信息缺失、作者未回应、实验缺陷：无法评估自然语言处理中既有人工评估的可复现性

Anya Belz[a,b](anya.belz@adaptcentre.ie), Craig Thomson[b], Ehud Reiter[b], Gavin Abercrombie[8], Jose M. Alonso-Moral[17],Mohammad Arvan[16],Anouck Braggaar[13],Mark Cieliebak[20], Elizabeth Clark[6], Kees van Deemter[19],Tanvi Dinkar[8],Ondˇrej Dušek[9],Steffen Eger[1], Qixiang Fang[19], Mingqi Gao[11],Albert Gatt[19],Dimitra Gkatzia[4],Javier González-Corbelle[17], Dirk Hovy[2], Manuela Hürlimann[20],Takumi Ito[10],John D. Kelleher[12],Filip Klubiˇcka[12], Emiel Krahmer[13], Huiyuan Lai[7],Chris van der Lee[13],Yiru Li[7],Saad Mahamood[14], Margot Mieskes[15], Emiel van Miltenburg[13],Pablo Mosteiro[19],Malvina Nissim[7], Natalie Parde[16], Ondˇrej Plátek[9],Verena Rieser[8], Jie Ruan[11],Joel Tetreault[3],Antonio Toral[7], Xiaojun Wan[11],Leo Wanner[18],Lewis Watson[4],Diyi Yang[5]

[a]ADAPT/DCU，爱尔兰； [b]阿伯丁大学，英国； [1]比勒费尔德大学，德国； [2]博科尼大学，意大利； [3]Dataminr，美国； [4]爱丁堡龙比亚大学，英国； [5]佐治亚理工学院，美国； [6]Google Research，美国； [7]格罗宁根大学，荷兰； [8]赫瑞瓦特大学，英国； [9]布拉格查理大学，捷克； [10]东北大学，日本； [11]北京大学，中国； [12]都柏林理工大学，爱尔兰； [13]蒂尔堡大学，荷兰； [14]trivago，德国； [15]达姆施塔特应用科学大学，德国； [16]伊利诺伊大学芝加哥分校，美国； [17]圣地亚哥-德孔波斯特拉大学，西班牙； [18]庞培法布拉大学，西班牙； [19]乌得勒支大学，荷兰； [20]苏黎世应用科学大学，瑞士

arXiv:2305.01633v2 [cs.CL] 7 Aug 2023

摘要

我们报告了为自然语言处理领域的一项协同研究所做的努力：我们旨在找出一组适合纳入研究的既有人工评估，以探究哪些因素会使自然语言处理领域的人工评估更易或更难复现。我们介绍了研究结果与发现，其中包括：仅有 13% 的论文（i）复现门槛足够低，且（ii）能够获取的信息足够充分，因此可纳入复现研究；我们选定进行复现的实验中，除一个以外，其他都被发现存在缺陷，以致开展复现是否有意义也成了疑问。因此，我们不得不将协同研究的设计从直接复现改为“先标准化，再复现两次”的方法。我们的总体（负面）发现是：绝大多数自然语言处理领域的人工评估无法重复和／或无法复现，和／或存在过多缺陷，不足以支持复现。这一发现描绘出严峻的图景，但也为重新思考如何设计和报告自然语言处理领域的人工评估提供了契机。

1 引言

自然语言处理（自然语言处理）领域日益认识到，结果的可复现性，尤其是系统评估结果的可复现性，至关重要，而目前该领域对结果可复现性的评估还不够严谨，也缺乏统一的方法。近期研究，尤其是在自动评估方面，已取得进展（Pineau，2020年；Whitaker，2017年），但人们普遍认为，人工评估是衡量自然语言处理质量的试金石，而其可复现性却鲜少受到关注。可以说，如果人们不了解人工评估的可复现性，就无法判断其可靠性；如果不了解其可靠性，也就无法判断以人工评估为基准进行元评估的自动评估是否可靠。

本文所报告的工作是 ReproHum 项目的一部分[1]，我们的目标是在现有工作基础上，以数据表的形式记录人工评估的属性（Shimorina 和 Belz，2022年），并评估复现研究的结果与原始研究有多接近（Belz 等人，2022年），从而系统地探究哪些因素会使人工评估更容易或更难复现。本文介绍了我们迄今为止在该项目中的研究成果，这些成果促使我们重新思考设计此类调查的整体方法。

