123	0.7025	0.0233	0.0233	0.0245
发明内容	0.1651	0.2094	0.2792	0.0114	0.0236	0.0365
附图说明	0.0764	0.0526	0.1364	0.1962	0.1263	0.0530
具体实施方式	0.0801	0.3316	0.2488	0.1923	0.1591	0.2343

综合上述结果，不同板块及板块内部不同位置呈现出不同的相似度分布特征。其中，背景技术整体具有较高的文本重合度，而发明内容、附图说明和具体实施方式表现出更明显的位置差异。这种非均匀分布说明，仅采用统一规则处理说明书各位置的句子难以反映真实“交底书—说明书”文档对的分布特征，因此后续伪交底书构造将进一步引入板块及位置相关的概率分布作为采样先验。

3.3　现有替代数据与说明书映射分布

为比较现有替代数据与真实“交底书—说明书”文档对在结构分布上的差异，本文进一步对 Knappich 等[1]构建的 1813 组“学术论文—说明书”语料进行局部发生概率分析。考虑到学术论文与专利说明书在篇章结构上并不完全对应，本文选取学术论文中的方法与结果部分，并将其与专利说明书的具体实施方式板块进行匹配分析。二者均集中承载技术方案、实现过程及实验结果等技术内容，因此选取该组对应关系作为比较对象，并采用与真实“交底书—说明书”相同的相似度度量与分布统计方法进行分析。

基于上述对齐方式，本文对该数据集进行统计，结果如表 5 所示。对比表 4 与表 5 可以发现，“学术论文—说明书”与真实“交底书—说明书”在局部相似度分布上存在明显差异。具体而言，学术论文—说明书的高、中、低相似度句子占比在前、中、后三个区间的变化幅度均较小，整体呈现较为平缓的位置分布；相比之下，真实交底书与说明书的具体实施方式板块具有更明显的位置差异。

表 5　学术论文—说明书局部发生概率

Tab. 5　Position-dependent Operation Probabilities for Academic Paper–Specification Pairs

区间	高于上限	中间范围	低于下限
0~30%（前）	0.0155	0.5333	0.4511
30~70%（中）	0.0293	0.5246	0.4461
70~100%（后）	0.0414	0.5213	0.4372

同时，两类数据在不同相似度层级上的占比也存在较大差异。“学术论文—说明书”数据中，低相似度句子在三个区间的占比分别为 45.11%、44.61%和 43.72%，均超过 40%；高相似度句子占比则分别仅为 1.55%、2.93%和 4.14%。相比之下，真实“交底书—说明书”具体实施方式板块的低相似度句子占比为 15.91%～23.43%，高相似度句子占比为 8.01%～33.16%。上述结果表明，以学术论文作为替代源文本时，其与说明书之间形成的相似度分布与真实技术交底书存在明显差异，难以直接复现真实“交底书—说明书”文档对的结构映射

特征。故有必要利用真实文档对中观测到的结构分布规律构造更接近目标分布的替代数据。

4 伪交底书生成框架

4.1 基于先验分布的采样生成机制

基于前述对真实“交底书—说明书”文档对的结构分布分析，本文提出一种基于先验分布的概率采样生成机制。该机制以真实文档对中观测到的句子相似度分布作为先验，对公开专利说明书不同板块及位置的句子进行概率化操作分配，从而逆向构造伪技术交底书。

根据 3.1.2 节定义的相似度层级，本文将高、中、低三类句子分别对应 KEEP、DISTILL 和 DELETE 三种处理操作（表 6）。需要说明的是，在伪交底书构造阶段，由于不存在对应的真实交底书，无法直接计算说明书句子的实际相似度类别。因此，本文并非对单个句子的真实类别进行判定，而是依据真实文档对中统计得到的局部发生概率，为不同位置的句子随机分配操作标签，以在整体分布层面模拟真实“交底书—说明书”的结构映射特征。

表 6 基于相似度层级的句子处理策略

Tab. 6 Sentence Processing Strategies Based on Similarity Levels

相似度层级	操作假设	处理方式
高	与交底书具有较高文本重合度	KEEP（保留）
中	与交底书存在一定文本对应，但表述差异较大	DISTILL（提炼）
低	与交底书文本重合度较低	DELETE（删除）

