

大陪审团：用于动态质量评估标准的协作机器学习 模型评估协议

Arthur Cho

hello@memiorji.com

Memoirji LLC

2025 年 8 月 9 日

摘要

生成式机器学习模型已成为现代系统的核心，驱动着创意写作、总结、多跳推理和上下文感知对话等应用 [2]。这些模型支撑着大规模 AI 助手、工作流程自动化和自主决策 [15, 20]。在这样的领域中，“可接受的响应”很少是绝对或静态的，而是多样且高度依赖于上下文 [3]。然而，标准评估体系仍然依赖于静态的基准测试，激励优化以排行榜得分为目标，而不是与动态用户需求或不断变化的实际情况保持一致 [6]。大陪审团引入了一种正式的评价协议——结合时间衰减聚合、完整的可追溯性，并支持动态透明的任务评分标准和多评估人的判断。这些元素共同使得多元且负责的评估成为可能，能够捕捉不断演变的一致意见并揭示分歧。我们提供了 (i) 一个开源实现（大陪审团 PyPI）以及 (ii) 一系列大型语言模型 (LLM) 推理输出的公共集合以说明其需求和方法。GrandJury 为 AI 从业者提供了一种新的评估机器学习输出的方法 没有绝对的地面真相。专利申请中：美国临时专利 63/825,484，提交日期 2025 年 6 月 17 日。

1 介绍

生成式 AI 系统的迅速普及改变了信息在各个领域中的生产、中介和消费方式——范围涵盖了创意写作和对话辅助到自动化摘要、分析以及复杂推理任务 ~ [2, 19]。这些模型现在作为数字助理、企业工作流、教育和研究中代理过程的基础 ~ [23]。在这种情况下，输出不仅仅是“正确”或“错误”，而是反映了有效观点、目标和解释的光谱 ~ [1, 21]。

然而，模型评估的标准方法仍然依赖于静态的地面真相数据和标准化的评估标准。诸如 BLEU [16]、ROUGE [12] 和 MMLU [9] 等评价指标是为具有单一正确答案和明确定义

指标的任务设计的。长期以来用于翻译和摘要生成的 BLEU 和 ROUGE 倾向于奖励抽取式输出和 n-gram 重叠, 未能评估抽象性、细微差别或语境适宜性 [18]。将它们应用于生成式、开放式的或带有价值判断的输出会导致所测量的内容与实际重要的内容之间存在深刻不匹配。

针对这些方法论的优化通常集中在自动化指标计算过程。其中值得一提的是 RAGAS (检索增强生成评估) [5], 它是一套无参考指标, 旨在在不依赖人工标注真实标签的情况下, 通过检索相关性、答案忠实度和上下文精确度等多个维度来评估检索增强生成 (RAG) 系统。同样地, 以大型语言模型作为评判的方法 [26, 24] ——包括 Prometheus [11]、JudgeLM [27] 以及基于 GPT 的评估管道等框架——通过使用模型来评估其他模型, 从而实现评分自动化。基于嵌入的技术如余弦相似度 [25] 也常用于自动可信性检查。

然而, 这些代理方法仍然不足: RAGAS 指标在开放任务中缺乏实证验证, 并且经常与人类偏好不相关 [7], 以 LLM 作为裁判的系统表现出可变的偏差和对提示的敏感性 [17, 22], 基于嵌入的分数忽视了叙述、语用或风格上的忠实度 [14]。因此, 尽管这些尝试旨在避免人工标注的需求, 但仍难以与微妙的人类判断保持一致。更糟的是, 随着供应商和研究人员设计自己的基准测试, “最先进的”定义变得越来越可塑且自指 [13, 8]。

