贾鲁：来自可信来源的合法巴西大型语言模型

Roseval Malaquias Junior[1,2][0000−0002−6005−0515], Ramon Pires[2][0000−0002−0023−1971], Roseli A. F. Romero[1][0000−0001−9366−2780], and Rodrigo Nogueira[2][0000−0002−2600−6035]

[1] Computer Science Department, University of São Paulo, São Carlos, São Paulo, Brazil
roseval@usp.br, rafrance@icmc.usp.br

[2] Maritaca AI, Campinas, São Paulo, Brazil
{ramon, rodrigo}@maritaca.ai

摘要 与大型语言模型的预训练相关的高昂计算成本限制了其研究。为了解决这个问题，出现了两种策略：领域专业化和使用高质量数据进行预训练。为了探索这些策略，我们利用来自信誉良好的巴西法律来源的 19 亿个独特标记对 Mistral-7B 模型进行了专门化，并在法律知识和通用知识测试套件上进行了少量样本评估。我们的模型 Juru 通过实现法律基准上的性能改进展示了领域专业化的优点，即使预训练数据量减少也是如此。然而，这种通过持续预训练实现的领域专业化带来了与无关领域的遗忘增加相关的成本，这体现在葡萄牙语和英语的一般知识测试套件上的表现下降。本研究为不断增长的科学证据体系做出了贡献，证明了预训练数据的选择可以提高大型语言模型的性能，从而使这些模型以更低的成本进行探索。Juru 在 https://huggingface.co/roseval/Juru-7B 公开可用。

Keywords: 领域专业化 · 持续预训练 · 法律语言模型 · 少资源预训练

1 介绍

大型语言模型（LLMs）通常在来自互联网快照的大量通用数据上进行训练，例如 CommonCrawl[3,4,32] 和 C4[32,34]。缩放定律 [14] 支持这一范式，表明随着数据和可训练参数的增加，模型的能力也会增强。因此，通过这种方法开发的 LLMs 展示了跨多种语言和领域 [20] 执行任务的能力。

预训练所需的巨大计算资源对研究构成了挑战，限制了对其能力的探索。然而，通过应用持续预训练，在较小规模上观察到具有竞争力的结果是可能的。在文献中，从可靠的来源选择特定领域的数据 [8,9,22] 至关重要 [12,18,21]。

大型语言模型的广泛预训练促进了通用事实知识和语言理解。因此，继续在同一规模和范围内进行预训练可能不是提高语言模型性能的必要条件。通过使用来自可靠来源的专业领域数据，我们假设最初在通用数据上训练的语言模型可以通过少量额外数据提升其在特定领域的表现，尽管这可能会以牺牲其他领域的表现为代价。

为了检验这一假设，我们将 Mistral-7B[15] 模型专门用于巴西法律领域，使用了主要来自信誉良好的学术研究的 1.9 亿个独特标记。生成的模型 Juru 在涵盖巴西法律知识和通用知识领域的多项选择测试套件上进行了评估。我们的目标是评估在持续预训练期间仅纳入特定领域数据对模型性能的影响。

在整个提议的预训练过程中，我们的模型在法律测试套件中表现出增强的性能，而在通用知识测试套件中的性能则有所下降。特别地，在英语测试套件上观察到的下降幅度比葡萄牙语套件更为明显，这表明先前获得的知识在与目标专业化领域具有更大相似性时能更好地保留下来。这些结果为支持通过专业预训练提高通用模型在特定领域表现的有效性的科学证据体系做出了贡献。这种专业化可能通过减少与培训相关的计算成本来促进他们的研究。此外，生成的模型 Juru 公开发布，以支持通过持续预训练进行领域的专业化进一步研究。据我们所知，Juru 是首个为巴西法律领域预训练的大语言模型。

2 相关工作

大语言模型 (LLM) 的规模扩大，无论是参数数量还是训练数据的数量，都导致了在较小规模模型中出现了前所未有的能力 [33]。最近的进步使这些模型能够在需要文本理解、专业知识、数学推理、多语言能力、编程和图像解释等各种任务上达到人类级别的表现 [20]。凭借这样的结果，可以说这些模型展示了将这些多样化技能以一种连贯的方式整合在一起的能力。

