arXiv:2601.00166v1 [cs.CL] 1 Jan 2026

Pat-DEVAL：面向专利说明书的法律思维链评估

Yongmin Yoo麦考瑞大学
yooyong
min91@gmail.com

KrisW Pan亚马逊
krisk
wpan@gmail.com

摘要

专利说明书必须提供全面的技术公开，同时满足充分公开和书面描述要求等严格的法律标准。尽管大语言模型已实现端到端的自动化专利撰写，但现有评估方法无法评估专利说明书特有的长篇结构连贯性和法定合规性。我们提出 Pat-DEVAL，这是首个专门针对专利说明书正文的多维评估框架。Pat-DEVAL 借助大语言模型评审范式，引入法律思维链（CoLT），这是一种受法律约束的推理机制，可强制执行按顺序进行的、针对专利法的分析。在 Pap2Pat-EvalGold 数据集上经专利专家验证的实验表明，Pat-DEVAL 的皮尔逊相关系数达到 0.69，显著优于基线指标和现有大语言模型评估器。值得注意的是，该框架在法律专业合规性方面的相关系数高达 0.73，这证明明确注入法定约束对于捕捉细微的法律有效性至关重要。
Pat-DEVAL 为确保技术可靠性和法律合规性确立了新标准，并为自动化专利撰写系统的实际部署提供了坚实的方法论基础。

1 引言

随着全球专利申请量在2024年达到约370万件的历史新高，并连续第五年增长（世界知识产权组织，2024年），为加速技术创新、减轻法律专业人士负担而开发的自动化专利撰写工具，其需求比以往任何时候都更加突出。尤其是，创新竞争日益激烈，这一点从美国专利商标局仅在2024年就收到逾79万件申请可见一斑（美国专利商标局，2024年）。专利说明书的撰写

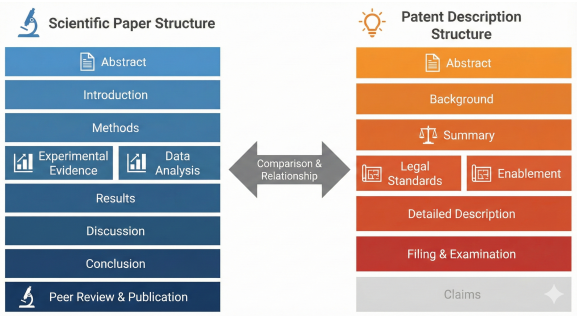

图 1：科学论文与专利说明书在结构和修辞范式上的概念性比较。

是一项知识密集型工作，不仅要求对发明进行详尽的技术阐述，还要求严格遵循法律标准，例如充分公开和充分披露（Jiang and Goetz，2024年）。尽管近期研究已利用大语言模型（LLM）将科学论文转化为专利说明书，但由于科学论文与专利说明书在修辞目的上存在根本差异，有效的知识转移仍面临挑战：前者强调实验性证据，而后者优先考虑法律保护和全面的技术公开（Murray and Stern，2007）。

尽管生成式模型取得了这些进展，但对生成的专利说明书的评估仍主要依赖传统的自然语言处理指标，而这些指标无法捕捉法律领域所必需的结构连贯性和修辞模式（Papineni 等人，2002；Lin，2004；Zhang 等人，2020b）。这些传统指标侧重于词汇重叠或基于嵌入的相似度，在评估专利文件的法律有效性或跨章节逻辑流方面存在根本局限。

近期的 LLM 即评审者方法（例如 PatentScore（Yoo 等人，2025年）通过处理专利权利要求独有的逻辑一致性和范围界定问题，推动了专利权利要求评估的发展。

然而，专利说明书面临着本质上不同的挑战：它们构成了文档的主体，必须提供充分的技术公开，以满足《美国法典》第35编第112条(a)款规定的充分公开和书面描述要求（美国众议院，2024年）。因此，评估专利说明书需要考察长篇文本的结构完整性、跨章节逻辑流，以及本领域普通技术人员（PHOSITA）能否据此复现，而这些方面在权利要求评估中基本不存在。因此，现有框架无法可靠判断生成的说明书在技术上是否合理、在法律上是否合规，从而阻碍了自动化专利撰写系统的实际应用。随着研究界转向端到端专利生成系统，这种缺乏专门评估方法的问题正变得日益严峻（Knappich 等人，2025年；Wang 等人，2024年）。这些研究通过提出各种自动化专利撰写框架，显著推动了该领域的发展；然而，它们仍缺乏一种全面的方法替代方案，以验证说明书正文的结构完整性和法律有效性，导致生成的结果未经法律核验。

为解决这些局限，我们提出 Pat-DEVAL，这是一种新的结构感知评估框架，旨在弥合专利说明书中技术精确度与法律要求之间的差距。Pat-DEVAL 使用大语言模型作为评估器，从四个关键维度分析生成描述：技术内容忠实度（TCF）和数据精确度（DP）用于验证技术完整性，结构覆盖度（SC）和法律专业合规性（LPC）用于确保符合专业撰写规范。

