Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2025) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

International Neural Network Society Workshop on Deep Learning Innovations and Applications

探索温度对大型语言模型的影响：热还是冷?

Lujun Li[a,*], Lama Sleem[a], Niccolo' Gentile[b], Geoffrey Nichil[b], Radu State[a]

[a] *University of Luxembourg*
[b] *Foyer S.A.*

Abstract

采样温度是大型语言模型 (LLMs) 中的一个关键超参数，它在 softmax 层之前修改 logits，从而重塑输出 token 的分布。最近的研究通过证明 LLMs 能够理解语义而不仅仅是记住数据，挑战了“随机鹦鹉”的类比，并指出随机性，由采样温度调节，在模型推理中起着重要作用。在本研究中，我们系统地评估了温度在 0 到 2 范围内对六种不同能力的数据集的影响，并对三种不同大小的开源模型进行了统计分析。小型 (1B-4B)、中型 (6B-13B) 和大型 (40B-80B)。我们的研究结果揭示了温度对模型性能的技能特定影响，凸显了在实际应用中选择最佳温度的复杂性。为了解决这一挑战，我们提出了一种基于 BERT 的温度选择器，利用这些观察到的效果来识别给定提示的最佳温度。我们证明这种方法可以显著提高中小型模型在 SuperGLUE 数据集中的性能。此外，我们的研究扩展到 FP16 精度推理，揭示了温度效应与在 4 比特量化模型中观察到的一致。通过评估三个量化模型中高达 4.0 的温度效应，我们发现“突变温度”——发生显著性能变化的点——随着模型大小 [1] 的增加而增加。

Keywords: Large Language Models; Sampling Temperature; Model Performance Evaluation; BERT-based Classifier; GPT-based Evaluation

1. 介绍

自从 ChatGPT 发布以来，大型语言模型（LLMs）显著地影响了学术界和工业界，彻底改变了人工智能的发展。不同规模的开源模型促进了多个领域的进步，[27] 包括问答和摘要。影响 LLMs 性能的一个关键因素是超参数调整。例如，Top-K 采样从 K 个最可能的候选项中选择下一个令牌，而 Top-P 采样则从其累积概率超过 P [9] 的最小令牌集合中进行采样。此外，重复惩罚减少已经出现的令牌的概率，有助于避免重复。在本文中，我们关注于温度，这是使用最频繁的超参数之一。在 LLMs 的推理过程中，该参数用于缩放输出层的 logits，有效地控制模型预测的随机性。温度的概念，记作 T [1] ，最初由 Ackley 提出，他强调其在塑造 Boltzmann 分布中的关键作用。形式上，i -th 令牌的概率 P_i 由 $P_i = \frac{e^{y_i/T}}{\sum_{j=1}^{V} e^{y_j/T}}$ 给出，其中 y_i 表示 i -th 令牌的预 softmax 激活（通常称为 logit），T 代表温度，V 是词汇表中的令牌总数。值 P_i 确定 i -th 令

[1] https://github.com/DobricLilujun/temperature_eval

* Corresponding author. Tel.: +0033-766636416.
E-mail address: lilujun588588@gmail.com

牌将被生成的概率，之后模型输出由采样算法产生。随着 T 增加，概率质量函数（PMF） [8] 变得更加均匀；相反，当 T 趋近于零时，分布坍缩为一个 函数，使得算法总是选择最可能的令牌并表现出贪婪行为。在每个生成步骤中，通过随机采样更新后的概率分布 [13] 来选择新令牌。

在这项研究中，我们关注三个关键研究问题（RQ）：（RQ1）温度在多大程度上影响 LLMs 在不同能力上的表现？（RQ2）温度对不同能力和模型是否有均一的影响，并观察到的主要差异是什么？（RQ3）是否存在针对每项能力的最佳温度，并且是否可以为特定提示确定最佳温度？本文的剩余部分组织如下：第 2 节回顾相关工作。第 3 节描述实验方法。第 4 节详细说明实验设置。第 5 节展示并分析结果。最后，第 6 节总结全文。

2. 相关工作

最近的文献中对温度影响的研究仍然有限。大多数研究只使用一个温度值进行报告，而没有系统地探索更广泛的范围，除了 Llama 模型系列在代码生成测试中使用了两个设置 [24] 。此外，旨在同时评估多种模型能力的一般数据集，例如用于人工通用智能（AGI; [28] ）的数据集，往往缺乏对单一基础能力的具体关注。例如，更高的温度增加创造力 [29] ，而较低的温度改善逻辑推理。对于像复杂数学问题这样需要逻辑和创造力的任务，这些影响可能会相互抵消。这使得很难看到温度的真正影响，我们称这种现象为“温度悖论”。

最近的研究检查了温度设置如何影响不同的任务和动态配置。[22] 使用提示工程和 0 到 1 的范围在多任务场景中探索了温度，发现对 LLM 的性能没有显著影响。[19] 在创意写作中研究了温度，测量困惑度和余弦相似性，发现对创造力只有微弱的影响。[29] 提出了一种用于代码生成的自适应温度策略，将较高的温度分配给较难的标记（如 Python 函数的开始），而将较低的温度分配给模型置信度更高的标记，显示出较高的温度可以帮助复杂任务。然而，尽管温度调整对 LLM 用户、RAG 系统和代理 AI 系统很重要，目前仍没有为不同的 LLMs、任务或提示选择温度的明确指导方针。

