arXiv:2304.13010v2 [cs.DB] 5 Jul 2023

非结构化数据与结构化数据：借助大语言模型，我们能否兼得二者之长?

Wang-Chiew Tan*Reality
Labs Research，Meta
wangchiew@meta.com

1 引言

我们正在见证大语言模型（LLMs）领域的快速发展。在 Google Scholar 上进行搜索可以发现，2022 年，标题中包含“大语言模型”的论文约有 3,910 篇。截至 2023年4月12日，标题中包含 LLMs 的文章已经有 1,700 篇。除了 Google Scholar 之外，近几个月来，我们还见证了围绕 LLMs 大量涌现的博客文章、新闻报道、推文和开源代码仓库。

或许，ChatGPT（于 2022年11月30日 发布）正是这场大语言模型革命的缩影。它真正释放并向大众展示了近年来自然语言和机器学习社区一直在酝酿的力量。ChatGPT[1] 能够完成许多任务，并且可以通过针对问题生成智能且类似人类的回答，表现得十分出色。以自然语言作为输入和输出[2], 时，它可以在一定程度上解决复杂的数学问题 [4, 21], ，将你的规格说明转换为你所选择的编程语言代码（例如， [11, 21]），帮助你以不同风格撰写文章，这份成就清单还在不断延长 [18]。

除了回答问题之外，还可以向它提供表格，例如 CSV文件，并用自然语言提出关于这些表的问题。更有趣的是，还可以同时向 ChatGPT 提供文本和表格，并以无缝的方式针对这两类信息提问。再多坚持一会儿

*文中所表达的观点仅代表我个人。[1]在本文其余部分中，我用“ChatGPT”泛指基于大语言模型的通用聊天系统。[2]GPT-4 可以接受图像和文本输入。

进行这项尝试很快就会发现，ChatGPT 对每次提示中可以输入的内容量以及它能够保留的信息量都有限，至少根据我在2023年3月底进行尝试时的体验是如此。

很自然地，我们会思考能否通过一些扩展，将 ChatGPT 用作存储和查询数据的系统，并以它擅长的自然语言作为主要的输入和输出媒介。这样做面临哪些挑战？在这一背景下，作为数据库实践者和理论研究者，我们应如何取得进展？

2 非结构化数据与结构化数据

如今，包括文本、图像、音频和视频在内的大量数据都处于“盒子之外”。通常，这类非结构化数据会同时包含多种模态。例如，我们经常能在图像中找到文本 [12],，也可能在网络上的视频和/或图像旁找到与之关联的文本。非结构化数据之所以在很大程度上普遍存在，是因为用户可以通过各种广泛可用的应用和创作工具轻松创建并分享这类数据 [8]。这类数据通常通过关键词搜索进行查询；如今，也可以使用大语言模型以自然语言进行查询。通常，设计模式并建立用于查询数据的数据库所需的成本，会阻碍人们一开始就使用数据库管理系统（DBMS）。然而，随着数据规模扩大，对结构和语义的需求变得越发重要，从而能够更快速、更准确地检索内容。例如，按年份、旅行或实体类型（如人物或宠物）整理照片，可以增加一定的结构，使回答某些类型的查询更加快速、准确。然而，要回答“我上次去 Italio 餐厅旁边的咖啡店是什么时候？”或“我和 Anna 一起用芒果蛋糕庆祝生日多少次？”这类复杂查询，就需要对数据进行超出当今大语言模型能力范围的复杂推理和计算 [1, 22]。

企业数据位于光谱的另一端。它大多不是由日常用户编写的，而是高度结构化的，通常“装在某个数据库管理系统的盒子里”。企业数据拥有经过精心设计的模式，并且通常经过高度优化，能够高效地服务于规模可观的查询工作负载。此类数据通常直接使用 SQL 进行查询；SQL 具有足够的表达能力，可以指定包含聚合和递归等复杂查询。人们还开展了大量研究，以支持使用自然语言查询数据库（例如 [17, 20]）：用户以自然语言提出问题，系统将其转换为 SQL，然后在数据库管理系统上执行。数据库管理系统经过高度优化，能够处理大量数据，还可以执行具有 ACID 保证的事务 [6]。然而，核心数据库管理系统并不理解自然语