具体而言，对于说明书中的任意板块及其归一化位置区间，分别记 3.2 节统计得到的高、中、低相似度句子局部发生概率为$h = P(High \mid Region)$、$m = P(Mid \mid Region)$和$l = P(Low \mid Region)$，满足：

$$h + m + l = 1 \tag{7}$$

对该区间内的每个句子$s_i^{(k)}$生成一个服从均匀分布的随机数$n \sim U[0,1]$，并根据其所在概率区间分配操作标签$y_i^{(k)}$：1）当$0 \le n < h$时，标记为 KEEP；2）当$h \le n < h + m$时，标记为 DISTILL；3）当$h + m \le n < 1$时，标记为 DELETE。通过上述采样过程，各板块及位置区间内三类操作的期望比例与真实文档对中对应的相似度分布保持一致，从而在数据集合层面模拟真实“交底书—说明书”的结构分布特征。

根据采样得到的操作标签执行以下处理：1）KEEP：对于标记为 KEEP 的句子集合$\mathcal{S}_{keep}$，保持原始文本不变；2）DELETE：对于标记为 DELETE 的句子集合$\mathcal{S}_{delete}$，直接删除，不进入后续生成过程；3）DISTILL：对于标记为 DISTILL 的句子集合$\mathcal{S}_{distill}$，输入大语言模型进行语义还原，在保留技术信息的基础上对其表述进行重写，得到改写后句子集合$\mathcal{S}'_{distill}$。最终，将$\mathcal{S}_{keep}$与$\mathcal{S}'_{distill}$按照其在原始说明书中的相对顺序重新组合，形成完整的伪交底书$\widehat{D}$。

4.2 伪交底书构造质量评价方案

4.2.1 结构分布一致性评价

基于前述的统计分析，本文获得了真实“交底书—说明书”文档对在不同板块及位置区

间的高、低相似度句子分布。为评价生成数据对上述结构分布的复现程度，本文将各板块的分布特征表示为向量进行比较。按照 3.1.3 节定义的归一化位置，将各板块划分为前段（0–30%）、中段（30–70%）和后段（70–100%）三个区间，并分别统计各区间内高相似度或低相似度句子的发生概率，构建三维结构特征向量$\boldsymbol{V} \in \mathbb{R}^3$：

$$\boldsymbol{V} = [p_{Front}, p_{Mid}, p_{Rear}] \tag{8}$$

其中，p表示该区间内特定类别句子的占比。

对于测试专利，分别计算真实“交底书—说明书”与伪“交底书—说明书”的结构特征向量V_{real}和V_{pseudo}，并在全部样本上取平均，得到真实结构分布向量$\bar{V}_{real}$与生成结构分布向量$\bar{V}_{pseudo}$。本文采用平均绝对误差（MAE）和均方根误差（RMSE）衡量二者的分布差异。

MAE 用于衡量各位置区间分布概率的平均绝对差异。设$k \in \{F, M, R\}$（分别表示$Front, Mid, Rear$），区间总数$K = 3$，其计算公式如下：

$$MAE = \frac{1}{K} \sum_{k \in \{F,M,R\}} \left| \bar{\boldsymbol{V}}_{pseudo,k} - \bar{\boldsymbol{V}}_{real,k} \right| \tag{9}$$

RMSE 进一步考虑较大偏差的影响，其计算公式如下：

$$RMSE = \sqrt{\frac{1}{K} \sum_{k \in \{F,M,R\}} \left(\bar{\boldsymbol{V}}_{pseudo,k} - \bar{\boldsymbol{V}}_{real,k} \right)^2} \tag{10}$$

两项指标越低，表明生成数据与真实数据在对应板块及位置区间上的相似度分布越接近。

4.2.2 内容语义一致性评价

为评价伪交底书对原始技术内容的保留程度，本文构建了基于句子级语义对齐的 F1 评价指标。该方法借鉴 BERTScore[20]的双向匹配思想，并针对专利长文本中句子之间非严格对齐的特点，将 Token 级匹配调整为句子级语义匹配，以衡量不同文档之间的技术内容对应程度。本文采用 text2vec-base-chinese[21]这一基于 BERT[22]的模型对专利文档各板块的句子进行编码，得到对应的高维向量矩阵M_A和M_B。随后计算两个句子集合之间的余弦相似度矩阵$S = M_A M_B^T$，并分别从两个方向为每个句子选取语义相似度最高的匹配句，据此计算召回率R与精确率P。最终语义一致性得分$Score$定义为二者的调和平均值，即 F1 分数：

$$R = \frac{1}{|\boldsymbol{\mathcal{A}}|} \sum_{a \in \boldsymbol{\mathcal{A}}} max_{b \in \boldsymbol{\mathcal{B}}} cos(a, b), \quad P = \frac{1}{|\boldsymbol{\mathcal{B}}|} \sum_{b \in \boldsymbol{\mathcal{B}}} max_{a \in \boldsymbol{\mathcal{A}}} cos(a, b), \quad Score = \frac{2PR}{P + R} \tag{11}$$