第 2节概述了我们开展这项工作的动机

[1]https://gow.epsrc.ukri.org/NGBOViewGrant.aspx?GrantRef=EP/V05645X/1

出一项关于影响自然语言处理领域可复现性因素的多实验室多测试（MLMT）研究，以及我们最初为该研究设计的方案。第3 节介绍了我们的论文筛选、标注和筛选流程，最终得到的可供复现的候选论文数量少得令人意外。在第4 节中，我们介绍了在与合作实验室共同开展复现工作时，所遇到的原始评估研究中的诸多其他问题。第6 节总结了我们的负面发现，即无法按照现有形式评估此前在自然语言处理领域开展的人工评估的可复现性，并概述了这些发现所迫使我们对多实验室多测试研究做出的调整。

2 研究动机与整体研究设计

单项研究可以告诉我们，复现研究的结果与原始研究的结果有多接近。大量此类研究能够揭示哪些类型的评估具有较好的可复现性这一总体趋势。然而，目前自然语言处理领域尚无大量复现研究，而且由于成本高、吸引力不足，这种情况不太可能改变。此外，汇总单项研究并不能提供适当条件，以衡量特定因素对可复现性的效应量和显著性，以及这些因素之间的交互作用。

为了创造这样的条件，需要开展一项受控研究，其中纳入数量相等的复现，并分别考察是否存在关注因素。此外，我们从现有研究（Belz 等人，2022年；Huidrom 等人，2022年）中得知，对同一项原始研究的不同复现可能产生截然不同的结果。最后，尽管在相同条件下检验可复现性很有启发意义，检验良好的可复现性能够经受多大程度的变化也同样值得关注——例如，将七点质量量表替换为五点量表是否会影响可复现性。

因此，一项旨在研究提高或降低复现性的因素的研究需要：（i）对每项原始研究进行不止一次复现；（ii）由不同团队组成的多样化团队组合来实施；（iii）纳入多轮实验，并逐步降低各轮实验条件的相似程度。开展此类研究的具体步骤如下：

1. 识别候选评估实验，从中筛选出因素平衡的实验纳入 MLMT 研究；2. 记录评估实验的属性，以便选择因素并

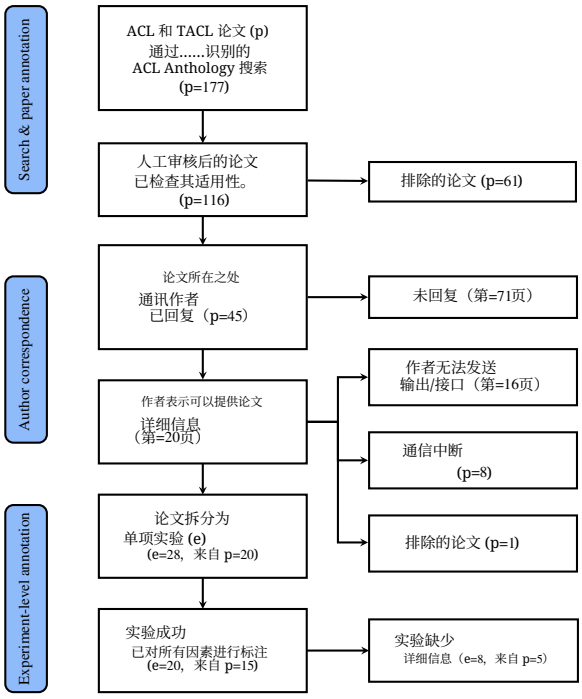


图1：论文筛选流程图，展示了随着信息收集的深入，符合要求的论文数量逐渐减少。

对其进行控制；

3. 选择需要控制的因素及相应的实验子集；以及 4. 对选定的评估研究和因素进行复现。

我们在第3.1 节和3.2节介绍步骤1和2，在第3.3节介绍步骤3，并在第4节介绍步骤4，直至我们中止原始研究设计。

3 候选评估实验的筛选与评估

图1以流程图的形式展示了筛选和标注过程，其中显示了剩余论文/实验数量逐步减少的情况。第一步是在 ACL Anthology 中检索 2018—2022 年间发表于 ACL（主会）或 TACL 的论文[2]，检索条件包含短语“人工评估”和“参与者”；我们共找到 177 篇此类论文。

