

对话助手支持心力衰竭患者： 比较一种神经符号架构与 ChatGPT

Anuja Tayal¹ Devika Salunke² Barbara Di Eugenio¹ Paula G Allen-Meares³
Eulalia P Abril⁴ Olga Garcia-Bedoya³ Carolyn A Dickens³ Andrew D. Boyd²

¹Department of Computer Science ²Department of Biomedical and Health Information Sciences

³Department of Medicine ⁴Department of Communications

{ atayal14, dsalun2, bdieugen, pameares, eulalia, ogarciab, cdickens, boyda } @uic.edu

Abstract

对话助手在各个领域变得越来越受欢迎，包括医疗保健领域，部分原因是大型语言模型的可用性和能力。需要进行有真实利益相关者参与的可控探究评估，以突出传统架构与基于生成式人工智能的架构之间的优缺点。我们进行了一个组内用户研究，以比较两个版本的对话助手，这些助手允许心力衰竭患者询问食物中的盐含量。其中一个版本是使用神经符号架构在内部开发的，另一个版本基于 ChatGPT。评估表明，内部系统更准确，完成更多任务，并且比基于 ChatGPT 的系统冗长程度更低；另一方面，基于 ChatGPT 的系统发生语音错误的情况更少，并且完成任务需要的澄清更少。患者对这两者没有表现出偏好。

1 介绍

自 1966 年 ELIZA 起，医疗领域中的对话助手就已有自然语言处理 (NLP) 那么久的历史，ELIZA 已经扮演了精神科医生的角色 (?)。在近年来，随着数据集和机器学习方法的可用性，这些对话助手迅速增多，甚至在大型语言模型 (LLMs) 崛起之前。同时，中的 (?) 提供了一份全面的调研，深入分析了这些多样化的医疗导向的对话系统，从计算的角度研究这些系统，并突出它们的不同终端用户。

传统的任务导向型对话系统通常使用诸如槽填充 F1-得分和意图识别准确性来评估自然语言理解 (NLU)，以及槽准确性和联合目标准确性来评估对话状态跟踪 (DST) (?)。然而，任务绩效仅是对话系统评估的一个方面，正如 Paradise 所已展示的那样，它是一个将用户满意度与任务成功和交互成本联系起来的框架。

基于大型语言模型 (LLM) 的对话系统的评估引发了额外的关注：LLM 并不在严格的任务边界内运作，使得应用标准的任务导向评估指标变得困难。此外，LLM 在数据来源上往往缺乏透明性，可能无法可靠地遵循用户的提示，在那些精准性和问责制至关重要的高风险领域中引发了重大担忧。这些限制在促进医疗对话时的实际现实应用中尤为关键。在这种情

况下，定义和评估什么构成“良好”性能变得更加复杂和细致。

因此，人类评估依然是评估医疗对话系统的黄金标准 (?)。人类评估者可以提供关于连贯性、信息量和用户满意度等主观质量的见解——这些因素仅用自动化指标是难以捕捉的。我们展示了一项用户研究，其中我们比较了一个传统的任务导向的对话系统 (ToDS) 与一个基于 LLMs 的系统在医疗保健中的表现。

该研究聚焦于患有心力衰竭的非裔美国患者，旨在通过提供有关各种食物中盐含量的信息，帮助他们管理盐摄入量。心力衰竭患者必须仔细监测和减少盐的摄入：非裔美国人更容易患上心力衰竭 (?)，对盐的敏感度更高，并且面临健康食品获取困难等挑战。此外，在 (?)，作者指出，患有心力衰竭的非裔美国患者经常在心力衰竭教育课程中探讨盐和食物，这表明在这一领域存在显著的兴趣和信息需求。通过提供一个工具来方便获取有关食物中盐含量的信息，我们旨在增强患者做出更健康饮食选择的能力，从而解决心力衰竭管理的关键方面。

这项研究采用受试者内设计比较了两种对话系统——一个内部的神经符号系统 (HFFoodNS) 和一个基于 ChatGPT 的系统 (HFFoodGPT)。我们通过使用互动前后的调查问卷进行内在 (任务表现) 和外在的分析 (?)，来评估这两个系统在非洲裔美国患者住院期间的表现。通过评估真实患者 (而不是像 Mechanical Turk 工人或学生这样的典型研究参与者) 如何讨论食物和偏好接收信息的方式，我们确立了该干预措施对真实利益相关者的相关性和价值。

2 相关工作

任务导向的对话系统 (ToDS) 通常采用流水线方法，有四个主要组件：自然语言理解 (NLU) (?)、对话状态跟踪 (DST) (?)、对话管理器 (?) 和自然语言生成 (NLG) (?)。NLU 模块解释用户输入并将对话状态表示为在对话中需要

