/burl@stx null def /BU.S /burl@stx null def def /BU.SS currentpoint /burl@lly exch def /burl@llx exch def burl@stx null ne burl@en

$=1$

GPT-5 在脑肿瘤 MRI 推理中的表现

Mojtaba Safari[1] **Shansong Wang**[1] **Mingzhe Hu**[1] **Zach Eidex**[1] **Qiang Li**[1] **Xiaofeng Yang**[1]

[1]Department of Radiation Oncology, Winship Cancer Institute, Emory University School of Medicine
Corresponding author: xiaofeng.yang@emory.edu

Abstract

在磁共振成像（MRI）上准确区分脑肿瘤的类型对于神经肿瘤学的治疗计划至关重要。近年来，大型语言模型（LLM）的进步已使视觉问答（VQA）方法能够将图像解释与自然语言推理相结合。在本研究中，我们在一个从 3 个脑肿瘤分割（BraTS）数据集（胶质母细胞瘤（GLI）、脑膜瘤（MEN）和脑转移瘤（MET））衍生的精选脑肿瘤 VQA 基准上评估了 GPT-4o、GPT-5-nano、GPT-5-mini 和 GPT-5。每个病例包括多序列 MRI 三平面图块和转换为标准化 VQA 项目的结构化临床特征。在视觉和推理任务的零样本思路链设定中评估模型的准确性。结果表明，GPT-5-mini 达到了最高的宏平均准确率（44.19 %），其次是 GPT-5（43.71 %）、GPT-4o（41.49 %）和 GPT-5-nano（35.85 %）。表现因肿瘤亚型而异，没有单一的模型在所有队列中占据主导地位。这些结果表明，GPT-5 系列模型在结构化神经肿瘤学 VQA 任务中可以达到中等准确率，但尚未达到可接受的临床使用水平。

关键词：脑肿瘤 MRI，大型语言模型，胶质母细胞瘤，GPT-5，脑膜瘤，转移瘤

脑肿瘤，包括胶质母细胞瘤 (GBM)、脑膜瘤和脑转移瘤，代表了一组多样的颅内肿瘤，这些肿瘤具有不同的放射学外观、生物学行为及预后意义 [1] 。准确和及时的诊断是至关重要的，因为对这些肿瘤类型的早期鉴别能够显著影响手术计划、辅助治疗和患者的治疗效果 [2] 。由于其优越的软组织对比度和多参数功能，磁共振成像 (MRI) 仍然是脑肿瘤评估的临床标准。然而，对 MRI 图像的解读需要专业的神经放射学知识，并且诊断的可变性可能因培训、经验和疲劳的差异而产生 [3, 4] 。因此，人工智能的整合，特别是大语言和多模态模型，进入诊断工作流程的趋势已经成为增强效率和标准化的一个有前途的途径。

医学影像中的视觉问答（VQA）范式旨在回答关于医学图像的临床相关自然语言问题，将视觉理解与医学推理相结合。在神经肿瘤学中，VQA 系统不仅需要识别放射学特征，如对比增强、坏死和肿瘤周围水肿，还需将这些发现进行情境化处理，以提供准确的定位、鉴别诊断及治疗影响。这使得脑肿瘤的 VQA 成为一个高复杂性任务，要求既有强大的视觉基础又有领域特定的推理能力。

最近在大型语言模型（LLMs）方面的进展加速了行业从狭隘训练的任务特定系统向框架的转变，在这些框架中，LLMs 作为中央推理引擎发挥作用 [5, 6] 。在临床应用中，相关信息通常来自不同的模态。例如，LLMs 中的临床文本和医学知识评估基准 [7] ，结构化表示和有根据的报告框架 [8] ，以及成像数据 [9, 10] 。GPT-4 展现了强大的普通医学推理能力，并在多个医学基准中优于之前的模型 [11, 12] 。然而，它在脑部肿瘤 MRI 方面的表现仍然受到限制，尤其是在细粒度定位和肿瘤类型区分方面。AutoRG-Brain [13] 的引入强调了像素级基础和脑部 MRI 报告生成的重要性，提供了使用 BraTS-like 数据集的结构化评估框架，将视觉特征明确链接到文本输出。在这些基础上，GPT-5 及其轻量化变体（GPT-5-mini，GPT-5-nano）在多模态推理方面引入了进一步改进 [14] [*] [†]，包括提高临床 VQA 准确性、推理一致性以及领域适应能力，超过了 GPT-4o。

