合成临床文本的生成：系统综述

Basel Alshaikhdeeb[1], Ahmed Abdelmonem Hemedan[1], Soumyabrata Ghosh[1], Irina Balaur[1], and Venkata Satagopam[1]

[1]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, 6, avenue du Swing, Esch-sur-Alzette, L-4367, Luxembourg

Corresponding email: basel.alshaikhdeeb@uni.lu

July 25, 2025

Abstract

生成临床合成自由文本是解决常见临床 NLP 问题（如稀疏性和隐私性）的有效方案。本文旨在通过对三个研究问题进行定量分析，即生成的目的、技术和评估方法，对生成合成的医学非结构化自由文本进行系统的回顾。我们在 PubMed、ScienceDirect、Web of Science、Scopus、IEEE、Google Scholar 和 arXiv 数据库中搜索了与生成医学合成非结构化自由文本相关的出版物。在审查标题、摘要和全文后，我们从 1398 篇文章中收集了 94 篇。从 2018 年起，合成医学文本生成受到了极大的关注，其主要目的是用于文本增强、辅助写作、语料库建设、保密性共享医学文本、注释和实用性。Transformer 架构是生成文本的主要技术，尤其是 GPT。另一方面，有四个主要的评价方面，包括相似性、隐私性、结构和实用性，其中实用性是评估生成的合成医学文本最常用的方法。虽然生成的合成医学文本展示出在不同下游 NLP 任务中代替真实医学文档的中等可能性，但其在增强和补充真实文档方面证明了其价值，有助于提高准确性和克服稀疏性/欠采样问题。然而，隐私仍然是生成合成医学文本的一个主要问题，需要更多的人为评估来检查是否存在敏感信息。尽管如此，合成医学文本生成的进步将大大加速工作流程和流水线开发的采用，省去数据传输繁琐的法律程序。

1 介绍

电子病历/健康记录（EMRs/EHRs）的广泛采用促使生成了大量与病人信息相关的临床文本。从临床笔记到出院小结以及实验室测试报告，为自然语言处理（NLP）任务打开了一扇大门。然而，非结构化的医学自由文本面临着不同方面的挑战 [1] 。首先，由于其非结构化的特性，相较于结构化的信息，它在存储过程中容易丢失。非结构化的医学自由文本提供了关于诊断、症状和药物的显著临床细节。更具体地说，自由文本在心理健康笔记中起着至关重要的作用，这些笔记高度依赖于医生和患者之间发生的叙述。综上所述，非结构化的医学自由文本在隐私方面极令人担忧，因为它包含了关于患者、当前疾病、家庭史以及其他敏感信息的内容。这使得共享此类医学自由文本或提供其公共访问权限极具风险，因为这可能导致个人的重新识别。另一方面，一般的医学领域面临稀疏性的问题。这是因为某些疾病很稀有且仅与有限数量的记录/文件相关联，这促使需要进行过抽样，尤其是针对少数类/疾病 [2] 。生成语言模型的出现成功地增加了非结构化自由文本的数量，其中文本生成不再局限于人类的手动数据输入。这种自动化文本的关键特征在于它模仿人类构造句子的方式，使文本看起来像是合成的，因为它偏离了用于训练模型的原始数据 [3] 。这可能是克服隐私限制的一个良机。虽然这种自动文本生成主要用于医疗保健以外的应用（如聊天机器人、机器翻译、客户服务等）。然而，在过去的几年中，医学领域受到了极大的关注。随着依赖这些生成语言模型的医疗应用数量的增加，对生成文本实用性的理解需求呈指数增长。另一方面，迫切需要确定这些语言模型的能力，并根据质量、隐私和实用性评估其生成的文本。本文旨在通过调查生成目

的、技术和评估方法，针对合成的医学/临床非结构化自由文本进行系统回顾。以下小节将确定研究问题、贡献和相关工作。

1.1 研究问题（RQs）

本研究的主要目的是回顾合成医疗文本生成。因此，指导本次回顾的研究问题如下：

- RQ1: 生成合成医学自由文本的目的是什么？已经处理了哪些语言和数据集？
- RQ2：有哪些技术用于生成合成医学文本？这些技术的性能和关键特征是什么？
- RQ3: 用于评估此类合成医学文本的方法是什么？这些方法如何分类？

1.2 贡献

本研究旨在进行广泛且叙述性的综述，以审查合成文本生成。这将包括生成的目的、生成中使用的技术以及评估这类生成文本所遵循的评估方面。据我们所知，这项工作是首次尝试全面审查合成医学文本的生成。

1.3 相关工作

研究合成医学数据的生成是一个引人注目的领域，如表格 1 所示，在过去的五年中已经出现了多次综述工作。这些综述大多数集中在多模态数据生成或特定类型的数据上，例如图像生成或表格/结构化数据生成，而对医学领域的自由文本生成关注较少。尽管 Murtaza 等人的研究 [4] 将其调查的大部分内容献给了医学文本数据生成，但他们严格地集中在隐私方面。因此，本研究将缩小焦点，专注于医学领域的合成非结构化（自由文本）生成。

Table 1: 类似的综述研究

Year	Author	Medical Synthetic Image Generation	Medical Tabular Synthetic Data Generation	Medical Multimodal Synthetic Data Generation
2022	Hernandez et al. [5]		✓	
2023	Murtaza et al. [4]			✓
2023	Eigenschink et al. [6]			✓
2024	Sherwani & Gopalakrishnan [7]	✓		
2024	Ghosheh et al. [8]		✓	
2024	Budu et al. [9]		✓	
2024	Pezoulas et al. [10]			✓
2024	Kim et al. [11]			✓
2025	Ibrahim et al. [12]			✓

1.4 论文大纲

本文的结构如下：第 2 节详细描述了进行系统回顾的过程。在第 3 节中，将分别回答本研究的研究问题。最后，第 4 节提供了广泛讨论，对重要发现进行了说明。

2 方法

我们遵循了 Khan 等人和 Uman 提出的五个阶段进行审查。第一阶段是确定搜索策略。而第二阶段旨在根据已确定的策略进行搜索，在该阶段中，根据包容标准收集相关文章，并根据排除标准丢弃其余文章。第三阶段旨在通过阅读和筛选过程从选定的文章中提取信息。在第四阶段，将对提取的信息进行总结。最后，第五阶段将旨在解释研究结果。这些阶段如图 1 所示。以下小节将讨论这些方法阶段。

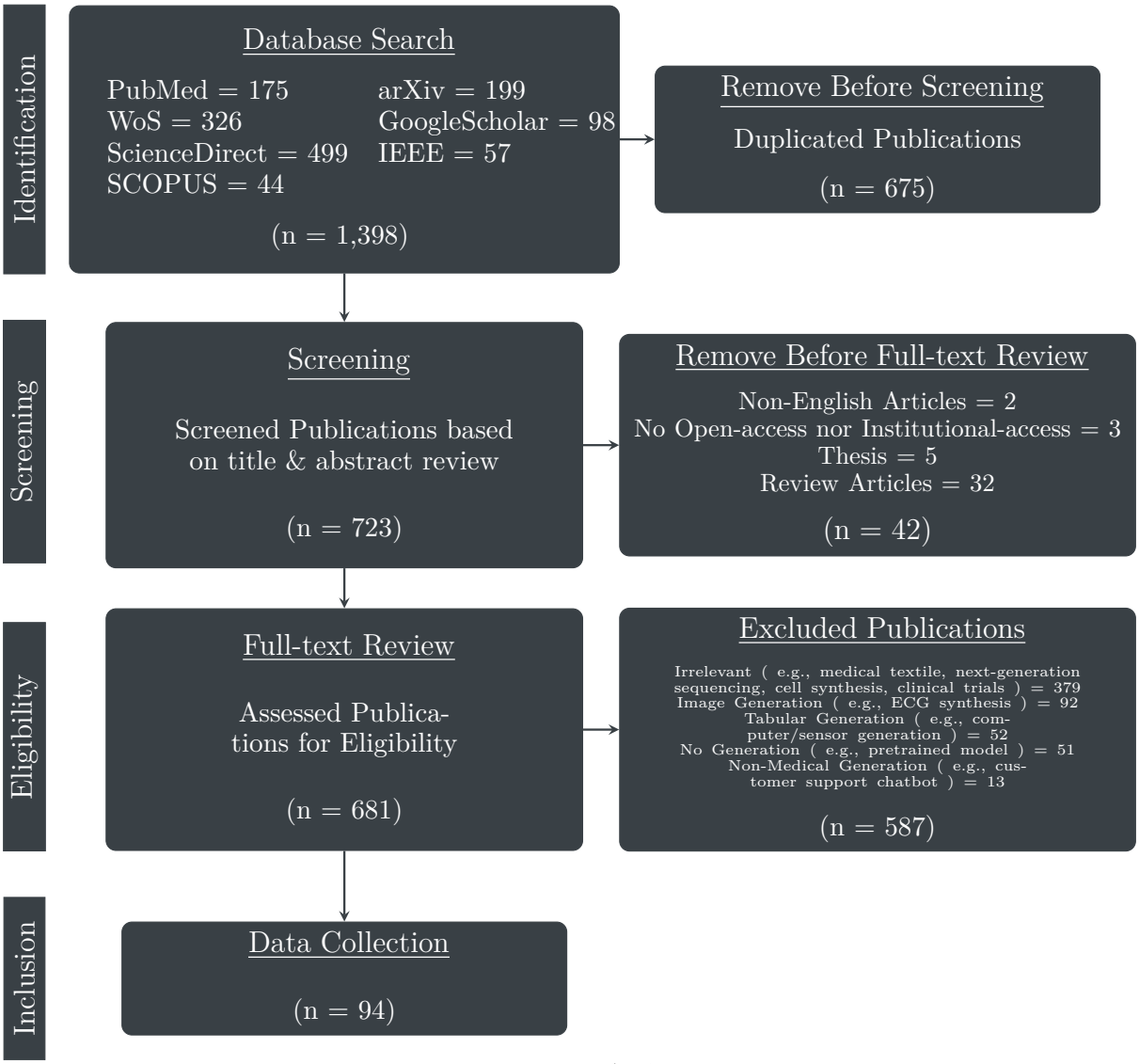

Figure 1: 选择流程图

为了制定搜索策略，首先有必要确定搜索引擎。实际上，我们为这次系统综述选择了七个搜索引擎，包括 PubMed [1] (RRID: SCR _004846)、ScienceDirect [2] (RRID: SCR _027174)、Web of

[1] https://pubmed.ncbi.nlm.nih.gov
[2] https://www.sciencedirect.com

Science [3] (RRID: SCR _022629)、Scopus [4] (RRID: SCR _022613)、IEEE [5] (RRID: SCR _008314)、Google Scholar [6] (RRID: SCR _008878) 和 arXiv [7] (RRID: SCR _005488)。这些搜索引擎将提供广泛的一般科学出版物覆盖范围。另一方面，我们设定了三个搜索限制，包括搜索词、日期和出版类型。首先，对于搜索词，由于我们的评审重点是医学领域的合成文本生成，因此选择了四个词根合成，医学，文本和生成在标题中进行搜索，如下所示：
(generat* OR augment*) AND (synthe* OR pseudo* OR artifici*) AND (medic* OR clinic* OR health*) AND (text* OR record* OR note*)
然后，使用了更丰富的关键短语和进一步的同义词来搜索标题、摘要、关键词和/或主题字段，如下所示：
("synthetic text" OR "synthetic free-text" OR "synthetic unstructured text" OR "synthetic natural language") AND ("health records" OR "medical records" OR "clinical notes") AND ("generative model" OR "data augmentation" OR "re-use") 关于出版日期，我们认为 "Attention is all you need" [15] 中变压器架构的突破（即从 2017 年起）显著促进了文本生成模型的兴起。然而，我们更倾向于将出版日期设置从 2015 年到文献检索时间，即 2024 年 8 月，以包括早期的努力。最后，关于出版类型，我们将搜索范围限定为用英文撰写的同行评审期刊和会议论文。我们还包括了一些来自 arXiv 的非同行评审文章，以获得一些新的见解。

就出版类型而言，我们主要关注的是同行评审的期刊和会议文章。但我们也添加了一些来自 arXiv 的非同行评审文章，因为它包含了最新的趋势。此外，我们包含了以英语撰写的文章，并且通过机构访问或开放访问获取。我们主要关注文章，而不是论文、学位论文和海报。另一方面，由于本次综述的目的是关注生成合成医学自由文本，因此所选择的文章必须满足三个要求。首先，文章必须包括生成医学非结构化自由文本的机制。其次，文章必须展示此类生成的目的。第三，生成的合成文本应该从 EMR/EHR 数据类型（即临床笔记、出院总结、患者记录、实验室报告等）推断出来。

2.1 排除标准

不符合纳入标准的文章将被丢弃。这包括：i) 合成的非医学文本生成，ii) 合成的医学非文本（结构化的、表格的、图像等）生成，iii) 没有生成方法，iv) 论文、学位论文或海报，v) 非英文书写的出版物，最后，vi) 既不是机构访问也不是开放访问的出版物。

2.2 数据提取

一旦根据纳入和排除标准收集了数据，将进行数据提取。提取的数据将反映 RQ1，其中确定生成合成医学文本的目的，以及涉及到的语言和数据集。同时，它还将包括针对 RQ2 的信息，其中讨论了生成技术，并附有每种技术的关键特征。最后，针对 RQ3，将提取和分类用于评估生成的合成医学文本的评估方法。

如图 1 所示，研究开始于流行的数据库，最终得到 1,398 篇文章。在应用标准后，选出了共计 94 篇相关文章（完整的表格列表参见附录 C ）。

[3] http://webofscience.com
[4] https://www.scopus.com
[5] https://ieeexplore.ieee.org
[6] https://scholar.google.com
[7] https://arxiv.org

3 结果

3.1 RQ1：生成合成医学自由文本背后的目的是些什么？涉及到了哪些语言和数据集？

合成医学文本必须为特定目标生成。实质上，文献展示了六个主要目标，包括隐私保护、增强、实用性、辅助写作、注释和语料库建设，如图 **??** 所示（详见附录 **??** ）。
尽管大多数研究使用英语进行生成，但也有一些努力是为了服务于中文 [16–21] 、德语 [22–24] 、日语 [25–27] 、挪威语 [28–30] 、法语 [31] 、荷兰语 [32] 、阿拉伯语 [33] 、印尼语 [2] 和保加利亚语 [34] 。
关于用于训练生成模型的数据，文献中描绘了五种来源，包括私有 EHR/EMR、手动收集/整理、在线资源、提示和公开可用来源，如图 **??** 所示。私有 EHR/EMR 指的是利用医院出院摘要或医疗记录的研究。而手动创建指的是手工收集和整理临床文本的研究。另一方面，在线资源指的是利用电子书、文献或医疗网站（例如，DailyMed [8] 、Mtsamples [9] 、Reddit [10] ）。提示指的是利用 AI 驱动工具（例如，Synthea [11] 、ChatGPT [12] ）的过程，其中一些研究通过这些工具获得了初始数据。最后，公开可用来源指的是黄金标准/基准数据集，包括 MIMIC-III [35] 、MIMIC-CXR [36] 和 IUX-RAY [37] 。

实际上，文献中用于生成合成医学文本的技术可以分为四类，包括手动方法、文本处理、知识源和神经网络模型，如图 2 所示（详见附录 A ）。

3.2 RQ3：用于评估这种合成医学文本的方法有哪些？这些方法如何分类？

为了评估生成的合成文本，文献中讨论了四个方面，包括结构、隐私、相似性和实用性，如图 3 所示。每个评估方面都有其自动（即，基于距离的、统计的和基于神经网络的）和手动（即，人类）指标（详见附录 B ）。

[8] https://dailymed.nlm.nih.gov
[9] https://mtsamples.com
[10] https://www.reddit.com
[11] https://synthea.mitre.org
[12] https://openai.com/index/chatgpt

