

不同的问题，不同的模型：使用 LLM 进行临床问答的不确定性和校准的精细化评估

Alberto Testoni^{1,2}, Iacer Calixto^{1,2,3}

¹Department of Medical Informatics, Amsterdam University Medical Center, University of Amsterdam, Amsterdam, The Netherlands,

²Amsterdam Public Health, Methodology, Amsterdam, The Netherlands,

³Amsterdam Public Health, Mental Health, Amsterdam, The Netherlands

Correspondence: a.testoni@amsterdamumc.nl; i.coimbra@amsterdamumc.nl

Abstract

在高风险领域如临床决策支持中，准确且良好校准的不确定性估计是至关重要的。我们对临床多项选择题回答的不确定性估计方法进行了细粒度评估，涵盖了两个数据集、十一种医学专业、六种问题类型的十种开源大型语言模型（通用、生物医学和推理模型）。我们比较了标准的单次生成和采样方法，并通过案例研究探索了基于推理轨迹中行为信号的简单单次估计方法。这些轻量级的方法在只需一次生成的情况下接近语义熵的性能。我们的结果揭示了各个专业和问题类型间的显著差异，强调了根据问题的性质和模型的具体优势来选择模型的重要性。

1 介绍

临床文档在现代医疗中起着核心作用，包含大量结构化和非结构化文本。其日益增加的复杂性和规模对临床医生造成了更大的负担，使得信息检索和解释变得困难 (Ehrenfeld and Wanderer, 2018; Gesner et al., 2019; Tierney et al., 2024)。这些挑战加剧了对自动化临床自然语言处理任务的努力，比如摘要、信息提取和医学问答。大型语言模型 (LLMs) 由于其生成流利文本的能力、跨领域迁移能力以及上下文推理能力，已成为处理复杂医疗文档的有前途工具 (Thirunavukarasu et al., 2023; Singhal et al., 2023; Van Veen et al., 2024)。

与那些受益于成熟的基准和已明确任务定义的传统自然语言处理任务不同，临床自然语言处理仍然是一个相对新兴的领域，存在有限的基准资源以及缺乏关于最佳实践的共识 (Blagec et al., 2023; Au Yeung et al., 2023; Thirunavukarasu, 2023)。尽管如此，大规模语言模型正越来越多地应用于临床环境中，通常未对其可靠性、校准或稳健性进行严格评估。这一验证上的差距对它们在医学问答等高风险应用中的可信性提出了严重的担忧，而在这些应用中，事实准确性和校准的不确定性对于安全和负责任的应用至关重要。

在自然语言生成任务中估计不确定性，特别是使用大型语言模型 (LLMs)，存在独特的挑战。与传统分类设置不同，输出空间是离散且一个明确定义的，而文本生成涉及一个开放、高维度的输出空间，其本质上是随机的且自回归的 (Baan et al., 2023)。早期的研究探索了不确定性量化 (UQ) 技术，比如使用标记级别的置信度 (Huang et al., 2023) 和通过采样进行语义一致性 (Kuhn et al., 2023; Farquhar et al., 2024)，并使用像期望校准误差和布里尔评分 (Van Calster et al., 2019) 这样的校准指标。然而，研究临床自然语言处理任务（例如医疗问答）的不确定性量化的研究相对较少。尽管已经有一些新发表的研究，但这一方向的现有工作 (Wu et al., 2024; Savage et al., 2025; Wang et al., 2025) 范围有限，仅评估了一小部分过时的模型，并没有包括最新的进展比如推理模型和领域特定的 LLMs。此外，这些评估缺乏细粒度分析，未能深入理解在医学专业或问题类型上的性能变异——这是两个临床相关的维度，能够提供对模型能力和局限性的更为细致的理解。

在我们的研究中，我们评估了不确定性量化 (UQ) 方法，涉及多个临床多项选择问答的大型语言模型 (LLM)，涵盖两个数据集 (S-MedQA 和 MedExQA)、11 个医学专业——其中六个是常见专业，五个是较少代表的专业——我们还引入了一种新的注释方式，根据问题类型（例如，诊断、定义等）对问题进行分类。我们评估了 10 个通用、推理和生物医学大型语言模型的不确定性估计。我们的结果表明，推理模型在准确性和其不确定性估计的区分/校准方面比其他模型类型具有优势。然而，这一优势并未在所有医学专业和问题类型中普遍存在。关键的是，我们经常观察到具有最高准确性的模型和校准最好的模型之间存在不匹配。

从观察到基于采样的方法持续产生更强的不确定性估计 (图 1) 开始，我们研究了单次生成方法作为基于采样的不确定性量化方法的有效替代品的潜力。为此，我们引入了一种方法，

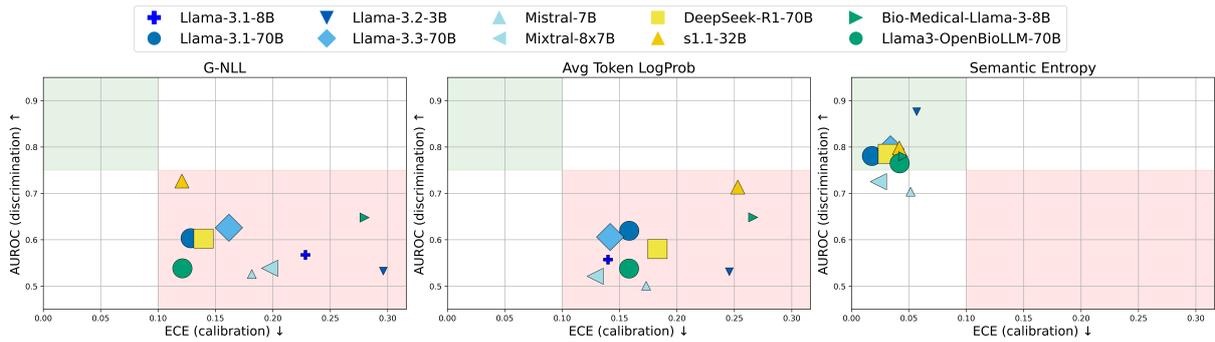


Figure 1: 跨 LLMs 不确定性估计方法的评估。每个子图显示 AUROC（辨别力）与 ECE（校准）的关系。标记的大小反映模型的大小；颜色和形状表示模型类型（通用型、推理型、生物医学型）。绿色阴影区域（左上方）标记了理想性能的区域。

该方法利用大型语言模型的推理痕迹中的行为信号——如响应长度、自我质疑和自我验证——作为轻量级不确定性替代，提供具备竞争力的性能，无需多个生成。总体而言，我们的工作强调了在临床问答中精细评估的重要性，并为根据专业和问题意识标准选择模型提供了实用指导。¹

2 相关工作

不确定性量化传统上在 NLP 任务中被研究，例如情感分析、机器翻译和命名实体识别 (Xiao and Wang, 2019; Zerva et al., 2022; Hashimoto et al., 2024)。在临床 NLP 中，确保预测的良好校准是至关重要的，因为过度自信可能会转化为对患者的风险或诊断错误 (Ghoshal et al., 2021; Kompa et al., 2021; Dawood et al., 2023)。随着 LLMs 的出现，近期的工作提出了用于评估不确定性的标记级信心水平 (Huang et al., 2023)、多样本的语义熵 (Farquhar et al., 2024) 和提示灵敏度分析 (Si et al., 2023; Zhou et al., 2023)。然而，这些方法中的很少一部分在临床领域得到了系统评估。最近的数据集如 MedQA (Jin et al., 2021) 和 MedExQA (Kim et al., 2024) 使得评估更加严格，但在此背景下检查不确定性估计的研究很少，且通常仅限于小模型集或过时的架构 (Wu et al., 2024; Savage et al., 2025)。此外，评估通常侧重于准确性，忽视了不确定性估计如何随临床专业或问题类型而变化——这两个维度对于在真实世界的医疗环境中部署至关重要。

最近的研究比较了用于一般目的的大型语言模型 (LLMs) 与特定领域的生物医学模型在临床任务上的表现，结果显示了关于生物医学 LLMs 相对于一般目的模型的优势的混合证据 (Dorfner et al., 2024)。然而，这些比较几乎没有关注不确定性校准和鲁棒性。Savage et al.