语言和领域专业化已成为确保大语言模型在实际应用中带来切实利益的有效策略。最近的研究表明，在特定语言的语料库中继续预训练，特别是在葡萄牙语 [1,17,22] 中，可以显著提高下游任务的表现，而无需增加参数数量。

巴西的法律领域作为这种专业领域的相关示例出现，正如先前的研究强调了在法律语料库上预训练的语言模型的表现 [9,24]. 然而，这些模型只有编码器，缺乏生成文本的能力，这对于许多法律应用是必不可少的。

SaulLM-7B[8] 是首个预训练的大型语言模型，用于处理法律语言的细微差别并遵循指令以支持对话互动。从 Mistral-7B[15] 开始，该模型在 30 亿个英语法律数据集标记上进行了预训练。在一个后续的研究中，Colombo 等人 [7] 将这一方法扩展到了专家混合架构。作者持续在 5400 亿个英语法律文本标记上对 Mixtral-54B 和 Mixtral-141B 进行预训练，生成了 SaulLM-54B 和 SaulLM-141B。虽然这些更大的模型在大多数领域的任务中表现出更高的性能，但它们的增大尺寸与某些法律任务上的较低性能相关。

据我们所知，Juru 是首个为巴西法律领域预训练的大型语言模型。我们的实验表明，相较于其基础模型，Juru 在领域内任务上表现出更优性能。然而，在超出巴西法律领域的任务中，其表现有所下降。

3 方法论

在本节中，我们描述了从巴西法律领域公共来源收集和整理数据的过程，以及 Juru 模型的预训练，包括其基础模型、超参数和优化方法。

3.1 预训练数据

我们进行了葡萄牙语学术论文的网络爬取，重点关注巴西法律领域的具有教育价值的数据。

学术论文从各种可靠的来源中被爬取，这些来源记录了巴西高等教育机构和国家期刊的研究。每个平台都提供了过滤器，用于专门提取仅用葡萄牙语撰写的巴西论文，可供非商业性分发使用。此外，我们还爬取了 LexML[3] 数据库，重点关注了一部分巴西联邦法律。

总共从学术论文数据库中提取了 13,278 份文档。近 99% 的这些文档是便携式文档格式（PDF）。因此，我们使用 Marker[4] 来从这些文档中提取文本。

从 LexML 数据库中提取了 14,743 部联邦法律的文本。此外，还向预训练数据集添加了 Sakiyama 等人研究中的数据集 [27]，该数据集包含 32,535 份法律文件，分布在巴西最高联邦法院的裁决和判决之间。

提取文档后，它们被自动整理使用了 Rae 等人提出的过滤器 [25]，并针对葡萄牙语进行了调整。该过滤器会移除不太可能是主要自然语言的文档，如表格或长列表，因为这些可能会妨碍 LLM 的学习过程。

[3] https://www.lexml.gov.br

[4] https://github.com/VikParuchuri/marker

表 1. 字节配对编码（BPE）[29] 词元在预训练数据集上的分布，使用 Mistral-7B 分词器。

Dataset	Tokens
Academic papers	1,805,059,922
Federal regulations	60,402,071
Decisions and rulings	133,295,895
Total	1,998,757,888

考虑到我们在预训练中使用的 Mistral-7B 的分词器，我们收集了 1.9 亿个令牌，如表 1 所示。每个文档被划分为包含 4,096 个令牌的序列。

3.2 模型师傅

Juru 模型扩展了 Mistral-7B 模型的因果语言建模预训练，没有进行指令微调。我们采用了框架 t5x 和 seqio [26] 进行建议的预训练。

尽管与 Mistral-7B 相比，近期发布的类似规模的模型 [10,11] 表现更优，但我们选择了较早发布的 Mistral-7B。这一决定旨在降低数据污染在我们测试套件中的风险。

对于预训练超参数，我们使用了无因子分解的 AdaFactor 优化器 [30]，一阶动量为 0.9，二阶动量为 $1 - k^{-0.8}$，其中 k 表示当前训练步骤。此外，我们通过 lr^2 应用了动态权重衰减，当前学习率值为 lr，并且使用了全局范数裁剪 1.0。

