Can LLMs Verify Arabic Claims? Evaluating the Arabic Fact-Checking Abilities of Multilingual LLMs

Ayushman Gupta[*], **Aryan Singhal**[*], **Thomas Law**[*], **Veekshith Rao**[*],
Evan Duan, Ryan Luo Li
Association of Students for Research in Artificial Intelligence (ASTRA)
astra.ai.lab@gmail.com

Abstract

Large language models (LLMs) have demonstrated potential in fact-checking claims, yet their capabilities in verifying claims in multilingual contexts remain largely understudied. This paper investigates the efficacy of various prompting techniques, viz. Zero-Shot, English Chain-of-Thought, Self-Consistency, and Cross-Lingual Prompting, in enhancing the fact-checking and claim-verification abilities of LLMs for Arabic claims. We utilize 771 Arabic claims sourced from the X-fact dataset to benchmark the performance of four LLMs. To the best of our knowledge, ours is the first study to benchmark the inherent Arabic fact-checking abilities of LLMs stemming from their knowledge of Arabic facts, using a variety of prompting methods. Our results reveal significant variations in accuracy across different prompting methods. Our findings suggest that Cross-Lingual Prompting outperforms other methods, leading to notable performance gains.

1 Introduction

Large language models (LLMs) have demonstrated remarkable proficiency in a wide range of tasks Minaee et al. [2024]. One particular area where LLMs have shown promising capabilities is in fact-checking and claim verification Choi and Ferrara [2024], Hoes et al. [2023], Lee et al. [2020], Zhang and Gao [2023]. The rise of fake news and misinformation in recent years has been well-documented, making fact-checking and claim verification essential to combat the rapid spread of misinformation.

However, previous work on fact-checking and claim verification using LLMs has primarily focused on English and Chinese facts and claims, leaving a significant gap in the exploration of multilingual fact-checking Cao et al. [2023], Quelle and Bovet [2024], Zhang et al. [2024]. This paper addresses this gap by focusing on fact-checking in Arabic, an inherently complex language due to its rich morphology, diverse dialects, and significant variation between written Modern Standard Arabic and spoken forms, using LLMs, which remains an under-explored domain. To this end, we benchmark LLM performance on a filtered dataset of 771 Arabic claims sampled from the X-fact dataset Gupta and Srikumar [2021a].

We utilize a variety of leading prompting techniques, including Zero-Shot (as a Baseline), English Chain-of-Thought Wei et al. [2023], Self-Consistency Wang et al. [2023], and Cross-Lingual Prompting Qin et al. [2023], to evaluate the effectiveness of LLMs in verifying Arabic claims. We present the variations in the accuracy of LLMs across different prompting methods. To our knowledge, this is the first work to evaluate the factual Arabic knowledge possessed by LLMs and their inherent Arabic fact-checking abilities based on this knowledge.

The remainder of this paper is organized as follows: In Section 2, we review related work. In Section 3, we define the problem of claim verification as explored in this paper. In Section 4, we describe the

[*] Equal contribution

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

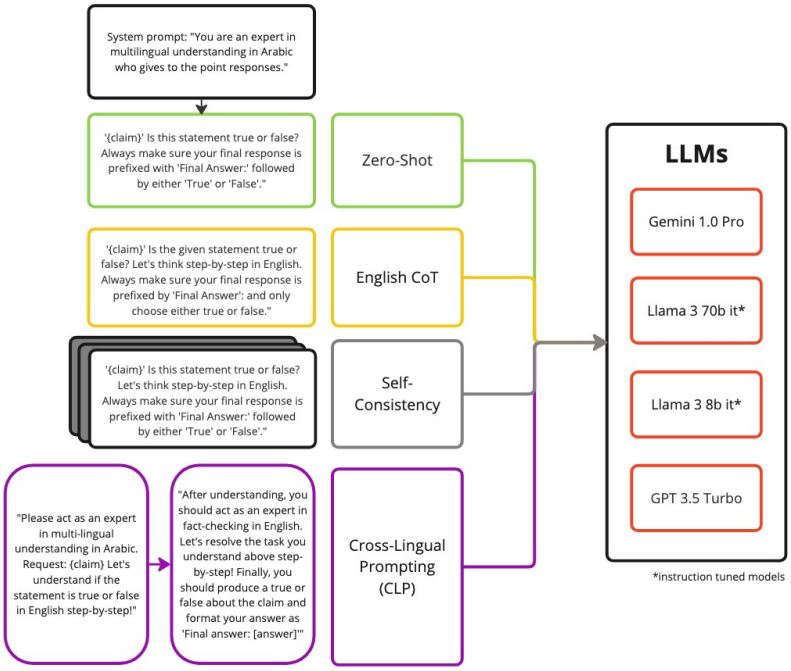

Figure 1: Workflow for comparing prompting strategies (Zero-Shot, English Chain-of-Thought (CoT), Self-Consistency, and Cross-Lingual Prompting (CLP)) used to evaluate the Arabic fact-checking capabilities of LLMs.

