从校准到合作：LLM 不确定性量化应更加以人为中心

Siddartha Devic **Tejas Srinivasan**
Jesse Thomason **Willie Neiswanger** **Vatsal Sharan**
University of Southern California
{ devic, tejas.srinivasan } @usc.edu

Abstract

大型语言模型（LLMs）越来越多地在现实世界中帮助用户，然而其可靠性仍然是一个问题。已有研究强调，不确定性量化（UQ）是一种可以通过使用户知道何时信任 LLM 预测，从而增强人类与 LLM 协作的工具。我们认为，当前在 LLM 中进行不确定性量化的实践，并不是为在现实任务中做决策的用户开发有用 UQ 的最佳方式。通过对 40 种 LLM UQ 方法的分析，我们识别了三种阻碍社区实现惠及下游用户这一目标的普遍实践：1）在生态有效性低的基准上进行评价；2）仅考虑认知不确定性；以及 3）优化的指标不一定能指示下游的效用。针对每个问题，我们提出了一些注重用户的实践和研究方向，LLM UQ 研究者应予以考虑。我们认为，相较于在不具代表性的任务上用不完美的指标进行攀升，社区应该采用更加以人为本的方法来量化 LLM 的不确定性。

1 介绍

经典的机器学习和预测领域长期以来一直关注于预测结果和未来事件的诱人可能性。一个关键且高度相关的目标是准确估计模型预测何时可以安全地被人类用户信任或忽略 (Guo et al., 2017; Valdenegro-Toro and Stoykova, 2024; Marusich et al., 2024) 。这个目标构成了估计机器学习模型预测不确定性的方法的基础，我们将其统称为不确定性量化（UQ）方法 (Sullivan, 2015) 。

虽然不确定性量化（UQ）传统上是在经典机器学习模型中进行研究的，但随着大型语言模型（LLMs）的出现，促使了许多研究来创建和调整适合这一新范式的 UQ 方法。量化 LLM 不确定性的主要原因是用户导向的：许多提出的 LLM UQ 方法的动机是使用户能够知道何时应该信任 LLM 的输出，更好地提高 LLM 的可靠性和可信度，并允许其扩展到高风险领域。LLM UQ 方法包括校准显示给用户的特定生成或生成中的声明的可信度分数，改变生成中使用的语言以更好地表示模型在其生成中的不确定性，以及在符合预测框架中生成具有覆盖率保证的预测集。

尽管 LLM UQ 方法已经展示出能够定量地改善某些 LLM 能力—例如，通过后处理 (Li et al., 2024) 或采样和汇总多次生成 (Kuhn et al., 2023) —我们认为，最初为了促进用户更好依赖的动机已经被搁置。尤其是在 LLM UQ 文献中，人类研究是少而又少的，因此我们缺乏对 LLM UQ 方法在帮助人类进行实际决策任务时有多大用处，或这些方法何时增加了 LLM 生成的推荐和决策的透明度的全面理解。相反，许多提出的 LLM UQ 方法专注于在缺乏生态有效性的狭窄基准上进行爬坡优化，但这些基准仅捕获了认知不确定性的有缺陷的校准度量 (De Vries et al., 2020) 。

目前，LLM UQ 主要是一个定量领域，每个月都有新的技术工具获得更好的基准分数（例如，更低的校准误差）。在这篇立场论文中，我们质疑当前用于基准测试 LLM UQ 方法的做法能否有效地推动我们走向最佳的人机协作。我们识别了社区中一些常见的做法，这些做法妨碍我们正确评估用于 LLM 辅助决策的 LLM UQ 方法的实用性。然后，我们提出了一些

Preprint. Under review.

建议，我们相信这些建议将推动 LLM UQ 提供对现实世界用户的价值，并可能使其融入到数百万人正在使用的许多基于 LLM 的产品中。

1.1 位置和主要论点

我们认为，NLP 社区目前用于评估 LLM UQ 方法的常规做法不足以在实际环境中惠及人类用户。从这个以用户为中心的观点来看，我们识别出 LLM UQ 方法广泛采用的三个主要障碍：

1. 在第 3 节中，我们认为 LLM UQ 方法主要是在生态效度较低的基准上进行评估，这些基准不能代表真实世界的决策场景。
2. 在 4 节中，我们指出常用于 LLM UQ 评估的基准测试没有考虑随机不确定性，并且在一定程度上也没有考虑分布不确定性。这些都是用户在使用大型语言模型（LLM）作为决策辅助工具时可能遇到的重要不确定性类型。
3. 在 5 节中，我们首先认为大模型不确定性量化方法与真实用户之间的互动研究不足且未被很好地理解。接下来，我们声称，大模型不确定性量化方法使用与用户下游效用关系不大的指标进行评估。最后，针对语言任务本身向大模型用户展示不确定性的可能方案也缺乏研究。

对于每一个问题，我们强调具体的补救措施和研究方向，这些方向可能会让 LLM UQ 方法的研究人员耳目一新。整体而言，我们希望这些现有 UQ 实践的明确缺陷能够指导该领域在真实世界的人机协作情境中提供有用的算法建议。我们还在第 6 节讨论了一些其他看法。

2 预备知识和方法学

在经典机器学习和大型语言模型（LLMs）中，不确定性量化（UQ）方法的目标是从系统中所有不确定性来源中提取估计。UQ 方法通常尝试量化三种类型的不确定性：认知不确定性、随机不确定性和分布不确定性 (Hüllermeier and Waegeman, 2021; Ben-David et al., 2010) 。

认知不确定性，也称为模型不确定性，表示由于模型的局限性而导致结果的不确定性。认知不确定性可能源于训练过程中的计算、数据或优化限制，或者仅仅是因为为特定任务选择了错误的模型。

另一方面，任意不确定性表示由于数据中的内在随机性或噪声所导致的不确定性。任意不确定性的典型例子是诊断流感患者：给定两个症状相同的患者，一个可能患有流感，另一个可能有过敏。任意不确定性的存在意味着自然界分配的结果在某种程度上是内在随机的，并且无法通过收集额外的训练数据来减少。分布不确定性表示一种常见的情况，即测试分布和训练分布在某些定量方面是不同的。例如，一个在其他选项中被选中的预训练 LLM，因为它在诸如 MMLU 或 TriviaQA (Hendrycks et al., 2021; Joshi et al., 2017) 这样的基准上取得了最高的评估得分，可能在部署时遇到真实用户查询时面临分布变化。关于不确定性类型的更全面概述被推迟到 He et al. (2023b); Smith (2024); Shorinwa et al. (2024) 。

不确定性量化方法通常试图估计不确定性的某些子集或所有来源。特别是，LLM 不确定性量化方法可分为几大类。首先，有一些方法是无监督的：这些方法中有的通过采样多个世代以测试响应的一致性，并通过变体熵 (Xiong et al., 2023; Lin et al., 2024; Kuhn et al., 2023; Duan et al., 2023; Malinin and Gales, 2021) 来衡量不确定性，还有的方法则以不同方式提示 LLM，以探测其对答案的信心 (Tian et al., 2023; Kadavath et al., 2022; Yang et al., 2024c) 。也有一些有监督的方法，这些方法首先尝试通过令牌级别的概率或特殊提示来引出置信水平，然后以事后方式校准它们 (Li et al., 2024; Ulmer et al., 2024; Chen et al., 2024; Detommaso et al., 2024; Mielke et al., 2022) 。最后，还有一些方法使用如保序预测这样的工具来修改 LLM 的生成，以更好地反映 LLM 的置信程度 (Quach et al., 2024; Jiang et al., 2025; Cherian et al., 2024; Wang et al., 2025b) 。有关相关工作的全面讨论，我们建议参考更全面的综述 (Shorinwa et al., 2024; Huang et al., 2024; Liu et al., 2025) 。

评估方法论。我们分析了 40 篇关于大型语言模型不确定性量化方法的论文及其相关的评估基准，以为我们的论点提供证据。我们从两项最近关于大型语言模型不确定性量化的调研中收集了方法论文及其中引用的文献。在 Appendix A 中，我们对每篇选定的论文进行了简要注释，重点关注与我们批评相关的方面，例如“这篇论文是否测试了方法对分布不确定性的鲁棒性?”以及“这篇论文使用了哪些基准来评估其大型语言模型不确定性量化方法?”

范围和局限性。在这项工作中，我们只考虑那些旨在通过传达不确定性来提高用户决策的 LLM UQ 方法。因此，我们不考虑主要目标是提高语言模型在推理、数学或科学基准测试能力的 UQ 方法。[1] 此外，我们的观点范围不包括那些可能允许 LLMs 在野外更好地作为完全自主代理行动的 LLM UQ 方法 (Liu et al., 2024b) 。关于我们分析局限性的额外讨论，请参阅 Section 6 。

3 UQ 基准的生态有效性

我们认为，大型语言模型的不确定性量化方法主要是在生态效度低的基准上进行评估 (Bronfenbrenner, 1977) ，并且在这些基准上的发现不太可能推广到用户在现实设置中提出的查询。如果我们希望大型语言模型的不确定性量化方法能够帮助用户解决决策问题，那么有必要对其在反映现实世界使用模式的任务上进行评估。

生态效度 (Bronfenbrenner, 1977) 指的是实验结果在多大程度上可以推广到现实世界的情境中。随着大型语言模型（LLMs）被部署以协助人们在现实环境中工作，NLP 数据集的生态效度问题受到关注 (De Vries et al., 2020) ，已经有多次努力开发反映真实用户与人工智能交互的基准测试 (Lin et al., 2023; Zhao et al., 2024) 。De Vries et al. (2020) 识别出 NLP 数据集缺乏生态效度的几种方式：合成语言、人工任务、未与潜在用户合作、预设的交互，以及限制为单回合交互。在本节中，我们评估了我们调查的 40 种 LLM 不确定性量化方法所评估的 22 个不同基准测试的生态效度。[2]

3.1 LLM UQ 基准测试代表了一小部分任务

我们首先观察到的现象如 **??** 所示，就是大多数 LLM UQ 方法仅在事实性问答或常识推理任务上进行评估。特别是，17 个基准测试都是问答（QA）或常识推理任务。真实世界中的用户查询要多样得多；例如，WildChat (Zhao et al., 2024) 发现他们的 LLM 查询中只有 6.3 % 是事实性问答。如果大多数 LLM UQ 方法的评估仅在这两个领域进行，我们就无法很好地了解这些方法在更广泛的决策应用中提供有用的不确定性估计的能力，用户在这些应用中使用 LLM。