(2025) 最近进行的一项密切相关的研究调查了临床问答的不确定性估计方法。作者发现，样本一致性——通过随机解码生成的多个答案的一致性来评估——在鉴别力和校准方面都优于其他方法，尽管后者仍然不理想。虽然他们的发现提供了有价值的初步见解，但他们的评估在几个关键方面是有限的。首先，它专注于一小部分过时的模型，忽略了诸如指导调优推理和特定领域 LLMs 等近期进展。其次，他们的分析仅限于诊断和治疗问题，忽略了其他临床相关的问题类型。第三，他们没有研究在不同的医学专业中，表现如何变化。

3 数据集和问题注释

我们关注的是在以往研究中基本被忽略的两个维度：医学专业和问题类型。

按医学专业 我们利用了两个英语多项选择题回答数据集：S-MedQA (Yan et al., 2025) 和 MedExQA (Kim et al., 2024)。S-MedQA 将 MedQA (Jin et al., 2021) 和 MedMCQA (Pal et al., 2022) 的问题结合在一起。S-MedQA 中介绍的数据集包括一个经过验证的注释流程，该流程将每个问题映射到欧盟认可的 55 个医学专业之一。我们考虑了 S-MedQA 测试集中最常见的六个专业——胃肠病学、心脏病学、产科和妇科、神经病学、传染病和儿科，我们在下面称之为“常见专业”。为了在不同专业之间确保评估的平衡，同时保持问题的总数在计算和分析方面可以管理，我们将每个专业的样本量固定为 145 个问题，总共 870 个问题。每个问题包括四个答案选项和一个正确答案。MedExQA 专注于具体冷门的医学专业：生物医学工程 (MedExQA 测试集中有 143 个问题)、临床实验科学 (372 个问题)、临床心理学 (106 个问题)、职业治疗 (189 个问题) 和言语语言病理学 (130 个问题)，总计 940 个问题。MedExQA 中的问题是从各种免费可访问

¹代码和说明将在接受后发布。

的在线资源中手动收集的，例如专门为某些医学专业设计的模拟测试和练习考试。每个问题包括四个答案选项和一个正确答案。

在对每个数据集中的 50 个问题的随机样本进行人工检查后，我们识别出六种临床上有意义的问题类型：诊断、治疗、诊断测试、定义、程序/手术和其他。我们使用 Llama-3.1-70B-Instruct (Meta AI, 2024b) 对所有 1,810 个问题进行标注，并在 100 个问题的样本上对其输出与其中一位作者的手动标注进行验证。结果的吻合度很高，准确率为 94%；少数不一致涉及到广泛的其他类别中的边缘情况。关于提示、问题类型定义和分布的详细信息，请参见附录。

4 方法论

4.1 提取不确定性估计

LLM 的不确定性量化方法大致可以分为两类：只需一次响应的单次方法和依靠为相同输入生成多个响应的基于采样的方法，后者受自我一致性理论 (Wang et al., 2023) 的启发。大多数单次方法利用令牌级别的概率。在我们的工作中，我们评估了该类中的两种代表性方法。首先，我们计算了模型响应的平均令牌级概率，参考 Huang et al. (2023)。与之前的发现一致 (Savage et al., 2025)，我们观察到使用最小或最大令牌概率时差异可以忽略不计。我们使用以贪婪解码生成的响应，因为 (a) 对于此指标，采样不会带来一致的性能提升，并且 (b) 贪婪解码确保与下一个方法兼容。其次，我们包括最近由 Aichberger et al. (2024) 引入的贪婪负对数似然 (G-NLL)，该方法计算通过贪婪解码生成的输出序列的负对数似然。作者为 G-NLL 提供了一个理论上的论证，作为模型不确定性的有效代理，显示其与信度和校准指标有很好的相关性。

对于依赖于对多个回答进行采样并评估其一致性的方法，我们采用了受 Kuhn et al. (2023); Farquhar et al. (2024) 启发的语义熵。尽管与之密切相关，但 Savage et al. (2025) 中的样本一致性方法通过测量多重生成输出的一致性来量化不确定性，而语义熵则直接捕捉响应分布的熵。鉴于我们的任务是多选题的性质，我们采用后者，将选择相同答案选项的响应归类在一起，并计算由此产生的答案分布的熵。潜在的假设是，高语义熵反映出更大的模型不确定性，因为这表明存在多种竞争的解释或答案；相反，低熵反映出语义意图的一致性，因此自信度更高。尽管这种方法依赖于重复采样来估计不确定性，但已有研究表明其在性能上优于基于标记级别的置信度指标。我们为每个输入

生成 10 个响应，与以往的工作和 Manakul et al. (2023) 中提出的分析一致。尽管之前的工作已研究了基于双向蕴涵或 GPT-4 注释的聚类策略——各自有其权衡和潜在噪音来源——我们利用了任务的结构：模型被提示将其选定的答案选项放在方括号中，这使我们能够可靠地对响应进行分组 (详细信息见第 4.2 节)。

我们评估了 10 个开源的经过指令调优的大语言模型。这组多样化的模型使得针对架构、规模和领域专业化的不确定性估计进行了全面评估。所有模型使用的提示以及其他详细信息在附录中报告。我们包括六个不同大小和架构的通用模型：LLaMA-3.1-8B、LLaMA-3.1-70B、LLaMA-3.2-3B、LLaMA-3.3-70B、(Meta AI, 2024a)，以及 Mistral-7B-v0.2 和 Mixtral-8x7B-v0.1 (Mistral AI, 2023)；两个优化推理的模型：DeepSeek-R1-70B (Guo et al., 2025) 和 s1.1-32B (Muennighoff et al., 2025)；最后，还有两个特定领域的生物医学模型：Bio-Medical-LLaMA-3-8B (ContactDoctor, 2024) 和 LLaMA3-OpenBioLLM-70B (Ankit Pal, 2024)，都是在生物医学文献和临床语料库上进行了微调的。

4.2 评估

答案提取和准确性 模型被提示在方括号中生成其选择的答案选项 (例如, [C])。我们使用正则表达式提取这个选项。在模型偏离预期格式的情况下 (大约发生在 5% 到 6% 的响应中——例如，在括号中重复完整答案文本)，我们使用 LLaMA-3.1-8B-Instruct 将模型响应映射到四个答案选项之一。对于这项任务，我们验证了 LLaMA-3.1-8B-Instruct 在随机抽取的 100 个模型响应上与人工标注的准确性，发现模型达到了 92% 的准确率。在提取预测的答案选项后，我们通过将其与数据集提供的真实标签进行比较来计算准确率。为了控制可能的位置信息偏差，我们在使用贪心解码时对每个问题的答案选项排序进行四次随机打乱；在采样期间 (用于语义熵)，我们在每次生成时 (每个问题十次) 对顺序进行打乱。

