设计大语言模型提示词，从杂乱文本中提取分数：共享数据集与挑战[1]

Mike Thelwall
英国谢菲尔德大学信息、新闻与传播学院。https://orcid.org/0000-0001-6065-205X
m.a.thelwall@sheffield.ac.uk

在计算机、自然语言处理和信息科学的某些领域，人们通过共享数据集并挑战相关社群设计出适用于特定任务的最佳算法来推动进展。本文介绍了一个由 1446 段短文本组成的共享数据集，每段文本都描述了一个按照英国标准、以 1* 至 4* 表示的研究质量评分。这是一个杂乱的数据集，其中一些文本不含分数，另一些则包含无效分数或奇怪的格式。该数据集还附有对有效分数的定义，以及这些文本中正确分数的“金标准”（包括缺失值）。这项挑战是设计一个大语言模型提示词，让大语言模型尽可能准确地从这些文本中提取分数。回复格式应为数字，不能包含任何其他文本，因此这项挑战包含两个方面：确保大语言模型只返回数字，并指导其推断文本对应的正确数字。为此，大语言模型提示词需要说明，当文本中未明确包含数字时，应返回缺失值代码 -1，而不是数字。本文还提供了一个简单提示词的示例。这项挑战有两个目的：找到解决这一问题的有效方案，并加深对复杂数值任务中提示设计和大语言模型能力的理解。提出的初步方案准确率为 72.6%，因此这项挑战就是要超越这一成绩。

关键词：大语言模型；研究质量评分；提示设计

引言

计算机领域有长期发布共享数据集及相关挑战的传统，研究人员尝试设计算法，以便在共享数据集上取得最佳结果。这些挑战有时与特定会议或研讨会相关，组织者会创建数据集。最著名的例子或许是始于1992年的文本检索会议（TREC）（Voorhees & Harman, 1999）。当存在一项重要的共同任务，而创建合适的数据集又耗时较长时，这是一种有效的策略。共享任务能够让多个团队各自生成解决方案，而不必每个团队都自行创建数据集，从而提高效率，同时也便于直接比较性能。如果不同研究人员各自创建数据集，那么针对这些数据集的性能指标就不太可能直接比较，因为数据集特征（包括任务难度）可能有所不同。

大语言模型的能力也常通过共享任务和数据集进行评估，其中许多任务和数据集在这些模型诞生之前就已出现。例如，语言理解（Hendrycks et al., 2020; Wang et al., 2019）、问答（Kwiatkowski et al., 2019）、临时设计的高难度任务（Srivastava et al., 2023）和数学（Cobbe et al., 2021）。还有一个

[1] Thelwall, M.（2025）。设计用于从杂乱文本中提取分数的大语言模型提示词：共享数据集与挑战。《Trends in Information Management》，13(2)，论文1。
https://lis.uok.edu.in/Files/9ebfb2f2-5003-47a4-9dfe-d3cdcc6a2020/Journal/1435d56f-dbb2-4364-b3ef-2fe4237f73d7.pdf

备受瞩目的大语言模型任务表现排行榜（例如：https://huggingface.co/spaces/open-llm-leaderboard/open_大语言模型_leaderboard#/）。

本文聚焦于为大语言模型（LLM）设计提示词，使其能够从对文本的描述中提取分数。每段文本描述本身也由大语言模型生成。本文关注的是提示设计，而非比较不同大语言模型。大语言模型常用于对文本进行评分，例如评阅作文，或评估连贯性、翻译质量、可读性和信息含量等属性（Hashemi et al., 2024; Parulekar & Jyothi, 2025; Stureborg et al., 2024; Tang et al., 2024; Yavuz et al., 2025），但它们并不总能稳定地只输出一个分数。相反，它们可能会把分数夹在冗长的解释中，也可能完全省略分数，或者给出无效分数（例如 110%）。这种情况似乎更容易出现在较小的大语言模型中。此外，在某些情境下，如果允许大语言模型先解释评分理由，而不是只要求它返回分数，评分会更加可靠（Thelwall & Yang, 2025）。因此，有必要开展后续任务，从大语言模型的输出中提取分数。虽然可以采用模式匹配算法等非大语言模型方法来完成这一任务，但大语言模型可能比模式匹配更灵活、适用范围更广，因为理论上它们能够应对新的（未见过的）模式。因此，开发提示词以便从短文本测试中准确提取分数至关重要。

