



HAL
open science

THE 20 QUESTIONS GAME TO DISTINGUISH LARGE LANGUAGE MODELS

Gurvan Richardeau, Erwan Le Merrer, Camilla Penzo, Gilles Trédan

► **To cite this version:**

Gurvan Richardeau, Erwan Le Merrer, Camilla Penzo, Gilles Trédan. THE 20 QUESTIONS GAME TO DISTINGUISH LARGE LANGUAGE MODELS. 2024. hal-04699271

HAL Id: hal-04699271

<https://hal.science/hal-04699271v1>

Preprint submitted on 16 Sep 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

THE 20 QUESTIONS GAME TO DISTINGUISH LARGE LANGUAGE MODELS

Gurvan Richardeau, Erwan Le Merrer, Camilla Penzo, Gilles Tredan

ENS Paris-Saclay, PEReN, INRIA, IRISA, LAAS-CNRS

ABSTRACT

In a parallel with the 20 questions game, we present a method to determine whether two large language models (LLMs), placed in a black-box context, are the same or not. The goal is to use a small set of (benign) binary questions, typically under 20. We formalize the problem and first establish a baseline using a random selection of questions from known benchmark datasets, achieving an accuracy of nearly 100% within 20 questions. After showing optimal bounds for this problem, we introduce two effective questioning heuristics able to discriminate 22 LLMs by using half as many questions for the same task. These methods offer significant advantages in terms of stealth and are thus of interest to auditors or copyright owners facing suspicions of model leaks.

Index Terms— LLMs, black-box distinguishability.

1. INTRODUCTION

Auditing AI models in a black-box interaction scheme is a difficult task attracting a considerable attention today [?, ?]. Although there is now an important set of works for auditing machine learning classifiers (whether for fairness [?], security [?], intellectual property matters [?] or others), the relative novelty of large language models (LLMs) calls for adapted auditing methods. In this paper, we tackle the possibility of assessing the simple, yet challenging question of efficiently differentiating LLMs only by prompting them. This task has important applications such as being a preamble for a potential prosecution (in the case of model stealing suspicion, see e.g. [?] for the same task with classifier models), or evaluating the convergence of these models towards a high accuracy on predefined set of regulatory prompts (thus leading in hardness to discriminate them).

In this paper, and after similar attempts in the domain of image classification, [?] for example, we take the step to address the issue using *benign* questions, i.e. questions that are not *adversarially* crafted [?]. These benign inputs have the advantage of being less prone to defenses on the model side [?], as they follow the genuine data distribution.

2. PROBLEM STATEMENT

We denote \mathcal{Q} the *infinite* set of all questions (i.e. prompts or queries) that have binary answers. In this context, we define a *LLM* or a *model* m as a deterministic map in $\mathcal{M} \stackrel{\text{def}}{=} \{0, 1\}^{\mathcal{Q}}$.

Let then μ denote the *empirical distribution of the models*. Specifically, a representative sample from μ is obtained by randomly selecting a model from the collection of all existing models, e.g. those available on Hugging Face.

2.1. The Problem of Distinguishing LLMs

Consider an auditor having a prompt-only access to two such LLMs (m, m') . Her objective is to know whether m and m' are the same model or if they differ. If m and m' are indeed the same model, the two would provide identical responses to any given prompt. However, even if m and m' differ, they may provide identical answers to (at least) some of the prompts. In the following, we will formalize the question of whether m and m' are different and the question of which prompts we should select to efficiently distinguish the two given models.

Definition 2.1 (Distinguishing LLMs). *Given two unknown models m, m' from \mathcal{M} , find a set of k requests that maximizes the probability of correctly rejecting hypothesis $\mathcal{H}_0 : m = m'$ when $\mathcal{H}_1 : m \neq m'$ is true. In other words, accurately distinguish two distinct LLMs within k prompts.*

This involves analyzing the following map:

$$S : \mathbb{N}^* \longrightarrow [0, 1] \\ k \mapsto \max_{Q \subset \mathcal{Q}, |Q|=k} \text{acc}(Q) \quad (1)$$

The accuracy being:

$$\text{acc}(Q) = \mathbb{P}(T_Q = 0, \mathcal{H}_0) + \mathbb{P}(T_Q = 1, \mathcal{H}_1), \quad (2)$$

with T_Q being the test such that the two models are detected as equal if and only if they answer the same on every questions of Q , i.e.:

$$T_Q : (m, m') \mapsto \begin{cases} 0 & \text{if } m(Q) = m'(Q) \quad (\text{accept } \mathcal{H}_0) \\ 1 & \text{if } m(Q) \neq m'(Q) \quad (\text{reject } \mathcal{H}_0) \end{cases}$$

Here, the game becomes devising algorithms that find the best sets of questions which solve (??) for a small number k of questions.

