Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 264 (2025) 242–251

Procedia Computer Science

www.elsevier.com/locate/procedia

International Neural Network Society Workshop on Deep Learning Innovations and Applications

Exploring the Impact of Temperature on Large Language Models: Hot or Cold?

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

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

Abstract

The sampling temperature, a critical hyperparameter in large language models (LLMs), modifies the logits before the softmax layer, thereby reshaping the distribution of output tokens. Recent studies have challenged the "Stochastic Parrots" analogy by demonstrating that LLMs are capable of understanding semantics rather than merely memorizing data and that randomness, modulated by sampling temperature, plays a crucial role in model inference. In this study, we systematically evaluated the impact of temperature in the range of 0 to 2 on data sets designed to assess six different capabilities, conducting statistical analyses on open source models of three different sizes: small (1B–4B), medium (6B–13B), and large (40B–80B). Our findings reveal distinct skill-specific effects of temperature on model performance, highlighting the complexity of optimal temperature selection in practical applications. To address this challenge, we propose a BERT-based temperature selector that takes advantage of these observed effects to identify the optimal temperature for a given prompt. We demonstrate that this approach can significantly improve the performance of small and medium models in the SuperGLUE datasets. Furthermore, our study extends to FP16 precision inference, revealing that temperature effects are consistent with those observed in 4-bit quantized models. By evaluating temperature effects up to 4.0 in three quantized models, we find that the "Mutation Temperature"—the point at which significant performance changes occur—increases with model size[1].

© 2025 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the IJCNN 2025

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

1. Introduction

Since the release of ChatGPT, LLMs have significantly impacted both academia and industry, revolutionizing the development of artificial intelligence. Open source models of different sizes have facilitated advances in various domains, [27] including question answering and summarization. A key factor influencing the performance of LLMs is hyperparameter adjustment. For example, Top-K sampling selects the next token from the K most probable candi-

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

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

1877-0509 © 2025 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the IJCNN 2025
10.1016/j.procs.2025.07.135

dates, while Top-P sampling samples from the smallest set of tokens whose cumulative probability exceeds P [9]. Additionally, the repetition penalty reduces the probability of tokens that have already appeared, helping to avoid repetition. In this paper, we focus on temperature, which is one of the most frequently used hyperparameters. During inference in LLMs, this parameter is used to scale the logits of the output layer, effectively controlling the randomness of model predictions. The concept of temperature, denoted as T [1], was first introduced by Ackley, who emphasized its crucial role in shaping the Boltzmann distribution. Formally, the probability P_i of the i-th token is given by $P_i = \frac{e^{y_i/T}}{\sum_{j=1}^{V} e^{y_j/T}}$, where y_i denotes the pre-softmax activation of the i-th token (commonly called the logit), T represents the temperature, and V is the total number of tokens in the vocabulary. The value P_i determines the probability that the i-th token will be generated, after which the model output is produced by a sampling algorithm. As T increases, the probability mass function (PMF) [8] becomes more uniform; conversely, as T approaches zero, the distribution collapses to a delta function, causing the algorithm to behave greedily by always selecting the most likely token. At each generation step, a new token is selected by randomly sampling the updated probability distribution [13].

In this study, we focus on three key research questions (RQs): (RQ1) **To what extent does temperature impact the performance of LLMs across different abilities?** (RQ2) **Does temperature have uniform effects across different abilities and models, and what are the main differences observed?** (RQ3) **Is there an optimal temperature for each capability, and can the best temperature be determined for a specific prompt?** The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 describes the experimental methods. Section 4 details the experimental settings. Section 5 presents and analyzes the results. Finally, Section 6 concludes the paper.

2. Related Work

Investigations of the effects of temperature remain limited in the recent literature. Most studies report results using only one temperature value, without systematically exploring a wider range, except for the series of Llama models, which tested two settings for code generation[24]. Furthermore, general data sets designed to evaluate multiple model capabilities simultaneously, such as those for Artificial General Intelligence (AGI; [28]), tend to lack a specific focus on a single foundational ability. For example, higher temperatures increase creativity [29], while lower temperatures improve logical reasoning. For tasks like complex math problems that need both logic and creativity, these effects may cancel each other out. This makes it difficult to see the real impact of temperature, a phenomenon we call the "Temperature Paradox".

