逻辑草图提示（LSP）：一种确定且可解释的提示方法

Satvik Tripathi[1]

[1]De美国宾夕法尼亚州费城，宾夕法尼亚大学放射科

2025年12月24日

摘要

大语言模型（LLM）擅长自然语言推理，但在需要严格遵循规则、具备确定性和可审计性的任务上仍不可靠。逻辑草图提示（LSP）是一种轻量级提示框架，通过引入类型化变量、确定性条件评估器和基于规则的验证器，生成可追溯且可重复的输出。我们采用两项药理学逻辑合规任务，在三种开放权重模型——Gemma 2、Mistral 和 Llama 3——上，将 LSP 与零样本提示、思维链（CoT）和简短提示进行基准比较。

在两项任务和所有模型中，LSP 始终取得最高的准确率（0.83–0.89）和 F1值（0.83–0.89），显著优于零样本（0.24–0.60）、简短提示（0.16–0.30）和思维链（0.56–0.75）。McNemar检验显示，在几乎所有比较中，LSP 都取得了具有统计学意义的提升（$p < 0$。01）。

这些结果表明，LSP 在不牺牲性能的前提下提升了确定性、可解释性和一致性，因此适用于临床、受监管及安全关键型决策支持系统。

GitHub： https://github.com/satviktri/LSP

1 引言

大语言模型已成为广泛应用的工具，可用于临床和非临床领域的分类、信息抽取和通用推理[1，2]。它们能够生成连贯的自然语言输出，因此在决策支持任务中颇具吸引力

而这些任务过去通常需要借助结构化的基于规则的系统。尽管具有这些优势，当前的提示策略在任务要求严格遵循规则、决策路径透明且在重复评估中表现可靠时，仍存在明显局限 [3, 4]。这些要求在医疗保健领域尤为重要，因为模型输出不一致可能会直接影响临床行动和患者结局 [5]。

零样本提示的结构极少，完全依赖模型隐含的推理。这种方法往往会产生不稳定的结果，容易受到措辞、上下文顺序或解码参数的细微变化影响。思维链提示试图引入中间推理步骤，但由此产生的解释无法验证，而且经常存在内部矛盾 [6, 4]。这些回答还依赖随机采样，这限制了它们在受监管环境中的可复现性 [3]。代码风格提示引入了更明确的规则结构，但需要模型解读以自然语言或类似编程语言的语法编写的伪代码 [7]。这种策略可能很脆弱，因为格式或措辞的细微变化往往会导致输出发生显著变化。此外，在同一个提示中结合检测规则与决策逻辑，使系统性调试变得困难。

这些局限促使人们开发能够提供更高结构性和稳定性的提示框架。近期监管指南，包括NIST人工智能风险管理框架、欧盟人工智能法案，以及 FDA、加拿大卫生部和 MHRA 联合制定的良好机器学习实践原则，一致强调透明度、可追溯性和系统行为可预测性的必要性 [8, 9, 10]。当前的提示技术尚未完全满足这些期望，尤其是在模型必须应用确定性逻辑或遵守特定领域安全要求时。

逻辑草图提示（LSP）旨在通过在自然语言输入 [11] 之上引入一种轻量且易于理解的结构，来弥补这些不足。LSP 通过使用显式条件函数、类型化变量以及将变量状态映射为结果的确定性验证器（如图 1 所示），将相关信息的检测与最终决策分离开来。这种设计具有多项优势，包括清晰呈现触发了哪些规则、重复运行时输出一致，以及将模式检测与逻辑决策明确分离。LSP 仍与现有模型兼容，无需对底层模型的参数或架构进行任何更改。