在这项研究中，我们在从 BraTS 数据集中提取的脑肿瘤 VQA 基准测试上评估了 GPT-4o-2024-11-20、GPT-5、GPT-5-mini 和 GPT-5-nano。每个病例的多模态 MRI 序列，包括 T1 对比（T1c）、T1 加权（T1w）、T2 液体衰减反转恢复（T2-FLAIR）、T2 加权（T2w），都被处理以提取质心对齐的三平面马赛克。关联的

[*]https://H.Bgithub.H.Bcom/H.Bwangshansong1/H.BGPT-5-EvaluationH.B

[†]https://H.Bgithub.H.Bcom/H.BTsinghuaC3I/H.BMedXpertQAH.B

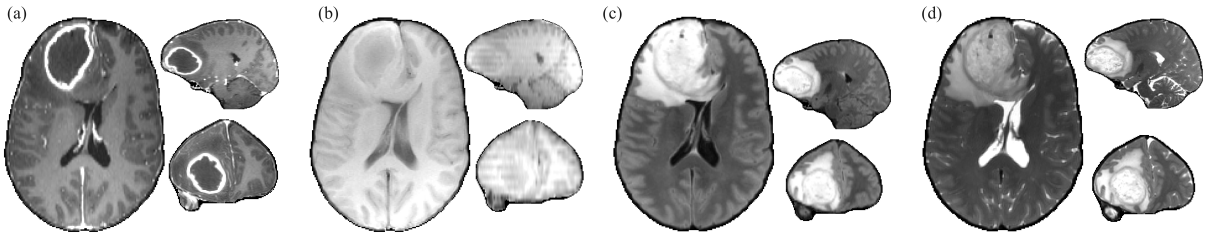


Figure 1: 一个 BraTS-GLI 案例的示例，包含选择的轴向、冠状面和矢状面切片，以展示 (a) T1c、(b) T1w、(c) T2-FLAIR 和 (d) T2w MRI 的肿瘤体积范围。给 GPT 模型提供了单个序列用于 VQA。

放射学发现被解析为结构化的临床概念，包括病变位置、增强模式、边界定义、水肿、中线移位、最大尺寸和脑室压缩，然后被转换为标准化的是/否以及包含干扰选项、答案关键和相关影像的多项选择题。这确保每个模型都在同样、基于临床的提示上得到评估，结合图像和文本输入，使性能可以在模型和问题类型之间进行直接比较。通过与 [14] 提示和评估协议对齐数据集构建，我们确保可重复性并隔离在多模态推理上模型特定的差异，包括胶质母细胞瘤、脑膜瘤和转移病例。

1 方法论

本研究使用了 BraTS 数据集，该数据集包含多参数 MRI 扫描，包括胶质母细胞瘤（GLI） [15]，脑膜瘤（MEN） [16]，以及脑转移瘤（MET） [17]。对于每个病例，四种标准 MRI 序列是可用的：T1 对比增强（T1c）、T1 加权（T1w）、T2 加权（T2w）和 T2 液体衰减反转恢复（T2-FLAIR）。数据集提供的肿瘤分割掩码用于定位病变，并且每个病例还配有生成的 AutoRG-Brain 风格或从等效临床报告来源获得的放射学风格文本发现。

在分析之前，所有 MRI 体积均被转换为标准的 RAS 方向，以确保受试者之间的空间一致性。肿瘤中心点直接从分割掩膜中计算得出。从每个体积中提取通过肿瘤中心点的轴向、矢状和冠状切片。使用基于百分位的截断（第 2 到第 98 百分位）对强度进行归一化，以抑制异常值并增强病灶的可视性。三个切片被连接形成一个三平面马赛克图像，为后续评估提供了一种紧凑但信息丰富的视觉表达。

