VeriMinder: 缓解 NL2SQL 中的分析脆弱性

Shubham Mohole
Cornell University
sam588@cornell.edu

Sainyam Galhotra
Cornell University
sg@cs.cornell.edu

Abstract

使用自然语言界面的数据库应用系统（NLIDBs）已经使数据分析大众化。虽然这一积极发展带来了便利，但也提出了一个迫切的挑战，即帮助那些可能没有统计分析背景的用户制定无偏见的分析性问题。尽管大量研究集中在文本到 SQL 的生成准确性上，但对于分析性问题中认知偏见的研究仍然不足。我们提出了 VeriMinder，[1]，一个用于检测和缓解此类分析脆弱性的交互系统。我们的方法引入了三个关键创新：(1) 针对特定分析情境中相关偏见的语境语义映射框架 (2) 实现了“难以变化”原则并指导用户进行系统化数据分析的分析框架 (3) 一个优化的基于 LLM 的系统，使用包含多个候选者、评论反馈和自我反思的结构化流程来生成高质量、针对特定任务的提示。用户测试证实了我们方法的优点。在直接用户体验评估中，82.5 % 的参与者报告称对分析质量产生了积极影响。在对比评估中，VeriMinder 的得分明显高于其他方法，在分析的具体性、全面性和准确性的指标上至少高出 20 %。我们的系统被实现为一个网络应用，旨在帮助用户在数据分析中避免“错误问题”的弱点。VeriMinder 的代码库及提示 [2] 作为 MIT 许可的开源软件可用，以促进进一步的研究和社区内的采用。

1 介绍

自然语言到 SQL（NL2SQL）系统作为一种关键技术出现，使得非技术用户能够在没有专业 SQL 知识的情况下查询复杂数据库，从而实现数据访问民主化。然而，这一积极发展并非没有显著风险。从一个根本上存在缺陷的分析问题中派生出的技术上完美的 SQL 查询将导致误导性结果。像 SQLPalm (Sun et al., 2023)、SPLASH (Elgohary et al., 2020) 和 DAIL-SQL (Gao et al., 2023) 这样的系统专注于 NL2SQL 的准确性，但并未考虑用户原始问题的分析质量。

[1] https://veriminder.ai
[2] https://reproducibility.link/veriminder

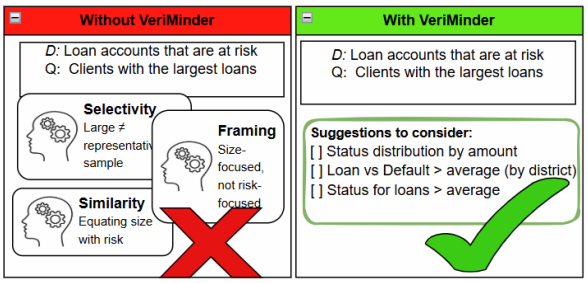

Figure 1: 实验数据集中的示例，展示了 VeriMinder 通过改进建议来减轻偏见

考虑图 1 中所示的这个例子：一位金融分析师的任务是识别“有风险的贷款账户”，但他却询问“贷款金额最大的客户”。这个查询表现出多种认知偏差：（1）相似性偏差——错误地假设“最大贷款”和“有风险贷款”是相似的类别，（2）框架偏差——将问题框架围绕在贷款金额而非风险因素上，从而完全改变了将会检索到的信息，（3）选择偏差——仅关注大额贷款选择了一个不具代表性的可能有风险账户的子集，因为小额贷款可能有更高的违约率。尽管最先进的 NL2SQL 系统可以为原始问题生成语法上正确的 SQL，但它无法解决这些分析盲点，从而留下一个未解决的重要漏洞。

研究表明，认知偏见显著影响医学、法律等领域的专业决策 (Berthet, 2022)。这些偏见，诸如锚定效应和可用性偏见，与不良结果（如健康诊断不准确）的一致关联，强调了需要像 VeriMinder 这样的缓解系统的重要性。正如彼得·德鲁克所说：“最严重的错误不是由于错误的答案造成的。真正危险的是提出了错误的问题。” (Drucker, 1971)。

传统的缓解此类问题的方法依赖于静态清单 (Lenders and Calders, 2025) 或教育干预措施 (Thompson et al., 2023)，但在实施时具有一致性挑战。虽然 FISQL (Menon et al., 2025) 和 SPLASH (Elgohary et al., 2020) 提供有限的反馈机制，它们主要关注 SQL 的改进而不是解决分析质量问题 (Qu et al., 2024)。为了应对这些挑战，我们推出了 VeriMinder，它可以在

NL2SQL 工作流中识别和缓解分析漏洞。我们的交互式网络应用通过三项创新来解决这些漏洞：(1) 一个系统检测分析问题中偏见和盲点的语义框架；(2) 一个基于“难以改变”原则的结构化分析过程 (Deutsch, 2011)；以及 (3) 一个与 NL2SQL 工作流集成的优化的 LLM 驱动优化界面。VeriMinder 通过简单的配置无缝集成到现有的 NL2SQL 系统中，为这些系统的用户提供支持，通过准确的 SQL 生成并制定稳健的分析问题。我们的评估显示，VeriMinder 显著增强了分析结果，在关键分析指标上优于基准方法。