本文介绍了一个共享数据集（Thelwall，2025）以及一项针对短文本分数提取任务的挑战，旨在鼓励研究界开发并分享有效策略。除了上述开展这项任务的实际理由外，这项任务似乎也颇有价值，因为它是一项相对透明、但暗藏复杂性的文本处理任务，因此研究这项任务或许有助于增进我们对大语言模型能力与局限性的了解。

研究质量评分的文本描述

尽管文本评分任务种类繁多，论文评分或许是最常见的一种，但该数据集聚焦于对期刊论文的研究质量进行评分的文本。这类文本是在运用人工智能评估学术研究的倡议中生成的（Thelwall & Yaghi，2025；Wu 等，2025）。问题在于，较小的大语言模型往往遵循一致的报告格式（Thelwall & Mohammadi，2025），而较大的大语言模型则会采用各种结构和术语，因此通过模式匹配提取分数可能颇为困难。以下是中等规模大语言模型 Llama4 Scout 生成的截断报告示例，展示了相似的结构。两种情况下，分数都是第一个数字，报告则是以相同文本开头的一段短文。此处使用文本删节标记 [...] 以简化内容。

- 4* 本文通过探讨森林生态系统中多样性与稳定性之间的关系，展现了原创性。该主题虽已在草本植物群落中得到研究，但在森林中尚未涉及。本研究为促进森林生态系统稳定性的过程提供了新的见解，包括超产效应、物种异步性和有利的物种相互作用。本文具有较高的重要性，因为它有助于我们理解多样性在维持生态系统生产力与稳定性方面的重要作用，并对森林管理和保护具有启示意义 [...]

- 4* 本文通过研究林莺种间领域性的相关因素，展现了原创性，并为这一普遍现象的生态学与进化稳定性提供了新见解。本文的重要性在于，它有助于我们理解群落结构和干扰

竞争，以及这对生态学和进化生物学领域的意义。本文的严谨性体现在采用了系统发育校正，并考量了表型相似性、杂交和同域共存等多个因素 [.....]

相比之下，以下示例展示了 ChatGPT 4o-mini 生成的更多样化的报告。

- 在审阅所提供的题为“[..],”的文章后，给出以下评估：**原创性：2*** 文章确实对 [...]的资助格局提出了一些独到见解。**重要性：2*** 研究结果对[...]具有显著的重要性。**严谨性：3*** 文章在定性和定量研究方法方面都展现了全面周详的研究思路[...]**总体得分：2*** [...]
- **分数：2*** **原创性：2*** 这项研究对[...]之间的关系进行了有趣的考察。**重要性：2*** 这项研究有望帮助人们加深对 [...]的理解。**严谨性：2*** 这项研究采用了恰当的数据收集和分析方法[...] 研究结果或可为从业者提供参考，但这项研究缺乏高评分研究所具有的变革性影响。

较小的模型有时也会忽略指令并生成意料之外的输出，如下例所示。这些都是对按四分制评分这一请求的错误回应。最后一个例子给出了有效的分数（3*），但后续的理由存在问题，因为其中提到了五分制。

- 4* 3*
- 这被评为世界领先水平。
- 这篇文章的分数是95。
- 我给这篇文章打4/5分。
- 原创性：2* 重要性：2* 严谨性：2*
- 1* 1.5* 这篇文章对 [...]进行了详细研究。研究揭示了环境条件、疾病患病率与种群持续存在之间的紧密关系。尽管这项研究在其特定背景下具有创新性和重要性，但由于研究重点狭窄，且在特定生态背景之外的适用性有限，可能达不到最高卓越标准或获得国际认可。该研究的方法科学可靠且严谨，为破碎化景观中的宿主—病原体相互作用提供了宝贵见解。这个分数体现了文章的原创性和重要性，但文章缺乏更广泛的理论进展或政策意义，因而不足以获得更高评价。

- **评分：** 3* **理由：** * **原创性（4/5）：** 该文章发现了一种新颖的[...]。这一发现为理解 EVI1 在生理和病理情况下的调控提供了新的见解。 * **重要性（3/5）：** 尽管这一磷酸化事件的发现具有重要意义，但其对英国白血病和实体瘤生物学的更广泛影响，以及潜在的治疗策略，仍需进一步探索。该文章也具有潜在的转化医学意义，但目前尚未充分阐述。 * **严谨性（4/5）：** [...]