除了与因果语言建模任务相关的损失外，我们还使用了辅助损失 $10^{-4} \log^2 \left(\sum_i e^{z_i}\right)$，其中 z 表示预测的 logits。此辅助损失有助于抑制预训练损失的值。学习率初始设置为 0 并在 250 个训练步骤的线性预热期间逐渐增加，直到达到 0.001。在此预热阶段之后，学习率保持不变。

该模型在 TPU v2-256 集群上进行了因果语言建模的预训练。每个批次包含 512 个序列，每个序列中有 4,096 个标记。训练进行了 3,800 步，相当于大约 4 个周期，总共处理了 79.6 亿个标记。整个预训练消耗了 3.35×10^{20} FLOPs 计算 [16]，耗时 30.61 小时完成。训练实现了模型浮点运算利用率（MFU）为 54.2%，不包括自注意力操作 [6]。

4 评估

评估 LLMs 的文本生成能力存在挑战，因为新兴技能如自然语言理解和领域专业知识具有主观性，并且它们将这些技能整合用于问题解决的能力也是如此。鉴于其广泛的能力，通过开放式提问进行评估可以洞察模型的知识表示及其生成类似人类响应的能力 [35]。

然而，对开放式问题的回答进行评估存在内在的困难。人工评估需要大量资源，要求在特定领域任务中具备专业知识，并且还需要接受评估方法论的培训以确保实验的稳健性。此外，仅依赖 SOTA 大语言模型进行评估可能因目标领域的特定词汇和知识而与人类评判者的关联度较低。因此，越来越多地强调使用标准化的选择题考试来评估这些模型，例如大学录取过程中使用的 [2,5,19,23,28]。这样的考试涵盖各种主题和技能，对于大语言模型的评估具有优势。

我们研究了通过领域专业化进行持续预训练如何影响知识学习和遗忘。因此，我们在三个标准化的选择题测试套件上评估了专业化的 Juru 模型：巴西法律知识，代表我们的专业化目标领域；葡萄牙通用知识，与目标领域使用相同的语言但不评估法律知识；以及英语通用知识，其分布更接近于 Mistral-7B 原始以英语为中心的训练数据。

对于所有测试套件，我们采用准确率作为主要评估指标。我们将同一年内进行的所有考试版本归为一个单一的评估组。在汇总多个考试时，我们报告平均准确率。最后，我们在评估中使用了三例少样本学习，因为该模型仅用于文本补全训练，并未经过任何指令微调。

4.1 巴西法律知识

为了评估 Juru 模型，我们使用了来自 2024 和 2023 的六次律师协会考试（OAB）的问题以及两次全国司法考试（ENAM）的 2024。OAB 考试是所有希望在巴西从事法律工作的法学毕业生必须参加的考试。它包括两个阶段；初始阶段包含 80 个选择题，而第二阶段则由 4 个论述题和程序文件的制作组成。ENAM 于 2023 年引入，旨在统一进入巴西司法系统的选拔过程。自实施以来，只有在过去两年内通过了 ENAM 考试的候选人有资格申请巴西的司法职位。此次考试包括一个阶段，包含 80 个选择题。我们仅包括了来自 2023 和 2024 的 OAB 以及 2024 的 ENAM 考试的选择题，总共包含 638 个问题，如表 2 所示。

表 2. 巴西法律知识测试套件的配置。

Benchmark	Domain	Exams	Task	Size
OAB-2023	Legal	3	Multiple-choice (4)	240
OAB-2024	Legal	3	Multiple-choice (4)	238
ENAM-2024	Legal	2	Multiple-choice (5)	160

4.2 葡萄牙通用知识

皮雷斯等人 [22] 在将一种以英语为中心的语言模型专门化为葡萄牙语后，在英语基准测试中观察到了性能下降。这些结果提出了以下问题：灾难性遗忘是否仅限于跨语言领域专门化，还是在一个单一语言内的特定领域进行专门化也会导致该语言的一般知识被遗忘?

