迈向美国国会的数字孪生

Hayden Helm
Helivan
hayden@helivan.io

Tianyi Chen
Johns Hopkins University

Harvey McGuinness
Johns Hopkins University

Paige Lee
Nomic AI

Brandon Duderstadt
Nomic AI

Carey E. Priebe
Johns Hopkins University

arxiv:2505.00006v1 中译本

摘要

在这篇论文中，我们提供了证据表明基于一组语言模型的美国国会议员虚拟模型符合数字孪生的定义。特别地，我们介绍了并提供了对一个每日更新的数据集的高层次描述，该数据集中包含了每位美国国会议员在其任期内发布的每一条推文。我们展示了配备有特定议员子集数据的现代语言模型能够生成几乎无法与物理对应者实际发布的推文区分开来的推文。我们说明了如何利用生成的推文来预测点名投票行为，并量化国会议员跨越党派界限的可能性，从而帮助利益相关方分配资源并可能影响现实世界的立法动态。我们在讨论我们的分析的局限性和重要扩展中得出结论。

数字孪生是捕获物理系统相关属性的虚拟模型。要被称为数字孪生，虚拟模型必须能够生成最新的推断，这些推断可以影响物理系统的运行行为。在过去五年中，数字孪生的发展和部署出现了显著增长。例如，患者的数字孪生已经使肿瘤学 (32; 29) 和心脏病学 (5; 27) 中的预测医学采取了高度个性化的途径成为可能。类似地，数字孪生也承诺改进风力发电的电网整合 (28; 9)；在制造环境中实现快速进步的加工和质量控制过程 (4; 10; 16)；以及提供有效的解决方案来应对社会问题如城市规划 (26) 和可持续发展 (31; 24; 25)。

随着数字孪生在众多科学和工业学科中的发展，大型语言模型（如 OpenAI 的 GPT-4(2)、Meta 的 LLaMA 3 系列 (6) 等）的生成能力也一直在进步。这些大型语言模型现在能够生产类似人类的内容 (13)，并在控制实验环境中表现出类似人类的行为。例如，GPT-4 已在经典经济实验、心理语言学和社交心理学实验（如最后通牒游戏、花园路径句子、米尔格拉姆电击实验以及群体的智慧 (3)）中展示了类似于人类的行为。此外，基于 LLM 参数化的交互生成代理的模拟，这些模型通过定制化系统提示和之前互动的数据进行参数化，显示出与人类社会系统的相似性更广泛地存在 (21; 12; 18)。虽然类似人类的内容生成、心理学和社会行为并不意味着一套语言模型就是任何特定人群的数字孪生——它们可以被视为 LLM 能够模仿个体的语言生产特异性能力的迹象，前提是给定正确的数据和环境。

本文通过引入并分析美国国会议员的虚拟模型，为不断增长的能够充分捕捉物理系统相关特性的虚拟模型集合做出了贡献。特别是，我们提供了证据表明，一组具有访问包含来自国会议员官方账户推文的个性化数据库的语言模型，超越了通用的“类人”生成、行为和社会学，并合理地满足数字孪生的定义。

1 定义数字孪生

我们采用国家科学院对数字孪生的定义作为衡量一个人或物体虚拟模型是否为孪生的标准 (20):

数字孪生是一组虚拟信息构建，模仿

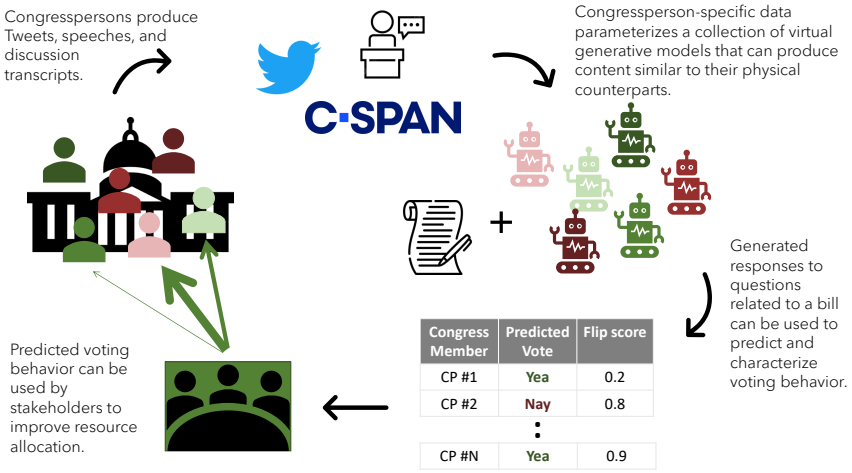

图 1: 一个包含国会议员数字孪生的系统的示例。

> 自然、工程或社会系统（或系统之系统）的结构、背景和行为，通过其物理孪生的数据进行动态更新，具有预测能力，并为实现价值的决策提供信息。虚拟与物理之间的双向互动是数字孪生的核心。

我们将这一定义分解为虚拟模型成为数字孪生所需的四个条件：A) 相关且动态的数据，B) 物理到虚拟的反馈，C) 有价值的推理能力，D) 虚拟到物理的反馈。

在一个国会议员群体的背景下，这些要求转化为证明存在一个与每位活跃国会议员相关的最新内容来源（针对 A），虚拟模型代表的每位国会议员产生的内容类似于他们的实体对应者（针对 B），对虚拟模型集合的推断可以预测实际国会议员的行为（针对 C），以及虚拟模型的输出可以影响利益相关者与国会议员互动的方式（针对 D）。

