这太耗时了——探究时间作为大语言模型能耗代理指标的可行性

Lars Krupp嵌入式智能 *DFKI* 与 *RPTU*德国凯撒斯劳滕
lars.krupp@dfki.de

Daniel Geißler嵌入式智能 *DFKI* 与 *RPTU*德国凯撒斯劳滕
daniel.geissler@dfki.de

Francisco M. Calatrava-Nicolas应用自主传感器系统中心（*AASS*）¨厄勒布鲁大学¨瑞典厄勒布鲁
francisco.calatrava-nicolas@oru.se

Vishal Banwari嵌入式智能 *DFKI* 与 *RPTU*德国凯撒斯劳滕
vi
shal.banwari@dfki.de

Paul Lukowicz嵌入式智能 *DFKI* 与 *RPTU*德国凯撒斯劳滕
paul.lukowicz@dfki.de

Jakob Karolus嵌入式智能 *DFKI* 与 *RPTU*德国凯泽斯劳滕
jakob.karolus@dfki.de

arXiv:2603.15699v1 [cs.PF] 16 Mar 2026

摘要—**大语言模型（LLM）的能耗及其对环境稳定性和资源使用造成的不利影响，正日益引发关注。然而，用户对这些能源成本仍知之甚少，尤其是在通过 API 访问模型时——API 如同一个黑箱，所有信息都取决于服务提供商选择披露的内容。在本研究中，我们探讨将推理时间测量作为代理指标，以估算基于 API 的大语言模型所产生的相关能源成本。我们通过将估算结果与本地托管的同类模型的实际能耗测量值进行比较，为这一方法提供依据。研究结果表明，时间测量有助于推断基于 API 的大语言模型所使用的图形处理器型号，从而为能源成本估算提供依据。本研究旨在为理解基于 API 的大语言模型所产生的相关能源成本提供方法，尤其是帮助最终用户了解这些成本。**

索引词—**大语言模型、能耗、能源估算、可持续性**

I. 引言

深度学习领域，尤其是大语言模型（LLM）的最新进展，改变了人工智能在众多领域的能力范围，从自然语言理解到医疗保健、机器人学等其他领域皆是如此 [1], [2]。社会正在经历重大变化，使大语言模型远远超越研究环境，走进日常生活 [3]。具有全国代表性的大型调查显示，消费者采用生成式人工智能工具来完成日常任务、获取信息以及支持工作和学习的比例迅速增长 [4]–[6]。近年来，提出的模型数量大幅增加 [7], ，进一步加速了生成式人工智能在社会各领域的普及。

这一技术发展得益于日益庞大的大规模架构、海量数据语料库以及高性能计算（HPC）基础设施。由此，许多公司和研究机构加入了竞争，竞相开发更强大的模型，力求突破当前最先进水平的边界。然而，这种发展也带来了高昂的计算成本，主要来自模型的训练和推理阶段。由图形处理器级功耗和系统级电力消耗所带来的能源需求 [8], [9], 已达到新的高度。一些公司，例如谷歌，甚至正在建设专用核反应堆，以维持其人工智能基础设施 [10]。除了技术和经济方面的挑战之外，不断攀升的能源消耗所带来的环境影响也日益令人担忧。

令人遗憾的是，公司政策对于能耗信息披露的规定往往含糊不清，并通过不起眼的网页用户界面掩盖真实成本，ChatGPT 即为一例 [11]。闭源模型以及关于数据中心基础设施的有限信息，使得可靠的能源估算无从实现 [7], [8]。无法直接获取模型权重及其推理配置，也进一步阻碍研究人员开展透明测量，以评估大语言模型的能源足迹，并最终衡量基于 API 的大语言模型交互的效率。

然而，仍有一项指标可以直接测量：任务完成时间，也就是模型处理单个词元所需的时间。因此，在估算基于 API 的大语言模型的总能源消耗时，计算时间可能很有价值。不过，所使用的 GPU、模型规模与架构、负载因子以及许多其他因素仍会发挥关键作用，并可能显著改变计算时间。通过采用预设选项的本地部署模型来验证结果，对于巩固基于时间的估算至关重要。

本文探讨了推理完成时间作为可靠代理指标，用于估算大语言模型（LLM）能耗的潜力。为此，我们提出了一项基准测试，旨在受控实验条件下分析本地部署模型和基于 API 的模型中推理时间与能源使用量之间的关系。我们重点研究 Mistral 模型 [12], ，因为它们同时提供开放权重版本和可通过 API 访问的版本。本文旨在评估推理时间能否作为估算基于 API 的 Mistral 模型能耗的有效代理指标

