arXiv:2303.04673v2 [cs.CL] 8 Aug 2023

面向大语言模型生成推理的经济高效的超参数优化

Chi Wang [1] **Susan Xueqing Liu** [2] **Ahmed H. Awadallah** [1]

摘要

大语言模型（LLM）的生成能力引发了广泛关注，推动了各种商业应用程序的发展。使用模型的高成本促使应用程序构建者在有限的推理预算下最大化生成的价值。本文研究了对响应数量、温度和最大令牌数等推理超参数的优化，这些超参数会显著影响文本生成的效用/成本。我们设计了一个名为 EcoOptiGen 的框架，该框架利用经济高效的超参数优化和基于成本的剪枝。针对多种任务在 GPT-3.5/GPT-4 模型上开展的实验验证了其有效性。EcoOptiGen 已在 FLAML 库的“autogen”包中实现：`https://aka.ms/autogen`。

1. 引言

大语言模型（LLM），如 GPT-3.5 和 GPT- 4（Brown 等，2020；Ouyang 等，2022；OpenAI，2023），已在广泛的生成任务中展现出令人印象深刻的能力，包括讲故事（Lucy 和 Bamman，2021；Chen，2022）、代码生成（Trummer，2022；Poesia 等，2022）、数学问题求解（Cobbe 等，2021；Zong 和 Krishnamachari，2022）以及许多其他任务（Wang 等，2022）。尽管大语言模型并不总能生成完美的答案，但它们已经推动了构建强大用户体验的趋势，例如编程助手和支持聊天的搜索引擎。随着人们对构建支持大语言模型的应用程序的兴趣不断增长，对充分发挥生成推理价值的技术的需求也将随之增加。

近期，研究界开始探讨各个超参数对推理性能的影响，例如提示词（Liu 等人，2021；Mishra 等人，2021；Shieh，2022）和温度（Branwen，2020；Nadeem 等人，2020）。然而，关于如何集体且系统地优化不同超参数，我们所知甚少。此外，货币成本是大多数应用程序构建者关心的问题。高昂的成本，以及对能源消耗和环境影响的影响（Schwartz 等人，2020）有力地推动了人们系统地优化超参数，以实现效用最大化、成本最小化。

本文首次系统研究了使用大语言模型进行文本生成推理时的超参数优化。考虑到成本问题，我们采用一种经济的超参数优化方法（Wang 等，2021），并提出一种基于成本的剪枝策略，以在预算约束下提高优化效率。我们将名为 *EcoOptiGen* 的优化框架用于联合调整多个超参数，包括响应数量、最大令牌数、温度、概率质量和提示词。

为研究 EcoOptiGen 的有效性，我们在以下数据集上对其进行了评估：APPS（Hendrycks 等，2021a）（用于代码生成）；HumanEval（Chen 等人，2021）（用于代码生成）；MATH（Hendrycks 等，2021b）（用于数学问题求解）；以及 XSum（Narayan 等，2018）（用于文本摘要）。在全部四个数据集上，我们发现，EcoOptiGen 能够找到质量高于近期大语言模型基准测试 HELM（Liang 等，2022）所建议的默认设置或同等预算下的简单修改方案的超参数设置。结果表明，我们的剪枝技术能够显著提升调优效果。我们还发现，整体性的超参数优化可以缓解个体特性带来的影响，从而避免次优结果。

2. 背景

2.1. 使用大语言模型进行文本生成

大语言模型文本生成的输入和输出 图 1 展示了一个输入提示词的示例。收到提示词后，大语言模型会进行推理，生成一个或多个输出响应。输入提示词还可以包含多个示例，用以说明什么

[1]Microsoft Research [2]Stevens Institute of Techology。通信请联系：Chi Wang <wang.chi@microsoft.com>。

版权所有 © 2023 作者。

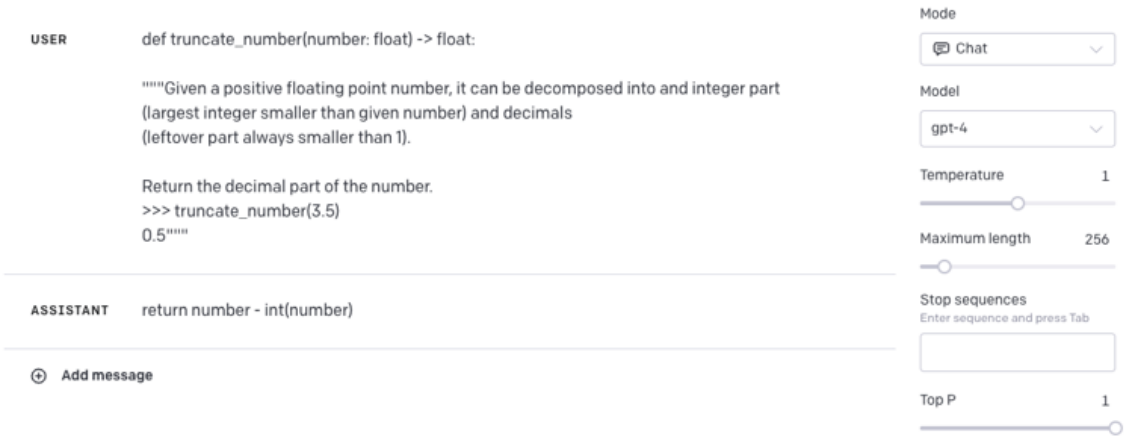


图*1.* 使用 GPT-4 进行代码生成的示例，以及用户可为推理设置的几个超参数示例，例如提示词、温度和最大令牌数。

样的响应才是理想的。应用程序可以通过多种方式使用或验证输出，例如，使用代码执行器（Hendrycks 等，2021a）或数学表达式检查器（Hendrycks 等，2021b）。推理过程中通常可以生成多个响应并搜索其中的最佳响应，例如代码生成（Chen 等人，2021）和机器翻译（Wullachand Chazan，2022）。生成文本的效用取决于使用它的应用程序。例如，如果在生成前提供了预定义的程序化测试，那么只要生成的响应中有一个能够通过测试，就可以认为代码生成成功。

使用大语言模型进行文本生成的成本 使用大语言模型处理文本生成请求的成本，与生成输出所需的计算量成正比。计算量由输入和输出中的词元数量决定。大语言模型通常以服务形式提供，并根据使用量收费。从使用大语言模型的应用程序开发者的角度来看，目标是在推理预算约束下（例如，以解决一个编码问题所需的平均美元成本来衡量）最大限度地提高生成文本的效用。这可以通过优化推理的超参数来实现，因为它们会显著影响生成文本的效用和成本。下一节将讨论超参数如何影响效用和成本。

2.2. 超参数如何影响文本生成性能?