Recent studies have examined how temperature settings affect different tasks and dynamic configurations. [22] explored temperature in multitask scenarios using prompt engineering and a range from 0 to 1, finding no significant effect on LLM performance. [19] investigated temperature in creative writing, measuring perplexity and cosine similarity, and found only a weak effect on creativity. [29] proposed an adaptive temperature strategy for code generation, assigning higher temperatures to harder tokens (such as the start of a Python function) and lower temperatures to tokens with greater model confidence, showing that higher temperatures can help with complex tasks. However, there is still no clear guideline for choosing temperature for different LLMs, tasks, or prompts, although temperature adjustment is important for LLM users, RAG systems, and agentic AI systems.

3. Approaches

To address the "Temperature Paradox" and more accurately measure the effect of temperature on each ability with minimal bias, we adopt datasets with clear capability preferences and employ a single-prompt format, querying the model only once to avoid multi-prompt assistance. This approach enables a more precise and unbiased assessment of the intrinsic abilities of LLMs. We hypothesize that temperature influences different model abilities in distinct ways. Therefore, our study focuses on six core intrinsic abilities that not only represent the primary competencies of LLMs but are also central to computational linguistics research.

3.1. Evaluating Intrinsic Abilities

Causal Reasoning (CR): A cognitive faculty historically ascribed solely to humans that consists in deriving conclusions from given premises by adhering to strict logical principles[14]. In this paper, we use CRASS [10], a publicly

available counterfactual reasoning data set that simplifies the evaluation process by requiring the model to select the correct answer rather than generate it. **Top-1 Accuracy (T1)** is used to quantify the frequency with which the model correctly predicts the true class by selecting the class with the highest confidence after multiple repetitions.

Creativity (CT): Creativity involves the generation of novel and valuable ideas, concepts, or products that require both originality and effectiveness [23]. For CT, we adopt a framework that assesses stories in four dimensions: fluency, flexibility, originality, and elaboration, using customized questions based on the Torrance Test of Creative Writing (TTCW) procedure [7]. In this framework, each category includes multiple standard evaluation questions, with true or false determined by expert judgment. Twelve publicly available New Yorker stories are used as plots. The **TTCW Accuracy** is then calculated by counting the number of positive evaluations among all Q&A pairs [12].

In-Context Learning (ICL): ICL reflects an LLM's ability to understand text and perform tasks using contextual information and a few examples [21]. In this study, we focus on the LongBench-TREC long-context task [3], where the model learns from a sequence of questions and answers and must classify a final question based on this context. **Classification Score (CLS)**, measured by accuracy, evaluates the model's ability to recognize patterns and make correct predictions compared to the ground truth.

Instruction Following (IF): IF measures the model's ability to follow instructions provided in the prompts, which is essential for effective LLM applications. For this study, we used InfoBench [20], which introduces the Decomposed Requirements Following Ratio (**DRFR**) as a metric to assess the performance of the follow-up of instruction. DRFR decomposes complex instructions for more granular evaluation and has demonstrated greater reliability and effectiveness.

Machine Translation (MT): MT evaluates an LLM's ability to translate text between languages, a key area where LLMs have shown strong performance. We use the FLORES-101 benchmark [11] for multilingual evaluation, adopting **spBLEU** as the metric. BLEU scores measure the similarity between model outputs and reference translations. To ensure comparability, we normalize the spBLEU scores by dividing by 100, so that all results fall within the range [0, 1]. Given the prevalence of English in the LLM training data, we focus on English-to-other-language translation, selecting diverse pairs (e.g., English-to-Maltesian, Indonesian, Latvian, Icelandic and Khmer) to cover varying levels of translation difficulty.