填写的槽位（例如，地点、价格范围）。DST 监控未填的槽位以通知对话管理器，由其决定下一步动作。然后将此动作传递给 NLG 模块以生成系统的响应。

早期的系统 (?) 使用精心设计的动作空间来管理对话状态，后来被神经网络 (??) 所取代。虽然管道方法整合了领域特定的知识和槽填充方法，但它通常需要额外的人为标注。

医疗保健中的对话系统。在医疗保健领域，对话系统 (ToDS) 取得了显著的增长。这些系统为各种疾病而开发，包括病人教育、心力衰竭、精神障碍、公众焦虑、癌症等。它们的应用扩展到多个领域，包括疾病诊断和健康指导等。

以患者为中心。(??) 强调开发一个对话代理，以满足心力衰竭患者自我护理的需求，利用教育课程的洞察力。(?) 研究使用巫师操作法 (WOZ) 技术创建一个基于角色的健康咨询对话数据集。最近的进展也整合了大型语言模型 (LLMs) 来解答患者询问 (?), 特别注重安全性。为了克服 LLMs 在医疗知识上的固有限制, (?) 专注于使用约 100,000 个医患对话的数据集来增强和细化 LLaMa 模型。(?) 采用以用户为中心的方法, 通过半结构化访谈进行用户反馈, 多次迭代改进他们的医学笔记生成模型。用户研究对于在特定使用案例数据缺乏的领域收集真实世界评估是很有益的 (??)。

3 两种对话系统架构

3.1 HFFood-NS

第一个系统, HFFood-NS, 是一个从 (Reference Withheld) 改编的神经符号会话系统。由于没有与食品含盐量相关的对话数据集, 我们创建了一个基于模板的对话数据集。我们利用了 USFDC (美国食品数据中心) (?) 数据集, 该数据集提供了详细的食品描述及其营养价值。我们根据食品、厨师、类型、食品重量和指标的食品槽对数据集进行了注释。创建数据集的完整细节可以在 (Reference Withheld) 中找到。最初, 我们训练了一个端到端的对话系统, 但在用户研究中, 我们只使用了其 DST 部分。

3.1.1 端到端对话系统

我们使用端到端对话系统 PPTOD (即插即用任务导向对话系统) (?) 来训练对话系统。PPTOD 是一个基于 T5 的模型, 通过为特定任务使用自定义提示, 在上下文学习中表现出色。

我们的实验表明, 仅对一个 transformer 模型进行微调以预测盐含量是不够的 (见表格 7)。这是由于两个关键原因。首先, 大型预训练语言模型 (PLMs), 如 GPT-3 和 T5 (??), 偶尔

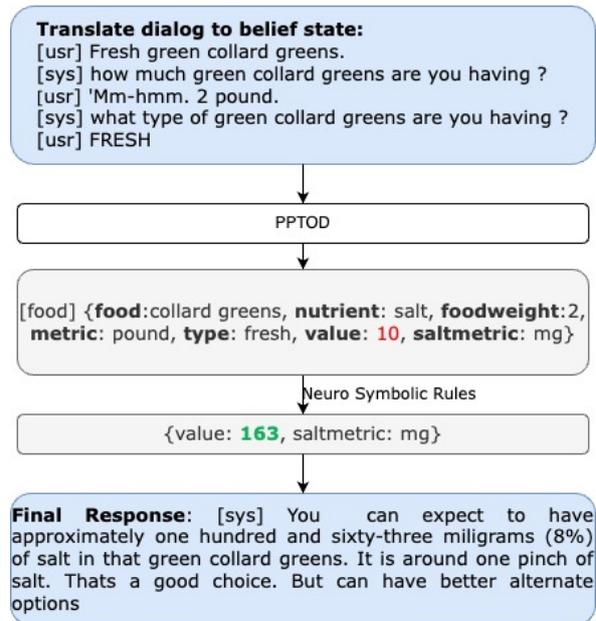


Figure 1: 带有交互功能的 HFFood-NS: 由于数据库中的值是以 100 克为单位提供的, 而用户查询的是 2 磅的量, 因此它提取了适当的值并计算出 2 磅的值。

会出现计算错误, 特别是在方程中的数学运算变得更复杂的时候 (?). 其次, 盐是一个多值槽, 其值会因食品制作方法或数量的细微变化而改变。

为了解决这些挑战, 我们将 PPTOD 模型与神经-符号规则结合。这些规则可以从数据库中检索准确的盐值, 并对特定食物重量进行数学计算, 使系统能够有效处理非标准食物数量。这种整合显著提升了系统性能, 使不同数据集大小上的联合目标准确率提高了 20% (如表 6 所示)。这些结果表明, 将预训练语言模型与神经-符号规则相结合能够实现更好的准确性。