3. 方法

为了解决“温度悖论”并以最小的偏倚更准确地衡量温度对每种能力的影响，我们采用具有明确能力偏好的数据集，并使用单一提示格式，仅查询模型一次以避免多提示协助。这种方法能够更精确和无偏倚地评估大型语言模型的内在能力。我们假设温度以不同的方式影响不同的模型能力。因此，我们的研究集中于六种核心内在能力，这些能力不仅代表大型语言模型的主要能力，也是计算语言学研究的核心。

3.1. 评估内在能力

因果推理 (CR)：一种认知能力，历史上仅被归于人类，指的是通过遵循严格的逻辑原则，从给定的前提中推导出结论 [14] 。在本文中，我们使用 CRASS [10] ，这是一种公开可用的反事实推理数据集，简化了评估过程，要求模型选择正确答案而不是生成答案。Top-1 精确度 (T1) 被用于量化模型在多次重复后通过选择置信度最高的类别正确预测真实类别的频率。创造力 (CT)：创造力涉及生成新颖且有价值的想法、概念或产品，需具备原创性和有效性 [23] 。对于 CT，我们采用评估故事的四个维度的框架：流畅性、灵活性、原创性和详尽性，使用基于 Torrance 创意写作测试 (TTCW) 的程序定制的问题 [7] 。在此框架中，每个类别包含多个标准评估问题，真假由专家判断确定。十二篇公开可获取的《纽约客》故事被用作情节，然后通过计算所有 Q & A 对中正面评价的数量来计算 TTCW 精确度 [12] 。上下文学习 (ICL)：ICL 反映了大型语言模型理解文本并使用上下文信息和少量示例执行任务的能力 [21] 。在本研究中，我们关注 LongBench-TREC 长上下文任务 [3] ，该任务中模型从一系列问题和答案中学习，必须基于此上下文对最后一个问题进行分类。分类分数 (CLS)，通过准确性测量，评估模型识别模式和做出正确预测的能力，与真实值进行比较。

指令跟随 (IF)：IF 衡量模型在提示中跟随指令的能力，这对于有效的 LLM 应用至关重要。在本研究中，我们使用了 InfoBench [20] ，它引入了分解需求跟随比率 (DRFR) 作为评估指令跟随性能的指标。DRFR 将复杂指令分解以进行更细致的评估，并已证明具有更高的可靠性和有效性。

机器翻译（MT）：MT 评估大型语言模型（LLM）在不同语言之间翻译文本的能力，这是 LLM 表现出强劲性能的关键领域。我们使用 FLORES-101 基准 [11] 进行多语言评估，采用 spBLEU 作为指标。BLEU 分数衡量模型输出与参考译文之间的相似性。为了确保可比性，我们通过除以 100 来规范化 spBLEU 分数，

使所有结果均在 [0, 1] 范围内。鉴于英文在 LLM 训练数据中的普遍性，我们专注于从英语到其他语言的翻译，选择不同类型的组合（例如，英语到马耳他语、印尼语、拉脱维亚语、冰岛语和高棉语），以涵盖不同级别的翻译难度。

摘要（SUMM）：摘要旨在将长文本压缩为简明的总结，同时保留关键信息和主要思想。此任务的一个主要挑战是可靠的评估。为了解决这个问题，我们使用 “benchmark_llm_summarization” 数据集 [26] ，该数据集提供了专家撰写的参考摘要。在评估中，我们采用基于参考的 Rouge-L F1 度量，该度量衡量生成的摘要与参考摘要之间最长公共子序列的重叠情况，平衡精确度和召回率，并已被证明与人工判断具有良好的相关性 [16] 。

3.2. 大型语言模型作为评审评估

Table 1: Selected datasets and evaluation metrics

Ability	Dataset	Samples	Metrics	Source	Evaluations
CR	CRASS	3500	Top-1 accuracy	[10]	GPT3.5
CT	Creativity_eval	84	TTCW Accuracy	[7]	GPT3.5
ICL	LongBench-TREC	1015	CLS score	[3]	Exact Matching
IF	InfoBench	3500	DRFR	[20]	GPT3.5
MT	Flore101	2100	Normalized spBLEU	[11]	SPM tokenizers
SUMM	benchmark_llm_summarization	2114	Rouge-L F1 Score	[26]	Exact Matching

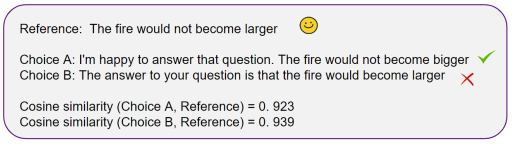

Fig. 1: Cosine Similarity Using BERT Embedding Model

“LLM 作为判断者” 已被广泛使用，并被证明与人类判断高度一致 [18] 。由于 LLM 固有的随机性和文本表达灵活性，即使在小型和中型模型中，往往会在目标回应之前或之后出现无意的结果。这使得基于参考的评估，例如精确匹配或相似度指标，尤其具有挑战性。例如，相似度指标有两个主要问题：(1) 很难为正确答案设定明确的分界线；(2) 基于嵌入的方法往往会遗漏关键词，例如 “不是”，如图 1 [2] 所示。

