arXiv:2304.01938v1 [physics.med-ph] 1 Apr 2023

大语言模型在高度专业化主题——放射肿瘤学物理领域的评估

Jason Holmes[1], Zhengliang Liu[2], Lian Zhang[1], Yuzhen Ding[1], Terence T. Sio[1], Lisa A. McGee[1],Jonathan B. Ashman[1], Xiang Li[3], Tianming Liu[2], Jiajian Shen[*1], 以及 Wei Liu[†1]

[1]梅奥诊所放射肿瘤科[2]佐治亚大学计算学院[3]麻省总医院放射科

摘要

我们开展了首项研究，探究大语言模型（LLM）回答放射肿瘤学物理问题的能力。由于 AP Physics、LSAT 和 GRE 等热门考试的应试者人数众多，且有大量备考资源流通，这些考试可能无法准确评估大语言模型的真实潜力。本文提出在一个高度专业化的主题——放射肿瘤学物理领域评估大语言模型；这一主题或许更贴近科学界和医学界的需求，也可作为一项有价值的大语言模型基准测试。我们基于在梅奥诊所积累的专业能力，编制了一份包含 100 道放射肿瘤学物理问题的试卷。我们评估了四种大语言模型：ChatGPT（GPT-3.5）、ChatGPT（GPT-4）、Bard（LaMDA）和 BLOOMZ，并将其表现与医学物理学家及非专业人士进行了比较。平均而言，ChatGPT（GPT-4）的表现优于所有其他大语言模型，也优于医学物理学家。提示 ChatGPT（GPT-4）先解释再作答后，其表现进一步提升。无论答案正确与否，ChatGPT（GPT-3.5 和 GPT-4）在多次试验中的选项表现出高度一致性，而人类测试组未呈现这一特征。我们采用一种新方法评估 ChatGPT（GPT-4）的演绎推理能力，即将正确答案替换为“以上选项均不正确”。ChatGPT（GPT-4）展现出令人意外的准确性，表明其可能具有涌现能力。最后，尽管 ChatGPT（GPT-4）总体表现良好，但其内在特性使其无法通过对多次试验采用多数投票来进一步提升表现。相比之下，医学物理学家团队通过多数投票，取得了远胜于 ChatGPT（GPT-4）的表现。本研究表明，大语言模型有望作为知识丰富的助手，与放射肿瘤学专家协同工作。

1 引言

大语言模型（LLM）的出现彻底改变了自然语言处理（NLP） [1]。传统的 NLP 范式遵循典型流程：通过监督训练，为下游应用创建定制解决方案。例如，预训练的 BERT[2] 模型必须添加额外的网络层，然后在带标签的训练数据上进行微调，才能执行序列分类或问答等任务。在某些情况下，为了

[*]共同通讯作者：shen.jiajian@mayo.edu
[†]共同通讯作者：liu.wei@mayo.edu

达到可接受的性能 [3, 4]。例如，AgriBERT [5]在农业相关文本数据上进行了预训练，以便妥善处理食品与农业领域的自然语言处理任务。然而，大语言模型规模庞大且具备出色的少样本学习能力，因此能够通过上下文学习解决自然语言处理问题，从而减少甚至消除对标注训练样本的需求[6, 7]。在上下文学习过程中，大语言模型根据提示，利用少量示例（甚至完全没有示例）进行泛化。这些提示通常是描述性用户输入，用于说明期望大语言模型给出的回答 [6, 8]。例如，“总结以下文本”就是一个简单明了的提示，要求语言模型对输入文本进行总结。总体而言，大语言模型为自然语言处理提供了一种新颖而简化的工作流程，有望免去监督微调及其相关复杂步骤，例如超参数调优和模型架构修改。此外，上下文学习显著减少了对耗时且成本高昂的人工标注的需求 [6, 9]。由于医学和科学领域可用的数据有限，这种方法尤其适用 [10, 11, 4, 12]。

近几个月来，ChatGPT 的兴起引起了全世界的关注[1]。凭借前所未有的语言能力，以及通过聊天机器人界面向公众开放的便利性，它在全球范围内广受欢迎。ChatGPT 基于强大的 GPT-3 模型 [6], ，这是历史上首批大语言模型之一。拥有 1750 亿个参数的 GPT-3 使用大规模数据集进行训练，其中涵盖了多种互联网数据（包括 Common Crawl [2] 和维基百科 [3]）。通过文本生成这一目标，它在从文本摘要到命名实体识别（NER）等各种自然语言处理任务中展现出卓越的性能（事实上，许多自然语言处理任务都可以转化为某种形式的文本生成）。ChatGPT 继承了 GPT-3 的这些能力，也继承了存储在参数空间中的、涵盖多种主题的海量知识。更重要的是，ChatGPT 通过基于人类反馈的强化学习（RLHF）进行训练。这是一种强化学习过程，通过用户反馈纳入人类偏好和人类排序的价值判断。该过程对模型进行调优，使其生成最能吸引人类用户且与之最相关的内容。ChatGPT 的能力推动了多种实际应用，从论文写作到代码生成 [13]。

迄今最强大的大语言模型之一是 GPT-4 [4]，它是 GPT-3 的继任者。虽然 OpenAI 尚未透露太多技术细节，但 GPT-4 在多种情境下的表现优于基于 GPT-3.5 的 ChatGPT [14, 9, 12, 15]。事实上，截至2023年3月，GPT-4 已为 Microsoft 的搜索引擎必应提供支持 [16], ，这展现了基于大语言模型的搜索潜力。此外，与前代模型不同，GPT-4 是一种多模态模型，能够接收图像输入，这无疑将带来更多有趣的应用。

GPT-4 在各种学术与职业基准测试中表现出色[14]。例如，GPT-4 通过美国执业医师资格考试的成绩高出及格线 >20% [17]。事实上，GPT-4 在 SAT、统一律师资格考试以及 GRE 的语言部分均取得了超过第 90 百分位的成绩（见《GPT-4 技术报告》中的图4 [14]），这些考试几乎都包含选择题部分。的确，选择题考试是评估大语言模型的常用方式 [18, 14, 19]。用于评估大语言模型的大多数选择题考试，都涉及在学术领域中最受广泛研究的主题。例如，2022年，参加 AP 物理考试的考生有 144,526 人 [20], ，参加 LSAT 的考生有 128,893 人 [21], ，参加 GRE 的考生约有 342,000 人 [22]。由于参加这些考试的考生人数众多

[1] https://openai.com/blog/chatgpt [2] http://commoncrawl.org/ [3] https://www.wikipedia.org/ [4] https://openai.com/research/gpt-4

，而且考试成绩对大学录取至关重要，因此相关资源（包括可在互联网上获取的文本数据）极为丰富。无论评估的是哪种大语言模型，这些考试及相关备考材料易于获取且无处不在，实际上都使得模型在这些考试中取得高分变得不足为奇。因此，也有必要研究大语言模型在较为冷门、专业性更强的主题上的表现，因为这类主题的训练数据规模可能小得多。2022年，只有 162 名医学院毕业生申请了放射肿瘤学住院医师项目 [23]。因此，放射肿瘤学物理是一个普通人群相对陌生的主题，与知识库中广泛涵盖的主题相比，它或许能更公平地用于评估大语言模型。在大语言模型的背景下，冷门主题可能带来最大的教育机会，同时也可能给普通人群带来最大的风险，因为人们可能会更依赖模型对此类主题的回答，而这些回答的准确性却较低，其中的错误也更不容易被察觉。

评估大语言模型准确性的一项重要因素，是确保测试题未包含在训练数据中 [24], ，即未受到污染。确保这一点的最佳方法是专门编写新的测试题。本研究为此设计了一场选择题考试。研究选取了四种基于 Transformer 的大语言模型进行评估：ChatGPT（GPT-3.5） [6], ChatGPT（GPT-4） [14], Bard（LaMDA） [25], 和 BLOOMZ [26]。研究将这些模型的结果与放射肿瘤学专家及非专家的结果进行比较。此外，研究还进一步探讨了如何改进 ChatGPT（GPT-4）的回答，以及其演绎推理能力。

2 相关工作

2.1 大语言模型

基于 Transformer 的预训练语言模型（PLM），如 BERT [2] 和 GPT [27], ，彻底改变了自然语言处理。它们在众多任务中超越了以往的方法（例如，基于循环神经网络的模型），推动了人们对语言模型的兴趣，也提高了语言模型的普及程度 [28]。通常，预训练语言模型可分为三类：自回归模型（如 GPT）、掩码语言模型（如 BERT）以及编码器-解码器模型（如 BART [29] 和 T5 [30]）。近年来，超大规模语言模型不断涌现，包括 GPT-3 [6], Bloom [31], PaLM [32], 和 OPT[33]。这些模型以变换器架构为基础，借鉴了 BERT 和 GPT 等模型的设计思路，但其规模大得多。

大语言模型的目标是从输入文本中准确学习上下文相关且特定于领域的潜在特征表示 [28]。例如，“discharge”的向量表示在医学领域和通用领域之间可能存在显著差异。较小的语言模型通常需要在下游任务上进行持续预训练和监督微调，才能达到可接受的性能 [3, 4]。然而，非常大的语言模型有望在保持具有竞争力的结果的同时，消除对微调的需求 [6]。

除了模型架构、规模和训练策略方面的进步外，大语言模型还可以通过基于人类反馈的强化学习（RLHF）进一步与人类偏好保持一致[34]。这种方法已应用于多种大语言模型，例如 Sparrow [35], InstructGPT[36], 和 ChatGPT。InstructGPT 基于 GPT-3，通过一套训练流程进行训练，在此过程中，借助人工生成的排序反馈优先考虑用户偏好。作为 InstructGPT 的后续模型，ChatGPT 同样采用基于人类反馈的强化学习，重点在于遵循提示并生成全面的回答。OpenAI 还实施了护栏机制，以防止生成带有偏见和不良的输出 [37]。ChatGPT 已成为一款非常成功的人工智能聊天机器人，利用

GPT-3.5 的能力来促进类似人类的互动。

基于人类反馈的强化学习通过根据人工标注者对生成结果的排名来训练奖励模型，将人类反馈融入最优结果的生成与选择中 [38]。随后，该奖励模型会奖励最符合人类偏好与价值观的输出。我们认为，这些突破性创新使 ChatGPT 成为本研究的理想候选对象。

GPT-4 的近期发展显著推动了语言模型的发展水平。GPT-4 展现出更强的推理能力、创造力、图像理解、上下文理解和多模态能力，因此能够生成更加复杂多样的回答。大型 GPT 模型的成功推动了针对特定领域开发专用模型的探索，例如面向医疗与健康应用的大语言模型，这有望彻底改变这些领域。

2.2 语言模型与考试

大语言模型具有卓越的自然语言理解能力。此外，它们经过海量数据训练，积累了丰富的知识。这些特点使大语言模型成为参加学术与职业基准测试的理想对象。

OpenAI 最近发布了文献中的首项研究，评估大语言模型在为受过良好教育的人类设计的学术与职业考试中的表现 [14]。结果表明，GPT-4 在多种科目考试中表现出色，范围从统一律师资格考试到 GRE。此外，Microsoft 的一项研究表明，GPT-4 在美国执业医师资格考试（住院医师的职业考试）中取得了大幅领先的成绩 [17]。