本研究的贡献如下：

- **首个面向专利说明书的结构感知框架**：我们提出 Pat-DEVAL，这是首个专为专利说明书设计的多维框架，弥补了评估长篇文本结构连贯性和法定充分公开方面的关键空白，而以往的指标主要聚焦于权利要求。

- **法律思维链（CoLT）**：我们提出一种受法律约束的推理策略，通过注入法定约束并强制进行顺序式专利法分析来扩展标准思维链，从而可靠地模拟本领域普通技术人员严谨的审查过程。

- **无参考评估范式**：我们建立了一个源锚定框架，直接依据源技术评估描述，并适应有效专利撰写中“一对多”的特性，而传统的基于参考文本的指标在这种情况下则会失效。

2 相关工作

2.1 自动化专利生成

自动化专利生成已从早期的统计摘要模型发展到由大语言模型驱动的完整说明书撰写。BigPatent 的发布表明，专利文档的语篇结构比普通新闻文本复杂得多（**?**）。因此，为提升长文档抽象式摘要的表现，人们提出了间隙句生成等预训练目标，通过掩蔽并还原关键句子来实现这一点（Zhang 等人，2020a）。

后续研究针对处理专利说明书篇幅过长所带来的技术局限性展开了探索。例如，BigBird 的提出（Zaheer 等人，2020年），其利用稀疏注意力机制高效处理长达数千个词元的上下文，使处理长序列专利文本成为可能。在这些基础技术的推动下，早期研究尝试对 GPT-2 进行微调，以根据发明摘要生成专利权利要求（Lee 和 Hsiang，2020年）；然而，这些研究大多仅限于生成相对较短的文本片段。

近年来，尝试从科学论文端到端生成完整专利说明书的方法逐渐受到关注（Knappich 等人，2025年）。尽管这些研究借助大语言模型显著提升了生成文本的流畅度，但在结构化验证贯穿完整说明书正文的逻辑一致性和法律上所要求的充分公开方面，仍然存在不足。

2.2 基于大语言模型的评估框架

数十年来，自然语言生成（NLG）的评估一直由基于 n 元语法重叠指标的方法主导，例如 BLEU、ROUGE 和 ME- TEOR（Papineni 等人，2002；Lin，2004；Banerjee 和 Lavie，2005年）。这些指标虽然计算效率高，却因无法与人类对流畅度和语义等价性的判断良好相关而广受诟病

。为解决这些不足，研究者提出了基于嵌入的指标，例如 BERTScore（Zhang 等人，2020b）和 MoverScore（Zhao 等人，2019），它们分别利用上下文嵌入和地球移动距离。尽管这些方法缓解了词汇不匹配的问题，但它们本质上仍依赖表层或向量相似度，无法捕捉长篇生成中固有的复杂逻辑矛盾或事实性幻觉。

近年来，评估范式已转向由大语言模型担任评审者，利用大语言模型的推理能力充当代理评估者。例如，GPTScore 被提出用于在不进行微调的情况下，依据模型的条件概率评估文本质量（Fu 等人，2023年）；G-Eval 则表明，在摘要任务中采用思维链（CoT）提示，能够与人类专家实现更高的一致性（Liu 和其他人，2023年）。此外，Prometheus 等专用评估模型也已开发出来，以提高反馈的可靠性（Kim 等人，2024年）。然而，这些通用框架针对标准领域进行了优化，缺少评估严格法律标准所需的领域专门知识，例如专利法中的充分公开要求。

2.3 专利专用评估指标

在生成式人工智能时代到来之前，专利质量评估主要依赖文献计量指标和检索指标。研究通常利用前向引文、专利族规模或权利要求数量等统计元数据，推断专利的经济价值或技术影响（Squicciarini 等人，2013）。在自然语言处理领域，评估重点在于下游任务的表现，例如自动专利分类或信息抽取，而非文本本身的质量（Trappey 等人，2017）。尽管这些方法有助于分析现有数据库，却未能提供任何方法论依据，以验证新生成专利草案的内部连贯性或法律有效性。

随着自动化撰写技术的出现，对生成内容进行评估的需求也随之产生。尤其值得一提的是，PatentScore 被提出作为一种框架，利用大语言模型评估专利权利要求的法律一致性和语义忠实度（Yoo 等人，2025年）。尽管这项工作在评估权利要求方面取得了重大进展，但权利要求是一种精炼且高度符合逻辑的结构，用于界定法律保护范围，而 PatentScore 并未涉及专利说明书所面临的根本不同的挑战。