2.2. Experimental Assumptions

To compute the accuracy (??), we need the probability that the two models are the same, that is, $\mathbb{P}(\mathcal{H}_0)$. This probability is either the result of our suspicion that the two models are different (e.g., *secret retraining* of a stolen model) or our suspicion that the two are the same (e.g., direct use of a stolen model). We set $\mathbb{P}(\mathcal{H}_0) = \frac{1}{2}$, indicating that we have no stronger belief that the models are identical or different.

We note that given the deterministic behaviours of our models, true positives are trivial in our experiment ($\mathbb{P}(T_Q = 0 \mid \mathcal{H}_0) = 1$). Therefore, (??) can be tackled from the true negatives point of view:

$$\arg \max_{Q \in \mathcal{Q}, |Q|=k} \text{acc}(Q) = \arg \max_{Q \in \mathcal{Q}, |Q|=k} \mathbb{P}(T_Q = 1 \mid \mathcal{H}_1). \quad (3)$$

We therefore focus our analysis on the case $m \neq m'$.

3. OPTIMAL NUMBER OF QUESTIONS

We define a query as optimal when it separates a set of models into two equal groups according to their answers. If the two groups are evenly divided, the number of differentiated pairs is maximized with a single question (see Appendix ?? for proof). Let X be the number of queries necessary to differentiate a pair of models (m, m') s.t. $m \neq m'$ and s.t. m and m' are randomly drawn uniformly from the infinite set of models \mathcal{M} where the queries are optimal (see Th. ??). Mathematically, let $(m, m') \sim \mathcal{U}(\mathcal{M})^{\otimes 2}$ s.t. $m \neq m'$. Let $X = \inf\{k \in \mathbb{N}^* \mid m(Q^*(k)) \neq m'(Q^*(k))\}$ where Q^* is \mathcal{Q} ordered such that the queries are optimal. The law of X is then $P(X \leq k) = 1 - (\frac{1}{2})^k, \forall k \in \mathbb{N}^*$ (see Appendix ??).

4. EXPERIMENTAL SETUP

To tackle (??), we need a set of questions \mathcal{Q} and the ability to sample from the distribution μ . To achieve this, we select 22 models (listed in Figure ?? in Appendix, set we coin \mathcal{M}) and K questions (see Table ??) from HuggingFace.

As seen in (??), we are interested in the true negatives, which we approximate by Monte Carlo as follows:

$$\mathbb{P}(T_Q = 1 \mid \mathcal{H}_1) \propto \sum_{\substack{(m, m') \in \mathcal{M}^2 \\ m \neq m'}} \mathbb{1}_{m(Q) \neq m'(Q)}.$$

Finally, we seek for maximizers over \mathcal{Q} of (??). An optimal, yet non-tractable, algorithm would require looking at as many sets as the binomial coefficient $\binom{|\mathcal{Q}|}{\text{max_size}}$, where max_size is the maximum size of the question set we consider. We will focus experimentally on the set sizes of max_size = 20, making such an optimal algorithm intractable. We thus propose different *heuristics* in Sect. ??.