本研究首次评估了大语言模型在放射肿瘤学和医学物理学领域的表现，我们相信，这项研究能够启发未来对大语言模型在高度专业化医学分支领域的评估研究。

2.3 提示工程

为训练或微调自然语言处理模型而收集和标注数据，可能需要大量资源且成本高昂，尤其是在医学领域 [4, 9, 12]。近期研究表明，通过使用提示，大规模预训练语言模型（PLM）无需微调即可适应下游任务 [6, 8]。

提示由一组指令构成，用于定制或优化大语言模型的回复。提示的作用不止于描述任务或指定输出格式。事实上，提示可以经过设计，用于创建新颖的交互方式。例如，可以提示 ChatGPT 通过模拟终端命令来模拟网络安全 breach [39]。此外，还可以通过自适应过程使用提示生成更多提示 [39]。

提示工程的兴起标志着自然语言处理新时代的开始 [8]。精心设计的提示在各种复杂应用中具有巨大潜力，这一点毋庸置疑。然而，在大语言模型时代，确定理想的提示是一项独特的挑战。目前，提示可以由人工设计，也可以自动生成 [40, 8]。尽管自动生成的提示在某些任务中的表现可能优于人工提示 [8], ，但它们往往可读性和可解释性较差 [41, 8]。因此，在可解释性至关重要的领域（如临床实践和研究）中，人工生成提示可能更受青睐。在本研究中，我们基于自身的

放射肿瘤学和医学物理学经验，设计了一套提示和思维链提示，并评估其对大语言模型的影响。

3 方法

本研究由一位经验丰富的医学物理学家编制了一份放射肿瘤学物理领域的100题选择题考试，旨在符合美国放射肿瘤学会（ASTRO）推荐的放射肿瘤学住院医师教育课程[42]。考试内容包括以下主题：基础物理（12题）、辐射测量（10题）、治疗计划（20题）、放射治疗中的成像方式及其应用（17题）、近距离放射治疗（13题）、高级治疗计划和特殊操作（16题），以及安全、质量保证（QA）和辐射防护（12题）。100道题中有17道需要进行数值计算（基于数学）。考试题目列于附录第B节。

3.1 大语言模型得分与人类得分的比较

每个大语言模型均接受了5次独立试验（试验1至试验5），每次输入的都是这份放射肿瘤学物理领域的100题选择题考试；BLOOMZ除外，仅接受了一次测试。每次试验均从新对话线程开始，或在重置后开始；首先向大语言模型发送初始化提示词，告知其即将接受测试。随后，向大语言模型发送指令，并分批提供20道题，直至完成考试。每次试验中的指令都要求大语言模型只返回正确答案，不作解释。由于观察到若不重复说明，大语言模型较少遵循指令，因此每批题目中都附有指令。对于无法一次接受20道题的大语言模型，则改为每批提供10道题（Bard）。如果大语言模型未返回所有答案，下一批会同时包含未作答题目和完整的下一批题目。这种情况很少发生。在每次测试试验中，全局提示和指令提示词的措辞均有所不同，以控制提示词差异造成的响应噪声。初始化提示词和指令提示词列于表1。

比较了大语言模型测试分数及其分布，并将其与两个人类群体（医学物理学家组和非专家组）的得分进行了比较。医学物理学家组包括四名经验丰富、经认证的医学物理学家、三名医学物理学住院医师和两名医学物理学研究员。非专家组包括六名拥有电气工程、计算机工程或计算机科学高级学位的人员，但他们此前均无已知的放射肿瘤学物理相关经验或教育背景。每位人类考生均有3小时参加考试，考试为闭卷，允许使用基础计算器。比较大语言模型得分与人类得分时，评估了平均得分、得分一致性以及答案置信度。

为量化评分成功的整体一致性，计算了标准差以及试验间的平均相关性；后者定义为试验间皮尔逊相关矩阵上三角值的平均值。平均相关性反映各次试验中正确得分的一致程度，其中1表示分布完全相同，0表示分布完全随机，-1则表示分布完全负相关。

为了量化大语言模型和人类群体给出的答案的置信度，

表1：每项试验中使用的大语言模型提示词。

试验	初始化提示词	指令提示词
试验1	我是一名放射治疗研究人员。My研究小组想研究 ChatGPT 针对这一主题给出的答复：放射肿瘤学物理。现在我将开始提问，内容涉及放射肿瘤学物理。	指令：对于每道选择题，请给出正确答案，不要提供任何理由。
试验 2	我想通过提出一些问题，来评估你在放射肿瘤学物理方面的知识。多项选择题。	请仅给出题目标签和正确答案对应的字母。
试验 3	请回答以下练习题，就像一名正在备考的住院医师专科认证考试。	只在你的回-答中给出正确答案。不要解释答案。
试验 4	我们想测试你对放-射肿瘤学物理学的理解。因此，我们准备了一些问题要问你。	请在回答中仅报告题目标签及其对应的答案。
试验 5	我会问你一些多项选择题——目。	指令：仅回复正确的字母选项。

统计了每道题在所有试验中的正确答案数量。例如，如果某个大语言模型对同一道题答对了 5 次，那么所有 5 次答案均正确的题目比例就增加 1%（因为共有 100 道题）。此外，还将测试结果与应试者随机猜测时预期出现的分布进行了比较。随机猜测时，5 次试验中的正确答案预期数量平均约为 $0.25 \times 5 = 1.25$（100 道题中，98 道有 4 个选项，2 道有 5 个选项）。利用这个数值，可以根据由此得到的泊松分布估算每道题的正确答案出现次数。

最后，将 ChatGPT（GPT-3.5 和 GPT-4）与 Bard 的得分和人类得分进行了比较，其中人类得分是通过多数投票计算得出的。

3.2 提高 ChatGPT（GPT-4）的准确率——先解释，再作答

由于基于 Transformer 的大语言模型会根据先前的上下文预测下一个词，因此已有研究表明，如果提示规模足够大的大语言模型以逐步方式推导答案，就能提高回答的准确率 [24, 43, 44]。研究采用这一策略评估 ChatGPT（GPT-4），以检验提示其先解释再作答是否能提高得分。初始化提示词与试验1中的相同，但试验1的指令提示词改为以下内容：“指令：对于每一道多项选择题，请先解释答案，然后给出正确答案（字母选项）。”随后将这些测试结果与原始的未要求解释的测试结果进行了比较。

3.3 测试 ChatGPT（GPT-4）的演绎推理能力

在选择题中，当题目和答案经常出现在同一语境中时，大语言模型的表现会最好。然而，如果正确答案与题目没有共同语境，会发生什么呢?

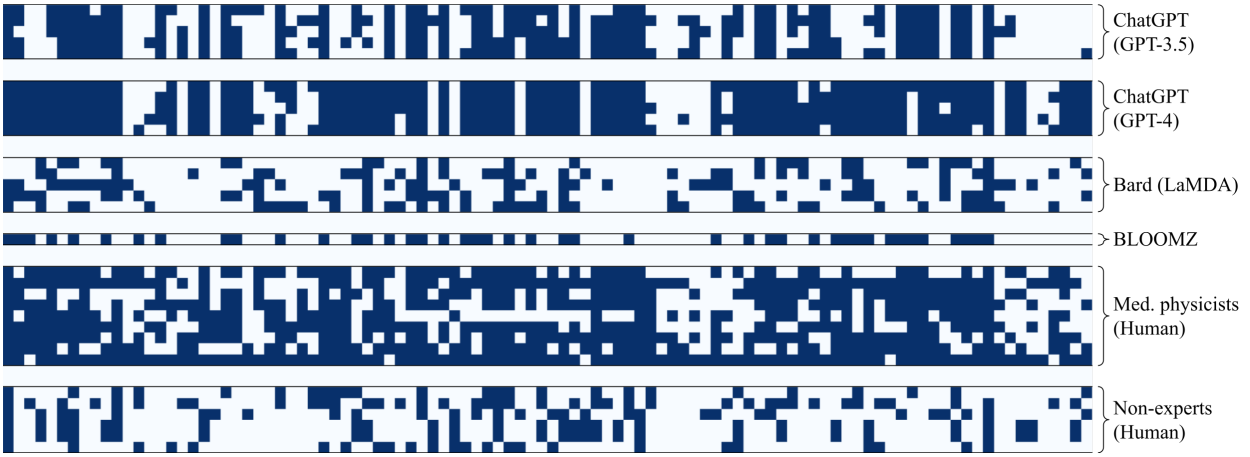


图1：各次测试的原始得分，行表示不同的测试，列表示测试题。深色方格表示答案正确。

例如，答案是“以上选项均不正确”时？在这种情况下，大语言模型必须通过排除其他所有答案来推断正确答案，而这些答案很可能都与题目共享语境。对于大语言模型而言，这种情况似乎尤其具有挑战性。为了研究 ChatGPT（GPT-4）的演绎推理能力，我们对100题选择题考试中的每道题都进行了修改。删除每道题的正确答案，并将其替换为“以上选项均不正确”。这种语境缩减转换不能用于人类，因为人类会察觉到其中的规律。因此，专为人类设计、后来又被用于大语言模型训练数据的测试中，很可能不会有这种转换的示例。据此，我们假设大语言模型不会察觉到这一规律。我们使用试验1提示词将修改后的考试交给 ChatGPT（GPT-4）作答，随后按照第3.2节中所述，通过先解释再作答的方式测试其准确率是否有所提高。

4 结果

4.1 大语言模型得分与人类得分的比较

原始得分和平均测试分数分别见图 1和 2 ，其中，大语言模型的平均测试分数为5次试验的平均值（BLOOMZ除外，仅进行1次试验），人类的平均测试分数则为各自群体的平均值（见第 3.1节）。总体而言，每个大语言模型的表现都优于非专家人类组，但只有ChatGPT（GPT-4）的表现优于医学物理学家组。在数学类问题上，医学物理学家的表现优于ChatGPT（GPT-4）。

从图1所示的原始得分可以看出，每个大语言模型和人类组的各次试验结果均存在差异；这种差异不仅体现在总分的不确定性上，也体现在各题答对次数上。各次试验之间的标准差和平均相关性见图 3a 和 3b。与人类群体相比，大语言模型的得分和答案一致性高得多，表现为评分的标准差较低，各次试验之间的平均相关性较高。