与权利要求相比，专利说明书通常篇幅仅有数段，侧重于演绎逻辑的严谨性以及前序基础的一致性；而专利说明书构成了整份文件的主体，往往包含数千字。这些说明书承担着独特的法律用途，即提供充分的技术公开，以满足《美国法典》第35编第112条(a)款规定的充分公开和书面描述要求（美国众议院，2024年）。因此，评估说明书需要考察其长篇结构的完整性，例如是否涵盖背景技术、发明内容、附图简要说明和详细说明等必备部分。此外，还需分析各部分之间的逻辑脉络，以及本领域普通技术人员能否重复实施；这些方面在权利要求评估中往往无关紧要或完全缺失。因此，现有框架在评估自动生成的专利说明书的技术可靠性和结构完整性方面存在重大缺口。

3 方法论：Pat-DEVAL

现有的自然语言生成（NLG）指标和通用的基于大语言模型的评估框架，在评估具有法律约束力的文档时存在根本性局限。专利说明书并非只是信息的序列，而是一种复杂的逻辑构造，能够将技术构想转化为法律权利。为弥合这一差距，我们提出了 Pat-DEVAL，这一框架将专利审查员的专业判断过程建模为可计算的形式。我们的方法论核心在于法律思维链（CoLT）机制，它确保法律领域知识与神经网络的推理轨迹相一致。

3.1 框架概述

Pat-DEVAL 旨在评估生成描述 D_{gen} 是否不仅准确反映源技术 R（此处定义为涵盖摘要、方法论和实验的完整学术论文），还评估其是否提供了足够的技术公开，以支持权利要求的保护范围。为此，我们模拟

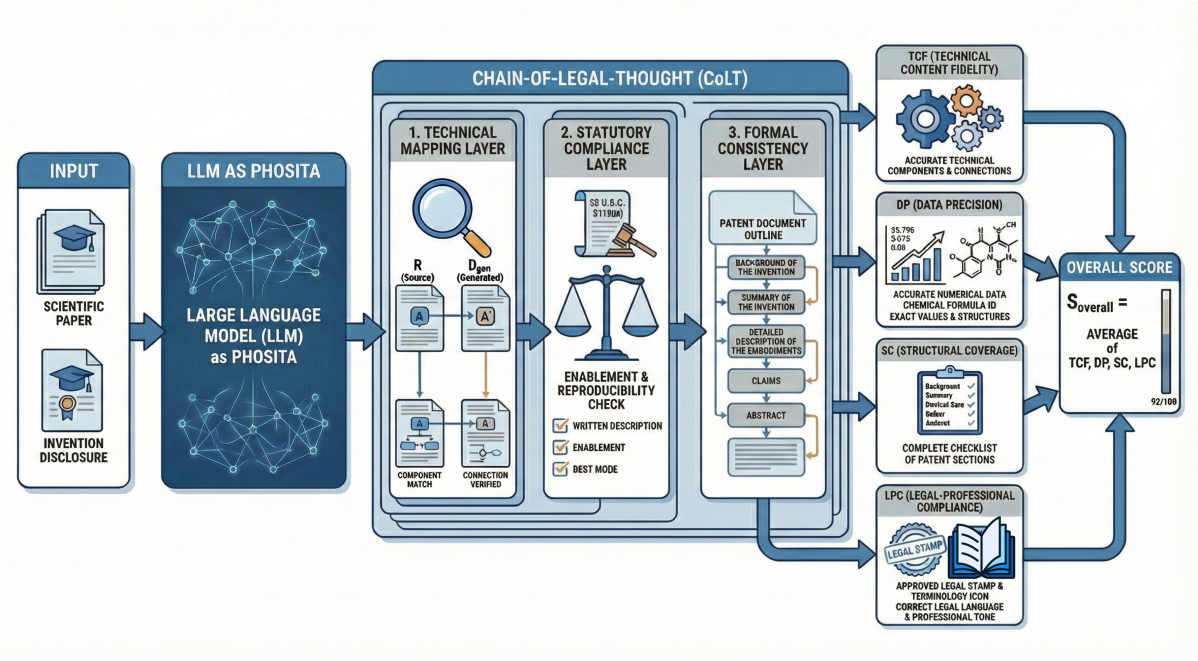

图2：Pat-DEVAL 框架概览。法律思维链（CoLT）机制利用模拟本领域普通技术人员（PHOSITA）的大语言模型，通过三个专利法特定层（技术映射、法定合规和形式一致性）强制进行顺序推理。该过程会生成四项维度评分（TCF、DP、SC、LPC），并将其汇总为总体质量评分。

大语言模型评估器 $\mathcal{M}$ 作为本领域普通技术人员（PHOSITA）。该评估构成一个结构化推理流程，其中$\mathcal{M}$ 通过 CoLT 进行法律推理，并随后从四个不同维度得出量化评分。