图 1 提供了一个信息流动的示例，其中一组国会议员生成数据（如推文、演讲和来自公共互动的记录等），特定国会议员的生成模型产生与其物理对应物相似的内容，这些模型的输出用于预测点名投票行为，并且这种预测的投票行为告知利益相关者如何最有效地分配资源。本文其余部分提供的证据表明，当前的数据集、语言模型以及数据处理方法足以合理地实现图 1 中所示的系统。

2 国会议员推特的数字孪生

2.1 相关和动态数据

政治人物的社交媒体习惯是政治和社会科学研究的重要资源。目前，大多数社交媒体数据被封存在主题或时间有限的数据集中。其中最广泛的数据集是由皮尤研究中心 (23) 整理的一系列 2015-2020 年间由美国政治人物生成的脸书和 X（原推特）数据。同样，哈佛大学的 Dataverse 项目维护着与第 115 届美国国会 (7) 相关的数据集，中国数据实验室则维护了近 83 万个国会议员关于中国的 X 帖子 (15)。

尽管这些数据集已经使各种主题的研究成为可能，但在构建国会议员的虚拟模型时，它们在主题上或时间上都是不够的。特别地，需要通过全面且最新的推文集合来充分捕捉每位国会议员所产生内容的复杂性。

为此，我们展示了 Nomic 国会数据库。Nomic 国会数据库是一个包含超过 300 万条 X 帖子和转发的新数据集，最早可追溯到 2011 年。该数据集包括所有美国国会议员官方账户发布的推文，这些议员在发布时正在任上，其中包括一些已被删除的帖子。该数据集还包括每条推文中议员的 X 用户名、党派归属、所代表的州以及推文发布时间。该数据集每天更新一次，在我们的语境下足够保持相关信息的相关性。因此，Nomic 国会数据库具有构建数字

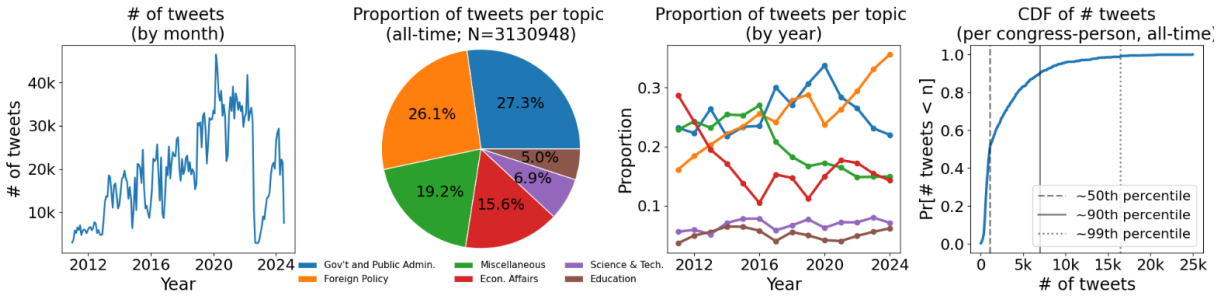

图 2: Nomic 国会 Twitter 数据集自 2024 年 10 月 10 日以来的高级特征。该数据集每天更新，可在 https://atlas.nomic.ai/data/hivemind/ 获取。

孪生所需的时空相关性。

我们注意到，由于存储了整个推文文本，数据集包括与提及和转发相关的信息。数据集不包含账户数据，如关注者或他们关注的人。我们在图 2 中展示了数据库的高级特征。左侧的图表显示了在职国会议员每月发布的推文数量或转发次数。请注意，推文数量的大规模减少（≈2022 年春季）对应于平台的私有化及其使用的一般犹豫。

我们还包括基于六个主题的简单话题分析：政府和公共服务、外交政策、经济事务、科学技术、教育和其他。这六个主题是通过结合众议院和参议院各自 20 个和 16 个常设委员会的主题而选定的。一旦确定了广泛的话题名称，我们就要求 ChatGPT 将 100,000 的推文归类到其中一个话题中。然后我们将推文嵌入到一个 768 维向量空间中，使用开源嵌入模型 (19)，并利用 100,000 个标注过的推文构建了一个线性分类器来对剩余的帖子进行分类。

图 2 中的中间两个图形显示了主题的分布以及主题相对分布随时间的变化。虽然我们的分析并未集中在主题相对分布的变化上，但我们注意到 2016 年被标记为“杂项”的推文数量减少——同时被标记为“对外政策”和“政府与公共管理”的推文数量增加——这恰好与特朗普当选总统的时间吻合，并有效地表明了数据集的内容随着国会议员行为的演变而变化。

最后，我们包括了每位国会议员推特数量的分布。大多数国会议员在整个任期内有 1000 条帖子或转发，而最活跃的 1%的议员则拥有超过 15,000 条帖子或转发。我们没有根据在职时间来标准化这些数量。[1]

2.2 物理到虚拟反馈

接下来，我们将使用国会数据库来验证一组生成模型是否能够生成与每位国会议员的真实推文相似的推文。特别是，我们将通过统计图灵测试 (13) 证明机器生成的推文的分布与国会议员生成的推文的分布彼此接近。统计图灵测试框架用一个统计量 τ 装饰了一个人类检测问题 (8)，该统计量代表了区分人类生成内容和机器生成内容的难度。$\tau = 0$ 表明人类生成内容和机器生成内容无法区分。在我们的案例中，每位国会议员都有一个对应的虚拟模型，因此，也有一个对应的 τ。