从4所示的结果来看，Bard 的表现略优于非专家组，但两组的表现都与随机猜测者相近。ChatGPT（GPT-3.5 和 GPT-4）与医学物理学家

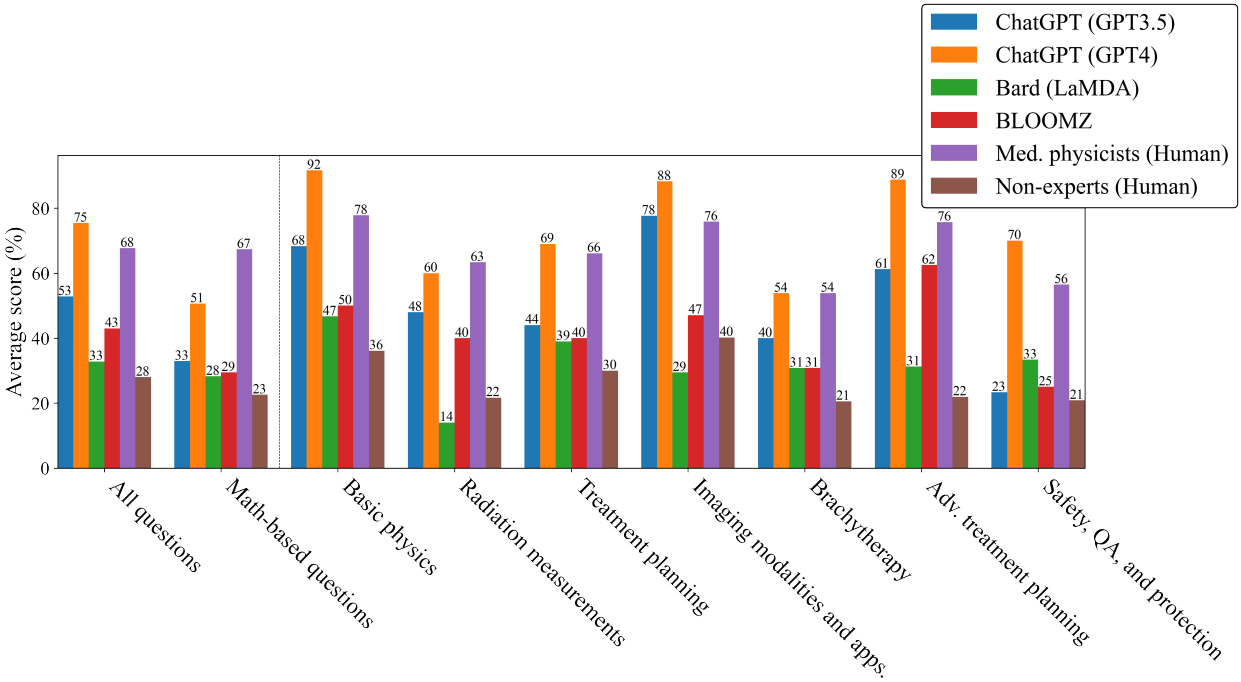
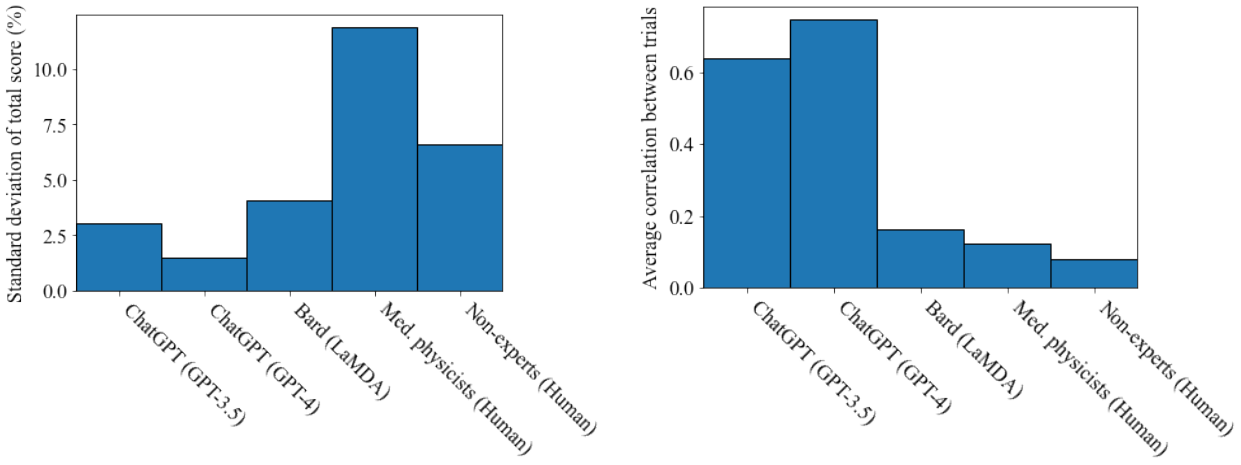


图 2：各大语言模型按类别划分的平均测试分数。



（a）总分的标准差。

（b）各试验之间的平均相关性。

图 3：大语言模型与人类评分的一致性。

结果与随机猜测毫无相似之处。ChatGPT（GPT-3.5）要么信心十足，每次试验中有 35% 的答案正确；要么感到困惑，每次试验中有 28% 的答案错误。ChatGPT（GPT-4）信心更强，每次试验中有 67% 的题目答对，但也表现出困惑倾向，每次试验中有 14% 的题目答错。作为一个群体，医学物理学家既没有表现出极强的信心，也没有显得困惑，不过在选择正确答案时往往意见一致。

尽管 ChatGPT（GPT-3.5 和 GPT-4）的总体得分不错，但图 4 所示的评分分布表明，如果这些大语言模型能够协同作答，评分几乎不会有所提升，因为它们往往不是信心十足，就是感到困惑，而且答案的变化很小。Bard（LaMDA）和非专家组协同作答也可能几乎没有提升，因为他们的答案成功率趋近于随机水平。然而，由于医学物理学家在正确答案上往往意见一致，因此可以预期，他们协同作答时得分会显著提高。为检验这一点，我们采用“多数投票”合并各组的答案。对于每道题，选择出现次数最多的选项作为该组的答案。如果出现平票，则从出现次数最多的选项中随机选择一个。图 5 展示了采用多数投票后的评分结果。如图所示，ChatGPT（GPT-3.5 和 GPT-4）的成绩仅略有提升，为 1%。Bard（LaMDA）和非专家组分别提升了 4% 和 3%。然而，医学物理学家组的成绩大幅提升了 23%。

4.2 Im ChatGPT（GPT-4）准确性提升——先解释后回答

图 6 展示了提示 ChatGPT（GPT-4）先解释后回答、从而允许答案逐步形成时的结果。ChatGPT（GPT-4）的总分提高了 5%，超过了此前的每一次试验。近距离放射治疗和数学类问题类别的提升最大。这些结果与先前研究一致；那些研究发现，对于足够大的大语言模型而言，这种能力是一种涌现特征 [44]。ChatGPT（GPT-4）的示例回答见附录第 A节。

4.3 测试 ChatGPT（GPT-4）的演绎推理能力

图 7 展示了演绎推理测试的结果：在全部 100 道题中，正确答案均被替换为“以上选项均不正确”。与原始测试相比，ChatGPT（GPT-4）的整体表现差得多。尽管总体表现较差，“先解释后回答方法”对提升其通过演绎推理解答问题的能力尤为重要。未先进行解释时，ChatGPT（GPT-4）数学类问题的答对率为 0%；采用“先解释后回答方法”后，答对率提高到 65%，仅比原试验中同样采用该方法的结果少答对一道题。

5 讨论

本研究旨在通过一项专为本研究设计的100题选择题考试，评估大语言模型在放射肿瘤学物理这一高度专业化领域的表现。考试内容见附录第 B 节。非专家组的评分结果表明，普通人群对放射肿瘤学物理知之甚少，其得分与随机猜测相近。Bard（LaMDA）的表现略优于非专家组，而BLOOMZ和ChatGPT（GPT-3.5和GPT-4）的表现则远优于非专家组。令人惊讶的是，GPT-4

