逻辑参数化神经符号自然语言推理：通过控制逻辑形式体系实现可验证的大语言模型推理

Ali Farjami[1], **Luca Redondi**[1,2],**MarcoValentino**[31]卢森堡大学 [2]波鸿鲁尔大学 [3]谢菲尔德大学 ali.farjami@uni.lu luca.redondi@ruhr-uni-bochum.de m.valentino@sheffield.ac.uk

arXiv:2601.05705v1 [cs.AI] 9 Jan 2026

摘要

大语言模型和定理证明器可以有效结合，用于可验证的自然语言推理。然而，现有方法依赖固定的逻辑形式体系，这一特点限制了其稳健性和适应性。我们提出一种用于神经符号自然语言推理的逻辑参数化框架，将底层逻辑视为可控组件，而非静态背景。借助 LogiKEy 方法论，我们将一系列经典与非经典形式体系嵌入高阶逻辑，从而系统比较推理质量、解释细化和证明行为。我们重点研究规范推理，因为逻辑的选择会产生重大影响。具体而言，我们比较逻辑外部方法（通过公理编码规范性要求）与逻辑内部方法（规范模式由逻辑内置的结构自然产生）。大量实验表明，逻辑内部策略能够持续提升自然语言推理的表现，并生成更高效的混合证明。此外，我们还表明，逻辑的有效性取决于领域：一阶逻辑更适用于常识推理，而道义逻辑和模态逻辑则在伦理领域表现出色。研究结果凸显了将逻辑作为一等参数化要素融入神经符号架构的价值，从而实现更稳健、模块化且适应性更强的推理。

1 引言

大语言模型（LLM）在自然语言推理（NLI）方面取得了令人瞩目的进展，能够针对各种任务给出合理流畅的解释（Liu 等人，2025；Cheng 等人，2025）。然而，对于法律、伦理或监管情境等要求推理具备逻辑有效性、可泛化性和可信度的场景，现有大语言模型系统往往表现不佳（Li，2023；Hadi 等人，2023；Bender 等人，2021年）。

神经符号架构提供了一种潜在解决方案（Bhuyan 等人，2024；Garcez 和 Lamb，2023），这类架构将大语言模型与外部定理证明器（TP）相结合，以实现形式化验证和细化（Quan 等人，2025b，a，2024；Pan 等人，2023；Olausson 等人，2023）。然而，现有方法的一个关键局限在于对逻辑的处理方式：大多数神经符号系统固定使用单一逻辑，通常是一阶逻辑（FOL），并将其视为静态的背景层，而非可调整的组件。

本文对这一假设提出质疑。我们认为，在神经符号推理中，逻辑本身应被视为一个参数。不同的逻辑系统支持不同的推理模式：模态逻辑可以原生表达义务和许可，条件逻辑能够处理例外和语境转换，而一阶事件逻辑则擅长编码特定实例。通过启用逻辑参数化架构，我们可以系统地探索逻辑结构如何影响大语言模型驱动的推理，以及一种逻辑在何时可能优于另一种逻辑。

为奠定本研究的基础，我们聚焦于规范推理，将其视为一个逻辑选择影响尤为显著的领域（Gabbay 等人，2013年）。在伦理学、法律和政策等情境中，推理涉及义务、许可、禁止、例外和违规行为。这些概念通常难以用传统的一阶逻辑来表达。例如，从“你有义务提交作业”推导出“你被允许提交作业”，虽然直觉上合理，但仅凭一阶逻辑无法推导得出，除非添加外部公理。相比之下，模态逻辑 KD 将这一推论作为内置公理（$O\varphi \to P\varphi$）（Chellas，1980）。由此可见一个核心区别：一类是逻辑外部推理，即在领域理论中将规范规则明确添加为公理（例如在一阶逻辑中）；另一类是逻辑内部推理，即将规则嵌入逻辑本身，作为结构原则（例如

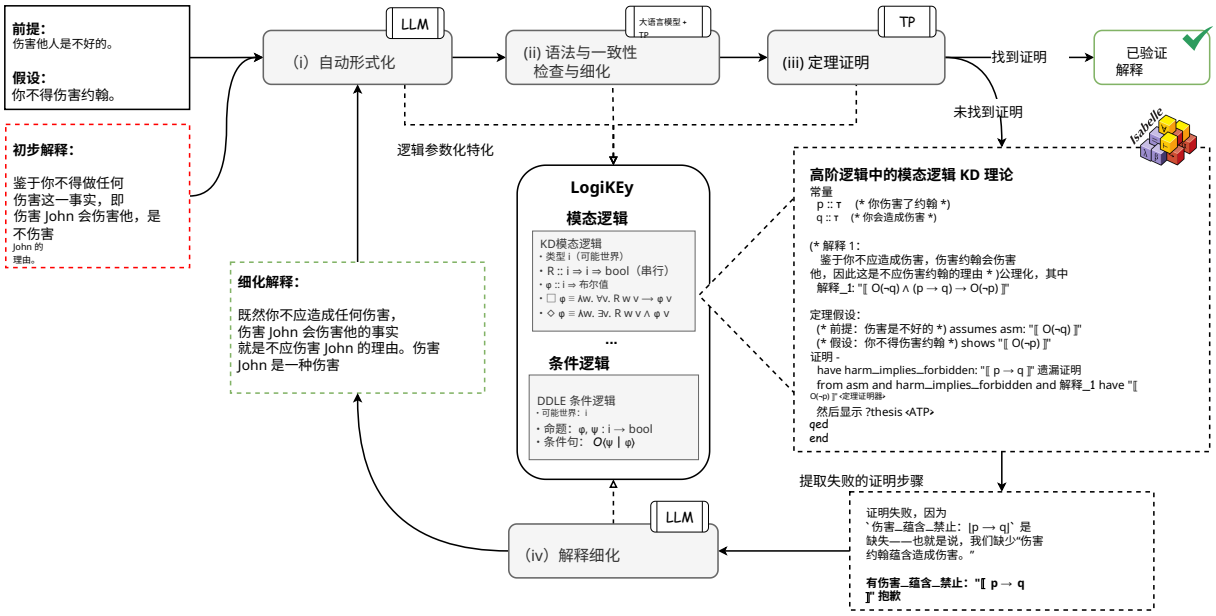

图 1：逻辑参数化神经符号自然语言推理框架与大语言模型的示意图。该框架通过 *LogiKEy* 推广神经符号架构，将经典与非经典逻辑嵌入高阶逻辑（HOL）。这使得我们能够利用多种逻辑形式体系，整合大语言模型与定理证明器（TP），在不同任务和领域中迭代细化解释。

模态公理、条件算子）。

我们基于 *LogiKEy* 方法（Benzmüller 等人，2020b年），该方法支持在高阶逻辑（HOL）中对经典和非经典逻辑进行语义嵌入。这一基础设施使我们能够将多种逻辑整合到混合大语言模型—定理证明器流程中，并比较它们在形式化和验证自然语言推理解释方面的表现。

我们的研究围绕以下问题展开：（*RQ1*）逻辑形式体系的选择如何影响大语言模型驱动的自然语言推理定理证明？（*RQ2*）逻辑内部推理策略与逻辑外部方法相比如何？（*RQ3*）逻辑参数化架构能否提升证明简约性、解释细化和验证效果?

为系统地研究这些问题，我们首先介绍一个用于自然语言推理中的道义解释的新数据集，称为 *B*ioethical *E*xplanations and*N*ormative *R*easoning (*BENR*)[1]，旨在揭示规范解释的内部结构（Goble，2013年）。利用 *BENR*，我们通过扩展并泛化用于自然语言推理解释验证和细化的既有大语言模型驱动的神经符号框架，开展逻辑参数化评估（Quan 等人，2025b，2024）。

我们的结果表明，逻辑内部方法在不同大语言模型（包括 GPT-4o（Bubeck 等人，2023）和 DeepSeek-V1（Bi等人，2024）上，始终在规范推理任务中优于逻辑外部方法。具体而言，采用模态逻辑 KD 可实现最高的解释细化率（最高达 77.67%），以更少的迭代次数收敛，并与一阶逻辑（FOL）相比大幅降低推理成本。相较之下，FOL 的句法鲁棒性更高，但验证通过的解释显著更少，这凸显了表达充分性与稳定性之间的权衡。此外，我们的结果表明，逻辑的有效性取决于领域：FOL 更适用于常识推理，而道义逻辑和模态逻辑更适用于伦理领域。

总而言之，我们的贡献如下：

1. 我们提出了一种逻辑参数化的神经符号框架，将逻辑的选择视为面向自然语言推理的大语言模型驱动定理证明中的显式设计维度。

2. 我们引入了*BENR*，这是一个用于自然语言推理中道义解释的新数据集，旨在揭示规范推理和解释的内部结构。

3. 我们提供了详尽的实证分析，表明逻辑内部推理能够提

[1]代码和数据集将在论文发表后发布。

流程阶段	FOL	模态逻辑	条件逻辑	目的
句法的 **句法分析**	识别主语、动词、宾—ject	识别模态关键词： 必须、可以、应该	识别 条件句 从句：如果—那么、除非	结构 抽取-tion
形式化	施事(e, x) ∧ 动词(e) ∧ 患者(e, y)	○（动词）或 P（动词）	$\bigcirc(\psi/\varphi)$ 或 $P(\psi/\varphi)$	自然语言 → 逻辑映射-ping
证明概要	使用 ∧、→、∀、∃	使用 ○、P、□、◇	使用 $\bigcirc(-/-)$、$P(-/-)$	引导式证明
细化	“缺少关于施事”	“模态公理未满-足：$\bigcirc\varphi \to P\varphi$”	“不可分离的条件规范附着：$\varphi \wedge \bigcirc(\psi/\varphi)$ ”	反馈，纠正 错误

表 1：流水线各阶段的逻辑参数化适配。流水线的每个阶段都针对特定逻辑量身定制，并遵循其语法和语义。

升鲁棒性、证明简约性和解释细化程度，并优于逻辑外部策略。总体而言，我们证明了逻辑的选择会对大语言模型驱动的神经符号系统产生决定性影响，并为实现更有效、适应性更强且模块化程度更高的集成奠定了基础。

