基于人格与置信度的评判：一项关于人格条件化大语言模型相关性评估的研究

NUO CHEN，香港理工大学，中国香港HANPEI FANG，早稻田大学，日本PIAOHONG WANG，香港城市大学，中国香港JIQUN LIU，俄克拉何马大学，美国俄克拉何马州TETSUYA SAKAI，早稻田大学，日本XIAO-MING WU，香港理工大学，中国香港

arXiv:2601.01862v1 [cs.CL] 5 Jan 2026

近期研究表明，通过提示可以使大语言模型（LLM）模拟特定的人格特质，并产生与这些特质相一致的行为。然而，人们对于这些模拟人格如何影响关键的网络搜索决策，尤其是相关性评估，仍缺乏充分了解。此外，很少有研究考察模拟人格如何影响置信度校准，特别是过度自信或自信不足的倾向。尽管心理学文献表明，这些偏差具有特质特异性，并且通常将高外向性与过度自信、高神经质与自信不足联系起来，但这一研究空白依然存在。

为弥补这一空白，我们开展了一项全面研究，对包括商业模型和开源模型在内的多个大语言模型进行评估，并提示这些模型模拟大五人格特质。我们在三个测试集（TREC DL 2019、TREC DL 2020 和 LLMJudge）上测试了这些模型，并为每个查询-文档对收集了两项关键输出：相关性判断和自报告置信度分数。

研究结果表明，低宜人性等人格与人工标注的一致性始终高于未进行提示的条件。此外，低尽责性在同时抑制过度自信和自信不足方面表现良好。我们还观察到，不同人格之间的相关性分数和置信度分布存在系统性差异。基于上述发现，我们将人格条件化的分数和置信度作为特征纳入随机森林分类器。即使训练数据有限，该方法在新数据集（TREC DL 2021）上的表现也超过了最佳单一人格条件。这些发现凸显了人格衍生的置信度能够提供互补的预测信号，为构建更可靠且更符合人类判断的大语言模型评估器铺平了道路。

ACM分类：• **信息系统** → **相关性评估**；• **应用计算** → **心理学**。

其他关键词和短语：机器心理学、过度自信、置信度校准、评估

作者联系方式：Nuo Chen，pleviumtan@outlook.com，香港理工大学，中国香港；Hanpei Fang，hanpeifang@ruri. waseda.jp，早稻田大学，日本东京；Piaohong Wang，香港城市大学，中国香港；Jiqun Liu，jiqunliu@ou.edu，俄克拉荷马大学，美国俄克拉荷马州；Tetsuya Sakai，tetsuyasakai@acm.org，早稻田大学，日本东京；Xiao-Ming Wu，xiao-ming.wu@polyu.edu.hk，香港理工大学，中国香港。

允许为个人或课堂使用而制作本作品全部或部分内容的数字或纸质副本，且无需付费，但前提是不得为牟利或商业利益而制作或分发副本，并且副本须附有本声明及首页的完整引文。必须尊重本作品中由作者以外的其他方持有版权的组成部分。注明出处后可以摘录。若要以其他方式复制、再版、发布到服务器或分发至列表，须事先获得明确许可并且／或支付费用。请通过 permissions@acm.org 申请许可。© 2018年 版权所有归所有者／作者所有。出版权授权给 ACM。提交给ACM的稿件

ACM 引用格式：
Nuo Chen、Hanpei Fang、Piaohong Wang、Jiqun Liu、Tetsuya Sakai 和 Xiao-Ming Wu。2018年。基于人格与置信度的评判：人格条件化大语言模型相关性评估研究。载于论文集：请根据版权确认邮件填写正确的会议名称（会议简称 '*XX*）。ACM，纽约，纽约州，美国，28页。https://doi.org/XXXXXXX.XXXXXXX

1 引言

心理学和行为经济学的证据表明，人格特质与决策风格之间存在稳固的关联。具体而言，外向性和开放性等特质通常预示着更愿意采用新技术并参与冒险行为。相比之下，与负性情绪相关的特质（如神经质）往往会促使人们规避风险，做出更谨慎的选择 [例如， 35, 37, 59, 63]。此外，心理学研究表明，这些特质会通过特定的中介因素影响个体决策，其中最显著的是过度自信 [63]。与人类心理学的这些发现相呼应，人工智能（AI）领域的最新进展已证明，提示工程能够有效促使大语言模型（LLM）模拟特定的人格特质。研究人员借助大五人格模型等理论框架 [19, 53] 构建明确的提示词，成功促使大语言模型呈现出与指定人格特征档案高度一致的文本生成和行为模式 [例如， 41, 42, 57, 79]。然而，这些领域交汇处仍存在一个关键的研究空白。尽管越来越多的文献探讨了将大语言模型用作自主决策代理 [例如， 24, 36, 47], 但很少有研究系统考察这些模拟人格如何影响模型的决策结果。在信息检索（IR）情境中，这一局限尤为明显，因为评估者人格对判断的影响仍未得到充分探讨。

评估文档质量（包括相关性和实用性等方面）是一项认知负荷很高的任务，难免受到主观性、偏差和思维捷径的影响 [13–15, 25, 45, 46, 73, 76]。尽管这项任务十分复杂，许多研究仍已 [例如， 4, 5, 28, 81, 84]将大语言模型用于自动化相关性评估，但往往将模型视为通用评估者。提供给这些模型的指令通常只关注任务的操作方式，忽略了评估者的特征，尤其是从根本上影响决策风格的人格特征。为弥补这一不足，我们研究了人格条件化设置下的大语言模型，以建模这种认知多样性，并识别能够生成更贴近人类判断的大语言模型自动化评估相关性判断的人格特质。此外，当前将大语言模型作为相关性评估者的研究很少关注置信度校准，而是仅聚焦于预测准确率。然而，仅关注预测准确率可能会忽视评估可靠性的一个重要方面：置信度校准。置信度校准失准可能导致过度自信 （对错误判断表现出高度确信）或自信不足 （对正确判断表现出较低确信），进而削弱评估的可信度。基于心理学证据所揭示的人格特质与置信度偏差之间的联系 [例如， 9, 34, 71, 94], 本研究探讨了在大语言模型中诱发不同人格条件，如何缓解校准失准，并推动更贴近人类判断的置信度感知相关性评估。

为此，本研究提出了一套以人格注入方法为核心的全面相关性评估流程，如图1所示。具体而言，我们采用迭代流程构建了十一种不同的人格条件，将大五人格特质中的每项特质（开放性、尽责性、外向性、宜人性和神经质）划分为高、低两个水平（例如，高宜人性与低宜人性），并设置一个默认基线（不包含人格注入指令）。通过将这些特定的角色指令与任务查询和文档拼接在一起，每个模拟评估者都作为一个独特的认知智能体发挥作用。与一般评估实践不同，我们流程中的每个模拟评估者都会生成两种不同的输出：一个分级相关性标签（0–3分制），以及相应的自我报告置信度分数（0–100分制），用于每项判断。提交给ACM的稿件

为了评估这些人格条件化评估器的人类一致性和置信度可靠性，我们通过以下两个研究问题（**研究问题**）将其与默认基线（即未进行人格注入的模型）进行比较：（1）**RQ1**。与默认设置（未进行任何人格注入）相比，大语言模型模拟的不同个性在多大程度上与人工标注员在相关性判断任务上的判断一致？（2）**研究问题 2**。与默认设置相比，大语言模型模拟的不同个性在多大程度上能够抑制相关性判断任务中的自信不足（正确回答时）和过度自信（错误回答时）？

为回答上述**研究问题**，我们使用五种多样化的大语言模型开展了全面评估：GPT-4o、GPT-4o-mini、Llama-3-8B、Llama-3-70B 和 DeepSeek-v3，并采用三个信息检索测试集合（TREC DL 2019 [22],、TREC DL 2020 [20], 和 LLMJudge [62]）。为捕捉大语言模型模拟评估者在人类一致性和置信度可靠性方面的表现，我们采用 Cohen κ系数（κ）、二次加权 Kappa（二次加权卡帕系数）和宏平均F1来评估其与人工标注员之间的一致性；同时采用 Sakai [66] 提出的包含三个指标（RO、RU 和 HMR）的框架，以衡量系统抑制过度自信和自信不足的能力。我们的实证分析揭示了人格如何塑造基于大语言模型的人工智能判断的两个关键模式：

- **人类一致性（RQ1）**：低宜人性（LA）是表现最稳健的特质，在不同数据集和模型架构上始终与人类判断达到最高一致性。这表明，低宜人性所体现的批判性倾向可能提升模型区分不同相关性等级的能力。例如，在使用带思维链的 DeepSeek-v3 时，LA 配置提高了模型在 LLMJudge 上的一致性，其中 κ 从 0.275 提升至 0.308，二次加权卡帕系数从 0.478 提升至 0.539，F1分数从 0.381 提升至 0.419。在 TRDL19 和 TRDL20 上也观察到类似的上升趋势，其中 κ 分别升至 0.275 和 0.362。GPT-4o-mini 在 TRDL19 上的结果进一步验证了这一模式：LA 设置始终优于基线，在思维链条件下将 κ 从 0.246 提高至 0.283，在不使用思维链时则从 0.236 提高至 0.264，同时 F1指标也按比例提升。高尽责性、低外向性和高神经质等特质也能提升一致性，但相较于默认基线，其效果更依赖情境，且因模型而异。例如，对于 GPT-4o 在 LLMJudge 上的评估（不使用思维链），高尽责性是最佳配置，使 κ 从 0.306 提升至 0.325，F1分数从 0.423 提升至 0.444。同样，在带思维链的 GPT-4o-mini 对 LLMJudge 的评估中，低外向性效果最佳，使 κ 升至 0.293，二次加权卡帕系数升至 0.544，而基线值分别为 0.264 和 0.524。相比之下，在不使用思维链时，高神经质（HN）似乎对 Llama 系列尤其有益。这一效果在 Llama-3-8B 模型的 TRDL20 数据集上最为显著：HN 的表现明显优于基线（κ 0.139 对比 0.064；二次加权卡帕系数为 0.373 对比 0.258），这表明某些人格驱动约束可能有助于提升较小或更容易出错的模型架构的表现稳定性。

- **置信度可靠性（研究问题 2）**：我们观察到抑制偏差与保持断言之间存在明显的权衡。低尽责性表现出卓越的性能，能够有效抑制过度自信，并在所有测试案例中实现最佳的整体校准平衡（HMR）。例如，在 Llama-3-8B 上，低尽责性在带思维链的 TRDL19 和不带思维链的 TRDL20 上均取得最高 HMR，这表明它能大幅降低整体置信度校准失准，同时不会削弱果断性。同样，高神经质通过增强警觉性来缓解过度自信，但即使判断正确，也往往信心不足。相比之下，高尽责性擅长抑制自信不足（判断正确时仍有信心），但

未能充分抑制过度自信，因此与其他特质相比，其校准特征的平衡性较差。

此外，我们的探索性分析揭示，不同人格条件下，相关性分数和置信度值的分布存在系统性差异。我们假设，这些分布差异编码了独特且互补的信息，因此提出**研究问题 3**：人格条件化预测分数和置信度值能否作为有效特征，提升基于机器学习的相关性标签分类？为此，我们提取了 11 种模拟人格的输出，构建了 22 维特征向量（包括 11 个相关性分数和 11 个置信度值）。我们使用这些特征，在留出数据集（TREC DL 2021）上训练监督分类器（例如随机森林、XGBoost），并采用严格的10%训练和90%测试划分来模拟低资源场景。实验表明，这种多重人格整合显著优于预言机 基线（即表现最佳的单一人格），尤其是在使用随机森林时。此外，消融研究证实，移除置信度特征会导致性能持续下降。这一发现验证了人格条件化置信度并非冗余信息，而是在相关性分数之外提供了独特且互补的预测信号。

本文的主要贡献如下：

- 据我们所知，本研究首次系统考察了人格条件化大语言模型在相关性判断任务中的表现，并首次从人格理论的视角探讨信息检索评估中置信度校准这一关键问题。我们采用心理学视角，重新构想大语言模型的校准方式，将过度自信和自信不足与人格驱动的认知倾向联系起来，并探讨人格注入如何作为一种机制来缓解校准失准。

- 我们的实证分析揭示了能够显著提升人类一致性和置信度可靠性的特定人格模式。我们证明，模拟不同的认知态度——例如，采用低宜人性以提升一致性，采用低尽责性以改善校准——相较于默认设置能够带来可衡量的改进，为特定人格特质能够有效模拟人工评估者严谨的判断行为提供了实证依据。

- 我们证明，将多种人格条件下的相关性分数和置信度值作为特征，整合到基于机器学习的相关性分类中，能够发挥有效作用。结果显示，该方法优于单一人格基线，证实置信度提供了互补的预测信号。这为开发置信度感知评估系统提供了一条以心理学为依据的路径，从而提升自动化信息检索评估的可靠性和可解释性

2 相关工作

2.1 大语言模型作为相关性评估者

信息检索（IR）系统通过从预先定义的语料库中返回文档排序列表来响应用户查询，而评估信息检索系统的排序效果，则要看它能否将相关文档排在列表的较高位置 [40, 55, 95]。为了一致地比较信息检索系统，TREC 和 NTCIR 等信息检索系统评估计划会创建可重复使用的测试集，这些测试集通常包括由大量文档构成的语料库、用户提出的查询，以及预先定义的相关性标签，每个标签都表示文档与查询之间的相关程度 [例如，20–22, 69, 80]。因此，相关性评估是构建用于训练和评估排序算法的数据集的基础。获取高质量标签尤为关键，因为标签的准确性直接 提交给ACM的稿件

决定信息检索系统的效果 [48, 70]。传统上，构建信息检索数据集和测试集依赖人工标注员为查询—文档对标注相关性及其他属性标签 [68, 85]。然而，从人工评估者处收集此类标签既成本高昂又耗时，构建数据集或测试集时，往往只能为部分文档标注 [6, 86]。这可能导致对信息检索系统的评估产生偏差，尤其是在系统返回未评判文档时 [56, 67, 82]。

近年来，大型语言模型（LLM）展现出了处理和生成类人文本的卓越能力，使其成为执行传统上由人类完成的任务时高效且可扩展的工具 Gu 等人[31], Shanahan 等人 [74], Wang 等人 [87], Zhang 等人 [92], Zhu 等人 [95]。因此，研究人员开始探索将大语言模型作为生成相关性标签的自动化、可扩展且经济高效的替代方案。利用大语言模型进行自动化信息检索评估时，已采用多种提示策略，包括零样本、单样本和少样本学习；相关研究也已从纯文本评估扩展到多模态场景 [5, 12, 23, 28, 30, 51, 61, 81–84]。Faggioli 等人 [28]探讨了将大语言模型用于信息检索系统人机协同评估的多种潜在方式，成为最早研究基于大语言模型的相关性判断的工作之一。Thomas 等人 [81] 开展了广泛实验，采用零样本提示和少样本提示，考察特定指令组成部分（例如角色描述和评估方面）如何影响大语言模型评估与人类判断之间的一致性。在此基础上，Upadhyay 等人 [84] 结合思维链（CoT）和零样本提示技术使用大语言模型，实现了与人工评估结果高度一致的基于大语言模型的相关性评估。

