Mirror：用于数据查询、摘要和可视化的自然语言界面

Canwen Xu 加利福尼亚大学圣迭戈分校 美国加利福尼亚州 cxu@ucsd.edu

Julian McAuley 加利福尼亚大学圣迭戈分校 美国加利福尼亚州 jmcauley@eng.ucsd.edu

Penghan Wang*思科 美国加利福尼亚州 pengha wa@cisco.com

arXiv:2303.08697v1 [cs.DB] 15 Mar 2023

摘要

我们介绍 Mirror，这是一个由大语言模型驱动的开源数据探索与分析平台。Mirror 提供直观的自然语言界面，用于查询数据库，并自动生成可执行的 SQL 命令，以检索相关数据并用自然语言进行摘要。此外，用户可以预览并手动编辑生成的 SQL 命令，以确保查询准确无误。Mirror 还会生成可视化内容，帮助用户理解数据。Mirror 的设计兼顾灵活性与人工参与，既适合经验丰富的数据分析师，也适合希望从数据中获取洞见的非技术专业人士。[1]

CCS 概念

• **信息系统** → 数据挖掘；• **以人为中心的计算** → 自然语言界面；可视化系统与工具；• **计算方法** → 自然语言处理。

关键词

语义解析、预训练语言模型、自动数据分析、自然语言界面

ACM 参考格式：

Canwen Xu、Julian McAuley 和 Penghan Wang。2023年 Mirror：用于数据查询、摘要和可视化的自然语言界面。载于*2023* 年 *ACM Web* 会议伴随论文集（*WWW '23 Companion*），*2023*年*4*月*30*日至*5*月*4*日，美国得克萨斯州奥斯汀。美国计算机协会，纽约，美国，4 页。

https://doi.org/10.1145/3543873.3587309

1 引言

数据探索和数据分析对于各种规模、各个行业的组织而言都是至关重要的任务。从大型复杂数据库中有效提取洞见的能力，能够带来显著的竞争优势，并为战略决策提供依据 [14]。然而，对许多个人和组织来说，获取和理解数据可能是一项艰巨的任务。传统的数据分析工具通常需要大量专业技术知识（例如，使用 SQL 进行查询；使用 Python 或 R 进行分析和可视化），非技术专业人士可能难以使用。

本文介绍了 Mirror，这是一个由大型语言和代码模型 GPT-3 [3] 和 Codex [4], 驱动的开源数据探索与分析平台，其灵感源于此前关于语义解析 [23, 24], 数据摘要 [1], 和自动可视化 [26]的研究。Mirror 的用户界面如图 1 所示。Mirror 以童话故事白雪公主 [6],中的魔镜命名，提供直观的自然语言界面来查询数据库，并自动生成可执行的 SQL 命令，以检索相关数据并用自然语言进行摘要。这种方法使用户无需编写复杂的 SQL 命令即可轻松提取相关信息，让数据探索变得更加便捷高效。除了自然语言查询功能外，Mirror 还允许用户预览并手动编辑生成的 SQL 命令，以确保查询的准确性。这一功能让用户能够更好地掌控所提取的数据，从而微调查询并确保获得最相关的信息。为了进一步帮助用户理解数据，Mirror 还会生成可视化内容，以清晰简洁的方式呈现信息。由于 GPT-3 和 Codex 等预训练大型代码和语言模型擅长零样本和少样本推理 [3]，，因此无需为新的数据库收集和标注监督数据，Mirror 是一款用于数据探索的“即插即用”工具。

Mirror 兼顾灵活性与人性化设计，既适合经验丰富的数据分析师，也适合希望从数据中获取洞见并据此做出明智决策的非技术专业人士。我们还展示了两个应用案例：一个可实时更新的体育自动问答应用，以及 OSS Insight Data Explorer——将 Mirror 应用于开源事件分析的工业级集成方案。

2 相关工作

语义解析 [23, 24] 是一项重要任务，旨在将自然语言转换为可由计算机执行的逻辑形式。作为语义解析的一项主要应用，文本转 SQL 旨在根据用户查询生成可执行程序。这些方法在特定数据集或领域上进行训练和测试时已取得成功 [2, 7, 8, 10, 18, 21, 22, 25]。然而，直到最近，如何设计一个稳健且适应性强、能够即插即用于任意数据源的系统，仍未得到探索。

近年来，预训练语言模型在零样本和少样本推理方面展现出巨大潜力 [3, 5, 17, 19]。一些同期研究也探索了使用预训练语言模型进行数据查询。Rajkumar 等人 [16] 测试了