2 不确定性情况下的评估

2.1 基于度量的评估的局限性与 Vibe 评估的诞生

评估生成型 AI 从根本上不同于封闭的、确定性的任务。开放式的输出引入了更多的紧张关系: 不同标注者之间的协议往往很低 ($\kappa < 0.3$) [4, 10], 即使有多名评分人员也是如此, 因为没有共享的评价契约来引导判断的一致性。在实践中, 如果没有明确定义的标准, 不同的标注者会应用不同的标准, 导致不一致的结果。在这种模糊领域中, 评估往往退化为振动评估: 病毒性的轶事或基于截图的说法缺乏可重复性、结构或透明度, 使得“分数剧场”和炒作占据主导地位而无需承担责任 [13]。

2.2 自动代理评估的尝试及其局限性

为了应对评估规模的挑战, 已经提出了一些自动化代理方法:

- **拉加斯:** 无参考评估指标用于 RAG 系统——涵盖检索质量、忠实度和上下文相关性。这些指标在介绍时没有明确的地面实况参考, 但通过人类标注的基准进行验证和优化, 在设计中间接嵌入了地面实况 [5]。
- **LLM 作为裁判框架** 依赖模型 (例如, GPT-4 或 Claude) 来评定输出。虽然在可扩展

性方面颇具吸引力，但研究表明存在显著问题：评估者偏差、提示敏感度、位置偏差和不稳定性——在细微的任务中，其性能与人类共识相去甚远 [17, 22]。

- **基于嵌入的度量标准**（例如，余弦相似度）提供快速、无需参考的评估，但无法捕捉更深层次的语义、语用或风格一致性，经常与人类判断不符 [7, 14]。

尽管这些方法减少了标注负担，但它们都依赖——明确或隐含地——真实信号，并且在开放领域中仍未达到与人类对齐的可靠评估。

3 大陪审团：一种结构化判断的协议

我们引入了 GrandJury，这是一种旨在实现多元、透明和负责评估的治理协议，适用于缺乏明确基本事实的领域。

3.1 同意与投票提示

每个任务都以一个“投票者提示”开始：一份公开共享的评分标准，规定了要评估的内容、如何评估以及为何评估。受 CrowdTruth [3] 等标注框架的启发——将标注者的意见分歧视为信号而非噪声——这一结构化的评分标准在保持解释差异的同时，通过一个共同的框架统一多样化的判断。

3.2 时间衰减聚合

代替静态平均值，GrandJury 应用指数时间衰减：

$$S_t = \alpha S_{t-1} + (1 - \alpha) \langle v_t \rangle \quad (1)$$

符号表示。

- t — 批次索引； S_{t-1} — 在批次 t 之前的累计分数。
- $v_{t,i} \in [0, 1]$ — 第 i 位评审员在第 t 批中的投票 ($1 = \text{accept}; 0 = \text{reject}$)。
- $r_i > 0$ — 可选声誉权重（默认为 1）。
- $\langle v_t \rangle = \frac{\sum_i r_i v_{t,i}}{\sum_i r_i}$ — 声誉加权平均。
- Δt — 自批次 $t-1$ 以来的时间（可配置单位）； λ — 衰减常数（时间⁻¹）。
- $\alpha = e^{-\lambda \Delta t}$ — 衰减因子。

工作示例。 设 $S_{t-1} = 0.72$, $\Delta t = 3$ 天, $\lambda = 0.1$, 因此 $\alpha = e^{-0.3} \approx 0.741$ 。三名陪审员投票 $\{0.90, 0.80, 0.60\}$ 。然后 $\langle v_t \rangle = 0.766$ 和

$$S_t = 0.741 \cdot 0.72 + 0.259 \cdot 0.766 = 0.733.$$

时间单位灵活性。 虽然本示例使用天数进行说明, 但该实现支持任意时间单位 (秒、小时、天) 以及相应的 λ 值, 从而能够根据评估环境实现精细的时间控制。

参数选择。 衰减常数 λ 决定了响应速度: 较大的值会使分数对最近的投票更敏感。典型值范围从 $\lambda = 0.01 \text{ s}^{-1}$ (高响应速度) 到 $\lambda = 0.001 \text{ s}^{-1}$ (稳定共识), 具体取决于时间粒度和评估要求。