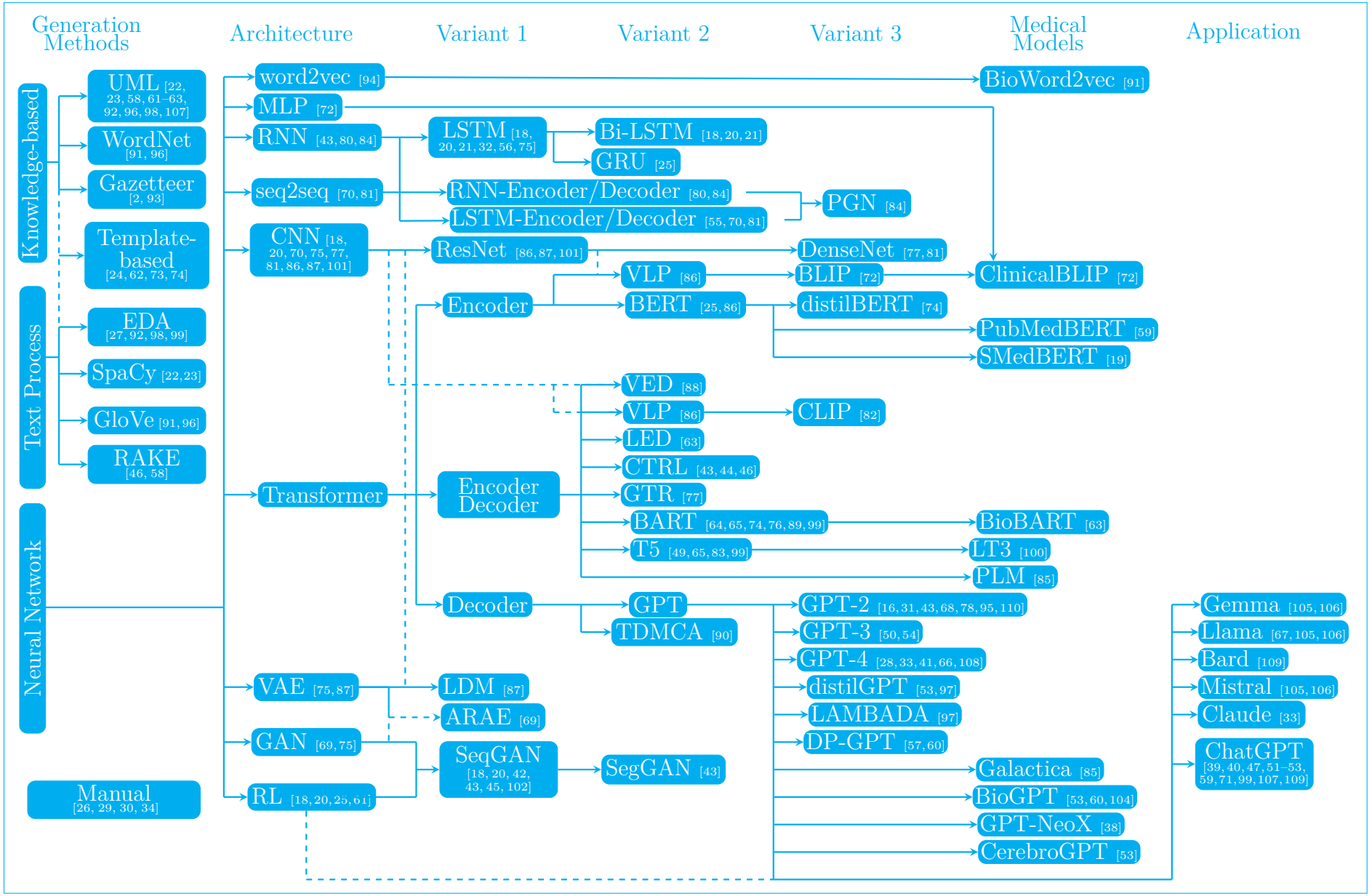

Figure 2: 生成技术

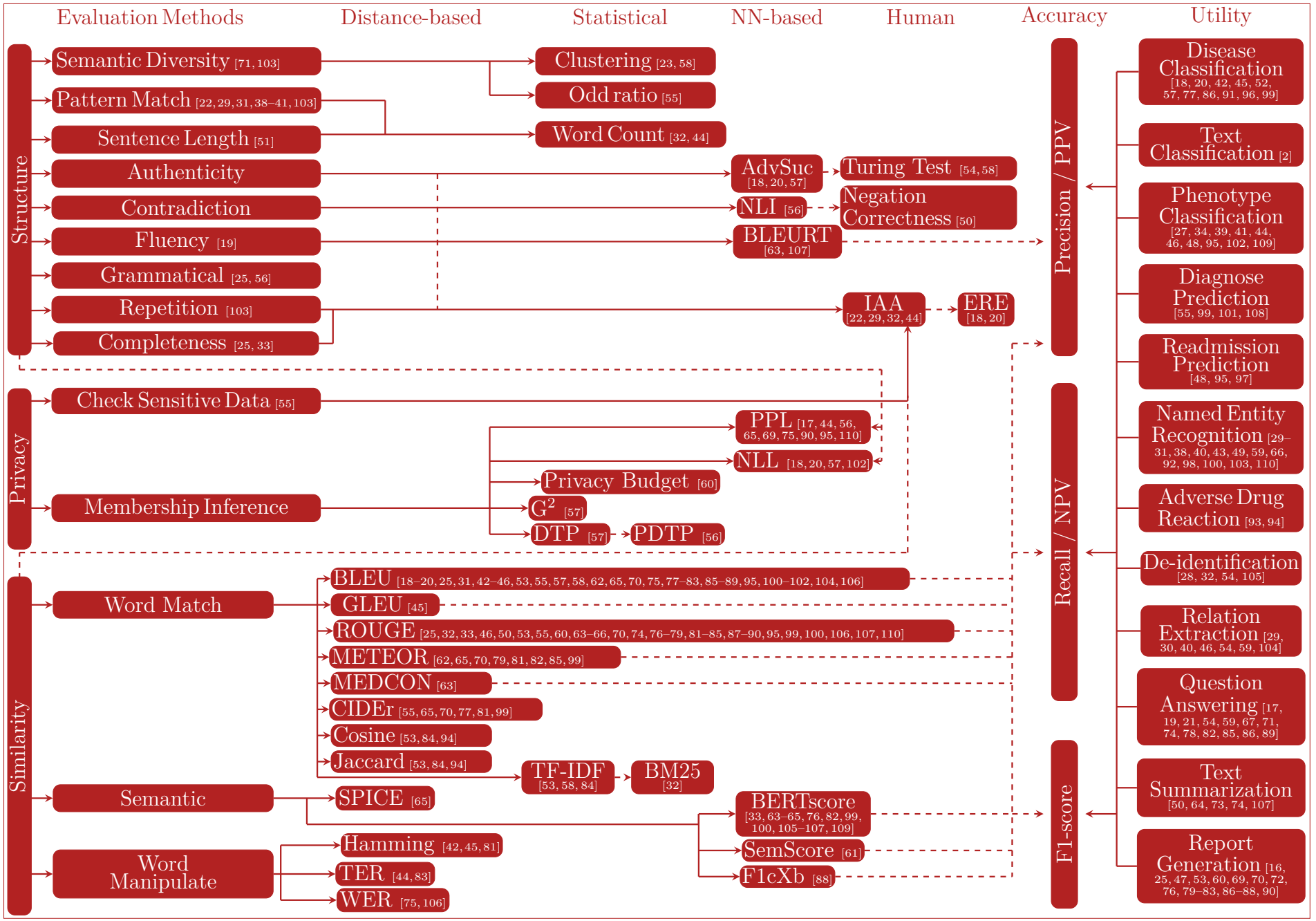

Figure 3: 评估方法

4 讨论

图 4 展示了所选文章的一般统计数据，接下来的小节将对生成医学合成文本的优点和挑战进行广泛讨论。

如文献中所述，生成合成医疗文本将带来宝贵的贡献。例如，临床报告的自动生成以及自动补全将节省临床医生和医师花费在手动填写临床笔记和生成报告上的数百小时。另一方面，合成医疗文本的生成在注释和标记领域具有潜力，它可以通过生成标记的文本来节省人工整理大规模文本语料库的时间。此外，合成生成的医疗文本在隐私能力上表现得比仅使用去识别方法更好。最后，当评估生成的合成医疗文本的实用性时，发现它在某些情况下可以替代原始文本；同时，它几乎在所有情况下展示出增强原始数据的强大能力。这可以作为解决处理不平衡医疗数据时样本不足问题的概念验证。最后，合成生成将显著加速研究进程，因为大多数时间通常消耗在各方之间的协议或甚至是遵守法规政策上。

尽管生成合成医学文本可以克服去识别方法的缺陷，但这种方法仍然重要，应该在生成之前应用。从一项研究中注意到，使用原始数据（即未经去识别）生成合成文本会导致生成现实的身份。即使研究人员得出结论认为这可能是一种针对再次识别的自然保护特征，但在生成之前进行去识别仍然更为安全。

即使是在去识别的文本上进行合成生成，隐私仍然是生成合成临床/医学文本的主要挑战，这归因于多种因素造成的重新识别和成员推测威胁。例如，最不常见和触发性的短语，（如“事故在媒体上广泛报道”）可以很容易地用于追踪个体的身份。此外，在生成过程中从原始文本中复制精确和较长的序列会增加重新识别的可能性。这是因为这些序列中可能涉及独特的治疗方案、药物和测试。在其他情况下，合成文本就像是原始文本的延续，这可能会揭示敏感信息。

然而，最关键的隐私问题是在生成合成文本的训练模型中记忆某些个体的信息。因此，我们还未能见证公开可用且保护隐私的患者数据的发布。

在生成合成医学文本的过程中，文献中描述了若干结构上的缺陷。生成过程中可能出现拼写错误、模糊缩写和重复术语。此外，还观察到语法和句法方面的不正确，例如，代词替换错误（如用“he”代替“she”），医学术语错误（如“丙型肝炎缺乏症”），以及出院记录顺序不正确，这些记录通常包含入院详情、病史、治疗和用药。尽管如此，文本的连贯性问题并未影响合成文本的用途，因为其目的是为了训练机器学习模型，而不是用于教学人类。最后，生成的合成文本中的多样性对后续任务有重大影响。注意到在训练过程中使用小型语料库生成合成医学文本会导致文本的泛化程度较低、文本多样性较低，因此质量较差。

虽然手动生成合成文本可以被视为准确的，但它非常昂贵、繁琐且耗时。相比之下，使用外部知识来源可以促进数据增强，但不能完全生成合成文本。在所有技术中，transformer 模型是最有前途的。然而，已经注意到，通过传统或所谓的 vanilla transformer 生成文本是不可行的，特别是当训练数据相对较小时。这是因为传统 transformer 在较长序列建模任务上存在困难，因为其最大处理能力限制为 512 个 tokens。这也适用于 BERT，它并非用于生成文本，而是通过微调来完成特定任务，例如提取问题的答案或预测文本的预定义标签。

由于使用了大量的数据进行训练，并且可以通过超参数（特别是所谓的温度）来修改其确定性，GPTs 被认为是最合适的模型。值得一提的是，没有观察到医学特定预训练模型在生成文本方面具有明显优势的确凿证据。这种限制的原因在于这些模型对于文本中口语化语气的理解不足。另一方面，GANs 显示出了一些潜力。GAN 的关键特性在于其判别器结构，该结构与生成器结构并行训练。在这种情况下，文本的生成将基于判别器区分真实和合成能力的提升进行调整。尽管一些研究表明 GPTs 的性能超过了 GAN，但 SeqGAN（通过强化学习提升了 GAN）和条件 GAN 仍然很有前景。条件生成的概念在文本生成由情境信息引导或条件化时表现出色。为 GPTs 补充情境信息将非常有前途。

虽然相似性度量是测量生成文本与真实文本接近程度的重要工具，但它们可能与隐私相矛盾。实际上，努力提高相似性可能会导致泄露与个人相关的独特医疗信息，这意味着需要一种折衷机制。另一方面，使用自动隐私度量也存在缺点，因为它们不能完全保证生成的合成文本可以保护隐私。在隐私评估过程中，需要人工评估，其中一组医疗或隐私专业人员可以仔细检查生成的文本。

这也适用于文本结构评估，专家可以通过后处理来纠正无效的语法。最后，需要测试多种实用性以研究生成合成文本的有用性。换句话说，某些任务或实用性可能会因为文本的随机生成而受到负面影响。例如，考虑生成的合成医疗文本在同一患者记录中混淆高血压和低血压等术语，这会导致为表型分类的下游任务生成不适当的数据实例。

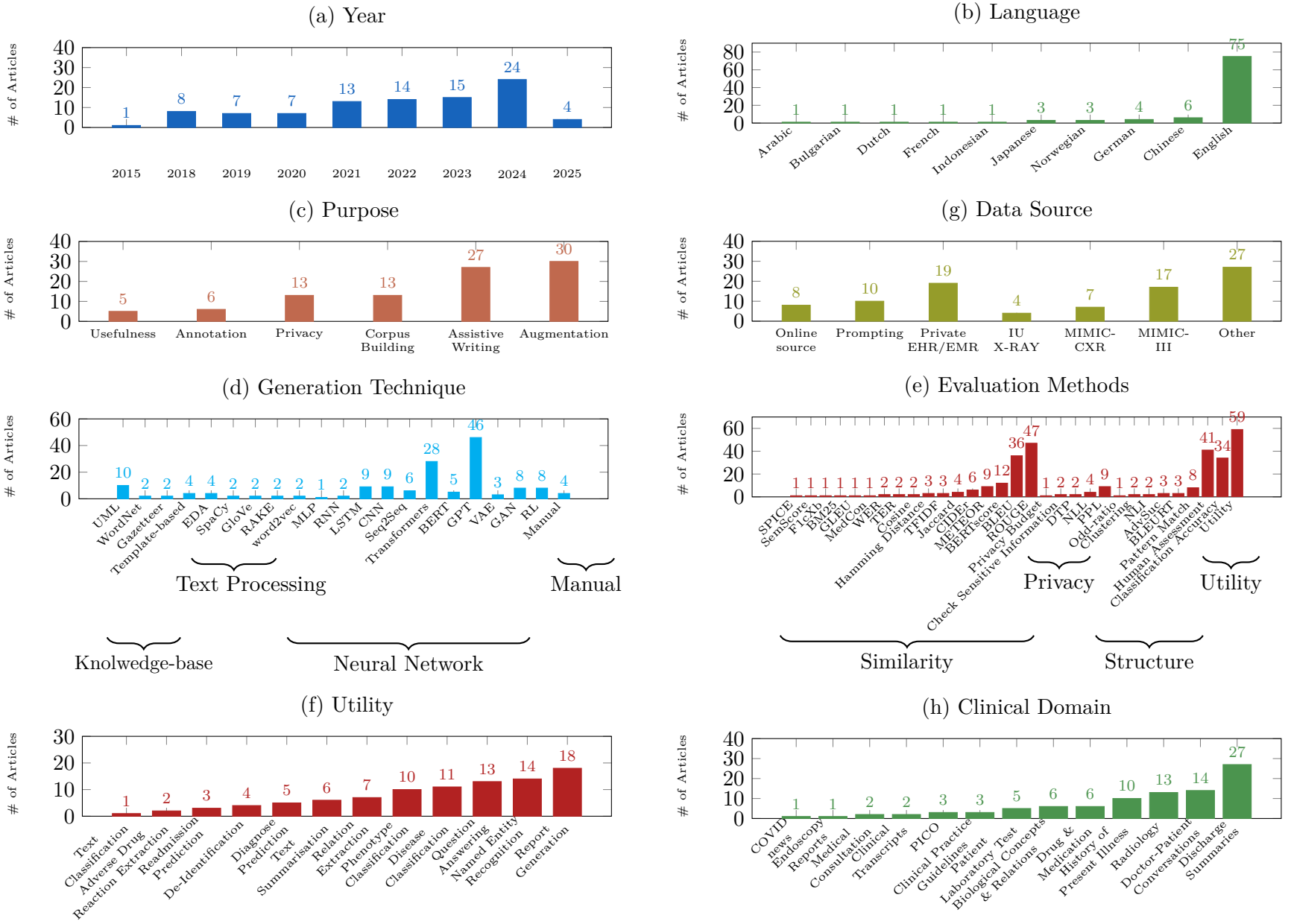

Figure 4: 总体统计

4.1 局限性

本研究中方法综述的主要限制在于搜索方法所固有的可能选择偏差。其中最显著的是“synthetic”一词的多义性，它在不同语境下针对临床试验有不同的含义，这可能会检索到大量与本综述相关的出版物。为此，我们尝试在搜索中使用不同的同义词，例如“artificial”和“pseudo”。另一方面，在搜索中使用如“generat” * 等词根，在某些数据库中效果不佳并且检索结果为零。因此，我们使用了词语本身的多种可能性，如“generate”、“generation”、“generative”和“generator”。在我们的搜索中，可能没有检查进一步的派生词形变化。另一个方法上的限制在于整个自动化医疗文本生成。本研究明确集中于“synthetic”即在文本上进行变更的研究，但实际相关的出版物会更多。所以，我们力求遵循那些在自动化过程中力图对文本进行改动的研究。

本研究综述的组织限制在于由于缺乏统一的评估指标，无法适当地比较生成技术之间的性能。一些研究集中于隐私，而另一些研究关注生成文本与原文本的相似程度。此外，由于隐私问题，缺乏公开可用的电子病历/电子健康记录的数据集也助长了这一限制。另一个限制是缺乏对非神经网络方法的说明，这与对神经网络结构的说明相比显得不足。原因是我们相信神经语言模型在未来潜力巨大，并且将比传统技术获得更多的关注。

在医学领域中，非结构化自由文本在分析电子病历/电子健康记录内的有用信息如表型预测、疾病分类，甚至提取诸如药物、实验室测试、症状等字级别的文物方面的作用从未如此重要。由于隐私问题，共享非结构化医学文本面临巨大挑战，合成生成此类文本可能是一个可替代的解决方案。这不仅能克服隐私问题，还能解决医学非结构化数据集主要面临的样本不平衡问题。本文提供了一份关于合成医学文本生成的全面分类法，其中详尽讨论了生成过程中使用的目的、技术和评估。我们认为，医学合成文本生成在未来不同的下游分析中将发挥重要作用。对于未来的方向，进行生成技术的实证比较将提供对其优缺点的更全面理解。

5 致谢

本工作由 CHIST-ERA 资助项目 CHIST-ERA-22-ORD-02，以及卢森堡国家研究基金（FNR, INTER/CHIST23/17882238/FAIRClinical）资助。

References

[1] Elizabeth Ford, Malcolm Oswald, Lamiece Hassan, Kyle Bozentko, Goran Nenadic, and Jackie Cassell. Should free-text data in electronic medical records be shared for research? a citizens' jury study in the UK. *J Med Ethics*, 46(6):367–377, may 2020.