为了评估不确定性估计的判别能力，我们计算接收者操作特征曲线下面积 (AUROC)。这可以衡量不确定性分数在多大程度上将正确预测与错误预测分开。AUROC 为 1.0 表示完美的判别能力，而 0.5 则对应于随机猜测。我们使用期望校准误差 (ECE) 和 Brier 得分来评估校准。ECE 测量预测信心与实际准确性之间的差异，量化过度或不足的信心。根据 Rivera et al. (2024); Savage et al. (2025)，我们在 ECE 的实现中使用 10 个区间。Brier 得分捕获预测信心

和正确性之间的平方误差，提供一个结合校准和锐度两个方面的正确评分规则。对于 ECE 和 Brier 得分，值为 0 表示完美校准。

为实现跨模型的不确定性估计质量的整体比较，我们引入了一个简单的综合指标——DC 分数，作为 AUROC, ECE 和 Brier 分数的平均值，后两个指标经过缩放，使得较高的数值始终表示更好的表现。正式来说：

$$\text{DC Score} = \frac{\text{AUROC} + (1 - \text{ECE}) + (1 - \text{Brier})}{3}$$

。设计上，DC 分数为 1 意味完美的不确定性估计：理想的区分能力 (AUROC = 1) 和完美的校准 (ECE = 0, Brier Score = 0)。

5 结果

我们比较了第 4.1 节中讨论的三种不确定性估计方法，以评估 (a) 哪种方法能产生最符合校准和最具辨别力的估计，以及 (b) 模型类型或大小是否系统地影响它们。图 1 显示了所有模型和方法的 AUROC (辨别能力) 与 ECE (校准能力)。模型类型用颜色标识 (蓝色：通用，黄色：推理，绿色：生物医学)，标记大小反映模型大小。左上象限标记了理想性能的区域 (高辨别能力，低校准错误)。如预期的那样，语义熵实现了最佳的整体性能，尽管它需要每个输入进行多次生成。无论是面向推理还是生物医学的大型语言模型，都没有持续地优于从通用模型提取的不确定性估计。模型大小同样未显示出对不确定性估计质量的持续影响。值得注意的是，最近的 s1.1-32B 推理模型中，G-NLL 估计表现相对较好。语义熵在几乎所有模型中 (包括最小的一个 LLaMA-3.2-3B) 呈现出强劲的结果。相比之下，Mistral-7B 和 Mixtral-8x7B 相对其他模型表现不佳。

基于这些发现，我们在本节余下部分采用语义熵作为主要的不确定性估计方法。在第 6 节中我们将重新审视单生成替代方法。虽然语义熵为广泛的模型范围提供良好的校准不确定性估计，但仍然存在重要问题：这种表现是否在所有医学专业领域中均匀保持？以及这些不确定性估计与实际问答准确性的一致程度如何？

5.1 医学专业的影响

表 1 报告了不同医学专业的模型准确性，每个专业中排名前三的模型以粗体突出显示。我们观察到模型规模的一个一致性效果：较大的模型通常比较小的模型具有更高的准确性，特别是在 LLaMA 系列中。推理模型 (s1.1-32B 和 DeepSeek-R1-70B) 经常位列表现最好的选手之中，这表明指令调优和推理踪迹在各个领

域中促进了稳健的表现。生物医学模型在较小规模时往往表现优于通用模型 (例如，8B)，但这种优势在较大规模时不再存在，此时像 LLaMA-3.1-70B 这样的模型超过了它们的特定领域对应物 (例如，Llama3-OpenBioLLM-70B)。值得注意的是，所有模型在语言病理学专业上的表现仅略高于偶然水平，这与原始数据集作者 (Kim et al., 2024) 的发现相呼应。

为了评估不确定性估计与模型准确性的一致性，图 2 显示了每个专业前三名模型的平均 DC 分数——一个结合了辨别力和校准能力的综合指标。那些在准确性上也排名前三的模型用红色边框标出，使我们能够评估可靠的不确定性估计与强大的 QA 性能同时出现的情况。我们观察到，推理模型通常在各个专业中都能获得较高的 DC 分数，并且常常与前三名的准确性相符 (例如，在心脏病学、神经病学和传染病领域)。然而，这一模式并不在所有专业中一致。在诸如儿科和职业治疗等领域，推理模型表现出相对较差的校准能力，而最佳校准的模型中没有一个在准确性上名列前茅。值得注意的是，在诸如言语语言病理学等具有挑战性的专业中——模型准确性较低使得可靠的不确定性估计尤为重要——校准最好的模型是评估中规模最小的模型，LLaMA-3.2-3B。

5.2 问题类型的影响

表格 2 报告了每个模型在问题类型上的准确率。在基于专业的分析中观察到的几个趋势仍然存在：在 LLaMA 系列中，较大的模型始终优于其较小的同类，而注重推理的模型通常在准确率方面名列前茅。我们新的问题类型注释还揭示了程序/操作是一个特别具有挑战性的类别，所有模型的表现都不佳，只有 OpenBioLLM-70B 的准确率超过 70%。这可能源于正确回答此类问题所需的程序和顺序推理，通常超出了表层事实检索。而相比之下，定义问题显得最为简单，几乎所有模型都达到高准确率。这可能反映出定义类查询在形式和内容上更接近于预训练数据，通常类似于词汇表条目或百科全书式的描述。

转向不确定性估计，图 3 揭示了几个显著的模式。同样，虽然推理模型通常能产生良好校准的估计，但这并不在所有问题类型中都适用。例如，在 Treatment 类别中，尽管 DeepSeek-R1-70B 和 s1.1-32B 都取得了接近 80% 的准确率，但它们都不属于校准最佳的模型。相反，最好的不确定性估计来自于小得多的 LLaMA-3.2-3B。同样的模式也能在 Definition 问题中观察到。相比之下，对于更复杂的 Procedure/Operation 问题，尽管推理模型的准确率低于一些大型通用模型，但它们却能够提

Model	Gastro.	Cardiology	Obs. and gyn.	Neurology	Infectious dis.	Pediatrics	Biomed. Eng.	Clinical Lab.	Clinical Psy.	Occ. Therapy	Speech-Lang. Path.
Llama-3.1-8B	0.623	0.633	0.651	0.617	0.601	0.616	0.714	0.742	0.785	0.784	0.286
Llama-3.1-70B	0.769	0.781	0.828	0.774	0.750	0.848	0.868	0.887	0.925	0.870	0.342
Llama-3.2-3B	0.538	0.484	0.512	0.543	0.534	0.512	0.572	0.661	0.665	0.713	0.285
Llama-3.3-70B	0.769	0.812	0.809	0.719	0.748	0.769	0.834	0.881	0.851	0.848	0.335
Mistral-7B	0.463	0.529	0.505	0.452	0.401	0.440	0.787	0.700	0.764	0.762	0.287
Mixtral-8x7B	0.590	0.607	0.641	0.573	0.512	0.584	0.811	0.794	0.835	0.792	0.285
DeepSeek-R1-70B	0.828	0.833	0.841	0.831	0.862	0.886	0.877	0.904	0.932	0.856	0.310
s1.1-32B	0.847	0.845	0.850	0.819	0.867	0.867	0.844	0.896	0.925	0.861	0.318
Llama3-OpenBioLLM-70B	0.719	0.757	0.759	0.777	0.723	0.779	0.880	0.897	0.897	0.865	0.345
Bio-Medical-Llama-3-8B	0.653	0.647	0.705	0.648	0.659	0.643	0.743	0.722	0.818	0.737	0.273