datasets, models, and evaluation methods used. We discuss our experiments in Section 5 and present our results in Section 6. Finally, we conclude in Section 7 and suggest directions for future research.

2 Related Work

Fact-Checking using LLMs With the rise of widespread misinformation, various studies have examined the capabilities of LLMs in fact-checking and claim verification. LLMs such as GPT-3 and GPT-4 excel in fact-checking when provided with sufficient contextual information, though they suffer from inconsistent accuracy Quelle and Bovet [2024]. Tian et al. 2023 suggests enhancing LLM factuality by fine-tuning models with automatically generated factuality preference rankings, which leads to improved factual accuracy without the need for human labeling. Cheung and Lam 2023 incorporates external evidence-retrieval to bolster fact-checking performance for the Llama model. Hu et al. 2023 examines the factual knowledge possessed by LLMs and their fact-checking capabilities using prompting techniques such as zero-shot, few-shot, and Chain-of-Thought.

Multilingual Fact-Checking using LLMs While there have been significant advancements in LLM-based fact-checking in English, multilingual fact-checking using LLMs remains relatively under-explored. Shafayat et al. 2024 examines the factual accuracy of LLMs across nine languages, including Arabic. Cekinel et al. 2024 explores cross-lingual learning and low-resource fine-tuning for fact-checking in Turkish, and uses in-context learning to evaluate LLMs' performance in this task.

Arabic and LLMs NLP in the Arabic language has seen significant advancements Darwish et al. [2021], Guellil et al. [2021] with Large Language Models (LLMs). Alyafeai et al. 2023 evaluates ChatGPT on a variety of Arabic NLP tasks. Pre-trained language models and language models fine-tuned on Arabic data have also demonstrated state-of-the-art performance in Arabic classification and generative tasks Alghamdi et al. [2023], Antoun et al. [2021], Deen et al. [2023]. Despite advancements in LLMs' capabilities in Arabic, fact-checking using LLMs remains under-explored.

Claim		Label
Arabic	**English Translation**	
وزيرة الصحة الفلسطينية تخرج عن طورها بسبب تفشي فايروس كورونا المستجد.	The Palestinian Minister of Health is out of her position due to the outbreak of the new Coronavirus.	0
طبيب مصري يقول إنّ مناعة التونسيين قد تكون علاجاً جديداً لفايروس كورونا (كوفيد-19).	An Egyptian doctor says that Tunisians' immunity may be a new treatment for the Coronavirus (COVID-19)	0
رئيس البرتغال يقف في المتجر وسط المواطنين ينتظر دوره.	The President of Portugal stands in the store among the citizens waiting for his turn	1
إصابة الفنانة رجاء الجداوي بفايروس كورونا المستجد(كوفيد_19) خلال تواجدها في مسقط رأسها بمحافظة الإسماعيلية	The artist, Ragaa Al-Jeddawi, was infected with the new Coronavirus (Covid_19) while she was in her home-town in Ismailia Governorate.	1

Figure 2: Examples of Arabic claims, their English translations, and ground-truth labels (0: false; 1: true) from the test data.

Althabiti et al. 2024 present Ta'keed: an LLM-based system for explainable Arabic fact-checking, and achieve promising results. In this work, we benchmark the Arabic fact-checking abilities of several multilingual LLMs using a variety of prompting methods.

3 Problem Definition

We treat claim verification as a binary classification task. For each claim x_i in our test dataset δ we prompt an LLM l to classify the claim as either 'true' ($\hat{y} = 1$) or 'false' ($\hat{y} = 0$), where $\hat{y}$ is the value predicted by l. In the case that l fails to return a binary value (inconclusive response) for $\hat{y}$, we take $\hat{y} = \neg y$.