3.2 LLM UQ 基准测试不反映以用户为基础的应用程序

在以人为中心的可解释人工智能（XAI）的背景下，Chaleshtori et al. (2024) 引入了用于确定基准适合性以评估模型解释对人类效用的标准。这些标准并非特定于解释，因此我们采用它们来评估 LLM UQ 基准的生态有效性：

C1. 该基准与用户可能寻求大型语言模型帮助的现实任务有着有意义的联系。为了使基准在生态上有效，它必须以实际应用为基础。不满足此标准的基准示例包括 CommonsenseQA (Talmor et al., 2019) 、StrategyQA (Geva et al., 2021) 和 TriviaQA (Joshi et al., 2017) 。[3]

C2. 基准测试的输入代表了现实世界中 LLM 使用中的任务输入。即使基准测试基于一个现实世界的应用，其输入也必须代表用户可能提供给 LLM 的真实输入。例如，WebQA (Chang et al., 2022) 由设计成能产生预先知道的特定答案的问题组成，这在用户实际上与 LLM 协作时不太可能发生。

C3. 基准实例需要人们付出相当的努力，或者人们在完成任务时表现较差（即使提供了合理的资源）。用户可以轻松准确完成的任务不需要他们与大型语言模型（LLM）合作，因此不适合用于研究 LLM 的不确定性量化方法。例如，研究 UQ 的最流行基准之一是 GSM8K (Cobbe et al., 2021) ，该基准涉及解决大多数用户可以轻松解决的小学数学问题。

C4. 由于错误决策，可能会发生一些不希望的事件。在用户容易受到一些不良后果影响的任务中，量化 LLM 不确定性显得尤为重要，因为这些后果是无效的人类与 LLM 合作的结果。

[1] 例如，使用 UQ (Ye et al., 2025) 改进 GRPO 风格的 LLM 推理链超出了本研究的范围。

[2] 我们只考虑至少被两篇方法论文使用的基准。

[3] 我们在判断基准是否满足这一标准时，会考虑其构建方式和输入域。例如，尽管用户可能会使用大型语言模型助手来回答事实性问题，但 TriviaQA 由小众的冷知识问题组成。在真实应用中，个人可能在不知道答案的情况下提问冷知识问题的唯一场景是参加竞赛，而在竞赛中他们不允许使用大型语言模型。

Benchmark	Category	C1: The benchmark has a meaningful connection to a real-world task	C2: The benchmark inputs are representative of real-world task inputs	C3: People find the benchmark to be challenging or require notable effort	C4: Undesirable events may occur due to bad decisions.
TriviaQA	Factual QA	✗	-	✓	✗
Natural Questions	Factual QA	✓	✓	✓	✗
MMLU	Factual QA	✗	-	✓	✗
CoQA	Open-book QA	✗	-	✓	✗
TruthfulQA	Factual QA	✗	-	✗	-
SciQ	Factual QA	✗	-	✓	✗
GSM8K	Math	✗	-	✗	✗
BIGBench	Other reasoning	✗	-	✓	✗
SST	Sentiment Analysis	✗	-	✗	✗
WebQA	Open-book QA	✓	✗	✗	✗
CommonSenseQA	Commonsense	✗	-	✗	✗
IMDB	Sentiment Analysis	✓	✓	✗	✗
AGNews	Topic Classification	✓	✓	✗	✗
Yelp Reviews	Sentiment Analysis	✓	✓	✗	✗
NQ-Open	Factual QA	✓	✓	✗	✗
AmbigQA	Factual QA	✗	-	✓	✗
MNLI	Commonsense	✗	-	✗	✗
SNLI	Commonsense	✗	-	✗	✗
SWAG	Commonsense	✗	-	✗	✗
HellaSwag	Commonsense	✗	-	✗	✗
BioASQ	MedicalQA	✗	-	✓	✗
MedMCQA	MedicalQA	✗	-	✓	✗

Table 1: 我们对 22 个基准进行了分析，并根据第 3.2 节中的标准进行了分类和注释。

Chaleshtori et al. (2024) 认为，只有满足标准 C1 到 C3 的基准才适合研究用户依赖性，而在这些基准中，还满足标准 C4 的应优先考虑。

在我们对大型语言模型（LLM）不确定性量化（UQ）基准是否满足这些标准的注释中（表格 1），我们发现只有 22 个基准中的 6 个（27.3 %）满足 C1 标准（基于实际应用），并且只有 5 个同时满足 C2 标准（真实输入）。22 个基准中有 9 个满足 C3 标准（对用户具有挑战性）。然而，只有 2 个基准满足所有 C1 到 C3 的标准：Natural Questions (Kwiatkowski et al., 2019) 和 NQ-Open (Lee et al., 2019)，它们由来自真正 Google 搜索查询的信息寻求问题组成。没有任何基准满足 C4 标准（高风险），这表明用于评估 LLM UQ 方法的流行基准对其在高风险决策任务中的适用性几乎没有启示。

最后，我们发现我们分析的 22 个基准测试中有 15 个（68.2%）是多项选择问题。众所周知，多项选择问答（MCQA）问题存在许多问题，例如在 Khatun and Brown (2024); Zheng et al. (2024); Balepur et al. (2025) 中详细描述的那些。从高层次上看，MCQA 评估的问题包括位置排序偏见、任务格式的刚性、存在对抗性选择的干扰项、对扰动的脆弱性等。最重要的是，现实世界的用户查询几乎从来不是多项选择题，正如在 ShareGPT (Ouyang et al., 2023) 和 WildChat (Zhao et al., 2024) 中所展示的。即使是由问答基准捕获的简单知识查询，现实世界的用户查询仍然是短文本或长文本生成。主要依靠 MCQA 基准评估大模型中的不确定性估计方法，并不一定能够反映它们在可能决策集未被枚举时估计不确定性的能力。

3.3 建议

我们提出了一些建议，以便选择更合适的评估基准：

R1. 大语言模型不确定性量化方法的开发者应该在更多元化的基准测试上进行评估，这些测试应包括超越事实性问答的应用，并且不附带保证包含正确答案的固定选项。在对多项选择问题回答基准进行评估时，我们建议使用由 Balepur et al. (2025) 规定的稳定评估协议。

R2. 在选择评估 LLM UQ 方法的基准时，我们建议根据 Section 3.2 中的 Chaleshtori et al. (2024) 采用的标准来评估基准。满足这些标准的基准将为 LLM UQ 方法在现实世界的人机协作中的适用性提供更有意义的见解。

R3. 我们呼吁社区更加关注理解生成任务中不确定性的本质，例如代码生成 (Vasconcelos et al., 2025a) 、摘要 (He et al., 2024) 、规划 (Hu et al., 2024) ，以及其他复杂推理任务，这些任务不符合事实问答任务中的简单“正确性”概念。

4 LLM UQ 中考虑的不确定性来源

在本节中，我们论述了用于衡量 LLM UQ 方法进展的许多评估基准未能解决两个关键方面的问题：随机不确定性和分布漂移。从用户中心的视角来看，一个理想的 LLM UQ 方法应该能够表示所有存在的不确定性来源，而不仅仅是认识论不确定性。然而，我们发现 LLM UQ 研究中存在倾向于喜好既定的 QA 基准的趋势，尽管它们缺乏对随机不确定性的考虑。此外，尽管 LLM UQ 方法在检测分布外例子 (Lin et al., 2022) 方面的重要性，我们发现几乎一半的受调研的有监督 LLM UQ 方法没有在分布漂移的情况下评估其性能。这些疏忽共同表明，一些 LLM UQ 方法在涉及多种不确定性相互作用的真实场景中可能缺乏可靠性。

4.1 随机不确定性

人类决策者经常面临可能本身具有不确定性的任务。例如，在紧急分诊情况下，可供临床医生使用的电子病人健康记录可能无法完全说明病人的状况，但临床医生仍需做出治疗决定 (Alur et al., 2023) 。一家银行可能只有关于潜在客户的少数特征可用——例如他们的信用评分、收入等等，但银行仍需要决定是否发放抵押贷款。在这些情况下，由于自然的固有随机性、特征的部分可观察性，或两者兼有，真实结果可能是非确定性的，并包含随机的不确定性。实际上，许多人正是在类似高风险的情况下设想机器辅助可能是最有影响的 (De-Arteaga et al., 2020; Brennan et al., 2009) 。

大多数 LLM UQ 方法未能解决和评估具有偶然性不确定性的任务。LLM UQ 论文中评估的最流行的数据集是问答数据集。例如，基准测试如 GSM8k (Cobbe et al., 2021) 、MMLU (Hendrycks et al., 2021) 和 TriviaQA (Joshi et al., 2017) 都收集了人类标记和检查的单一正确答案的问题。单一黄金标签的存在从本质上消除了估计偶然性不确定性的可能性，因为偶然性不确定性依赖于结果本身并不总是有相同的正确答案。[4]

事实上，在我们评估的 LLM UQ 论文中使用的 22 个基准中，我们发现只有 1 个基准有意包含了偶然不确定性 (Min et al., 2020, AmbigQA) 。评估含有偶然不确定性的基准的不足并不是基准创建者的过错。实际上，存在许多专门设计考虑了偶然不确定性的基准。例如，SST-5（有全部五个类别，Socher et al. (2013) ）、AmazonReviews (Hou et al., 2024) 、像 ToxiGen (Hartvigsen et al., 2022) 这样的有毒文本分类数据集，以及像 ChaosNLI (Nie et al., 2020) 这样的意见聚合数据集，每个实例都包含多个标签注释，标签有变化，因此从根本上在潜在群体中编码了偶然不确定性。

最近提出的一个专门用于在偶然性不确定性下评估 LLM 不确定性量化方法的数据集是 FolkTexts (Cruz et al., 2024) ，该数据集要求 LLM 根据公开的美国人口普查数据回答有关个人的问题。由于该数据集对每个任务有数十万个数据点，给定一个特定的任务和特征集，数据集中通常包含多个可能互相矛盾的结果。因此，一个校准过的 LLM 应该对每个个体输出合理的概率估计。重要的是，(Cruz et al., 2024) 发现并不是所有的 LLM 不确定性启发和校准方法在处理偶然性不确定性时都表现良好，这进一步强调了明确测试这种不确定性的必要性。此外，已经显示自回归 LLM 在这种多标签情境下表现出不一致的标签预测和概率分布 (Ma et al., 2025) 。