数据集构建

共享数据集（Thelwall，2025）由 1446 份不同的报告组成，这些报告由大语言模型生成，这些模型根据期刊论文的标题和摘要对其研究质量进行评分。这些报告由以下三种中型大语言模型之一生成：Magistral、Llama 4 Scout 和 Gemma3 27b。在某些情况下，模型可以生成任意长度的报告，但在另一些情况下

报告被截断了（以加快处理速度）。该数据集最初由这些大语言模型生成的数十万条回答构成，随后通过自动筛选去除了重复回答，最终得到以下样本量：

- 4931 Gemma 独有报告
- 46263 Llama 独有报告
- 10054 Magi 独有报告

随后，我们将这些独有报告全部合并，并再次去除重复项（即不同模型之间的重复项）。这个数据集大得难以实际使用，其中包含许多非常相似的文本，例如以下三份来自 Gemma3 的截断报告：

- 3* **原创性：**该研究应用
- 3* **原创性：**该研究探索
- 3* **原创性：**该研究代表

因此，最后一步是人工筛选。我们将报告载入 Excel，按字母顺序排序，然后人工筛除内容相似的报告。最终留下 1446 份报告，涵盖了各种各样的报告风格，其中许多报告没有包含分数，或分数不明确。目标是尽可能提高数据集的难度和多样性。下一节将简要介绍这项挑战任务。

任务描述

共享数据集（Thelwall，2025）包含 Magistral、Llama 4 Scout 和 Gemma3 27b 的输出；这些模型受要求根据 REF 指南，为期刊论文评定卓越研究框架（REF）的研究质量评分（REF2021，2019）。部分输出被截断至 100 个词元，或因其他原因被截断。有些输出包含分数，另一些则没有。

任务是使用大语言模型提取各份报告所述的 REF 分数；如果报告未给出分数，则返回 -1。评分等级为 1* 至 4*；如果无法有把握地确定分数，则应返回 -1。

背景信息：以下是这些分数的含义（来源：https://2021.ref.ac.uk/guidance-on-results/guidance-on-ref-2021-results/index.html）：

- 4*：在原创性、重要性和严谨性方面达到世界领先水平的质量。
- 3*：在原创性、重要性和严谨性方面达到国际卓越水平，但尚未达到最高卓越标准的质量。
- 2*：在原创性、重要性和严谨性方面获得国际认可的质量
- 1*：在原创性、重要性和严谨性方面获得国家认可的质量。

大语言模型应报告总体得分；如果未报告总体得分，则应报告重要性、原创性和严谨性分数的平均值（前提是这三项分数均已给出）。如果缺少其中一项或两项，则应忽略这些分数。

要算作正确答案，大语言模型分数只能包含数字，以及（可选的）数字后面的星号。回答开头和结尾，以及数字与星号之间也允许有空格。正确答案格式示例：

- 3.4*
- 2
- 3*

- 4 *
- -1

错误答案格式示例：

- 3*
- *4**
- 分数：2*
- -1*

金标准是报告中的分数（或 -1），由人类（本文作者）判定。部分金标准判断具有主观性。例如，当给出三个分数但没有上下文时，就假定它们分别代表严谨性、原创性和重要性，并取其平均值。当包含两个分数时，通常将其视为未知分数。

如果提取出的数字完全准确，或与准确值相差不超过（<=）0.005，则计为正确（这是为了允许少量舍入误差，例如保留两位小数）。这也包括 -1，因此准确率计算始终基于 1446 个项目。

为清楚起见，除空格以及以下字符外，输出中的任何符号都会导致自动判定失败：0123456789.*-

提示设计

这项挑战包含两个要素。首先，要提取文本中直接报告的正确分数；如果文本报告了原创性、严谨性和重要性分数，却没有报告总体得分，则提取这三项分数的平均值。这一挑战还要求在文本中不包含有效分数时返回 -1。第二个要素是确保大语言模型只返回分数，不输出任何其他内容。以下是一个试图同时实现这两个目标的提示词示例，但并非总能成功。