Table 1: 各医学专科的准确性。每个专科的前三名模型以粗体显示。

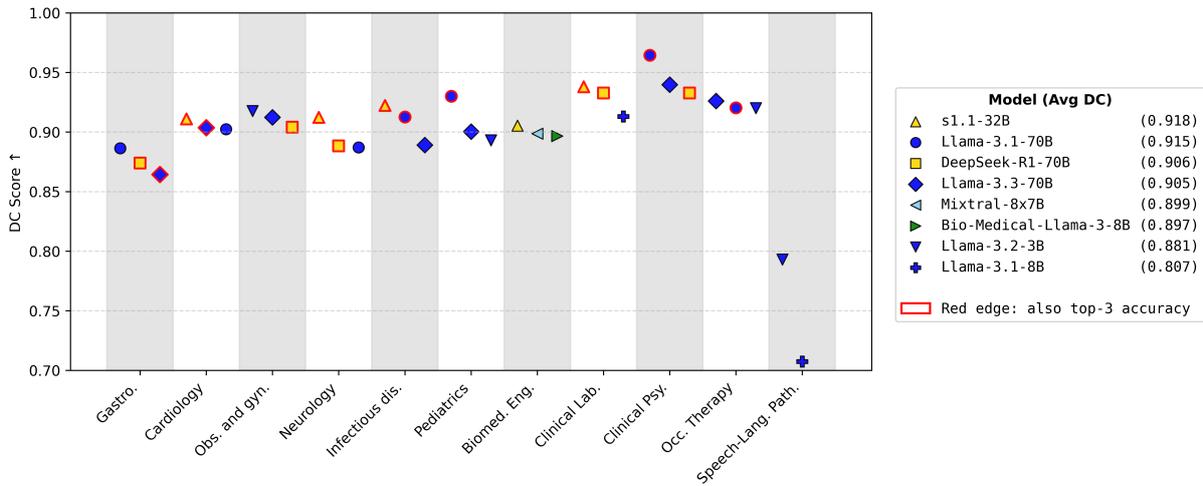


Figure 2: 每个医学专业中由 DC 得分 (准确率、1-ECE 和 1-Brier 的平均值) 排名的前三个模型。红色边框的标记表示在同一专业中以准确率排在前三名的模型。对于言语语言病理学，我们仅报告其不确定性估计达到 DC 得分超过 0.70 的模型。

Model	Diagnosis	Treatment	Diagnostic Test	Definition	Procedure/Operation	Other
Llama-3.1-8B	0.620	0.646	0.659	0.730	0.618	0.558
Llama-3.1-70B	0.790	0.772	0.818	0.853	0.686	0.733
Llama-3.2-3B	0.538	0.543	0.570	0.645	0.443	0.448
Llama-3.3-70B	0.766	0.761	0.805	0.835	0.633	0.688
Mistral-7B-v0.2	0.482	0.503	0.618	0.720	0.549	0.530
Mixtral-8x7B-v0.1	0.588	0.598	0.719	0.776	0.611	0.616
DeepSeek-R1-70B	0.839	0.806	0.863	0.853	0.631	0.772
s1.1-32B	0.848	0.795	0.833	0.845	0.617	0.819
Llama3-OpenBioLLM-70B	0.739	0.775	0.825	0.869	0.715	0.765
Bio-Medical-Llama-3-8B	0.657	0.639	0.627	0.737	0.591	0.597

Table 2: 每种问题类型的准确率。每种问题类型的前 3 名模型以粗体显示。

供一些校准最佳的估计。这种脱节现象证实了一个关键点：高准确率并不保证能够提供可靠的不确定性估计，特别是在需要细致推理或程序性知识的问题类型中。

总之，我们的研究结果强调了超越整体模型性能的重要性，强调需要进行精细化、专业意

识和问题意识的评估，以确保不确定性估计在不同的临床领域和问题类型中保持可靠和有信息性。没有这样的细粒度，模型在平均上可能显得稳健，但在那些准确和校准预测最为关键的特定高风险领域中却会失效。此分析也为实践者提供了实用的指导，能够在考虑的专业领域基础上进行更明智的模型选择，并建议使用结合模型间互补优势的集成方法的机会。

6 推理模型中不确定性的行为信号：一个案例研究

在第 5 节中，我们发现语义熵可以在各种模型类型和规模中产生良好校准的不确定性估计。虽然在医学专业和问题类型之间性能有所不同，但语义熵在校准和区分方面明显优于基于句子级别对数概率的方法。尽管有效，这种方法需要针对每个输入进行多次生成（10 个样本）：这一特性可能会阻碍其在实际应用中的部署，特别是在对延迟敏感或资源有限的环境中，例如临床决策支持。相比之下，像 G-NLL

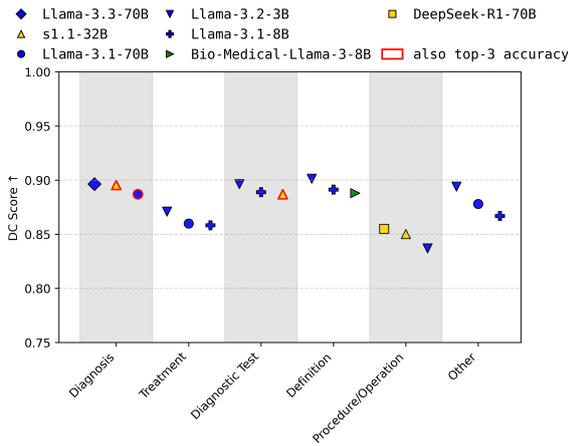


Figure 3: 按 DC 分数排名的每种医疗问题类型的前 3 名模型。

和平均标记概率这样的单一传递方法每个输入只需一次前向传递，但其区分和校准性能有所下降。

为了解决这些限制，我们研究了是否可以将观察到的大规模语言模型 (LLM) 推理过程中行为信号充当轻量级的不确定性代理。在我们的实验中，我们注意到诸如 DeepSeek-R1-70B 和 s1.1-32B 的模型经常产生包含可能与不确定性相关的语言线索的中间推理步骤。例如，我们观察到经常使用自我导向的问题和诸如“Wait ...”之类的标记，这些标记似乎会引发自我验证或重新考虑。此外，我们还假设更长的回答可能表明在得出最终答案时经历了更大的困难。有关这些特征的一个示例，请参见附录 C。有趣的是，s1.1-32B 的作者展示了在测试阶段使用“Wait”标记进行缩放的有效性（有关更多细节，请参见 Muennighoff et al. 2025）。

为了操作化这些假设，我们从每个模型响应中提取三个表层行为特征：(1) 推理过程中的标记数量，(2) 自问问题的数量，以及 (3) 标记“Wait”出现的次数。然后，我们探讨两种将这些特征转化为不确定性估计的策略。

