Explaining LLM-based Question Answering via the self-interpretations of a model*

Darío Garigliotti[1 [0000−0002−0331−000X]]

University of Bergen, Norway - dario.garigliotti@uib.no

Abstract. As Large Language Models (LLMs) become increasingly ubiquitous in data-driven methods for multiple information processing tasks, so is also more significant the need to provide explainability mechanisms for these methods. In this work, we tackle a paradigmatic instance of the family of Question Answering problems by the means of a general approach based on Retrieval-augmented Generation (RAG). We focus not only on the performance for different parameter configurations but, in particular, on augmentation strategies that inquire the very generator LLM about its own interpretations behind the answer that it provides for a question.

Keywords: Interpretability · Question Answering · LLMs.

1 Introduction

In recent years, the development of —and accompanying body research on— language models (LMs) has taken a significant step forward with the appearance of so-called Large Language Models (LLMs). These LLMs are trained with state-of-the-art technologies in elements such as the learner architecture —distinguishedly the transformer— and the training regime —including multi-tasking, fine-tuning, and Reinforcement Learning with Human Feedback—, autoregressively over vast amounts of information typically crawled from the Web (Touvron et al., 2023; OpenAI, 2024). With a seemingly always-increasing hype for the applicability of LLMs, which has already shown state-of-the-art performance in several tasks (Radford et al., 2019; Si et al., 2023), come also their studied drawbacks (Dodge et al., 2021; Sainz et al., 2023). Beyond the issues with feasibility for making the construction of these vast models reproducible outside of very few dedicated environments, and the implications of commercial-only availability of closed LLMs (Jacovi et al., 2023), there is also a series of interests for understanding its intricacies and challenges for its expected usability (Elazar et al., 2024; Anwar et al., 2024). An intertwined kind of phenomena are the hallucinations that characterize most of the well-known LLMs. These are particularly crucial in applications where there is a need for ensuring the truthfulness of the textual content generated by an LLM (Liu et al., 2023; Menick et al., 2022). A paradigmatic task with these needs is Question Answering (QA), when

* Part of NRF project 329745 Machine Teaching For XAI.

it is instantiated in a way that also requires from an answering method to provide evidence that supports the obtained answer (Bohnet et al., 2022). This work addresses Self-supported Question Answering (SQA) (Menick et al., 2022), a problem where the answer to an input question must be complemented with one or more pointers to textual excerpts from a given collection as supporting references. SQA is related to several other similar tasks (Asai et al., 2022; Liu et al., 2023).

Arguably, SQA contributes to model interpretability, as just like interpretability, evidence helps increase trust in the outputs of a model (Menick et al., 2022). Moreover, by providing references with its responses, the model is implictely attempting to explain the rationale behind the answers. We approach SQA via Retrieval-augmented Generation (RAG) (Lewis et al., 2020), a general framework that suits well the scenario where the parametric knowledge of an LLM should be complemented with explicit knowledge. RAG allows for this by integrating selected contexts by retrieval to the text generator at prompting, with which achieves state of the art in evidence-aware QA (Gao et al., 2023; Garigliotti, 2024). Our experiments test methods within the RAG umbrella by setting relevant parameters. Beyond implicit explainability, we focus in particular on explicit mechanisms to inquire the interpretation of the model's rationales.

In the rest of the paper, we describe the dataset and approach we use in our experimental setup, and then address our research questions by analyzing the experimental results.

2 Approach

2.1 Methodology

Following a very recent benchmark in the literature Gao et al. (2023), we address SQA via a series of methods all within the same general RAG paradigm (Lewis et al., 2020). According to a particular configuration set for each of the parameters of interest, the configuration corresponds to a specific '(SQA) method' as we refer to these. Each RAG-based SQA method is made of the same three distinguished components. Firstly, retrieval obtains relevant contexts or passages for a question, from a given collection. After that, the passages are integrated into a well-engineered prompt that also contains the question and, possibly, examples for few-shot prediction. Finally, the generation stage takes the prompt as input for an LLM to generate the desired answer with evidence.
We carry out the designed experiments aiming to answer these research questions:

- **RQ1**: How do the explicit interpretability mechanisms impact the performance of the LLM-powered RAG approach for SQA?
- **RQ2**: What is the relation between self-interpretation inquiry and the zero- or few-shot strategies augmented in prompt?
- **RQ3**: How does the awareness about need for interpretability prompted alongside the question affect the performance on SQA when the passages are provided in non-standard orders in the prompt?