我们考虑三种基于 Meta 的 `LLaMa-3-8B-Instruct` (6) 越来越复杂的虚拟模型集合：i) 带有通用系统提示且无增强的基础模型 (−SP−RAG)，ii) 带有简单的国会议员特定系统提示且无增强的基础模型 (+SP−RAG)，iii) 带有简单的国会议员特定系统提示并带有增强的基础模型 (+SP+RAG)。通用系统提示是“您是一个乐于助人的助手。”国会议员特定的系统提示是“您是美国国会议员{name}”。对于模型 iii)，我们通过在由 `nominal-嵌入文本-v1.5` (19) 诱导的 768 维向量空间中与目标推文余弦相似度最高的国会议员发布的推文来增强提示。我们将此过程称为检索增强生成

[1] Nomic 国会数据库的交互式可视化集合可以在这里找到：https://atlas.nomic.ai/data/hivemind/

("RAG") (17)。表 1 显示了三个虚拟模型集合的系统提示和查询结构。

我们将每位国会议员的推文根据发布时间分为两类：一类是发布于 2023 年 1 月 1 日之前的，另一类是之后发布的。我们随意选择了这个日期，在此之前和之后都有一定数量的推文存在。生成模型可以访问 2023 年 1 月 1 日之前的推文作为潜在的提示增强内容。从 2023 年 1 月 1 日后随机抽取了 200 条推文来进行统计图灵测试：其中 100 条用作国会议员生成内容的例子，其余 100 条则用于虚拟模型生成的内容基础。特别是，虚拟模型生成的内容是通过给模型提供一条推文的前 20 个字符并要求其完成整个推文来产生的。在虚拟模型生成了这 100 条推文后，我们通过多维缩放（MDS）将生成和真实的推文集合嵌入到一个低维度欧几里得空间 (30) 中，该表示来源于规约嵌入文本-v1.5。然后使用费舍尔线性判别分析（FLD）来分类推文是国会议员生成的还是虚拟模型生成的。

我们在图 3 左侧的图表中展示了一个可检测性分析示例。我们为每个生成系统使用相同的 100 条推文开头。直方图是通过 MDS 学习后投影到 d 维空间中的 1-d FLD 投影，其中 d 是数据相关的 (34)。报告的 $\hat{\tau}$-不可检测性是 FLD 的经验风险。我们还在左侧图形右下角面板中包括了从代表性人物 Morgan Griffith 独立抽取的两组 100 条真实推文的可检测性分析作为对照。这个 $\hat{\tau}$ 为可检测性提供了一个实用的下限——尽管报告值可能小于它，但它捕捉到了分类这两组推文的基本难度。

我们报告了每个生成系统和右图 3 中的实际下限的 $\hat{\tau}$，这些数据是通过对每位国会议员进行 100 次随机采样得出的。我们删除了模型拒绝生成推文的所有实例，例如，"我不能创建诽谤或骚扰他人的内容。还有其他我可以帮忙的事情吗?" 对于那些我们删除了生成的推文的模型，我们也移除了相同数量的真实推文以保持分类问题的平衡。我们删除的最大推文数量是 95（总共 100 条）。被删除的推文中位数为 0。如 (+SP−RAG) 相比 (−SP−RAG)，以及 (+SP+RAG) 相比 (+SP−RAG) 所示，增加系统提示和查询中的议员特定信息量平均降低了 $\hat{\tau}$。

值得注意的是，100 个具有特定系统提示并可以访问 2023 年 1 月 1 日之前推文的虚拟模型中，有 16 个对应的 $\hat{\tau}$ 等于或小于它们对应的实际下限。这些虚拟模型中有 63 个比最大的实际下限更难以检测。

尽管可以考虑更复杂的生成系统，例如在检索分数 (33) 中添加时间相关性组件或从国会议员关注的账户中检索相关的推文，但我们的可检测性分析表明，带有国会议员特定系统提示和检索增强的生成模型集合足以捕捉大多数国会议员的特有推文复杂性。此外，这些模型可以很容易地与 Nomic 国会 Twitter 数据集一起更新。在本文的其余部分，我们将只考虑这一系列虚拟模型。

2.3 有价值的推理能力

我们接下来展示由虚拟国会议员生成的推文可以用来对其现实中的对应人物进行预测。我们关注国会在我国社会中扮演的一个最重要角色：立法。我们以第 117 届国会的议案 1319（"117-HR-1319"），即《2021 美国救援计划法》，作为示例法案来描述我们的实验设置。

我们首先通过提示 ChatGPT 使用公共信息（如法案的摘要（例如，"根据 S. Con. Res. 5 第二标题的规定，就协调问题提供支持。"）和由国会研究服务处撰写的法案简短总结（例如，"该法案提供了额外的支持以应对 COVID-19（即 2019 冠状病毒病）对经济、公共健康、州和地方政府、个人及企业的持续影响。"）来生成与该法案相关的 20 个问题。例如，针对 117-HR-1319 生成的两个相关问题是 "您是否支持该法案中提出的额外 COVID-19 救济措施?" 以及 "该法案是否为抗击 COVID-19 的公共卫生倡议提供了足够的支持?"

