Comparative Analysis of Human and Large Language Model Performance in Pharmacology Multiple-Choice Questions

Ricardo Rodriguez[1], Stéphane Huet[1], Benoît Favre[2], Mickael Rouvier[1]

[1]LIA - Avignon Université, [2]LIS - Université d'Aix Marseille

Correspondence: firstname.lastname@{univ-avignon.fr, lis-lab.fr}

Abstract

In this article, we study the answers generated by a selection of Large Language Models to a set of Multiple Choice Questions in Pharmacology, and compare them to the answers provided by students, to understand which questions in this clinical domain are difficult for the models when compared to humans and why. We extract the internal logits to infer probability distributions and analyse the main features that determine the difficulty of questions using statistical methods. We also provide an extension to the FrenchMedMCQA dataset, with pairs of question-answers in pharmacology, enriched with student response rate, answer scoring, clinical topics, and annotations on question structure and semantics.

1 Introduction

Large Language Models (LLM), such as ChatGPT (OpenAI et al., 2024) or Llama (Meta, 2024), have led to significant advances in language generation. These breakthroughs have had a transformative effect on numerous Natural Language Processing (NLP) tasks like question answering, text summarisation, and machine translation.

In order to evaluate the capabilities of LLMs, significant efforts have been dedicated to measuring their performance across various benchmarks. The evaluations have evolved from testing specific skills (Liu et al., 2023; Zhang et al., 2024) to assessing them on complex, expert-level tasks such as medical exams (Pal et al., 2022; Labrak et al., 2022) or law exams (Guha et al., 2023).

Performance scores resulting from benchmarking, such as accuracy, remain a common method for assessing the capabilities of LLMs. However, they provide a rather surface-level understanding, primarily indicating whether a model can complete a given task like question answering without offering deeper insights into the underlying causes of failures or the subtleties of performance degradation (Liang et al., 2022; Ribeiro et al., 2020). The difficulties LLMs face in answering questions vary depending on several factors. On one hand, some questions are inherently more complex because they require an advanced understanding of context, cross-references, or multi-step reasoning. On the other, the frequency of concepts in training datasets plays a key role: LLMs struggle more with processing rare or highly specific answers.

In this paper, we propose to explore why these models fail in certain situations and whether they encounter the same obstacles as humans. We consider answering these questions essential for refining LLM design and enhancing their robustness. To evaluate our approach, we chose the FrenchMedMCQA dataset (Labrak et al., 2022) (multi-choice question-answering task), that contains publicly available question-answer pairs in pharmacology, and for which we could obtain student answer rates enabling us to do a comparison of systems against humans.

The novel contributions in our work are the following:

- We propose an original comparative analysis between human responses and LLM output on pharmacology multiple-choice questions (FrenchMedMCQA dataset).
- We enrich the FrenchMedMCQA dataset with additional annotations including student response rates, manually labelled tags about syntactic features and question structure (e.g., negation, question mode), and clinical topics. This new dataset is freely available online [1].
- We provide an in-depth analysis that identifies the most influential features that deter-

[1]Dataset accessible on HuggingFace: https://huggingface.co/datasets/uy-rrodriguez/FrenchMedMCQA-extended

Proceedings of Recent Advances in Natural Language Processing,pages 1022–1029
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-117

mine question difficulty for both humans and LLMs.

The paper is organised as follows. In Section 2, we describe the FrenchMedMCQA dataset along with the additional data we have incorporated. In Section 3, we present the experimental framework used in our study. We detail the comparative results between human and LLM performance across different difficulty levels in Section 4. Finally, we conclude in Section 5.

2 The FrenchMedMCQA dataset

In this section, we describe the FrenchMedMCQA dataset and introduce the supplementary data points that we collected to enrich its contents and support our analysis.

2.1 Description

We propose to use the FrenchMedMCQA (Labrak et al., 2022) dataset that contains around 3,000 multiple choice questions (MCQs) in the pharmacology domain, taken from exams of the French Diploma of Pharmacy Specialisation and downloaded from the website MedShake.net [2]. This corpus is similar to those found in other languages, such as the MedMCQA (Pal et al., 2022) and SciQ (Welbl et al., 2017) corpora.