，并通过在具有代表性的条件下本地测得的计算时间，整合基于时间的估算结果。我们的代码和评估结果可在 GitHub[1] 上获取。

尽管我们的工作并非首个提出基于 API 的能耗估算的方法，参见 [13],但我们通过在具有代表性的计算基础设施上运行相同模型并进行本地测量，为基于 API 的能耗估算提供了依据。我们的结果表明，基于 API 的大语言模型的推理时间测量，结合本地模型的对应测量，可以帮助我们推断外部服务器所使用的图形处理器配置。我们的工作证明，基于 API 的大语言模型与其本地对应模型在计算时间乃至能耗方面存在直接对应关系。

II. 相关工作

自注意力机制问世以来， [14],ChatGPT 等大语言模型日益受到关注，促使 OpenAI [11],Llama [15], DeepSeek [16],通义千问 [17], 和 Mistral [18]等不同模型系列之间的竞争日趋激烈。然而，除了在不同基准测试上的竞争，开源模型与专有模型也一直是人们讨论的话题。尽管专有模型在基准测试中往往表现更优， [19], 但与开源大语言模型相比，其内部运作机制通常不透明。尤其是关于专有模型能耗的信息，若有披露也往往十分有限且不可靠。

大语言模型的能源足迹涵盖训练和推理两个阶段，是人工智能领域最紧迫的可持续发展挑战之一。据估计，LLaMA-65B 等近期大规模系统在训练期间的能耗约为 5×10^5kWh [9], [20]。尽管现代数据中心能效和人工智能优化硬件取得了显著进步，但如果使用化石能源供电，这类运行仍可能排放数十至数百吨 CO_2 [20], [21]。模型稀疏化（如混合专家）、高效加速器架构以及低碳数据中心等方面的进展，可将能源强度降低一个至三个数量级 [20], [22]。

对于推理，尽管单次大语言模型查询可能仅消耗 0.3 至 1Wh，但每天数百万次请求的累计需求到 2026 [23], [24]时可能达到每年拍瓦时的规模。GPU 闲置时的功率消耗、吞吐量限制以及能耗非比例增长造成的低效，会进一步加剧排放；而量化、批处理和动态调度等技术则可减轻这些影响 [23], [25]。

准确评估大语言模型的能源足迹仍然颇具挑战。近期推出的 Carbon- Tracker [26] 和 CodeCarbon [27]等工具，旨在通过整合硬件层面的能耗数据与区域电网碳强度，实时监测功耗和 CO_2 排放。

[1]https://github.com/DFKIEI/LLM-Energy

近期研究开始量化深度学习系统在训练和评估阶段产生的能源与环境成本。在训练方面，Geißler 等人 [28]分析了超参数对神经网络能耗的影响，发现即使准确率相近，不理想的超参数选择也会增加能耗。随着模型规模不断扩大，这些问题也从传统深度网络延伸到了大语言模型（LLM）。Fernandez 等人 [29] 对大语言模型推理的能耗影响进行了全面研究，考察了模型架构、解码策略和软件框架如何影响功耗。Jegham 等人 [13] 进一步提出基于时间的估算方法，用于评估 API 服务模型的能耗，但缺乏实际测量数据的验证。在此，Krupp 等人 [30] 展示了一种对本地运行的网络智能体进行能耗基准测试的方法，并揭示了估算使用基于 API 的大语言模型的网络智能体能耗所面临的复杂性。

III. 方法论

基于这些研究成果，我们提出了一种方法论：利用本地执行的模型，为基于 API 的系统进行能源估算提供依据，并将推理计算时间作为两种设置下能耗的共同因素。我们的核心 **假设**是，当计算硬件以最佳利用率或接近最佳利用率运行时，其功率消耗基本保持不变，因此可以根据执行时间估算总能耗。然而，由于基于 API 的模型的具体硬件规格未公开，我们还在不同 GPU 架构上对模型进行本地基准测试，以建立能源估算值与 GPU 集群之间合理的映射关系。随后，我们将这些结果与从基于 API 的模型测得的推理完成时间进行比较。

我们的基准测试由合成提示组成（使用 llama3-70b 生成），这些提示来自四种输入输出配置（涵盖技术、创意、教育和面向商业的任务，并注入相关主题），其序列长度各不相同（短= 2,048 个词元，长 = 8,192 个词元）。教育类提示示例可以是：“请详细解释高等数学”。