3.1 论文高层级标注

在第一轮为论文标注人工评估相关属性时，我们使用了以下论文级属性，且仅依据论文或补充材料中的信息进行标注：

[2]检索于 2022年 7 月进行，因此未收录该年稍后发表的一些 TACL 论文。

1. 评估了多少个系统；2. 使用了多少个数据集；3. 参与者类型（例如 MTurk）；4. 参与者总数；5. 判断次数的大致估计；6. 所评估系统实现的自然语言处理任务类型（例如摘要生成）；7. 所使用的输入/输出语言（例如英语）。

在首次标注过程中，我们人工筛除了仅讨论人工评估、而未实际开展人工评估的论文（例如，关于人工评估的综述）、纵向研究、使用高度专业化参与者（如医生）的研究，以及根据粗略估算、重复成本过高的研究（评估人员报酬的门槛为 $2,000）。筛选后剩下 116 篇论文。对于这些论文，附录中的表 3列出了所标注的各项属性中最常见取值的数量[3]。英语是占主导地位的系统语言，出现在超过 90% 的论文中。第二常见的语言是中文，在略低于 10% 的实验中使用。语言生成任务最为常见，其中摘要生成是最常见的任务，其次是对话和机器翻译（MT）。

约三分之一的论文未说明参与者类型。在说明了参与者类型的论文中，60% 使用众包，其中绝大多数通过 Mechanical Turk 开展。参与者相关信息通常难以查找，约有一半论文未报告参与者总人数。很少有论文清楚说明系统、数据集、项目与参与者之间的关系，因此判断数量只能估算。

在高层级标注过程中，我们发现，筛选后剩余的 116 篇论文中，仅有不到 5% 能够仅凭公开信息加以复现。评估者的数量和类型、指导与培训，以及所评估的数据等基本详细信息往往都未予说明。因此，我们下一步便联系作者，希望获得缺失的信息。近期已有多次评论指出人工评估相关信息不足（van der Lee 等人，2019年；Howcroft 等人，2020年；Belz 等人，2020年）。

[3]由于一些论文包含多个属性，例如机器翻译系统中的多种语言，因此某些行的总和不会达到 116。

培训或专业知识	两者皆非	仅一个	both
	11	13	4
参与者人数	小		不小
	14		14
复杂度	low	中	high
	9	11	8

表1：控制因素标注的频率。

3.2 对实验进行更详细的标注

在下一阶段，我们对评估属性进行了详细标注，以便为多实验室多测试研究挑选一部分属性加以控制。我们给 116 篇论文各自的通讯作者（若未指定，则联系第一作者）发送电子邮件，询问他们是否愿意支持复现研究，以及能否提供有关其实验的更多详细信息。

所请求的信息包括评估过程中使用的用户界面，以及向评估者展示的输出集合（完整清单见附录A.2）。即使发送了提醒，我们也只收到了 39% 论文作者的回复。许多回复者无法提供所需信息。最终，只有 20 位作者（涉及 20 篇论文中的 28 个实验）向我们提供了足够的信息，使论文得以进入详细标注阶段。[4] 作者回复但无法提供信息，最常见的原因是他们已离开原先的职位（通常是研究生职位），而相关文件未予保留。在某些情况下，来自商业研究团队的作者因业务原因无法提供信息。此外，还有 8 篇论文的作者最初作出了回复，但之后通信便陷入停滞。

我们结合作者提供的信息、论文、补充材料和在线资源，对进入此阶段的20篇论文进行标注，记录第A.4节所示的评估的详细属性，并以单个实验为单位进行标注（28项），因为在这一更细粒度的标注层面，同一篇论文中的不同实验可能具有不同属性。

本文前三位作者之一对这28项实验的详细属性进行了标注；另外两位作者则各自核查了一半的标注。对于任何差异，作者们都进行了讨论并

[4]另有一位作者确实提供了足够的信息，但在进一步分析论文及其发送的资源后，我们认为其中报告的评估实验与其他20篇论文中的实验差异太大；这些系统检测的是语言使用随时间的变化。