The dataset was chosen for its high quality for question-answering in the French language and medical domain, and because the MedShake platform also provides student answer rates, which enables our comparative analysis of system behaviour against humans.

Each sample contains five possible answers and their associated choice letter a-e, among which one or more choices determine the correct answer.

2.2 Data annotation

The MedShake platform enables students to practice exams online and evaluate their knowledge, obtaining a final score based on the real scoring scale used during the exam. The platform also provides the correct and partially correct answers across possible choices, the number of points allocated to each combination of choices, the number of students having answered each combination, the question's clinical topic(s) [3], and the year the exam took place.

The availability of student responses differs from one question to another, as it depends on the specific questions that users answered. Besides, it is impossible to pinpoint the responses of an individual. There are more than 2.4 million answers, with an average of 664 student answers per question.

For the present work, we have enriched the FrenchMedMCQA dataset with the student responses, clinical topics, and years. Additionally, we manually tagged each question with a handful of characteristics based on its semantic content and the way it is formulated. This new dataset is made freely available online under the name "FrenchMedMCQA-extended" [1]. These features were added to enrich the statistical analysis to follow; notably, they help us identify particular characteristics in our corpus that impact a question's difficulty. A selection of statistics from the resulting corpus is available in Table 1.

The manual annotations describe the following aspects:

Negation: Indicates if a negated phrase is present anywhere in the question.

Composition required: Indicates when the question is a partial sentence and needs to be combined with one of the choices to form a full correct sentence. *E.g.: "'Crack' is a form of:". Choices: (a) heroin; (b) cocaine.*

Identification of intruder: Indicates whether the question requires the student to identify the choice(s) that do not respect a certain condition. *E.g.: "Which proposition does not apply to norfloxacin?".*

Sentence mode: Categorises the "question" as a true question, an instruction, or an affirmation. *E.g.: Instruction: "Concerning misoprostol, give its action mechanism."; Affirmation: "Anaemia is generally observed under the following parasitic infections:".*

Explicit number of choices: Indicates whether the number of expected answers is explicitly provided (single, multiple, or undefined). *E.g.: Single: "Only one proposition is exact. Serotonin is:";*

[2] MCQ exams of the French Diploma of Pharmacy Specialisation ("Annales QCM des concours d'internat en pharmacie"): https://www.medshake.net/pharmacie/concours-internat/annales/qcm/. Last accessed: 2025-03-01.

[3] Clinical topics: pharmacology, physiology, bacteriology, analytical chemistry, toxicology, haematology, clinical biochemistry, immunology, public health, virology, parasitology, biophysics, epidemiology, galenic, mycology, pharmacokinetics, genetics, statistics, enzymology.

A.

	Num. Samples	Avg. Answers	Num. Expected Choices				
			1	2	3	4	5
Train	2170	740	27%	24%	33%	14%	2%
Dev	312	601	52%	14%	23%	10%	1%
Test	622	650	52%	15%	23%	9%	1%

B.

	Negation		Composition		Intruder		Sentence Mode			Num. Choices		
	no	yes	no	yes	no	yes	Q	I	A	S	M	U
Overall	94%	6%	69%	31%	82%	18%	72%	18%	10%	53%	30%	17%

Table 1: A) Number of samples and average of student answers per sample in the FrenchMedMCQA corpus. Followed by distribution of samples per number of correct choices (each question expects that 1-5 choices are selected for a correct answer). B) Distribution of manual annotations added to the corpus. Clarification of column names: **Sentence mode:** Q=Question; I=Instruction; A=Affirmation. **Num. choices:** (Explicit number of choices) S=Single; M=Multiple; U=Undefined.

Multiple: "Which ones of the following propositions apply to IL-2?"; Undefined: "What happens during ventricular systole?".

3 Experimental settings

The LLMs employed in this study, along with their fine-tuning and inference methods, are described in Section 3.1. In Section 3.2, we describe the metrics used to assess these models, and in Section 3.3 we present a method to assess question difficulty.