各项超参数的影响 我们选取 OpenAI 的一个典型 API，即补全 API（ope，2023年），分析可调节的超参数及其对成本和效用（例如准确率、成功率）的影响：（1）模型——这是必填输入，用于指定要使用的模型 ID。（2）提示词——提供文本生成任务上下文的模型输入提示。（3）最大令牌数——输出中要生成的最大词元数（单词或子词）。（4）温度——介于 0 和 1 之间的值，用于控制生成文本的随机性。温度越高，生成的文本越随机、越多样；温度越低，生成的文本越可预测。（5）top p——介于 0 和 1 之间的值，用于控制每次生成词元时的采样概率质量。top p 值越低，生成文本越倾向于依据概率最高的词元；值越高，模型则可以探索更广泛的可能词元。（6）n——针对给定提示词要生成的响应数量。生成多个响应可以带来更多样且可能更有用的输出，但也会增加请求成本。（7）停止序列——字符串列表；生成文本中遇到列表中的字符串时，生成过程便会停止。此项可用于控制输出的长度或有效性。（8）存在惩罚和频率惩罚——用于控制生成文本中特定单词或短语的存在情况及出现频率相对重要性的值。这些超参数有助于控制生成文本的重点和均衡性。（9）最佳候选数——为给定提示词在服务器端生成的响应数量，系统将从中选出“最佳”响应（即每个词元的对数概率最高的响应）。

多个超参数的联合影响 可以看出，文本生成的成本和效用与这些超参数的共同作用交织在一起。超参数的不同子集之间也存在复杂的相互作用。例如，不建议同时调整温度和 top p，使其偏离默认值，因为二者都会控制

生成文本的随机性，同时更改二者可能会产生相互冲突的效果（ope，2023年）；n 和最佳候选数通常不会一起调整，因为如果应用程序能够处理多个输出，在服务器端进行筛选会造成不必要的信息损失；n 和最大令牌数都会影响生成的词元总数，进而影响请求成本。这些相互作用和权衡使得人们难以针对特定的文本生成任务，手动确定最优的超参数设置。

3. EcoOptiGen

我们首先介绍 EcoOptiGen 中使用的以下符号和定义：

调优数据 D，即一小组示例，可用于衡量每种超参数配置的优劣。每个数据示例包含若干文本字段。例如，HumanEval（Chen 等人，2021）数据集包含一个输入字段，由 Python 函数签名与文档字符串拼接而成；此外还包含一个测试字段，其中包含测试代码。

效用函数 U，即用于表示生成文本效用（一个实数分值）的函数。例如，对于代码生成，效用是通过测试的成功率；对于文本摘要，效用是摘要效果，例如 ROUGE 分数。多个可验证响应（即应用程序可以从中选出最佳响应）的效用，定义为所有响应中的最高效用分数。

预算约束 $B = (B.i, B.o)$，这是一个由两个值组成的元组：$B.i$ 表示调优数据中每个示例的平均推理预算，D 而 $B.o$ 表示总优化预算。通常，这些预算以美元成本衡量，其与输入和输出中的词元数量成正比（参见第2节）。当输入和输出的每词元价格恒定时，美元成本可转换为输入和输出中的词元数量。为便于说明，本节用词元数量表示预算：$B.i$ 表示每个请求允许消耗的平均词元数量，例如 1K；$B.o$ 表示调优过程中允许消耗的词元总数，例如 1M。当 $B = (1K, 1M)$ $|D| = 20$时，如果每个请求恰好消耗 1K 个词元，则优化过程中允许尝试的超参数配置数量等于 $1M/1K/20 = 50$。

搜索空间 S，这是一个字典，其中键为每个超参数的名称，值为对应超参数的待搜索取值范围。根据第 2.2 节的分析，如果用户未提供此输入，我们会设计如表 1 所示的默认搜索空间。是否需要调整某些超参数取决于应用程序。

配置 x**在** D**中的实例上的平均效用和成本消耗**：$U_x(D)$ 和 $C_x(D)$。如果 x是 无效，则配置 $C_x(D) > B.i$。

在指定的优化预算 $B.o$内，框架会在给定的搜索空间中迭代尝试不同的配置 S，并输出一个配置 x^*，使其在调优数据 $U_{x^*}(D)$ 上的平均效用最大 D，同时受平均推理预算 $C_{x^*}(D) \leq B.i$的约束。其架构如图2所示。 超参数搜索器 负责提出配置，并调用 配置评估器 评估每个配置的有效性和效用。

3.1. 超参数搜索器

我们选择黑盒优化方法，因为我们的目标是让该框架能够通用地应用于：(1) 作为服务的LLM；(2) 复杂的效用函数，这些函数可能涉及对 LLM 返回结果进行黑盒评估。我们的框架将配置效用的计算过程抽象出来；该过程包括向 LLM 发出请求，并使用特定于应用程序的流程评估其响应。

黑盒优化技术有多种，例如随机搜索（Bergstra 和 Bengio，2012）、贝叶斯优化（Bergstra 等，2011）、进化搜索（Goldberg 和 Deb，1991）和局部搜索（Wu 等，2021）。考虑到成本效率，我们选择了一种结合贝叶斯优化与局部搜索的方法，称为 BlendSearch（Wang 等，2021）。BlendSearch 中的局部搜索方法采用随机直接搜索，并具有可证明的收敛速率和成本界。贝叶斯优化用于为局部搜索生成起始点，并自适应地为不同的局部搜索线程分配优先级。原始的 BlendSearch 专为训练超参数优化而设计，并使用训练时间作为成本的度量。我们对其进行调整，以优化推理超参数，并推广成本指标。

3.2. 配置评估器

简单评估器以配置 x 作为输入，并输出要优化的指标。它会遍历 D中的数据样本。对于每个样本 $d_i \in D$，都会使用该配置和输入字段 $d_i.in$向 LLM 服务发送请求。随后，利用 LLM服务返回的响应计算效用 U，并衡量成本消耗 C。遍历结束后，将平均成本$C_x(D)$与用户提供的上限 $B.i$进行比较。如果$C_x(D) \leq B.i$，则返回 D 中所有数据的平均效用 $U_x(D)$；否则返回零值，以表示该配置无效。下文将介绍对简单评估器的改进。

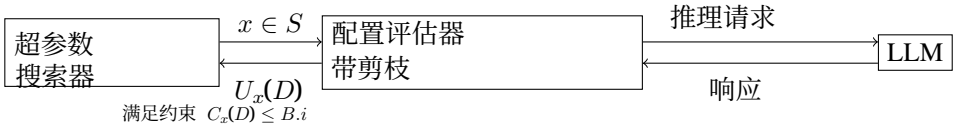


图*2*。架构。

表1。默认搜索空间　　用于优化框架。默认情况下，部分超参数固定为常量，但用户可以　　n 根据需要覆盖所有这些配置 领域知识覆盖所有这些设置。

超参数	默认搜索范围	Note
模型	[“text-ada-001”，　“text-babbage-001”、“text-davinci-003”、“gpt-3.5-turbo”、“gpt-4”]	具有不同成本-质量权衡的 GPT 模型
提示词	[“$\{prompt\}$”]	提示模板列表。“$\{prompt\}$”将替换为一个每个实例中名为”prompt”的输入字段，以生成实际的实例提示词。通常，用户会根据领域知识覆盖此设置。用户也可以使用此列表从不同的提示策略中进行选择，例如是否使用思维链或上下文学习示例。
最大令牌数	lograndint(100, 1000)	随机整数，按对数分布
温度或 top_p	[{“温度”：uniform(0, 1)}, {“top p”：uniform(0, 1)}]	分层搜索空间：一个配置将选择温度或 top p
n	randint(1, 100)	1 到 100 之间的随机整数
stop	None	用户可以指定依赖应用程序的停止选项
存在惩罚	0	用户可以根据需要指定一个介于 [-2, 2]之间的浮点数范围
频率惩罚	0	用户可以根据需要指定一个介于 [-2, 2]之间的浮点数范围
最佳候选数	1	用户可以指定一个下界为 1 的整数范围，并将 n 固定为 1