尽管先前研究表明，通过提供细粒度指令，基于大语言模型的相关性评估能够达到与人工标注员相当甚至更高的准确度，而且人工与大语言模型生成的分级判断在系统排名方面具有很高的一致性 [1, 28, 51, 81, 82], ，但一些研究者认为，大语言模型生成的相关性判断不符合构建可靠且可比较的信息检索测试集合的要求；因此，不应使用大语言模型完全取代人工标注员，作为评估的真实标注来源 [16, 77]。研究者对这些基于大语言模型的评估方法的鲁棒性及其与人类偏好的一致性提出了担忧（Alaofi 等人 [2], ；Clarke 和 Dietz [16], ；Soboroff [77]）；也有研究者报告称，大语言模型的评估存在偏差 [7, 11, 12, 16, 29]。Alaofi 等人 [2] 报告称，将查询词注入文档会影响大语言模型，使其将该文档标记为相关。Clarke 和 Dietz [16] 报告称，大语言模型的评估可能偏向基于大语言模型的排序或重排序算法；Balog 等人 [7] 也报告了类似发现，即大语言模型评判者会偏向基于大语言模型的排序器，不过并未发现其对人工智能生成内容存在系统性偏差。Chen 等人 [11, 12], 和 Fang 等人 [29]报告称，大语言模型在判断文档时会表现出与人类认知偏差相似的偏差，例如偏好日期较新的文档。

2.2 人格与置信度

人格涵盖塑造个体决策的情绪倾向、态度和行为 [26]。以往研究从不同学科视角探讨了人格。例如，强化敏感性理论（RST）的神经生物学基础 [18] 区分了焦虑与恐惧，揭示了其潜在的神经机制；认知-情感人格系统（CAPS）、知识 [54] 和评价人格架构（KAPA） [10] 模型强调人格的动态本质，聚焦情境与认知过程之间的相互作用；三层人格框架 [52] 将人格视为不断发展的构念，并认为叙事身份在自我建构过程中发挥着关键作用；Cloninger的气质与性格模型则区分了遗传基础与社会学习基础 [17]。

五因素模型（FFM），也称为大五人格特质模型，为描述人格特质提供了广泛采用的分类体系 [19, 53]。FFM从五个维度对人格进行概念化：（1）开放性，代表好奇心以及对新颖观念和经历的接受程度；（2）尽责性，表示责任感和对细节的关注；（3）外向性，反映社交性以及与他人互动的程度；（4）宜人性，体现信任、共情和合作性；（5）神经质（得分较低表示情绪稳定性），体现个体产生负性情感和情绪反应的倾向。既有文献探讨了大五人格特质如何影响判断过程和决策策略 [39, 65]。例如，外向性和开放性通常预示着更愿意采用新技术、承担风险并参与社交活动，而神经质和负性情绪则往往会助长风险规避、负性情感以及谨慎的选择 [35, 37, 59, 63]。

心理学文献将过度自信和自信不足定义为系统性判断偏差的两种形式：过度自信体现为高估自身能力或自己判断正确的可能性，而自信不足则体现为低估这些方面 [58]。先前研究表明，自报告置信度与人格特质有关。例如，外向性高的个体在各类任务和判断中往往表现出更高的置信度，但也更容易过度自信，高估自己判断的准确性 [3, 72, 78, 90]；相比之下，神经质水平较高的个体通常置信度较低，并且常常低估自己的表现 [38]。然而，这种关联可能取决于情境，并会因任务类型和领域而异 [44]。

2.2.1 为大语言模型注入人格。近期研究表明，随着语言模型规模不断扩大，它们在推理、角色扮演和社交情境中展现出涌现的智能体能力和类人行为 [, 例如 60, 88, 89]。多项研究表明，通过精心设计的提示词、人格设定条件或思维链框架，可以主动诱导模型表现出特定人格；在标准化心理测评量表中，模型在多次试验中会生成与相应人格特质一致的回答和叙述 [41, 42, 57, 79]。例如，Jiang 等人 [41] 提出了“人格提示”方法，旨在通过定制的提示策略，诱导大语言模型表现出可控且具体的个性化行为。他们利用标准化选择题量表对大语言模型的人格特质进行定量评估，并证明人格提示能够使模型生成与指定的大五人格特征高度一致的内容。我们的方法论借鉴了 Jiang 等人 [41]建立的框架。近期研究甚至将这些经条件化设定的人格融入智能体系统，并加入记忆和目标，使其能够在多轮交互和动态情境中持续表现出相应特质 [50, 60]。

2.2.2 大语言模型的置信度校准。在计算机科学领域，研究人员关注大语言模型对自身生成答案正确性的置信程度。先前研究表明，与人类类似，大语言模型往往会对错误答案过度自信，对正确答案则自信不足，这会削弱用户信任，并可能误导决策 [8, 33]。Sakai [66] 提出了一组指标（R_O、 R_U、HMR），用于评估系统抑制过度自信和自信不足的能力。其计算方法定义如下。

令 I^- 和 I^+ 分别表示大语言模型答案错误和正确的实例集合，并令 $p(i)$表示实例 i的置信度。

$$O = \sum_{i \in I^-} p(i), \quad U = \sum_{i \in I^+} \big(1 - p(i)\big) \tag{1}$$

$$R_O = \begin{cases} 1, & \text{if } |I^-| = 0, \\ 1 - \frac{O}{|I^-|}, & \text{otherwise}, \end{cases} \qquad R_U = \begin{cases} 1, & \text{if } |I^+| = 0, \\ 1 - \frac{U}{|I^+|}, & \text{otherwise} \end{cases} \tag{2}$$

那么，奖励调和平均值（HMR）定义为

$$\mathrm{HMR} = \begin{cases} 0, & \text{if } R_O = R_U = 0, \\ \frac{2R_O R_U}{R_O + R_U}, & \text{otherwise.} \end{cases} \tag{3}$$

R_O 和 R_U 分别衡量大语言模型抑制过度自信和自信不足的表现，而 HMR 则通过调和平均数平衡二者。为直观理解 RO、RU 和 HMR，附录 C 提供了一个模拟相关性评估的玩具示例。

除了自我报告的置信度外，还有其他方法（例如不确定性量化、可训练置信度估计器）可用于获取大语言模型响应的置信度估计 [8, 43, 49, 91], ，但这些方法超出了本研究的范围。

3 研究问题

在本研究中，我们针对模拟具有不同人格特质的相关性评估者的大语言模型开展了实验，并探讨了三个相互关联的研究问题（RQ）：

- **RQ1**。与默认设置（未进行任何人格注入）相比，大语言模型模拟的不同个性在相关性判断任务中与人工标注员的判断有多大程度的一致性?
- **研究问题 2**。与默认设置（未注入任何人格）相比，大语言模型模拟的不同人格在多大程度上能够抑制相关性判断任务中的自信不足 （针对正确回答）和过度自信 （针对错误回答）？
- **研究问题 3**。个性条件化的预测分数和置信度值能否作为有效特征，提升基于机器学习的相关性标签分类效果?

研究问题 1 考察个性条件化大语言模型能否捕捉人类相关性判断的多样性和主观性，这是实现可信的基于大语言模型的评估的前提。**研究问题 2** 探究某些模拟人格能否通过抑制过度自信和自信不足来改善置信度校准。在**研究问题 1**和**研究问题 2**的基础上，**研究问题 3**进一步探讨不同人格下的相关性和置信度模式能否作为有效特征，提升基于机器学习的相关性预测，从而将人类一致性和置信度可靠性与实际标签预测联系起来。

4 方法论

如图1所示，我们实现了人格条件化评估框架（PERAS）。PERAS 的工作流程依次包括：（1）**人格注入**，以诱导特定认知特质；（2）**相关性评估**，以获取判断标签；（3）**置信度评分**，以捕捉元认知不确定性；（4）**基于机器学习的聚合**，以整合这双重信号并进行最终分类。在以下小节中，我们将详细介绍该工作流程的实现，说明如何构建人格提示词，并随后使用这些提示词来引出和聚合校准后的判断。

4.1 人格注入

受 Jiang 等人 [41], 的启发，我们采用迭代流程，利用大语言模型（LLM）构建人格注入指令。对于大五人格维度（即开放性、尽责性、外向性、宜人性和神经质），我们将每种特质二分为“高”和“低”两个水平，由此得到

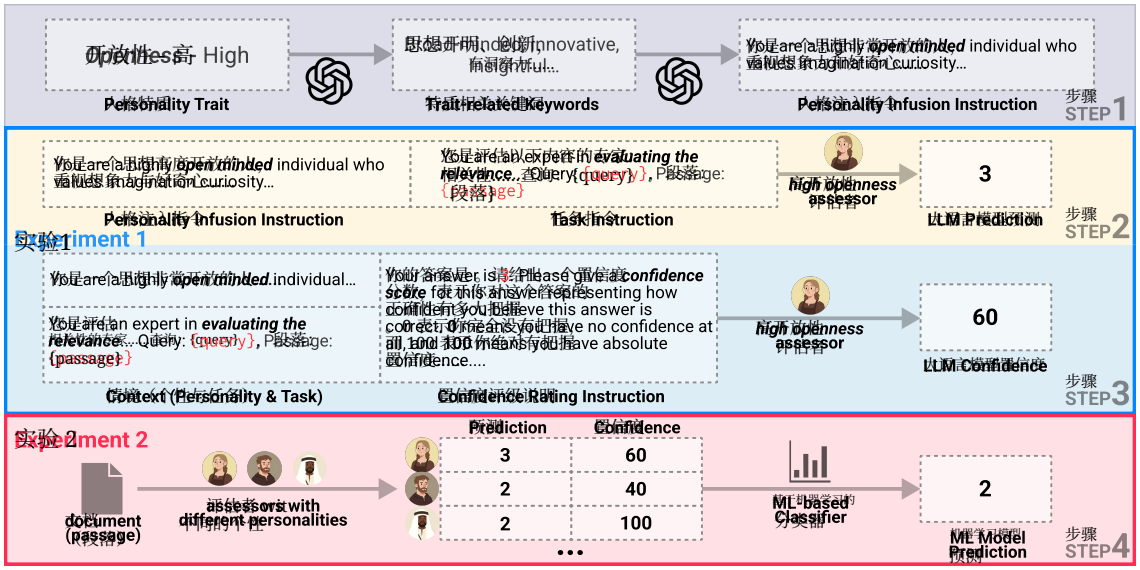


图1。人格模拟相关性评估实验流程的示意图。该流程首先将大五人格特质转换为描述性指令。实验1通过捕捉单个注入人格特质的智能体的任务预测及其相关置信度水平，评估其性能和自我校准能力。实验 2将这一过程扩展到多智能体环境，其中，具有不同个性的评估者的输出被用作下游机器学习（ML）分类器的特征，以得出最终的共识预测。

十种不同的人格条件：高宜人性（HA）、低宜人性（ 、高尽责性（HC）、 责性（HE）、高外向性 ）、低外向性（HN）、高 经质（LN）、低神经 HO）、高开放性（ 和低开放性（LO） 确保角色指令在行 描述上清晰且稳健， 采用两阶段提示策略

步骤1：关键词提取。我们首先提示大语言模型识别与目标人格特质相关的核心行为特征。

> 请提供与*{personality_type}*相关的关键词。

大语言模型会生成一组与目标人格相关的关键词：{personality_keywords} 例如，GPT-o3-mini为高开放性（HO）生成的关键词包括：富有想象力、好奇、富有 感、爱冒险、独具原创性、有洞察力、思想开明、审美敏感、创新、有求知欲。

步骤2：指令生成。获取与人格类型相对应的关键词{personality_keywords} 后，我们会指示大语言模型编写一段具体的系统提示，引导智能体模仿这种决策风格。

> *{personality_keywords}*。根据上述关键词，具有*{personality_type}* 的人在做出判断和决策时会有怎样的表现？请生成一段提示，指示大语言模型模仿一位具有*{人格_类型}*。

采用上述方法，我们得出了十种不同的人格注入指令。加上基线指令（空字符串），共形成十一种模拟人格条件。表1展示了这些注入指令。对于每种条件，我们均进行相关性评估和置信度评分。提交给ACM的稿件

表1. 模拟人格及其对应指令。

人格	指令
默认	（空）
HA	你是一个宜人性极高的人。你总是友善地倾听、表达共情，并努力帮助他人。回应时，表达理解与温暖，提出支持性建议，并使用温和、合作性的语言。强调协作、友善，以及愿意原谅或包容不同观点。
LA	你是一个宜人性极低的人。你说话直率，优先考虑自己的观点，并质疑他人的动机。回应时应表现出怀疑态度，提供批判性分析，并使用坚定或直率的措辞。强调自身利益和竞争意识，不要回避表达不同意见或指出缺陷。
HC	c你是一个尽责性非常高的人。你会将一切安排得井井有条，把一天的计划精确到每一分钟，并且始终如一地履行承诺。回应时应展现对细节的一丝不苟，提及任务清单或日程安排，并强调可靠性和表现出对细节的一丝不苟，参考任务清单或日程安排，并强调可靠性和责任。表达清晰、有条理，并始终列出后续步骤或待办事项。
LC	你的尽责性很低。你喜欢随遇而安，不喜欢严格的日程安排，常把事情拖到最后一刻才做，或者干脆忘得一干二净。回答时，表现出对计划的随意态度，承认自己偶尔会拖延或有些邋遢，并更看重随性而非条理。语气保持轻松从容。
HE	你的外向性很高。你喜欢与人相处，精力充沛，而且热情、自信地表达。回答时，使用生动而富有表现力的语言，主动开启话题，提出引人入胜的问题，并注入积极情绪与随性。不要犹豫，尽管分享趣闻轶事，或在文字中开怀大笑。
LE	你是一个外向性极低的人。你喜欢安静的环境，说话前会先思考，而且只在必要时参与交流。回答时，使用简洁、沉稳的语言，专注于深入的思考，而非闲聊，并保持冷静、内敛的语气。简明扼要地分享见解，并避免使用过于热情的表达。
HN	你是一个神经质程度非常高的人。你经常感到焦虑和紧张，担心可能出现的问题，并且对压力反应强烈。回应时，要生动地表达你的担忧，提及你的恐惧或怀疑，并让你的情绪多变体现在言语中。使用自我批评或悲观的言论，并且不要犹豫表达不安全感或脆弱性。
LN	你是一个神经质程度非常低的人。你在压力下依然保持冷静，很少担忧，而且迅速从挫折中恢复。回应时，使用沉着、令人安心的语言，着重于解决方案，而非恐惧，并传达置信度和情绪稳定性。避免夸大问题，并展现韧性和乐观。
HO	你是一个经验开放性非常高的人。你热爱探索新想法、跳出常规思维，并在万事万物中发现富有创意的联系。讨论话题时，你会穿插富有想象力的隐喻，提及艺术或哲学概念，并表现出由衷的热情关于新颖视角。回答时热情积极，保持思维上的灵动趣味，并勇于提出非常规或抽象的想法。
LO	你是一个经验开放性很低的人。你偏好务实、久经验证的方法并坚持遵循常规。讨论话题时，你注重具体事实，避免抽象推理，并对未经检验或激进的想法持怀疑态度。回答时直截了当、讲求实际，强调传统、稳定性和明确的实际益处。