3.1 Model selection and training approach

In our experiments, we select a series of Large Language Models based on their results on the shared task TALN-DEFT 2023 (Labrak et al., 2023), to have a reference baseline. They represent a mix of general-purpose models: Llama-3-8B and 70B (Touvron et al., 2023), and Mistral-7B (Jiang et al., 2023); and others specialised for the medical domain: BioMistral-7B (Labrak et al., 2024), and Apollo-7B (Wang et al., 2024).

The models are loaded in 4-bit precision and then fine-tuned on the FrenchMedMCQA training dataset for 1 epoch using Low Rank Adaptation (LoRA) (Hu et al., 2022) to fit our infrastructure.

The response template used at this stage is the same as the prompt for inference, based on our previous works to obtain comparable results: a short description of the task asking the model to answer to a question from a pharmacology exam, then the question and the choices, and finally a simple format to introduce the correct answer *("Response(s):")* immediately followed by one of two formats of expected response, one including the full text of the choices *(e.g., "(a) text a; (b) text b")*, and the other only providing the choice letters *(e.g., "(a) (b)")*. It's worth noting that our goal is not to optimise the prompt and the model's response to the given task, but to analyse its behaviour.

Another variation evaluated was having simple and structured task descriptions. The simple prompt uses natural language and line breaks to separate sections of the prompt, while the more structured prompt uses special handles: *"### Instruction:", "### Input:" (i.e., question and choices), and "### Response:"*.

For loss calculation, early experiments showed that only considering the answer text after "Response(s):" gave the best results.

Following fine-tuning, we evaluate their performance on the FrenchMedMCQA test corpus to assess both its generalisation capabilities and overall effectiveness.

For each LLM, we run the inference 4 times and average the scores to obtain the best per model. These results can be seen in Table 2.

3.2 Metrics

Unlike traditional classification tasks, MCQs can have partially correct responses. For example, if a question requires selecting two correct options but only one is identified, the answer is incomplete. Two metrics initially proposed in (Labrak et al., 2023), EMR and Hamming score, and an original, MedShake score, have been employed to measure the proportion of correct answers while penalising incorrect ones.

In the formulas below, N is the number of questions, y_i is the set of correct answers for the i-th question, and $\hat{y}_i$ is the set of predicted answers for

	Medical LLM	MedShake	EMR	Hamming	Prompt
Student responses		0.594	0.517	0.677	
LLaMa-3-8B	-	0.366	0.295	0.522	1
LLaMa-3-70B	-	0.189	0.138	0.381	1
Mistral-7B-v0.3	-	0.391	0.318	0.539	1
BioMistral-7B	✓	0.289	0.224	0.475	2
Apollo-7B	✓	0.413	0.333	0.557	2
LLaMa-3-8B*	-	0.453	0.373	0.575	1
LLaMa-3-70B*	-	0.419	0.345	0.554	1
Mistral-7B-v0.3*	-	**0.491**	**0.418**	**0.626**	1
BioMistral-7B*	✓	0.404	0.326	0.551	2
Apollo-7B*	✓	0.417	0.339	0.575	2

Table 2: Summary of the best results per model showing the average rates in all difficulty classes combined. Fine-tuned models are marked with *. Specialised models are identified with a check mark. Prompt "1" corresponds to the simple natural language prompt while "2" corresponds to the more structured format, as described in Section 3.1

the i-th question.

Exact Match Ratio (EMR): checks if the set of predicted answers exactly matches the correct answers for each question.

$$\text{EMR} = \frac{1}{N}\sum_{i=1}^{N}[y_i = \hat{y}_i]$$

Hamming Score: inspired by the Hamming distance, measures the overlap between predicted and correct answers by comparing the size of their intersection with their union.

$$\text{Hamming} = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i \cap \hat{y}_i|}{|y_i \cup \hat{y}_i|}$$

MedShake Score: is the original point-based metric used in the exams, where scores are assigned based on how many answers match the key, awarding full points for complete correctness and reducing points for errors or omissions (max=2.0, one error=1.0, two errors=0.4, 0 otherwise).