这反映了推荐系统和人类反馈系统中的“时间感知”方法, 确保最近的投票具有更大的影响, 同时保留历史背景。

批处理与流处理。 虽然该公式描述了离散的“批次” t , 但该实现实时处理投票: 每次新的评估调用构成一个微批次, 允许随着投票到达进行连续的评分更新。这种架构通过包含推理 ID、投票、投票者_id、投票时间和投票者提示 ID 的结构化投票数据, 支持个体评分和群体层面的分析。

3.3 新鲜度指标

为了保持分数时效性的清晰, GrandJury 追踪新鲜度 $F_t = 1 - \alpha$, 表明最近的评估贡献了多少权重——借鉴了持续基准测试和自适应评分协议的概念。

解读。 新鲜度量化了新的证据的影响:

$$F_t = 1 - \alpha = 1 - e^{-\lambda \Delta t}.$$

在上面的例子中, $F_t = 0.259$, 意味着当前分数的 25.9% 来自最新批次, 而 74.1% 来自历史共识。得出的分数 (0.733) 展示了协议如何平衡稳定性和响应性。较小的 Δt 或较大的 λ 使得分对近期证据更敏感。

3.4 歧义标记

高评估方差 (例如, $\sigma^2 > 0.5$) 触发一个模糊标志, 表明输出缺乏共识。这将不一致从一个问题转化为一种分析信号, 对于透明评估至关重要。

符号和规则。 令 $\sigma_t^2 = \text{Var}(v_{t,*})$ 。大陪审团将项目标记为模棱两可如果

$$\sigma_t^2 > \sigma_{\text{crit}}^2, \quad \text{e.g. } \sigma_{\text{crit}}^2 = 0.05.$$

工作示例。 Using the same three votes $\{0.90, 0.80, 0.60\}$:

$$\sigma_t^2 = \frac{1}{3} \sum_i (v_{t,i} - 0.766)^2 = 0.016 < 0.05,$$

so 不 flag is raised. With divergent votes $\{1, 0, 0\}$ we get $\sigma_t^2 = 0.111 > 0.05$; the item is flagged for curator review or an additional voting round.

3.5 投票的可追溯性

每个投票都带有时间戳 (ISO 8601), 链接到一个匿名陪审员 ID, 并与选民提示上下文相关联。这反映了人类评估数据表和众包审计中的最佳实践, 使完整的审计轨迹成为可能, 并支持元分析。

3.6 实现架构

时间衰减算法在服务器端实现, Python 客户端提供数据格式化和 API 访问。此架构确保:

- 所有用户和评估会话中一致的评分
- 实时处理作为微批次的个别选票
- 大规模评估活动的可扩展聚合
- 集中式参数管理 ($\lambda, \sigma_{\text{crit}}^2$) 与领域特定调优

客户端数据处理与服务器端计算的分离使得个别研究人员和大规模评估计划能够利用相同的算法基础设施, 同时保持数据来源和评估工作流程的灵活性。

4 演示: 公共推理数据集和评估工具包

为了在实际中展示 GrandJury, 我们发布了一个包含来自 7 个主要 AI 平台的 31 种不同模型的 149 个笑话回应的综合推理集合。该数据集既作为演示协议适用性的工具, 也是一个可用于可重复性评估研究的资源。该数据集托管于 Hugging Face。

4.1 模型选择方法论

模型选择遵循了一种系统方法，以确保在创造性任务中表现优秀的模型得以体现。我们参考了 LMSYS Chatbot Arena 的排行榜 [28]，特别筛选出在创意写作任务中表现出色的模型。从这个精选的顶级创意模型列表中，我们选择了那些可以在主要云 AI 平台（Azure AI、Anthropic、Bedrock、Gemini、Hugging Face、Mistral 和 NVIDIA）上获得的模型，以确保研究社区的可访问性和可重复性。