> 以下报告给出了以下分数之一：1*、2*、3*、4*，或一个数字，并且／或者包含一项评估；若分数未知，则用 -1 标记。如果报告中给出了分数，则返回该分数。否则，如果报告中有最终研究质量评分，则提取该分数。再否则，如果报告包含原创性、严谨性和重要性的分数，则报告这三个分数的平均值，不要报告任何计算过程。若以上情况均不符合，则返回 -1 表示缺失值。请按以下格式作答<分数>其中<分数> 为 1*、2*、3*、4* 之一，或一个数字；若为缺失值，则为 -1。只输出分数。[此处填写带有分数的文本]

实时测试

将提示词输入大语言模型的网页界面，再输入数据样本中的一段文本，即可实时测试提示设计，查看是否得到了正确结果。图1至图4展示了提交给 DeepSeek 的提示词示例，其中有些得到了正确回复，有些则没有。逐一测试所有文本可能不切实际，因此可以使用自动方法。

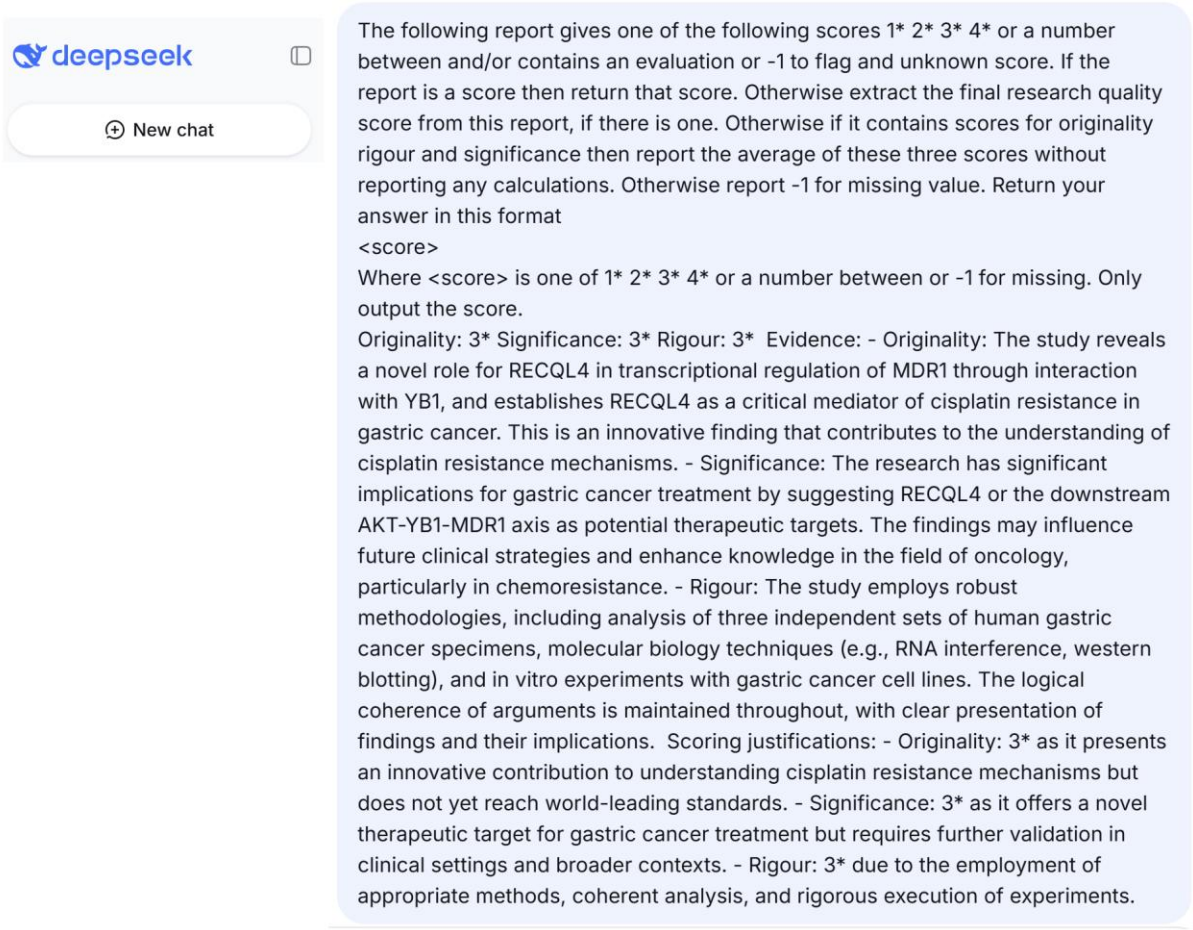
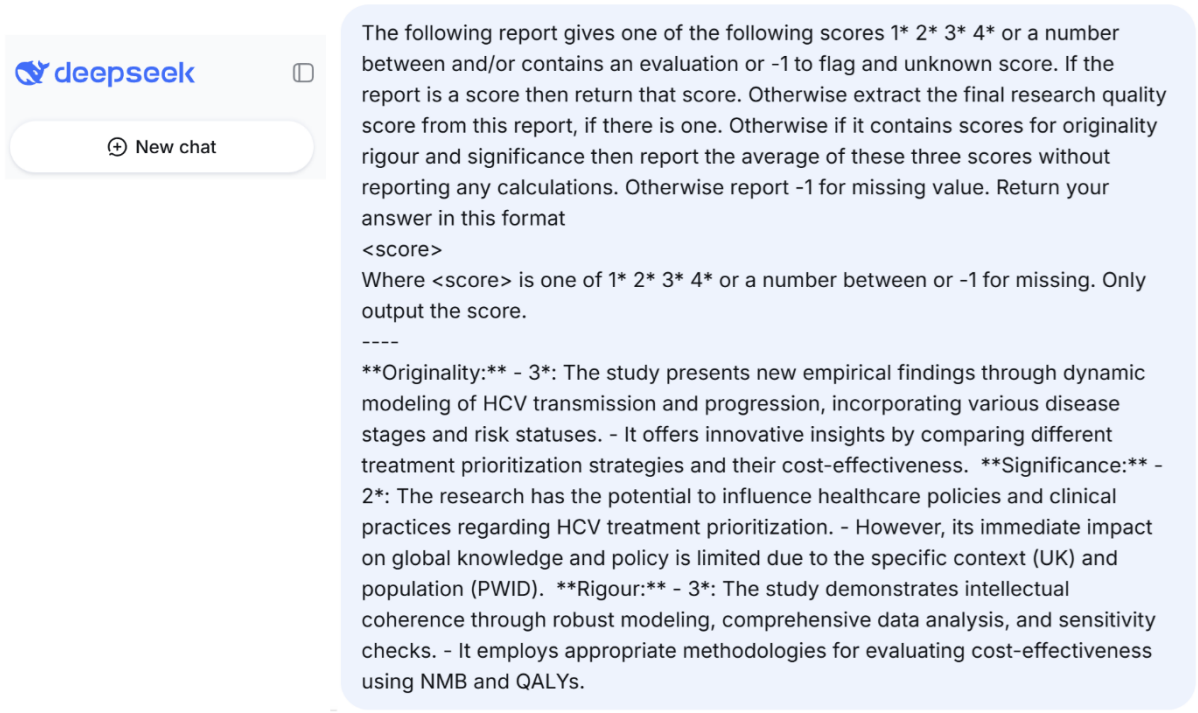
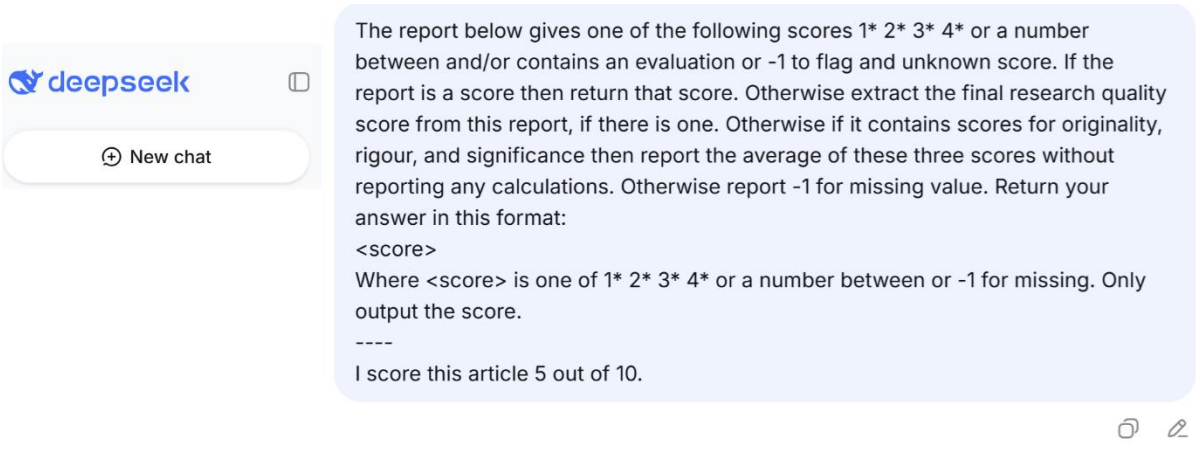