[2] Febi Siti Sutria Ningsih, Purnomo Husnul Khotimah, Andria Arisal, Andri Fachrur Rozie, Devi Munandar, Dianadewi Riswantini, Ekasari Nugraheni, Wiwin Suwarningsih, and Dian Kurniasari. Synonym-based text generation in restructuring imbalanced dataset for deep learning models. In *2022 5th International Conference on Networking, Information Systems and Security: Envisage Intelligent Systems in 5g//6G-based Interconnected Digital Worlds (NISS)*, pages 1–6, 2022.

[3] Yankun Ren, Jianbin Lin, Siliang Tang, Jun Zhou, Shuang Yang, Yuan Qi, and Xiang Ren. Generating natural language adversarial examples on a large scale with generative models, 2020.

[4] Hajra Murtaza, Musharif Ahmed, Naurin Farooq Khan, Ghulam Murtaza, Saad Zafar, and Ambreen Bano. Synthetic data generation: State of the art in health care domain. *Computer Science Review*, 48:100546, 2023.

[5] Mikel Hernandez, Gorka Epelde, Ane Alberdi, Rodrigo Cilla, and Debbie Rankin. Synthetic data generation for tabular health records: A systematic review. *Neurocomputing*, 493:28–45, 2022.

[6] Peter Eigenschink, Thomas Reutterer, Stefan Vamosi, Ralf Vamosi, Chang Sun, and Klaudius Kalcher. Deep generative models for synthetic data: A survey. *IEEE Access*, 11:47304–47320, 2023.

[7] Moiz Khan Sherwani and Shyam Gopalakrishnan. A systematic literature review: deep learning techniques for synthetic medical image generation and their applications in radiotherapy. *Frontiers in Radiology*, 4:1385742, 2024.

[8] Ghadeer O. Ghosheh, Jin Li, and Tingting Zhu. A survey of generative adversarial networks for synthesizing structured electronic health records. *ACM Comput. Surv.*, 56(6), January 2024.

[9] Emmanuella Budu, Kobra Etminani, Amira Soliman, and Thorsteinn Rögnvaldsson. Evaluation of synthetic electronic health records: A systematic review and experimental assessment. *Neurocomputing*, 603:128253, 2024.

[10] Vasileios C. Pezoulas, Dimitrios I. Zaridis, Eugenia Mylona, Christos Androutsos, Kosmas Apostolidis, Nikolaos S. Tachos, and Dimitrios I. Fotiadis. Synthetic data generation methods in healthcare: A review on open-source tools and methods. *Computational and Structural Biotechnology Journal*, 23:2892–2910, 2024.

[11] Kiduk Kim, Gil-Sun Hong, and Namkug Kim. Primer on generative artificial intelligence and large language models in medical imaging. *Journal of the Korean Society of Radiology*, 85(5):848–860, 2024.

[12] Mahmoud Ibrahim, Yasmina Al Khalil, Sina Amirrajab, Chang Sun, Marcel Breeuwer, Josien Pluim, Bart Elen, Gökhan Ertaylan, and Michel Dumontier. Generative ai for synthetic data across multiple medical modalities: A systematic review of recent developments and challenges. *Computers in Biology and Medicine*, 189:109834, 2025.

[13] Khalid S Khan, Regina Kunz, Jos Kleijnen, and Gerd Antes. Five steps to conducting a systematic review, 2003.

[14] Lindsay S Uman. Systematic reviews and meta-analyses, February 2011.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[16] Junkun Peng, Pin Ni, Jiayi Zhu, Zhenjin Dai, Yuming Li, Gangmin Li, and Xuming Bai. Automatic generation of electronic medical record based on gpt2 model. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 6180–6182, 2019.

[17] Zhijie Qu, Juan Li, Zerui Ma, and Jianqiang Li. Cmed-gpt: Prompt tuning for entity-aware chinese medical dialogue generation, 2023.

[18] Jiaqi Guan, Runzhe Li, Sheng Yu, and Xuegong Zhang. Generation of synthetic electronic medical record text. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, dec 2018.

[19] Fei Xia, Bin Li, Yixuan Weng, Shizhu He, Kang Liu, Bin Sun, Shutao Li, and Jun Zhao. Medconqa: Medical conversational question answering system based on knowledge graphs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, page 148–158. Association for Computational Linguistics, 2022.

[20] Jiaqi Guan, Runzhe Li, Sheng Yu, and Xuegong Zhang. A method for generating synthetic electronic medical record text. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 18(1):173–182, January 2021.

[21] Keyi Huang, Fenda Ji, Wei Lu, and Yue Xiao. Research on text generation of medical intelligent question and answer based on bi-lstm and neural network technology. In *2022 IEEE/ACIS 22nd International Conference on Computer and Information Science (ICIS)*, pages 54–59, 2022.

[22] Florian Borchert, Christina Lohr, Luise Modersohn, Thomas Langer, Markus Follmann, Jan Philipp Sachs, Udo Hahn, and Matthieu-P. Schapranow. Ggponc: A corpus of german medical text with rich metadata based on clinical practice guidelines, 2020.

[23] Luise Modersohn, Stefan Schulz, Christina Lohr, and Udo Hahn. GRASCCO — the first publicly shareable, multiply-alienated german clinical text corpus. In *Studies in Health Technology and Informatics*. IOS Press, aug 2022.

[24] Christina Lohr, Sven Buechel, and Udo Hahn. Sharing copies of synthetic clinical corpora without physical distribution — a case study to get around IPRs and privacy constraints featuring the German JSYNCC corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[25] Toru Nishino, Ryota Ozaki, Yohei Momoki, Tomoki Taniguchi, Ryuji Kano, Norihisa Nakano, Yuki Tagawa, Motoki Taniguchi, Tomoko Ohkuma, and Keigo Nakamura. Reinforcement learning with imbalanced dataset for data-to-text medical report generation. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2223–2236, Online, November 2020. Association for Computational Linguistics.

[26] Rina Kagawa, Yukino Baba, and Hideo Tsurushima. A practical and universal framework for generating publicly available medical notes of authentic quality via the power of crowds. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 3534–3543, 2021.

[27] Toshiharu Igarashi and Misato Nihei. Cognitive assessment of japanese older adults with text data augmentation. *Healthcare*, 10(10):2051, oct 2022.

[28] Jørgen Aarmo Lund, Karl Øyvind Mikalsen, Joel Burman, Ashenafi Zebene Woldaregay, and Robert Jenssen. Instruction-guided deidentification with synthetic test cases for norwegian clinical text. In Tetiana Lutchyn, Adín Ramírez Rivera, and Benjamin Ricaud, editors, *Proceedings of the 5th Northern Lights Deep Learning Conference (NLDL)*, volume 233 of *Proceedings of Machine Learning Research*, pages 145–152. PMLR, 09–11 Jan 2024.

[29] Taraka Rama, Pål Brekke, Øystein Nytrø, and Lilja Øvrelid. Iterative development of family history annotation guidelines using a synthetic corpus of clinical text. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*. Association for Computational Linguistics, 2018.

[30] Pål H. Brekke, Taraka Rama, Ildikó Pilán, Øystein Nytrø, and Lilja Øvrelid. Synthetic data for annotation and extraction of family history information from clinical text. *J Biomed Semant*, 12(1), jul 2021.

[31] Nicolas Hiebel, Olivier Ferret, Karen Fort, and Aurélie Névéol. Can synthetic text help clinical named entity recognition? a study of electronic health records in French. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2320–2338, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[32] Claudia Alessandra Libbi, Jan Trienes, Dolf Trieschnigg, and Christin Seifert. Generating synthetic training data for supervised de-identification of electronic health records. *Future Internet*, 13(5):136, may 2021.

[33] Mariam ALMutairi, Lulwah AlKulaib, Melike Aktas, Sara Alsalamah, and Chang-Tien Lu. Synthetic Arabic medical dialogues using advanced multi-agent LLM techniques. In Nizar Habash, Houda Bouamor, Ramy Eskander, Nadi Tomeh, Ibrahim Abu Farha, Ahmed Abdelali, Samia Touileb, Injy Hamed, Yaser Onaizan, Bashar Alhafni, Wissam Antoun, Salam Khalifa, Hatem Haddad, Imed Zitouni, Badr AlKhamissi, Rawan Almatham, and Khalil Mrini, editors, *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 11–26, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[34] Boris Velichkov, Kristina Ivanova, Valeri Hristov, Ivan Borisov, Alexander Peychev, Ivan Koychev, and Svetla Boytcheva. AI-driven approach for automatic synthetic patient status corpus generation. In *2020 4th International Conference on Artificial Intelligence and Virtual Reality*. ACM, oct 2020.

[35] Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1), May 2016.

[36] Alistair E. W. Johnson, Lucas Bulgarelli, and Tom J. Pollard. Deidentification of free-text medical records using pre-trained bidirectional transformers. In *Proceedings of the ACM Conference on Health, Inference, and Learning*. ACM, apr 2020.

[37] Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 07 2015.

[38] Johann Frei and Frank Kramer. Annotated dataset creation through general purpose language models for non-english medical nlp, 2022.

[39] Isabelle Lorge, Dan W. Joyce, Niall Taylor, Alejo Nevado-Holgado, Andrea Cipriani, and Andrey Kormilitzin. Detecting the clinical features of difficult-to-treat depression using synthetic data from large language models, 2024.

[40] Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. Does synthetic data generation of llms help clinical text mining?, 2023.

[41] Rumeng Li, Xun Wang, and Hong Yu. Two directions for clinical data generation with large language models: Data-to-label and label-to-data. In *Findings of the Association for*

Computational Linguistics: EMNLP 2023, page 7129–7143. Association for Computational Linguistics, 2023.

[42] Suranga N Kasthurirathne, Gregory Dexter, and Shaun J Grannis. An adversorial approach to enable re-use of machine learning models and collaborative research efforts using synthetic unstructured free-text medical data, August 2019.

[43] Jianfu Li, Yujia Zhou, Xiaoqian Jiang, Karthik Natarajan, Serguei Vs Pakhomov, Hongfang Liu, and Hua Xu. Are synthetic clinical notes useful for real natural language processing tasks: A case study on clinical entity recognition. *Journal of the American Medical Informatics Association*, 28(10):2193–2201, jul 2021.

[44] Julia Ive, Natalia Viani, Joyce Kam, Lucia Yin, Somain Verma, Stephen Puntis, Rudolf N. Cardinal, Angus Roberts, Robert Stewart, and Sumithra Velupillai. Generation and evaluation of artificial mental health records for natural language processing. *npj Digit. Med.*, 3(1), may 2020.

[45] Suranga N Kasthurirathne, Gregory Dexter, and Shaun J Grannis. Generative adversarial networks for creating synthetic free-text medical data: A proposal for collaborative research and re-use of machine learning models, May 2021.

[46] Zixu Wang, Julia Ive, Sumithra Velupillai, and Lucia Specia. Is artificial data useful for biomedical natural language processing algorithms? In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 240–249, Florence, Italy, aug 2019. Association for Computational Linguistics.

[47] Reece Alexander James Clough, William Anthony Sparkes, Oliver Thomas Clough, Joshua Thomas Sykes, Alexander Thomas Steventon, and Kate King. Transforming healthcare documentation: harnessing the potential of ai to generate discharge summaries. *BJGP Open*, 8(1), 2024.

[48] Ali Amin-Nejad, Julia Ive, and Sumithra Velupillai. Transformer models trained on mimic-iii to generate synthetic patient notes, 2020.

[49] Anthony Hughes and Xingyi Song. Identifying and aligning medical claims made on social media with medical evidence, 2024.

[50] Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. Medically aware gpt-3 as a data generator for medical dialogue summarization, 2021.

[51] Isidoro Calvo-Lorenzo and Iker Uriarte-Llano. Generación masiva de historias clínicas sintéticas con chatgpt: un ejemplo en fractura de cadera. *Medicina Clínica*, 162(11):549–554, 2024.

[52] Onkar Litake, Brian H Park, Jeffrey L Tully, and Rodney A Gabriel. Constructing synthetic datasets with generative artificial intelligence to train large language models to classify acute renal failure from clinical notes. *Journal of the American Medical Informatics Association*, 31(6):1404–1410, 04 2024.

[53] Byoung-Doo Oh, Gi-Youn Kim, Chulho Kim, and Yu-Seop Kim. How to use language models for synthetic text generation in cerebrovascular disease-specific medical reports. In Ameet

Deshpande, EunJeong Hwang, Vishvak Murahari, Joon Sung Park, Diyi Yang, Ashish Sabharwal, Karthik Narasimhan, and Ashwin Kalyan, editors, *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 10–17, St. Julians, Malta, March 2024. Association for Computational Linguistics.

[54] Cheng Peng, Xi Yang, Aokun Chen, Kaleb E. Smit h, Nima PourNejatian, Anthony B. Costa, Cheryl Martin, Mona G. Flores, Ying Zhang, Tanja Magoc, Gloria Lipori, Duane A. Mitchell, Naykky S. Ospina, Mustafa M. Ahmed, William R. Hogan, Elizabeth A. Shenkman, Yi Guo, Jiang Bian, and Yonghui Wu. A study of generative large language model for medical research and healthcare. *npj Digital Medicine*, 6(1), November 2023.

[55] Scott H. Lee. Natural language generation for electronic health records. *npj Digital Med*, 1(1), nov 2018.

[56] Oren Melamud and Chaitanya Shivade. Towards automatic generation of shareable synthetic clinical notes using neural language models. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 35–45, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[57] Md Momin Al Aziz, Tanbir Ahmed, Tasnia Faequa, Xiaoqian Jiang, Yiyu Yao, and Noman Mohammed. Differentially private medical texts generation using generative neural networks. *ACM Transactions on Computing for Healthcare*, 3(1):1–27, jan 2022.

[58] Nina Zhou, Qiucheng Wu, Zewen Wu, Simeone Marino, and Ivo D. Dinov. DataSifterText: Partially synthetic text generation for sensitive clinical notes. *J Med Syst*, 46(12), nov 2022.

[59] Ran Xu, Hejie Cui, Yue Yu, Xuan Kan, Wenqi Shi, Yuchen Zhuang, May Dongmei Wang, Wei Jin, Joyce Ho, and Carl Yang. Knowledge-infused prompting: Assessing and advancing clinical text data generation with large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15496–15523, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[60] Agathe Zecevic, Xinyue Zhang, Sebastian Zeki, and Angus Roberts. Generation and evaluation of synthetic endoscopy free-text reports with differential privacy. In Dina Demner-Fushman, Sophia Ananiadou, Makoto Miwa, Kirk Roberts, and Junichi Tsujii, editors, *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 14–24, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[61] Simon Meoni, Éric De la Clergerie, and Théo Ryffel. Generating synthetic documents with clinical keywords: A privacy-sensitive methodology. In Dina Demner-Fushman, Sophia Ananiadou, Paul Thompson, and Brian Ondov, editors, *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024*, pages 115–123, Torino, Italia, may 2024. ELRA and ICCL.

[62] Edmon Begoli, Kris Brown, Sudarshan Srinivas, and Suzanne Tamang. SynthNotes: A generator framework for high-volume, high-fidelity synthetic mental health notes. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, dec 2018.

[63] Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific Data*, 10(1), September 2023.

[64] Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. An empirical study of clinical note generation from doctor-patient encounters. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[65] Steven Y. Feng, Vivek Khetan, Bogdan Sacaleanu, Anatole Gershman, and Eduard Hovy. Chard: Clinical health-aware reasoning across dimensions for text generation models, 2022.

[66] Avijit Mitra, Emily Druhl, Raelene Goodwin, and Hong Yu. Synth-sbdh: A synthetic dataset of social and behavioral determinants of health for clinical text, 2024.

[67] Konstantin Kotschenreuther. Ehr-ds-qa: A synthetic qa dataset derived from medical discharge summaries for enhanced medical information retrieval systems, 2024.

[68] Jason Walonoski, Dylan Hall, Karen M. Bates, M. Heath Farris, Joseph Dagher, Matthew E. Downs, Ryan T. Sivek, Ben Wellner, Andrew Gregorowicz, Marc Hadley, Francis X. Campion, Lauren Levine, Kevin Wacome, Geoff Emmer, Aaron Kemmer, Maha Malik, Jonah Hughes, Eldesia Granger, and Sybil Russell. The "coherent data set" : Combining patient data and imaging in a comprehensive, synthetic health record. *Electronics*, 11(8):1199, April 2022.

[69] Graham Spinks and Marie-Francine Moens. Generating continuous representations of medical texts, 2018.

[70] Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2018.

[71] Kelly Reynolds, Daniel Nadelman, Joseph Durgin, Stephen Ansah-Addo, Daniel Cole, Rachel Fayne, Jane Harrell, Madison Ratycz, Mason Runge, Amanda Shepard-Hayes, Daniel Wenzel, and Trilokraj Tejasvi. Comparing the quality of chatgpt- and physician-generated responses to patients' dermatology questions in the electronic medical record. *Clinical and Experimental Dermatology*, 49(7):715–718, 01 2024.

[72] Jia Ji, Yongshuai Hou, Xinyu Chen, Youcheng Pan, and Yang Xiang. Vision-language model for generating textual descriptions from clinical images: Model development and validation study. *JMIR Form Res*, 8:e32690, Feb 2024.

[73] Goldstein Ayelet and Shahar Yuval. Generation of natural-language textual summaries from longitudinal clinical records. In *MEDINFO 2015: eHealth-enabled Health*. IOS Press, 2015.

[74] Binh-Nguyen Nguyen, Hoang-Quynh Le, and Duy-Cat Can. Enhancing clinical note generation from doctor-patient conversations through semantic partition-oriented summarization. In *2023 15th International Conference on Knowledge and Systems Engineering (KSE)*, pages 1–6, 2023.

