arXiv:2602.14349v1 [stat.AP] 15 Feb 2026

相同提示词，不同结果：评估大语言模型数据分析的可重复性*

Jiaxin (Allyson) Cui RohanAlexander

2026年2月15日

我们系统地评估了大语言模型（LLM）进行数据分析的可重复性。我们评估了两种提示策略、六种模型和四种温度设置，每种配置进行了十次独立执行，共计 480 次尝试。我们评估了每次尝试的完成度、一致性、有效性和一致性，发现即使配置相同，分析结果也存在显著差异。这表明，与人类进行的数据分析一样，大语言模型进行的数据分析也可能发生变化，即使任务、数据和设置都相同。我们的结果意味着，如果使用大语言模型进行数据分析，就应独立运行多次，并考量结果的分布。

*感谢Monica Alexander和Jamie Stafford提出宝贵建议。我们衷心感谢多伦多大学数据科学研究所和NSERC Alliance提供的资金支持。Cui在调查新闻基金会（IJF）工作期间开展了这项分析，我们衷心感谢IJF提供任命数据库。联系方式：rohan.alexander@utoronto.ca。

关键词：人工智能；可重复性；代码生成；数据分析；基准测试；大语言模型

CRediT 贡献： *Cui*：概念化；数据整理；正式分析；调查；方法论；软件；可视化；撰写——初稿；撰写——审阅与编辑。*Alexander*：概念化；经费获取；方法论；项目管理；监督；可视化；撰写——初稿；撰写——审阅与编辑。

人工智能使用情况： 研究中使用了人工智能辅助编程相关工作，包括与应用程序接口交互、数据清理以及制作图表，但所有代码和结果均由作者审阅和评估。人工智能还用于推荐相关文献，并审阅论文草稿，以发现错误、不一致之处和措辞不当的问题。

意义声明：大语言模型越来越多地被用于生成数据分析代码，但人们较少考察其估计值的一致性。我们针对一项贴近现实的数据分析任务，运行了48种不同的模型、提示词和温度配置，每种配置重复运行十次。我们发现，即使配置完全相同，得出的结果在实质上也可能有所不同。即使温度为零，有些估计值也会导致人们对同一研究问题得出不同结论。这些发现表明，仅进行一次大语言模型生成的数据分析并不足够，重复进行独立执行应成为标准做法。

1 引言

科学的一项基础是能够按照他人所采用的相同步骤操作，并得到相同结果。不同学科对这一概念的称呼各不相同（Barba 2018），但对可重复性的一种定义是“使用相同的数据和方法获得一致的结果”（美国国家科学、工程与医学研究院 2019）。为提高计算可重复性，如今许多期刊都要求共享代码和数据（Stodden、Seiler 和 Ma 2018；Vilhuber 2020）。

与这些举措并行的，还有对所谓“多分析者”方法的探索：不同研究团队独立分析同一个数据集，并比较各自的发现。例如，29个团队分析了同一份数据，研究足球裁判是否会以更高的比率向深色皮肤球员出示红牌。结果发现，其中9个团队未发现具有统计显著性的效应，而其余20个团队则发现了（Silberzahn 等，2018）。其他研究也发现，不同研究团队之间存在类似差异（Botvinik-Nezer 等，2020；Ostropolets 等，2023；Ortloff 等，2023；Huntington-Klein 等，2025）。一个关键的担忧是，许多差异源于“……数据准备和分析决策方面的不同，而这些差异很可能不会在发表的论文中报告”（Huntington-Klein 等，2021，944）。

如今，大语言模型（LLM）能够编写出性能超越人类代码，或至少在某些维度上与人类代码难以区分的代码（Hou and Ji 2024；Xu et al. 2025）。早期迹象表明，大语言模型可以运用这一能力编写代码，完成数据分析中的某些环节（Bubeck et al. 2023）；如今，人们正在更贴近现实的情境中探索这一能力（Li et al. 2026）。随着大语言模型越来越具备自主行动能力，也就是无需直接指示便能追求目标（Grootendorst and Alammar 2026），它们将能够自主开展数据准备与分析。

本文关注的是：与人类团队类似，大语言模型在面对相同的任务和数据时，多次运行是否会得出不同结论。迄今为止，许多大语言模型评估关注的是最佳表现，而非表现的分布。这对于软件工程等领域是合适的，因为这些领域可能关注代码是否通过相应测试，而不在意需要尝试多少次。例如，`pass@k`指标（Kulal et al. 2019）曾用于衡量语言模型 Codex 根据文档字符串生成的 k 个函数中是否至少有一个正确（Chen et al. 2021）。大语言模型在根本上具有统计特性，即使温度为零，也可能表现出非确定性（Atil et al. 2025；He and Lab 2025）。实际上，这意味着重复运行同一提示词时，几乎总有可能得到不同结果。这也是 `pass@k` 指标在这些领域如此实用的原因之一：只要曾经达到某一表现水平，就意味着已经达到该水平，即使其他尝试未能达到。

如今，大语言模型被用于生成支撑多个领域科学论断的分析代码，包括传染病（Kraemer 等人（2025））、科学发现（Wang 等人（2023））和材料发现（Cheng 等人（2026））。与软件

工程环境中的代码不同，用于科学发现和数据分析的代码可能不会被反复运行，因为它可能仅仅是某项发现的依据。但如果变异性很重要——例如，我们想要判断某种差异是否具有统计显著性时——那么，大语言模型进行的数据分析仅仅有一次得出特定结论，并不一定足够。

