Jev 与 Open-Weight System One 决策模型的分诊准确性及其在症状评估应用程序中用于早期急症识别的作用

Marvin Kopka[1,2,3*], Austen El-Osta[3], Christoph Kollwitz[1], Felix Balzer[2] 和 Florian Bontrup[1,4]

[1] DOCYET GmbH，德国莱比锡

[2] 柏林夏里特医学院医学信息学研究所（柏林自由大学和柏林洪堡大学的法人成员），德国柏林

[3] 伦敦帝国理工学院自我护理学术研究中心 (SCARU)、初级保健与公共卫生系，英国伦敦

[4] 吕讷堡大学信息系统研究所，德国吕讷堡 ny

[*] 通讯作者

摘要

普通民众越来越多地使用大语言模型（LLM）获取分诊建议，但大语言模型的回答以自由文本形式呈现且不尽相同，因此难以评估其作为医疗器械的适用性。决策模型为每个预定义的分诊类别赋予一个概率，可能更容易评估。为检验其分诊准确性，我们在233个合成病例和10,000个真实世界病例上评估了16个决策模型（14个开放权重模型、2个专有模型），共进行了163,728次评估。表现最佳的模型对72.1%的合成病例做出了正确分诊，并在56.1%的真实世界病例中与现有算法意见一致。这些模型在尽早识别急症方面更有前景：仅根据人口统计学信息、主诉、风险因素和危险信号问题的回答，最佳模型的AUROC达到0.87。在5%的误差范围内，该模型能够对约半数用户判断是否存在急症。因此，决策模型不应取代分诊引擎，但在针对具体场景设定的误差范围内，或可缩短症状评估所需时间。

引言

判断是否需要就医以及应多紧急地就医，是最常见的健康决策之一[1,2]，普通民众越来越多地借助大语言模型（LLM）做出这类决定[3,4]。一项美国调查显示，使用 ChatGPT 获取健康信息的人中，约有一半是为了决定是否就医3</style}；而在澳大利亚，超过三分之一的 ChatGPT 健康服务用户会借助它评估自己的症状[4]。由于这类建议可能直接影响人们是否就医以及去哪里就医，错误可能造成两方面的危害：建议用户在急症情况下进行自我护理，会直接危及他们的安全（“分诊不足”）；而建议几乎所有人都去就医（“过度分诊”），则可能导致有限资源被过度使用，损害其他患者的利益[5]。

鉴于这些风险，《欧盟医疗器械法规》（MDR）将支持分诊和诊断决策的软件归类为至少IIa类医疗器械，要求制造商在产品上市前证明其安全性和性能[6]。专门的症状评估应用程序（SAA）通常采用透明的算法，例如贝叶斯网络或基于规则的系统[7–9]，并据此获得认证。为普通民众提供医疗建议的大语言模型在技术上可能也适用相同规定，但往往无需医疗器械认证即可使用[10,11]。若仅以准确率衡量，大语言模型的表现并不逊于经认证的症状评估应用程序：其准确率约为58%–76%，处于症状评估应用程序准确率的宽泛范围（12%–90%）之内[5,12]。然而，与往往会过度分诊的症状评估应用程序不同[5,7,12]，大语言模型似乎会犯上述两类错误：在近期两项评估中，它们很少建议自我护理，对超过一半的真实紧急情况则分诊不足[13,14]。

即使这些错误得以减少，批准大语言模型仍会面临两大难题。首先，它们生成的是自由文本，而非规定的某一分诊级别，例如德国《医疗保健数据与数字创新法案草案》（GeDIG）[15]中规定的级别，因此在使用或评估其建议之前，必须先将其映射到这些级别。其次，它们具有非确定性，也就是说，相同的输入可能产生不同的输出。ChatGPT 对相同病例进行重复分诊时，即使提示词完全相同，结果也会有很大差异，有时甚至会针对同一病例建议所有可能的护理级别[13,16]。这与 MDR 的要求相冲突：软件必须经过设计，以确保重复性和可靠性[6]；同时也使评估更加复杂，因为单次运行无法说明模型下次面对同一病例时会如何响应。因此，如果某个系统既像大语言模型一样灵活（例如接受自由文本输入），又能返回规定的某一分诊级别，并对相同病例作出一致响应，就会更容易集成到经认证的软件中，也更便于评估。

决策模型（也称为“系统一模型”），例如 Jev——一种于2026年9月发布的新型模型——可能具备这一组合。它们不生成文本，而是从预先定义的多个回答选项中选择一个，为每个选项返回一个概率，并且通常能在不到一秒的时间内作出响应[17]。其优势之一是，相较于大语言模型，它们给出的建议变化较少[18,19]。由于有多种决策模型采用开放权重，其中一些仅包含数十亿个参数，制造商还可以在本地运行这些模型，而无需将健康数据传输给外部提供商[20,21]。固定的选项和概率也为症状评估带来了新的设计可能：可以从自由文本中提取症状、缩短症状评估流程、通过及早识别紧急情况来提高安全性，还可以利用概率阈值，将建议调整到可接受的误差范围内。

然而，决策模型能否提供准确的分诊建议仍不得而知。在医学领域的早期评估中，其表现取决于任务类型：尽管在研究问题上可与前沿大语言模型匹敌，但在复杂诊断病例[17]和数值任务上却处于劣势[18]。因此，其他（医学）任务的结果不能直接用于评估其分诊准确性。

为弥补这一知识空白，我们开展了一项探索性诊断准确性研究。我们的主要研究问题是决策模型的分诊建议有多准确。为评估这一问题，我们评估了 14 个开放权重模型和两个专有

决策模型的分诊准确性，所用数据包括来自各种既往评估的合成病例[9,22–33]，以及来自德国普通民众使用的一款SAA的真实世界数据。我们重点关注开放权重模型，因为它们可以在本地运行，并纳入了两个专有模型作为对照。

除主要研究问题外，即使决策模型对整体分诊决策的准确率不足，它们仍可能有助于解决特定子决策中的两个常见症状评估问题：（i）建议不准确且往往过于谨慎；（ii）评估过程过长。每个问题都引出了一个探索性研究问题。首先，错误可能部分源于需要一次性选择多个分诊级别之一；这比回答诸如某病例是否属于急症这样的二元问题更为复杂。因此，既往研究将分诊决策划分为一系列二元问题，并分别评估每个问题的准确率[25,34]。由于决策模型会从预定义的选项中进行选择，因此可以通过这两种方式向其提出分诊任务。由此，我们的第一个探索性研究问题是：与直接进行分诊决策相比，由三个二元问题组成的决策树是否会改变准确率？作为该问题的一部分，我们还评估了所有模型在每个子决策上的敏感度和特异度，以确定分诊过度和分诊不足发生在哪个环节。其次，症状评估应用通常会在给出建议前提出许多问题，这会使评估过程变得冗长[35]，也意味着处于急症状态的部分用户可能要到最后才能获得建议。如果决策模型能够及早识别急症，就可以为这些用户缩短评估过程。这引出了我们的第二个探索性研究问题：决策模型能否仅依据人口统计信息、主诉、风险因素和危险信号问题识别急症？在可接受的误差范围内，有多少用户可以据此及早作出这一决策？因此，在这一探索性设计中，我们不仅比较了不同模型，还比较了提出分诊任务的不同方式：涵盖所有分诊级别的直接提问、由二元问题组成的决策树，以及用于早期急症识别的精简输入。

方法

研究设计

本研究设计为一项针对决策模型的回顾性诊断准确性研究。我们使用233个合成病例和10,000个真实世界病例，对16个模型进行了评估。主要结局为分诊准确性（分为3至4个分诊级别），次要结局包括：（a）分诊过度和分诊不足的程度；（b）敏感度；以及（c）特异度。由于数据源之间的分诊级别存在差异，我们纳入了一项敏感性分析，依据SCARF清单，将具有四级分诊的病例映射至三级分诊级别系统[36]。对于真实世界病例，我们还将分诊决策拆分为二元问题，并测试能否根据精简后的输入识别紧急情况。所有分析均为探索性分析。结果依据STARD和SCARF清单进行报告[36,37]。

病例与指标检测

合成病例

我们使用了此前系统综述中确定的五组已发表合成病例集[5,7,12,38]。Semigran等的第一组包含45个病例，采用三级分诊[22]。另外两组对原始病例描述进行了改编，增加了病例，并采用四级分诊（每组48个病例）[23,28]。这三组中的所有病例均予以保留，因为每个病例都根据其来源数据集的参考标准和分诊量表进行了评估。Delshad等的另一组病例与Semigran等的病例集无关，包含50个采用四级分诊的病例，其参考标准由医生多数投票确定[29]。其中有三个病例因投票未能得出明确的分诊级别而被排除，该病例集最终保留47个病例。我们还使用了Kopka等的另一组无关病例集，其中包含45个采用四级分诊的病例，这些病例基于真实患者的问题，并以通俗语言撰写[24]。其他所有病例集均以面向临床医生的方式表述。我们未排除非典型病例。