2 系统架构

VeriMinder 根据 Deutsch 的“难以更改”原则 (Deutsch, 2011) 通过一个系统化的架构来识别和缓解用户问题 (Q) 中的分析漏洞，将潜在偏见的查询转化为稳健的分析解释 (E)，在给定的领域 (D) 和决策背景 (C) 下。该原则认为，好的解释是受约束的，以至于改变其组成部分会削弱解释或导致不一致。应用于数据分析中，一个稳健的解释 E ，通常通过 SQL 查询 (S) 实现，如果其组成部分在上下文 C 中必然和连贯地解决 Q ，且没有移除不会降低质量的任意元素，则难以更改。相反，容易更改的解释允许不具特定角色的可替换组成部分，可能导致因有缺陷的问题而出现误导性结果（例如，在决定政府削减开支措施时分析广泛的支出类别而非特定成本因素）。VeriMinder 强制执行这一原则，确保分析指出特定因素，从而提供数据支持的、可证伪的解释，抵抗变化。

2.1 核心模块和架构

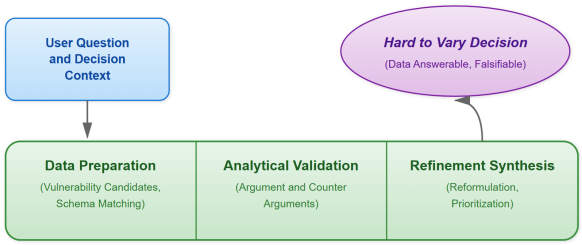

Figure 2: 三阶段框架落实难变原则。

VeriMinder 系统实施了一种系统的方法，帮助分析人员将脆弱的问题细化为稳健的数据分析，以使难以改变原则具体化。如图 2 所示，我们的架构通过三个连续阶段处理自然语言问题：数据准备、分析验证和精炼综合。

在数据准备阶段，系统分析问题和决策背景，以识别潜在的分析漏洞和相关的模式元素。在分析验证过程中，漏洞被检测出来，并且通过使用论证组件和反论证测试进行结构分析，以验证其重要性。在精炼合成阶段，系统生成有针对性的精炼建议，以帮助分析与特定决策背景下的数据支持解释的难以改变的方法保持一致。

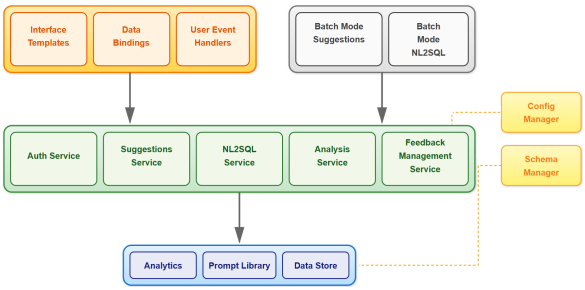

Figure 3: 模块化架构支持可扩展性和灵活的部署模式

VeriMinder 通过模块化服务架构（图 3）实现该框架，以便灵活性，具有五个核心服务通过标准化接口进行通信：Auth（用户提供/访问，未来的企业插件），Suggestion（实施核心框架分析），NL2SQL（将 (Qu et al., 2024) 的方法扩展到带有元数据和数据集特定分布信息，并使用 Gemini Flash 2.0 (Google DeepMind, 2025)），分析（比较初始与精炼结果以便用户反思），和用户反馈（收集改进数据）。底层分析框架组件（详见附录 A.1）包括 53 种分类的认知偏差（例如，记忆、统计、框架），数据模式（临时的、分类的、数值的，详见附录 A.2），用于论证结构评估的 Toulmin 模型 (Toulmin, 1958)（附录 A.3），以及用于帮助解决挑战和完善解释的问题的反驳框架 (Greitemeyer, 2023)（附录 A.4）。

在我们的系统实现中，我们开发了一个实验性的 NL2SQL 组件，该组件基于 LLM 文本到 SQL 生成的最佳实践 (Qu et al., 2024; Sun et al., 2023; Gao et al., 2023)。VeriMinder 旨在补充现有的 NL2SQL 系统，而不是替代它们，专注于分析性问题的提出这一正交问题。

2.2 提示制定方法

VeriMinder 通过一个三阶段的工作流程（图 4）为用户的自由形式分析问题提供偏差缓解的替代方案。该流程由一个形式定义的难以改变的目标驱动，但通过实用的近似实现，以尊重 LLM 限制和推理延迟。

VeriMinder 的架构遵循一个核心原则：一个稳健的分析问题应该在最大化关于决策的预测洞察的同时，最小化其自身的描述复杂性，并受限于交互延迟预算。本节概述了激励我们系统设计的理想理论框架（§2.2.1），并详细介绍

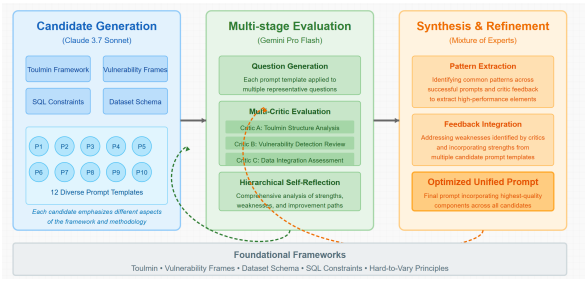

Figure 4: 具有评论反馈和自我反思的多候选提示工程流水线。

了其转化为一个实用的多阶段 LLM 流程（§?? - ??），最后讨论其范围和局限性（§??）。

2.2.1 理想化的理论动机

我们使用难变性 (HV) 分数形式化稳健探究的原理，这一指标受 Deutsch 的良好解释概念 (Deutsch, 2011) 和最小描述长度 (MDL) 原理 (Rissanen, 1978; Grünwald, 2007) 的启发。对于一组选定的分析变量 S 和一个决策目标 T，HV 分数为：

$$HV(S) = \frac{I(T;S)}{DL(S)} \tag{1}$$

这里，$I(T;S)$ 是互信息 (Cover and Thomas, 2006)，而 $DL(S)$ 是模型的描述长度。此公式扩展了如信息增益率 (Quinlan, 1993) 等归一化信息指标，奖励解释密度（高信息每单位复杂度），并呼应信息瓶颈理论 (Tishby et al., 2000) 的目标。