[75] Anjanava Biswas and Wrick Talukdar. Enhancing clinical documentation with synthetic data: Leveraging generative models for improved accuracy. *International Journal of Innovative Science and Research Technology (IJISRT)*, page 1553–1566, June 2024.

[76] Wang Zhao, Dongxiao Gu, Xuejie Yang, Meihuizi Jia, Changyong Liang, Xiaoyu Wang, and Oleg Zolotarev. Medt2t: An adaptive pointer constrain generating method for a new medical text-to-table task. *Future Generation Computer Systems*, 161:586–600, 2024.

[77] Christy Y. Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. Knowledge-driven encode, retrieve, paraphrase for medical image report generation, 2019.

[78] Anirban Karak and Kaustuv Kunal. Implementation of gpt models for text generation in healthcare domain. In *2023 International Conference on New Frontiers in Communication, Automation, Management and Security (ICCAMS)*, volume 1, pages 1–6, 2023.

[79] Santosh Sanjeev, Fadillah Adamsyah Maani, Arsen Abzhanov, Vijay Ram Papineni, Ibrahim Almakky, Bartłomiej W. Papież, and Mohammad Yaqub. Tibix: Leveraging temporal information for bidirectional x-ray and report generation, 2024.

[80] Litton J Kurisinkel and Nancy Chen. Set to ordered text: Generating discharge instructions from medical billing codes. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6165–6175, Hong Kong, China, November 2019. Association for Computational Linguistics.

[81] Yi Zhou, Lei Huang, Tao Zhou, Huazhu Fu, and Ling Shao. Visual-textual attentive semantic consistency for medical report generation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3965–3974, 2021.

[82] Fan Bai, Yuxin Du, Tiejun Huang, Max Q. H. Meng, and Bo Zhao. M3d: Advancing 3d medical image analysis with multi-modal large language models, 2024.

[83] Heng-Yi Wu, Jingqing Zhang, Julia Ive, Tong Li, Vibhor Gupta, Bingyuan Chen, and Yike Guo. Medical scientific table-to-text generation with human-in-the-loop under the data sparsity constraint, 2022.

[84] Claudia Meyer, Daniel Adkins, Koyena Pal, Ruggero Galici, Augusto Garcia-Agundez, and Carsten Eickhoff. Neural text generation in regulatory medical writing. *Frontiers in Pharmacology*, 14, February 2023.

[85] Zhiyuan Liu, An Zhang, Hao Fei, Enzhi Zhang, Xiang Wang, Kenji Kawaguchi, and Tat-Seng Chua. Prott3: Protein-to-text generation for text-based protein understanding, 2024.

[86] Jong Hak Moon, Hyungyung Lee, Woncheol Shin, Young-Hak Kim, and Edward Choi. Multimodal understanding and generation for medical images and text via vision-language pretraining. *IEEE Journal of Biomedical and Health Informatics*, 26(12):6070–6080, 2022.

[87] Chenlu Zhan, Yu Lin, Gaoang Wang, Hongwei Wang, and Jian Wu. Medm2g: Unifying medical multi-modal generation via cross-guided diffusion with visual invariant, 2024.

[88] Daniel Parres, Alberto Albiol, and Roberto Paredes. Improving radiology report generation quality and diversity through reinforcement learning and text augmentation. *Bioengineering*, 11(4):351, April 2024.

[89] Wei Sun, Mingxiao Li, Damien Sileo, Jesse Davis, and Marie-Francine Moens. Generating explanations in medical question-answering by expectation maximization inference over evidence, 2023.

[90] Peter J. Liu. Learning to write notes in electronic health records, 2018.

[91] Mahdi Abdollahi, Xiaoying Gao, Yi Mei, Shameek Ghosh, and Jinyan Li. A dictionary-based oversampling approach to clinical document classification on small and imbalanced dataset. In *2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. IEEE, dec 2020.

[92] Abdul Majeed Issifu and Murat Can Ganiz. A simple data augmentation method to improve the performance of named entity recognition models in medical domain. In *2021 6th International Conference on Computer Science and Engineering (UBMK)*. IEEE, sep 2021.

[93] Carson Tao, Kahyun Lee, Michele Filannino, and Özlem Uzuner. An exploratory study on pseudo-data generation in prescription and adverse drug reaction extraction, August 2019.

[94] Tomoki Ishikawa, Takahiro Yakoh, and Hisashi Urushihara. An NLP-inspired data augmentation method for adverse event prediction using an imbalanced healthcare dataset. *IEEE Access*, 10:81166–81176, 2022.

[95] Ali Amin-Nejad, Julia Ive, and Sumithra Velupillai. Exploring transformer text generation for medical dataset augmentation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4699–4708, Marseille, France, May 2020. European Language Resources Association.

[96] Mahdi Abdollahi, Xiaoying Gao, Yi Mei, Shameek Ghosh, Jinyan Li, and Michael Narag. Substituting clinical features using synthetic medical phrases: Medical text data augmentation techniques. *Artificial Intelligence in Medicine*, 120:102167, oct 2021.

[97] Qiuhao Lu, Dejing Dou, and Thien Huu Nguyen. Textual data augmentation for patient outcomes prediction. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, dec 2021.

[98] Tian Kang, Adler Perotte, Youlan Tang, Casey Ta, and Chunhua Weng. UMLS-based data augmentation for natural language processing of clinical research literature. *Journal of the American Medical Informatics Association*, 28(4):812–823, dec 2020.

[99] Atif Latif and Jihie Kim. Evaluation and analysis of large language models for clinical text augmentation and generation. *IEEE Access*, 12:48987–48996, 2024.

[100] Samuel Belkadi, Nicolo Micheletti, Lifeng Han, Warren Del-Pinto, and Goran Nenadic. Generating medical prescriptions with conditional transformer, 2023.

[101] Artur Gomes Barreto, Juliana Martins de Oliveira, Francisco Nauber Bernardo Gois, Paulo Cesar Cortez, and Victor Hugo Costa de Albuquerque. A new generative model for textual descriptions of medical images using transformers enhanced with convolutional neural networks. *Bioengineering*, 10(9), 2023.

[102] ML Tlachac, Walter Gerych, Kratika Agrawal, Benjamin Litterer, Nicholas Jurovich, Saitheeraj Thatigotla, Jidapa Thadajarassiri, and Elke A. Rundensteiner. Text generation to aid depression detection: A comparative study of conditional sequence generative adversarial networks. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 2804–2813, 2022.

[103] Robert Horton, Maryam Hosseinabadi, Alexandre Vilcek, Wolfgang Pauli, and Mario Inchiosa. Generating and evaluating simulated medical notes: Getting a natural language

generation model to give you what you want. In *Document Intelligence Workshop at KDD, 2021, Virtual Event*, pages 1–5, 2021.

[104] Maxime Delmas, Magdalena Wysocka, and André Freitas. Relation extraction in underexplored biomedical domains: A diversity-optimized sampling and synthetic data generation approach. *Computational Linguistics*, 50(3):953–1000, 09 2024.

[105] Sanjeet Singh, Shreya Gupta, Niralee Gupta, Naimish Sharma, Lokesh Srivastava, Vibhu Agarwal, and Ashutosh Modi. Generation and de-identification of indian clinical discharge summaries using llms, 2024.

[106] Kuluhan Binici, Abhinav Ramesh Kashyap, Viktor Schlegel, Andy T. Liu, Vijay Prakash Dwivedi, Thanh-Tung Nguyen, Xiaoxue Gao, Nancy F. Chen, and Stefan Winkler. Medsage: Enhancing robustness of medical dialogue summarization to asr errors with llm-generated synthetic dialogues, 2024.

[107] Viktor Schlegel, Hao Li, Yuping Wu, Anand Subramanian, Thanh-Tung Nguyen, Abhinav Ramesh Kashyap, Daniel Beck, Xiaojun Zeng, Riza Theresa Batista-Navarro, Stefan Winkler, and Goran Nenadic. Pulsar at mediqa-sum 2023: Large language models augmented by synthetic dialogue convert patient dialogues to medical records. volume 3497, page 1668 –1679, 2023. Cited by: 1.

[108] Jihye Kim Scroggins, Maxim Topaz, Jiyoun Song, and Maryam Zolnoori. Does synthetic data augmentation improve the performances of machine learning classifiers for identifying health problems in patient–nurse verbal communications in home healthcare settings? *Journal of Nursing Scholarship*, 57(1):47–58, July 2024.

[109] İrfan AYGÜN and Mehmet KAYA. Use of large language models for medical synthetic data generation in mental illness. In *7th IET Smart Cities Symposium (SCS 2023)*, volume 2023, pages 652–656, 2023.

[110] Karan Aggarwal, Henry Jin, and Aitzaz Ahmad. ECG-QALM: Entity-controlled synthetic text generation using contextual Q & A for NER. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5649–5660, Toronto, Canada, July 2023. Association for Computational Linguistics.

[111] Mhairi Aitken, Jenna de St. Jorre, Claudia Pagliari, Ruth Jepson, and Sarah Cunningham-Burley. Public responses to the sharing and linkage of health data for research purposes: a systematic review and thematic synthesis of qualitative studies. *BMC Medical Ethics*, 17(1), November 2016.

[112] Jessica Stockdale, Jackie Cassell, and Elizabeth Ford. "giving something back": A systematic review and ethical enquiry into public views on the use of patient data for research in the united kingdom and the republic of ireland. *Wellcome Open Research*, 3:6, January 2019.

[113] Hassan Ramchoun, Mohammed Amine, Janati Idrissi, Youssef Ghanou, and Mohamed Ettaouil. Multilayer perceptron: Architecture optimization and training. *International Journal of Interactive Multimedia and Artificial Intelligence*, 4(1):26, 2016.

[114] Yoav Goldberg and Omer Levy. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method, 2014.

[115] Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific Data*, 6(1), May 2019.

[116] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization, 2014.

[117] Alex Graves. *Long Short-Term Memory*, pages 37–45. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[118] Ebru Arisoy, Abhinav Sethy, Bhuvana Ramabhadran, and Stanley Chen. Bidirectional recurrent neural network language models for automatic speech recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5421–5425, 2015.

[119] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014.

[120] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation, 2015.

[121] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks, 2017.

[122] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013.

[123] Xianxu Hou, Linlin Shen, Ke Sun, and Guoping Qiu. Deep feature consistent variational autoencoder. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1133–1141, 2017.

[124] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks, 2015.

[125] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[126] Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. Ctrl: A conditional transformer language model for controllable generation, 2019.

[127] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm, 2022.

[128] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, 2019.

[129] Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J. Kim. Graph transformer networks, 2020.

[130] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020.

[131] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019.

[132] Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. Biobart: Pretraining and evaluation of a biomedical generative language model, 2022.

[133] Jorge Gabín, M. Eduardo Ares, and Javier Parapar. Enhancing automatic keyphrase labelling with text-to-text transfer transformer (t5) architecture: A framework for keyphrase generation and filtering, 2024.

[134] Samuel Belkadi, Nicolo Micheletti, Lifeng Han, Warren Del-Pinto, and Goran Nenadic. Generating medical prescriptions with conditional transformer, 2023.

[135] Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos Jr, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. Large language models generate functional protein sequences across diverse families. *Nature biotechnology*, 41(8):1099–1106, 2023.

[136] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm, 2021.

[137] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[138] Jia Ji, Yongshuai Hou, Xinyu Chen, Youcheng Pan, and Yang Xiang. Vision-language model for generating textual descriptions from clinical images: Model development and validation study. *JMIR Formative Research*, 8:e32690, February 2024.

[139] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[140] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2019.

[141] Qing Han, Shubo Tian, and Jinfeng Zhang. A pubmedbert-based classifier with data augmentation strategy for detecting medication mentions in tweets, 2021.

[142] Taolin Zhang, Zerui Cai, Chengyu Wang, Minghui Qiu, Bite Yang, and Xiaofeng He. Smedbert: A knowledge-enhanced pre-trained language model with structured semantics for medical text mining, 2021.

[143] Tianda Li, Yassir El Mesbahi, Ivan Kobyzev, Ahmad Rashid, Atif Mahmud, Nithin Anchuri, Habib Hajimolahoseini, Yang Liu, and Mehdi Rezagholizadeh. A short study on compressing decoder-based language models, 2021.

[144] Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. Not enough data? deep learning to the rescue!, 2019.

[145] Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating wikipedia by summarizing long sequences, 2018.

[146] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409, 09 2022.

[147] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science, 2022.

[148] Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. Gpt-neox-20b: An open-source autoregressive language model, 2022.

[149] Byoung-Doo Oh, Gi-Youn Kim, Chulho Kim, and Yu-Seop Kim. How to use language models for synthetic text generation in cerebrovascular disease-specific medical reports. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 10–17, 2024.

[150] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[151] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient, 2016.

[152] Yuxi Li. Deep reinforcement learning: An overview, 2017.

[153] Xingyuan Chen, Yanzhe Li, Peng Jin, Jiuhua Zhang, Xinyu Dai, Jiajun Chen, and Gang Song. Adversarial sub-sequence for text generation, 2019.

[154] Jake Zhao, Yoon Kim, Kelly Zhang, Alexander M. Rush, and Yann LeCun. Adversarially regularized autoencoders, 2017.

[155] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[156] Stanislau Semeniuta, Aliaksei Severyn, and Sylvain Gelly. On accurate evaluation of gans for language generation, 2018.

[157] Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models, 2018.

[158] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation, 2016.

[159] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[160] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.

[161] Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. CharacTer: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 505–510, Berlin, Germany, August 2016. Association for Computational Linguistics.

[162] Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA, August 8-12 2006. Association for Machine Translation in the Americas.

[163] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation, 2014.

[164] Stephen Robertson. Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*, 60(5):503–520, October 2004.

[165] Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

[166] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[167] Ansar Aynetdinov and Alan Akbik. Semscore: Automated evaluation of instruction-tuned llms based on semantic textual similarity. *arXiv preprint arXiv:2401.17072*, 2024.

[168] Daniel Parres, Alberto Albiol, and Roberto Paredes. Improving radiology report generation quality and diversity through reinforcement learning and text augmentation. *Bioengineering*, 11(4):351, 2024.

[169] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer, 2016.

[170] Marry L. McHugh. Interrater reliability: the kappa statistic. *Biochemia Medica*, pages 276–282, 2012.

[171] Paul Rayson, Damon Berridge, and Brian Francis. Extending the cochran rule for the comparison of word frequencies between corpora. In *7th International Conference on Statistical analysis of textual data (JADT 2004)*, pages 926–936, 2004.

[172] Vlado Keselj. Speech and language processing (second edition) daniel jurafsky and james h. martin (stanford university and university of colorado at boulder) pearson prentice hall, 2009, xxxi+988 pp; hardbound, ISBN 978-0-13-187321-6, $115.00. *Computational Linguistics*, 35(3):463–466, September 2009.

[173] Yunhui Long, Vincent Bindschaedler, and Carl A. Gunter. Towards measuring membership privacy, 2017.

[174] Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. Bleurt: Learning robust metrics for text generation, 2020.