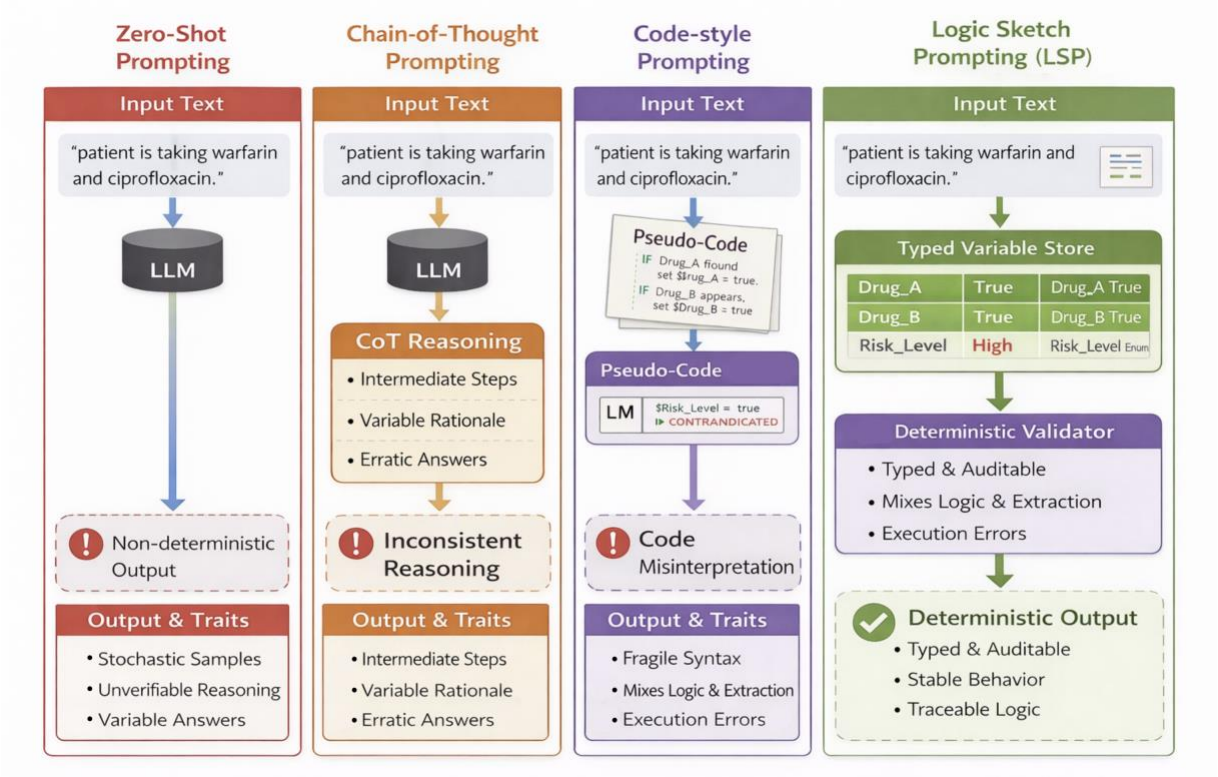

图 1：提示策略对比。零样本提示完全依赖模型的隐式推断，因此输出具有非确定性且无法验证。思维链提示引入了中间推理步骤，但这些步骤仍具有随机性、不一致性，且难以审查。代码风格提示试图通过伪代码施加结构，但往往混淆信息提取与决策逻辑，容易产生误解和执行错误。相比之下，逻辑草图提示（LSP）通过确定性条件评估器填充类型化变量存储，并使用基于规则的验证器生成稳定、可审查且可追溯的输出，从而将学习到的模式检测与符号决策分开。

本研究使用三种开放权重大语言模型 Gemma、Mistral 和 Llama，在两项药理学逻辑合规任务上评估 LSP。我们将 LSP 与零样本提示、简短自然语言提示和思维链提示进行比较。评估指标包括准确率、精确率、召回率、F1分数，并使用 McNemar检验进行显著性检验。在所有模型和任务中，LSP 均表现出持续优异的性能和更高的稳定性。这些发现支持在需要确定性规则评估的应用中采用最简化的结构化提示层，以提升可靠性和可解释性 [12]。

2 方法

2.1 逻辑草图提示框架

逻辑草图提示（LSP）将基于提示的推理形式化为一系列应用于自然语言文本的确定性逻辑变换。设输入文本表示为一个词元序列

$$T = (t_1, t_2, \ldots, t_n).$$

LSP 在 T 上评估一组条件函数，更新类型化变量存储，并应用确定性验证器，将变量状态映射为最终决策。该过程产生透明且可复现的推理轨迹。LSP 的核心组成部分汇总于表 1。

2.1.1 类型化变量存储

变量存储被定义为一个映射

$$V : \{v_1, v_2, \ldots, v_k\} \to \mathrm{D},$$

其中 D 表示所支持的数据类型集合，包括布尔型、整数型、浮点型以及枚举类别值。每个变量 v_i 被初始化为默认值 $v^{(0)}$。在整个评估过程中，变量更新遵循如下规则 $_i$

$$v_i^{(t+1)} = f_i(v_i^{(t)}, c_i(T)),$$

其中 $c_i(T)$ 是与 v_i 相关联的条件函数的输出。

组成部分	描述	在 LSP 中的作用
输入文本	非结构化自然语言输入 例如临床病历、报告或 句子	提供原始语言 信号，用于下游 评估
条件 评估器	应用于输入的确定性检查- 入内容执行的确定性检查，包括正则表达式匹配、关 键词检测、语义相似度， 以及数值比较	检测是否预定义的 逻辑条件 已满足
类型化变量	具有明确数据类型的显式变量 （布尔型、整数型、浮点型、枚举型） 初始化为默认值	存储条件结果 以结构化且 可解释的形式
变量 更新 规则	确定性更新函数将 条件触发器映射到变量 状态	确保一致且 可重复的状态转换
执行轨迹	触发条件和 变量更新	实现透明度、 可审计性和误差 分析
确定性 验证器	基于规则的决策函数（例如， IF–THEN 逻辑、合取范式/析取范式 表达式），作用于最终的 变量状态	生成稳定且可 复现的决策 边界
输出	结构化模型输出，例如带有支 持性证据的标签	提供确定性的预 测，并附带可 解释的推理 轨迹

表 1：逻辑草图提示（LSP）的核心组成部分。LSP 将学习到的模式检测与确定性决策逻辑分离，从而使语言模型行为变得可解释、可审计且可复现。

2.1.2 条件函数

条件函数表示为

$$c_j : T \rightarrow \{0, 1\} \times C_j,$$

其中，第一个组成部分表示条件是否已触发，而C_j 包含可选的捕获值，例如匹配的片段或提取出的数值。

实现了四类条件：

$$\text{Regex}(p), \quad \text{Keywords}(A), \quad \text{NumericCompare}(x, \theta), \quad \text{SemanticSim}(q, \tau),$$

其中p 是编译模式，A 是关键词集合，θ 和τ 表示数值阈值。触发条件后会产生一项更新

$$v_j \leftarrow 1,$$

并将该更新追加到执行轨迹 E 中。

2.1.3 确定性验证器

所有条件评估完毕后，变量存储的最终状态

$$V^* = (v_1^*, v_2^*, \ldots, v_k^*)$$

会传递给确定性验证器

$$g : V \rightarrow Y,$$

其中Y 是结果标签的集合。使用了两种验证器表述：

IF–THEN 链。

$$\text{If } \phi_1(V^*) \text{ then } y_1, \quad \text{else if } \phi_2(V^*) \text{ then } y_2, \ldots$$