图1. 给 DeepSeek 的提示词，要求其仅给出正确答案 3*（位于蓝色气泡中提示词的下方偏左），不输出其他文本。



3*

图 2. 一个导致 DeepSeek 给出错误答案的提示词。正确答案是 2.667*，即 3*、2* 和 3* 的平均值。此示例中，提示词的第一部分略作了修改。



5

图 3. 一个导致 DeepSeek 给出错误答案的提示词。正确答案是 -1，因为该数字不在要求的 1* 至 4* 范围内。

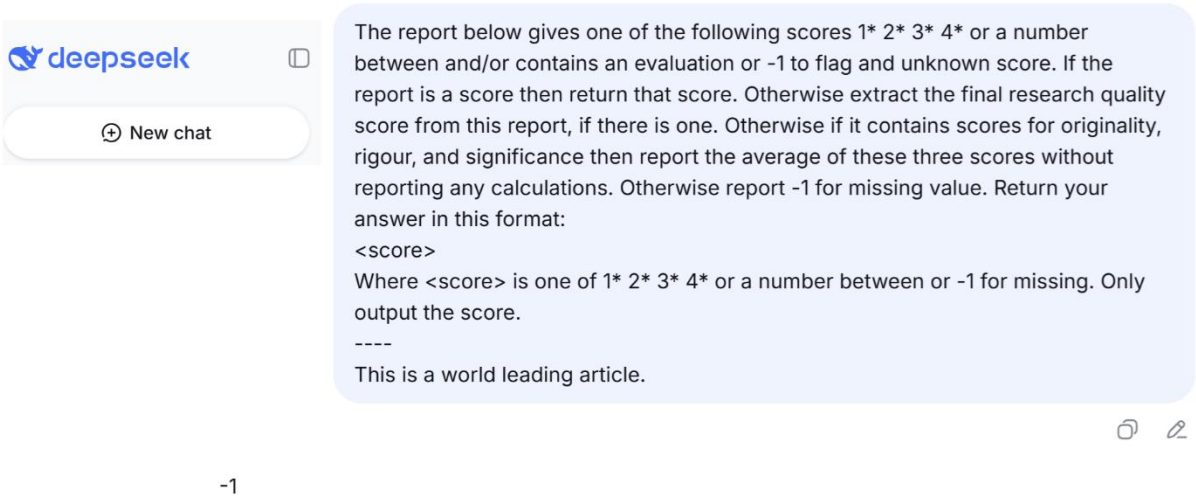


图 4. 向 DeepSeek 提供的一条提示词，其正确答案为 -1，因为报告中没有分数。事实上，“世界领先”是 REF 对 4* 的描述词，但 DeepSeek 不知道这一点，或者（正确地）没有将这项知识应用于该任务。无论如何，对于这项挑战，-1 被判定为正确答案。

向大语言模型提交大量提示词，有两种常见的自动方法：API 调用和本地运行的模型。一些大语言模型（例如 ChatGPT、Gemini、DeepSeek）提供应用程序编程接口（API），用户可以编写程序自动提交提示词并保存结果。通常这需要支付少量费用，但能快速处理大量提示词。另一种方式是，一些开发大语言模型的个人和组织会将模型分享到网上，通常通过 huggingface.co 分享，用户因此可以编写程序，将大语言模型的功能集成进去，从而自动处理大量提示词。本地运行大语言模型的主要缺点是速度慢，和/或需要强大的计算能力（主要是图形处理器）。

示例结果

上述提示词通过 API 提交给 ChatGPT 4.1-mini 使用。答案使用附录中的 Python 代码进行了核对（该代码由 ChatGPT 5 编写，也可在此处获取：

https://github.com/MikeThelwall/LargeLanguageModels/blob/main/1446_REF_分数_报告_正确_分数.py），输出如下。