言，而且在查询非结构化数据方面往往能力不足；非结构化数据通常以数据块的形式存储，而为这些数据块添加语义则需要额外的工作。例如，“告诉我去年黑色星期五和网络星期一期间的销售数字”这一请求，需要具备关于“销售数字”含义的常识知识，并了解它如何对应于相关表属性；这些属性在不同数据库中可能以不同名称存储，还需要知道黑色星期五和网络星期一分别是何时。另一个例子是“找出所有评价良好且与这些图像相似的商品”，这需要对图像进行（语义上的）匹配，并根据数据理解“评价良好”的含义。

尽管如今非结构化数据与结构化数据的管理方式存在差异，但用自然语言进行查询是二者共同的需求。与此同时，要求所有非结构化数据都适应某种结构，或反过来为了查询而使用同一个系统，都是不合理且不自然的。那么，借助拥有海量知识和语言理解能力的大语言模型，我们该如何有效查询这两类数据呢?

3 借助大语言模型兼得二者之长?

大语言模型的模型参数中包含海量参数化知识，但缺乏整合外部数据（即模型之外的数据）的能力。因此，如果某个模型基于截至2021年的数据进行训练，那么对于需要2021年之后的知识才能回答的事件或事实，它将无法提供正确答案。例如，如果有人在包含文本的网页上提出问题”2022年的中期选举结果如何？”，text-davinci-003 [15], 返回的答案是”2022年的中期选举结果尚不可用，因为选举尚未举行。”这是因为text-davinci-003使用截至2021年6月的数据进行训练。

检索增强语言模型（例如 [5, 10, 25]）通过采用半参数化方法来回答查询，克服了这一局限。它们使用外部数据：先从外部数据存储中检索相关数据，然后尝试结合检索到的数据和自身的参数化知识来回答问题。然而，正如（表 5, [1]）所示，此类系统在处理涉及聚合和某些类型时间查询的复杂查询时，表现仍可能不佳。这是因为，对于这类查询，往往无法将回答问题所需的全部数据装入语言模型所规定的有限长度词元输入中。然而，随着大语言模型变得更大，或随着人们进一步改进以提高语言模型所规定的词元限制，我们可以预期，未来大语言模型将接受越来越大的输入，有限长度的词元限制或许很快就不再是问题。同时，可能会出现规模更大的数据集需要管理，也需要更多数据来计算正确答案。因此，这个问题至少在一段时间内仍将持续存在。

而且，可能会出现规模更大的数据集需要管理，也需要更多数据来计算正确答案。因此，这个问题至少在一段时间内仍将持续存在。

文献 [2] 提出了一种方案，以克服上述局限对某些类型查询造成的影响。该论文提出了一种设想：利用视图（例如表）对底层数据源中的部分数据进行结构化。数据源可以包含不同模态的数据，例如文本、图像、视频，甚至表。视图用于呈现数据的重要属性，或不同模态数据之间的关联；大语言模型则用于将自然语言查询转换为查询（例如此处的 SQL），并尽可能在视图上执行。采用这一方案时，一个关键问题是要弄清楚哪些自然语言查询可以通过视图回答。如果无法使用视图回答某个查询，系统就会转而使用检索增强语言模型，尽力回答该查询。或者，这两个组件（基于视图的查询问答组件和基于检索的查询问答组件）也可以协同生成最终答案。论文还提出了其他几个问题，例如应该具体化哪些视图？给定预期的查询工作负载，如何自动选择“合适”的视图进行具体化？又该如何判断一个问题更适合通过视图、检索系统，还是两者结合来回答？

除了上述内容外，下面我还将重点介绍一些我认为可能是数据库领域目前最关心的问题：

使用不同资源并考虑预算约束进行查询问答[2]中已经讨论过如何为给定查询找到合适的查询计划。该系统的核心依赖两个组件——视图，以及用于回答查询的检索增强语言模型。可以设想为系统添加其他组件，例如根据自然语言提示生成图像的组件；有时，这可能有助于回答某些查询[3]。于是，一个关键问题是：我们如何判断应利用哪个组件来回答查询，或让多个组件协作得出答案？此外，大语言模型计算密集型，生成答案可能较慢。有些大语言模型也并非免费。除了要决定如何利用所有可用资源尽可能有效地回答查询之外，如何考虑每个组件的优势和局限，以列举并比较用于计算给定查询答案的计划，和／或满足给定预算的计划？我们也能否使用语言模型来生成查询计划或其中的一部分，类似于语言模型如何被用于自行推理一系列步骤，以根据问题推导答案（例如， [23, 24]）？