3.2 法律思维链（CoLT）

法律思维链（CoLT）代表着对标准思维链（CoT）提示方式的根本性转变。不同于会带来高延迟的多轮对话方法，CoLT 采用单次生成、强制约束的提示策略。它引入了结构化的推理轨迹，要求在单次生成过程中依次经过专利法特定的分析层。

我们明确将外部法定约束（例如《美国法典》第35编第112条(a)款（美国众议院，2024年））注入系统提示词中，防止模型在推理轨迹完整阐述之前输出最终分数。为实现这一点，CoLT 要求评估器大语言模型在评分前，依次完成三个特定领域的分析层次。表 1 展示了这些层次如何检测源技术与生成输出之间具体的技术和法律差异R D_{gen}。

技术映射层 这一层对源文档 R 与生成描述 D_{gen}进行细致的逐项对比分析。它会识别事实性幻觉或遗漏，为忠实度评分提供证据基础。

法定合规层 通过模拟本领域普通技术人员（PHOSITA）的视角，该层评估充分公开要求。它会检查是否仅凭 D_{gen}即可在无需过度试验的情况下复现发明，并对缺失的实施例或实施细节予以扣分。

形式一致性层 该层评估各章节之间的逻辑连贯性，以及对专业修辞规范的遵循情况，确保文档符合可专利性所要求的结构标准。

3.3 评估分类体系

CoLT 得出的推理结果被归纳为四项形式化指标。每个维度均采用 1 至 5 分的李克特量表进行评估，其中 1 分表示未达到法律或技术标准的致命缺陷，5 分则代表专业水准的撰写质量。

- **技术内容忠实度（TCF）：** 衡量

层	源技术 (R)	生成专利 (D_{gen})	CoLT 推理（内部）
技术映射	提出了一种“动态 Gating 模块”，使用 Gumbel-Softmax 实现稀疏性。	提到了一个“过滤单元”，用于减少计算。	遗漏之处：具体的数学机制（Gumbel-缺少 Softmax）；已替换为通用术语。
法定合规性 ance	超参数 λ 设置为 0.01，以实现稳定收敛。	建议使用“适当的定值”以确保稳定性。	充分公开问题：“适当值”对于本领域普通技术人员来说过于模糊，无法复现结果。
形式上的 性 **一致性**	包含一张流程图，见图 3，展示了该流程。	简要说明提及图3，但具体说明中省略了它。明。	章节间不一致：图 3在附图列表中列出，但正文中从未对其进行描述。

表1：A 评估过程中识别具体缺陷的CoLT层代表性示例 s.

源技术核心机制的失真或遗漏。

- **数据精确度（DP）：** 评估实验值、参考标号和化学式是否完全匹配。

- **结构覆盖度（SC）：** 确定专利法规要求的强制性部分是否完整。

- **法律专业合规性（LPC）：** 验证是否符合充分公开要求以及是否采用适当的法律措辞。

形式上，每个得分 $s_i \in \{1, \ldots, 5\}$ 及其对应的理由 ρ_i 均通过函数f_{CoLT}得出，该函数以提示词 P、法律约束 L以及源文件和生成文件作为输入：

$$s_i, \rho_i = f_{CoLT}(P, L, R, D_{gen}) \quad (1)$$

总体 Pat-DEVAL 评分通过四个维度评分的非加权平均值计算得出。尽管法律有效性通常是二元的，但我们采用这种平均方式，以便从专利撰写的不同方面对生成质量进行全面评估。详细的系统提示词，包括具体的法定约束以及每个维度的 5 分制评分标准，见附录A。

4 实验

在本节中，我们将介绍用于验证 Pat-DEVAL 框架有效性的实验设置和评估结果。我们的首要目标是评估该框架衡量生成的专利说明书在法律和技术方面质量的准确性，并量化其与人类专家判断之间的相关性。为此我们构建了一个涵盖多个技术领域的基准数据集，并将其与传统的 n 元语法指标以及最先进的基于大语言模型的评估器进行比较分析。我们旨在重点揭示仅依赖文本相似度的现有方法的局限性，并证明 CoLT 机制是捕捉专利说明书独特结构特征和法定约束的关键组成部分。此外，来自认证专利专业人员的定性评估数据作为金标准，用于全面考察 Pat-DEVAL 的可靠性和可解释性。

4.1 实验设置与数据集

为评估 Pat-DEVAL 的有效性，我们建立了一个稳健的实验流程，包括数据整理、候选生成和专家标注。

数据集构建 我们构建了 Pap2Pat-EvalGold，这是一个源自 Pap2Pat（Knappich 等人，2025 年）语料库的高精度数据集。为确保真实标注的可靠性，我们使用 Sentence-BERT 测量标题与摘要之间的语义相似度，并设置最低 BERTScore 阈值为 0.8，以此筛选语料库。我们还进一步验证了作者重合率，确保其达到或超过 0.5，最终选出了 146 组高质量的论文 — 专利配对。