高级模型如 GPT-4o 和 DeepSeek [5, 4] 在许多基准测试中取得了强劲的结果。对于复杂答案或评估具有挑战性的任务，我们使用 LLM-as-a-Judge。对于 CR、CT 和 IF，我们使用精心设计的提示对 ChatGPT 进行代替精确匹配或人工注释。这些判断用于计算第 3.1 节中描述的指标。对于具有简单参考答案的其他三种能力，我们使用标准评估方法，如表 1 所示。

4. 实验

4.1. 总体实验设置

在本研究中，实验从选择旨在挑战最新技术（SOTA）模型的多样化基准数据集开始，如表 2 所示。评估主要通过问答格式或匹配函数进行，并对每个数据集应用特定指标。所有模型都采用 AWQ 方法 [17] 量化为 4 位，并使用 vLLM [15] 作为默认的推理加速框架。每个问题在 12 个模型中被测试三次。

Small Size Models (1B - 4B)			Medium Size Models (6B - 13B)			Large Size Models (40B - 80B)		
Model	Size	Date	Model	Size	Date	Model	Size	Date
Llama-3.2-1B-Instruct	1.2B	Sep 2024	Llama-2-7b-chat-hf	6.7B	Jul 2023	Llama-2-70b-chat-hf	69.0B	Jul 2023
Llama-3.2-3B-Instruct	3.2B	Sep 2024	Llama-2-13b-chat-hf	13.0B	Jul 2023	Meta-Llama-3-70B-Instruct	70.6B	Apr 2024
Phi-3.5-mini-instruct	3.8B	Jun 2024	Mistral-7B-Instruct-v0.2	7.2B	Mar 2024	Mixtral-8x7B-Instruct-v0.1	46.7B	Dec 2023
Qwen2.5-1.5B-Instruct	1.5B	Sep 2025	Meta-Llama-3-8B-Instruct	8.0B	Apr 2024	–	–	–
Qwen2.5-3B-Instruct	3.1B	Sep 2025	–	–	–	–	–	–

Table 2: 调查了具有各自尺寸和发布时间的小型、中型和大型模型。

温度设置范围从 0.1 到 1.9，递增为 0.3，导致每个模型有七种不同的配置。排除了高于 2.0 的温度，因为先前的研究显示较高的值往往会产生没有信息和过于不连贯的文本 [6] 。每个模型只用一个测试问题进行评估。我们始终使用 gpt-3.5-turbo-0125 作为评估模型，温度设置为 0.01，并选择了如表格 2 所列的开源模型。采用了核采样，因为它产生的混淆度值最接近于人类文本 [13] ，具体参数如下：max_length = 4096, Top_P = 0.9, 重复惩罚（RP）= 1.0, 和 max_new_tokens = 1024。

4.2. 补充实验

4.2.1. SuperGLUE 上最佳温度选择

SuperGLUE 是一个由八个任务组成的基准，用于评估词义消歧、自然语言推理、指代消解和问答。上一节的主要结果可以用来确定给定提示和模型的最佳温度，前提是已知提示所需的主要能力。为此，提出了一个基于微调 BERT 框架的分类模型，称为“基于 BERT 的选择器”，该模型在专为主要实验定制的实验提示上进行训练，并最终用于对每个输入提示进行分类。另一种选择是基于将向 GPT 模型询问的精心设计的提示来获得所需的提示能力，这种方法被称为“基于 GPT 的选择器”。

该选择器的基本思想如下：设 F 为一个选择模型，可以是 BERT 或 GPT，基于给定的提示 x_p 充当能力预测器。模型输出 $F(x_p)$ 表示与给定提示 x_p 最密切相关的预测能力。例如，给定一个训练提示，如 x_p = “将 ‘J’aime le chat’ 从法语翻译成英语”，模型 $F(x_p)$ 预计会预测 “MT”（机器翻译）为最需要的能力。因此，选择最佳的温度参数 $T^* = \arg\max_T \mathcal{D}(T, F(x_p), M)$ ，以在不同温度设置下最大化模型 M 在任务 $F(x_p)$ 上的估计性能，其中 $\mathcal{D}()$ 表示从之前主要实验结果中获得的给定模型 M 在不同温度下的性能分布。

为了测试这个框架，我们评估了三个不同大小的模型——Llama-3.2-1B-Instruct（小）、Llama-3-8B-Instruct（中），和 Mixtral-8x7B-Instruct-v0.1（大） [25] ——在 SuperGLUE 基准测试上的表现。基准测试中的每个问题都被问了三次，每个实例都是通过递增随机种子（一开始设为 42）生成的，每次迭代增加 1。

4.2.2. 扩展设置实验

我们还在保持 4-bit 量化的同时研究了温度范围 $[0, 4]$ 。虽然我们观察到在高于 2 的温度时性能下降和生成不一致的现象，但我们没有发现大型模型中特定 “突变温度” 的证据。为了进一步探讨推理精度的影响，我们在同样的三个模型上使用 FP16 精度重复了主要实验，重点考察了从 0 到 2 的温度范围，以确定当推理精度改变时温度影响是否不同。此外，由于 Top- K 和 Top- P 采样在候选选择层面影响输出分布，而 RP 则在 logits 层面操作，我们进一步评估了 Top- K （2, 5, 10）、Top- P （0.8, 0.9, 1.0）和 RP（0.0, 1.0, 2.0）的各种设置。这些实验在上述所有三个模型上进行，以系统评估这些参数对性能的影响。