如果配置 x 无效，则提前终止试验有助于节省不必要的成本。我们设计了一种剪枝策略，在试验过程中谨慎调整两个与成本相关的因素：调优数据示例的数量和响应数量。对配置进行完整评估x 需要发送 $|D|$ 个大语言模型请求，每个请求要求返回 $x.n$个响应，其中 $x.n$ 是配置 x 中超参数 n 的设置（或者，如果搜索的是最佳候选数而非 n，则为最佳候选数的设置）。我们的目标是在无效试验上花费更少的成本。完整流程详见附录中的算法 1。

初始有效性检查 对于给定配置 x，在昂贵的评估开始之前，我们首先检查是否可以直接剪枝该配置（算法第 4–10 行 1）。此检查基于一个针对我们优化问题的特定假设。

假设 3.1。 给定两个配置 x_1 和 x_2，它们的模型、提示词和停止序列设置相同。如果 x_1 中的响应数量和最大令牌数均大于或等于 x_2 中的对应值，那么我们预期 x_1 的平均令牌消耗不低于 x_2。

这一假设的结果是：如果 x_2 无效，那么x_1 也无效。如果 x_1 有效，那么 x_2 也有效。我们的剪枝利用这一假设，找出已知有效的最大值 n 和某个配置 x的已知无效最小值 n，利用已尝试配置的有效集和无效集 X_{valid}和 X_{invalid}；这些配置与 x 具有相同的模型、提示词和停止序列设置。

$$\text{max_valid_n} = \max_{x' \in X_{\text{valid}}, x'.\text{max_tokens} \ge x.\text{max_tokens}} x'.n \quad (1)$$

$$\text{min_invalid_n} = \min_{x' \in X_{\text{invalid}}, x'.\text{max_tokens} \le x.\text{max_tokens}} x'.n \quad (2)$$

接下来，根据最大有效 n、最小无效 n 和 $x.n$ 之间的关系，我们执行以下操作：

1. 如果 $x.n \le$ max valid n，则我们按原样评估该试验。这对应于以下情况： x 根据假设，预计为有效3.1。2. 否则（$x.n$ > max valid n），如果 $x.n$ ≥min invalid n，则我们不进行任何评估，直接剪枝该试验。这对应于以下情况： x 根据假设，预计为无效3.1。3. 否则（max valid n < $x.n$ < min invalid n），我们从响应数量等于 max valid n 时开始评估该试验。这对应于以下情况： x 预计可能有效，也可能无效，而 max valid n 预计是一个有效的响应数量，可用于 x。

我们的检查顺序考虑到了假设3.1偶尔可能被违反，即 max valid n 偶尔可能等于或大于 min invalid n。通过先检查条件 1 再检查条件 2，我们确保在违反假设时仍有机会评估该试验。

算法的外循环 1：改变响应数量 通过初始检查（情况 1 或 3）后，我们逐步将响应数量加倍，直至达到 $x.n$ (第 11 行至末尾）。 n 的起始值根据上述规则确定。为了评估特定的 n，我们暂时将$x.n$ 修改为 $n - n'$，其中 n' 是当前配置在 D 中最后一次评估的响应数量（在起始点或发生数据跳过时为 0，下一段将对此进行说明）。这样就能利用为当前配置评估较小 n 时所发出的请求结果，与直接请求 n 个响应相比，可以节省成本。请注意，虽然逐步增加 n 的数量可以在响应数量较少时提前终止一次试验，但由于输入令牌在每次请求中都会重复消耗，这也可能增加评估有效试验的总成本。我们的设计通过以下几种方式缓解了这一问题：(a) 从已知的最大有效 n 值开始，而不是从 1 开始；(b) n 的几何增长将改变 n 的次数从线性增长（线性调度）降为对数级；(c) 下一段所述的数据跳过机制有助于在试验有效时减少较小 n 的请求次数。

算法 1 的内循环：改变数据样本数量 对于每个固定的 n个响应，我们在调优数据 D 上采用渐进式子采样（Provost 等，1999）来剪枝试验（第 12–32 行）。获得 D中 k 个样本的响应后，我们可以计算其成本和效用的均值。我们将这 k 个样本的子集记为 D_k。 C_x (D_k)是 C_x(D)的估计值。如果 C_x(D)确实低于（分别高于） $B.i$，则可以使用 Hoeffding-Serfling 不等式（Bardenet 和 Maillard，2015）计算C_x(D_k)的上界（分别为下界）。如果C_x(D_k)大于上界，我们就终止该试验，并更新 X_{invalid}。如果 C_x(D_k)小于下界，且当前 n 小于原始$x.n$，我们就跳过 D中剩余的数据点，重置 $n' = 0$，并更新 X_{valid}。数量 k 每次翻倍，直至达到 $|D|$。 k 的几何增长限制了每次试验中进行此假设检验的次数，从而降低错误剪枝的可能性。

4. 实验

我们希望探究以下研究问题：首先，对于文本生成任务，EcoOptiGen通过调优超参数集能够获得多大的提升，且需要满足预算约束？其次，改变推理预算会如何影响优化结果？第三，改变模型会如何影响优化结果？本节首先在第4.1节介绍实验设置。然后，我们在第4.2节至第4.4节探讨这三个研究问题，并在第4.6节讨论局限性及未来工作。

4.1. 设置

数据集为评估 EcoOptiGen 的性能，我们从 HELM 基准测试（Liang 等，2022）（v1.0）中选取了多种文本生成任务：代码生成、数学问题求解和文本摘要。对于代码生成，我们在两个数据集上评估 EcoOptiGen：APPS（Hendrycks 等，2021a）（给定编程题描述后生成 Python 代码的数据集）和 HumanEval（Chen 等人，2021）（根据函数名称和文档字符串生成 Python 代码的数据集）。对于数学问题求解，我们使用 MATH 数据集（Hendrycks 等，2021b）（包含思维链的数学问题数据集，即解题推导步骤）。对于文本摘要，我们在 XSum 数据集（Narayan 等，2018）（用于总结新闻文章的大型数据集）上评估 EcoOptiGen 的性能。