3.2 最终模型-HFFood-NS

对于 HFFood-NS 的用户研究版本, 实施了几项措施以修改系统以在患者中心应用中部署。首先, 仅使用 NS-PPTOD 的对话状态跟踪 (DST) 模块, 以简化和提高可靠性。其次, 如果某个槽未填充, 系统会尝试询问用户该槽的信息最多两次。第三, 系统响应是基于模板的响应, 这些模板是为训练模型而创建的 (章节 3.1.1)。第四, 由于 USFDC 数据集包含详细信息, 而用户通常提供的食物描述不够精确, 系统选择第一个符合所有最低要求关键词的食物项, 以适应食物项的多值特性, 当提供盐值时。

最后, 我们精心设计了最终系统响应, 以有效呈现盐值。我们使用两种正式单位——毫克 (mg) 和百分比 (%)——来表示盐值, 这些通常可以在食品标签上找到, 同时我们还包括了一

种非正式测量——少许/一点点。我们还将盐值与推荐的每日摄入量 2000mg 进行了比较。我们将最终响应结构化为激励性的，按将盐含量分类为四个范围 (<5 % , <20 % , <50 % , <100 % , 以及其他)，并编写激励性的信息。图 1 显示了一个用户研究互动的示例和神经符号规则的应用。

3.3 HFFood-GPT

我们使用 GPT 助手创建了 HFFood-GPT，因为这样更容易与用户界面集成。由于我们的目标是教育患者而不是提供医疗建议，我们提示它（提示在附录 9 中提供）不要给出健康建议或建议就饮食指导咨询专业人士。我们将该系统命名为 Sodium Scout，并提示其分析食物中的盐含量。它建议超过 20 % 摄入量的食物不推荐，而低于 5 % 的食物则是理想选择。我们启用了代码解释器和检索功能以实现数据访问。这使得助手能够从 USFDC 数据集获取信息，该数据集也用于创建 HFFood-NS 数据集。我们提示它不要向患者提及数据集或在网络上搜索信息。

为了将 HFFood-GPT 与我们的内部系统 HFFood-NS 进行比较，我们提示 GPT-4 (?) 尽可能与 HFFood-NS 的功能相似。我们引导钠侦察员提出关于食物类型、烹饪方法和分量大小的澄清问题，并将问题限制为一次一个。此外，为了解决 GPT-4 倾向于产生冗长回答的问题，我们提示它将回答限制在 40 个词以下。

4 用户研究

设置我们招募了 23 名非裔美国 (AA) 患者，年龄在 18 至 89 岁之间，他们能够说英语并有心力衰竭的病史。所有参与者都获得了时间补偿。研究是在他们住院期间，在医院病房内进行的。

共有 20 名参与者（13 名男性和 7 名女性，年龄范围为 18-89 岁，平均年龄 $\mu = 58.75$ ，标准差 $\sigma = 14.32$ ）完成了这项需要大约 6 个月的研究。研究遵循了适当的 IRB 指导原则。如果参与者符合纳入标准并愿意参与研究，他们会收到一份知情同意书和预调查问卷。每位参与者以随机的方式依次与两个版本的对话代理 (DA) 进行互动，以减少回忆偏差。

为帮助参与者思考问题，我们提示他们思考早餐、午餐或晚餐所吃的食物。然后，参与者向两个对话系统询问关于相同食物的问题。这么做是为了确保在第二次互动时参与者不会没有问题可问，同时也为了更加方便地比较两个对话代理。然而，这些互动并不完全相同，因为每个系统提出的澄清性问题不同，这导致对

话朝不同的方向发展。

在本研究中，所有与系统的互动均通过口头进行——对话系统是基于语音的。这包括通过口述获得口头同意和用口头方式进行调查问卷的前后部分。选择使用口头交流是为了方便可能身体不适或连接了医疗设备的参与者，确保他们有一个更易接触和舒适的体验。

在住院患者中进行用户研究具有独特的挑战和考虑。例如，患者 15 的伴侣作为负责管理饮食的主要照顾者，也参与了会议并一同提问。患者 17 尽管有视觉障碍，但由于系统是基于语音的，没有参与研究的困难。有一位患者使用了气管切开管，另一位是孕妇，这突出显示了参与者的多样性。此外，三位患者由于疼痛，起初在理解系统上遇到了一些挑战，但最终能够舒适地使用它。