我们从每个国会议员投票前的时间段中检索与每个问题最相似的推文。否则，我们将使

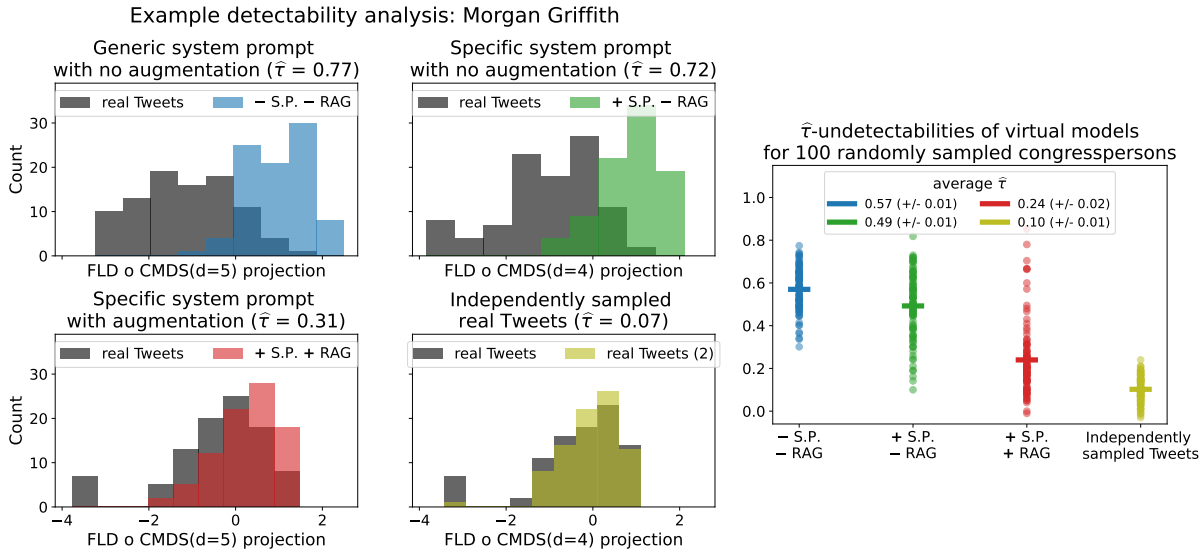

图 3: 不同生成系统产生美国国会议员摩根 · 格里菲斯推文的可检测性分析（左）以及不同系统对 100 个随机抽取的国会议员的可检测性分布（右）。通过 RAG 包含先前撰写的推文显著降低了可检测性。

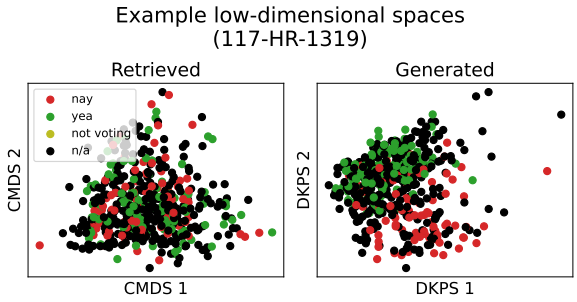

图 4: 由检索到的推文（左）和生成的推文对应于 117-HR-1319 的数据内核视角空间（右）诱导出的二维欧几里得空间。每个点代表一名国会议员。颜色对应国会议员在该议案上的投票情况。生成的推文几何结构比检索到的推文几何结构包含更多的与投票相关的信息。我们在图 5 中验证了这一观察结果。

用上述相同的检索过程，并构建结构类似于 (+SP+RAG) 的查询——唯一的区别在于查询结构中的第一句话，在这里我们用“撰写一条回复以下问题的推文：{question}”替换“完成……”。我们将适当格式化的查询提示每个虚拟国会议员二十次。

我们将每个响应嵌入到 `nominal-嵌入文本-v1.5` 中，并对副本的嵌入进行平均，以获得每位国会议员的 20×768 矩阵表示。成对距离矩阵（其条目等于这些矩阵表示之间的差的弗罗贝尼乌斯范数）的多维缩放产生每位国会议员的低维表示。这些数字国会议员的低维表示是相对于 20 个与法案相关的查询而言每位国会议员相对位置的一个总结，并且随着每个问题的问题数量和副本数量的增长，这种表示对于“真实”的国会议员表示是一致的 (1)。此外，使用这些低维表示进行模型级别的推断——例如预测数字国会议员的个人投票行为——是有原则的并且明显有效的 (11)。根据 (12)，我们将这个低维子空间称为*数据核视角空间*（DKPS）。

对应的 117-HR-1319 的二维 DKPS 显示在图 4 的右侧。每个点代表一位虚拟国会议员，并根据投票情况着色。所有在投票前发布过推文的国会议员都被包括在内。那些在第 117 届国会期间不是众议院成员的议员被分配了“n/a”的投票。我们在图 4 的左侧展示了使用每个问题检索到的推文来构建低维表示的类似代表。生成的推文的几何结构与投票行为有明显的关联，而检索到的推文的几何结构则显得相对无信息量。