2 自然语言推理的逻辑参数化解释验证

给定一个由假设 h、前提 p和解释 e组成的自然语言推理问题，这些内容均以自然语言表述。若满足 $p \cup e \models h$，则将 e 定义为逻辑有效解释。为验证这一点，我们将三元组$\{p, e, h\}$ 映射为一组逻辑公式。 p 和 e 对应的公式构成理论 Θ，而 h 则映射为目标公式 ψ。随后可使用定理证明器判断 $\Theta \models \psi$是否成立，从而验证该解释。

在这项工作中，我们扩展并推广了Quan 等人(2025b提出的神经符号框架，并提出一种模块化、逻辑参数化的流水线，旨在支持多种形式体系，例如模态逻辑和条件逻辑，并在 Isabelle/HOL 中实现。如图1所示，该流水线通过以下四个关键阶段协调大语言模型与定理证明器之间的交互：

(i) 自动形式化大语言模型将自然语言三元组 $\{p, h, E\}$ 转换为一组公式 Φ。此阶段包括句法解析，随后将其转换为目标形式语言，并定义一个可由定理证明器验证的证明概要（ Quan 等人，2024）。与以往仅限于一阶逻辑的工作不同，我们的框架通过目标逻辑 $\mathcal{L}$对这一阶段进行参数化。这会生成形式理论 $\Theta_{\mathcal{L}}$ 和一个表示假设的$\psi_{\mathcal{L}}$ 目标公式。

(ii) 语法与一致性检查形式化之后， $\Theta_{\mathcal{L}}$ 和 $\psi_{\mathcal{L}}$ 会接受自动句法与一致性检查。这可确保大语言模型生成的形式化表达在Isabelle/HOL环境中语法有效，并且前提之间不存在矛盾（即，$p \cup e \not\models \bot$），从而避免空洞蕴涵。

(iii) 定理证明 形式理论由 Isabelle/HOL 交互式定理证明器（Nipkow 等人，2002）进行处理。系统根据所选逻辑模块使用专门的公理化体系。证明器尝试推导$\Theta_{\mathcal{L}} \vdash \psi_{\mathcal{L}}$，并使用 Isabelle 中集成的自动定理证明工具。

(iv) 解释细化 如果证明失败，该框架会提取定理证明器返回的失败证明步骤。这种符号反馈可以识别缺失前提或逻辑缺口（例如，缺失的桥接规则 $p \to q$ 如图 1所示）。随后，系统会将反馈提供给大语言模型，以便按照细化策略生成修订后的解释 e'。这一循环会持续进行 t 步，或直到解释成功通过验证。

2.1 通过 LogiKEy 实现语义嵌入

我们的实现基于*LogiKEy*方法论（Benzmüller 等人，2020b年），该方法支持将各种逻辑系统语义嵌入高阶逻辑（HOL）中。这些嵌入在统一的形式元语言中保留了每种目标逻辑的语义，使它们能够共享定理证明器（例如 Isabelle/HOL）、模型查找器（例如 Nitpick（Blanchette 和 Nipkow，2010年））以及证明助手（例如 Sledgehammer（Blanchette 等人，2013年））等基础设施。LogiKEy 并非将单个公式从一种逻辑转换为另一种逻辑，而是将每种逻辑视为一等模块，各自拥有

公理、运算符和推理约束，并在语义上全部嵌入 HOL 中。这使得研究者能够在统一且可验证的框架内开展逻辑参数化实验。

2.2 支持的逻辑

LogiKEy 框架支持多种与规范推理及其他领域相关的逻辑（Benzmüller 等人，2020a年）：

模态逻辑 KD 能够有效表达义务、许可与禁止之间的逻辑关系，并支持基本的模态推理（vonWright，1951）。在命题逻辑（PL）的语言中加入模态算子 ◯即可得到 **K** 的语言。其生成规则如下：

$$\varphi ::= p|\neg\varphi|\varphi \vee \varphi|\bigcirc \varphi$$

P是 ◯的对偶，即 $P\varphi =_{df} \neg \bigcirc \neg\varphi$。模态逻辑 **KD** 是模态逻辑**K**加入公理**D**后的扩展：$\bigcirc\varphi \rightarrow P\varphi$ 该公理体现了义务蕴含许可这一直觉。

条件逻辑 DDLE 以基于偏好的语义取代 KD 所采用的标准可能世界语义（Åqvist，1984年；Parent，2021年）。在可能世界语义中，我们指定从每个世界可达的可接受世界，并据此定义义务。另一方面，在基于偏好的语义中，所有世界都按从理想世界到（道德上）最差世界的顺序排列。这使得我们能够表达违反义务情形下的义务（Chisholm，1963），即：当某项其他义务被违反时，在次优世界中变得具有约束力的义务。DDLE 的语言是在命题逻辑的语言中加入以下算子得到的：$\Box$（表示必然性）；$\Diamond$（表示可能性）；以及$\bigcirc(-/-)$（表示条件义务）；$P(-/-)$（表示条件许可）。$\bigcirc(\psi/\varphi)$读作“如果φ，那么 ψ 是义务的”，而 $P(\psi/\varphi)$读作“如果 φ，那么 ψ 是允许的。”

带事实脱离的条件逻辑违反义务并不会使义务消失。因此，在责任冲突情境中，“应当”呈现出两种不同的含义：一方面，它指理想情况下应当发生的事；另一方面，在违规已经发生的前提下，它指实际上应当发生的事。Carmo 和 Jones 二元道义逻辑（Carmo 和 Jones，2013年），下文将其称为**DDL_CJ**，能够无歧义地表示这两者。**DDL_CJ**公式集扩展了条件逻辑公式集（如系统**DDLE**中所述），具体如下：

- $\Box\varphi$ —在所有可能世界中
- $\Box_a\varphi$ —在当前世界的所有实际版本中
- $\Box_p\varphi$ —在当前世界的所有潜在版本中
- $\bigcirc_a\varphi$ — 表示实际义务的一元道义算子
- $\bigcirc_p\varphi$ —表示主要义务的一元道义算子

通过将这些逻辑嵌入高阶逻辑中（Benzmüller 等人，2018年、2019年；Farjami，2020年），我们在统一的推理框架中将它们整合起来，同时保留各自不同的推理属性。

3 实证设置

3.1 BENR 数据集

为了测试模型生成有效义务逻辑解释的能力，我们构建了一个名为 *B*ioethical*E*xplanationsand*N*ormative *R*easoning(*BENR*) 的数据集。BENR 旨在探索伦理推理中发挥作用的不同推理模式。与现有替代方案相比，它有一个主要的显著特点。数据集通常追求简洁：案例以高度抽象的方式描述，并置于特定情境中（Hendrycks 等人，2021年；Forbes 等人，2020年）。相比之下，我们关注的不是伦理推理所应用的情境，而是推理本身的结构，[2] 包括不同的组成模式，以及在进行道德评价时这些模式的组合方式。

为实现这一目标，我们聚焦于应用伦理学中一个尤为复杂的子领域：（生物）伦理学。该数据集共包含 103 个示例。其中相当一部分（47 个案例）包含（生物）伦理推理中典型的推理模式，例如从一般原则中实例化表面理由，以及解决这些理由之间的冲突（Goble，2013年）。另外 56 个案例包含的推理模式虽与伦理推理相关，但并非伦理学所独有。这些模式包括认识论默认推理以及关于道义模态的推理。对于第二个子集，我们的数据集范围与其他

[2](Emelin 等人，2021年) 例如，呈现了丰富的情境，但伦理推理却很简单。

数据集有所重叠。因此，我们借鉴了以往的资源，并将现有示例调整为我们的格式。具体而言，我们改编了（Holliday 等人，2024）中的经典逻辑问题和模态相关问题，以及 e-SNLI 中的常识与默认推理问题（Camburu 等人，2018年）。

总体而言，我们的数据集中的案例呈现以下格式：

示例 （自主性需要胜任的选择）。前提。 一名患者拒绝接受一种简单且能挽救生命的治疗。患者因高烧而严重意识混乱。你应该尊重他人的自主性。增进他人的福祉是好的。

假设。 你应该给予这种治疗 *t.*

解释。

1. 鉴于你应该尊重自主性，患者的拒绝通常会构成不予治疗的理由。*2.* 鉴于增进他人的福祉是好的，挽救生命的益处构成治疗的理由。*3.* 但是，患者是在缺乏心智能力的情况下作出拒绝的，因此这并未体现自主决定。*4.* 因此，不予治疗的理由被削弱，而治疗的理由仍然成立。

解释指出了从前提到假设的推理步骤。在本例中，这些步骤包括从一般原则中导出初步理由，以及通过削弱来化解初步理由之间的冲突。

3.2 模型

为支持逻辑参数化定理证明，我们扩展了 Faithful-Refiner（Quan 等人，2025b）。Faithful-Refiner 是一个面向自然语言推理的解释细化框架，最初旨在 Isabelle/HOL 中实现 Neo-Davidsonian一阶逻辑（一阶逻辑）形式化。尽管其提示能够有效引导大语言模型采用基于事件的表示，但它们假定逻辑基底固定。我们的扩展将该框架推广至多种经典与非经典逻辑，这些逻辑通过 LogiKEY 框架嵌入其中。表1总结了不同逻辑形式体系之间的主要差异。

我们评估了两种最先进的大语言模型：*GPT-4o*(Bubeck等人，2023）和*DeepSeek-V1* (Bi等人，2024）。两个模型均采用默认配置，并使用相同的提示，以确保公平地

