Fact-checking: Generative LLMs Don’t Pay Attention to Details

Anonymous ACL submission

Abstract

Fact-checking is an established knowledge-intensive natural language processing (NLP) task, including evidence retrieval and claim verification steps. Meanwhile, generative large language models (LLMs) increasingly memorize facts as the model sizes grow. This memorization ability of language models leads to the question of whether extractive retrieval is still necessary if the facts required to check a claim have been seen during pre-training. Consequently, this paper evaluates generative LLMs’ closed-book fact-checking performance on two Wikipedia-based datasets, FEVER and HOVER. Instead of retrieving the evidence from external knowledge bases, we let the model generate rationales and verify the claim by itself in a few-shot setup. For the simple dataset, the best verification performance of selected open-source LLMs achieves an F1 score up to 89% (GPT-4 93%); for the complex dataset, the performance reaches 62% (GPT-4 70%). Compared to the claim-only verification, the extra rationale generation step has boosted the verification performance by 4.05 percentage points on FEVER and 5.12 on HOVER.

1 Introduction

With the increasing spread of misinformation, fact-checking has become an urgent topic in our daily life. In academics, the research on fact-checking spans from political science to computer science (Shu et al., 2017). The standard pipeline for automated fact-checking consists of three stages: claim detection, evidence retrieval, and claim verification with the retrieved evidence (Guo et al., 2022). The stage of evidence retrieval is to retrieve relevant information from external structured and unstructured knowledge bases, typically done in an extractive and open-book way. Given a claim and a corpus, the task is to select relevant documents and sentences from the corpus for further claim verification. Several benchmark datasets for fact-checking in the research community are based on the Wikipedia corpus, e.g., FEVER (Thorne et al., 2018) and HOVER (Jiang et al., 2020). The Wikipedia corpus is often part of the training data of large language models (LLMs). In the pre-training stage, LLMs can memorize large amounts of facts from the training data in their parameters. Given a prompt, generative LLMs can then generate relevant texts, raising the question of whether we can retrieve the needed evidence in a generative and closed-book way from LLMs instead of the standard extractive way of the above-mentioned fact-checking datasets. Given retrieved evidence, we check whether the claim is entailed in the given evidence for the claim verification step. Often, the task is implemented as a classification task with fine-tuned BERT-based models (Soleimani et al., 2020; DeHaven and Scott, 2023). Recent research shows increasing reasoning abilities in generative LLMs. Thus, this poses the question of whether we can verify the claims in a generative way instead of traditional classification with BERT-based models.

This paper investigates state-of-the-art open-source generative models with Wikipedia-based fact-checking datasets in a closed-book style. As the open-book fact-checking, we verify the claims in a pipeline. Given a claim, we retrieve the needed rationales in a generative way from pre-trained LLMs with in-context learning. In this stage, the LLMs generate relevant rationales about the claim from the knowledge learned during the pre-training phase. Based on the generated rationales, the generative LLMs will verify the claim’s truthfulness in a few-shot setup. Since fact-checking real-world claims often requires multiple pieces of evidence and reasoning steps (Pan et al., 2023), i.e., multi-hop, we include datasets with simple and complex claims.

Our contributions include: (1) We are the first to evaluate the generative LLMs for fact-checking with a closed-book setup, including rationale generation and claim verification. (2) We measure the knowledge and reasoning capabilities of representative generative LLMs, not only GPT-4 but also open-source LLMs, and compare them with traditional retrieval-based verification. We find that instruction-tuned models gain more reasoning capabilities at the cost of factuality. (3) We show that as claims become complex, current generative LLMs cannot pay attention to details in the rationale generation and claim verification phases. (4) We demonstrate the extra rationale generation step can boost the verification performance of generative LLMs, especially for complex claims.