我们通过比较使用两种表示方法训练的分类器在预测个别国会议员投票行为方面的表现来量化这一观察结果。特别是，我们考虑了使用检索到的推文或生成的推文所诱导出的议

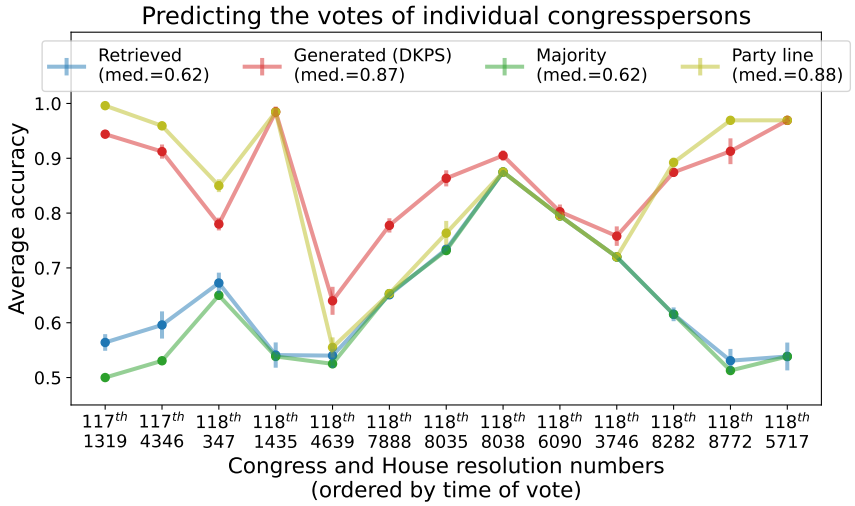

图 5: 四种不同方法在预测个别国会议员对各种立法的投票行为方面的平均性能。平均值是通过 10 折交叉验证计算得出的。误差棒代表一个标准误。这些立法按照众议院投票的时间顺序排列。对于“检索”和“生成”方法，我们报告的是最高性能的 k-近邻分类器在 $k \in \{1, 5, 9, 19, 49\}$ 中的平均准确率。不同的法案可能具有不同的最优 k。

员表示进行训练的 k-最近邻分类器。基于这些表示，我们在图 5 中报告了在 10 折交叉验证中对 13 项不同立法达到最高准确率的 $k \in \{1, 5, 9, 19, 49\}$ 分类器的平均准确度。用于交叉验证的训练和测试分割仅包含对法案投“赞成”或“反对”的众议员，或者当适用时，在参议院关联法案上投票“赞成”或“反对”的参议员。对于两种表示方法以及不同的法案，表现最好的 k 可能会有所不同。

我们还包括了在训练集中预测最流行投票的分类器（“多数派”）和在测试国会议员所属政党中最受欢迎投票中进行预测的分类器(“党线”）的表现。使用 DKPS 表示的最佳分类器的中位数性能为 0.87，而使用从检索到的推文中获得的表示的最佳分类器的中位数性能为 0.62。[2] 使用 DKPS 表示带来的改进在统计上是显著的——生成与检索在每项法案上的双侧 Wilcoxon 检验得到 p 值 < 0.001——并且在操作上也是显著的——使用从生成推文中获得的表示相比使用从检索到的推文中获得的表示，提供了平均 36.7%的改进。

[2]最高表现的 k-近邻分类器在交叉验证中的平均准确率通常是使用最优 k 的 k-近邻分类器时准确性的一个正偏估计。这种偏差影响了这两种方法。

最后，我们注意到使用 DKPS 表示的分类器的表现有时比仅仅根据党派预测更差。在某些情况下，例如第 117 届国会中的 HR 1319（《2021 年美国救援计划法案》），这可能是因为国会议员事先“知道”投票的结果，并且为了与党派领导保持一致而违背了自己的政策偏好进行投票。对于《2021 年美国救援计划法案》，该议案是民主党执政三联体（总统及众议院和参议院的控制权）的重要立法，使得共和党的反对在政策上功能有限，但从维护党派忠诚度的角度来看可能是有用的。因此，共和党可以宣传成功实施后的法案的好处，与其他党成员保持良好的一致性和对另一方敢于对抗的意愿进行竞选。

2.4 虚拟到物理反馈

在图 5 中分析的 13 项法案中，有 12 项在众议院通过。在众议院通过的法案通常会被送往参议院。有些被送往参议院的法案争议不大——在那 12 项通过众议院的法案中，只有 4 项在进入参议院后出现了非琐碎的动作 [3]。我们将分析的重点放在这四项法案上：117-HR-

[3] “非琐屑”的行动是指任何需要记录个别参议员投票情况的行动。

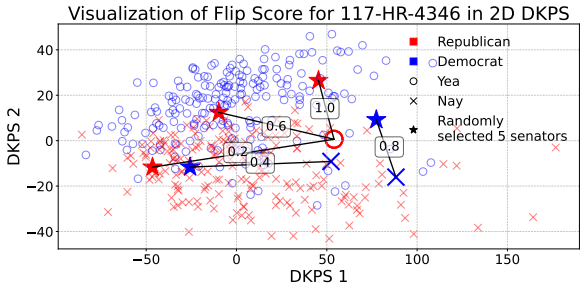

图 6: 所提议的“翻转分数”在 2-d DKPS 中对 117-HR-4346 进行可视化。每个标记代表一名国会议员。颜色对应于政党归属。投票为“赞成”的众议员用圆圈（○）表示，投票为“反对”的众议员则用“x”（×）表示。一颗星（⋆）代表一个未知投票情况的参议员。放大的 × 和 ○ 符号表示用于计算每位参议员翻转分数的众议院成员。每位参议员的翻转分数提供在其连接到最近同一党派跨党派线选民的连线上。靠近同一党派中跨党派线选民的参议员被赋予更高的翻转分数。