为进一步比较不同文本之间的语义关系，本文在具有真实交底书的配对样本上设置三类文本对：真实交底书—说明书（类型 1）、伪交底书—说明书（类型 2）以及真实交底书—伪交底书（类型 3）。其中，类型 1 用于表征真实文档对之间的语义对应水平，类型 2 用于考察伪交底书对源说明书技术内容的保留程度，类型 3 则直接衡量伪交底书与真实交底书之间的语义接近程度。三类结果共同用于分析伪交底书的语义一致性。

5 实验设计与结果分析

5.1 构造方案可行性评价

基于 4.1 节提出的采样生成机制，本文利用 50 篇测试说明书构造对应的伪交底书，并依据 4.2 节的评价体系，从结构分布一致性与内容语义一致性两个维度评价构造效果。考虑到技术领域板块通常仅包含单个句子，不适用于板块内部位置采样，因此对高相似度句子直接保留，其余情况采用语义重写，以避免该板块整体缺失。

表 7 给出了生成数据与真实数据之间的结构分布误差。在高相似度分布方面，四个评价板块的 MAE 介于 0.0267 至 0.0787 之间，RMSE 均低于 0.08，表明生成数据能够较好复现真实文档对的高相似度分布。在低相似度分布方面，背景技术和发明内容的 MAE 分别为 0.0232 和 0.0227，附图说明与具体实施方式的误差相对较高。这可能与逆向构造文本仍保留较强的源说明书语义关联有关，使部分低相似度句子的分布难以完全复现。总体而言，多数结构指标保持较低误差，说明所提出的采样机制能够较好地复现真实文档的结构分布特征。

表 7　核心数据集结构一致性评价

Tab. 7　Structural Consistency Evaluation of the Core Dataset

板块	高相似度 MAE	高相似度 RMSE	低相似度 MAE	低相似度 RMSE
背景技术	0.0267	0.0376	0.0232	0.0235
发明内容	0.0787	0.0788	0.0227	0.0248
附图说明	0.0457	0.0521	0.1200	0.1309
具体实施方式	0.0579	0.0645	0.1930	0.1954

在内容语义一致性方面，表 8 显示，伪交底书与真实交底书（类型 3）的平均 F1 为 0.8219，伪交底书与源说明书（类型 2）的平均 F1 为 0.9095，真实交底书与说明书（类型 1）的平均 F1 为 0.8508。结果表明，伪交底书在保留源说明书主要技术语义信息的同时，与对应真实交底书也具有较高的语义一致性。结合结构分布评价结果，验证了所提出方法在结构特征复现与技术语义保持方面的可行性。

表 8　核心数据集语义相似度计算结果

Tab. 8　Semantic Similarity Results for the Core Dataset

类型	技术领域	背景技术	发明内容	附图说明	具体实施方式	平均
1	0.7894	0.9516	0.8412	0.8213	0.8557	0.8508
2	0.8597	0.9846	0.9149	0.8861	0.9133	0.9095
3	0.7788	0.9408	0.8257	0.7940	0.8194	0.8219

注：表中平均值依据表 3 中各板块的字数占比进行加权计算。

5.2　结构先验消融实验与结果分析

为进一步验证结构先验及其建模粒度对伪交底书构造效果的影响，本文设置三种对比方法：Direct LLM、Global Prior 和 Section-only Prior。其中，Direct LLM 不引入显式的句级操作分布约束，直接生成技术交底书；Global Prior 采用统一的全局操作概率分布；Section-only Prior 根据不同板块的统计特征分别配置操作概率，但不考虑板块内部的位置差异。三种方法均采用与主实验一致的评价体系，并与本文完整方法进行比较。

由表 9 可见，本文完整方法在相似度结构一致性方面取得了最优结果，其 MAE 和 RMSE 低于三种对比方法。与 Section-only Prior 相比，高相似度 MAE 降低约 86.8%，表明在板块级先验的基础上进一步考虑板块内部的位置分布，能够有效提高生成文本与真实数据之间的