我们考察在一项贴近现实的数据分析任务中，同一提示词多次执行是否会得出截然不同的结论。我们评估了来自三家提供商的六种模型：Anthropic 的 Claude Haiku 4.5 和 Claude Opus 4.6、OpenAI 的 GPT-5-mini 和 GPT-5.2，以及 Google 的 Gemini Flash 3 和 Gemini Pro 3。我们考察了四种温度设置：0.0、0.3、0.7 和 1.0；较低的设置通常会返回概率更高的词元（Alammar 和 Grootendorst 2024, 171）。我们还考察了两种不同的提示方式：单步和多步。前者使用一个提示词生成完整的分析流程；后者则通过一系列连续的提示词，逐一完成各个阶段，并进行中间验证。多步提示应有助于及时纠正方向，而单步提示更贴近现实世界中人类数据分析师的工作方式，也关系到人工智能可能带来的经济效益。对于每种模型和设置，我们针对每种提示策略进行了十次独立执行。这总计产生了 480 次执行尝试。

我们评估了一个五步骤数据分析流程的表现，该流程处理加拿大某省的政府任命记录：合并多个数据文件、识别再次任命模式、汇总至分析单位、进行普通最小二乘回归，并生成可视化图表。该任务类似于初级数据科学家或初级分析师所承担的工作。

我们记录了三项主要发现。第一，单步提示的完成率往往高于多步提示，因为在多步方法中，任何阶段出错都会导致后续所有阶段无法运行。第二，生成的代码几乎总是在结构上有效，但可能与人类进行的数据分析有所不同，尤其是在数据准备步骤中；这些步骤对于排序、缺失值处理以及纳入哪些记录等选择，规定得并不明确。第三，这些数据准备差异会传递到分析中，产生不同的估计值簇。尽管大多数配置得出的 t 统计量均表明不存在具有统计显著性的关系，但估计斜率在正负方向和大小上都有差异。这意味着，相同的数据和任务可能会因模型采用的分析路径不同而得出不同结论。

2 结果

2.1 总体

我们根据四项标准评估大语言模型的执行尝试：完成度、一致性、有效性和一致性（图 1）。完成度指生成的代码能否无误执行，并且

生成预期的输出文件。一致性是将输出与人类进行的分析进行比较。有效性指输出是否具有预期的列、姓名、数据类型和值域。一致性则指特定配置的 10 次执行之间的相似程度。附录 A2 中的图 C 提供了更多细节。

2.2 完成度

图 1 (a) 展示了两种提示策略在不同流水线步骤、温度和模型下的完成率。完成度衡量生成的代码是否生成了所有预期输出文件。

图 1 (a) 显示，单步提示方法（模型一次性接收所有指令）的完成率往往高于多步提示方法（将任务拆分为较小的提示）。这似乎有些违反直觉。人们可能会认为，将任务拆分成若干独立步骤会让模型更容易完成任务。但图 1 (a) 展示了相反的结果，这是因为错误传播。

在多步提示方法中，每个步骤都会生成一个独立的脚本，该脚本必须成功执行，下一步才能运行。任何环节出错都会导致错误传播；例如，如果步骤2失败，那么步骤3到5都不会执行，即使这些步骤本来可以正常运行也是如此。在单步提示方法中，模型能够生成一个内部一致的脚本。由于所有代码都是作为一个整体生成并执行的，因此各步骤之间出现故障传播的可能性较小。GPT-5-mini 和 Gemini Flash 3 体现了这一点：多步与单步模式下的完成率存在显著差异。Gemini Flash 3 在单步模式下的完成率达到100%，但在多步模式下为0%，因为每次执行都会在步骤2因文件路径处理错误而失败。

对于单步提示，在所有温度值下，大多数模型的完成度都很高。对于多步提示，温度似乎也与完成度没有稳定的关联，不过 GPT-5-mini 不支持自定义温度设置。

2.3 一致性

考察大语言模型估计值与人类估计值之间一致性的方式有很多。一种方式是查看所有步骤的一致性。图 1 (b) 中的一致率基于成功完成的执行结果。对于某个步骤，如果模型未能生成输出文件，则该模型会被排除在该步骤的一致性计算之外。例如，GPT-5-mini 的某些多步提示配置在步骤5的完成度仅为 60%，尽管成功完成步骤5的执行结果可能具有较高的一致率。这一模式表明，模型在正常运行时能够生成一致的输出，但经常完全无法生成可用的代码。



图 1：不同流水线步骤、模型、温度和提示策略的评估指标。每个色块显示一个模型-步骤组合的比率。行按提示策略分组，列按温度分组。颜色深浅表示指标值，从 0（红色）到 1（绿色）。GPT-5-mini 不支持自定义温度，因此仅在其默认温度 1 下进行评估。

考察一致性的另一种方式是只关注一个方面。例如，考察各院系估计的年度续聘率。将大语言模型生成的比率与人类生成的比率进行比较，我们会发现两者存在显著差异（图 2）。除少数例外，大语言模型对再任命率的估计几乎都低于实际再任命率。

图 2 比较了机构-年份层级的各项再任命率估计值，显示许多大语言模型输出与实际比率相同或接近。图 1 (b) 的判定标准更严格——只有数值完全相同（四舍五入后）才算一致；任何数值、结构或纳入观测值方面的差异，都视为不一致。图 1 (b) 中的一致率，是每种配置下十次执行中实现完全匹配的比例。因此，图 2 与图 1 (b) 之间的差异，源于粒度层级不同。图 1 (b) 展示了模型表现较好和较差的环节。步骤2和步骤3的一致率始终最低，而步骤4和步骤5的一致率往往较高。这可能是因为回归计算和可视化属于定义明确的任务，因此模型更容易生成与人类输出相匹配的结果。步骤5的一致性只检查是否生成了PNG文件，而不检查图像内容是否匹配，因此其较高的比率应理解为完成度，而非视觉相似性。