1319、117-HR-4346、118-HR-7888 和 118-HR-3746。

通常，在众议院通过一项法案与参议院投票之间会有一段相当长的时间。对于正在审议的四项法案，两院投票之间的天数分别为 10 天、364 天、5 天和 1 天。在此期间，利益相关者（活动家、选民、游说者等）有机会联系参议员的办公室，试图影响他们的投票方式。

没有其他信息的情况下，默认预测一名参议员投票的方式是按照党派立场进行。例如，如果该参议员所在政党的大多数众议院成员投票“赞成”，那么这位参议员也很可能会投票“赞成”。如图 5 所示，当按党派立场预测准确时，使用特定法案的 DKPS 分类可以获得相当的性能。重要的是，当按照党派立场预测不准确时，使用 DKPS 分类效果更好。因此，在一项法案已经收到投票——例如在众议院通过或未通过的情况下——可以利用 DKPS 的几何结构来量化一名国会议员跨党派投票（或“翻转”）的可能性。例如，如果一位未知投票情况的共和党参议员的 DKPS 表示形式接近于一个跨越了党派线的共和党众议员的 DKPS 表示形式，那么他们更有可能跨越党派立场。相反，如果这位共和党参议员的 DKPS 表示形式远离所有跨过党派线的共和党众议员，则他们不太可能跨越党派线。

我们引入了“翻转得分”来量化这一理念。对于给定的法案，我们将识别每个参议员在适当 DKPS 中的最近同党跨党派成员。如果参议员所在的政党中没有跨党派投票者，则他们被分配一个翻转得分为 0。当参议员所在政党内有跨党派投票者时，则该参议员的翻转得分与其到最近同党跨党派选民的距离（在 DKPS 中的距离）成反比。对于给定的参议员 S 并定义

$$\begin{aligned}\mathcal{H}(S) := \{&H : H \in \text{House},\\ &H \text{ and } S \text{ in same party},\\ &H \text{ voted across party lines}\}\end{aligned}$$

则，如果 $\mathcal{H}(S)$ 非空，

$$\text{flip score}(S) = \frac{1}{\min_{H\in\mathcal{H}(S)} \|X_H - X_S\|},$$

其中 $X_C \in \mathbb{R}^d$ 是国会议员 C 的 DKPS 表示。

图 6 展示了与 117-HR-4346 对应的 DKPS。该图包括所有在投票前至少发布过一条推文且投票为“赞成”或“反对”的众议院成员。它还包括五位参议员，并用线将他们连接到用于计算其翻转得分的众议院成员。

我们通过将其与给定分数下翻转的参议员的比例进行比较来验证所提出的翻转得分。为此，我们将某项法案内的翻转得分量化。得分为 0 的分配一个量化的翻转得分为 0。对于该法案而言，在第 0-20 百分位的翻转得分被分配一个量化的翻转得分为 0.2，处于第 20-40 百分位的翻转得分被分配一个量化的翻转得分为 0.4，依此类推。

图 7 展示了量化翻转得分与观察到的翻转比例之间的关系。观察到的翻转比例是在所有四个两院法案中计算得出的。通过卡 endall 的 τ 对翻转得分和观察到的翻转比例之间无序关联性的假设检验得到的 p 值，可以确定这种关系具有统计显著性（p 值 < 0.05）。此外，这种关系相当线性：线性拟合优度测量 R^2 为 > 0.8。

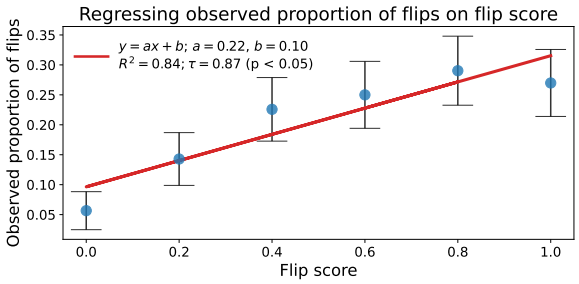

图 7: 参议员跨越党派界限的经验似然与其提议的翻转分数在四份两院法案上存在相关性。估计翻转的比例是通过首先将翻转分数量化为 6 个区间来计算的：一个用于零，其他五个基于 20%分位数（即 0-20%，20-40%等）去除零值后。误差棒表示一个标准误。红线是最优线性拟合。R^2 值表明存在强烈的线性关系，而对应于肯德尔的 τ 的 p 值则表明存在统计学意义上的等级关联。

因此，利益相关者可以使用翻转分数来优先考虑如何分配他们的沟通资源：在具有高翻转分数的参议员身上投入更多。如果利益相关者利用翻转分数来优先分配其沟通资源，则他们提供了一种机制，使得虚拟模型上的推断可以为物理双胞胎提供反馈并实现价值。

3 讨论与限制

我们已经提供了证据，证明一组配备有特定国会议员数据集的语言模型满足作为国会议员集合的数字孪生虚拟模型所需的四个条件。实际上，此处的结果展示了实现类似图 1 所示系统的可行性。尽管这些虚拟模型集合符合数字孪生的定义，但我们的工作仍有一些细节和限制需要讨论和进一步研究。