4 Experimental Setup

4.1 Datasets

We utilize the X-fact dataset Gupta and Srikumar [2021a] as the source for the Arabic claims. The dataset is organized into several splits: Train, Development (Dev), In-domain Test (α_1), Out-of-domain Test (α_2), and Zero-Shot Test (α_3). We filter out those claims whose ground truth labels differ from either 'true' or 'false' from the Train, Dev, and In-domain Test (α_1) splits to create a test dataset δ containing 771 claims in Arabic:

$$\delta = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$$

where x_i is a claim in Arabic and $y_i \in \{0, 1\}$ is its ground truth label.

We note that 730 of the claims in the test dataset are false, while 41 are true. A sample from the test dataset is presented in Figure 2. Appendix A.1 contains further details about the test dataset.

4.2 Models

We conduct our experiments on Meta AI's Llama 3 8B and Llama 3 70B MetaAI [2024], Google DeepMind's Gemini 1.0 Pro Anil et al. [2023], and OpenAI's GPT-3.5-turbo. [2] For all models included in the study, we set the temperature to 0.7. The maximum possible token length for the outputs was set for each model given their respective context lengths.

[2] https://platform.openai.com/docs/models/gpt-3-5-turbo

4.3 Evaluation

We calculate an accuracy score for each LLM tested in each experiment. This accuracy score s is expressed as a percentage value as follows:

$$s = \frac{n_c}{n} \times 100\%$$

where n_c is the number of correct class predictions made by the LLM and n is the size of the test dataset. As mentioned in Section 3, inconclusive responses are treated as incorrect classifications.

5 Experiments

Figure 1 depicts the four prompting techniques used.

Zero-Shot Prompting We employ zero-shot prompts to gauge the baseline performance of the LLMs on the test data. A zero-shot prompt simply contains an Arabic claim x_i from the test dataset δ and an instruction Z to classify the claim as either 'true' or 'false'. As such, the LLM l's response is:

$$\hat{y} = l(x_i, Z)$$

English Chain-of-Thought Chain-of-Thought (CoT) prompting has been shown to significantly improve performance across various tasks Wei et al. [2023], including claim verification Hu et al. [2023]. This method enables models to articulate a clear, human-like, step-by-step reasoning process before arriving at a conclusion. Typically, in a zero-shot CoT prompt, the instruction "Let's think step by step" is added to the original instruction Z to create a new instruction Z_{CoT}. The response r_i of the LLM l to an Arabic claim x_i from the test dataset δ is computed as follows:

$$r_i = l(x_i, Z_{\text{CoT}})$$

$$r_i = (p_i, \hat{y}_i)$$

where p_i represents the reasoning path followed by the language model to arrive at the final answer $\hat{y}_i$.

We explore English Chain-of-Thought Qin et al. [2023], i.e. we add the instruction "Let's think step-by-step in English" to the original instruction Z. Since the test data is in Arabic, we hypothesize that prompting the model to reason out the answer in English would increase the likelihood of the LLM understanding the Arabic claim, thereby leading to performance gains.

Self-Consistency Wang et al. 2023 shows that replacing the greedy decoding used in Chain-of-Thought with 'self-consistency' significantly improves CoT reasoning. Self-consistency involves prompting a language model to generate a variety of reasoning paths to arrive at an answer and marginalizing these reasoning paths to choose the most consistent answer as the final answer.

We add Self-Consistency to Cross-Lingual CoT. For an Arabic claim x, we prompt the LLMs to generate *three* reasoning paths in English and obtain three responses such that $r_i = (p_i, \hat{y}_i)$. We choose the most consistent value of $\hat{y}_i$ as the final answer.

Cross-Lingual Prompting Qin et al. 2023 leverage Cross-Lingual Prompting (CLP) to enhance zero-shot Chain-of-Thought reasoning in language models in multilingual settings. They show that CLP outperforms popular prompting techniques including English Chain-of-Thought.

CLP involves two steps: **(i)** Cross-Lingual Alignment Prompting, where the language model is prompted to understand the Arabic claim verification task step-by-step in English, and **(ii)** Task-specific Solver Prompting, where the language model is prompted to solve the task using CoT reasoning.