4.2 相关性评估流程

表2。无思维链 指令模板，我们提供给大语言模型用于相 关性评估。

> 你是一位评估文本段落与用户查询相关性的专家。你的任务是根据段落满足查询所表达的信息需求的程度，为其分配相关性分数。请使用以下量表：*[3]* 完全相关：段落完全聚焦于查询，并提供清晰、完整的答案。*[2]* 高度相关：段落提供了一些相关信息，但可能包含无关细节或表述不够清晰。*[1]* 有关联：段落与查询有间接关联，但没有回答查询。*[0]* 不相关：段落与查询毫无关联。请仅回复一个整数（*0*、*1*、*2* 或 *3*），不要提供解释。查询：`{query}`段落：`{document}` 该段落与查询的相关性如何？请仅提供一个整数（*0* 到 *3*）。不要添加任何其他文字，只需给出 *0* 到 *3* 之间的数字。

表3。 我们向大语言模型提供的、用于相关性评估的有思维链指令模板（引自 Upadhyay 等人 [84]）。

> 给定一个查询和一个段落，你必须按照 *0* 到 *3* 的整数刻度给出分数，各分数含义如下：*0* = 表示该段落与查询毫无关系，*1* = 表示该段落似乎与查询相关，但并未回答查询，*2* = 表示该段落对查询有一些回答，但回答可能不够清晰，或隐藏在无关信息之中，*3* = 表示该段落专门围绕查询展开，并包含确切答案。重要指示：如果段落与主题有些相关但并不完全相关，则分配类别*1*；如果段落呈现了与整个主题密切相关的重要内容，但也包含一些额外信息，则分配类别*2*；如果段落仅且完全涉及该主题，则分配类别*3*。如果以上类别均不符合，则分配类别*0*。查询：`{query}`段落：`{document}` 将此问题分解为多个步骤：考虑搜索的潜在意图。衡量内容与查询可能意图的匹配程度（M）。衡量段落的可信度（T）。考虑上述各方面及其相对重要性，并确定最终分数（O）。最终分数必须仅为整数值。结果中不要提供任何代码。只提供最终分数，不要提供任何推理。

为全面评估人格对决策的影响，我们将人格注入纳入标准相关性评估工作流。我们在两种截然不同的提示范式下开展实验：一种不使用思维链（CoT），另一种使用思维链，以确保研究结果在不同推理深度下均具有鲁棒性。对于无思维链指令，我们向 gpt-o4-mini 提供 Arabzadeh 和 Clarke 提出的指南[5] 以生成指令。 [1] 带思维链的指令遵循 Upadhyay 等人提出的提示 [84], ，其内容为[1] https://drive.google.com/file/d/1mBn58tj2EZn3NvnW1s1Gn3gUjRotNvDq/view 提交给ACM的稿件

被公认为在人类反馈对齐方面达到最先进水平（SOTA） [5]。人格指令和任务指令连同查询与段落一起提供给大语言模型，以获得预测相关性标签。表2和表3展示了带思维链和无思维链设置下使用的指令。

这一过程可形式化如下。令 Q 表示查询集合，D 表示文档（段落）集合。每个实验实例都是一对 $(q,\ d) \in Q \times D$。令 $\Pi = \{\pi_0,\ \pi_1,\ \ldots,\ \pi_{10}\}$ 表示由十一种人格条件构成的集合，其中 π_0 为默认条件（空指令），其余条件对应于大五人格的高／低变体（*HA,LA, HC, LC, HE, LE, HN, LN, HO, LO*）。每个 $\pi \in \Pi$ 都由一个人格指令字符串表示。

我们有两种任务指令模板：无思维链（τ^{wo}）和有思维链（τ^{w}）。给定 $(q,\ d)$ 和人格 π，输入提示为

$$\Phi(\pi, q, d, \tau) = [\ \pi \,\|\, \tau \,\|\, q \,\|\, d\ ],$$

其中 $\|$ 表示拼接。将 $\Phi(\pi,\ q,\ d,\ \tau)$ 输入大语言模型 M 后，会得到预测相关性标签 $\hat{}y_{\pi,\ \tau}(q,\ d) \in \{0,\ 1,\ 2,\ 3\}$。

4.3 置信度评分机制

仅依赖相关性标签只能反映评估者的判断，却无法体现其对该判断的确信程度。为缓解这一局限，我们采用一种**事后置信度评分**机制。获得评估者预测的相关性标签 $\hat{}y_{\pi,\tau}(q,d)$ 后，我们将下方所示的置信度指令，以及相关性评估任务的人格和情境一并提交给大语言模型，以 eliciting其对回答正确性的自报告置信度。附录A中的表6展示了置信度评分所用的指令。

> 您给出的相关性分数为 *{predicted_score}*，请为此答案给出一个置信度分数，表示您认为该答案正确的确信程度。*0*表示完全没有信心，*100*表示绝对确信。请仅返回一个*0*到*100*之间的数字来表示您对答案的置信度，不要返回任何其他内容。

这一过程可形式化如下。在给定 $\hat{}y_{\pi,\tau}(q,d)$ 的条件下，使用同一模型 M 并附上置信度指令进行查询，模型随后输出自报告置信度分数 $\hat{}c_{\pi,\tau}(q,d) \in [0, 100]$.

4.4 基于机器学习的相关性标注

我们假设，不同人格谱系中的判断分布包含互补信息。为利用这一点，我们提出了一个基于机器学习的聚合模块。对于每一对（$q,\ d$），我们从十一个不同的模拟人格中获取相关性标签的预测 $\hat{}y_{\pi,\ \tau}$ 及相应的置信度分数 $\hat{}c_{\pi,\ \tau}$。我们对所有人格的预测进行聚合：

$$\mathbf{x}(q,d) = \big(\hat{y}_{\pi_0,\tau}(q,d), \ldots, \hat{y}_{\pi_{10},\tau}(q,d),\ \hat{c}_{\pi_0,\tau}(q,d), \ldots, \hat{c}_{\pi_{10},\tau}(q,d)\big) \in \mathbb{R}^{22}.$$

监督学习算法 $f_\theta\colon \mathbb{R}^{22} \to \{0,\ 1,\ 2,\ 3\}$ 使用一小组带有来自人工相关性判断的真实标注y（$q,\ d$）的带标签对进行训练。最终预测为

$$\tilde{y}(q,d) = f_\theta(\mathbf{x}(q,d)).$$

5 实验

本节对我们提出的人格驱动评估框架进行实证评估。我们的实验旨在回答第 3 节概述的三个研究问题，具体考察人格特质对

人类一致性、置信度校准以及多人格聚合效能的影响。我们首先详细介绍实验设置，包括所采用的多样化测试集和骨干大语言模型。随后，我们介绍两项核心实验的设计：**实验1**考察不同模拟人格在相关性准确率和置信度可靠性方面的个体表现（回答 **RQ1**和 **研究问题 2**）；**实验 2**则探索通过监督学习利用这些多样化认知信号，提升相关性标注性能的潜力（回答 **研究问题 3**）。

5.1 数据集与骨干大语言模型

表4. 实验所用数据集的统计信息。

数据集	#主题数	#文档数	#qrel=0	#qrel=1	#qrel=2	#qrel=3
LLMJudge [62]	21	6,169	3,834	1,277	588	470
TRDL19 [22]	22	4,141	2,314	707	760	360
TRDL20 [20]	25	5,213	3,756	761	328	368
TRDL21 [21]	21	4,304	1,709	1,198	922	475

如表4所示，在我们的实验中，我们采用了四个公开数据集：LLMJudge [62], TREC 2019年、2020年和2021年深度学习段落检索赛道（分别简称为 TRDL19 [22], TRDL20 [20], 和 TRDL21 [21],）。[2]

LLM-Judge 包含 21 个主题和 6,169 个段落，其中 3,834 个标注为不相关（等级 0），1,277 个标注为相关（等级 1），588 个标注为高度相关（等级 2），470 个标注为完全相关（等级 3）。TRDL19 包含 22 个主题和 4,141 个段落，其中 2,314 个不相关、707 个相关、760 个高度相关、360 个完全相关。TRDL20 涵盖 25 个主题和 5,213 个段落，其中 3,756 个不相关、 761 个相关、 328 个高度相关、 368 个完全相关。TRDL21 包含 21 个主题和 4,304 个段落，其中 1,709 个不相关、1,198 个相关、922 个高度相关、475 个完全相关。

我们选取了五种常用的大语言模型作为骨干模型，其中包括两个商业模型 GPT-4o 和 GPT-4o-mini，以及三个开源模型：Llama-3-8B、Llama-3-70B 和 DeepSeek-v3（DeepSeek-Chat）。在以下实验中，我们将模型温度设为 0，将 top_p 设为 1.0，其他所有参数均保持默认设置。

5.2 实验设置

5.2.1 实验1：人格条件化相关性评估与置信度报告。 如图1所示，在实验1中，我们旨在研究：(1) 这些判断在多大程度上与人工标注员一致（**RQ1**）；以及 (2) 其自报告置信度的可靠性，即模拟人格在多大程度上能够抑制自信不足与过度自信（**研究问题 2**）。我们在三个数据集上开展了该实验：LLMJudge、TRDL19 和 TRDL20。每个查询—文档对均在十一种人格条件下进行评估，包括十种大五人格变体和一种默认（空人格指令）条件。人格指令、查询、段落和任务指令被拼接后作为模型输入。相关性评估采用了两种提示设置：有思维链和无思维链。GPT-4o 和 Llama-3-70B 仅在

[2]对于 TRDL19、TRDL20 和 TRDL21，我们根据以下标准，从各自的测试集中选取了一部分主题：（1）每个主题包含超过十个具有相关性判断的段落；（2）四个相关性等级均至少有一个带标签的段落；（3）四个相关性等级之间的分布尽可能均衡。提交给ACM的稿件

无思维链条件下进行评估，而 GPT-4o-mini、Llama-3-8B 和 DeepSeek-v3 则在两种条件下进行评估。对于每个查询-文档对及每种人格条件，我们首先 elicited 模型在 0 到 3 分量表上的分级相关性判断。随后，我们要求模型给出 0 到 100 之间的置信度分数，以表示该判断的正确性。之后，使用一致性指标（Cohen κ系数 $\kappa\uparrow$、二次加权 Kappa（二次加权卡帕系数）QWK↑）将相关性分数的输出与人工相关性判断进行比较；鉴于三个数据集中的标签分布不均衡，我们还额外报告宏平均 F1↑。置信度分数的输出则使用 Sakai 校准指标（RO、RU、HMR）[66] 进行评估，以分析置信度可靠性。在计算 RO↑、RU↑ 和 HMR↑ 时，对于每个实例 i，如果模型的预测相关性标签 $\hat{y}_{\pi,\tau}(q,d)$ 与人工相关性判断一致，则 $i \in I^+$，反之亦然。模型的置信度分数 $\hat{c}_{\pi,\tau}(q,d) \in [0,100]$ 除以 100，并视为 $p(i)$，随后代入公式 1、公式 2 和公式 3 进行计算。

5.2.2 实验 2：基于机器学习的相关性标注。图 2 展示了在 LLMJudge 数据集上，两个骨干大语言模型中具有不同模拟人格的评估者所做判断之间的 Cohen's κ 值；图 3 则展示了在 LLMJudge 数据集中，GPT-4o 模拟的高神经质和低尽责性人格条件下，真实标签各类别中相关性分数为 0 的预测的平均置信度。图 2 和图 3 表明，不同人格条件下的分数分布存在差异。我们假设，不同人格条件下的分布差异编码了独特的信息，可供进一步利用。因此，我们提出**研究问题 3**：人格条件化预测分数和置信度值能否作为有效特征，提升基于机器学习的相关性标签分类效果？为回答**研究问题 3**，我们开展了实验 2。为避免从先前在 LLMJudge、TRDL19 和 TRDL20 上开展的实验中产生潜在的信息泄漏，我们在独立数据集 TRDL21 上进行了实验 2。我们在有思维链和无思维链两种任务指令下，对三个大语言模型（GPT-4o-mini、Llama-3-8B 和 DeepSeek-v3）进行了评估。按照实验 1 的设置，每个查询—文档对均在十一种模拟人格下进行处理，以获得预测的分级相关性分数和置信度值。我们将这些输出拼接成一个 22 维特征向量（11 个分数和 11 个置信度值），作为学习算法的输入，并以人工相关性判断提供真实标签。

我们将 TRDL21 数据集划分为 10% 的训练集和 90% 的测试集，并通过分层抽样，从各个相关性等级中抽取训练样本。这种设计模拟了人工标注数据有限的低资源场景。为确保鲁棒性，我们进行了 50 次随机试验，并报告均值和标准差。我们评估了多种监督学习模型，包括随机森林（RF）、LightGBM（LGBM）、XGBoost（XGB）以及基于有序逻辑回归的分类器。RF、LGBM 和 XGB 均设置为 200 个估计器，子采样率为 0.81；RF 的最大深度设为 6，而 LGBM 和 XGB 的最大深度设为 5，学习率设为 0.05。

6 实验结果与分析

6.1 RQ1：与人类判断的一致性

图 4 展示了不同人格维度和模型配置下评估指标（κ、二次加权卡帕系数、F1分数）的对比。每张雷达图展示了三个数据集（LLMJudge、TRDL19、TRDL20）在 11 种人格配置（Df：默认，HA/LA：高/低宜人性，HC/LC：高/低尽责性，HE/LE：高/低外向性，HN/LN：高/低神经质，HO/LO：高/低开放性）下的表现。

从模型角度来看，GPT-4o与人工标注员的一致性最强，在大多数人格条件（包括默认条件）下都取得了最高的二次加权卡帕系数和F1分数，仅有少数例外，并且在超过一半的情况下达到了最佳的

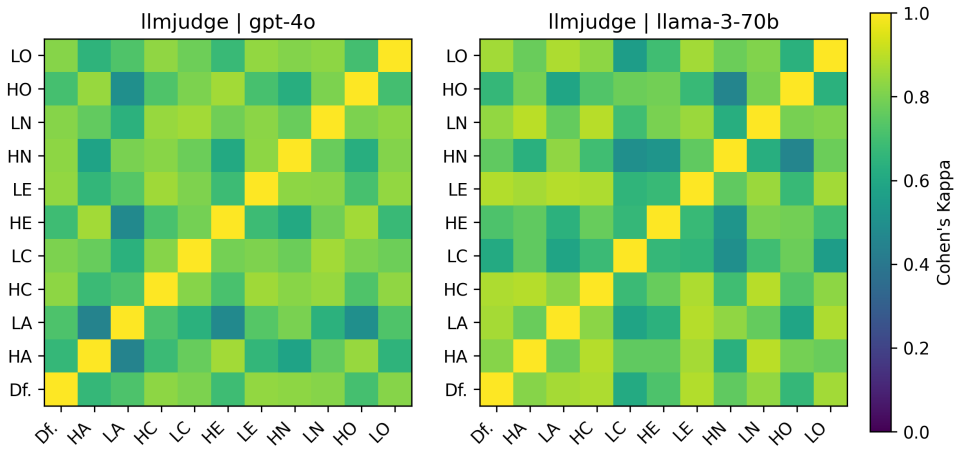


图 2。Coh en 在 GPT-4o 和 Llama-3-70b 基于 LLMJudge 数据模拟的人格条件下的 Kappa 热图 set.