2 Related Work

Generative Large Language Models Currently, generative large language models are gaining a lot of attention due to the popularity of ChatGPT. The sizes of these models keep increasing, e.g., from GPT-1 with 117 million parameters to GPT-3 with 175 billion parameters (Radford et al., 2018; Brown et al., 2020; Zhao et al., 2023). The training method, reinforcement learning from human feedback (RLHF) is applied to tune pre-trained generative models (Ouyang et al., 2022). With RLHF, LLMs get a better understanding of human instructions. Due to the large size of generative models, further fine-tuning of pre-trained generative models is computationally expensive. Thus, in-context learning (Brown et al., 2020), which learns from a few examples and does not update model parameters during the learning process, is widely applied for downstream NLP tasks (Min et al., 2022; Zhou et al., 2023).

Gehrmann et al. (2023) define natural language generation (NLG) tasks, including summarization, machine translation, dialog generation, and data-to-text generation, as those in which a machine learning model can be trained to maximize the conditional probability $p(y|x)$ where y is natural language, and x is an input that provides information about what should be generated. Hallucination is a well-known problem in NLG. Ji et al. (2023) classify the contributors to hallucination in NLG into two groups: hallucination from data, hallucination from training and inference. Baan et al. (2023) analyze the uncertainty of NLG and identify the main sources of uncertainty as input ambiguity, errors and complexity, the open-endless of the communicative task, the agent's personal perspective, and the final linguistic realization and modeling. Factuality and faithfulness in NLG have been addressed in several studies. Lee et al. (2022) evaluate the factuality in open-ended text generation with various metrics, e.g., the overlap of named entities, entailment rate, perplexity, and diversity repetition. Tam et al. (2023) propose a factual inconsistency benchmark for evaluating the factual consistency of LLMs through summarization. Min et al. (2023) propose FActScore for evaluating the factuality of LLM-generated texts with atomic facts.

Fact-checking Language models (BERT-based) are used as fact-checkers in a closed-book style, where certain facts (e.g., entities) are masked and the claim is verified against the predicted facts (Lee et al., 2020). In our work, we also utilize the memorization ability of language models. However, we let the LLMs generate all necessary rationales instead of entities. Dense passage retrieval from question & answering (Karpukhin et al., 2020) is widely applied for document retrieval with a fine-tuned bi-encoder for selecting top-k candidates. Cross-encoders have often been used to further re-rank documents and select sentences in the documents as evidence (Soleimani et al., 2020; DeHaven and Scott, 2023). Other architectures like Poly-encoder (Humeau et al., 2020), ColBERT (Khattab and Zaharia, 2020) are proposed to balance prediction quality and speed with late interaction mechanisms.

The stage of claim verification is treated as a natural language inference (NLI) task with retrieved rationales (sentences) as the premise and the claim as the hypothesis. The target is to check whether a claim is entailed in the retrieved rationales. Pre-trained language models, mostly BERT-based, which are further fine-tuned with NLI datasets, are typically applied to the task (Lewis et al., 2020; Liu et al., 2019; Williams et al., 2018). Several fact-checking studies have applied fine-tuned BERT-based models for verifying claims given retrieved evidence (Martín et al., 2022; Arana-Catania et al., 2022). Hansen et al. (2021) show that machine learning models, e.g., random forest, LSTM, and BERT-based models, do not really learn to reason: Given only evidence without claims of political fact-checking datasets, the models achieve the highest effectiveness. Recently, we have seen increasing reasoning abilities in generative models as the sizes of models grow (Wei et al., 2022a). The phenomenon of increasing abilities with growing model sizes is referred to as the emergent abilities of LLMs (Wei et al., 2022a). Huang and Chang

(2023) summarize techniques applied to improve or elicit reasoning in LLMs as fully supervised fine-tuning, prompting & in-context learning, hybrid method. Chain-of-thought prompting is proposed in Wei et al. (2022b), an example of in-context reasoning. The authors show that with intermediate reasoning steps, LLMs perform better in various reasoning tasks, e.g., arithmetic reasoning, commonsense reasoning, symbolic reasoning, etc. Pan et al. (2023) tackle the complex multi-hop fact-checking problem with claim decomposition, leveraging the in-context learning ability of LLMs.

3 Methodology

In this section, we explain the theory behind in-context learning with generative LLMs. Based on this concept, the templates for rationale generation and claim verification are introduced.