不确定性和选择性预测。模糊问答 (Cole et al., 2023; Zhang and Choi, 2021) 和选择性预测 (Geifman and El-Yaniv, 2017; Chen et al., 2023) 等领域也在大规模语言模型 (LLMs) 中体现了偶发不确定性元素。例如，模糊 QA 数据集如 AmbigQA (Min et al., 2020) 和模糊 Trivia QA (Kuhn et al., 2022) 通过查询的语言模糊性引入了偶发不确定性。然而，评估通常与识别能将原始查询转换为具有单一正确答案的新查询的澄清性问题相结合，而不是要求 LLM 表示原始查询的概率性不确定性。在选择性预测领域的额外工作也强调了需要更多的 LLM UQ 方法来进行测试和验证，以应对偶发不确定性。例如，Vazhentsev et al. (2023) 表明，那些仅设

[4]唯一可能剩下的偶然不确定性来源是标注者错误导致的无意噪声，这通常在后续旨在改进基准的工作中被去除，例如 Rücker and Akbik (2023) 。

计和测试捕捉知识性不确定性的 UQ 方法在包含偶发不确定性的选择性预测任务中可能表现不佳。

这些相关的发展强调了一个重要的观点：如果 LLM UQ 方法仅仅被用于评估认知不确定性——而不是测试它们量化和表达许多现实世界情境中不可减少的数据固有不确定性的能力——那么 UQ 方法最终可能对用户中心的下游应用，如选择性预测或模糊 QA，作用不大。这个盲点强化了需要将不确定性作为一个中心原则进行测试的必要性。

4.2 分布偏移

利用大型语言模型的真实用户在许多方面具有多样性，包括地理上 (Zhao et al., 2024; Dunn et al., 2024) 和文化上 (Hershcovich et al., 2022)，这对下游应用有实质性的影响。例如，Xing et al. (2022) 证明了用户与对话搜索代理互动的方式在复杂性和长度上因用户年龄而异。当评估一个大型语言模型的度量方法时，对于单一基准的训练/测试分割，这种变异性的鲁棒性通常是不可测量的。相反，对查询变异性的鲁棒性要求大型语言模型的度量方法在面对第三种不确定性：分布偏移时仍保持有用。

许多大型语言模型（LLM）的不确定性量化（UQ）方法集中于对模型输出进行后处理，以改进校准或不确定性估计。这些方法通常属于监督参数学习范畴，即该方法需要从一组数据中学习一些参数。例如，知名的 Temperature 或 Platt 标度 UQ 方法（其变体常用于 LLM UQ 中）需要保留数据以拟合模型 logits 或概率预测的重新标度参数。即使是无监督无参数的 LLM UQ 方法，如口头表达的置信评分，通常也会应用事后校准来改进特定任务上的最终不确定性估计。对于这样的方法，所使用的确切保留数据会对 UQ 方法的学习参数及其表现产生影响。

因此，评估数据分布外（OOD）的表现是至关重要的：也就是说，这些查询看起来与用于学习 LLM UQ 方法最初参数的训练数据不同。如果 LLM UQ 方法始终无法识别或适应分布变化，可能会削弱用户对 LLM 可靠性的信任，从而导致人机联合系统 (Srinivasan and Thomason, 2025) 的实用性下降。

不幸的是，我们发现，在我们评估的 19 篇监督式参数学习 LLM UQ 方法的论文中，只有 10 篇讨论和测试了对分布偏移的鲁棒性。我们认为，在分布偏移下测试事后校准方法应该是任何 LLM UQ 论文的要求。在分布偏移下作为标准实践进行评估，将有利于 LLM UQ 方法在真实世界环境中应用，因为用户查询可能会随时间变化或随新的人口统计转变，并且当 LLM UQ 方法用于下游任务时——例如幻觉检测 (Detommaso et al., 2024; Zhang et al., 2023) 或选择性预测 (Chen et al., 2023) ——OOD 性能对可靠性至关重要。

为了解决本节提出的问题，我们提供了两个简单且可操作的建议：

R4. 在具有内在的、故意加入的随机不确定性的数据集（例如 FolkTexts (Cruz et al., 2024) 或 ChaosNLI (Nie et al., 2020)）上评估方法。

R5. 对于提出监督 LLM UQ 方法的工作，确保您在有分布变动的情况下进行测试。因为使用保形预测保证的方法在理论上通常在分布变动情况下不成立，所以在分布变动下的评估尤其重要 (Cherian et al., 2024; Quach et al., 2024) (Aolaritei et al., 2025)。

5 指标和强调人类研究的必要性

我们认为，用于优化和评估 LLM UQ 方法的指标可能无法充分反映 LLM 推动有效和有意义的人机协作的能力。特别是，我们调查的大多数论文都集中在简单问答任务上的标准校准测量，如 Brier 得分、ECE 或 AURUC。[5] 我们主张关注与现实世界决策任务中人类提升相关的指标。这个以用户为中心的转变需要在人机协作理论的进步以及算法辅助工具的实用性研究和不确定性量化方面的突破。尽管一些研究已经在经典 UQ 设置中进行过 (Corvelo Benz and Gomez Rodriguez, 2025; Alur et al., 2023)，生成语言、UQ 和对用户的实用性之间的相互作用却被严重忽视。在我们每个部分提出的问题和方向中，这个问题需要更多的思考、专注和谨慎的研究来执行。

[5] 在 **??** 中，我们还提出从 ECE 切换到 smECE (Błasiok and Nakkiran, 2024)。

5.1 源于经典不确定性量化的启示：用于实际应用的校准指标

当大型语言模型（LLM）帮助医生诊断复杂病例或帮助软件工程师评估代码中的潜在错误时，不确定性量化（UQ）的根本目标超出了在任务上获得良好的 ECE 或 Brier 分数。相反，LLM UQ 文献的一个普遍期望是使语言模型更有用和透明，作为人类辅助工具 (Yang et al., 2024c; Tian et al., 2023; Sung et al., 2025; Li et al., 2024) 。计算模型在特定任务或数据点上的不确定性是这一目标的一个方面。然而，如上述例子所示，简单地计算校准概率往往不是故事的结尾。为了宣告成功，LLM UQ——就像许多其他与现代 LLM 应用相关的研究领域 (Kapoor and Narayanan, 2025) ——需要进行人类提升研究。特别地，有算法建议并量化不确定性的人类在现实任务中的表现是否优于以下两种情况：(i) 有算法建议但没有量化不确定性；(ii) 完全没有算法建议？

理论界和人机交互（HCI）社区都探讨了如何向用户呈现不确定性以提供提升的各个方面；即，通过获取算法建议来帮助用户改善决策。在经典（非大型语言模型）的不确定性量化（UQ）中，虽然研究结果仍在不断涌现，但它们已经指向一个有些出人意料的趋势：仅仅校准概率似乎并不是向用户呈现不确定性的最佳方式。例如，Donahue et al. (2022) 在一个简单的理论框架中证明了人类和算法预测的互补性是发生提升的必要和充分条件。重要的是，模型校准（或准确性）并不总是能满足所需的条件。

在不确定性量化（UQ）领域还有大量的其他工作反映了类似的结果：Vodrahalli et al. (2022) 证明在人类从事一项真实任务时，若获得未经校准的算法建议，其表现优于获得已校准的建议。Cao et al. (2024) 也表明，模型校准并不是在人类研究中提升表现的充分条件，并集中于设计更有效的不确定性展示方案。最后，Corvelo Benz and Rodriguez (2023) 证明在聚合人类与算法的预测时，校准模型总是次优的。因此，他们建议获取不一定在总体上校准的模型，而是从正式意义上与人类对齐的模型，这种对齐是由人类和算法预测之间的关系定义的。随后，他们证明在人类研究中运行的合成任务中，人类对齐在经验上与提高的算法辅助人类绩效相关。最后，Vasconcelos et al. (2025b) 表明，LLM 不确定性量化方法可能需要进一步修改，以便为程序员在现实世界的编码任务中提供提升。

总的来说，这些研究揭示了在人机协作和 LLM 不确定性量化（UQ）研究之间的显著脱节。从表面上看，LLM UQ 研究声称专注于使语言模型更加可靠和值得信赖的目标 (Li et al., 2024) 。然而，在更深层次上，几乎没有直接确认 LLM UQ 是否真的对人类在实际决策任务中有用的研究。特别是，来自 HCI 和理论社区的人类能力提升研究似乎表明，校准概率——LLM UQ 的黄金标准——可能不是实现最佳人机协作的正确度量标准。尽管如此，许多 LLM UQ 论文仍然专注于在诸如事实问答或知识检索等任务上提供简单、校准的概率。虽然优化诸如 ECE 和 Brier 分数等 LLM UQ 指标可能已经对选择性预测 (Li et al., 2023) 或诸如推理 (Taubenfeld et al., 2025) 之类的纯能力研究有用，但优化这些相同指标并不意味着我们会“免费”获得更好的协作性人机模型性能。

LLM 的不确定性呈现新方案。正如在 Liao and Vaughan (2023, Section 3.5) 中所述，除了简单地向用户呈现概率或置信度之外，LLM 的不确定性量化技术可以通过特定利用语言媒介而显得与经典的不确定性量化不同。例如，LLM 不确定性量化方法可能会利用独特的方式引入更模糊的声明，以避免在高度不确定时完全放弃 (Jiang et al., 2025) ，或者使用语言嵌套 (Mielke et al., 2022) 以及拟人化的语言 (Kim et al., 2024) ，以非数值方式向用户传达不确定性。这些方法是 LLM 不确定性量化所独有的，应该进一步研究以了解它们如何影响人类在决策任务上的提升。

5.2 建议

为了改善当前 LLM UQ 评估指标存在的上述不足，我们倡导以下三个研究方向。通过这些提出的方法，我们推动 NLP 研究人员开发 LLM UQ 方法与 HCI 社区之间的更好结合。理想情况下，这将创建并加速两个社区之间的反馈循环，使得对可提供可衡量人类进步的 LLM UQ 方法进行更快速的迭代和研究。

R6. 借鉴现有经典不确定性量化研究中的关于人类校准对齐的发现，应用于不确定性增强的 LLM 建议。

R7. 开展研究，尝试理解 LLM 不确定性量化指标（如 ECE, Brier 分数）与人机团队表现之间的关系。

R8. 激励更多的研究，探索 LLM 不确定性呈现如何影响用户依赖。

建议 R6 相对简单，并基于 Section 5.1 中提到的众多涉及人类研究的 UQ 论文。通过 LLM UQ 方法展示生成文本的概率可能会与经典的 UQ 设置对用户产生不同的影响。语言回应可能提供解释和拟人化线索，而之前人类研究中使用的经典预测模型则没有。了解展示不确定性估计与 LLM 生成的回应同时对人类-AI 协作结果的影响至关重要。