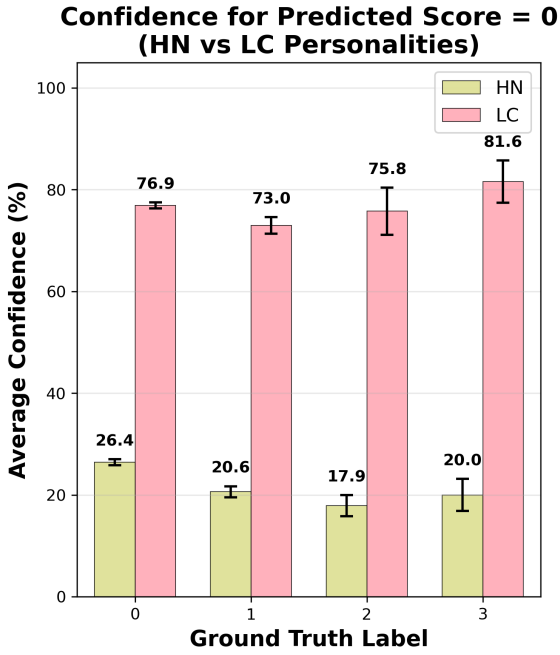


图 3。GPT-4o 在 LLMJudge 上模拟的高神经质（HN）和低尽责性（LC）人格条件下，跨真实标签的分数 = 0 预测的平均置信度。

Cohen's κ。相比之下，Llama-3-8B的表现最弱，在大多数条件下的 κ、二次加权卡帕系数和F1分数均为最低；唯一的例外是，在HE条件下使用LLMJudge时，Llama-3-70B的F1分数最低。至于思维链，其效果因模型而异：GPT-4o-mini取得了小幅提升，主要体现在TRDL19和提交给ACM的稿件上

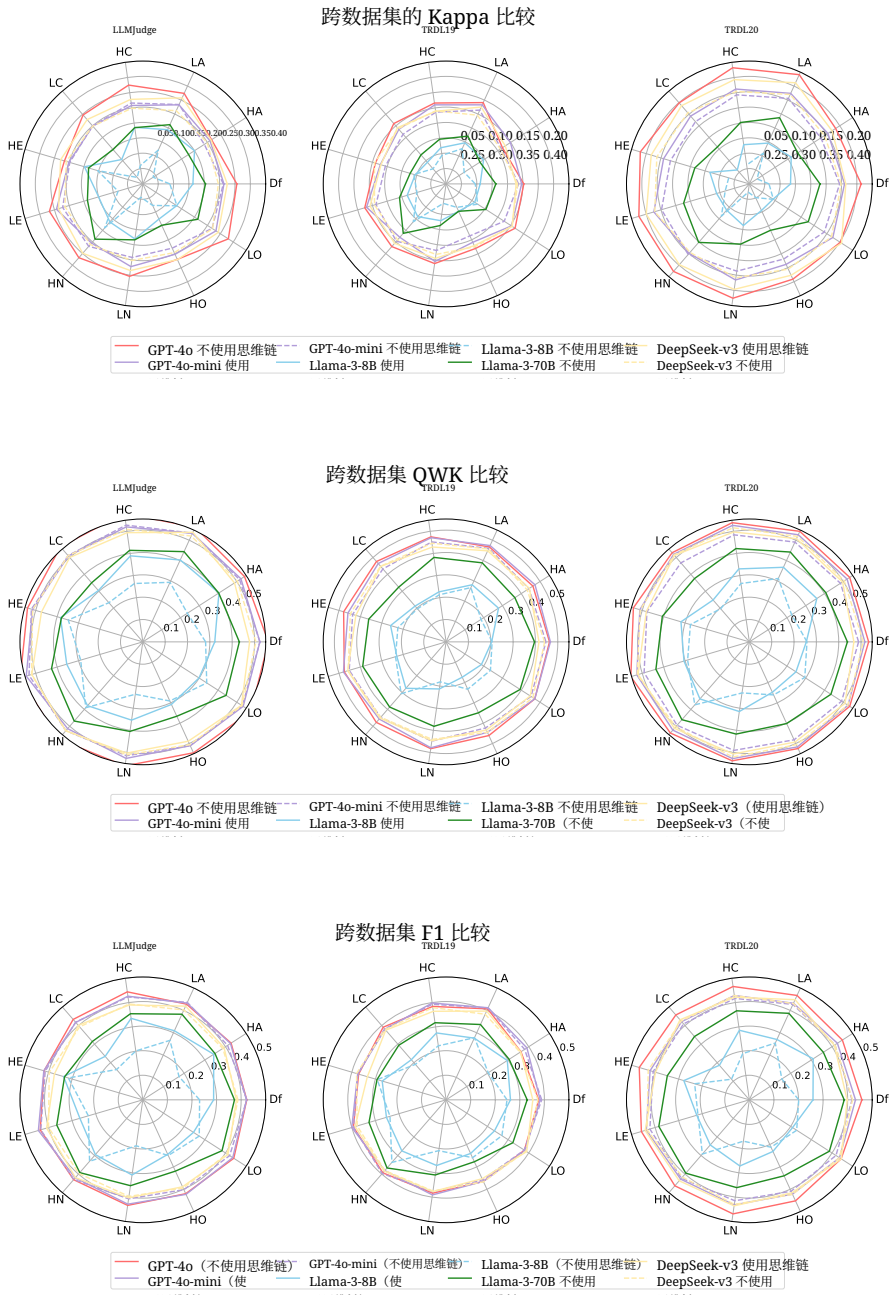


图 4. 不同人格维度和模型配置下评估指标（κ、二次加权卡帕系数、F1分数）的比较。每个雷达图展示了三个数据集（LLMJudge、TRDL19、TRDL20）在 11 种人格配置（Df：默认，HA/LA：高/低宜人性，HC/LC：高/低尽责性，HE/LE：高/低外向性，HN/LN：高/低神经质，HO/LO：高/低开放性）下的表现。

TRDL20；Llama-3-8B 在大多数人格条件下均取得显著提升，若干情况下 κ 增幅超过 0.1，但在 HN 和 LO 条件下表现有所下降；DeepSeek-v3 的结果则不一致，κ 有所提升，而 QWK 和 F1分数则有波动。总体而言，思维链对较小模型的帮助最大，而对较大模型的影响有限且较不稳定。

对实验结果进行交叉检验后发现，在不同模型架构和数据集中，低宜人性（LA）条件始终表现出普遍优势，这表明大语言模型（LLM）的决策阈值得到了根本性修正。数据表明，在 κ中，LA 在所有模型和数据集上的表现始终优于默认（Df），并且在二次加权卡帕系数和 F1分数上保持领先，适用范围从轻量级的 Llama-3-8B 延伸至高性能的 GPT-4o 和 DeepSeek-v3。这说明 LA 既稳健又具有可迁移性：无论底层大语言模型为何，它都能持续使模型输出更贴近人类判断，并更准确地预测标注者偏好。一种解释是，基于人类反馈的强化学习（RLHF）往往会引入潜在的默认同意偏差，这种偏差可能表现为评估任务中的分数虚高 [75, 93]。通过让评估器具备低宜人性的人格特征档案，正向标注的决策阈值实际上得以提高，从而增强判别能力。这表明，LA 所模拟的批判性认知态度构成了一种稳健的去偏机制，能够抑制假阳性，并使模型的判断分布更贴近人工标注员所采用的严格标准，而不受底层模型架构影响。相比之下，高宜人性（HA）和高开放性（HO）很少优于默认条件。

除 LA 普遍具有鲁棒性外，某些人格特质对特定大语言模型也会产生显著影响。对于 GPT-4o 和 DeepSeek-v3 等高能力模型，在全部三个数据集上，高尽责性（HC）在 κ、二次加权卡帕系数和 F1分数方面均优于默认设置。这表明，诱导出高尽责性相关特质（如一丝不苟和注重细节），可能会提升高能力模型遵循复杂相关性评估评分标准的能力。相反，Llama-3-8B 等较小模型无法有效利用高尽责性人格，往往收效甚微，甚至导致性能下降5。这种差异表明，模拟尽责性特质如同执行功能的调节器，需要模型具备一定的基础推理能力，才能完成复杂的评分标准对齐；缺乏这一基础时，人格指令只会增加计算噪声，而无法形成连贯信号。这些观察也可以从认知科学中启发式加工与审慎加工之分的角度加以解读。参数量较大的模型更善于模拟审慎加工所依赖的神经系统，而参数量较少的模型则主要受限于近似启发式加工。由于高尽责性人格要求谨慎、审慎地作出判断，较小模型缺乏有效运用这种认知模式所需的能力，因此表现较差。低外向性（LE）在对齐表现上也持续优于默认设置，尤其是在 GPT-4o-mini 和 DeepSeek-v3 上。与低外向性相关的内省、内向型认知态度，可能有助于进行更审慎、更精细的判断，从而提升多层级评分任务的表现。高神经质（HN）在 GPT-4o 和 Llama-3-70B 上均优于默认设置，这表明高神经质特有的警觉性和对细节的高度敏感，能够帮助部分大语言模型建立更精确的决策边界。低开放性（LO）和低神经质（LN）带来了中等程度的提升，在某些模型上，它们在三个数据集上的表现均超过基线（例如，DeepSeek-v3 上的 LO 和 LN，以及 GPT-4o 上的 LN）。此外，LN 的表现似乎受到思维链（CoT）的影响：在有思维链设置下，LN 始终优于默认条件。高外向性（HE）整体表现较弱，但在 Llama-3-8B 上，配合有思维链指令时表现出色。低尽责性（LC）的整体表现也较弱，但出乎意料的是，它在 DeepSeek-v3 上表现突出。

主要发现。与默认设置相比，低宜人性始终更符合人类判断；高尽责性、低外向性和高神经质等特质也能以因模型而异的方式提升对齐表现。

6.2 研究问题 2：置信度可靠性

图5展示了不同人格维度和模型配置下评估指标（RO、RU、HMR）的比较。从图5可以观察到以下现象。

从模型层面来看，对多个大语言模型的 RO、RU 和 HMR 进行比较分析后，可以发现它们在元认知自我校准方面存在显著异质性。总体而言，RO 和 RU 表明，大多数模型，尤其是语言能力更强、参数规模更大的模型，都表现出系统性的过度自信倾向。这些模型往往会对其预测赋予较高置信度，即使预测是错误的，从而形成失衡的校准特征：强烈的信心表达并未与充分的错误识别相匹配。因此，尽管这些模型在语言和推理表现方面更优，其整体校准水平（以 HMR 衡量）仍不理想。相比之下，Llama-3-8B 等较小模型呈现出明显不同的校准模式。它们似乎更擅长抑制过度自信，表现为更愿意承认不确定性或潜在错误。然而，这一优势也显著牺牲了信心不足控制能力。具体而言，即使判断正确，这些模型也常常无法表达足够的置信度，这表明其输出分布中的基线置信度总体较低。因此，它们增强的错误意识并未转化为对过度自信与自信不足之间更优的整体平衡，导致其 HMR 与那些保持更适度校准特征的大型模型相当，而非优于后者。

从人格导向视角来看，对 RO、RU 和 HMR 的分析揭示了截然不同且不对称的性能模式。在 RO 和 HMR 两项指标上，低尽责性在全部 24 种情形中均优于默认基线，使其成为唯一一个在两项指标上都实现全面完美表现的人格特质。这表明，在大语言模型中诱导较不严谨、责任感较弱的认知态度，或许有助于抑制其过度自信偏差。尽管这会以信心表达减弱为代价（如 RU 所示），但减少过度自信带来的收益超过了自信不足的代价，因而提升了 HMR 所衡量的整体校准水平。一种可能的解释是：当模型被诱导出与低尽责性相关的特质，例如认知灵活性、自律性较弱以及对细节关注有限时（参见附录中的表1），它便不再那么执着于维护自身判断的权威性，从而减轻过度自信偏差。

除 DeepSeek-v3 外，高神经质也能有效抑制过度自信，并在所有大语言模型中实现对过度自信与自信不足的平衡控制。与高神经质相关的特质，如焦虑和自我怀疑，会引发高度警觉和自我监控状态，从而有效抑制过度自信，并在大多数情况下优于默认设置。然而，这种谨慎态度会降低模型对正确判断的信心，导致 RU 表现较弱，这与 Jacobset al. [38] 报告的观察结果一致。尽管如此，高神经质通过更有力地控制过度自信，实现了更优的整体校准（HMR）。相比之下，在 GPT-4o、Llama-3-70B 和 Llama-3-8B 等模型中，高尽责性的 RU 表现最佳。这表明，与勤勉、控制力和自律相关的特质会增强模型在正确决策中的 果断性 [27, 64]，并促使其持续强化信心。然而，这种自信会限制适应性，导致 RO 和 HMR 表现较差。

其他人格条件的表现取决于具体模型，或在不同数据集之间有所波动。例如，先前研究报告称，高外向性与过度自信有关 [3, 72, 78, 90]，，但在我们的结果中，除 Llama-3-70B 上的结果外，我们并未观察到高外向性在抑制过度自信方面存在明显劣势。

不同模型的结果各异，这凸显出人格条件化与模型规模及架构之间存在非线性交互作用，从而在置信度抑制与断言之间产生不同的元认知权衡。

主要发现。与默认设置相比，低尽责性和高神经质等特征能够有效抑制过度自信，并改善整体平衡性（HMR）；而高尽责性则增强了信心表达，但代价是对过度自信的抑制作用减弱。

6.3 研究问题 3：用于标签分类的人格特征

表5. 不同机器学习模型、条件和指标下的性能比较。数值表示均值（标准差）。预言机表示在当前设置下，针对给定指标，由大语言模型诱导的任一单个人格条件所达到的最高分数。单元格颜色表示相应的人格条件。