用户界面 (UI) 采用简约设计，旨在便于使用，包含两个单选按钮用于选择系统，还有一个按钮用于指示系统是正在录音还是在听。为了构建用户界面，我们使用了由 Hugging Face 提供的框架 Gradio (?)，因为它提供了简单而有效的解决方案。用户界面并不是“始终倾听”的，只能在某一时间段内倾听或说话。这样做是为了优先考虑患者的健康，并在互动过程中照顾到护士和医生的存在。我们将两个系统命名为 Lion 和 Shark，以便参与者能够轻松记住正在使用的系统的名称。选择这两个名字是因为它们代表了众所周知的强大的动物。

在这项研究中，我们利用了 OpenAI 技术实现文本转语音 (TTS) 和语音转文本 (STT) 功能。语音转文本方面，我们采用了 Whisper-1 (?)，而文本转语音方面，我们使用了 TTS-1。我们记录了音频并收集了转录文本用于错误分析。

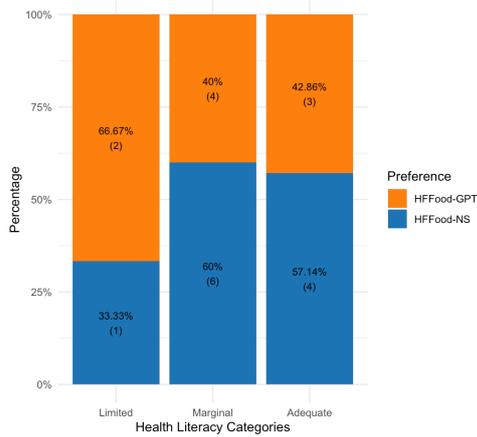
我们使用自我报告的测量工具评估了参与者的健康素养和数字素养，以更好地了解他们获取和处理健康信息的能力。可以在附录 10 中评估问卷的预调查。

健康素养是使用简短健康素养筛查工具 (?) 进行测量的。此工具包含 6 个问题，用于评估个人“获取、处理和理解做出适当健康决策所需的基本健康信息和服务”的能力。

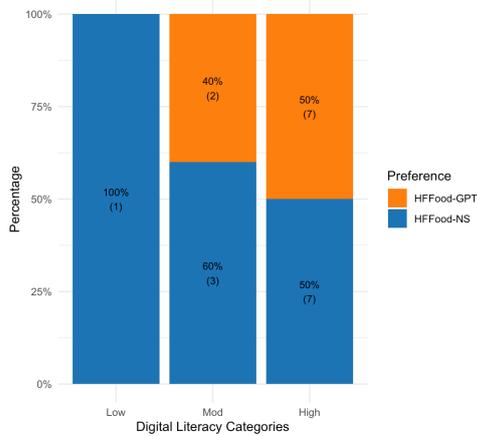
数字素养通过 e 健康素养量表 (eHEALS) (?) 进行测量。该工具包括 8 个问题，用于评估参与者感知到的寻找、评估和应用电子健康信息以解决健康相关问题的能力。

在用户体验过两个系统后，我们请他们通过问卷来评估他们的体验。问卷调查的内容见附录 11。

我们询问用户关于这两个系统的以下方面：理解答案的难易程度、问题的实用性以及理解问题的难易程度。我们还询问了他们更喜欢哪



(a) Health Literacy Vs Preference



(b) Digital Health Literacy Vs Preference

Figure 2: 按健康素养和数字健康素养水平分类的用户偏好比较，括号中的数字表示用户数量。

个系统、使用这些系统时最具挑战性的方面、他们将来使用该系统的可能性、该系统是否有帮助，以及他们是否有任何其他反馈。

5 评估

我们使用 USFDC 数据集作为基准来评估这两个系统的准确性。我们将 HFFood-GPT 视为一个任务导向的对话系统，通过将提供的值分类到预定义的槽里（如 Food、Cook、Type、FoodWeight 和 metric），类似于 HFFood-NS。

表 ?? 总结了两个系统的任务完成和准确性分析。任务完成被定义为系统提供一个值，无论正确或错误，而准确性是指系统提供正确的盐值。语音错误未计入准确性计算中。对于任务完成，我们考虑了模型提供盐值（无论是具体数值还是一个范围）的情况，而不是像 HFFood-GPT 中那样的分类描述。

评估 HFFood-GPT 特别具有挑战性，因为其具备黑箱特性。为了保持一致性，我们通过比

较提供的盐值与 USFDC 数据库中满足所有最低槽值的第一个食品项目的盐值，类似于评估 HFFood-NS 的方式评估它。然而，目前仍然不清楚 HFFood-GPT 实际访问了哪些数据或其用来计算最终答案的过程。