调优时，我们从所有数据集中随机抽取 20 个样本用于调优；XSum 除外，因为其数据量更大，我们抽取 60 个样本。对于每个数据集，我们随机抽取数百个样本用于测试。测试集的选择流程遵循 HELM。对于 MATH，我们使用第 4.2 至4.4节中符合 HELM 标准的“Level 1”问题。

评估指标 对于两项代码生成任务，数据集已提供测试用例，用于在推理时验证响应。对于每个代码生成实例，只要任意一个响应通过测试用例，EcoOptiGen 在该实例上的得分就是 1，否则为 0。对于 MATH，我们采用两种评估方式。在第 4.2节至第 4.4节中，我们将“成功”定义为：如果返回的思维链响应中有一个的最终答案与真实答案等价，则 EcoOptiGen 在该实例上的得分为 1，否则为 0。在第 4.5节中，我们将“成功投票”定义为：如果基于多数投票得出的响应，其最终答案与真实答案等价，则 EcoOptiGen 在该实例上的得分为 1，否则为 0。对于 XSum，我们使用“最佳候选数”按平均对数概率对生成的响应重新排序，并采用排名最高的响应的 Rouge-2 分数（Lin，2004）。

比较方法 我们评估以下替代方案，并与 EcoOptiGen 进行比较：

- HELM。我们检查在 HELM 基准测试下评估得出的最佳分数。对于每个数据集，我们找出表现最佳的

表 2. 根据 HELM 评估结果，在 code-davinci-002、text-davinci-002 和 text-davinci-003 中选出的“最佳” GPT-3.5 模型的结果。

方法	APPS	HumanEval	MATH	XSum
HELM	0.03	0.465	0.378	0.140
EcoOptiGen（HELM 预算）	0.05	0.521	0.414	0.144
搜索	0	0.493	0.769	0.136
Search+PSR	0	0.493	0.739	-
EcoOptiGen	0.05	0.792	0.771	0.144
HELM（已修改）	0.03	0.701	0.403	0.140

GPT-3.5 模型（code-davinci-002、text-davinci- 002 和 text-davinci-003），这些模型由 HELM 报告。随后，我们使用相同的模型和超参数设置（在 APPS 和 MATH 上采用经过修改的提示词，具体说明见下一段），基于 HELM 的设置在我们的测试数据上重新评估。 这些配置的完整详情列于附录表 4 中。之所以采用这一基线，是因为 HELM 广泛覆盖了最新的大语言模型，并为每个任务指定了特定的超参数设置，而其他地方很少能找到这样的资源。

• Search。这是一种使用与 EcoOptiGen 相同的超参数搜索器，但不使用剪枝，也不更改优化指标的方法。

• Search + PSR。这种方法与 Search 相同，但使用概率成功率而非成功率作为优化指标。与 XSum 数据集无关。

默认情况下，所有基于搜索的方法的输入均设为 $B.i = 1K$， $B.o = 1M$。搜索空间遵循表 1，而“model”和“stop”则由 HELM 配置覆盖。对于 HumanEval，我们在四种模板中搜索提示词：“{定义}”，“# Python 3{定义}请补全以下 Python 函数：{定义}请补全以下 Python 函数，并在函数内包含必要的导入语句：{定义}”。为节省推理成本，在 APPS 上我们使用零样本提示，而非 HELM 使用的两样本提示。对于 MATH，我们在提示词中对所有类别都只使用一个固定的示例，而不是像 HELM 那样每个类别使用八个示例。对于 XSum，使用与 HELM 相同的提示词、n 和最大令牌数，同时在 randint(1, 100) 范围内搜索最佳候选数；预算设为$B.i = 2K$， $B.o = 4M$。

4.2. EcoOptiGen 的性能

EcoOptiGen 的性能得分如表 2 所示，并列出了其他比较方法。在全部 4 个数据集上，EcoOptiGen 的表现均优于 HELM 基准测试中未经调优的 GPT-3.5 模型。为验证性能提升是否仅仅源于响应数量增加，我们在表 2中加入了“HELM（修改版）”方法，将响应数量调整为与 EcoOptiGen 最佳配置所消耗的推理预算相匹配。我们还加入了“EcoOptiGen（HELM 预算）”，表示 EcoOptiGen 在每项任务上使用与 HELM 相同的推理预算时的性能。表格前两行之间或后两行之间的比较 2 表明，HELM 基准测试中的超参数调优不足，而联合调优所有超参数优于单纯增加响应数量。

在表 2中，我们还将 EcoOptiGen 的性能与其他采用超参数搜索的方法进行了比较。可以看到，采用剪枝技术后，EcoOptiGen 的表现始终优于其他未采用剪枝技术的方法；在 APPS 和 HumanEval 上，优势尤为显著。这证实了剪枝技术的有效性。我们还在附录的图 5 中比较了 EcoOptiGen 与其他搜索方法搜索到的试验次数。可以观察到，在相同的优化预算下，EcoOptiGen 搜索到的试验次数多出 2-27×，这有助于它在所有任务上取得更好的结果。

本研究发现：(a) 与直接采用 HELM 评估得出的最佳配置相比，通过调整推理超参数，有可能为特定应用程序找到更好的配置；(b) 剪枝可以大幅提升优化效率。

4.3. 推理预算的影响

为了了解推理预算如何影响 EcoOptiGen 的性能，我们进一步将 APPS、HumanEval 和 MATH 上的推理预算 $B.i$ 从 500 个词元调整到 2000 个词元，同时固定总优化预算 $B.o = 1M$。附录中的表 5 展示了性能得分，图6 展示了在优化预算内完成的试验次数。在 APPS 上，输入令牌的平均数量超过 500，因此该情况下没有可用的性能得分。在 HumanEval 上，随着推理预算从 500 增至 2000，优化后的性能得分从 0.653 提升至 0.819。在 MATH 上，随着推理预算从 500 增至 2000，性能得分从 0.398 提升至 0.863。在 APPS 上，当推理预算从 1500 增至 2000 时，性能得分从 0.10 降至 0.07。根据图 6，我们推测，性能下降是因为随着每次试验的平均成本增加，总优化预算内的试验次数减少。我们进行了一项额外实验来验证这一假设：将 $B.o$ 提高到 2M，用于 $B.i = 2000$上的 APPS。随后，优化后的性能得分从 0.07 提升至 0.12，试验次数从 86 增至 165。结果支持了这一假设。

本研究的核心结论是，EcoOptiGen 能够在推理预算增加时找到明显更优的配置，

表 3 $2K\ B.o = 1M$ 35 $B.i =$ 。使用不同 GPT-. 模型针对 , 调优的结果。* 表示根据 HELM，该模型在相应数据集上的表现最佳。

模型	APPS	HumanEval	MATH	XSum
code-davinci-002	0.07*	0.819*	0.856	**0.198**
text-davinci-002	0.20	0.847	0.785	0.144*
text-davinci-003	**0.21**	**0.861**	**0.863***	0.119