Summarization (SUMM): Summarization aims to condense long texts into concise summaries while preserving key information and main ideas. One of the main challenges in this task is the reliable evaluation. To address this, we use the "benchmark_llm_summarization" dataset [26], which provides expert-written reference summaries. For evaluation, we adopt the reference-based metric **Rouge-L F1**, which measures the overlap of the longest common subsequence between generated and reference summaries, balancing precision and recall, and has been shown to correlate well with human judgments [16].

3.2. LLM-as-a-Judge Evaluation

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

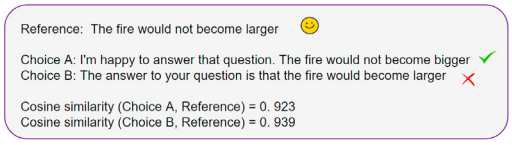

Fig. 1: Cosine Similarity Using BERT Embedding Model

LLM-as-a-Judge has been widely used and has been proven to be highly aligned with human judgment [18]. Due to the inherent stochasticity of LLMs and the flexibility in textual expressions that convey identical meanings, particularly in small and medium models, unintended results often appear preceding or following the target response. This makes reference-based evaluations, such as exact matching or similarity metrics, particularly challenging. For example, similarity metrics have two main issues: (1) It is hard to set a clear cutoff for correct answers; (2) Embedding-based methods often miss key words like "not," as shown in Figure 1 [2].

Advanced models such as GPT-4o and DeepSeek [5, 4] have achieved strong results on many benchmarks. For tasks with complex answers or challenging evaluation, we use LLM-as-a-Judge. For CR, CT, and IF, we use ChatGPT with carefully designed prompts instead of exact matching or human annotation. These judgments are used to calculate

the metrics described in Section 3.1. For the other three abilities with simple reference answers, we use standard evaluation methods, as shown in Table 1.

4. Experiments

4.1. General Experiment Settings

In this study, experiments begin with the selection of diverse benchmark datasets designed to challenge state-of-the-art (SOTA) models, as shown in Table 2. Evaluation is carried out primarily in a question-answer format or through matching functions, with specific metrics applied to each dataset. All models are quantized to 4 bits using the AWQ method [17], and vLLM [15] is used as the default inference acceleration framework. Each question is tested three times across 12 models.

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: Investigated Small, Medium, and Large models with their respective sizes and release dates.

The temperature settings range from 0.1 to 1.9 in increments of 0.3, resulting in seven distinct configurations for each model. Temperatures above 2.0 are excluded, as previous research has shown that higher values tend to produce non-informative and excessively incoherent text [6]. Each model is evaluated using only one question per test. We consistently used gpt-3.5-turbo-0125 as the evaluation model with a temperature setting of 0.01, and selected open source models as listed in Table 2. Nucleus sampling was adopted, as it yields perplexity values closest to human text [13], with the following parameters: max_length = 4096, Top_P = 0.9, repetition penalty (RP) = 1.0, and max_new_tokens = 1024.

4.2. Supplementary Experiment

4.2.1. Best Temperature Selection On SuperGLUE

SuperGLUE is a benchmark consisting of eight tasks for evaluating word-sense disambiguation, natural language inference, coreference resolution, and question answering [25]. The main results from the previous section can be used to identify the optimal temperature for a given prompt and model, provided that the primary ability required for the prompt is known. To this end, a classification model based on a fine-tuned BERT framework, denoted as “BERT-based Selector", is proposed, which is trained on the experimental prompts tailored for the main experiments and will finally be used in classifying every input prompt. Another option is to obtain the required ability for the prompt based on the well-designed prompt that will be asked to GPT models, denoted as “GPT-based Selector".

The basic idea of this selector is the following: Let F be a selection model, either BERT or GPT, acting as an ability predictor based on a given prompt x_p. The model output $F(x_p)$ represents the predicted ability most closely associated with the given prompt x_p. For example, given a training prompt such as x_p = *“Translate ‘J’aime le chat’ from French to English”*, the model $F(x_p)$ would be expected to predict “MT” (Machine Translation) as the most required ability. As such, the optimal temperature parameter $T^* = \arg\max_T \mathcal{D}(T, F(x_p), M)$ is selected to maximize the estimated performance of the model M on the task $F(x_p)$ under different temperature settings, where $\mathcal{D}()$ represents the performance distribution of a given model M over temperatures obtained from previous main experimental results.