我们根据大语言模型是否同时提供开放权重版本和基于 API 的版本，以及模型规模来进行选择。这些标准使我们能够通过在本地和通过 API 测试模型，更可靠地评估基于时间的代理能耗估算方法的准确性，同时确保模型足够轻量，能够在本地测试时部署在单个图形处理器上。具体而言，我们使用了 *Mistral-7B-Instruct-v0.3*（本地版），其对应的基于 API 的版本为*Open-Mistral-7B* ，以及*Mistral-NeMo-Instruct-2407* （本地版），其对应的基于 API 的版本为*Open-Mistral-NeMo*。

A. 实验方案

为确保结果可复现且一致，所有测试均按照结构化实验方案进行。

这适用于本地执行的运行，也适用于通过基于 API 的模型发起的运行。每次运行都会处理来自我们基准测试的固定序列，其中包含 100 条合成提示。这个确定性的顺序消除了抽样方差，确保所有模型和硬件平台接收完全相同的工作负载。

执行过程分为多个结构化轮次，每种配置都会重复执行，以确保统计结果可靠。对于每种配置，都会独立运行完整的基准测试十次，从而计算所有运行时间和能耗指标的均值与标准差。具体而言，对于基于 API 的模型，我们会在一天中的不同时段安排运行，以捕捉高负载时段造成的波动。对于每次本地运行，所有提示词都会以每批八条的方式并行处理，以确保部署时图形处理器利用率充足。我们会尽量减少查询之间的空闲周期，以确保每令牌能耗具有一致的可比性。

B. 本地硬件架构选择

为评估架构对运行时间和能效的影响，我们在六款英伟达 GPU 上开展本地实验，这些 GPU 代表了两个世代（Ampere、Hopper）和两种外形规格（SXM、PCIe）：A100-40GB、A100-80GB、A100-PCI、H100、H100-PCI 和 H200。如表 I 所示，这些设备在热设计功耗（TDP）和最佳持续负载范围方面差异显著。

比较重点关注服务器级 SXM 模块与 PCIe 版本（A100-PCI、H100-PCI）之间的差异。服务器版本受益于更高的 TDP 上限、外部散热以及高带宽 NVLink 互连；而 PCIe 版本的功耗较低，且存在更大的辅助损耗，例如电源转换和散热，这些损耗由芯片承担。

公开资料并未披露为 Mistral API 提供服务的确切 GPU 集群。不过，Mistral 表示其使用 H100 GPU 进行训练 [31], , 而 MLPerf v4.1 显示，H100/H200 级加速器在数据中心大语言模型推理中占据主导地位 [32]。由于 Mistral 模型通过主要云平台提供 [33], , 托管于 API 的部署最有可能使用 H100/H200 级 GPU。

C. 本地大语言模型初始化

评估的模型 Mistral-7B-v0.3 和 Mistral-NeMo-Instruct-2407（12B）在 Ampere 和 Hopper GPU 上以 FP16 精度运行，并启用 TF32 加速；使用 device-map="auto" 以均衡分配内存。采用 KV 缓存和左填充来优化批次一致性并减少冗余计算。

为确保结果可比，所有运行中的温度均固定为 0.7，涵盖本地模型和基于 API 的模型，以模拟真实使用场景。Python、NumPy、PyTorch 和 CUDA 的随机种子（42）保持同步。提示词直接传递给模型，不包含系统指令，从而确保运行时间和能耗的差异完全源自模型架构、精度和解码配置。

D. 本地能耗跟踪

能耗由 CarbonTracker [26],记录，该工具通过英伟达系统管理接口（SMI）采样图形处理器功耗，并整合整块电路板的总能耗。相比之下，热设计功耗（TDP）仅反映芯片的功耗范围，可作为理论下界。两者之间的差异表明辅助组件额外消耗的功率。

这种差异在 PCIe GPU 上最为明显：其电力输送效率较低、转换损耗较高，导致 CarbonTracker 的数值超过热设计功耗估算值；而服务器级 GPU 的两项数值则更为接近。对比这两项指标，可以看出芯片级规格会低估实际能耗，尤其是在 PCIe 配置下。

E. 基于 *API* 的能耗估算

按照实验方案，我们在本地开放权重部署 和基于 *API* 的部署上分别进行了$N = 10$ 次独立执行。对于每次本地运行$i \in \{1, 2, \ldots, N\}$，我们跟踪了代码执行期间的能耗。由此得到的能耗记为 E_i^{loc}，每次运行的实际运行时长记为 T_i^{loc}。因此，本地运行的平均功耗 可按以下方式计算：$P_{\text{loc}_i} = \frac{E_i^{\text{loc}}}{T_i^{\text{loc}}}$，对所有运行取平均后得到 P_{loc}。