我们的下一个推荐建议 R7 更加基础：它表明需要更好地理解 LLM UQ 希望优化的指标。对指标如 ECE 或 Brier 分数是否与用户的下游性能增加相关性进行全面分析，将会使社区受益。理解指标与下游实用性之间的这种关系，可以为 LLM UQ 社区提供信息，以确定是否继续优化诸如 ECE 之类的指标，或者可能转向通过人类研究识别为重要的其他可测量指标。例如，开创性的自然语言摘要指标 ROUGE 被证明与人类摘要判断相关，使研究人员更有信心认为它是一个合理的指标，可以在 (Lin, 2004) 上做优化。此外，在机器翻译中，其他指标如 BLEU 被证明表现出相反的行为——即，有时与人类判断相关性较差 (Callison-Burch et al., 2006) ——这将注意力转向开发新的指标 (Freitag et al., 2022) 。如果没有证据表明校准指标与用户的实用性相关，我们的社区在选择临时评估指标时就有可能“盲飞”。

最后，R8 指出了需要更好地理解大型语言模型不确定性呈现对用户依赖性的影响。研究人员需要了解哪种类型的大型语言模型不确定性量化（UQ）呈现方案（例如，显示声明级的数值概率 (Detommaso et al., 2024) ，使用模糊的语言不确定性 (Band et al., 2024) ，结合具有人性化暗示的缓和措施 (Kim et al., 2024) ）在不同的人类决策背景下表现最佳。此类研究将直接引导大型语言模型不确定性量化社区中的进一步研究，因此，应鼓励其在机器学习 / 自然语言处理会议上发表，而不应仅限于 HCI 场合，如 CHI 。

6 替代观点和结论

我们提出了两种可能的替代观点，即从事 LLM 不确定性量化研究的研究人员可能会对我们呼吁将该子领域重新聚焦于以人为中心的实用性做出不同看法。

当前的 LLM UQ 研究正在通过技术指标和较简单的基准建立必要的基础。复杂的评估可以稍后进行。

回复。仅依赖特定的指标或简单的基准测试，在研究子领域的初始阶段，可能导致其发展轨迹不完全符合人类为中心的实用性在下游决策中的最终目标。作为一个正在进行中的例子，我们依靠过去十年在可解释人工智能（XAI）方面的发展。在 XAI 中，最初以技术性和定量性为主的研究领域，转变为一个以人机交互为核心的领域。这种视角的改变只有在大量工作认识到以下几方面后才实现：1）总是提供模型解释有时可能会降低 AI 辅助人类 (Paleja et al., 2021; Buçinca et al., 2020; Alufaisan et al., 2021; Liao et al., 2022; Joshi et al., 2023) 的性能；2）学习何时及如何共同建模 AI 辅助和人类行为并不能总是通过纯技术解决方案来解决，而且高度依赖于具体情境 (Rong et al., 2023) 。确实，大语言模型的不确定性量化（LLM UQ）可以被视为 XAI 广泛领域中的一个研究途径，因此，我们应在此应用 XAI 中的许多见解。

当前的基准是可扩展的，并允许快速迭代。您提出的评估过于耗费资源，会减缓创新和评估的进程。

回应。虽然评估当前大型语言模型不确定性量化（LLM UQ）基准的成本肯定比进行人工研究更便宜，但部署那些削弱用户信任或导致决策系统效用降低的不确定性量化方案的长期成本可以说要大得多。过度激进和仓促的技术部署历史上对用户信任有持久影响。例如，2021 年，Wong et al. (2021) 显示医疗软件公司 Epic 的败血症分类工具在部署时表现不佳（部分原因是由于类似于 **??** 中描述的分布变动）。此外，由于系统实施的方式，临床医生和护士对系统的信任如此之低，以至于他们试图尽量减少系统在决策中的参与 (Gichoya et al., 2023; Habib et al., 2021) 。此类例子强调了以人为中心的技术开发的必要性，以便在重要的现实任务中实现有效和高效的人类提升 (Shneiderman, 2020) 。结论。在这项工作中，我们列出了我们认为是 LLM UQ 方法评估中的基础性问题。特别是，我们强调了三种问题实践：1）在生态有效性低的基准上进行评估；2）仅考虑认知不确定性；3）在不一定能反映下游效用的指标上进行爬山演算法。我们进一步提供了克服这些问题的具体建议，无论是通过修订的实践还是新的研究方向。尽管问题众多，但我们相信，通过许多研究人员采用更以人为本的视角，这些问题可以得到解决。

References

Aichberger, L., Schweighofer, K., Ielanskyi, M., and Hochreiter, S. (2025). Improving uncertainty estimation through semantically diverse language generation. In *The Thirteenth International Conference on Learning Representations*.

Alufaisan, Y., Marusich, L. R., Bakdash, J. Z., Zhou, Y., and Kantarcioglu, M. (2021). Does explainable artificial intelligence improve human decision-making? In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Alur, R., Laine, L., Li, D., Raghavan, M., Shah, D., and Shung, D. (2023). Auditing for human expertise. *Advances in Neural Information Processing Systems*, 36:79439–79468.

Aolaritei, L., Jordan, M. I., Marzouk, Y., Wang, Z. O., and Zhu, J. (2025). Conformal prediction under l\'evy-prokhorov distribution shifts: Robustness to local and global perturbations. *arXiv preprint arXiv:2502.14105*.

Bakman, Y. F., Yaldiz, D. N., Buyukates, B., Tao, C., Dimitriadis, D., and Avestimehr, S. (2024). MARS: meaning-aware response scoring for uncertainty estimation in generative llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*.

Balepur, N., Rudinger, R., and Boyd-Graber, J. L. (2025). Which of these best describes multiple choice evaluation with llms? a) forced b) flawed c) fixable d) all of the above. *arXiv preprint arXiv:2502.14127*.

Band, N., Li, X., Ma, T., and Hashimoto, T. (2024). Linguistic calibration of long-form generations. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. (2010). A theory of learning from different domains. *Machine learning*, 79:151–175.

Błasiok, J. and Nakkiran, P. (2024). Smooth ECE: principled reliability diagrams via kernel smoothing. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.

Brennan, T., Dieterich, W., and Ehret, B. (2009). Evaluating the predictive validity of the compas risk and needs assessment system. *Criminal Justice and behavior*, 36(1):21–40.

Bronfenbrenner, U. (1977). Toward an experimental ecology of human development. *American psychologist*, 32(7):513.

Buçinca, Z., Lin, P., Gajos, K. Z., and Glassman, E. L. (2020). Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th international conference on intelligent user interfaces*, pages 454–464.

Callison-Burch, C., Osborne, M., and Koehn, P. (2006). Re-evaluating the role of bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256. Association for Computational Linguistics.

Cao, S., Liu, A., and Huang, C.-M. (2024). Designing for appropriate reliance: The roles of ai uncertainty presentation, initial user decision, and user demographics in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–32.

Chaleshtori, F. H., Ghosal, A., Gill, A., Bambroo, P., and Marasović, A. (2024). On evaluating explanation utility for human-ai decision making in nlp. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7456–7504.

Chang, Y., Narang, M., Suzuki, H., Cao, G., Gao, J., and Bisk, Y. (2022). Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*.

Chen, J. and Mueller, J. (2023). Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. *arXiv preprint arXiv:2308.16175*.

Chen, J., Yoon, J., Ebrahimi, S., Arik, S. O., Pfister, T., and Jha, S. (2023). Adaptation with self-evaluation to improve selective prediction in llms. *arXiv preprint arXiv:2310.11689*.

Chen, L., Perez-Lebel, A., Suchanek, F., and Varoquaux, G. (2024). Reconfidencing llm uncertainty from the grouping loss perspective. In *The 2024 Conference on Empirical Methods in Natural Language Processing*. arXiv.

Cherian, J., Gibbs, I., and Candes, E. (2024). Large language model validity via enhanced conformal prediction methods. *Advances in Neural Information Processing Systems*, 37:114812–114842.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Cole, J. R., Zhang, M. J., Gillick, D., Eisenschlos, J. M., Dhingra, B., and Eisenstein, J. (2023). Selectively answering ambiguous questions. *arXiv preprint arXiv:2305.14613*.

Corvelo Benz, N. and Rodriguez, M. (2023). Human-aligned calibration for ai-assisted decision making. *Advances in Neural Information Processing Systems*, 36:14609–14636.

Corvelo Benz, N. L. and Gomez Rodriguez, M. (2025). Human-alignment influences the utility of ai-assisted decision making. *arXiv preprint arXiv:2501.14035*.

Cruz, A. F., Hardt, M., and Mendler-Dünner, C. (2024). Evaluating language models as risk scores. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

De-Arteaga, M., Fogliato, R., and Chouldechova, A. (2020). A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–12.

De Vries, H., Bahdanau, D., and Manning, C. (2020). Towards ecologically valid research on language user interfaces. *arXiv preprint arXiv:2007.14435*.

Detommaso, G., Lopez, M. B., Fogliato, R., and Roth, A. (2024). Multicalibration for confidence scoring in llms. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*.

Donahue, K., Chouldechova, A., and Kenthapadi, K. (2022). Human-algorithm collaboration: Achieving complementarity and avoiding unfairness. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1639–1656.

Duan, J., Cheng, H., Wang, S., Zavalny, A., Wang, C., Xu, R., Kailkhura, B., and Xu, K. (2023). Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. *arXiv preprint arXiv:2307.01379*.

Dunn, J., Adams, B., and Madabushi, H. T. (2024). Pre-trained language models represent some geographic populations better than others. *arXiv preprint arXiv:2403.11025*.

Freitag, M., Rei, R., Mathur, N., Lo, C.-k., Stewart, C., Avramidis, E., Kocmi, T., Foster, G., Lavie, A., and Martins, A. F. (2022). Results of wmt22 metrics shared task: Stop using bleu–neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68.

Geifman, Y. and El-Yaniv, R. (2017). Selective classification for deep neural networks. *Advances in neural information processing systems*, 30.

Geva, M., Khashabi, D., Segal, E., Khot, T., Roth, D., and Berant, J. (2021). Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*.

Gichoya, J. W., Thomas, K., Celi, L. A., Safdar, N., Banerjee, I., Banja, J. D., Seyyed-Kalantari, L., Trivedi, H., and Purkayastha, S. (2023). Ai pitfalls and what not to do: mitigating bias in ai. *The British Journal of Radiology*, 96(1150):20230023.