每个病例的附带临床文本使用定制的解析管道处理，以提取肿瘤外观的结构化描述符。这个过程识别了主要病灶位置、增强模式、边界定义、肿瘤周围水肿的存在、中线移位的有无、以毫米为单位的最大病灶尺寸、跨越中线的涉及情况以及任何脑室结构的压迫。解析器结合了基于规则的术语匹配、同义词标准化以及用于数值的正则表达式，以确保报告中一致的提取。

然后将这些结构化特征按照 [14] 中描述的风格协议转换为 VQA 项。每个特征被映射到一个自然语言问题，旨在评估视觉理解或临床推理。例如，增强特征产生了是/否的问题，例如 “病变是否显示环形增强?”；位置特征被转换为具有合理干扰项的多项选择题；病变尺寸测量则被设计为数字多项选择题。每个 VQA 项目包括问题文本和标记的答案选项、正确答案、相关的三面拼接图和指定医学任务、身体系统和问题类型的元数据。

对生成的 VQA 数据集进行了四个大型多模态语言模型的评估：GPT-4o，GPT-5，GPT-5-mini 和 GPT-5-nano。每个模型接收到三向平面拼贴和附有答案选项的问题文本作为单一的多模态输入（见图 1）。模型预测的选择与参考答案进行比较，并计算准确率作为正确回答问题的比例。

1.1 提示设计

每个模型都接收一个包含单一序列的三平面镶嵌图像和相关问题文本的组合输入，其中包含标记的答案选项。为了避免来自任务特定调试的偏差，同时保留清晰、独立的临床措辞，提示中的额外说明被保持在最低限度。

在每次交互中，图像被提供作为主要视觉上下文，紧接着是文本问题。答案选项被明确地列举为 “答案选项：(A) …(B) …(C) …(D) …”，以标准化模型间的解析。使用 “让我们一步一步地思考” 明确提示模型以引出中间推理过程（思维链），记录在 助手 _ 基本原理，然后进行最终的强制选择（ASSISTANT_FINAL），由单个字母（A-D）组成。只有 最终助手 用于计算准确性。中间理由被记录用于定性分析但不评分。该设计确保了一致的评估，同时允许对推理行为的检查。

VQA 的提示设计模板如图 2 所示，具体示例如图 3 所示。每个问题的提示措辞直接取自特征提取步骤，无需重新表述或简化。例如，检测到的环形增强特征会生成问题 “病变显示环形增强吗?” 而不是一个改

Zero-shot + CoT for VQA

```
[
  {
    "role": "system",
    "content": "You are a helpful medical assistant."
  },
  {
    "role": "user",
    "content": [
      {
        "type": "text",
        "text": "Q: {QUESTION_TEXT}\nA: Let's think step by step."
      },
      {
        "type": "image_url",
        "image_url": { "url": "{IMAGE_URL_1}" }
      }
    ]
  },
  {
    "role": "assistant",
    "content": "{ASSISTANT_RATIONALE}"
  },
  {
    "role": "user",
    "content": "Therefore, among A through {END_LETTER}, the answer is"
  },
  {
    "role": "assistant",
    "content": "{ASSISTANT_FINAL}"
  }
]
```

Figure 2: 使用零样本链式思维引导的提示设计。 助手 _ 推理 是模型的中间推理； 助理 _ 最终 是用于评分的最终字母。

述的变体。位置问题保留了解析算法生成的解剖术语，干扰选项来自数据集构建过程中使用的相同控制词汇表。

通过保持固定的输入顺序，包括首先输入图像，然后是问题文本，再然后是答案选项，我们控制了位置变化，并确保所有模型以可比较的方式处理多模态背景。相同的提示结构、标记限制和格式被完全应用于 GPT-4o、GPT-5、GPT-5-mini 和 GPT-5-nano，使性能差异可以归因于模型的内在能力，而不是提示的变化。

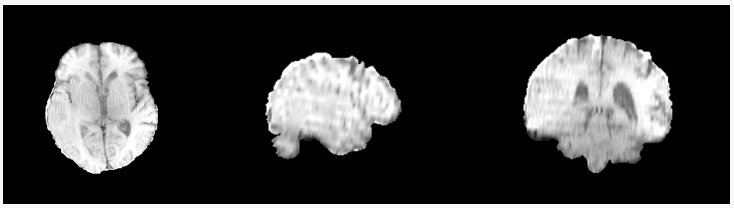
在三个 BraTS 队列（MET，GLI，MEN）中，模型准确率集中在一个狭窄的范围内，绝对值约在三分之一到二分之一之间。每个队列的结果（正确百分比）在表 1 中总结如下：