The Similarity of LLMs on Their Responses Figure ?? represents the distribution of models answers for each question of the set K . For instance, the point $x = 20, y = 0.043$

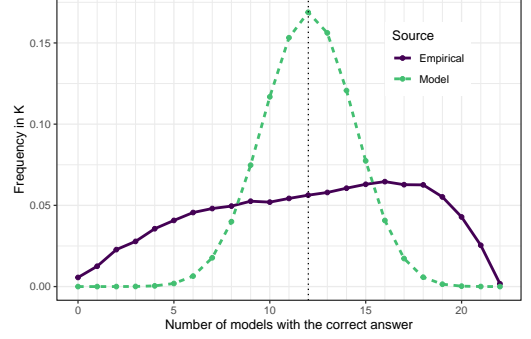


Fig. 1: (plain/purple) Distribution of correct answers for questions in K , compared to (dashed/green) a random binomial model. Vertical line: average of both distributions.

means that 4.3% of the questions of K were correctly answered by 20 of the 22 tested models. Using the average correct answer rate ($p = 54.6\%$), we present the corresponding binomial distribution (number of success with 22 coin tosses at probability p). This model assumes that each LLM answers independently correctly. The resulting bell shaped distribution differs drastically from the empirical observation: this highlights a lack of independence from the models. This means that in practice some questions are easy and correctly answered by most models, while almost all models fail at correctly answering some other (hard) questions. Motivated by this lack of independence between models, we now look at an experimental approach to identify good question sets.

Experimental method We consider heuristic algorithms of the form $\mathcal{A} : (\mathcal{M}, \mathcal{Q}), \mapsto \lambda \in [0, 1]^{\mathcal{Q}}$ that yields a *discriminating score* λ such that for $q \in \mathcal{Q}$, the closer $\lambda(q)$ is to 1, the better q is supposed to distinguish models of μ .

Then, for each pair (m, m') of \mathcal{M} models, we query both models with the questions from \mathcal{Q} ordered decreasingly by λ , until they answer differently (i.e. we distinguished them) and record the corresponding number of questions. From this we plot the cumulative distribution function that indicates how many model pairs can be distinguished as a function of the number of questions. See Alg. ??, where: $t(Q, m, m') = \min_{j \in [1, \dots, |\mathcal{Q}|]} \{j \mid m(Q[j]) \neq m'(Q[j])\}$.

A Baseline: Selecting Questions at Random We define the Random heuristic $\mathcal{A}_{\text{rand}}$ as $\mathcal{A}_{\text{rand}}(\mathcal{M}, \mathcal{Q}) = q \mapsto \mathcal{U}([0, 1])$. Fig ?? (a) illustrates that by randomly selecting questions uniformly across all datasets, we achieve an average accuracy of 95% with 6 questions. The best case reached within the 2000 runs (green curve) is based on AUC (Area Under Curve). The figure representing each question dataset is omitted, as the differences are not substantial: achieving 95% accuracy requires 5 questions for the “best” dataset vs 7 for the “worst”.

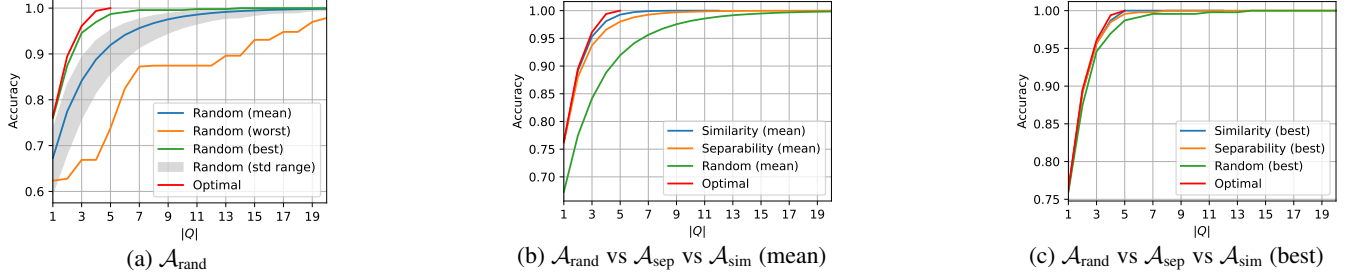


Fig. 2: Map $S : k \mapsto \max_{Q \in \mathcal{Q}, |Q|=k} \text{acc}(Q)$, for all questions datasets combined and the 22 LLMs. Each heuristic has been run 2000 times, and we present the mean, std, best and worst cases (based on AUC).

5. TWO HEURISTICS FOR DISTINGUISHABILITY

In the following, $Q \subset \mathcal{Q}$ and $M \subset \mathcal{M}$ denote finite subsets of questions and models. N is a min-max normalization.