2.2 Experimental Setup

Dataset. QAMPARI (Amouyal et al., 2023) is a publicly available QA dataset based on Wikipedia as corpus. Each question in QAMPARI requires as answer a list of entities that occur in passages to the question. Also following the benchmark (Gao et al., 2023), we randomly select 60 questions from QAMPARI, and refer to these as our QAMPARI instances. For an instance to be selected, it must have at least one of its possibly multiple correct answers occurring in the top 3 ranked passages obtained with dense retrieval in the benchmark.

RAG components. In **retrieval** phase, we index a collection of passages from the benchmark associated to all the selected 60 QAMPARI instances. Then, we retrieve the top 10 passages for each question with a dense retrieval method, and obtain the subsequences of top 3 and top 5 results to also experiment with. During **augmentation**, we instantiate a general prompt template with the actual question and retrieved contexts, as well as one or more possible examples —each made of a question, contexts and the correct answer with reference(s)— if not in zero-shot mode. The basic prompt template —referred to as XAI-agnostic since interpretability is only implicit in the request for evidence— is presented in Table 1 (Garigliotti et al., 2024). In order to obtain an XAI-aware prompt, we first enable the [XAI instr.] part by replacing it with the further instruction "You are also asked about why you are giving this answer to the question. Please respond to it right after." We then obtain two variants of a XAI-aware prompt, a *direct* one —where [XAI Q.] becomes "Why do you think that this is the answer to the question?"— and a *counterfactual* one —by replacing [XAI Q.] with "What would you have answered to the same question if the order of the passages in the prompt was different?"—. In the final stage, **generation**, we input the prompt into the GPT-3.5 (gpt-3.5-turbo-0125) LLM (Radford et al., 2019).
This list summarizes our experimental parameters:

- Retrieval: **cut-off** —top 3, 5, or 10 passages—.
- Augmentation: **order** of the passages in the prompt —as retrieved in ranking, top ranked result goes last, or random—; **number** of few-shot examples —0, 1, or 2—; and **XAI** prompting —basic, direct or counterfactual—.

Evaluation metrics. We evaluate answer correctness by verifying whether any of the collected possible expressions of the correct answer is an exact sub-string of the generation —answer recall or exact match recall, following (Stelmakh et al., 2022)—. Answer support is evaluated by applying standard retrieval metrics of precision and recall with respect to the retrieved and relevant passage sets. For a given method, we report the average performance across all the questions.

3 Experimental Results

Tables 2, 3 and 4 present all our experimental results. Each table corresponds to one of the XAI-oriented prompting strategies: basic or implicit, direct, and counterfactual, resp.

Table 1: Template to build the basic prompt during augmentation. The templates for the XAI-aware prompts are almost identical except for the enabled XAI components **[XAI instr.]** and **[XAI Q.]** omitted in the basic prompt.

Prompt template	Prompt template (ctd.)
You are an assistant for question-answering tasks. Use the pieces of context provided by the user to ANSWER the QUESTION to the best of your ability. If you don't know the answer, just say that you don't know. Keep the answer concise. Always cite one or more corresponding context IDs as your sources (which must be among the given CONTEXTS) between square brackets (e.g. [a1b2x34d]), as it's done in each example. Examples are given below, each example between the '⟨example⟩' and '⟨/example⟩' tags. After that, you are given the actual question with contexts so that you answer it. **[XAI instr.]**	⟨example⟩ ... ⟨/example⟩ ... QUESTION: ... **[XAI Q.]** CONTEXTS: Context ID: ... Context: ... ... ANSWER:

RQ1: In general, we observe a clear increase in the performance for several methods in Table 2 when compared with its respective counterparts in Tables 3 and 4, especially in the few-shot scenarios. As a qualitative example, for the question "Which FA Cup Final did Manchester United win?" a basic method that prompts with the top of its 5 retrieved passages at the bottom, while doing 1-example shot, answers "1990", while its XAI-direct counterpart correctly says "1990 FA Cup Final."

RQ2: The results are mixed. Some increments in the absolute performances for the best measurements are observed across Tables 3 and 4.

RQ3: The awareness of being inquired about explaining its own mechanisms at promting —i.e. question— time seems to influence variations in the best performing methods, in terms of their characterization by the order of the passages in their prompt. In particular, our last XAI-explicit prompt, counterfactual, challenges an alternative scenario not necessarily about which the question or answer was or could be, but rather about the augmentation strategy itself.