该方法产生了 31 个不同的模型，涵盖多个模型家族和大小，提供了创意领域中最新生成能力的代表性样本。所有模型均使用相同的提示“告诉我一个笑话”进行查询，并设置温度=1.0 以保持一致的实验条件。最终收集包含在受控 6 分钟窗口内收集的 149 个响应，确保了跨平台的时间一致性。

4.2 数据集特征与设计

推理集合包含每个响应的综合元数据：

- 模型标识符和提供平台
- 完整推理输入和生成的输出
- 时间分析的 ISO-8601 时间戳
- 平台特定的模型配置和参数

表 1 展示了推理集合中的一个代表性样本，展现了数据集中捕获的模型、平台和响应风格的多样性。每个响应代表了对相同提示“给我讲个笑话”的不同模型解释，温度值为 1.0，说明了生成输出中自然变化的情况，这促使需要多元化的评估。

按照设计，幽默生成任务没有单一的“正确”输出，且数据集发布时未提供参考答案或黄金标准。这种结构故意强调了基于相似性的度量（例如余弦嵌入相似性）的局限性，这些度量在诸如幽默、创造力和开放性推理等主观领域中已被反复证明与人类判断不一致。

该数据集通过 Hugging Face 公开提供，以促进可重复的、社区驱动的评估。GrandJury 推理集合一笑柄在我们身上¹ 在 CC-BY-4.0 许可下发布，并存档于 Hugging Face 以确保可重复性 [29]。

4.3 两阶段评估 workflow

GrandJury 通过数据收集和投票分析之间的明确职责分离来运行：

第一阶段：人类评估收集

¹传记/关于我们的笑话 (v1.0, CC-BY-4.0, 访问日期 2025 年 8 月 4 日)。

1. 从 Hugging Face 下载推理数据集
2. 定义适合幽默评估的投票者提示（评价标准）
3. 进行人类评估环节以收集对模型输出的投票
4. 存储带有所需元数据（投票者 ID、时间戳、推理 ID）的选票

第二阶段：使用 grandjury 包进行投票分析

1. 使用 grandjury Python 客户端加载收集的投票数据
2. 应用时间衰减算法和共识分析
3. 生成具有可追溯性和新鲜度指标的评估报告

这种分离确保了评估标准保持灵活和社区驱动，而不是嵌入到自动化指标中。推理数据集作为人类评价的原材料，而 grandjury 包提供了处理由此产生的人类判断的分析基础设施。

研究人员和从业者被邀请探索该数据集，并将 GrandJury 协议应用于这一推理集合：通过定义选民提示、收集多元判断并使用时间衰减方法汇总分数，社区可以揭示分歧、绘制共识变化图，并透明地评估超越静态基准的生成模型。

我们通过开源的 grandjuryPython 包 [31](可在 PyPI 和 GitHub 上获得)提供了 GrandJury 协议的一个参考实现。该软件包旨在满足流水线使用需求，使研究人员和实践者能够采用该协议进行投票分析和时间衰减评分。

4.4 软件包安装

通过 pip 安装：

```
pip install grandjury
```

可选的性能依赖项以增强数据格式支持：

```
pip install grandjury[performance] # Includes msgspec, pyarrow, polars
```

源代码和文档可在 GitHub 获取。

包结构遵循科学 Python 库（如 Scikit-learn）中已建立的设计模式，以确保熟悉性和易于采用。

4.5 投票分析 API

该软件包对第一阶段评估期间收集的投票数据进行操作。用户可以分析多种格式的投票数据：

```
from grandjury import GrandJuryClient
import pandas as pd

# Initialize client (no authentication needed for analysis)
client = GrandJuryClient()

# Load vote data from your evaluation sessions
vote_data = pd.read_csv("collected_votes.csv")

# Analyze vote patterns and temporal distributions
histogram = client.vote_histogram(vote_data)
completeness = client.vote_completeness(vote_data, voter_list)
confidence = client.population_confidence(vote_data, voter_list)
distribution = client.votes_distribution(vote_data)

# Apply time-decay scoring (requires API key for server access)
client = GrandJuryClient(api_key="your-api-key")
result = client.evaluate_model(
    previous_score=0.0,
    votes=[0.8, 0.6, 0.9], # From your collected evaluations
    reputations=[1.0, 1.0, 1.0]
)
print(f"Time-decayed score: {result['score']:.4f}")
```