Algorithm 1 \mathcal{A}_{sep} : The Separability Heuristic

Require: $\mathcal{M}, \mathcal{Q}, \mathcal{A}$ (a heuristic).
1: $\lambda \leftarrow \mathcal{A}(\mathcal{M}, \mathcal{Q})$
2: $\mathcal{Q}_{\mathcal{A}} \leftarrow \text{sort}(\mathcal{Q}; \text{by decreasing } \lambda)$
3: Let \mathcal{D} an empty dict with default values equal to 0.
4: **for** each $(m, m') \in \mathcal{M}^2$, s.t. $m \neq m'$ **do**
5: $n_{\text{queries}} \leftarrow t(\mathcal{Q}_{\mathcal{A}}, m, m')$
6: $\mathcal{D}[n_{\text{queries}}] \leftarrow \mathcal{D}[n_{\text{queries}}] + 1$
7: **end for**
8: **return** \mathcal{D}

5.1. The Separability Heuristic

We here focus on individual questions: what makes it a good question? An initial measure of distinguishing power is how evenly the binary question divides the two groups of models. The heuristic is given in Algorithm ??, here is a formal definition:

Definition 5.1 (Subset Separability). *Separability of $X \subset M$:*

$$\Delta_M : \mathcal{P}(M) \longrightarrow [0, 1]$$

$$X \mapsto N(|M| - ||X| - |\bar{X}||),$$

where $\bar{X} = M \setminus X$.

We denote $M_q \stackrel{\text{def}}{=} \{m \in M \mid m(q) = 1\}$ such that the highest $\Delta_M(M_q)$ is, the better q divides the models evenly. Finally, we define the heuristic \mathcal{A}_{sep} as follows:

$$\mathcal{A}_{\text{sep}}(M, Q) = q \mapsto \begin{cases} \mathcal{U}([0, 1]) & \text{if } q \in \arg \max_{q \in \mathcal{Q}} \Delta_M(M_q) \\ 0 & \text{else} \end{cases} \quad (4)$$

5.2. The Recursive Similarity Heuristic

We now consider sets of questions: what makes it a good question set? Sampling from questions with maximum separability makes it possible for two questions to obtain the same

partitions (i.e. the same two groups of models), making it pointless to use them both. Hence, the need for a metric that compares the similarity between the two partitions. We propose a heuristic that constructs a sequence of questions using a recursive approach, ensuring that each subsequent question is as dissimilar as possible from the previous.

Definition 5.2 (Similarity of two partitions of same separability). *For $X, Y \subset M$ s.t. $\Delta_M(X) = \Delta_M(Y)$, we define the similarity Γ_M , as:*

$$\Gamma_M : \mathcal{P}(M) \times \mathcal{P}(M) \longrightarrow [0, 1]$$

$$(X, Y) \mapsto N(\max\{|X \cap Y|, |\bar{X} \cap Y|, |X \cap \bar{Y}|, |\bar{X} \cap \bar{Y}|\})$$

Then we define the heuristic \mathcal{A}_{sim} with Algorithm ?? such that $\mathcal{A}_{\text{sim}}(M, Q) = \text{RecursiveSim}(M, Q, \text{empty list})$. This algorithm initially selects two questions with the least similarity (line 3), a process that is tractable for two questions. Subsequently, it iteratively adds a question that is the least similar to the existing ones (line 8). Each search for a new question incurs a linear cost, ensuring the algorithm remains tractable. Finally, as this process yields very good discriminating questions, we can define a limit of iterations (`max_iter`) making the algorithm even more efficient (the questions that have not been added have a score set to 0).