予以解决。为完成这些标注，除两项实验外，我们都不得不向作者追加提问（通常需要经过多轮问答）。最终，对于28项实验中的8项，我们未能获得上述属性所需的全部信息。

请注意，A 节A.4中的最后两个属性（评估任务复杂度、界面复杂度）与其他属性的性质不同：它们是次要属性，在标注过程中通过主观评估得出，而非源自作者提供的信息。我们发现，对于这两个属性的取值，我们通常能够达成一致；即使存在分歧，取值也相邻。出现分歧时，我们通过讨论而非试图将规则形式化来解决，因为要穷尽地涵盖所有分歧似乎是不可能的。

表 1以及附录中的表 4展示了 28 个实验中最常见属性取值的频率（此处包括不明确的取值）。我们发现，大多数已标注属性都有一两个取值的出现频率远高于其他取值。例如，在 28 个实验中，评估结果有 26 个是内在的，26 个是主观的，20 个是绝对的。只有两个实验采用了外在且客观的评估，其余 26 个采用了内在且主观的评估。参与者人数差异很大，最少为 2 人，最多为 233 人。没有实验为参与者提供明确的培训环节，只有一个实验包含练习环节。约四分之三的实验提供了说明和/或评判标准定义。[5] 约有一半实验使用了具备专业知识的受试者，其专业领域通常是语言学或自然语言处理。

3.3 选择需要控制的属性

前几节讨论的问题给受控研究的论文筛选带来了严重困难：我们只有 20 个经过完整标注的实验；而且，我们标注的许多属性分布极不均衡，许多属性组合根本没有出现，或只出现过一两次。鉴于上述问题，我们显然只能选取少数属性进行控制。因此，我们将已标注的属性缩减为三项，这些属性既切实可行又根据现有研究，有较大可能影响可复现性。对于这些属性，我们根据原始取值范围划分为两到三个区间，具体如下：

1. 评估者数量（小、不小）：评估者人数为 1–5 人的实验被赋予小 这一取值，评估者人数超过 5 人的实验则被赋予不小 这一取值。

2.评估者所执行评估的认知复杂度（低、中、高）：根据第 A.4 节所列的任务复杂度和界面复杂度属性，为实验赋予三个可能取值中的一个。

3. 评估者的培训和/或专业知识（两者皆有、任一、两者皆非）：对于评估者既接受过培训又需要具备特定专业知识的实验，赋值为两者皆有；对于评估者接受过培训或需要具备专业知识（但非两者皆有）的实验，赋值为任一；其余实验则赋值为两者皆非。

即使控制因素集大幅缩减，我们仍没有足够的实验覆盖所有2×3×3取值组合，因此最终选定了 6 项实验，使评估者数量与培训/专业知识属性的两两组合数量相等，同时评估者数量与复杂度属性的两两组合数量也相等。

4 设置复现

开始复现最终选定的六个实验（有关复现的通用约定方法，见附录 A.5）必然需要深入了解每个实验的完整实施细节。在此过程中，我们发现的实验缺陷、错误和漏洞之多，尤其令人担忧。我们越深入研究重复评估实验所需的评估实验属性，就越多地发现一些缺陷；在某些情况下，任何基于这些存在缺陷的实验得出的结论都不可靠，这使我们开始质疑是否还有必要重复该实验。第 A.6 节列出了六个具体问题。[6] 注

[5]由于在某些情况下，即使我们与作者沟通过，仍未能获得这些信息，因此我们无法做到精确。

[6]请注意，考虑到相关的声誉风险，我们以匿名形式报告这些内容。另请参阅附录中所附的《负责任研究检查清单》。

Task	评估者数量		认知复杂度			培训和/或专业知识		
	小	不小	low	中	high	两者皆非	任一	both
对话	1	0	0	1	0	0	1	0
生成	6	5	4	5	2	4	5	2
摘要生成	3	1	2	1	1	1	3	0
其他	2	2	1	0	3	2	0	2