一个宏平均（MET/GLI/MEN 的未加权平均）得到结果：GPT-5-mini 44.19 %，GPT-5 43.71 %，GPT-4o-2024-11-20 41.49 %，以及 GPT-5-nano 35.85 %。因此，观察到 GPT-5-mini 具有最高的宏平均值，但它相对于 GPT-5 的优势仅为 0.48 百分点，这在 VQA 设置中相对于预期抽样变异性来说非常小。值得注意的是，GPT-4o-2024-11-20 在 GLI 准确性上达到了最佳成绩（49.80 %），而 GPT-5 在 MET（42.68 %）和 MEN（42.12 %）上领先。GPT-5-nano 在所有群体中表现不佳（见表 1 ）。

本研究评估了四种大型语言模型在从 BraTS 数据集中提取的、具有临床依据的脑肿瘤 VQA 基准上的表现。在所有肿瘤亚型中，各模型准确率聚集在一个相对较窄的范围内，其中 GPT-5-mini 实现了最高的宏平均准确率（44.19 %），紧随其后的是 GPT-5（43.71 %）和 GPT-4o（41.49 %），而 GPT-5-nano 则表现不佳。GPT-5 与 GPT-5-mini 之间的微小差异表明，除模型规模之外的因素也可能影响该领域的性能。

GPT-5-mini 的轻微优势的一个可能解释是，小规模模型可能展示出更加保守的决策过程，减少在强迫选择格式中受到干扰选项的影响。较大的模型，如 GPT-5，虽然通常表现出更强的综合推理能力，但也可能更容易过度解读模糊的特征或对不相关的图像和文本相关性进行过拟合，特别是在提示上下文极少的高度结构化任务中。不同肿瘤类型的性能变化也可能反映每个队列不同的放射特征。胶质母细胞瘤，由于其异质增强、坏死和瘤周水肿，可能有利于使用更通用推理策略的模型，这可能解释了 GPT-4o 在此类别

A Sample from BraTS-GLI (case BraTS-GLI-00017-000-t1n)



```
[
  {
    "role": "system",
    "content": "You are a helpful medical assistant."
  },
  {
    "role": "user",
    "content": [
      {
        "type": "text",
        "text": "
            Q: Are the lesion boundaries clear or ill-defined
            Answer Choices:
                (A) clear                (B) ill-defined
                (C) mixed/ambiguous
            A: Let's think step by step."
      },
      {
        "type": "image_url",
        "image_url": { "url": "slices_out_gl/BraTS-GLI-00017-000-t1n_tri.png" }
      }
    ]
  },
  {
    "role": "assistant",
    "content": "{ASSISTANT_RATIONALE}"
  },
  {
    "role": "user",
    "content": "Therefore, among A through C, the answer is"
  },
  {
    "role": "assistant",
    "content": "B"
  }
]
```

Figure 3: BraTS-GLI 案例的示例 VQA 提示。首先提供三平面图像拼接，然后是问题和标记的答案选项。模型输出一个中间推理步骤（`ASSISTANT_RATIONALE`），然后输出一个对应于其选择答案的最终字母（`ASSISTANT_FINAL`）。准确性仅从 `ASSISTANT_FINAL` 计算得出。

Table 1: 在包括脑转移瘤（MET）、胶质母细胞瘤（GLI）和脑膜瘤（MEN）的三个 BraTS 肿瘤队列中，准确性（%）和队列的未加权宏平均。" Δ vs GPT-5" 报告了相对于 GPT-5 的宏平均值的百分比点绝对差异。

Cohort	GPT-5	GPT-5-mini	GPT-5-nano	GPT-4o
MET	42.68	42.09	35.95	38.48
GLI	46.34	48.97	38.00	49.80
MEN	42.12	41.52	33.60	36.19
Average	43.71	44.19	35.85	41.49
Δ vs GPT-5	0.00	0.48	-7.86	-2.22