3.3 Question classification by difficulty

A naive way of assessing question difficulty would be to use the percentage of students who selected the correct set of options. However, this does not take into account partial answers and the number of correct choices, which affect the chances of answering the question correctly at random. In order to consider these factors, we measure question difficulty with the Shannon entropy (Shannon, 1948):

$$H(P) = \frac{-\sum_{i=1}^{n} p_i log(p_i)}{log(n)}$$

where n is the number of possible answers and p_i is the proportion of students who chose answer i. Thus, a value of 0 means that the question is obvious for all students (everyone gives the same answer), and a value of 1 means the question is extremely hard (the response rate is equivalent to a random selection).

Based on this approach, we choose to classify the corpus in five categories, each with the same number of items: Very Easy, Easy, Medium, Hard, and Very Hard. H is computed for each sample and the corpus is then divided into five equal-sized buckets corresponding to our difficulty classes.

4 Results

In Section 4.1 we describe the main observations of LLM responses. Then, in Section 4.2 we compare the performance of the best model against humans. Later, in Section 4.3 we introduce the strategy to build a probability distribution out of model logits, which we finally use for the linear regression and feature importance analysis of Section 4.4.

4.1 LLM results and best model

Table 2 shows the results for all models. The LLMs were evaluated according to three metrics: MedShake, EMR (Exact Match Ratio), and Hamming.

We observe that fine-tuned models consistently outperform their non-fine-tuned counterparts across every model configuration. For instance, Mistral-7B-v0.3 achieves a MedShake score of 0.391 without fine-tuning versus 0.491 after fine-tuning. This model actually attains the highest overall performance across all metrics (MedShake: 0.491; EMR: 0.418; Hamming: 0.626), thereby

Model$_{Metric}$	**Very easy**	**Easy**	**Medium**	**Hard**	**Very hard**	**Overall**
Mistral-7B$_{MedShake}$	0.780	0.599	0.453	0.289	0.336	0.492
Mistral-7B$_{EMR}$	0.750	0.524	0.355	0.196	0.262	0.418
Mistral-7B$_{Hamming}$	0.814	0.706	0.637	0.506	0.467	0.626
Human$_{MedShake}$	0.891 $^{+11\%}$	0.724 $^{+13\%}$	0.565 $^{+11\%}$	0.438 $^{+15\%}$	0.352 $^{+2\%}$	0.594 $^{+10.3\%}$
Human$_{EMR}$	0.872 $^{+12\%}$	0.642 $^{+12\%}$	0.452 $^{+10\%}$	0.324 $^{+13\%}$	0.295 $^{+3\%}$	0.517 $^{+9.9\%}$
Human$_{Hamming}$	0.903 $^{+9\%}$	0.775 $^{+7\%}$	0.667 $^{+3\%}$	0.579 $^{+7\%}$	0.460 $^{-0.7\%}$	0.677 $^{+5.1\%}$

Table 3: Comparison of performance between the fine-tuned model Mistral-7B-v0.3 and humans across all difficulty levels and evaluation metrics. The reported scores represent the mean over 4 independent runs. Human scores correspond to student response rates from the "test" dataset, accompanied by the performance gain over the model as absolute percentages.

demonstrating its robustness once adapted to the task, and making it our best candidate for the feature importance analysis in Section 4.4.

Among the models specialised in the medical domain, Apollo-7B yields the best results among the non-fine-tuned models (MedShake: 0.413; EMR: 0.333; Hamming: 0.557). Nevertheless, even after fine-tuning, its performance remains inferior to that of Mistral-7B. It is also noteworthy that, despite their biomedical specialisation, the fine-tuned domain-specific models still lag behind the fine-tuned general-purpose models.

4.2 Human vs LLM evaluation

Table 3 provides a detailed comparison of performance between the fine-tuned model Mistral-7B-v0.3 and human respondents, across all five levels of difficulty and the three evaluation metrics. Model scores represent the mean of four independent runs, whereas human performance is derived from the additional data obtained from MedShake.net as described in Section 2.2.

Overall, performance declines consistently with increasing difficulty for both humans and the model. This trend corroborates the validity of the proposed categorisation and confirms that the perceived difficulty by students is broadly mirrored by the model.