尽管 HFFood-GPT 在任务完成准确率方面低于 HFFood-NS，但它实现了较高的槽位准确率。表 1 和表 2 分别展示了 HFFood-NS 和 HFFood-GPT 的槽位准确率分析。我们将不正确的槽位分为语音错误和部分语音错误 (P-SE)，这可能是由于嘈杂的医院环境、患者的口音或医院环境的整体条件导致的。

尽管我们为这两个系统使用了相同的语音合成和语音转文字系统，但 HFFood-NS 记录的词错误率 (WER) (?) 高于 HFFood-GPT (表 ??)。HFFood-GPT 的不正确槽位错误完全来自语音错误和部分语音错误。

	Correct	Incorrect	Speech Error	P-SE
Food	86	6	4	1
Cook	39	21	11	2
Type	44	31	17	2
Foodweight	29	56	18	0
Metric	27	60	20	0

Table 1: HFFood-NS 槽位准确性分析：错误情况包括语音错误和部分语音错误 (P-SE)

	Correct	Incorrect	Speech Error	P-SE
Food	94	7	3	4
Cook	46	5	5	0
Type	58	15	9	6
FoodWeight	59	5	5	0
Metric	59	5	5	0

Table 2: HFFood-GPT 槽位准确性分析：错误包括语音错误和部分语音错误 (P-SE)

5.1 外部评价：用户感知

图 3 展示了参与者在调查问卷之后如何评价这两个系统，重点在于理解答案、理解问题以及有用的问题。我们可以看到超过 50% 的人认为所有方面都很有用。

在 20 名患者中，11 人更喜欢 HFFood-NS，而 9 人更喜欢 HFFood-GPT。由于这是一个组内研究，存在回忆偏差的可能性；然而，我们没有观察到任何证据。在首次尝试 HFFood-NS 的 10 个病例中，7 名用户更喜欢它，而在首次尝试 HFFood-GPT 的 10 个病例中，6 名用户更喜欢它。但是，对于第一个系统的偏好在统计上没有显著性，Pearson 相关系数显示 $r = 0.3$ ， $p = 0.19$ 。

基于健康和数字健康素养的偏好比较，通过预调查问卷评估，如图 2 所示。数量太小，无法得出任何确定的结论。

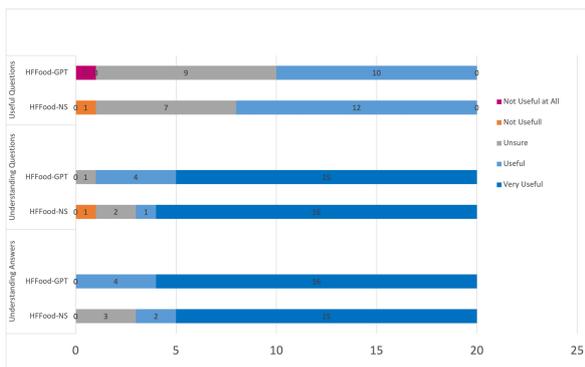


Figure 3: 调查后：对两个系统在答案和问题的有用性与理解方面的外部评估。

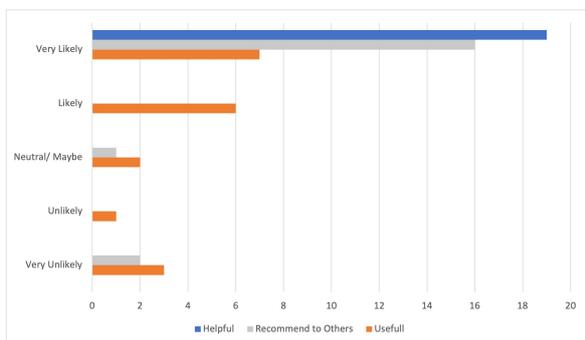


Figure 4: 调查后分析：用户如何看待会话系统

当我们询问用户对其偏好的原因时，那些喜欢 HFFood-NS 的人强调其精确和直接的回答、较快的流畅性、和简明的应答。另一方面，喜欢 HFFood-GPT 的用户则赞赏其问题易于理解、解释清晰详细以及问题的更好构建。

虽然一些用户发现这两个系统都没有问题，并认为他们对系统理解良好，但其他用户报告了一些挑战。这些挑战包括系统重复、难以理解问题以及不知道如何措辞自己的问题。

患者会在日常生活中使用决策辅助工具吗？

图 4 突出了所有用户都发现该系统很有帮助，其中大多数表示他们会推荐给其他人。大多数人认为两个系统都很有用。然而，由于患者 P5 的后续调查问题由于预定程序被中断，数据仅对 19 名患者可用。