结构一致性。在低相似度分布方面，各方法的误差整体较为接近，本文方法的 MAE 为 0.1388，RMSE 为 0.1411，同样取得了最优结果。

表 9　不同基线方法的结构分布一致性比较

Tab. 9　Comparison of Structural Distribution Consistency Across Baseline Methods

方法	高相似度 MAE	高相似度 RMSE	低相似度 MAE	低相似度 RMSE
Direct LLM	0.5102	0.5139	0.1419	0.1432
Global Prior	0.5624	0.5692	0.1472	0.1491
Section-only Prior	0.4507	0.4621	0.1465	0.1485
Full Prior	**0.0596**	**0.0651**	**0.1388**	**0.1411**

注：表中结果根据表 3 背景技术、发明内容、附图说明和具体实施方式四个板块的字数占比加权计算。

对比三种基线方法可以发现，Section-only Prior 的高相似度 MAE 和 RMSE 分别为 0.4507 和 0.4621，相较于 Global Prior 分别降低约 19.9% 和 18.8%，且均优于 Direct LLM。这说明结构先验的建模粒度对生成效果具有重要影响。全局先验能够提供整体分布约束，而板块级先验进一步考虑了不同技术板块的内容与表达差异；本文完整方法则在此基础上引入板块内部的位置特征，从而更准确地刻画真实文档对的结构分布。

从表 10 可以看出，Section-only Prior 在发明内容、附图说明和具体实施方式三个板块均优于其他基线方法，其中附图说明和具体实施方式相较于 Global Prior 的高相似度 MAE 分别降低约 33.0%和 25.7%。结合表 7 可见，本文完整方法在四个板块的高相似度 MAE 均保持在 0.08 以下，进一步体现了细粒度结构先验在不同板块中的适用性。

表 10　不同板块的高相似度 MAE

Tab. 10　High-Similarity MAE Across Sections

板块	Direct LLM	Global Prior	Section-only Prior
背景技术	0.1103	0.0861	0.1330
发明内容	0.6742	0.6349	0.5960
附图说明	0.7486	0.6242	0.4180
具体实施方式	0.4992	0.5973	0.4438

在语义一致性方面，各方法均能够较好地保留技术内容，整体表现较为接近。综合而言，消融实验表明，结构先验的建模粒度是影响伪交底书构造效果的重要因素。由全局分布约束逐步细化至板块级及板块内部位置级约束，能够更充分地刻画真实文档对的结构特征。本文提出的完整采样生成机制在保持技术语义一致性的同时，提高了结构分布的复现精度，进一步验证了所提出方法的有效性。

5.3　规模化伪交底书数据集构建

基于前述真实配对样本的验证结果，本文进一步开展规模化伪交底书数据构建。从 Google Patents 数据库中获取 150 篇已授权、优先权日期介于 2023 年至 2025 年间的人工智能技术专利说明书，经结构化解析、文本清洗及句级切分后，按照 4.1 节提出的采样生成机制，结合真实文档对的结构分布先验与 DeepSeek v3.2 模型，逆向生成对应的伪交底书。最终，结合前述 50 篇验证样本，构建了包含 200 组“伪交底书—说明书”文档对的数据集。表 11 给出了数据集的文本规模统计。经逆向构造后，规模化数据集伪交底书平均包含 174.0 个句子和 8210.1 个字符，文本规模得到压缩。

表 11　规模化数据集平均句数和字数

Tab. 11　Average Sentence and Character Counts in the Large-Scale Dataset

数据集	文件	数量	平均句数/篇	平均字数/篇
核心数据集	交底书	50	82.9	4889.5
	说明书	50	179.0	12040.5
规模化数据集	伪交底书	200	174.0	8210.1
	说明书	200	219.9	13967.2

表 12 给出了规模化数据集的结构分布一致性评价结果。对于除技术领域外的四个板块，高相似度分布的 MAE 和 RMSE 均低于 0.10；背景技术和发明内容的低相似度分布误差同样较小。总体来看，在数据规模扩展至 200 篇后，多数结构分布指标仍保持较低误差，说明基于先验分布的采样机制在规模化构造过程中能够较为稳定地复现目标结构分布。