在医学领域，隐私起着至关重要的作用，其中保护个人信息是一个重要任务。像英美和欧洲国家这样的西方文化认为个人有权拥有自己的个人数据，因此制定了相关法规来保护这种所有权。在这种情况下，隐私侵犯仅仅是指通过多种威胁披露个人信息，包括重新识别、成员推断和属性披露。重新识别（也称为去匿名化）是指能够从个人的医疗数据中揭示其身份。成员推断是一种攻击，旨在测试某个人是否是数据收集的一部分，使用该个人的特征组合。这不仅限于数据集，也可能涉及预训练的机器学习模型。属性披露指的是揭示个人的敏感属性。

美国健康保险便利与责任法案（HIPAA）以及欧洲联盟的通用数据保护条例（GDPR）共同规定了一项规则，针对患者电子健康记录（包括临床文本）的发布/共享。这样的法规在严格的去识别条件下允许发布/共享这些记录。去识别是消除个人识别信息的过程，例如姓名、地点、电话号码、日期（例如，出生日期、入院日期等）、医院等，这些可以被称为个人可识别信息（PII）或受保护的健康信息（PHI）。公众对结构化医疗数据的看法已被广泛研究，不同的研究表明参与者更愿意通过统计图表分享和发布他们的医疗信息。然而，大多数关于分享医疗数据的公众调查研究并没有明确区分数据是结构化还是非结构化。在这方面，Ford 等人进行了一项研究，明确了公众对在英国公民陪审团研究中分享非结构化医疗自由文本的看法。研究结果显示，公众更同意出于研究目的分享医疗自由文本，但需要在一个严格条件下进行，其中包括适当的去身份化过程。然而，在某些情况下，去身份化可能仍不足以防止通过独特的临床事件组合或独特的临床数据元素重新识别个体。因此，合成生成的医疗文本可能是一个替代解决方案。

.1 增强

NLP 方法通常被认为是数据密集型或数据贪婪型，这表明为了获得更好的性能，需要更多的样本。此外，共享语言资源，如语料库、本体、词典、注释和词汇资源，是 NLP 基础设施向其任务进展的基石。从这个意义上说，隐私会妨碍通过共享 NLP 资源可以实现的宝贵贡献。另一方面，处理医疗文本数据，例如用于诊断/疾病分类的任务通常需要处理欠采样的问题。这是因为某些疾病是罕见的，并且与有限数量的记录/文档相关联，这就产生了过采样的需求，尤其是针对少数类/疾病 [2] 。因此，生成合成医疗文本可以通过增加与少数疾病类相关的实例数量来促进解决这个问题。

一些自然语言处理任务，如命名实体识别 (NER)、不良反应 (ADR)、强化学习 (RL)，甚至自动去标识化，通常需要由专家进行的真实数据标注。假设文本中存在一个生物医学关系，比如蛋白质-蛋白质相互作用。识别这种关系需要生物学领域的专家，专家需要标注出表达该关系的具体句子以及涉及的实体（即蛋白质 A 和蛋白质 B）。这可以通过机器学习模型对标注实例的学习来实现。这被认为是训练过程的基础，并会显著影响模型的质量。然而，由专家来标注或整理这些材料被视为一项繁琐且耗时的任务，特别是当文本量很大时。因此，生成已经标注好的医学文本将节省大量时间，专家/整理者只需要审核生成的标签，而不是从头开始创建它们。

.2 语料库构建

类似于注释的目的，语料库构建是为训练机器学习模型准备文本数据样本的过程。语料库可能用于如命名实体识别（NER）和强化学习（RL）这样的任务，但也可以被扩展以服务于更详细的任务，如文本摘要和问答。这肯定需要进行文本样本的获取、清理和整理，但这些工作可以通过生成合成文本来避免。

.3 辅助写作

医疗人员，包括医疗接待员、医疗秘书、医疗行政助理，甚至临床医生，通常需要撰写关于特定患者的报告。这样一项任务可能看起来耗时且易出错。因此，通过合成医学文本生成临床报告可能会节省大量时间，医疗人员只需审核这些文本，而无需从头开始创建。这将通过加快服务速度显著提高医疗质量，从而缩短患者的等待时间。

.4 有用性

生成新知识主要与测试该知识的实用性有关。因此，此目的旨在观察生成的医学合成文本对 NLP 任务的益处。请注意，它可能会与其他目的重叠，例如注释、增强，甚至是保护隐私。

A 附录 B：生成技术

A.1 手册

人工技术是指在医学领域中使用人类专家来制定、管理和审查临床文本 [26, 29, 30, 34] 。这些技术中常用的方法是众包和人机交互，其中生成或管理文本的任务被分解为子任务，这些子任务整合了一组人以获得他们的干预和反馈。

文本处理技术指的是采用传统的半自动化方法来重建原始医学文本。这种方法包括易数据增强 (EDA)、SpaCy、全词同现向量 (GloVe)、基于模板的技术和快速自动关键词提取 (RAKE)。EDA 依赖于四个主要步骤：同义词替换 (SR)、随机插入 (RI)、随机交换 (RS) 和随机删除 (RD)。SR 旨在通过使用外部知识源将非停用词术语替换为其语义同义词。而 RI 指的是在随机位置为随机非停用词术语添加同义词的过程。RS 指的是交换两个随机选择术语的过程。最后，RD 旨在消除一个随机术语。在 EDA 中，只有文本输入 x 被更改，而目标 y 将被保留。而 SpaCy 是一个用于执行 NLP 任务（如分词、词性标注和实体识别）的 Python 包。同样地，RAKE 是一种基于频繁出现来提取关键词和关键短语的算法。GloVe 将大文本语料库（例如，英语维基百科）中的全局同现术语因子化以生成嵌入向量。最后，基于模板的方法旨在识别预定义的模式，如日期、命名实体和事件。

A.2 基于知识的

知识来源技术依赖于医学领域的外部词典或本体，包括地名词典 [2, 93] 、WordNet [91, 96] 和统一医学语言 UML [22, 23, 58, 61–63, 92, 96, 98, 107] 。WordNet 是一个开放领域的英语词库，它编码了术语之间的语义关系，如同义词（即相似）、反义词（即相反）、上义词（即广义术语或是 - a）和下义词（即组内项目或部分 - of）。UML 是一种医学本体，其中包含诸如疾病、综合症、症状、药理物质以及其他与医学和生物医学互动相关的定性和功能性概念的语义医学注释。

传统的神经网络（Neural Network, NN）或多层感知机（Multilayer Perceptron, MLP，也称为 vanilla）由三个主要层组成：输入层、隐藏层和输出层（见图 5 (a)）。输入层类似于输入数据 x 的特征，而输出层则代表所需的目标 y，对应于类别标签。隐藏层代表特征之间关系的编码。需要注意的是，隐藏层可以从单层到多层（例如，两层或三层）[113] 。在隐藏层内的特征编码是通过生成随机权重和输入特征值的乘积来实现的，并添加一个称为偏置的常数值。这将在层与层之间重复（即从输入到隐藏，从隐藏到隐藏，以及从隐藏到输出）。最终结果将被输入一个激活函数，以便为结构添加非线性，并视为预测的目标。激活函数有多种，如双曲正切（Hyperbolic Tangent, TanH）、Sigmoid、Softmax 等。将对实际目标和预测目标进行比较以计算错误率。因此，NN 的训练或学习机制通过一种称为反向传播（Backpropagation）的概念来实现，旨在向后更新权重值。这将反复进行，直到错误率变得极小（即，实际目标和预测目标之间的差异最小）。在那一刻，最终的权重值将被存储并用于未来对未见数据（即测试数据）的预测。虽然 MLP 在临床文本生成中尚未被频繁使用，但它已被整合到更加复杂的架构中，这将在以下小节中描述。

A.2.1 Word2vec

这一版本的神经网络是一种全连接的前馈架构，旨在处理文本数据。考虑到独热编码，它旨在通过稀疏矩阵（即，充满零向量的矩阵，只有一个元素为 1，对应于匹配项）来表示分类标签。文本数据可以通过包含唯一术语的词汇稀疏矩阵在独热编码中处理。因此，每个术语的独热向量将被输入到 word2vec 架构中，以预测另一个术语的独热向量 [114] 。类似地，通过随机生成权重并通过反向传播进行更新的训练过程将进行。一旦实际目标和预测目标之间的差异最小化，隐藏神经元的值将作为词嵌入（WEs）存储。结果嵌入向量将包含上下文和关系信息，其中单词 male 和 king 之间的关系类似于单词 female 和 queen 之间的关系。

事实上，Word2Vec 有两种不同的架构：Skip-gram 和连续词袋（CBOW）（见 Fig. 5 （b））。Skip-gram 旨在处理目标术语的独热向量，试图预测其上下文术语。这对于识别单词的语义十分有用。而 CBOW 旨在处理上下文术语的独热向量，试图预测一个目标术语，这对于识别单词周围的

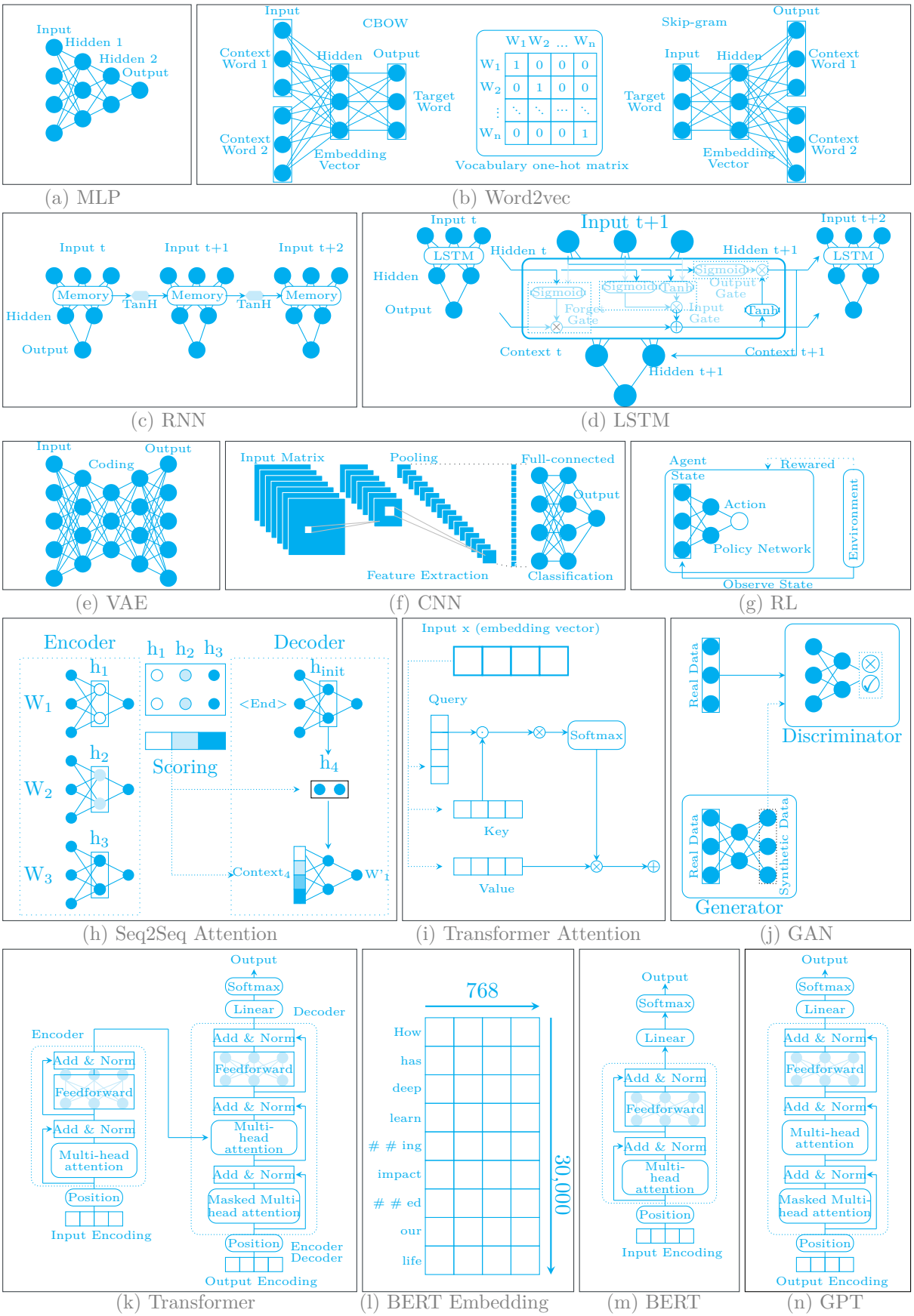

Figure 5: 神经网络架构

语法十分有用。结果嵌入应该在语法和语义上表现出相似性。这意味着相似的单词在其嵌入向量中应该具有更接近的值。然而，word2vec 架构高度依赖于训练文本的大小，它需要训练中有大量文本以获得准确的嵌入。因此，有多项研究努力旨在在大量文本上训练 word2vec，然后提供可用于进一步下游任务（例如，分类、预测或聚类）的预训练向量。BioWord2Vec 是一个尝试解决领域特定嵌入问题的努力，通过在来自 PubMed 的大量生物医学/医学文本上训练 word2vec 架构并生成预训练向量。然而，word2vec 有一个显著的问题称为词汇外问题（OOV）。该问题指的是在训练期间缺少某个单词时，该单词将没有嵌入的情况。

A.2.2 循环神经网络 (RNN)

为了处理时间序列或序列数据，引入了 RNN。RNN 旨在保留一些来自时间点 t 的输入信息，并尝试将其传递到未来的输入（即，输入 t_i）。实际上，RNN 具有与传统神经网络相同的层结构，但它具有一个额外的层，称为记忆或上下文层（见图 5 (c)）。这种记忆将通过一个激活函数（通常是 TanH）传递这样的上下文信息 [116] 。在 NLP 的背景下，RNN 在所谓的 CharRNN 架构中被使用，以生成字符级的嵌入，而不是在 word2vec 中的词级嵌入。由于字符的数量是固定的，这种机制在缓解 OOV 问题上表现更好。Li 等人 [43] 已经使用 CharRNN 架构生成合成的临床文本，以实现隐私保护的目的。

在某些场景中，可能需要长期的上下文信息来预测当前的标签。例如，考虑一个大文本语料库中预测后续单词的场景，预测将在语料库末尾进行，而所需的上下文信息位于语料库的最开头。在这种情况下，RNN 架构很难保持长期的上下文信息。此外，长序列是导致被称为消失梯度问题的主要原因，该问题指的是由于通过许多隐藏层的导数连续乘积导致神经元值周期性减少，从而使值接近零。这里引入了 LSTM，它是 RNN 的扩展，用以克服保持长期上下文信息的问题。LSTM 具有与 RNN 类似的结构，只是其记忆单元要复杂得多。LSTM 的记忆单元包含三个主要组成部分，称为门：遗忘门、输入门和输出门（见图 5 (d)）。遗忘门包含一个 sigmoid 函数，用于决定是否遗忘或保留信息，其生成的值介于 0（即完全遗忘）到 1（即完全保留）之间。另一方面，输入门包含两个激活函数，包括 sigmoid 和 TanH。Sigmoid 将决定要更新上下文中的哪个值，而 TanH 将用新的候选向量替换上下文中的旧信息（即需要更新的）。最后，输出门旨在准备输出的内容，其中包含相同的激活函数。然而，它首先应用 sigmoid 以确定上下文的哪个部分将对当前输出作出贡献，然后应用 TanH 将该部分与输出连接。

LSTM 有不同的变体；其中之一是双向 LSTM Bi-LSTM，它将前向和后向的概念顺序地应用。这种机制已证明在处理自然语言时具有高效性，因为它利用了词语的顺序并获得了更多的上下文信息 [118] 。

LSTM 的另一个变体是门控循环单元 GRU，它通过结合遗忘门和输入门简化了 LSTM 的记忆单元 [119] 。

然而，在生成文本时，RNN 和 LSTM 都遇到了一个显著的缺陷。这个问题被称为曝光偏差，即模型试图使用错误的前一个词来预测下一个词。前一个词的错误性来源于训练过程中模型并未见过真实的前一个词，因此用其生成的一个词来替代。这个问题在生成较长文本时会被放大，导致结果不准确。

虽然 LSTM 大大缓解了梯度消失的问题，但随着词汇数量的增加，这个问题仍然存在。换句话说，给 LSTM 的文本越长，长期上下文信息的损失就会越多。另一方面，在处理较长序列时，RNN 和 LSTM 在训练中显示出了非常慢的性能。为此，Luong 等人提出了一种称为 Seq2Seq 的架构，由编码器层和解码器层组成，用于机器翻译，其中 RNN 被应用于这两个层（见图 (h)）。Seq2Seq 旨在使学习更加集中于输入序列的特定部分，而不是过多关注信息。与传统的 RNN 中累积隐藏状态通过时间间隔传递不同，Seq2Seq 旨在让编码器将所有时间间隔的各个隐藏状态传递给解码器。因此，每个隐藏状态的向量都会被评分，以识别输入序列之间的关系；换句话说，确定哪些隐藏状态能够在当前时间间隔提供最需要的信息。然后，解码器将在每个隐藏状态上执行 Softmax 操作并将其乘以分数。生成的向量将作为输入用于预测下一个标记。在文献中，Seq2Seq 被描绘成在某些情况下其编码器/解码器可能是 RNN、LSTM，甚至是 GRU。Seq2Seq 的一个修改版本是指针生

成网络 PGN，其旨在通过采用指针机制解决 OOV 问题，该机制旨在通过生成从输入中复制文本部分。

另一种特殊的编码器-解码器架构是 VAE，它旨在处理一个输入

A.2.3 变分自编码器 (VAE)

并尝试在输出处预测相同的 [122] 输入数据。它由三个内部层组成，包括编码、编码和解码（见图 5 （e））。编码层旨在学习输入数据的低维或高维结构。然后，编码将学习数据的更低维结构，其中连续的潜在变量被保留。之后，解码将学习从编码中的潜在变量重建更高维度。最后，输出将类似于输入数据的实际/确切维度。由于其学习数据低维度的机制，VAE 已被广泛用于降维和特征提取 [123] 。

另一种广泛用于图像处理、模式识别和字符到字符语言模型的深度神经网络架构是 CNN [124] 。其结构通过引入一个新组件——特征提取，并结合传统的用于分类的全连接层，偏离了传统的神经网络 (见图 5 (f))。特征提取组件旨在通过三种机制执行过滤过程：卷积层、池化和步长。卷积层旨在生成具有特定二维尺寸（2×2 或 3×3）的滤波器。因此，步长将决定移动滤波器的所需步数。最后，池化旨在对滤波器的值进行平均，其中可以进行最大化或最小化。然后，将得到的矩阵展平并输入到全连接层中进行分类。文献中描述了一种集成了 CNN 和 VAE 的架构称为潜在扩散模型 LDM [125] ，用于生成与医学图像对齐的文本，利用图像的潜在变量而非其像素。