基于这一项本地功耗计算，我们可以预测在可比计算配置下运行时，等效的基于 API 的模型的能耗：

$$\widehat{E}^{\text{api}} = P^{\text{loc}} \cdot \bar{T}^{\text{api}}, \tag{1}$$

其中，$\bar{T}^{\text{api}}$ 表示所有 API 运行的平均 API 执行时间。

为便于比较不同运行和模型[2]，我们按词元数量对结果进行归一化。因此，对于每次运行，我们都将能耗 E_i 和时间 T_i 除以该次运行的词元数量 $N_{tokens}^{(i)}$。由此得到平均每词元能耗 E_{token} 和平均每词元推理时间 T_{token}。

IV. 结果

本节评估 Mistral 系列大语言模型的能耗，这些模型既在本地运行，也通过基于 API 的服务运行。我们首先分析在这两种场景（本地与 API）下完成基准测试所需的计算时间，因为这是后续能源估算的基础。接下来，我们分析能耗，并报告总能耗和每词元能耗（见第 III-E 节）。最后，我们通过将本地配置下的能耗预测与根据各图形处理器的热设计功耗（TDP）参数得出的估算值进行比较，验证这些预测。

A. 计算时间基准测试

表 II 展示了在我们的基准测试中，针对两种场景（本地和 API）各运行十次的运行时间。本地执行与基于 API 的执行之间的比较凸显出计算效率存在显著差异。免费和付费 API 运行基准测试的速度都快于任何本地图形处理器配置。与性能最佳的本地配置相比，

[2]不同模型的输出词元数量并不一致。

表 I 热设计功耗和最佳工作负载范围

GPU	VRAM	热设计功耗 (TDP) [W]	最佳 TDP 负载范围 [W]
A100-40GB	40 GB	400	340–380
A100-80GB	80 GB	400	340–380
A100-PCI	40 GB	250	213–238
H100	80 GB	700	595–665
H100-PCI	94 GB	400	340–380
H200	140 GB	700	595–665

表 II 基准测试计算时间（秒）的均值和标准差表。GPU 按发布日期的先后顺序排列。

模型	Type	GPU	均值 ± 标准差
Mistral-7B	本地	A100-40GB	2447. ± 8.85
		A100-80GB	2308. ± 8.36
		A100-PCI	2525. ± 14.7
		H100	1662. ± 22.5
		H100-PCI	1820. ± 9.24
		H200	1559. ± 4.72
	免费 API	-	718. ± 43.8
	付费 API	-	727. ± 28.8
Mistral-NeMo	本地	A100-40GB	3865. ± 8.13
		A100-80GB	3612. ± 14.0
		A100-PCI	4003. ± 16.4
		H100	2140. ± 23.1
		H100-PCI	2353. ± 11.9
		H200	2018年. ± 8.17
	免费 API	-	1260. ± 161.
	付费 API	-	1159. ± 31.2

API 执行使 Mistral-7B 的计算时间缩短约 54%，使 Mistral-NeMo 的计算时间缩短约 43%，这表明不同设置（例如系统提示词或并行化）会带来差异。免费 API 与付费 API 之间的差异很小（仅几秒，且在一个标准差范围内），这表明两者依赖的后端资源相似，差异可能源于服务器负载或排队效应。相比之下，本地执行的运行时间更长，但更可预测（标准差更小），其性能取决于底层硬件。在本地配置中，呈现出明显的性能层级：较新的图形处理器（H100 和 H200）始终比较旧的 Ampere 系列模型更快地完成基准测试。

此外，我们还使用¯每个令牌的平均时间进行了细粒度比较。如图 1 所示，对于本地运行，不同配置的 T_{token} 按 GPU 代际聚类。对于这两种大语言模型，我们观察到两个本地集群（见图 1），即 A 集群和 H 集群。Mistral-NeMo 的基于 API 的 T_{token} 接近 H 集群，但 Mistral-7B 并非如此，其运行时间略长于 H 集群 GPU。第 IV-C 节从能耗角度阐释了这些结果。