在后续调查中，55% 的用户（11 名参与者）更喜欢以非正式术语接收有关盐含量的信息，例如“捏”或“少许”。只有 20%（4 名参与者）更喜欢毫克或克等正式单位，而其余的人选择了百分比或每日摄入量。在之前的调查中，参与者报告说他们通过撒盐、用“捏”或者凭眼睛判断来非正式地测量盐。这些非正式的方法与食物标签上出现的毫克或每日摄入量百分比等正式值不同。

6 讨论：两个系统之间的比较

表 3 总结了这两个系统，通过比较性能、设计和可用性、可靠性以及灵活性来突出其优缺点。

	HFFood-NS	HFFood-GPT
Task completion	✓	✗
Accuracy	✓	✗
Slot Accuracy	✗	✓
Less Speech Error	✗	✓
Less Processing Time	✓	✗
More Constrained	✓	✗
Error Analysis	✓	✗
Reliable	✓	✗
Predictable	✓	✗
Handling Complex query	✗	✓
Gave Options	✗	✓
Fluent	✗	✓
Concise	✓	✗
Create easily with less time	✗	✓

Table 3: HFFood-NS 和 HFFood-GPT 在性能、设计和可用性、可靠性和灵活性方面的优缺点。

6.1 HFFood-NS

使用神经符号系统结合传统的 ToDS 而不是直接提示 LLM 的一个显著优势在于能够进行更全面和精确的错误分析。这种能力在以患者为中心的系统中特别重要。通过采用这种方法，我们能够准确定位系统不足之处。

表 4 详细列出了对 HFFood-NS 进行的错误分析。这项分析包括多个重叠的类别，突出了遇到的错误的细微性。通过系统地解决这些问题，我们可以迭代地开发一个更强大、更可靠的对话系统。

	HFFood-NS
Missed Slot	27
Wrong Food Identified	9
System Error	8
Internet	6
No Data Fetched	5
Wrong Math	5
Food Not in USFDC	4
Complex Query	3
Missed Slot not in USFDC	2
Wrong Food Fetched	1

Table 4: 通过和精确的误差分析 HFFood-NS

此外，神经符号规则使我们能够为系统添加保险机制，当模型无法推断出食物重量和度量的空位时，它会假定为 100 克，从而提高任务完成率。

此外，对系统拥有更大的控制权提供了显著的优势。首先，它确保了系统始终与其主要目标一致，而这是 HFFood-GPT 难以始终如一地保持的。例如，在一个实例中，HFFood-GPT 询问了一个关于使用中甜椒颜色的澄清问题——这一细节在决定盐的用量时被视为无关紧要的。这样的控制水平帮助最小化了干扰，并使系统专注于提供准确且以目标为导向的响应。

其次，通过提供以毫克 (mg) 为单位的值来确保响应的一致性，使系统更加用户友好和可靠。相比之下，HFFood-GPT 在其响应中显示出一定的不一致性，其中 10% 的值以克 (g) 为单位给出，10% 则为分类描述符 (例如，低、高、显著)，8% 为范围而非精确值，2% 则以茶匙 (tsp) 为单位，其他则用毫克表示。这种不一致可能会给患者带来挑战，因为他们可能由于不同的数值素养水平而忽视了 1 克等于 1000 毫克的事实 (?)。保持响应的一致性是提高可访问性和确保用户能够轻松解读所提供信息的关键步骤。

6.2 HFFood-GPT

HFFood-GPT 有效处理了不同的用户描述，能够解释非标准输入并保持流畅的对话。患者以各种不寻常的方式描述他们的食物份量。许多人以“供 6 人食用”或“1 人食用份量”等份数来报告，而不是使用标准单位如盎司、克或磅。有两位患者使用独特的测量方式，如“一满勺”和“一满叉”，还有两人依靠大小比较，比如将热狗描述为“手机大小”或者量化食物重量为“3 英寸的培根”。HFFood-GPT 展示了处理这些输入的灵活性，而 HFFood-NS 在槽位识别方面遇到了困难，如表 1 所反映，大多数错误发生在食物重量和度量槽位。

当 HFFood-GPT 无法直接确定食物中的盐含量时，它会尝试收集食谱，并根据成分的盐含量来估算总盐量。虽然这种方法有时是有帮助的，但在某些情况下，系统会提出过于具体的问题，例如每种成分的确切重量或品牌，这可能对用户来说比较麻烦。

在某些情况下，品牌名称广为人知，以至于食物直接以品牌本身被识别，比如“Chick-fil-A”或“Hurricane Whopper”。在这些情况下，HFFood-GPT 成功推断出食品项，而无需额外的描述性细节，展示了其有效处理常见品牌食品引用的能力。