人工标注 为建立高保真金标准，我们对整个 Pap2Pat-EvalGold 数据集（$N = 146$）进行了全面的人工评估，每个样本均由三位认证专利专业人士独立评分。我们使用克里彭多夫阿尔法系数和组内相关系数（ICC）验证了这些评判的可靠性。分析结果显示，阿尔法系数平均值为 0.764，且

ICC 为 0.81。该 ICC 分数超过了通常用于判定可靠性极佳的 0.80 阈值。值得注意的是，考虑到专利说明书固有的语言歧义和复杂的法律技术要求，能够实现如此高的一致性意义重大。这种高度一致性证实，我们的 Pap2Pat-EvalGold 数据集可作为高保真基准，从而最大限度地降低后续将自动化指标与其进行相关性分析时出现个体偏差的风险。

候选生成 我们使用该数据集中的学术论文作为源输入，并采用 Llama-3.1-70B 生成相应的专利描述。为确保可重复性并尽可能减少随机性，所有生成任务的解码温度均固定为 0.3，Top-k 参数设为 10。这些生成描述将作为评估的候选对象。

4.2 主要结果

表2 展示了自动评估指标与专利质量四个维度的人类专家判断之间的皮尔逊相关系数。结果表明，Pat-DEVAL 与专业标准的吻合度最高，相关系数达到 0.69。这一表现显著优于所有基线方法，且具有统计显著性（p 值小于 0.01）。

指标	TCF	DP	SC	LPC	平均值
传统指标与嵌入指标					
BLEU	0.12	0.08	0.05	0.03	0.07
ROUGE-L	0.15	0.10	0.07	0.04	0.09
BERTScore	0.31	0.27	0.18	0.15	0.23
LLM 即评审者（骨干模型：***Qwen3-32B***）					
GPTScore	0.38	0.35	0.29	0.22	0.31
标准思维链	0.45	0.41	0.38	0.32	0.39
Prometheus-2	0.51	0.48	0.44	0.39	0.46
G-Eval	0.58	0.54	0.51	0.45	0.52
所提框架（骨干模型：***Qwen3-32B***）					
Pat-DEVAL（我们的方法）	**0.68**	**0.72**	**0.64**	**0.73**	**0.69**

表 2：在 Pap2Pat-EvalGold 上与人工判断的皮尔逊相关系数（r）（$N = 146$）。在统一采用 Qwen3-32B 作为骨干模型的条件下，Pat-DEVAL 的表现显著优于基线方法，尤其是在法律专业合规性（LPC）维度上（$p < 0.01$）。

表层指标的失效传统的 n 元语法指标（如 BLEU 和 ROUGE-L ）与人类专家的评价几乎没有相关性，相关系数仅为 0.09 或更低。这证实了词汇重叠并不能有效地代表专利质量，因为针对同一发明的有效专利说明书往往采用多样的术语和结构，这体现了专利撰写中固有的一对多映射。尽管基于嵌入的指标（如 BERTScore）通过捕捉语义相似性，将相关系数提升至 0.23，表现略有改善，但仍不足以评估可专利性所要求的结构连贯性和法定合规性。

通用及微调大语言模型的局限性尽管基于大语言模型的评估器总体上优于传统指标，但通用提示策略在法律领域存在明显局限。标准思维链和 GPTScore 等方法难以区分语言流畅的文本与具有法律效力的充分公开，其相关系数分别为 0.39 和 0.31。一个值得注意的发现是，Prometheus-2 是专门针对评估任务进行微调的模型。尽管经过了专门训练，Prometheus-2 的相关系数仅为 0.46，低于基于提示的 G-Eval 框架所达到的 0.52。这一差异表明，通用评估任务的微调并不会自动迁移到专利领域，除非具备特定的法律依据。因此，这些模型在法律专业合规性维度上的表现有限，相关系数介于 0.39 至 0.45 之间。

Pat-DEVAL 的优势 Pat-DEVAL 与人工判断的吻合度更高，比最强基线 G-Eval 高出 0.17。法律专业合规性维度上的性能差距最为显著：Pat-DEVAL 的相关性达到 0.73，而 G-Eval 为 0.45。这验证了我们的假设：法律思维链机制能够成功模拟本领域普通技术人员的推理过程，使模型得以严格评估充分公开和书面描述要求。由于所有基于大语言模型的评估器均采用相同的 Qwen3-32B 主干模型，这一性能提升应归功于所提出的领域专用方法，而非底层模型能力的差异。

4.3 消融研究

为验证构成 Pat-DEVAL 框架的关键组件各自的贡献，我们使用

Qwen3-32B 主干模型。表3总结了移除法律思维链逻辑、本领域普通技术人员角色和维度分解结构后观察到的性能变化。

配置	TCF	DP	SC	LPC	Avg.
Pat-DEVAL（完整版本）	**0.68**	**0.72**	**0.64**	**0.73**	**0.69**
不含法律思维链逻辑	0.49	0.45	0.41	0.35	0.43
不含本领域普通技术人员角色	0.61	0.65	0.58	0.62	0.62
不含分解	0.52	0.49	0.46	0.41	0.47