第一种方法是简单地对三个特征进行归一化求和，称为线性求和法。令 x_1 、 x_2 和 x_3 表示三个特征的归一化值，并将不确定性估计为 $u = x_1 + x_2 + x_3$ 。第二种方法将这三个特征作为一个监督模型的输入：无论是逻辑回归还是校准回归（前者建模为 S 型似然，后者通过 L2 正则化拟合成岭回归的线性函数）。为了训练这些模型，我们使用与每个响应关联的实际正确性标签（正确为 1，错误为 0）。形式上，模型学习一个函数 $f: \mathbb{R}^3 \rightarrow [0, 1]$ 。尽管回归模型是为了预测正确性而训练的，我们假设其输出可以作为信心水平的代理。我们将不确定

性定义为它们的补充，即 $u = 1 - f(\mathbf{x})$ ，从而能够与其他不确定性估计方法进行直接比较。为了确保公平比较并避免过拟合，我们在 40% 的数据集上训练回归模型，同时在剩下的 60% 上进行评估，以保持医学专业的原始分布。所有的响应都是通过贪婪解码生成的，以便能够直接与单通道基线（如 G-NLL 和平均标记概率）进行比较。我们使用 AUROC、ECE 和 Brier Score 评估行为不确定性信号的质量。为了进一步解析模型行为，我们分别报告常见和不常见医学专业的结果（有关后者，见附录 C.3）。我们包括了语义熵的比较结果，注意到它依赖于重复采样来估计不确定性。

如图 4 所示，结果表明，我们基于轻量级行为特征的方法——线性和、逻辑回归和校准回归——优于其他句子级方法，并且性能与语义熵相当，而每个输入只需一次前向传递。对于两种推理 LLM，DeepSeek-R1-70B 和 s1.1-32B，这种模式均成立。即使是简单的线性和方法，在大多数情况下也优于 G-NLL 和平均标记对数概率等标记级基线，特别是在区分性 (AUROC) 方面。校准回归表现始终良好，特别是在校准方面。在附录 C.2 中，我们提供了关于回归模型系数的更多细节。

总体而言，这些结果支持我们的假设，即推理轨迹中的行为线索编码了关于模型不确定性的相关信息，并指出了一种无需采样或逐词概率的高效、轻量且可解释的不确定性估计的有前途方向。这对于需要可靠且高效地估计不确定性的现实应用具有重要意义——例如，在临床决策支持等场景中，延迟和可解释性都至关重要。关于较少代表性专业的结果在附录 C.3 中报告。它们证实了我们的方法对 DeepSeek-R1-70B 的有效性，同时对 s1.1-32B 的改善有限，强调了在资源较少的环境中需要针对模型的调整或更有表现力的行为特征。我们的案例研究为未来验证这种方法在更广泛的任务、数据集和模型中提供了起点。我们将我们的贡献视为朝着建立一个利用最近发布的推理大型语言模型输出的、更一般的行为不确定性估计框架迈出的初步步骤。

7 讨论与结论

我们的研究对用于解决临床多项选择题的 LLM 的不确定性估计方法进行了细粒度的评估。我们的研究跨越了包括通用、生物医学和推理模型在内的十个 LLM，评估了两个数据集、十一种医学专科和六种问题类型。我们的结果揭示了不确定性估计校准在专科和问题类型上的显著差异，高准确率模型并不总能产生良好校准的估计。尽管准确率较低，我们强调

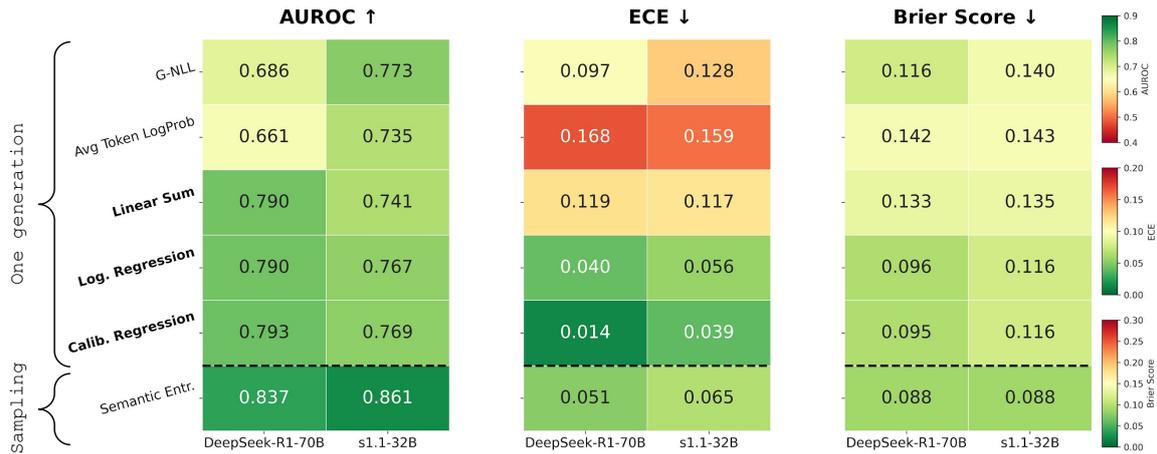


Figure 4: DeepSeek-R1-70B 和 s1.1-32B 的不确定性估计方法的比较。我们提出的方法基于行为特征（用粗体显示），在 AUROC、ECE 和 Brier Score 方面表现强劲，接近语义熵的方法，而后者则依赖于通过采样进行的多次生成。

了小型 LLM，例如 LLaMA-3.2-3B，在生成良好校准的不确定性估计方面的潜力，这是临床支持应用的一个重要特征。此外，我们发现最新推出的 s1.1-32B，一个只需 1K 精心挑选样本训练的高样本效率推理模型，在其他模型中表现突出，展示了战略性数据整理和以推理为重点的目标的价值。我们的研究结果强调了综合报告的局限性，并突出了在临床 NLP 及其他领域中自适应、上下文感知策略对不确定性估计的价值，同时为从业者提供了利用不同模型优势的可行指南。

尽管语义熵在校准和辨别中实现了最佳的整体表现，其对多次生成的依赖可能在某些场景中限制其实用性。为了解决这一问题，我们展示了一个案例研究，在其中我们引入了一种基于推理轨迹中的行为线索（如长度、自我提问和自我验证标记）的单次传递、可解释的方法，其表现与语义熵相匹配，并且优于标准的标记概率基准，为轻量级集成和模型选择策略铺平了道路。

8

局限性

在下文中，我们将讨论我们实验设置的一些局限性。

- 首先，我们关注多项选择题的问答数据集。虽然这种设置确保了明确的正确性标记和可扩展的评估，但它并未涵盖临床自然语言处理任务的全部范围，特别是那些涉及自由形式生成的任务。我们认为这种关注作为第一步是合理的：据我们所知，这是首次在这种规模上系统地评估不同模型类型、医学专业和问题类型的不确定性估计。然而，未来的

工作应该将这一分析扩展到更加开放的临床场景中。

- 其次，我们依赖于模型预测和问题类型注释的自动提取。尽管我们的启发式方法和模型辅助的注释经过手动标签验证并达到了高一致性，但仍可能存在一些剩余噪声，特别是在模糊问题或边界分类（例如，“其他”）的情况下。
- 第三，尽管我们的模型集是多样化的（涵盖通用、推理和生物医学 LLMs），但它仍然仅限于撰写时可用的代表性开源模型，并不包括闭源的专有模型。
- 第四，虽然我们展示了行为特征提供了一种有前景、可解释的替代方案，可以替代像语义熵这样的多样本方法，但我们的特征空间是故意设计得简单且任务特定的。将这些发现推广到其他任务、语言或推理范式中，将需要更广泛的验证以及可能更复杂的特征设计。
- 最后，我们没有应用诸如温度缩放之类的事后校准技术。我们的重点是比较未经调优的原始不确定性估计。在我们的设置中，一致地应用温度缩放是有挑战性的，因为模型在分词方案和输出结构上有所不同——例如，它们在格式化答案时（直接标签与更复杂的句子）的不同会影响可能性的计算方式，并使得校准变得不简单。我们将模型特定校准的整合留作未来工作的重要方向。至于温度设置和解码细节的其他考虑，请参见附录 ??。