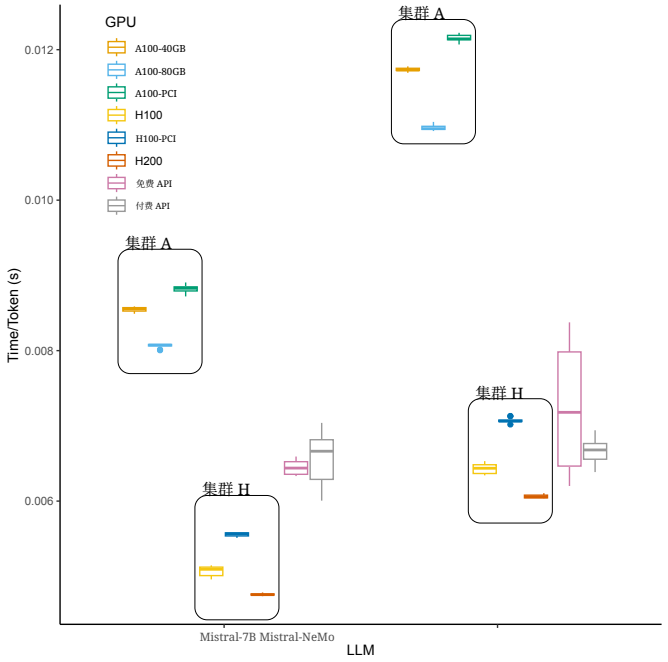

图 1. 箱线图显示，在不同本地 GPU 和 API 执行环境下，对两个模型运行相同基准测试时，每个词元的计算时间（以秒为单位） $\bar{T}_{token}$ 。集群 A 包含 A100 GPU，而集群 H 包含 H100 和 H200 GPU。

B. 将时间转换为能量

本节介绍与 API 模型能源估算中时间的用途相关的结果（参见第 III-E 节）。此外，我们还将该值与根据每个 GPU 的热设计功耗 (TDP) 计算的能耗进行比较。我们采用 0.9 × TDP 作为平均持续功率，以此作为保守估算，其中未计入冷却成本和热辐射造成的能量损失。根据实证得出的高吞吐推理实际能耗估算值介于 85% 和 95% 之间，参见 [8], [31], [34]。基于这些发现，表 I 报告了本分析采用的最佳 TDP 负载值。

将这些信息应用于计算得出的运行时间后，我们可以得出能耗的下界。如表III所示，热设计功耗与本地测量所得能耗之间的差异可能很大，具体取决于 GPU 架构（SXM 或 PCIe），详见第 III-B 节。

C. 估算基于 *API* 的大语言模型的能耗

¯通过对比我们的 GPU 配置测得的每词元时间 T_{token} 与基于 API 的模型的结果，我们可以推断出适用于给定大语言模型的可能 GPU 配置。我们¯确实要指出，尽管 T_{token} 可能具有可比性，但我们无法根据证据断定基于 API 的模型背后的实际 GPU 配置。因此，这种比较是基于有关计算集群的报告，以及我们通过推理时间测量得出的佐证性发现。

表III 平均持续功率与能耗：通过热设计功耗估算与本地测量（Carbon Tracker）的比较。各超大规模群组的每词元能耗及总能耗。

模型	GPU	令牌总数	热设计功耗（计算值）		Carbon Tracker（测量值）		差异（占本地的百分比）
			每个词元（毫瓦时）	总计（瓦时）	每个词元（毫瓦时）	总计（瓦时）	平均值 [%]
Mistral-7B	A100-40GB	286219	0.86 ± 0.003	245. ± 0.885	1.06 ± 0.023	303. ± 6.54	19
	A100-80GB	286219	0.81 ± 0.003	231. ± 0.836	1.10 ± 0.024	314. ± 6.75	26
	A100-PCI	286219	0.55 ± 0.003	158. ± 0.921	0.949 ± 0.007	272. ± 1.95	42
	H100	327674	0.88 ± 0.012	291. ± 3.94	1.07 ± 0.016	350. ± 5.37	16
	H100-PCI	327674	0.56 ± 0.003	182. ± 0.924	0.904 ± 0.009	296. ± 2.89	39
	H200	327674	0.83 ± 0.003	273. ± 0.827	1.04 ± 0.011	340. ± 3.54	20
Mistral-NeMo	A100-40GB	329345	1.15 ± 0.002	387. ± 0.813	1.47 ± 0.027	485. ± 8.75	20
	A100-80GB	329345	1.10 ± 0.004	361. ± 1.40	1.54 ± 0.015	507. ± 4.83	29
	A100-PCI	329345	0.75 ± 0.003	250. ± 1.02	1.32 ± 0.007	435. ± 2.23	43
	H100	332860	1.13 ± 0.012	375. ± 4.04	1.37 ± 0.019	456. ± 6.28	18
	H100-PCI	332860	0.71 ± 0.004	235. ± 1.19	1.17 ± 0.013	388. ± 4.25	39
	H200	332860	1.06 ± 0.004	353. ± 1.43	1.33 ± 0.021	444. ± 7.00	20