To test this framework, we evaluated three models of different sizes—Llama-3.2-1B-Instruct (Small), Llama-3-8B-Instruct (Medium), and Mixtral-8x7B-Instruct-v0.1 (Large) [25]—on the SuperGLUE benchmark. Each question in the benchmark was asked three times, and each instance was generated by incrementing the random seed, initially set to 42, by 1 for each successive iteration.

4.2.2. Experiments On Extended Settings

We also investigated the temperature range [0, 4] while maintaining 4-bit quantization. Although we observed performance degradation and inconsistent generations at temperatures above 2, we did not find evidence for a specific "Mutation Temperature" in large models. To further explore the effect of inference precision, we repeated the main experiments on the same three models using FP16 precision, focusing on the temperature range from 0 to 2, to determine whether temperature effects differ when inference precision is altered. Additionally, since Top-K and Top-P sampling influence the output distribution at the candidate selection level, while RP operates at the logits level, we further evaluated various settings for Top-K (2, 5, 10), Top-P (0.8, 0.9, 1.0), and RP (0.0, 1.0, 2.0). These experiments were conducted on all three models mentioned above to systematically assess the impact of these parameters on performance.

5. Results and Analysis

5.1. Findings from Statistical Analysis

Table 3 provides a summary of the temperature-performance correlations in six abilities, based on results from three categories of models. "P. Coef." and "S. Coef." correspond to the Pearson and Spearman correlation coefficients, respectively. "Range (%)Max" may be interpreted as the relative performance variation across temperatures, while "Range Max (%)" refers to the maximum relative ranges within the size category. "CV" (Coefficient of variation) represents the ratio of average performance to standard deviation, and "CV Max" indicates the highest CVs observed within the size category. The average accuracy and standard deviation for the temperatures and models within each group are also reported.

Table 3: Comparison of temperature-performance correlations for six abilities across three model categories.

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**

In this table, it can be observed that the performance of IF, MT, and SUMM exhibits relatively strong correlations with temperature, as indicated by both correlation coefficients. The statistical significance of these correlations is further supported by p-values of zero. Furthermore, both "Range Max" and "CV Max" decrease as the size of the model increases, which statistically suggests that larger models are more robust to temperature-induced variations. The average accuracy metric further demonstrates that larger models achieve higher statistical performance across all six abilities. In particular, performance differences among models of different sizes are relatively small for CT, IF, and SUMM, but much more pronounced for CR, ICL, and MT. These findings provide practical guidance for selecting the model size according to specific functional requirements.

5.2. Temperature Effects on Different Abilities

Figure 2 illustrates the impact of temperature on models of varying sizes across a range of evaluated abilities. Lines show the mean performance for each model size, while shaded regions correspond to ±0.2 standard deviations. For the sake of consistency, all evaluation metrics enumerated in Table 1—including spBLEU for machine translation—are uniformly referred to as "accuracy" and have been normalized to the interval [0, 1].

Causal Reasoning (CR). CR questions are counterintuitive and require logical reasoning, each with three options. Medium and large models exceed the 33.3% random baseline, while small models perform only slightly above this chance level—by approximately 7%—across most temperature settings, indicating limited reasoning ability. The large and medium models show slight improvement at a temperature of 1.3, suggesting that higher temperatures may help to address complex problems. The optimal temperature for CR is not always zero and an increase in temperature does