A.2.4 变压器

Vaswani 等人利用注意力机制引入了一种称为 Transformer 的新架构。该架构由一个编码器和一个解码器组成，它们都应用了注意力机制。关键区别在于 Transformer 中的注意力利用了并行化来同时处理多个单词。与双向 LSTM 按顺序处理两个方向 (即前向和后向) 的单词顺序，或者 Seq2Seq 按顺序关注不同单词相比，Transformer 中的注意力可以以并行模式处理序列输入的多个部分。这缓解了 LSTM 和 RNN 中训练速度慢的问题，为 Transformer 提供了高效的处理能力。Transformer 由两个组件组成，包括编码器和解码器，每个组件都接收一个嵌入向量；编码器接收输入，而解码器接收相应的输出（例如，句子及其翻译）。然后，在嵌入向量上添加位置嵌入。这种位置编码旨在为词语的上下文理解增加一个标记的位置信息。因此，会进行注意力操作，之后是一个前馈神经网络，输出注意力权重。注意有两个编码器；一个在左侧，另一个在右侧，称为编码器-解码器（见图 5 (k)）。编码器-解码器注意力执行额外的掩码任务，其中下一个标记被掩盖，以便让模型预测这样的词语。两个编码器的输出将被传递给解码器，解码器旨在执行另一种注意力机制，以映射两个编码器输出之间的关系，然后是一个前馈神经网络。最后，将应用线性变换和 Softmax 来进行所需的下游任务（无论是文本生成、文本翻译、文本分类还是实体识别）。实际上，transformer 中的注意力单元旨在将输入嵌入向量转换为三个子向量，称为查询、键和值（参见图 5 (i)）。这是通过分别采用三个参数化矩阵 XMATHX_w_4、XMATHX_w_5 和 XMATHX_w_6 对输入向量 XMATHX_x 进行分解来实现的。然后，查询向量和键向量之间将进行点积运算，接着是 Softmax 缩放和另一个 Softmax，随后这些将被值向量分解并最终相加。

条件变压器语言 CTRL 是一种基于 transformer 的架构，已补充了如风格、内容以及特定任务特征等附加信息，以控制生成的文本 [126] 。

编码器-解码器 transformer 架构经历了变体的显著扩展。其中一些变体用于视觉内容，例如与 CNN 集成的视觉编码器解码器 VED [127] ，视觉语言预训练 VLP [128] ，以及图变压器 GTR [129] 。其他变体用于文本生成目的，例如长文档编码器解码器 LED [130] ，双向和自回归变压器 BART [131] 及其生物医学变体 BioBART [132] ，文本到文本转换变压器 T5 [133] 及其医学变体标签到文本转换器 LT3 [134] ，以及蛋白质语言模型 PLM [135] 。尤其是 VLP，有两种知名架构，分别是对比语言图像预训练 CLIP [136] 和自举语言图像预训练 BLIP [137] 及其医学模型 ClinicalBLIP [138] 。这两种架构都有图像编码器和文本解码器。

BERT 是一种重要的架构，基于 transformer 的编码器构建。BERT 由两个主要部分组成：预

训练和微调。预训练只采用了 transformer 架构中的编码器层（见图 5 (m)），并增加了两个额外任务，包括掩码和句子预测。掩码任务旨在预测句子中的下一个词，而句子预测旨在预测一个句子是否与其后续句子相关。这些任务旨在让模型学习通用的语言特征。BERT 的第二个组件是微调，这是模仿所需下游任务的部分（例如，文本分类、实体检测等）。注意，BERT 可用于端到端分类，在这种情况下整个模型（预训练和微调）将根据特定任务进行更新和调整，也可以用于特征提取，在这种情况下，将预训练组件的输出嵌入输入到另一个模型，如 LSTM 或 RNN。在特征提取的情况下，模型无需更新。实际上，BERT 已经在大规模文本上进行过预训练，使用了 BookCorpus（8 亿词）和 Wikipedia（25 亿词）等来源。此外，BERT 在固定词汇表（30,000 个英语单词）上进行了预训练，固定嵌入维度是 768。这个固定词汇表处理词根或词根词，其中任何派生变化如 ing 或 ed 将分别作为单独的标记处理（见图 5 (l)）。在这一方面，OOV 问题已被显著克服。由于预训练机制导致模型计算量大，阻碍了这些模型的转移过程，因此出现了一种知识蒸馏的概念，用以将这些大型模型压缩成更小的模型。蒸馏包括两个模型：教师模型和学生模型。学生模型将被训练来模仿老师的较大模型。通常通过转移隐藏层或所谓的提示层中的知识来实现，这些层中描述了输入特征之间的潜在上下文信息和关系。这也催生了蒸馏 BERT [140] 。

另一方面，BERT 在医学领域通过不同的预训练模型得到了体现，包括 PubMedBERT [141] 和 SMedBERT [142] 。

A.2.5 生成式预训练变换模型（GPT）

GPT 是另一种基于 Transformer 解码器构建的架构，也被称为自回归模型。GPT 借用了 Transformer 架构的解码器组件（见图 5 (n)），形成了解码器的堆栈。类似于 BERT，GPT 包含预训练和微调部分，但预训练主要用于掩码语言建模，在这种建模中，句子中的下一个词被预测，而不进行句子预测。此预训练被称为无监督生成模型或自监督，指的是为了自动标注而进行的词语预测任务。GPT 在 BookCorpus（8 亿词）上进行了预训练，并且有 1.1 亿个参数。这些统计数据在 GPT-2、GPT-3 和 GPT-4 中得到了扩充（分别为 15 亿、1750 亿和 1 万亿参数）。类似于 distilBERT，一个更小/压缩版本的 GPT 在 distilGPT [143] 中被描述。

GPT 在文本数据增强方面进行了研究，其中提出了一种基于语言模型的数据增强 LAMBADA [144] 。LAMBADA 基于微调的 GPT-2 架构构建，输入是句子 x 数据及其目标 y。LAMBADA 的目的是在保留目标 y 的同时综合生成句子 x。一旦生成新的合成句子后，将进行另一层的二元分类。这个任务被称为噪声控制或过滤，其中 GPT 架构将选择最准确的合成句子。这样的选择将基于分类器对特定目标 y 给出的最高概率。换句话说，与特定类别标签一致而被分类为最高概率的句子将被选中。另一种称为差分隐私 GPT（DP-GPT）的 GPT 变体在文献中有所描述，其目标是通过最小化个体的概率对数来提供数学界限。这可以通过添加随机性或噪声来实现，以避免特定个体（即患者）的目标类别（例如，疾病/诊断）获得更高概率，同时确保隐私。使用内存压缩注意力的 Transformer 解码器（T-DMCA）是一种只使用变压器结构解码器且具有压缩内存的 GPT，可以处理更长的序列。在医疗领域，GPT 通过不同的预训练模型如 BioGPT、Galactica、GPT-Neox、CerebroGPT 等得以体现。此外，GPT 还推动了生成应用程序的出现，例如 Meta Llama、Google Bard、Mistral、Claude、Gemma 和 ChatGPT。

一种近期被研究用于生成合成数据的架构是 GAN（生成对抗网络）。这种架构由两个主要组成部分构成：生成器和判别器（见图 5 (j)）。生成器将类似于一个网络，它接受输入 x 并尝试预测/生成 x 的合成副本 [150] 。而判别器是一个网络，接受输入 x 以及由生成网络生成的合成 x，并尝试预测一个对应真实或合成的二元分类的正确或错误。在这方面，GAN 描绘了一种猫捉老鼠的游戏，其中生成器不断产生新的数据，而判别器则接受测试以识别哪些数据是原始的，哪些数据是合成的。在这种情况下，优化一个网络优于另一个网络并不是理想的情况。例如，精确的生成器意味着判别器会被过多的实例欺骗；反之，精确的判别器意味着由生成器生成的任何数据都会被判别器检测到。因此，理想的情况是两个网络的准确性之间的权衡（对两个网络的 $\approx 50\%$ ）。GAN 是一种概念，其中前馈（例如，RNN）或卷积（例如，CNN）层可以用于生成器或判别器中。
GAN 在缓解 RNN 和 LSTM 描述的暴露偏置问题上有所贡献，但仍面临一些问题。首先，从判别

器到生成器的反向传播方面存在显著的困难。其次，由于生成器旨在通过对数据进行微小更改来创建逼真的复制品，将其应用于离散令牌（即生成文本）会给判别器带来问题。这是因为在有限的词典空间中没有对应于这种微小更改的令牌。这时引入了序列 GAN 或所谓的 SeqGAN，在 GAN 架构中引入了强化学习 RL 的使用 [151] 。RL 是一种除了监督和无监督之外的新范式，其中一个代理被编程为从环境中获得观察状态（例如，游戏中的位置）并尝试预测任意值。然后，代理预测的这个任意值将被评估以改变环境中的动作 [152] 。假设一场国际象棋比赛，对手走了一步；代理将预测反映随机移动的任意值。如果此移动有任何好处，代理将获得分数奖励（见图 5 (g)）。否则，它将受到惩罚分数的惩罚。因此，随着时间的推移，代理将学会做出准确的动作。RL 已广泛用于掌握不同的游戏，并在机器人领域具有潜力。然而，在 SeqGAN 的背景下，生成器作为代理工作，判别器作为评估环境。由于 GAN 判别器进行的整个序列评估，出现了一个显著的问题，称为模式崩溃。评估整个序列不是一个准确的策略，尤其是当序列过长时。因此，基于子序列评估的改进的 SeqGAN，即 SegGAN，被引入 [153] 。

最近，由于 GAN 和 VAE 在生成诸如文本序列等离散变量方面面临挑战，提出了一种将这两种架构结合起来的新方法，形成了对抗性正则化自动编码器 ARAE [154] 。这种架构包含了传统的 VAE 层，其中包括输入、编码器、编码器、解码器和输出。此外，编码生成的向量连同用户指定的样本分布会作为输入送入一个判别器层，该层将区分这两个样本。

B 附录 C：评估方法

B.1 相似性

一个好的合成文本应该在一定程度上反映出与原始文本的相似性，以便在隐私保护场景中实现两者之间可能的替代，以及用于数据扩充。在本节中，将对文献中描绘的基于距离的相似性度量进行全面讨论。

B.1.1 双语评估助手 (BLEU)

文献中使用的最常见指标是 BLEU。最初，BLEU 是用于评估机器翻译文本的 [155] 。最近，由于语言生成模型的出现，它在评估生成文本与原始文本的相似性方面获得了很高的知名度。要理解 BLEU，需要讨论精确度，它对应于合成文本和原始文本中匹配词的数量除以合成文本中词的总数，这可以通过下列公式计算：

$$\text{Precision} = \frac{\#Matching(W_s, W_o)}{W_s} \tag{1}$$

这可以用于单个匹配词或多个连续词组，即称为 N-gram。N-gram 指的是连续词的数量，例如 unigram（即单个词）、bigram（即词对）、trigram（即词三）、quadgram（即词四）。精确度也称为正预测值（Positive Predictive Value，PPV）。事实上，精确度不能估计合成文本的准确性，因为词的重复匹配会导致 100 % 的准确性，而合成文本可能会没有任何意义。因此，出现了一个精确度的修正版，称为剪裁精确度，以考虑词的唯一出现次数及匹配情况，如下列公式所示：

$$P_n = \frac{\sum_{C \in (Candidates)} \sum_{ngram \in C} Count_{clip}(ngram)}{\sum_{\bar{C} \in (Candidates)} \sum_{n\bar{g}ram \in \bar{C}} Count(n\bar{g}ram)} \tag{2}$$

因此，对于一个文本序列，每个 n-gram 将被计算并相乘以计算几何平均精确度，如下列公式所示：

$$\text{Geometric Average Precision (N)} = exp\left(\sum_{n=1}^{N} w_n \, log \, p_n\right) \tag{3}$$

通常，N 被认为是 4，对应可能的 n-gram（即 unigram、bigram、trigram、quadgram），而 w_n 是等于 1/N 的统一权重。计算 BLEU 之前的一步是处理合成文本和原始文本的长度，以准确估计相似性。这可以通过下面的公式计算：

$$\text{Brevity Penalty} = \begin{cases} 1, & \text{if } c > r \\ e^{\frac{1-c}{r}}, & \text{if } c \leq r \end{cases} \tag{4}$$

，其中 c 指的是合成文本的长度，r 反映了原始文本的长度。现在，BLEU 可以通过几何平均精度乘以简短惩罚来计算，如下公式所示：

$$\text{BLEU} = BP * N \tag{5}$$

BLEU 被多项研究用于比较合成医学文本与原始文本之间的相似性 [42–46, 57, 58, 62, 70, 83, 95] 。然而，BLEU 存在多种缺陷，例如忽略术语的意义，其中同义词可能被错误处理（例如，medical ≠ clinical）。这同样适用于同一术语的任何词形变化（例如，clinical ≠ clinicals）[156] 。
自我-BLEU 是另一种度量，它扩展了原始 BLEU 用于评估合成文本的多样性，可以根据与其他句子的相似性来评估生成的合成句子 [157] 。自我-BLEU 计算每个生成句子的 BLEU，然后计算整个生成合成文档的平均值。自我-BLEU 值越高，合成文本的多样性就越低。在评估合成医学文本的背景下，该度量已被 Guan 等人使用 [18, 20] 。

尽管 BLEU 仅考虑精度，另一个可以评估检索术语的度量是 Recall 或 Sensitivity，有时称为 Negative Predictive Value NPV 。召回率对应于原始 W $_{o}$ 和合成 W $_{s}$ 文本之间匹配的单词数量除以原始文本中的单词总数，如下公式所示：因此，谷歌推出了一种新的 BLEU 版本，被称为 GLEU。GLEU 简单地取生成文本和原始文本之间计算出的精确度和召回率的最小值。GLEU 在语料库级别上的表现与 BLEU 相同，但在句子级别上表现更好。GLEU 已被 Kasthurirathne 等人的研究所使用。

ROUGE 是另一个用于评估合成文本的指标。ROUGE 已被广泛用于评估机器翻译和自动文本摘要 [159] 。类似于 BLEU，ROUGE 使用多个 n-gram 的精度，其中 ROUGE-1 将关注于 unigram，ROUGE-2 关注于 bigram，以此类推。此外，ROUGE 可以扩展评价范围以包括召回率和 F1-score ，后者是精度和召回率之间的和谐，计算方式如下所示：在评估合成医学文本的背景下，多个研究已使用 ROUGE [32,70,83,90,95] 。另一种已被修改的 ROUGE 版本，称为 ROUGE-L，通过最长公共子序列 LCS 来扩展匹配分析。LCS 指的是即使被不同词语打断，原始文本与合成文本之间最长的连续匹配的单词。ROUGE-L 已被 Amin-Nejad 等人 [95] 和 Ive 等人 [44] 用于合成医学文本评估。

B.1.2 具有显式排序的翻译评价指标 (METEOR)

METEOR 是用于评估合成文本的另一种指标。与 BLEU 和 ROUGE 类似，METEOR 在评估机器翻译和自动文本摘要或生成方面得到了广泛应用。它只使用单词匹配来计算原始文本和合成文本之间的精度和召回率。然而，METEOR 的调和平均稍有不同，其中召回率比精度的权重更高，如以下公式所示 [160]：在合成医学文本评估的背景下，多个研究已经使用了 METEOR [62,70] 。

B.1.3 翻译编辑率 (TER)

TER 是一种简单的度量，它检查将合成文本转换为原始文本所需的最少操作次数。它可以应用于字符级别（识别将一个单词转换成另一个单词所需的字符编辑次数）[161] 或者单词级别（识别将一个句子转换成另一个句子所需的单词编辑次数）[162] 。在假设单词级别的情况下，TER 衡量的是单词编辑次数 W $_{edits}$ 与原始文本中单词平均数量 W $_{o}$ 之间的比率，如下方公式所示：

$$\text{TER} = \frac{W_{edits}}{Avg(W_o)} \tag{6}$$

TER 指标由 Ive 等人 [44] 的研究描述，他们评估了合成心理健康文本与原始文本之间的差异。此外，Wu 等人 [83] 使用 TER 来评估合成文本生成用以总结医疗报告表格的效果，与人工生成的总结进行比较。

另一方面，还有一些度量指标用于检查单词向量之间的相似性。在这种类型的评估中，单词之间的相似性是通过将单词自身转换为向量来实现的。这些向量或者是通过字符出现频率形成的，或者是通过词嵌入形成的。第一个也是最流行的基于向量的度量指标是余弦相似性，其中两个向量之间角度的余弦值反映了它们之间的相似性。假设有两个向量 A 和 B，余弦值可以通过如下公式计算：

Melamud 和 Shivade 使用余弦来计算合成医学文本和真实文本之间的相似性。同样，Jaccard 是另一种基于向量的度量，检查两个向量 A 和 B 之间的相似性，通过它们之间的交集被两个向量的并集所除的比例来实现，如下公式所示：

Al Aziz 等人使用 Jaccard 来估计合成医学文本和原始文本之间的相似性。

B.1.4 基于共识的图像描述评估 (CIDEr)

CIDEr 是通过应用两个句子内各个词向量之间的 n-gram 余弦相似性来检查两个句子 C 和 R 之间相似性的另一度量标准 [163] 。为了形成词向量，CIDEr 利用词频-逆文档频率 (TF-IDF)，其中生

成文本中的一个词 t 将按照其在原始文本中的频率乘以包含 t 的句子数 N 除以句子总数 N 的对数来表示，具体如以下公式 [164] 所示：因而，CIDEr 将考虑平均余弦相似性如以下公式所示：

$$CIDEr(C,R) = \sum \frac{g^n(C) * g^n(R)}{\|g^n(C)\| \, \|g^n(R)\|} \tag{7}$$

g ^{n} (C) 为第一句中由 n-gram 字词构成的 TF-IDF 向量，而 g ^{n} (R) 为第二句中由 n-gram 字词构成的 TF-IDF 向量。Jing 等人 [70] 和 Lee [55] 使用 CIDEr 来估计合成医学文本与原始医学文本之间的相似性。类似于 TFIDF，最佳匹配 BM25 添加可调参数与平滑对数以获得概率和非线性相似性排名 [165] 。有些研究探讨了原始和合成医学文本之间的词级特征选择的作用 [42, 45] 。随后检查了所选特征/词之间的重叠。