Model	Correct	Incorrect	Inconclusive	Accuracy %	% Increase
Llama 3 8B-instruct					
Zero-Shot (Baseline)	455	305	11	59.01	–
English Chain-of-Thought	500	209	38	66.93	13.42
Self-Consistency	529	201	41	68.61	16.27
Cross-Lingual Prompting	664	91	9	**86.55**	46.67
Llama 3 70B-instruct					
Zero-Shot (Baseline)	310	438	23	40.21	–
English Chain-of-Thought	472	265	34	61.22	52.25
Self-Consistency	460	247	64	59.66	48.37
Cross-Lingual Prompting	620	134	17	**80.42**	100.00
Gemini 1.0 Pro					
Zero-Shot (Baseline)	236	531	5	30.60	–
English Chain-of-Thought	383	307	81	49.68	62.35
Self-Consistency	405	322	44	**52.53**	71.67
Cross-Lingual Prompting	381	385	5	49.41	61.47
GPT-3.5-turbo					
Zero-Shot (Baseline)	468	279	21	60.94	–
English Chain-of-Thought	461	244	66	59.79	-1.89
Self-Consistency	491	235	45	63.68	4.50
Cross-Lingual Prompting	603	116	2	**78.21**	28.34

Table 1: Results for each prompting method and LLM. '% Increase' denotes the percentage increase in model performance from the baseline (zero-shot).

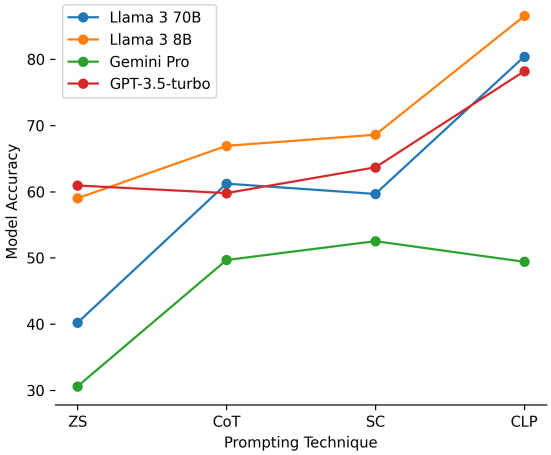

Figure 3: Model Accuracy vs Prompting Method

6 Results and Analysis

Our findings for each prompting approach are presented in Table 1. Figure 3 shows the relation between the prompting technique and model accuracy for each model. The percentage increase in accuracy from the baseline for each prompting method and model is shown in Figure 4. Generally, we find that the model accuracy increases from zero-shot to Cross-Lingual CoT to Self-Consistency, and typically reaches its maximum value in the CLP setting.

Figure 6 shows the relation between the prompting technique and the number of inconclusive answers for each LLM. As shown in the figure, the number of inconclusive responses, on average, increases when going from zero-shot to Cross-Lingual CoT or Self-Consistency. This number decreases in the CLP setting, in which the fewest inconclusive responses are returned.

Figure 5 shows a mostly linear relationship between the prompting technique and the number of correct answers for each LLM.

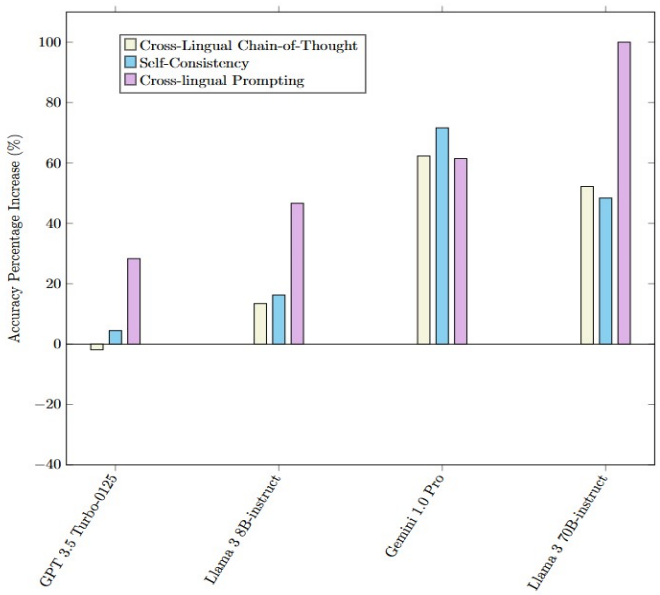

Figure 4: Percentage Increase from the Baseline (Zero-Shot) for each Prompting Method and LLM.