为了验证这一度量的行为，我们开发了一个数值验证套件。如我们的代码库详细所述，对合成贝叶斯网络的实验展示了 HV 分数在理想条件下的关键特性。所有模拟使用了精确的互信息计算，并将复杂性定义为变量集合的基数，即 $DL(S) = |S|$。这为 HV 分数是一个合理的理论目标提供了经验支持。

直接优化方程 1 即使在结构化特征空间 (Nguyen et al., 2014) 中也是计算上不可行的，并且在开放性自然语言领域中复杂性呈指数增长，因为搜索空间包括所有可能的问题构拟。VeriMinder 因此采用基于 LLM 的启发式代理，由 HV 公式的直觉指导。我们认识到这不是正式的等价；HV 分值的理想特性仅在正式定义下准确保持，而我们的代理旨在通过实证来近似它们。

为了探索分析空间，系统使用十二个提示模板来生成一组多样化的候选者。这些模板本身是基于 Claude 3.7 Sonnet 模型的一个自动化元层提示工程过程的输出，选择该模型是因为其在智能类别中的排名，确保每个目标指向一个独特的分析角度（例如，漏洞检测、模式验证）。这种集成方法确保了广泛的覆盖，这是在机器学习中兼具装袋法和现代大语言模型提示的一种技术。

2.2.2 阶段 2: 分布式评论评价

生成的候选项由一个由三名专业 LLM 评论家组成的小组进行评估（基于 Claude 3.7 Sonnet 模型）。为了提高效率，两名评论家的随机子集评估每个候选项。这实现了类似于提升算法的分布式评估，其中一个弱学习者委员会形成稳健的判断 (Schapire, 1990)。这与现代方法一致，使用自一致性和多代理共识来改进 LLM 评估 (Wang et al., 2023; Li et al., 2024b)。

最后，系统执行一个自我反思过程，通过评论反馈来改进提示。这类似于通过自我完善技术来提高 LLM 性能。目前我们只执行了一次迭代，但多次自我反思循环将是现有流程的一个可能自然扩展。

我们的方法有三个主要限制。首先，我们的生产系统依赖于启发式搜索，与在我们验证套件中的穷举搜索不同。其次，评论评分和我们的分析流程阶段是务实的替代品，而不是 $I(T;S)$ 和 $DL(S)$ 的正式等价物。最后，我们当前的成本模型仅限于响应结构，尚未纳入计算延迟。

2.3 交互式用户界面

VeriMinder 的用户界面（图 5）采用渐进披露模式来引导工作流程：用户提供他们的问题和背景，系统执行查询并分析漏洞，建议细化以供用户选择，呈现结果的并排比较，并解释检测到的问题和修复。为了在密集计算期间增强用户体验，服务器发送的事件（SSE）提供流式更新和教育性见解。该系统具有可插拔接口和统一抽象层，以支持多种数据库类型，利用 SQLite（通过 BIRD-DEV 基准 (Li et al., 2023)）进行执行，并使用 MySQL 跟踪应用程序状态。

3 实验

3.1 实验装置

为了全面评估 VeriMinder，我们设计了一个多步骤评估框架，针对关键研究问题：(1) VeriMinder 解决方案在使用 NL2SQL 界面改进分析方面有多大效果？(2) 我们的方法如何与其他方法相比，以增强在关键准确性、具体性和全面性指标上的分析质量 (Zhu et al., 2024b)？

评估数据集来源于 BIRD-DEV 基准问题。为了创建真实的决策环境，我们采用案例研究法 (Ellet, 2007) 手动设计了 164 个决策场景，确保选择、评估和诊断类型的平衡覆盖。数据分析

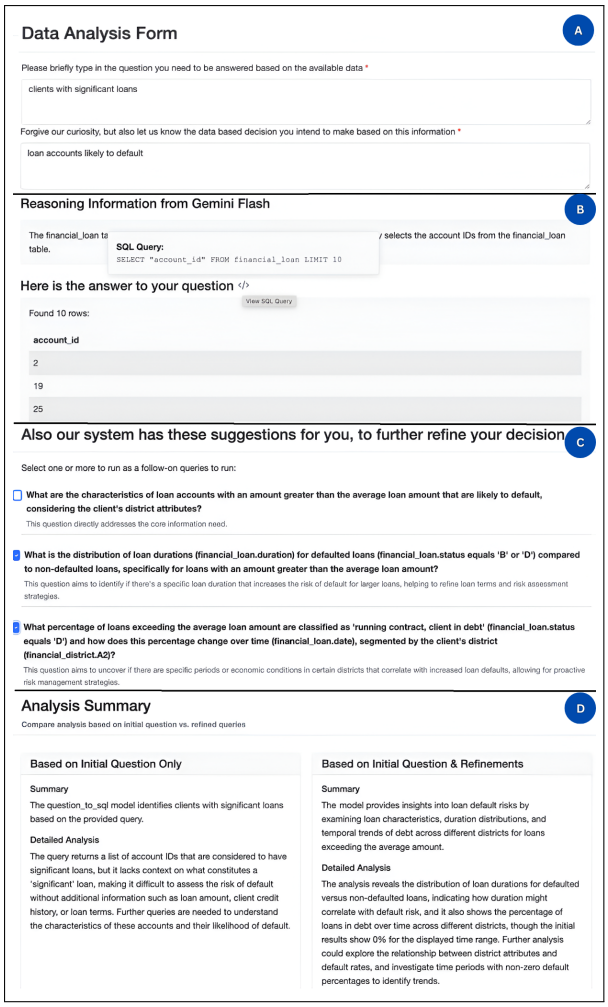

Figure 5: VeriMinder 的用户界面工作流程：（A）初始问题，（B）查询结果，（C）优化建议，（D）比较分析

专家设计了这些场景来表示分析脆弱性可能显著影响结果的环境。我们使用 TF-IDF 向量化方法将每个决策与 BIRD-DEV 中语义最相关的问题进行匹配，形成二分关系。最终的决策文本经过轻微编辑以确保在用户研究期间的语法和句子结构一致，而不改变决策环境的分析重点。这种系统的方法产生了 164 个问题-决策对，分为三个子集：64 对（DS1）用于人工评估，100 对（DS2）用于自动评估。另一个由 DS1 创建的较小子集 DS1-T1 包含 36 对。所有划分均为随机进行。