3.1 In-context Learning

In the following, we briefly describe the in-context learning method for rationale generation and claim verification. Xie et al. (2022) explain in-context learning from a Bayesian perspective. Their assumption on the structure of pre-training documents is that a document is generated by first sampling a concept, and then the document is generated by conditioning on the latent concept. The prompts of in-context learning are lists of training examples and one test example, where the training examples are independent and identically distributed. Each example is a sequence conditioned on the same prompt concept, which describes the task to be learned. The process of locating learned capabilities is the Bayesian inference of a prompt concept that all examples in the prompt share. Mathematically, the posterior predictive distribution can be formulated as $\mathrm{p}(out|prompt) = \int_c \mathrm{p}(out|c, prompt)\mathrm{p}(c|prompt)\, d(c)$, where c is the latent concepts and out is the generated output conditioned on the shared concept in the prompt.

3.2 Rationale Generation

Our closed-book evaluation pipeline consists of rationale generation and claim verification. Evidence-based fact-checking datasets consist of pairs of $(claim, rationales)$, where $rationales$ consists of corresponding facts, i.e., sentences, supporting or refuting the $claim$. Figure 1 shows the template for generating rationales with LLMs. In the template, n is the number of examples we show to the generative LLMs. m, p, q are the number of corresponding rationales of each claim. With n examples, the generative LLMs should infer that the task is to generate relevant factual rationales (sentences) for verifying the claim. In the test example t, only a claim is given, and the LLMs should generate the corresponding rationales.

To verify the factuality of claim <$claim_1$>, following factual evidence is needed: <rat_{11}, ..., rat_{1m}> ###
To verify the factuality of claim <$claim_2$>, following factual evidence is needed: <rat_{21}, ..., rat_{2p}> ###
...
To verify the factuality of claim <$claim_n$>, following factual evidence is needed: <rat_{n1}, ..., rat_{nq}> ###
To verify the factuality of claim <$claim_t$>, following factual evidence is needed: ____________________

Figure 1: Prompt template for rationale generation

Given the premise <rat_{11}, ..., rat_{1m}>, is the hypothesis <$claim_1$> true? Yes.###
Given the premise <rat_{21}, ..., rat_{2p}>, is the hypothesis <$claim_2$> true? No.###
...
Given the premise <rat_{n1}, ..., rat_{nq}>, is the hypothesis <$claim_n$> true? Not enough information.###
Given the premise <rat_{t1}, ..., rat_{tq}>, is the hypothesis <$claim_t$> true? _____________

Figure 2: Prompt template for claim verification with rationales

3.3 Claim Verification

To verify the claim, we also apply in-context learning with LLMs. We build the template in the traditional natural language inference (NLI) format. A standard NLI task consists of a premise, a hypothesis, and a label. The task is to verify whether the hypothesis is entailed in the premise. The label for verification can be SUPPORTS, REFUTES, and NEI (Not Enough Information). We use the prompt template in Figure 2 for claim verification with rationales. Accordingly, we show the LLMs n examples to infer the in-context learning task. For comparison, we also evaluate the claim-only setup, where the generative LLMs verify the claims only based on the claims without any rationales. The prompt template is shown in Figure 3.

4 Experiments

In the following, we describe the experimental setup. In the first step, we introduce selected fact-checking datasets and generative LLMs for evaluation. We then evaluate the LLMs' rationale generation and claim verification capabilities with the

Is the following claim <$claim_1$> true? Yes.###
Is the following claim <$claim_2$> true? No.###
...
Is the following claim <$claim_n$> true? Not enough information.###
Is the following claim <$claim_t$> true? ____________

Figure 3: Prompt template for claim-only verification

selected datasets. The evaluation focuses on the knowledge and reasoning capabilities of generative LLMs.