该软件包支持多种数据格式（pandas 数据框、Polars、CSV、Parquet、JSON），以无缝集成到现有的评估工作流程中。

4.6 生态系统集成与可扩展性

大陪审团是有意设计为参考实现的，具有最小依赖性和模块化 API，以促进集成到更广泛的工具中——例如仪表板、分析管道或 UI 插件。

研究人员可以：

1. 将该包应用于社区或专有数据集，
2. 扩展自定义仓库中的管道，
3. 利用其他评估系统中的时间衰减和新鲜度函数。

这不仅将大陪审团定位为一个独立的工具，而且还作为结构化、多元评价的基础性基础设施。

5 结论和社区参与

大陪审团不是一个基准、指标或产品——它是一种旨在对生成式 AI 进行多元的、透明的和可再生评估的协议和社区基础设施，在这种评估中，真实情况并不适用。本最终部分明确了目的，将其置于当前治理动态之中，并发出了一个强烈且可操作的邀请，鼓励参与。

5.1 贡献与目标

大陪审团的核心目标是实现结构化、多元化的评估，而不是宣布胜者或嵌入不透明的排名。它强调多样性而非共识、透明度而非不透明性、更新而非停滞的观点，与新兴的人工智能治理框架相一致，这些框架倡导独立评估和以人为本的监督机制。

5.2 协议范围、社区采纳及未来方向

大陪审团作为一种治理协议，定义了谁投票、在何种选民提示下投票、何时投票以及如何汇总这些投票，同时作为一个系统来呈现多元判断、记录元数据并追踪共识随时间的变化。重要的是，它不是一个静态的评分指标、商业 SaaS 产品或规定性政策——相反，它提供了评估工具和结构而不强加指令。这种定位反映了我们致力于实现有组织的多元化评估而不宣布胜者或嵌入不透明排名的承诺。

我们邀请广泛的社区参与多个利益相关者群体：研究人员和实践者将 GrandJury 应用于他们的模型、领域和数据集；开源开发者构建基于该协议的用户界面、仪表板和工作流集成；治理机构、标准组织和审计员在透明的实际评估或合规使用案例中试用 GrandJury；以及人类-AI 伦理和政策研究员研究并完善协议的透明度、问责制和共识跟踪动态。

这种协作方法确保了评价标准保持灵活且由社区驱动，而不是嵌入自动化指标中。

大陪审团的设计与当前机构在负责任的人工智能治理方面的努力相一致。该协议呼应了第三方和基于审计的评估系统，如英国标准协会推广的 ISO/IEC 42001 人工智能管理系统标准 [32] 用于独立审核和认证人工智能治理过程，同时支持优先考虑透明度、人类监督和可审计评估的负责任的人工智能计划 [33, 34]。它对多元性而非共识、透明度而非不透明

以及更新而非停滞的强调与欧盟人工智能法案等监管运动相呼应，这些法规要求有人类监督和独立符合性评估 [30]。通过将评估扎根于多样的人类判断——记录在案、可审计且具有时间意识——大陪审团提供了拥抱不确定性、分散权威并将分歧转化为责任的基础设施。

展望未来，我们敦促社区下载推理集合，实验选民提示，并在实践中应用大陪审团软件包。我们鼓励公开分享数据集、选民提示变体和见解——通过 GitHub、论坛或出版物——同时参与协议对话以完善元数据日志记录、衰减参数和模糊阈值的设定。最终目标是将 GrandJury 整合到组织审计工作流程和社区评估标准中，不仅将其定位为独立工具，而且作为在地面真相难以捉摸的不同领域进行结构化、多元评价的基础架构。