Humans outperform Mistral-7B on every metric and across all difficulty levels. For instance, on the "Very Easy" questions, humans achieve a MedShake score of 0.891 versus 0.780 for the model. This gap persists as difficulty increases. For "Hard" questions, EMR scores drop to 0.324 for humans and 0.196 for the model, illustrating the model's limitations when faced with complex queries requiring nuanced understanding or advanced reasoning. Interestingly, for "Very Hard" questions the model results improve compared to the "Hard" ones, and the difference with humans drops to a minimum, with students obtaining an EMR score of just 3.3% above the model (0.295 vs 0.262). Although we haven't explored in detail why specific questions in this class are less difficult for Mistral, the overall results underline their high difficulty, often requiring multiple choices to be answered correctly.

Finally, the Hamming metric, which measures partial overlap between expected and generated answers, reveals a narrower gap between LLM and human performance. This suggests that, even in the absence of fully correct responses, the model is often able to identify some of the relevant elements.

These findings highlight both the current capabilities of LLMs in specialised tasks and indicate that improvements are needed to meet or exceed human performance in domain-specific challenges.

4.3 Model-based probability distribution

To compare human responses, expressed as rates, with LLMs, we decide to derive a comparable output for the model by converting its internal logits into a probability distribution.

We derive the model's learned probabilities by extracting token-level probabilities (via softmax on internal logits) from sequences formatted like the fine-tuning prompt (including task introduction, question, and choices). For a sequence S, the log probability $log\_prob(S)$ is computed as the sum of the logarithms of individual token probabilities. We evaluated various strategies for computing log_prob (using the full sequence, just the answer segment, and only the choice letters) to predict which answer the model would select, and compared their scores to the results of the inference.

Finally, we decided to use the probability of the choice letters only as a good enough measure of the relative probabilities given by the model, and leveraged this result to construct a probability distribution.

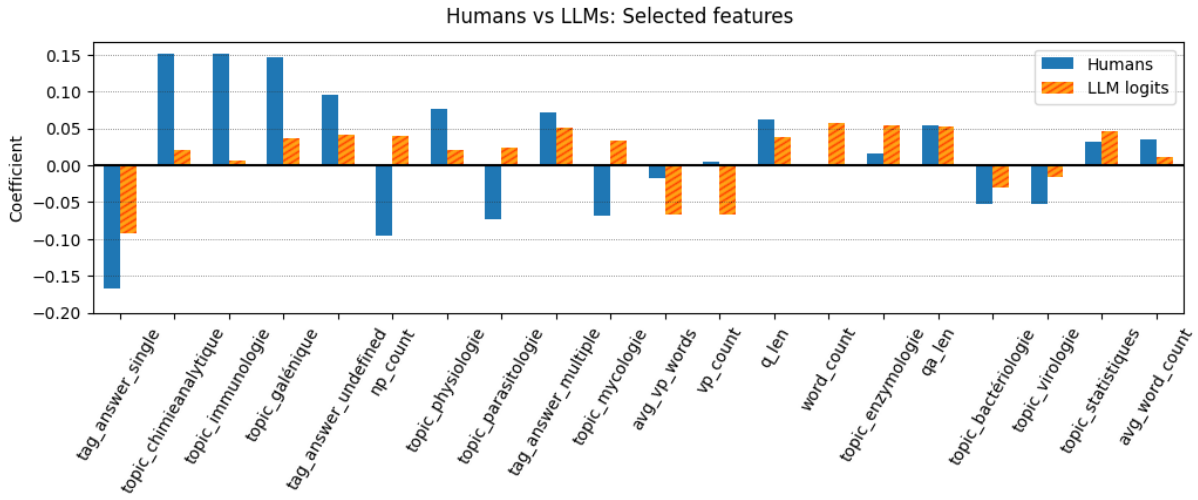

Figure 1: Comparison of the variance in feature coefficients derived from linear regression models for human and LLM scores. Only features with the highest absolute coefficients are displayed.

4.4 Feature importance

To identify the most important features that determine question difficulty for both humans and LLMs, we trained a linear regression model using five-fold cross-validation with Ridge (L2) regularisation, on a merged corpus of "train" and "test" (approx. 1,600 items).

It is worth clarifying that our goal is not to develop a perfect predictor of question difficulty. This means that semantic features like question embeddings, used in other works such as Yaneva et al. (2024), are not included here, keeping all features explicable and comparable with humans. This decision likely reduces the ability of the regression model to predict the difficulty for LLMs, which is an acceptable limitation for our work.