B.1.5 基于语义的度量

所有上述指标都是基于词级相似性；有一些指标考虑了语义相似性，比如 BERTscore [166] 、SemScore [167] 和 F1cXb [168] ，这些是使用 BERT 架构通过它们的 BERT 标记嵌入的余弦相似总和来计算两个句子相似程度的指标。另一个指标是语义命题图像字幕评估 SPICE [169] ，它也是基于对象（例如名词）、属性（即形容词）和关系（即动词和介词）的元组来计算两个句子之间的语义相似性。另一个医学特定指标是医学概念相似性（MEDCON），其目的是通过识别和比较医学概念来计算两个句子之间的相似性 [63] 。

B.1.6 人工评估

一些研究已经讨论了用于相似性评估的人类评估，其中要求多名标注者标记合成医学文本。然后，通过 Cohen’s Kappa 计算跨标注者一致性 IAA，其可以通过以下公式计算 [170] ：

$$\text{Kappa} = \frac{P_o - P_e}{1 - P_e} \tag{8}$$

，其中 P $_{o}$ 是标注者之间观察到的相对一致性，而 P $_{e}$ 是假设的偶然一致性的概率。Brekke 等人的研究 [30] 和 Rama 等人的研究 [29] 使用人类评估来评估合成临床文本与原始文本的相似性。在相同的背景下，评估者可靠性误差 ERE 是另一种用于计算来自不同标注者的误差的度量 [18] 。

除相似性外，医学生成的合成文本还应防范诸如重新识别或成员推断之类的威胁。确保合成医学文本的安全将有利于提供公开可用的医学自由文本，并促进机构之间的共享，这将为重要贡献打开大门。评估安全性的最简单方法由 Lee 等人的研究描绘，其中作者检查了生成的合成文本中是否存在 PII 或 PHI。然而，这些敏感标识符的不存在不会阻止某些攻击者使用隐式特征重新识别个人的威胁。

另一种方法是检查合成个人记录与原始记录之间的相似性。Zhou 等人反向使用 BLEU（见公式 5 ）和余弦（见公式 **??** ），其中高相似性表明隐私性较低。其他研究，如 Kasthurirathne 等人，使用了哈明距离，这是一种类似于 TER（见公式 6 ）的度量，用于计算将一个句子转换为另一个句子所需的操作数，目的相同。

文献中描绘了更复杂的隐私评估度量。这些度量提供了一个数学边界，以检查个体的隐式信息是否仍存在于合成文本中。以下小节将描述这些度量。

B.1.7 标准对数似然比检验（ G^2 ）

这种统计检验可以用于确定某些词是否属于某个文本集合 [171] 。在这方面，它可以用来计算一个术语在原始文本或合成文本中存在的统计显著性。一个术语在某个文本集合中取得的概率越高，意

味着该术语来自于这样的集合。因此，G ^{2} 的较小值表明归属的不确定性。设 E1 和 E2 分别为原始文本和合成文本的两个语料库。两者中一个术语 t 的期望频率可以通过以下公式计算：

$$\mathrm{E1} = c * \frac{a+b}{c+d} \tag{9}$$

$$\mathrm{E2} = d * \frac{a+b}{c+d} \tag{10}$$

其中 a 和 b 分别对应 t 在原始语料库和合成语料库中的出现次数。而 c 和 d 分别对应原始语料库和合成语料库中的总词数。因此，G ^{2} 可以通过以下公式计算：

$$G^2 = 2 * \left(a\, ln\, \left(\frac{a}{E1} \right) + b\, ln\, \left(\frac{b}{E2} \right) \right) \tag{11}$$

考虑到语料库大小的差异，有必要通过以下公式来检查效应大小：

$$\text{Effect Size} = \frac{G^2}{c+d} * ln\, min(E1, E2) \tag{12}$$

G ^{2} 已被 Al Aziz 等进行的研究使用，用于评估合成医学文本的隐私性。

B.1.8 负对数似然 (NLL)

负对数似然（NLL）是评估隐私的另一个指标。这个评估测试由 Yu 等人提出，用来评估用于文本生成的 SeqGAN 方法。其目的是通过导出似然性的负自然对数来估计生成模型的“坏”程度。预测越好，得到的 NLL 值就越低。因此，它可以用于评估生成模型在从真实数据中预测单词序列方面的能力，如下方方程所示：

$$NLL_{test} = -E_{Y_{1:T} \sim G_{real}} \left[\sum_{t=1}^{T} log\left(G_\theta(y_t \mid Y_{1:t-1}) \right) \right] \tag{13}$$

，其中 G $_{\text{real}}$ 是真实数据单词的分布，G $_\theta$ 是生成模型。NLL 已被 Guan 等人的研究所使用。

B.1.9 困惑度 (PPL)

PPL 是另一种度量标准，旨在内在地评估语言模型以便更好地调整参数。给定一个包含单词序列的真实数据集，一个好的语言生成模型在这些单词上测试时，会给出较高的概率。这意味着这些单词对模型来说是熟悉的。然而，随着测试数据的增加，概率会开始下降。因此，困惑度作为一个大小无关的评估指标出现，通过如下公式 [172] 将概率标准化为测试集中的总单词数：

$$PPL(W) = \sqrt[N]{\frac{1}{P(W_1, W_2, \ldots, W_N)}} \tag{14}$$

其中 N 是测试集中的单词数。更低的困惑度值表示是一个好的模型。PPL 已被多项研究用于评估合成医学文本 [44, 56, 95] 的隐私性。

尽管 PPL 和 NLL 仅限于深度学习架构，但 DTP 是一种衡量标准，用于利用任何分类模型的局部属性来研究模型如何揭示敏感信息。给定一个模型 M，它有可能的预测目标 Y 和一个训练集 T，设 t 为属于 T 的一个单独记录。为了保护 t 的隐私，有必要使用一个包含 t 训练的模型和一个不包含 t 训练的模型来检查 t 的预测概率。只要差异被最小化，t 就会得到保护，如下公式所示：其中 M(T) 是使用 t 训练的模型，而 $M(T \setminus \{t\})$ 是不使用 t 训练的模型。而 P($y \mid t$) 是预测 t

记录的目标类别的概率。Melamud 和 Shivade 提出了一个称为序列逐点差别训练隐私的 S-PDTP，以适应语言模型生成的情况。在这方面，个体 t 被认为是个人临床记录中出现的单词序列，其可以通过以下公式计算：另一个称为隐私预算的指标在文献中被描述，其旨在计算 psilon ε，指示数据集查询允许的隐私损失。向数据实例添加更多噪声将导致更低的隐私损失，这意味着更好的隐私保护。已有一些研究涉及通过人工评估隐私性进行评估，在这种评估中，多位标注者被要求将合成的医疗文本标记为" 暴露-识别" 和" 未暴露" 两类，然后使用 Cohen's Kappa 计算 IAA。Libbi 等人的研究使用了这种评估方法来评估合成文本的隐私性，并发现了多个有趣的发现。首先，生成的合成文本可能会产生真实身份，但它们呈现的信息是不一致的，并带有虚假的细节（例如，组合不同的名字）。这可能提供了一种自然的隐私保护特性。然而，合成文本有时复制了与真实药物相关的长序列。如果这些药物是特定于某个人的，可能会损害隐私。因此，合成文本中对较长序列的精确复制必须得到仔细审查。

B.2 结构

一些研究非常关注评估合成医学文本的一致性。Lee [55] 使用了奇异比，这是一种统计测量，用于检查原始和合成文本中的词语分布。它通过以下方程计算：

$$\text{Odd ratio(w)} = \frac{D_{wi} \setminus R}{D_{wj} \setminus S} \tag{15}$$

，其中 D $_{\text{wi}}$ 和 D $_{\text{wj}}$ 分别是词 w 出现的次数，而 R 和 S 分别指的是原始和合成语料库中的词的总数。在这种情况下，作者选择了一些术语，如怀孕或过量药物，并根据其他标准如年龄或性别检查它们在原始和合成文本中的分布。结果显示原始和合成文本集之间达到令人满意的分布相似性。例如，怀孕的词根 pregnancy 在原始文本中与女性一起出现了 134 次，而与男性一起出现为零。在合成文本中，这个词根与女性一起出现了 44 次，而与男性一起出现为零。
聚类是另一种用于评估语义多样性的方法 [23, 58] 。
另一方面，模式匹配专门用于评估文本生成是否符合注释，其中简单词数可以用于匹配注释标签。这种词数度量还帮助评估句子长度 [31, 38] 。

B.2.1 对抗成功

AdvSuc 是一个非常复杂的测试，非常类似于图灵测试，其中合成文本和原始文本在真实性和合成性方面进行测试。然而，该测试通常由在 GAN 架构内被称为判别器的单个分类器来执行。一旦生成器部分生成出合成文本，合成文本和原始文本将一起输入给判别器进行测试。判别器将被训练来区分真实文本和合成文本。理想的情况是判别器获得 50 % 的准确率，这意味着它在真实和合成之间会产生混淆，从而反映出合成文本的高质量。所有采用 GAN 架构进行合成医学文本生成的研究都在其实验中考虑了 AdvSuc [18, 20, 57, 58] 。

为了解释矛盾，文献中描述了否定正确性，其中进行了一种简单的否定术语搜索，并从正确性角度进行了评估。NLI 是一个自动任务，用于评估合成文本的矛盾性。该任务采用两个句子，并尝试判断这些句子是矛盾的、蕴涵的还是中立的 [56] 。

B.2.2 基于迁移学习的 BLEU（BLEURT）

为了评估流畅性，文献中描述了 BLEURT [174] 。这种指标基于 BERT 架构，旨在通过采用两个句子（即合成和原始）作为输入来进行回归任务（即预测一个分数），以指示生成的句子是否流畅并传达原句的意思 [63, 107] 。值得注意的是，PPL 也被用于评估合成文本的结构，一些研究使用它来评估生成模型本身的准确性 [69, 90] 。最后，IAA 和 ERE 的人类评估也被用来评估生成的合成文本的结构。

在这个评估中，将检查生成的合成医学文本的有用性。这可以通过下游 NLP 任务来展示。在接下来的内容中，将解释这些任务。

B.2.3 文本分类

这是一个流行的任务，旨在将一组文本分类到预定义的类别标签中。在医疗背景下，[2] 已经将这一下游任务作为分类 COVID 新闻的工具。

B.2.4 疾病分类

该任务旨在根据特定疾病 [42,52,57,91] 将患者记录和健康对照分类为二元类别。

B.2.5 表型分类

该任务旨在将患者记录分类为一个或多个预定义的症状 [27,34,39,44,95] 。

B.2.6 诊断预测

该任务旨在处理患者记录以预测特定疾病的程度/进展（例如，轻度、中度或重度）[55,99,101,108] 。

B.2.7 再入院预测

该任务旨在处理患者记录，并尝试预测该患者是否会再次入院 [48,95,97] 。

B.2.8 命名实体识别 (NER)

该任务旨在提取医疗实体，如药物、疾病、实验室测试或症状 [31,38,40,43,110] 。

B.2.9 不良药物反应/事件提取 (ADR/ADE)

这是命名实体识别的一个子任务，它将提取范围缩小到通过医学文本包括不良事件或反应的提及 [93,94]

B.2.10 关系抽取 (RL)

这是一个与实体识别（NER）互补的任务，其目的在于识别通过 NER 提取的医学实体之间的关系，例如蛋白质-蛋白质相互作用、药物-药物相互作用以及化学物质诱发的疾病 [30,54,59,104] 。

B.2.11 自动去识别

这个任务也被称为安全港，其目的是将 PII/PHI 替换为随机代码。去标识化有时通过向 PII/PHI 添加噪音来进行，以此来防止重新识别。然而，手动去标识化似乎是一项繁琐且耗时的任务。因此，多个研究尝试将去标识化自动化为一个命名实体识别（NER）任务，其中训练一个模型来识别 PII/PHI 并将其替换为安全的代理（例如，将 John 替换为 Name）[28,32,105] 。

B.2.12 问答系统 (QA)

此任务旨在提取满足输入问题的准确文本部分。此任务在医疗领域变得非常有价值，因为它用于医疗咨询 [17,67,71,74] 。

B.2.13 报告生成和文本摘要

这个任务旨在为特定目标生成文本。一些研究使用这个任务来处理医学图像（例如，X 光），并尝试生成合成文本来描述诊断 [69, 70] 。而 Liu 的研究 [90] 则使用这个任务来自动补全临床笔记。最后，Wu 等人 [83] 通过生成医疗报告的表格摘要来使用这个任务。

C 附录 D：精选文章

Table 2: Summary of selected articles

Article	Year	Purpose	Generation Method	Approach / Architecture	Data Source	Language	Evaluation Method	Evaluation Paradigm	Utility	Clinical Source
[73]	2015	Assistive Writing	Knowledge Source, Text Processing	Template-based	Publicly Available	English	Human Assessment	Structure	Text Summarisation	Discharge Summaries
[69]	2018	Assistive Writing	Neural Network	ARAE	Publicly Available (IU X-RAY)	English	PPL	Structure, Test on Utility	Report Generation	Radiology
[18]	2018	Privacy-preserving	Neural Network	SeqGAN (LSTM as generator, CNN and Bi-LSTM as discriminator), RL	Private EHR/EMR	Chinese	BLEU, NLL, AdvSuc, Human Assessment (ERE)	Similarity, Structure, Test on Utility	Disease Classification	Discharge Summaries
[70]	2018	Assistive Writing	Neural Network	Seq2Seq (CNN Encoder and LSTM Decoder)	Publicly Available (IU X-RAY)	English	BLEU, ROUGE, CIDEr, METEOR	Similarity, Test on Utility	Report Generation	Radiology
[90]	2018	Assistive Writing	Neural Network	Transformers (TDMCA)	Publicly Available (MIMIC-III)	English	PPL, ROUGE, Human Assessment	Similarity, Structure, Test on Utility	Report Generation	Discharge Summaries
[62]	2018	Corpus Building	Text Processing, Knowledge Source	UML, Template-based	Publicly Available (MIMIC-III)	English	BLEU, METEOR	Similarity	-	Discharge Summaries
[24]	2018	Corpus Building	Text Processing	Template-based	Online source	German	Human Assessment	Structure	-	Clinical Practice Guidelines
[55]	2018	Privacy-preserving, Augmentation	Neural Network	Seq2Seq (LSTM Encoder and LSTM Decoder)	Private EHR/EMR	English	BLEU, ROUGE, Odd-ratio, CIDEr, Human Assessment (Check Sensitive Information), Classification Accuracy	Similarity, Structure, Privacy, Test on Utility	Diagnose Prediction	Discharge Summaries
[29]	2018	Corpus Building, Annotation	Manual	-	Private EHR/EMR	Norwagian	Human Assessment (IAA), Pattern Match, Classification Accuracy	Structure, Test on Utility	Named Entity Recognition, Relation Extraction	History of Present Illness (HPI)
[46]	2019	Usefulness	Neural Network	Transformers (CTRL), RAKE	Publicly Available (MIMIC-III)	English	BLEU, ROUGE	Similarity, Test on Utility	Phenotype Classification, Relation Extraction	Discharge Summaries
[56]	2019	Privacy-preserving	Neural Network	LSTM	Publicly Available (MIMIC-III)	English	PPL, PDTP, Human Assessment, NLI	Similarity, Structure, Privacy	-	Discharge Summaries
[42]	2019	Usefulness, Privacy-preserving	Neural Network	SeqGAN	Private EHR/EMR	English	BLEU, Hamming Distance	Similarity, Privacy, Test on Utility	Disease Classification	Patient Laboratory Test
[93]	2019	Augmentation	Knowledge Source	Gazetteers	Publicly Available	English	Classification Accuracy	Test on Utility	Adverse Drug Reaction/Event Extraction	Drug and Medication
[80]	2019	Assistive Writing	Neural Network	Seq2Seq (RNN Encoder and RNN Decoder)	Publicly Available (MIMIC-III)	English	BLEU, Human Assessment	Similarly, Structure	Report Generation	Discharge Summaries
[16]	2019	Assistive Writing	Neural Network	GPT-2	Private EHR/EMR	Chinese	Human Assessment	Similarity	Report Generation	Discharge Summaries
[77]	2019	Assistive Writing	Neural Network	GTR, CNN (DenseNet)	Publicly Available (IU X-RAY)	English	BLEU, ROUGE, CIDEr, Human Assessment, Classification Accuracy	Similarity, Structure, Test on Utility	Disease Classification	Radiology

Continued on next page

Table 2: Summary of selected articles (Continued)