表 3：在 Pap2Pat-EvalGold 上进行的消融分析（骨干模型：Qwen3-32B）。数值表示与人工判断的皮尔逊相关系数（r），突显了各框架组件的贡献。

法律思维链的影响移除法律推理机制后，观察到的性能下降幅度最大。在缺少法律思维链逻辑的配置中，我们将基于《美国法典》第35编的具体法条推理步骤替换为通用提示语“让我们一步一步地思考”。因此，平均相关性从 0.69 骤降至 0.43。值得注意的是，法律专业合规维度的相关性降至 0.35，与标准思维链基线相当。这有力表明，Pat-DEVAL 的性能提升并非源于大语言模型的一般推理能力，而是源于明确注入的特定领域法律知识。

本领域普通技术人员角色的作用如果未赋予评估器本领域普通技术人员这一角色，平均相关性会从 0.69 适度下降至 0.62。尽管模型仍保留了一定的技术内容评估能力，但其表现下降显而易见。明确设定专家视角有助于使模型的判断与专利律师采用的审查标准保持一致，尤其是在判断技术术语是否恰当以及发明水平时。

维度分解的效果最后，我们验证了将专利质量划分为四个维度的有效性。若要求模型不区分维度，直接给出整体评分，相关性仅为 0.47。这表明，专利说明书的质量过于复杂，无法压缩为单一标量值。将评估目标分解为具体维度的策略，确保模型关注各个方面，因此成为一种重要机制，可减少抽象评分中常见的幻觉，并提高精确度。

5 主要发现与讨论

基于实验结果和消融研究，我们得出了关于自动化专利生成评估的以下主要学术见解。

词汇相似度与法律有效性的脱钩 我们的结果表明，传统的 n 元语法指标（BLEU、ROUGE）与人类专家判断之间不存在有意义的相关性。这证实，词汇重叠并不能有效衡量专利质量。与通用摘要不同，专利说明书呈现一对多映射关系：同一发明构思可以通过多种语言结构和策略性法律术语来表述。因此，较高的文本相似度并不保证技术准确性或法定合规性，这使得 Pat-DEVAL 这样的领域特定评估框架不可或缺。

明确法定约束的效力 消融研究中最重要的发现是法律思维链（CoLT）机制发挥着关键作用。移除 CoLT 后，法律专业合规性（LPC）维度的相关性从 0.73 大幅下降至 0.35，这表明大语言模型的一般推理能力不足以胜任高风险法律评估。明确注入法定要求（例如《美国法典》第35篇第112条）至关重要，这能将通用评估器转变为专业法律裁判者，使其能够评估充分公开和书面说明标准。

通过本领域普通技术人员角色实现专家对齐将本领域普通技术人员角色分配给评估者后，其与专利专业人士的判断显著更加一致。这表明，模型的判断对所定义的专家视角较为敏感，使其能够超越语言流畅性，评估技术公开内容的适当性以及本领域普通技术水平。这一发现意味着，在专业领域评估中，提供正确的专业背景与模型自身的能力同等重要。

通过维度分解实现精准性从整体评分转向多维度方法（TCF、描述充分性、法定合规性和LPC）

带来了显著的性能提升（$r = 0.47 \rightarrow$0.69）。这种分解机制发挥了扎根作用，降低了抽象幻觉的风险。通过强制模型针对每个具体的法律和技术维度生成结构化推理轨迹，Pat-DEVAL 确保最终评估具有更高的精准性和更好的可解释性。

6 结论

本文介绍了 Pat-DEVAL，这是首个专为专利说明书正文设计的多维评估框架。该框架通过 CoLT （法律思维链）机制模拟本领域普通技术人员（PHOSITA），成功弥合了技术摘要与法律有效性评估之间的关键鸿沟。不同于依赖词汇重叠的传统表层指标，Pat-DEVAL 强制采用结构化推理路径，并纳入明确的法定约束，例如《美国法典》第35编第112条(a)款规定的充分公开和书面描述要求。

我们在 Pap2Pat-EvalGold 数据集上的实验结果表明，Pat-DEVAL 与人类专家判断的皮尔逊相关系数达到 0.69，显著优于所有基线指标。消融研究进一步凸显了法律思维链逻辑不可或缺，尤其是在法律专业合规性（LPC）维度，其相关系数达到 0.7。这些发现证实，要确保自动化专利撰写在法律层面 sound，不能仅依靠一般的语言流畅度；还必须严格引入特定领域的法律推理。