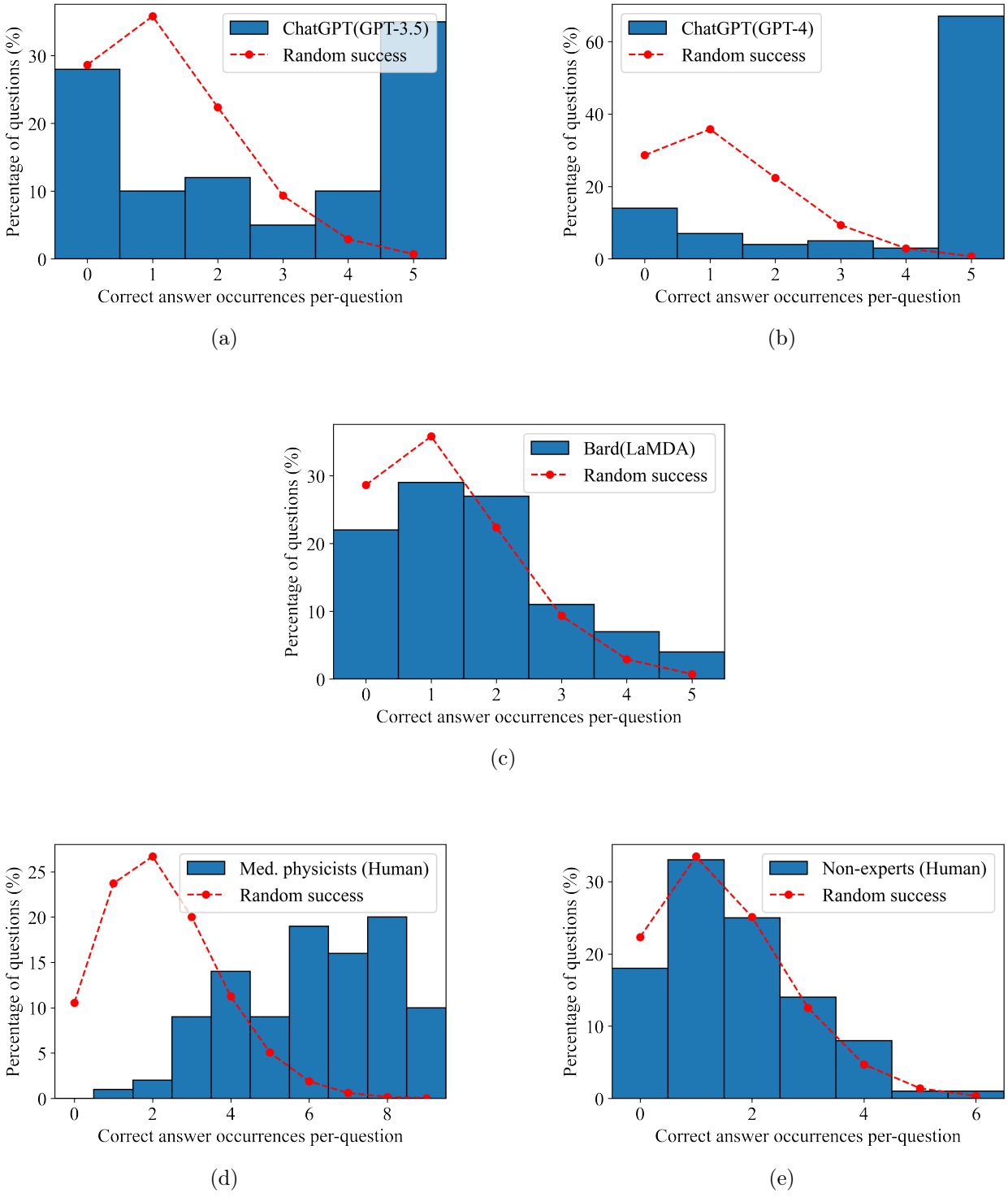


图4：答案置信度。各大语言模型和人类组在每个问题上的正确答案出现次数。红色虚线曲线表示根据泊松分布随机选择答案时的预期分布。

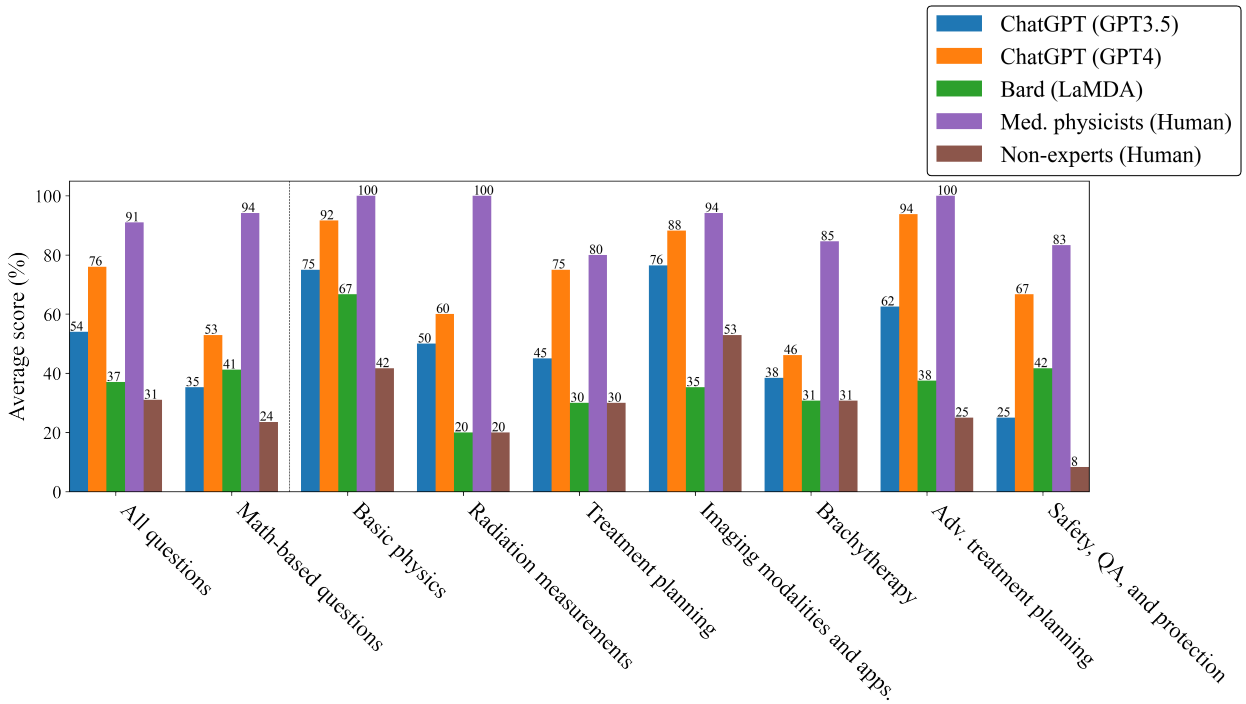


图5：按类别统计的得分；大语言模型的得分按各试验中的多数投票结果汇总，人类的得分则按组内多数投票结果汇总。

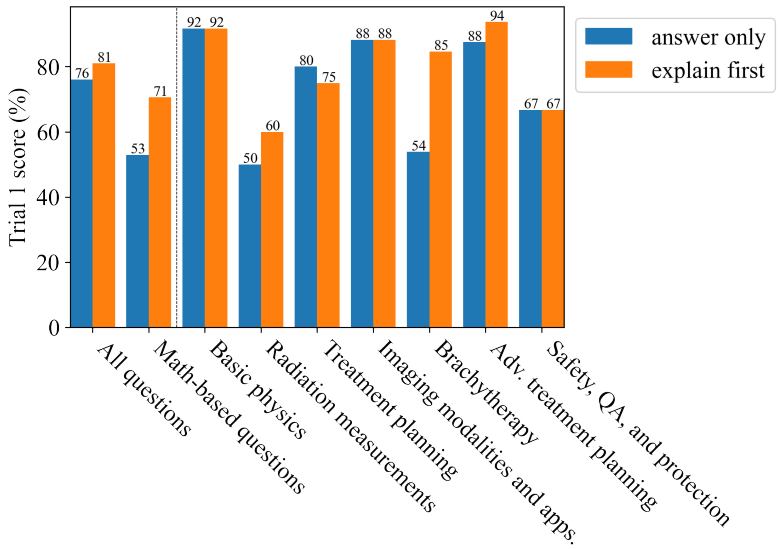


图6：　　试验1的改进归因于采用了先解释、再回答的方法　　d.

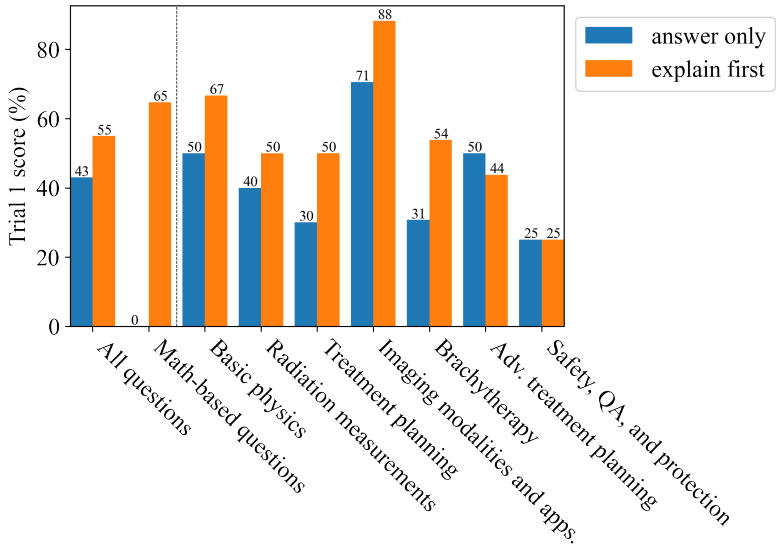


图7：将正确答案替换为“以上选项均不正确”后，试验1的评分。这种方法用于测试演绎推理能力；随后使用先解释后回答方法，表现有所提升。

能够在几乎所有子类别中超越普通医学物理学家的表现，并且在被要求先解释推理过程再作答时，答案准确率有所提高（图2 和6）。ChatGPT（GPT-4）展现出令人惊讶的演绎推理能力：在正确答案均被改为“以上选项均不正确”的100道题中，它都进行了作答；尤其是在采用先解释后回答方法时，总体得分达到55%。这一结果令人困惑，也可能是一种涌现属性。随着大语言模型参数数量增加，已知会出现涌现属性 [44]。这种新方法或可用于判断演绎推理能力是否会随着参数数量增加而提升。

尽管 ChatGPT（GPT-4）的总体表现优于医学物理学家，但本研究也提供了证据，表明单个大语言模型无法与少数几位协同工作的医学物理学家相匹敌（图 5）。可能的原因是，即使专业背景相似，不同个体的人类在能力和知识方面也存在显著差异。此外，虽然选择题的答案非对即错，但图 4 所示的评分次数分布表明，医学物理学家不太容易产生困惑；将整个医学物理学家群体的表现汇总后，他们通过多数投票选出正确答案的概率便高得多。ChatGPT（GPT-3.5 和 GPT-4）答错时，往往错得很有把握（表现出困惑）；答对时，也往往答得很有把握。我们的结果表明，在高度专业化主题方面具备专业能力的人类知道何时猜测、如何进行明智的猜测，而且即使没有选中正确答案，在推理过程中出错的可能性也较低。这种比较或许并不完全公平，因为如果以与 ChatGPT（GPT-3.5 和 GPT-4）完全相同的方式反复测试同一个人类受试者，他们也可能重复作答，并在个体层面表现出一定程度的困惑。不过，这一点或许无关紧要，因为经验丰富的医学物理学家为数众多，而能力可与 GPT-4 媲美的大语言模型却寥寥无几。ChatGPT（GPT-3.5 和 GPT-4）在正确与错误答案上表现出的高度一致性，可能意味着其在放射肿瘤学物理知识方面存在过拟合（或记忆）。无论如何，放射肿瘤学物理是一门

高度专业化的学科，ChatGPT（GPT-4）的表现令人瞩目，且在不久的将来很可能会继续提升。从实践角度来看，本研究表明，放射肿瘤学专家与 ChatGPT（GPT-4）协同工作具有很大潜力，可将其用作知识渊博的助手。

使用本研究所采用的这类考试来评估大语言模型，有一个缺点：这类考试并不能代表医学物理学家和放射肿瘤学专家日常临床工作中所处理的细致而复杂的问题。因此，本研究报告的大语言模型与医学物理学家在放射肿瘤学物理考试中的相对表现，可能会误导人们对大语言模型与个别医学物理学家能力等效程度的判断。此外，GPT-4在这项涵盖高度专业化主题的类似认证的考试中表现出色，这表明受评估的知识可能具有一定的表面性。否则，我们就不得不考虑GPT-4是否有能力胜任医学物理学家的工作，而这似乎极不可能。因此，随着大语言模型不断进步，放射肿瘤学界乃至更广泛的医学界可能需要重新评估认证程序，因为人类投入大量精力获取这类表层知识的必要性将会降低。鉴于此，大语言模型或许可以用来检验知识的表面性。也许应更加重视那些大语言模型尚未掌握的知识。

5.1 在放射肿瘤学中应用大语言模型

本研究延续了一系列将先进的自然语言处理方法应用于放射肿瘤学的研究。例如，Rezayi 等人 [11] 使用大规模放射肿瘤学文献语料库，以及来自某医院、经过筛选和匿名化的文本数据集，对 BioBERT 进行训练，从而构建了专用于放射肿瘤学的语言模型 ClinicalRadioBERT。Liao 等人 [45] 提出了一种框架，将基于 Transformer 的语言模型的注意力引导至对分类决策有显著影响的重要输入词元。这种方法对于仅有少量已标注样本时的少样本学习尤为重要。在放射肿瘤学领域，由于患者隐私方面的顾虑，很难收集和整理大量符合特定要求的多机构患者数据，因此少样本学习是一项常见挑战。相关地，ChatGPT 在测试准确率和增强样本分布方面表现出优于当前先进文本数据增强方法的效果，可作为一种有效的文本数据增强方法 [9], ，也可用于应对少样本学习的挑战。

此外，大语言模型还可用于数据去标识化等创新应用。例如，在对临床笔记进行去标识化时，GPT-4 的准确率达到 99%，优于 ChatGPT 和其他语言模型竞争产品 [12]。这对放射肿瘤学以及整个医学领域的各个专业都极为重要，因为在跨机构临床协作和学术研究中，对数据进行匿名化往往十分繁琐。语言模型的其他应用还包括为肿瘤学构建领域专用知识图谱 [46]，而无需临床医生或其他领域专家进行人工标注。

5.2 放射肿瘤学中的多模态模型

多模态模型是语言模型的未来 [1]，在医学诊断中也至关重要 [47]。早期一些使用多模态数据的大语言模型研究包括 ChatCAD [48],，这是一个整合图像和文本以用于计算机辅助诊断的框架。它支持多种诊断网络，例如用于病灶分割和报告生成的网络。在该框架中，可以使用 ChatGPT 来增强这些网络的输出。