2.4 有效性

图 1 (c) 显示了在文件存在的条件下的有效率。有效性用于评估生成的输出是否符合任务规范：列是否存在且命名正确、数据类型是否恰当，以及值是否处于有效范围内。几乎所有配置的有效率都很高。

2.5 一致性

图 1 (d) 显示了每种配置下重复执行的一致性。对于每种配置，我们找出十次执行中最常见的输出，并计算与之匹配的比例。比率为 1.0 表示该步骤中所有产生输出的执行结果完全相同；数值较低则表示各次执行结果存在差异。由于可视化输出无法进行数值比较，因此不评估步骤5的一致性。结果表明，即使温度为零，也不能假设模型会提供相同的结果。

图 3 展示了每次完成步骤4的独立执行所得到的回归斜率、截距、标准误和 t统计量。每个点代表一次执行。配置按模型、提示策略和温度分组；这四个面板中的分组顺序保持一致，因此每一行在所有面板中都对应同一种配置。y 轴括号中的数字表示产生有效回归输出的执行次数。由于 GPT-5-mini 不支持自定义温度，其四个温度组按每种提示策略合并为一行，总计尝试40次。配置中较少

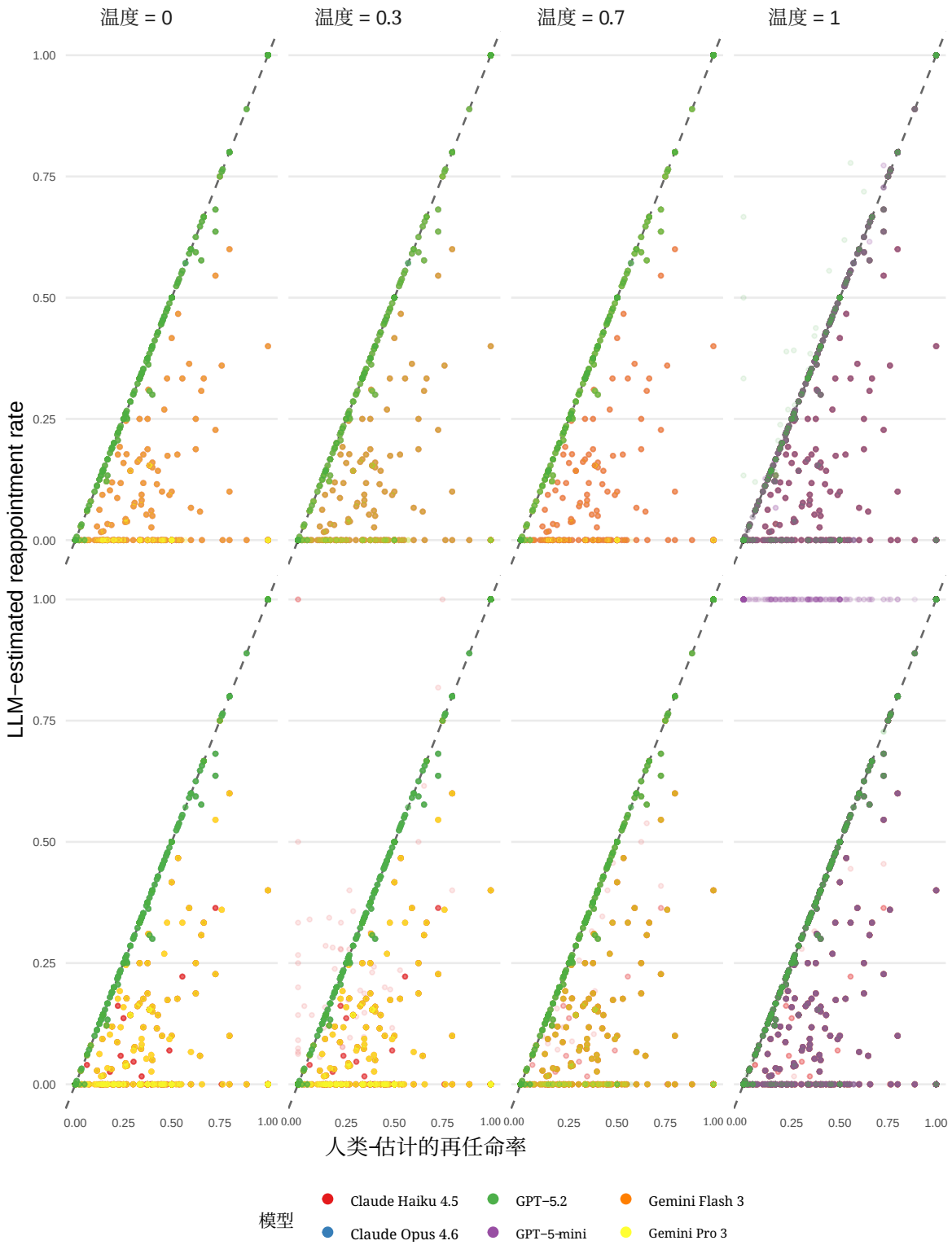


图2：比较大语言模型在院系—年份层级估计的再任命率与人工分析得出的比率。每个点代表一次执行中的一个院系—年份观测值。虚线45度线表示估计值相同。GPT-5- mini 仅在其默认温度为1时进行评估。[8]