Groot, T. and Valdenegro-Toro, M. (2024). Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models. *arXiv preprint arXiv:2405.02917*.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Habib, A. R., Lin, A. L., and Grant, R. W. (2021). The epic sepsis model falls short—the importance of external validation. *JAMA Internal Medicine*, 181(8):1040–1041.

Hansen, D., Devic, S., Nakkiran, P., and Sharan, V. (2024). When is multicalibration post-processing necessary? In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., and Kamar, E. (2022). Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*.

He, G., Chen, J., and Zhu, J. (2023a). Preserving pre-trained features helps calibrate fine-tuned language models. *arXiv preprint arXiv:2305.19249*.

He, J., Yang, R., Yu, L., Li, C., Jia, R., Chen, F., Jin, M., and Lu, C.-T. (2024). Can we trust the performance evaluation of uncertainty estimation methods in text summarization? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16514–16575.

He, W., Jiang, Z., Xiao, T., Xu, Z., and Li, Y. (2023b). A survey on uncertainty quantification methods for deep learning. *arXiv preprint arXiv:2302.13425*.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2021). Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Heo, J., Xiong, M., Heinze-Deml, C., and Narain, J. (2024). Do llms estimate uncertainty well in instruction-following? *arXiv preprint arXiv:2410.14582*.

Hershcovich, D., Frank, S., Lent, H. C., de Lhoneux, M., Abdou, M., Brandl, S., Bugliarello, E., Piqueras, L. C., Chalkidis, I., Cui, R., Fierro, C., Margatina, K., Rust, P., and Søgaard, A. (2022). Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*. Association for Computational Linguistics.

Hou, Y., Li, J., He, Z., Yan, A., Chen, X., and McAuley, J. (2024). Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*.

Hu, Z., Liu, C., Feng, X., Zhao, Y., Ng, S.-K., Luu, A. T., He, J., Koh, P. W. W., and Hooi, B. (2024). Uncertainty of thoughts: Uncertainty-aware planning enhances information seeking in llms. *Advances in Neural Information Processing Systems*, 37:24181–24215.

Huang, H.-Y., Yang, Y., Zhang, Z., Lee, S., and Wu, Y. (2024). A survey of uncertainty estimation in LLMs: Theory meets practice. *arXiv preprint arXiv:2410.15326*.

Hüllermeier, E. and Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506.

Jiang, Z., Liu, A., and Van Durme, B. (2025). Conformal linguistic calibration: Trading-off between factuality and specificity. *arXiv preprint arXiv:2502.19110*.

Joshi, B., Liu, Z., Ramnath, S., Chan, A., Tong, Z., Nie, S., Wang, Q., Choi, Y., and Ren, X. (2023). Are machine rationales (not) useful to humans? measuring and improving human utility of free-text rationales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*.

Joshi, M., Choi, E., Weld, D. S., and Zettlemoyer, L. (2017). Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., et al. (2022). Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Kapoor, S. and Narayanan, A. (2025). Ai as normal technology. *25-09 Knight First Amend. Inst.*

Khatun, A. and Brown, D. G. (2024). A study on large language models' limitations in multiple-choice question answering. *arXiv preprint arXiv:2401.07955*.

Kim, S. S., Liao, Q. V., Vorvoreanu, M., Ballard, S., and Vaughan, J. W. (2024). " i'm not sure, but...": Examining the impact of large language models' uncertainty expression on user reliance and trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 822–835.

Kuhn, L., Gal, Y., and Farquhar, S. (2022). Clam: Selective clarification for ambiguous questions with generative language models. *arXiv preprint arXiv:2212.07769*.

Kuhn, L., Gal, Y., and Farquhar, S. (2023). Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*.

Kumar, A., Morabito, R., Umbet, S., Kabbara, J., and Emami, A. (2024). Confidence under the hood: An investigation into the confidence-probability alignment in large language models. *arXiv preprint arXiv:2405.16282*.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., et al. (2019). Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Lee, K., Chang, M.-W., and Toutanova, K. (2019). Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Li, M., Shi, T., Ziems, C., Kan, M.-Y., Chen, N. F., Liu, Z., and Yang, D. (2023). Coannotating: Uncertainty-guided work allocation between human and large language models for data annotation. *arXiv preprint arXiv:2310.15638*.

Li, Y., Qiang, R., Moukheiber, L., and Zhang, C. (2025). Language model uncertainty quantification with attention chain. *arXiv preprint arXiv:2503.19168*.

Li, Y., Wang, S., Huang, L., and Liu, L.-P. (2024). Graph-based confidence calibration for large language models. *arXiv preprint arXiv:2411.02454*.

Liao, Q. V. and Vaughan, J. W. (2023). Ai transparency in the age of llms: A human-centered research roadmap. *arXiv preprint arXiv:2306.01941*, 10.

Liao, Q. V., Zhang, Y., Luss, R., Doshi-Velez, F., and Dhurandhar, A. (2022). Connecting algorithmic research and usage contexts: a perspective of contextualized evaluation for explainable ai. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, pages 147–159.

Lin, B. Y., Deng, Y., Chandu, K., Ravichander, A., Pyatkin, V., Dziri, N., Le Bras, R., and Choi, Y. (2023). Wildbench: Benchmarking llms with challenging tasks from real users in the wild. In *The Thirteenth International Conference on Learning Representations*.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Lin, S., Hilton, J., and Evans, O. (2022). Teaching models to express their uncertainty in words. *Trans. Mach. Learn. Res.*, 2022.

Lin, Z., Trivedi, S., and Sun, J. (2024). Generating with confidence: Uncertainty quantification for black-box large language models. *Trans. Mach. Learn. Res.*, 2024.

Liu, L., Pan, Y., Li, X., and Chen, G. (2024a). Uncertainty estimation and quantification for llms: A simple supervised approach. *arXiv preprint arXiv:2404.15993*.

Liu, O., Fu, D., Yogatama, D., and Neiswanger, W. (2024b). Dellma: A framework for decision making under uncertainty with large language models. *arXiv preprint arXiv:2402.02392*, 1.

Liu, T. and Wu, Z. S. (2024). Multi-group uncertainty quantification for long-form text generation. *arXiv preprint arXiv:2407.21057*.

Liu, X., Chen, T., Da, L., Chen, C., Lin, Z., and Wei, H. (2025). Uncertainty quantification and confidence calibration in large language models: A survey. *arXiv preprint arXiv:2503.15850*.

Ma, M., Chochlakis, G., Pandiyan, N. M., Thomason, J., and Narayanan, S. (2025). Large language models do multi-label classification differently. *arXiv*.

Malinin, A. and Gales, M. J. F. (2021). Uncertainty estimation in autoregressive structured prediction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Manakul, P., Liusie, A., and Gales, M. J. F. (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, EMNLP 2023, Singapore, December 6-10, 2023*.

Manggala, P., Mastakouri, A. A., Kirschbaum, E., Kasiviswanathan, S., and Ramdas, A. (2025). Qa-calibration of language model confidence scores. In *The Thirteenth International Conference on Learning Representations*.

Marusich, L., Bakdash, J. Z., Zhou, Y., and Kantarcioglu, M. (2024). Using AI uncertainty quantification to improve human decision-making. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*.

Mielke, S. J., Szlam, A., Dinan, E., and Boureau, Y.-L. (2022). Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872.

Min, S., Michael, J., Hajishirzi, H., and Zettlemoyer, L. (2020). AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Nie, Y., Zhou, X., and Bansal, M. (2020). What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Nikitin, A., Kossen, J., Gal, Y., and Marttinen, P. (2024). Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. *Advances in Neural Information Processing Systems*, 37:8901–8929.

Ouyang, S., Wang, S., Liu, Y., Zhong, M., Jiao, Y., Iter, D., Pryzant, R., Zhu, C., Ji, H., and Han, J. (2023). The shifted and the overlooked: A task-oriented investigation of user-gpt interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Paleja, R., Ghuy, M., Ranawaka Arachchige, N., Jensen, R., and Gombolay, M. (2021). The utility of explainable ai in ad hoc human-machine teaming. *Advances in neural information processing systems*, 34:610–623.

Quach, V., Fisch, A., Schuster, T., Yala, A., Sohn, J. H., Jaakkola, T. S., and Barzilay, R. (2024). Conformal language modeling. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Rong, Y., Leemann, T., Nguyen, T.-T., Fiedler, L., Qian, P., Unhelkar, V., Seidel, T., Kasneci, G., and Kasneci, E. (2023). Towards human-centered explainable ai: A survey of user studies for model explanations. *IEEE transactions on pattern analysis and machine intelligence*, 46(4):2104–2122.

Rücker, S. and Akbik, A. (2023). Cleanconll: A nearly noise-free named entity recognition dataset. *arXiv preprint arXiv:2310.16225*.

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6):495–504.

Shorinwa, O., Mei, Z., Lidard, J., Ren, A. Z., and Majumdar, A. (2024). A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *arXiv preprint arXiv:2412.05563*.

Smith, R. C. (2024). *Uncertainty quantification: theory, implementation, and applications*. SIAM.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Srinivasan, T. and Thomason, J. (2025). Adjust for trust: Mitigating trust-induced inappropriate reliance on ai assistance. *arXiv preprint arXiv:2502.13321*.

Stengel-Eskin, E., Hase, P., and Bansal, M. (2024). Lacie: Listener-aware finetuning for confidence calibration in large language models. *arXiv preprint arXiv:2405.21028*.

Sullivan, T. J. (2015). *Introduction to uncertainty quantification*, volume 63. Springer.

Sung, Y. Y., Fleisig, E., Hou, Y., Upadhyay, I., and Boyd-Graber, J. L. (2025). Grace: A granular benchmark for evaluating model calibration against human calibration. *arXiv preprint arXiv:2502.19684*.

Talmor, A., Herzig, J., Lourie, N., and Berant, J. (2019). Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158.

Taubenfeld, A., Sheffer, T., Ofek, E., Feder, A., Goldstein, A., Gekhman, Z., and Yona, G. (2025). Confidence improves self-consistency in llms. *arXiv preprint arXiv:2502.06233*.

Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C., and Manning, C. D. (2023). Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*.

Ulmer, D., Gubri, M., Lee, H., Yun, S., and Oh, S. J. (2024). Calibrating large language models using their generations only. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*.

Valdenegro-Toro, M. and Stoykova, R. (2024). The dilemma of uncertainty estimation for general purpose ai in the eu ai act. *arXiv preprint arXiv:2408.11249*.