逻辑范式。

$$g(V^*) = y_c \quad \text{if} \quad \psi_c(V^*) = 1,$$

其中每个 ψ_c 都以合取范式或析取范式表示。验证器没有副作用，也不包含任何随机组件。图 2 示意性地概述了 LSP 流水线，重点展示了条件评估、类型化状态跟踪和确定性验证之间的分离。

2.2 基线提示策略

使用了三种比较条件。零样本提示采用直接的分类指令，不包含中间结构。简短提示提供了任务和标签空间的简洁自然语言描述。思维链提示要求模型在给出最终决策前生成中间推理步骤。所有基线都完全依赖自然语言推理，不使用外部逻辑或变量存储。

2.3 模型

评估了三种开放权重大语言模型：Gemma、Mistral 和 Llama。所有方法均使用相同的提示处理各个模型。推理采用确定性解码，并固定温度、top-p 和 top-k 参数，以确保结果可复现。

2.4 任务与数据

实验采用 ADE-Corpus-V2 基准语料库，该语料库包含描述药物不良事件（ADEs）的标注句子 [12]。评估了两项监督任务。

任务1：句子级ADE检测。给定输入句子s，目标是估计二元变量

$$Y = \begin{cases} 1, & \text{if the sentence reports a patient-level ADE,} \\ 0, & \text{otherwise.} \end{cases}$$

使用了“分类”划分，其中包含 17，637 个训练实例和 5，879 个测试实例。在注明的情况下，使用了规模为 5，000 的子样本。此任务要求识别药物提及、临床效应，以及隐含的因果或时间关联是否存在。

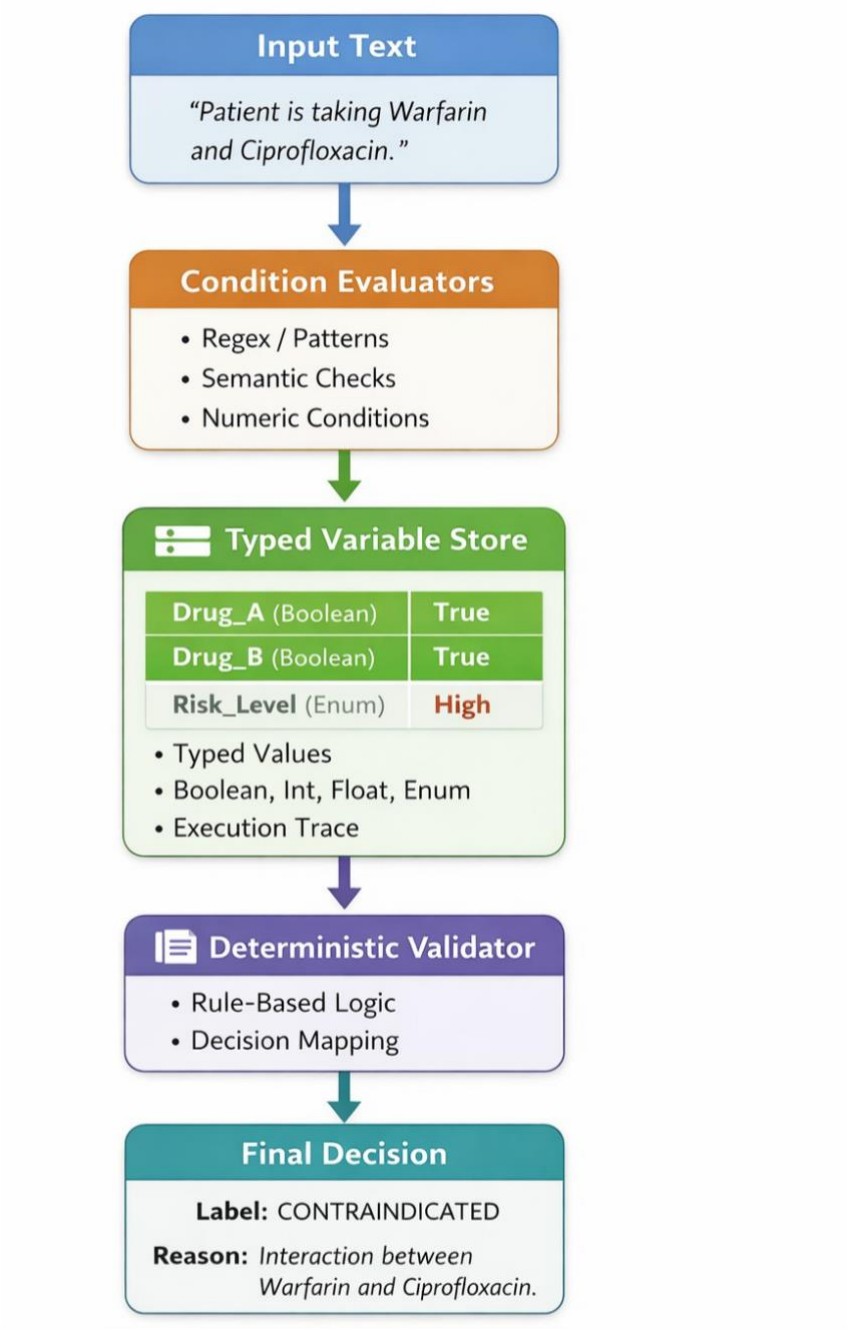

图 2：逻辑草图提示（LSP）框架。自然语言输入经过确定性条件评估器（例如模式匹配、语义检查和数值约束）处理，并据此填充类型化变量存储。随后，变量状态和执行轨迹被传递给基于规则的确定性验证器，由该验证器将最终变量配置映射为稳定的决策，并提供可解释的依据。通过将学习到的模式检测与符号决策逻辑分离，LSP 为安全关键型分类任务提供了可复现、可审计且模型无关的推理。