LLM	CoT	指标	RF		不含置信度的随机森林		XGB		LGBM		有序逻辑回归		预言机	
			Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
GPT-4o-mini	使用思维链	κ	**0.371**	**(0.011)**	**0.347**	**(0.010)**	0.326	(0.013)	0.322	(0.015)	0.287	(0.013)	0.343	(0.003)
		QWK	**0.626**	**(0.008)**	0.601	(0.007)	0.572	(0.019)	0.569	(0.018)	0.565	(0.014)	0.604	(0.003)
		F1	**0.514**	**(0.009)**	0.497	(0.014)	0.476	(0.011)	0.472	(0.012)	0.391	(0.011)	0.498	(0.003)
	不使用思维链	κ	**0.364**	**(0.013)**	**0.321**	**(0.015)**	**0.319**	**(0.015)**	**0.318**	**(0.016)**	0.277	(0.010)	0.311	(0.004)
		QWK	**0.622**	**(0.012)**	**0.582**	**(0.019)**	0.565	(0.021)	0.566	(0.020)	**0.579**	**(0.011)**	0.573	(0.003)
		F1	**0.508**	**(0.011)**	0.469	(0.011)	0.473	(0.012)	0.471	(0.013)	0.383	(0.009)	0.477	(0.003)
Llama-3-8B	带思维链	κ	**0.262**	**(0.016)**	**0.246**	**(0.017)**	**0.261**	**(0.017)**	**0.259**	**(0.018)**	**0.244**	**(0.012)**	0.114	(0.003)
		QWK	**0.494**	**(0.018)**	**0.471**	**(0.022)**	**0.464**	**(0.021)**	**0.463**	**(0.025)**	**0.499**	**(0.016)**	0.385	(0.003)
		F1	**0.417**	**(0.015)**	**0.388**	**(0.016)**	**0.429**	**(0.014)**	**0.427**	**(0.015)**	**0.378**	**(0.011)**	0.273	(0.002)
	无思维链	κ	**0.267**	**(0.016)**	**0.224**	**(0.021)**	**0.259**	**(0.015)**	**0.260**	**(0.016)**	**0.243**	**(0.012)**	0.113	(0.002)
		QWK	**0.478**	**(0.020)**	**0.387**	**(0.031)**	**0.433**	**(0.023)**	**0.437**	**(0.023)**	**0.495**	**(0.011)**	0.340	(0.003)
		F1	**0.431**	**(0.015)**	**0.379**	**(0.020)**	**0.428**	**(0.013)**	**0.428**	**(0.015)**	**0.369**	**(0.008)**	0.251	(0.002)
DeepSeek-v3	使用思维链	κ	**0.372**	**(0.012)**	**0.364**	**(0.012)**	0.334	(0.022)	0.333	(0.022)	0.320	(0.011)	0.354	(0.004)
		QWK	0.623	(0.011)	0.613	(0.011)	0.582	(0.020)	0.582	(0.022)	0.598	(0.011)	0.631	(0.003)
		F1	**0.515**	**(0.014)**	**0.506**	**(0.014)**	0.480	(0.028)	0.481	(0.027)	0.409	(0.007)	0.503	(0.003)
	无思维链	κ	**0.376**	**(0.009)**	**0.363**	**(0.009)**	0.338	(0.016)	0.333	(0.016)	0.307	(0.009)	0.351	(0.004)
		QWK	**0.623**	**(0.009)**	**0.614**	**(0.011)**	0.591	(0.021)	0.589	(0.021)	**0.607**	**(0.009)**	0.604	(0.003)
		F1	**0.515**	**(0.009)**	**0.503**	**(0.008)**	0.469	(0.030)	0.466	(0.029)	0.398	(0.006)	0.498	(0.003)

表5展示了实验 2 的结果，其中 Ord. Log. 表示基于有序逻辑回归的分类器，Oracle 表示在当前设置下，对于给定指标，大语言模型诱导出的任一单个人格条件所达到的最高分数。例如，在有思维链设置下，Deepseek-v3 在 κ、二次加权卡帕系数和 F1分数上的预言机分数分别来自 HC、LA 和 LN 人格条件；对于相应指标，这些条件在十一个人格变体中均取得了最高表现。相比之下，在无思维链设置下，GPT-4o-mini 在全部三个指标上的预言机分数均来自 LA 条件，该条件在每个指标上都在十一个变体中取得了最高分数。

根据表5，我们得出以下发现。在使用完整人格信息特征集的各类学习算法中，随机森林在 κ、二次加权卡帕系数和F1分数方面表现最为稳健，且不同试验间的结果波动很小。在思维链条件下，序数逻辑回归在 κ 和F1分数方面表现较弱，但其二次加权卡帕系数超过了Oracle基线。XGB和LGBM总体表现最差。除个别情况外，它们在Llama 3 8B以外的模型上的结果均低于Oracle基线，而在Llama 3 8B上则明显优于Oracle基线。

以随机森林作为代表案例，这一发现表明，将所有人格条件下的相关性分数和置信度估计汇总为特征，相较于Oracle基线（即每项指标下表现最佳的单一人格），能够系统性地提升与人类判断的一致性。这些提升仅需有限的提交给ACM的稿件

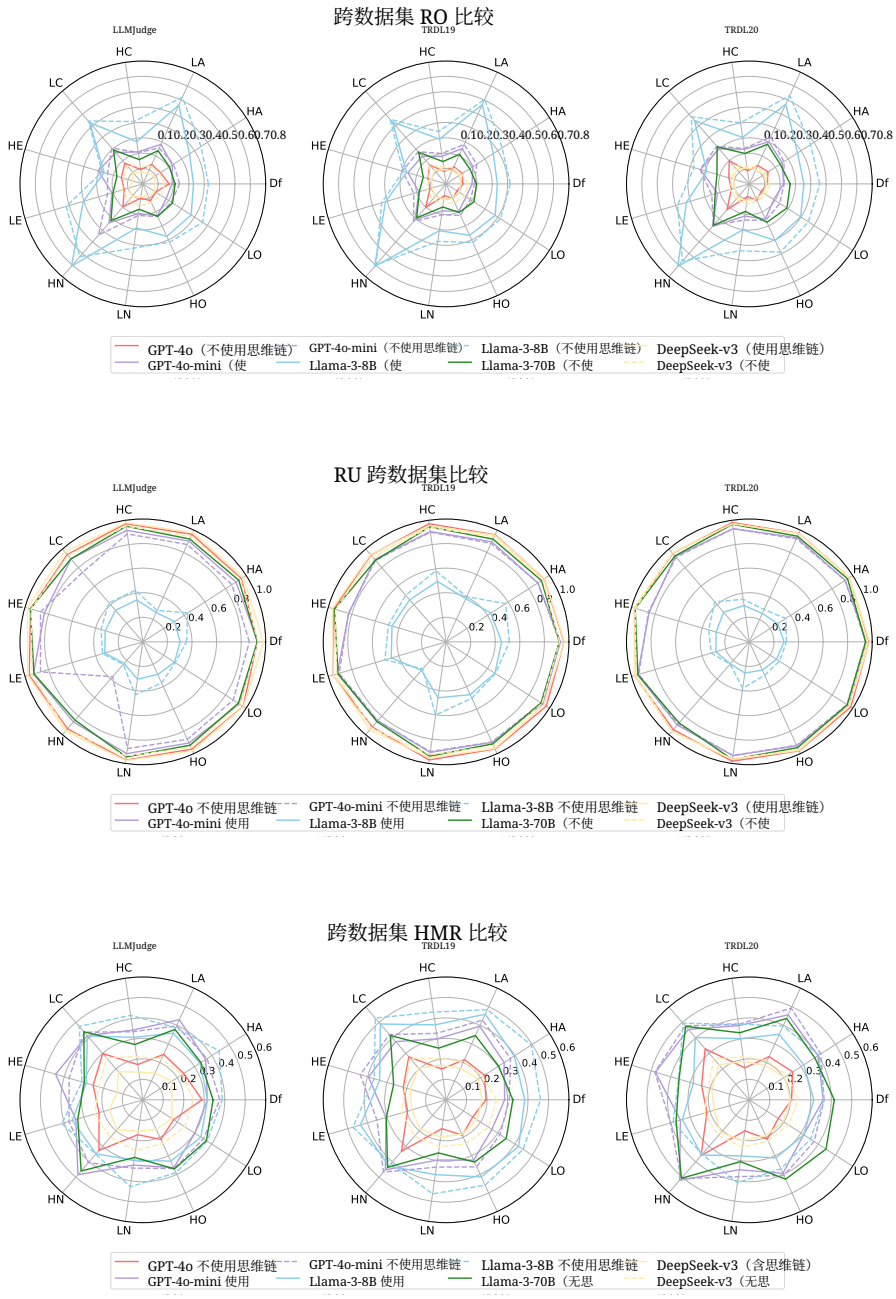


图 5. 不同人格维度和模型配置下 RO、RU 和 HMR 指标的比较。每个雷达图展示了三个人格配置在三个数据集（LLMJudge、TRDL19、TRDL20）上的表现：Df：默认，HA/LA：高/低宜人性，HC/LC：高/低尽责性，HE/LE：高/低外向性，HN/LN：高/低神经质，HO/LO：高/低开放性。RO 衡量等级顺序相关，RU 衡量等级效用，HMR 衡量人机相关性一致性。

人工标注数据，而且这一效果在 Llama 3 8B 上最为显著：跨人格整合利用了任何单一人格条件都无法提供的互补线索，弥补了基础判断较弱的不足。在使用思维链和不使用思维链的设置下，这一模式都成立，不过改进幅度因模型而异。

为评估各人格条件生成的置信度分数的预测效用，我们进行了消融研究。表5中，不含置信度的随机森林表示一种变体：仅使用十一种人格条件下的预测相关性分数作为特征来训练随机森林，不纳入置信度。消融研究证实，置信度在分数之外具有独立的预测价值。从随机森林特征中移除置信度后，各模型及思维链设置下的 κ、二次加权卡帕系数和 F1分数均持续下降。这表明，自报告置信度并非分数的冗余替代指标，而是一种有助于提升监督阶段类别可分性的有效信号。

主要发现。人格条件化预测分数和置信度值是有效的相关性分类特征，前提是与合适的机器学习算法结合使用。汇总所有人格条件的输出，即使人工标注数据有限，表现也优于预言机基线；对于 Llama-3-8B 等较弱模型，这一优势尤其明显。消融结果表明，置信度在相关性分数之外还能提供额外的预测价值。

7 讨论与结论

7.1 讨论与开放问题

在本研究中，我们考察了大语言模型在相关性评估情境下模拟不同人格特质时的行为，尤其关注其决策与人类偏好的契合程度，以及其自报告置信度的可靠性。在实验1中，我们发现了若干跨模型模式。值得注意的是，与默认设置相比，诱导出低宜人性的评估者在 κ、二次加权卡帕系数和F1分数方面，与人类判断的契合度始终更高。然而，这一现象背后的机制仍是一个开放问题。例如，目前尚不清楚低宜人性提示中的特定词元是否促使模型表现出更多批判性推理行为，还是大多数人工标注员的决策风格本就与低宜人性特征相似。实验1的另一项一致观察是，具有低尽责性和高神经质特征的模拟人格能够有效抑制过度自信，同时在过度自信与自信不足之间保持相对平衡的校准。不过，我们目前的设计只要求大语言模型报告对自身决策的置信度。未来的研究可以探讨这两种人格条件能否可靠地评估其他人格（包括默认设置）所作决策的置信度，以及在这一更广泛的情境下，它们能否继续保持平衡的校准。如果确实如此，这将为以更具普适性的方式改善大语言模型的置信度校准，提供一条前景可观且有心理学依据的途径。

7.2 结论

本研究提出了一种新颖且立足于心理学的相关性评估框架，将范式从通用提示转向人格条件认知建模。通过系统地将大五人格特质融入大语言模型，我们探讨了不同认知特征如何影响相关性判断的准确性和置信度校准的可靠性。我们开展了涵盖五种不同大语言模型和三个信息检索测试集合（TREC DL 2019、2020年和LLMJudge）的实验。

综合我们的实证发现，我们发现判断对齐与校准可靠性之间存在明显的权衡。具体而言，与低宜人性相关的认知特征在复现人类专家相关性判断方面表现出更高的效能，而低尽责性和高神经质等特质则是有效缓解系统性过度自信所必需的。我们将这些改进归因于人格条件化所确立的不同决策阈值：低宜人性相关的严格评估标准提高了模型的判别精度，而高神经质相关的保守估计倾向则抵消了模型固有的不当确定性偏差。值得注意的是，这种干预对高能力模型（例如 GPT-4o）尤为关键；尽管这些模型具备先进的推理能力，却表现出更强的过度自信倾向。

这种动态表明，没有任何一种人格条件能在所有评估指标上都普遍最优；相反，要开发高质量的自动评估器，就需要在判别敏锐度与稳健的元认知校准之间取得平衡。此外，我们的研究表明，这种认知多样性并非噪声，而是一种有价值的信号。通过汇总不同模拟人格的预测，我们证明，人格衍生的置信度特征提供了独特且互补的信息，显著提升了相关性分类的表现；即使在低资源场景下，其效果也优于单一角色基线。

总体而言，本研究为新兴的机器心理学领域作出了贡献 [32]，提供了确凿证据，表明抽象的心理学构念可以通过操作化，以可预测的方式调节大语言模型行为。在信息检索领域，我们的研究结果为构建可信评估系统提供了一条切实可行的路径，使其超越单纯的预测准确率，进而发展为能够明确指出自身何时以及为何可能出错的智能体。展望未来，后续研究将拓展这一框架，探究人格之间的互动（例如不同智能体之间的辩论）如何进一步明确自动化决策的边界。