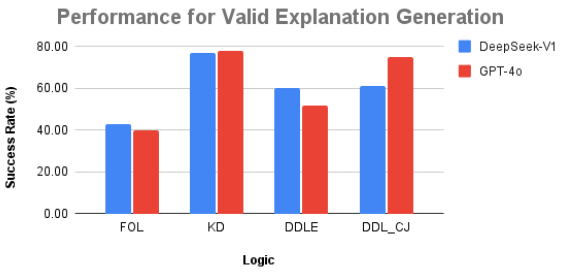

图 2：有效解释生成的成功率。

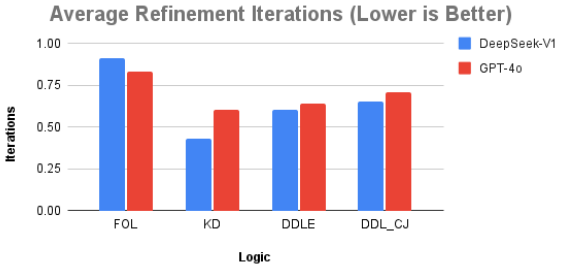

图 3：达到有效解释所需的平均细化迭代次数。

比较它们在不同逻辑框架下的推理能力。我们对模型进行评估，细化迭代次数最多为 $t = 3$次。

3.3 逻辑形式体系

我们在 BENR 的多种推理任务和领域上评估该框架，涵盖经典逻辑、常识推理、默认推理、模态和生物伦理推理；评估涉及四种逻辑形式体系——一阶逻辑、知识蒸馏、*DDLE* 和 *DDL_CJ* 。

4 结果

我们从四个互补维度分析性能：整体解释成功率（图2）、细化效率（图3）、计算效率（图5）、对句法失败的鲁棒性（图6）以及领域特定行为（图4）。

解释成功率 图2 展示了 DeepSeek 和 GPT-4o 在四种逻辑框架下的成功率。成功率表示每种模型与逻辑的组合成功为自然语言推理问题生成有效解释的测试用例百分比。我们发现，采用知识蒸馏的 GPT-4o 达到了最高成功率（77.67%），而两个模型在处理一阶逻辑时都面临困难。

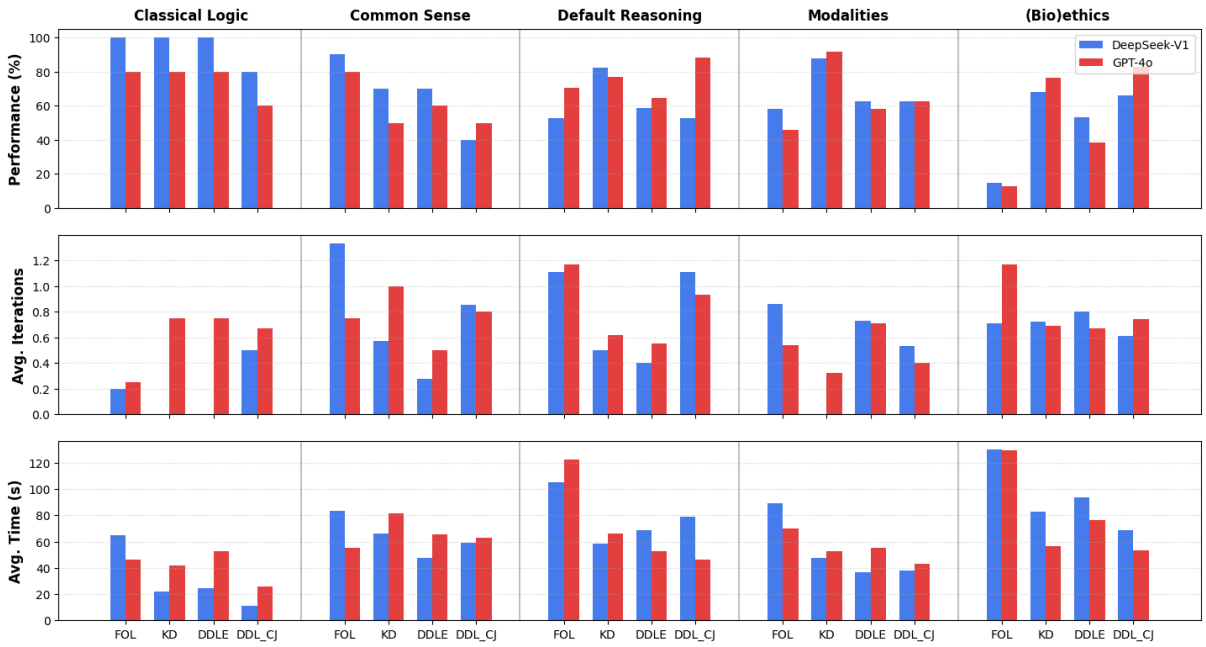

图 4：解 释在不同逻辑框架和领域中的细化表现，涵盖 DeepSeek-V1 和 d GPT-4o。求 解时间仅对成功运行取平均值。

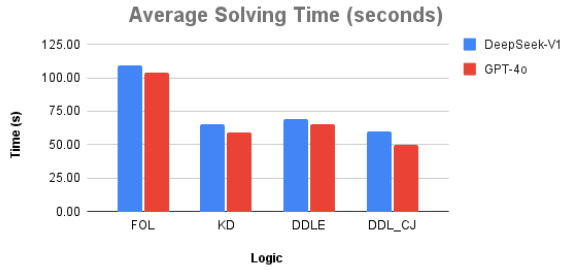

图 5：成功完成解释细化的平均求解时间（秒）。

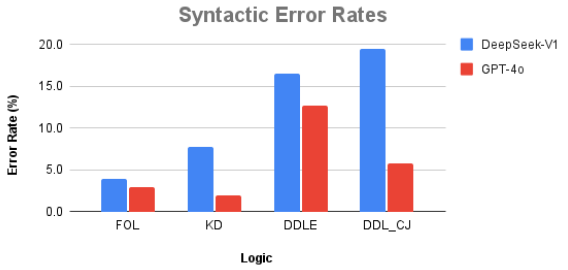

图 6：语法错误率。

细化效率 细化效率反映了逻辑收敛到有效解释的速度。图3仅关注细化深度。知识蒸馏 始终能以比其他逻辑更少的细化步骤得出有效解释。

计算效率 图 5 报告了各逻辑的平均求解时间。一阶逻辑 表现出最高的计算成本，这反映了一阶推理与细化相结合所带来的开销。知识蒸馏显著缩短了运行时间，而*DDL_CJ*尽管采用了表达能力更强的道义逻辑框架，却实现了最低的平均求解时间。值得注意的是，细化深度与运行时间并非完全相关：表达能力强的逻辑可能需要更多细化步骤，但在句法有效的示例中仍能实现有竞争力的求解时间.....

获得解释。

对句法失败的鲁棒性 除了成功细化之外，语法错误也可能构成一种主要的失败模式。图6总结了观察到的各逻辑鲁棒性。道义逻辑，尤其是*DDLE* 和*DDL_CJ*，对格式错误的中间表示更为敏感，尤其是在默认推理、情态推理和生物伦理情境中。 一阶逻辑 表现出更高的句法鲁棒性，但总体上验证的解释更少，凸显了稳定性与表达覆盖范围之间的权衡。

特定领域性能 图 4报告了所有任务的汇总性能，并按照推理领域和逻辑进行了细分。更详细的考察揭示出一种清晰的领域依赖模式：一阶逻辑在常识推理任务中的表现始终最佳

规范／义务焦点	描述／事实焦点
DDL_CJ：细化策略：“证明步骤失败表明，从拒绝不表达自主决定到提供治疗的义务之间的逻辑链并不完整。我们需要细化解释性句子，以在以下各项之间建立清晰的联系：缺乏心智能力、削弱拒绝作为不予治疗的理由，以及提供治疗的义务。”	**一阶逻辑**：细化策略：“证明步骤失败表明，从给出的解释中推导出‘挽救生命的益处是治疗的理由’时存在逻辑错误。问题似乎源于患者的心智能力与将挽救生命的益处作为治疗理由之间缺乏直接联系。我们需要细化这些解释，以建立患者的精神状态与治疗的正当理由之间更清晰的联系。”
知识蒸馏：细化策略：“证明步骤失败表明，逻辑上的联系缺失：挽救生命的益处是治疗的理由，而义务是提供治疗。这种联系缺失或不完整。我们需要细化解释性句子，以明确建立挽救生命的益处、对拒绝的削弱以及给予治疗的义务。”	**DDLE**：细化策略：“证明步骤失败表明，假设（患者不具备心智能力）无法成立。患者的心智能力与拒绝并非自主决定之间的逻辑联系缺乏充分支持。我们需要细化解释性句子，使这种联系更加明确，并确保从患者的心智能力出发的逻辑链对治疗决定而言，是清晰且完整的。”
规范性细化	**描述性细化**
DDL_CJ：“当拒绝并非出于自主意愿时，促进他人福祉的义务通过提供救命治疗而优先。”	**一阶逻辑**：“鉴于你应该尊重自主性，患者的拒绝通常会构成不予治疗的理由，除非患者是在不具备心智能力的情况下作出拒绝。”
知识蒸馏：“当挽救生命的益处是首要考量，且理由并非不予治疗的理由被削弱时，你就应该给予治疗。”	**DDLE**：“未体现自主决定的拒绝，不构成有效的不予治疗理由。”

表2：（上）按逻辑类型划分的细化策略及其产生的陈述，并根据其关注规范关系还是描述性条件进行分类。（下）按逻辑类型划分的细化陈述，并根据其关注规范关系还是描述性条件进行分类。

，而模态逻辑和非经典逻辑（*KD*、*DDLE* 和 *DDL_CJ*）在涉及模态、默认推理以及（生物）伦理推理的领域中表现更优。仅考察汇总结果时，这一区别会被掩盖——例如，尽管*KD*的总体表现可能优于 一阶逻辑 ，但这一优势并非在所有领域中都一致。相反，结果表明，不同逻辑各具独特优势，而逻辑的选择是一项根本性的设计决策，应由目标推理领域来指导。

