Testing Uncertainty of Large Language Models for Physics Knowledge and Reasoning

Elizaveta Reganova
Helmholtz AI Team Matter
Helmholtz-Zentrum Dresden-Rossendorf
01328 Dresden Germany
lisa.reganova@gmail.com

Peter Steinbach
Helmholtz AI Team Matter
Helmholtz-Zentrum Dresden-Rossendorf
01328 Dresden Germany
p.steinbach@hzdr.de

Abstract

Large Language Models (LLMs) have gained significant popularity in recent years for their ability to answer questions in various fields. However, these models have a tendency to "hallucinate" their responses, making it challenging to evaluate their performance. A major challenge is determining how to assess the certainty of a model's predictions and how it correlates with accuracy. In this work, we introduce an analysis[1] for evaluating the performance of popular open-source LLMs, as well as gpt-3.5 Turbo, on multiple choice physics questionnaires. We focus on the relationship between answer accuracy and variability in topics related to physics. Our findings suggest that most models provide accurate replies in cases where they are certain, but this is by far not a general behavior. The relationship between accuracy and uncertainty exposes a broad horizontal bell-shaped distribution. We report how the asymmetry between accuracy and uncertainty intensifies as the questions demand more logical reasoning of the LLM agent, while the same relationship remains sharp for knowledge retrieval tasks.

1 Introduction

Large language models (LLMs) have demonstrated remarkable performance across various text generation tasks, including question answering [1–3]. However, despite their impressive power and complexity, the capabilities of LLMs are inherently limited. These limitations stem from the finite nature of their training data, as well as the models' intrinsic memorization and limited reasoning capacities. Reliability is a critical component of LLM trustworthiness [4]. To build user trust, it is essential that models provide clear and accurate answers, preventing the spread of misinformation. One of the most significant challenges facing LLMs is the tendency to generate hallucinated responses [5]. In tasks like question answering, it is crucial to determine when we can trust the outputs of these models. Despite the recent advancements in natural language generation, there remains limited understanding of uncertainty in foundation models [6].

Uncertainty estimation involves quantifying the degree of confidence in the predictions made by a machine learning model [7–9]. Without proper measures of uncertainty, it is difficult to rely on generated text as a trustworthy source of information. A common approach to evaluating model performance is through the use of question-answering (QA) benchmarks, which come in various formats [10]. Among available formats, multiple-choice questions, which present multiple candidate answers alongside the input question, are the most popular. They offer a straightforward and efficient means of assessing model performance [10].

[1] All code and data is made available. For details, see Appendix D.

Machine Learning and the Physical Sciences Workshop, NeurIPS 2024.

Uncertainty estimation for machine learning has progressed to a well-studied field, particularly in the context of classification and regression tasks [8, 11]. In this work, we make an effort to assess the uncertainty of answers by an LLM agent on a physics specific multiple choice question and answer dataset [12]. Our contributions are as follows: (a) we obtain answers to high-school grade physics questions by 3 open-source and one closed-source LLM, (b) we compare the variation of answers between abovementioned LLMs and (c) we analyse the accuracy-certainty trade-off for each LLM in five question categories.

To our knowledge, this is the first publication which focuses on the trustworthiness of LLM answers in physics reasoning and physics related knowledge retrieval.

2 Method

2.1 Dataset

The `mlphys101` dataset [12] for our study consists of 823 university-level physics multiple-choice questions in English, each with five possible answers, among which one is correct. Corresponding one-letter answers are provided. The questions are classified into five categories:

- **D** Replication of Definitions, 153 question-answer pairs
- **F** Replication of Physical Facts, 138 question-answer pairs
- **C** Conceptual Physics and Qualitative Reasoning, 238 question-answer pairs
- **S** Single-Step Reasoning, 223 question-answer pairs
- **M** Multi-Step Reasoning, 71 question-answer pairs

We emphasize that this is the first dataset to our knowledge, which contains a large number of question and answer pairs tackling modern day physics topics up to the proficiency level of a physics bachelor degree or advanced high school degree. In this way, the dataset can help to push our understanding of how well LLMs are informed about the physical world in a language aligned with our current scientific description of it. Two datasets coming close to this in spirit are [13, 14]. However they both focus on situative physics effects rather than a vast range of topics in physics as a science. With this, the dataset used here provides a unique opportunity to study the compressed physics knowledge of LLMs and (hypothetically) their reasoning capabilities thereof.