任务2：药物–效应关系分类。每个输入都由一个三元组（*s, d, e*）组成，其中 *d* 是药物跨度，*e* 是效应跨度。目标是预测

$$y = 1 \quad \text{iff the sentence asserts that drug } d \text{ caused effect } e.$$

正例来自药物 ADE 关系子集。语料库中还纳入自然负例（即同时包含 *d* 和 *e* 、但不存在因果关系的句子），以保持分布的真实性。

2.5 模型

评估了三个经过指令微调的开放权重大语言模型：

$$\text{Gemma}, \qquad \text{Mistral-Instruct}, \qquad \text{Llama 3.2}.$$

所有模型均采用相同的解码参数进行查询。未使用微调、检索或外部上下文注入。

2.6 提示条件

令p 表示一种提示方法，令$\hat{y}^{(m,p)}$ 表示模型m 在提示p下的预测标签。评估了三种提示策略。

LSP提示（我们的方法）。模型会收到一份简洁的规则结构化清单，其中包含五个二元问题。最终决策由以下符号规则定义：

$\hat{y} = 1 \Leftrightarrow d \in s \wedge e \in s \wedge$患者范围($s$)$\wedge$ 显式关联(s)$\vee$时间关联(s) $\wedge\neg$否定(s). 提示要求采用以下形式的严格JSON输出

$$\{\text{"label"} : \hat{y}, \ \text{"evidence"} : \ldots\}.$$

其中恰好包含一个正例和一个负例。

零样本控制。 此基线刻意采取保守策略。只有出现明确的因果触发短语（例如“由……引起”“由于”“继发于”“由……诱发”）时，才允许预测为 1。没有明确因果关系的时间关系（“开始后”“服用期间”）不算作因果证据。如有不确定性，指令默认采用$\hat{y} = 0$。这种做法以牺牲召回率为代价，提高精确率。

简短基线。 一条最小化指令，指导模型检查药物提及、效果提及、关联以及是否不存在否定。与LSP提示一样，需要严格JSON输出。不提供示例。

2.7 评估指标

对于大小为N 的测试集，令$\hat{y}_i$ 表示预测结果，令y_i 表示相应的真实标签。计算准确率、精确率、召回率和F1值，以评估性能。使用包含1,000次重采样的95%自助法置信区间汇总不确定性。使用McNemar检验分析配对系统比较。令n_{01} 表示方法A错误而方法B正确的实例数，令n_{10} 表示相反情况的实例数。统计量定义为

$$\chi^2 = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}}.$$

此测试用于评估分歧模式是否对称，从而支持或否定错误率相等的假设。

2.8 可复现性详情

所有实验均采用确定性解码配置进行。温度设为零，并禁用 top-p 和 top-k 采样，以消除词元生成中的随机变化。所有提示方法均使用完全相同的提示和固定的解码参数查询各模型。实验未使用微调、检索增强、外部工具或额外的上下文注入。基于模式匹配、关键词检测和数值比较的条件评估器均以确定性方式实现。语义相似度检查使用与分类相同的模型实例进行，以确保每个实验条件内的一致性。所有报告的指标均在固定的测试划分上计算，配对统计比较则基于相同的预测集进行。

3 结果

对 ADE-Corpus-V2 两项任务的评估表明，与零样本和简短提示基线相比，LSP提示在所有三种模型上均持续表现更优。任务1和任务2的结果呈现出稳定的模式，说明观察到的改进源于提示的结构，而非特定模型的行为。下文将详细介绍各项任务的实证发现，重点关注准确率、F1表现，以及通过配对麦克尼马尔分析检验的统计显著性。

3.1 任务1：句子级ADE检测

任务1旨在判断句子中是否包含患者级药物不良事件。在 Gemma、Mistral 和 Llama 上，LSP提示均取得了最佳表现。对于 Gemma，LSP提示的准确率达到 0.85，相应的 F1分数为 0.85；而零样本控制的准确率为 0.58，F1值为 0.45。简短基线的准确率与零样本控制相近，但召回率极低，F1分数仅为 0.17。这些结果表明，尽管简短提示包含了必要的语义标准，但它无法持续检测药物—效应关联，尤其是在关联以间接方式表达或置于患者病史背景下描述时。

对于 Mistral，LSP提示的优势依然明显。LSP提示的准确率达到 0.83

，F1值为 0.83。零样本控制的准确率为 0.54，F1值为 0.48；简短基线的准确率达到 0.65，但 F1值仅为 0.15。Mistral表现出与 Gemma相同的措辞敏感性：当通过“开始服药后”或“服药期间”等时间线索来描述不良药物事件时，除非存在明确的因果触发因素，否则零样本控制会系统性地默认判定为否定。相比之下，LSP提示凭借决策规则中的明确条件，能够可靠地捕捉时间关系。