4.1 不同形式系统中的细化策略

表 2 并报告模型针对如何细化 3.1 中示例里的解释所提供的反馈。每种方法都需要以不同方式将初步解释与明确的桥接陈述相结合，以完成证明。我们聚焦于 GPT-4o模型和第一次细化。在该示例中，两条伦理原则在初步理由之间产生了冲突。[3] 然而，其中一条被削弱了（即证明其因某种特殊情况而不具相关性），因此剩下的一条具有约束力。

道义／规范逻辑（DDL_CJ 和知识蒸馏）关注从初步道德考量（即理由）到假设中的道义判定这一推理步骤。两种细化策略都涉及理由之间冲突的解决。*DDL_CJ* 引入了一条偏好规则，规定在某些条件下，其中一个理由“优先于”另一个理由。知识蒸馏则采取了不同的路径：它在削弱与假设中所表达的义务之间建立了逻辑蕴含。两种细化都强调义务，而非选择情境的描述性特征。此外，两种细化策略都针对最后一个推理步骤：解决冲突并推出综合考量后的义务。

描述性／事实逻辑（一阶逻辑和 DDLE）关注通过识别某些道德相关事实来确定道德理由的推理步骤。因此，它们强调条件和例外，试图通过明确某条理由在特定特殊条件下不成立来解决冲突。采用新戴维森语义学的 一阶逻辑 需要明确的例外条款 和限定语（“除非”“尤其是在.....时”），以便在基于事件的表示中处理可废止推理；其细化过程通过向现有陈述中添加明确的例外条件展示了这一点。与此同时，*DDLE*要求明确的有效性条件，规定理由在审议中何时算作有效的考量，并要求陈述是什么使某个理由“无效”。两种方法都揭示出，自然语言中看似简单的事实陈述（“不具备能力的拒绝不算数”）若要在证明系统中发挥作用，就需要经过多层次的明确形式编码。

5 讨论

综合来看，我们的定性和定量结果表明，逻辑形式体系对.....具有

[3]“理由”这一概念是当代元伦理学的核心（Schroeder 和 Howard，2024；Tucker，2025）。初步义务的思想可以追溯到（Ross，2002）。

显著影响，既影响大语言模型驱动的神经符号推理的结构，也影响其效率。

针对*RQ1*，我们观察到，不同的逻辑形式系统往往会将缺失的推理步骤定位在解释链条中的不同位置：KD 和 DDL_CJ 等道义逻辑侧重于细化从相互竞争的理由到综合考量后的义务这一规范性转变，而 FOL 和 DDLE 等逻辑则倾向于强调额外的事实条件或有效性约束，以决定某项考量是否能算作理由。这一区别构成了*RQ2*的基础。在逻辑外部方法中，尤其是 FOL，仅将道德理由隐式表示为谓词或命题常元，从而削弱理由与义务之间的逻辑联系，并使得脆弱且临时拼凑的细化成为必要。相比之下，逻辑内部方法，尤其是 DDL_CJ，将理由直接表示为条件规范，从而能够解决冲突，并有助于导出无条件义务（更多详情见附录 A.2）。就*RQ3*而言，这些结构性差异转化为可衡量的收益：具有更强内部规范结构的逻辑收敛得更可靠，所需的细化步骤更少，并且在规范性案例中的解释成功率更高。

更广泛地说，我们的结果表明，逻辑参数化架构不仅能够提升性能指标，还揭示了不同形式系统如何突出实践推理的不同阶段，从而推动未来的系统动态选择或组合不同的逻辑形式体系。

6 相关工作

神经符号自然语言推理 当代神经符号自然语言推理系统旨在结合大语言模型（大语言模型）的语言流畅性和语境感知能力，以及形式推理的严谨性和透明性（Quan 等人，2025b,a，2024；Pan 等人，2023；Olausson 等人，2023；Ye 等人，2023；Ranaldi 等人，2025；Arakelyan 等人，2025；Tan 等人，2025；Qi 等人，2025）。近期研究探索了利用大语言模型执行自动形式化（Wu 等人，2022；Zhang 等人，2025）和解释生成（Quan 等人，2024；Dalal 等人，2024）等任务，通常与自动定理证明器（定理证明器）协同工作，以验证推理有效性（Pan 等人，2023；Olausson 等人，2023；Jiang 等人，2023；Quan 等人，2024）。在这种设置下，大语言模型生成候选形式化表示，这些表示随后由基于逻辑的组件（如 Isabelle/HOL 或 Lean）进行检查、完善或补全。尽管前景可期，但这些架构大多假设采用固定的逻辑框架——通常是一阶逻辑或命题逻辑——来开展推理。这限制了系统适应需要更专门推理结构的领域的能力。我们的工作与Xu 等人（2025）的研究动机相似；然而，他们关注的是动态求解器组合，而非不同逻辑形式体系的影响。

道义解释与形式逻辑 解释这一概念可以非正式地定义为对“为什么”问题的回答。在道义逻辑的语境中，“为什么”问题可能涉及某些道义判定，例如：“为什么 A 是义务的（/允许的/禁止的）？”道义逻辑中的解释也可以采取对比形式：“尽管对于 B 存在初步义务，为什么 A 是义务的而不是 B？”对比解释通常涉及处理道德冲突、例外、偏好以及违反义务的情形。从形式逻辑的角度来看，提供道义解释的目的在于满足系统的某些要求。具体而言，这意味着我们不仅希望系统给出正确的输出，还希望它能够透明、准确且令人信服地说明为何会得到这些输出。关于道义解释的大多数形式化研究都与形式论证领域相关（Gov- ernatori 等人，2022年；Rotolo 和 Sartor，2023；van Berkel 和 Straßer，2024）。

7 结论

本文提出了一种用于神经符号自然语言推理的逻辑参数化框架，借助 LogiKEy 方法论将多种逻辑系统嵌入高阶逻辑。我们的核心主张是，逻辑的选择并非中性的设计决策，而是影响推理泛化能力、证明行为和解释质量的关键因素。

这项工作的更广泛意义在于，有可能构建逻辑自适应推理架构——这类系统能够根据特定任务或输入的需求，动态选择、切换或组合逻辑。此类流水线不会假定所有推理过程都采用固定逻辑，而是将逻辑选择视同机器学习中的模型选择，并依据结构、语义或语境线索进行引导。

8 局限

尽管我们的分析揭示了不同逻辑形式体系如何与大语言模型交互并处理规范性推理和伦理推理，但仍须承认其中存在若干局限。

我们的研究使用了两种语言模型（GPT-4o 和 DeepSeek-V1）来生成和细化逻辑形式化。我们还需要进一步研究，以了解我们的发现能否推广到不同模型。例如，较小的模型可能难以应对所需的复杂多步推理，而其他大型模型可能会采用不同的推理策略。

此外，我们的实证研究考察了四种特定的逻辑系统（知识蒸馏、一阶逻辑、*DDLE*和*DDL_CJ*）。这些系统代表了规范推理的特定方法，但还存在许多其他形式系统，包括量化模态逻辑和条件逻辑、默认逻辑、论证框架、非单调逻辑以及概率推理系统。未来的研究可以探讨我们的观察结果如何适用于其他系统；每种系统可能需要不同类型的细化，也可能呈现不同的失效模式。

此外，我们分析的细化策略由自动化系统生成，或通过结构化的交互流程制定。这些细化代表完成每个证明的一种可能路径，但其他细化方案也可能弥补逻辑缺口。我们尚未系统地探索每种形式系统所有可能的细化方案，也未评估所选方案是否从人类推理的角度来看最为自然。

最后，我们的研究使用相对简单、结构化的自然语言解释，以实现全面的实验控制。现实世界推理可能涉及更复杂、更细致或更隐含的推理模式，这些模式可能会给形式化带来额外挑战。当自然语言输入更加复杂时，尤其是涉及歧义术语、语境依赖或特定文化的规范概念时，我们观察到的局限可能会更加明显。

参考文献

Lennart Åqvist。1984年。《道义逻辑》。载于 D. Gabbay 和 F. Guenthner 编，《哲学逻辑手册：第二卷：经典逻辑的扩展》，第 605–714 页。施普林格，多德雷赫特。

Erik Arakelyan、Pasquale Minervini、Patrick Lewis、Pat Verga 和 Isabelle Augenstein。2025。FLARE：忠实的逻辑辅助推理与探索。载于 *2025* 年自然语言处理实证方法会议论文集，第 23396–23414 页，中国苏州。计算语言学协会。

Pietro Baroni、DovGabbay、Massimilino Giacomin 和 Leendert Van der Torre。2018年。《形式论证手册》。College Publications。

Tom L. Beauchamp 和 James F. Childress。1979年。《生物医学伦理学原理》。牛津大学出版社。

Emily M Bender、Timnit Gebru、Angelina McMillan- Major 和 Shmargaret Shmitchell。2021年。论随机鹦鹉的危险：语言模型会不会过于庞大？载于*2021*年*ACM*公平、问责与透明度会议论文集，第610–623页。C. Benzmüller 和 L.C. Paulson。2010年。简单类型论中的多模态逻辑与直觉主义逻辑。《*IGPL*逻辑期刊》，18(6):881–892。C. Benzmüller 和 L.C. Paulson。2013年。简单类型论中的量化多模态逻辑。《普遍逻辑》（多模态逻辑特刊），7(1):7–20。Christoph Benzmüller。2019年。普遍（元）逻辑推理：近期成果。《计算机编程科学》，172:48–62。Christoph Benzmüller、Ali Farjami、David Fuenmayor、Paul Meder、Xavier Parent、Alexander Steen、Leendert van der Torre 和 Valeria Zahoransky。2020a年。LogiKEy工作台：道义逻辑、逻辑组合以及富有表现力的伦理与法律推理（Isabelle/HOL数据集）。《数据简报》，33:106409。Christoph Benzmüller、Ali Farjami 和 Xavier Parent。2018年。高阶逻辑中的二元道义逻辑。载于《道义逻辑与规范系统——第*14*届国际会议 *DEON 2018*》，*2018*年*7*月*3*日至*6*日，荷兰乌得勒支，第33–50页，伦敦。College Publications。Christoph Benzmüller、Ali Farjami 和 Xavier Parent。2019年。高阶逻辑中的Åqvist二元道义逻辑E。《应用逻辑期刊——*IfCoLoG*逻辑及其应用期刊》（法律人工智能推理特刊），6(5):733–755。Christoph Benzmüller、Xavier Parent 和 Leendert van der Torre。2020b年。为伦理与法律推理设计规范理论：LogiKEy框架、方法论与工具支持。《人工智能》，237:103348。Selim Berker。2022年。道义者、评价者与适切者。载于《适切性：规范性哲学论文集》。牛津大学出版社。