除非优化预算不足。

4.4. 模型的影响

在之前的实验中，我们根据 HELM 基准测试，为每个数据集固定了模型。在本小节中，我们首先在每个数据集上将 EcoOptiGen 应用于 GPT-3.5 系列中的其他模型，使用 2K 的推理预算和 1M 的总优化预算。在进行这项实验时，OpenAI 通常建议代码生成使用“code-davinci-002”，其他文本生成任务使用“text-davinci-003”（Shieh，2022）。

表 3总结了实验结果。在 APPS、HumanEval 和 MATH 上，调优后 text-davinci-003 的表现最佳。在 XSum 上，调优后 code-davinci-002 的表现最佳。在三个数据集上，调优后的最佳模型与 HELM 基准测试所示的最佳模型并不相同，这一点可从各列中星号（HELM）与粗体（调优后）标记不一致看出。在 APPS 和 HumanEval 上，我们发现 text-davinci-003 在两个代码生成任务中表现出色。在 MATH 上，调优后 code-davinci-002 的表现可接近 text-davinci- 003。在 XSum 上，我们发现 code-davinci- 002 的表现出人意料地远超 text-davinci 系列模型。这些结果与关于模型选择的普遍看法相当矛盾（Shieh，2022）。

我们进一步在图 3 中绘制了模型在各个独立推理预算下的表现。我们发现，对于 HumanEval，text-davinc-003 模型始终表现最佳。在 MATH 上，尽管表 3 显示 text-davinci-003 模型略微优于 code-davinci-002，但图 3 显示后者实际上在低推理预算范围内表现更优。

本研究的核心结论是，借助 EcoOpti-Gen，表现最佳的模型并不总是通常推荐的模型。这揭示了超参数优化的一项优势：避免因模型的独特特性而做出次优选择。较新的模型并不一定比旧模型表现更好。

4.5. ChatGPT 模型

在本文初始版本发布后，为 ChatGPT 提供支持的两种聊天优化模型（gpt-3.5-turbo 和 gpt-4）相继发布论文。我们将在本节评估 EcoOptiGen 在这两种模型上的性能。

我们使用 MATH 数据集进行评估。为增加评估场景的多样性，本节采用的设置与前几节有所不同。我们使用“成功投票”来比较每道题的多数投票结果与真实答案，而不使用“成功”这一指标；后者用于检查是否至少有一个回答正确。我们使用以下提示模板：“{问题}请仔细解题。尽可能简化答案。将最终答案写在 \方框{} 中。”我们使用代数类别中从第2级到第5级的所有级别。我们按级别进行调优，并将 gpt-3.5-turbo 和 gpt-4 都纳入“模型”的搜索空间。通常认为，gpt-4 在数学问题上的表现远胜 gpt-3.5-turbo（OpenAI，2023年）。我们使用默认推理超参数与 gpt-4 进行比较。

图 4 展示了各配置的平均准确率和平均推理成本。在第2级，令人惊讶的是，调优后的 gpt-3.5-turbo 模型被选为更优模型，其准确率远超未经调优的 gpt-4（92% 对 70%），且推理预算相同或高出 2.5 倍。在第3级也观察到相同的情况。到了第4级和第5级，选中的模型发生了变化。调优后的 gpt-4 准确率显著更高（第4级为 56% 对 44%，第5级为 35% 对 20%），且成本低于未经调优的 gpt-4。这些结果进一步验证了 EcoOptiGen 的稳健有效性，也强化了前文的结论。随着大语言模型不断发展，性能调优仍有发挥作用的空间。

4.6. 讨论与未来工作

对于摘要任务，调优带来的提升不如代码和数学任务那么显著。一个可能的原因是，从$best\ of$响应中选出一条响应的排序标准与最终评估指标并不一致。

未来一个有趣的研究方向是开发有助于理解优化后超参数选择的方法（参见附录中的表6 ）。此外，还可以进一步实现调优自动化。例如，当前的解决方案将用户指定的提示词选择作为输入搜索空间。自动搜索示例的最佳数量和选择方式，有望更有效地利用推理预算。

5. 相关工作

通用机器学习模型的超参数优化方法已被研究了十年（Feurer 和 Hutter，2019；Bergstra 等人，2011）。由于训练深度神经网络的成本非常高，人们提出了新的超参数优化方法，以降低所需成本。早期