为了解决这个问题，我们考察了将 Juru 模型专用于巴西法律领域对其在葡萄牙语通用知识任务中的表现的影响，使用了表 3 中列出的选择题测试，共计 2,123 个问题。

表 3. 葡萄牙通用知识测试套件的配置。

Benchmark	Domain	Exams	Task	Size
ENEM-2024	General	1	Multiple-choice (5)	179
BLUEX-2024	General	2	Multiple-choice (5)	172
CPNU-2024	General	7	Multiple-choice (5)	320
BNDES-2024	General	12	Multiple-choice (5)	420
REVALIDA-2024	Medical	2	Multiple-choice (4)	187
MREX-2024	Medical	6	Multiple-choice (4)	452
CFCEQ-2024	Accounting	12	Multiple-choice (5)	294
CFCES-2024	Accounting	2	Multiple-choice (4)	99

遵循 Almeida 等人的做法，[1]，我们忽略所有需要理解图像的问题。有些考试有描述图像内容的图注。在这种情况下，我们将它们放在图像本来会出现的确切位置。

4.3 英语通用知识

与经过广泛持续预训练的 Sabiá[22] 模型不同，由于其法律专业化数据集较小，Juru 模型的训练步数较少。Pires 等人 [22] 发现在葡萄牙语数据上进行扩展训练后，他们的模型失去了英语知识，但他们的实验涉及一个显著更长的训练方案和更多样化的数据集。这引发另一个问题：较短的专业领域预训练能否减轻或防止先前学习知识的灾难性遗忘?

为了研究这一点，我们在广泛用于评估英语中的通用知识和推理能力的 MMLU 基准测试套件 [13] 上评估了 Juru 模型 [15,20,31]。虽然 MMLU 包括广泛的科目，如 STEM、人文、法律和医学，但我们仅限于大学水平和高中水平的子集进行评估，共 4,369 个问题，详见表 4。

表 4. 英语通用知识测试套件的配置。

Benchmark	Domain	Exams	Task	Size
MMLU-College	General	6	Multiple-choice (4)	719
MMLU-High School	General	14	Multiple-choice (4)	3,650

5 结果

提出的测试套件用于在持续预训练期间跟踪模型在各个检查点上的性能。图 1 展示了准确性作为持续预训练过程中处理的标记数量的函数。每个纪元包含 1.9 亿个来自巴西法律文本的标记。0B 点对应于基础 Mistral-7B 模型，而 7.1B 标记处的检查点代表专门的 Juru 模型。

结果支持了研究的主要假设：当 LLM 进行领域专业化时，其在领域内任务上的表现会得到提升，但在领域外的一般知识任务上表现则会下降。通过对曲线的分析，我们观察到该模型在法律测试套件中的准确率稳步上升，在 71 亿个标记处达到了 49.2% 的峰值。相比之下，在一般知识测试套件上

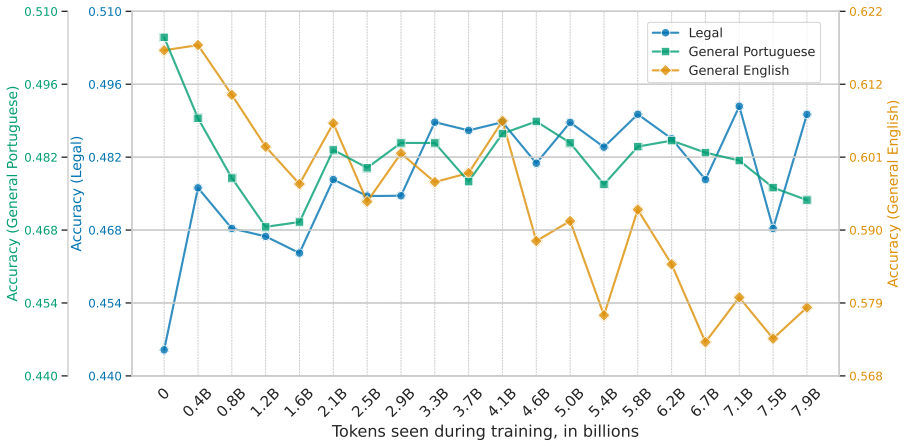

图 1. 在持续预训练过程中，在巴西法律知识、英语通用知识和葡萄牙语通用知识测试套件上的准确性。

的性能相对于基础模型有所下降，葡萄牙语知识测试套件的最终准确率为 48.1%，英语知识测试套件的最终准确率为 58.0%，这两个数据是在同一检查点获得的。

总体而言，Juru 的最佳版本在 71 亿个令牌时获得，对应于法律知识测试套件的最高准确性。所有后续分析均基于此检查点，我们将其称为 Juru。下一子部分将比较最终模型与基础 Mistral-7B 模型在法律和通用知识测试套件上的性能。