对于 Llama，结果呈现相同趋势。LSP提示的准确率为 0.84，F1值为 0.84，优于零样本配置（准确率为 0.60，F1值为 0.49）。简短基线再次表现出尚可的准确率（0.62），但 F1值非常低（0.16），这主要是因为漏掉了因果陈述。三个模型系列的结果一致，表明 LSP提示的性能提升源于规则结构化提示的优势，而非特定模型的优势。

统计分析进一步证实了这些差异。对于所有三个模型，配对 McNemar 检验表明，LSP提示与各基线之间的预测分歧分布显著不对称。对于 Gemma，LSP提示与零样本配置、简短基线的比较所得近似 p 值分别为 0.006 和 0.003。对于 Mistral，相应数值分别为 0.0047 和 0.0021；对于 Llama，这些差异对应的数值为 $p \approx 0.0080$ 和 $p \approx 0.0032$。这些结果有力地支持了如下结论：与自然语言提示策略相比，LSP提示能够持续减少分类错误。

3.2 任务2：药物–效应关系分类

任务2评估句子是否明确断言某种特定药物会导致某种特定效应。此任务需要谨慎处理词汇变体、时间表达、提及匹配、否定和替代解释。LSP风格关系提示分别检查是否存在明确的因果短语以及时间性因果措辞，将推断限制在患者级范围内，并确保药物和效应提及都与待评估的文本片段完全对应。这些限制会显著影响性能。

在全部三个模型中，LSP提示再次取得了最佳表现。零样本提示的表现符合其定义，即需要明确的因果标记，例如“由……导致”“诱发”或“继发于”。因此，零样本提示取得了

相对较高的精确率，但召回率明显较低，并且遗漏了许多通过时间框架表达的有效不良药物事件关系。简短提示缺乏这些结构性保障，常常将适应证陈述或药物获益描述误解为不良结局，从而产生假阳性。

因此，在任务1中观察到的经验模式在任务2中依然成立。LSP提示在 Gemma、Mistral 和 Llama 上均取得最高的原始准确率，其次是表现明显逊色的零样本，而简短提示则是表现最弱的基线。McNemar检验比较证实，在多数按模型划分的配对比较中，LSP提示的表现显著优于两种基线，这进一步表明，其优势源于提示的规则化架构，而非模型特有的行为。LSP提示对时间关联的显式处理是性能提升的重要来源，尤其是在 ADE-Corpus-V2 中，许多因果陈述是通过时间顺序而非明确的因果连接词来表达的。

3.3 各任务综合结果

对两项任务和全部三个模型的结果取平均后，得出的排名始终一致，并适用于所有实验条件：

LSP提示 > 零样本 > 简短(+思维链)

无论是在句子级不良药物事件识别，还是在结构化药物—效应关系推理任务中评估模型，这一排序都保持稳定。与基线相比，LSP提示的综合性能提升在F1分数上尤为显著，体现出精确率与召回率之间更均衡的权衡。简短基线在两项任务中的召回率始终不理想，而零样本基线主要在任务2、部分在任务1中召回率不佳。

与基线相比，LSP提示在不同模型上的表现方差也更低。零样本提示和简短提示对措辞、语境歧义以及因果关系的表达方式都高度敏感。相比之下，LSP提示将推理路径限定为一系列固定顺序的检查，因此在不同模型系列中的表现更加一致。

模型	提示方法	平均准确率	平均 F1
Gemma	零样本	0.58	0.47
	简短	0.52	0.22
	思维链	0.68	0.64
	LSP（我们的方法）	**0.86**	**0.86**
Mistral	零样本	0.55	0.50
	简短	0.61	0.21
	思维链	0.70	0.65
	LSP（我们的方法）	**0.83**	**0.83**
Llama	零样本	0.60	0.51
	简短	0.63	0.23
	思维链	0.74	0.73
	LSP（本文方法）	**0.84**	**0.84**

表2：ADE-Corpus-V2两个任务的综合表现（任务1：句子级ADE检测；任务2：药物-效应关系分类）。报告的数值表示两个任务的平均准确率和平均 F1分数。逻辑草图提示（LSP）在所有受评估模型中始终取得最强的总体表现。

3.4 定性错误模式

剩余错误的性质与各类提示的结构设计相符。对于LSP提示，最常见的假阳性出现在这样的句子中：句子同时提及药物和效应，却描述的是预期的治疗结局，而非不良反应。这些假阳性之所以出现，是因为尽管上下文线索表明其表达的是治疗意图，词汇模式有时仍与不良药物事件陈述相似。

对于零样本提示，假阴性主要集中在以时间顺序表达因果关系的句子中，例如“患者开始服用药物后出现皮疹”或“开始用药后不久出现症状”。由于零样本提示不允许将时间关联作为因果关系的证据，模型在这类情况下会系统性地预测为负类。简短提示则因缺乏结构约束而同时出现假阳性和假阴性；由于没有明确规定作用范围、否定或证据类型，模型经常会将提及共现误解为因果关系。

3.5 研究结果总结

总体结果表明，在 ADE-Corpus-V2 的两项任务以及所有测试的模型系列中，简洁且受规则约束的提示结构显著提升了准确率、F1值、稳定性和可解释性。这些改进具有统计显著性且表现一致，说明 LSP提示方法为基于自然语言临床文本中的药物安全陈述进行推理提供了稳健且可泛化的框架。

4 讨论

在 ADE-Corpus-V2 的两项任务和全部三个模型系列中，LSP提示始终优于自然语言提示方法。这种提升不仅体现在准确率和 F1值上，也体现在 Gemma、Mistral 和 Llama [13, 14] 的预测稳定性上。这种一致性表明，性能提升源于提示结构本身，而非特定模型的训练方式或架构。