4 Conclusion and Future Work

We have studied strategies of self-interpretation for an LLM within the general RAG framework for a series of configured methods, as a mechanism to make operational an explicit explainability of the rationales behind answering questions with evidence. In future work, we plan to study other possible strategies for interpretability in SQA, such as example-based XAI. Another aspect to work further in is the evaluation of these observed strategies. A third line of future investigation deals with extending the space of choices for selected parameters, such as the actual LLM used in a framework like RAG and experimenting with more advanced RAG-based approaches. Finally, an additional aspect to analyze the experimental results in terms of question types from the dataset here used.

Table 2: Experimental results over the QAMPARI instances, for the basic prompt (i.e. without XAI component). In all these experiments, retrieval method is dense, and the generator LLM is GPT-3.5. In each block of this table, the best performance on a metric is shown in **bold**.

Number of few-shot examples in prompt: zero					
Retrieval cutoff	Passage order	Answer Recall	Citation Precision	Citation Recall	Citation F-score
3	As in ranking	**0.5717**	0.7861	0.6833	0.71
	Top result last	0.555	**0.8167**	**0.6972**	**0.7306**
	Random	0.5083	0.7528	0.6417	0.6711
5	As in ranking	0.4731	0.7542	0.5261	0.5875
	Top result last	0.4352	0.7083	0.4942	0.5506
	Random	0.4713	0.75	0.5428	0.5932
10	As in ranking	0.3887	0.775	0.4167	0.4946
	Top result last	0.3396	0.7189	0.3764	0.4521
	Random	0.3679	0.7208	0.385	0.458

Number of few-shot examples in prompt: one					
Retrieval cutoff	Passage order	Answer Recall	Citation Precision	Citation Recall	Citation F-score
3	As in ranking	**0.4917**	0.7194	0.6389	0.6517
	Top result last	0.45	**0.7444**	**0.65**	**0.6739**
	Random	0.4517	0.6611	0.5917	0.6
5	As in ranking	0.4196	0.7117	0.4956	0.5461
	Top result last	0.3088	0.6208	0.4114	0.4659
	Random	0.3852	0.7417	0.4872	0.5522
10	As in ranking	0.3565	0.7353	0.43	0.4842
	Top result last	0.2853	0.6642	0.3378	0.4052
	Random	0.3114	0.7356	0.3959	0.4587

Number of few-shot examples in prompt: two					
Retrieval cutoff	Passage order	Answer Recall	Citation Precision	Citation Recall	Citation F-score
3	As in ranking	0.5067	0.6417	0.6111	0.5961
	Top result last	0.5372	**0.7736**	0.7	0.7026
	Random	**0.5408**	0.7389	**0.7361**	**0.7206**
5	As in ranking	0.4596	0.7056	0.5425	0.5882
	Top result last	0.4254	0.7167	0.5464	0.5912
	Random	0.4018	0.6403	0.4956	0.5186
10	As in ranking	0.4085	0.7303	0.4706	0.5193
	Top result last	0.3601	0.6611	0.3717	0.441
	Random	0.3393	0.6315	0.3795	0.4276

Table 3: Experimental results over the QAMPARI instances, for the direct XAI-aware prompt. In all these experiments, retrieval method is dense, and the generator LLM is GPT-3.5. In each block of this table, the best performance on a metric is shown in **bold**.

Number of few-shot examples in prompt: zero					
Retrieval cutoff	Passage order	Answer Recall	Citation Precision	Citation Recall	Citation F-score
3	As in ranking	**0.5528**	0.7444	0.6417	0.6683
	Top result last	0.5106	**0.7583**	**0.65**	**0.6772**
	Random	0.5306	**0.7583**	0.6333	0.6661
5	As in ranking	0.4717	0.7242	0.5108	0.5586
	Top result last	0.496	0.7111	0.4747	0.5406
	Random	0.4815	0.7417	0.5183	0.5756
10	As in ranking	0.4332	**0.7583**	0.4094	0.4931
	Top result last	0.3915	0.7514	0.4037	0.4799
	Random	0.4037	0.7417	0.3932	0.4709