据我们所知，没有直接可比的系统专注于优化用户提出的问题以及解决偏见和盲点。因此，除了 Direct NL2SQL（标准的文本到 SQL 生成，没有分析增强功能）之外，我们通过使研究界考虑用于偏差减缓或整体分析的三种替代方法来评估 VeriMinder：决策为焦点的查询生成（直接从决策上下文生成问题（Zhang et al., 2025）），问题扰动（PerQS）（创造原始问题的变体（Zhu et al., 2024a）），以及批评代理反馈（CAF）（实现批评代理提供反馈（Li et al., 2024a））。我们在所有基准测试中使用相同的 LLM（Gemini Flash 2.0）与 VeriMinder，并计划在代码发布的一部分中发布它们。

我们评估方法的一个关键方面是确保所有比较系统之间的一致 SQL 生成。为了隔离分析性问题表述（我们关注的重点）对 NL2SQL 准确性的影响，我们为所有基线系统和 VeriMinder 实施了相同的实验性 NL2SQL 组件。在我们的评估中，我们验证了所有生成的 SQL 查询在评估前都能正确执行，使我们能够专注于分析质量而非技术上的 SQL 正确性。

我们使用 DS1-T1 数据集进行了一个互动用户研究，从 Prolific 招募了 63 名背景各异的参与者。对于 30 个场景，每个场景我们收到了来自两个用户的提交，另外三个场景则收到了一个用户的提交（共 63 名独特参与者）。附录 B1 展示了提供给参与者的反馈表。我们方案在提高分析质量方面的整体效果获得了 82.5% 的正面评价（评分为 4 或 5），Gwet's AC1 为 0.766。建议的有效性获得了 74.6% 的正面评价，Gwet's AC1 为 0.670。理由的清晰度获得了 66.7% 的正面评价（Gwet's AC1 为 0.479），场景的真实度获得了 61.9% 的正面评价（Gwet's AC1 为 0.457）。这些可靠性得分，特别是在清晰度和真实性上，可能反映了来自 Prolific 的多样化用户群体的影响。此外，场景真实性得分可能受到实验设置的影响，其中决策背景受限于将其与现有的 BIRD-DEV 数据集问题相匹配。

从 DS1 数据集中，我们对生成的分析问题进行了比较评估，参与评估的包括两家总部位于美国的软件公司各一名数据分析员，他们对我们的请求做出了回应。附录 B.2 展示了这些数据分析员用户用来评估在决策情境中分析问题比较强度的界面的截图。与之前的测试一样，我们的分析中只包括成功完成的部分（由于一个无关系统故障问题，我们未能获取五个条目的提交）。对于 59 个场景，我们从两位用户那里收到了提交。VeriMinder 在所有维度上均表现出色：准确性（平均 =7.87/10，95% CI [7.57, 8.18]）、具体性（平均 =7.79/10，95% CI [7.47, 8.10]）和全面性（平均 =8.05/10，95% CI [7.74, 8.36]）。

图 6 展示了 VeriMinder 相对于每个基线系统的百分比改进。相对于 Direct NL2SQL，观察到最显著的改进，准确性提高了 60.4 %，具体性提高了 63.2 %，全面性提高了 86.9 %。即便是与最强基线系统（问题扰动）相比，VeriMinder 也显示出准确性提高了 22.1 %，具体性提高了 28.4 %，全面性提高了 21.2 %。

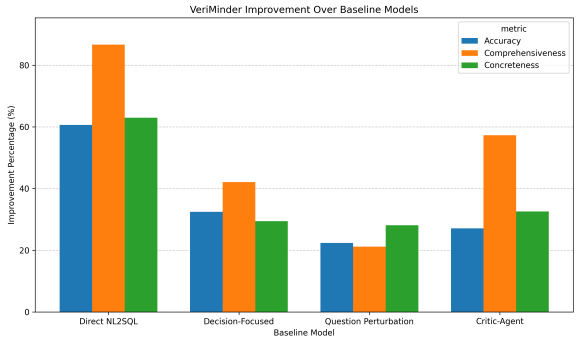

Figure 6: VeriMinder 在关键分析维度上相对于基线系统的百分比提升

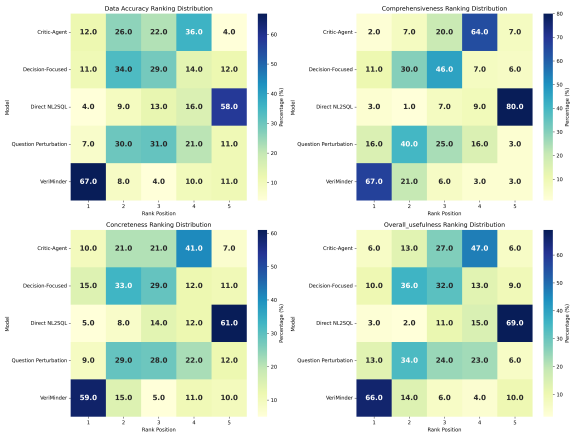

Figure 7: 在分析维度上的排名分布；VeriMinder 始终获得最高排名

统计分析证实了这些改进是显著的（$p < 0.001$），在所有维度和基线比较中进行了配对 t 检验。胜率进一步说明了 VeriMinder 的质量，分别在 83.9 % 的准确性比较、86.4 % 的具体性比较和 97.5 % 的全面性比较中优于 Direct NL2SQL。基于模型等级的评分者间信度指标在我们的评估中展示了强一致性，其中 Gwet 的 AC1 系数为 0.941（准确性）、0.960（具体性）和 0.862（全面性）。