表 2：在所有属性都明确的 20 个实验（来自 15 篇论文）中，各自然语言处理任务的控制属性值计数。

需要注意的是，我们选定的六个实验中只有一个不存在上述任何问题。我们仍在发现更多问题。

我们为原始研究设计的结构见附录 A.1 节，图2。

5 讨论

我们决定放弃原有研究设计的原因如下。第一，我们很难找到足够多的论文，这些论文不存在（i）妨碍复现的重大障碍，和/或（ii）无法获取、但重复实验所必需的信息，和/或（iii）实验缺陷与错误。第二，无论我们多么努力地向作者索取完整的实验细节，仍有一些问题没有答案，尽管问题越来越细致，例如评估项目的呈现顺序是否经过随机化，或参与者接受了哪些说明/培训。在某些情况下，我们偶然得知此前提供的信息无法推断出的、关于其他已实施事项的信息，因此有必要进一步调整实验设计。

解决末尾论文数量不足的一个潜在办法，是在开始时选择更多论文（涵盖更多年份、更多会议）。然而，鉴于我们为从作者处获取足够信息投入了极其大量的工作，简单地将最初的论文池扩大三倍或四倍并不可行。同样，对于过高的成本和高度专业化的评估者所造成的复现障碍，我们也几乎无能为力。

另一方面，接受基于不完整实验信息开展工作也会带来问题，因为不同论文的信息以不同方式残缺，而我们将无法进行同类比较。

同样，修正缺陷和错误也会在原始研究与复现研究之间引入差异，而且不同案例中的差异各不相同。严格来说，在这种情况下，我们将不再是在进行复现。

我们曾考虑从头设计新的评估，以满足 MLMT 研究所需的特性。然而，要确保新设计的研究在某种程度上能够代表自然语言处理中实际开展的研究类型，将会非常困难。

现在，我们选择了一种融合上述大多数方案要素的方法：按照类似之前的流程，选取一组规模稍大的现有研究，减少我们所控制的不同因素取值数量，然后在复现前对研究进行标准化，并在必要时加以修正。这样一来，可复现性衡量的是两项新研究之间的情况，而不是它们与原始研究之间的情况。

6 结论

自然语言处理领域在记录人工评估实验信息方面的历史表现目前十分糟糕（Howcroft 等人，2020年）。从论文层面的标注（附录表 3）中可以看到，在 116 篇论文中，有 37 篇未明确说明参与者的类型，59 篇未明确说明参与者人数，15 篇未明确说明判断的数量。即使在实验层面的详细标注阶段与作者进行了长时间交流，有些情况下仍无法获得一些非常基本的详细信息：参与者人数、培训的详细信息、指导内容和练习项目、是否要求参与者为母语者，甚至是接受评估的输出集合。

我们的总体结论是，基于在我们抽取的177篇论文中实验信息无法获取、复现障碍和/或实验缺陷等问题，先前自然语言处理领域的人工评估中，只有很小一部分能够在相同条件下重复进行，因此无法通过重复实验来检验其可复现性。未来的解决途径似乎是接受详细记录实验细节所带来的额外工作量，例如采用 HEDS（Shimorina 和 Belz，2022年），同时大幅提高实验设计各个方面的标准化程度。

致谢

ReproHum 项目由 EPSRC 资助，资助编号为EP/V05645X/1。我们谨向所有抽出时间回应信息请求的作者表示感谢，也感谢 Jackie Cheung。

局限性

我们少数基于从作者处获得的信息得出的发现必然存在局限，因为这些发现无法反映那些未回复的作者可能提供的信息。

此外，我们通过搜索短语“人工评估”和“参与者”来筛选初始论文集。虽然这个短语很常用于指代非自动化评估，但我们仍有可能遗漏使用其他术语的论文。

基于我们的实验样本得出的少数结论也受到样本量的限制，因此未必能代表更广泛的自然语言处理领域当前的人工评估情况。

伦理声明

作为一篇对其他学术出版物进行元评审的论文，本文可被视为低风险。除汇集出版物中的信息外，我们还对论文进行了标注，分析了结果，并根据标注得出了描述性统计数据。第5节总结了我们在为复现研究做准备的实验中发现的缺陷、漏洞和错误。我们决定不引用发现这些问题的论文，因为重要的是此类问题确实存在，而不是应由哪些研究人员负责。