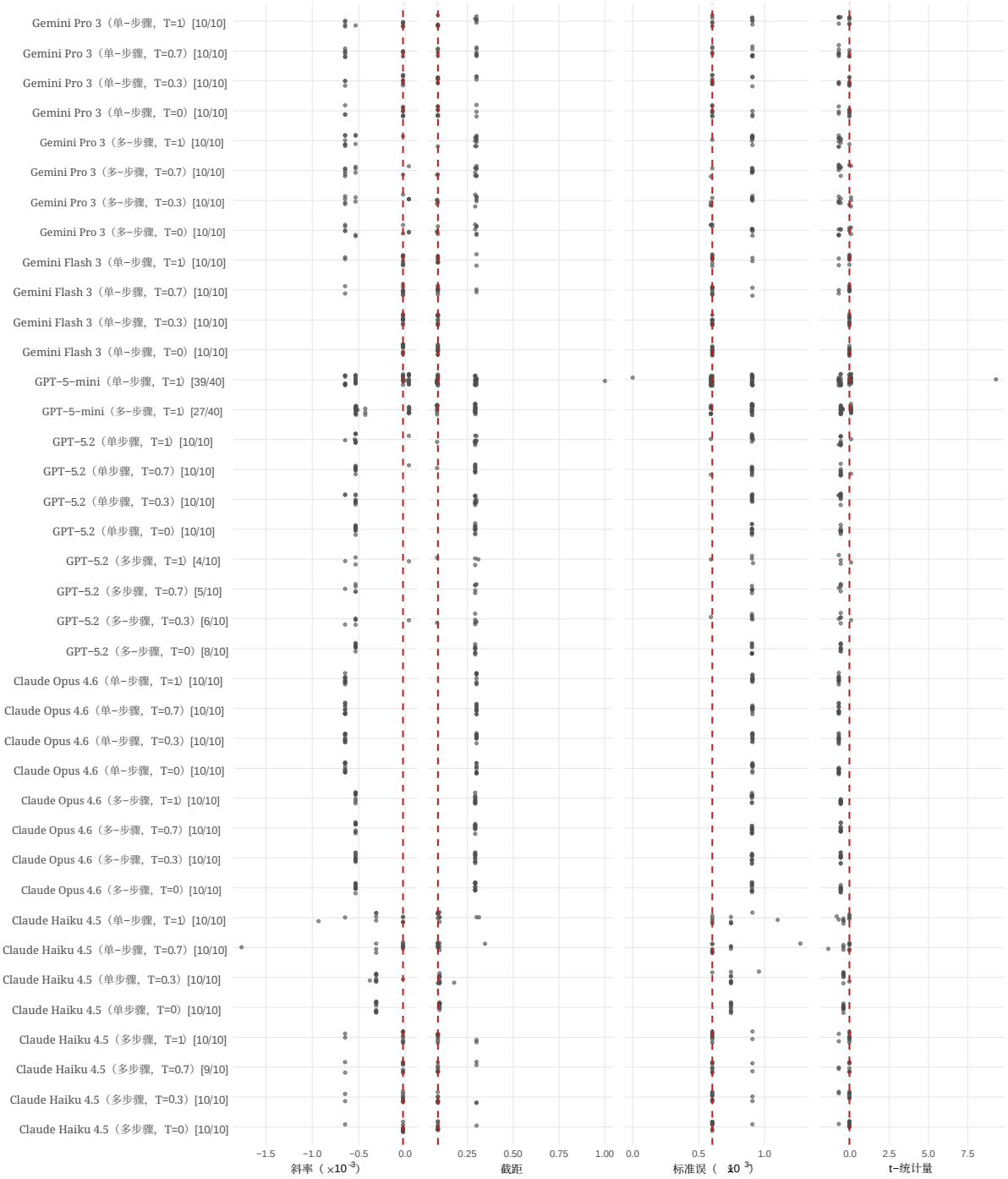


图 3：按配置显示每次单独执行的输出。每个点代表一次独立执行。括号中的数字表示总尝试次数中产生有效回归输出的执行次数。由于 GPT-5-mini 不支持自定义温度，其配置按温度设置合并显示。配置按模型、提示策略、温度的顺序分组。虚线表示人类生成的值。

已完成的执行次数较少，例如 GPT-5-mini（多步，T=1）完成了 40 次尝试中的 33 次；这并非因为一致性，而是因为部分尝试在生成回归结果之前就失败了。虚线表示人类生成的数值。

估计值呈现出若干特定的聚类，这或许表明大语言模型遵循着特定的分析路径，且数据准备的选择对结果有很大影响。标准误面板显示，不同配置的精度存在差异。这可能表明数据清理后保留的观测值数量有所不同。大多数执行得到的 t 统计量接近零，与人类分析结果一致；多步提示下，没有任何 t 统计量的绝对值超过 2，不过少数单步配置得到的 t 统计量可能会导向具有统计显著性差异的结论。

3 讨论

我们评估了大语言模型代码生成在两种提示策略、六种模型和四种温度设置下的可靠性与一致性。我们发现，即使设置相同，不同执行的输出也会有所不同，结果也可能与人类得出的结果不同。尽管如此，大语言模型几乎总能生成结构有效的代码；尤其是在一次性向其呈现完整任务，而非拆分任务时。人们原本期待代码和数据共享能够促进不同工作流程之间的方法论验证（Donoho 2017），而我们的研究表明，人工智能的最新进展已让我们更接近这一愿景。

我们的结果表明，在使用大语言模型进行数据分析时，有必要对同一任务执行多次。不同模型的表现各异，因此我们的结果也强调，使用至少两家不同提供商的模型开展任何数据分析都很重要。至少，如果研究重点是推断，就应展示结果中的任何变异；但如果重点是预测，也可以考虑采用某种集成方法。

尽管如此，大语言模型给出的估计值往往会以特定方式产生差异。这表明，大语言模型生成的代码会导致分析工作流程中的某些环节出现差异。这与多分析师研究中的发现相似：相关研究认为，许多差异源于数据准备和清理方面的不同选择（Huntington-Klein 等，2025）。与早期关于人工多分析师分析的研究（Silberzahn 等，2018）一样，这些系统性差异会导致估计出的斜率系数不同。