Number of few-shot examples in prompt: one					
Retrieval cutoff	Passage order	Answer Recall	Citation Precision	Citation Recall	Citation F-score
3	As in ranking	**0.57**	0.7361	0.6306	0.6578
	Top result last	0.4967	0.7278	0.6194	0.6417
	Random	0.555	0.7483	**0.6472**	**0.6706**
5	As in ranking	0.4439	0.7292	0.5011	0.5657
	Top result last	0.4171	0.6861	0.4858	0.5384
	Random	0.449	0.745	0.54	0.5811
10	As in ranking	0.3737	0.7117	0.3912	0.4558
	Top result last	0.3472	**0.7611**	0.3781	0.4554
	Random	0.392	0.6808	0.3774	0.4402

Number of few-shot examples in prompt: two					
Retrieval cutoff	Passage order	Answer Recall	Citation Precision	Citation Recall	Citation F-score
3	As in ranking	0.5439	0.7083	0.6472	0.655
	Top result last	0.58	0.7472	0.675	0.6911
	Random	**0.6106**	**0.8056**	**0.7306**	**0.7417**
5	As in ranking	0.4833	0.735	0.5497	0.5969
	Top result last	0.4699	0.7556	0.5344	0.5933
	Random	0.4421	0.7389	0.5414	0.5943
10	As in ranking	0.4195	0.7694	0.467	0.5334
	Top result last	0.3812	0.7292	0.4104	0.4779
	Random	0.3854	0.7322	0.4139	0.4808

Table 4: Experimental results over the QAMPARI instances, for the counterfactual XAI-aware prompt. In all these experiments, retrieval method is dense, and the generator LLM is GPT-3.5. In each block of this table, the best performance on a metric is shown in **bold**.

Number of few-shot examples in prompt: zero					
Retrieval cutoff	Passage order	Answer Recall	Citation Precision	Citation Recall	Citation F-score
3	As in ranking	0.5467	0.7444	0.6611	0.68
	Top result last	**0.555**	0.75	**0.6667**	**0.6828**
	Random	0.505	0.7583	0.65	0.68
5	As in ranking	0.5022	0.7542	0.5331	0.5892
	Top result last	0.481	0.7194	0.4933	0.5502
	Random	0.4796	0.73	0.5289	0.5759
10	As in ranking	0.4387	**0.7806**	0.4584	0.5292
	Top result last	0.3551	0.6764	0.3509	0.4252
	Random	0.3482	0.7211	0.41	0.4677

Number of few-shot examples in prompt: one					
Retrieval cutoff	Passage order	Answer Recall	Citation Precision	Citation Recall	Citation F-score
3	As in ranking	0.5339	0.6972	0.6972	0.6656
	Top result last	0.555	0.7944	0.7111	0.7217
	Random	**0.6078**	**0.8083**	**0.7639**	**0.7544**
5	As in ranking	0.4369	0.6917	0.5664	0.5856
	Top result last	0.4435	0.7375	0.5581	0.5925
	Random	0.426	0.7611	0.5719	0.6133
10	As in ranking	0.4139	0.7667	0.48	0.5358
	Top result last	0.3492	0.7403	0.4307	0.4935
	Random	0.3821	0.7542	0.4287	0.5023

Number of few-shot examples in prompt: two					
Retrieval cutoff	Passage order	Answer Recall	Citation Precision	Citation Recall	Citation F-score
3	As in ranking	0.5667	0.6722	0.7444	0.6817
	Top result last	0.5633	0.7139	0.7778	0.7083
	Random	**0.6194**	0.7361	**0.8167**	**0.7394**
5	As in ranking	0.5444	0.6656	0.6469	0.6211
	Top result last	0.4796	0.7125	0.5831	0.613
	Random	0.4963	0.7367	0.6608	0.6596
10	As in ranking	0.4568	**0.7536**	0.541	0.5897
	Top result last	0.3835	0.5972	0.368	0.4105
	Random	0.3895	0.7506	0.4392	0.5019

References

S. J. Amouyal, T. Wolfson, O. Rubin, O. Yoran, J. Herzig, and J. Berant. QAMPARI: An open-domain question answering benchmark for questions with many answers from multiple paragraphs, 2023.