例如，我们的分析完全集中在 Twitter 数据上。Twitter 绝对不是国会议员与公众或彼此交流的唯一媒介。其他数据来源，如竞选演讲、发给选民的电子邮件、C-SPAN 文本记录、以前的投票记录等，应在声称一组虚拟模型是一个合适的数字孪生之前被包含并进行分析——尽管在虚拟模型保真度和可解释性之间很可能始终存在权衡。

沿着同样的思路，我们用来研究虚拟模型集合的数据不包括充分模拟国会议员之间或国会议员与公众之间互动所需的信息。因此，我们不对对话进行模拟或近似更复杂的行为，例如起草法案。如果要声称某些国会流程的数字孪生体，则可能需要一个能够充分捕捉这些行为的虚拟模型。

我们注意到之前的研究已经证明了使用关于法案的信息来预测国会议员投票情况的能力 (14; 22)。例如，(22) 使用一个分类器通过基于特定国会议员的新闻文章手动整理的知识库来预测投票行为。他们的分类器在第 106-109 届国会会期的所有法案上的平均准确率达到 0.9。我们的 DKPS 方法表现相当。重要的是，基于 DKPS 的方法使用了每个国会议员虚拟模型的输出，这些输出明显类似于实际国会议员的输出。

图 5 比较了检索到的推文的几何结构与生成的推文的几何结构在预测国会议员如何对特定法案投票方面的效用。生成的推文的几何结构比检索到的推文的几何结构更有助于预测国会议员将如何对特定法案进行投票。这种性能差距表明 `LLaMa-3-8B-Instruct` 能够有效利用包含少量与投票相关信息(根据准确性)的检索到的推文来生成更多与投票相关的信息。我们注意到，我们使用的基线模型由 Meta 于 2024 年中四月发布，并且我们考虑的大多数法案是在 2024 年中四月或之后进行投票的。因此，我们不期望该模型在给定低信号推文时生成高度相关内容的能力随着感兴趣的法案与基线模型训练时间点的距离越来越远而大幅恶化。

最后，我们论证了提出的翻转分数可以被利益相关者用于改善资源配置。如上所述，翻转分数需要现有的投票信息，例如来自众议院的信息。我们怀疑无监督的翻转分数类似物在优化资源分配时也将是有用的。例如，如果一个共和党的 DKPS 表示周围是民主党的 DKPS 表示，那么有理由相信这个共和党人可能会倾向于与民主党人一起投票。

总体而言，我们提供了实证和统计证据，证明一组能够访问包含国会议员官方账户推文的个性化数据库的语言模型超越了通用的“类

人”生成、行为和社会学定义，并合理满足数字孪生的定义。

致谢

我们想感谢 Henry Farrell、Hahrie Han、Ben Johnson、Connor Pollak 和 Jeremy Ratcliff 在本文手稿开发过程中提供的有益讨论和反馈。H.M. 得到了约翰霍普金斯大学 SNF Agora 学院的支持。

References

[1] Aranyak Acharyya, Michael W. Trosset, Carey E. Priebe, and Hayden S. Helm. 2024. Consistent estimation of generative model representations in the data kernel perspective space. *Preprint*, arXiv:2409.17308.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

[3] Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR.

[4] Jinsong Bao, Dongsheng Guo, Jie Li, and Jie Zhang. 2019. The modelling and operations for the digital twin in the context of manufacturing. *Enterprise Information Systems*, 13(4):534–556.

[5] Genevieve Coorey, Gemma A Figtree, David F Fletcher, Victoria J Snelson, Stephen Thomas Vernon, David Winlaw, Stuart M Grieve, Alistair McEwan, Jean Yee Hwa Yang, Pierre Qian, et al. 2022. The health digital twin to tackle cardiovascular disease—a review of an emerging interdisciplinary field. *NPJ digital medicine*, 5(1):126.

[6] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

[7] Tobias Eichinger, Felix Beierle, Sumsam Ullah Khan, Robin Middelanis, Veeraraghavan Sekar, and Sam Tabibzadeh. 2019. Tweets - US Senators of 115th Congress until Sept 2018.

[8] Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy. Association for Computational Linguistics.

[9] Amirashkan Haghshenas, Agus Hasan, Ottar Osen, and Egil Tennfjord Mikalsen. 2023. Predictive digital twin for offshore wind farms. *Energy Informatics*, 6(1):1.

[10] Albrecht Hänel, Thorben Schnellhardt, Eric Wenkler, Andreas Nestler, Alexander Brosius, Christian Corinth, Alexander Fay, and Steffen Ihlenfeldt. 2020. The development of a digital twin for machining processes for the application in aerospace industry. *Procedia Cirp*, 93:1399–1404.

[11] Hayden Helm, Aranyak Acharyya, Brandon Duderstadt, Youngser Park, and Carey E. Priebe. 2024. Embedding-based statistical inference on generative models. *Preprint*, arXiv:2410.01106.

[12] Hayden Helm, Brandon Duderstadt, Youngser Park, and Carey Priebe. 2024. Tracking the perspectives of interacting language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1508–1519, Miami, Florida, USA. Association for Computational Linguistics.

[13] Hayden Helm, Carey E. Priebe, and Weiwei Yang. 2023. A statistical turing test for generative models. *Preprint*, arXiv:2309.08913.