中的相对优势。相比之下，脑膜瘤和转移瘤的更刻板的影像表现可能让能够更有效利用结构化视觉线索的模型受益。

尽管在各个模型中使用了相同的提示结构和三平面马赛克表示，多模态整合仍然是一个显著的挑战。相对接近的准确率聚类表明，在当前基准测试中，没有哪个模型在所有神经肿瘤任务中表现出一致且优越的视觉-文本对齐能力。强制选择格式进一步增加了认知负担，需要在微妙的视觉特征和离散的分类反应之间进行精确映射，这可能限制所有评估系统的性能上限。

重要的是，我们的评估采用了零样本链式思维（CoT）提示，其中模型在生成最终答案之前会生成一个中间推理步骤。虽然 CoT 可以在许多通用基准中提高推理性能，但在这种环境下的影响不一定是积极的。在高度结构化的强迫选择医疗视觉问答（VQA）任务中，CoT 可能会导致对模糊或低信号特征的过度解读，增加对干扰选项的敏感性，或加强虚假的图像-文本关联。我们没有系统地将 CoT 与直接回答提示进行比较，因此它对本基准中性能的净影响仍不确定。先前的研究表明，在类似设置中，CoT 可能只会带来边际收益，或者在某些情况下，由于对无关细节的过度展开，精度会有小幅下降 [18, 19, 20] 。

1.2 局限性

需要承认几个局限性。首先，用于结构化特征提取的放射学风格报告是由自动化管道生成的，可能无法完全反映自然临床文档的细微差别和变异性。其次，缺乏人类专家对照限制了报告准确度的临床可解释性，因为神经放射学家之间的观察者间变异是一个相关的参考点。第三，准确率是唯一的评估指标；其他重要维度，如推理透明性、不确定性校准和错误分析等均未评估。第四，所有报告的结果均使用零样本链式思维（CoT）提示获得。虽然这种方法可以促进更结构化的推理，但也可能导致对模糊特征的过度解释或依赖于虚假的图像-文本关联，可能在某些情况下降低准确性。我们没有评估不使用 CoT 的直接回答基线，因此这种提示策略对性能的净影响仍不确定。最后，所有模型都在没有领域特定微调的零样本设置下进行评估，这可能低估了在应用临床场景中可实现的性能。

1.3 未来方向

未来的研究应通过多种途径解决这些局限性。涉及人类专家的比较研究将有助于阐明这些系统在诊断工作流程中的实际效用。本研究中使用的数据集已经提供了与图像配对的文本信息，这可以使设计需要模型生成自由文本应答的开放式 VQA 任务成为可能。这种方法将有助于对 GPT 模型进行更细致的评估，尤其是在直接与人类专家推理对比时。此外，整合模型校准指标和视觉基础评估，将通过识别模型何时因错误原因而正确或在错误答案中过于自信，从而增强可解释性和可信度。最后，针对特定领域的微调，利用神经肿瘤学影像资料库，可能会进一步提高性能，特别是在处理微妙或模糊的影像特征时。

2 结论

本研究表明，GPT-5 系列模型在一个标准化、临床基础的脑肿瘤 VQA 基准测试中达到了中等准确性，各变体之间的性能差异较小。这些结果表明模型具备有意义但仍然有限的结构化神经肿瘤学推理能力。重要的是，基准测试本身提供了一个可再现、可扩展的测试平台，我们预计它将成为未来研究的宝贵资源，使得跨模型、提示和训练方案的系统比较成为可能。同时，要将其应用于实践，还需要更广泛的验证：人类专家比较器、超越准确性的评估（例如校准和推理透明度）、使用配对文本的开放形式 VQA 以及跨多样化数据集和成像协议的领域特定微调。总之，这些步骤将澄清此类系统何时以及如何支持真实世界中的临床工作流程。

References

[1] J Ricardo McFaline-Figueroa and Eudocia Q Lee. Brain tumors. The American journal of medicine, 131(8):874–882, 2018.

[2] David N Louis, Arie Perry, Pieter Wesseling, Daniel J Brat, Ian A Cree, Dominique Figarella-Branger, Cynthia Hawkins, H K Ng, Stefan M Pfister, Guido Reifenberger, Riccardo Soffietti, Andreas von Deimling, and David W Ellison. The 2021 who classification of tumors of the central nervous system: a summary. Neuro-Oncology, 23(8):1231–1251, 06 2021.