For examples of the dataset, we provide one question and answer pair for each question category in the appendix A.1.

2.2 Models

To facilitate our analysis, we accessed open-source models on the BlaBlador server infracture provided by FZ Jülich (Germany). The infrastructure stores and runs a variety of LLM models. We accessed these models through a REST API mechanism compliant with the `openai-python` library. In this fashion, we were able to compare the performance of four LLMs: Llama3.1-8B-Instruct, Mixtral-8x7B-Instruct-v0.1, Mistral-7B-Instruct-v0.3, and GPT-3.5-turbo. We have set the temperature parameter for all models to a fixed value of 0.7.

All models were evaluated using a fixed few-shot prompting approach. They were asked one question at a time as the user input field. For each question or repetition of a question, the chat session was reset. To reduce the complexity of evaluation, we instructed the models to respond with only the letter corresponding to the correct answer.

We manually created examples for few-shot prompting similar to those in the dataset, see appendix B. To achieve this, we followed guidelines in [15]. After testing zero-shot, one-shot, two-shot, and three-shot prompting, we chose the three-shot approach. This method proved effective as it allowed not only Llama but also other models to follow instructions accurately. To create examples, we requested GPT-4 to generate similar samples based on the real dataset.

Since different models are sensitive to various styles of prompting, we filtered out whitespace, newline characters, and any potential explanations using regular expressions. Additionally, in a few instances, the Mistral model returned two letters as a response. In such cases, or when the response did not match the expected pattern, the replies were replaced with `None` and excluded from further analysis.

2.3 Uncertainty Estimation

In this work, we aim to evaluate the uncertainty of a LLM when generating answers to specific questions. To achieve this, we prompt the model with each question from the `mlphys101` dataset, repeating each prompt $N = 20$ times to gather N responses on the same question. For each question, we then assess the diversity in the model's answers by calculating the frequency with which each answer choice y_i ($y_i \in \{A, B, C, D, E\}$) appears across the $N = 20$ responses. This frequency serves as an approximation of the probability for each response choice in our discrete setting: $p(y_i|x, h) = \frac{count(y_i)}{N}$.

Using these probabilities, we compute the entropy $H(Y|x, h)$ of the model's responses Y to quantify the uncertainty for each question x when prompted with a specific three-shot prompt h:

$$H(Y|x, h) = -\sum_i p(y_i|x, h) * ln[p(y_i|x, h)]$$

This entropy measure allows us to gauge the level of consistency or uncertainty in the model's answers to individual questions.

3 Results and Discussion

In Figure 1, we show an overview of diversity of answers in all models regardless of whether the answers are correct or incorrect. For Llama3.1-8B-Instruct, Mixtral-8x7B-Instruct-v0.1, and Mistral-7B-Instruct-v0.3, our analysis shows a concentration of replies at entropy close to 0. Thus, these models appear more likely to provide responses with low entropy (low diversity, high consistency). In contrast, GPT-3.5-turbo tends to produce responses with higher diversity, i.e. more entries with entropy $H \geq 1$. In addition, Mixtral-8x7B-Instruct-v0.1 (being the largest model in use for our analysis) demonstrates a majority of entries in the lowest entropy bin and thus provides answers with minimal (if not vanishing) diversity. We can hypothesis at this point, that the degree of reliability is lowest with GPT-3.5-turbo as all question-answer pairs only exhibit one correct answer.