这些差异往往出现在特定的流水线阶段。一些模型会排除文件，或以不同方式解析年份。时间排序和平局处理会在步骤2产生差异，不同的稳定排序会改变哪些记录被视为首次出现。缺失值处理决定空组织值是否会在聚合过程中保留下来，从而直接影响回归样本构成。这表明提示词可能缺乏足够的具体性。但这正是数据分析的现实：数据分析是一个从数据中学习的交互式过程。

我们表明，两个人工智能智能体可以仅凭相同的数据，自主地针对同一任务得出不同结论。随着智能体数据分析的成本降低，这表明有必要制定相关指南，正如在人类多分析师研究的语境中所讨论的那样（Auspurg 和 Brüderl 2024），以便随着探索“不确定性的隐藏宇宙”变得更加廉价，能够进行系统性比较（Breznau 等，2022）。

我们的研究存在若干局限。我们使用一个数据集，在一个领域内评估单一分析任务，而我们的研究发现能在多大程度上推广到其他任务、学科或数据结构，尚不得而知。我们的一致性度量依赖于一名人类分析师，但多分析师方法表明，对于同一任务，可能存在多种有效的方法。话虽如此，我们并不是说与人类生成的结果不同的结果就是错误的，只是说它们有所不同。我们在一个时间点测试了来自三家提供商的六个模型，但模型行为会随着更新而变化，因此我们的结果只是一个快照。我们只考虑不经过迭代的单轮代码生成，但在实践中，分析师会审查并修改大语言模型输出，而允许模型执行、观察错误并重试的智能体框架可能会产生不同的结果。

考虑不同的数据结构和领域，以及更具体或不那么具体的提示是否会影响我们的结果，将会很有意义。尤其值得考察智能体工作流，在这种工作流中，模型可以检查中间输出，并在继续之前纠正错误。最重要的是，我们的工作表明，有必要开发能够自动实现多次独立执行的方法，并识别这些执行是否已沿着不同的分析路径收敛，同时考察估计值的分布。

4 方法

底层实质性数据集是加拿大新不伦瑞克省的年度公共任命记录。这些数据来自调查新闻基金会，是真实数据，包含现实世界数据中常见的杂乱情况，例如姓名不一致、缺失值和错误。有关这些数据的更多详情见附录 A。我们希望回答以下研究问题：新不伦瑞克政府部门的重新任命比率在不同部门之间有何差异，又如何随时间变化？为此，我们建立了一个包含五个步骤的数据处理与分析流程：

1. 将 12 个年度 CSV 文件（2013年至2024年，含首尾年份）合并为一个数据集。2. 通过匹配姓名—职位—组织组合，识别个人再任命。3. 将个人再任命汇总到组织—年份层面，计算任命总数、再任命数和比率。4. 以再任命率为结果变量、任命总数为预测变量，估计回归模型。5. 绘制散点图，并叠加回归线。

每个步骤都以自然语言描述，类似于经理向初级数据分析师下达指示。这意味着并未完全明确所有细节，因此可能存在一定差异。

我们考察了两种提示方法。第一种是“多步”方法：模型依次接收五个步骤各自对应的提示词。每条提示词都会说明该步骤的任务、预期输入文件以及所需输出。大语言模型会生成各自独立的 Python 脚本，然后按顺序运行。第二种是“单步”方法：大语言模型接收一条包含全部五个步骤及其输入和输出的提示词，并根据要求编写一个执行所有环节的 Python 脚本。两种提示方法的任务规范相同；若执行成功，产出的结果也相同。关于提示词的更多详情见附录 B。

我们评估了来自三家提供商的六个模型：Claude Haiku 4.5 和 Claude Opus 4.6（Anthropic）、GPT-5-mini 和 GPT-5.2（OpenAI），以及 Gemini Flash 3 和 Gemini Pro 3（Google）。每个模型均采用两种提示策略进行测试，并在四种温度设置（0.0、0.3、0.7、1.0）下，每种配置独立执行十次，共计尝试执行 480 次。GPT-5-mini 不支持自定义温度设置，因此其 40 次执行均采用模型默认温度 1（OpenAI 2026）。GPT-5.2 需要关闭内置推理才能接受温度参数。Gemini Pro 3 则必须启用思考模式，且无法关闭；我们将思考预算设为最低，以近似标准生成。作为背景说明，温度与返回较低概率词元的可能性有关：温度为零意味着返回高概率词元，而温度为一则意味着有时会返回概率较低的词元（Alammar 和 Grootendorst 2024, 171）。

我们使用统计编程语言 `R` (R 核心团队 2025)，以及 `tidyverse` (Wickham 等人，2019)和 `tinytable`(Arel-Bundock，2025) 软件包完成所有分析并生成所有图表。

附录

A 底层数据

我们的基础数据来自加拿大各省通过枢密院令（OiC）发布的文件。枢密院令正式确认了对政府机构、董事会和委员会的任命，并形成关于公共人员任免决定的行政记录。调查新闻基金会将这些记录汇编成标准化的“任命数据库”，记录任命属性，包括受任者姓名、组织、职位、任命日期以及是否续任（调查新闻基金会，2025）。我们聚焦于新不伦瑞克省。表 A1 列出了描述性统计数据。

表 A1：新不伦瑞克任命数据集（2013–2024）的汇总统计。该数据集包含 43 个组织的 3,718 条任命记录。再任命率按原始行政数据中标记为再任命的记录所占比例计算。

特征	值
任命总数（2013-2024）	3,718
不同被任命者	2,357
不同组织	43
涵盖年份	12
每年任命次数（均值）	310
每年任命次数（标准差）	72
每年组织数（均值）	17
每年组织数（标准差）	3
组织信息缺失的记录	1.5%
标记为再次任命的记录	874
再任命率（汇总）	23.5%