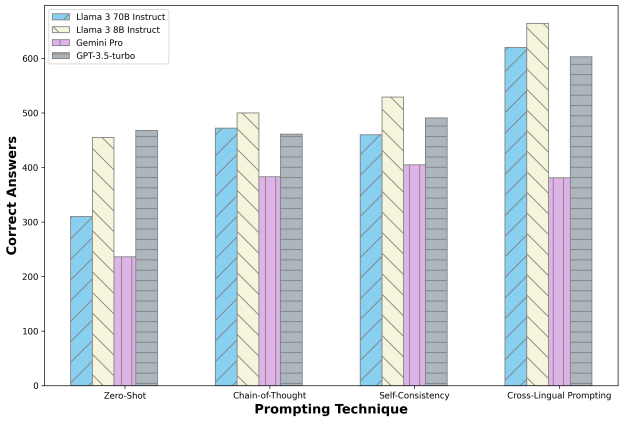

Figure 5: Variation of the number of correct answers with prompting method for each model.

6.1 Zero-Shot

Accuracy We find that Llama 3 70B Instruct achieves an accuracy of 40.21%, and Llama 3 8B achieves a higher accuracy of 59.01%. GPT-3.5-turbo achieves the second-best accuracy of 60.94% while Gemini Pro performs the worst with an accuracy of 30.60%.

Inconclusive Responses The language models show varying levels of inconclusive responses, with Llama 3 70B, Llama 3 8B, and GPT-3.5-turbo recording 23, 11, and 21 inconclusive responses respectively. Interestingly, despite a lower overall accuracy, Gemini 1.0 Pro returns only 5 inconclusive responses, which could indicate a propensity to deliver more decisive answers, albeit incorrect.

We observe that in the zero-shot setting, the LLMs are not effective fact-checkers and have room for improvement.

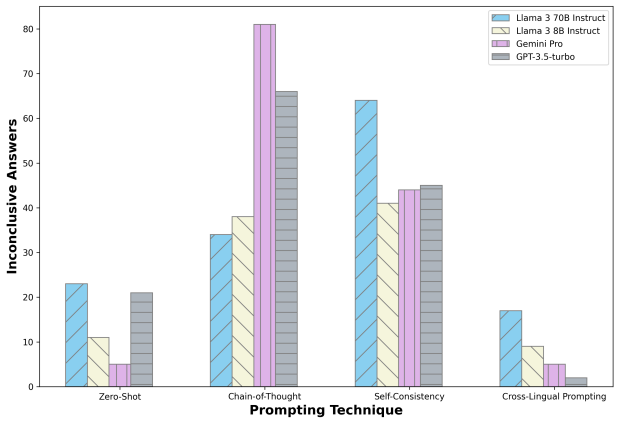

Figure 6: Variation of inconclusive answers for each model with different prompting techniques.

6.2 English Chain-of-Thought

Accuracy We observe that the English Chain-of-Thought (CoT) approach generally improves accuracy across most models compared to the zero-shot baseline. Llama 3 70B Instruct's accuracy increases by 52.25% (from 40.21% to 61.22%) in the CoT setting. Llama 3 8B Instruct's accuracy increases from 59.01% to 66.93%, a 13.42% increase. Gemini Pro's performance rises by 62.35% (49.68% from 30.60%).

In contrast, GPT-3.5-turbo performs with similar accuracy in the Cross-Lingual CoT setup, with a 1.89% drop in accuracy from its zero-shot performance.

Inconclusive Responses Despite the increase in accuracy for most LLMs, there was a significant rise in inconclusive responses across all models when applying the Cross-Lingual CoT method. This was particularly marked in Gemini Pro and GPT-3.5-turbo where inconclusive responses shot up to 61, 81, and 66 respectively. We find that while Cross-Lingual CoT appears to improve accuracy by allowing the LLMs to reason out the answers in English, it also seems to introduce greater uncertainty, leading to a higher number of inconclusive responses.

We find that generally, while English Chain-of-Thought leads to a rise in the number of inconclusive responses, the LLMs mostly return more correct answers, leading to a net increase in accuracy.

6.3 Self-Consistency

Accuracy We find that implementing Cross-Lingual CoT with Self-Consistency enhances model performance beyond Cross-Lingual CoT. For Llama 3 8B Instruct and Llama 3 70B Instruct, the accuracy increases by 16.27% and 48.37%, respectively. Gemini Pro's accuracy rises significantly, by 71.67%. GPT-3.5-turbo's accuracy increases by 4.50%. Llama 3 70B Instruct performs worse in the Self-Consistency setting than in the Cross-Lingual CoT setting.

Inconclusive Responses As shown in Figure 6, Self-Consistency leads to the highest number of inconclusive responses out of all the prompting methods. Llama 3 70B Instruct returns the highest number of inconclusive responses (64). We hypothesize that because the model is prompted to generate three lines of reasoning, it is susceptible to hallucinations and indeterminate chains of thought.

We observe that integrating Self-Consistency with Cross-Lingual CoT leads to an increase in the number of inconclusive responses returned by the LLMs. However, due to a rise in the number of correct answers, there is a net increase in model accuracy.

6.4 Cross-Lingual Prompting

Accuracy We find that cross-lingual prompting (CLP) often leads to the best model performance out of all the four prompting techniques. Llama 3 8B Instruct's accuracy improves by 46.67% over the baseline to achieve an accuracy of 86.55%, the highest among all tested models and methods. Similarly, GPT-3.5-turbo's performance also benefits from CLP, with its accuracy rising to 78.21% from a baseline of 60.94%. Llama 3 70B's performance reaches 80.42% from its baseline of 40.21%, a 100% improvement.

Inconclusive Responses Interestingly, while CLP improved accuracy across the board, it also led to a reduction in inconclusive responses for most models, indicating an increase in decisiveness. We observe a reduction in inconclusive responses from 11 to 9 for Llama 3 8B, 23 to 17 for Llama 3 70B, and 21 to 2 for GPT-3.5-turbo from zero-shot to CLP. The number of inconclusive responses remains unchanged for Gemini Pro.

Our findings suggest that CLP is extremely effective in clarifying the decision-making processes for these LLMs in an Arabic context while maintaining accuracy.

7 Conclusion and Future Work

In this study, we examined the Arabic fact-checking and claim verification capabilities of four LLMs: Llama 3 8B Instruct, Llama 3 70B Instruct, Gemini 1.0 Pro, and GPT-3.5-turbo. We employed four prompting techniques: Zero-Shot, English Chain-of-Thought, Self-Consistency, and Cross-Lingual Prompting. Our findings reveal that although these LLMs perform inadequately in a zero-shot setting, prompting techniques that engage reasoning capabilities significantly enhance their performance. In particular, Cross-Lingual Prompting showed substantial improvement in accuracy, suggesting that leveraging the reasoning capabilities of LLMs through sophisticated prompting strategies can effectively address the challenges posed by the complex morphology and diverse dialects of the Arabic language.

In future work, we aim to expand our dataset to establish a comprehensive benchmark for Arabic claim verification that includes diverse claims from various domains. Additionally, a future study could investigate how LLMs perform on fact-checking for claims in various independent Arabic dialects. Given the promising results of Cross-Lingual Prompting, we plan to explore other advanced prompting strategies, including few-shot prompting and Cross-Lingual Prompting with Self-Consistency, to further enhance performance.

Limitations

The scope of our analysis is restricted to a select group of LLMs. It would be interesting to investigate the Arabic fact-checking abilities of other leading models such as OpenAI's GPT-4 and Anthropic's Claude 3 series. Additionally, our dataset mainly comprises claims labeled as ground-truth false (730) as opposed to true (41). While this skew does not compromise the assessment of the LLMs' verification abilities, a more balanced distribution could provide deeper insights into their fact-checking capabilities in Arabic.

References

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey, 2024.

Eun Cheol Choi and Emilio Ferrara. Fact-gpt: Fact-checking augmentation via claim matching with llms. *arXiv preprint arXiv:2402.05904*, 2024.

Emma Hoes, Sacha Altay, and Juan Bermeo. Leveraging chatgpt for efficient fact-checking. *PsyArXiv. April*, 3, 2023.

Nayeon Lee, Belinda Z Li, Sinong Wang, Wen-tau Yih, Hao Ma, and Madian Khabsa. Language models as fact-checkers? *arXiv preprint arXiv:2006.04102*, 2020.

Xuan Zhang and Wei Gao. Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method. *arXiv preprint arXiv:2310.00305*, 2023.

Han Cao, Lingwei Wei, Mengyang Chen, Wei Zhou, and Songlin Hu. Are large language models good fact checkers: A preliminary study, 2023.