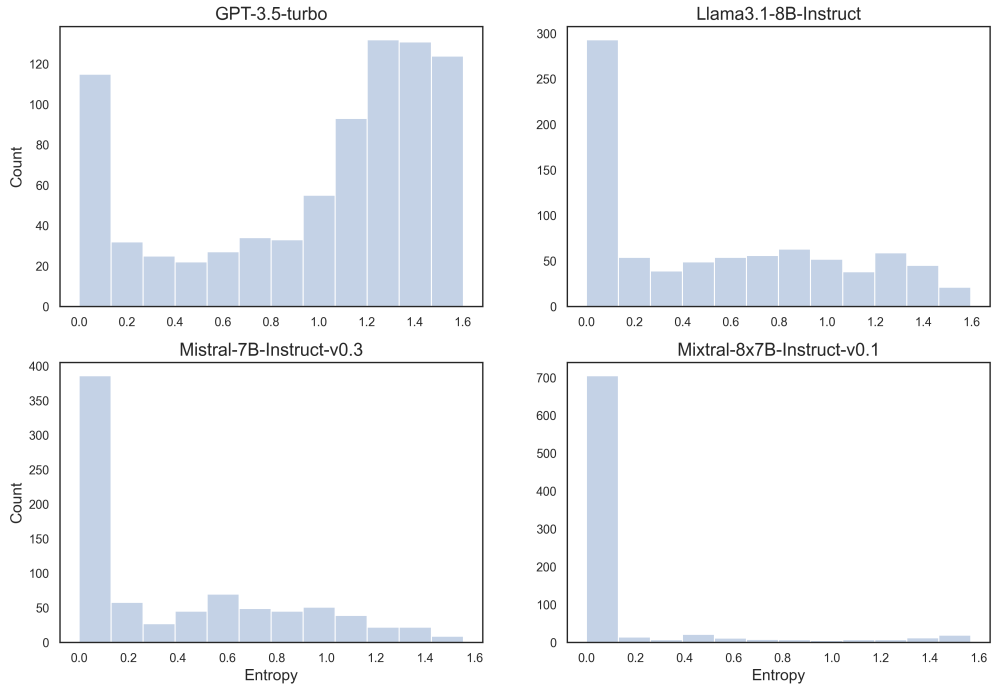

Figure 1: Entropy obtained from the distribution of answers to single questions of the `mlphys101` dataset [12] for all four models.

In order to study this hypothesis, we calculated and compared the error rate in responses versus the entropy of responses. This can provide direct insights in the degree of hallucination as LLM users assume the model to be correct no matter how often a question is raised. Figure 2 summarizes the results of this study.

The 2D histograms of "1-Accuracy" (or Error Rate) versus Entropy for all models exhibit a similar bell-shaped distribution (for a detailed mathematical explanation of the curve's shape, refer to Appendix E). In the bottom-left corner are questions where the models provide highly accurate answers with very low uncertainty. Conversely, in the top-right corner, questions are located where models give low accuracy responses with high uncertainty—indicating hallucinated answers. However, questions in the top-left corner reveal instances where models provide incorrect results with high certainty - which can also be attributed to the effect of hallucination. Figure 2 also illustrates that not all models are created equal in this regard. Diversity is lowest in the results of Mixtral-8x7B-Instruct-v0.1. Diversity is highest with GPT-3.5-turbo. We refer the curious reader to Figure 4 for a detailed account of this analysis.

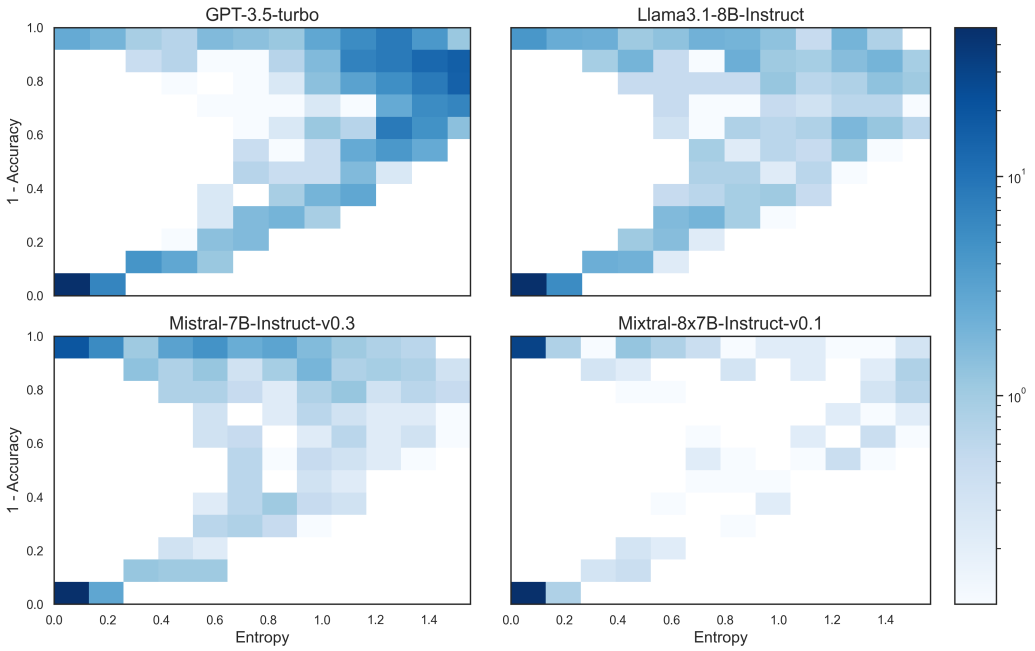

Figure 2: Two-dimensional Histogram of Error Rate (1 - Accuracy) vs. Entropy across Models. The binning of entropy is identical to Figure 1.

Additionally, we examined the accuracy-certainty trade-off for each LLM across the five question categories. Figure 3 summarizes our findings. The overall shape of the accuracy-certainty curve (Figure 2) remains consistent across a majority of categories for all models below 10 billion parameters. A distinct behavior is visible, where diversity increases with the question category becoming more complex (complexity increases from **D** to **M**). For single-step **S** and multi-step reasoning questions **M**, GPT-3.5-turbo yields a minimal number of correct replies with high diversity. Mixtral-8x7B-Instruct-v0.1 can provide correct and incorrect answers at low diversity. The remaining models perform in the middle ground between these two extremes. We further suggest, that failure of LLMs in single and multi-step reasoning questions is inline with findings from other fields [16, 17].

One of the main limitations of this work stems from the prompting approach. Previous studies have shown that response accuracy is highly dependent on the prompting context and style used [18]. This might explain the high diversity observed in the entropy plot for the GPT-3.5-turbo model, suggesting the need for further experimentation with different prompting techniques. However, we hypothesize that the observed variability is unlikely to affect the overall shape of the error rate versus entropy curve. This consideration requires validation through additional testing and comparisons to other datasets.

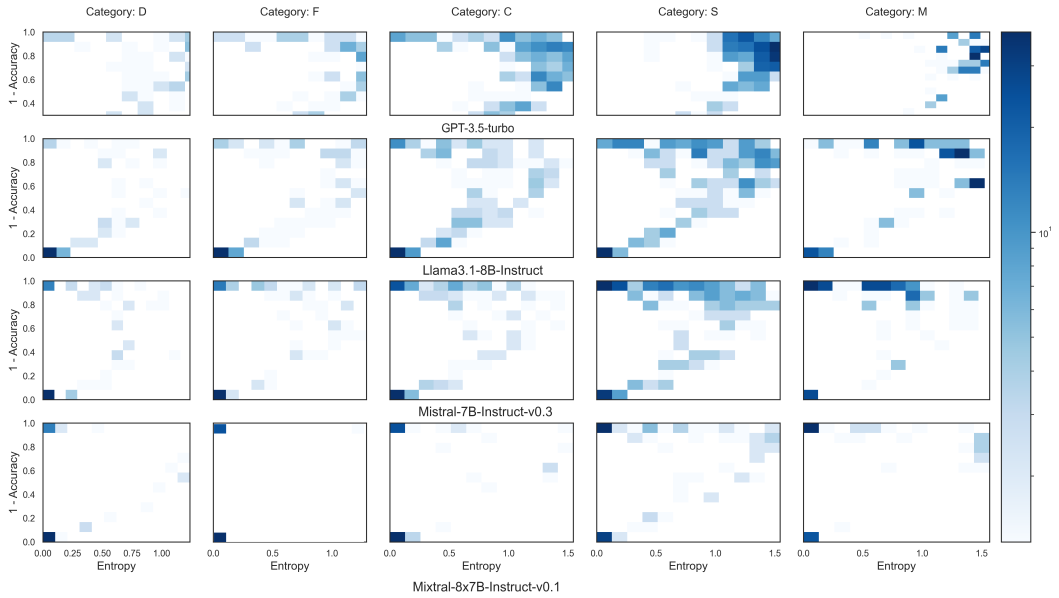

Figure 3: Accuracy-certainty trade-off for each LLM in five question categories

4 Summary

Reliability of LLMs is an important component towards their trustworthiness. Hallucinations and inconsistency of model's reply may lead to incorrect replies and losing users' trust. Hallucinations may often appear in narrow domains possibly due to the lack of training data. For this reason we have proposed a pipeline for evaluating an accuracy-uncertainty trade-off. We tested it on a physics MCQ dataset for four popular LLMs. The dataset exposes questions and answer pairs at different levels of complexity and reasoning demand. The experiment has shown the difference in consistency of responses and hallucinating depending on model size and question complexity. A downstream analysis has to be undertaken, to identify the root cause of these observations and compare these findings to different datasets of similar nature.

References

[1] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL `https://arxiv.org/abs/2303.08774`.

[2] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023.

[3] Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. Toolqa: A dataset for llm question answering with external tools, 2023. URL `https://arxiv.org/abs/2306.13304`.

[4] Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. Trustworthy llms: a survey and

guideline for evaluating large language models' alignment, 2024. URL https://arxiv.org/abs/2308.05374.

[5] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

[6] Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. Lm-polygraph: Uncertainty estimation for language models, 2023. URL https://arxiv.org/abs/2311.07383.

[7] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110(3):457–506, March 2021. ISSN 1573-0565. doi: 10.1007/s10994-021-05946-3. URL http://dx.doi.org/10.1007/s10994-021-05946-3.

[8] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. A survey of uncertainty in deep neural networks. URL http://arxiv.org/abs/2107.03342.

[9] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. 76:243–297. ISSN 1566-2535. doi: 10.1016/j.inffus.2021.05.008. URL https://www.sciencedirect.com/science/article/pii/S1566253521001081.

[10] Wangyue Li, Liangzhi Li, Tong Xiang, Xiao Liu, Wei Deng, and Noa Garcia. Can multiple-choice questions really be useful in detecting the abilities of llms?, 2024. URL https://arxiv.org/abs/2403.17752.

[11] Yarin Gal et al. Uncertainty in deep learning. 2016.

[12] Marcel Völschow, P. Buczek, C. Carreno-Mosquera, E. Reganova, J. Roldan-Rodriguez, P. Steinbach, and A. Strube. mlphys101 - exploring the performance of large-language models in multilingual undergraduate physics education. publication submitted, but unpublished, 2024. URL https://rodare.hzdr.de/record/3137.

[13] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language, 2019. URL https://arxiv.org/abs/1911.11641.

[14] Yi Wang, Jiafei Duan, Dieter Fox, and Siddhartha Srinivasa. NEWTON: Are large language models capable of physical reasoning? In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9743–9758, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.652. URL https://aclanthology.org/2023.findings-emnlp.652.

[15] Meta. Prompting. https://llama.meta.com/docs/how-to-guides/prompting/, 2024. Accessed: 2024-09-10.

[16] Marianna Nezhurina, Lucia Cipolina-Kun, Mehdi Cherti, and Jenia Jitsev. Alice in wonderland: Simple tasks showing complete reasoning breakdown in state-of-the-art large language models, 2024. URL https://arxiv.org/abs/2406.02061.

[17] Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. Are emergent abilities in large language models just in-context learning?, 2024. URL https://arxiv.org/abs/2309.01809.

[18] Daniel Machlab and Rick Battle. Llm in-context recall is prompt dependent, 2024. URL https://arxiv.org/abs/2404.08865.

A Appendix

A.1 Dataset Examples

As discussed in section 2.1, our dataset consists of 5 classes of questions. We provide one example for each here to provide a more comprehensive insight into the variety of topics and requirements for an answer.

A.1.1 Replication of Definitions, D

Question The property of a moving object to continue moving is what Galileo called

Answer A velocity.

Answer B speed.

Answer C acceleration.

Answer D inertia.

Answer E direction.