Bikram Pratim Bhuyan、Amar Ramdane-Cherif、Ravi Tomar 和 TP Singh。2024。神经符号人工智能：综述。《神经计算与应用》，36(21):12809–12844。

Xiao Bi、Deli Chen、Guanting Chen、Shanhuang Chen、Damai Dai、Chengqi Deng、Honghui Ding、Kai Dong、Qiushi Du、Zhe Fu 等 1 人。2024。Deepseek 大语言模型：通过长文本扩展开源语言模型。*arXiv* 预印本 *arXiv:2401.02954*。

Jasmin Christian Blanchette、Sascha Böhme 和 Lawrence C Paulson。2013。使用 SMT 求解器扩展 sledgehammer。自动推理期刊，51(1)：109–128。

J.C. Blanchette 和 T. Nipkow。2010。Nitpick：一种基于关系模型查找器、用于高阶逻辑的反例生成器。载于第一届国际交互式定理证明会议，《计算机科学讲义》第 6172 卷，第 131–146 页，柏林。施普林格。

Sébastien Bubeck、VarunChandrasekaran、Ronen Eldan、Johannes Gehrke、Eric Horvitz、Ece Kamar、Peter Lee、Yin Tat Lee、Yuanzhi Li、Scott Lundberg 等 1 人。2023。通用人工智能的火花：使用 GPT-4 进行的早期实验。*arXiv* 预印本 *arXiv:2303.12712*。

Oana-Maria Camburu、Tim Rocktäschel、Thomas Lukasiewicz 和 Phil Blunsom。2018年。e-SNLI：带有自然语言解释的自然语言推理。《神经信息处理系统进展》，31。J. Carmo 和 A.J.I. Jones。2013年。反义务条件句逻辑的完备性与可判定性结果。《逻辑与计算杂志》，23(3):585–626。Brian F Chellas。1980年。模态逻辑。剑桥大学出版社，剑桥。Fengxiang Cheng、Haoxuan Li、Fenrong Liu、Robert van Rooij、Kun Zhang 和 Zhouchen Lin。2025。以逻辑推理赋能大语言模型：全面综述。*arXiv*预印本 *arXiv:2502.15652*。Roderick M. Chisholm。1963年。反义务命令与道义逻辑。《*Analysis*》，24：33–36。Dhairya Dalal、Marco Valentino、Andre Freitas 和 Paul Buitelaar。2024。大语言模型中的最佳解释推理。载于《计算语言学协会第*62*届年会论文集（第*1*卷：长篇论文）》，第217–235页，泰国曼谷。计算语言学协会。Denis Emelin、Ronan Le Bras、Jena D Hwang、Maxwell Forbes 和 Yejin Choi。2021。道德故事：关于规范、意图、行动及其后果的情境推理。载于《*2021*年自然语言处理经验方法会议论文集》，第698–718页。

Ali Farjami。2020年。话语输入/输出逻辑：道义情态与计算。博士论文，卢森堡大学，卢森堡。

Maxwell Forbes、Jena D Hwang、Vered Shwartz、Maarten Sap 和 Yejin Choi。2020年。《社会化学101：学习推理社会规范与道德规范》。*arXiv* 预印本 *arXiv:2011.00620*。

D.M. Gabbay、J. Horty、X. Parent、R. van der Meyden 和 L. van der Torre，编。2013年。道义逻辑与规范系统手册。College Publications。

Artur d'Avila Garcez 和 Luis C Lamb。2023。神经符号人工智能：第三次浪潮。人工智能评论，56(11)：12387–12406。

Lou Goble。2013。初步规范、规范冲突与两难困境。载于 D.M. Gabbay、J. Horty、X. Parent、R. van der Meyden 和 L. van der Torre 编，《style id='1'>道义逻辑手册》，第1卷，第499–544页。College Publications，伦敦。

Guido Governatori、Francesco Olivieri、Antonino Ro- tolo 和 Matteo Cristani。2022年。稳定规范解释。载于《法律知识与信息系统》，第43–52页。IOS Press。

Muhammad Usman Hadi、Rizwan Qureshi、Abbas Shah、Muhammad Irfan、Anas Zafar、Muhammad Bilal Shaikh、Naveed Akhtar、Jia Wu、Seyedali Mirjalili及另外1位作者。2023年。大语言模型：其应用、挑战、局限与未来前景的全面综述。*Authorea*预印本，1(3):1–26。D Hendrycks、C Burns、S Basart、A Critch、J Li、D Song和J Steinhardt。2021年。让人工智能与人类共同价值观保持一致，29。Wesley H. Holliday、Matthew Mandelkern和Cede- gao E. Zhang。2024年。大语言模型中的条件句与模态推理。载于*2024*年自然语言处理实证方法会议论文集，第3800–3821页，佛罗里达州迈阿密，美国。计算语言学协会。Albert Qiaochu Jiang、Sean Welleck、Jin Peng Zhou、Timothee Lacroix、Jiacheng Liu、Wenda Li、Mateja Jamnik、Guillaume La mple和Yuhuai Wu。2023年。起草、构思并证明：用非形式化证明指导形式化定理证明器。载于第十一届国际学习表征会议。Zihao Li。2023年。ChatGPT的阴暗面：随机鹦鹉与幻觉带来的法律和伦理挑战。*arXiv*预印本 *arXiv:2304.14347*。Hanmeng Liu、Zhizhang Fu、Mengru Ding、Ruoxi Ning、Chaoli Zhang、Xiaozhang Liu和Yue Zhang。2025年。大语言模型中的逻辑推理：综述。*arXiv*预印本 *arXiv:2502.09100*。

T. Nipkow、L.C. Paulson和M. Wenzel。2002年。*Is- abelle/HOL*——用于高阶逻辑的证明助手，第2283卷，出自《计算机科学讲义》。施普林格，柏林。

Theo Olausson、Alex Gu、Ben Lipkin、Cedegao Zhang、Armando Solar-Lezama、Joshua Tenenbaum 和 Roger Levy。2023年。LINC：结合语言模型与一阶逻辑证明器进行逻辑推理的神经符号方法。载于*2023*年自然语言处理实证方法会议论文集，第5153–5176 页，新加坡。计算语言学协会。

Liangming Pan、Alon Albalak、Xinyi Wang 和 William Wang。2023。Logic-LM：利用符号求解器赋能大语言模型，实现可靠的逻辑推理。载于计算语言学协会论文集：*EMNLP2023* 精选论文，第 3806–3824 页，新加坡。计算语言学协会。

Xavier Parent。2021年。汉松型二元义务逻辑的偏好语义：成果综述。载于 D.M. Gabbay、J. Horty、X. Parent、R. van der Meyden 和 L. van der Torre 编，《道义逻辑手册》第 2 卷，第 7–70 页。College Publications，伦敦。

Chengwen Qi、Ren Ma、Bowen Li、He Du、Binyuan Hui、Jinwang Wu、Yuanjun Laili 和 Conghui He。2025。大语言模型与符号证明器相结合，用于逻辑推理评估。载于第十三届国际学习表征会议。

Xin Quan、Marco Valentino、Danilo Carvalho、Dhairya Dalal 和 Andre Freitas。2025a。PEIRCE：通过大语言模型驱动的神经符号精化统一实质推理与形式推理。载于计算语言学协会第六十三届年会论文集（第三卷：系统演示），第11–21页，奥地利维也纳。计算语言学协会。

Xin Quan、Marco Valentino、Louise A. Dennis 和 Andre Freitas。2024。通过大语言模型符号定理证明验证并细化自然语言解释。载于*2024*年《自然语言处理中的经验方法》会议论文集，第2933–2958页，美国佛罗里达州迈阿密。计算语言学协会。

Xin Quan、Marco Valentino、Louise A. Dennis 和 An- dre Freitas。2025b。面向自然语言推理解释的忠实且稳健的大语言模型驱动定理证明。载于 第*63*届计算语言学协会年会论文集（第*1*卷：长篇论文），第17734–17755页，奥地利维也纳。计算语言学协会。

Leonardo Ranaldi、Marco Valentino 和 Andre Fre- itas。2025年。通过准符号抽象改进思维链推理。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17222–17240页，奥地利维也纳。计算语言学协会。

William David Ross。2002年。《正当与善》。牛津大学出版社。

Antonino Rotolo 和 Giovanni Sartor。2023。法律中的论证与解释。《人工智能前沿》，6:1130559。

Mark Schroeder 和 Nathan Howard。2024。理由的基础。牛津大学出版社。

Xingwei Tan、Marco Valentino、Mahmud Elahi Akhter、Maria Liakata 和 Nikolaos Aletras。2025。 通过符号引导的蒙特卡洛过程监督增强语言模型的逻辑推理能力。载于*2025*年自然语言处理实证方法会议论文集，第31874–31888页，中国苏州。计算语言学协会。

Chris Tucker. 2025.《理由的分量：伦理学框架》。牛津大学出版社。

Kees van Berkel 和 Christian Straßer。2024年。通过对话实现道义解释。载于 第二届可解释人工智能论证国际研讨会（*ArgXAI*）。*CEUR*研讨会论文集（即将发表）。

Georg Henrik von Wright。1951。道义逻辑。《心灵》，60(237):1–15。

Yuhuai Wu、Albert Qiaochu Jiang、Wenda Li、Markus Norman Rabe、Charles E Staats、Mateja Jamnik 和 Christian Szegedy。2022年。利用大语言模型进行自动形式化。载于 《神经信息处理系统进展》。