5.1 每个基准的结果

表 5 显示，Juru 在所有法律基准测试中均优于基础 Mistral-7B 模型，平均准确率总体提升约 4.7%。最显著的改进出现在 ENAM-2024 基准测试中，提升了 5.6%。这一结果证明了通过持续预训练实现领域专精在提高法律知识任务性能方面的有效性。

相比之下，表 6 表明在领域专业化后，在英语通用知识基准上的性能下降。虽然 Mistral-7B 在 MMLU 的大学和高中子集上都获得了更高的分数，但 Juru 全面表现不佳，导致整体平均准确率降低了 3.6%。

类似的趋势在表 7 中的葡萄牙通用知识基准测试中也被观察到。Juru 在 8 个单独的基准测试中有 5 个准确率较低，导致所有考试的平均准确率下降了 2.4%。

表 5. Mistral-7B 与 Juru 在巴西法律知识基准测试中的比较。

	准确率	
Benchmark	Mistral-7B	Juru
OAB-2023	48.3%	**54.5%**
OAB-2024	49.6%	**52.0%**
ENAM-2024	31.2%	**36.8%**
Mean (8 Exams)	44.5%	**49.2%**

表 6. Mistral-7B 与 Juru 在英语通用知识基准测试中的比较。

	准确性	
Benchmark	Mistral-7B	Juru
MMLU-College	**54.2%**	49.4%
MMLU-High School	**64.8%**	61.6%
Mean (20 Exams)	**61.6%**	58.0%

图 2 展示了 Juru 和 Mistral-7B 在 BNDES-2024 基准测试的 13 个知识领域中的比较。总体而言，Juru 在 13 次考试中的 8 次表现不如基础模型。最大的下降出现在 IT 支持方面，下降了大约 11.4%。相比之下，Juru 在 IT 网络安全、数据科学和会计方面的准确率与 Mistral-7B 相同，并在建筑学和经济学方面有所提升。尽管有这些特定的改进，整体趋势显示，在经过领域特定的持续预训练后，一般知识领域的表现出现了下降。

5.2 结果分析

结果展示了学习特定领域知识以实现 Juru 模型专业化的动态过程。在巴西法律领域的专业化使得与 Mistral-7B 相比，在巴西法律知识测试套件中准确率提高了 4.7%，尽管训练数据集相对较小，仅有 1.9 亿个唯一标记。

根据本研究的假设，专业化导致模型解决其专业领域之外任务的能力下降。尽管在法律和葡萄牙语通用测试套件中评估的任务之间共享语言，但仅在一个知识域中的专业化导致了其他领域区域性能下降。如图 2 所示，不同

表 7. Mistral-7B 与 Juru 在葡萄牙语通用知识基准测试中的比较。

	准确性	
Benchmark	Mistral-7B	Juru
ENEM-2024	**62.0%**	61.4%
BLUEX-2024	**60.7%**	58.4%
CPNU-2024	56.8%	56.8%
BNDES-2024	**59.1%**	56.0%
REVALIDA-2024	50.7%	**51.7%**
MREX-2024	**42.2%**	36.7%
CFCEQ-2024	**40.7%**	38.0%
CFCES-2024	40.3%	**41.3%**
Mean (44 Exams)	**50.5%**	48.1%