GPT-4 支持图像等多模态输入，这进一步释放了大语言模型在放射肿瘤学领域的潜力。我们有必要研究未来能够将文本、图像、剂量学数据及其他模态整合到诊断与治疗流程中的模型和应用。我们认为，此类多模态模型与人脑具有内在的亲和性 [1]，未来面向医学的大语言模型可以从神经科学和自然语言处理的进展中获得启发。

6 结论

本研究表明，大语言模型，尤其是 ChatGPT（GPT-4），不仅能够出色地回答放射肿瘤学物理这类高度专业化主题的问题，还展现出令人惊讶的演绎推理能力。虽然 GPT-4 的平均表现优于单个医学物理学家，但一小组医学物理学家协同工作仍取得了更好的结果，这彰显了人类专业人士知识与专业能力多样性的重要性——或许对大语言模型而言亦是如此。这进一步凸显了将大语言模型的能力与人类洞见和经验相结合，以提升知识水平和决策能力的潜在益处。总体而言，研究结果表明，大语言模型在专业领域的应用前景广阔。

致谢

我们谨向许多自愿参加放射肿瘤学物理考试的人士（Yuzhen Ding、Yiran Cui、Yonghui Fan、Yikang Li、Riti Paul、Nupur Thakur、Sachin Chhabra、Chenbin Liu、Yang Li、Man Zhao、Yuchao Hu、Shumei Jia、Lian Zhang、Yao Xu、Hongying Feng 和 Yunze Yang）表示感谢。

参考文献

[1] Lin Zhao、Lu Zhang、Zihao Wu、Yuzhong Chen、Haixing Dai、Xiaowei Yu、Zhengliang Liu、Tuo Zhang、Xintao Hu、Xi Jiang 等人。脑启发式人工智能与通用人工智能相遇。*arXiv* 预印本 *arXiv:2303.15935*，2023年。[2] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。BERT：用于语言理解的深度双向变换器预训练。*arXiv* 预印本 *arXiv:1810.04805*，2018年。[3] Yu Gu、Robert Tinn、Hao Cheng、Michael Lucas、Naoto Usuyama、Xiaodong Liu、Tristan Naumann、Jianfeng Gao 和 Hoifung Poon。用于生物医学自然语言处理的领域专用语言模型预训练。*ACM 医疗保健计算汇刊（HEALTH）*，3(1):1–23，2021年。[4] Zhengliang Liu、Mengshen He、Zuowei Jiang、Zihao Wu、Haixing Dai、Lian Zhang、Siyi Luo、Tianle Han、Xiang Li、Xi Jiang 等人。医学图像分析中的自然语言处理综述。中南大学学报（医学版）= 中南大学学报（医学版），47(8):981–993，2022年。[5] Saed Rezayi、Zhengliang Liu、Zihao Wu、Chandra Dhakal、Bao Ge、Chen Zhen、Tianming Liu 和 Sheng Li。AgriBERT：用于食品与营养匹配的知识增强型农业语言模型。国际人工智能联合会议，2022年。

[6] Tom B. Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell、Sandhini Agarwal、Ariel Herbert-Voss、Gretchen Krueger、Tom Henighan、Rewon Child、Aditya Ramesh、Daniel M. Ziegler、Jeffrey Wu、Clemens Winter、Christopher Hesse、Mark Chen、Eric Sigler、Mateusz Litwin、Scott Gray、Benjamin Chess、Jack Clark、Christopher Berner、Sam McCandlish、Alec Radford、Ilya Sutskever 和 Dario Amodei。《语言模型是少样本学习器》，2020年。

[7] Qingxiu Dong、Lei Li、Damai Dai、Ce Zheng、Zhiyong Wu、Baobao Chang、Xu Sun、Jingjing Xu 和 Zhifang Sui。《上下文学习综述》。*arXiv* 预印本 *arXiv:2301.00234*，2022年。[8] Pengfei Liu、Weizhe Yuan、Jinlan Fu、Zhengbao Jiang、Hiroaki Hayashi 和 Graham Neubig。《预训练、提示与预测：自然语言处理领域提示方法的系统综述》。*ACM* 计算调查，55(9):1–35，2023年。[9] Haixing Dai、Zhengliang Liu、Wenxiong Liao、Xiaoke Huang、Zihao Wu、Lin Zhao、Wei Liu、Ninghao Liu、Sheng Li、Dajiang Zhu 等。《ChatAug：利用 ChatGPT 进行文本数据增强》。*arXiv* 预印本 *arXiv:2302.13007*，2023年。[10] Rina Kagawa、Masaru Shirasuna、Atsushi Ikeda、Masaru Sanuki、Hidehito Honda 和 Hirokazu Nosato。《一秒提升法：一种简单且经济高效的机器学习数据标注干预方法》。2022年。[11] Saed Rezayi、Haixing Dai、Zhengliang Liu、Zihao Wu、Akarsh Hebbar、Andrew H Burns、Lin Zhao、Dajiang Zhu、Quanzheng Li、Wei Liu 等。《ClinicalRadioBERT：融合知识的少样本临床笔记命名实体识别》。载于医学影像中的机器学习：第*13*届国际研讨会 *MLMI 2022*，与 *MICCAI 2022* 同期于新加坡举行，*2022*年*9*月*18*日，论文集，第269–278页。施普林格，2022年。

[12] Zhengliang Liu、Xiaowei Yu、Lu Zhang、Zihao Wu、Chao Cao、Haixing Dai、Lin Zhao、Wei Liu、Dinggang Shen、Quanzheng Li 等。《Deid-gpt：利用 GPT-4 实现零样本医学文本去标识化》。*arXiv* 预印本 *arXiv:2303.11032*，2023年。[13] Chengwei Qin、Aston Zhang、Zhuosheng Zhang、Jiaao Chen、Michihiro Yasunaga 和 Diyi Yang。《ChatGPT 能否解决通用自然语言处理任务？》*arXiv* 预印本 *arXiv:2302.06476*，2023年。[14] OpenAI。《GPT-4 技术报告》，2023年。

[15] Anis Koubaa。GPT-4 与 GPT-3.5：简明对决。2023年。

[16] The Verge。必应 AI 机器人一直在秘密运行 GPT-4，2023年。

[17] Harsha Nori、Nicholas King、Scott Mayer McKinney、Dean Carignan 和 Eric Horvitz。Capabilities of GPT-4 on 医学挑战题。*arXiv* 预印本 *arXiv:2303.13375*，2023年。[18] Andrew K Lampinen、Ishita Dasgupta、Stephanie CY Chan、Kory Matthewson、Michael Henry Tessler、Antonia Creswell、James L McClelland、Jane X Wang 和 Felix Hill。语言模型能否从上下文中的解释中学习？*arXiv* 预印本 *arXiv:2204.02329*，2022年。

[19] Jaromir Savelka、Arav Agarwal、Christopher Bogart 和 Majd Sakr。大语言模型（GPT）难以回答有关代码的多项选择题。*arXiv* 预印本 *arXiv:2303.08033*，2023年。

[20] 美国大学理事会。2022年学生成绩分布。访问日期：2023年3月31日。

[21] 法学院招生委员会。2023年考试报名者和考生。访问日期：2023年3月31日。[22] Achievable。2022年 ETS GRE 概览报告中的 GRE 关键统计数据，2023年。访问日期：2023年3月31日

[23] 美国国家住院医师匹配计划。匹配结果分析：美国MD医学院大四学生，2023年。访问日期：2023年3月31日。

[24] Sébastien Bubeck、Varun Chandrasekaran、Ronen Eldan、Johannes Gehrke、Eric Horvitz、Ece Kamar、Peter Lee、Yin Tat Lee、Yuanzhi Li、Scott Lundberg、Harsha Nori、Hamid Palangi、Marco Tulio Ribeiro 和 Yi Zhang。《通用人工智能的火花：GPT-4 的早期实验》，2023年。

[25] Romal Thoppilan、Daniel De Freitas、Jamie Hall、Noam Shazeer、Apoorv Kulshreshtha、Heng-Tze Cheng、Alicia Jin、Taylor Bos、Leslie Baker、Yu Du、YaGuang Li、Hongrae Lee、Huaixiu Steven Zheng、Amin Ghafouri、Marcelo Menegali、Yanping Huang、Maxim Krikun、Dmitry Lepikhin、James Qin、Dehao Chen、Yuanzhong Xu、Zhifeng Chen、Adam Roberts、Maarten Bosma、Vincent Zhao、Yanqi Zhou、Chung-Ching Chang、Igor Krivokon、Will Rusch、Marc Pickett、Pranesh Srinivasan、Laichee Man、Kathleen Meier-Hellstern、Meredith Ringel Morris、Tulsee Doshi、Renelito Delos Santos、Toju Duke、Johnny Soraker、Ben Zevenbergen、Vinodkumar Prabhakaran、Mark Diaz、Ben Hutchinson、Kristen Olson、Alejandra Molina、Erin Hoffman-John、Josh Lee、Lora Aroyo、Ravi Rajakumar、Alena Butryna、Matthew Lamm、Viktoriya Kuzmina、Joe Fenton、Aaron Cohen、Rachel Bernstein、Ray Kurzweil、Blaise Aguera-Arcas、Claire Cui、Marian Croak、Ed Chi 和 Quoc Le。《LaMDA：面向对话应用的语言模型》，2022年。[26] Niklas Muennighoff、Thomas Wang、Lintang Sutawika、Adam Roberts、Stella Biderman、Teven Le Scao、M Saiful Bari、Sheng Shen、Zheng-Xin Yong、Hailey Schoelkopf、Xiangru Tang、Dragomir Radev、Alham Fikri Aji、Khalid Almubarak、Samuel Albanie、Zaid Alyafeai、Albert Webson、Edward Raff 和 Colin Raffel。《通过多任务微调实现跨语言泛化》，2022年