Lei Xu、Pierre Beckmann、Marco Valentino 和 André Freitas。2025。通过动态逻辑求解器组合实现自适应大语言模型符号推理。*arXiv*预印本 *arXiv:2510.06774*。

Xi Ye、Qiaochu Chen、Isil Dillig 和 Greg Durrett。2023。SatLM：使用声明式提示的可满足性辅助语言模型。载于第三十七届神经信息处理系统大会。

Lan Zhang、Marco Valentino 和 AndreFreitas。2025。《真实场景中的自动形式化：评估大语言模型对现实世界数学定义的处理能力》。载于*2025* 年自然语言处理实证方法会议论文集，第 1720–1738 页，中国苏州。计算语言学协会。

A 附录

A.1 逻辑参数化形式化的提示适配

为支持逻辑参数化的定理证明，我们改编了Quan等人(2025b) 提出的提示框架。该框架最初旨在用于 Neo-Davidsonian一阶逻辑（FOL）的形式化。虽然他们的提示能有效引导大语言模型生成忠实的

基于事件的表示，但其假设逻辑基底固定不变。我们的扩展修改了这些提示，以适应通过 LogiKEy 框架嵌入的多种经典逻辑和非经典逻辑。

HOL 中的模态逻辑与条件逻辑 所谓的 浅层语义嵌入 方法由 Benzmüller（Benzmüller，2019年）提出，用于将经典与非经典逻辑的语义转换为 HOL。文献（Benzmüller 和 Paulson，2010年、2013年）讨论了在 HOL 中嵌入一大类模态逻辑。二元义务逻辑（系统**DDLE**）的语义嵌入见文献（Benzmüller 等人，2019年），而 Carmo 和 Jones 条件逻辑的嵌入则见文献（Benzmüller 等人，2018年）。图 7 更详细地讨论了这些在 HOL 中的逻辑嵌入。

提示适配概述。 我们没有引入全新的提示，而是保留原有流水线的整体结构，并有选择地调整行为取决于底层逻辑的提示。这样既能保留逻辑不变的语义提取，又能实现特定逻辑的推理、细化和证明构造。具体而言，我们区分 *与逻辑无关* 和 *与逻辑相关* 的提示；切换目标逻辑时，只修改后者。

逻辑无关提示 第一类提示负责从自然语言中提取语义内容，不受目标逻辑的影响。这类提示包括：

- **句法解析提示** – 提取语法结构，以指导谓词—论元映射。
- **生成与细化解释提示**：要求大语言模型利用因果知识和常识，为所提供的因果推理情境给出逻辑解释。

由于这些提示纯粹作用于自然语言推理，并处于语义解析层面，因此其输出在不同逻辑设定下保持稳定。因此，在所有实验中都会原样重复使用这些提示。这一设计选择确保，不同逻辑之间观察到的差异归因于推理和证明机制，而非语义解释的变化。

逻辑敏感编码提示 第二类提示将语义表示转换为 Isabelle/HOL 公理和定理陈述。这些提示直接受到逻辑选择的影响，因为它们决定了形式化中使用的逻辑算子、模态或义务逻辑构造以及公理模式。

重要的是，尽管生成的公理的句法形式会随逻辑而变化，但从输入中提取的底层语义内容保持不变。这种分离使我们能够系统地研究逻辑选择对下游推理的影响。如图8所示，将自然语言自动形式化为模态逻辑，依赖于在高阶逻辑中对**知识蒸馏**的编码，而 Isabelle 公理提示则基于同一底层的**知识蒸馏**理论。

逻辑敏感的细化与修复提示第三类提示用于处理解释细化、矛盾消解，以及证明尝试失败时的语法修复。这类提示高度依赖逻辑，因为不同逻辑中矛盾的概念和可采用的修复策略有很大差异。这些调整对于不同逻辑在细化效率和稳健性方面呈现出的差异起着关键作用；模态逻辑和条件逻辑需要更细致的错误修正策略。

逻辑敏感的证明构造提示最后，证明构造提示会根据流水线前序阶段生成的公理和假设，生成 Isabelle/HOL 证明草图。这类提示对逻辑选择最为敏感，因为证明策略直接取决于目标逻辑的推理规则和公理。对于每种逻辑，我们都会修改这些提示，以确保生成的证明遵循 LogiKEy 中编码的相应模态或道义原则。

A.2 使用 MoralReasons 进行推理：四种逻辑的比较

在对示例的解释中3.1，核心推理模式是削弱。它由两个步骤构成：首先，实例化两个表面理由；其次，其中一个理由被削弱。引入这些理由的前提体现了生物伦理推理中的一种典型模式：“根据 Principle1，Fact1($F1$) 是支持 Act1($A1$) 的理由”。四种逻辑如何形式化这一命题，颇为有趣。值得注意的是，它们都没有用于表达理由的原始算子。

```
theory  KD   imports Main
begin
  typedecl i — ‹type for possible worlds›
  type_synonym τ = "(i⇒bool)"
  consts r_t :: "i⇒i⇒bool" (infixr "rt" 70)(*relation for a modal logic KD*)

  abbreviation serial  where "serial  r ≡ (∀x.∃y. r x y)"

  axiomatization where ax_serial_rt : "serial  r_t"

  definition KDnot :: "τ⇒τ" ("¬_"[52]53) where "¬φ ≡ λw. ¬φ(w) "
  definition KDor :: "τ⇒τ⇒τ" (infixr "∨"50) where "φ∨ψ ≡ λw. φ(w) ∨ ψ(w)"
  definition KDand :: "τ⇒τ⇒τ" (infixr "∧"51) where "φ∧ψ ≡ λw. φ(w) ∧ ψ(w)"
  definition KDimp :: "τ⇒τ⇒τ" (infixr "⟶" 49) where "φ⟶ψ ≡ λw. φ(w) ⟶ ψ(w)"
  definition KDtrue  :: "τ" ("⊤") where "⊤ ≡ λw. True"
  definition KDfalse :: "τ" ("⊥") where "⊥ ≡ λw. False"

  definition KDobligatory :: "τ⇒τ" ("O") where "O φ ≡ λw. ∀v. w rt v ⟶ φ(v)"
  definition KDforbidden :: "τ⇒τ" ("F") where "F φ ≡ O(¬φ)"
  definition KDpermitted :: "τ⇒τ" ("P") where "P φ ≡ ¬(O(¬φ))"
  definition KDvalid :: "τ⇒bool" ("⌊_⌋" [8]109)  where "⌊p⌋ ≡ ∀w. p(w)"
  lemma True nitpick [satisfy,user_axioms,expect=genuine,show_all,format=2] oops
end
```

(a) 高阶逻辑中的模态逻辑 KD。• 第 3 行引入了用于表示可能世界的原始类型 i。• 第 4 行引入了用于表示公式的类型 τ。• 第 5 行引入了 r_t，用于编码可及关系。• 第 7 行通过串行性限制可及关系。• 第 11–16 行定义了布尔联结词。

• 第 18–20 行定义了一元道义算子（义务、禁止、许可）。• 第 21 行引入了全局有效性。• 第 22 行使用 Nitpick 确认一致性。

```
theory DDLE imports Main
begin
typedecl i (*type for possible worlds*)
type_synonym τ = "(i⇒bool)" (*type for formulas*)
consts aw::i (*actual world*)
consts r :: "i⇒τ"  (infixr "rt" 70) (*comparative goodness relation*)

definition ddetop          :: "τ" ("⊤")              where "⊤  ≡ λw. True"
definition ddebot          :: "τ" ("⊥")              where "⊥  ≡ λw. False"
definition ddeneg          :: "τ⇒τ" ("¬_"[52]53)     where "¬φ ≡ λw. ¬φ(w)"
definition ddeand          :: "τ⇒τ⇒τ" (infixr"∧"51) where "φ∧ψ ≡ λw. φ(w)∧ψ(w)"
definition ddeor           :: "τ⇒τ⇒τ" (infixr"∨"50) where "φ∨ψ ≡ λw. φ(w)∨ψ(w)"
definition ddeimp          :: "τ⇒τ⇒τ" (infixr"⟶"49) where "φ⟶ψ ≡ λw. φ(w)⟶ψ(w)"
definition ddeequivt       :: "τ⇒τ⇒τ" (infixr"↔"48) where "φ↔ψ ≡ λw. φ(w)⟷ψ(w)"

definition ddebox          :: "τ⇒τ" ("□")  where "□φ ≡ λw. ∀v. φ(v)" (*necessity operator*)
definition ddediomond      :: "τ⇒τ" ("◇")  where "◇φ ≡ λw. ∃v. φ(v)" (*possibility operator*)

definition ddeopt :: "τ⇒τ" ("opt<_>")
  where "opt<φ> ≡ (λv. ( (φ)(v) ∧ (∀x. ((φ)(x)  ⟶  v r x)) ) )"
abbreviation(input) msubset :: "τ⇒τ⇒bool" (infix "⊆" 53)
  where "φ ⊆ ψ ≡ ∀x. φ x ⟶ ψ x"
definition ddecond :: "τ⇒τ⇒τ" ("O<_|_>")
  where "O<ψ|φ> ≡ λw. opt<φ> ⊆ ψ" (*conditional obligation*)
definition ddeperm :: "τ⇒τ⇒τ" ("P<_|_>")
  where "P<ψ|φ> ≡ ¬O<¬ψ|φ>" (*conditional permission*)
definition ddeproh :: "τ⇒τ⇒τ" ("F<_|_>")
  where "F<ψ|φ> ≡ O<¬ψ|φ>" (*conditional prohibition*)

definition ddevalid :: "τ⇒bool" ("⌊_⌋"[8]109) (*global validity*)
  where "⌊p⌋ ≡ ∀w. p w"
definition ddeactual :: "τ⇒bool" ("⌊_⌋"[7]105) (*local validity*)
  where "⌊p⌋ ≡ p(aw)"

lemma True nitpick [satisfy,user_axioms,show_all,expect=genuine] oops (*consistency check*)
end
```