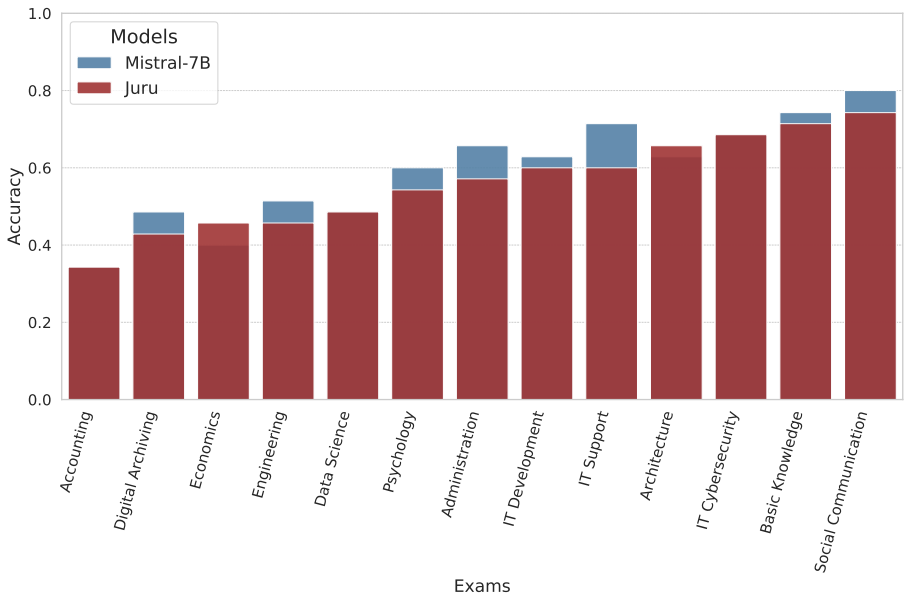

图 2. Juru 与 Mistral-7B 在 BNDES-2024 考试中的比较。

类型的知识领域与退化之间的明确相关性并不明显。此外，在比较葡萄牙语和英语通用知识测试套件上的性能下降时，Juru 模型在英语套件上表现出更大的遗忘程度。这种模式表明，知识遗忘受到评估领域与专业化领域之间相似性的影响：先前学习的知识与目标域越不相似，则遗忘的程度越大。

我们假设进一步预训练 Juru 模型可能会增加法律知识曲线与一般知识曲线之间的差异。因此，虽然在与法律相关任务中的表现可能提高，但在需要不同技能集的具体知识领域中，性能可能会显著下降。

6 限制

虽然取得了积极的成果，但在预训练数据集中存在可能的污染风险。这种风险源自于在 2023 的后半段通过网络抓取预训练数据，此时 OAB-2023 和 MMLU 考试已经发布。尽管该数据集主要来源于信誉良好的巴西学术和法律文件，并且不太可能包含测试题目，但间接污染的风险无法完全排除。为了解决这个问题，我们计划在未来评估中纳入 2025 年和 2026 年新发布的考试。

本文中报告的所有实验都基于巴西法律数据的一个特定子集。观察到的趋势是否适用于其他子领域或其他类型的非权威来源的数据仍然是一个开放性问题。探索更广泛的法律语料库，包括新闻文章、博客和司法评论，可能会揭示不同的专业化行为。

我们的工作面临与先前研究 [8,10,11] 中所述的在评估大语言模型时相同的有效性挑战。在巴西背景下，什么是法律熟练度的标准？是能够回忆立法、生成合法摘要还是预测司法决定的能力？我们主张这些技能是相互关联的。因此，在一项法律任务上的表现提高可能与在其他任务上有所改善相关联，因为它们往往依赖于重叠的能力。我们并不声称 Juru 在所有法律语言处理任务中都表现出色。相反，我们将使用巴西法律标准化考试的表现作为衡量专业化效果的实际替代指标，这些考试旨在评估人类的法律能力。我们在这些基准测试上的改进表明，领域专长可能转移到更广泛的法律任务上。

7 结论与未来工作

在这项研究中，我们介绍了 Juru，首个专注于巴西法律领域的大型语言模型，它在 1.9 亿个唯一标记上进行了预训练。尽管其规模较小且数据有限，我们在该模型解决巴西法律领域多项选择题的能力方面观察到了显著的

改进。我们的研究结果证实了专业领域专精可以成为一种有效策略，在较低计算成本下提升大型语言模型的表现。然而，这也带来了一个权衡：在与目标领域无关的任务上的表现会下降。

此外，葡萄牙语和英语通用测试套件之间的比较显示了英语基准的更明显下降。这表明遗忘的程度可能与领域相似性有关：先前获得的知识与专业领域的相似度越低，忘记的风险就越高。