[27] Alec Radford、Karthik Narasimhan、Tim Salimans、Ilya Sutskever 等。《通过生成式预训练提升语言理解》。2018年。[28] Katikapalli Subramanyam Kalyan、Ajit Rajasekharan 和 Sivanesan Sangeetha。《Ammus：自然语言处理领域基于Transformer的预训练模型综述》。*arXiv*预印本 *arXiv:2108.05542*，2021年。[29] Mike Lewis、Yinhan Liu、Naman Goyal、Marjan Ghazvininejad、Abdelrahman Mohamed、Omer Levy、Ves Stoyanov 和 Luke Zettlemoyer。《BART：用于自然语言生成、翻译和理解的去噪序列到序列预训练》。*arXiv*预印本 *arXiv:1910.13461*，2019年。[30] Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li 和 Peter J Liu。《探索统一文本到文本Transformer的迁移学习极限》。《机器学习研究杂志》，21(1):5485–5551，2020年。[31] Teven Le Scao、Angela Fan、Christopher Akiki、Ellie Pavlick、Suzana Ilić、Daniel Hesslow、Roman Castagné、Alexandra Sasha Luccioni、François Yvon、Matthias Gallé 等。《BLOOM：拥有1760亿参数的开放访问多语言语言模型》。*arXiv*预印本 *arXiv:2211.05100*，2022年。[32] Aakanksha Chowdhery、Sharan Narang、Jacob Devlin、Maarten Bosma、Gaurav Mishra、Adam Roberts、Paul Barham、Hyung Won Chung、Charles Sutton、Sebastian Gehrmann 等。《PaLM：利用Pathways扩展语言建模》。*arXiv*预印本 *arXiv:2204.02311*，2022年。[33] Susan Zhang、Stephen Roller、Naman Goyal、Mikel Artetxe、Moya Chen、Shuohui Chen、Christopher Dewan、Mona Diab、Xian Li、Xi Victoria Lin 等。《OPT：开放式预训练Transformer语言模型》。*arXiv*预印本 *arXiv:2205.01068*，2022年。[34] Daniel M Ziegler、Nisan Stiennon、Jeffrey Wu、Tom B Brown、Alec Radford、Dario Amodei、Paul Christiano 和 Geoffrey Irving。《基于人类偏好对语言模型进行微调》。*arXiv*预印本 *arXiv:1909.08593*，2019年。[35] Amelia Glaese、Nat McAleese、Maja Trębacz、John Aslanides、Vlad Firoiu、Timo Ewalds、Maribeth Rauh、Laura Weidinger、Martin Chadwick、Phoebe Thacker 等。《通过有针对性的人类判断提升对话智能体的对齐》。*arXiv*预印本 *arXiv:2209.14375*，2022年。[36] Long Ouyang、Jeffrey Wu、Xu Jiang、Diogo Almeida、Carroll Wainwright、Pamela Mishkin、Chong Zhang、Sandhini Agarwal、Katarina Slama、Alex Ray 等。《通过人类反馈训练语言模型遵循指令》。《神经信息处理系统进展》，35:27730–27744，2022年。[37] OpenAI。《ChatGPT简介》，2023年。访问日期：2023年3月31日。[38] Yuntao Bai、Andy Jones、Kamal Ndousse、Amanda Askell、Anna Chen、Nova DasSarma、Dawn Drain、Stanislav Fort、Deep Ganguli、Tom Henighan 等。《通过基于人类反馈的强化学习训练有帮助且无害的助手》。*arXiv*预印本 *arXiv:2204.05862*，2022年。[39] Jules White、Quchen Fu、Sam Hays、Michael Sandborn、Carlos Olea、Henry Gilbert、Ashraf Elnashar、Jesse Spencer-Smith 和 Douglas C Schmidt。《用于借助ChatGPT改进提示工程的提示模式目录》。*arXiv*预印本 *arXiv:2302.11382*，2023年。

[40] Tianyu Gao、Adam Fisch 和 Danqi Chen。让预训练语言模型成为更好的少样本学习器。*arXiv* 预印本 *arXiv:2012.15723*，2020年。[41] Niall Taylor、Yi Zhang、Dan Joyce、Alejo Nevado-Holgado 和 Andrey Kormilitzin。利用冻结语言模型进行临床提示学习。*arXiv* 预印本 *arXiv:2205.05535*，2022年。[42] Jay Burmeister、Zhe Chen、Indrin J Chetty、Sonja Dieterich、Anthony Doemer、Michael M Dominello、Rebecca M Howell、Patrick McDermott、Adrian Nalichowski、Joann Prisciandaro、Tim Ritter、Chadd Smith、Eric Schreiber、Timothy Shafman、Steven Sutlief 和 Ying Xiao。美国放射肿瘤学会面向放射肿瘤学住院医师的2015年核心物理学课程。国际放射肿瘤学、生物学与物理学杂志，95(4):1298—1303，2016年7月。[43] Noah Shinn、Beck Labash 和 Ashwin Gopinath。Reflexion：一种具备动态记忆和自我反思能力的自主智能体，2023年。[44] Jason Wei、Yi Tay、Rishi Bommasani、Colin Raffel、Barret Zoph、Sebastian Borgeaud、Dani Yogatama、Maarten Bosma、Denny Zhou、Donald Metzler、Ed H. Chi、Tatsunori Hashimoto、Oriol Vinyals、Percy Liang、Jeff Dean 和 William Fedus。大语言模型的涌现能力，2022年。[45] Wenxiong Liao、Zhengliang Liu、Haixing Dai、Zihao Wu、Yiyang Zhang、Xiaoke Huang、Yuzhong Chen、Xi Jiang、Dajiang Zhu、Tianming Liu 等。用于少样本文本分类的掩码引导的BERT。*arXiv* 预印本 *arXiv:2302.10447*，2023年。[46] Homgmin Cai、Wenxiong Liao、Zhengliang Liu、Xiaoke Huang、Yiyang Zhang、Siqi Ding、Sheng Li、Quanzheng Li、Tianming Liu 和 Xiang Li。基于远程监督迭代训练的由粗到细的知识图谱领域适应。*arXiv* 预印本 *arXiv:2211.02849*，2022年。[47] Xiaoyan Cai、Sen Liu、Junwei Han、Libin Yang、Zhenguo Liu 和 Tianming Liu。ChestXrayBERT：一种用于胸部放射学报告摘要生成的预训练语言模型。*IEEE* 多媒体汇刊，2021年。[48] Sheng Wang、Zihao Zhao、Xi Ouyang、Qian Wang 和 Dinggang Shen。ChatCAD：使用大语言模型对医学图像进行交互式计算机辅助诊断。*arXiv* 预印本 *arXiv:2302.07257*，2023年。

附录

A 先解释后回答的 ChatGPT (GPT-4) 响应

为了展示“先解释再回答”提示在ChatGPT（GPT-4）中的工作方式，我们选取了10个问题进行演示，首先从 61st 附录B节中100个问题里的第B题开始。

61. 1居里 (Ci) 等于 3.7×10^{10} 贝可勒尔 (Bq)。因此，10 Ci 的 ^{192}Ir 相当于37×10^{10} Bq，即 370 吉贝可勒尔。答案：C。其衰变率为 37 GBq。62. 有效半衰期 (T_{eff}) 的计算公式为 $\frac{1}{T_{\text{eff}}} = \frac{1}{T_{\text{phys}}} + \frac{1}{T_{\text{bio}}}$，其中 T_{phys} 为物理半衰期，T_{bio} 为生物半衰期。求解 T_{bio} 可得 $T_{\text{bio}} = \left(\frac{1}{T_{\text{eff}}} - \frac{1}{T_{\text{phys}}}\right)^{-1} = \left(\frac{1}{2} - \frac{1}{6}\right)^{-1} = 3$ 小时。答案：D。~~3 小时~~ 63. 初始剂量率 (D_0) 可用公式 $D_0 = \frac{D_{\text{total}}}{(T \times \ln(2))}$ 计算，其中 D_{total} 为总剂量，T 为该同位素的半衰期。~~对于 ^{125}I，~~ 其半衰期约为 ~~60~~ 天。 $D_0 = \frac{^{120}\text{戈瑞 (60 天}}{\times \ln(2))} \approx 5.9$ 厘戈瑞/小时。答案：B。5.9 厘戈瑞/小时 64. 根据 AAPM TG-43，在所列同位素中，125I 的各向异性校正最大。答案：D。 ^{125}I 65. AAPM TG-43 确实考虑了源中的吸收和散射，因此这并非该协议的局限。~~答案：~~ A。未考虑源中的吸收和散射。66. 对于 ^{192}Ir 源，当 r 小于 5 cm 时，径向剂量函数 $g(r)$ 近似恒定，因为衰减被增加的散射所~~补偿。答案：~~ C。衰减被增加的散射所补偿。67. 距圆柱表面 5 mm（0.5 cm）处的剂量率为 700 厘戈瑞。可用平方反比定律估算圆柱表面处的剂量（r $= 1.5$ cm）：剂量 $_{\text{surface}} =$剂量$_{5\text{mm}} \times \left(\frac{^{5}\text{mm}}{15\text{ mm}}\right)^{2} \approx 700$厘戈瑞$\times \left(\frac{1}{9}\right) = 77.8$厘戈瑞。答案：D。500 厘戈瑞（最接近计算值）68. 在高剂量率阴道治疗中，为了使距圆柱表面 5 mm 处的剂量均匀，圆柱两端的放射源驻留时间较长。答案：B。两端的驻留时间较长。69. 使用 ^{103}Pd 而非 ^{125}I 时，永久性前列腺植入粒子的处方剂量较低，因为 ^{103}Pd 的半衰期较短。答案：A。 ^{103}Pd 的半衰期较短。70. 高剂量率每日质量保证不进行放射源校准，因为校准通常频率较低（例如每月或每季度一次）。答案：D。放射源校准。

B 辐射肿瘤物理学 100 题选择题考试

正确答案 =C, A, B, A, B, D, B, A,C,A, D,A,D, C, A,A, E,C, C, D, D,C, A,A, D,C, B, D, C, A, D, D, A, B, C, A, D, C, A, C, D, C, A, C, D, B, A, B, C, B, B, D, C, A, C, A, C, B, C, C, A, D, B, D, A, C, B, B, A, D, C, D, C, C, A, B, D, C, A, B, C, D, C, A, A, B, C, A, D, C, B, C, A, C, A, A, C, D, C, B

1. 下列哪种粒子无法被电场加速?

(a) 质子 (b)
电子 (c) 中子
(d) 正电子

2. 其质量　　碳-12（^{12}C）比6个质子和6个中子的质量总和　　s.

(a) 略小 (b) 略大 (c) 等于 (d) 以
上情况都有可能。

3. 钨的K、L和M壳层的结合能分别为70、11和2 keV。一个50 keV的光子通过光电效应与钨相互作用。发射出的光电子可能具有哪些能量?

(a) 从2 keV到70 keV的连续能谱 (b) 39 keV和
48 keV (c) 9 keV、59 keV和68 keV (d) 9 keV

4. 以下哪　　项陈述正确描述了X射线管的X射线能谱?

(a) 最大光子能量由kVp决定。(b) 最大光子能量由靶材料决定。(c) 特征光
子能量由管电流决定。(d) 特征光子能量由滤过决定。

5. X射线束的半值层（HVL）为2毫米铝。将X射线照射量降低50%所需的铝的最小厚度是多少?

(a) 1毫米
(b) 2毫米
(c) 3毫米
(d) 4毫米

6. 以下哪种部件为直线加速器提供高功率射频场，以加速电子?

(a) 闸流管 (b)
靶 (c) 电子枪
(d) 磁控管

7. 什么是 多叶准直器 (MLC) 的“榫舌与凹槽”设计有什么作用?

(a) 减少叶片之间的摩擦。(b) 减少叶片
之间的漏射剂量。(c) 提高叶片的移动速
度。(d) 增大治疗照射野尺寸。

8. 使用直线加速器以电子模式治疗患者时，以下哪种硬件是不需要的?

(a) 靶 (b) 电子枪 (
c) 监测电离室 (d) 散
射箔

9. 与均整束流相比，无均整器束流 。 _____

(a) 平均能量更高。(b) 穿透能
力更强。(c) 剂量率更高。(
d) 治疗时间延长。

10. 对于以下哪种效应，质量衰减系数对介质原子序数和入射光子能量的依赖性最强?

_____

(a) 光电效应 (b) 康普
顿散射 (c) 电子对产生
(d) 瑞利散射

11. 以下哪项不是光子与介质之间的相互作用?

(a) 光电效应 (b) 康普
顿散射 (c) 电子对产生
(d) 轫致辐射

12. 在康普顿相互作用中，当散射光子的能量最小时，发射出的康普顿电子相对于初始光子方向的角度是多少?