表 12　规模化数据集结构一致性评价

Tab. 12　Structural Consistency Evaluation of the Large-Scale Dataset

板块	高相似度 MAE	高相似度 RMSE	低相似度 MAE	低相似度 RMSE
背景技术	0.0215	0.0335	0.0232	0.0235
发明内容	0.0954	0.0956	0.0208	0.0219
附图说明	0.0317	0.0354	0.1153	0.1249
具体实施方式	0.0714	0.0726	0.1920	0.1943

在语义一致性方面，表 13 显示，规模化数据集中伪交底书与对应源说明书的平均 F1 分数为 0.9074，其中背景技术、发明内容和具体实施方式分别达到 0.9847、0.9127 和 0.9112。该结果表明，在规模化构造过程中，生成的伪交底书与源说明书仍保持较高的句子级语义一致性，说明经过句子筛选与语义重写后，源说明书中的主要技术语义信息得到了较好的保留。

表 13　规模化数据集语义一致性评价

Tab. 13　Semantic Consistency Evaluation of the Large-Scale Dataset

技术领域	背景技术	发明内容	附图说明	具体实施方式	平均
0.8498	0.9847	0.9127	0.8777	0.9112	0.9074

综合上述结果，在数据规模由 50 篇扩展至 200 篇后，所构建的伪交底书在多数结构分布指标上仍与真实数据的目标分布保持较低误差，同时与源说明书保持较高的语义一致性，表明本文方法具有进一步用于规模化伪交底书数据构建的可行性。

6　结语

本文针对专利自动撰写中高质量“交底书—说明书”配对语料稀缺的问题，提出了一种基于结构分布先验的伪交底书逆向构造方法。首先，基于真实文档对分析句级结构映射特征，并与现有替代数据进行比较；随后，结合“双指标—双阈值”分级机制、概率采样与大语言模型语义重写，实现伪交底书构造。实验结果表明，该方法能够较好地复现真实文档对的主要结构分布特征，并保持较高的技术语义一致性。基线对比与消融实验进一步验证了细粒度结构先验在改善相似度分布一致性方面的作用。在此基础上，本文构建了 200 组“伪交底书—说明书”文档对，为后续专利自动撰写研究提供数据支持。未来将进一步探索内容感知的结构建模方法，以提升该方法在不同技术领域及单篇专利上的适应性与泛化能力。

参考文献

[1] Knappich V, Hätty A, Razniewski S, et al. PAP2PAT: benchmarking outline-guided long-text patent generation with patent-paper pairs[C]//Findings of the Association for Computational Linguistics: ACL 2025. 2025: 9524-9554.

[2] Wang Q, Ni S, Liu H, et al. Autopatent: a multi-agent framework for automatic patent generation[J]. arXiv preprint arXiv:2412.09796, 2024.

[3] Jiang L, Goetz S M. Natural language processing in the patent domain: a survey[J]. Artificial Intelligence Review, 2025, 58(7): 214.

[4] Lee J S, Hsiang J. Patent claim generation by fine-tuning OpenAI GPT-2[J]. World Patent Information, 2020, 62: 101983.

[5] Christofidellis D, Torres A B, Dave A, et al. Pgt: a prompt based generative transformer for the patent domain[C]//ICML 2022 workshop on knowledge retrieval and language models. 2022.

[6] Jiang L, Zhang C, Scherz P A, et al. Can Large Language Models Generate High-quality Patent Claims?[C]//Findings of the Association for Computational Linguistics: NAACL 2025. 2025: 1272-1287.

[7] Wang J, Mudhiganti S K R, Sharma M. Patentformer: A novel method to automate the generation of patent applications[C]//Proceedings of the 2024 conference on empirical methods in natural language processing: industry track. 2024: 1361-1380.

[8] Yang R, Mudhiganti S K R, Sharma M. PatentVision: A multimodal method for drafting patent applications[C]//Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track). 2026: 397-405.

[9] Sharma E, Li C, Wang L. BIGPATENT: A large-scale dataset for abstractive and coherent summarization[C]//Proceedings of the 57th annual meeting of the association for computational linguistics. 2019: 2204-2213.

[10] Suzgun M, Melas-Kyriazi L, Sarkar S, et al. The harvard uspto patent dataset: A large-scale, well-structured, and multi-purpose corpus of patent applications[J]. Advances in neural information processing systems, 2023, 36: 57908-57946.

[11] Marco A C, Sarnoff J D, Charles A W. Patent claims and patent scope[J]. Research Policy, 2019, 48(9): 103790.