我们使用基于 LLM 的评估器对数据集 DS2（100 个场景）（Gemini Flash 2.0）进行了评估。考虑到 LLM 在定量评分方面的已知限制 (OpenAI et al., 2024; Bubeck et al., 2023) 但在文本分析和相对排名方面表现更佳 (Zheng et al., 2023; Gilardi et al., 2023)，我们的测试重点在于 LLM 在文本理解和比较定性评估方面的技能。在附录 B.3 中，我们讨论了提示设计的方法。对于基于 LLM 的评估，我们首先将自动评估器（基于 Gemini 2.0 Flash）与 DS1 的 15 个示例子集中人类判断的比较排名进行了校准，找到了一个 m（Pearson's $r = 0.74, p < 0.001$），这使我们对自动化结果有了信心。

如图 7 所示，VeriMinder 在各个指标中始终获得最高的第一名排名：数据准确性为 67.0 %，全面性为 67.0 %，具体性为 59.0 %，整体有用性为 66.0 %。相反，Direct NL2SQL 在所有指标中收到最多的最后一名排名，这突显了超越原始 SQL 生成的分析增强的重要性。

3.2 偏见缓解效果分析

Figure 8: 关键分析能力推动 VeriMinder 中的认知偏差缓解

图 8 中的词云可视化突出了通过对 LLM 响应的定性分析识别出的 VeriMinder 的关键分析能力。该可视化是通过对整个数据集的改进建议进行自动化内容分析生成的。如图 8 所示，比分析、模式识别和关系探索作为关键能力出现，使 VeriMinder 能够减轻认知偏差。

3.3 局限性

几个限制需要注意。首先，在特定领域的应用可能需要对分析组件进行定制。其次，系统的有效性取决于底层 NL2SQL 引擎的质量，这里实现为一个简化的服务模块。我们主要在 BIRD-DEV 上评估 VeriMinder，该数据可能在大型语言模型的训练中已经被看到，这引发了关于信息泄漏和在真正未见过的数据库上过高估计 SQL 成功率的担忧。界面经过桌面优化，但未进行可访问性测试。在全面发布之前，关键改进包括移动支持、可访问性功能、多查询处理以及在以前未见过的数据库上进行验证以确认泛化能力。

4 相关工作

我们的工作建立在认知偏差缓解、自然语言数据库接口和非真实情况中的大型语言模型推理技术的研究之上 - 这一分析上下文中没有单一的“正确”答案，而是基于全面性、准确性和与决策目标的一致性而具有不同程度的分析质量。之前在认知偏差缓解方面的研究已检查了数据驱动情境中的偏差 (Kahneman, 2011; Tversky and Kahneman, 1974; Sumita et al., 2024; Ke et al., 2024)，但主要集中在偏差意

识上，而不是在分析工作流程中进行积极的缓解。Spider 2 (Lei et al., 2025) 等基准推动了 NL2SQL 生成的最近进展 (Deng et al., 2025; Wang and Liu, 2025)，基于 LLM 的系统实现了高执行精度。然而，这些系统主要解决技术性的 SQL 问题，而非分析脆弱性。

虽然 VeriMinder 主要关注分析性问题的制定，我们的评估采用了一个简化的 NL2SQL 服务。该服务为我们的设置中 SQL 生成引入了元数据和特定数据集的分布信息，灵感来自于近期关于缓解 NL2SQL 幻觉的工作，例如 (Qu et al., 2024) 提出的任务对齐策略和通过 (Mohole and Galhotra, 2025) 增强数据集的列统计的 LLM 基于表格学习任务。包括响应选择在内的 LLM 提示技术 (Zhao et al., 2025)，提升了推理能力，但在需要互动体验的非真实情况中可能不适用。通过 Deutsch 的框架 (Deutsch, 2011) 启发和多候选的精化过程，我们提供了一个轻量但系统化的框架，以优化 LLM 对下游 NL2SQL 和分析任务的响应。

5 未来工作和结论

虽然 VeriMinder 目前针对 NL2SQL 交互，但其分析核心是模态无关的，这使得将来可以扩展到 Python/pandas 代码生成以进行统计探索。基于 Self-RAG (Asai et al., 2023)，我们计划将自反思阶段发展为多头的、偏见感知的量规，输出校准的概率以评估证据充足性、认知偏见标志和统计有效性。这些概率不仅将引导自适应检索器-生成器循环，还将作为 Conformal LM (Quach et al., 2024) 的偏见感知非一致性评分，允许设定拒绝阈值，在减少偏见的同时保持覆盖率。我们的信息理论框架自然扩展到这一校准重点——通过最大化 HV(S) 在反思头输出的表现，信息理论引导的修剪可以保证最小因果充分性，同时保持校准的简洁性。

通过 VeriMinder，我们展示了一个从头到尾的系统，用于缓解自然语言查询中的分析漏洞。通过将“难以变更”的解释付诸实践，我们展示了其对 NL2SQL 用例的有效性。结合 SELF-RAG 原则和偏差感知的保形预测，这项研究可以为自然语言接口数据库打开新的途径，这些数据库不仅提供可能正确的答案，而且无偏见，并以证据为基础。