另请参阅本文填写的负责任自然语言处理研究清单（附录A.7）。

参考文献

Anya Belz、Simon Mille 和 David M. Howcroft。2020年。厘清人工评估方法的属性：支持可比性、元评估和可复现性测试的分类体系。载于第*13*届国际自然语言生成会议论文集，第183–194页，爱尔兰都柏林。计算语言学协会。

Anya Belz、Maja Popovic 和 Simon Mille。2022年。自然语言处理结果的量化可复现性评估。载于第*60*届年度会议论文集计算语言学协会（第*1*卷：长篇论文），第16–28页，都柏林，爱尔兰。计算语言学协会。

David M. Howcroft、Anya Belz、Miruna-Adriana Clinciu、Dimitra Gkatzia、Sadid A. Hasan、Saad Mahamood、Simon Mille、Emiel van Miltenburg、Sashank Santhanam 和 Verena Rieser。2020年。人工评估中二十年的困惑：自然语言生成需要评估表和标准化定义。载于第*13*届国际自然语言生成会议论文集，第169–182页，爱尔兰都柏林。计算语言学协会。

Rudali Huidrom、Ondˇrej Dušek、Zdenˇek Kasner、Thiago Castro Ferreira 和 Anya Belz。2022年。对语义错误检测系统人工评估比较评估的两次复现。载于第*15*届国际自然语言生成会议论文集：生成挑战，第52–61页，美国缅因州沃特维尔及线上会议。计算语言学协会。

Joelle Pineau。2020年。机器学习可复现性清单 v2.0。

Anastasia Shimorina 和 Anya Belz。2022年。人工评估数据表：用于记录自然语言处理领域人工评估实验详细信息的模板。载于 第二届自然语言处理系统人工评估研讨会（*HumEval*）论文集，第 54–75 页，爱尔兰都柏林。计算语言学协会。

Chris van der Lee、Albert Gatt、Emiel van Miltenburg、Sander Wubben 和 Emiel Krahmer。2019年。自动生成文本人工评估的最佳实践。载于 第十二届自然语言生成国际会议论文集，第 355–368 页，日本东京。计算语言学协会。

KirstieWhitaker。2017年。《机器翻译可复现性检查清单》。https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist。pdf。

A 附录

A.1 原始研究设计

图2 展示了多实验室多测试研究的原始设计。

A.2 向作者索取的初始信息

我们在最初发给作者的电子邮件中询问，他们是否能够提供以下信息：

1. 向参与者展示的系统输出。2. 参与者填写的界面、表单或文档；如果能提供实际使用的确切文档或表单则更理想。3. 参与研究的参与者人数及类型（学生、研究人员、Mechanical Turk 等）的详细信息。4. 原始研究的总成本。

A.3 高层级标注计数

表3 展示了第一轮论文级属性标注的计数。

A.4 实验级标注的详细信息

下方列出了详细标注中的所有属性名称和值，以及每项属性所记录内容的描述：

1. 使用的具体数据集；2. 使用的具体评估标准名称；如果可能，采用论文中所述的标准名称，否则采用能代表评估内容的标准名称。3. 系统语言；系统用作输入或输出的语言。4. 系统任务；系统所要处理的自然语言处理任务。28 项实验涉及的任务包括跨语言摘要生成、数据到文本生成、可控复杂度定义生成、对话摘要生成、对话轮次生成、解释生成、事实核查理由生成、机器翻译错误预测、提示式生成、问题生成、问答生成、指代表达生成、简化、摘要生成、文本转语音。