(b) HOL 中的条件逻辑 DDLE。

```
theory CJ_DDL imports Main
begin
 typedecl i (*type for possible worlds*)
 type_synonym τ = "(i⇒bool)" type_synonym ο = "τ⇒τ"  type_synonym ρ = "τ⇒τ⇒τ"
 consts av::"i⇒τ" ob::"τ⇒(τ⇒bool)" (*accessibility, resp. neighborhood, relations*)
 axiomatization where
  ax_3a: "∀w.∃x. av(w)(x)" and
  ax_4a: "∀w x. av(w)(x) ⟶ pv(w)(x)" and
  ax_4b: "∀w. pv(w)(w)" and
  ax_5a: "∀X.¬ob(X)(λx. False)" and
  ax_5b: "∀X Y Z. (∀w. ((Y(w) ∧ X(w)) ⟷ (Z(w) ∧ X(w)))) ⟶ (ob(X)(Y) ⟷ ob(X)(Z))" and
  ax_5c: "∀X Y Z. (((∃w. (X(w) ∧ Y(w) ∧ Z(w))) ∧ ob(X)(Y) ∧ ob(X)(Z))
                  ⟶ ob(X)(λw. Y(w) ∧ Z(w)))" and
  ax_5d: "∀X Y Z. ((∀w. Y(w) ⟶ X(w)) ∧ ob(X)(Y) ∧ (∀w. X(w) ⟶ Z(w)))
                  ⟶ ob(Z)(λw. (Z(w) ∧ ¬X(w)) ∨ Y(w))" and
  ax_5e: "∀X Y Z. ((∀w. Y(w) ⟶ X(w)) ∧ ob(X)(Z) ∧ (∃w. Y(w) ∧ Z(w))) ⟶ ob(Y)(Z)"

 abbreviation ddltop::τ ("⊤") where "⊤ ≡ λw. True"
 abbreviation ddlbot::τ ("⊥") where "⊥ ≡ λw. False"
 abbreviation ddlneg::ο ("¬_"[52]53) where "¬φ ≡ λw. ¬φ(w)"
 abbreviation ddland::ρ (infixr"∧"51) where "φ∧ψ ≡ λw. φ(w)∧ψ(w)"
 abbreviation ddlor::ρ (infixr"∨"50) where "φ∨ψ ≡ λw. φ(w)∨ψ(w)"
 abbreviation ddlimp::ρ (infixr"→"49) where "φ→ψ ≡ λw. φ(w)⟶ψ(w)"
 abbreviation ddlbox::ο ("□") where "□φ ≡ λw.∀v. φ(v)"
 abbreviation ddlboxa::ο ("□ₐ") where "□ₐφ ≡ λw. (∀x. av(w)(x) ⟶ φ(x))"
 abbreviation ddlboxp::ο ("□ₚ") where "□ₚφ ≡ λw. (∀x. pv(w)(x) ⟶ φ(x))"
 abbreviation ddldia::ο ("◇") where "◇φ ≡ ¬□(¬φ)"
 abbreviation ddldiaa::ο ("◇ₐ") where "◇ₐφ ≡ ¬□ₐ(¬φ)"
 abbreviation ddldiap::ο ("◇ₚ") where "◇ₚφ ≡ ¬□ₚ(¬φ)"
 abbreviation ddlo::ρ ("O(_|_)"[52]53) where "O(ψ|φ) ≡ λw. ob(φ)(ψ)"
 abbreviation ddlobl::ο ("O<_>") where "O<φ> ≡  O(φ|⊤)"
 abbreviation ddloa::ο ("Oₐ") where "Oₐφ ≡ λw. ob(av(w))(φ) ∧ (∃x. av(w)(x) ∧ ¬φ(x))"
 abbreviation ddlop::ο ("Oₚ") where "Oₚφ ≡ λw. ob(pv(w))(φ) ∧ (∃x. pv(w)(x) ∧ ¬φ(x))"
 abbreviation ddlvalid::"τ ⇒ bool" ("⌊_⌋"[7]105) where "⌊φ⌋ ≡ ∀w. φ w"
 lemma True nitpick [satisfy,user_axioms,show_all] oops (*consistency*)
end
```

(c) HOL 中的条件逻辑 DDL_CJ。

图 7：在我们的逻辑参数化设置中使用的 Isabelle/HOL 嵌入：KD（上方）和两种条件逻辑（下方）。

• 在 **一阶逻辑** 中，“……是……的理由”被视为一个谓词。

• 在 **知识蒸馏**中，道义算子用于表达尊重自主性的总体命令以及假设中的道义判定，但不用于表达 $F1$ 与 $A1$ 之间的理由关系。“$F1$ 是 $A1$ 的理由”通过命题常量来表示。

• 在 **DDLE** 中，第一次迭代时，二元道义算子用于将理由表达为条件规范： $\bigcirc(A1/F1)$。然而，细化过程再次将理由表示为命题常量，从而产生了扁平表示。

当理由以如此抽象的方式表示时，它们与综合考量后的义务之间的关系便会丢失。模型需要捕捉这种关系

这需要通过引入新的解释性句子来建立逻辑联系，但这一过程可能既漫长又不可靠。事实上，迄今提到的三种逻辑都无法针对当前案例提供有效解释。

- **DDL_CJ**则不同，它成功地运用二元道义算子来表示理由。支持 $A1$ 的理由和反对 $A1$ 的理由都被表示为条件规范。

这使得表示理由成为可能，但如何得出综合考量后的义务呢？在当前示例中，关键在于削弱，它作为一种冲突解决技术，用来消解两个理由之间的冲突。这种推理步骤是可废止逻辑和论证的典型做法，但并非**DDL_CJ** 的设计初衷。不过，混合模型成功克服了这一障碍，将解释重新表述为一种条件偏好："当 $F2$ 成立时，支持 $A1$ 的理由优先。"这一命题继而被表示为一个条件义务：$\bigcirc(A1/F2)$；也就是说，"已知 $F2$，就应当 $A1$。"这一策略奏效了：**DDL_CJ** 能够将两个理由都表示为条件规范，并推出一项给予治疗的无条件义务，从而得到有效解释。

这项分析表明，大语言模型——至少在某些情况下——能够将逻辑拓展到其偏好的解释领域之外。道义算子可用于重构规范概念，例如道德理由。在这种情况下，原本繁琐的任务——手动构建不同规范概念之间的关系（例如，最有理由去做的事与应该做的事之间的联系）——便可部分交由逻辑定理处理，从而提高快速得出有效解释的可能性。

A.3 关于数据集的更多详情

正如我们在第 3 节中所述，该数据集探讨道义推理，尤其是伦理推理。我们关注的是构成这类推理的不同推理模式。数据集的组织方式也反映了这一重点：每个文件夹对应一种特定的推理模式或多种模式的组合。在本小节中，我们将介绍所有文件夹，并逐一说明和讨论相应的推理模式。

在介绍这些文件夹之前，先回顾一下案例的结构会很有帮助。每个案例由以下部分组成：一组前提、一个假设和一段解释。解释的作用是通过一系列推理步骤，将前提与假设联系起来。这些步骤决定了该案例属于哪个文件夹。

1. **经典逻辑（5 个案例）**。[4] 此文件夹中的案例仅需对前提应用经典逻辑，即可推导出假设。

2. **常识（10 个案例）**。[5] 这个文件夹分为两个子文件夹：前者探讨认识论常识推理，后者则侧重于实践常识推理。在这两类案例中，推导假设的关键在于理解两个术语含义相同。请注意，在自然语言中，这种情况也可能出现在道义陈述中：例如，可以说"杀人是不对的"，也可以说"人不应杀人"。[6]

3. **默认推理（17 个案例）**。[7] 这个文件夹分为两个子文件夹："认识论默认推理"和"实践默认推理"。前者包含一些案例，在这些案例中，必须使用关于世界的某种默认假设来推导假设。后者则包含两类案例。首先，我们纳入认识论案例在实践方面的对应案例，也就是必须使用关于道德的默认假设（例如，不应做坏事）来推导假设的案例。其次，我们引入道德理由的概念。理由是支持或反对某一特定行动的事实（Schroeder 和 Howard，2024；Tucker，2025）。在第二类案例中，要解释假设中的道义判断，必须认识到前提中的某个事实是一项支持该判断的理由。

4. **模态（24 个案例）**。[8] 真理模态

[4]在此文件夹中，我们改编了 3 个案例，取自（Holliday 等人，2024）。

[5]在此文件夹中，我们改编了 5 个案例，取自 e-SNLI 数据集（Camburu 等人，2018年）。

[6]元伦理学家（Berker，2022年）对这些不同表达进行了研究，并将其归为不同类别。在此文件夹中，我们只使用属于义务类别的成对术语。在下一个关于默认推理的文件夹中，我们使用的成对术语则横跨不同类别。

[7]在此文件夹中，我们改编了 5 个 e-SNLI 数据集中的案例。[8]在此文件夹中，我们改编了 2 个案例，取自（Holliday 等人，2024）。

模态（必然性、可能性、不可能性）和道义模态（义务、许可、禁止）呈现出某些逻辑关系：例如，如果 φ 是必然的，那么 $\neg\varphi$ 就不可能；同样，如果 φ 是义务的，那么$\neg\varphi$ 就不被允许。在此文件夹的案例中，解释指的是模态之间的逻辑关系。由于我们的重点是实践推理，因此我们会更多地讨论道义模态。特别是，我们引入条件义务这一概念，它是二元义务逻辑的核心，通常用以下符号形式化表示： $\bigcirc(\psi/\varphi)$。我们探讨与条件应当相关的两种推理模式：事实脱离（即由 $\bigcirc(\psi/\varphi)$和 φ推得 $\bigcirc(\psi/\top)$），以及道义脱离（即由 $\bigcirc(\psi/\varphi)$和 $\bigcirc(\varphi/\top)$推得 $\bigcirc(\psi/\top)$）。