一项关键发现是，LSP提示缩小了较小模型与较大模型之间的性能差距。在零样本和简短提示条件下，较小模型难以推断隐含的因果规则、理清时间关联或正确应用否定，因此预测波动较大且经常出现分类错误 [15, 13]。LSP提示通过将决策过程外显化，缓解了这些局限。它不要求模型在内部重建多步骤推理链，而是通过一系列固定的二元检查，引导模型判断药物检测、效应检测、患者范围、关联类型和否定。这种结构减轻了潜在推理的负担，使较小模型的表现更接近较大的指令微调模型。

更大型的模型也能从这种结构中获益。即使指令微调提升了零样本推理能力，不受约束的自然语言提示仍会引入歧义，尤其是在句子存在多种合理解读时 [14]。LSP提示通过定义确定性的决策边界，优先考虑临床相关线索而非措辞风格，从而缩小这种歧义。因此，在不牺牲精确率的情况下，召回率显著提高。

统计分析进一步支持了这一解读。麦克尼马尔检验显示，在所有模型配对中，意见分歧均明显倾向于 LSP提示 [16]。这些差异的

幅度在较小模型和较大模型中相近，这表明 LSP提示 能够以不受规模影响的方式提升不良药物事件推理的可靠性。

该方法也很好地契合临床文本的语言特性。许多真实的不良药物事件陈述依赖时间关系，而非明确的因果性措辞 [15]。自然语言提示，尤其是零样本提示，往往会忽略这些模式。LSP提示将时间关系编码为有效证据，使各种规模的模型都能捕捉这一重要的信息渠道。LSP下剩余的错误往往源于治疗适应证陈述，它们在表面形式上可能与不良药物事件描述相似，因此可能需要额外的上下文特征来加以区分[17, 14]。

LSP确实能在变量更新和最终决策验证层面确保确定性，但无法使整个推理流程完全独立于底层语言模型。具体而言，某些条件评估器（例如语义相似度检查）依赖模型内部的表征，而这些表征可能因架构或实现方式不同而有所差异。不过，这种对模型的依赖局限于检测阶段；一旦变量状态确定，就不会影响基于规则的验证器的确定性执行。因此，虽然无法保证端到端的完全确定性，但对于一组固定的条件输出，LSP可确保决策逻辑、规则应用和结果映射保持稳定、可追溯且可复现。

总体而言，这些发现表明，对推理路径进行结构化，是一种有效且轻量的方式，能够稳定语言模型在生物医学信息提取任务中的行为 [13]。通过明确规则来约束决策，并将模式识别交由模型完成，LSP 提示创建了一种兼具可解释性和稳健性的混合推理环境，且适用于不同规模的模型。

5 局限性与适用范围

本研究存在一些局限性，这些局限性界定了 LSP 的适用范围。首先，LSP 依赖于预定义条件评估器的覆盖范围和质量，包括模式匹配、关键词检测和语义检查。在这一阶段出现的错误或遗漏可能导致变量状态不正确，并以确定性的方式传递至最终决策。因此，LSP 并不能消除检测错误，而是限制这类错误对下游推理的影响。

其次，构建 LSP 规则需要领域知识，以明确相关变量、条件和决策边界。虽然这种显式结构提升了透明度和可审计性，但也增加了人工设计工作，并可能限制其迁移到新任务的能力，除非投入额外的规则工程。因此，该方法更适用于任务要求和逻辑标准能够清晰表述的场景。

第三，尽管 LSP 在变量更新和最终验证中强制执行确定性，但某些条件评估器仍可能依赖基于模型的组成部分，例如语义相似度判断。在这些情况下，随机性被限制在局部检测步骤中，不会影响决策逻辑的确定性执行，但无法保证完整的端到端确定性。

最后，本研究中的实验聚焦于涉及不良药物事件推理的规则约束型生物医学分类任务。尚未评估 LSP 在开放式生成、创造性推理或缺乏明确定义逻辑结构的功能中的有效性。因此，对于优先重视可复现性、可解释性和稳定规则遵循，而非不受约束的语言生成的决策支持场景，LSP 最为适用。

6 结论

本研究表明，紧凑的规则结构化提示策略能够显著提升开放权重语言模型的不良药物事件提取效果。在两项任务和三个模型系列中，LSP提示的表现始终优于零样本提示和简短提示，准确率更高、F1分数更出色，决策边界也更稳定。这些提升不受模型规模影响。较小的模型受益于对隐式多步推理依赖的降低，较大的模型则受益于更清晰、歧义更少的决策标准。该方法无需微调、外部工具或额外训练数据，却能显著改善错误特征，并更有效地捕捉显式线索和时间因果线索。总体而言，这些发现表明，轻量级符号支架能够引导语言模型进行更可靠、可解释的临床推理，为生物医学文本分析提供一种实用且可扩展的方法，适用于确定性、透明度和跨模型稳定性至关重要的场景。

参考文献

[1] Tom B. Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell、Sandhini Agarwal、Ariel Herbert-Voss、Gretchen Krueger、Tom Henighan、Rewon Child、Aditya Ramesh、Daniel M. Ziegler、Jeffrey Wu、Clemens Winter、Christopher Hesse、Mark Chen、Eric Sigler、Mateusz Litwin、Scott Gray、Benjamin Chess、Jack Clark、Christopher Berner、Sam McCandlish、Alec Radford、Ilya Sutskever 和 Dario Amodei。语言模型是少样本学习者。载于神经信息处理系统进展第 *33* 卷：*NeurIPS 2020*，神经信息处理系统年度会议 *2020*，第 1877–1901 页。Curran Associates 公司，2020 年。另可通过 arXiv:2005.14165 获取，数字对象标识符为 10.48550/arXiv.2005.14165。