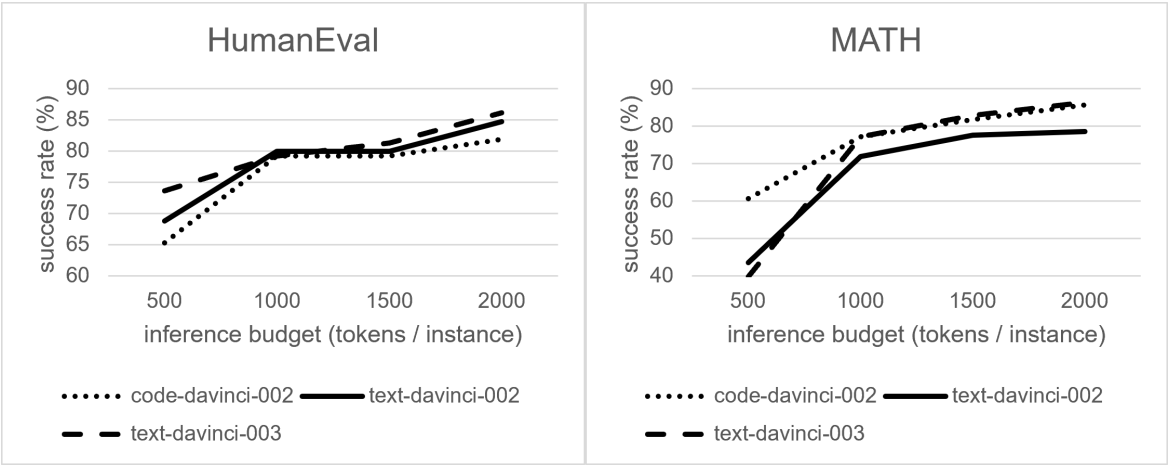


图*3*。 在固定优化预算下，使用不同推理预算的 GPT-3.5 模型所得结果。

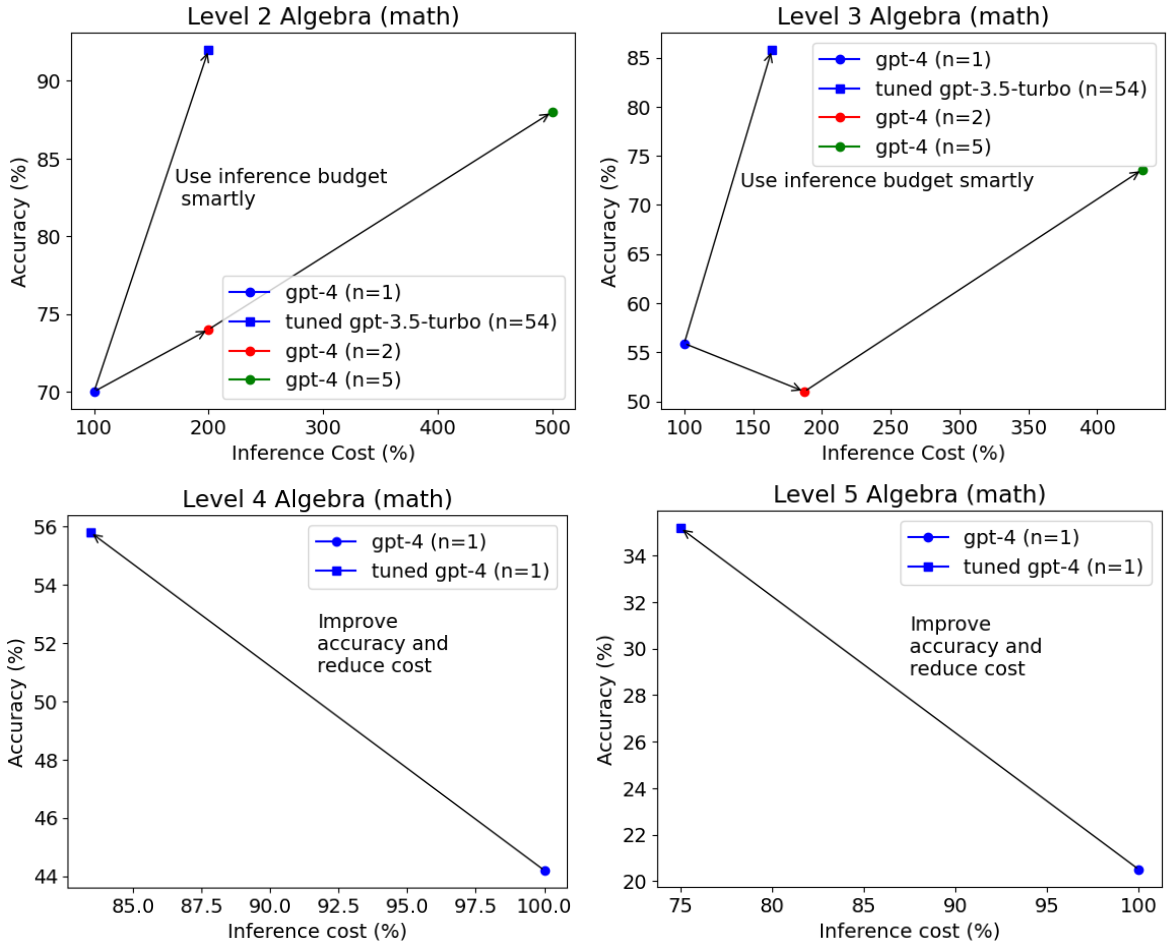


图*4*。 针对 MATH 调优 ChatGPT 模型。

停止方法（Li 等人，2017）通过将低保真度（例如训练轮数）下配置的训练表现与同等保真度下其他配置的训练表现进行比较，停止使用表现不佳的配置继续训练。后来，人们提出了经济高效的超参数优化方法。例如，在 BlendSearch（Wang 等人，2021）中，提出了一种经济高效的混合搜索策略，用于处理异构的评估成本，并在对 Transformer 模型 Turing-NLRv2 进行微调时验证了其有效性。针对自然语言处理任务的自动超参数优化也得到了专门研究，例如，对类似 BERT 的语言理解模型进行微调（Liu 和 Wang，2021）以及对神经机器翻译系统进行微调（Zhang 和 Duh，2020）。

6. 更广泛影响声明

大语言模型的用户在确定推理预算和优化预算时，应将环境影响纳入考虑。我们的工作通过帮助用户在预算上限内找到可部署配置，降低了能源消耗；否则，这类配置很难找到，并可能导致环境影响加剧。如果预算设置不当，优化过程可能会造成过高的能源消耗。

参考文献

(2023年)。OpenAI API。
https://beta.openai.com/docs/api-reference/completions/create。
Bardenet, R. 和 Maillard, O.-A. (2015)。无放回抽样的集中不等式。伯努利，21(3):1361–1385。Bergstra, J. 和 Bengio, Y. (2012)。用于超参数优化的随机搜索。《机器学习研究杂志》，13:281–305。Bergstra, J. S., Bardenet, R., Bengio, Y. 和 K´egl, B. (2011)。超参数优化算法。载于神经信息处理系统进展。Branwen, G. (2020)。GPT-3 创意小说。Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. 等 (2020)。语言模型是少样本学习器。神经信息处理系统进展，33:1877–1901。Chen, M. (2022)。利用自然监督进行语言表征学习与生成。*arXiv* 预印本 *arXiv:2207.10617*。Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G. 等 (2021)。评估基于代码训练的大语言模型。*arXiv* 预印本 *arXiv:2107.03374*。

Cobbe, K.、Kosaraju, V.、Bavarian, M.、Chen, M.、Jun, H.、Kaiser, L.、Plappert, M.、Tworek, J.、Hilton, J.、Nakano, R. 等（2021）。训练验证器以解决数学文字题。*arXiv* 预印本 *arXiv:2110.14168*。

Feurer, M. 和 Hutter, F.（2019）。超参数优化。载于 自动化机器学习，第 3–33 页。施普林格，尚。

Goldberg, D. E. 和 Deb, K.（1991）。遗传算法中所用选择方案的比较分析。载于 遗传算法基础，第 1 卷，第 69–93 页。爱思唯尔。

Hendrycks, D., Basart, S., Kadavath, S., Mazeika, M., Arora, A., Guo, E., Burns, C., Puranik, S., He, H., Song, D., 等（2021a）。使用 APPS 衡量编程挑战能力。*arXiv* 预印本 *arXiv:2105.09938*。Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., 和 Steinhardt, J.（2021b）。使用 MATH 数据集衡量数学问题求解。*arXiv* 预印本 *arXiv:2103.03874*。Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., 和 Talwalkar, A.（2017）。Hyperband：一种用于超参数优化的新型基于多臂老虎机的方法。机器学习研究杂志，18(1)：6765–6816。Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., 等（2022）。语言模型的整体评估。*arXiv* 预印本 *arXiv:2211.09110*。Lin, C.-Y.（2004）。ROUGE：用于自动评估摘要的软件包。载于文本摘要百花齐放，第 74–81 页。Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., 和 Neubig, G.（2021）。预训练、提示与预测：自然语言处理领域提示方法的系统综述。*arXiv* 预印本 *arXiv:2107.13586*。Liu, X. 和 Wang, C.（2021）。关于微调预训练语言模型的超参数优化的实证研究。*arXiv* 预印本 *arXiv:2106.09204*。Lucy, L. 和 Bamman, D.（2021）。GPT-3 生成的故事中的性别与表征偏见。载于第三届叙事理解研讨会论文集，第 48–55 页。Mishra, S., Khashabi, D., Baral, C., Choi, Y., 和 Hajishirzi, H.（2021）。重新构建指令提示，以适配 GPTk 的语言。*arXiv* 预印本 *arXiv:2109.07830*。

Nadeem, M., He, T., Cho, K., 和 Glass, J.（2020）。开放式语言生成采样算法的系统性特征分析。*arXiv* 预印本 *arXiv:2009.07243*。

Narayan, S., Cohen, S. B. 和 Lapata, M.（2018）。别告诉我细节，只要摘要！用于极限摘要的主题感知卷积神经网络。*arXiv*预印本 *arXiv:1808.08745*。

OpenAI（2023年）。GPT-4技术报告。

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A. 等（2022）。通过人类反馈训练语言模型遵循指令。*arXiv*预印本 *arXiv:2203.02155*。

Poesia, G., Polozov, O., Le, V., Tiwari, A., Soares, G., Meek, C. 和 Gulwani, S.（2022）。Synchromesh：基于预训练语言模型的可靠代码生成。*arXiv*预印本 *arXiv:2201.11227*。Provost, F., Jensen, D. 和 Oates, T.（1999）。高效渐进式采样。载于第五届 *ACM SIGKDD* 知识发现与数据挖掘国际会议论文集，第 23–32 页。Schwartz, R., Dodge, J., Smith, N. A. 和 Etzioni, O.（2020）。绿色人工智能。《*ACM*通讯》，63(12)：54–63。Shieh, J.（2022）。使用 OpenAI API 进行提示工程的最佳实践。Trummer, I.（2022）。CodexDB：利用 GPT-3 Codex 根据自然语言指令合成用于查询处理的代码。《*VLDB*基金会论文集》，15(11)：2921–2928。Wang, C., Wu, Q., Huang, S. 和 Saied, A.（2021）。采用混合搜索策略进行经济高效的超参数优化。载于国际学习表征会议。Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Arunkumar, A., Ashok, A., Dhanasekaran, A. S., Naik, A., Stap, D. 等（2022）。Super-NaturalInstructions：通过声明式指令在 1600+ 自然语言处理任务上实现泛化。网址 *https://arxiv.org/abs/2204.07705*。Wu, Q., Wang, C. 和 Huang, S.（2021）。针对成本相关超参数的节俭优化。载于*AAAI'21*。Wullach, T. 和 Chazan, S. E.（2022）。别那么自信！通过置信度松弛提升自动语音识别解码效果。*arXiv*预印本 *arXiv:2212.13378*。

Zhang, X. 和 Duh, K.（2020）。神经机器翻译系统超参数优化的可复现且高效的基准测试。计算语言学协会会刊，8:393–408。Zong, M. 和 Krishnamachari, B.（2022）。使用 gpt- 3 求解涉及方程组的数学文字题。载于第十三届*AAAI*人工智能教育进展研讨会论文集。

表 4. HELM 基线的配置。

数据集	配置
APPS	{‘模型’：‘code-davinci-002’，‘提示词’：‘{输入}’，‘最大令牌数’：600，‘温度’：0.2，‘停止序列’：[‘’’,– – –,””””，\n\n\n]}
HumanEval	{‘模型’: ‘code-davinci-002’, ‘提示词’: ‘{定义}’, ‘最大令牌数’: 600, ‘温度’: 0.2, ‘停止序列’：[\nclass, \ndef, \nif, \nprint]}
MATH	{‘模型’: ‘text-davinci-003’，‘提示词’: 1 次思维链演示*，‘最大令牌数’：400，‘温度’：0，‘停止序列’：[###]}
XSum	{‘模型’：‘text-davinci-002’，‘提示词’：5 次演示**，‘最大令牌数’：400，‘温度-度’：0.3，‘停止序列’：[###]}

* MATH 中的 1 次思维链演示:

给定一道数学题，求出答案。请尽可能简化你的答案。###

问题：$\sqrt{3! \cdot 3!}$ 的值用正整数表示是多少？答案：$\sqrt{3!\cdot 3!}$ 等于 $\sqrt{(3!)^2}=3!=3\cdot 2\cdot 1=\boxed{6}$。
问题：{proble
m} 答案:

** XSum 中的 5-shot 示例:

###
文章：在圣诞节和除夕庆典期间为期六周的节庆活动中，几乎有一百万人到访这座城市。组织者表示，2014/15 年度有近 89 万人参加了爱丁堡圣诞节活动，为当地经济贡献了 \u00a3199.5m。为期三天的除夕庆典吸引了超过 15 万人，带来了 \u00a341.8m 的经济影响。爱丁堡圣诞节庆典负责人 Charlie Wood 说：“这对爱丁堡来说是个好消息。产生的收入并不会流向活动本身、活动组织者或爱丁堡市议会。“这些钱将流入爱丁堡的各类企业，包括零售、住宿、餐饮、饮品、购物和娱乐业。”

请用 1 句话概括上述文章。
据组织者称，爱丁堡的冬季节庆活动为该市创造了超过 \u00a3241m 的收入。

###
文章：威尔士北部一家企业希望将 PooPrints 服务从美国引入英国，据称最多有 15 个地方议会对这项计划感兴趣。地方议会可以要求问题区域的狗主人登记他们的狗，并采集口腔拭子样本建立数据库。随后，可以从街道、人行道或草地上的粪便中提取 DNA，并在数据库中进行比对。位于登比郡鲁辛的 Streetkleen Bio 董事总经理 Gary Downie 认为，地方政府可以利用《2014 年反社会行为与警务法》赋予的新权力，强制狗主人配合。他说：“这套系统的目的是让公共空间更加清洁、安全。”该公司正在洽谈的地方议会包括伦敦西南部的泰晤士河畔金斯顿、阿伯丁和东柴郡。

狗粪中的 DNA 可用于抓出那些不清理宠物粪便的主人。

###
文章：诺丁汉城堡的作品包括一幅圣安妮的粉笔肖像画、人体和植物的素描，以及一些技术图纸。这位艺术家仅创作了约 20 幅画作

算法 1带剪枝的配置评估器 1：**输入：**配置 x、调优数据 D、目标平均令牌数 $B.i$、有效和无效试验 X

2：**输出：**效用 $U_x(D)$ 若成本 $C_x(D) \leq B.i$；否则为 0。3：**初始化：**根据公式（1）和（2）获取最大的有效 n 和最小的无效 n。 $N \leftarrow x.n,\ n' \leftarrow 0, R \leftarrow []$ **4：如果** $x.n \leq$ **最大有效 n，则 5：** $n \leftarrow x.n$ **6：否则，如果** $x.n \geq$ **最小无效 n，则 7：返回 0 8：否则 9：** $n \leftarrow$ **最大有效 n 10：结束 if 11：当 True 时执行 12：** $x.n \leftarrow n - n', k \leftarrow 1, k' \leftarrow 0$ **13：当 True 时执行 14：对于** $i \in [k', k)$ **执行 15：从大语言模型获取数据点** d_i **的响应，使用配置** x **并追加到** R **16：结束 for 17：** $\rho \leftarrow \begin{cases} 1 - \frac{k}{|D|} & 1 + \frac{1}{k} \quad 2k > |D| \\ 1 - \frac{k-1}{|D|} & 2k \leq |D| \end{cases}$ **( ) 18：如果** $C_x(D_k) > B.i(1 + .1\sqrt{\frac{\rho}{k}})$**，则 19：更新** X_{invalid} **并返回 0 20：结束 if 21：如果** $C_x(D_k) \leq B.i(1 - .1\sqrt{\frac{\rho}{k}})$ **且（**$n < N$ **或** $k = |D|$**），则 22：更新** X_{valid} **23：如果** $n < N$ **则 24：** $R \leftarrow [], n' \leftarrow 0$**，跳出 25：结束 if 26：结束 if 27：如果** $k < |D|$ **则 28：** $k \leftarrow \min(2k, |D|)$ **29：否则 30：跳出 31：结束 if 32：结束 while 33：如果** $n < N$ **则 34：如果** $R \neq []$ **则 35：** $n' \leftarrow n$ **36：结束 if 37：** $n \leftarrow \min(2n, N)$ **38：否则 39：返回** $U_x(D)$ **40：结束 if 41：结束 while**

有效且 X_{invalid}。

$\leftarrow []$

他在世时创作了《蒙娜丽莎》和《最后的晚餐》等作品，但留下了更多的素描。英国皇家收藏中共有近600幅达·芬奇的素描。它们最初装订成一本册子，据认为是查理二世于17世纪购得的。专家认为，列奥纳多的素描是所有艺术家作品中最丰富、涵盖面最广且技艺最精湛的。展览正在诺丁汉城堡博物馆和美术馆展出，持续至10月9日。

请用一句话总结上述文章。
列奥纳多·达·芬奇的珍贵素描作品现已展出，这些作品属于女王皇家收藏。

###
文章：帝亚吉欧集团旗下的 Distill Ventures 表示，该公司已向总部位于墨尔本的 Starward 威士忌投资一笔金额未公开的资金。这是 Distill Ventures 的第二笔威士忌投资。该公司成立的目的是扶持早期品牌并助其发展。上周，该公司宣布投资总部位于丹麦的 Stauning Whisky。帝亚吉欧高端核心烈酒全球负责人 David Gates 表示：“澳大利亚威士忌近年来在全球获得越来越多的认可，这理所当然，而 Starward 已打造出独具特色的威士忌，能够把握这一机遇。”Distill Ventures 联合创始人 Frank Lampen 补充道：“Starward 团队正是我们乐于合作的创业者类型。他们对未来的愿景令人振奋，这笔投资将助力提升其标志性单一麦芽威士忌的产量，并持续推进创新产品线的开发。”去年，按销量计算，帝亚吉欧在苏格兰威士忌市场占有 37%% 的份额。

用一句话概括上文。全球最大的苏格兰威士忌蒸馏商帝亚吉欧投资了一家澳大利亚蒸馏厂，以助其拓展新的出口市场。

###
文章：此前有报道称，坎珀当和凯尔德公园球场出现狗粪污染和损坏情况。人们仍然可以带狗在球场上散步，但如果主人当时正在打球，则不能带狗同行。邓迪休闲与文化局的一位发言人表示，相关规定于4月20日进行了调整。他说：“这项调整反映了许多球员和工作人员的担忧，他们担心狗粪污染和损坏球场，尤其是果岭和沙坑。” “新的管理规定不影响《自由通行法》，并已在球场和邓迪休闲与文化局网站上清楚标明。” “苏格兰大多数高尔夫球场都不允许球员带狗同行。”

用一句话概括上文。
在邓迪的公共球场，因其他球员和工作人员提出投诉，高尔夫球手已被禁止带狗同行。

###
文章：{input}

请用一句话总结上述文章。

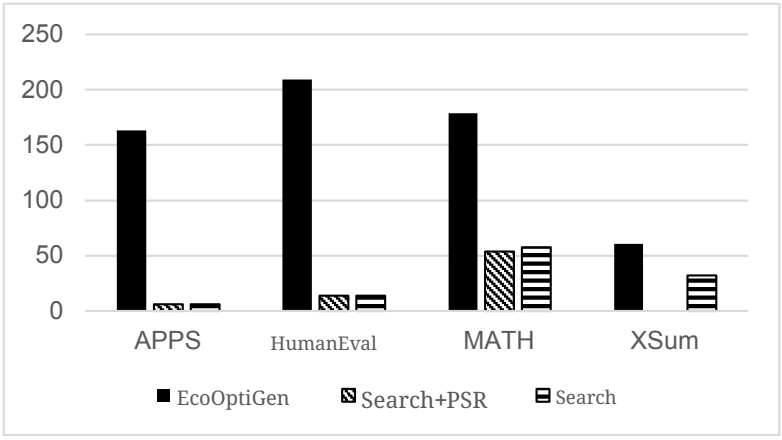


图 *5.* 总优化预算内的试验次数。

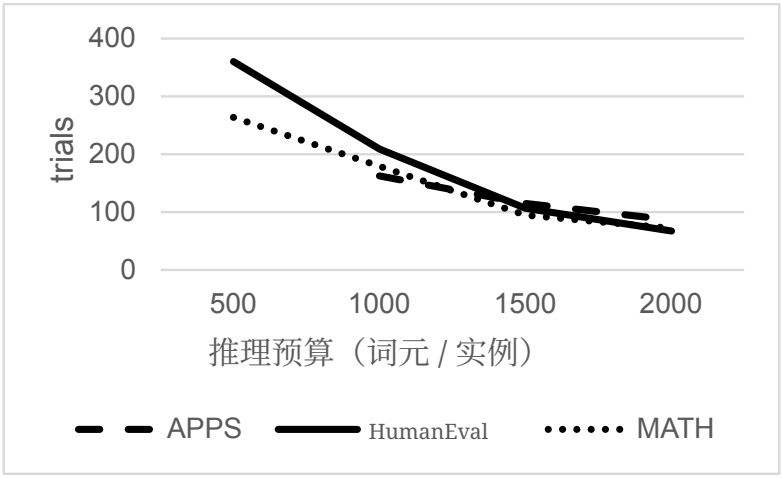


图*6.* 次　　在总优化预算固定为 ($B.o$) 时，不同推理预算 ($B.i$) 下的试验次数　　.

表5. 在总优化预算固定为 $B.o$ 时，使用不同推理预算（$B.i$）的结果。

$B.i$ ($B.o = 1M$)	APPS	HumanEval	MATH
500	-	0.653	0.398
1000	0.05	0.792	0.771
1500	0.10	0.792	0.828
2000	0.07	0.819	0.863

表6. text-davinci-003 的优化超参数配置，采用 $B.i = 2K, B.o = 1M$。

Task	最大令牌数	温度或 top_p	n
APPS	176	top_p: 0.982	15
HumanEval	517	top_p: 0.682	18
MATH	193	温度: 1	26