(a) 0°
(b) 45°
(c) 90°
(d) 180°

13. 质量衰减系数 μ/ρ、质量能量转移系数（μ_{tr}/ρ）与质能吸收系数（μ_{en}/ρ）之间有什么关系?

(a) $\mu_{\mathrm{tr}}/\rho \geq \mu_{\mathrm{en}}/\rho \geq \mu/\rho$
(b) $\mu_{\mathrm{en}}/\rho \geq \mu_{\mathrm{tr}}/\rho \geq \mu/\rho$
(c) $\mu/\rho \geq \mu_{\mathrm{en}}/\rho \geq \mu_{\mathrm{tr}}/\rho$
(d) $\mu/\rho \geq \mu_{\mathrm{tr}}/\rho \geq \mu_{\mathrm{en}}/\rho$

14. 在水模体中 10 cm 深处，15 MV光子束的比释动能与吸收剂量相比如何?

(a) 比释动能大于吸收剂量。(b) 比释动能等于吸收剂量。(c) 比释动能小于吸收剂量。(d) 无法确定比释动能与吸收剂量之间的关系。

15. 以下哪种射线不会在百分深度剂量曲线上产生建成效应?

(a) 来自100 kVp X射线管的X射线。
(b) 来自 ^{60}Co的伽马射线。(c) 来自直线加速器的6兆伏X射线。(d) 来自直线加速器的18兆伏X射线。

16. 以下哪种设备最适合测量横向射束剖面?

(a) 二极管 (b) 法默电离室 (c) 平行板电离室 (d) 盖革-穆勒计数器

17. 四张已曝光胶片的光密度均为1.25。如果将四张胶片叠放在一起，透过它们的光线比例是多少?

(a) 20%
(b) 1%
(c) 0.1%

(d) 0.01%
(e) 0.001%

18. 以下哪种设备不适用于体内剂量测量? ?

(a) 二极管 (b) 热释光剂量计 (TLDs) (c) 电离室 (d) 光释光剂量计 (OSLDs)

19. 根据 TG-51，将电离室中的测量电荷转换为产生电荷时，以下哪种修正因子不需要?

(a) 温度和气压修正 (b) 复合修正 (c) 散射修正 (d) 极性修正

20. 对于在标准温度和气压下校准的 Farmer电离室，当当地温度为 23°C、气压为 720 mmHg 时，PTP 因子是多少?

（a）0.97 （b）0.99 （c）1.03 （d）1.06

21. TG-51 中如何定义射束质量?

(a) 标称射束能量 (b) 十分之一值层 (c) 平均射束能量 (d) %dd(10)x

22. 根据 TG-51，直线加速器应在水箱中进行剂量校准。 _____

(a) 每天 (b) 每月
(c) 每年 (d) 每两年

23. 为什么在大于最大剂量深度的深度处，组织最大比总是大于百分深度剂量?

（a）百分深度剂量包括射束衰减和射束发散。（b）百分深度剂量包括射束衰减，但不包括射束发散。（c）组织最大比包括射束衰减和射束发散。

（d）组织最大比不包括射束衰减，但包括射束发散。

24. 以下哪项变化会导致几何半影增大？

(a) 增大 SSD。(b) 增大源准直器距。(c) 减小源尺寸。(d) 将患者移至更靠近准直器的位置。

25. 在 SSD=100 cm、深度=1.5 cm、照射野尺寸为 10 cm x10 cm 的条件下，6 MV射束的输出为 1 cGy/机器跳数。在 SSD=98.5 cm、深度=1.5 cm、照射野尺寸为 10 cm x10 cm 的条件下，输出是多少？

(a) 0.985 cGy/机器跳数 (b) 0.97 cGy/机器跳数 (c) 1.015 cGy/机器跳数 (d) 1.03 cGy/机器跳数

26. 对于单野源轴距治疗，最大剂量深度处的输出为 1.02 cGy，深度 10 cm 处的组织最大比为 0.7。在深度 10 cm 处给予 100 cGy 需要多少机器跳数？

(a) 100
(b) 102
(c) 140
(d) 160

27. 两个治疗照射野之间相隔 100° 时，什么楔形角能产生最均匀的剂量分布？

(a) 30°
(b) 40°
(c) 50°
(d) 60°

28. 患者的脊柱采用两个相邻照射野进行治疗，源皮距为=100 cm。头脚方向的照射野长度（对称野）分别为 20 cm 和 24 cm。在 10 cm 深度处使两个脊柱照射野匹配时，皮肤表面的间隙应为多少？

(a) 1.0厘米 (b) 1.5 厘米 (c) 2.0 厘米 (d) 2.2 厘米

29. 对于采用对称野的颅脊治疗，上部脊柱照射野的长度为 25 cm，侧向颅部照射野的照射野尺寸为 20 cm x 20 cm。颅部照射野的准直器旋转角度应为多少，才能与脊柱照射野的发散对齐？

(a) 3°
(b) 5°
(c) 7°
(d) 9°

30. 剂量体积直方图 (DVH) 不包含以下哪项信息?

(a) 热点的位置 (b) 平均剂量 (c) 最小剂量 (d) 最大剂量

31. DVH 曲线上的心脏剂量 D5%=20戈瑞表示什么?

(a) 5% 的心脏恰好接受 20 戈瑞。(b) 5% 的心脏接受的剂量低于 20 戈瑞。(c) 95% 的心脏接受的剂量高于 20 戈瑞。(d) 95% 的心脏接受的剂量低于 20 戈瑞。

32. 剂量师在制定治疗计划时有时会更改以下哪项参数?

(a) 每次治疗分次的剂量 (b) 总治疗剂量 (c) 治疗分次数量 (d) 射线束数量

33. 在电子束放射治疗中，哪种电子能量在表面的PDD最低?

(a) 6MeV
(b) 9 MeV
(c) 12 MeV (d) 16MeV (e) 20MeV

34. 电子束的PDD在实际射程之后有一个低剂量尾部。哪种物理过程与这一尾部的产生有关?

(a) 康普顿散射 (b) 轫致辐射
(c) 特征X射线 (d) 电子对产生

35. 以下哪种治疗适合采用源皮距设置?

(a) 盆腔治疗的四野盒式照射 (b) 多野调强放射治疗 (c) 乳腺癌胸壁电子线加量照射 (d) 采用两弧的容积旋转调强放射治疗

36. 以下关于电子线有效源皮距的说法，哪一项是正确的?

(a) 它会随电子能量增加而增大。(b) 它等于X射线源到等中心的距离。(c) 它大于X射线源到等中心的距离。(d) 它与电子能量无关。

37. 6 MeV电子束在肺（0.25 g/cm^3）中的射程是多少?

(a) 3厘米
(b) 6 厘米
(c) 9 厘米
(d) 12厘米

38. 什么是调强放射治疗中的逆向计划?

(a) 选择最佳射束角度。(b) 选择合适的射束修饰器，例如楔形滤过板和补偿器。(c) 根据预期的剂量约束优化射束注量。(d) 添加射束挡块以保护危及器官。

39. 叶片序列算法在调强放射治疗计划中的作用是什么?

(a) 它将优化后的注量转换为多叶准直器可实现的注量。(b) 它将代价函数转换为优化后的注量。(c) 它将可实现注量转换为剂量。(d) 它将预期的约束转换为代价函数。

40. 调强放射治疗计划中可实现的剂量梯度约为。_____

(a) 0.1%/mm (b) 1%/mm (c) 10%/mm (d) 50%/mm

41. 以下哪种设备不适用于调强放射治疗患者质量保证?

(a) 电子射野成像装置 (b) 胶片和电离室

(c) 二极管阵列 (d) 单电
离室

42. 以下哪个参数通常不是剂量师在调强放射治疗计划中设定的? ?

(a) 射束能量 (b) 射束角度 (c) 射束权
重 (d) 上限和下限剂量约束

43. 对于亨斯菲尔德单位为-10的材料，该材料与水的线性衰减系数之比是多少?

(a) 1.01
(b) 1.1
(c) 0.99
(d) 0.9

44. 以下哪种方法可以提高计算机断层扫描的空间分辨率?

(a) 提高管电压。(b) 增大视
野。(c) 缩小视野。(d) 降低
管电流。

45. 为什么磁共振图像不能替代计算机断层扫描，用于放射治疗中的精确剂量计算?

(a) 磁共振图像的软组织对比度低于计算机断层扫描。(b) 磁共振图像
的空间分辨率低于计算机断层扫描。(c) 磁共振图像对患者造成的电
离辐射剂量高于计算机断层扫描。(d) 磁共振图像不包含电子密度信
息。

46. 超声在软组织中的传播速度为 1500 m/s。发出脉冲后，接收到信号的时间为 0.1 ms。解剖界面的深度是多少?

(a) 5厘米
(b) 7.5 厘
米 (c) 10
厘米 (d)
15厘米

47. 以下哪种成像方式使用非电离辐射?

(a) 磁共振成
像 (b) 正电
子发射断层
成像 (c) 计算
机断层扫描

(d) 射线照片

48. 以下　哪项陈述正确描述了正电子发射断层成像图像的空间分辨率?

（a）正电子发射断层成像图像的空间分辨率优于计算机断层扫描和磁共振成像。（b）发射正电子的平均能量降低时，正电子发射断层成像图像的空间分辨率会提高。（c）环形探测器的数量减少时，正电子发射断层成像图像的空间分辨率会提高。（d）正电子发射断层成像图像的空间分辨率优于 1 毫米。

49. 以下哪项信息未存储在 DICOM 放射治疗计划文件中?

（a）患者姓名（b）治疗角度（c）计算机断层扫描仪参数，例如 kVp、毫安秒（d）处方剂量

50. 一组四维 CT 有 10 个时相，每个时相有 200 个切片。每个切片有 512×512 个像素，每个像素占用 2 字节。以 2 GB/秒的传输速度通过互联网传输这组四维 CT 需要多长时间?

(a) 0.2 秒
(b) 0.5 秒
(c) 1.0 秒
(d) 2.0 秒

51. 与　诊断 CT 相比，以下哪项关于 CT 模拟定位机的说法不正确?

(a) CT 模拟定位机的孔径通常比诊断 CT 更大。(b) CT 模拟定位机的图像空间分辨率和对比度优于诊断 CT。(c) CT 模拟定位机需要额外的激光来定位治疗等中心。(d) CT 模拟定位机需要全身范围内准确的几何参数和亨斯菲尔德单位。

52. 以下哪种技术不会降低计划靶区（PTV）的摆位边界?

(a) 将每周一次的图像引导放射治疗改为每日一次。(b) 将摆位方式从激光改为图像引导放射治疗。(c) 在靶区添加标记物。(d) 将三维适形放射治疗计划改为调强放射治疗计划。

53. 对于头颈部癌症治疗，哪种图像引导放射治疗方法能提供最佳软组织对比度?

(a) 兆伏级射野影像 (b) 千伏级正交平面影像

(c) 千伏级锥形束
CT (d) 表面成像

54. 以下哪种图像引导放射治疗方法不用于调强放射治疗?

(a) 兆伏级射野影像。(b) 千伏级
正交平面影像。(c) 千伏级锥形
束CT。(d) 轨道式CT。

55. 哪种图像引导放射治疗方法产生的剂量最低?