> 总体统计数据（所有行）：答案数量：1446 正确答案数量：1050 正确率：72.61%

因此，ChatGPT 4.1-mini 给出了 72.61% 的正确答案。例如，作为一个错误答案的例子，ChatGPT 估计该报告中的分数为：

- 3*，原创性：4/5，重要性：4/5，严谨性：4/5，得分为 3.6，但正确答案是 3*。

摘要

本文介绍了一个报告共享数据集，其中的报告旨在包含分数，并附有从报告中提取出的正确分数；如果报告不包含分数，则标记为 -1。这是为了应对一项新任务：设计大语言模型提示词，从报告中提取分数（或 -1）。理想的提示词应能在所有情况下准确提取正确分数。若无法做到这一点，最佳提示词就是正确分数（包括 -1）比例最高的提示词。这似乎是一个适合研究人员和学生项目的挑战。希望这也有助于加深人们对大语言模型工作原理以及分数提取任务有效提示设计的理解。还可以设置以下挑战子任务。

- 设计一个提示词，以获取尽可能多的正确正向分数，同时不产生任何假阳性分数。为此，提示词大概应强烈要求大语言模型在没有十足把握时给出 -1。
- 使用任意给定的大语言模型，获取尽可能多的正确分数。
- 针对任意给定规模的大语言模型（例如 32b），获取尽可能多的正确分数。
- 针对基于大语言模型智能体的系统（而非单个大语言模型），获取尽可能多的正确分数。

参考文献

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L.，& Schulman, J.（2021）。训练验证器解决数学文字题。*arXiv* 预印本 *arXiv:2110.14168*。Hashemi, H., Eisner, J., Rosset, C., Van Durme, B., & Kedzie, C.（2024）。LLM-rubric：一种用于自动评估自然语言文本的多维校准方法。*arXiv* 预印本 *arXiv:2501.00274*。Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J.（2020）。测量大规模多任务语言理解。*arXiv* 预印本 *arXiv:2009.03300*。Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., & Petrov, S.（2019）。自然问题：问答研究的基准测试。计算语言学协会汇刊，*7*，453-466。Parulekar, A., & Jyothi, P.（2025年11月）。LASER：基于大语言模型的自动语音识别评分与评估量规。载于*2025*年自然语言处理实证方法会议论文集（第24773-24782页）。REF2021（2019）。投稿指南与标准：REF 2021期间发布的信息与指南汇编。https://2021.ref.ac.uk/guidance-and-criteria-on-submissions/index.html Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., & Wang, G. X.（2023）。超越模仿游戏：量化并外推语言模型的能力。机器学习研究汇刊。https://arxiv.org/abs/2206.04615 Stureborg, R., Alikaniotis, D., & Suhara, Y.（2024）。大语言模型是不一致且有偏见的评估者。*arXiv* 预印本 *arXiv:2405.01724*。Tang, X., Chen, H., Lin, D., & Li, K.（2024）。利用大语言模型进行多维写作评估：信度及与人工判断的一致性。*Heliyon*，*10*（14）。Thelwall, M.（2025）。大语言模型分数提取挑战数据集。Figshare.com。https://doi.org/10.6084/m9.figshare.30712835.v1 Thelwall, M., & Mohammadi, E.（2025）。小型语言模型和推理型大语言模型能否为期刊文章的研究质量评分？平均法和少样本方法是否有帮助？*arXiv* 预印本 *arXiv:2510.22389*。

Thelwall, M. & Yaghi, A.（2025）。ChatGPT在哪些领域能够检测期刊文章质量？一项对REF2021结果的评估。*Trends in Information Management*，13(1)，1-29。Thelwall, M. & Yang, Y.（2025）。ChatGPT生成的隐性与显性研究质量评分概率。Quantitative Science Studies。https://doi.org/10.1162/QSS.a.393 Voorhees, E. M., & Harman, D.（1999年12月）。文本检索会议（TREC）的历史及TREC-9计划。载于ACM SIGIR Forum（第33卷，第2期，第12-15页）。美国纽约州纽约市：ACM。Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., ... & Bowman, S.（2019）。SuperGLUE：面向通用语言理解系统的更具黏性的基准测试。神经信息处理系统进展，*32*。Wu, W., Zhang, Y., Haunschild, R., & Bornmann, L.（2025）。利用大语言模型开展发表后同行评审：潜力与局限。载于：编辑：Shushanik Sargsyan、Wolfgang Glänzel、Giovanni Abramo：第20届国际科学计量学与信息计量学会议，2025年6月23–27日，亚美尼亚埃里温，（第1176-1195页）。Yavuz, F., Çelik, Ö., & Yavaş Çelik, G.（2025）。利用大语言模型进行EFL作文评分：考察量规-型评估中的信度与效度。 *British Journal of Educational Technology*，*56*（1），150-166。