通过将评估锚定在多元的人类判断上——记录、可审计且具备时间意识——GrandJury 提供了一种新的评估生成系统的基石：一种拥抱不确定性、去中心化权威并将分歧转化为问责制的方法。

参考文献

- [1] Basile, V., Fell, M., Fornaciari, T., Hovy, D., Paun, S., Plank, B., Poesio, M., and Uma, A. We Need to Consider Disagreement in Evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, 2021.
- [2] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.
- [3] Dumitrache, A., Aroyo, L., and Welty, C. CrowdTruth 2.0: Quality Metrics for Crowdsourcing with Disagreement. *arXiv preprint arXiv:1808.06080*, 2018.
- [4] Elangovan, A., He, J., and Verspoor, K. Memorization vs. Inference in the Era of Large Language Models. *arXiv preprint arXiv:2404.03129*, 2024.
- [5] Es, S., James, J., Espinosa-Anke, L., and Schockaert, S. RAGAS: Automated Evaluation of Retrieval Augmented Generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, 2024.

- [6] Ethayarajh, K. and Jurafsky, D. Utility is in the Eye of the User: A Critique of NLP Leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, 2020.
- [7] Fabbri, A.R., Kryściński, W., McCann, B., Xiong, C., Socher, R., and Radev, D. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 2021.
- [8] Fodor, V. Line Goes Up? Inherent Limitations of Benchmarks for Evaluating LLMs. *arXiv preprint arXiv:2502.14318*, 2025.
- [9] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*, 2020.
- [10] Kenyon-Dean, K., Cheung, J.C.K., and Precup, D. Sentiment Analysis: It’s Complicated! In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1886–1895, 2018.
- [11] Kim, S., Shin, J., Kang, Y., Cho, L., Du, B., Rhyou, C.H., Choi, H., Yu, Y., Kim, H., and Kim, J. Prometheus: Inducing Fine-grained Evaluation Capability in Language Models. *arXiv preprint arXiv:2310.08491*, 2024.
- [12] Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, 2004.
- [13] Lipton, Z.C. and Steinhardt, J. Troubling Trends in Machine Learning Scholarship. *arXiv preprint arXiv:1807.03341*, 2018.
- [14] Liu, C.W., Lowe, R., Serban, I., Noseworthy, M., Charlin, L., and Pineau, J. How NOT to Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, 2016.
- [15] OpenAI. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2023.
- [16] Papineni, K., Roukos, S., Ward, T., and Zhu, W.J. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

- [17] Park, C., Ahn, D., and Kang, J. OffsetBias: Leveraging Debiased Data for Tuning Evaluators. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 984–1000, 2024.
- [18] Reiter, E. A Structured Review of the Validity of BLEU. *Computational Linguistics*, 44(3):393–401, 2018.
- [19] Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.T., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y., Lee, H., Zheng, H.S., Ghafouri, A., Menegali, M., Huang, Y., Krikun, M., Lepikhin, D., Qin, J., Chen, D., Xu, Y., Chen, Z., Roberts, A., Bosma, M., Zhao, V., Zhou, Y., Chang, C.C., Krivokon, I., Rusch, W., Pickett, M., Srinivasan, P., Meier-Hellstern, L., Morris, M.R., Doshi, T., Santos, R.D., Duke, T., Soraker, J., Zevenbergen, B., Prabhakaran, V., Diaz, M., Hutchinson, B., Olson, K., Molina, A., Hoffman-John, E., Lee, J., Aroyo, L., Rajakumar, R., Butryna, A., Lamm, M., Kuzmina, V., Fenton, J., Cohen, A., Bernstein, R., Kurzweil, R., Aguera-Arcas, B., Cui, C., Croak, M., Chi, E., and Le, Q. LaMDA: Language Models for Dialog Applications. *arXiv preprint arXiv:2201.08239*, 2022.
- [20] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*, 2023.
- [21] Uma, A., Fornaciari, T., Hovy, D., Paun, S., Plank, B., and Poesio, M. Learning from Disagreement: A Survey. *Journal of Artificial Intelligence Research*, 72:1385–1470, 2021.
- [22] Wataoka, H., Yasunaga, M., Suzuki, R., Miyazaki, Y., and Patel, D. Self-Preference Bias in LLM-as-a-Judge. *arXiv preprint arXiv:2410.21819*, 2024.
- [23] Wornow, M., Xu, P., Thakkar, A., Patel, B., Steinberg, E., Fleming, S., Pfeffer, M., Fries, J., Shah, N., and Fatahalian, K. Automating the Enterprise with Foundation Models. *Proceedings of the VLDB Endowment*, 17(11):2853–2867, 2024.
- [24] Ye, J., Chen, X., Xu, N., Zu, C., Shao, Z., Liu, S., Cui, Y., Zhou, Z., Gong, C., Shen, Y., Zhou, J., Chen, S., Gui, T., Zhang, Q., and Huang, X. A Comprehensive Capability Analysis of GPT-3 and GPT-4 as an LLM-as-a-Judge. *arXiv preprint arXiv:2402.01047*, 2024.