(a) 兆伏级射野影像 (b) 兆伏级
锥形束计算机断层扫描 (c) 千伏
级正交平面影像 (d) 千伏级锥形
束CT

56. 前列腺运动无法通过 进行追踪。 _____

(a) 红外表面成像。(b) 植入式射频信标。(c) 植入式定
位标记和透视成像。(d) 磁共振直线加速器中的磁共振
图像。

57. 以下哪一项 不是治疗肺肿瘤时采用自由呼吸门控的优势 ?

(a) 减小内靶区体积。(b) 降低危及器官的
剂量。(c) 缩短治疗时间。(d) 患者在自由呼
吸时感到舒适。

58. 以下哪种情况不需要进行四维 CT?

(a) 膈肌附近的肿瘤 (b) 脑部
肿瘤 (c) 肺部肿瘤 (d) 肝脏肿
瘤

59. 不是放射治疗中管理呼吸运动的方法。

(a) 深吸气屏气（DIBH）。(b) 实时肿瘤
追踪。(c) 使用锥形束计算机断层扫描的
图像引导放射治疗（IGRT）。(d) 自由呼
吸呼吸门控。

60. 某放射性核素的半衰期为 69 天。一天内它会衰变多少?

(a) 0.5%
(b) 0.7%
(c) 1.0%
(d) 1.5%

61. 对于活度为 10 Ci 的高剂量率源 192Ir，以下哪项陈述正确?

(a) 其衰变率相当于约 10 克 226Ra。(b) 其衰变率相当于约 1 克 226Ra。(c) 其衰变率为 37 吉贝可勒尔。(d) 其衰变率为 3.7 吉贝可勒尔。

62. 某非密封放射性同位素的物理半衰期和有效半衰期分别为 6 小时和 2 小时。其生物半衰期是多少?

(a) 12 小时
(b) 8 小时
(c) 4 小时
(d) 3 小时

63. 什么是 前列腺125I粒子植入的初始剂量率，才能达到120戈瑞的总剂量 ?

(a) 4.9 厘戈瑞/小时 (b) 5.9厘戈瑞/小时 (c) 7.0厘戈瑞/小时 (d) 7.9厘戈瑞/小时

64. 根据 AAPM TG-43，哪种具有最大的各向异性校正 .

(a) 192Ir
(b) 137Cs (c) 226Ra
(d) 125I

65. 以下哪项不是 AAPM TG-43 的局限性?

(a) 不考虑源中的吸收和散射。(b) 将施源器视为水。(c) 通常忽略源在移动过程中的剂量输送。(d) 将患者视为水。

66. 为什么对于 192Ir 源，当 r 小于 5 cm 时，径向剂量函数 g(r) 近似为常数?

(a) 前 5 cm 内没有衰减。(b) 衰减已计入几何因子。(c) 衰减由增加的散射补偿。(d) 衰减已计入各向异性因子。

67. 一个直径为 3 cm 的高剂量率阴道圆柱体病例，计划向距离圆柱表面 5 mm 处的阴道组织给予 700 cGy。圆柱表面的剂量约为多少?

(a) 1244 厘戈瑞 (b) 933 厘戈瑞 (c) 700 厘戈瑞 (d) 500 厘戈瑞

68. 在高剂量率阴道治疗中，若要使距离圆柱表面 5 mm 处的剂量均匀，放射源驻留时间应如何设置?

_____

(a) 所有驻留点的时间相同。(b) 两端的时间更长。(c) 中间的时间更长。(d) 上端的时间比下端更长。

69. 使用 103Pd 而非 125I 作为永久性前列腺植入粒子时，为什么处方剂量较低?

(a) 103Pd 的半衰期较短。(b) 103Pd 发射光子的能量较高。(c) 103Pd 的径向剂量函数 g(r) 较大。(d) 103Pd 的各向异性函数较大。

70. 不会在高剂量率每日质量保证中进行。

(a) 放射源位置准确性检查 (b) 巡测仪功能检查 (c) 辐射监测仪功能检查 (d) 放射源校准

71. _____ 是　　在手术室中寻找丢失的放射性核素粒子最合适的设备。

(a) 井型电离室 (b) Farmer电离室 (c) 盖革-穆勒计数器 (d) 二极管

72. 以下哪项测试不是高剂量率近距离治疗的典型流程?

(a) 使用新源时对源进行校准。(b) 每隔 3–4 个月更换一次源。(c) 仅在安排患者治疗的日期进行每日质量保证。(d) 仅在治疗前对患者进行调查。

73. 以下哪项关于全身照射治疗的说法是正确的?

(a) 使用侧向对穿射束时始终需要肺挡块。(b) 使用无平坦滤过器模式以提高剂量率。(c) 延长源皮距可提高剂量均匀性。(d) 该治疗比常规患者的治疗更快。

74. 直线加速器的最大照射野尺寸为 40 cm x 40 cm，某全身照射治疗患者身高为 180 cm。准直器角度为 0° 时，该治疗所需的最小源皮距是多少?

(a) 250厘米
(b) 350 厘
米 (c) 450
厘米 (d)
550厘米

75. 在全皮肤电子束治疗（TSET）中，为什么机架角度约为20o，使中心轴位于患者头部上方、脚部下方?

(a) 尽量减少X射线污染造成的剂量。(b) 尽量增加X射线污染造成的剂量。(c) 增加源皮距。(d) 缩短治疗时间。

76. 哪种方式不适用于术中放射治疗（IORT）？

(a) 兆电子伏特电子 (
b) 兆伏级光子 (c) 高剂
量率近距离治疗 (d) 千
伏级X射线

77. 以下哪项陈述正确描述了治疗质子束的侧向半影?

(a) 它们始终大于直线加速器中兆伏级X射线的半影。(b) 它们始终小于直线加速器中兆伏级X射线的半影。(c) 它们不随深度变化。(d) 它们随深度增加而增大。

78. 专用于质子笔形束扫描的喷嘴不需要。

(a) 剂量监测电离室 (b) 扫描磁铁 (c) 射程调制轮 (d) 束斑位置监测器

79. 关于质子治疗中的射程涨落，以下哪项陈述正确?

(a) 它是由能量损失的统计不确定性造成的。(b) 它会导致约 3.5 (c) 可通过鲁棒优化将其降至最低。(d) 它与射程不确定性相同。

80. 为什么质子设备需要大型机架来弯转质子，而直线加速器却不需要?

(a) 能量为 250 MeV 的质子比能量为 6-20 MeV 的电子速度更快。(b) 质子的质量约为电子质量的 1800 倍。(c) 质子带正电，而电子带负电。(d) 质子不会通过轫致辐射产生 X 射线。

81. 关于调强质子治疗（IMPT），以下哪项陈述不正确?

(a) 通过稳健性优化，IMPT计划可以具有稳健性。(b) 与双散射相比，IMPT的适形性更好。(c) 与双散射相比，IMPT对运动不那么敏感。(d) 铅笔束扫描IMPT所需的硬件比双散射少。

82. 关于质子的相对生物效应（RBE），以下哪项陈述正确?

(a) 在所有深度上始终为常数1.1。(b) 在浅层深度处最高。
(c) 在展宽布拉格峰中心处最高。(d) 在远端跌落深度处最高。

83. 哪种设备不适合测量立体定向放射外科（SRS）中小野的横向剖面?

(a) 二极管 (b) 针形电离室 (c) Farmer电离室 (d) 金刚石探测器

84. 在立体定向放射外科技术中使用三级准直器的目的是什么?

(a) 减小半影。(b) 缩短治疗时间。(c) 减少碰撞。(
d) 将多个等中心合并为一个等中心。

85. 关于立体定向放射外科，以下哪项陈述不正确?

(a) 无均整器（FFF）不能用于立体定向放射外科技术。(b) 必须进行Winston-Lutz测试，以验证辐射等中心与机械等中心是否重合。(c) 立体定向放射外科照射野可能不满足横向带电粒子平衡（CPE）。(d) 有框架系统和无框架系统均可用于立体定向放射外科。

86. 在进行计划 时，对于立体定向放射外科/立体定向体部放射治疗，以下哪项通常不是优先考虑的事项 ?

(a) 靶外剂量快速跌落。(b) 靶内剂量均匀。
(c) 适形指数接近1。(d) 梯度指数尽可能小。

87. 与常规分次治疗相比，立体定向放射外科技术不需要。

(a) 更小的CT层厚 (b) 更小的剂量计算网格 (c) 更高的输出精度 (d) 更高的射束等中心与机架/治疗床等中心重合精度。

88. 哪种立体定向放射外科照射技术可以使用单等中心治疗计划同时治疗两个脑转移瘤?

(a) 使用动态多叶准直器的弧形计划。(b) 使用锥形准直器的弧形计划。(c) 动态适形弧。(d) 使用静态多叶准直器的弧形计划。

89. 是一种用于测试多叶准直器叶片位置的测试。

(a) 用于机架等中心性的星形射野测试。(b) Winston-Lutz测试。(c) 用于多叶准直器等中心性的星形射野测试。(d) 栅栏测试。

90. 根据 TG-142，哪项测试不需要每天进行?

(a) 机房门联锁

(b) 输出 (c) 光野与射野重合 (d) 激光精度

91. 以下哪项参数对于用于图像引导放射治疗（IGRT）的锥形束CT（CBCT）最重要?

(a) 亨氏单位准确度 (b) CBCT等中心与辐射束等中心之间的一致性 (c) 空间分辨率 (d) 视野

92. 根据TG-142，每月输出稳定性的容差是多少?

(a) 0.5%
(b) 1%
(c) 2%
(d) 3%

93. TG-142未涉及哪项等中心的质量保证测试?

(a) CT 模拟定位机
(b) 多叶准直器 (c)
CBCT (d) 平板兆伏级
成像仪

94. 次级屏障屏蔽计算中不需要以下哪项参数?

(a) 距离。(b) 居留因
子。(c) 使用因子。
(d) 工作负荷。

95. 直线加速器治疗室旁护士站在屏蔽设计中采用的居留因子是多少?

(a) 1 (
b) 1/5
(c) 1/20
(d)
1/40

96. 直线加速器治疗室门的材料由内向外的正确排列顺序是_____。
_____

(a) 钢、含硼聚乙烯、铅

(b) 铅、含硼聚乙烯、钢 (c) 含硼聚乙烯、铅、钢 (d) 含硼聚乙烯、钢、铅

97. 需要多少个十分之一值层才能将照射率降至原值的1/5?

(a) 0.3
(b) 0.5
(c) 0.7
(d) 1.0

98. 具有最高的辐射权重因子。

(a) 6兆伏X射线 (b) 250兆电子伏质子 (c) 6兆电子伏电子 (d) 100千电子伏中子

99. 对于5次分割前列腺立体定向体部放射治疗，以下哪种错误不被视为放射医疗事件?

(a) 总剂量偏差为 25%。(b) 治疗了错误的患者。(c) 单次分割剂量偏差为 45%。(d) 剂量被施加到了肝脏。

100. 肺癌CT模拟定位的失效模式与影响分析流程图不需要以下哪项因素?

(a) 记录CT模拟定位中使用的固定装置。(b) 验证CT模拟定位设备的孔径。(c) 在每日质量保证中验证CT模拟定位设备的亨氏单位准确度。(d) 在CT模拟定位前确认患者身份正确。