表格A2 摘录自原始任命记录。它展示了模型必须分析的数据结构。

表A2：新不伦瑞克任命数据库（2013年）中的示例记录。字段包括受任者姓名、职位、组织、连任状态及行政元数据。记录显示了行政数据中常见的完整度和格式差异。

Name	职位	Org	重新任命	开始日期
Cindy Howe	成员	卫生	否	2013年1月9日
Floyd Haley	成员	卫生	否	2013年1月9日
Frank B. Trevors	成员	卫生	否	2013年1月9日
Gerald Richard	副部长...	执行委员会...	否	2013年1月21日
Harry Doyle	成员	卫生	否	2013年1月9日
Jeffrey Beairsto	成员	卫生	否	2013年1月9日

图 A1 展示了任命数量随时间的分布。这些数据用于与模型的估计值进行比较。

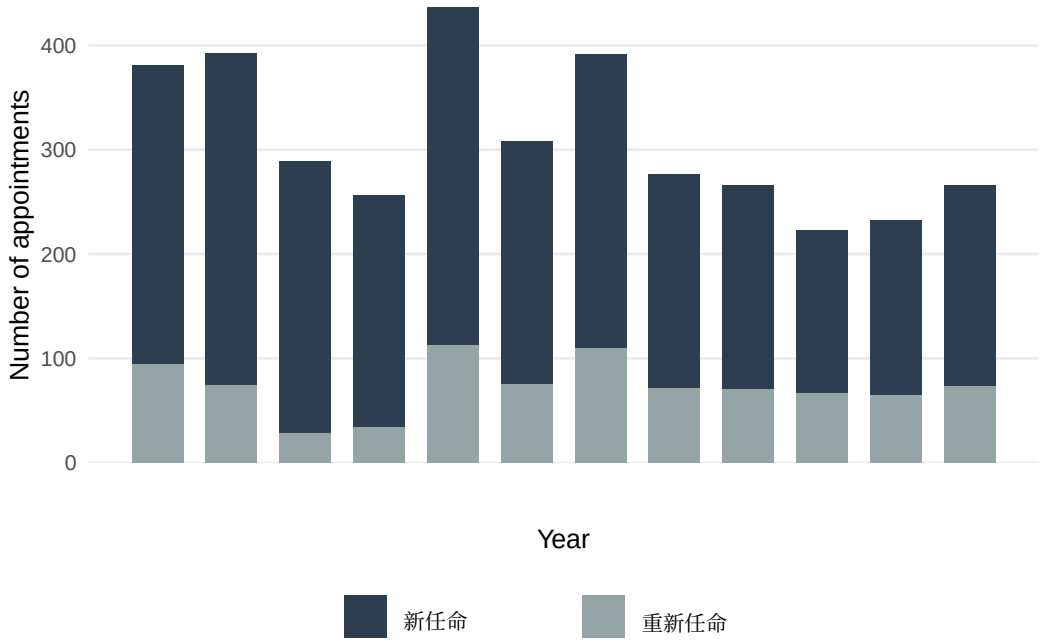


图 A1：新不伦瑞克的年度任命数量，按连任状态划分（2013年–2024年）。深色柱表示新任命，浅色柱表示再任命。任命总数和再任命数量均存在逐年变化。

B 提示词详情

B.1 多步提示

多步提示方法依次使用五个提示。每个提示前都会加上：

```
Only respond with the python script that can be directly used to run the task,
do not include any explanation or other text. Therefore, your response should
start with "import...", and end with "if __name__ == '__main__': main()"
```

步骤 1： 提示词指定了位于 `prompt_raw_data/appointments_{year}.csv` 的多个 CSV 文件，每个文件包含 16 列（姓名、职位、组织、地点、地区、发布日期、开始日期、结束日期、任期长度、法案、薪酬、重新任命、代任负责人、href、正文、链接）。模型需要合并这些文件，添加年份列，并仅保留姓名、职位、组织、重新任命和年份。预期输出为 `step1.csv`，包含这五列，且均为可空。

步骤2： 提示词将 `step1.csv` 作为输入，并要求模型按姓名、职位、组织、年份和重新任命排序，然后在相同的姓名-职位-组织组合并非首次出现时，将重新任命列更新为 true。预期输出为 `step2.csv`，包含相同的五列。

步骤3： 提示词将 `step2.csv` 作为输入，并要求模型删除姓名和职位列，将数据集压缩为按年份和组织分组的数据，其中包含任命总数和再任命总数，并添加再任命率列。预期输出为 `step3.csv`，包含五列：年份、组织、任命总数_、再任命总数_和再任命_率。