5. **（生物）伦理学**。假设你认同一些一般性的道德原则，例如仁慈是好的、自主性应受到尊重。现在考虑一个生物伦理案例：一位患者拒绝接受有益治疗。你应该强迫患者接受治疗吗？要回答这个问题，你需要运用一般性的伦理原则来确定表面理由（步骤 1）。例如，鉴于自主性应受到尊重，患者的拒绝就是一个不应强迫其接受治疗的理由。接着，如果你发现理由相互冲突，就必须找到解决冲突的方法[9] 并推断出综合考量后的义务（步骤 2）。[10] 基于对伦理推理的这一描述，我们对文件夹进行组织，以探究其中涉及的不同推理模式：

（a）**从原则到表面理由（12 个案例）**。这个子文件夹中的案例都只有一个相关理由。这使得伦理推理的步骤 2 变得不言而喻，让我们可以专注于步骤 1。

(b) **削弱（3 种情况）**。回想一下患者拒绝治疗的例子，但现在假设患者因高烧而没有行为能力。在这种情况下，不应将患者的拒绝视为有效理由。借用论证学界的术语（Baroni 等人，2018年），我们称这个理由被削弱。本文件夹中的案例聚焦于削弱。

(c) **同一原则内的冲突（9个案例）**。试想，你认为唯一的道德原则就是仁慈。即便如此，仍可能出现理由冲突：有时你无法对所有人都施以仁慈，尽管理想情况下你应该如此（想想资源分配相关的伦理问题）。此文件夹中的案例探讨了这类冲突，并采用不同的解决方法（理由权衡、基于案例的推理、削弱）。

(d) **不同原则间的冲突（10个案例）**。患者拒绝治疗的示例，就是一个涉及指向不同原则的理由之间冲突的案例。在这个子文件夹中，我们将探讨这类冲突，并采用不同的解决方法。

(e) **案例研究：安乐死（13 个案例）**。此子文件夹中的案例列举了围绕安乐死实践所涉及伦理问题的各种可能情境。它们对资料库的贡献在于其复杂性：这些案例旨在近似呈现现实生活中的选择所具有的丰富性与模糊性。

[9]这里有必要说明一点：我们并未对这些案例应如何评判作出任何道德主张。相反，无论最终作出何种决定，我们关注的都是对该决定的解释。[10]尽管这一框架受到原则主义启发（Beauchamp 和 Chil- dress，1979年），但它具有很强的普适性：它可以表达任何一种伦理推理理论，只要该理论提出了一些一般原则、规则或职责，用于在具体选择情境中识别表面义务或理由。

```
SYSTEM: You are an expert in modal logic, working specifically within the KD modal logic.

---

### 🧩 Translate the natural language sentence into a logical form

Use the following KD modal logic notation:

- Negation:           \<^bold>\<not> p
- Conjunction:        p \<^bold>\<and> q
- Disjunction:        p \<^bold>\<or> q
- Implication:        p \<^bold>\<longrightarrow> q
- Obligation:         O p
- Prohibition:        F p  (defined as O(\<^bold>\<not> p))
- Permission:         P p  (defined as \<^bold>\<not> O(\<^bold>\<not> p))
- Validity:           \<lfloor>p\<rfloor>


#### ❗ Instructions for logical form:
1. Use propositional constants (e.g., `p`, `q`, `r`) for atomic actions.
2. **Reuse propositions** if they refer to the same action/concept across sentences.
3. **Ignore tense and auxiliary verbs** unless they encode deontic modality.
4. Use `\<^bold>\<and>` only when both parts are true; `\<^bold>\<or>` for disjunction.
5. Use implication for conditionals, definitionals, or explanatory links.
6. Label your logical forms with comments as in the example.


USER: Here are some formalisation examples:
###
Sentence: You must stop the car.
Defined Propositions:
  p: you stop the car
Logical Form:
  O p
```

(a) KD 语法提示。

```
begin
  typedecl i — ‹type for possible worlds ›
  type_synonym τ = "(i⇒bool)"
  consts r_t :: "i⇒i⇒bool" (infixr "rt" 70)(*relation for a modal logic KD*)

  abbreviation serial  where "serial  r ≡ (∀x.∃y. r x y)"

  axiomatization where ax_serial_rt : "serial  r_t"

  definition KDnot :: "τ⇒τ" ("¬_"[52]53) where "¬φ ≡ λw. ¬φ(w) "
  definition KDor :: "τ⇒τ⇒τ" (infixr "∨"50) where "φ∨ψ ≡ λw. φ(w) ∨ ψ(w)"
  definition KDand :: "τ⇒τ⇒τ" (infixr "∧"51) where "φ∧ψ ≡ λw. φ(w) ∧ ψ(w)"
  definition KDimp :: "τ⇒τ⇒τ" (infixr "⟶" 49) where "φ⟶ψ ≡ λw. φ(w) ⟶ ψ(w)"
  definition KDtrue  :: "τ" ("⊤") where "⊤ ≡ λw. True"
  definition KDfalse :: "τ" ("⊥") where "⊥ ≡ λw. False"

  definition KDobligatory :: "τ⇒τ" ("O") where "O φ ≡ λw. ∀v. w rt v ⟶ φ(v)"
  definition KDforbidden :: "τ⇒τ" ("F") where "F φ ≡ O(¬φ)"
  definition KDpermitted :: "τ⇒τ" ("P") where "P φ ≡ ¬(O(¬φ))"
  definition KDvalid :: "τ⇒bool" ("⌊_⌋" [8]109)  where "⌊p⌋ ≡ ∀w. p(w)"
  lemma True nitpick [satisfy,user_axioms,expect=genuine,show_all,format=2] oops

  consts
 (* Declare each atomic proposition used in axioms or theorems.
All constants must be of type \<tau> = "i ⇒ bool" *)

(* Explanation 1: [provided sentence 1 in natural language] *)
axiomatization where
  explanation_1: [Transfer the logical form into isabelle code here,
non-bracketed of the predicate-argument form]

(* Explanation 2: [provided sentence 2 (if any) in natural language]  *)
axiomatization where
  explanation_2: [Transfer the logical form into isabelle code here,
 non-bracketed of the predicate-argument form]
end
```

(b) KD Isabelle 公理提示（包含 HOL 中的 KD 理论片段）。

图 8：用于知识蒸馏的提示：语法检查提示（上）和 Isabelle 公理提示（下），后者将高阶逻辑中的理论嵌入其中。

逻辑	DeepSeek (%)	ChatGPT(%)
FOL	42.72	39.81
KD	76.70	77.67
DDLE	6019.0	51.46
DDL_CJ	61.17	74.76

表 3：有效解释生成的成功率。

逻辑	DeepSeek	ChatGPT
FOL	0.91	0.83
KD	0.43	0.60
DDLE	0.60	0.64
DDL_CJ	0.65	0.71

表 4：达到有效解释所需的平均细化迭代次数。

逻辑	DeepSeek（秒）	ChatGPT（秒）
FOL	108.99	103.74
KD	65.25	58.83
DDLE	69.45	65.34
DDL_CJ	60.18	49.80

表5：成功的解释细化的平均求解时间（秒）。

逻辑	DeepSeek（%）	ChatGPT（%）
FOL	3.90	2.90
KD	7.80	1.90
DDLE	16.50	12.60
DDL_CJ	19.40	5.80

表6：语法错误率。ChatGPT-KD 的错误率最低（低于 2%），展现出卓越的鲁棒性。

领域	逻辑	模型	有效率(%)	平均迭代时间（秒）	
经典逻辑（5）	FOL	DeepSeek	100.0	0.20	64.83
	FOL	ChatGPT	80.0	0.25	46.54
	KD	DeepSeek	100.0	0.00	22.09
	KD	ChatGPT	80.0	0.75	41.64
	DDLE	DeepSeek	100.0	0.00	24.65
	DDLE	ChatGPT	80.0	0.75	52.49
	DDL_CJ	DeepSeek	80.0	0.50	11.03
	DDL_CJ	ChatGPT	60.0	0.67	26.11
常识意义（10）	FOL	DeepSeek	90.0	1.33	83.53
	FOL	ChatGPT	80.0	0.75	55.60
	KD	DeepSeek	70.0	0.57	66.44
	KD	ChatGPT	50.0	1.00	81.30
	DDLE	DeepSeek	70.0	0.28	47.36
	DDLE	ChatGPT	60.0	0.50	65.49
	DDL_CJ	DeepSeek	40.0	0.85	59.33
	DDL_CJ	ChatGPT	50.0	0.80	63.05
默认推理（17）	FOL	DeepSeek	52.94	1.11	105.21
	FOL	ChatGPT	70.59	1.17	122.81
	KD	DeepSeek	82.35	0.50	58.39
	KD	ChatGPT	76.74	0.62	65.96
	DDLE	DeepSeek	58.82	0.40	68.61
	DDLE	ChatGPT	64.71	0.55	52.51
	DDL_CJ	DeepSeek	52.94	1.11	78.73
	DDL_CJ	ChatGPT	88.24	0.93	46.49
模态（24）	FOL	DeepSeek	58.33	0.86	89.41
	FOL	ChatGPT	45.83	0.54	70.10
	KD	DeepSeek	87.50	0.00	47.45
	KD	ChatGPT	91.67	0.32	53.01
	DDLE	DeepSeek	62.50	0.73	36.56
	DDLE	ChatGPT	58.33	0.71	55.45
	DDL_CJ	DeepSeek	62.50	0.53	38.36
	DDL_CJ	ChatGPT	62.50	0.40	43.46
（生物）伦理学（47）	FOL	DeepSeek	14.89	0.71	130.01
	FOL	ChatGPT	12.77	1.17	129.46
	KD	DeepSeek	68.09	0.72	82.62
	KD	ChatGPT	76.60	0.69	56.91
	DDLE	DeepSeek	53.19	0.80	94.05
	DDLE	ChatGPT	38.30	0.67	76.75
	DDL_CJ	DeepSeek	65.96	0.61	69.05
	DDL_CJ	ChatGPT	82.98	0.74	53.17

表7：DeepSeek 和 ChatGPT 模型在各逻辑框架中的总体解释细化性能。求解时间仅对成功运行取平均值。