Algorithm 2 \mathcal{A}_{sim} : The Recursive Similarity Heuristic

Require: M, Q, Q_{sim}
1: **if** Q_{sim} is empty **then**
2: $Q_{\text{sep}}^* \leftarrow \arg \max_{q \in Q} \Delta_M(M_q)$
3: Let $q_1, q_2 \in \arg \min_{(q, q') \in Q^2} \Gamma_M(M_q, M_{q'})$
4: $Q_{\text{sim}} \leftarrow \text{append}(q_1, q_2)$
5: **return** $\text{RecursiveSim}(M, Q_{\text{sep}}^*, Q_{\text{sim}})$
6: **end if**
7: **if** $\text{len}(Q_{\text{sim}}) \leq \text{max_iter}$ **then**
8: Let $q \in \arg \min_{q \in Q \setminus Q_{\text{sim}}} \{\sum_{q' \in Q_{\text{sim}}} \Gamma_M(M_q, M_{q'})\}$
9: $Q_{\text{sim}} \leftarrow \text{append}(q)$
10: **return** $\text{RecursiveSim}(M, Q, Q_{\text{sim}})$
11: **else**
12: $\lambda \leftarrow q \mapsto \begin{cases} 1 - \frac{\text{rank}(q)}{\text{max_iter}} & \text{if } q \in Q_{\text{sim}} \\ 0 & \text{else} \end{cases}$
13: **return** λ
14: \triangleright where $\text{rank}(q)$ is such that $Q_{\text{sim}}[\text{rank}(q)] = q$
15: **end if**

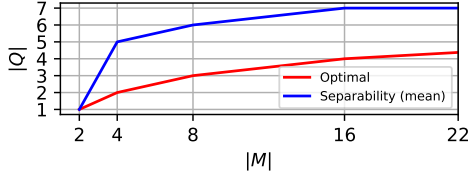


Fig. 3: Scalability: number of questions $|Q|$ needed to distinguish 99% of the pairs of a set of $|M|$ models, for the Separability heuristic and the optimal case.

6. EXPERIMENTAL RESULTS

Figure ?? (a) shows how the random baseline performs. Figure ?? (b) shows that selecting questions with the highest separability at random produces significantly better results than a purely random selection. It also shows that the \mathcal{A}_{sim} heuristic performs better, although with modest improvement, suggesting that our set of questions with maximum separability (in our experiments, we got 4,000 such questions) is quite diverse in the partitions the questions yield. While achieving 95% accuracy requires on average 6 questions using random selection, our heuristics only require 3. Fig ?? (c) represents the best set of questions obtained by each heuristic against our specific set of models.

7. CONCLUSION

Our framework shows that properly selecting from binary questions shows promising results on the task of distinguishing LLMs. This encourages further investigation; in particular a generalization to wider sets of models. Also, focusing on differentiating models that are by construction close to each other would constitute an interesting angle, e.g. models with different training parameters or training data. Furthermore, investigating the robustness of our approach with non-deterministic models is futurework.

A. DATASETS AND MODELS

Dataset	Number of choice	Experiment Subset Size	Original Size
hellaswag	4	20k	40k
MedMCQA	4	20k	182k
mmlu	4	20k	115k
piqa	2	16k	16k

Table 1: Question Datasets for Experiments (details in [?] [?] [?]). Note that their cropping was not performed randomly, which may result in a biased representation of the datasets. We binarized the datasets that were not, by mapping to 0 a wrong answer and 1 a correct one.

B. PROOFS

Theorem B.1. *Let M be a finite set such that $|M| = L$. A question that maximally differentiates pairs in M is one that*

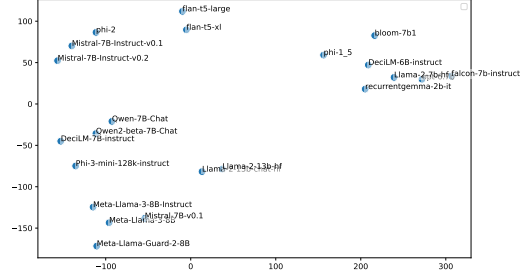


Fig. 4: Proximity (t-sne on response vectors) of the 22 LLMs. Those of same *family* i.e. from same company and differing by version appear globally close, except the DeciLM models.

splits the set into the most equal groups.