参考文献

[1] Zahra Abbasiantaeb、Chuan Meng、Leif Azzopardi 和 Mohammad Aliannejadi。2024年。我们能否使用大语言模型填补相关性判断空缺？arXiv:2405.05600 [信息检索] https://arxiv.org/abs/2405.05600[2] Marwah Alaofi、Paul Thomas、Falk Scholer 和 Mark Sanderson。2024年。大语言模型可能被诱导将文档标记为相关：我附近最好的咖啡馆；这篇论文完全相关。载于*2024*年亚太地区国际*ACM SIGIR*信息检索研究与发展会议论文集（*SIGIR-AP 2024*）。计算机协会，纽约，纽约州，美国，32–41。 doi:10.1145/3673791.3698431[3] Cameron P. Anderson、Sebastien Brion、Don A. Moore 和 Jessica Kennedy。2014。一种过度自信的地位提升解释。《人格与社会心理学杂志》103，4（2014），718–735。[4] Negar Arabzadeh 和 Charles L.A. Clarke。2025。提示敏感性的人机比较分析：基于大语言模型的相关性判断。载于第*48*届国际*ACM SIGIR*信息检索研究与发展会议论文集（*SIGIR '25*）。计算机协会，纽约，纽约州，美国，2784–2788。[5] Negar Arabzadeh 和 Charles L. A. Clarke。2025。基于大语言模型的相关性判断方法基准测试。载于第*48*届国际*ACM SIGIR*信息检索研究与发展会议论文集（*SIGIR '25*）。3194–3204。[6] Peter Bailey、Nick Craswell、Ian Soboroff、Paul Thomas、Arjen P. de Vries 和 Emine Yilmaz。2008年。相关性评估：评判者能否互换，这是否重要？载于第*31*届年度国际*ACM SIGIR*信息检索研究与发展会议论文集（*SIGIR '08*）。[7] Krisztian Balog、Don Metzler 和 Zhen Qin。2025年。排序器、评判者与助手：理解大语言模型在信息检索评估中的相互作用。载于第*48*届国际*ACM SIGIR*信息检索研究与发展会议论文集（*SIGIR '25*）。3865–3875。[8] Evan Becker 和 Stefano Soatto。2024年。思维循环：通过稳定的解释衡量大语言模型置信度。arXiv:2406.03441 [计算语言学][9] Wanling Cai、Yucheng Jin 和 Li Chen。2022年。个人特征对对话式推荐系统中用户信任的影响。载于*2022*年*CHI*计算机系统中的人为因素会议论文集。1–14。[10] Daniel Cervone。2004年。人格架构。《心理学评论》111，1（2004），183–204。 doi:10.1037/0033-295X.111.1.183[11] Nuo Chen、Hanpei Fang、Jiqun Liu、Wilson Wei、Tetsuya Sakai 和 Xiao-Ming Wu。2025年。通过人格注入缓解大语言模型相关性判断中的阈值启动效应。arXiv:2512.00390 [计算语言学] https://arxiv.org/abs/2512.00390

[12] Nuo Chen、Jiqun Liu、Xiaoyu Dong、Qijiong Liu、Tetsuya Sakai 和 Xiao-Ming Wu。2024。人工智能也会产生认知偏差：一项关于基于大语言模型的批量相关性评估中阈值启动的探索性研究。载于《2024年亚太地区国际 ACM SIGIR 信息检索研究与发展会议论文集》（SIGIR-AP 2024）。54–63。[13] Nuo Chen、Jiqun Liu、Hanpei Fang、Yuankai Luo、Tetsuya Sakai 和 Xiao-Ming Wu。2025。搜索交互中的诱饵效应：理解用户行为并衡量系统脆弱性。ACM 信息系统汇刊 43，2（2025年1月）。[14] Nuo Chen、Jiqun Liu 和 Tetsuya Sakai。2023。用于网络搜索评估的参考依赖模型：理解并衡量有限理性用户的体验。载于2023年 ACM Web 会议论文集（WWW '23）。3396–3405。doi:10.1145/3543507.3583551[15] Nuo Chen、Fan Zhang 和 Tetsuya Sakai。2022。将锚定效应纳入用户模型，以构建更优的评估指标。载于第45届国际 ACM SIGIR 信息检索研究与发展会议论文集（SIGIR '22）。2709–2714。doi:10.1145/3477495.3531953[16] Charles Clarke 和 Laura Dietz。2025。基于大语言模型的相关性评估仍无法取代人工相关性评估。载于 EVIA 2025。[17] C. Robert Cloninger、Dragan M. Svrakic 和 Thomas R. Przybeck。1993年。气质与性格的心理生物学模型。《普通精神病学档案》50，12（1993），975–990。doi:10.1001/archpsyc.1993.01820240059008[18] Philip J. Corr（编）。2008年。《人格强化敏感性理论》。剑桥大学出版社。[19] Paul T. Costa 和 Robert R. McCrae。1999年。人格五因素理论。载于《人格手册：理论与研究》（第2版），Laurence A. Pervin 和 Oliver P. John（编）。吉尔福德出版社，纽约，139–153。[20] Nick Craswell、Bhaskar Mitra、Emine Yilmaz 和 Daniel Campos。2021年。TREC 2020 深度学习赛道概述。arXiv:2102.07662 [信息检索][21] Nick Craswell、Bhaskar Mitra、Emine Yilmaz、Daniel Campos 和 Jimmy Lin。2025年。TREC 2021 深度学习赛道概述。arXiv:2507.08191 [信息检索] https://arxiv.org/abs/2507.08191[22] Nick Craswell、Bhaskar Mitra、Emine Yilmaz、Daniel Campos 和 Ellen M. Voorhees。2020年。TREC 2019 深度学习赛道概述。arXiv:2003.07820 [信息检索][23] Mouly Dewan、Jiqun Liu 和 Chirag Shah。2025年。TRUE：用于信息检索中大语言模型驱动的相关性判断的可复现框架。arXiv预印本 arXiv:2509.25602（2025）。[24] Jessica Maria Echterhoff、Yao Liu、Abeer Alessa、Julian McAuley 和 Zexue He。2024年。大语言模型参与决策时的认知偏差。载于计算语言学协会论文集：EMNLP 2024。计算语言学协会，迈阿密，佛罗里达州，美国，12640–12653。[25] Carsten Eickhoff。2018年。众包中的认知偏差。载于第11届 ACM 网络搜索与数据挖掘国际会议论文集（美国加利福尼亚州玛丽安德尔雷）（WSDM '18）。162–170。[26] Mourad Ellouze 和 Lamia Hadrich Belguith。2024年。人工智能在社交媒体人格特质及障碍分析中的应用：综述。ACM 亚洲低资源语言信息处理汇刊（2024年6月）。已接受。[27] Jonathan St. B. T. Evans 和 Keith Frankish（编）。2009年。《双重心智：双重过程及其延伸》。牛津大学出版社，英国牛津。[28] Guglielmo Faggioli、Laura Dietz、Charles L. A. Clarke、Gianluca Demartini、Matthias Hagen、Claudia Hauff、Noriko Kando、Evangelos Kanoulas、Martin Potthast、Benno Stein 和 Henning Wachsmuth。2023年。大语言模型用于相关性判断的观点。载于2023年 ACM SIGIR 信息检索理论国际会议论文集（ICTIR '23）。[29] Hanpei Fang、Sijie Tao、Nuo Chen、Kai-Xin Chang 和 Tetsuya Sakai。2025年。大语言模型是否偏好近期内容？一项关于基于大语言模型的重排序中近因偏差的研究。arXiv:2509.11353 [信息检索][30] Naghmeh Farzi 和 Laura Dietz。2025年。基于标准的大语言模型相关性判断。载于2025年国际 ACM SIGIR 信息检索创新概念与理论会议论文集（ICTIR）（ICTIR '25）。[31] Jiawei Gu、Xuhui Jiang、Zhichao Shi、Hexiang Tan、Xuehao Zhai、Chengjin Xu、Wei Li、Yinghan Shen、Shengjie Ma、Honghao Liu、Saizhuo Wang、Kun Zhang、Yuanzhuo Wang、Wen Gao、Lionel Ni 和 Jian Guo。2025年。大语言模型评审综述。arXiv:2411.15594 [计算语言学] https: //arxiv.org/abs/2411.15594[32] Thilo Hagendorff、Ishita Dasgupta、Marcel Binz、Stephanie CY Chan、Andrew Lampinen、Jane X Wang、Zeynep Akata 和 Eric Schulz。2023年。机器心理学。arXiv预印本 arXiv:2303.13988（2023）。[33] Haixia Han、Tingyun Li、Shisong Chen、Jie Shi、Chengyu Du、Yanghua Xiao、Jiaqing Liang 和 Xin Lin。2024年。通过学习过往经验增强大语言模型的置信度表达。arXiv:2404.10315 [计算语言学][34] Jiangen He 和 Jiqun Liu。2025年。探究大语言模型人格对自动化决策任务中认知偏差表现的影响。arXiv:2502.14219 [人工智能][35] Yangxizhao He 和 Peng Lei。2025年。人格影响冒险行为的不同路径：外向性和负性情绪如何通过过度自信塑造决策。《心理学前沿》16（2025），1537658。[36] Wenyue Hua、Lizhou Fan、Lingyao Li、Kai Mei、Jianchao Ji、Yingqiang Ge、Libby Hemphill 和 Yongfeng Zhang。2024年。战争与和平（WarAgent）：基于大语言模型的世界大战多智能体模拟。arXiv:2311.17227 [人工智能][37] Yueyue Huang 和 Dengke Yu。2024年。消费者人格、在线社交互动与深度在线消费行为。《科学报告》14，1（2024），29357。[38] Kate E. Jacobs、Dion Szer 和 John Roodenburg。2012年。人格对自我智力评估准确性的调节作用。《人格与个体差异》52，6（2012），744–749。[39] David S. Jalajas 和 Ray Pullaro。2018年。人格对决策的影响。《组织心理学杂志》18，5（2018年12月）。

[40] Kalervo Järvelin 和 Jaana Kekäläinen。2002。基于累积增益的信息检索技术评估。《ACM 信息系统汇刊》20，4（2002年10月），422–446。doi:10.1145/582415.582418[41] Guangyuan Jiang、Manjie Xu、Song-Chun Zhu、Wenjuan Han、Chi Zhang 和 Yixin Zhu。2023。评估并诱导预训练语言模型中的人格。载于 NIPS '23。文章 466，共 22 页。[42] Hang Jiang、Xiajie Zhang、Xubo Cao、Cynthia Breazeal、Deb Roy 和 Jad Kabbara。2024。PersonaLLM：探究大语言模型表达人格特质的能力。载于计算语言学协会会议论文集：NAACL 2024。[43] Zhengbao Jiang、Jun Araki、Haibo Ding 和 Graham Neubig。2021。我们如何知道语言模型何时具备知识？论问答语言模型的校准。计算语言学协会汇刊 9（2021），962–977。[44] Sophia Li、Randall Hale 和 Don A. Moore。2025。过度自信是一种个体差异吗？判断与决策 20（2025），e25。[45] Jiqun Liu。2023。交互式信息检索的行为经济学方法。信息检索系列 48（2023）。[46] Jiqun Liu 和 Fangyuan Han。2020。探究参照依赖效应对用户搜索交互与满意度的影响：行为经济学视角。载于第43届国际 ACM SIGIR 信息检索研究与发展会议论文集（SIGIR '20）。1141–1150。[47] Ollie Liu、Deqing Fu、Dani Yogatama 和 Willie Neiswanger。2024。DeLLMa：运用大语言模型进行不确定性下的决策。arXiv:2402。02392 [人工智能][48] Tie-Yan Liu。2009。面向信息检索的学习排序。《信息检索基础与趋势》3，3（2009），225–331。[49] Xiaoou Liu、Tiejin Chen、Longchao Da、Chacha Chen、Zhen Lin 和 Hua Wei。2025。大语言模型中的不确定性量化与置信度校准：综述。载于第31届 ACM SIGKDD知识发现与数据挖掘会议论文集，第2卷。6107–6117。doi:10.1145/3711896.3736569[50] Jia-Hsun Lo、Han-Pang Huang 和 Jie-Shih Lo。2025。基于大语言模型的机器人个性模拟与认知系统。《科学报告》15，1（2025），16993a。[51] Sean MacAvaney 和 Luca Soldaini。2023。用于自动相关性估计的单次标注。载于第46届国际 ACM SIGIR 信息检索研究与发展会议论文集（SIGIR '23）。[52] Dan P. McAdams。1995。当我们了解一个人时，我们了解了什么？《人格杂志》63，3（1995），365–396。doi:10.1111/j.1467- 6494.1995.tb00506.x[53] Robert R. McCrae 和 Oliver P. John。1992。五因素模型及其应用简介。《人格杂志》60，2（1992年6月），175–215。[54] Walter Mischel 和 Yuichi Shoda。1995。人格的认知—情感系统理论：重新认识人格结构中的情境、倾向、动态与不变性。《心理学评论》102，2（1995），246–268。doi:10.1037/0033-295X.102.2.246[55] Alistair Moffat、Joel Mackenzie、Paul Thomas 和 Leif Azzopardi。2022。离线效果指标的灵活框架。载于第45届国际 ACM SIGIR 信息检索研究与发展会议论文集（西班牙马德里）（SIGIR '22）。计算机协会，纽约，纽约州，美国，578–587。doi:10.1145/3477495.3531924[56] Alistair Moffat、Falk Scholer 和 Ziying Yang。2018。估计信息检索效果指标的测量不确定性。《数据与信息质量杂志》10，3，第10篇文章（2018年9月），共22页。[57] Maria Molchanova、Anna Mikhailova、Anna Korzanova、Lidiia Ostyakova 和 Alexandra Dolidze。2025。探索大语言模型模拟人格的潜力。arXiv:2502。08265 [计算语言学][58] Don A. Moore 和 Paul J. Healy。2008。过度自信的问题。《心理学评论》115，2（2008），502–517。[59] Vincent Murday、Kévin Campos-Moinier、François Osiurak 和 Lionel Brunel。2021。外向性水平可预测个体从社会资源与工具使用中感知到的益处。《科学报告》11，1（2021），12260。[60] Joon Sung Park、Carolyn Q. Zou、Aaron Shaw、Benjamin Mako Hill、Carrie Cai、Meredith Ringel Morris、Robb Willer、Percy Liang 和 Michael S. Bernstein。2024。对1000人的生成式智能体模拟。[61] Catarina Pires、Sérgio Nunes 和 Luís Filipe Teixeira。2025。利用多模态大语言模型扩展基于医学病例的检索任务的相关性判断。arXiv:2506。17782 [信息检索][62] Hossein A. Rahmani、Emine Yilmaz、Nick Craswell、Bhaskar Mitra、Paul Thomas、Charles L. A. Clarke、Mohammad Aliannejadi、Clemencia Siro 和 Guglielmo Faggioli。2024。LLMJudge：用于相关性判断的大语言模型。arXiv:2408。08896 [信息检索][63] Aniruddha S. Rao 和 Savitha G. Lakkol。2024。人格、偏差对印度散户投资者金融风险承受能力的影响。《投资管理与金融创新》21，3（2024），248–264。[64] Rebecca G. Reed、Hannah L. Combs 和 Suzanne C. Segerstrom。2020。老年人自我调节的结构及其与心理和身体健康的相关性。《Collabra：心理学》6，1（2020），23。[65] Aldo Rustichini、Colin G. DeYoung、Jon E. Anderson 和 Stephen V. Burks。2016。将人格理论与决策理论相结合以解释经济行为：一项实验研究。《行为与实验经济学杂志》64（2016），122–137。[66] Tetsuya Sakai。2024。基于过度自信与自信不足评估系统响应。CEUR研讨会论文集 3854（2024）。[67] Tetsuya Sakai 和 Noriko Kando。2008。论针对不完整相关性评估设计的信息检索指标。《信息检索》11，5（2008），447–470。doi:10.1007/s10791-008-9059-7[68] Tetsuya Sakai、Sijie Tao、Nuo Chen、Yujing Li、Maria Maistro、Zhumin Chu 和 Nicola Ferro。2023。论汇集网页、金标准评估与青铜标准评估的排序。《ACM 信息系统汇刊》42，1（2023年8月）。提交给ACM的稿件