A.1.2 Replication of Physical Facts, F

Question ____ are examples of vector quantities..

Answer A Acceleration and time

Answer B Velocity and acceleration

Answer C Volume and velocity

Answer D Mass and volume

Answer E Time and mass

A.1.3 Conceptual Physics and Qualitative Reasoning, C

Question If an object is moving, then the magnitude of its ____ cannot be zero.

Answer A speed

Answer B velocity

Answer C acceleration

Answer D A and B

Answer E A, B, and C

A.1.4 Single-Step Reasoning, S

Question A firefighter with a mass of 70 kg slides down a vertical pole, accelerating at $2m/s^2$. The force of friction that acts on the firefighter is

Answer A $70\,N$.

Answer B $560\,N$.

Answer C $140\,N$.

Answer D $700\,N$.

Answer E $0\,N$.

A.1.5 Multi-Step Reasoning, M

Question A bowling ball at a height of 36 meters above the ground is falling vertically at a rate of 12 meters per second. Which of these best describes its fate?

Answer A It will hit the ground in exactly three seconds at a speed of $12m/s$.

Answer B It will hit the ground in less than three seconds at a speed greater than $12m/s$.

Answer C It will hit the ground in more than three seconds at a speed less than $12m/s$.

Answer D It will hit the ground in less than three seconds at a speed less than $12m/s$.

Answer E It will hit the ground in more than three seconds at a speed greater than $12m/s$.

B Few-shot prompting

We use few-shot prompting to trigger the LLM for an answer. Here is an example prompt to illustrate our strategy.

```
system(''You're a highly knowledgeable physics tutor. For each message,
give only the letter of the correct answer without any
explanations or additional information.''),

user(''A ball rolls down a slope and accelerates uniformly
at 2 m/s^2. If it starts from rest, what will be its speed after
3 seconds? A. 3 m/s, B. 4 m/s, C. 5 m/s, D. 6 m/s, E. 7 m/s''),

assistant(''D''),

user(''A cyclist accelerates uniformly from rest to a speed
of 10 m/s in 5 seconds. What is their acceleration?
A. 1 m/s^2, B. 2 m/s^2, C. 3 m/s^2, D. 4 m/s^2, E. 5 m/s^2''),

assistant(''B''),

user(''A rocket accelerates from rest at a constant rate of
6 m/s^2. What speed will it reach after 4 seconds?
A. 12 m/s, B. 18 m/s, C. 24 m/s, D. 30 m/s, E. 36 m/s''),

assistant(''C''),

user(''<question to evaluate goes here>''),
```

Details on how the few shot prompts were created are given in section 2.2.

C Error rate vs. Entropy 2D Histogram with exact values

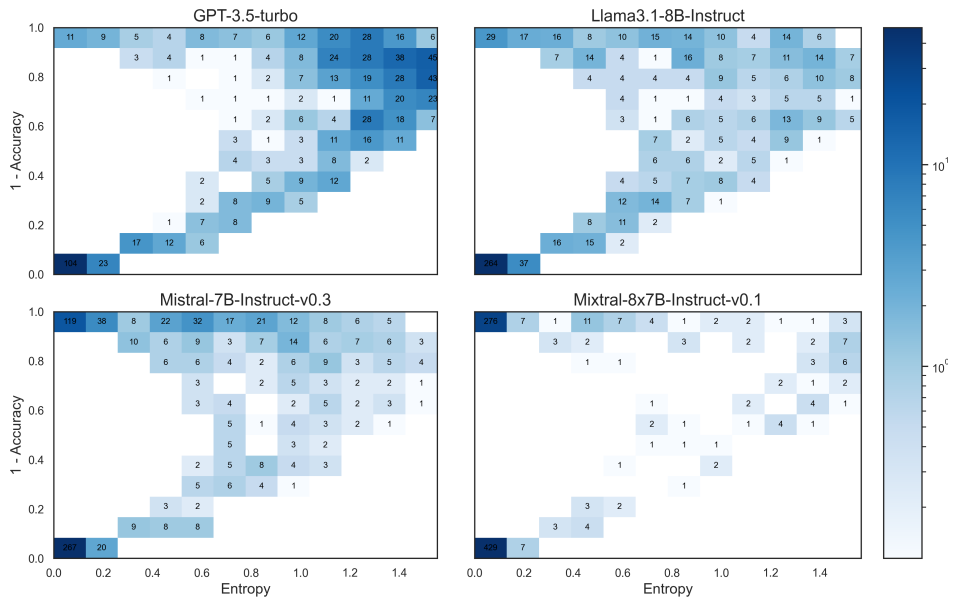

Figure 4: Two-dimensional Histogram of Error Rate (1 - Accuracy) vs. Entropy across Models with counts per bin. Entries are identical to Figure 2.

D Code and Data Availability

We make our analysis software available at here. The `mlphys101` dataset is available here until it has been published in a peer-reviewed journal.

E Shape of the Curves

To analyze the shape of the curve, let us consider different scenarios of a model's response patterns separately:

1. The model generates only correct responses for a given question.
2. The model generates only incorrect responses.
3. The model generates two distinct responses, one of which is correct and the other incorrect.
4. The model generates three or more distinct responses, with at least one being correct.

When the model consistently produces correct responses, both the entropy and the error rate are zero (accuracy equals 1). On the curve, this scenario corresponds to the bottom-left corner. A higher density of points in this region indicates that the model often generates accurate and consistent responses.

When the model fails to generate any correct responses, all points align along the horizontal line where the error rate equals 1. The top-left corner of the curve represents scenarios where the model's responses are consistently incorrect.

In the case where the model produces only two different responses - one correct and one incorrect — the entropy $H(Y|x,h)$ is given by:

$$H(Y|x,h) = -p(y_{\text{correct}}|x,h) * \ln[p(y_{\text{correct}}|x,h)] - p(y_{\text{incorrect}}|x,h) * \ln[p(y_{\text{incorrect}}|x,h)].$$

Substituting $p(y_{\text{correct}}|x,h) =$ accuracy and $p(y_{\text{incorrect}}|x,h) = 1 -$ accuracy, the entropy can be expressed as:

$$\begin{aligned} H(Y|x,h) &= -(\text{accuracy}) * \ln(\text{accuracy}) - (1 - \text{accuracy}) * \ln(1 - \text{accuracy}), \\ H(Y|x,h) &= -(1 - \text{error\_rate}) * \ln(1 - \text{error\_rate}) - \text{error\_rate} * \ln(\text{error\_rate}). \end{aligned} \tag{1}$$

The resulting curve is illustrated in Figure 5(A). This pattern can also be observed in the curves for the Mistral and Llama models (see Figure 1).

When a model's responses include correct replies and more than one distinct incorrect replies (i.e., two or more versions of incorrect output), we obtain families of parameterized curves with $(\#$ of incorrect types $- 2)$ parameters.

For example, in the simplest case, let us consider three distinct responses: $p(y_{\text{correct}}|x,h)$, $p(y_{\text{incorrect\_1}}|x,h)$, and $p(y_{\text{incorrect\_2}}|x,h)$. The probabilities can be defined as follows:

$$\begin{aligned} p(y_{\text{correct}}|x,h) &= \text{accuracy} = 1 - \text{error\_rate}, \\ p(y_{\text{incorrect\_1}}|x,h) &= p(y_{\text{incorrect\_1}}|x,h) = p\_\text{i1}, \\ p(y_{\text{incorrect\_2}}|x,h) &= \text{error\_rate} - p\_\text{i1}. \end{aligned}$$

In this case, the entropy $H(Y|x,h)$ is given by:

$$\begin{aligned} H(Y|x,h) = &-(1 - \text{error\_rate}) * \ln(1 - \text{error\_rate}) - p\_\text{i1} * \ln(p\_\text{i1}) - \\ &-(\text{error\_rate} - p\_\text{i1}) * \ln(\text{error\_rate} - p\_\text{i1}). \end{aligned} \tag{2}$$

Examples of such curves, for $p\_\text{i1} = \{0.1, 0.3, 0.5, 0.7, 0.9\}$, are shown in Figure 5(B).