5. 结果与分析

5.1. 统计分析结果

表 3 总结了基于三类模型结果的六种能力的温度-性能相关性。“P. Coef.” 和 “S. Coef.” 分别对应皮尔逊和斯皮尔曼相关系数。“Range (%)Max” 可以解释为不同温度下相对性能的变化，而 “Range Max (%)” 指的是尺寸类别内最大相对变化范围。“CV” (变异系数) 表示平均性能与标准差的比率，而 “CV Max” 表示在尺寸类别内观察到的最高变异系数。每组中温度和模型的平均准确率和标准差也被报告。

Table 3: 三种模型类别中六种能力的温度-性能相关性的比较。

Ability	P. Coef.	P. p-value	S. Coef.	S. p-value	Range Max (%)			CV Max			Average Accuracy			Standard Deviation		
					Small	Medium	Large	Small	Medium	Large	Small	Medium	Large	Small	Medium	Large
CR	-0.07	0.00	-0.07	0.00	146.02	49.37	19.41	58.79	14.88	6.43	0.41	0.52	0.82	0.05	0.03	0.02
CT	-0.14	0.00	-0.10	0.00	186.81	154.55	82.02	82.64	72.90	28.07	0.36	0.45	0.47	0.27	0.12	0.08
ICL	-0.10	0.00	-0.09	0.00	122.04	55.52	20.19	48.83	21.66	7.20	0.38	0.26	0.49	0.06	0.04	0.01
IF	-0.40	0.00	-0.37	0.00	154.65	116.63	22.03	72.22	47.64	8.04	0.49	0.68	0.73	0.26	0.08	0.02
MT	-0.216	0.00	-0.40	0.00	192.32	162.59	76.86	91.09	72.14	27.35	4.72	5.95	11.55	3.19	2.54	1.96
SUMM	-0.51	0.00	-0.45	0.01	154.29	89.20	4.35	72.89	32.70	1.57	0.16	0.21	0.23	0.09	0.02	0.00

在这个表中可以观察到，IF、MT 和 SUMM 的性能与温度之间表现出较强的相关性，如两个相关系数所示。这些相关性的统计显著性得到了零 p 值的进一步支持。此外，随着模型规模的增大，“范围最大值” 和 “CV 最大值” 均减少，这在统计上表明较大的模型对温度诱导的变化更加稳健。平均准确性指标进一步显示，较大的模型在所有六种能力上均达到更高的统计性能。特别是，对于 CT、IF 和 SUMM，不同规模模型之间的性能差异相对较小，但对于 CR、ICL 和 MT，这种差异更加显著。这些发现为根据具体功能需求选择模型规模提供了实际指导。

5.2. 温度对不同能力的影响

图 2 展示了温度对不同大小模型在各种评估能力上的影响。线条表示每个模型大小的平均性能，而阴影区域对应于 ±0.2 个标准差。为了保持一致性，表 1 中列出的所有评估指标——包括机器翻译的 spBLEU——都统一称为“准确度”，并已归一化至区间 $[0, 1]$ 。

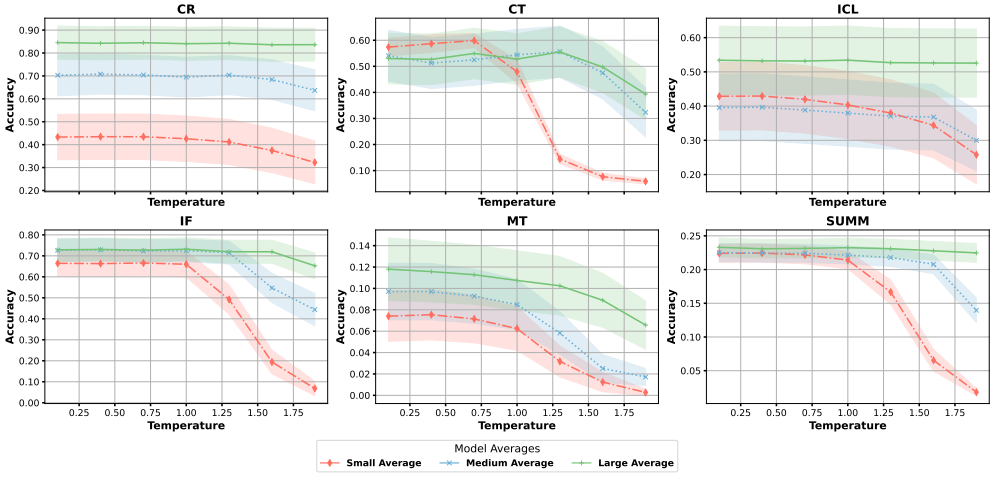

Fig. 2: 不同模型规模的平均性能趋势，阴影区表示变异性。

因果推理 (CR). CR 问题是违背直觉的，并且需要逻辑推理，每个问题有三个选项。中型和大型模型超过了 33.3 % 的随机基准线，而小型模型的表现仅略高于这个概率水平——大约高出 7 %——在大多数温度设置下，表明其推理能力有限。大型和中型模型在温度为 1.3 时显示出轻微的改善，提示更高的温度可能有助于解决复杂问题。CR 的最佳温度并不总是零，而且温度的升高不一定会降低性能。相反，在本研究范围内，小型模型并未表现出显著的因果推理能力。