对于未来的工作，我们的主要目标是将 Juru 模型的评估与超出模型知识截止点的新基准结合，以减轻数据污染的可能性。一个有前景的方向是在持续预训练期间调查聚合高相似性领域是否可以增强专业化性能而不会遗忘之前学习的知识。对这些动态的更深入了解可能会通过有针对性的专业化来推动计算需求更低的大规模语言模型的发展。我们还发布了本文中展示的所有 Juru 实验检查点，以进一步支持未来在这些主题上的研究，特别是在巴西法律领域的研究。

Acknowledgments. 本工作部分由巴西的 Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) 和 INCT 资助（CAPES #88887.136349/2017-00, CNPQ #465755/2014-3）。感谢 Google Cloud Platform 提供 TPU 赠款。

Disclosure of Interests. 作者声明与本文内容无关的利益冲突。

参考文献

1. Almeida, T.S., Abonizio, H., Nogueira, R., Pires, R.: Sabiá-2: a new generation of Portuguese large language models. arXiv preprint arXiv:2403.09887 (2024), https://arxiv.org/abs/2403.09887
2. Almeida, T.S., Laitz, T., Bonás, G.K., Nogueira, R.: BLUEX: a benchmark based on Brazilian leading universities entrance exams. In: Naldi, M.C., Bianchi, R.A.C. (eds.) Intelligent Systems. pp. 337–347. Springer Nature Switzerland (2023). https://doi.org/10.1007/978-3-031-45368-7_22
3. Brown, T., et al.: Language models are few-shot learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 1877–1901. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

4. Carlini, N., et al.: Extracting training data from large language models. In: 30th USENIX Security Symposium (USENIX Security 21). pp. 2633–2650. USENIX Association (2021), https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting
5. Cataneo Silveira, I., Deratani Mauá, D.: Advances in automatically solving the ENEM. In: 2018 7th Brazilian Conference on Intelligent Systems (BRACIS). pp. 43–48 (2018). https://doi.org/10.1109/BRACIS.2018.00016
6. Chowdhery, A., et al.: PaLM: scaling language modeling with pathways. Journal of Machine Learning Research **24**(240), 1–113 (2023), http://jmlr.org/papers/v24/22-1144.html
7. Colombo, P., et al.: SaulLM-54B & SaulLM-141B: scaling up domain adaptation for the legal domain. In: Globerson, A., et al. (eds.) Advances in Neural Information Processing Systems. vol. 37, pp. 129672–129695. Curran Associates, Inc. (2024), https://proceedings.neurips.cc/paper_files/paper/2024/file/ea3f85a33f9ba072058e3df233cf6cca-Paper-Conference.pdf
8. Colombo, P., et al.: SaulLM-7B: a pioneering large language model for law. arXiv preprint arXiv:2403.03883 (2024), https://arxiv.org/abs/2403.03883
9. Garcia, E., et al.: RoBERTaLexPT: a legal RoBERTa model pretrained with deduplication for Portuguese. In: Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1. pp. 374–383. Association for Computational Linguistics (2024), https://aclanthology.org/2024.propor-1.38/
10. Grattafiori, A., et al.: The LLaMa 3 herd of models. arXiv preprint arXiv:2407.21783 (2024), https://arxiv.org/abs/2407.21783
11. Groeneveld, D., et al.: OLMo: accelerating the science of language models. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 15789–15809. Association for Computational Linguistics (2024). https://doi.org/10.18653/v1/2024.acl-long.841
12. Gunasekar, S., et al.: Textbooks are all you need. arXiv preprint arXiv:2306.11644 (2023), https://arxiv.org/abs/2306.11644
13. Hendrycks, D., et al.: Measuring massive multitask language understanding. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=d7KBjmI3GmQ
14. Hoffmann, J., et al.: Training compute-optimal large language models. arXiv preprint arXiv:2203.15556 (2022), https://arxiv.org/abs/2203.15556
15. Jiang, A.Q., et al.: Mistral 7B. arXiv preprint arXiv:2310.06825 (2023), https://arxiv.org/abs/2310.06825