Vasconcelos, H., Bansal, G., Fourney, A., Liao, Q. V., and Wortman Vaughan, J. (2025a). Generation probabilities are not enough: Uncertainty highlighting in ai code completions. *ACM Trans. Comput.-Hum. Interact.*

Vasconcelos, H., Bansal, G., Fourney, A., Liao, Q. V., and Wortman Vaughan, J. (2025b). Generation probabilities are not enough: Uncertainty highlighting in ai code completions. *ACM Transactions on Computer-Human Interaction*, 32(1):1–30.

Vazhentsev, A., Fadeeva, E., Xing, R., Panchenko, A., Nakov, P., Baldwin, T., Panov, M., and Shelmanov, A. (2024). Unconditional truthfulness: Learning conditional dependency for uncertainty quantification of large language models. *arXiv preprint arXiv:2408.10692*.

Vazhentsev, A., Kuzmin, G., Tsvigun, A., Panchenko, A., Panov, M., Burtsev, M., and Shelmanov, A. (2023). Hybrid uncertainty quantification for selective text classification in ambiguous tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11659–11681.

Vodrahalli, K., Gerstenberg, T., and Zou, J. Y. (2022). Uncalibrated models can improve human-ai collaboration. *Advances in Neural Information Processing Systems*, 35:4004–4016.

Wang, P., Lam, B. D., Liu, Y., Asgari-Targhi, A., Panda, R., Wells, W. M., Kapur, T., and Golland, P. (2025a). Calibrating expressions of certainty. *International Conference on Learning Representations*.

Wang, Z., Wang, Q., Zhang, Y., Chen, T., Zhu, X., Shi, X., and Xu, K. (2025b). Sconu: Selective conformal uncertainty in large language models. *arXiv preprint arXiv:2504.14154*.

Wong, A., Otles, E., Donnelly, J. P., Krumm, A., McCullough, J., DeTroyer-Cooley, O., Pestrue, J., Phillips, M., Konye, J., Penoza, C., et al. (2021). External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA internal medicine*, 181(8):1065–1070.

Xiao, Y., Liang, P. P., Bhatt, U., Neiswanger, W., Salakhutdinov, R., and Morency, L.-P. (2022). Uncertainty quantification with pre-trained language models: A large-scale empirical analysis. *arXiv preprint arXiv:2210.04714*.

Xing, Z., Yuan, X., and Mostafa, J. (2022). Age-related difference in conversational search behavior: Preliminary findings. In *CHIIR '22: ACM SIGIR Conference on Human Information Interaction and Retrieval, Regensburg, Germany, March 14 - 18, 2022*.

Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., and Hooi, B. (2023). Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*.

Yaldiz, D. N., Bakman, Y. F., Buyukates, B., Tao, C., Ramakrishna, A., Dimitriadis, D., Zhao, J., and Avestimehr, S. (2024). Do not design, learn: A trainable scoring function for uncertainty estimation in generative llms. *arXiv preprint arXiv:2406.11278*.

Yang, A., Chen, C., and Pitas, K. (2024a). Just rephrase it! uncertainty estimation in closed-source language models via multiple rephrased queries. *arXiv preprint arXiv:2405.13907*.

Yang, A. X., Robeyns, M., Wang, X., and Aitchison, L. (2024b). Bayesian low-rank adaptation for large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.

Yang, D., Tsai, Y.-H. H., and Yamada, M. (2024c). On verbalized confidence scores for llms. *arXiv preprint arXiv:2412.14737*.

Ye, Z., Melo, L. C., Kaddar, Y., Blunsom, P., Staton, S., and Gal, Y. (2025). Uncertainty-aware step-wise verification with generative reward models. *arXiv preprint arXiv:2502.11250*.

Yona, G., Aharoni, R., and Geva, M. (2024). Can large language models faithfully express their intrinsic uncertainty in words? *arXiv preprint arXiv:2405.16908*.

Yuksekgonul, M., Zhang, L., Zou, J. Y., and Guestrin, C. (2023). Beyond confidence: Reliable models should also consider atypicality. *Advances in Neural Information Processing Systems*, 36:38420–38453.

Zhang, M. J. and Choi, E. (2021). Situatedqa: Incorporating extra-linguistic contexts into qa. *arXiv preprint arXiv:2109.06157*.

Zhang, T., Qiu, L., Guo, Q., Deng, C., Zhang, Y., Zhang, Z., Zhou, C., Wang, X., and Fu, L. (2023). Enhancing uncertainty-based hallucination detection with stronger focus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*.

Zhao, W., Ren, X., Hessel, J., Cardie, C., Choi, Y., and Deng, Y. (2024). Wildchat: 1m chatgpt interaction logs in the wild. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Zheng, C., Zhou, H., Meng, F., Zhou, J., and Huang, M. (2024). Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

可能有其他原因使得在衡量大型语言模型（LLM）的校准误差时，避免仅仅优化 ECE。近期一系列相关研究探讨了衡量校准误差的基本问题。从总体上看，这些研究表明，ECE——最流行的校准误差概念——可能非常不稳定且不健壮。这是因为在计算 ECE 时，由于实值概率预测的可测性问题，必须进行离散化/分箱过程。相反，Błasiok and Nakkiran (2024) 提出了一种相关但更为稳健的概念，称为 SmoothECE (smECE)，通过应用核平滑来绕开分箱要求。Hansen et al. (2024) 进一步表明，使用 smECE 而不是 ECE 可以改变关于哪个模型在特定任务上更好校准的结论，尤其是在小样本情况下（<5k 数据点）。

然而，ECE 仍然是衡量 LLM UQ 方法校准误差的一个流行指标。我们发现有 16 篇论文使用 ECE 来衡量校准误差，此外，其中许多论文的评估集小于 5000 个数据点。例如，Li et al. (2024) 和 Yang et al. (2024b) 都使用 ECE 来衡量约 2000 个数据点的评估集的性能。在这种数据集规模下，应该优先使用 smECE。在我们分析的所有论文中，唯一利用 smECE 来测量校准误差的论文是 Ulmer et al. (2024) 。实际上，Ulmer et al. (2024, Table 3) 表明，用 smECE 与 ECE 测量校准误差会对应用于 Vicuna 和 GPT-3.5 的 LLM UQ 方法的质量排序产生不同的结果。由于在更大数据集规模下获得 LLM 预测的难度和成本，我们认为社区将受益于将 smECE 作为 ECE 的直接替代品。这样做可能有助于确保在测试数据集规模受限的情况下，校准误差的任何技术进步仍然具有鲁棒性和统计显著性。

A 大型语言模型 UQ 方法注释

在本节中，我们为每篇包含在我们调查中的 LLM UQ 方法论文提供注释。

P1. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs (Xiong et al., 2023)

(a) 不确定性量化的类型（短/长形式生成、保形等）。
- 通过提示、采样和聚合技术实现口头化信心。

(b) 用于评估所提出的 LLM UQ 方法的数据集。
- SportUND（常识）、StrategyQA（常识）、GSM8K、SVAMP、日期理解（符号推理）、物体计数（符号推理）、MMLU-法律、MMLU-商业伦理

(c) 在分布偏移下测试表现？（是/否）。对于无监督的方法，也可以是 N/A。
- 不适用

(d) 有监督、半监督或无监督的方法？
- 无监督，但有可能影响方法结果的超参数。

(e) 用于评估性能的指标。
- ECE, AUROC, AUPRC

(f) 人类提升研究？
- 否

P2. Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models (Lin et al., 2024)

(a) 不确定性量化的类型（短/长形式生成、保形等）。
- 一种用于衡量生成短格式回复集合的“语义分散性”的简单方法。可以作为预测 LLM 回复质量的可靠指标。

(b) 用于评估所提出的 LLM UQ 方法的数据集。
- CoQA，TriviaQA，Natural Questions

(c) 在分布偏移下测试表现？（是/否）。对于无监督的方法，也可以是 N/A。
- N/A

(d) 有监督、半监督或无监督的方法？
- 无监督

(e) 用于评估性能的指标。
- AUROC, AUARC

(f) 人类提升研究？
- 不

P3. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation (Kuhn et al., 2023)

(a) 不确定性量化的类型（短/长形式生成、保形等）。
- 通过聚类多代并计算熵的简化不确定性量化。

(b) 用于评估所提出的 LLM UQ 方法的数据集。
- CoQA, TriviaQA

(c) 在分布偏移下测试表现？（是/否）。对于无监督的方法，也可以是 N/A。
- 不适用

(d) 有监督、半监督或无监督的方法？
- 无监督

(e) 用于评估性能的指标。
- 曲线下面积（AUROC）

(f) 人类提升研究？
- 否

P4. Shifting Attention to Relevance: Towards the Predictive Uncertainty Quantification of Free-Form Large Language Models (Duan et al., 2023)

(a) 不确定性量化的类型（短/长形式生成、保形等）。
- 通过多个生成进行简化形式的不确定性量化，并在标记/句子级别重新加权。

(b) 用于评估所提出的 LLM UQ 方法的数据集。
- CoQA，TriviaQA，SciQ，MedQA，MedMCQA

(c) 在分布偏移下测试表现？（是/否）。对于无监督的方法，也可以是 N/A。
- 不适用

(d) 有监督、半监督或无监督的方法？
- 无监督

(e) 用于评估性能的指标。
- 受试者工作特征曲线下面积

(f) 人类提升研究？
- 否

P5. Kernel Language Entropy: Fine-grained Uncertainty Quantification for LLMs from Semantic Similarities (Nikitin et al., 2024)

(a) 不确定性量化的类型（短/长形式生成、保形等）。
- 通过对多代进行聚类和以比语义熵更精细的方法计算熵来实现短格式的不确定性量化（是该工作的推广）

(b) 用于评估所提出的 LLM UQ 方法的数据集。
- TriviaQA，SQuAD，BioASQ，NQ，SVAMP

(c) 在分布偏移下测试表现？（是/否）。对于无监督的方法，也可以是 N/A。
- 不适用

(d) 有监督、半监督或无监督的方法？
- 无监督

(e) 用于评估性能的指标。
- ROC 曲线下面积, AR 曲线下面积

(f) 人类提升研究？
- 否

P6. Mars: Meaning-aware response scoring for uncertainty estimation in generative llms (Bakman et al., 2024)

(a) 不确定性量化的类型（短/长形式生成、保形等）。
- 短问答回复的无监督不确定性量化。

(b) 用于评估所提出的 LLM UQ 方法的数据集。
- TriviaQA，自然问题，WebQA

(c) 在分布偏移下测试表现？（是/否）。对于无监督的方法，也可以是 N/A。
- N/A

(d) 有监督、半监督或无监督的方法？

- 无监督

(e) 用于评估性能的指标。
- 曲线下面积（AUROC）

(f) 人类提升研究?
- 否

P7. Uncertainty Estimation in Autoregressive Structured Prediction (Malinin and Gales, 2021)