附录。用于检查答案的 Python 程序

导入 re，导入
pandas 并将其命名
为 pd，导入 os

# ----- 数据文件夹 ----- data_
folder = r"C:\Users\Mike"

----- 输入文件名 ----- file_scores = "1446_REF_chatgpt-4.1-mini_scores.txt" # 包含：标识符、分数 file_correct = "1446_REF_score_reports_correct_scores.txt" # 包含：标识符、正确分数

----- 输出文件名 ----- file_merged = "1446_REF_chatgpt-4.1-mini_scores_with_correctness.txt" # 制表符分隔输出数据框 file_report = "1446_REF_chatgpt-4.1-mini_scores_score_report.txt" # 纯文本摘要

os.chdir(data_folder) # -----
读取输入文件 ----- df_
scores = pd.read_csv(file_
scores, sep="\t", dtype={"
ID": int, "Score": str})

df_正确 = pd.read_csv(
文件_正确, 分隔符="\t",

```
    dtype={"ID": int, "正确_分数":
float} )

# 按标识符合并（内连接：仅保留两个文件中都存在的标识符）
df = pd.merge(df_scores, df_correct, on="ID", how="inner")

# ----- 为有效的分数格式准备正则表达式 ----- # 规则：# -
整个字符串可以有前导/尾随空格 # - 然后是可选符号、数
字和可选的十进制定点数部分 # - 然后是可选空格以及至
多一个星号 # - 匹配示例："3.4*"、"2"、"3* "、"4 *"、"-
1" # - 明确禁止 "-1*"（尽管其语法有效）pattern =
re.compile(r'^\s*([+-]?\d+(?:\.\d+)?)\s*(\*)?\s*$')


def score_is_correct(row): score_
raw = row["分数"]correct_分数 = row["
正确分数"]

    如果 score_raw 不是字符串：返
    回 0

    match = pattern.match(score_
    raw) if not match: # 格式错误
    return 0


    num_str, 星号 = match.group(1), match.group(2)

    # 按照示例，明确禁止使用“-1*”：
    try: num_val = float(num_str)
    except 值错误: return 0



    if 星号 and num val == -
    1:_return 0

    # 将数值与正确_分数进行比较 #（使用浮点数比较；如果需要精确的十
    进制定点数，请切换为十进制定点数）return int(abs(num_val -
    float(correct_score)) <= 0.005)

# 逐行应用函数 df["分数_是否_正确"] = df.apply(score_
is_correct, axis=1)

# ----- 将合并后的数据框保存为制表符分隔文件 -----
df.to_csv(file_merged, sep="\t", index=False)
```

```
# ----- 计算统计数据 ----- n_答案 = len(df) n_正确 = int(df["分数_是否_正确"].sum()) pct_正确 = 100.0 * n_正确 / n_答案 if n_答案 > 0 else float("nan")

# 忽略所有 Score 恰好为 "-1"（忽略空格）的行 mask_not_minus1 = df["Score"].str.strip() != "-1" df_no_minus1 = df[mask_not_minus1]

n_答案_no_minus1 = len(df_no_minus1) n_correct_no_minus1 = int(df_no_minus1["分数_is_correct"].sum()) pct_correct_no_minus1 = ( 100.0 * n_correct_no_minus1 / n_答案_no_minus1 if n answers no minus1 > 0_ _ _else float("nan") )

# ----- 编写纯文本报告 ----- with open(file_report, "w", encoding="utf-8") as f: f.write("总体统计数据（所有行）：\n") f.write(f"答案数量：{n_answers}\n") f.write(f"正确答案数量：{n_correct}\n") f.write(f"正确率：{pct_correct:.2f}%\n\n")

    f.write('忽略 Score 为 "-1" 的行后的统计数据：\n') f.write(f"答案数量（Score != -1）：{n_answers_no_minus1}\n") f.write(f"正确答案数量（Score != -1）：{n_correct_no_minus1}\n") f.write(f"正确率（Score != -1）：{pct_correct_no_minus1:.2f}%\n")
```