16. Kaplan, J., et al.: Scaling laws for neural language models. arXiv preprint arXiv:2001.08361 (2020), https://arxiv.org/abs/2001.08361
17. Larcher, C., Piau, M., Finardi, P., Gengo, P., Esposito, P., Caridá, V.: Cabrita: closing the gap for foreign languages. arXiv preprint arXiv:2308.11878 (2023), https://arxiv.org/abs/2308.11878
18. Li, Y., Bubeck, S., Eldan, R., Del Giorno, A., Gunasekar, S., Lee, Y.T.: Textbooks are all you need II: phi-1.5 technical report. arXiv preprint arXiv:2309.05463 (2023), https://arxiv.org/abs/2309.05463
19. Nunes, D., Primi, R., Pires, R., Lotufo, R., Nogueira, R.: Evaluating GPT-3.5 and GPT-4 models on Brazilian university admission exams. arXiv preprint arXiv:2303.17003 (2023), https://arxiv.org/abs/2303.17003
20. OpenAI: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023), https://arxiv.org/abs/2303.08774
21. Penedo, G., et al.: The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. arXiv preprint arXiv:2306.01116 (2023), https://arxiv.org/abs/2306.01116
22. Pires, R., Abonizio, H., Almeida, T.S., Nogueira, R.: Sabiá: Portuguese large language models. In: Naldi, M.C., Bianchi, R.A.C. (eds.) Intelligent Systems. pp. 226–240. Springer Nature Switzerland (2023). https://doi.org/10.1007/978-3-031-45392-2_15
23. Pires, R., Almeida, T.S., Abonizio, H., Nogueira, R.: Evaluating GPT-4's vision capabilities on Brazilian university admission exams. arXiv preprint arXiv:2311.14169 (2023), https://arxiv.org/abs/2311.14169
24. Polo, F., et al.: LegalNLP - natural language processing methods for the Brazilian legal language. In: Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional. pp. 763–774. SBC (2021). https://doi.org/10.5753/eniac.2021.18301
25. Rae, J.W., et al.: Scaling language models: methods, analysis & insights from training Gopher. arXiv preprint arXiv:2112.11446 (2021), https://arxiv.org/abs/2112.11446
26. Roberts, A., et al.: Scaling up models and data with t5x and seqio. Journal of Machine Learning Research **24**(377), 1–8 (2023), http://jmlr.org/papers/v24/23-0795.html
27. Sakiyama, K., Montanari, R., Malaquias Junior, R., Nogueira, R., Romero, R.A.F.: Exploring text decoding methods for Portuguese legal text generation. In: Naldi, M.C., Bianchi, R.A.C. (eds.) Intelligent Systems. pp. 63–77. Springer Nature Switzerland (2023). https://doi.org/10.1007/978-3-031-45368-7_5

28. Sayama, H.F., Araujo, A.V., Fernandes, E.R.: FaQuAD: reading comprehension dataset in the domain of Brazilian higher education. In: 2019 8th Brazilian Conference on Intelligent Systems (BRACIS). pp. 443–448 (2019). https://doi.org/10.1109/BRACIS.2019.00084

29. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: Erk, K., Smith, N.A. (eds.) Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1715–1725. Association for Computational Linguistics (2016). https://doi.org/10.18653/v1/P16-1162

30. Shazeer, N., Stern, M.: Adafactor: adaptive learning rates with sublinear memory cost. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 4596–4604. PMLR (2018), https://proceedings.mlr.press/v80/shazeer18a.html

31. Touvron, H., et al.: LLaMa 2: open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023), https://arxiv.org/abs/2307.09288

32. Touvron, H., et al.: LLaMa: open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023), https://arxiv.org/abs/2302.13971

33. Wei, J., et al.: Emergent abilities of large language models. Transactions on Machine Learning Research (2022), https://openreview.net/forum?id=yzkSU5zdwD

34. Xue, L., et al.: mT5: a massively multilingual pre-trained text-to-text transformer. In: Toutanova, K., et al. (eds.) Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 483–498. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.naacl-main.41

35. Zheng, L., et al.: Judging LLM-as-a-judge with MT-Bench and chatbot arena. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 46595–46623. Curran Associates, Inc. (2023), https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf