Small Models Exhibit Limited Answer Consistency in Repetition Trials of the Multiple-Choice MMLU-Redux and MedQA Benchmarks

Claudio Pinhanez, Paulo Cavalin, Cassia Sanctos, Marcelo Carpinette Grave

IBM Research Brazil
Rua Tutóia 1157, São Paulo, SP, CEP 01403-001, Brazil
claudio@pinhanez.com

Abstract

This work explores the consistency of LLMs in answering multiple times the same question. In particular, we study how known, open-source LLMs respond to 10 repetitions of questions from the multiple-choice benchmarks MMLU-Redux and MedQA, considering different inference temperatures, small (2B-10B parameters) vs. medium models (50B-80B), finetuned vs. base models, and other parameters. The paper also examines the effects of requiring answer consistency in repetitive inferences on accuracy and the trade-offs involved in deciding which model best provides both of them, for what we propose some new representations. Results show that the number of questions which can be answered consistently vary wildly among models but typically is in the 50%-85% range for small models and that accuracy among consistent answers correlates to overall accuracy at low inference temperatures. Results for medium-sized models seem to indicate much higher levels of answer consistency.

1 Introduction

In this paper, we investigate the non-determinism of small LLMs (2 to 8 billion parameters) by examining their consistency when answering multiple-choice questions, using standard benchmarks of both general knowledge, *MMLM-Redux* (Gema et al. 2025)), and medical expertise, *MedQA* (Jin et al. 2021), and show that they rarely display anything close to determinism or high levels of consistency.

Recent work has explored some aspects of answer consistency but mostly in the context of non-multiple choice answers and using closed commercial LLMs, often with fixed inference temperatures (Atil et al. 2025; Nalbandyan, Shahbazyan, and Bakhturina 2025; Patwardhan, Vaidya, and Kundu 2024; Lee, Hong, and Thorne 2024; Ouyang et al. 2025). Unlike those, this work focuses on multiple-choice benchmarks using only open source models. In particular, we look into the issue how the inference temperature affects the accuracy of the consistently-produced answers. The use of multiple-choice benchmarks avoids the noise created by measuring similarity between answers to the same question. Also, the use of open source models assures that our results are reproducible since commercial LLMs may include cache/retrieval systems which may affect the outputs.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

We start by defining *answer consistency* as oracle guessing at the a given level and show how it is translated into an objective criteria in the case of repeated multiple-choice evaluations. In this process, we propose a simple but inefficient method to define the consistency in answers. With those tools, we explore answer consistency at different temperatures in 23 small models of about 8 billion parameters (2B-8B) and in 3 medium models in the range of 70 billion parameters (50B-80B), including finetuned vs. base models, and a commercial family of models, for the two benchmarks.

Answer consistency is often a requirement of most non-creative applications of LLMs and it is either wrongly assumed by developers or not explicitly declared, although expected. Consider, for example, the common application of LLMs in chatbots for customer service. In many countries, answering differently to customers with identical questions is considered deception and/or discrimination and may entitle legal reparations. Similarly, patients and physicians would be very suspicious of an AI radiologist which produced contradictory readings and diagnoses of the exactly same X-ray. Also, answer consistency is an important feature when considering *value alignment* and *safety* of LLMs. In many situations, it is not only required that the model almost never produces an undesirable output, but also, for some questions, that it never produces a particular answer.

The key contributions of this paper are:

- The **definition of answer consistency at a level** c as equivalent to oracle guessing at the same level and its derivation to the context of multiple-choice evaluation.
- **Experimental studies with the MMLU-Redux and MedQA benchmarks** across 26 different small and medium models, with and without finetuning, showing that **small models produce consistent answers in the range of about 50% to 85%** of the times, often at low temperatures, and with accuracy of the consistent answers correlating to the average benchmark accuracy.
- Evidence from the results with 3 medium-sized models (50B-80B) that **limited consistency is mostly an issue of small models (2B-8B)**.
- Evidence that in the general knowledge benchmark MMLU-Redux, **increasing temperature yields a higher degree of accuracy among answers which are consistently given**.

2 Related Work

There is very limited work on answer consistency in the context of repetition of questions. In (Nalbandyan, Shahbazyan, and Bakhturina 2025), an evaluation score is proposed with one component where the same question was asked using different seeds but only the average accuracy is reported. Conversely, (Patwardhan, Vaidya, and Kundu 2024) looked into consistency over identical and semantically similar prompts with non-multiple choice cybersecurity benchmarks. Likewise, (Lee, Hong, and Thorne 2024) evaluated commercial LLMs consistency to follow instructions using common benchmarks, performing 5 repetitions at a high 1.0 temperature and found limited consistency (Jiang et al. 2023). Also, (Atil et al. 2025) looked into commercial models and explored consistency in both zero- and few-shot scenarios. Finally, (Song et al. 2024) focused on the differences between greed and sampling decoding.

Our work differs from those because we measure consistency with multiple mechanisms, report accuracy along with its standard deviation, look at more than one temperature, and do not use closed, commercial LLMs to avoid unknown mechanisms like caching systems. *Multiple-choice evaluation* is often used to evaluate LLMs (Singhal et al. 2023; Jiang et al. 2023; Nori et al. 2023; Dubey and et al 2024) since it can be more objectively evaluated as opposed to open questions. This is the primary reason we used multiple-choice contexts in this work.

A related area of work has explored sensitivity of LLMs according to changes in the input and, in particular, on simple variations on the multiple-choice answers, such as reordering. In (Mirzadeh et al. 2024), it was shown that LLMs are negatively impacted by even small changes in the input question such as when only having choice numbers modified in math-related questions. The work described in (Ackerman et al. 2024) proposed a metric for LLM robustness to input changes and reported important impacts on the accuracy.

Another line of research has focused on investigating the *consistency* of LLMs in providing a response when the question is kept intact but with variations in other factors, such as the set of choices and parameters of the inference algorithm. An investigation on the sensitivity of choice order was reported in (Li et al. 2024) according to different values for the temperature parameter. In (Wei, Chen, and Luo 2024), the authors compared the results of multiple-choice evaluations and open-ended answers and found low consistency between these two methods. In (Pezeshkpour and Hruschka 2024), the *MV* metric based on majority voting was proposed in what could be considered as a simplified, non-generic version of the methods proposed here, a theme also explored in (Wang et al. 2024).

3 Defining Answer Consistency Metric

Most LLMs use a *softmax* function at the top of the decoder stack of a Transformer-based system (Vaswani et al. 2017) which selects the next token to be generated, considering the probabilities of each token as computed by the preceding layers as part of a weighted random choice process controlled by the *temperature* of the inference process. When the temperature is zero, the most likely token is, at least theoretically, produced and therefore the generation of the output token in each cycle of the inference process is, or should be, deterministic. However, when the temperature is greater than zero the random weighted drawing of the next token makes the process intrinsically non-deterministic.

Notice that in the context of this study, it is trivial to determine whether the same answer was produced from the same input and whether the answer is correct or not, since we use *multi-choice benchmarks*, where an answer is just one letter identifying one of the answers provided by the question.

3.1 Consistency and Oracle Guessing

A way to characterize how consistent a model is in answering a set of questions is to compare it to an *oracle machine* which answers questions at a certain rate c of success. Oracles were introduced by Alan Turing in 1939 (Turing 1939) and are a cornerstone of complexity theory, leading to:

Definition: *given a question, an LLM-model has* c-answer consistency *when the model is equivalent to an oracle guessing the question at a* c *level of correctness.*

In this work, we explore the characteristics of systems which exhibit c-answer consistency in M repetitions of a set with Q multiple-choice questions of k choices. We start by considering a single multiple-choice question q of k choices which is repeatedly evaluated by a model M times, yielding answers $llm(q_i), 1 \leq i \leq M$, where $llm(q_i) = 1$ if, and only if, the answer is correct. In M repetitions, the number of possible arrangement of choices where exactly p are correct, $C_M(p)$ is, trivially:

$$C_M(p) = \left(\begin{array}{c} M \\ p \end{array} \right) = \frac{M!}{(M-p)!\,p!} \tag{1}$$

Now consider a system with a success guessing rate r, where $r = 1/k$ if it were a purely random guess. It is easy to see that the probability of guessing correctly exactly p of the M repetitions, $T_r^M(p)$ is, by basic probability:

$$P(T_r^M(p)) = C_M(p) r^p (1-r)^{(M-p)} \tag{2}$$

Following, the probability of guessing correctly p or more answers in M repetitions, $\overline{T}_r^M(p)$ is, just by summing up:

$$P(\overline{T}_r^M(p)) = \sum_{j=p}^{M} C_M(j) r^j (1-r)^{(M-j)} \tag{3}$$

For instance, let us consider the value of $\overline{T}_r^M(p)$ for $M = 10$ repetitions of $k = 5$ choices, when the guessing rate is purely random, $r = 1/k$. In this case, the probability of obtaining 10 correct answers in 10 repetitions of a question, if the model is randomly guessing, is 0.0000001.

Conversely, now imagine the model as an oracle which guesses the correct answer at a certain success rate, SGR. We can then compute the minimum success needed to always get at least p correct answers in M, what we call the *minimum success guessing rate*, $MSGR(p)$. We computed numerically such values, for $M = 10$ repetitions of $k = 4$ and $k = 5$ choices. For the former, a model has to be guessing at least of a success rate of 0.93 to achieve 6 out of 10

correct answers ($MSGR(6)$), which is equivalent to the requirements of the metric MV proposed in (Pezeshkpour and Hruschka 2024). In our view, a $MSGR(6) = 0.93$ is still insufficient to guarantee that a model is actually consistent in a multiple-choice benchmark. However, requiring that the model is consistent in 10 out of 10 repetitions ($MSGR(10)$) warrants that it can only successfully guess if its success rate is above 0.9999, which for us seem to be an excessive requirement.

For the experiments in this paper, a model is answer consistent when it is equivalent to an oracle guessing correctly at a 0.99 rate, or when it has *0.99-answer consistency* or, simply, *0.99-consistency*. As shown by the derivation we have just done, for $M = 10$ repetitions of $k = 4$ or $k = 5$ choices, this is equivalent to answering at least 9 of the 10 repetitions with the same choice. This requirement thus covers the studies with the two benchmarks used in this paper.

3.2 Assuring Consistency

We are now in position to propose a formal method to determine whether a model has 0.99-consistency when answering a 4- or 5-choice question in a context of 10 repetitions. We say the model is *SURE* of its answers to a question when it shows, experimentally, 0.99-consistency, and *UNSURE* when it is not able to produce evidence of it. This is accomplished by:

1. Asking the question to the model 10 times.
2. If the model answers identically 9 or 10 times, the question is SURE.
3. If not, the question is UNSURE.

Notice that we do not need to actually call 10 times the model, since in some cases only 3 answers already warrant the model is UNSURE of the question, and, similarly, 9 identical answers are sufficient for SURE. However, being SURE does not mean being correct. In the case of benchmarks, where the correct answer is known, we consider that a SURE question is answer correctly, *right*, only if it matches the correct choice for the 9 or 10 times it responds identically, and *wrong* if the converse has happened. In the case of UNSURE questions, we consider here that the model is *right* if the correct answer is among the choices with highest number of answers and *wrong* otherwise.

To characterize differences and trade-offs among models, we propose a representation for *c-consistency* of a model by a pair of numbers separated by the symbol "|":

- **RWS**, or "right when SURE", is the ratio of right SURE questions to the total number of SURE questions.
- **S/T**, is the percentage of SURE questions in relation to the total number T of questions.
- **c-consistency of a model: RWS | S/T**.

We understand that repeating a question multiple times is not an efficient way to achieve consistency but, for the purposes of this work, it is a safe method. The proposal of efficient consistency methods is beyond the scope of this paper.

4 Experiments with MMLU-Redux

We first explored answer consistency by running 10 repetitions of the **MMLU-Redux** general knowledge benchmark (Gema et al. 2025), considering three general-purpose small models in the range of 8 billion parameters: *LLama3 8B* (Llama-3-8B) (Aaron Grattafiori et al 2024), *LLama3 8B Instruct* (Llama-3-8B-instruct) (Aaron Grattafiori et al 2024), and *DeepSeek v3 7B* (DeepSeek-AI and Aixin Liu et al 2024). We also explored three medium-size models, *Llama3 v3 70B* (llama-3.3-70b) (Aaron Grattafiori et al 2024), *Mistral 8x7B Instruct* (mixtral-8x7b-instruct) (Jiang et al. 2024), and *Qwen 2.5 72B Instruct* (qwen2-5-72b-instruct) (Qwen et al. 2025), all in the range of 50 to 70 billion parameters. The labeling of the models as *small* and *medium* follows practices of the LLM research community.

The experiment consisted on performing 30 repetitions of each question of the benchmark, in three sets of 10 identical repetitions, with *inference temperatures* of 0.3, 0.7, and 1.0. MMLU-Redux is a 4-choice benchmark, so we computed the correct answer by prompting the model to answer the question as the letter of the correct choice, using the prompt described in section 4.1 of (Pinhanez et al. 2025), with top-K sampling decoding. The same work provides more technical details of the experiment, including information about infrastructure, environment, and parsing.

4.1 Small vs. Medium Models

Results for the 3 small and 3 medium models are shown on the top part of table 1. For each evaluation we computed the ratio of questions for which the model provided the correct alternative, yielding 10 results of accuracy. The average of those 10 accuracy results is reported as the *accuracy average* together with the standard deviation, *accuracy stdev*.

We then determined, examining the 10 answers for each question in an evaluation, whether the model answered the question in a SURE or UNSURE way and whether the set of 10 answers was right or wrong as described before. We followed by determining the ratio of correct questions to the number of SURE questions, RWS, and the percentage of SURE questions, S/T, as described previously.

For small models, the best temperature was $t = 0.3$, in which the percentage of SURE questions (S/T) was in the 53% to 79% range and the accuracy on consistent answers (RWS) was better than the average accuracy. Notice also a pattern where the best S/T scores were at the low temperatures while the best RWSs happened at high temperatures.

The medium models had very high levels of percentage of SURE questions (S/T), from 96% to 99% at their best temperatures, suggesting that bigger models may have less issues with consistency. Possibly because of that, their accuracy average and RWS scores were very similar, and we saw again the best RWS scores happening at high temperatures. However, bigger models have a higher chance of benchmark contamination, since they are trained with much larger amounts of data. We will further explore possible contamination issues in large models in future works.

MMLU-Redux - 10 trials 0.99-consistency	temp	SURE & right	UNSURE & right	UNSURE & wrong	SURE & wrong	accuracy average	accuracy stdev	RWS	S/T
SMALL MODELS ($\leq 8B$ parameters)									
Llama-3-8B	0.3	43%	18%	29%	10%	**0.590**	0.004	0.81	**53%**
Llama-3-8B	0.7	24%	36%	39%	2%	0.520	0.005	0.93	25%
Llama-3-8B	1.0	10%	47%	43%	1%	0.428	0.006	**0.94**	11%
Llama-3-8B-instruct	0.3	58%	7%	14%	22%	**0.645**	0.002	0.73	**79%**
Llama-3-8B-instruct	0.7	49%	15%	24%	12%	0.634	0.004	0.80	62%
Llama-3-8B-instruct	1.0	43%	21%	29%	7%	0.616	0.003	**0.86**	50%
deepseek-llm-7b	0.3	41%	8%	18%	33%	**0.485**	0.002	0.55	**74%**
deepseek-llm-7b	0.7	32%	17%	34%	17%	0.478	0.004	0.65	49%
deepseek-llm-7b	1.0	24%	25%	42%	10%	0.465	0.003	**0.71**	34%
MEDIUM MODELS ($\geq 50B$ and $\leq 80B$ parameters)									
llama-3.3-70b	0.3	80%	1%	1%	18%	**0.805**	0.001	0.81	**98%**
llama-3.3-70b	0.7	79%	2%	2%	17%	0.805	0.001	0.82	96%
llama-3.3-70b	1.0	78%	2%	4%	16%	0.805	0.001	**0.83**	94%
mixtral-8x7b-instruct	0.3	70%	0%	1%	29%	**0.700**	0.093	0.71	**99%**
mixtral-8x7b-instruct	0.7	70%	1%	1%	29%	0.687	0.100	0.71	98%
mixtral-8x7b-instruct	1.0	69%	1%	2%	29%	0.679	0.105	**0.71**	98%
qwen2-5-72b-instruct	0.3	81%	2%	2%	15%	**0.826**	0.001	0.84	96%
qwen2-5-72b-instruct	0.7	79%	4%	5%	12%	0.825	0.001	0.87	91%
qwen2-5-72b-instruct	1.0	77%	6%	7%	10%	0.823	0.001	**0.88**	87%
SMALL GRANITE MODELS ($\leq 8B$ parameters)									
3.2-8b-instruct	0.3	56%	8%	14%	22%	**0.636**	0.002	0.72	**78%**
3.2-8b-instruct	0.7	48%	16%	24%	12%	0.626	0.004	0.80	60%
3.2-8b-instruct	1.0	40%	24%	28%	8%	0.611	0.003	**0.84**	48%
3.1-8b-instruct	0.3	56%	8%	14%	21%	**0.640**	0.002	0.73	**78%**
3.1-8b-instruct	0.7	46%	18%	25%	11%	0.624	0.003	0.81	57%
3.1-8b-instruct	1.0	37%	26%	30%	7%	0.592	0.006	**0.84**	44%
3.1-8b-base	0.3	44%	14%	24%	18%	**0.575**	0.002	0.71	**63%**
3.1-8b-base	0.7	24%	34%	38%	4%	0.535	0.005	0.84	28%
3.1-8b-base	1.0	10%	48%	41%	1%	0.462	0.007	**0.90**	11%
3.0-8b-base	0.3	42%	19%	28%	11%	**0.591**	0.003	0.79	**52%**
3.0-8b-base	0.7	21%	39%	37%	2%	0.537	0.005	0.90	24%
3.0-8b-base	1.0	10%	49%	40%	1%	0.457	0.004	**0.90**	11%
3.1-2b-instruct	0.3	56%	8%	14%	22%	**0.636**	0.002	0.72	**78%**
3.1-2b-instruct	0.7	48%	16%	24%	12%	0.626	0.004	0.80	60%
3.1-2b-instruct	1.0	40%	24%	28%	8%	0.611	0.003	**0.84**	48%
3.1-2b-base	0.3	21%	28%	44%	7%	**0.465**	0.005	0.76	**24%**
3.1-2b-base	0.7	7%	41%	52%	1%	0.394	0.006	0.91	7%
3.1-2b-base	1.0	2%	42%	55%	0%	0.331	0.004	**0.91**	3%

Table 1: Results of 0.99-consistency for the MMLU-Redux benchmark with 10 repetitions for 3 small models (2B-8B), 3 medium models (50B-80B), and for 6 small models of the granite family.

4.2 Exploring Multiple Versions of a Model

We also performed an identical study using 6 different models of the *granite* family (Granite Team 2024): *3.2 8B instruct* (3.2-8b-instruct), *3.1 8B Instruct* (3.1-8b-instruct), *3.1 8B Base* (3.1-8b-base), *3.0 8B Base* (3.0-8b-base *3.1 2B Instruct* (3.1-2b-instruct), and *3.1 2B Base* (3.1-2b-base), comprising models with between 2 and 8 billion parameters. We used the granite models because they were trained with carefully curated data and with a strong attention to avoid incorporating benchmarks (Granite Team 2024).

Results are shown on the bottom of table 1. We observed very similar results as in the previous models: the percentage of SURE questions at the best temperature for each model going from 52% to 78% (we ignored the results of the 3.1-2b-base as outliers); the accuracy among SURE questions (RWS) a little higher then the average accuracy; and the best performance for S/T at the temperature 0.3. Similarly, the best RWS scores happened at high temperatures suggesting theres is a trade-off between being SURE often and being right when SURE.

4.3 Aggregated Results over All Models

To provide an overall understanding of the impact of assuring consistency, we first considered the RWS results of all models at different temperatures and computed the correlation with the average accuracy. We obtained correlation values of 70%, 8%, and 18% for the results with temperatures 0.3, 0.7, and 1.0, respectively. In the scatter plot depicted on the top of figure 1, we can see that the blue markers corresponding to the evaluations with $t = 0.3$ are reasonably aligned. Using basic regression, we obtained a coefficient of 0.563 and an intercept value of 0.374, with $R^2 = 0.488$.

The scatter plot on the bottom of figure 1 plots the pairs of RWS and S/T values for all models and temperatures and the averages of the S/T value for each of the inference temperatures. We considered the percentage of SURE questions (S/T) for each model at the $t = 0.3$ for all 9 small models and obtained an average across models of 69% with an standard deviation of 11%, therefore exhibiting a tendency to be in the range between 54% and 80%. We did not consider the medium models because, as observed before, they

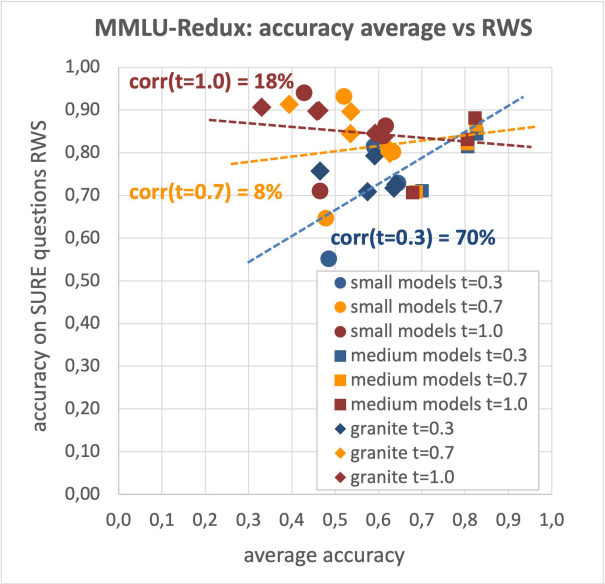

Figure 1: Scatter plot graphs showing the correlation between average accuracy and RWS for small and medium models (top) and between RWS and S/T for small models (bottom) tested with MMLU-Redux benchmark.

exhibit very strong consistency, in the range of the upper 90%s. For $t = 0.7$, we obtained an average of 46% $\pm$ 17% and for $t = 1.0$, 32% $\pm$ 18%. This seems to indicate that small models, even at the low temperature 0.3, tend to exhibit high levels of inconsistency: in none of our experiments with MMLU-Redux, any of them showed more than 80%. Notice also that the corresponding RWS values stayed in a small range, from 0.55 to 0.94.

5 Experiments with MedQA

To explore whether those results were particular to the MMLU-Redux benchmark, we ran similar studies with the health-related *MedQA* benchmark (Jin et al. 2021) on 4 finetuned medical models and their respective base models, again using 3 different temperatures. Instead of exploring the differences between small and medium models, as we did with the MMLU-Redux benchmark, we explored the differences in answer consistency among the *finetuned* vs. *base* models. The objective was to see how the finetuning process may affect consistency. We also performed evaluations with the same 6 granite models used before, taking in account that MedQA is beyond their areas of expertise.

Here, we considered as benchmark the test set of MedQA with 1,273 5-choice questions extracted from the USMLE exam. We chose 4 LLMs finetuned on medical data which had reported scores on this benchmark: *MedLlama3 7B* (medllama3-v20) (Medical 2024), *BioMedical Llama3 8B* (Bio-Medical-Llama-3-8B) (Medical 2024), *BioMistral 7B* (BioMistral-7B) (Labrak et al. 2024), and *MedAlpaca 7B* (medalpaca-7b) (Han et al. 2025). The corresponding 4 base models from where these models were finetuned are: *LLama3 8B* (Llama-3-8B) (Aaron Grattafiori et al 2024), *LLama3 8B Instruct* (Llama3-8B-instruct) (Aaron Grattafiori et al 2024), *Mistral 7B Instruct* (Mistral-7B-Instruct) (Jiang et al. 2023), and *Llama1 7B* (llama1-7b) (Touvron et al. 2023). The granite models were the same used in the MMLU-Redux experiment. We used exactly the same methodology as in the MMLU-Redux case.

5.1 Finetuned vs. Base Models

The top of table 2 displays the results of the 10 evaluations of the 8 finetuned and base models at 3 inference temperatures. First, all the accuracy standard deviations are extremely small, agreeing to the results reported in (Nalbandyan, Shahbazyan, and Bakhturina 2025). Among the finetuned models, medllama3-v20 had the best accuracy average, 0.736, at $t = 0.3$. The second best model was Bio-Medical-Llama-3-8B, with 0.717 average accuracy, also at 0.3 temperature.

As for answer consistency, the finetuned medical models produced consistent answers from 49% to 96% of the benchmark questions, considering the output at the best temperature but it was as low as 17% for the worst model at $t = 1$. The best finetuned model was medllama3-v20 at 0.3 temperature which produced the highest percentage of SURE (S/T) questions, an impressive 96%, with 0.75 SURE accuracy (RWS), what was higher than its own accuracy average.

However, the highest ratio of correct SURE questions (RWS) was yielded by the Bio-Medical-Llama-3-8B model at temperatures 0.7 and 1.0, of 0.79, better by more than 5% over the best average accuracy, 0.736, although only in about 60% of the answers. Here is an interesting case where overall accuracy can be increased by filtering out UNSURE answers, albeit at the cost of 40% of the questions. In some highly critical situations where certainty and correctness are top requirements, this would be the best model. The other two models, BioMistral-7B and medalpaca-7b, had much lower performances across all metrics, although BioMistral-7B showed a high number of consistent questions, but with less than one third correct.

If we look into the S/T column of table 2, we see the same pattern observed in the MMLU-Redux in which, as the

MedQA - 10 trials 0.99-consistency	temp	SURE & right	UNSURE & right	UNSURE & wrong	SURE & wrong	accuracy average	accuracy stdev	RWS	S/T
SMALL FINETUNED MODELS									
medllama3-v20	0.3	71%	2%	2%	24%	**0.736**	0.002	0.75	**96%**
medllama3-v20	0.7	67%	6%	5%	21%	0.733	0.004	**0.76**	89%
medllama3-v20	1.0	67%	7%	5%	21%	0.730	0.004	**0.76**	88%
Bio-Medical-Llama-3-8B	0.3	66%	7%	5%	23%	**0.717**	0.003	0.74	**88%**
Bio-Medical-Llama-3-8B	0.7	47%	25%	15%	13%	0.701	0.007	**0.79**	60%
Bio-Medical-Llama-3-8B	1.0	48%	23%	16%	13%	0.691	0.007	**0.79**	61%
BioMistral-7B	0.3	25%	12%	14%	49%	**0.371**	0.004	**0.34**	**75%**
BioMistral-7B	0.7	7%	26%	40%	27%	0.329	0.011	0.20	34%
BioMistral-7B	1.0	7%	25%	40%	28%	0.327	0.011	0.20	35%
medalpaca-7b	0.3	14%	18%	32%	35%	**0.342**	0.012	**0.29**	**49%**
medalpaca-7b	0.7	2%	21%	60%	17%	0.289	0.015	0.11	19%
medalpaca-7b	1.0	2%	20%	64%	14%	0.290	0.008	0.14	17%
SMALL BASE MODELS									
Llama3-8B	0.3	38%	14%	14%	34%	**0.512**	0.008	**0.53**	**72%**
Llama3-8B	0.7	19%	30%	31%	20%	0.472	0.006	0.49	39%
Llama3-8B	1.0	10%	33%	43%	14%	0.424	0.007	0.41	25%
Llama3-8B-instruct	0.3	50%	8%	6%	37%	0.405	0.007	0.39	**87%**
Llama3-8B-instruct	0.7	42%	15%	15%	29%	**0.559**	0.007	0.59	70%
Llama3-8B-instruct	1.0	33%	23%	22%	22%	0.541	0.013	**0.60**	55%
Mistral-7B-Instruct	0.3	23%	10%	11%	54%	**0.321**	0.008	**0.29**	**79%**
Mistral-7B-Instruct	0.7	14%	16%	26%	44%	0.310	0.009	0.24	58%
Mistral-7B-Instruct	1.0	8%	21%	36%	35%	0.300	0.010	0.19	43%
llama1-7b	0.3	0%	9%	70%	20%	**0.201**	0.009	**0.00**	**20%**
llama1-7b	0.7	0%	6%	78%	16%	0.187	0.014	0.00	16%
llama1-7b	1.0	0%	5%	77%	18%	0.165	0.012	0.00	18%
SMALL GRANITE MODELS									
3.2-8b-instruct	0.3	30%	13%	13%	44%	**0.422**	0.005	**0.41**	**74%**
3.2-8b-instruct	0.7	19%	21%	28%	32%	0.393	0.008	0.38	51%
3.2-8b-instruct	1.0	11%	27%	38%	25%	0.376	0.007	0.32	36%
3.1-8b-instruct	0.3	32%	12%	12%	44%	**0.433**	0.006	**0.42**	**75%**
3.1-8b-instruct	0.7	19%	22%	28%	30%	0.405	0.007	0.39	49%
3.1-8b-instruct	1.0	12%	26%	36%	26%	0.368	0.011	0.32	39%
3.1-8b-base	0.3	11%	23%	37%	29%	**0.349**	0.008	**0.27**	**40%**
3.1-8b-base	0.7	2%	21%	60%	16%	0.283	0.007	0.13	18%
3.1-8b-base	1.0	0%	17%	67%	15%	0.234	0.010	0.00	16%
3.0-8b-base	0.3	17%	23%	49%	11%	**0.369**	0.008	0.60	**29%**
3.0-8b-base	0.7	4%	33%	61%	1%	0.293	0.010	0.75	5%
3.0-8b-base	1.0	1%	36%	63%	0%	0.244	0.010	**0.79**	1%
3.1-2b-instruct	0.3	22%	14%	11%	53%	**0.296**	0.011	**0.29**	**75%**
3.1-2b-instruct	0.7	11%	21%	27%	40%	0.263	0.010	0.22	52%
3.1-2b-instruct	1.0	6%	22%	38%	34%	0.244	0.009	0.15	40%
3.1-2b-base	0.3	9%	19%	33%	39%	**0.271**	0.010	**0.19**	**49%**
3.1-2b-base	0.7	1%	19%	59%	21%	0.239	0.011	0.04	21%
3.1-2b-base	1.0	0%	14%	68%	18%	0.217	0.010	0.03	19%

Table 2: Results of 0.99-consistency of models on the MedQA benchmark with 10 repetitions for 4 finetuned models, 4 base models, and for 6 models of the granite family.

temperature increases, the number of SURE questions decreases. However, unlike the MMLU-Redux case, the number of correct answers among SURE questions (RWS) tends to remain stable or with small decreases, with a few exceptions. The middle of table 2 displays the results for the 4 base models from where the medical models were finetuned. Overall, as expected, the average accuracy is considerably lower. However, the LLama3-8B-instruct model, at 0.7 temperature, had an impressive RWS of 0.59 with 70% SURE answers, results much better than the worst two finetuned models. Also, the same model at temperature 0.3 yielded a considerable 87% percentage of SURE questions.

5.2 Exploring Multiple Versions of a Model

We also performed an identical study using the same 6 different versions of the *granite* model (Granite Team 2024). Results are shown on the bottom of table 2. We observed here very similar results as in the previous models: the percentage of SURE questions at the best temperature for each model, going from 29% to 75%; the accuracy among SURE questions (RWS) similar to the average accuracy, with some exceptions. However, unlike in the base models, the best RWS performance were at the low temperature 0.3.

5.3 Aggregated Results over All Models

As we did with the MMLU-Redux benchmark, we looked into the correlation between the RWS results and the accuracy average. With MedQA we obtained much higher corre-

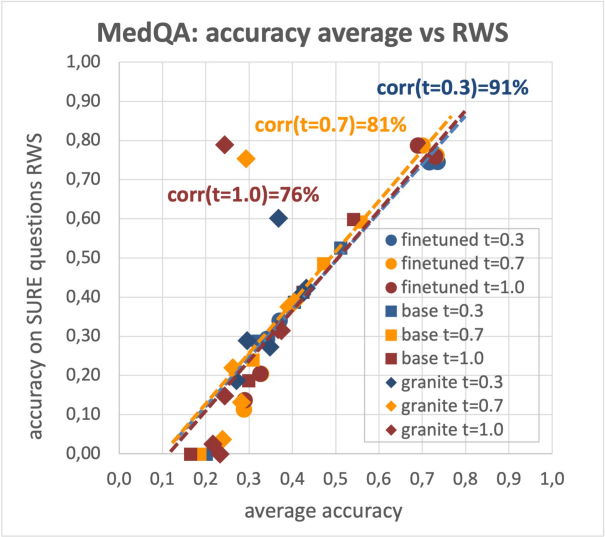

Figure 2: Scatter plot graphs showing the correlation between RWS and average accuracy (left) and between S/T and average accuracy (right) for the 4 finetuned and base models tested with MedQA benchmark across 3 temperatures.

lation values of 91%, 81%, and 76%, for temperatures 0.3, 0.7, and 1.0, respectively. As seen in the scatter plot depicted on the top of figure 2, except for the 3 granite models (corresponding to the results of 3.0-8b-base), the values of average accuracy and RWS are mostly aligned. Regression for $t = 0.3$ values yielded a coefficient of 1.234 and an intercept value of -0.119, with $R^2 = 0.831$.

Similarly as we did with MMLU-Redux, we computed the percentage of SURE questions (S/T) for each model at the best temperature level and obtained averages of $64\% \pm 14\%$, $42\% \pm 25\%$, and $35\% \pm 23\%$, for temperatures of 0.3, 0.7. an 1.0, respectively. The scatter plot on the bottom of figure 2 shows the values for all models. Unlike the case of MMLU-Redux, RWS values occupied a large range from 0% to 85%.

6 Limitations

The results and conclusions drawn from these studies may have been impacted by two key limitations. First, we are considering only multiple-choice benchmarks in this study, under the top-K sampling decoding method. Although this methodology warrants good precision in the reported results, it may have impacted them as discussed, for instance, in (Song et al. 2024). Second, we are not considering possible effects of benchmark contamination (Xu et al. 2024; Cheng, Chang, and Wu 2025; Ravaut et al. 2024) which may have inflated the number of consistently corrected answers due to *rogue memorization* (Cavalin et al. 2024), especially for medium models.

7 Conclusion and Future Work

This paper proposed a generic definition of answer consistency by an equivalence to oracle guessing at the same level, and then derived a formula for answer consistency in the case of repetitions of multiple-choice questions. We then suggested a compact representation of answer consistency as a pair of numbers, RWS | S/T, the proportion of correct answers among the consistent ones (RWS) followed by the percentage of consistent answers (S/T).

More details can be found in the extended version of this paper (Pinhanez et al. 2025) where we also propose a way to visualize the RWS — S/T representation, called *consistency plot*, which helps understanding phenomena like the effects of finetuning and of different inference temperatures.

In this work we have performed experimental studies with the MMLU-Redux and MedQA benchmarks using 26 models at 3 different temperatures. Our results indicated: (i) most small models produce only between 50% and 85% of consistent answers (S/T); (iii) medium size models display high level of consistency, above 95%; (iii) there is a high correlation between model average accuracy and the proportion of correct answers among the consistent ones (RWS) at low temperatures; and (iv) in the more generic benchmark MMLU-Redux, increasing temperature yielded a higher degree of accuracy among SURE answers, RWS. Visualizations of those results as well as detailed information about the studies are also available in the extended version of this paper (Pinhanez et al. 2025).

Several aspects remain to be explored. First, we did not consider benchmark contamination through memorization, what may have had significant impacts in the consistency measurements. Although we explored the granite family where the risk of contamination is likely to be smaller, we are still working reliable ways to filter out contaminated questions. Also, we want to go beyond simple repetition and use equivalent wordings for questions and answers, aiming to apply the methods to contexts beyond multiple-choice benchmarks. Finally, we also want to study alternative methods to determine whether a question is consistently answered by a model, during runtime, which are not based on repetitive API calls, a method which is often too expensive to be used in practice.

Acknowledgments

We would like the acknowledge the important participation in this paper of our intern Yago Primerano who was responsible for running many of the experiments of the paper. We would like also to acknowledge the support of the São Paulo Research Foundation (FAPESP, processo 2019/07665-4).

References

Aaron Grattafiori et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Ackerman, S.; Rabinovich, E.; Farchi, E.; and Anaby Tavor, A. 2024. A Novel Metric for Measuring the Robustness of Large Language Models in Non-adversarial Scenarios. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2794–2802. Miami, Florida, USA: Association for Computational Linguistics.

Atil, B.; Aykent, S.; Chittams, A.; Fu, L.; Passonneau, R. J.; Radcliffe, E.; Rajagopal, G. R.; Sloan, A.; Tudrej, T.; Ture, F.; Wu, Z.; Xu, L.; and Baldwin, B. 2025. Non-Determinism of "Deterministic" LLM Settings. arXiv:2408.04667.

Cavalin, P.; Domingues, P. H.; Pinhanez, C.; and Nogima, J. 2024. Fixing Rogue Memorization in Many-to-One Multilingual Translators of Extremely-Low-Resource Languages by Rephrasing Training Samples. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 4503–4514.

Cheng, Y.; Chang, Y.; and Wu, Y. 2025. A Survey on Data Contamination for Large Language Models. *arXiv preprint arXiv:2502.14425*.

DeepSeek-AI; and Aixin Liu et al. 2024. DeepSeek-V3 Technical Report. arXiv:2412.19437.

Dubey, A.; and et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Gema, A. P.; Leang, J. O. J.; Hong, G.; Devoto, A.; Mancino, A. C. M.; Saxena, R.; He, X.; Zhao, Y.; Du, X.; Madani, M. R. G.; Barale, C.; McHardy, R.; Harris, J.; Kaddour, J.; van Krieken, E.; and Minervini, P. 2025. Are We Done with MMLU? arXiv:2406.04127.

Granite Team, I. 2024. Granite 3.0 Language Models. https://github.com/ibm-granite/granite-3.0-language-models/. Accessed: 2025-12-02.