Dorian Quelle and Alexandre Bovet. The perils and promises of fact-checking with large language models. *Frontiers in Artificial Intelligence*, 7, February 2024. ISSN 2624-8212. doi: 10.3389/frai.2024.1341697. URL http://dx.doi.org/10.3389/frai.2024.1341697.

Caiqi Zhang, Zhijiang Guo, and Andreas Vlachos. Do we need language-specific fact-checking models? the case of chinese, 2024.

Ashim Gupta and Vivek Srikumar. X-fact: A new benchmark dataset for multilingual fact checking. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 675–682, Online, August 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.86. URL https://aclanthology.org/2021.acl-short.86.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2695–2709, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.163. URL https://aclanthology.org/2023.emnlp-main.163.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D. Manning, and Chelsea Finn. Fine-tuning language models for factuality, 2023.

Tsun-Hin Cheung and Kin-Man Lam. Factllama: Optimizing instruction-following language models with external knowledge for automated fact-checking, 2023.

Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip S. Yu, and Zhijiang Guo. Do large language models know about facts?, 2023.

Sheikh Shafayat, Eunsu Kim, Juhyun Oh, and Alice Oh. Multi-fact: Assessing multilingual llms' multi-regional knowledge using factscore, 2024.

Recep Firat Cekinel, Pinar Karagoz, and Cagri Coltekin. Cross-lingual learning vs. low-resource fine-tuning: A case study with fact-checking in turkish, 2024.

Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T. Al-Natsheh, Samhaa R. El-Beltagy, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Wassim El-Hajj, Mustafa Jarrar, and Hamdy Mubarak. A panoramic survey of natural language processing in the arab world, 2021.

Imane Guellil, Houda Saâdane, Faical Azouaou, Billel Gueni, and Damien Nouvel. Arabic natural language processing: An overview. *Journal of King Saud University - Computer and Information Sciences*, 33(5):497–507, June 2021. ISSN 1319-1578. doi: 10.1016/j.jksuci.2019.02.006. URL http://dx.doi.org/10.1016/j.jksuci.2019.02.006.

Zaid Alyafeai, Maged S. Alshaibani, Badr AlKhamissi, Hamzah Luqman, Ebrahim Alareqi, and Ali Fadel. Taqyim: Evaluating arabic nlp tasks using chatgpt models, 2023.

Asaad Alghamdi, Xinyu Duan, Wei Jiang, Zhenhai Wang, Yimeng Wu, Qingrong Xia, Zhefeng Wang, Yi Zheng, Mehdi Rezagholizadeh, Baoxing Huai, Peilun Cheng, and Abbas Ghaddar. Aramus: Pushing the limits of data and model scale for arabic natural language processing, 2023.

Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language understanding, 2021.

Mohammad Majd Saad Al Deen, Maren Pielka, Jörn Hees, Bouthaina Soulef Abdou, and Rafet Sifa. Improving natural language inference in arabic using transformer models and linguistically informed pre-training, 2023.

Saud Althabiti, Mohammad Ammar Alsalka, and Eric Atwell. Ta'keed: The first generative fact-checking system for arabic claims, 2024.

MetaAI. Introducing meta llama 3: The most capable openly available llm to date, Apr 2024. URL `https://ai.meta.com/blog/meta-llama-3/`.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Ashim Gupta and Vivek Srikumar. X-fact: A new benchmark dataset for multilingual fact checking, 2021b.

A Appendix

A.1 Dataset Creation

A.2 Dataset Statistics

The X-fact dataset Gupta and Srikumar [2021b] was utilized as our primary data source. The claims in the dataset are sourced from `https://misbar.com`.

A.3 Preprocessing Steps

1. Filtering: We filtered the dataset to include only claims that were labeled as either "true" or "false". Claims with other labels or those lacking verification were excluded from the finalized dataset.
2. Combining Splits: After filtering, the claims from the Train, Dev, and In-domain Test (α_1) splits were combined to form a single dataset for our experiments.

A.4 Dataset Composition

Table 2 shows the total number of Arabic claims and the number of Arabic claims filtered. After pre-processing, the test dataset contained a total of 771 Arabic claims.