Similarly, the equations for four and five distinct responses among replies to a single query can be expressed as follows:

For four types of responses:

$$H(Y|x,h) = -(1 - \text{error\_rate}) * \ln(1 - \text{error\_rate}) - p\_\text{i1} * \ln(p\_\text{i1}) -$$

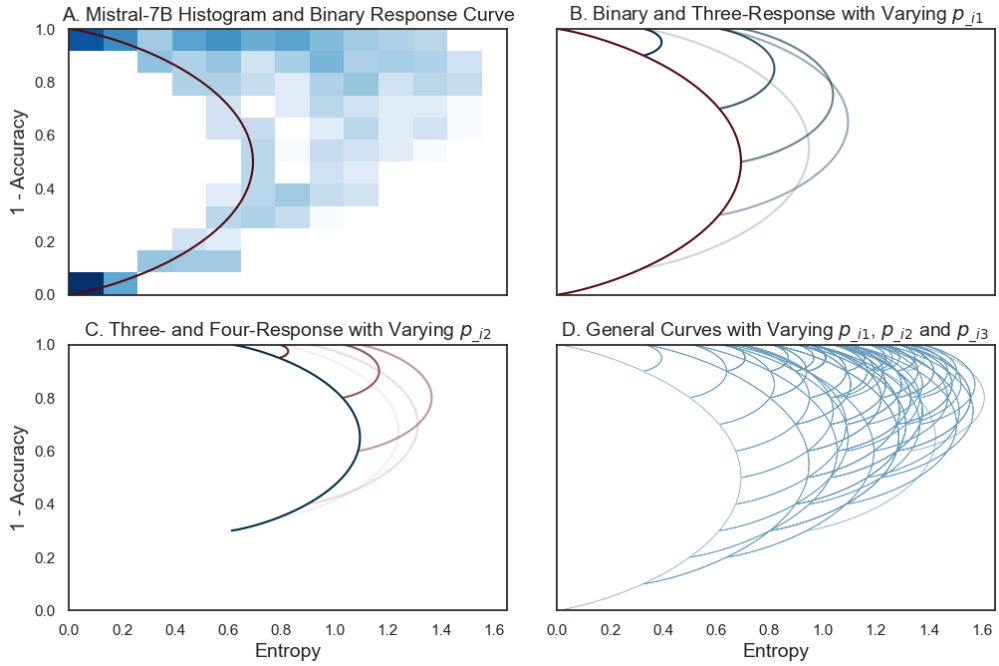

Figure 5: A. Two-dimensional histogram of (1 - Accuracy) vs. Entropy for the Mistral 7B model, shown alongside the theoretical curve (red) representing the scenario where the model provides only two distinct responses, one of which is correct (see Equation1). B. Theoretical curves for binary responses (red) and three distinct responses (blue) with p_{i1} = {0.1, 0.3, 0.5, 0.7, 0.9} (see Equation 2). Curve intensity increases as p_{i1} increases. C. Theoretical curves (Equation 3) for three (blue) and four (red) distinct responses, with p_{i1} = 0.3 and p_{i2} = {0.05, 0.1, 0.3, 0.5, 0.65}. Curve intensity increases as p_{i2} increases. D. Theoretical curves based on the general equation 4 with parameters p_{i1}, p_{i2}, and p_{i3} varying within {0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}.

$$-p_{\text{i2}} * \ln(p_{\text{i2}}) - (\text{error\_rate} - p_{\text{i1}} - p_{\text{i2}}) * \ln(\text{error\_rate} - p_{\text{i1}} - p_{\text{i2}}). \tag{3}$$

For five types of responses:

$$H(Y|x,h) = -(1-\text{error\_rate}) * \ln(1-\text{error\_rate}) - p_{\text{i1}} * \ln(p_{\text{i1}}) - p_{\text{i2}} * \ln(p_{\text{i2}}) - p_{\text{i3}} * \ln(p_{\text{i3}}) - \\ -(\text{error\_rate} - p_{\text{i1}} - p_{\text{i2}} - p_{\text{i3}}) * \ln(\text{error\_rate} - -p_{\text{i1}} - p_{\text{i2}} - p_{\text{i3}}). \tag{4}$$

All the equations previously described in this section are special cases of Equation 4, where some of the probabilities are equal to zero (see Figure 5(D)).