6 广泛影响声明

虽然 VeriMinder 解决了分析漏洞，但关键的限制和伦理问题依然存在：

该系统提供指导，但不提供保证，旨在增强用户的批判性思维，而不是取代它。漏洞检测可能不够全面。

商业 API 依赖 对商业大型语言模型的依赖限制了可访问性；未来的工作应该探索开源的替代方案。

文化和领域偏见 偏见分类法主要基于西方，并可能需要进行特定领域或文化的适应。

滥用潜力 分析增强工具可能被滥用；需要治理框架来确保完整性。

增强与自动化 VeriMinder 增强了人类分析，保持用户主动性而不是完全自动化该过程。

我们认为，随着数据访问的民主化，解决分析漏洞是至关重要的。VeriMinder 是一个初步步骤，旨在激发在认知科学、数据分析和自然语言处理交叉领域的进一步研究。

References

Artificial Analysis. 2025. Ai model & api providers analysis. https://artificialanalysis.ai. Accessed: June 10, 2025.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *Preprint*, arXiv:2310.11511.

Vincent Berthet. 2022. The impact of cognitive biases on professionals' decision-making: A review of four occupational areas. *Frontiers in Psychology*, 12:802439.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *Preprint*, arXiv:2303.12712.

Jean-Paul Caverni, Jean-Marc Fabre, and Michel Gonzalez. 1990. *Cognitive biases*. Advances in psychology. North Holland.

Thomas M. Cover and Joy A. Thomas. 2006. *Elements of information theory*, 2nd edition. Wiley-Interscience.

Minghang Deng, Ashwin Ramachandran, Canwen Xu, Lanxiang Hu, Zhewei Yao, Anupam Datta, and Hao Zhang. 2025. Reforce: A text-to-sql agent with self-refinement, format restriction, and column exploration. *Preprint*, arXiv:2502.00675.

David Deutsch. 2011. *The Beginning of Infinity: Explanations that Transform the World*. Penguin UK.

Evanthia Dimara, Steven Franconeri, Catherine Plaisant, Anastasia Bezerianos, and Pierre Dragicevic. 2020. A task-based taxonomy of cognitive biases for information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 26(2):1413–1432.

Peter F. Drucker. 1971. *Men, Ideas, and Politics*. Harper & Row, New York.

Joyce Ehrlinger, Wilson Readinger, and Bora Kim. 2016. Decision-making and cognitive biases. In Howard S. Friedman, editor, *Encyclopedia of mental health*, 2 edition, pages 5–12. Academic Press, Oxford.

Ahmed Elgohary, Saghar Hosseini, and Ahmed Hassan Awadallah. 2020. Speak to your parser: Interactive text-to-SQL with natural language feedback. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2065–2077, Online. Association for Computational Linguistics.

William Ellet. 2007. *The Case Study Handbook: How to Read, Discuss, and Write Persuasively About Cases*. Harvard Business Review Press, Boston, Massachusetts. Accessed: March 26, 2025.

Dawei Gao, Haibin Wang, Yaliang Li, Xiuyu Sun, Yichen Qian, Bolin Ding, and Jingren Zhou. 2023. Text-to-sql empowered by large language models: A benchmark evaluation. *Preprint*, arXiv:2308.15363.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30).

Google DeepMind. 2025. Gemini: A family of highly capable multimodal models. Accessed: 2025-03-26.

Tobias Greitemeyer. 2023. Counter explanation and consider the opposite: Do corrective strategies reduce biased assimilation and attitude polarization in the context of the COVID-19 pandemic? *Journal of Applied Social Psychology*, 53(5):306–322.

Peter D. Grünwald. 2007. *The Minimum Description Length Principle*. MIT Press.

Martin Hilbert. 2012. Toward a synthesis of cognitive biases: How noisy information processing can bias human decision making. *Psychology Bulletin*, 138(2):211–237.

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Yuhe Ke, Rui Yang, Sui An Lie, Taylor Xin Yi Lim, Yilin Ning, Irene Li, Hairil Rizal Abdullah, Daniel Shu Wei Ting, and Nan Liu. 2024. Mitigating cognitive biases in clinical decision-making through multi-agent conversations using large language models: Simulation study. *Journal of Medical Internet Research*, 26:e59439.

Fangyu Lei, Jixuan Chen, Yuxiao Ye, Ruisheng Cao, Dongchan Shin, Hongjin Su, Zhaoqing Suo, Hongcheng Gao, Wenjing Hu, Pengcheng Yin, Victor Zhong, Caiming Xiong, Ruoxi Sun, Qian Liu, Sida Wang, and Tao Yu. 2025. Spider 2.0: Evaluating language models on real-world enterprise text-to-sql workflows. *Preprint*, arXiv:2411.07763.

Daphne Lenders and Toon Calders. 2025. Users' needs in interactive bias auditing tools introducing a requirement checklist and evaluating existing tools. *AI Ethics*, 5:341–369.

Jinyang Li, Binyuan Hui, Ge Qu, Jiaxi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Rongyu Cao, Ruiying Geng, Nan Huo, Xuanhe Zhou, Chenhao Ma, Guoliang Li, Kevin C. C. Chang, Fei Huang, Reynold Cheng, and Yongbin Li. 2023. Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls. *Preprint*, arXiv:2305.03111.

Michael Y. Li, Vivek Vajipey, Noah D. Goodman, and Emily B. Fox. 2024a. Critical: Critic automation with language models. *Preprint*, arXiv:2411.06590.

Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. 2024b. Improving multi-agent debate with sparse communication topology. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Bangkok, Thailand. Association for Computational Linguistics. ArXiv:2406.11776.