创造力 (CT). 建议中大型模型设定在 1.3 的最佳温度以最大化创造力。小型模型在 $T = 1.0$ 时显示出明显的创造力下降，而中大型模型只在 $T = 1.7$ 时受到影响。通常情况下，温度首先会增加然后减少创造力。小型模型在较低温度时更具创造力，而大型模型则更稳定并产生更多样化的输出，如其更宽的阴影区域所示。总的来说，温度对创造力有显著影响，中等值最为有益。

上下文学习 (ICL). 大型模型获得最佳的平均性能，而中型和小型模型之间的差异最小。这表明，作为大型语言模型的一种新兴特性，ICL 需要足够大的模型尺寸，这突显了规模法则的重要性。中型模型的性能下降比小型模型更少。在温度为 1.7 时，小型模型比中型模型退化得更快，尽管在较低温度下表现优于中型模型。大型模型在 0 到 2 的温度范围内保持稳定的性能，未观察到性能的突然下降。尽管大型模型在较高温度下可能会有轻微的性能提升，提高温度通常会降低性能。

指令遵循 (IF). IF 的行为尤其值得注意。随着温度从 0 增加到 1，IF 的性能几乎没有变化。然而，当温度超过 1 时，不同模型会经历相对明显的负面影响，且模型尺寸越大，这些负面影响出现的时间越晚。随着温度的变化，性能改变是突然的：小型模型在 1.0 到 1.3 之间表现出突变，中型模型在 1.3 到 1.6 之间，而大型模型则在 1.6 到 1.9 范围内表现出温和的突变。因此，对于需要严格遵循指令的 LLM 用户，建议将温度设置在 1 以下。

机器翻译 (MT). 在较低范围内稍微增加温度仅略微改善了小型和中型模型的翻译性能。温度的升高对机器翻译的影响是最具破坏性的，如表格 3 中所示的性能和 CV 的最高范围所示。这种趋势可以归因于翻译固有的确定性性质，所有模型在性能上的下降均是相似的。最佳温度接近于零（$0 + \epsilon$），而语言理解性能主要依赖于训练数据的广度和模型的参数规模。

总结（*SUMM*）。温度效应曲线在初始阶段是稳定的，但在较高温度下急剧下降，尤其是对于小型模型。统计分析显示性能与温度之间存在很强的负相关性。SUMM 任务与 IF 任务呈现类似趋势，但中型模型的突变温度更高（大约 1.7 ），而大型模型没有明显的突变温度。

5.3. 补充实验

5.3.1. SuperGLUE 上的最佳温度选择

我们在三个模型上进行了实验，如表 4 所示。该表显示了在不同温度设置下的 SuperGLUE 验证精度：ACC_D 表示默认温度为 1.0 时的准确率，ACC_B 和 ACC_C 分别表示通过动态选择最佳温度使用我们微调的 BERT 模型和基于 ChatGPT 的提示实现的准确率。此比较清楚地显示了从最佳温度选择中得到的性能差异。

Table 4: 默认和动态选择温度设置下的 SuperGLUE 验证准确率。

Model	Type	COPA	WIC	WSC	Average
Llama-3.2-1B-Instruct	ACC_D	0.510	0.196	0.346	0.252
	ACC_B	0.600	0.500	0.356	0.494
	ACC_C	0.600	0.477	0.365	0.477
Meta-Llama-3-8B-Instruct	ACC_D	0.860	0.547	0.673	0.600
	ACC_B	0.900	0.556	0.673	0.612
	ACC_C	0.900	0.549	0.664	0.605
Mixtral-8x7B-Instruct-v0.1	ACC_D	0.800	0.608	0.298	0.593
	ACC_B	0.800	0.608	0.298	0.593
	ACC_C	0.800	0.608	0.298	0.593

调整温度可以大大提高 Llama-3.2-1B 和 Meta-Llama-3-8B-Instruct 在 WIC（SuperGLUE 任务之一）上的表现。这表明最优温度选择器提供了稳定的性能，避免了使用固定温度时可能出现的性能下滑。在处理小模型时，这确实是需要考虑的必要参数之一，尤其是在资源受限的场景中。考虑到 SuperGLUE 主要评估一系列不同的能力，我们的最优温度选择器仍能展示出一致的改进。需要指出的是，这个选择器并不能从根本上提高性能——监督微调（SFT）仍然是主要的方法——但它通过避免次优设置来确保模型达到最佳可能性能。对于大模型，我们没有观察到显著的性能差异，这表明优化温度设置对较大模型来说一般不太关键。然而，正如之前的发现所建议的，当大型模型用于解决复杂推理任务时，较高的温度有时可以带来性能提升。因此，在这种情况下，调整温度可能仍然是必要的。

5.3.2. 扩展设置的结果

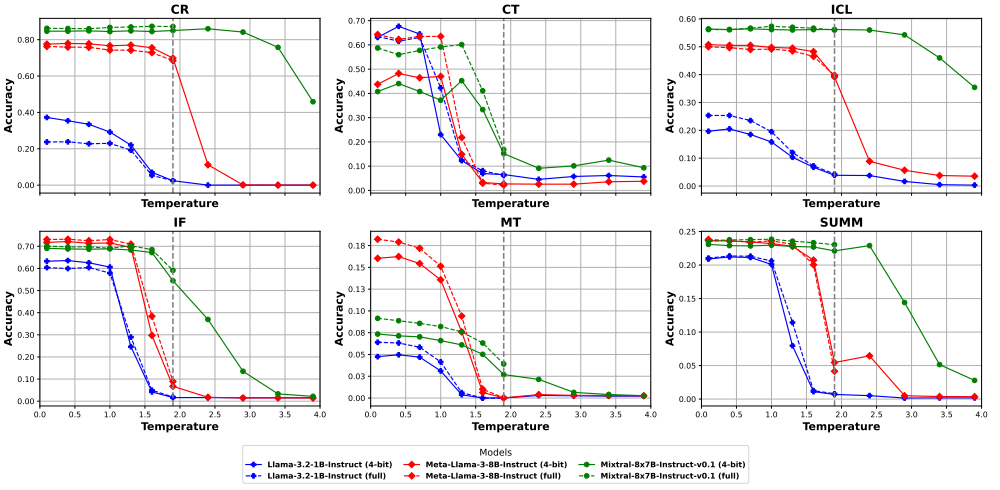

Fig. 3: 性能在温度扩展到 4.0 时

图 3 展示了随着温度变化，4-bit 精度模型在六项能力上的性能曲线；“Full”指的是 FP16 精度推理。扩展温度范围有助于识别“突变温度”（即性能急剧下降的温度）和温度设置的潜在上限。大模型对于温度的

增加更具鲁棒性，而小模型在较低温度下经历显著的性能损失，这与我们的预期一致。尤其是，每个模型都有其自己的突变温度，而大模型通常表现出更高的阈值。

FP16 精度. 如图所示，3 显示，模型的最佳温度在整个温度范围内没有出现显著变化。总体上，结果表明在两个精度级别之间的最佳温度没有实质性区别，尽管性能存在 10 % –20 % 的差异。由于我们的主要实验是在 4 位量化下进行的，这些发现也验证了我们结果的有效性，并展示了它们在推理精度方面的可扩展性。

Experiment	CR			CT			ICL			IF			MT			SUMM		
	p_1	p_2	p_3	p_1	p_2	p_3	p_1	p_2	p_3	p_1	p_2	p_3	p_1	p_2	p_3	p_1	p_2	p_3
(Top-P, RP) = (0.9, 1.0) ; Top-K ∈ { 2, 5, 10 }	1.00	0.99	1.00	0.27	0.46	0.57	1.00	0.99	0.99	0.90	0.86	0.99	0.98	0.98	0.99	0.86	0.83	0.99
(Top-K, RP) = (5, 1.0) ; Top-P = [0.8, 0.9, 1.0]	1.00	0.99	1.00	0.53	-0.29	-0.33	0.98	0.96	0.98	0.98	0.91	0.91	0.98	0.99	0.99	0.97	0.97	0.99
(Top-P, Top-K) = (0.9, 5) ; RP = [0.0,1.0,2.0]	1.00	1.00	1.00	0.18	-0.08	0.47	1.00	1.00	1.00	0.86	0.58	0.90	0.99	0.99	1.00	0.98	0.99	0.99

Table 5: 具有不同参数值的性能曲线的成对皮尔逊相关系数。

其他参数. 我们通过计算产生的性能曲线之间的皮尔逊相关系数，分析了变化单个参数的效果，同时保持所有其他参数不变。具体而言，对于每个参数，我们选择三个不同的值，并计算它们对应的温度-性能曲线之间的成对皮尔逊相关系数。这些值 p_1 、p_2 和 p_3 分别表示第一个与第二个、第一个与第三个、以及第二个与第三个参数设置的相关性。这些结果总结在表 5 中。可以清楚地观察到，除 CT 之外，不同参数设置之间温度效应的相关性非常高，表明影响最小。相反，RP 对 CT 和 IF 有显著影响。对于 CT，Top- K 的影响小于 RP，而 Top- P 具有最大的影响，这可以从相对较低甚至负的相关系数中看出。因此，当模型需要执行创造性任务时，必须适当地设置 Top- P ，并且 RP 也应小心配置。然而，这三个参数对温度性能曲线的影响最小，仅在 CT 和 IF 任务中需要特别考虑。

6. 结论

在本研究中，我们通过检查六种能力在更广泛的温度范围内的表现，扩展了先前关于温度效应在语言模型中的工作的研究。我们还改进了 GPT 评估方法和测试协议，重点关注温度效应的统计和实证分析。我们还比较了模型在 FP16 和 4 位量化下的性能，发现温度效应差异最小。通过将温度范围扩展到 4.0，我们观察到大模型仍然有一个突变温度，但高于 2.0。此外，我们引入了基于 BERT 和 GPT 的温度选择器，展示了它们在 SuperGLUE 数据集上的有效性，特别是对于小型模型。

为回答 RQ1，“温度在多种能力下对 LLMs 的性能影响到什么程度?” 我们发现温度对上下文学习和因果推理有适度影响，但对小模型中的机器翻译（最多 192.32 %）、创造力（最多 186.81 %）有显著影响（见表 3 ）。斯皮尔曼系数表明温度与性能之间普遍存在负相关关系。

为了解答 RQ2 “温度对不同模型中的能力是否产生一致的影响，以及观察到的主要差异是什么?” 我们发现大型模型对温度变化更具弹性。温度升高稍微提高了因果推理、上下文学习和指令遵循能力，但随后性能下降。对于摘要和机器翻译，较高的温度通常会产生负面影响，尤其是对于较小的模型而言。温度的影响在不同的能力和模型大小之间变化很大，很难总结出一致的模式；最佳温度在不同模型之间也有显著差异。此外，较高的温度并不总是对因果推理、指令遵循和上下文学习等逻辑导向能力有害。

为了回答 RQ3：“是否每种能力都有一个最佳温度，并且能否为特定提示找到最佳温度?”，我们发现没有单一温度对所有任务都是最佳的。更高的温度并不总是对创意写作最优，零温度也并不总是最适合遵循指令。然而，通过使用我们的 BERT 模型识别每个提示所需的能力，并参考我们的实验结果，我们可以为每个提示选择最佳温度。在三个 SuperGLUE 任务中，这种方法提高了小型和中型模型的性能。

本研究存在一些局限性。我们仅考察了温度对六种基于文本的能力的影响；其他能力，如计划和编码，也可以在未来的工作中进行评估。若不对每个案例进行人工核查，很难验证 GPT 模型评估的准确性。需要更多测试来确定最佳温度选择器是否对其他相似或更大规模的模型有效。此外，还需进一步研究以数学方式解释温度如何影响模型性能。

7.

致谢

这项研究极大地受益于我们工业合作伙伴 Foyer S.A. 的合作和支持，以及学术机构的支持，特别是安全、可靠性和信任跨学科中心（SnT）和许多个人贡献者。我们要感谢“AI & 数据工作室”团队，他们的宝贵见解、指导以及提供的基本计算资源（NVIDIA H100 GPUs），这些对于进行我们的实验至关重要。由教师和顾问提供的指导和辅导，再加上博士后研究人员分享的建设性建议和技术专长，显著提高了这项工作的质量和影响。此外，我们仅使用了公开可用的数据集和模型，并确认在准备这篇手稿时没有使用 AI 生成的文本。在整个研究过程中，我们遵守了道德标准并保持透明，确保我们的所有方法和结果得到清晰沟通。

References

[1] Ackley, D.H., Hinton, G.E., Sejnowski, T.J., 1985. A learning algorithm for boltzmann machines. Cognitive Science 9, 147–169. URL: https://www.sciencedirect.com/science/article/pii/S0364021385800124, doi:https://doi.org/10.1016/S0364-0213(85)80012-4.

[2] Anschütz, M., Miguel Lozano, D., Groh, G., 2023. This is not correct! negation-aware evaluation of language generation systems, in: Keet, C.M., Lee, H.Y., Zarrieß, S. (Eds.), Proceedings of the 16th International Natural Language Generation Conference, Association for Computational Linguistics, Prague, Czechia. pp. 163–175. URL: https://aclanthology.org/2023.inlg-main.12/, doi:10.18653/v1/2023.inlg-main.12.

[3] Bai, Y., Lv, X., Zhang, J., Lyu, H., Tang, J., Huang, Z., Du, Z., Liu, X., Zeng, A., Hou, L., Dong, Y., Tang, J., Li, J., 2024. LongBench: A bilingual, multitask benchmark for long context understanding, 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), Association for Computational Linguistics, Bangkok, Thailand. pp. 3119–3137. URL: https://aclanthology.org/2024.acl-long.172/, doi:10.18653/v1/2024.acl-long.172.

[4] Bi, X., Chen, D., Chen, G., Chen, S., Dai, D., Deng, C., Ding, H., Dong, K., Du, Q., Fu, Z., et al., 2024. Deepseek llm: Scaling open-source language models with longtermism. arXiv preprint arXiv:2401.02954 .

[5] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D., 2020. Language models are few-shot learners, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA.

[6] Caccia, M., Caccia, L., Fedus, W., Larochelle, H., Pineau, J., Charlin, L., 2018. Language gans falling short. CoRR abs/1811.02549. URL: http://arxiv.org/abs/1811.02549, arXiv:1811.02549.

[7] Chakrabarty, T., Laban, P., Agarwal, D., Muresan, S., Wu, C.S., 2024. Art or artifice? large language models and the false promise of creativity, in: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA. URL: https://doi.org/10.1145/3613904.3642731, doi:10.1145/3613904.3642731.

[8] Chang, C.C., Reitter, D., Aksitov, R., Sung, Y.H., 2023. Kl-divergence guided temperature sampling. arXiv:2306.01286.

[9] Fan, A., Lewis, M., Dauphin, Y., 2018. Hierarchical neural story generation, in: Gurevych, I., Miyao, Y. (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia. pp. 889–898. URL: https://aclanthology.org/P18-1082/, doi:10.18653/v1/P18-1082.

[10] Frohberg, J., Binder, F., 2022. CRASS: A novel data set and benchmark to test counterfactual reasoning of large language models, in: Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., Piperidis, S. (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France. pp. 2126–2140. URL: https://aclanthology.org/2022.lrec-1.229.