表 IV：Mistral-7B 和 Mistral-NeMo 的 API 版本（免费和付费）实测与计算总能耗成本对比。

模型	API 成本		令牌数量	H100-PCI 总能耗（瓦时）
Mistral-7B	free	M	111272. ± 6351。	100.60 ± 6.00
Mistral-7B	free	C	111272. ± 6351.	62.31 ± 3.56
Mistral-7B	paid	M	110882. ± 5922.	100.23 ± 5.60
Mistral-7B	paid	C	110882. ± 5922.	62.09 ± 3.32
Mistral-NeMo	free	M	173743 ± 4967.	203.28 ± 5.81
Mistral-NeMo	free	C	173743 ± 4967。	123.36 ± 3.53
Mistral-NeMo	paid	M	174496. ± 5815。	204.16 ± 6.80
Mistral-NeMo	paid	C	174496. ± 5815.	123.89 ± 4.13

M = 实测；C = 计算。

GPU 配置。同样，图 1 表明，H100-PCI 是适用于 Mistral 两种模型的良好代表性配置。

最后，我们使用付费和免费 API 计算得出的平均令牌数，以及热设计功耗计算和本地测量两种方法所得的每词元能耗结果，按照公式（1）计算了基于 API 的模型的平均能耗估计值 $\widehat{E}^{\text{api}}$ 。Mistral-7B 和 Mistral-NeMo 的结果如表 IV 所示，展示了不同的能耗。请注意，热设计功耗计算可能被低估，参见第III-D节。我们将在下一节讨论影响这些结果的其他因素。

V. 讨论

我们的结果证实，严格遵循实验方案十分必要。部署相关因素的差异，例如系统提示词的变化，以及 API 与开源版本之间细微的模型版本差异，都可能轻易导致时间测量结果出现偏差，如表III与表 IV 的对比所示。在此，API 与本地执行产生的词元数量不同，因此直接比较不同设置下的原始总能耗值会产生误导。因此，根据处理的词元数量对能耗值进行归一化，对于获得准确结果至关重要。

图 1 展示了基于 API 的模型和本地模型的这些归一化值。具体而言，图中显示，对于 Mistral-NeMo，基于 API 的部署与集群 H（配备 NVIDIA Hopper系列GPU 的本地部署）结果十分接近。不过，这并不意味着这两种部署的每词元能耗必然相同。服务器端优化技术（例如张量并行或流水线并行）可能会被用于降低延迟，但代价是功耗更高。例如，在多块GPU上运行 Mistral-7B 可以缩短运行时间，但会按比例增加每秒的总能耗。

尽管如此，我们知道大多数数据中心使用的是 Nvidia Ampere 和 NVIDIA Hopper系列GPU 的混合配置。此外，我们本地测得的每词元时间结果与基于 API 的结果相差不超过一个标准差。由于我们选择的模型可以装入单个 GPU，因此可以合理地认为这两种测量结果直接对应，最终证实了我们的假设：对于基于 API 的大语言模型，每词元推理时间是能源估算的合理代理指标。

A. 局限性

本研究基于这样一个假设：我们测得的基于 API 的模型的时间主要由 GPU计算时间决定，其他因素的贡献可以忽略不计。这一假设使我们能够降低实验评估的复杂性，例如仅使用一个图形处理器和单一批量大小。然而，在实际应用中，实际部署可能会引入其他延迟和能耗来源，而这些来源无法通过我们提出的实验设置直接观测到。其中包括多GPU执行，以及张量并行、流水线并行和专家并行策略；这些因素可能改变运行时间与能耗的权衡（例如，以增加功率消耗为代价降低延迟）。因此，尽管我们的方法能够对通过 API 访问的大语言模型的能耗提供合理的粗略估算，但对这些影响进行更细粒度的表征可以提高准确性。未来有必要开展系统性的消融研究，包括对并行策略、硬件配置和批量大小的研究，以量化这些因素对运行时间和能耗的影响。

VI. 结论

在本研究中，我们探讨了将推理时间作为代理指标，用于估算基于 API 的语言模型能耗的潜力。结果表明，每个词元所需的时间可用于推断基于 API 的模型背后所采用的 GPU 配置。同样，也可以对这些特定模型进行粗略的能耗估算，从而为研究人员评估基于 API 的模型的能效提供依据。

VII. 致谢

本研究由欧盟“地平线欧洲”研究与创新计划通过“SustainML”项目（101070408）资助并支持；由卡尔蔡司基金会通过“可持续嵌入式人工智能”项目、德国联邦研究、技术与航天部通过“Cross-Act”项目（01IW25001），以及克努特与爱丽丝·瓦伦堡基金会旗下的瓦伦堡人工智能、自主系统与软件计划（WASP）资助并支持。