Number of claims from Train set: 643

Number of claims from Dev set: 88

Number of claims from In-domain Test (α_1) set: 40

A.5 Label Distribution

TRUE Claims: 41 claims (5.32%)
FALSE Claims: 730 claims (94.68%)

Dataset Split	Total Number of Claims	Filtered Number of Arabic Claims (True & False)
Train	18246	643
Dev	3657	88
In-domain Test (α_1)	2406	40
Total	24309	771

Table 2: Summary of the dataset splits before and after filtering claims labeled as 'TRUE' or 'FALSE'.

Compute Resources

All experiments were conducted using a combination of cloud-based GPU instances and local compute resources. The specific details of the compute setup are outlined below:

GPU Resources

For training and evaluating the LLMs, we utilized the following GPU configurations:

- **Cloud GPU Instances:** Experiments were primarily conducted on NVIDIA A100 40GB GPUs hosted on cloud providers (e.g., AWS EC2, Google Cloud Platform). Each instance included 8 A100 GPUs with 320GB of total VRAM. The experiments on these instances ran across multiple GPUs in parallel for faster throughput.
- **Local GPU Instances:** Some experiments were run locally on a system equipped with 2 NVIDIA RTX 3090 GPUs, each with 24GB of VRAM.

Compute Time

- **Zero-Shot Prompting:** Each model required approximately 1 hour of compute time on a single GPU for evaluating the 771 claims using zero-shot prompting.
- **Chain-of-Thought Prompting:** English Chain-of-Thought and Cross-Lingual Chain-of-Thought evaluations required about 3 hours per model per experiment, as generating reasoning chains increased compute time.
- **Self-Consistency:** The self-consistency experiments, which required generating multiple reasoning paths for each claim, took approximately 6 hours per model.

Total Compute Resources

The total compute time across all models and experiments was approximately 100 GPU hours. Most of this time was spent on the Self-Consistency and Cross-Lingual Prompting experiments due to the additional reasoning paths generated.

Memory and Storage

Each experiment required at least 200GB of storage for caching intermediate results and model checkpoints. The average memory usage was 120GB during peak execution of the larger models (e.g., Llama 3 70B).

Software Environment

All experiments were run using the following software stack:

- **Operating System:** Ubuntu 20.04 LTS
- **Deep Learning Framework:** PyTorch 2.0
- **CUDA Version:** 11.7
- **Other Dependencies:** Transformers (Hugging Face), Python 3.9, and specific drivers for NVIDIA GPUs.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the contributions, focusing on the performance of multilingual LLMs in Arabic fact-checking using various prompting techniques. The experimental results in the paper match these claims.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a "Limitations" section, discussing the restricted scope of models tested and the class imbalance in the dataset (730 false claims vs. 41 true claims). It acknowledges the need for further exploration of other LLMs and more balanced datasets.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper is empirical and does not present any new theoretical results or proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The experimental setup, models, dataset, and evaluation methods are clearly detailed in the "Experimental Setup" and "Experiments" sections. The test dataset, model parameters, and evaluation metrics are all described in a reproducible manner.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: While the data and models used are publicly available, the paper does not provide access to the exact code used in the experiments. Code will be released in the final version.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides detailed information about the dataset used (X-fact), the specific LLMs tested, the hyperparameters (e.g., temperature of 0.7), and the methods used for evaluation, as seen in the "Experimental Setup" section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper does not provide statistical significance measures, such as error bars or confidence intervals, for the reported accuracy scores.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provides details in the appendix, including the types of GPUs used (NVIDIA A100 and RTX 3090), total compute time (approximately 100 GPU hours), memory usage (120GB), storage requirements (200GB), and the software environment (PyTorch, CUDA, Ubuntu). This ensures that the experiments can be reproduced accurately.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research adheres to ethical standards, focusing on benchmarking LLMs for Arabic fact-checking, with no risks of harm or misuse identified. No privacy, fairness, or ethical violations were raised during the research.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The potential positive societal impact of improving fact-checking in Arabic is discussed. The paper does not foresee negative impacts, but it acknowledges the risks associated with LLMs, including hallucinations and factual inconsistencies.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [No]

Justification: The paper does not address this directly, as it does not release a new model or dataset. The work primarily evaluates existing models, and no high-risk release is involved.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets (X-fact) and models (Llama 3, Gemini, GPT-3.5-turbo) are properly cited, and their original sources and licenses are mentioned in the references.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not introduce new assets such as datasets or models.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: No crowdsourcing or research with human subjects is involved in this paper.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The research did not involve human subjects and therefore did not require IRB approval.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.