真实世界病例

我们从2026年1月1日至2026年10月20日期间在DOCYET SAA中完成的病例中随机抽取了10,000例。德国的普通民众使用该SAA进行就医决策；它已整合至多个医疗场景，包括健康保险机构、全科医生（GP）诊所和急诊科。我们没有刻意排除任何非典型病例。我们认为10,000例的样本量适当，因为它既能提供足够的精确度，又能使API调用次数保持在可行范围内。

参考标准

对于合成病例，我们采用了原始出版物中所分配的分诊级别。在所有病例集中，这些分诊级别均由至少两名执业医师进行分配。对于真实世界病例，我们采用 DOCYET SAA 在实际使用过程中前瞻性分配的分诊级别作为参考标准，因为让医生评估 10,000 个病例被认为不可行。因此，对于这些病例，准确率、敏感度和特异度所表示的是与 SAA 的一致性，而非临床正确性。我们区分了四级分诊：（1）急症；（2）紧急护理，定义为在 24 小时内咨询医生；（3）非急诊咨询；以及（4）自我护理。

对于合成病例和真实世界病例，我们采用了既往研究中的以下定义 [24,28]：

- 自我护理：让健康问题自行好转。这种情况很可能不需要就医。- 非急诊：尽量在一周内就医。问题可能不会自行好转，但也不属于急症。- 紧急：尽量在一天内就医。问题很紧急，但不属于急症。- 急症：拨打 112 或直接前往急诊室。问题需要立即处理。

由于各病例集的分诊级别数量不同，我们用三级分诊重新查询了所有模型，以便在同一尺度上进行比较 [36]。为此，我们将紧急和非急诊咨询合并为非急诊护理，并据此调整了参考标准。我们采用了以往研究 [13,28] 对这三个级别的定义：

- 自我护理：让健康问题自行好转。这种情况很可能不需要看医生。- 非急诊：尽量去看医生，但症状无需在急诊室立即处理。例如，可以在 24 小时内或下周去看医生。- 急症：拨打 112 或直接前往急诊室。该问题需要立即处理。

模型

为确定开放权重决策模型，我们于 2026年11月22日 搜索了 OpenRouter 模型目录，并纳入其中列出的所有开放权重决策模型。如果某个模型有多个版本，我们采用最新版本。此次搜索共纳入 14 个模型：clef、clef-flash、clef-omni、drex-max、bison-1、groundhog-1、elephant-1、joey-1、kev-4b、mammoth-1、pplx-decider-v1.3、pplx-decider-flash-v1.3、span-02 和实验性 tev1-4b。作为对照，我们还纳入了两个专有模型：用于决策的 gpt-6-luna 和 jev-1.14。

程序

程序如图1所示。第一步，所有模型均按照各自原有的分诊量表对每个病例进行分诊。作为敏感性分析，我们将分诊级别改为三级，并重复了这一过程。第二步，我们仅使用真实世界病例，并测试了一种决策树，其中包含

若干二元问题，以确定四级分诊，而不是通过一个包含全部四个分诊级别的问题来确定。最后，我们使用在急诊问题中表现最佳的三个模型，并测试它们是否可用于急诊预筛查——也就是说，在症状评估早期信息有限的情况下，它们能否准确判断是否需要急诊护理。



图1. 数据收集与分析流程

我们将每个合成病例以原始文本形式输入模型，并将真实世界病例整理为德语结构化JSON对象。每个JSON病例包含年龄和性别，随后是用户的主诉、危险信号问题、风险因素，以及有关用户症状的其他后续问题。对于每个问题，均列明用户是确认、否认，还是回答“不知道”（结构及虚构示例见附录A）。

所有模型均通过OpenRouter API由一个使用R（4.6.1版）编写的自定义脚本调用。由于脚本输入了所有病例，因此没有人工输入人员参与。提示词采用了此前用于评估常规大语言模型分诊准确性的指令[13,24]：“请为以下病例给出分诊分类”，并将上一节所述的分诊级别作为回答选项。模型的回答，包括预测的分诊级别和类别概率，均存储在数据框中。在收集数据之前，我们进行了数次预检验，以确保病例传输正确且数据完整。所有数据均于2026年11月22日收集。

第二步，我们将真实世界病例的分诊决策拆分为三个二元问题，见图2

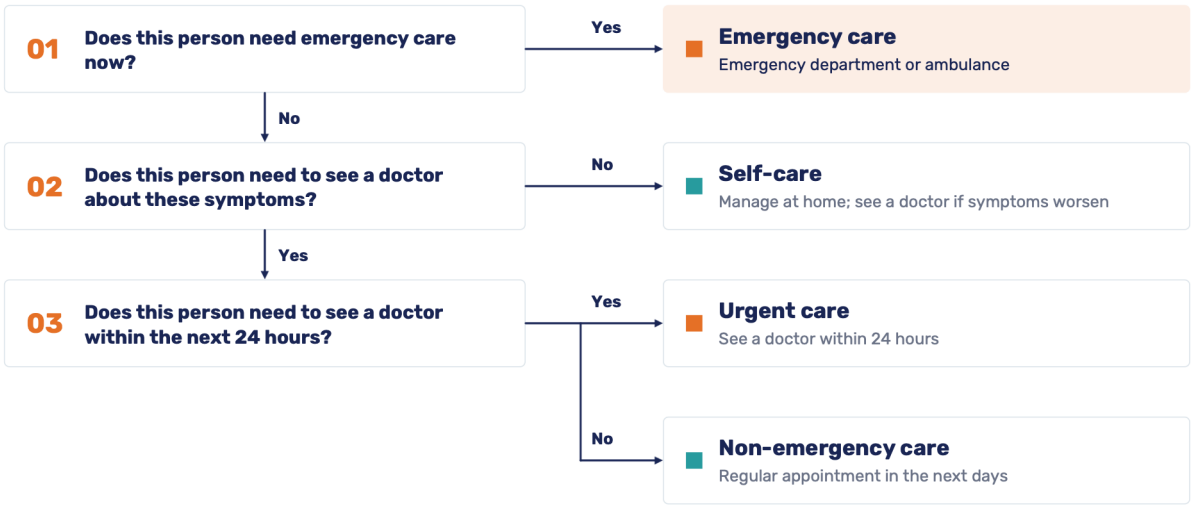


图2. 决策树

由此形成的决策树首先询问“此人现在是否需要急诊护理？”，随后询问“此人是否需要就这些症状就医，还是这些症状最有可能不需要就医？”，以及“此人是否需要在接下来的*24*小时内就医？”。每个问题都有两个回答选项，其中一个会升级护理级别。如果第一个问题的回答是升级护理级别的选项，则该病例被分诊为急症；如果第二个问题的回答不是该选项，则被分诊为自我护理。否则，第三个问题将决定是紧急护理还是非急诊护理。作为敏感性分析，我们再次使用前两个问题，将该决策树映射为三个分诊级别（急症、非急诊、自我护理）。

最后，我们探讨了决策模型能否在评估早期作出急症判定。为此，我们使用精简后的输入重新询问了三个开放权重模型：在第二步识别出的、针对二元急诊问题敏感度最高、特异度最高和平衡准确率最高的模型。精简后的输入仅包含后续问题之前收集的信息：年龄、性别、风险因素、主诉以及对危险信号问题的回答。

分析

由于这是一项诊断准确性研究，因此所有数据均采用描述性分析，并辅以置信区间（CI）作为推断性分析。

首先，我们计算了模型分诊级别与参考标准一致的病例比例，以此作为总体准确率，并分别计算各参考级别的准确率 [39]。为评估错误的方向，我们计算了所有错误中属于过度分诊的比例，并展示模型输出与参考标准的混淆矩阵。

其次，对于决策树，我们计算了每个二元问题的准确率、敏感度和特异度，并将其准确率与同一批病例中直接提问的准确率进行了比较。

最后，对于急症识别，我们计算了每个模型自身决策的敏感度、特异度和平衡准确率，以及基于急症预测概率计算的 AUROC。接下来，我们用概率阈值取代模型的决策，将低于某一阈值的病例归为非急症，将高于该阈值的病例归为急症。为考察阈值的选择如何影响分诊错误，我们绘制了每个模型的 ROC曲线，并展示了从 0 到 100% 的所有阈值下的两种错误类型：漏诊的紧急情况所占比例

（1 – 敏感度，即分诊不足）以及被归为急症的非急症所占比例（1 – 特异度，即过度分诊）。通过这一可视化，可以确定调整阈值如何以增加一种错误为代价来减少另一种错误。