Han, T.; Adams, L. C.; Papaioannou, J.-M.; Grundmann, P.; Oberhauser, T.; Figueroa, A.; Löser, A.; Truhn, D.; and Bressem, K. K. 2025. MedAlpaca – An Open-Source Collection of Medical Conversational AI Models and Training Data. arXiv:2304.08247.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.-A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. arXiv:2310.06825.

Jiang, A. Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Hanna, E. B.; Bressand, F.; Lengyel, G.; Bour, G.; Lample, G.; Lavaud, L. R.; Saulnier, L.; Lachaux, M.-A.; Stock, P.; Subramanian, S.; Yang, S.; Antoniak, S.; Scao, T. L.; Gervet, T.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2024. Mixtral of Experts. arXiv:2401.04088.

Jin, D.; Pan, E.; Oufattole, N.; Weng, W.-H.; Fang, H.; and Szolovits, P. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14): 6421.

Labrak, Y.; Bazoge, A.; Morin, E.; Gourraud, P.-A.; Rouvier, M.; and Dufour, R. 2024. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. arXiv:2402.10373.

Lee, N.; Hong, J.; and Thorne, J. 2024. Evaluating the Consistency of LLM Evaluators. *arXiv preprint arXiv:2412.00543*.

Li, W.; Li, L.; Xiang, T.; Liu, X.; Deng, W.; and Garcia, N. 2024. Can Multiple-choice Questions Really Be Useful in Detecting the Abilities of LLMs? In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2819–2834. Torino, Italia: ELRA and ICCL.

Medical, P. 2024. MedLlama3 v20. https://huggingface.co/ProbeMedicalYonseiMAILab/medllama3-v20. Accessed: 2025-12-01.

Mirzadeh, I.; Alizadeh, K.; Shahrokhi, H.; Tuzel, O.; Bengio, S.; and Farajtabar, M. 2024. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. arXiv:2410.05229.

Nalbandyan, G.; Shahbazyan, R.; and Bakhturina, E. 2025. SCORE: Systematic COnsistency and Robustness Evaluation for Large Language Models. *arXiv preprint arXiv:2503.00137*.

Nori, H.; Lee, Y. T.; Zhang, S.; Carignan, D.; Edgar, R.; Fusi, N.; King, N.; Larson, J.; Li, Y.; Liu, W.; Luo, R.; McKinney, S. M.; Ness, R. O.; Poon, H.; Qin, T.; Usuyama, N.; White, C.; and Horvitz, E. 2023. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. arXiv:2311.16452.

Ouyang, S.; Zhang, J. M.; Harman, M.; and Wang, M. 2025. An empirical study of the non-determinism of chatgpt in code generation. *ACM Transactions on Software Engineering and Methodology*, 34(2): 1–28.

Patwardhan, A.; Vaidya, V.; and Kundu, A. 2024. Automated Consistency Analysis of LLMs. In *2024 IEEE 6th International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS-ISA)*, 118–127. IEEE.

Pezeshkpour, P.; and Hruschka, E. 2024. Large Language Models Sensitivity to The Order of Options in Multiple-Choice Questions. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Findings of the Association for Computational Linguistics: NAACL 2024*, 2006–2017. Mexico City, Mexico: Association for Computational Linguistics.

Pinhanez, C.; Cavalin, P.; Sanctos, C.; Grave, M.; and Primerano, Y. 2025. The Non-Determinism of Small

LLMs: Evidence of Low Answer Consistency in Repetition Trials of Standard Multiple-Choice Benchmarks. arXiv:2509.09705.

Qwen; :; Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Tang, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2025. Qwen2.5 Technical Report. arXiv:2412.15115.

Ravaut, M.; Ding, B.; Jiao, F.; Chen, H.; Li, X.; Zhao, R.; Qin, C.; Xiong, C.; and Joty, S. 2024. How Much are Large Language Models Contaminated? A Comprehensive Survey and the LLMSanitize Library. *arXiv preprint arXiv:2404.00699*.

Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Hou, L.; Clark, K.; Pfohl, S.; Cole-Lewis, H.; Neal, D.; Schaekermann, M.; Wang, A.; Amin, M.; Lachgar, S.; Mansfield, P.; Prakash, S.; Green, B.; Dominowska, E.; y Arcas, B. A.; Tomasev, N.; Liu, Y.; Wong, R.; Semturs, C.; Mahdavi, S. S.; Barral, J.; Webster, D.; Corrado, G. S.; Matias, Y.; Azizi, S.; Karthikesalingam, A.; and Natarajan, V. 2023. Towards Expert-Level Medical Question Answering with Large Language Models. arXiv:2305.09617.

Song, Y.; Wang, G.; Li, S.; and Lin, B. Y. 2024. The good, the bad, and the greedy: Evaluation of llms should not ignore non-determinism. *arXiv preprint arXiv:2407.10457*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Turing, A. M. 1939. Systems of logic based on ordinals. *Proceedings of the London Mathematical Society, Series 2*, 45: 161–228.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, H.; Zhao, S.; Qiang, Z.; Xi, N.; Qin, B.; and Liu, T. 2024. Beyond the Answers: Reviewing the Rationality of Multiple Choice Question Answering for the Evaluation of Large Language Models. arXiv:2402.01349.

Wei, F.; Chen, X.; and Luo, L. 2024. Rethinking Generative Large Language Model Evaluation for Semantic Comprehension. arXiv:2403.07872.

Xu, C.; Guan, S.; Greene, D.; Kechadi, M.; et al. 2024. Benchmark data contamination of large language models: A survey. *arXiv preprint arXiv:2406.04244*.