[12] Risch J, Alder N, Hewel C, et al. Patentmatch: a dataset for matching patent claims & prior art[J]. arXiv preprint arXiv:2012.13919, 2020.

[13] Jiang L, Scherz P A, Goetz S. Patent-cr: A dataset for patent claim revision[C]//Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2025: 2300-2314.

[14] Li S, Hu J, Cui Y, et al. DeepPatent: patent classification with convolutional neural networks and word embedding[J]. Scientometrics, 2018, 117(2): 721-744.

[15] Arts S, Hou J, Gomez J C. Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures[J]. Research policy, 2021, 50(2): 104144.

[16] Lee J S, Hsiang J. Patent classification by fine-tuning BERT language model[J]. World Patent Information, 2020, 61: 101965.

[17] Miao R, Chen X, Hu L, et al. PatSTEG: Modeling Formation Dynamics of Patent Citation Networks via The Semantic-Topological Evolutionary Graph[C]//2023 IEEE International Conference on Data Mining (ICDM). IEEE, 2023: 1229-1234.

[18] Pairolero N A, Giczy A V, Torres G, et al. The artificial intelligence patent dataset (AIPD) 2023 update[J]. The Journal of Technology Transfer, 2025: 1-24.

[19] Papineni K, Roukos S, Ward T, et al. Bleu: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002: 311-318.

[20] Zhang T, Kishore V, Wu F, et al. Bertscore: Evaluating text generation with bert[J]. arXiv preprint arXiv:1904.09675, 2019.

[21] Xu M. Text2vec: Text to vector toolkit[J]. https://github. com/shibing624/text2vec, 2023.

[22] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). 2019: 4171-4186.

附录 A：提示词模板

A.1 句子复述提示词

你是一名具有丰富经验的专利代理人，请对提供的专利申请文本句子进行改写。在改写过程中，应在不改变原句技术含义和完整语义信息的前提下，对表达方式进行重构，避免与原句在措辞或句式上的直接重复。改写后的句子需保持内容准确、逻辑严谨，并符合专利文本的专业表述风格。

待改写文本：{sentence}
输出：

A.2 句子判定提示词

你是一位资深的审查员，具备极强的语言逻辑分析能力。接下来我将给你一系列句子，请对比“原句”与“待测句”，判断待测句是否为原句的复述，即两者表达的实质语义是否完全相同。判断语句相同要求：

-核心事实一致：句中的技术特征、动作关系、逻辑顺序必须保持一致。

-允许表达差异：允许同义词替换、句式变换、或修辞。

-排除干扰：如果待测句只是包含相似的关键词，但描述的是不同的过程、结构或逻辑，则判定为不同。

你只需要回答[相同]或者[不相同]，不需要提供理由。

原句：{sentence1}
待测句：{sentence2}
输出：

A.3 大模型直接生成交底书提示词

你是一名熟悉中国发明专利技术文档的专业撰写人员。请依据给定专利说明书原文，逆向整理为技术交底书中的对应板块。准确保留技术事实，不得编造原文未提供的技术特征、实验结果或技术效果。请将下面专利说明书的{section}板块，直接改写为技术交底书的{section}板块。要求：

1. 准确保留本板块的技术信息、技术术语、参数、结构和逻辑关系；不得引入原文没有的技术事实。
2. 使用适合技术交底书的客观、清晰的技术语言组织内容，可调整表达方式。
3. 根据内容本身组织文本。

4. 仅输出该板块正文，不要输出标题、说明或 Markdown 代码块。

原始说明书{section}正文：{content}

输出：

A.4 句子改写提示词

你是一名具有多年经验的研发工程师，正在整理一份技术交底书初稿。你的任务是：将一段已经被专利代理人“法律化”的文本，还原为工程师撰写的技术交底表达。

改写原则

1. 风格定位：
   - 保持技术书面语，而不是口语
   - 表达应清晰、克制、偏说明性、较原句更为简练
2. 去法律化：
   - 删除“本发明提供了一种…”等专利常用句式
   - 避免权利要求式表达
   - 保留技术要点和逻辑严谨性
3. 逻辑重构：
   - 可以改变原句顺序
   - 但必须保证逻辑连贯、工程上可理解
4. 表达约束：
   - 优先用技术等价改写，而不是简单同义词替换
   - 短句之间用逗号连接，不允许使用分号或句号，只有最后可以用句号结尾
5. 信息完整性：
   - 不得遗漏原文中的关键技术信息
   - 不得随意引入新技术内容