(a) 不确定性量化的类型（短/长形式生成、保形等）。
- 无监督 UQ 用于短回答问题。

(b) 用于评估所提出的 LLM UQ 方法的数据集。
- newstest14 和 LibriSpeech（翻译数据集）

(c) 在分布偏移下测试表现?（是/否）。对于无监督的方法，也可以是 N/A。
- 是的（超出预训练领域）

(d) 有监督、半监督或无监督的方法?
- 无监督

(e) 用于评估性能的指标。
- NLL, AUPR, PRR

(f) 人类提升研究?
- 否

P8. Quantifying Uncertainty in Answers from Any Language Model and Enhancing Their Trustworthiness (Chen and Mueller, 2023)

(a) 不确定性量化的类型（短/长形式生成、保形等）。
- 无监督短问答响应的不确定性量化。

(b) 用于评估所提出的 LLM UQ 方法的数据集。
- GSM8k、SVAMP、CSQA、TriviaQA

(c) 在分布偏移下测试表现?（是/否）。对于无监督的方法，也可以是 N/A。
- 不适用

(d) 有监督、半监督或无监督的方法?
- 无监督

(e) 用于评估性能的指标。
- 曲线下面积（AUROC）

(f) 人类提升研究?
- 否

P9. Enhancing Uncertainty-Based Hallucination Detection with Stronger Focus (Zhang et al., 2023)

(a) 不确定性量化的类型（短/长形式生成、保形等）。
- 无监督、无参考的 UQ 方法，用于在句子级别检测幻觉。

(b) 用于评估所提出的 LLM UQ 方法的数据集。
- WikiBio

(c) 在分布偏移下测试表现?（是/否）。对于无监督的方法，也可以是 N/A。
- 不适用

(d) 有监督、半监督或无监督的方法?
- 无监督

(e) 用于评估性能的指标。
- AUC-PR

(f) 人类提升研究?
- 否

P10. Language Models (Mostly) Know What They Know (Kadavath et al., 2022)

(a) 不确定性量化的类型（短/长形式生成、保形等）。
- 短形式 UQ 中的 P(True) 标记概率，也即 UQ 微调。

(b) 用于评估所提出的 LLM UQ 方法的数据集。
- TriviaQA、Lambada、算术、GSM8k 和 Codex HumanEval

(c) 在分布偏移下测试表现?（是/否）。对于无监督的方法，也可以是 N/A。
- 是

(d) 有监督、半监督或无监督的方法?
- 有监督和无监督

(e) 用于评估性能的指标。
- 电子与计算机工程

(f) 人类提升研究?
- 否

P11. SELFCHECKGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models (Manakul et al., 2023)

(a) 不确定性量化的类型（短/长形式生成、保形等）。
- 通过 UQ 方法检测长篇生成文本中的幻觉。

(b) 用于评估所提出的 LLM UQ 方法的数据集。
- WikiBio

(c) 在分布偏移下测试表现?（是/否）。对于无监督的方法，也可以是 N/A。
- 不适用

(d) 有监督、半监督或无监督的方法?
- 无监督

(e) 用于评估性能的指标。
- PR 曲线下面积

(f) 人类提升研究?
- 否

P12. Graph-based confidence calibration for large language models (Li et al., 2024)

(a) 不确定性量化的类型（短/长形式生成、保形等）。
- 生成简短形式的置信度。

(b) 用于评估所提出的 LLM UQ 方法的数据集。
- CoQA, TriviaQA

(c) 在分布偏移下测试表现?（是/否）。对于无监督的方法，也可以是 N/A。
- 是

(d) 有监督、半监督或无监督的方法?
- 监督

(e) 用于评估性能的指标。
- ECE, AUROC, Brier

(f) 人类提升研究?
- 否

P13. Calibrating Large Language Models Using Their Generations Only (Ulmer et al., 2024)

(a) 不确定性量化的类型（短/长形式生成、保形等）。
- 训练辅助模型以预测大型语言模型的短格式生成的置信度

(b) 用于评估所提出的 LLM UQ 方法的数据集。
- TriviaQA

(c) 在分布偏移下测试表现?（是/否）。对于无监督的方法，也可以是 N/A。
- 否

(d) 有监督、半监督或无监督的方法?
- 监督的

(e) 用于评估性能的指标。
- Brier，ECE，smECE，AUROC

(f) 人类提升研究?
- 否

P14. Bayesian Low-Rank Adaptation for Large Language Models (Yang et al., 2024b)

(a) 不确定性量化的类型（短/长形式生成、保形等）。

- 训练在目标任务上微调的 LoRA 贝叶斯集成。

(b) 用于评估所提出的 LLM UQ 方法的数据集。
- MC / 常识推理任务：WG-S, ARC-C, ARC-E, OBQA, WG-M, BoolQ, MMLU

(c) 在分布偏移下测试表现？（是/否）。对于无监督的方法，也可以是 N/A。
- 是

(d) 有监督、半监督或无监督的方法？
- 监督的

(e) 用于评估性能的指标。
- 准确率，期望校准误差，负对数似然

(f) 人类提升研究？
- 没有

P15. Reconfidencing LLM Uncertainty from the Grouping Loss Perspective (Chen et al., 2024)

(a) 不确定性量化的类型（短/长形式生成、保形等）。
- 答案的不确定性量化简表

(b) 用于评估所提出的 LLM UQ 方法的数据集。
- 提出新的数据集，简短答案但出于公平目的而衍生（具有组信息）

(c) 在分布偏移下测试表现？（是/否）。对于无监督的方法，也可以是 N/A。
- 否

(d) 有监督、半监督或无监督的方法？
- 监督

(e) 用于评估性能的指标。
- 布里尔、CL、GL

(f) 人类提升研究？
- 否

P16. Reducing conversational agents' overconfidence through linguistic calibration (Mielke et al., 2022)

(a) 不确定性量化的类型（短/长形式生成、保形等）。
- 训练一个“校准器”，该校准器试图根据 LLM 的内部状态预测其正确性。

(b) 用于评估所提出的 LLM UQ 方法的数据集。
- TriviaQA

(c) 在分布偏移下测试表现？（是/否）。对于无监督的方法，也可以是 N/A。
- 否

(d) 有监督、半监督或无监督的方法？
- 监督的

(e) 用于评估性能的指标。
- ECE, 最大 CE, NLL

(f) 人类提升研究？
- 否

P17. Multicalibration for Confidence Scoring in LLMs (Detommaso et al., 2024)

(a) 不确定性量化的类型（短/长形式生成、保形等）。
- 短问答的监督不确定性量化

(b) 用于评估所提出的 LLM UQ 方法的数据集。
- BigBench、MMLU、triviaQA、OpenBookQA、TruthfulQA 和 MathQA

(c) 在分布偏移下测试表现？（是/否）。对于无监督的方法，也可以是 N/A。
- 否

(d) 有监督、半监督或无监督的方法？
- 监督学习

(e) 用于评估性能的指标。
- 组平均平方校准误差（gASCE），ASCE（类似于 ECE）。

(f) 人类提升研究？

- 否

P18. Teaching Models to Express Their Uncertainty in Words (Lin et al., 2022)

(a) 不确定性量化的类型（短/长形式生成、保形等）。
- 提示 LLM 生成一系列符号作为置信度

(b) 用于评估所提出的 LLM UQ 方法的数据集。
- 校准数学

(c) 在分布偏移下测试表现？（是/否）。对于无监督的方法，也可以是 N/A。
- 是

(d) 有监督、半监督或无监督的方法？
- 监督

(e) 用于评估性能的指标。
- 布里尔, 期望分类误差（ECE）

(f) 人类提升研究？
- 否

P19. Do Not Design, Learn: A Trainable Scoring Function for Uncertainty Estimation in Generative LLMs (Yaldiz et al., 2024)

(a) 不确定性量化的类型（短/长形式生成、保形等）。
- 可训练/可学习的响应评分函数用于 LLMs。后续工作在 (Bakman et al., 2024) 。

(b) 用于评估所提出的 LLM UQ 方法的数据集。
- TriviaQA NaturalQA WebQA GSM8K

(c) 在分布偏移下测试表现？（是/否）。对于无监督的方法，也可以是 N/A。
- 是

(d) 有监督、半监督或无监督的方法？
- 监督

(e) 用于评估性能的指标。
- AUROC, PRR

(f) 人类提升研究？
- 否

P20. Conformal Linguistic Calibration: Trading-off between Factuality and Specificity (Jiang et al., 2025)

(a) 不确定性量化的类型（短/长形式生成、保形等）。
- 修改输出，使所用语言更加模糊，以更好地表示模型的不确定性。

(b) 用于评估所提出的 LLM UQ 方法的数据集。
- SimpleQA，自然问题

(c) 在分布偏移下测试表现？（是/否）。对于无监督的方法，也可以是 N/A。
- 否

(d) 有监督、半监督或无监督的方法？
- 监督

(e) 用于评估性能的指标。
- 保形覆盖/事实性要求。

(f) 人类提升研究？
- 否

P21. Calibrating Expressions of Certainty (Wang et al., 2025a)

(a) 不确定性量化的类型（短/长形式生成、保形等）。
- 校准人类和语言模型中的确定性语言表达，例如，“也许”和“可能”。

(b) 用于评估所提出的 LLM UQ 方法的数据集。
- SciQ, TruthfulQA

(c) 在分布偏移下测试表现？（是/否）。对于无监督的方法，也可以是 N/A。
- 否

(d) 有监督、半监督或无监督的方法？

- 监督

(e) 用于评估性能的指标。
- ECE 和 Brier 分数。

(f) 人类提升研究?
- 在人类上的测试方法。

P22. QA-Calibration of Language Model Confidence Scores (Manggala et al., 2025)

(a) 不确定性量化的类型（短/长形式生成、保形等）。
- 问答的组别校准保证

(b) 用于评估所提出的 LLM UQ 方法的数据集。
- MMLU，TriviaQA，SciQ，BigBench，OpenBookQA

(c) 在分布偏移下测试表现?（是/否）。对于无监督的方法，也可以是 N/A。
- 否

(d) 有监督、半监督或无监督的方法?
- 监督

(e) 用于评估性能的指标。
- AUAC，校准误差

(f) 人类提升研究?
- 否

P23. Unconditional Truthfulness: Learning Conditional Dependency for Uncertainty Quantification of Large Language Models (Vazhentsev et al., 2024)