[2] Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Brian Ichter、Fei Xia、Ed H. Chi、Quoc V. Le 和 Denny Zhou。思维链提示能够激发大语言模型的推理能力。载于神经信息处理系统进展第 *35* 卷：*NeurIPS 2022，2022*年神经信息处理系统年度会议，第 5482–5496 页。Curran Associates 公司，2022年。另见 arXiv:2201.11903，数字对象标识符 10.48550/arXiv.2201.11903。

[3] Antonia Creswell 和 Murray Shanahan。使用大语言模型进行忠实推理。*arXiv* 预印本，2022年。预印本；数字对象标识符 10.48550/arXiv.2208.14271。

[4] Miles Turpin、Julian Michael、Ethan Perez 和 Samuel Bowman。语言模型并非总会说出自己的想法：思维链提示中的不忠实解释。载于神经信息处理系统进展第 *36* 卷：*NeurIPS 2023，2023*年神经信息处理系统年度会议。Curran Associates 公司，2023年。另可见 arXiv:2305.04388，数字对象标识符为 10.48550/arXiv.2305.04388。

[5] Andrew L. Beam 和 Isaac S. Kohane. 医疗保健中的大数据和机器学习。*JAMA*，319(13):1317–1318，2018。

[6] Alon Jacovi 和 Yoav Goldberg。迈向忠实且可解释的自然语言处理系统：我们应如何定义和评估忠实性？载于计算语言学协会第*58*届年会论文集，第4198–4205页，2020年。

[7] Tushar Khot、Daniel Khashabi、Kyle Richardson、Peter Clark 和 Ashish Sab harwal。文本模块化网络：用现有模型的语言学习任务分解。载于*2021*年计算语言学协会北美分会会议论文集：人类语言技术，第1264–1279页，2021年。

[8] 美国国家标准与技术研究院。《人工智能风险管理框架》（AI RMF 1.0）。NIST 出版物，2023年1月。

[9] 美国食品药品监督管理局、加拿大卫生部和英国药品和医疗保健产品管理局。《医疗器械开发中的良好机器学习实践：指导原则》。监管机构于2021年10月发布的指导文件。该联合指导文件概述了医疗器械开发中良好机器学习实践的原则。

[10] 欧盟。《欧洲议会和欧盟理事会关于制定人工智能统一规则的《欧盟第2024/1689号条例》（《人工智能法案》）。《欧盟官方公报》，2024年6月。于2024年6月13日通过。

[11] Tarek R. Besold、Artur d’Avila Garcez、Sebastian Bader、Howard Bowman、Pedro Domingos、Pascal Hitzler、Kai-Uwe Kohnberger、Luis C. Lamb、Daniel Lowd、Priscila Machado Vieira Lima、Leo de Penning、Gadi Pinkas、Hoifung Poon 和 Gerson Zaverucha。《神经符号学习与推理：综述与阐释》。*arXiv* 预印本，2017年。预印本；数字对象标识符 10.48550/arXiv.1711.03902。

[12] Harsha Gurulingappa、Abdul Mateen Rajput、Angus Roberts、Juliane Fluck、Martin Hofmann- Apitius 和 Luca Toldo。开发基准语料库，以支持从医学病例报告中自动提取药物相关不良反应。《生物医学信息学杂志》，45(5):885–892，2012年。

[13] Sina Shool、Sara Adimi、Reza Saboori Amleshi、Ehsan Bitaraf、Reza Golpira 和 Mahmood Tara。临床医学领域大语言模型（llm）评估的系统综述。《*BMC*医学信息学与决策制定》，25:117，2025年3月。

[14] Sonish Sivarajkumar、Mark Kelley、Alyssa SamolykMazzanti、Shyam Visweswaran 和 Yan- shan Wang。大语言模型在

零样本临床自然语言处理中的提示策略实证评估：算法开发与验证研究。*JMIR* 医学信息学，12:e55318，2024年4月。

[15] Yohan Bonescki Gumiel、Lucas Emanuel Silva e Oliveira、Vincent Claveau、Natalia Grabar、Emerson Cabrera Paraiso、Claudia Moro 和 Deborah Ribeiro Carvalho。《临床文本中的时间关系抽取：系统综述》。*ACM* 计算调查，54(7):1–36，2022年。在线发表于2021年9月；最终版本发表于2022年。

[16] Thomas G. Dietterich。用于比较监督分类学习算法的近似统计检验。神经计算，10(7):1895–1923，1998年9月。

[17] Salisu Modi、Khairul Azhar Kasmiran、Nurfadhlina Mohd Sharef 和 Mohd Yunus Sharum。从临床病历中提取药物不良事件：所用方法的系统综述。《生物医学信息学杂志》，151：104603，2024年3月。