输出要求：

- 仅输出要求改写的句子文本
- 不要添加解释说明或额外注释

待改写句子：{sentence}

输出：

附录 B：Full Prior 采样结果统计

本附录对 Full Prior 在结构先验指导下的操作采样结果进行补充统计。

表 B.1 汇总不同板块的句子数量及 KEEP、DISTILL 和 DELETE 三类操作的分布情况。其中，由于技术领域板块存在特殊处理，不在此处统计。

表 B.1　不同板块的实际操作比例

板块	总句子数	KEEP (%)	DISTILL (%)	DELETE (%)
背景技术	677	82.52	17.18	0.30
发明内容	1652	22.34	77.24	0.42
附图说明	174	2.30	89.08	8.62
具体实施方式	3165	23.10	57.69	19.21

注：各板块的操作比例根据对应板块的实际采样数量计算，总体比例根据全部句子计算。

表 B.2 汇总了各相似度类别及采样操作的总体分布。统计共涉及 5718 个句子（包括技术领域板块），其中原始高、中、低相似度句子分别为 1641、3380 和 697 句。在操作采样结果中，DISTILL 占比最高，共 3416 句，其次为 KEEP（1670 句）和 DELETE（632 句）。

表 B.2　相似度类别与采样操作的总体统计

指标	句子数
总体	5718
原始高相似度	1641
KEEP 操作	1670
原始中相似度	3380
DISTILL 操作	3416
原始低相似度	697
DELETE 操作	632

附录 C：实验实现细节与参数配置

本附录对实验涉及的文本预处理、相似度计算、模型生成及概率采样等关键环节进行补充说明。

C.1 文本预处理与相似度计算

实验首先依据专利文档的章节结构提取各板块文本，并以句号、问号、感叹号及分号等标点进行句子切分。句子切分后，过滤长度不足 8 个字符的句子。对于切分后的句子，采用 Jieba 工具进行中文分词。

句子相似度采用 BLEU 和 Jaccard 两项指标进行度量。其中，BLEU 基于分词后的词序列计算，使用 NLTK 提供的 sentence_bleu 函数，采用默认的四元语法权重及 SmoothingFunction().method1 平滑方法；Jaccard 则基于分词去重后的词集合计算。对于待评价句子，分别计算其与候选句子集合之间的最大 BLEU 和最大 Jaccard，并结合双指标阈值确定相似度类别。

相似度阈值依据正文第 3.1.2 节所述方法确定。其中，BLEU 的高、低阈值分别为 0.4381 和 0.0300，Jaccard 的高、低阈值分别为 0.6339 和 0.1383。当两项指标均达到对应的高阈值时，判定为高相似度；当两项指标均低于对应的低阈值时，判定为低相似度；其余情况归为中相似度。

C.2 大语言模型生成配置

实验采用 DeepSeek v3.2 完成复述样本生成、语义关系判定及伪交底书构造过程中的语义重写。模型采用对话式调用方式，通过系统提示词设定专利撰写任务角色，并结合用户提示词提供任务要求及相关文本。生成过程中将 temperature 设置为 0.3，其余参数均采用 API 默认配置。各任务通过相应的提示词模板进行约束，具体内容见附录 A。

C.3 概率采样与数据构造

伪交底书构造阶段采用基于板块与位置分布的概率采样策略。首先，根据句子在所属板块中的相对位置，将各板块划分为0–30%、30–70%和70–100%三个区间，并依据真实文档对中统计得到的相似度分布确定各区间的操作概率。

对于各区间内的句子，按照相应概率随机分配 KEEP、DISTILL 或 DELETE 操作。KEEP 直接保留原始句子，DISTILL 通过大语言模型进行语义重写，DELETE 则移除相应句子。完成处理后，按照原始句序重新组织保留及改写后的文本，形成伪技术交底书。

概率采样采用 Python 内置的 random 模块实现，并设置随机种子为 42。

C.4 实验环境与数据说明

实验在 Windows 11 环境下开展，Python 版本为 3.11.9。主要使用 Jieba、NLTK、PyMuPDF、python-docx 和 Pandas 等工具完成文本处理、相似度计算及结果统计。

实验使用 50 组真实技术交底书—专利说明书配对文档开展结构分布分析，并结合公开专利说明书扩展构造数据集，最终形成 200 组伪交底书—说明书配对数据。涉及真实技术交底书的数据按照相应保密要求进行管理。