这篇出版物是 CaRe-NLP 项目的一部分，其文件编号为 NGF.1607.22.014，属于 AiNed Fel-

lowship Grants 研究项目, (部分) 由荷兰研究理事会 (NWO) 资助。我们要感谢阿姆斯特丹大学的 ILLC 的 DMG 小组和阿姆斯特丹 UMC 的 NLP4Health 实验室的成员, 在项目过程中提供了宝贵的反馈和建议。

References

- Lukas Aichberger, Kajetan Schweighofer, and Sepp Hochreiter. 2024. Rethinking uncertainty estimation in natural language generation. *arXiv preprint arXiv:2412.15176*.
- Malaikannan Sankarasubbu Ankit Pal. 2024. Openbiollms: Advancing open-source large language models for healthcare and life sciences. <https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B>.
- Joshua Au Yeung, Zeljko Kraljevic, Akish Luintel, Alfred Balston, Esther Idowu, Richard J Dobson, and James T Teo. 2023. Ai chatbots not yet ready for clinical use. *Frontiers in digital health*, 5:1161098.
- Joris Baan, Nico Daheim, Evgenia Iliia, Dennis Ulmer, Haau-Sing Li, Raquel Fernández, Barbara Plank, Rico Sennrich, Chrysoula Zerva, and Wilker Aziz. 2023. Uncertainty in natural language generation: From theory to applications. *arXiv preprint arXiv:2307.15703*.
- Kathrin Blagec, Jakob Kraiger, Wolfgang Frühwirth, and Matthias Samwald. 2023. Benchmark datasets driving artificial intelligence development fail to capture the needs of medical professionals. *Journal of Biomedical Informatics*, 137:104274.
- ContactDoctor. 2024. Bio-medical: A high-performance biomedical language model. <https://huggingface.co/ContactDoctor/Bio-Medical-Llama-3-8B>.
- Tareen Dawood, Chen Chen, Baldeep S Sidhu, Bram Ruijsink, Justin Gould, Bradley Porter, Mark K Elliott, Vishal Mehta, Christopher A Rinaldi, Esther Puyol-Antón, and 1 others. 2023. Uncertainty aware training to improve deep learning model calibration for classification of cardiac mr images. *Medical Image Analysis*, 88:102861.
- Felix J Dorfner, Amin Dada, Felix Busch, Marcus R Makowski, Tianyu Han, Daniel Truhn, Jens Kleesiek, Madhumita Sushil, Jacqueline Lammert, Lisa C Adams, and 1 others. 2024. Biomedical large languages models seem not to be superior to generalist models on unseen medical data. *arXiv preprint arXiv:2408.13833*.
- Jesse M Ehrenfeld and Jonathan P Wanderer. 2018. Technology as friend or foe? do electronic health records increase burnout? *Current Opinion in Anesthesiology*, 31(3):357–360.
- Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.
- Emily Gesner, Priscilla Gazarian, and Patricia Dykes. 2019. The burden and burnout in documenting patient care: an integrative literature review. *MED-INFO 2019: Health and Wellbeing e-Networks for All*, pages 1194–1198.
- Biraja Ghoshal, Allan Tucker, Bal Sanghera, and Wai Lup Wong. 2021. Estimating uncertainty in deep learning for reporting confidence to clinicians in medical image segmentation and diseases detection. *Computational Intelligence*, 37(2):701–734.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Wataru Hashimoto, Hidetaka Kamigaito, and Taro Watanabe. 2024. Are data augmentation methods in named entity recognition applicable for uncertainty estimation? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18852–18867.
- Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. 2023. Look before you leap: An exploratory study of uncertainty measurement for large language models. *arXiv preprint arXiv:2307.10236*.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.
- Yunsoo Kim, Jinge Wu, Yusuf Abdulle, and Honghan Wu. 2024. Medexqa: Medical question answering benchmark with multiple explanations. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 167–181.
- Benjamin Kompa, Jasper Snoek, and Andrew L Beam. 2021. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, 4(1):4.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.
- Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

- Meta AI. 2024a. Llama 3 models. <https://ai.meta.com/blog/meta-llama-3>.
- Meta AI. 2024b. Meta llama 3.1 70b instruct. <https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct>.
- Mistral AI. 2023. Mistral-7b and mixtral-8x7b models. <https://mistral.ai/news/mixtral-of-experts>.
- Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikanan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.
- Mauricio Rivera, Jean-François Godbout, Reihaneh Rabbany, and Kellin Pelrine. 2024. [Combining confidence elicitation and sample-based methods for uncertainty quantification in misinformation mitigation](#). In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertainNLP 2024)*, pages 114–126, St Julians, Malta. Association for Computational Linguistics.
- Thomas Savage, John Wang, Robert Gallo, Abdesslem Boukil, Vishwesh Patel, Seyed Amir Ahmad Safavi-Naini, Ali Soroush, and Jonathan H Chen. 2025. Large language model uncertainty proxies: discrimination and calibration for medical diagnosis and treatment. *Journal of the American Medical Informatics Association*, 32(1):139–149.
- Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Lee Boyd-Graber, and Lijuan Wang. 2023. Prompting gpt-3 to be reliable. In *The Eleventh International Conference on Learning Representations*.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.
- Arun James Thirunavukarasu. 2023. Large language models will not replace healthcare professionals: curbing popular fears and hype.
- Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.
- Aaron A Tierney, Gregg Gayre, Brian Hoberman, Britt Mattern, Manuel Balleca, Patricia Kipnis, Vincent Liu, and Kristine Lee. 2024. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. *NEJM Catalyst Innovations in Care Delivery*, 5(3):CAT–23.
- Ben Van Calster, David J McLernon, Maarten Van Smeden, Laure Wynants, Ewout W Steyerberg, Topic Group ‘Evaluating diagnostic tests, and prediction models’ of the STRATOS initiative Bossuyt Patrick Collins Gary S. Macaskill Petra McLernon David J. Moons Karel GM Steyerberg Ewout W. Van Calster Ben van Smeden Maarten Vickers Andrew J. 2019. Calibration: the achilles heel of predictive analytics. *BMC medicine*, 17(1):230.
- Dave Van Veen, Cara Van Uden, Louis Blanke-meier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, and 1 others. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.
- Zhiyuan Wang, Jinhao Duan, Chenxi Yuan, Qingyu Chen, Tianlong Chen, Yue Zhang, Ren Wang, Xiaoshuang Shi, and Kaidi Xu. 2025. Word-sequence entropy: Towards uncertainty estimation in free-form medical question answering applications and beyond. *Engineering Applications of Artificial Intelligence*, 139:109553.
- Jiaxin Wu, Yizhou Yu, and Hong-Yu Zhou. 2024. Uncertainty estimation of large language models in medical question answering. *arXiv preprint arXiv:2407.08662*.
- Yijun Xiao and William Yang Wang. 2019. Quantifying uncertainties in natural language processing tasks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7322–7329.
- Xinlan Yan, Di Wu, Yibin Lei, Christof Monz, and Iacer Calixto. 2025. [What does neuro mean to cardio? investigating the role of clinical specialty data in medical llms](#). *Preprint*, arXiv:2505.10113.
- Chrysoula Zerva, Taisiya Glushkova, Ricardo Rei, and André FT Martins. 2022. Disentangling uncertainty in machine translation evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8622–8641.
- Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. [Navigating the grey area: How expressions of uncertainty and overconfidence affect language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5506–5524, Singapore. Association for Computational Linguistics.

A

附录

B 模型细节和许可

LLaMA 模型系列和 LLaMA3-OpenBioLLM-70B 根据 Meta AI 社区许可协议进行分发。Mistral-7B-v0.2、Mixtral-8x7B-v0.1、s1.1-32B 根据 Apache 许可证 2.0 发布。DeepSeek-R1-Distill-Llama-70B 根据 MIT 许可证进行分发。Bio-Medical-Llama-3-8B 根据生物医学 Llama-3-8B LLM 许可证发布（仅限非商业用途）。所有模型通过 Hugging Face 接入 (<https://huggingface.co/>)。我们在一个包含 4 个 A100 GPU 的集群上运行推理，批量大小为 8。我们在 2 个 GPU 上运行小型和中型模型 (3B-32B 参数)，在 4 个 GPU 上运行大型模型。我们使用 vLLM 库。运行小型模型推理平均需要几个小时，而运行大型模型和推理模型则需要 24 小时，因为它们产生较长的模型响应。我们根据其许可和同意使用政策使用所有数据集。

为了获得模型对输入问题的响应，我们采用了以下提示（括号中例举了一个实例）：

问题：(一名 59 岁的超重女性因过去两小时前出现严重的腹痛而到急诊诊所就诊。她还抱怨背部有钝痛，伴有恶心和多次呕吐。她的疼痛与食物无关。她的既往病史显著的是胆石症引起的反复腹痛。她的父亲在 60 岁时因某种腹部癌症去世。她的体温为 37°C (98.6°F)，呼吸为 15 次/分，脉搏为 67 次/分，血压为 122/98 mm Hg。体格检查没有发现异常。然而，腹部 CT 扫描显示胆囊附近有一个钙化肿块。在这名患者中，首先应排除以下哪种诊断?) 请提供简洁的答案，并以方括号中的选项字母开始您的回答。选项：([A] 急性胆囊炎 [B] 胆囊癌 [C] 胆总管结石 [D] 胰腺炎) 您的答案：

对于那些附带推荐系统提示的模型，我们按模型开发者/作者提供的使用。

将响应映射到相应的选项 我们使用以下提示将不符合要求格式（即，选项字母用方括号括起）的模型响应映射到相应的答案选择。在执行标注之前，我们检查了 100 个格式错误的响应的随机样本，发现没有必要为此任务包含“其他”/“无效”类。花括号中的占位符。

您将获得一个选择题，四个答案选项标记为 A、B、C 和 D，以及一个自由格式的回答。问题：问题 选项：A. { 选项 a } B. { 选项 b } C. { 选项 c } D. { 选项 d } 回应：{ 回应 } 您的任务是确定回应指的是四个选项中的哪一个。只输出对应的字母，像 [X]。不要包含任何解释或额外的文本。您的标注必须是以下之一：[A]、[B]、[C] 或 [D]。如果回应包含不同的字母，您仍然必须根据回应的内容将其映射到上述四个有效选项中的一个。请记住，回应

可能格式不佳或包含无关字母——仅专注于识别最可能的意图选项。如果多个选项似乎都合理，则选择由回应本身最强烈暗示的一个，而不依赖于外部知识或上下文。您的标注：

问题类型标注 我们使用以下提示将每个问题分配到特定的问题类型（大括号中的占位符）：

你是一名医学专家助手。你的任务是根据以下类别对每个医学问题的类型进行分类：1. 诊断 - 确定疾病或状况。2. 治疗 - 干预措施，如药物、手术或治疗。3. 诊断测试 - 实验室测试、成像或诊断程序。4. 定义 - 询问医学概念的解释。5. 程序/操作 - 有关医学或外科程序的问题。6. 其他根据问题的意图分配最合适的类型。对以下问题进行分类。提供唯一对应的标签数字 (1-6)，并以方括号的形式作为你的回答。问：{ } 问题类型：

对于需要多个样本的实验，我们回顾了通常建议用于核采样的解码配置，以近似语言模型在实际中的使用方式。基于此回顾，我们将 $\text{top-}p$ 值设为 0.9，温度设为 0.6。作为探索性分析，我们也使用更高的温度 1.0（表 3）重复了实验。我们发现，这一变化对模型在准确性方面的总体排名或从语义熵导出的不确定性估计的校准和判别性能影响不大。然而，对于 DeepSeek-R1-70B，我们观察到准确性和校准均有所下降。这一效应值得进一步研究，这需要一个保留的验证集和对更广泛温度值范围的系统分析。这样的分析最终可能支持模型特定甚至任务特定的解码配置，变化各异于模型，医疗专科或问题类型。在这项工作中，我们坚持反映典型实际使用情况的解码设置——温度和 $\text{top-}p$ ——尤其是对于技术知识有限的用户。至于标注问题类型和将响应映射到相应选项，我们采用贪婪解码。

C 推理轨迹中的行为特征

在本节中，我们提供了使用行为特征进行推理模型实验的额外细节，包括基于回归的结果。在项目的早期阶段，我们还测试了其他方法，例如随机森林、简单的 MLP 和来自 SentenceTransformer 的句子嵌入，所有这些方法的表现都不如这里报告的模型，因此被舍弃。我们还实验了三个行为特征子集，但观察到一致的性能下降。

C.1 模型回应示例

在下面由 DeepSeek-R1-70B 生成的示例模型响应中，我们用粗体突出显示了我们工作中研究的一些行为特征，包括自我验证问题和“wait”标记（我们在下面突出显示相关句子）。

问题：一名 28 岁的男性因胸痛来到急诊科。他报告说大约一小时前，他在家里爬楼梯时

Model	Accuracy	AUROC	ECE	Brier Score
Llama-3.1-8B-Instruct	0.583	0.806	0.048	0.176
Llama-3.1-70B-Instruct	0.784	0.873	0.040	0.104
Llama-3.2-3B-Instruct	0.422	0.868	0.101	0.153
Llama-3.3-70B-Instruct	0.774	0.858	0.036	0.112
Mistral-7B-Instruct-v0.2	0.474	0.665	0.078	0.293
Mistral-8x7B-Instruct-v0.1	0.602	0.708	0.040	0.225
DeepSeek-R1-70B	0.767	0.898	0.119	0.113
s1.1-32B	0.837	0.852	0.081	0.090
Llama3-OpenBioLLM-70B	0.699	0.856	0.093	0.140
Bio-Medical-Llama-3-8B	0.666	0.781	0.030	0.169

Table 3: 使用 $T = 1$ 评估结果（不确定性估计作为语义熵提取）。

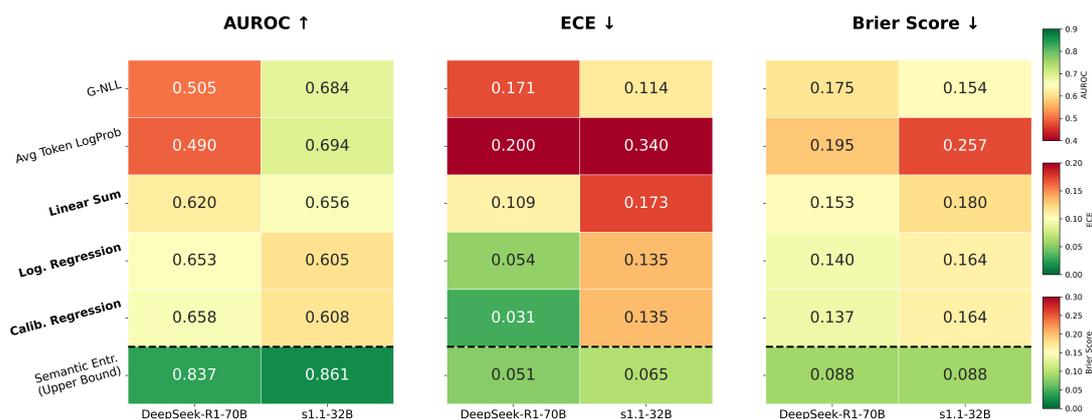


Figure 5: 针对不常见医疗专业的 DeepSeek-R1-70B 和 s1.1-32B 的不确定性估计方法比较。

突然感到剧烈的撕裂样胸痛，并放射到背部。患者的既往病史中有明确心脏杂音，正在门诊心脏科医生处随访。他没有服用任何药物。患者的职业是木匠，并且报告说他对妻子保持单一伴侣关系。他否认有静脉注射药物使用史或近期旅行。在急诊科，患者的体温为 99.4°F (37.4°C)，血压为 94/56 mmHg，心率为 121 次/分钟，呼吸为 14 次/分钟。体检时，患者看起来处于中度痛苦中，并捂住胸口。他身材高大并有关节活动过度。双侧肺呼吸音对称并且充足，但远端脉搏较弱。以下哪项最可能在该患者中被发现？[A] 心尖处有晚收缩期渐强型杂音及中收缩期喀啦音 [B] 左下胸骨缘有中舒张期杂音 [C] 右侧第二肋间隙有收缩期喷射型杂音 [D] 左下胸骨缘有全收缩期高音吹风样杂音模型回答：[C] <think> 好的，我正在尝试弄清楚这个医学问题。让我再读一遍，然后一步一步分析。患者是一名 28 岁的男性，出现了胸痛。在他爬楼梯时，突然出现极其剧烈的撕裂样胸痛，并向后放射。[...] 等一下，但是主动脉瓣返流通常表现为舒张期杂音，不是收缩期。那么也许我把事情搞混了。或者，如果患者有二叶主动脉瓣 (Marfan 综合征中常见)，可能会导致收缩期喷射型杂音。此外，在主动脉夹层中，除非伴有主动脉瓣返流，否则可能不会出现杂音。D. 左胸骨下缘的全收缩期高调吹风样杂音：这通常见于二尖瓣返流。全收缩期特性表明血液在整个收缩期通过二尖瓣回流。但再次强调，患者的表现更倾向于主动脉问题。等等，但患者的胸痛和体检提示主动脉夹层。那么，和主动脉夹层相关的杂音是什么？如果夹层导致主动脉返流，那么杂音将是舒张期的，比如一个舒张早期杂音，最佳听诊点在左胸骨

缘。但选项中没有舒张期的杂音。或者，如果病人有一个主动脉二叶瓣，这是一种常见的先天性心脏缺陷，可能会在右侧第二肋间隙出现收缩期喷射性杂音。主动脉二叶瓣与主动脉夹层的风险增加有关，特别是在马凡综合征或其他结缔组织疾病的背景下。[……]

等等，但之前我认为主动脉返流会出现舒张期杂音。所以，如果夹层导致主动脉返流，杂音将是舒张期的，但这不在选项中。那么杂音可能来自潜在的二叶瓣，这是一个收缩期喷射性杂音。是的，这说得通。患者的心脏杂音很可能源于主动脉二叶瓣，这是一种在右侧第二肋间隙听到的收缩期喷射性杂音。所以答案应该是选项 C。患者突然出现的严重撕裂样胸痛放射到背部，伴随高个子、关节过度活动性以及远端脉弱，强烈提示主动脉夹层。这种情况通常与潜在的主动脉二叶瓣相关，典型表现为右侧第二肋间隙的收缩期喷射性杂音。[C] 右侧第二肋间隙的收缩期喷射性杂音

C.2 回归系数

表 4 列出了基于推理模型行为特征的方法的回归系数。对于这两种模型，较高的标记数量与较低的精度和较差的校准相关，这表明冗长可能反映了不确定性或困惑。这种效应在 s1.1-32B 中更加明显，表明即使在信心不足的情况下，它也可能过度生成。

Feature	Logistic Regr.		Calibration Regr.	
	DeepSeek	s1.1	DeepSeek	s1.1
tokens	-0.6476	-1.4611	-0.0925	-0.1527
questions	-0.1052	0.0755	-0.0196	-0.0181
waits	-0.3456	0.3961	-0.0839	0.0708
bias	1.8559	1.9407	0.8269	0.8556

Table 4: 回归系数。

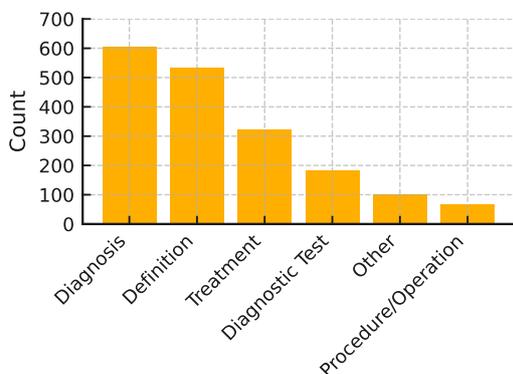


Figure 6: 问题类型的频率。

C.3 不常见专业结果

图 5 显示了在处理不常见医疗专科时使用推理模型行为特征的结果。回归方法优于其他方法，采用 DeepSeek-R1-70B，而 s1.1-32B 显示的改善不太一致。

D 不确定性量化方法的附加细节

首先，我们选择不包括直接提示大型语言模型报告信心水平的方法。尽管先前的研究已探索这些方法，但其可靠性仍然有限。尤其是，Savage et al. (2025) 指出，依赖模型生成的信心水平时，表现和校准性很差。

其次，对于基于对数概率 (logprobs) 的方法，我们仔细考虑了如何处理包含明确推理过程的模型输出。一个显著的例子是 DeepSeek-R1-70B，其响应由清晰分隔的推理段构成，并由诸如 <think> 这样的特殊标记标记出来。在这些情况下，一个关键决策是确定是否应对整个响应（包括推理过程）还是仅对最终答案部分计算不确定性分数。经过实证检查和初步分析，我们发现将对数概率的计算限制在最终答案上——排除推理过程——会产生更可靠和更具可比性的不确定性估计。

E 问题类型分布

图表 6 显示了我们实验中分析的问题类型的频率。

F AI 助手

我们使用了 OpenAI 的 ChatGPT 来支持小的风格修订并协助完成与推理和分析相关的编码任务。所有输出都经过作者的彻底审查和验证。