随着大语言模型驱动的专利生成转向端到端生产，Pat-DEVAL 为验证生成的说明书的结构完整性和技术忠实度确立了新标准。该框架提供了一个稳健的方法论基础，以确保自动生成的专利草案在技术上准确且具有法律效力。

局限性

本研究仅聚焦于专利说明书的说明书正文，以评估技术忠实度和法定充分公开，并有意将权利要求排除在评估范围之外。这种区分源于权利要求与说明书在语言属性和法律属性方面存在根本差异：权利要求需要分析逻辑包含关系和法律边界的清晰度，而说明书正文则必须涵盖广泛的技术披露，并提供足够的信息，使本领域普通技术人员（PHOSITA）能够复现该发明。因此，我们的框架通过为专利说明书的结构完整性和法律完整性建立首个专门的验证机制，成为现有以权利要求为中心的评估方法的关键且不冗余的补充。

此外，为验证这些复杂的评估标准而开发的 Pap2Pat-EvalGold 数据集由 146 对样本组成，与大规模通用大语言模型数据集相比，规模相对有限。然而，这是根据专利领域的专业性质，在定性忠实度与定量规模之间做出的战略性取舍。由于评估充分公开等复杂法律要求需要具备领域专门知识的认证专利专业人员进行严格判断，因此无法外包给通用众包。尽管样本量有限，但 0.81 的高组内相关系数（ICC）凸显了我们数据集的稳健性，使其成为自动化专利评估模型进行基准测试的可靠金标准。未来的研究将致力于将 Pat-DEVAL 与权利要求评估模型相结合，并通过人在回路流程扩大领域覆盖范围，从而开发一个全面且端到端的专利评估系统。

参考文献

Satanjeev Banerjee 和 Alon Lavie。2005年。Meteor：一种用于机器翻译评估的自动指标，与人工判断的相关性更高。载于*ACL 机器翻译内在与外在评估指标研讨会论文集*，第 65–72 页。Jinlan Fu、See-Kiong Ng、Zhengbao Jiang 和 Pengfei Liu。2023年。GPTScore：按需评估。载于*2023*年自然语言处理实证方法会议论文集，第 6593–6623 页。Lekang Jiang 和 Stephan Goetz。2024年。专利中的自然语言处理：综述。*arXiv* 预印本 *arXiv:2403.04105*。Seungone Kim、Jamin Shin、Yejin Cho、Joel Jang、Shayne Longpre、Hwaran Lee、Sangdoo Yun、Seongjin Shin、Sungdong Kim 和 Minjoon Seo。

2024年。Prometheus：赋予语言模型细粒度评估能力。载于国际学习表征会议。

Valentin Knappich、Anna Hätty、Simon Razniewski 和 Annemarie Friedrich。2025年。Pap2Pat：使用专利论文配对对大纲引导的长文本专利生成进行基准测试。载于计算语言学协会：*ACL 2025* 年会论文集，第 9524–9554 页。

Jieh-Sheng Lee 和 Jieh Hsiang。2020年。通过微调 OpenAI GPT-2 生成专利权利要求。世界专利信息，62:101983。

Chin-Yew Lin。2004年。ROUGE：用于摘要自动评估的工具包。载于文本摘要的拓展，第 74–81 页。

Yang Liu 等人。2023年。G-Eval：使用 GPT-4 并提升与人工评估一致性的自然语言生成评估。载于神经信息处理系统进展，第 36 卷。

Fiona Murray 和 Scott Stern。2007年。正式的知识产权是否会阻碍科学知识的自由流动？对反公地假说的实证检验。经济行为与组织杂志，63(4):648–687。此前发表于 HBS 工作论文 05-034（2005年）。

Kishore Papineni、Salim Roukos、Todd Ward 和 Wei- Jing Zhu。2002年。BLEU：一种机器翻译自动评估方法。载于计算语言学协会第*40*届年会论文集，第311–318页。

Mariagrazia Squicciarini、Hélène Dernis 和 Chiara Criscuolo。2013年。衡量专利质量：技术和经济价值指标。经济合作与发展组织科学、技术与工业工作论文，（2013/03）。

Amy JC Trappey、Charles V Trappey、Usha H Govin- darajan、AC Chuang 和 John J Sun。2017年。物联网基本标准和专利布局综述：工业4.0的关键推动因素。《先进工程信息学》，33：208–229。

美国众议院。2024年。《美国法典》第35编第112条 - 说明书。《美国法典》。可访问 https://www.govinfo.gov/content/pkg/USCODE-2024-title35/pdf/USCODE-2024-title35-partII-chap11-sec112.pdf。

美国专利商标局。2024年。专利数据仪表板：2024财年专利绩效数据。https://www.uspto.gov/dashboard/patents/。访问日期：2025年-12-18。

Qiyao Wang、Shiwen Ni、Huaren Liu 等。2024年。Au-toPatent：一种用于自动专利生成的多智能体框架。*arXiv* 预印本 *arXiv:2412.09796*。