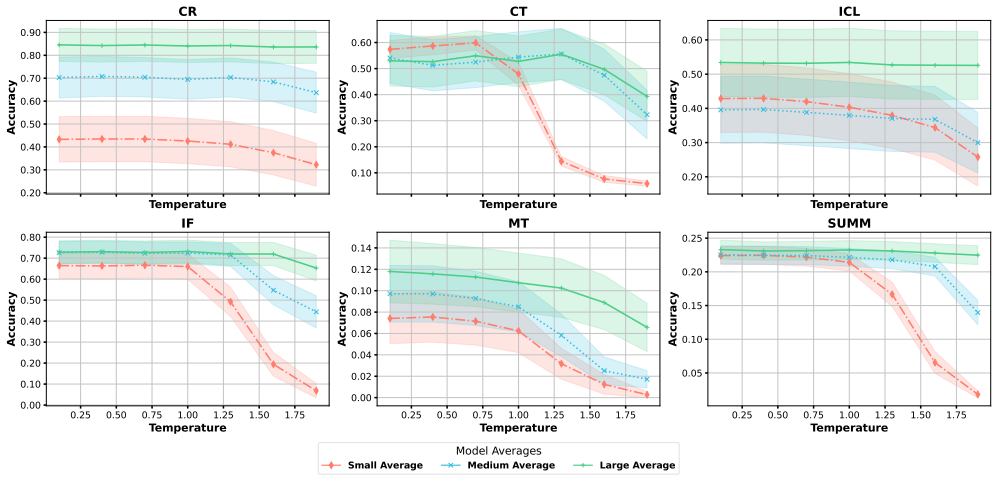

Fig. 2: Average performance trends for different model sizes, with shaded bands indicating variability.

not necessarily reduce performance. In contrast, small models do not demonstrate substantial causal reasoning ability within the scope of this study.

Creativity (CT). An optimal temperature of 1.3 is recommended for medium and large models to maximize creativity. Small models show a marked decline in creativity at $T = 1.0$, while medium and large models are only affected at $T = 1.7$. Generally, temperature first increases and then decreases creativity. Small models are more creative at lower temperatures, but large models are more robust and generate more diverse outputs, as indicated by their wider shaded regions. In general, temperature strongly influences creativity, with moderate values being the most beneficial.

In-Context Learning (ICL). Large models achieve the best average performance, while the difference between medium and small models is minimal. This indicates that ICL, as an emerging property of LLMs, requires a sufficiently large model size, highlighting the significance of scaling laws. Medium models show less performance decline than small models. At a temperature of 1.7, small models degrade faster than medium models, despite outperforming them at lower temperatures. Large models maintain stable performance across temperatures from 0 to 2, with no abrupt performance drop observed. Increasing temperature generally reduces performance, although large models may experience slight improvements at higher temperatures.

Instruction Following (IF). The behavior of IF is particularly noteworthy. As the temperature increases from 0 to 1, IF performance remains largely unchanged. However, when the temperature exceeds 1, different models experience relatively pronounced negative effects, and the larger the model size, the later these negative effects emerge. Performance changes with temperature are abrupt: small models exhibit a mutation between 1.0 and 1.3, medium models between 1.3 and 1.6, and large models demonstrate a moderate mutation temperature from 1.6 to 1.9. Therefore, for users of LLMs who require strict adherence to instructions, it is advisable to set the temperature below 1.

Machine Translation (MT). Slightly increasing the temperature within the low range marginally improves translation performance for small and medium models only. The rise in temperature has the most detrimental effect on MT, as indicated by the highest range of performance and CV in Table 3. This trend can be attributed to the inherently deterministic nature of translation, and all models exhibit comparable declines in performance. The optimal temperature is close to zero ($0 + \epsilon$), and language understanding performance depends primarily on the breadth of the training data and the model's parameter size.

Summarization (SUMM). Temperature effect curves are initially stable but drop sharply at higher temperatures, especially for small models. Statistical analysis shows a strong negative correlation between performance and temperature. SUMM tasks follow a similar trend to IF tasks, but the mutation temperature for medium models is higher (about 1.7), and large models show no clear mutation temperature.

5.3. Supplementary Experiment

5.3.1. Best Temperature Selection on SuperGLUE

We conducted experiments on three models, as shown in Table 4. The table presents the SuperGLUE validation accuracy under different temperature settings: ACC_D denotes the accuracy with the **D**efault temperature of 1.0, while ACC_B and ACC_C represent the precision achieved by dynamically selecting the optimal temperature using our fine-tuned **B**ERT model and **C**hatGPT-based prompting, respectively. This comparison clearly demonstrates the performance difference from optimal temperature selection.