[11] Goyal, N., Gao, C., Chaudhary, V., Chen, P.J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzmán, F., Fan, A., 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. Transactions of the Association for Computational Linguistics 10, 522–538. URL: https://aclanthology.org/2022.tacl-1.30/, doi:10.1162/tacl_a_00474.

[12] Guzik, E.E., Byrge, C., Gilde, C., 2023. The originality of machines: Ai takes the torrance test. Journal of Creativity 33, 100065. URL: https://www.sciencedirect.com/science/article/pii/S2713374523000249, doi:https://doi.org/10.1016/j.yjoc.2023.100065.

[13] Holtzman, A., Buys, J., Du, L., Forbes, M., Choi, Y., 2020. The curious case of neural text degeneration, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net. URL: https://openreview.net/forum?id=rygGQyrFvH.

[14] Kiciman, E., Ness, R.O., Sharma, A., Tan, C., 2024. Causal reasoning and large language models: Opening a new frontier for causality. Transactions on Machine Learning Research (TMLR) URL: https://www.microsoft.com/en-us/research/publication/causal-reasoning-and-large-language-models-opening-a-new-frontier-for-causality/. selected for presentation at ICLR 2025.

[15] Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C.H., Gonzalez, J., Zhang, H., Stoica, I., 2023. Efficient memory management for large language model serving with pagedattention, in: Proceedings of the 29th Symposium on Operating Systems Principles, pp. 611–626.

[16] Lin, C.Y., Och, F.J., 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics, in: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), Barcelona, Spain. pp. 605–612. URL: https://aclanthology.org/P04-1077, doi:10.3115/1218955.1219032.

[17] Lin, J., Tang, J., Tang, H., Yang, S., Xiao, G., Han, S., 2025. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. GetMobile: Mobile Comp. and Comm. 28, 12–17. URL: https://doi.org/10.1145/3714983.3714987, doi:10.1145/3714983.3714987.

[18] Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., Zhu, C., 2023. G-eval: NLG evaluation using gpt-4 with better human alignment, in: Bouamor, H., Pino, J., Bali, K. (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore. pp. 2511–2522. URL: https://aclanthology.org/2023.emnlp-main.153/, doi:10.18653/v1/2023.emnlp-main.153.

[19] Peeperkorn, M., Kouwenhoven, T., Brown, D., Jordanous, A., 2024. Is temperature the creativity parameter of large language models?, in: Grace, K., Llano, M.T., Martins, P., Hedblom, M.M. (Eds.), Proceedings of the 15th International Conference on Computational Creativity, ICCC 2024, Jönköping, Sweden, June 17-21, 2024, Association for Computational Creativity (ACC). pp. 226–235. URL: https://computationalcreativity.net/iccc24/papers/ICCC24_paper_70.pdf.

[20] Qin, Y., Song, K., Hu, Y., Yao, W., Cho, S., Wang, X., Wu, X., Liu, F., Liu, P., Yu, D., 2024. InFoBench: Evaluating instruction following ability in large language models, in: Ku, L.W., Martins, A., Srikumar, V. (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand. pp. 13025–13048. URL: https://aclanthology.org/2024.findings-acl.772/, doi:10.18653/v1/2024.findings-acl.772.

[21] Radford, A., Narasimhan, K., 2018. Improving language understanding by generative pre-training. URL: https://api.semanticscholar.org/CorpusID:49313245.

[22] Renze, M., 2024. The effect of sampling temperature on problem solving in large language models, in: Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics. p. 7346–7356. URL: http://dx.doi.org/10.18653/v1/2024.findings-emnlp.432, doi:10.18653/v1/2024.findings-emnlp.432.

[23] Runco, M., 1988. Creativity research: Originality, utility, and integration. Creativity Research Journal - CREATIVITY RES J 1, 1–7. doi:10.1080/10400418809534283.

[24] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al., 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 .

[25] Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S., 2019. Superglue: A stickier benchmark for general-purpose language understanding systems, in: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf.

[26] Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., Hashimoto, T.B., 2024. Benchmarking large language models for news summarization. Transactions of the Association for Computational Linguistics 12, 39–57. URL: https://aclanthology.org/2024.tacl-1.3/, doi:10.1162/tacl_a_00632.

[27] Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.Y., Wen, J.R., 2023. A survey of large language models. arXiv:2303.18223.

[28] Zhong, W., Cui, R., Guo, Y., Liang, Y., Lu, S., Wang, Y., Saied, A., Chen, W., Duan, N., 2024. AGIEval: A human-centric benchmark for evaluating foundation models, in: Duh, K., Gomez, H., Bethard, S. (Eds.), Findings of the Association for Computational Linguistics: NAACL 2024, Association for Computational Linguistics, Mexico City, Mexico. pp. 2299–2314. URL: https://aclanthology.org/2024.findings-naacl.149/, doi:10.18653/v1/2024.findings-naacl.149.

[29] Zhu, Y., Li, J., Li, G., Zhao, Y., Li, J., Jin, Z., Mei, H., 2024. Hot or cold? adaptive temperature sampling for code generation with large language models, in: Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence, AAAI Press. URL: https://doi.org/10.1609/aaai.v38i1.27798, doi:10.1609/aaai.v38i1.27798.