[14] Peter Kraft, Hirsh Jain, and Alexander M Rush. 2016. An embedding model for predicting roll-call votes. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 2066–2070.

[15] China Data Lab. Congress tweets on china by china data lab @ ucsd. https://chinadatalab.ucsd.edu/tweets/. Accessed: 2024-12-01.

[16] Jinfeng Liu, Xiaojian Wen, Honggen Zhou, Sushan Sheng, Peng Zhao, Xiaojun Liu, Chao Kang, and Yu Chen. 2022. Digital twin-enabled machining process modeling. *Advanced Engineering Informatics*, 54:101737.

[17] Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2020. Generation-augmented retrieval for open-domain question answering. *arXiv preprint arXiv:2009.08553*.

[18] Harvey McGuinness, Tianyu Wang, Carey E Priebe, and Hayden Helm. 2024. Investigating social alignment via mirroring in a system of interacting language models. *arXiv preprint arXiv:2412.06834*.

[19] Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic embed: Training a reproducible long context text embedder. *Preprint*, arXiv:2402.01613.

[20] National Academy of Engineering and National Academies of Sciences. 2024. *Foundational Research Gaps and Future Directions for Digital Twins*. The National Academies Press, Washington, DC.

[21] Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

[22] Pallavi Patil, Kriti Myer, Ronak Zala, Arpit Singh, Sheshera Mysore, Andrew McCallum, Adrian Benton, and Amanda Stent. 2019. Roll call vote prediction with knowledge augmented models. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 574–581.

[23] Pew Research. 2020. Congress soars to new heights on social media.

[24] Delf Rothe. 2024. When the world is an object: On the governmental promise of a digital twin earth. *International Political Sociology*, 18(3):olae022.

[25] Andrea Saltelli, Gerd Gigerenzer, Mike Hulme, Konstantinos V Katsikopoulos, Lieke A Melsen, Glen P Peters, Roger Pielke Jr, Simon Robertson, Andy Stirling, Massimo Tavoni, et al. 2024. Bring digital twins back to earth. *Wiley Interdisciplinary Reviews: Climate Change*, 15(6):e915.

[26] Gerhard Schrotter and Christian Hürzeler. 2020. The digital twin of the city of zurich for urban planning. *PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 88(1):99–112.

[27] Kaan Sel, Deen Osman, Fatemeh Zare, Sina Masoumi Shahrbabak, Laura Brattain, Jin-Oh Hahn, Omer T Inan, Ramakrishna Mukkamala, Jeffrey Palmer, David Paydarfar, et al. 2024. Building digital twins for cardiovascular health: From principles to clinical impact. *Journal of the American Heart Association*, 13(19):e031981.

[28] Florian Stadtmann, Adil Rasheed, Trond Kvamsdal, Kjetil André Johannessen, Omer San, Konstanze Kölle, John Olav Tande, Idar Barstad, Alexis Benhamou, Thomas Brathaug, et al. 2023. Digital twins in wind energy: Emerging technologies and industry-informed future directions. *IEEE Access*.

[29] Eric A Stahlberg, Mohamed Abdel-Rahman, Boris Aguilar, Alireza Asadpoure, Robert A Beckman, Lynn L Borkon, Jeffrey N Bryan, Colleen M Cebulla, Young Hwan Chang, Ansu Chatterjee, et al. 2022. Exploring approaches for predictive cancer patient digital twins: Opportunities for collaboration and innovation. *Frontiers in digital health*, 4:1007784.

[30] Warren S Torgerson. 1952. Multidimensional scaling: I. theory and method. *Psychometrika*, 17(4):401–419.

[31] Asaf Tzachor, Soheil Sabri, Catherine E Richards, Abbas Rajabifard, and Michele Acuto. 2022. Potential and limitations of digital twins to achieve the sustainable development goals. *Nature Sustainability*, 5(10):822–829.

[32] Chengyue Wu, Guillermo Lorenzo, David A Hormuth, Ernesto ABF Lima, Kalina P Slavkova, Julie C DiCarlo, John Virostko, Caleb M Phillips, Debra Patt, Caroline Chung, et al. 2022. Integrating mechanism-based modeling with biomedical imaging to build practical digital twins for clinical oncology. *Biophysics reviews*, 3(2).

[33] Zihan Zhang, Meng Fang, and Ling Chen. 2024. Retrievalqa: Assessing adaptive retrieval-augmented generation for short-form open-domain question answering. *Preprint*, arXiv:2402.16457.

[34] Mu Zhu and Ali Ghodsi. 2006. Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics & Data Analysis*, 51(2):918–930.

表 1: 系统提示和查询设计用于评估的三个生成系统。实验中真实推文的开始是前 20 个字符。(“SP” = 系统提示；“RAG” = 增强检索生成)

Model name	**System prompt**	**Query**
Generic system prompt with no augmentation (−SP −RAG)	You are a helpful assistant.	Complete the following Tweet: {start of real Tweet}. Respond with the full Tweet.
Specific system prompt with no augmentation (+SP −RAG)	You are U.S. congressperson {name}.	Complete the following Tweet: {start of real Tweet}. Respond with the full Tweet.
Specific system prompt with augmentation (+SP +RAG)	You are U.S. congressperson {name}.	Complete the following Tweet: {start of real Tweet}. Here is an example Tweet potentially related to the to-be-completed Tweet: “{retrieved Tweet}". Respond with the full Tweet.