参考文献

[1] J. Wang, E. Shi, H. Hu, C. Ma, Y. Liu, X. Wang, Y. Yao, X. Liu, B. Ge 和 S. Zhang，“用于机器人学的大语言模型：机遇、挑战与展望”，《自动化与智能期刊》，第 4 卷第 1 期，第 52–64 页，2025年。[2] J. Haltaufderheide 和 R. Ranisch，“医学与医疗保健中的 ChatGPT 伦理：关于大语言模型（大语言模型）的系统综述”，*NPJ*《数字医学》，第 7 卷第 1 期，第 183 页，2024年。[3] J. Howarth，“全球访问量最高的网站（2025年8月）”，网页，2025年。 [在线]。可访问：https://explodingtopics.com/blog/ most-visited-websites[4] F. Simon、R. K. Nielsen 和 R. Fletcher，“《2025 年生成式人工智能与新闻报告》：人们如何看待人工智能在新闻业和社会中的作用”，2025年。[5] O. McClain，“34%的美国成年人使用过 ChatGPT，约为2023年的两倍”，皮尤研究中心，网址：所引文章链接，2025年。[6]英国通信管理局，“《2024 年在线社会报告》”，英国通信管理局，英国伦敦，技术报告，2024年11月，访问日期：2025年10月23日。[在线]。可访问：https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/online-research/online-nation/2024/online-nation- 2024 =386238 -report.pdf?v N. [7] Maslej、L. Fattorini、R. Perrault、Y. Gil、V. Parli、N. Kariuki、E. Capstick、A. Reuel、E. Brynjolfsson、J. Etchemendy 等，“《2025 年人工智能指数报告》”，*arXiv* 预印本 *arXiv:2504.07139*，2025年。[8] S. Samsi、D. Zhao、J. McDonald、B. Li、A. Michaleas、M. Jones、W. Bergeron、J. Kepner、D. Tiwari 和 V. Gadepally，“从文字到瓦特：大语言模型推理能耗的基准测试”，载于*2023 IEEE* 高性能极限计算会议（*HPEC*）。IEEE，2023年，第 1–9 页。[9] E. Strubell、A. Ganesh 和 A. McCallum，“现代深度学习研究中的能源与政策考量”，载于《人工智能促进协会人工智能会议论文集》，第 34 卷第 09 期，2020年，第 13 693–13 696 页。[10] A. Hunt，“谷歌将资助三个核电站选址的开发”，2025年。 [在线]。可访问：https://www.world-nuclear-news.org/articles/google-to-fund-elementl-to-prepare-three-nuclear-power-sites [11] O. 等，“GPT-4 技术报告”，2024年。 [在线]。可访问：https://arxiv.org/abs/ 2303 08774 。M. AI，“谷歌将资助三个核电站选址的开发”，[12] 网页，2025年。 [在线]。可访问：https://mistral.ai/ [13] N. Jegham、M. Abdelatti、L. Elmoubarki 和 A. Hendawi，“人工智能有多‘耗能’？ 大语言模型推理的能源、水资源与碳足迹基准测试”，*arXiv* 预印本 *arXiv:2505.09598*，2025年。[14] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、Ł. Kaiser 和 I. Polosukhin，“注意力机制就是你所需要的一切”，《神经信息处理系统进展》，第 30 卷，2017年。[15] A. Dubey、A. Jauhri、A. Pandey、A. Kadian、A. Al-Dahle、A. Letman、A. Mathur、A. Schelten、A. Yang、A. Fan 等，“Llama 3 模型家族”，*arXiv* 电子预印本，第 arXiv–2407 页，2024年。