由于单一截断值无法同时将两类错误都控制在较低水平，因此我们采用两个截断值，将病例分为三组：低于排除阈值（例如 1%）的病例归类为无急症，高于确诊阈值的病例归类为急症，介于两者之间的病例则归为不确定。我们设定排除阈值，使漏诊的急症比例至多为 1%、5% 或 10%；同时设定确诊阈值，使被归类为急症的非急症病例比例与之相同。高于确诊阈值的病例会立即被转至急诊护理，评估也会随之结束。低于排除阈值的病例，其急症问题虽已判定，但仍会继续回答其余问题，以决定采取自我护理还是非急诊护理。在只需作出急症判定的情境下，例如从急诊科转诊，这些病例的评估也可以结束。因此，本分析展示了在可接受的误差范围内，有多少用户的评估可以缩短。由于不同样本可能得出不同的概率阈值，我们采用 10 折交叉验证评估这两个阈值：在九折数据上确定阈值，再应用于留出折，并汇总所有留出病例的比例。交叉验证比例的置信区间采用 Clopper-Pearson方法计算。

结果

常规分诊决策

合成病例

我们使用来自五个病例集的 233 个合成病例（其中四个病例集采用四级分诊，另一个采用三级分诊），评估了 16 个决策模型，共进行了 3,728 次评估。

当每个病例均按照其原有的分诊量表（即包含三个或四个选项）进行分诊时，开放权重模型 mammoth- 1 的准确率最高，正确分诊了所有病例的 72.1%（168/233；95%置信区间 65.9–77.8%）；其次是表现最佳的专有模型 gpt-6-luna-decisions（71.2%（166/233；95%置信区间 65.0–77.0%））。tev1-4b- experimental 的准确率最低（52.8%（123/233；95%置信区间 46.2–59.3%））。各模型在每个病例集上的准确率详见附录表 S1。

急症准确率最高的是 span-02（91.8%（56/61；95%置信区间 81.9–97.3%））；紧急病例准确率最高的是 clef-flash（82.5%（33/40；95%置信区间 67.2–92.7%））；非紧急病例准确率最高的是 bison-1（88.5%（69/78；95%置信区间 79.2–94.6%）；自我护理病例准确率最高的是 mammoth-1（61.1%（33/54；95%置信区间 46.9–74.1%））。所有模型均倾向于过度分诊，但在所有错误中，过度分诊所占比例从 drex-max 的 52.9%（37/70；95%置信区间 40.6–64.9%）到 span-02 的 82.7%（67/81；95%置信区间 72.7–90.2%）不等，见表 2。最常见的错误分类是将自我护理病例分诊为非急诊（占所有自我护理评估的 49.3%；426/864），且没有任何模型曾将急症病例分诊为自我护理，见附录图 S1–S5。

在敏感性分析中，当所有病例均采用相同的三种选项（自我护理、非急诊、急症）进行分诊时，准确率从 tev1-4b-experimental 的 60.5%（141/233；95%置信区间 53.9–66.8%）到 mammoth-1 的 80.3%（187/233；95%置信区间 74.6–85.2%）不等。再次强调，模型倾向于过度分诊（过度分诊占所有错误的比例为 60.9–91.4%），见附录表 S2。将自我护理病例分诊为非急诊仍是最常见的错误分类（68.9%；595/864），见附录中的图 S6–S9。

表 2. 各模型在所有合成病例中的分诊准确性（n = 233）

模型	总体	急症	紧急	非急诊	自我护理	过度分诊错误占比错误
mammoth-1	72.1%（168/233）	80.3%（49/61）	52.5%（21/40）	83.3%（65/78）	61.1%（33/54）	58.5% (38/65)
gpt-6-luna-决策	71.2%（166/233）	90.2%（55/61）	55.0%（22/40）	78.2%（61/78）	51.9%（28/54）	65.7% (44/67)
bison-1	70.8%（165/233）	72.1%（44/61）	75.0%（30/40）	88.5%（69/78）	40.7%（22/54）	60.3% (41/68)
drex-max	70.0%（163/233）	75.4%（46/61）	60.0%（24/40）	83.3%（65/78）	51.9%（28/54）	52.9% (37/70)
pplx-decider-v1.3	69.5%（162/233）	72.1%（44/61）	75.0%（30/40）	84.6%（66/78）	40.7%（22/54）	60.6% (43/71)
groundhog-1	67.4%（157/233）	77.0%（47/61）	67.5%（27/40）	80.8%（63/78）	37.0%（20/54）	61.8% (47/76)
jev-1.14	65.7%（153/233）	83.6%（51/61）	67.5%（27/40）	80.8%（63/78）	22.2%（12/54）	80.0% (64/80)
span-02	65.2%（152/233）	91.8%（56/61）	62.5%（25/40）	78.2%（61/78）	18.5%（10/54）	82.7% (67/81)
clef	64.8%（151/233）	75.4%（46/61）	72.5%（29/40）	71.8%（56/78）	37.0%（20/54）	63.4% (52/82)
pplx-decider-flash-v1.3	63.9%（149/233）	80.3%（49/61）	60.0%（24/40）	74.4%（58/78）	33.3%（18/54）	66.7% (56/84)
clef-flash	59.2%（138/233）	78.7%（48/61）	82.5%（33/40）	67.9%（53/78）	7.4%（4/54）	81.1%（77/95）
kev-4b	58.8%（137/233）	75.4%（46/61）	75.0%（30/40）	67.9%（53/78）	14.8%（8/54）	75.0% (72/96)
elephant-1	58.8%（137/233）	70.5%（43/61）	67.5%（27/40）	61.5%（48/78）	35.2%（19/54）	62.5% (60/96)
clef-omni	57.5%（134/233）	67.2%（41/61）	72.5%（29/40）	70.5%（55/78）	16.7%（9/54）	72.7% (72/99)
joey-1	54.1%（126/233）	86.9%（53/61）	12.5%（5/40）	62.8%（49/78）	35.2%（19/54）	65.4% (70/107)
tev1-4b-实验性	52.8%（123/233）	82.0%（50/61）	60.0%（24/40）	53.8%（42/78）	13.0%（7/54）	81.8% (90/110)

真实世界病例

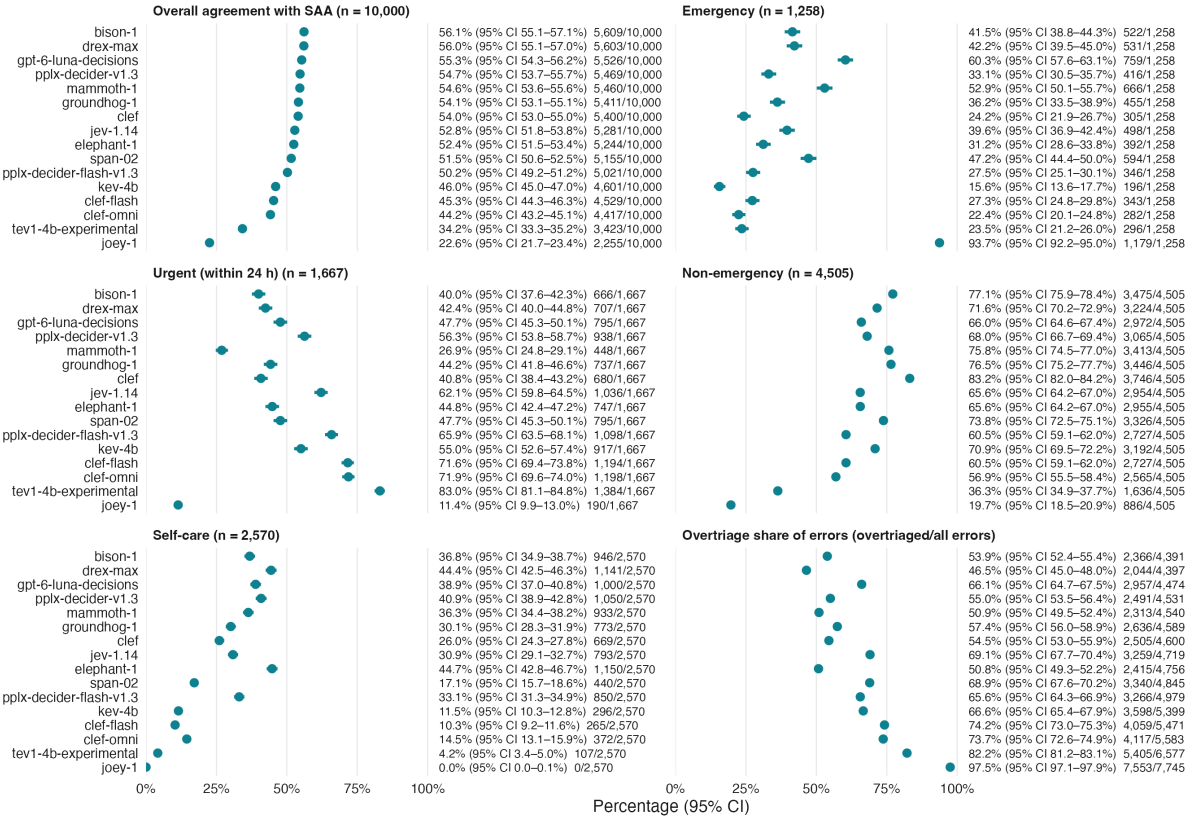
每个模型均对全部 10,000 个真实世界病例进行了分诊，共产生 160,000 次评估。

使用四级分诊对真实世界病例进行分诊时，开放权重模型 bison-1 与 SAA参考标准的一致性最高，其分诊级别匹配率为 56.1%（5,609/10,000；95%置信区间 55.1–57.1%）；相比之下，表现最佳的专有模型 gpt-6-luna-decisions 的匹配率为 55.3%（5,526/10,000；95%置信区间 54.3–56.2%）。joey-1 的一致性最低，为 22.6%（2,255/10,000；95%置信区间 21.7–23.4%）。

在紧急案例中，与SAA一致性最高的是joey-1（93.7%（1,179/1,258；95%置信区间92.2–95.0%））；在紧急（24小时内）病例中，最高的是tev1-4b-

experimental（83.0%（1,384/1,667；95%置信区间81.1–84.8%））；在非紧急病例中，最高的是clef（83.2%（3,746/4,505；95%置信区间82.0–84.2%））；在自我护理病例中，最高的是elephant-1（44.7%（1,150/2,570；95%置信区间42.8–46.7%））。16个模型中的15个倾向于过度分诊，而不是分诊不足；在所有错误中，过度分诊所占比例从drex-max的46.5%（2,044/4,397；95%置信区间45.0–48.0%）到joey-1的97.5%（7,553/7,745；95%置信区间97.1–97.9%）不等，见图3。最常见的错误分类是将自我护理病例分诊为非急诊（占所有自我护理评估的51.6%；21,236/41,120），最少见的错误分类是将紧急案例分诊为自我护理（占所有急症评估的0.6%；126/20,128），见图4。

在敏感性分析中，当使用三种选项（自我护理、非急诊、急症）重新询问模型时，与SAA的一致率从joey-1的28.1%（2,805/10,000；95%置信区间27.2–28.9%）到pplx-decider-v1.3的70.3%（7,033/10,000；95%置信区间69.4–71.2%）不等。再次强调，所有模型都倾向于过度分诊（在所有错误中，过度分诊所占比例为54.0%至99.0%），见附录表S3。被分诊为非急诊的自我护理病例仍是最常见的错误分类（73.8%；30,359/41,120），见图S10。



注释。点表示与 SAA 的一致性以及错误中过度分诊所占比例，线条表示 95% 置信区间。

图3。各模型与 SAA 的一致性（真实世界病例）

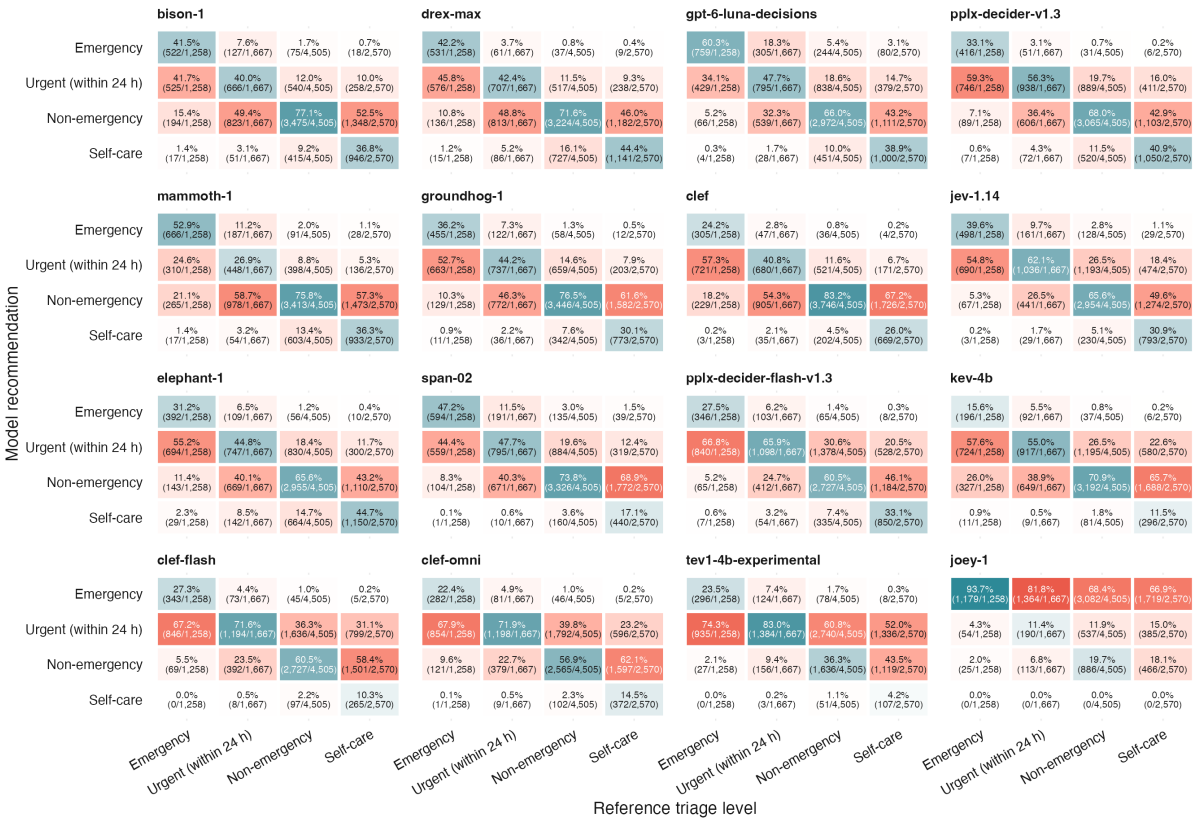


图4. 各模型的混淆矩阵（真实世界病例）

真实世界病例中的二元分诊决策

除gpt-6-luna-的决策外，每个模型都为全部10,000个真实世界病例返回了二元分诊决策；gpt-6-luna-反复拒绝为其中一个病例提供决策。

当将分诊决策拆分为包含三个二元问题的决策树时，groundhog-1 与 SAA 的一致性最高，其分诊级别有56.5%的病例与SAA一致（5,653/10,000；95%置信区间55.6–57.5%）；而 joey-1 的一致性最低（18.1%（1,807/10,000；95%置信区间17.3–18.8%））。与对相同病例直接提出四选一问题相比，决策树使16个模型中的5个模型的一致性更高（groundhog-1提高+2.4个百分点（95%置信区间+1.6至+3.2），kev-4b提高+5.8个百分点（95%置信区间+4.7至+6.8）），使7个模型的一致性更低（clef降低1.8个百分点（95%置信区间-2.9至-0.7），clef-omni降低9.7个百分点（95%置信区间-10.6至-8.8）），并且对4个模型没有影响（drex-max、bison-1、pplx-decider-v1.3和jev-1.14），见表3。当仅使用前两个问题、由此产生三级分诊（自我护理、非急诊、急症）时，groundhog-1与SAA的一致性最高，其分诊级别有69.2%的病例与SAA一致（6,915/10,000；95%置信区间68.2–70.1%）；而 joey-1 的一致性最低（23.8%（2,380/10,000；95%置信区间23.0–24.6%））。与对相同病例直接提出三选一问题相比，决策树使16个模型中的5个模型与SAA的一致性更高，使7个模型的一致性更低，并且对4个模型没有影响，见附录表S4。

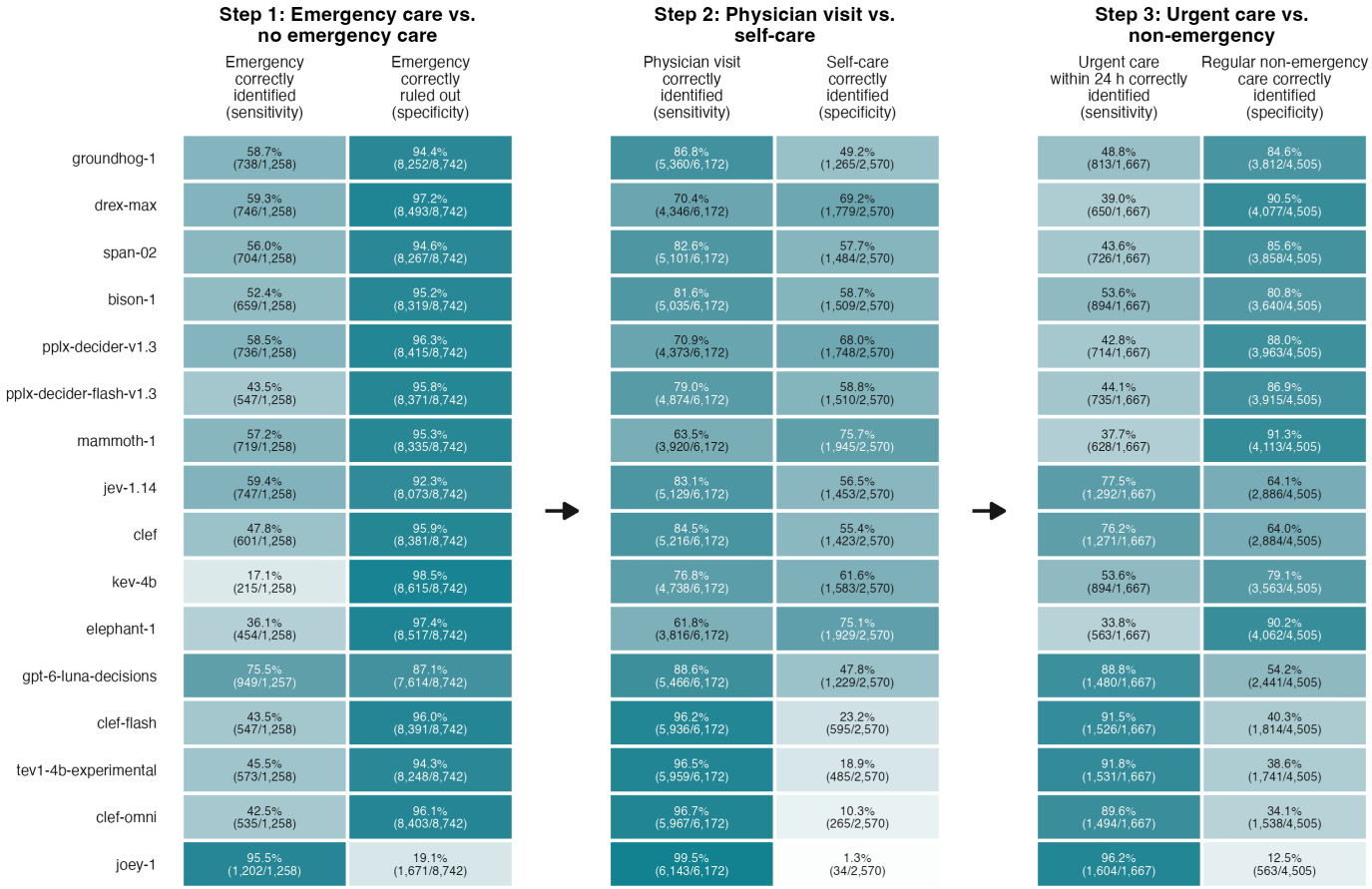
在评估紧急情况时，敏感度方面，kev-4b 为 17.1%（215/1,258），joey-1 为 95.5%（1,202/1,258）；特异度方面，joey-1 为 19.1%（1,671/8,742），kev-4b 为 98.5%（8,615/8,742）。在随后决定就诊还是自我护理时，敏感度方面，elephant-1 为 61.8%（3,816/6,172），joey-1 为 99.5%（6,143/6,172）；特异度方面，joey-1 为 1.3%（34/2,570），mammoth-1 为 75.7%（1,945/2,570）。在第三步，即决定选择紧急护理还是常规非紧急护理时，敏感度范围

从 elephant-1 的 33.8%（563/1,667）到 joey-1 的 96.2%（1,604/1,667），特异度则从 joey-1 的 12.5%（563/4,505）到 mammoth-1 的 91.3%（4,113/4,505），见图5。

表3。决策树与SAA的一致性

模型	决策树 %（95%置信区间） [一致/n]	急症	紧急	非急诊	自我护理	直接提问（4 个选项） %（95%置信区间） [同意/n]	差异（百分点；95%置信区间）
groundhog-1	56.5%（55.6–57.5%） [5,653/10,000]	58.7% [738/1,258]	32.0% [533/1,667]	69.2% [3,117/4,505]	49.2% [1,265/2,570]	54.1%（53.1–55.1%） [5,411/10,000]	+2.4 （+1.6 至 +3.2）
drex-max	56.1%（55.1–57.1%） [5,611/10,000]	59.3% [746/1,258]	29.5% [492/1,667]	57.6% [2,594/4,505]	69.2% [1,779/2,570]	56.0%（55.1–57.0%） [5,603/10,000]	+0.1 （-0.8 至 +1.0）
span-02	55.9%（54.9–56.8%） [5,585/10,000]	56.0% [704/1,258]	27.7% [462/1,667]	65.3% [2,942/4,505]	57.5% [1,477/2,570]	51.5%（50.6–52.5%） [5,155/10,000]	+4.3 (+3.2 至 +5.4)
bison-1	55.2%（54.2–56.2%） [5,519/10,000]	52.4% [659/1,258]	37.4% [624/1,667]	60.6% [2,728/4,505]	58.7% [1,508/2,570]	56.1%（55.1–57.1%） [5,609/10,000]	-0.9 (-1.8 至 0.0)
pplx-decider-v1.3	55.0%（54.1–56.0%） [5,504/10,000]	58.5% [736/1,258]	31.0% [517/1,667]	55.6% [2,503/4,505]	68.0% [1,748/2,570]	54.7%（53.7–55.7%） [5,469/10,000]	+0.4 （-0.7 至 +1.4）
pplx-decider-flash-v1.3	54.3%（53.4–55.3%） [5,434/10,000]	43.5% [547/1,258]	31.1% [519/1,667]	63.4% [2,858/4,505]	58.8% [1,510/2,570]	50.2%（49.2–51.2%） [5,021/10,000]	+4.1 (+3.1 至 +5.2)
mammoth-1	52.7% (51.8–53.7%) [5,274/10,000]	57.2% [719/1,258]	22.3% [371/1,667]	49.7% [2,239/4,505]	75.7% [1,945/2,570]	54.6%（53.6–55.6%） [5,460/10,000]	-1.9 （-2.9 至 -0.8）
jev-1.14	52.5% (51.5–53.5%) [5,247/10,000]	59.4% [747/1,258]	56.6% [944/1,667]	46.7% [2,104/4,505]	56.5% [1,452/2,570]	52.8%（51.8–53.8%） [5,281/10,000]	-0.3 （-1.2 至 +0.6）
clef	52.2%（51.2–53.2%） [5,219/10,000]	47.8% [601/1,258]	64.4% [1,074/1,667]	47.1% [2,121/4,505]	55.4% [1,423/2,570]	54.0%（53.0–55.0%） [5,400/10,000]	-1.8 (-2.9 至 -0.7)
kev-4b	51.8% (50.8–52.8%) [5,179/10,000]	17.1% [215/1,258]	42.1% [702/1,667]	59.5% [2,679/4,505]	61.6% [1,583/2,570]	46.0%（45.0–47.0%） [4,601/10,000]	+5.8 （+4.7 至 +6.8）
elephant-1	50.1%（49.2–51.1%） [5,014/10,000]	36.1% [454/1,258]	25.3% [422/1,667]	49.0% [2,209/4,505]	75.1% [1,929/2,570]	52.4% (51.5–53.4%) [5,244/10,000]	-2.3 （-3.3 至 -1.3）
gpt-6-luna-decisions	49.7%（48.8–50.7%） [4,973/9,999]	75.5% [949/1,257]	58.8% [980/1,667]	40.3% [1,816/4,505]	47.8% [1,228/2,570]	55.3%（54.3–56.2%） [5,525/9,999]	-5.5 (-6.4 至 -4.6)
clef-flash	40.9%（40.0–41.9%） [4,093/10,000]	43.5% [547/1,258]	78.9% [1,316/1,667]	36.3% [1,635/4,505]	23.2% [595/2,570]	45.3%（44.3–46.3%） [4,529/10,000]	-4.4 （-5.3 至 -3.5）

模型	决策树 %（95%置信区间） [同意/n]	急症	紧急	非急诊	自我护理	直接提问 （4 个选项） %（95%置信区间） [同意/n]	差异 （百分点； 95%置信区间）
tev1-4b-experimental	38.7%（37.7–39.7%） [3,870/10,000]	45.5% [573/1,258]	74.4% [1,241/1,667]	34.9% [1,571/4,505]	18.9% [485/2,570]	34.2%（33.3–35.2%） [3,423/10,000]	+4.5 （+3.7 至 +5.3）
clef-omni	34.5%（33.5–35.4%） [3,446/10,000]	42.5% [535/1,258]	77.4% [1,291/1,667]	30.1% [1,355/4,505]	10.3% [265/2,570]	44.2%（43.2–45.1%） [4,417/10,000]	-9.7 （-10.6 至 -8.8）
joey-1	18.1%（17.3–18.8%） [1,807/10,000]	95.5% [1,202/1,258]	7.6% [127/1,667]	9.9% [446/4,505]	1.2% [32/2,570]	22.6%（21.7–23.4%） [2,255/10,000]	-4.5 （-5.0 至 -4.0）



Fi图 5。决策树中每个二元问题的敏感度和特异度

利用危险信号和主诉进行急症识别

仅提供人口统计信息、主诉、危险信号问题的回答和风险因素时，joey-1识别出95.7%的急症（1,204/1,258；95%置信区间94.4–96.8%），并排除其他病例中17.2%的急症（1,501/8,742；95%置信区间16.4–18.0%）（平衡准确率为56.4%（95%置信区间55.8–57.1%））；kev-4b识别出14.0%的急症（176/1,258；95%置信区间12.1–16.0%），并排除其他病例中98.1%的急症（8,580/8,742；95%置信区间97.8–98.4%）（平衡准确率为56.1%（95%置信区间55.1–57.0%））；drex-max识别出49.7%的急症（625/1,258；95%置信区间46.9–52.5%），并排除其他病例中95.7%的急症（8,369/8,742；95%置信区间95.3–96.1%）（平衡准确率为72.7%（95%置信区间71.3–74.1%））。joey-1的AUROC为0.698（95%置信区间0.683–0.713），kev-4b的AUROC为0.772（95%置信区间0.757–0.786），drex-max的AUROC为0.874（95%置信区间0.863–0.884），见图6A。

当使用预测概率并设置排除阈值和确诊阈值时，留出病例中的错误率接近目标值1%、5%和10%（漏诊的紧急情况分别为0.9–1.0%、4.8–5.0%和9.6–10.1%；被标记为急症的非急症分别为1.0%、4.9–5.0%和9.9–10.0%）。在这些误差范围内，对于所有病例，使用joey-1时，6.0%、24.8%和38.8%的病例可在回答危险信号问题后作出急症判定；使用kev-4b时，这一比例分别为8.9%、28.2%和49.3%；使用drex-max时，则分别为20.5%、51.8%和71.5%。评估还可分别在1.7%、6.7%和12.3%；2.1%、7.8%和14.1%；以及4.1%、10.9%和17.0%的病例中以急症转诊结束，见图6C（所有95%置信区间均在±1.0个百分点以内；阈值见图6B和表S5）。

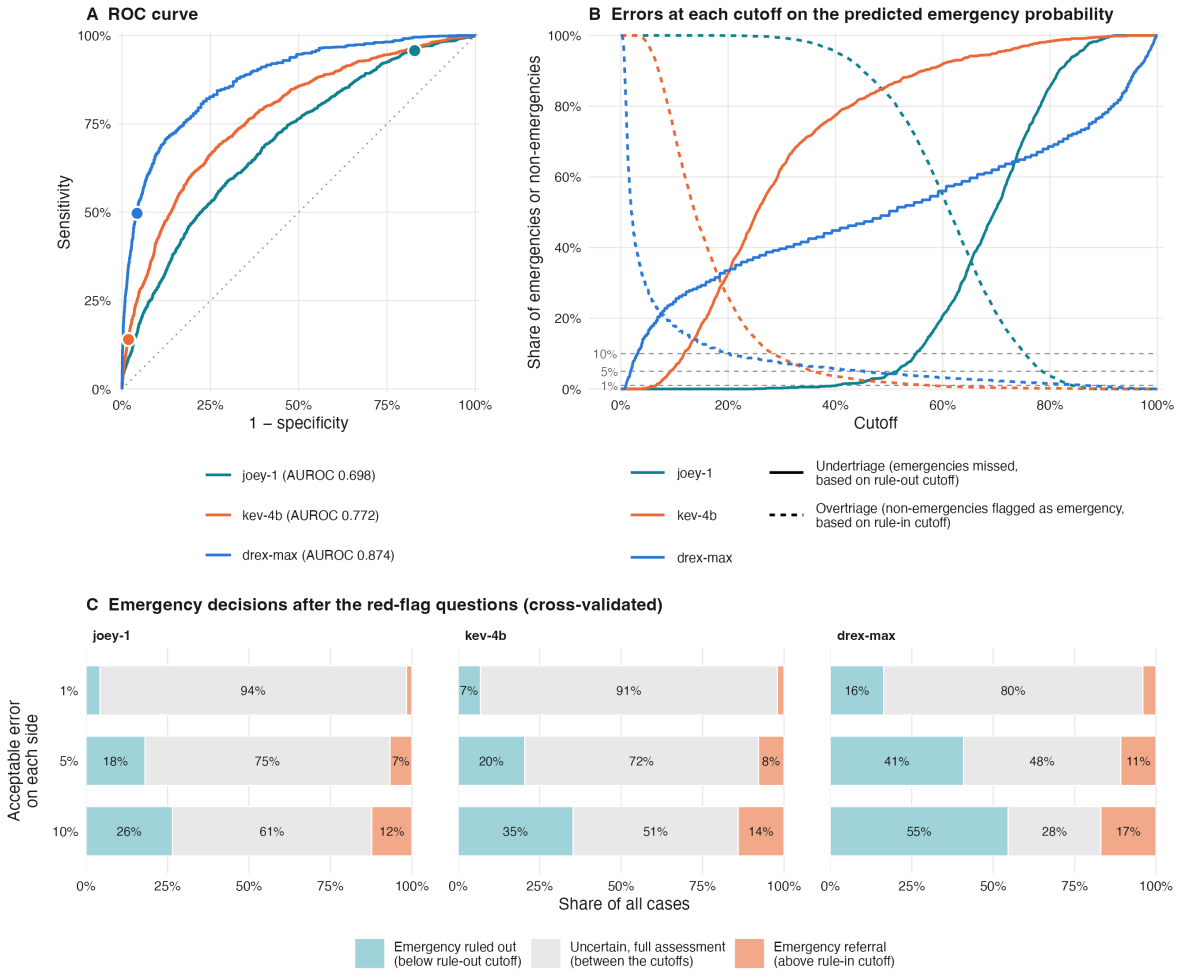


图 6. 基于主诉、危险信号和风险因素的急诊决策

讨论

据我们所知，这是首项评估决策模型分诊准确性的研究。我们发现，目前决策模型的准确性还不足以取代症状评估应用中现有的分诊引擎，但在一项特定子决策上表现出潜力：在评估早期识别急症，且误差范围可预先设定。总体而言，表现最佳的开放权重模型在四级分诊中的表现与最佳专有模型相当：正确分诊了72%的合成病例，并在56%的真实世界病例中与症状评估应用的判断一致（最佳专有模型的相应比例为71%和55%）；在三级分诊中则优于该模型（按相同顺序分别为80%对73%，以及70%对67%）。大多数错误属于过度分诊错误，而几乎从未（0.6%）建议急症病例进行自我护理。将分诊决策拆分为二元问题，也未能持续提高与症状评估应用判断的一致性：七个模型的一致性有所提高，五个则有所降低。能否缩短评估取决于如何使用模型的输出。使用精简输入时，模型自身的决策要么在适用情况下更善于排除急症，但总体上识别出的急症不足一半；要么像joey-1那样，几乎识别出所有急症，但代价是仅排除极少数非急症。然而，由于模型会为每个选项给出一个概率，因此可以预先设定可接受的误差范围，分别控制漏诊的紧急情况和误报。这种方法已被证明可行：在设定此类误差范围后，对于相当一部分用户，模型可在回答危险信号问题后判定其是否需要急诊护理；情况不明确者则继续完成全面评估。下文将讨论最重要的研究发现。

结合本研究的发现，决策模型以目前的形式不应作为独立分诊工具使用。在合成病例中，表现最好的模型正确分诊了72.1%的病例，这一结果处于类似病例中大语言模型（58–76%）[12]和症状评估应用（12–90%）[5,12]所报告的范围内。相反，在真实世界病例中，与症状评估应用的一致性更低，而且几乎没有模型能够识别出超过60%的紧急情况——joey-1除外，但它为此付出了总体一致性极低的代价。因此，决策模型在分诊决策方面似乎并不比大语言模型更准确，而公众已经在使用大语言模型做出就医决策。不过，决策模型在现有软件中的部署具有实际优势，因为它们不会生成输出词元[17]：如近期一项基准研究所示，它们生成响应的速度显著更快，每项决策约需0.3秒，而前沿大语言模型需要1.2–3.8秒，成本则高出50至80倍[17]。此外，由于一些开放权重决策模型的规模小得多，参数量仅为20亿或40亿，因此可以在性能较弱的设备上以更低成本本地运行。因此，它们的优势可能不在于取代症状评估应用中的分诊引擎，而在于支持其中的特定子决策，例如及早识别紧急情况以缩短评估时间；它们可以承担原本分配给大语言模型的任务，而成本仅为其一小部分。

最重要的实践意义在于，决策模型能够在评估早期作出急症判定（作为一项子决策），并且可以针对每种医疗场景预先设定误差范围。当仅提供人口统计信息、主诉、风险因素和危险信号问题时，在误差范围分别为 1%、5% 和 10% 的情况下，AUROC 最高的模型可以为约五分之一、一半以及超过三分之二的所有用户作出这一判定；其余用户则继续进行完整评估。由于决策模型会为每个回答选项返回一个概率，因此可以分别设定可接受的漏诊紧急情况比例和可接受的误报比例：前者通过排除阈值（敏感度）设定，后者通过确诊阈值（特异度）设定。大多数症状评估应用和大语言模型只会返回单一建议，因此其误差范围虽可测量，却很难针对具体医疗场景进行调整。

可接受的误差范围取决于医疗场景。这开启了一条更偏向临床和监管、而非技术层面的研究路径：应探讨在何种医疗接触点需要多高的准确率，以及可接受多大风险，例如健康保险服务、全科医生诊所或急诊科。举例来说，在急诊科的自助评估终端中，

首要目标应是确保没有任何真正的急症患者被拒之门外，而误报的代价很小：未被排除的用户会留在急诊科，并像往常一样由工作人员进行分诊。因此，只有排除阈值至关重要；该阈值可以设定为仅允许极少比例的紧急情况漏诊（即非常高的敏感度，例如 99%）。在此阈值下的特异度决定了可以为多少非急症患者提供门诊护理。即使如此严格的阈值所对应的特异度不高，每位被分流的患者都能在不增加额外风险的情况下节省资源；由于近一半的急诊科患者是在没有转诊的情况下就诊，其中一些人甚至可能不需要急诊护理，因此节省的资源可能相当可观[40,41]。现有的分流解决方案，例如急诊科中的数字服务终端，迄今显示出较低的分诊不足率，但过度分诊率较高（59– 65%），且尚未证实具有运营效益[42]。将排除阈值设定为明确的误差范围，可以使这种权衡更加清晰，也可能更有助于实现运营效益。在另一种场景中，日常使用的常规症状评估应用中的决策模型则可以作为第二意见，与基于规则的急症退出机制配合使用；当出现特定的危险信号和风险因素组合时，这些机制会终止评估。具有预定义误差范围的模型可能会标记出这些规则遗漏的急症，并对不必要触发的退出提出质疑，从而减少分诊不足和过度分诊。然而，如何最好地整合这种第二意见，取决于每个系统的评估逻辑。

另一个影响涉及如何提出分诊问题：只有当各项决策能够清晰区分时，将分诊决策拆分为二元问题，且仅对部分模型而言，似乎才能提升性能。在四级和三级分诊中，决策树都提高了部分模型与参考标准的一致性，却降低了其他模型的一致性。这在一定程度上与有关大语言模型的研究结果一致：将复杂任务分解为更简单的子任务可提升部分模型的性能[43]；同时也与以下发现相符：当问题形式发生改变时，决策模型给出的建议也会随之改变[18]。尽管在三级和四级分诊中，性能提升或下降的模型数量相同，但模型本身的性能仍存在部分差异，这可能是因为第三个问题区分了 24 小时内接受护理与接受常规护理。各模型针对该问题的敏感度和特异度差异很大，这表明，相较于是否需要急诊护理，或是否需要就医这一决策，该决策的界限较不明确。此外，错误会在决策树中累积，因为只有在所有先前问题均得到正确回答时，一个病例才能被正确分诊。哪些分解策略能够提升哪些模型的准确率，仍有待进一步研究。

未来研究可从三个方向拓展我们的研究结果。首先，我们仅评估了未经任务特定适配的通用决策模型。由于决策模型是一类新兴模型，未来版本很可能会有所改进；而且在分诊数据上进行微调或许能显著提升结果，因此，应使用更新的模型和经过微调的模型，重新评估本文考察的子决策。其次，除了本文考察的应用场景，决策模型还可能以其他多种方式支持症状评估：与大多数症状评估应用一样，我们的评估重点关注护理需求，即根据患者症状判断其所需的护理级别；这与患者需求不同，后者指用户实际希望通过评估获得什么[44,45]。决策模型可以将自由文本输入归类为此类患者需求，例如判断用户是想获得就医建议，还是只需要病假证明或处方，而这些情况可能并不需要完整的症状评估。在此类情况下，可以将 ICPC-3[46] 等针对这类就诊制定的成熟分类用作回答选项。同样，决策模型可以从自由文本输入中提取症状，并将其归入预定义类别，从而缩短症状评估时间。还可以测试评估中的其他简化方式，例如，当增加回答后分诊级别的概率不再变化时即结束评估，或尽早将用户引导至药房、远程医疗等特定服务。最后，由于决策模型会返回概率，因此研究用户如何理解每个选项（或最可能的选项）的概率，而非仅接收单一建议，也会很有意义。

局限性

首先，这两个数据集在效度方面都存在不足。合成病例的临床参考标准更可靠，由多位医生共同确定；但这些病例并非真实病例，因此限制了研究结果的外部效度。相比之下，真实世界病例描述了实际的用户交互及其症状，但其参考标准是症状评估应用所分配的分诊级别，而非临床决策，因此限制了内部效度。因此，这些病例的结果反映的是与该症状评估应用的一致性，部分分歧可能源于症状评估应用的错误，而非模型的错误。不过，对这两种方法进行比较仍能得出有参考价值的结果，因为两个数据集呈现的总体趋势相似：最常见的错误是过度分诊，表现最好和最差的也都是同一批模型。真实世界病例的表现明显较低，这或许是因为与教科书式合成病例相比，这些病例不够明确，且更为非典型[24]。

另一项局限性涉及真实世界病例的来源。所有病例均来自单个症状评估应用（SAA），该应用决定了症状和风险因素的本体，以及向用户提出的问题。因此，模型仅接收到该 SAA 认为相关的信息，而这些信息可能不同于其他症状评估应用或医疗从业者所收集的信息。然而，主诉、危险信号、风险因素和后续问题的结构在症状评估应用中较为常见，并且这一局限性同样适用于所有模型；因此，模型之间的比较和内部效度不应受到影响。不过，这确实限制了绝对准确率和紧急情况捷径向其他系统迁移的效果，今后应使用来自其他症状评估应用的数据进行检验。

与此相关的是，真实世界病例以德语提供。有研究表明，大语言模型对健康问题的回答会因语言而异，其中英语表现最佳[47]，决策模型可能也存在同样的情况。由于合成病例使用英语，两组数据集之间的部分差异也可能由语言造成。因此，在英语或其他语言的真实世界病例上的表现可能有所不同，未来研究应对此进行检验。

最后，我们采用了既往研究中的提示词和回答选项[13,28]，以便将我们的结果与此前对大语言模型的评估进行比较。这些未必是能够实现最高准确率的提示词。对于大语言模型而言，反映人类推理策略的提示词在一项既往研究中提高了分诊准确性[48]；其他提示词或回答类别也可能提高决策模型的准确率。这一点适用于对语言模型的总体评估，未来研究应对此进行探讨，以确定最适合分诊问题的提示词。

结论

新一代决策模型有助于及早识别急症，但就目前形式而言，不应取代确定性分诊引擎。不过，它们可用于在症状评估的早期阶段作出急症判定：由于模型会为每个回答选项返回一个概率，因此可以预先设定可接受的漏诊紧急情况和误报率；在这项回顾性研究中，这使得能够为很大一部分用户及早作出急症判定。可接受的误差范围需要针对每种医疗场景，与临床医生和监管机构共同确定。加之其输出仅限于预先定义的选项，这种对错误的控制也可能使决策模型比大语言模型更容易通过 MDR 认证。它们能否同时缩短症状评估的其他环节，仍有待探讨。

参考文献

1. McAteer, A., Elliott, A. M. & Hannaford, P. C. 在英国范围内开展的社区调查中，确定症状冰山的规模。英国全科医学杂志 **61**, e1–e11 (2011年)。2. Sætre, L. M. S. 等 基于丹麦症状队列重新审视症状冰山——2022年丹麦普通人群的症状体验与就医行为。*Heliyon* **10**, e31090 (2024)。3. Ayo-Ajibola, O., Davis, R. J., Lin, M. E., Riddell, J. & Kravitz, R. L. 美国用户对人工智能生成的健康信息的采用与体验特征：横断面问卷研究。*J. Med. InternetRes.* **26**, e55138 (2024)。4. Ayre, J., Cvejic, E. & McCaffery, K. J. 2024年澳大利亚人使用ChatGPT 获取健康信息：全国代表性调查的见解。*Med. J. Aust.* **222**, 210– 212 (2025年)。5. Wallace, W. 等 数字和在线症状检查工具的诊断与分诊准确性：系统综述。*Npj Digit. Med.* **5**, 118 (2022年)。6. 欧洲议会和欧盟理事会于*2017*年*4*月*5*日颁布的欧盟法规 *2017/745*，涉及医疗器械，修订指令 *2001/83/EC*、条例（*EC*）第*178/2002*号和条例（*EC*）第*1223/2009*号，并废止理事会指令 *90/385/EEC* 和 *93/42/EEC*（与欧洲经济区相关文本）。。《欧盟官方公报》L辑《欧盟官方公报》*L*辑第117卷 (2017年)。7. Chambers, D. 等 针对紧急健康问题的数字和在线症状检查工具及健康评估／分诊服务：系统综述。《英国医学杂志开放版》 **9**, e027743 (2019年)。8. Ćirković, A. 对四款人工智能辅助自我诊断应用在三种诊断方面的评估：两年随访研究。*J. Med. Internet Res.* **22**, e18097 (2020)。9. Baker, A. 等 为分诊和诊断而比较人工智能与人类医生。《人工智能前沿》 **3**, 543405 (2020)。10. Gilbert, S., Harvey, H., Melvin, T., Vollebregt, E. & Wicks, P. 大语言模型人工智能聊天机器人需要获得医疗器械批准。*Nat. Med.* **29**, 2396–2398 (2023年)。