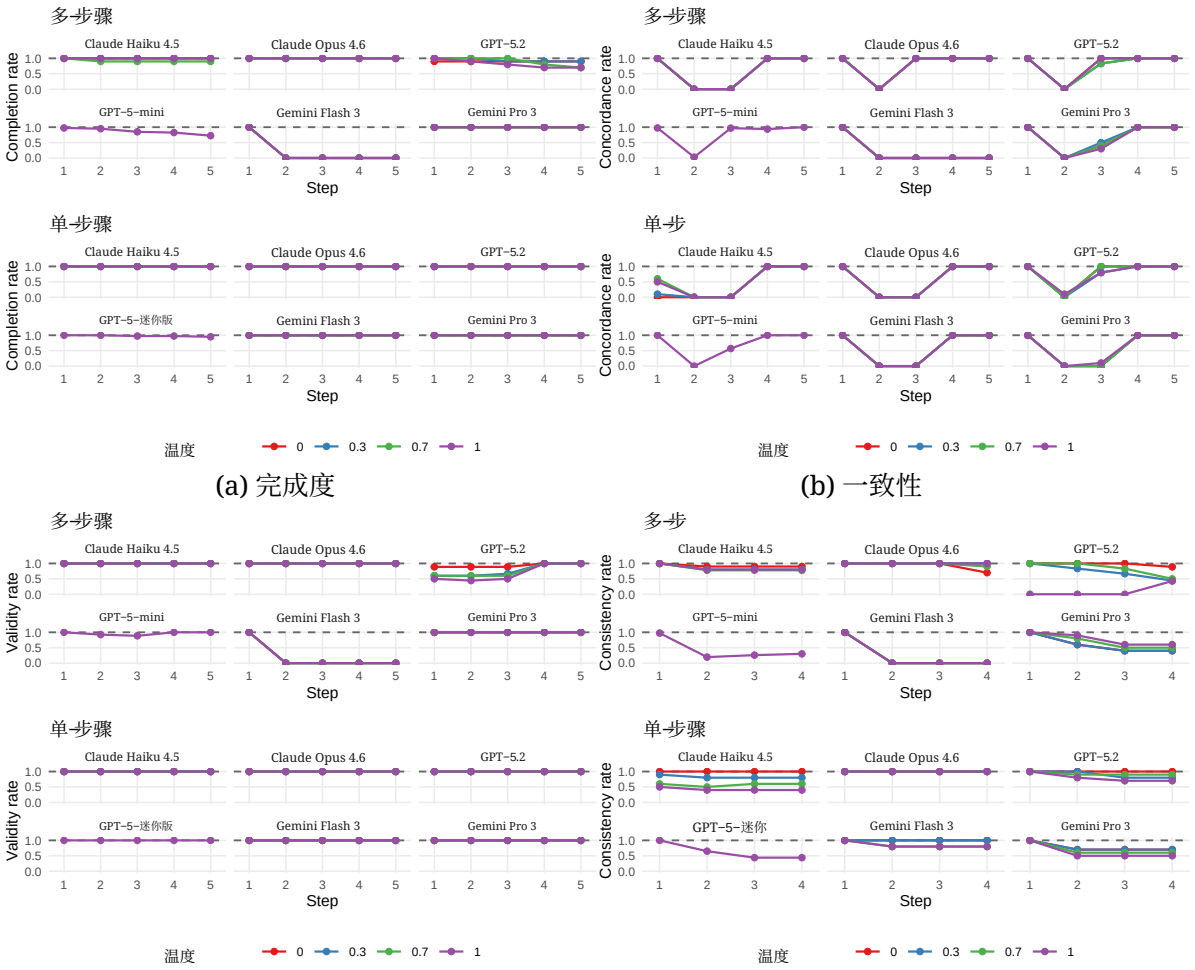
步骤4： 提示词将 `step3.csv` 作为输入，并要求模型对再任命_率与任命总数_进行普通最小二乘回归，记录斜率、截距、标准误和 t 统计量，并评论统计显著性和经济显著性。预期输出为 `step4.csv`，包含八列，其中包括回归系数、显著性标记和文字评论。

步骤5： 提示词将 `step3.csv` 和 `step4.csv` 作为输入，并要求模型创建散点图，以任命总数_为 x 轴、再任命_率为 y 轴，使用斜率和截距绘制回归线，并按年份使用 viridis 色图。预期输出为 `step5.png`。

B.2 单步提示

单步提示方法只是将全部五个步骤整合到一个具有相同前缀的提示词中。因此，任务描述被合并，并明确要求生成一个完整的流水线脚本。

C 更多结果详情



(a) 完成度 (b) 一致性

(c) 有效性 (d) 一致性

图 A2：按模型细分的各流水线步骤评估指标。每个面板展示一个模型在所有温度下的变化轨迹。线条颜色表示温度。前两行展示多步结果，后两行展示单步结果。GPT-5-mini 仅在其默认温度 1 下进行评估。

参考文献

Alammar, Jay, and Maarten Grootendorst. 2024. 《Hands-on Large Language Models》。O’Reilly。https://www.oreilly.com/library/view/hands-on-large-language/9781098150952/。Arel-Bundock, Vincent. 2025. Tinytable：适用于“HTML”、“La- TeX”、“Markdown”、“Word”、“PNG”、“PDF”和“Typst”格式的简单且可配置表格。https://doi.org/10.3 2614/CRAN.package.tinytable。Atil, Berk, Sarp Aykent, Alexa Chittams, Lisheng Fu, Rebecca J. Passonneau, Evan Radcliffe, Guru Rajan Rajagopal, et al. 2025. “「确定性」大语言模型设置的非确定性。”https://arxiv.org/abs/2408.04667。Auspurg, Katrin, and Josef Brüderl. 2024. “通过多分析师研究，更可信地评估科学的可信度。”美国国家科学院院刊 121 (38)。https://doi.org/10.1073/pnas.2404035121。Barba, Lorena A. 2018. “可重复研究术语。”https://arxiv.org/abs/18 02.03311。Botvinik-Nezer, Rotem, Felix Holzmeister, Colin F. Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, et al. 2020. “多个团队对单个神经影像数据集进行分析的变异性。”《自然》582 (7810)：84–88。https://doi.org/10.1 038/s41586-020-2314-9。Breznau, Nate, Eike Mark Rinke, Alexander Wuttke, Hung H. V. Nguyen, Muna Adem, Jule Adriaans, Amalia Alvarez-Benjumea, et al. 2022. “观察众多研究者使用相同的数据和假设，揭示了不确定性的隐藏宇宙。”美国国家科学院院刊 119 (44)。https://doi.org/10.1073/pnas.2203150119。Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, et al. 2023. “通用人工智能的曙光：GPT-4 的早期实验。”https://arxiv.org/abs/2303.12712。Chen, Mark, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, et al. 2021. “评估基于代码训练的大语言模型。”https://arxiv.org/abs/2107.03374。Cheng, Mouyang, Chu-Liang Fu, Ryotaro Okabe, Abhijatmedhi Chotrattanapituk, Artittaya Boonkird, Nguyen Tuan Hung, and Mingda Li. 2026. “人工智能驱动的材料设计与发现方法。”《自然·材料》25（1月）：174–90。https://doi.org/10.1038/s41563-025-02403-7。Donoho, David. 2017. “数据科学 50 年。”《计算与图形统计学杂志》26 (4)：745–66。https://doi.org/10.1080/10618600.2017.1384734。Grootendorst, Maarten, and Jay Alammar. 2026. 《人工智能智能体图解指南》。O’Reilly。https://www.oreilly.com/library/view/an-illustrated-guide/9798341662681/ch01.html。He, Horace, and Thinking Machines Lab. 2025. “战胜大语言模型推理中的非确定性。”https://doi.org/10.64434/tml.20250910。Hou, Wenpin, and Zhicheng Ji. 2024. “比较大语言模型与人类程序员生成程序代码的能力。”《先进科学》12 (8)。https://doi.org/10.1002/advs.202412279。