A 附录

A.1 术语和缩略语

表 3 总结了本文中使用的主要缩写和专业术语。

定义的表述旨在反映药物不良事件提取和基于大语言模型的临床推理中的实际用法。

Term	定义
ADE	药物不良事件；由药物引起的不良临床效应
药物–效应关系 tion	。 一种结构化断言，表明药物 *d* 为 一名患者。
LSP	逻辑草图提示；一种规则结构化提示框架 使用类型化变量并进行确定性验证。
显式关联	因果短语，例如“由.....引起”、“由.....诱发”或“继发于”。 继发于”。
时间关联	暗示因果关系的时间描述，例如“在.....之后”。 治疗开始后”或“开始治疗后不久”。
否定	任何能够反转或取消因果关系的语言线索 （例如，“并未导致”、“没有证据表明”）。
患者范围	限定药物不良事件与所述患者相关 句子，而非药物类别或一般知识。
零样本 提示	使用最少指令且不提供特定示例的分类 示例。
CoT	思维链提示；一种会进行推理的提示风格 要求模型生成明确的中间推理
简短提示	步骤。 不含明确逻辑脚手架的最简结构化指令—— 搭建。
验证器	从变量存储 V 到 最终标签的确 定性映射。

表 3：缩略语词汇表和特定任务术语。

A.2 提示技术示例

为说明不同提示系列之间的差异，以下示例展示如何分别使用零样本、简短、思维链和 LSP提示风格来指示同一项 ADE 分类查询。

零样本。

判断该句是否报告药物导致了该效果 t. 输出

1 或 0。 句子：“患者在开始服用阿莫西林后出现了皮疹。”

简短。

如果药物存在、效应存在、该效应由药物引起，且该陈述未被否定，则标记为 1。输出 JSON：{"label": ...}。句子：“...”

思维链。

请逐步推理。这句话是否提及药物？是否描述了效应？是否将两者联系起来？最后，仅根据最终推理输出带有标签的 JSON。句子：“...”

LSP提示（我们的方法）。

请默默回答以下二元问题。1. 是否提及了药物？2. 是否提及了效应？3. 该陈述是否与患者有关？4. 是否存在明确的因果联系？5. 如果没有，是否存在时间关联？当且仅当满足以下条件时，将标签设为 1：（1）&（2）&（3）&（（4）或（5））&不存在否定。返回 JSON：{"label": ..., "evidence": ...}。句子：“...”

A.3 C. 如何编写 LSP

LSP 可视为一种结构化程序，由一系列类型化变量、条件函数和确定性验证器定义。通用的 LSP规范可写成以下形式。

步骤 1：定义变量。令变量存储为

$$V = \{ v_1, v_2, \ldots, v_k \}.$$

每个变量都有类型、默认状态和允许的更新。

步骤 2：定义 个条件评估器。 对于每个变量v_i，指定一个条件函数 n

$$c_i(T) \rightarrow \{0, 1\} \times C_i,$$

其中 第一个输出表示触发状态，第二个输出可选地返回捕获的文本片段 s.

步骤 3：指定规则更新。规则应用更新

$$v_i \leftarrow 1 \quad \text{if } c_i(T) = 1,$$

并且每次更新都会记录在执行轨迹中。

步骤 4：定义验证器。最终决策规则定义为

$$g(V^*) = \begin{cases} 1, & \text{if } \phi(V^*) \text{ is true,} \\ 0, & \text{otherwise,} \end{cases}$$

其中，ϕ 可以写成有序的 IF–THEN 子句集合，也可以写成合取范式/析取范式形式。

示例。 用于药物-效应关系分类的 LSP 可以写成：

$$\begin{aligned} v_{\text{drug}} &= 1 \quad \text{iff a drug name is detected,} \\ v_{\text{effect}} &= 1 \quad \text{iff an effect keyword is detected,} \\ v_{\text{scope}} &= 1 \quad \text{iff patient-level reference is present,} \\ v_{\text{explicit}} &= 1 \quad \text{iff an explicit causal phrase is matched,} \\ v_{\text{temporal}} &= 1 \quad \text{iff a temporal causal construction is matched,} \\ v_{\text{negated}} &= 1 \quad \text{iff negation cues are detected.} \end{aligned}$$

验证器应用：

$$\hat{y} = 1 \quad \text{iff} \quad v_{\text{drug}} \wedge v_{\text{effect}} \wedge v_{\text{scope}} \wedge (v_{\text{explicit}} \vee v_{\text{temporal}}) \wedge \neg v_{\text{negated}}.$$

A.4 D 。将任意提示转换为 LSP 格式的提示 t

以下元提示可用于将任何自然语言提示自动转换为 LSP 风格的规则结构化提示。该元提示指示模型识别分类标准、提取相关变量、形式化条件，并构建一个确定性的验证器。

1. 提取决策所需的所有二元条件。2. 将每个条件改写为一个静默式是非问题。3. 为每个条件分配一个变量名。4. 构建一条决策规则，使用 AND、OR 和 NOT 将这些变量组合起来。5. 要求严格JSON输出：{”label”: ..., ”evidence”: ...}。6. 如果可能，恰好包含一个正例和一个负例。

返回最终的 LSP 格式提示。输入提示：$_{P}$ *ROMPT* > > > *Y ouareanLSP* − *compiler.Givenanypromptthatrequiresclassificationordecision−making, convertitintoanLSP −stylepromptasfollows* ：

1. 提取决策所需的所有二元条件。2. 将每个条件改写为一个静默式是非问题。3. 为每个条件分配一个变量名。4. 构造一条决策规则，使用 AND、OR 和 NOT 组合这些变量。5. 要求严格JSON输出：{”label”: ..., ”evidence”: ...}。6. 如有可能，恰好包含一个正例和一个负例。

Return 最终的 LSP 格式提示。输入提示：¡¡¡USER$_{P}$ *ROMP T* >>>

此元提示提供了一个模板，可将高层指令转换为确定性逻辑程序，适用于生物医学信息抽取及其他结构化推理场景。