然而，由于 HFFood-GPT 系统具有不可预测的特性，有时系统会特别要求提供以克或盎司为单位的食物重量。例如，在表 5 中显示的对话中，当被问及克的重量时，用户并不了解这一点。HFFood-GPT 努力进行教育是不错的。用户可能并不总是知道这些准确的测量值，更可能以他们自己熟悉的单位提供数量。

在 8 个 % 实例中，HFFood-GPT 试图自行填补空位，而不是询问一个澄清问题。这种方法有时与用户意图一致，但目标应该是避免做出假设，而是提供用户选择。例如，在一个案例中，用户提到“鸡胸肉”，尽管存在潜在的语音错误，系统还是正确识别出来了，因为鸡胸肉是最常见的鸡肉食用方式。然而，如果用户

	Turn
System	Thank you for specifying the preparation method. Now, could you please tell me the portion size of the grilled turkey sausage in grams?
User	5 milligrams.
System	It seems there might be a misunderstanding with the portion size. Typically, food portion sizes are measured in grams (g) and would be much larger than 5 milligrams (mg). Could you please provide the portion size again in grams? For reference, a common serving size for sausage might be around 50 to 150 grams.
User	100 milligrams
System	It looks like we're still not quite accurate on the portion size. When we're talking about turkey sausage, the realistic portion size is usually around 100 grams (g), not milligrams (mg). One gram is equal to 1000 milligrams, so a portion size of turkey sausage might be 100 grams (which is 100,000 milligrams), for instance. Could you clarify the portion size in grams for the turkey sausage?

Table 5: HFFood-GPT 对话摘录

询问的是“鸡柳”，这种假设就会导致错误的响应。

在构建 HFFood-GPT 时，我们设定了特定的规则来约束其行为。我们将回复限制为 40 个词，但如表 ?? 所示，系统平均每个回复 54.5 个词。我们还指导它每回合只问一个问题，但大约有 7% 的回合中它违反了这一规则，同时提出了多个问题。

我们曾要求 HFFood-GPT 仅参考提供的数据库，并且不要将其引用给用户。然而，在系统的 38.5% 发言中，HFFood-GPT 参考了数据库，并表示该食物项目不在引用的数据库中，要求提供更多的澄清。这可能是用户不喜欢 HFFood-GPT 的众多原因之一，因为这一点可能让人觉得不舒服。当我们询问用户他们希望如何改进系统时，他们的主要建议是扩展知识库，包括更多的成分，并增加更多种类的食物项目，以使系统更全面和用户友好。

GPT 通过提供选项表现良好，例如说：“这听起来很美味！你能告诉我你使用的是新鲜的还是干的意大利面吗？”这种情况在大约 30% 的系统轮次中发生。这表明 HFFood-GPT 对不同的食物有一些了解。然而，在某些情况下，即便选择了选项，HFFood-GPT 仍表示在提供的数据库中缺少该信息。这种表现让用户感到沮丧，因为 GPT 依赖于自身的知识而不是数据集，从而影响了系统的一致性和可靠性。

我们进行了一个用户研究，对象是 20 名非洲裔美国心力衰竭住院患者。采用被试内设计，我们比较了两个对话系统：一个内部开发的神经符号系统 (HFFood-NS) 和一个基于 ChatGPT 的系统 (HFFood-GPT)。评估使用了内在和外在外在的衡量标准，虽然两个系统都没有成为明确的赢家，但研究突出了两者之间的关键差异。

评估表明，HFFood-NS 比 HFFood-GPT 更准确，完成的任务更多，并且提供的回答更简洁。

Train Size	Epochs	Joint Accuracy	
		PPTOD	NS-PPTOD
100	6	55.56	73.08
300	4	51.92	72.8
500	6	58.75	83.2
1000	6	58.53	85.2

Table 6: 在不同的训练规模中, 与 PPTOD 相比, 使用 NS-PPTOD 可以提高联合精度

	Train Size	Epochs	Inform	Success	BLEU
PPTOD	100	8	71.43	0	24.99
NS-PPTOD	100	-	88.90	77.80	22.50
PPTOD	300	7	75.00	5.00	34.30
NS-PPTOD	300	-	81.50	63.00	26.90
PPTOD	500	9	82.86	2.86	29.81
NS-PPTOD	500	-	74.50	58.10	28.90
PPTOD	1000	7	93.50	2.70	29.00
NS-PPTOD	1000	-	85.90	71.70	30.00