5. 评估者类型；评估者的类型包括同事、商业机构内部评估者、众包评估者、作者与同事的组合、同事与学生的组合、专业评估者、学生。

6. 评估模式（Belz 等人，2020年）：

（a）内在与外在；（b）绝对与相对；（c）客观与主观。

7. 参与者数量；参与研究的唯一参与者总数。8. 评估的项目数量；对于绝对评估，这是一个系统输出；对于相对评估，则指正在比较的一组输出，例如一对输出。9. 评估每个项目的参与者人数；在某些实验中，这一人数有所变化。10. 每位参与者评估的项目数量；在某些实验中，这一数量有所变化。具体而言，在28项实验中，有13项通过人群众包进行，其中5项的数量为已知整数，4项有所变化，另有4项无法确定（我们怀疑这4项的数量也有所变化）。11. 是否为参与者提供了培训和/或练习环节；见下文讨论。12. 是否向参与者提供了说明？是否向他们说明了评估标准的定义；见下文讨论。13. 是否要求参与者具备特定的专业知识？如果是，具体是什么类型？这一点是由参与者自我报告还是由外部评估；见下文讨论。14. 是否要求参与者是母语者？如果是，这一点是由参与者自我报告还是由外部评估？对于第一部分，我们采用了“是”“否”“按众包地区筛选”这几种选项；另有一项实验是在一所大学由该语言为母语的学生参与完成。后两种情况本质上都是自我报告，不过研究人员进行了一定程度的控制。只有一项母语者实验的研究人员表示，他们已确认参与者的母语身份；其余实验均依靠自我报告。

用于研究影响可复现性的实验因素的多实验室、多测试对照研究的结构设计：

第***1***轮： 在重复性测量条件下测试精密度。

- 每项实验的重复次数：由两个不同的实验室各进行2次复现；
- 要改变的条件（实验因素）：评估者群体；
- 如果复现足够接近，则进入第2轮；否则，在改进实验设计后重复第1轮，改进措施包括增加评估者数量并降低评估任务的认知复杂度；
- 对于第1轮的重复实验，如果复现研究之间的可复现性有所提高（彼此比较，而非与原始研究比较），则进入第2轮；否则停止。

第***2***轮： 测试不同条件下的可重复性。

- 每项实验的重复次数：2 次，由两个不同的实验室完成；
- 需要改变的条件（实验因素）：评估者群体，以及评估者数量或任务复杂度；
- 如果复现结果足够接近，则进入第3轮；否则，在改进实验设计后重复第2轮，改进方式包括增加评估者数量，以及降低评估任务的认知复杂度。
- 对于第2轮的重复实验，如果复现研究之间的可复现性有所提高（彼此比较，而非与原始研究比较），则进入第3轮；否则停止。

轮 ***3***：在条件日益多样的情况下检验可复现性 条件。

- 每项实验的重复次数：由两个不同的实验室各进行 2 次；• 要改变的条件（实验因素）：评估者群体、评估者数量和 复杂度。

图 2：多实验室、多测试对照研究的原始设计，其中包含一组实验因素均衡的原始人工评估实验。

系统语言	英语	中文	德语	其他
	109	11	9	5
自然语言处理任务	摘要生成	对话系统	机器翻译	其他
	33	22	9	55
系统数量	*1-5*	*6-7*	*> 7*	不明确
	89	14	13	0
数据集数量	*1*	*2*	*> 3*	不明确
	83	25	8	0
参与者类型	人群（例如 *MTurk*）	作者/同事/学生	其他	不明确
	47	21	14	37
唯一参与者数量	*< 5*	*5-20*	*> 20*	不明确
	27	19	11	59
判断数量	*< 100*	*100-1000*	*> 1000*	不明确
	1	34	66	15

表3：116篇论文中高层级实验属性的频率（按论文统计）。部分论文包含多个类别属性，因此部分行的合计不会等于116。

15. 评估任务的复杂程度如何（低、中、高）；由本文作者评估。16. 界面的复杂程度如何（低、中、高）；由本文作者评估。

对参与者类型、培训、指导和专业知识进行分类非常困难。首先，并非所有实验都必然需要详细指导，但很难设定一个界限，以判断指导是否已超出敷衍应付的程度。培训也是如此。最终，我们决定记录是否提供了非敷衍的培训、指导、练习或标准定义。