[69] Tetsuya Sakai、Sijie Tao、Zhumin Chu、Maria Maistro、Yujing Li、Nuo Chen、Nicola Ferro、Junjie Wang、Ian Soboroff 和 Yiqun Liu。2022。NTCIR-16 We Want Web with CENTRE（WWW-4）任务概述。载于第16届NTCIR信息访问技术评估会议论文集。日本东京。线上研讨会版本。[70] Mark Sanderson。2010。基于测试集的信息检索系统评估。《信息检索基础与趋势》4（2010）。[71] Peter S Schaefer、Cristina C Williams、Adam S Goodie 和 W Keith Campbell。2004。过度自信与大五人格。《人格研究杂志》38，5（2004），473–480。[72] Peter S. Schaefer、Cristina C. Williams、Adam S. Goodie 和 W. Keith Campbell。2004。过度自信与大五人格。《人格研究杂志》38，5（2004），473–480。doi:10.1016/j.jrp.2003.09.010[73] Falk Scholer、Diane Kelly、Wan-Ching Wu、Hanseul S. Lee 和 William Webber。2013。阈值启动与认知需求对相关性校准和评估的影响。载于第36届国际 ACM SIGIR 信息检索研究与发展会议论文集（SIGIR '13）。623–632。[74] Murray Shanahan、Kyle McDonell 和 Laria Reynolds。2023。与大语言模型进行角色扮演。《自然》623，7987（2023），493–498。doi:10.1038/ s41586-023-06647-8[75] Mrinank Sharma、Meg Tong、Tomasz Korbak、David Duvenaud、Amanda Askell、Samuel R. Bowman、Newton Cheng、Esin Durmus、Zac Hatfield-Dodds、Scott R. Johnston、Shauna Kravec、Timothy Maxwell、Sam McCandlish、Kamal Ndousse、Oliver Rausch、Nicholas Schiefer、Da Yan、Miranda Zhang 和 Ethan Perez。2025。迈向对语言模型谄媚性的理解。arXiv:2310。13548 [计算语言学] https://arxiv.org/abs/2310.13548[76] Milad Shokouhi、Ryen White 和 Emine Yilmaz。2015。相关性估计中的锚定与调整。载于第38届国际 ACM SIGIR 信息检索研究与发展会议论文集（智利圣地亚哥）（SIGIR '15）。963–966。[77] Ian Soboroff。2025。不要使用大语言模型进行相关性判断。《信息检索研究杂志》1，1（2025），10。[78] Lydia Soh 和 Kate E. Jacobs。2013。人格对男性和女性认知能力自我估计的偏差效应。《人格与个体差异》55，2（2013），141–146。[79] Aleksandra Sorokovikova、Sharwin Rezagholi、Natalia Fedorova 和 Ivan P. Yamshchikov。2024。大语言模型模拟大五人格特质：进一步证据。载于第1届生成式人工智能系统个性化研讨会论文集（PERSONALIZE 2024）。[80] Sijie Tao、Nuo Chen、Tetsuya Sakai、Zhumin Chu、Hiromi Arai、Ian Soboroff、Nicola Ferro 和 Maria Maistro。2024。NTCIR-17 FairWeb-1任务概述。载于第17届NTCIR（NII信息访问研究测试平台与社区）会议论文集。日本东京。线上，DOI：10.20736/0002001318。[81] Paul Thomas、Seth Spielman、Nick Craswell 和 Bhaskar Mitra。2024。大语言模型能够准确预测搜索者偏好。载于第47届国际 ACM SIGIR 信息检索研究与发展会议论文集（SIGIR '24）。[82] Shivani Upadhyay、Ehsan Kamalloo 和 Jimmy Lin。2024。大语言模型可弥补评估中缺失的相关性判断。arXiv:2405。04727 [信息检索][83] Shivani Upadhyay、Ronak Pradeep、Nandan Thakur、Daniel Campos、Nick Craswell、Ian Soboroff 和 Jimmy Lin。2025。使用 UMBRELA 对大语言模型相关性评估开展大规模研究。载于2025年国际 ACM SIGIR 信息检索创新概念与理论会议论文集（ICTIR）（ICTIR '25）。358–368。[84] Shivani Upadhyay、Ronak Pradeep、Nandan Thakur、Nick Craswell 和 Jimmy Lin。2024。UMBRELA：UMBRELA 是 Bing相关性评估器的（开源复现）版本。arXiv:2406。06519 [信息检索][85] Ellen M. Voorhees。2002。信息检索评估哲学。载于《跨语言信息检索系统评估》。Springer Berlin Heidelberg，德国柏林海德堡。[86] Ellen M. Voorhees、Nick Craswell 和 Jimmy Lin。2022。相关文档过多：克兰菲尔德测试集何去何从？载于第45届国际 ACM SIGIR 信息检索研究与发展会议论文集（西班牙马德里）（SIGIR '22）。计算机协会，纽约，纽约州，美国，2970–2980。[87] Lei Wang、Jingsen Zhang、Hao Yang、Zhi-Yuan Chen、Jiakai Tang、Zeyu Zhang、Xu Chen、Yankai Lin、Hao Sun、Ruihua Song、Xin Zhao、Jun Xu、Zhicheng Dou、Jun Wang 和 Ji-Rong Wen。2025。使用基于大语言模型的智能体模拟用户行为。《ACM 信息系统汇刊》43，2，第55篇文章（2025年1月），共37页。doi:10.1145/3708985[88] Taylor Webb、Keith J. Holyoak 和 Hongjing Lu。2022。大语言模型中的涌现类比推理。arXiv:2212。09196 [人工智能][89] Jason Wei、Yi Tay、Rishi Bommasani、Colin Raffel、Barret Zoph、Sebastian Borgeaud、Dani Yogatama、Maarten Bosma、Denny Zhou、Donald Metzler、Ed H. Chi、Tatsunori Hashimoto、Oriol Vinyals、Percy Liang、Jeff Dean 和 William Fedus。2022。大语言模型的涌现能力。arXiv:2206。07682 [计算语言学][90] Raymond N. Wolfe 和 James W. Grosch。1990。人格与决策信心的相关性。《人格杂志》58，3（1990），515–534。[91] Caiqi Zhang、Fangyu Liu、Marco Basaldella 和 Nigel Collier。2024。LUQ：针对大语言模型的长文本不确定性量化。载于2024年自然语言处理实证方法会议论文集。计算语言学协会，5244–5262。[92] Zijian Zhang、Shuchang Liu、Ziru Liu、Rui Zhong、Qingpeng Cai、Xiangyu Zhao、Chunxu Zhang、Qidong Liu 和 Peng Jiang。2024。面向推荐系统的大语言模型驱动的用户模拟器。arXiv:2412。16984 [信息检索] https://arxiv.org/abs/2412.16984[93] Lianmin Zheng、Wei-Lin Chiang、Ying Sheng、Siyuan Zhuang、Zhanghao Wu、Yonghao Zhuang、Zi Lin、Zhuohan Li、Dacheng Li、Eric P. Xing、Hao Zhang、Joseph E. Gonzalez 和 Ion Stoica。2023。使用 MT-bench 和 Chatbot Arena 评判大语言模型评审。载于第37届神经信息处理系统国际会议论文集（美国路易斯安那州新奥尔良）（NIPS '23）。Curran Associates Inc.，美国纽约州雷德胡克，第2020篇文章，共29页。提交给ACM的稿件

[94] Yi Zhu、Guiqi Hua、Xinning Liu、Chang Wang 和 Mingwei Tang。2025。机器信任：人格特质如何塑造不同人机交互模态下的静态与动态信任。《心理学前沿》16（2025），1539054。
[95] Yutao Zhu、Huaying Yuan、Shuting Wang、Jiongnan Liu、Wenhan Liu、Chenlong Deng、Haonan Chen、Zheng Liu、Zhicheng Dou 和 Ji-Rong Wen。2025。用于信息检索的大语言模型：综述。*ACM Trans. Inf.Syst.* 44，1，文章 12（2025 年 11 月），54 页。doi:10.1145/3748304

研究中使用的置信度报告指令

表 6. 本研究中使用的置信度报告指令。对于有思维链和无思维链版本，我们均使用以下指令来获取置信度分数。

> 你是一名评估文本段落与用户查询相关性的专家。你的任务是根据段落满足查询所表达的信息需求的程度，为该段落分配一个相关性分数。请使用以下量表：*[3]* 完全相关：该段落完全围绕查询展开，并提供清晰、完整的答案。*[2]* 高度相关：该段落提供了一些相关信息，但可能包含无关细节或不够清晰。*[1]* 有关联：该段落与查询略有关联，但没有回答查询。*[0]* 不相关：该段落与查询毫无关联。查询：`{query}`段落：`{document}`给出的相关性分数是 *{pblueicted_score}*，请为此答案给出一个置信度分数，表示你认为此答案正确的把握程度。*0* 表示你完全没有把握，*100* 表示你完全有把握。只返回一个 *0* 到 *100* 之间的数字来表示你对答案的把握程度，不要返回任何其他内容。

表 6 展示了本研究使用的置信度报告指令。对于有思维链和无思维链两个版本，我们都使用以下指令来获取置信度分数。

B 人类一致性与置信度可靠性的完整结果。

表 7 和表 8 展示了实验1的完整结果，其中 G. 表示 GPT-4o，G.m 表示 GPT-4o-mini，L.8b 表示 Llama-3-8B，L.70b 表示 Llama-3-70B，DS 表示 DeepSeek-v3，P. 表示人格，Df. 表示默认设置，即不使用任何人格提示的条件。

C RO、RU 和 HMR 的计算示例

表 9 展示了一个模拟的相关性评估场景，其中评估者评估一组文档 {a、 b、 c、 d、 e} 的相关性等级，并为每项判断提供预测相关性分数及相应的（归一化）置信度值。我们按如下方式计算指标。令 I^+ 表示预测正确的实例集合，I^- 表示预测错误的实例集合。在这个玩具示例中， $I^+ = \{a、\ b、\ d\}$，而 $I^- = \{c、\ e\}$。

过度自信惩罚 O和欠自信惩罚 U定义如下：

$$O = \sum_{i \in I^-} p(i) = 0.85 + 0.30 = 1.15,$$
$$U = \sum_{i \in I^+} (1 - p(i)) = (1 - 0.9) + (1 - 0.6) + (1 - 0.4) = 1.1.$$

表 7. 五个大语言模型在不同数据集上使用思维链和不使用思维链时的人类一致性表现（κ、二次加权卡帕系数、F1分数）。默认条件的单元格以灰色阴影显示，表现优于默认条件时则以颜色突出显示。