[3] Aske Foldbjerg Laustsen, Rob Dineen, Jurgita Ilginiene, Jonathan Kjær Grønbæk, Astrid Marie Sehested, Kjeld Schmiegelow, René Mathiasen, Marianne Juhler, and Shivaram Avula. Interobserver variability in assessing preoperative imaging biomarkers for cerebellar mutism syndrome: a multiobserver pilot study. Pediatric Radiology, pages 1–12, 2025.

[4] Emily M Crowe, William Alderson, Jonathan Rossiter, and Christopher Kent. Expertise affects inter-observer agreement at peripheral locations within a brain tumor. Frontiers in psychology, 8:1628, 2017.

[5] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437, 2024.

[6] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.

[7] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. Nature, 620(7972):172–180, 2023.

[8] Jiayan Guo, Lun Du, Hengyu Liu, Mengyu Zhou, Xinyi He, and Shi Han. Gpt4graph: Can large language models understand graph structured data ? an empirical evaluation and benchmarking, 2023.

[9] Shansong Wang, Zhecheng Jin, Mingzhe Hu, Mojtaba Safari, Feng Zhao, Chih-Wei Chang, Richard LJ Qiu, Justin Roper, David S Yu, and Xiaofeng Yang. Unifying biomedical vision-language expertise: Towards a generalist foundation model via multi-clip knowledge distillation. arXiv preprint arXiv:2506.22567, 2025.

[10] Shansong Wang, Mojtaba Safari, Qiang Li, Chih-Wei Chang, Richard LJ Qiu, Justin Roper, David S Yu, and Xiaofeng Yang. Triad: Vision foundation model for 3d magnetic resonance imaging. Research Square, pages rs–3, 2025.

[11] Mingzhe Hu, Joshua Qian, Shaoyan Pan, Yuheng Li, Richard LJ Qiu, and Xiaofeng Yang. Advancing medical imaging with language models: featuring a spotlight on chatgpt. Physics in Medicine & Biology, 69(10):10TR01, 2024.

[12] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems, 2023.

[13] Jiayu Lei, Xiaoman Zhang, Chaoyi Wu, Lisong Dai, Ya Zhang, Yanyong Zhang, Yanfeng Wang, Weidi Xie, and Yuehua Li. Autorg-brain: Grounded report generation for brain mri, 2024.

[14] Shansong Wang, Mingzhe Hu, Qiang Li, Mojtaba Safari, and Xiaofeng Yang. Capabilities of gpt-5 on multimodal medical reasoning. arXiv preprint arXiv:2508.08224, 2025.

[15] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging, 34(10):1993–2024, 2014.

[16] Dominic LaBella, Maruf Adewole, Michelle Alonso-Basanta, Talissa Altes, Syed Muhammad Anwar, Ujjwal Baid, Timothy Bergquist, Radhika Bhalerao, Sully Chen, Verena Chung, et al. The asnr-miccai brain tumor segmentation (brats) challenge 2023: Intracranial meningioma. arXiv preprint arXiv:2305.07642, 2023.

[17] Ahmed W Moawad, Anastasia Janas, Ujjwal Baid, Divya Ramakrishnan, Rachit Saluja, Nader Ashraf, Nazanin Maleki, Leon Jekel, Nikolay Yordanov, Pascal Fehringer, et al. The brain tumor segmentation-metastases (brats-mets) challenge 2023: Brain metastasis segmentation on pre-treatment mri. ArXiv, pages arXiv–2306, 2024.

[18] Lennart Meincke, Ethan Mollick, Lilach Mollick, and Dan Shapiro. Prompting science report 2: The decreasing value of chain of thought in prompting. arXiv preprint arXiv:2506.07142, 2025.

[19] Ryan Liu, Jiayi Geng, Addison J Wu, Ilia Sucholutsky, Tania Lombrozo, and Thomas L Griffiths. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. arXiv preprint arXiv:2410.21333, 2024.

[20] Sohyeon Jeon and Hong-Gee Kim. A comparative evaluation of chain-of-thought-based prompt engineering techniques for medical question answering. Computers in Biology and Medicine, 196:110614, 2025.