专业知识也难以分类。一些论文最初可能会报告“专家 an-

标注者，但根据我们的询问，参与者表示自己是研究生或同事。这类参与者通常被称为“自然语言处理专家”。最终，如果原始研究的作者指出参与者是专家，我们就将其视为专家。

A.5 复现的常见方法

为确保研究之间具有可比性，我们就开展复现研究的以下共同方法达成了一致：

1. 制定完全重复原实验的计划，然后向研究伦理委员会申请批准。2. 如果原实验中向参与者支付了报酬，

质量标准名称	流畅性 10	连贯性 5	信息量 3	其他 54
系统语言	英语 26	中文 3	德语 2	其他 0
自然语言处理任务	摘要生成 6	问答 3	解释 3	其他 16
参与者类型	人群 13	学生 8	同事 7	其他 4
内在或外在	内在 26		外在 2	
绝对或相对	绝对 20		相对 8	
客观或主观	客观 2		主观 26	
独立参与者数量	*< 5* 11	*5–20* 4	*> 20* 8	不明确 5
评估项目数量	*< 200* 9	*200–1000* 10	*> 1000* 7	不明确 2
每个项目的参与者数量	*< 4* 17	*4–9* 3	*> 9* 3	不固定 5
每位参与者的项目数量	*< 50* 5	*50–200* 5	*> 200* 7	不固定/不明确 11
已提供培训。	*no* 24		不明确 4	
提供的说明	*yes* 8	*no* 15	不明确 5	
提供的标准定义	*yes* 17	*no* 3	*n/a* 4	不明确/混合 4
举行练习环节	*yes* 1	*no* 23	不明确 4	
参与者专业类型	*none* 16	研究人员 9	语言学家 2	领域 1
参与者 母语者	*yes* 2	*no* 12	的地区 10	未知 4

表4：28项实验中详细实验属性的频次。

则按照计算合理报酬的通用流程确定报酬（见附录）。

3. 填写完整的HEDS数据表。

4. 确定原论文中针对该实验报告了以下哪类结果：

(a) I类结果：单一数值评分，例如平均质量评分、错误计数等。(b) II类结果：一组数值评分，例如一组I类结果。(c) III类结果：附加在任意长度文本片段上的类别标签。(d) 原论文中明确陈述的定性结论/发现。

5. 严格按照 HEDS表中的描述开展分配给您的实验。6. 按照以下方式报告 8a–c 的量化可复现性评估：

(a) I类结果：变异系数（对小样本进行偏差校正）。(b) II类结果：皮尔逊相关系数、斯皮尔曼ρ。(c) III类结果：多评定者：Fleiss 的 κ；多评定者、多标签：Krippendorff 的α。(d) 结论/发现：并列总结在重复实验中得到确认和未得到确认的结论/发现。

A.6 发现的问题、缺陷和错误

1. 已发表论文中报告的人工评估数据存在错误，导致系统被报告为比实际表现更好或更差。2. 论文中报告的项目总数与所发送的文件不符。

3. 由于随机化操作不当，未能对待评估项目的顺序进行随机化（尽管声称要进行随机化）。4. 报告称评估者完成的评估数量相同，但从文件中可以清楚看出，他们完成的数量差异很大。5. 对部分参与者而非所有参与者进行了临时注意力检查（作者无法说明检查的具体内容）；未通过检查的参与者被排除在后续实验之外，但已完成的工作仍被保留。6. 聚合判断的方法有偏差（选择偏好的参与者，而非采用某种平均值）。

更一般地说，报告中的歧义可能会造成问题。即使对照 HEDS表进行检查，作者也可能觉得自己已经提及 HEDS 要求的所有实验细节，但这些内容往往描述得过于笼统，仍有被误解的空间。这意味着作者仍需确认论文是否得到了正确解读。自然语言处理作者可以考虑让第三方填写 HEDS表，看看他们在哪些地方会遇到困难，但这确实会增加额外负担。

A.7 ARR 负责任研究检查清单

A. **对于每篇投稿：**

A1.**您是否描述了您工作的局限性？** 是，例如，我们讨论了可供分析的论文是一个自我选择的子集（即作者作出了回应的论文），而非完整论文集，这一情况所带来的局限性。 A2. **您是否讨论了您工作中存在的任何潜在风险？** 本研究分析了此前经过同行评审并已发表的人工评估实验。尽管通常意义上的风险考量并不适用，但我们确实提到了以非匿名方式报告实验缺陷和/或其工作可复现性较低，可能对个体作者造成的潜在伤害。 A3. **摘要和引言是否概述了论文的主要论点？** 是，摘要、引言和结论都进行了概述。

总结这项工作的主要目标和结论。B. **您是否使用或创建了科学产物？** 未创建任何新数据或计算资源。C. **您是否运行了计算实验？** 未运行任何实验。D. **您是否使用人工标注员（例如人群工作者）或开展了涉及人类参与者的研究？** 本文未进行人工标注或评估（作者所做的除外）。