Table 7: 使用 NS-PPTOD 与 PPTOD 相比性能提升。

另一方面, HFFood-GPT 语音错误更少, 完成任务时需要的澄清更少, 并且更有效地处理复杂查询。

大型语言模型 (LLMs) 的广泛使用, 如 Chat-GPT, 往往缺乏审查, 尤其在医疗环境中引发担忧。需要更强的控制, 仅依靠提示是远远不够的。神经符号方法能够提供更大的透明度、可靠性和可解释性, 应该进一步探索并整合到未来的系统中。

展望未来, 我们的目标是开发结合两种系统优势的混合对话系统。

7 局限性和伦理声明

尽管我们认为用户研究/人工评估是评估医疗对话系统的最佳方法, 但它们有局限性。我们的研究仅招募了 20 名患者, 这是一个相对较小的样本量, 因此很难得到显著的结果。

为了构建这两个系统, 我们依赖于美国农业部食品成分数据库 (USFDC)。虽然该数据库是一个标准参考, 但它既不是完全详尽的, 也不是完全准确的。例如, 一些食物, 如本地连锁餐厅的产品或特定的小吃品牌, 在 USFDC 数据库中缺失。此外, 在关于油炸食品的讨论中, 用户经常提到使用空气炸锅的方法, 而这在数据库中没有被考虑。

8 NS-PPTOD 评估结果

9 HFFood-GPT 提示

为了开发 HFFood-GPT, 我们在 GPT-4 上使用了零样本提示。

钠侦察员热情地欢迎用户, 并帮助分析各种食物中的盐含量。它会系统地询问用户有关食物种类、烹饪方法和份量大小的问题, 一次一个, 以准确确定盐含量。钠侦察员使用提供的

JSON 文件中的数值来计算估计的盐含量, 并将其与推荐的每日摄入量 2000mg 进行比较。建议超过此摄入量 20% 的食物不推荐食用, 而低于 5% 的食物是理想的选择。钠侦察员不提供健康建议, 而是建议用户咨询专业人员以获得饮食指导。界面友好且简单明了。它专注于用清晰的方式告知用户他们的餐点中的盐含量, 确保每次只问一个问题。回答限制在 40 个词以内, 并且只搜索提供的 JSON 文件中的数据。用户不知道数据文件, 因此不要讨论它。只关注与食物及其盐含量相关的信息, 不要在网上查找信息。

10 调查前问题

• 电子健康素养问题

1. 您觉得互联网在帮助您做出健康决策方面有多有用?
2. 能够在互联网上获取健康资源对您来说有多重要?
3. 我知道互联网上可用的健康资源
4. 我知道在哪里可以在互联网上找到有用的健康资源
5. 我知道如何在互联网上找到有用的健康资源
6. 我知道如何使用互联网来回答我关于健康的问题
7. 我知道如何使用我在互联网上找到的健康信息来帮助自己
8. 我具备评估我在互联网上找到的健康资源的技能
9. 我可以辨别互联网上高质量的健康资源和低质量的健康资源
10. 我对使用来自互联网的信息做出健康决定感到有信心

• 简要健康素养筛查问题

1. 你有多经常让人帮助你阅读医院材料?
2. 您有多频繁因为难以理解书面信息而在了解自己的健康状况时遇到问题?
3. 您有多少次难以理解关于您的病情的说明?
4. 你在独自填写医疗表格时有多大的信心?

- 你是否注意食物中的盐 [是, 否, 可能] 你如何测量它?

11 调查后问题

- 1. 从 [插入系统名称: 狮子/鲨鱼] 收到的答案容易理解吗? 请从 1 到 5 进行评分, 其中 5 是很容易理解, 1 是非常难以理解。
- 您认为 [插入系统名称: Lion/Shark] 提出了有用的问题吗? (a. 大多数时候是, b. 是, c. 大多数时候不是, d. 不是)
- 理解来自 [插入系统名称 Lion/Shark] 的问题有多容易? 请从 1 到 5 进行评分, 其中 5 是很容易理解, 1 是很难理解。
- 您希望如何接收食品中盐含量的信息? 您是希望以毫克表示, 还是以占每日摄入量的百分比表示, 或者以其他方式? 请分享您的偏好。
- 您更喜欢哪个系统, Lion 还是 Shark?
- 用你自己的话来说, 使用这两个系统最困难的方面是什么?
- 您使用该系统的可能性有多大? (对每个系统进行评分: 1-非常不可能使用, 2-不太可能使用, 3-中立, 4-可能使用, 5-非常可能使用) a. 为什么? b. 如果参与者的回答是肯定的, i. 您会多常使用它? ii. 您会推荐给别人吗? (是/否)
- 这个系统有帮助吗? a. 你会如何改进这个系统? (如果有一个首选的系统), b. 如果没有首选的系统, 你会如何改进这些系统?
- 您还有其他反馈或评论吗?