(a) 不确定性量化的类型（短/长形式生成、保形等）。
- 学习生成步骤之间的条件依赖性，以估计不确定性。

(b) 用于评估所提出的 LLM UQ 方法的数据集。
- 文本生成：CNN/DailyMail、XSUM、SamSum 长篇问答：MedQUAD、PubMedQA、TruthfulQA 短篇问答：SciQ、CoQA、TriviaQA

(c) 在分布偏移下测试表现?（是/否）。对于无监督的方法，也可以是 N/A。
- 否

(d) 有监督、半监督或无监督的方法?
- 监督的

(e) 用于评估性能的指标。
- 预测拒绝比率（PRR，选择性预测指标）。

(f) 人类提升研究?
- 否

P24. Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback (Tian et al., 2023)

(a) 不确定性量化的类型（短/长形式生成、保形等）。
- 无监督提示方法，以及使用校准保留集来选择置信水平的有监督方法。

(b) 用于评估所提出的 LLM UQ 方法的数据集。
- TriviaQA、SciQA、TruthfulQA

(c) 在分布偏移下测试表现?（是/否）。对于无监督的方法，也可以是 N/A。
- 否

(d) 有监督、半监督或无监督的方法?
- 监督/无监督混合

(e) 用于评估性能的指标。
- ECE, ECE-t, BS-t, AUC

(f) 人类提升研究?
- 否

P25. Do LLMs Estimate Uncertainty Well in Instruction Following? (Heo et al., 2024)

(a) 不确定性量化的类型（短/长形式生成、保形等）。
- 指令跟随中大型语言模型不确定性的系统评估。

(b) 用于评估所提出的 LLM UQ 方法的数据集。
- IFEval

(c) 在分布偏移下测试表现？（是/否）。对于无监督的方法，也可以是 N/A。
- 不适用

(d) 有监督、半监督或无监督的方法？
- 无监督

(e) 用于评估性能的指标。
- 曲线下面积（AUROC）

(f) 人类提升研究？
- 否

P26. Improving Uncertainty Estimation through Semantically Diverse Language Generation (Aichberger et al., 2025)

(a) 不确定性量化的类型（短/长形式生成、保形等）。
- 通过引导多次生成来估计随机语义不确定性

(b) 用于评估所提出的 LLM UQ 方法的数据集。
- TruthfulQA, CoQA, TriviaQA

(c) 在分布偏移下测试表现？（是/否）。对于无监督的方法，也可以是 N/A。
- 不适用

(d) 有监督、半监督或无监督的方法？
- 无监督

(e) 用于评估性能的指标。
- Rouge-L、Rouge-1、BLEURT、AUROC

(f) 人类提升研究？
- 否

P27. Can Large Language Models Faithfully Express Their Intrinsic Uncertainty in Words? (Yona et al., 2024)

(a) 不确定性量化的类型（短/长形式生成、保形等）。
- 使用一种称为忠实度的新指标来评估大语言模型（LLM）的不确定性量化。论证大语言模型忽略了内在的不确定性。同时测试了一种基于少样本提示的大语言模型不确定性量化方法——Uncertainty+。

(b) 用于评估所提出的 LLM UQ 方法的数据集。
- 自然问题, PopQA

(c) 在分布偏移下测试表现？（是/否）。对于无监督的方法，也可以是 N/A。
- 不适用

(d) 有监督、半监督或无监督的方法？
- N/A

(e) 用于评估性能的指标。
- 忠实性

(f) 人类提升研究？
- 不适用

P28. Confidence Under the Hood: An Investigation into the Confidence-Probability Alignment in Large Language Models (Kumar et al., 2024)

(a) 不确定性量化的类型（短/长形式生成、保形等）。
- 将 LLM 表述的置信度（在模型响应中表达的）与通过标记概率量化的置信度联系起来。

(b) 用于评估所提出的 LLM UQ 方法的数据集。
- CommonsenseQA, QASC, RiddleQA, OpenBookQA, ARC

(c) 在分布偏移下测试表现？（是/否）。对于无监督的方法，也可以是 N/A。
- 不适用

(d) 有监督、半监督或无监督的方法？
- 无监督

(e) 用于评估性能的指标。
- 斯皮尔曼相关

(f) 人类提升研究?
- 否

P29. Language Model Uncertainty Quantification with Attention Chain (Li et al., 2025)

(a) 不确定性量化的类型（短/长形式生成、保形等）。
- 使用注意力权重在回溯过程中找到语义上关键的标记。允许估计答案标记的边际概率。

(b) 用于评估所提出的 LLM UQ 方法的数据集。
- GSM8K，MATH，BigBench

(c) 在分布偏移下测试表现?（是/否）。对于无监督的方法，也可以是 N/A。
- 不

(d) 有监督、半监督或无监督的方法?
- 无监督

(e) 用于评估性能的指标。
- AUROC, ECE

(f) 人类提升研究?
- 否

P30. Preserving Pre-trained Features Helps Calibrate Fine-tuned Language Models (He et al., 2023a)

(a) 不确定性量化的类型（短/长形式生成、保形等）。
- 微调后的语言模型在分布变化下的校准表现不佳。提出了一种方法来保存预训练特征，以改善校准效果。

(b) 用于评估所提出的 LLM UQ 方法的数据集。
- WikiText-103, SWAG, HellaSWAG, MNLI, SNLI

(c) 在分布偏移下测试表现?（是/否）。对于无监督的方法，也可以是 N/A。
- 是

(d) 有监督、半监督或无监督的方法?
- 监督

(e) 用于评估性能的指标。
- ECE + 可靠性图

(f) 人类提升研究?
- 否

P31. Uncertainty Estimation and Quantification for LLMs: A Simple Supervised Approach (Liu et al., 2024a)

(a) 不确定性量化的类型（短/长形式生成、保形等）。
- 以监督方式使用隐藏层特征进行不确定性量化。

(b) 用于评估所提出的 LLM UQ 方法的数据集。
- CoQA, TriviaQA, MMLU, WMT14 机器翻译

(c) 在分布偏移下测试表现?（是/否）。对于无监督的方法，也可以是 N/A。
- 是

(d) 有监督、半监督或无监督的方法?
- 监督的

(e) 用于评估性能的指标。
- 受试者工作特征曲线下面积

(f) 人类提升研究?
- 否

P32. Just rephrase it! Uncertainty estimation in closed-source language models via multiple rephrased queries (Yang et al., 2024a)

(a) 不确定性量化的类型（短/长形式生成、保形等）。
- 重新表述提示以采样多个输出，聚合以获得置信度。

(b) 用于评估所提出的 LLM UQ 方法的数据集。
- ARC，开放书问答，MMLU

(c) 在分布偏移下测试表现？（是/否）。对于无监督的方法，也可以是 N/A。
- 不适用

(d) 有监督、半监督或无监督的方法？
- 无监督

(e) 用于评估性能的指标。
- 准确性，ECE，TACE，布里尔分数，AUROC

(f) 人类提升研究？
- 否

P33. Uncertainty Quantification with Pre-trained Language Models: A Large-Scale Empirical Analysis (Xiao et al., 2022)

(a) 不确定性量化的类型（短/长形式生成、保形等）。
- 基准测试在 NLP 任务上的各种 LLM UQ 方法。

(b) 用于评估所提出的 LLM UQ 方法的数据集。
- 情感分析：IMDB，Yelp 评论。自然语言推理：MNLI，SNLI。常识推理：SWAG，HellaSWAG

(c) 在分布偏移下测试表现？（是/否）。对于无监督的方法，也可以是 N/A。
- 是

(d) 有监督、半监督或无监督的方法？
- 监督的

(e) 用于评估性能的指标。
- 电子与计算机工程

(f) 人类提升研究？
- 不

P34. Beyond Confidence: Reliable Models Should Also Consider Atypicality (Yuksekgonul et al., 2023)

(a) 不确定性量化的类型（短/长形式生成、保形等）。
- 估计输入的非典型性以改进大型语言模型不确定性量化。

(b) 用于评估所提出的 LLM UQ 方法的数据集。
- IMDB、TREC、AG 新闻

(c) 在分布偏移下测试表现？（是/否）。对于无监督的方法，也可以是 N/A。
- 是

(d) 有监督、半监督或无监督的方法？
- 监督

(e) 用于评估性能的指标。
- ECE

(f) 人类提升研究？
- 否

P35. Overconfidence is Key: Verbalized Uncertainty Evaluation in Large Language and Vision-Language Models (Groot and Valdenegro-Toro, 2024)

(a) 不确定性量化的类型（短/长形式生成、保形等）。
- 提示大型语言模型输出置信区间

(b) 用于评估所提出的 LLM UQ 方法的数据集。
- SST（情感分析）GSM8k（数学）CoNLL（命名实体识别）日语不确定图像（仅用于 VLMs）

(c) 在分布偏移下测试表现？（是/否）。对于无监督的方法，也可以是 N/A。
- 不适用

(d) 有监督、半监督或无监督的方法？
- 无监督

(e) 用于评估性能的指标。
- ECE, 最大 CE, 正规化 CE

(f) 人类提升研究?
- 否

P36. Multi-group Uncertainty Quantification for Long-form Text Generation (Liu and Wu, 2024)

(a) 不确定性量化的类型（短/长形式生成、保形等）。
- 长文本生成中事实正确性的不确定性量化

(b) 用于评估所提出的 LLM UQ 方法的数据集。
- BIO-NQ（与传记相关的自然问题子集）

(c) 在分布偏移下测试表现?（是/否）。对于无监督的方法，也可以是 N/A。
- 否

(d) 有监督、半监督或无监督的方法?
- 监督

(e) 用于评估性能的指标。
- ASCE（平均平方校准误差）、最大 gASCE、平均 gASCE、Brier 分数

(f) 人类提升研究?
- 否

P37. LACIE: Listener-Aware Finetuning for Confidence Calibration in Large Language Models (Stengel-Eskin et al., 2024)

(a) 不确定性量化的类型（短/长形式生成、保形等）。
- 将校准描述为具有说话者和听者的偏好优化问题。

(b) 用于评估所提出的 LLM UQ 方法的数据集。
- TriviaQA TruthfulQA

(c) 在分布偏移下测试表现?（是/否）。对于无监督的方法，也可以是 N/A。
- 是

(d) 有监督、半监督或无监督的方法?
- 监督

(e) 用于评估性能的指标。
- 用户的 ECE、AUROC、精准率、召回率

(f) 人类提升研究?
- 是的，人类评估是由标注者接受或拒绝 LLM 的回答。