[3]俗话说，一图胜千言。

数据溯源对于某些类别的 SQL 查询已有深入研究，其大致定义为能够解释某个元组为何出现在查询结果中的源元组。正如 [2], 中所述，只要条件允许，就应尝试使用表视图回答查询，方法是将自然语言查询转换为可在这些视图上执行的 SQL 查询。这样一来，便有可能“免费”获得数据溯源。然而，大语言模型生成的 SQL 查询可能很复杂，例如在 FROM 或 WHERE 子句中包含嵌套 SQL 查询和／或聚合。对于这种情况，我们能否将生成的 SQL 查询分解为一个或多个“更简单的查询”组成的序列？对于这些更简单的查询，数据溯源已有充分研究，也能轻松推导。如果无法做到这一点，我们能否制定另一种回答查询的计划，以便推导数据溯源？寻找替代查询计划的问题与前一段的讨论相关，但此处重点在于推导出由足够简单的步骤构成的计划，从而实现数据溯源。

检索增强语言模型还能够为其答案提供证据带来更多保障。我们也开始看到这样的实现：检索增强语言模型所提供答案的来源会作为答案的一部分返回 [13]。然而，仍需开展更多研究，以确定训练数据的数据溯源（例如， [3, 9, 16]），从而更全面地了解所提供答案的数据溯源。

提示工程分析 大语言模型能够接受的输入词元数量有限。即使充分利用可用的所有词元，用更多信息向其发出提示也不一定会带来更好的答案，因为有时向语言模型提供更多信息反而会使其感到困惑。这意味着，为了让语言模型得出高质量的正确答案，我们需要谨慎选择发送给它的内容[4]。

语言模型返回的答案还容易受到提示方式的影响。有时，仅仅以略有不同的措辞提出同一个问题，就会得到完全不同的答案。在提示工程中，目标是为当前任务找到最佳提示。

鉴于语言模型的最大词元限制，我们能否通过策略性地总结相关数据和/或删除无关数据，来优化语言模型返回的答案？例如， [14],的实体匹配系统使用语言模型，只需采用一个简单技巧，策略性地缩短数据的文本描述，仅保留价值较高、TD/IDF 较高的词元，就能立即更准确地执行实体匹配。从理论层面来说，如果对提示作出适当定义，并对语言模型作出一些假设，我们能否刻画出在词元预算限制下，需要访问

[4]正确回答问题的方法多种多样。有些方法优于其他方法。

外部数据的查询，分别可以通过一个提示、有限个提示或渐近数量的提示来回答?

4 结论

大语言模型领域正在快速发展，无论是在研究方面还是工业界方面。除 [2]–[7] 之外，相关工作还描述了在增强型语言模型和数据集成的背景下，通过单一数据源或串联多个数据源回答查询所面临的挑战。在 [19], 中，作者描述了如何扩展数据库管理系统，以利用大语言模型改进查询问答，并指出了构建同时涉及数据库管理系统和大语言模型的混合查询问答系统这一方向。我相信，这只是开始；在这一查询问答系统领域，我们很快将看到更多兼顾非结构化数据与结构化数据、并通过使用大语言模型实现的构想、研究和实践。

致谢 上述许多观点都源于与 Alon Halevy 和 Yuliang Li 的讨论。我还要感谢 Meta 的许多同事——Lambert Mathias、Richard Newcombe 和 Luna Dong——感谢他们积极参与围绕大语言模型的讨论，让我从中受益良多；同时也感谢 Lucian Popa 对本文提出的反馈。