Rakesh R. Menon, Kun Qian, Liqun Chen, Ishika Joshi, Daniel Pandyan, Jordyn Harrison, Shashank Srivastava, and Yunyao Li. 2025. Fisql: Enhancing text-to-sql systems with rich interactive feedback. In *Proceedings of the 2025 International Conference on Extending Database Technology (EDBT)*, pages 1032–1038.

Shubham Mohole and Sainyam Galhotra. 2025. Sifotl: A principled, statistically-informed fidelity-optimization method for tabular learning. KDD'25 (UMC).

Xuan Vinh Nguyen, Jeffrey Chan, Simone Romano, and James Bailey. 2014. Effective global approaches for mutual information based feature selection. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14)*, pages 512–521, New York, NY, USA. Association for Computing Machinery.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Eoin D. O'Sullivan and Susie J. Schofield. 2019. A cognitive forcing tool to mitigate cognitive bias –a randomised control trial. *BMC Medical Education*, 19(12).

Ge Qu, Jinyang Li, Bowen Li, Bowen Qin, Nan Huo, Chenhao Ma, and Reynold Cheng. 2024. Before generation, align it! a novel and effective strategy for mitigating hallucinations in text-to-sql generation. *Preprint*, arXiv:2405.15307.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. 2024. Conformal language modeling. *Preprint*, arXiv:2306.10193.

J. Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Jorma Rissanen. 1978. Modeling by shortest data description. *Automatica*, 14(5):465–471.

Robert E. Schapire. 1990. The strength of weak learnability. *Machine learning*, 5(2):197–227.

Michael Soprano, Kevin Roitero, David La Barbera, Davide Ceolin, Damiano Spina, Gianluca Demartini, and Stefano Mizzaro. 2024. Cognitive biases in fact-checking and their countermeasures: A review. *Information Processing & Management*, 61(3):103672.

Yasuaki Sumita, Koh Takeuchi, and Hisashi Kashima. 2024. Cognitive biases in large language models: A survey and mitigation experiments. *Preprint*, arXiv:2412.00323.

Ruoxi Sun, Sercan Ö. Arik, Alex Muzio, Lesly Miculicich, Satya Gundabathula, Pengcheng Yin, Hanjun Dai, Hootan Nakhost, Rajarishi Sinha, Zifeng Wang, and Tomas Pfister. 2023. Sql-palm: Improved large language model adaptation for text-to-sql (extended). *arXiv preprint arXiv:2306.00739*.

John Thompson, Helena Bujalka, Stephen McKeever, Adrienne Lipscomb, Sonya Moore, Nicole Hill, Sharon Kinney, Kwang Meng Cham, Joanne Martin, Patrick Bowers, and Marie Gerdtz. 2023. Educational strategies in the health professions to mitigate cognitive and implicit bias impact on decision making: a scoping review. *BMC Medical Education*, 23(455).

Naftali Tishby, Fernando C. Pereira, and William Bialek. 2000. The information bottleneck method. *Preprint*, arXiv:physics/0004057.

Stephen E. Toulmin. 1958. *The Uses of Argument*. Cambridge University Press.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Yihan Wang and Peiyu Liu. 2025. Linkalign: Scalable schema linking for real-world large-scale multi-database text-to-sql. *Preprint*, arXiv:2503.18596.

Yueheng Zhang, Xiaoyuan Liu, Yiyou Sun, Atheer Alharbi, Hend Alzahrani, Basel Alomair, and Dawn Song. 2025. Can llms design good questions based on context? *Preprint*, arXiv:2501.03491.

Eric Zhao, Pranjal Awasthi, and Sreenivas Gollapudi. 2025. Sample, scrutinize and scale: Effective inference-time search by scaling verification. *Preprint*, arXiv:2502.01839.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Gong, and Xing Xie. 2024a. Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts. In *Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis*, LAMPS '24, page 57–68, New York, NY, USA. Association for Computing Machinery.

Zining Zhu, Haoming Jiang, Jingfeng Yang, Sreyashi Nag, Chao Zhang, Jie Huang, Yifan Gao, Frank Rudzicz, and Bing Yin. 2024b. Situated natural language explanations. *Preprint*, arXiv:2308.14115.

Appendix A 分析框架组成部分

我们的框架通过优化的 LLM 提示整合了四种互补的分析视角，以在 SQL 生成之前识别和减轻自然语言查询中的脆弱性（偏见、数据不匹配、逻辑缺陷、框架问题）。

A。1 认知偏见框架

包含了 53 种与数据分析相关的认知偏见 (Soprano et al., 2024; Dimara et al., 2020; Hilbert, 2012; Caverni et al., 1990; Ehrlinger et al., 2016)，将自然语言查询模式映射到潜在推理陷阱。类别包括：

1. 记忆偏差 (8): 事后偏差, 想象力偏差, 回忆偏差, 搜索偏差, 相似性偏差, 证词偏差, 虚假记忆, 易得性偏差。

2. 统计偏差（9）：基础率忽视、机遇、结合、相关、析取、样本量忽视、子集偏差、赌徒谬误、概率忽视。

3. 信心偏见 (8): 完整错觉，控制错觉，确认偏差，欲望偏差，自信过度，冗余错觉，达克效应，偏见盲点。

方法偏误 (12)：数据质量忽视、多重检验谬误、选择偏倚、方法固着、工具自信过度、选择性、成功/自利偏倚、测试无能、锚定、保守主义、参照依赖、回归到均值。

5. 框架 & 上下文偏见 (16): 框架效应, 线性假设, 模式影响, 顺序效应, 量表扭曲, 首位效应, 新近效应, 粒度错觉, 衰减偏见, 复杂性回避, 承诺升级, 习惯, 不一致性, 规则遵循, 基本归因错误, 从众效应。

检查自然语言查询与数据类型的对齐情况。自然语言到 SQL 的关键注意事项：时间性：处理日期/时间格式（例如，‘DATEPART‘），一致的聚合。

A。2 图尔敏论证结构

根据图尔敏模型 (Toulmin, 1958) 评估 NL 查询/SQL 中的隐含参数：

1. 声明的清晰度/相关性：SQL 是否捕捉到自然语言断言并与上下文对齐？（‘SELECT‘，‘WHERE‘）。
2. 证据充分性/有效性：是否检索到足够可靠的数据？（‘COUNT‘，‘LEFT JOIN‘）。来源是否可信？
3. 保证有效性/适用性：NL 到 SQL 的逻辑是否合理？是否遵循约束条件？（CTEs，领域检查）。
4. 支撑：由标准做法/定义支持的逻辑？
5. 限定词精度/范围：是否承认限制（置信度、范围"WHERE"、舍入）？
6. 答辩考虑因素：替代查询、解释（‘JOIN‘混淆因素）、例外（‘EXCLUDE‘）？

A。3 反论证框架

系统性地质疑 NL 查询/表述以进行分析上的严谨：

1. 结论反驳者：需要限制范围吗？替代查询会产生不同的结论？
2. 前提反驳者：依赖于不准确/不完整（‘IS NULL‘）/非代表性的数据？指标是否合适？
3. 论点反驳者：隐藏的假设是否值得质疑？替代解释（通过 ‘JOIN‘ 的混淆因素）？
4. 框架挑战：问题是否正确？是否忽视了观点/时间框架？聚合水平是否合适？
5. 实施挑战：数据表明的可行性问题或意外后果？

Appendix B 实验设置细节

B。1 互动用户研究问卷

我们设计了一份直观的问卷，以评估用户在四个关键维度上对 VeriMinder 的体验：情境真实性、建议的有效性、理由的清晰度，以及对分析的影响。用户在每个维度上使用 5 点 Likert 量表进行评分。图 9 展示了我们互动研究中使用的反馈表。

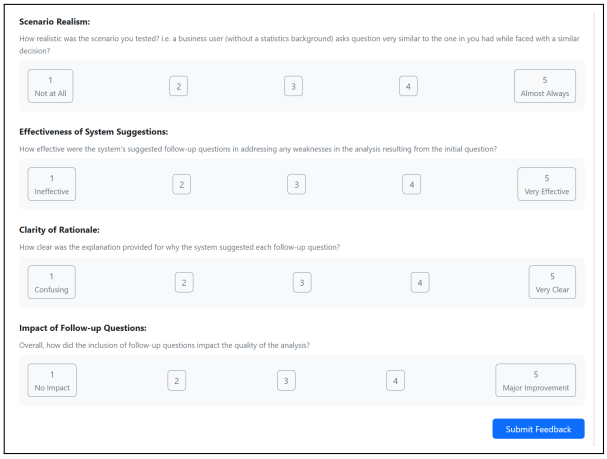

Figure 9: 互动用户研究反馈界面

B。2 系统比较评估

比较评估要求参与者在三个分析维度上对所有五个系统（VeriMinder、Direct NL2SQL、决策导向查询生成、问题扰动和评论代理反馈——测试期间名称匿名化）进行评分：准确性、具体性和全面性。参与者对每个系统的每个维度进行 10 分制评分，以便直接比较。图 10 显示了评估界面。

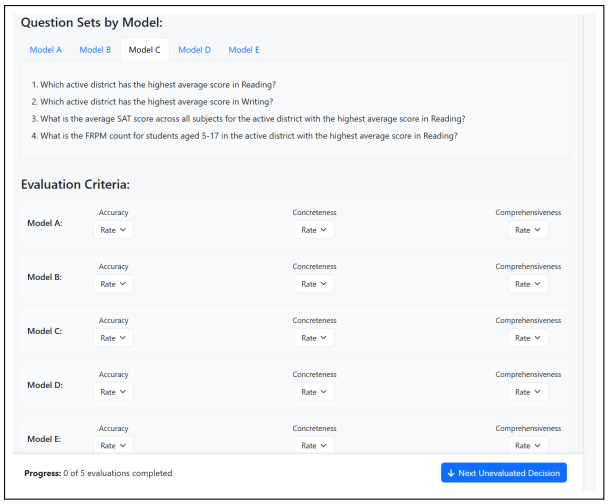

Figure 10: 用于评估不同方法分析质量的比较评估界面

B 。3 自动化评估程序

1. 目标：评估由 VeriMinder 和四个基线系统生成的查询集与大规模数据集（100 对）相比的分析质量。
2. 方法：使用结构化提示，使用 LLM 评估器（Gemini Flash 2.0）（Google DeepMind, 2025），包括：
   (a) 决策背景和原始自然语言问题。
   (b) 数据库模式片段和相关的证据上下文。
   (c) 每个系统（VeriMinder、Direct NL2SQL、Decision-Focused、PerQS、CAF）针对给定决策场景生成并成功执行的 SQL 查询结果的完整集合。我们选择 LLM 主要是基于响应时间（Artificial Analysis, 2025）和由用户界面需求决定的流支持。
3. 评估任务：LLM 被指示去：
   (a) 在决策背景中全面评估每个系统的整个查询和结果集。
   (b) 基于数据准确性、与自然语言问题意图的一致性、全面性、具体性以及在决策目标背景下的整体实用性评估每个系统。
   (c) 应用 SLOW 框架（S ure, L ook, O pposite, W orst）（O'Sullivan and Schofield, 2019）来识别每个系统输出和组合分析中的不确定性、缺失信息、替代解释和潜在问题结论。
4. 输出：该过程为每个系统产生了结构化评估和比较性评估，包括在指定分析维度上的相对排名。