Table 4: SuperGLUE validation accuracy under default and dynamically selected temperature settings.

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

Adjusting the temperature can greatly improve the performance of WIC (one of the tasks in SuperGLUE) for Llama-3.2-1B and Meta-Llama-3-8B-Instruct. This shows that the optimal temperature selector provides stable performance, avoiding potential performance drops that can occur when using a fixed temperature. When working with small models, this is indeed one of the necessary parameters to consider, especially in resource-constrained scenarios. Considering that SuperGLUE primarily evaluates a range of different capabilities, our optimal temperature selector still demonstrates consistent improvements. It is important to mention that this selector does not inherently boost performance—Supervised Fine-Tuning (SFT) remains the primary method—but it ensures that the model achieves the best possible performance by avoiding suboptimal settings. For large models, we did not observe significant performance differences, indicating that optimizing the temperature setting is generally less critical for larger models. However, as suggested in previous findings, when large models are used to solve complex reasoning tasks, a higher temperature can sometimes lead to performance gains. Therefore, adjusting the temperature may still be necessary in such scenarios.

5.3.2. Results on Extended Settings

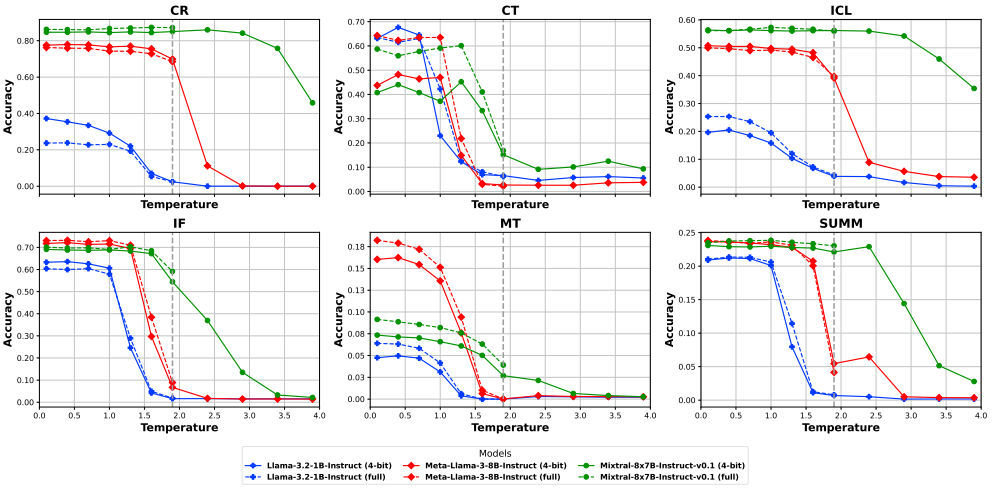

Fig. 3: Performance with extended temperature to 4.0

Extension to 4.0. Fig. 3 presents the performance curves of models with 4-bit precision across six capabilities as temperature varies; "Full" refers to FP16 precision inference. Extending the temperature range helps identify both the "mutation temperature," where performance drops sharply, and the potential upper limit for temperature settings. Large models are more robust to increasing temperatures, while small models experience significant performance loss at lower temperatures, which aligns with our expectations. In particular, each model has its own mutation temperature, with larger models generally exhibiting a higher threshold.

FP16 Precision. Fig. 3 also shows that the optimal temperature of the models does not exhibit significant changes across the entire temperature range. Overall, the results indicate that there is no substantial difference in optimal temperature between the two levels of precision, although there is a 10%–20% difference in performance. Since our main experiments were conducted under 4-bit quantization, these findings also validate the effectiveness of our results and demonstrate their scalability with respect to inference precision.

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: Pairwise Pearson correlation coefficients for performance curves with varying parameter values.