世界知识产权组织。2024年。《2024年世界知识产权指标》：全球专利和外观设计申请量再创新高。技术报告，世界知识产权组织。

Yongmin Yoo、Qiongkai Xu 和 Longbing Cao。2025年。PatentScore：对大语言模型生成的专利权利要求进行多维度评估。*arXiv* 预印本 *arXiv:2505.19345*。

Manzil Zaheer、Guru Guruganesh、Kumar Avinava Dubey、Joshua Ainslie、Chris Alberti、Santiago On-tanon、Philip Pham、Anirudh Ravula、Qifan Wang、Li Yang 等。2020年。Big Bird：面向更长序列的 Transformer。载于神经信息处理系统进展，第33卷，第17283–17297页。

Jingqing Zhang、Yao Zhao、Mohammad Saleh 和 Pe-ter Liu。2020a。PEGASUS：通过提取间隙句进行预训练，以实现抽象式摘要。载于国际机器学习会议，第11328–11339页。PMLR。

Tianyi Zhang、Varsha Kishore、Felix Wu、Kilian Q Weinberger 和 Yoav Artzi。2020b。BERTScore：使用 BERT 评估文本生成。载于国际学习表征会议（*ICLR*）。

Wei Zhao、Maxime Peyrard、Fei Liu、Yang Gao、Christian M Meyer 和 Steffen Eger。2019。MoverScore：利用上下文化嵌入和推土机距离评估文本生成。载于*2019* 年自然语言处理实证方法会议论文集，第 563–578 页。

A Pat-DEVAL 提示词详情

A.1 CoLT 的系统提示词设计

以下系统提示词用于模拟本领域普通技术人员评估器。为确保评估以法律标准为依据，我们要求在给出量化评分之前，必须完成结构化推理过程。

A.2 评估评分标准

为保持评估者间一致性，指示大语言模型遵循下文定义的 5 点李克特量表评分标准。这些标准适用于所有四个维度（TCF、描述充分性、法定合规性、LPC）：

评分 5（专业）该描述与源技术 R 完全一致。它展现了无懈可击的法律结构、专业术语、精准的数据复现，以及对所有**已公开实施例**的充分公开。

评分 4（高质量）大多数核心机制和实验值均得到准确呈现，仅有少量非关键内容遗漏。它满足充分公开要求，并且在所有必备章节中保持了高度的形式一致性。

Pat-DEVAL 评估器的完整系统提示词

[角色定义]
您是一位资深专利审查员，在相关技术领域拥有专业知识。请根据学术论文中提供的源技术（R），评估生成的专利说明书（D_{gen}）。您必须以专利法理所界定的**本领域普通技术人员（PHOSITA）**的身份行事。

[Req要求：法律思维链 g维 ght (CoLT)]
在给出评分之前，必须明确记录你在以下三个层面的推理：

1. 技术映射层：对 R 和 D_{gen}进行逐项比较。识别所有核心新颖机制、实验数据和技术实施例。指出任何事实性幻觉、失真或遗漏。

2. 法定合规层：评估 D_{gen} 是否满足**充分公开要求（《美国法典》第35编第112条(a)款）**。判断本领域普通技术人员能否在无需“过度试验”的情况下，根据该披露复现发明。此外，还要检查**书面描述要求**，以确保发明人实际拥有该发明。

3. 形式一致性层：评估文件的结构完整性（例如背景技术、发明内容、具体实施方式），以及其是否遵循专业的专利撰写惯例和法律术语。

[评分和输出格式]之后，n分析lysis，提供id

以下项目的评分（1-5）：(1) 技术内容忠实度（TCF）、(2) 数据精确度（DP）、(3) 结构覆盖度（SC）和 (4) 法律专业合规性（LPC）。

输出模板：-推理轨迹： [在此插入步骤 1、2、3 的分析]**-评分**：TCF： [X],DP： [X],SC： [X],LPC： [X]**-最终评分理由**： [对最终评分的简要说明]

表4：thePat-DEVAL框架的完整系统提示词。该提示词旨在通过要求先提供顺序法律推理轨迹（CoLT），再进行定量评估，从而最大限度地减少启发式偏差。

评分 3（可接受）主要发明已予披露，但描述可能缺少具体实施细节、存在轻微数值差异，或术语使用不一致，因此在提交申请前需要稍作修订。

评分 2（平庸）存在明显的技术缺陷或数据歧义。本领域普通技术人员仅凭所提供的文本，很可能需要进行过度试验才能复现该发明；或者关键章节（例如具体实施方式）的内容不够完善。

评分 1（致命）该文件包含严重的事实性幻觉，或未能描述发明的核心内容。由于存在严重的结构性或技术缺陷，该文件不符合《美国法典》第35篇第112条规定的法定要求。