[16] A. Liu、B. Feng、B. Xue、B. Wang、B. Wu、C. Lu、C. Zhao、C. Deng、C. Zhang、C. Ruan 等，“DeepSeek-v3 技术报告”，arXiv 预印本 arXiv:2412.19437，2024年。[17] A. Yang、A. Li、B. Yang、B. Zhang、B. Hui、B. Zheng、B. Yu、C. Gao、C. Huang、C. Lv 等，“Qwen3 技术报告”，arXiv 预印本 arXiv:2505.09388，2025年。[18] Mistral，“Mistral Nemo”，2024年。 [在线]。可访问：https://mistral.ai/ news/mistral-nemo[19] C. White、S. Dooley、M. Roberts、A. Pal、B. Feuer、S. Jain、R. Shwartz-Ziv、N. Jain、K. Saifullah、S. Dey 等，“Livebench：污染受限的大语言模型基准测试”，arXiv 预印本 arXiv:2406.19314，2024年。[20] D. Patterson、J. Gonzalez、Q. Le、C. Liang、L.-M. Munguia、D. Rothchild、D. So、M. Texier 和 J. Dean，“碳排放与大型神经网络训练”，arXiv 预印本 arXiv:2104.10350，2021年。[21] P. Henderson、J. Hu、J. Romoff、E. Brunskill、D. Jurafsky 和 J. Pineau，“推动系统性报告机器学习的能源与碳足迹”，《机器学习研究杂志》，第 21 卷第 248 期，第 1–43 页，2020年。[22] D. Zhao、N. C. Frey、J. McDonald、M. Hubbell、D. Bestor、M. Jones、A. Prout、V. Gadepally 和 S. Samsi，“面向人工智能的更绿色世界”，载于 2022 IEEE国际并行与分布式处理研讨会专题会议（IPDPSW）。IEEE，2022年，第 742–750 页。[23] M. ¨Ozcan、P. Wiesner、P. Weiß 和 O. Kao，“通过模拟量化大语言模型推理的能耗与碳排放”，arXiv 预印本 arXiv:2507.11417，2025年。[24] R. Desislavov、F. Mart´ınez-Plumed 和 J. Hern´andez-Orallo，“人工智能推理能耗趋势：超越深度学习的性能与参数规模定律”，《可持续计算：信息学与系统》，第 38 卷，第 100857 页，2023年。[25] C.-J. Wu、R. Raghavendra、U. Gupta、B. Acun、N. Ardalani、K. Maeng、G. Chang、F. Aga、J. Huang、C. Bai 等，“可持续人工智能：环境影响、挑战与机遇”，《机器学习与系统论文集》，第 4 卷，第 795–813 页，2022年。[26] L. F. W. Anthony、B. Kanding 和 R. Selvan，“CarbonTracker：跟踪和预测深度学习模型训练的碳足迹”，arXiv 预印本 arXiv:2007.03051，2020年。[27] B. Courty、V. Schmidt、S. Luccioni、Goyal-Kamal、MarionCoutarel、B. Feld、J. Lecourt、LiamConnell、A. Saboni、Inimaz、supatomic、M. L´eval、L. Blanche、A. Cruveiller、ouminasara、F. Zhao、A. Joshi、A. Bogroff、H. de Lavoreille、N. Laskaris、E. Abati、D. Blank、Z. Wang、A. Catovic、M. Alencon、M. Stechly、C. Bauer、L. O. N. de Ara´ujo、JPW 和 MinervaBooks，“mlco2/codecarbon：v2.4.1”，2025年5月。在线。可访问：https://doi.org/ 。 /zenodo。D. Geißler、B. Zhou、M. Liu、S. Suh 和 P. Lukowicz，“训练的能耗：不同神经网络配置如何影响能源需求”，载于国际计算机系统架构会议。施普林格，2024年，第 33–47 页。

2024 [ ] 10 5281 11171501 J. [28] Fernandez、C. Na、V. Tiwari、Y. Bisk、S. Luccioni 和 E. Strubell，“大语言模型推理的能源考量与效率优化”，arXiv 预印本 arXiv:2504.17674，2025年。[29] L. Krupp、D. Geißler、V. Banwari、P. Lukowicz 和 J. Karolus，“促进可持续网络智能体：通过实证与理论分析对能耗进行基准测试和估算”，arXiv 预印本 arXiv:2511.04481，2025年。[30] 英伟达，“Mistral AI 与英伟达发布 Mistral Nemo 12B”，https://blogs. nvidia.com/blog/mistral-nvidia-ai-model/，2024年，访问日期：2025年10月29日。[31] ——，“英伟达 Blackwell 在 MLPerf推理 v4.1 中树立生成式人工智能新标准”，https://blogs.nvidia.com/blog/mlperf-inference-benchmark-blackwell/，2024年，访问日期：2025年10月29日。[32]亚马逊云科技，“Mistral AI 模型现已在 Amazon Bedrock 上线”，https://aws.amazon.com/blogs/aws/mistral-ai-models-now-available-on-amazon-bedrock/，2024年，访问日期：2025年10月29日。[33] 英伟达，“英伟达 A100 Tensor Core GPU 架构”，英伟达公司，技术报告，2020年。[34]在线。可访问：https://images.nvidia.com/aem-dam/en-zz/Solutions/data-center/nvidia-ampere-architecture-whitepaper.pdf