Other Parameters. We analyze the effect of varying a single parameter by computing the Pearson correlation coefficients between the resulting performance curves, while keeping all other parameters fixed. Specifically, for each parameter, we select three distinct values and calculate the pairwise Pearson correlation coefficients among their corresponding temperature-performance curves. The values p_1, p_2, and p_3 denote the correlations for the first vs. second, first vs. third, and second vs. third parameter settings, respectively. These results are summarized in Table 5. It can be clearly observed that, except for CT, the correlation of temperature effects between different parameter settings is very high, indicating minimal impact. In contrast, RP has a significant effect on CT and IF. For CT, the impact of Top-K is smaller than that of RP, while Top-P exerts the greatest influence, as indicated by the relatively low or even negative correlation coefficients. Therefore, when the model is required to perform creative tasks, it is essential to set Top-P appropriately, and RP should also be configured with care. However, these three parameters exert minimal influence on the temperature performance curve and warrant particular consideration only for CT and IF tasks.

6. Conclusion

In this study, we extend previous work on temperature effects in language models by examining six capabilities over a wider temperature range. We also improve the GPT evaluation methods and the testing protocol, focusing on both statistical and empirical analysis of temperature effects. We also compare the performance of the model under FP16 and the quantization of 4-bits, finding minimal differences in the temperature effect. By extending the temperature range to 4.0, we observe that large models still have a mutation temperature, but it is higher than 2.0. Additionally, we introduce BERT-based and GPT-based temperature selectors, demonstrating their effectiveness on the SuperGLUE dataset, especially for small models.

To answer RQ1, **"To what extent does temperature affect the performance of LLMs across multiple abilities?"**, we find that temperature has a modest effect on In-Context Learning and Causal Reasoning, but can significantly impact Machine Translation (up to 192.32%), Creativity (up to 186.81%) in small models (see Table 3). Spearman coefficients indicate a generally negative correlation between temperature and performance.

To answer RQ2, **"Does temperature exert uniform effects on different abilities across models, and what are the primary differences observed?"**, we find that large models are more resilient to temperature changes. Increasing the temperature slightly improves Causal Reasoning, In-Context Learning, and Instruction Following, followed by a decline in performance. For Summarization and Machine Translation, higher temperatures generally have a negative impact, especially on smaller models. The influence of temperature varies substantially across different abilities and model sizes, making it difficult to generalize a consistent pattern; the optimal temperature also differs significantly

between models. Furthermore, high temperatures are not always detrimental to logic-oriented abilities such as Causal Reasoning, Instruction Following, and In-Context Learning.

To answer RQ3, **"Is there an optimal temperature for each ability, and can the best temperature be found for a specific prompt?"**, we find that no single temperature is optimal for all tasks. Higher temperatures are not always best for creative writing, nor is zero always best for following instruction. However, by identifying the required ability for each prompt using our BERT model and referencing our experimental results, we can select the optimal temperature for each prompt. In three SuperGLUE tasks, this approach improves performance for small and medium-sized models.

This study has some limitations. We examined only the effect of temperature on six text-based ability; additional abilities, such as planning and coding, could also be evaluated in future work. It is difficult to verify the accuracy of the GPT model evaluations without manually checking each case. More testing is needed to determine whether the optimal temperature selector is effective for other models of similar or larger sizes. Furthermore, a mathematical explanation of how temperature affects model performance requires further investigation.

Acknowledgements

This research has greatly benefited from the collaboration and support of our industrial partner, Foyer S.A., as well as from academic institutions, particularly the Interdisciplinary Centre for Security, Reliability and Trust (SnT), and numerous individual contributors. We would like to thank the "AI & Data Studio" team for their valuable insights, guidance, and for providing essential computational resources (NVIDIA H100 GPUs), which were crucial for conducting our experiments. The guidance and mentorship provided by faculty and advisors, together with constructive suggestions and technical expertise shared by postdoctoral researchers, have significantly improved the quality and impact of this work. Furthermore, we exclusively used publicly available datasets and models, and we affirm that no AI-generated text was used in the preparation of this manuscript. Throughout the research process, we have adhered to ethical standards and maintained transparency, ensuring that all of our methods and results are clearly communicated.

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.