*通信请联系作者。[1]代码可在 https://github.com/mirror-data/mirror 获取。

允许免费制作本作品部分或全部内容的数字或纸质副本，供个人或课堂使用，前提是不得为营利或商业利益而制作或分发副本，并且副本首页须保留此声明和完整引文。本作品中第三方组件的版权必须得到尊重。其他用途请联系所有者/作者。*WWW '23 Companion*，*2023*年*4*月*30*日至*5*月*4*日，美国得克萨斯州奥斯汀 © 2023年版权归所有者/作者所有。美国计算机协会 ISBN 978-1-4503-9419-2/23/04. https://doi.org/10.1145/3543873.3587309

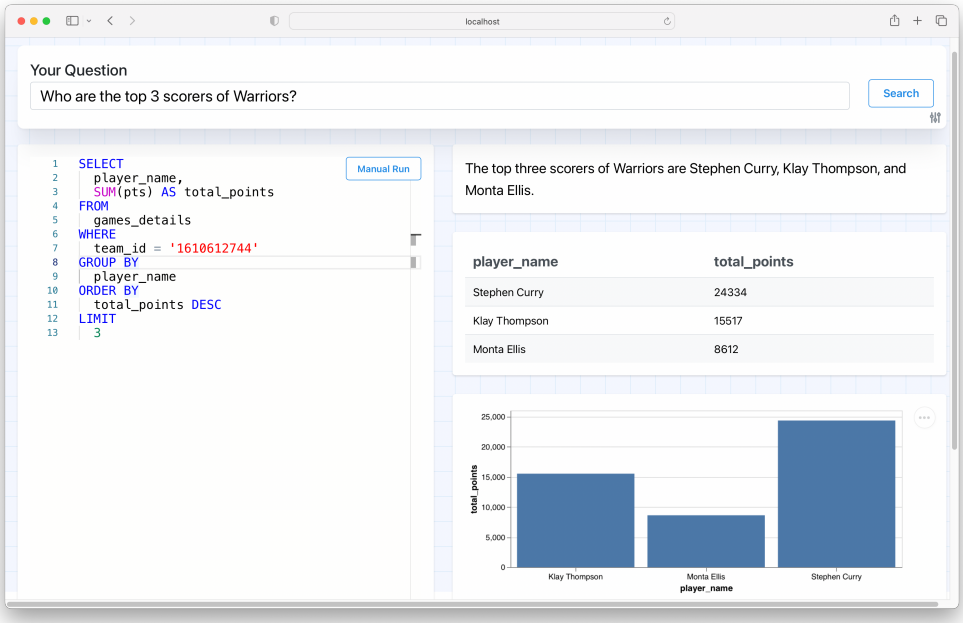


图 1：Mirror 的用户界面。（上方）输入框允许用户用自然语言输入任意查询。系统可根据元数据为用户输入提供自动补全建议。（左下方）代码编辑器显示预训练代码模型生成的 SQL 命令。用户可以编辑并重新运行 SQL 命令。（右下方）自动生成的可视化内容具有交互性。用户可以通过嵌入式菜单轻松导出生成的可视化内容。

GPT-3 [3] 和 Codex [4]在 Spider [22]基准测试上的表现。无需任何训练，Codex 的表现已接近最先进水平，可与微调后的 T5 模型相媲美 [15]。值得注意的是，这项工作作为一项概念验证，为我们的研究奠定了坚实基础。Binder [11] 是一条用于少样本数据查询的流水线，它利用 Codex API 填充指定槽位来生成 SQL。请注意，与这项工作不同，Mirror 专注于构建一个用于数据查询和分析的人在回路系统。因此，Mirror 与 Binder 相互正交，旨在通用化，并能以最少的改动兼容不同的提示技术。

商业智能（BI）工具是一类数据分析应用，具有友好的用户界面，允许人们连接并查询数据源、构建仪表板以可视化数据并分享结果。Redash2 就是一个例子，它使用 SQL 检索数据，并支持多种可视化格式。Metabase[3] 则为不会编写 SQL 查询的用户提供图形用户界面，作为一种替代方案。与这些工具不同，Mirror 使用自然语言界面，进一步降低了数据探索的门槛。Mirror 还实现了更高程度的自动化，能够完成数据查询、摘要和可视化的整个流程。

3 系统实现

我们分两个阶段介绍 Mirror，即数据查询以及摘要和可视化。第二阶段摘要和

可视化会在每次成功查询数据源并返回结果表时自动触发。请注意，所有 API 调用，包括 OpenAI Codex [4] 和 GPT-3 [3]，都可以轻松替换为其他 API 服务（例如 Hugging Face API [20]）或自托管模型。为简单起见，我们将预训练的代码生成模型称为 Codex，将文本生成模型称为 GPT-3。

3.1 数据查询

数据查询阶段的组件和工作流程如图 2 所示。

数据源 初始化 Mirror 时，服务器端后端会创建一个数据库连接，该后端由 Node.js 驱动。它支持大多数类型的关系型数据库。值得注意的是，我们还可以通过在服务器上将逗号分隔值（CSV）表转换为临时 SQLite 实例，从而支持查询该表。

提示词构建 提供给预训练代码模型的提示词由三个部分组成。(1) 模板 是由用户设计的、特定于数据库的格式字符串，其中包含用于填入元数据 和查询的占位槽。它还可以包含其他指令，例如指定使用哪种 SQL 方言、表名的别名、值映射（例如，将机场代码映射为名称）或任何有助于提升特定数据集上的性能的内容。(2) 元数据 是可从数据源自动读取的数据库模式，会被插入模板中相应的占位槽。(3) 查询 是以自然语言形式输入的用户内容。Mirror 还提供数据感知查询自动补全功能，可在用户输入时结合元数据补全查询。

[2]https://redash.io/ [3]https://www.metabase.com/

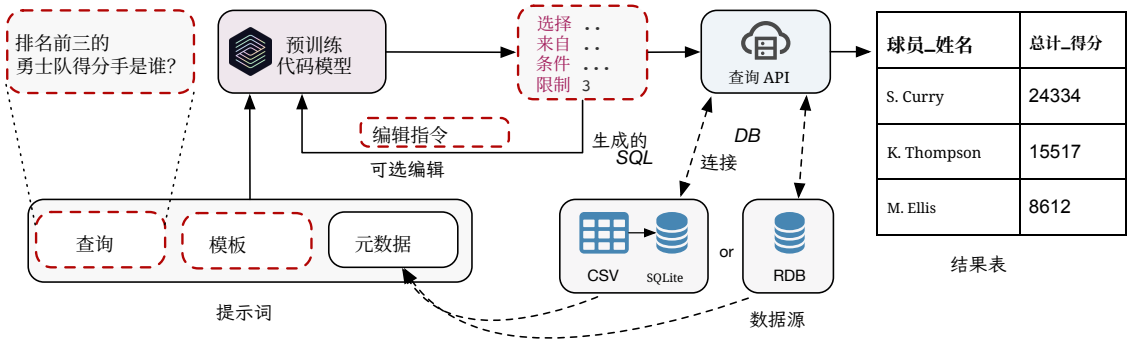


图 2：Mirror 在数据查询阶段的组件和工作流程。红色虚线表示人在回路组件，这些组件允许用户进行交互或编辑。此处省略了带元数据的查询自动补全。

SQL生成 构建好提示词后，我们会调用 Codex API，它会返回一段应可执行的 SQL 片段。获得 SQL 命令后，我们会调用后端查询 API，从数据源获取数据。请注意，为了提高安全性，数据库连接应设置为只读，以防止 SQL 注入或意外修改（尽管这种情况不太可能发生）。如果 SQL 执行失败，Mirror 会重试调用 Codex API。[4] 如果 API 成功从数据源获取数据，随后便会触发摘要和可视化流程（稍后将详细介绍）。

原始与指令式SQL编辑 我们提供了一个带自动补全功能的SQL编辑器，允许用户直接编辑SQL，以融入人工专业知识。对于不熟悉SQL的用户，如果对返回结果不满意，可以使用自然语言指令来编辑SQL 命令。例如，在图2所示的示例中，如果用户想从结果表及后续分析中排除退役球员，只需添加指令“排除已退役的球员”即可。此功能由OpenAI Codex编辑API提供支持。[5]

3.2 摘要和可视化

如图 3 所示，我们使用摘要模板将查询和结果组织成提示词，并使用预训练语言模型 GPT-3 [3], 生成文本形式的答案来回应查询。同时，我们使用模板格式化用于可视化的提示词，并将结果表作为输入。我们让 Codex 模型生成 Vega 描述。Vega[6] 是一种用于创建交互式可视化设计的声明式语言。它支持多种图表，包括条形图、折线图和面积图、饼图、散点图、分布图，甚至地理地图。这使 Mirror 能够针对任何数据集和用途生成丰富多样的可视化。生成的 Vega 描述采用 JSON格式，可由 Vega 运行时渲染。与 SQL生成类似，如果 Vega 代码编译失败，系统将重试生成。

[4]Codex API 使用随机采样 [9] 生成代码，因此重试时可能返回不同的结果。[5] https://beta.openai.com/docs/api-reference/edits [6]https://vega.github.io/

4 案例研究

4.1 体育领域的自动问答

体育数据自动问答系统对体育迷、媒体和管理者都很有用。体育数据会实时更新。因此，基于大语言模型的问答系统（例如 ChatGPT [12]）无法应对如此频繁的更新。Mirror 可以连接到实时体育数据库，为球迷提供快速、准确的信息，让他们及时了解体育界的最新动态并保持关注。它还可以成为研究和报道体育报道的宝贵工具，为相关人员提供最新数据，以便向读者和观众提供信息。对于管理者而言，该系统可以通过表格和图表提供有关团队和球员的宝贵见解，帮助他们就战略和战术做出更加明智的决策。

4.2 OSS Insight 数据探索器

OSS Insight[7] 是一款通过分析数十亿条 GitHub 事件（例如提交、仓库创建和拉取请求），为开源世界提供全面、有价值且紧跟趋势洞察的工具。它可深入分析单个 GitHub 仓库和开发者，也能使用相同指标比较两个仓库。OSS Insight 是发现开源社区新趋势的绝佳平台，还能助力发掘新的投资和合作机会。

OSS Insight Data Explorer[8] 是 Mirror 的一项工业应用。它通过集成让用户能够使用 Mirror 的自然语言界面查询数据。Mirror 让用户（包括非 SQL 使用者，例如投资者）轻松获取数据、洞察和图表。在线演示已公开提供。[8]

5 局限性

域外输入。 Mirror 的一个局限是对域外输入的处理。例如，当用户尝试从体育数据库中查询歌手的薪资时，Mirror 可能会生成随机或无法执行的 SQL，而不是告知用户数据库中没有这类信息。因此，允许

[7]https://ossinsight.io/ [8] https://ossinsight.io/explore

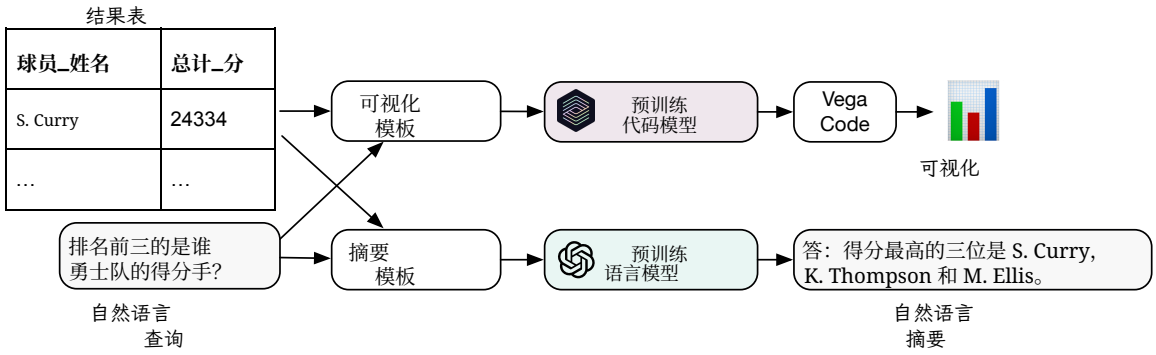


图 3：Mirror 在摘要和可视化阶段的组件与工作流程。

生成模型访问 SQL 执行引擎并启用多轮查询或许是可行的。借助对话式语言模型（例如 ChatGPT [12]），这或许可行。

安全性。 Mirror 是一款商业智能（BI）工具，因此主要用于内部场景或公共数据。先前的研究 [13] 表明，通过文本转 SQL 模型进行 SQL 注入是可行的。因此，我们建议授予 Mirror 只读访问权限，以防生成的 SQL 意外修改数据库。此外，只有应当拥有数据库完整读取权限的人员才应获得 Mirror 的访问权限。

6 结论与未来工作

本文提出了 Mirror，这是一个由语言模型驱动的数据分析开源平台。凭借人在环路中的设计和零样本能力，Mirror 是一款实用的数据探索与分析工具，适用于许多应用场景。未来，我们将探索结合 ChatGPT [12] 与 SQL 引擎进行基于对话的多轮查询，以进一步提升 Mirror 的自我纠正能力，并通过类对话界面实现更多用户交互。

致谢

Mirror 是 PingCAP TiDB Hackathon 2022 的参赛项目，并荣获奖项。我们感谢 PingCAP, Inc. 提供丰厚的奖项与支持。OSS Insight 的知识产权（包括但不限于其 Mirror 集成）归 PingCAP, Inc. 所有，不属于本文所述的开源项目 Mirror。

参考文献

[1] Mohiuddin Ahmed. 2019. 数据摘要：一项综述。知识与信息系统58, 2 (2019), 249–273. [2] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. 基于问答对的 Freebase 语义解析。载于*EMNLP*。ACL，1533–1544. [3] Tom B. Brown,Benjamin Mann,Nick Ryder, 等. 2020. 语言模型是少样本学习器。载于*NeurIPS*。[4] Mark Chen, Jerry Tworek, Heewoo Jun, 等. 2021. 对使用代码训练的大语言模型进行评估。*arXiv* 预印本 *arXiv:2107.03374* (2021). [5] Hyung Won Chung, Le Hou, Shayne Longpre, 等. 2022. 扩展指令微调语言模型。*arXiv* 预印本 *arXiv:2210.11416* (2022). [6] Jacob Grimm and Wilhelm K Grimm. 1991. 白雪公主。Andrews McMeel Publishing.

[7] Daya Guo, Yibo Sun, Duyu Tang, 等. 2018. 从 SQL 查询生成问题可改善神经语义解析。载于*EMNLP*。计算语言学协会，1597–1607. [8] Daya Guo, Duyu Tang, Nan Duan, 等. 2019. 结合检索与元学习进行上下文相关语义解析。载于*ACL*。计算语言学协会，855–866. [9] Ari Holtzman, Jan Buys, Li Du, 等. 2020. 神经文本退化的奇特案例。载于*ICLR*。OpenReview.net. [10] Chen Liang, Mohammad Norouzi, Jonathan Berant, 等. 2018. 用于程序合成与语义解析的记忆增强策略优化。载于*NeurIPS*。10015–10027.[11] Aman Madaan, Shuyan Zhou, Uri Alon, 等. 2022. 代码语言模型是少样本常识学习者。*arXiv* 预印本 *arXiv:2210.07128* (2022). [12] OpenAI. 2023. ChatGPT：优化用于对话的语言模型。https: //openai.com/blog/chatgpt/ [13] Xutan Peng, Yipeng Zhang, Jingfeng Yang, and Mark Stevenson. 2022. 文本转 SQL 模型的安全漏洞。*arXiv* 预印本 *arXiv:2211.15363* (2022). [14] Foster J. Provost and Tom Fawcett. 2013. 数据科学及其与大数据和数据驱动决策的关系。大数据1, 1 (2013), 51–59. [15] Colin Raffel, Noam Shazeer, Adam Roberts, 等. 2020. 探索统一文本到文本转换器的迁移学习极限。《机器学习研究杂志》21 (2020), 140:1–140:67. [16] Nitarshan Rajkumar, Raymond Li, and Dzmitry Bahdanau. 2022. 评估大语言模型的文本转 SQL 能力。*arXiv* 预印本 *arXiv:2204.00498* (2022). [17] Victor Sanh, Albert Webson, Colin Raffel, 等. 2022. 多任务提示训练实现零样本任务泛化。载于*ICLR*。OpenReview.net. [18] Peter Shaw, Ming-Wei Chang, Panupong Pasupat, and Kristina Toutanova. 2021. 组合泛化与自然语言变异：语义解析方法能否兼顾二者？载于*ACL-IJCNLP*。计算语言学协会，922–938. [19] Jason Wei, Maarten Bosma, Vincent Y. Zhao, 等. 2022. 微调语言模型是零样本学习者。载于*ICLR*。OpenReview.net. [20] Thomas Wolf, Lysandre Debut, Victor Sanh, 等. 2020. Transformers：自然语言处理的最新进展。载于*EMNLP*（演示）。计算语言学协会，38–45. [21] Pengcheng Yin and Graham Neubig. 2017. 用于通用代码生成的句法神经模型。载于*ACL*。计算语言学协会，440–450.[22] Tao Yu, Rui Zhang, Kai Yang, 等. 2018. Spider：用于复杂跨领域语义解析和文本转 SQL 任务的大规模人工标注数据集。载于*EMNLP*。计算语言学协会，3911–3921. [23] John M. Zelle and Raymond J. Mooney. 1996. 使用归纳逻辑编程学习解析数据库查询。载于*AAAI/IAAI, Vol. 2*。AAAI 出版社 / 麻省理工学院出版社，1050–1055. [24] Luke S. Zettlemoyer and Michael Collins. 2005. 学习将句子映射为逻辑形式：基于概率范畴语法的结构化分类。载于*UAI*。AUAI 出版社，658–666. [25] Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql：使用强化学习从自然语言生成结构化查询。*arXiv* 预印本 *arXiv:1709.00103* (2017). [26] Sujia Zhu, Guodao Sun, Qi Jiang, 等. 2020. 自动信息图表与可视化推荐综述。《可视化信息学》4, 3 (2020), 24–40.