Huntington-Klein, Nick, Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey R. Bloem, Pral- had Burli, Naibin Chen, et al. 2021. “隐藏的研究者决策对应用微观经济学的影响。”《经济学探究》59 (3)：944–60。https://doi.org/10.1111/ec in.12992。
Huntington-Klein, Nick, Claus C Portner, Ian McCarthy, and The Many Economists Collabo- rative on Researcher Variation. 2025. “经济学中研究者差异的来源。”工作论文33729。工作论文系列。美国国家经济研究局。https://doi.org/10.3386/w33729。调查新闻基金会。2025. “任命数据库。”https://theijf.org/a ppointments。Kraemer, Moritz U. G., Joseph L.-H. Tsui, Serina Y. Chang, Spyros Lytras, Mark P. Khurana, Samantha Vanderslott, Sumali Bajaj, et al. 2025. “用于传染病流行建模的人工智能。”《自然》638（2月）：623–35。https://doi.org/10.1038/ s41586-024-08564-w。Kulal, Sumith, Panupong Pasupat, Kartik Chandra, Mina Lee, Oded Padon, Alex Aiken, and Percy Liang. 2019. “SPoC：基于搜索的伪代码转代码。”载于第33届神经信息处理系统国际会议论文集，11906–17。https://dl.acm.org/doi/10.5555/3454287.3455353。Li, Yiyang, Zheyuan Zhang, Tianyi Ma, Zehong Wang, Keerthiram Murugesan, Chuxu Zhang, and Yanfang Ye. 2026. “LongDA：面向长文档数据分析的大语言模型智能体基准测试。” https://arxiv.org/abs/2601.02598。美国国家科学、工程与医学研究院。2019.《科学中的可重复性与可复现性》。华盛顿特区：美国国家科学院出版社。https://doi.org/10.17226 / 25303。OpenAI。2026. “GPT-5 Mini 模型。” https://platform.openai.com/docs/models/gpt-5-mini。Ortloff, Anna-Marie, Matthias Fassl, Alexander Ponticello, Florin Martius, Anne Mertens, Katharina Krombholz, and Matthew Smith. 2023. “不同的研究者，不同的结果？分析研究者经验和数据类型对一项有关安全建议的访谈与调查研究进行定性分析的影响。”载于2023年计算机系统中的人为因素国际会议论文集，1–21。CHI ’23。美国计算机协会。https://doi.org/10.1145/3544548.3580766。Ostropolets, Anna, Yasser Albogami, Mitchell Conover, Juan M Banda, William A Baumgart- ner, Clair Blacketer, Priyamvada Desai, et al. 2023. “可复现的变异性：评估 9 个研究团队在尝试复现同一项观察性研究时的研究者意见不一致。”《美国医学信息学会杂志》30 (5)：859–68。https://doi.org/10.1093/jamia/ocad009。R 核心团队。2025. R：用于统计计算的语言和环境。奥地利维也纳：统计计算基金会。https://www.R-project.org/。Silberzahn, R., E. L. Uhlmann, D. P. Martin, P. Anselmi, F. Aust, E. Awtrey, Š. Bahník, et al. 2018. “众多分析者，一个数据集：透明呈现分析选择的差异如何影响结果。”《心理科学方法与实践进展》1 (3)：337–56。https://doi.org/10.1177/2515245917747646。Stodden, Victoria, Jennifer Seiler, and Zhaokun Ma. 2018. “对同行……”

计算可重复性的国家政策成效。”美国国家科学院院刊 115(11): 2584–89. https://doi.org/10.1073/pnas.1708290115。Lars Vilhuber。2020。“经济学中的可重复性与可复现性。”《哈佛数据科学评论》 2(4). https://doi.org/10.1162/99608f92.4f6b9e67。Hanchen Wang、Tianfan Fu、Yuanqi Du、Wenhao Gao、Kexin Huang、Ziming Liu、Payal Chandak 等。2023。“人工智能时代的科学发现。”《自然》620(August): 47–60. https://doi.org/10.1038/s41586-023-06221-2。Hadley Wickham、Mara Averick、Jennifer Bryan、Winston Chang、Lucy D’Agostino McGowan、Romain François、Garrett Grolemund 等。2019。“欢迎使用 tidyverse。”《开源软件期刊》 4(43): 1686. https://doi.org/10.21105/joss.01686。Xiaodan Xu、Chao Ni、Xinrong Guo、Shaoxuan Liu、Xiaoya Wang、Kui Liu 和 Xiaohu Yang。2025。“通过对比学习区分大语言模型生成的代码与人类编写的代码。”《*ACM*软件工程与方法学汇刊》 34(4): 1–31.https://doi.org/10.1145/3705300。