U. Anwar, A. Saparov, J. Rando, D. Paleka, M. Turpin, P. Hase, E. S. Lubana, E. Jenner, S. Casper, O. Sourbut, B. L. Edelman, Z. Zhang, M. Günther, A. Korinek, J. Hernandez-Orallo, L. Hammond, E. Bigelow, A. Pan, L. Langosco, T. Korbak, H. Zhang, R. Zhong, S. O. hEigeartaigh, G. Recchia, G. Corsi, A. Chan, M. Anderljung, L. Edwards, Y. Bengio, D. Chen, S. Albanie, T. Maharaj, J. Foerster, F. Tramer, H. He, A. Kasirzadeh, Y. Choi, and D. Krueger. Foundational challenges in assuring alignment and safety of large language models, 2024. URL `https://arxiv.org/abs/2404.09932`.

A. Asai, M. Gardner, and H. Hajishirzi. Evidentiality-guided generation for knowledge-intensive NLP tasks. In M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2226–2243, Seattle, United States, July 2022. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2022.naacl-main.162`. URL `https://aclanthology.org/2022.naacl-main.162`.

B. Bohnet, V. Tran, P. Verga, R. Aharoni, D. Andor, L. B. Soares, M. Ciaramita, J. Eisenstein, K. Ganchev, J. Herzig, K. Hui, T. Kwiatkowski, J. Ma, J. Ni, T. Schuster, L. S. Saralegui, W. W. Cohen, M. Collins, D. Das, D. Metzler, S. Petrov, and K. Webster. Attributed question answering: Evaluation and modeling for attributed large language models. 2022. URL `https://arxiv.org/abs/2212.08037`.

J. Dodge, M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell, and M. Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2021.emnlp-main.98`. URL `https://aclanthology.org/2021.emnlp-main.98`.

Y. Elazar, A. Bhagia, I. Magnusson, A. Ravichander, D. Schwenk, A. Suhr, P. Walsh, D. Groeneveld, L. Soldaini, S. Singh, H. Hajishirzi, N. A. Smith, and J. Dodge. What's in my big data?, 2024.

T. Gao, H. Yen, J. Yu, and D. Chen. Enabling large language models to generate text with citations. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore, Dec. 2023. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2023.emnlp-main.398`. URL `https://aclanthology.org/2023.emnlp-main.398`.

D. Garigliotti. SDG target detection in environmental reports using Retrieval-augmented Generation with LLMs. In *Proceedings of the Workshop on Natural Language Processing meets Climate Change (ClimateNLP 2024)*. Association for Computational Linguistics, 2024.

D. Garigliotti, B. Johansen, J. V. Kallestad, S.-E. Cho, and C. Ferri. EquinorQA: Large Language Models for Question Answering over proprietary data. In *European Conference on Artificial Intelligence*, 2024.

A. Jacovi, A. Caciularu, O. Goldman, and Y. Goldberg. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5084, Singapore, Dec. 2023. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.308. URL https://aclanthology.org/2023.emnlp-main.308.

P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

N. Liu, T. Zhang, and P. Liang. Evaluating verifiability in generative search engines. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, Singapore, Dec. 2023. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.467. URL https://aclanthology.org/2023.findings-emnlp.467.

J. Menick, M. Trebacz, V. Mikulik, J. Aslanides, F. Song, M. Chadwick, M. Glaese, S. Young, L. Campbell-Gillingham, G. Irving, and N. McAleese. Teaching language models to support answers with verified quotes. *ArXiv*, abs/2203.11147, 2022. URL https://api.semanticscholar.org/CorpusID:247594830.

OpenAI. GPT-4 Technical Report, 2024.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners, 2019. URL https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe.

O. Sainz, J. Campos, I. García-Ferrero, J. Etxaniz, O. L. de Lacalle, and E. Agirre. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore, Dec. 2023. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.722. URL https://aclanthology.org/2023.findings-emnlp.722.

Q. Si, T. Wang, Z. Lin, X. Zhang, Y. Cao, and W. Wang. An empirical study of instruction-tuning large language models in Chinese. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4086–4107, Singapore, Dec. 2023. Association for Computational Linguistics. https://doi.org/10.18653/

v1/2023.findings-emnlp.269. URL https://aclanthology.org/2023.findings-emnlp.269.

I. Stelmakh, Y. Luan, B. Dhingra, and M.-W. Chang. ASQA: Factoid questions meet long-form answers. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8273–8288, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.566. URL https://aclanthology.org/2022.emnlp-main.566.

H. Touvron, L. Martin, K. R. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. M. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. S. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. M. Kloumann, A. V. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288, 2023. URL https://api.semanticscholar.org/CorpusID:259950998.