Proof. Let q_k be the query that splits M into two groups of size $2 \leq k \leq L - 1$ and $L - k$. We have:

$$\begin{aligned} |\{\text{pairs split by } q_k\}| &= |\{\text{all pairs}\}| - (|P_1| + |P_2|) \\ &= \binom{L}{2} - \left(\binom{k}{2} + \binom{L-k}{2} \right) \end{aligned}$$

where P_1, P_2 are the pairs in the first and second groups. Next, we aim to maximize the number of split pairs, which are the pairs that are differentiated. Therefore, we seek for the maximizer of $k \mapsto |\{\text{pairs split by } q_k\}|$ that is a second order polynomial with maximum reached in $\lfloor \frac{L}{2} \rfloor$. \square

Lemma B.2. *For a finite set of models M s.t. $|M| = 2^n$. The law of X is given by:*

$$\mathbb{P}(X \leq k) = 1 - \frac{2^{n-k} - 1}{2^n - 1}, \quad 1 \leq k \leq n. \quad (5)$$

Proof. At first step, i.e. $k = 1$, there are 2^n models, the first query $\mathcal{Q}^*(1)$ will divide in two groups of size 2^{n-1} . In each of these groups, there are $\binom{2^{n-1}}{2}$ possible pairs, whereas there are $\binom{2^n}{2}$ different pairs before querying. Therefore, we deduce that $\mathbb{P}(X \leq 1) = \frac{|\{\text{pairs that the query splits}\}|}{|\{\text{all possible pairs of } M\}|} = \frac{\binom{2^n}{2} - 2 \cdot \binom{2^{n-1}}{2}}{\binom{2^n}{2}}$. At step $k \in \mathbb{N}^*$, there are 2^{k-1} groups of $2^{n-(k-1)}$ models, query $\mathcal{Q}^*(k)$ will divide each of these groups by two. In each of these 2^k groups, there are $\binom{2^{n-k}}{2}$ possible pairs. With similar thinking as first step, we deduce that: $\mathbb{P}(X \leq k) = \frac{|\{\text{pairs that the queries split}\}|}{|\{\text{all possible pairs of } M\}|} = \frac{\binom{2^n}{2} - 2^k \cdot \binom{2^{n-k}}{2}}{\binom{2^n}{2}} = 1 - 2^k \cdot \frac{2^{n-k} \cdot (2^{n-k} - 1)}{2^n (2^n - 1)} = 1 - \frac{2^{n-k} - 1}{2^n - 1}$. \square

Theorem B.3. *For a infinite and countable set of models M . The law of X is given by:*

$$P(X \leq k) = 1 - \left(\frac{1}{2} \right)^k, \quad \forall k \in \mathbb{N}^*. \quad (6)$$

Proof. The result is obtained by taking the limit of n towards infinity of $\mathbb{P}(X \leq k)$ of Lemma ???. \square

C. REFERENCES

- [1] Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou, “Distill-and-compare: Auditing black-box models using transparent model distillation,” in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 303–310.
- [2] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi, “A survey of methods for explaining black box models,” *ACM computing surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.
- [3] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani, “Aequitas: A bias and fairness audit toolkit,” *arXiv preprint arXiv:1811.05577*, 2018.
- [4] Seong Joon Oh, Bernt Schiele, and Mario Fritz, “Towards reverse-engineering black-box neural networks,” *Explainable AI: interpreting, explaining and visualizing deep learning*, pp. 121–144, 2019.
- [5] Erwan Le Merrer, Patrick Perez, and Gilles Trédan, “Adversarial frontier stitching for remote neural network watermarking,” *Neural Computing and Applications*, vol. 32, no. 13, pp. 9233–9244, 2020.
- [6] Daryna Oliynyk, Rudolf Mayer, and Andreas Rauber, “I know what you trained last summer: A survey on stealing machine learning models and defences,” *ACM Comput. Surv.*, vol. 55, no. 14s, jul 2023.
- [7] Thibault Maho, Teddy Furon, and Erwan Le Merrer, “Model fingerprinting with benign inputs,” in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.
- [8] Aounon Kumar, Chirag Agarwal, Suraj Srinivas, Soheil Feizi, and Hima Lakkaraju, “Certifying llm safety against adversarial prompting,” *arXiv preprint arXiv:2309.02705*, 2023.
- [9] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi, “Hellswag: Can a machine really finish your sentence?,” 2019.
- [10] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi, “Piqa: Reasoning about physical commonsense in natural language,” 2019.
- [11] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt, “Measuring massive multitask language understanding,” 2021.
- [12] Ankit Pal, Logesh Kumar Umaphathi, and Malaikannan Sankarasubbu, “Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering,” 2022.