Article	Year	Purpose	Generation Method	Approach / Architecture	Data Source	Language	Evaluation Method	Evaluation Paradigm	Utility	Clinical Source
[48]	2020	Usefulness	Neural Network	Transformers (Encoder-Decoder), GPT-2 (Decoder)	Publicly Available (MIMIC-III)	English	Classification Accuracy	Test on Utility	Phenotype Classification, Readmission Prediction	Discharge Summaries
[22]	2020	Corpus Building	Knowledge Source, Text Processing	UML, SpaCy	Online source	German	Human Assessment (IAA), Pattern Match	Structure	-	Clinical Practice Guidelines
[25]	2020	Assistive Writing, Augmentation	Neural Network	Seq2Seq (GRU Encoder and GRU Decoder), RL, BERT	Publicly Available (MIMIC-CXR)	Japanese, English	BLEU, ROUGE, Human Assessment	Similarity, Structure, Test on Utility	Report Generation	Radiology
[34]	2020	Corpus Building	Manual	-	Private EHR/EMR	Bulgarian	Classification Accuracy	Test on Utility	Phenotype Classification	Discharge Summaries
[91]	2020	Augmentation	Knowledge Source, Neural Network	WordNet, GloVe, word2vec, BioWord2Vec	Publicly Available	English	Classification Accuracy	Test on Utility	Disease Classification	Discharge Summaries
[95]	2020	Augmentation	Neural Network	Transformers, GPT-2	Publicly Available (MIMIC-III)	English	PPL, BLEU, ROUGE	Similarity, Test on Utility	Readmission Prediction, Phenotype Classification	Discharge Summaries
[44]	2020	Privacy-preserving, Usefulness	Neural Network	Transformers (CTRL)	Private EHR/EMR	English	BLEU, PPL, TER, Human Assessment (IAA), Classification Accuracy	Similarity, Structure, Privacy, Test on Utility	Phenotype Classification	Discharge Summaries
[98]	2021	Augmentation	Text Processing, Knowledge Source	UML, EDA	Publicly Available	English	Classification Accuracy	Test on Utility	Named Entity Recognition	Population, Intervention, Comparison, and Outcome (PICO)
[81]	2021	Assistive Writing	Neural Network	Seq2Seq (DenseNet CNN Encoder and LSTM Decoder)	Publicly Available (IU X-RAY, MIMIC-CXR)	English	BLEU, ROUGE, CIDEr, METEOR, Hamming Distance	Similarity, Test on Utility	Report Generation	Radiology
[26]	2021	Privacy-preserving, Corpus Building	Manual	crowdsourcing human-in-the-loop	Private EHR/EMR	Japanese	Human Assessment	Similarity, Structure, Privacy	-	Patient Laboratory Test
[50]	2021	Augmentation, Usefulness	Neural Network	GPT-3	Manual Collection and Curation	English	ROUGE, Negation Correctness, Human Assessment	Similarity, Structure, Test on Utility	Text Summarisation	Medical Conversations
[20]	2021	Privacy-preserving	Neural Network	SeqGAN (LSTM as generator, CNN and Bi-LSTM as discriminator), RL	Private EHR/EMR	Chinese	BLEU, AdvSuc, NLL, Human Assessment (ERE)	Similarity, Structure, Test on Utility	Disease Classification	Discharge Summaries
[43]	2021	Usefulness	Neural Network	Transformers (CTRL), GPT-2, CharRNN, SeqGAN	Publicly Available	English	BLEU	Similarity, Structure, Test on Utility	Named Entity Recognition	History of Present Illness (HPI)
[96]	2021	Augmentation	Knowledge Source	GloVe, WordNet, UMLs	Publicly Available	English	Classification Accuracy	Test on Utility	Disease Classification	Discharge Summaries
[30]	2021	Augmentation	Manual	-	Private EHR/EMR	Norwegian	Classification Accuracy	Test on Utility	Named Entity Recognition, Relation Extraction	History of Present Illness (HPI)
[32]	2021	Annotation, Privacy	Neural Network	LSTM, GPT-2	Private EHR/EMR	Dutch	ROUGE, BM25, Human Assessment (IAA), Pattern Match	Similarity, Structure, Test on Utility	De-Identification	Discharge Summaries

Continued on next page

Table 2: Summary of selected articles (Continued)

Article	Year	Purpose	Generation Method	Approach / Architecture	Data Source	Language	Evaluation Method	Evaluation Paradigm	Utility	Clinical Source
[45]	2021	Usefulness, Privacy-preserving	Neural Network	SeqGAN	Private EHR/EMR	English	BLEU, GLEU, Hamming Distance	Similarity, Privacy, Test on Utility	Disease Classification	Patient Laboratory Test
[97]	2021	Augmentation	Neural Network	distil-GPT-2, LAMBADA	Publicly Available (MIMIC-III)	English	Classification Accuracy	Test on Utility	Readmission Prediction	Discharge Summaries
[92]	2021	Augmentation	Text Processing, Knowledge Source	UML, EDA	Publicly Available	English	Classification Accuracy	Test on Utility	Named Entity Recognition	Biological Concepts and Relations
[57]	2021	Privacy-preserving	Neural Network	DP-GPT	Publicly Available (MIMIC-III)	English	NLL, BLEU, Jaccard, AdvSuc, DTP	Similarity, Structure, Privacy, Test on Utility	Disease Classification	Discharge Summaries
[103]	2021	Augmentation	Neural Network	Transformers	Online source	English	Pattern Match, Classification Accuracy	Structure, Test on Utility	Diagnose Prediction, Named Entity Recognition	History of Present Illness (HPI)
[2]	2022	Augmentation	Knowledge Source	Gazetteers	Online source	Indonesian	Classification Accuracy	Test on Utility	Text Classification	COVID News
[38]	2022	Annotation, Corpus Building	Neural Network	GPT-neox	Prompting	German	Pattern Match, Classification Accuracy	Test on Utility	Named Entity Recognition	Drug and Medication
[102]	2022	Augmentation	Neural Network	SeqGAN	Publicly Available	English	NLL, BLEU, Classification Accuracy	Structure, Similarity, Test on Utility	Phenotype Classification	Doctor-Patient Conversations
[21]	2022	Assistive Writing	Neural Network	Bi-LSTM	Publicly Available	Chinese	Classification Accuracy	Similarity	Question Answering	Medical Consultation
[19]	2022	Assistive Writing	Neural Network	Transformers (Encoder-Decoder), SMedBERT	Publicly Available	Chinese	BLEU, Human Assessment	Similarity, Structure, Test on Utility	Question Answering	Medical Conversations
[83]	2022	Augmentation, Assistive Writing	Neural Network	Transformers (T5)	Publicly Available	English	ROUGE, BLEU, TER, Human Assessment	Similarity, Structure, Test on Utility	Report Generation	Patient Laboratory Test
[86]	2022	Assistive Writing	Neural Network	VLP (CNN ResNet, BERT)	Publicly Available (MIMIC-CXR)	English	BLEU, Classification Accuracy	Similarity, Test on Utility	Disease Classification, Report Generation, Question Answering	Radiology
[68]	2022	Corpus Building	Neural Network	GPT-2	Publicly Available (MIMIC-III)	English	Human Assessment	Structure	nan	Discharge Summaries (Cardiovascular)
[23]	2022	Corpus Building	Knowledge Source, Text Processing	UML, SpaCy	Online source	German	Clustering	Structure	nan	Clinical Practice Guidelines
[27]	2022	Augmentation	Text Processing	EDA	Manual Collection and Curation	Japanese	Classification Accuracy	Test on Utility	Phenotype Classification	Doctor-Patient Conversations
[94]	2022	Augmentation	Neural Network	word2vec, Cosine Similarity	Private EHR/EMR	English	Classification Accuracy	Test on Utility	Adverse Drug Reaction/Event Extraction	Drug and Medication
[58]	2022	Privacy-preserving	Knowledge Source, Text Processing	UML, RAKE	Publicly Available (MIMIC-III)	English	Clustering, BLEU, TFIDF, Human Assessment (Turing Test)	Similarity, Structure, Privacy	nan	Discharge Summaries

Continued on next page

Table 2: Summary of selected articles (Continued)

Article	Year	Purpose	Generation Method	Approach / Architecture	Data Source	Language	Evaluation Method	Evaluation Paradigm	Utility	Clinical Source
[65]	2022	Corpus Building	Neural Network	Transformers (T5 and BART)	Manual Collection and Curation	English	BLEU, ROUGE-L, CIDEr, METEOR, PPL, SPICE, BERTscore, Human Assessment	Similarity, Structure	-	History of Present Illness (HPI)
[100]	2023	Augmentation	Neural Network	Transformers (T5, LT3)	Publicly Available (MIMIC-III)	English	BLEU, ROUGE, Jaccard, BERTscore	Similarity	Named Entity Recognition	Drug and Medication
[63]	2023	Corpus Building	Knowledge Source, Neural Network	UML, BioBART, LED	Manual Collection and Curation	English	ROUGE, BERTscore, BLEURT, MedCon	Similarity	-	Doctor-Patient Conversations
[101]	2023	Augmentation	Neural Network	Transformers, CNN-ResNet	Publicly Available	English	BLEU, Classification Accuracy	Test on Utility	Diagnose Prediction	Radiology
[74]	2023	Assistive Writing	Neural Network, Text Processing	BART, distilBERT, Template-based	Publicly Available	English	ROUGE	Similarity	Question Answering, Text Summarization	Doctor-Patient Conversations
[78]	2023	Assistive Writing	Neural Network	GPT-2	Publicly Available	English	BLEU, ROUGE	Similarity	Question Answering	Drug and Medication
[110]	2023	Augmentation	Neural Network	GPT-2	Publicly Available	English	ROUGE-L, PPL	Similarity, Structure	Named Entity Recognition	Biological Concepts and Relations
[84]	2023	Assistive Writing	Neural Network	Seq2Seq (RNN Encoder and RNN Decoder) Point Generator Network	Online source	English	ROUGE, ROUGE-L, Cosine, Jaccard, TFIDF	Similarity	Text Summarisation	Drug and Medication
[64]	2023	Corpus Building, Augmentation	Neural Network	BART	Private EHR/EMR	English	ROUGE, BLEURT, BERTScore, Human Assessment	Similarity, Structure	Text Summarisation	Doctor-Patient Conversations
[40]	2023	Augmentation, Annotation, Usefulness	Neural Network	ChatGPT	Prompting	English	Classification Accuracy, Human Assessment, Pattern Match	Similarity, Structure, Test on Utility	Named Entity Recognition, Relation Extraction	Biological Concepts and Relations
[107]	2023	Augmentation	Neural Network, Knowledge Source	ChatGPT-3.5-turbo, UML	Publicly Available (MIMIC-III)	English	ROUGE, BERTscore, BLEURT	Similarity, Test on Utility	Text Summarisation	Medical Conversations
[31]	2023	Usefulness	Neural Network	GPT-2	Publicly Available	French	BLEU, Pattern Match, Human Assessment	Similarity, Structure, Test on Utility	Named Entity Recognition	Biological Concepts and Relations
[54]	2023	Usefulness	Neural Network	GPT-3	Private EHR/EMR	English	Human Assessment (Turing Test), Classification Accuracy	Structure, Test on Utility	De-Identification, Relation Extraction, Question Answering	History of Present Illness (HPI)
[109]	2023	Augmentation	Neural Network	GPT (ChatGPT, Bard)	Publicly Available	English	BERTScore, Classification Accuracy	Similarity, Test on Utility	Phenotype Classification	Doctor-Patient Conversations (Mental Health)
[41]	2023	Augmentation, Annotation	Neural Network	GPT-4	Manual Collection and Curation, Public Available. (MIMIC-III)	English	Human Assessment (Pattern Match), Classification Accuracy	Structure, Test on Utility	Phenotype Classification	History of Present Illness (HPI)
[59]	2023	Augmentation, Privacy-preserving	Neural Network	ChatGPT, PubMedBERT	Prompting	English	Classification Accuracy	Test on Utility	Relation Extraction, Named Entity Recognition, Question Answering	Clinical Transcripts

Continued on next page

Table 2: Summary of selected articles (Continued)

Article	Year	Purpose	Generation Method	Approach / Architecture	Data Source	Language	Evaluation Method	Evaluation Paradigm	Utility	Clinical Source
[71]	2024	Assistive Writing	Neural Network	ChatGPT	Prompting	English	Human Assessment	Structure, Test on Utility	Question Answering	Medical Consultation
[72]	2024	Assistive Writing	Neural Network	ClinicalBLIP	Publicly Available (IU X-RAY, MIMIC-CXR)	English	ROUGE, METEOR	Simlarity	Report Generation	X-Ray
[99]	2024	Augmentation	Neural Network, Text Processing	ChatGPT, BART, T5, EDA	Publicly Available	English	ROUGE, CIDEr, METEOR, BERTscore	Similarity	Diagnose Prediction, Disease Classification	Population, Intervention, Comparison, and Outcome (PICO)
[47]	2024	Usefulness	Neural Network	ChatGPT	Prompting	English	Human Assessment	Structure Comparison (Human vs. Synthetic text)	Report Generation	Discharge Summaries
[17]	2023	Assistive Writing	Neural Network	GPT	Online source	Chinese	PPL	Structure, Test on Utility	Question Answering	Doctor-Patient Conversations
[75]	2024	Assistive Writing	Neural Network	VAE, GAN (LSTM Generator, CNN Discriminator)	Publicly Available (MIMIC-III)	English	BLEU, PPL, WER, Human Assessment	Structure, Similarity	-	Clinical Transcripts
[76]	2024	Assistive Writing	Neural Network	BART	Publicly Available	English	ROUGE-L, BERTscore	Similarity	Report Generation	Patient Laboratory Test
[104]	2024	Augmentation	Neural Network	BioGPT	Publicly Available	English	BLEU, Classification Accuracy	Similarity, Test on Utility	Relation Extraction	Biological Concepts and Relations
[60]	2024	Assistive Writing, Privacy-preserving	Neural Network	DP-GPT, BioGPT	Private EHR/EMR	English	Human Assessment, ROUGE-L, Privacy Budget	Similarity, Structure, Privacy	Report Generation	Endoscopy Reports
[105]	2024	Augmentation	Neural Network	Mistral, Llama, Gemma	Publicly Available	English	Jaccard, BERTscore, Human Assessment	Similarity, Structure	De-Identification	Discharge Summaries
[79]	2024	Assistive Writing	Neural Network	Transformers	Publicly Available (MIMIC-CXR)	English	ROUGE, BLEU, METEOR, Human Assessment	Similarity, Structure	Report Generation	Radiology
[82]	2024	Assistive Writing	Neural Network	Transformers (CLIP)	Publicly Available	English	BLEU, ROUGE, METEOR, BERTScore	Similarity, Test on Utility	Report Generation, Question Answering	Radiology
[61]	2024	Privacy-preserving	Neural Network, Knowledge Source	GPT (Mistral), RL, UML	Publicly Available (MIMIC-III)	English	SemScore	Similarity, Privacy	-	Discharge Summaries
[85]	2024	Assistive Writing	Neural Network	Transformers (PLM), GPT (Galactica)	Publicly Available	English	BLEU, ROUGE, ROUGE-L, METEOR	Similarity	Question Answering	Biological Concepts and Relations
[49]	2024	Usefulness	Neural Network	Transformers (T5)	Online source	English	Classification Accuracy, Human Assessment	Structure, Test on Utility	Named Entity Recognition	Population, Intervention, Comparison, and Outcome (PICO)
[87]	2024	Assistive Writing	Neural Network	LDM (VAE, CNN ResNet)	Publicly Available (IU X-RAY, MIMIC-CXR)	English	BLEU, ROUGE-L	Similarity, Test on Utility	Report Generation	X-Ray, MRI, CT-scan
[51]	2024	Usefulness	Neural Network	ChatGPT-3.5-turbo	Prompting	English	Human Assessment	Structure	-	History of Present Illness (HPI)

Continued on next page

Table 2: Summary of selected articles (Continued)

Article	Year	Purpose	Generation Method	Approach / Architecture	Data Source	Language	Evaluation Method	Evaluation Paradigm	Utility	Clinical Source
[28]	2024	Annotation, Corpus Building	Neural Network	GPT-4	Private EHR/EMR	Norwegian	Human Assessment, Classification Accuracy	Similarity, Structure, Test on Utility	De-Identification	Discharge Summaries
[88]	2024	Assistive Writing	Neural Network	Transformers (VED)	Publicly Available (MIMIC-CXR)	English	BLEU, ROUGE-L, F1cXb	Similarity, Test on Utility	Report Generation	Radiology
[52]	2024	Usefulness	Neural Network	ChatGPT-3.5	Publicly Available (MIMIC-III), Prompting	English	Classification Accuracy	Test on Utility	Disease Classification	Discharge Summaries
[53]	2024	Usefulness	Neural Network	ChatGPT-3.5-Turbo, BioGPT, GPT-2, distil-GPT2, CerebroGPT	Private EHR/EMR	English	BLEU, ROUGE, Cosine, TF-IDF, Human Assessment	Similarity, Structure, Test on Utility	Report Generation	Radiology (Cerebrovascular)
[33]	2024	Corpus Building	Neural Network	Claude-3-Opus, GPT-4	Publicly Available	English, Arabic	ROUGE-L, BERTScore, Human Assessment	Similarity, Structure	-	Doctor-Patient Conversations
[66]	2024	Corpus Building	Neural Network	GPT-4	Manual Collection and Curation, Prompting	English	ROUGE-L, Human Assessment, Classification Accuracy	Similarity, Structure, Test on Utility	Named Entity Recognition	History of Present Illness (HPI)
[67]	2024	Corpus Building	Neural Network	GPT (Llama)	Publicly Available	English	Human Assessment	Structure, Test on Utility	Question Answering	Discharge Summaries
[89]	2025	Assistive Writing	Neural Network	Transformers (BART)	Publicly Available	English	ROUGE, BLEU	Similarity	Question Answering	Doctor-Patient Conversations
[106]	2025	Augmentation	Neural Network	Llama, Mistral, Gemma	Publicly Available	English	BLEU, ROUGE, ROUGE-L, WER, BERTScore, Human Assessment	Similarity, Structure	-	Doctor-Patient Conversations
[108]	2024	Augmentation	Neural Network	GPT-4	Prompting	English	Classification Accuracy	Test on Utility	Diagnose Prediction	Medical Conversations
[39]	2025	Augmentation, Annotation	Neural Network	ChatGPT-3.5	Private EHR/EMR, Prompting	English	Classification Accuracy, Pattern Match	Structure, Test on Utility	Phenotype Classification	History of Present Illness (HPI)