For human data, we trained a linear model to predict the correct response rate from students. The feature set comes from the data points presented in Section 2.2 as well as several syntactic features: question length; sum of answers length; total length; question word count (number of words in the question); avg. sentence word count; avg. depth of tree; number and avg. word count of noun, prepositional, and verb phrases. For the LLM, we followed the same approach with a linear model trained with the probability distribution based on the internal model logits.

Due to the limited size of the data, we selected features based solely on the magnitude of their coefficients instead of relying on p-tests, which proved unstable across runs. Figure 1 shows the comparison of feature coefficients between humans and Mistral-7B. Globally, we observe that most features have a similar relationship with question difficulty, both for humans and models, even though the coefficients differ. Notably, the annotations *tag_answer_single* and *tag_answer_undefined*, describing whether the question explicitly indicates the number of correct choices, have the same relationship with the difficulty perceived by humans and the LLM. On the other hand, some topics such as *immunology* and *analytical chemistry* are better predictors of difficulty for humans, while syntactic features like *avg. words in verb phrases* are better predictors for LLMs. This analysis suggests that humans and LLMs might generally assess question difficulty in an equivalent way, since most features relate to difficulty similarly.

5 Conclusion

This paper proposes a comparative analysis of MCQ answering behaviour for humans and LLMs in a medical corpus on pharmacology, which suggests that the factors contributing to question difficulty are similar for both categories. Both tend to struggle with questions which are considered to have a high level of difficulty. The feature importance analysis highlights that most features have a similar relationship with question difficulty.

Our experiments are conducted with a handful of LLMs and only on the French corpus FrenchMedMCQA, potentially excluding other corpora and thereby limiting the generalisability of results. Complementary analysis should be conducted to evaluate these results in other settings.

Acknowledgments

We thank Pierre-Michel Bousquet for his help and wise advice during this work.

This work benefitted from access to the supercomputers provided by the CNRS IDRIS via the resources allocation number 2024-A0161014871, and financial support from the project ANR MALADES (ANR-23-IAS1-0005, https://anr-malades.github.io/).

References

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, and 1 others. 2023. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36:44123–44279.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. LoRA: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *Preprint*, arXiv:2310.06825.

Yanis Labrak, Adrien Bazoge, Béatrice Daille, Richard Dufour, Emmanuel Morin, and Mickael Rouvier. 2023. Tâches et systèmes de détection automatique des réponses correctes dans des QCMs liés au domaine médical: Présentation de la campagne DEFT 2023. In *18e Conférence en Recherche d'Information et Applications–16e Rencontres Jeunes Chercheurs en RI–30e Conférence sur le Traitement Automatique des Langues Naturelles–25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, pages 57–67. ATALA.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Beatrice Daille, Pierre-Antoine Gourraud, Emmanuel Morin, and Mickael Rouvier. 2022. FrenchMedMCQA: A french multiple-choice question answering dataset for medical domain. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 41–46, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864, Bangkok, Thailand. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, and 1 others. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Yixin Liu, Alexander R Fabbri, Jiawen Chen, Yilun Zhao, Simeng Han, Shafiq Joty, Pengfei Liu, Dragomir Radev, Chien-Sheng Wu, and Arman Cohan. 2023. Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization. *arXiv preprint arXiv:2311.09184*.

AI Meta. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date. *Meta AI*, 2(5):6.

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, and 262 others. 2024. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with checklist. *arXiv preprint arXiv:2005.04118*.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Xidong Wang, Nuo Chen, Junyin Chen, Yan Hu, Yidong Wang, Xiangbo Wu, Anningzhe Gao, Xiang Wan, Haizhou Li, and Benyou Wang. 2024. Apollo: Lightweight multilingual medical LLMs towards democratizing medical AI to 6B people. *Preprint*, arXiv:2403.03640.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

Victoria Yaneva, Kai North, Peter Baldwin, Saed Rezayi, Yiyun Zhou, Sagnik Ray Choudhury, Polina Harik, Brian Clauser, and 1 others. 2024. Findings from the first shared task on automated prediction of difficulty and response time for multiple-choice questions. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 470–482.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.