11. Freyer, O.、Wiest, I. C.、Kather, J. N. 和 Gilbert, S. 未来大语言模型在医疗领域的应用取决于监管机构执行安全标准。*Lancet Digit.Health* **6**，e662–e672（2024）。

12. Kopka, M.、von Kalckreuth, N. 和 Feufel, M. A. 在线症状评估应用程序、大语言模型和普通民众进行自我分诊决策的准确率。*Npj Digit.Med.* **8**，1–14（2025年）。

13. Kopka, M.、He, L. 和 Feufel, M. A. 评估 ChatGPT 不同模型版本提供就医建议的准确率。*Commun. Med.* **6**，171（2026）。

14. Ramaswamy, A. 等 ChatGPT Health 在分诊建议结构化测试中的表现。*Nat. Med.* https://doi.org/10.1038/s41591-026-04297-7（2026）doi:10.1038/s41591-026-04297-7。

15. 德国联邦卫生部。《医疗保健数据与数字创新法》。*BMG*https://www.bundesgesundheitsministerium.de/service/gesetze-und-verordnungen/guv-21-lp/gedig（2026）。

16. Franc, J. M., Cheng, L., Hart, A., Hata, R. & Hertelendy, A. 商业大语言模型（ChatGPT）使用加拿大分诊与病情严重程度量表进行急诊科分诊的重复性、再现性和诊断准确性。*Can. J. Emerg.Med.* **26**, 40–46（2024）。

17. Madrid-García, A. & Merino-Barbancho, B. Jev 在医学中的应用：基准评估。预印本，见 https://doi.org/10.48550/arXiv.2609.34024（2026）。

18. Du, Y. 等。医学中的 System-One 决策有多可靠？Jev、类 Jev 模型和大语言模型的比较研究。预印本，见 https://doi.org/10.48550/arXiv.2610.18529（2026）。

19. Freyer, O. 等。针对医学多项选择题，对决策模型与生成式大语言模型进行基准测试。预印本，载于 https://doi.org/10.48550/arXiv.2610.38448（2026 年）。

20. Wiest, I. C. 等。用于结构化医疗信息检索的隐私保护型大语言模型。*Npj Digit. Med.* **7**，257（2024）。

21. Riedemann, L., Labonne, M. & Gilbert, S. 医学领域大语言模型的未来在于开放。*Npj Digit. Med.* **7**, 339 (2024).

22. Semigran, H. L., Linder, J. A., Gidengil, C. & Mehrotra, A. 症状检查器在自我诊断和分诊中的评估：审计研究。*BMJ* **351**, 1–9 (2015).

23. Hill, M. G., Sim, M. & Mills, B. 澳大利亚免费在线症状检查工具和应用程序提供的诊断和分诊建议质量。*Med. J. Aust.* **212**, 514–519 (2020).

24. Kopka, M. 等 用于设计特定用例代表性案例并评估普通民众和症状评估应用程序分诊准确性的 RepVig 框架。*Sci. Rep.* **14**, 30614 (2024).

25. Schmieding, M. L. 等人症状检查应用程序的分诊准确性：五年随访评估。*J. Med. Internet Res.* 1–13 (2021) doi:http://dx.doi.org/10.2196/31810.

26. Schmieding, M. L., Mörgeli, R., Schmieding, M. A. L., Feufel, M. A. & Balzer, F. 症状检查器分诊能力基准测试：与非专业人士的分诊能力比较——调查研究。*J. Med. Internet Res.* **23**, e24475 (2021).

27. Ceney, A. 等人在线症状检查工具的准确性及其对服务利用的潜在影响。*PLOS ONE* **16**, e0254088 (2021).

28. Levine, D. M. 等人GPT-3 人工智能模型的诊断和分诊准确性：一项观察性研究。*Lancet Digit. Health* **6**, e555–e561 (2024).

29. Delshad, S.、Dontaraju, V. S. 和 Chengat, V. 与医疗服务提供者的共识决策相比，人工智能应用程序可提供准确的医疗分诊建议。*Cureus* https://doi.org/10.7759/cureus.16956 (2021) doi:10.7759/cureus.16956。

30. Ito, N. 等 GPT-4 在健康状况诊断和分诊中的准确率及潜在种族和族裔偏见：评估研究。*JMIR Med. Educ.* **9**，e47532（2023）。

31. Cross, S.、Mourad, A.、Zuccon, G. 和 Koopman, B. 搜索引擎与症状检查器：在线健康建议效果比较。收录于*Web Conference 2021* 论文集，206–216 页（ACM，斯洛文尼亚卢布尔雅那，2021）。doi:10.1145/3442381.3450140。

32. Gellert, G. A. 等 使用基于规则的分诊协议与基于人工智能的自动化虚拟分诊进行实时临床分诊的比较性表现分析。*J.Hosp. Adm.* **13**，8（2023）。

33. Rutledge, G. W. GPT-4 在常见临床情景和具有挑战性的病例中的诊断准确性。*Learn. Health Syst.* **8**，e10438（2024）。

34. Kopka, M.、Feufel, M. A.、Balzer, F. 和 Schmieding, M. L. 普通人的分诊能力：回顾性探索性分析。*JMIR Form. Res.* **6**，e38977（2022年）。

35. Ben-Shabat, N. 等 评估基于聊天机器人的症状检查器的数据收集：一项临床病例短文研究。*Int. J. Med. Inf.* **168**，104897（2022年）。

36. Kopka, M. 和 Feufel, M. A. 如何评估症状检查器和诊断决策支持系统的准确率：症状检查器准确性报告框架（SCARF）。*JMIR Hum. Factors* **13**，e76168（2026）。

37. Cohen, J. F. 等 STARD 2015 指南：诊断准确性研究报告规范的说明与详解。《英国医学杂志开放版》 **6**, e012799 (2016)。

38. Riboli-Sasco, E. 等 在线症状检查工具的分诊与诊断准确性：系统综述。*J. Med. Internet Res.* **25**, e43803 (2023)。

39. Kopka, M. & Feufel, M. A. 软件 symptomcheckR：用于分析和可视化症状检查器分诊性能的 R 软件包。*BMC Digit. Health* **2**, 43 (2024)。

40. Gries, A., Schrimpf, A. M. & Von Dercks, N. 医院急诊科。*Dtsch.Ärztebl. Int.* https://doi.org/10.3238/arztebl.m2022.0276 (2022) doi:10.3238/arztebl.m2022.0276。

41. Slagman, A., Fischer-Rosinský, A., Legg, D., Schmieder, K. & Möckel, M. 德国急诊科常规临床信息中记录的低紧急度就诊识别。*BMC Emerg. Med.* **23**, 64（2023）。

42. Lammila-Escalera, E. 等 急诊科数字化登记与分诊亭的安全性和有效性：系统综述。*J. Med. Internet Res.* **27**, e69528（2025年）。

43. Khot, T. 等 分解式提示：解决复杂任务的模块化方法。预印本, https://doi.org/10.48550/arXiv.2210.02406（2023）。

44. Bontrup, F.、Kollwitz, C.、Biller, L.、Wache, H. 与 Hönigsberg, S. 医疗保健生态系统编排：一项关于跨层级患者导航的行动设计研究。载于《面向更美好未来的设计：超越人工科学。已完成研究》（编者：vom Brocke, J. 等）333–350（施普林格·自然瑞士，查姆，2026）。doi:10.1007/978-3-032-28316-0 19。_

45. Wright, J.、Williams, R. 与 Wilkinson, J. R. 健康需求评估的发展与重要性。*BMJ* **316**，1310–1313（1998）。

46. Napel, H. T. 等。 使用 *ICPC-3* 改进初级卫生保健数据：从医学视角转向以人为中心的视角。《家庭医学年鉴》**20**，358–361（2022年）。

47. Jin, Y. 等。 最好用英语提问：针对医疗保健查询的大语言模型跨语言评估。载于 *2024*年 *ACM Web*会议论文集，2627–2638（ACM，新加坡，新加坡，2024）。 doi:10.1145/3589334.3645643。

48. Kopka, M. & Feufel, M. A. 使用反映现实世界人类推理策略的提示词，提高大语言模型提供就医建议的准确率：验证研究。《JMIR生物医学工程》**11**，e88053（2026）。