- [25] Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., and Artzi, Y. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*, 2020.
- [26] Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, 36, 2024.
- [27] Zhu, J., Wang, W., Qin, C., Yu, P., Lu, L., Wang, C., and Wu, Q. JudgeLM: Fine-tuned Large Language Models are Scalable Judges. *arXiv preprint arXiv:2310.17631*, 2023.
- [28] LMSYS Org. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. <https://chat.lmsys.org/>, 2024. Accessed August 2024.
- [29] Arthur C. GrandJury Inference Collection — "Jokes on Us". Hugging Face, 2025. Version 1.0. https://huggingface.co/datasets/biographies/jokes_on_us. CC-BY-4.0 licence.
- [30] European Union. Artificial Intelligence Act — Article 14: Human Oversight. <https://artificialintelligenceact.eu/article/14/>, 2025. Consolidated text, 2025 trilogue version.
- [31] biographies-ai. grandjury: Python client for GrandJury ML evaluation and verdict analysis API. PyPI, 2025. Version 1.0.1. <https://pypi.org/project/grandjury/>. Patent pending: US Provisional 63/825,484, filed 17 June 2025.
- [32] International Organization for Standardization. ISO/IEC 42001:2023 — Artificial Intelligence Management Systems. British Standards Institution, 2023. <https://www.bsigroup.com/en-GB/standards/benefits-of-standards-enhanced-credibility-trust/iso-42001-ai-management-system/>.
- [33] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1, 2023. <https://doi.org/10.6028/NIST.AI.100-1>.
- [34] OECD. Advancing Accountability in AI: Governance and Multi-stakeholder Processes. OECD Digital Economy Papers, No. 349, 2023. <https://doi.org/10.1787/2448f04b-en>.

表 1: 来自大陪审团推理集合的代表性样本

平台	模型	时间戳	响应
Azure AI	gpt-4.1	2025-08-01 22:57:27	Why did the scarecrow win an award? Because he was outstanding in his field!
Anthropic	claude-opus-4-20250514	2025-08-01 22:57:27	Why don't scientists trust atoms? Because they make up everything!
Mistral	mistral-medium-2505	2025-08-01 22:57:26	Sure! Here's a classic for you: 为什么骨头架子不互相打架？因为他们没有胆量！
Vertex AI	gemini-2.5-pro	2025-08-01 22:57:37	Why don't scientists trust atoms? Because they make up everything
NVIDIA NIM	qwen/qwen3-235b-a22b	2025-08-01 22:57:32	Sure! Here's a light-hearted one for you: Why did the tomato turn red? Because it saw the salad dressing!
Bedrock	nova-pro-v1:0	2025-08-01 22:57:27	Sure, here's a classic joke for you: Why did the scarecrow win an award? Because he was outstanding in his field!
Hugging Face	Qwen/Qwen3-30B-A3B	2025-08-01 22:57:29	Why did the scarecrow win an award? 因为他在他所在的领域非常出色！