参考文献

[1] 待发布。[2] 待发布。[3] E. Akyurek、T. Bolukbasi、F. Liu、B. Xiong、I. Tenney、J. Andreas 和 K. Guu。将语言模型中的知识追溯至训练数据。在 EMNLP，页码 2429–2446，2022年12月。[4] S. Frieder、L. Pinchetti、R.-R. Griffiths、T. Salvatori、T. Lukasiewicz、P. C. Petersen、A. Chevalier 和 J. Berner。ChatGPT 的数学能力，2023。[5] K. Guu、K. Lee、Z. Tung、P. Pasupat 和 M.-W. Chang。Realm：检索增强语言模型预训练。在 ICML，2020。[6] T. Haerder 和 A. Reuter。面向事务的数据库恢复原则。ACM Computing Surveys，15(4)：287–317，1983年12月。[7] A. Y. Halevy 和 J. Dwivedi-Yu。数据集成对增强型语言模型的启示。CoRR，abs/2304.04576，2023。[8] A. Y. Halevy、O. Etzioni、A. Doan、Z. G. Ives、J. Madhavan、L. K. McDowell 和 I. Tatarinov。跨越结构鸿沟。在 CIDR，2003。

[9] X. Han、B. C. Wallace 和 Y. Tsvetkov。通过影响函数解释黑箱预测并揭示数据伪影。在 ACL，页码 5553–5563，2020年7月。[10] G. Izacard、P. Lewis、M. Lomeli、L. Hosseini、F. Petroni、T. Schick、J. Dwivedi-Yu、A. Joulin、S. Riedel 和 E. Grave。基于检索增强语言模型的少样本学习。2022。[11] A. Kashefi 和 T. Mukerji。使用 ChatGPT 编程数值方法，2023。[12] D. Kiela、H. Firooz、A. Mohan、V. Goswami、A. Singh、P. Ringshia 和 D. Testuggine。仇恨表情包挑战赛：检测多模态表情包中的仇恨言论。在 NeurIPS，2020。[13] LangChain。使用来源进行检索式问答。https://python.langchain.com/en/latest/modules/chains/index examples/vector db qa with sources.html。[14] Y. Li、J. Li、Y. Suhara、A. Doan 和 W. Tan。使用预训练语言模型进行深度实体匹配。Proc. VLDB Endow.，14(1)：50–60，2020。[15] OpenAI。text-davinci-003。2023年4月12日引用，Playground 地址为 https://platform.openai.com/playground，训练数据信息见 https://help.openai.com/en/articles/6643408-how-do-davinci-and-text-davinci-003-differ。[16] G. Pruthi、F. Liu、S. Kale 和 M. Sundararajan。通过追踪梯度下降估计训练数据影响。在 NeurIPS，第 33 卷，页码 19920–19930，2020。[17] J. Qi、J. Tang、Z. He、X. Wan、Y. Cheng、C. Zhou、X. Wang、Q. Zhang 和 Z. Lin。RASAT：将关系结构集成到用于文本到 SQL 的预训练序列到序列模型中。在 EMNLP，2022。[18]K. Rose。我该如何使用像 ChatGPT 这样的人工智能聊天机器人？《纽约时报》（2023年3月30日）https://www.nytimes.com/2023/03/30/technology/ai-chatbot-chatgpt-uses-work-life.html。[19] M. Saeed、N. D. Cao 和 P. Papotti。使用 SQL 查询大语言模型，2023。[20] T. Scholak、N. Schucher 和 D. Bahdanau。PICARD：对语言模型进行增量解析，以实现受约束的自回归解码。在 EMNLP，页码 9895–9901，2021年11月。[21] P. Shakarian、A. Koyyalamudi、N. Ngu 和 L. Mareedu。对 ChatGPT 在数学文字题（数学文字题）上的独立评估，2023。[22] J. Thorne、M. Yazdani、M. Saeidi、F. Silvestri、S. Riedel 和 A. Y. Levy。从自然语言处理到神经数据库。VLDB Endow.，14(6)：1033–1039，2021。[23] J. Wei、X. Wang、D. Schuurmans、M. Bosma、brian ichter、F. Xia、E. H. Chi、Q. V. Le 和 D. Zhou。思维链提示能够在大语言模型中激发推理。在 NeurIPS，2022。

[24] S. Yao、J. Zhao、D. Yu、N. Du、I. Shafran、K. Narasimhan 和 Y. Cao。《ReAct：在语言模型中协同推理与行动》。发表于 ICLR，2023。[25] M. Yasunaga、A. Aghajanyan、W. Shi、R. James、J. Leskovec、P. Liang、M. Lewis、L. Zettlemoyer 和 W. tau Yih。《检索增强多模态语言建模》，2022。