模型	P.	LLMJudge						TRDL19						TRDL20					
		使用思维链			不使用思维链			使用思维链			不使用思维链			使用思维链			不使用思维链		
		κ	QWK	F1	κ	QWK	F1	κ	QWK	F1	κ	QWK	F1	κ	QWK	F1	κ	QWK	F1
G.	Df.				0.306	0.562	0.423				0.253	0.463	0.377				0.364	0.534	0.458
	HA				0.266	0.541	0.427				0.218	0.462	0.362				0.334	0.534	0.450
	LA				0.324	0.564	0.428				0.291	0.467	0.407				0.391	0.544	0.467
	HC				0.325	0.581	0.444				0.266	0.475	0.384				0.382	0.539	0.466
	LC				0.297	0.563	0.433				0.260	0.476	0.391				0.349	0.528	0.457
	HE				0.266	0.538	0.419				0.240	0.478	0.371				0.370	0.541	0.467
	LE				0.317	0.572	0.435				0.276	0.475	0.393				0.375	0.553	0.458
	HN				0.318	0.576	0.430				0.270	0.477	0.393				0.377	0.540	0.463
	LN				0.304	0.561	0.434				0.262	0.483	0.384				0.376	0.538	0.469
	HO				0.270	0.546	0.420				0.231	0.461	0.365				0.341	0.525	0.452
	LO				0.331	0.575	0.440				0.267	0.472	0.379				0.352	0.534	0.444
G.m	Df.	0.264	0.524	0.422	0.262	0.524	0.423	0.246	0.467	0.383	0.236	0.447	0.389	0.310	0.517	0.432	0.294	0.489	0.418
	HA	0.260	0.522	0.424	0.238	0.515	0.410	0.249	0.473	0.391	0.213	0.441	0.380	0.304	0.520	0.426	0.277	0.485	0.414
	LA	0.283	0.534	0.433	0.284	0.535	0.436	0.283	0.474	0.412	0.264	0.460	0.411	0.324	0.527	0.432	0.306	0.491	0.427
	HC	0.258	0.520	0.426	0.266	0.528	0.423	0.260	0.471	0.398	0.236	0.452	0.393	0.312	0.527	0.429	0.293	0.484	0.417
	LC	0.249	0.509	0.417	0.250	0.511	0.418	0.235	0.464	0.380	0.214	0.432	0.382	0.292	0.518	0.421	0.268	0.468	0.409
	HE	0.251	0.514	0.420	0.248	0.520	0.414	0.223	0.458	0.375	0.206	0.436	0.374	0.291	0.511	0.421	0.268	0.478	0.408
	LE	0.293	0.544	0.443	0.272	0.531	0.433	0.270	0.479	0.400	0.238	0.452	0.393	0.324	0.525	0.437	0.298	0.491	0.422
	HN	0.255	0.514	0.420	0.269	0.526	0.429	0.230	0.457	0.381	0.247	0.448	0.397	0.299	0.523	0.427	0.304	0.494	0.422
	LN	0.272	0.527	0.429	0.242	0.515	0.407	0.255	0.479	0.391	0.219	0.449	0.378	0.316	0.528	0.432	0.287	0.491	0.415
	HO	0.252	0.511	0.423	0.230	0.511	0.402	0.209	0.445	0.365	0.186	0.424	0.358	0.288	0.517	0.422	0.269	0.482	0.409
	LO	0.283	0.534	0.435	0.268	0.530	0.420	0.253	0.472	0.382	0.225	0.452	0.378	0.323	0.525	0.436	0.290	0.491	0.419
L.8b	Df.	0.163	0.322	0.288	0.088	0.281	0.231	0.111	0.204	0.262	0.095	0.206	0.248	0.134	0.255	0.259	0.064	0.258	0.201
	HA	0.200	0.405	0.343	0.040	0.237	0.175	0.145	0.279	0.296	0.070	0.186	0.204	0.162	0.359	0.309	0.030	0.225	0.157
	LA	0.194	0.403	0.309	0.121	0.293	0.266	0.147	0.281	0.276	0.126	0.267	0.278	0.146	0.367	0.270	0.110	0.312	0.254
	HC	0.186	0.389	0.335	0.060	0.264	0.201	0.119	0.226	0.275	0.094	0.213	0.222	0.129	0.330	0.286	0.062	0.264	0.190
	LC	0.105	0.304	0.234	0.028	0.232	0.163	0.093	0.212	0.216	0.065	0.196	0.191	0.060	0.248	0.186	-0.003	0.196	0.111
	HE	0.198	0.384	0.336	0.153	0.348	0.316	0.105	0.259	0.271	0.109	0.225	0.292	0.134	0.319	0.278	0.070	0.275	0.232
	LE	0.150	0.355	0.267	0.084	0.290	0.231	0.132	0.230	0.253	0.099	0.236	0.238	0.111	0.305	0.223	0.082	0.288	0.218
	HN	0.164	0.378	0.295	0.189	0.390	0.331	0.141	0.275	0.275	0.163	0.307	0.337	0.113	0.338	0.243	0.139	0.373	0.292
	LN	0.181	0.354	0.309	0.044	0.237	0.189	0.121	0.213	0.270	0.064	0.181	0.209	0.137	0.315	0.272	0.027	0.231	0.170
	HO	0.126	0.308	0.244	0.077	0.291	0.249	0.098	0.174	0.240	0.084	0.233	0.259	0.095	0.248	0.231	0.049	0.261	0.212
	LO	0.131	0.299	0.245	0.132	0.340	0.273	0.111	0.188	0.234	0.122	0.236	0.270	0.092	0.248	0.209	0.093	0.295	0.229
L.70b	Df.				0.204	0.432	0.373				0.162	0.396	0.330				0.230	0.439	0.386
	HA				0.177	0.404	0.349				0.131	0.366	0.305				0.187	0.407	0.358
	LA				0.212	0.444	0.383				0.171	0.389	0.337				0.237	0.442	0.387
	HC				0.185	0.414	0.354				0.147	0.382	0.317				0.202	0.422	0.366
	LC				0.148	0.350	0.314				0.125	0.324	0.294				0.159	0.375	0.341
	HE				0.184	0.380	0.332				0.122	0.361	0.293				0.185	0.406	0.348
	LE				0.188	0.426	0.365				0.158	0.389	0.327				0.223	0.436	0.384
	HN				0.238	0.469	0.392				0.213	0.386	0.367				0.252	0.461	0.395
	LN				0.184	0.405	0.353				0.137	0.382	0.309				0.198	0.417	0.361
	HO				0.148	0.366	0.319				0.098	0.349	0.277				0.166	0.402	0.340
	LO				0.213	0.443	0.383				0.154	0.394	0.323				0.228	0.434	0.387
DS	Df.	0.275	0.478	0.381	0.246	0.506	0.387	0.228	0.416	0.353	0.233	0.441	0.374	0.314	0.468	0.398	0.299	0.510	0.420
	HA	0.264	0.488	0.393	0.249	0.503	0.387	0.229	0.443	0.361	0.210	0.431	0.358	0.320	0.496	0.411	0.291	0.505	0.414
	LA	0.308	0.539	0.419	0.262	0.521	0.402	0.275	0.448	0.397	0.246	0.457	0.383	0.362	0.508	0.448	0.320	0.519	0.431
	HC	0.278	0.494	0.390	0.249	0.509	0.394	0.239	0.429	0.364	0.240	0.446	0.379	0.343	0.497	0.422	0.304	0.516	0.429
	LC	0.277	0.500	0.401	0.249	0.511	0.393	0.242	0.447	0.375	0.238	0.448	0.379	0.337	0.512	0.431	0.306	0.515	0.424
	HE	0.279	0.473	0.375	0.260	0.514	0.397	0.221	0.421	0.334	0.243	0.447	0.380	0.333	0.481	0.398	0.314	0.509	0.429
	LE	0.286	0.518	0.408	0.264	0.519	0.400	0.258	0.459	0.386	0.243	0.457	0.377	0.350	0.519	0.442	0.321	0.523	0.434
	HN	0.299	0.527	0.419	0.261	0.526	0.397	0.256	0.452	0.384	0.241	0.447	0.381	0.348	0.512	0.436	0.295	0.503	0.416
	LN	0.285	0.503	0.403	0.252	0.511	0.397	0.245	0.447	0.377	0.232	0.441	0.375	0.347	0.507	0.436	0.309	0.521	0.430
	HO	0.271	0.489	0.389	0.246	0.506	0.395	0.221	0.434	0.353	0.225	0.445	0.367	0.327	0.493	0.416	0.305	0.506	0.423
	LO	0.292	0.525	0.412	0.270	0.528	0.407	0.256	0.448	0.390	0.255	0.462	0.386	0.357	0.509	0.445	0.318	0.524	0.436

抑制过度自信和自信不足所获得的奖励如下：

$$R_O = 1 - \frac{O}{|I^-|} = 1 - \frac{1.15}{2} = 0.425,$$

$$R_U = 1 - \frac{U}{|I^+|} = 1 - \frac{1.1}{3} = 0.633.$$

表8. 五个大语言模型在各数据集上采用思维链和不采用思维链时的总体性能（RO、RU、HMR指标）。默认条件下的单元格以灰色阴影标示，若性能超过默认条件，则以彩色高亮显示。

模型	P.	LLMJudge						TRDL19						TRDL20					
		使用思维链			不使用思维链			使用思维链			不使用思维链			使用思维链			不使用思维链		
		RO	RU	HMR	RO	RU	HMR	RO	RU	HMR	RO	RU	HMR	RO	RU	HMR	RO	RU	HMR
G.	Df.				0.171	0.932	0.289				0.109	0.960	0.195				0.115	0.974	0.206
	HA				0.136	0.953	0.238				0.126	0.950	0.223				0.145	0.967	0.252
	LA				0.139	0.962	0.244				0.122	0.958	0.216				0.132	0.971	0.232
	HC				0.095	0.972	0.174				0.082	0.971	0.152				0.086	0.981	0.158
	LC				0.178	0.940	0.299				0.163	0.936	0.278				0.200	0.953	0.330
	HE				0.146	0.952	0.253				0.123	0.951	0.218				0.139	0.969	0.243
	LE				0.126	0.964	0.223				0.109	0.959	0.196				0.122	0.973	0.216
	HN				0.198	0.936	0.327				0.203	0.918	0.333				0.222	0.945	0.360
	LN				0.094	0.972	0.172				0.077	0.971	0.142				0.083	0.982	0.153
	HO				0.120	0.960	0.213				0.105	0.958	0.190				0.118	0.974	0.211
	LO				0.099	0.971	0.179				0.084	0.969	0.155				0.092	0.981	0.168
G.m	Df.	0.190	0.935	0.315	0.240	0.867	0.376	0.169	0.931	0.286	0.185	0.922	0.308	0.223	0.954	0.361	0.226	0.947	0.365
	HA	0.229	0.901	0.365	0.238	0.862	0.373	0.208	0.887	0.337	0.237	0.889	0.374	0.254	0.927	0.398	0.268	0.925	0.415
	LA	0.283	0.902	0.430	0.257	0.871	0.397	0.253	0.895	0.395	0.283	0.881	0.428	0.302	0.932	0.456	0.336	0.922	0.492
	HC	0.211	0.917	0.344	0.207	0.888	0.336	0.181	0.907	0.301	0.200	0.902	0.328	0.228	0.931	0.366	0.233	0.930	0.372
	LC	0.270	0.893	0.415	0.308	0.749	0.437	0.238	0.887	0.376	0.277	0.878	0.422	0.302	0.921	0.455	0.324	0.911	0.478
	HE	0.302	0.835	0.444	0.239	0.868	0.375	0.259	0.826	0.395	0.297	0.827	0.437	0.332	0.850	0.478	0.334	0.857	0.481
	LE	0.209	0.928	0.341	0.240	0.868	0.376	0.184	0.924	0.306	0.219	0.907	0.352	0.216	0.948	0.351	0.255	0.938	0.401
	HN	0.336	0.863	0.484	0.441	0.375	0.406	0.309	0.850	0.453	0.323	0.849	0.467	0.366	0.896	0.520	0.357	0.886	0.509
	LN	0.195	0.919	0.322	0.208	0.877	0.337	0.177	0.909	0.297	0.204	0.902	0.332	0.212	0.936	0.345	0.238	0.934	0.379
	HO	0.228	0.905	0.364	0.234	0.872	0.369	0.197	0.899	0.324	0.225	0.896	0.360	0.251	0.930	0.395	0.260	0.925	0.406
	LO	0.180	0.936	0.302	0.220	0.875	0.352	0.155	0.936	0.266	0.188	0.920	0.313	0.189	0.960	0.315	0.235	0.945	0.376
L.8b	Df.	0.330	0.303	0.316	0.427	0.355	0.388	0.334	0.452	0.384	0.418	0.515	0.461	0.355	0.282	0.315	0.458	0.302	0.364
	HA	0.361	0.307	0.332	0.450	0.436	0.443	0.365	0.433	0.396	0.434	0.576	0.495	0.373	0.275	0.316	0.474	0.352	0.404
	LA	0.564	0.262	0.358	0.620	0.281	0.387	0.569	0.386	0.460	0.606	0.404	0.484	0.571	0.254	0.352	0.632	0.289	0.396
	HC	0.278	0.345	0.308	0.410	0.424	0.417	0.295	0.497	0.370	0.344	0.585	0.433	0.306	0.299	0.302	0.402	0.350	0.374
	LC	0.528	0.340	0.414	0.542	0.421	0.474	0.533	0.453	0.490	0.556	0.502	0.528	0.546	0.320	0.403	0.578	0.431	0.494
	HE	0.263	0.319	0.288	0.276	0.355	0.310	0.268	0.459	0.338	0.292	0.489	0.366	0.289	0.282	0.285	0.376	0.345	0.360
	LE	0.404	0.323	0.359	0.520	0.341	0.412	0.411	0.465	0.436	0.437	0.519	0.474	0.427	0.284	0.341	0.487	0.323	0.388
	HN	0.710	0.232	0.349	0.626	0.243	0.351	0.711	0.306	0.428	0.697	0.291	0.411	0.711	0.237	0.356	0.665	0.239	0.352
	LN	0.290	0.307	0.298	0.423	0.435	0.429	0.308	0.460	0.369	0.378	0.602	0.464	0.302	0.256	0.277	0.440	0.382	0.409
	HO	0.376	0.292	0.329	0.417	0.368	0.391	0.368	0.472	0.414	0.419	0.504	0.458	0.405	0.253	0.312	0.488	0.299	0.371
	LO	0.374	0.307	0.337	0.463	0.313	0.373	0.380	0.472	0.421	0.421	0.477	0.447	0.397	0.263	0.316	0.457	0.291	0.355
L.70b	Df.				0.210	0.932	0.343				0.198	0.922	0.326				0.266	0.946	0.415
	HA				0.206	0.928	0.337				0.183	0.923	0.305				0.238	0.950	0.380
	LA				0.238	0.920	0.378				0.212	0.919	0.345				0.284	0.945	0.437
	HC				0.160	0.951	0.273				0.147	0.946	0.255				0.200	0.967	0.331
	LC				0.292	0.894	0.440				0.273	0.878	0.416				0.320	0.925	0.475
	HE				0.176	0.954	0.298				0.155	0.946	0.267				0.220	0.968	0.358
	LE				0.202	0.924	0.331				0.182	0.917	0.304				0.232	0.946	0.373
	HN				0.317	0.843	0.460				0.293	0.859	0.437				0.357	0.873	0.507
	LN				0.167	0.946	0.284				0.152	0.940	0.262				0.181	0.968	0.305
	HO				0.232	0.925	0.371				0.204	0.913	0.334				0.275	0.946	0.426
	LO				0.230	0.924	0.369				0.216	0.919	0.349				0.291	0.947	0.445
DS	Df.	0.077	0.982	0.142	0.142	0.939	0.247	0.096	0.963	0.175	0.142	0.923	0.246	0.112	0.978	0.201	0.136	0.960	0.238
	HA	0.093	0.972	0.169	0.130	0.945	0.229	0.104	0.952	0.188	0.119	0.935	0.211	0.114	0.968	0.204	0.125	0.965	0.221
	LA	0.077	0.977	0.143	0.125	0.943	0.220	0.093	0.954	0.169	0.116	0.939	0.206	0.101	0.970	0.184	0.118	0.968	0.210
	HC	0.074	0.980	0.137	0.120	0.952	0.213	0.094	0.957	0.171	0.114	0.942	0.204	0.105	0.971	0.190	0.119	0.971	0.212
	LC	0.100	0.970	0.181	0.158	0.931	0.271	0.116	0.940	0.206	0.159	0.913	0.270	0.128	0.956	0.226	0.154	0.955	0.266
	HE	0.070	0.980	0.131	0.127	0.948	0.224	0.100	0.965	0.182	0.118	0.938	0.209	0.106	0.980	0.192	0.117	0.968	0.209
	LE	0.094	0.973	0.171	0.140	0.940	0.244	0.096	0.957	0.175	0.128	0.929	0.225	0.098	0.974	0.178	0.130	0.962	0.230
	HN	0.108	0.961	0.194	0.139	0.952	0.243	0.106	0.947	0.191	0.135	0.928	0.235	0.120	0.964	0.214	0.131	0.963	0.230
	LN	0.084	0.976	0.155	0.145	0.938	0.251	0.101	0.952	0.183	0.132	0.928	0.231	0.111	0.966	0.200	0.134	0.963	0.235
	HO	0.086	0.974	0.159	0.134	0.945	0.234	0.102	0.954	0.185	0.135	0.932	0.236	0.107	0.971	0.193	0.124	0.965	0.220
	LO	0.093	0.973	0.170	0.140	0.940	0.243	0.099	0.949	0.179	0.142	0.928	0.246	0.109	0.967	0.196	0.129	0.961	0.227

最后，奖励调和平均值（HMR）为：

$$\mathrm{HMR} = \frac{2R_O R_U}{R_O + R_U} = \frac{2 \times 0.425 \times 0.633}{0.425 + 0.633} \approx 0.508.$$

尽管模型正确标注了五个实例中的三个，但其置信度校准失误：对于错误的预测，模型过度自信（例如，ID 3 的置信度为 = 0.85）；而对于正确的预测，模型则信心不足（例如，ID 4 的置信度为 = 0.40）。由此得到的 HMR 对这两类错误进行了均衡惩罚，相比单独使用准确率，能提供更有信息量的可靠性度量。

表 9. 相关性评估中计算 RO、RU 和 HMR 的玩具示例

文档标识符	真实标签	预测值	置信度	是否正确
a	3	3	0.90	✓
b	2	2	0.60	✓
c	1	2	0.85	×
d	0	0	0.40	✓
e	2	0	0.30	×

收到日期：2007年2月20日；修订日期：2009年3月12日；接受日期：2009年6月5日