4.1 Datasets

We follow two criteria to select datasets for the evaluation of generative LLMs. Since our evaluation focus is evidence-based fact-checking, we only consider datasets with labeled rationales. Since we do a closed-book generation of rationales, the corpus from which the claims and corresponding rationales originate should be part of the pretraining data of LLMs. The Wikipedia corpus is in the standard training data of many LLMs, so Wikipedia-based datasets can be used for this evaluation. FEVER (Thorne et al., 2018) is one of the earliest and most well-known datasets for evidence-based fact-checking. The claims in the FEVER dataset are generated from the Wikipedia corpus. The study (Jiang et al., 2020) shows that 87% of FEVER claims require information from a single Wikipedia article, and manual inspection of the dataset shows that many claims in the FEVER dataset are very simple. HOVER (Jiang et al., 2020) is another fact-checking dataset based on Wikipedia corpus. This dataset focuses on the multi-hop problem in the fact-checking pipeline. The number of hops in the dataset ranges from 2 to 4. Therefore, the HoVER dataset is a good candidate for evaluating generative LLMs' capabilities in a complex setup.

In summary, FEVER and HOVER have fulfilled the selection criteria for our evaluation. For both datasets, we randomly sample 300 representative test examples from each dataset's development set. Only examples with verifiable claims are included since it's hard to evaluate the performance of rationale generation on non-verifiable claims, which do not have labeled rationales as references for evaluation. We focus on verifiable claims with rationales supporting or refuting them. Concretely, for the FEVER dataset, there are 100 test samples with 1 rationale, 100 with 2 rationales, and 100 with 3 rationales. In each 100 test samples, they are evenly distributed between 2 categories, SUPPORTS and REFUTES, each category with 50 samples. We have the same distribution for the HOVER dataset according to the number of hops, namely 2, 3, and 4. Each group has also 50 supported and 50 refuted claims. After publication, we will make the used split and the generated data publicly available.

4.2 Generative LLMs for Evaluation

Since the number of open-source LLMs is large (Zhao et al., 2023), we use HuggingFace's Open LLM leaderboard[1] as the reference for choosing LLMs. We select three representative open-source LLM families, BLOOM (Scao et al., 2023), Falcon (Almazrouei et al., 2023) and Llama2 (Touvron et al., 2023). For each model family, we include two versions, namely the original pretrained LLM version and the instruction-tuned version. Since there are various versions of the instruction-tuned models, we only consider the officially released instruction-tuned version. We evaluate the largest version of each model family, namely BLOOM 176B, Falcon-180B, and Llama2-70B. The corresponding instruction-tuned versions are BLOOMZ-176B, Falcon-180B-Chat, and Llama2-70B-Chat. In addition to these model families, we also add GPT-4 from OpenAI as a benchmark for evaluation. To evaluate the effect of model size on performance, we include the Llama2 family with extra sizes 7B and 13B. The implementation details are described in Appendix A.1.

4.3 Rationale Generation

In the first step, we generate rationales from LLMs for claim verification. As introduced in Figure 1, we use few-shot in-context learning for rationale generation. Considering the characteristics of the two datasets, we show each LLM 6 examples: 3 supported and 3 refuted claims with corresponding gold rationales. We randomize the 6 selected examples for each dataset so that they do not follow specific patterns. Appendix D.1 describes the details about selected example prompts. For comparison, we also include the retrieved rationales given the Wikipedia corpus, i.e., the open-book setup. We use the popular document retriever from (Hanselowski et al., 2018) for the FEVER dataset and train a cross-encoder for sentence selection (Soleimani et al., 2020). For the HOVER dataset, we use checkpoints of the retrieval pipeline

[1] https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

from (Khattab et al., 2021), which achieves state-of-the-art performance. The implementation details is included in Appendix A.2

Evaluation of Generated Rationales For evaluating the generated rationales, we use the labeled rationales (gold) in each dataset as the reference. The evaluation focuses on the memorization ability of LLMs and the factuality of the generation. Several evaluation metrics exist for generative downstream tasks, e.g., machine translation and summarization. Following Lee et al. (2022), we extract named entities in generated rationales and gold rationales, utilizing the spaCy package for named entity recognition (NER) (Honnibal et al., 2020), and report the F1 score of common named entities. We also report other overlap based metrics BLEU-4[2] and ROUGE-L[3] (Shuster et al., 2021). In addition to the overlap-based metrics, we include a semantic metric BERTScore (Zhang et al., 2020).

Table 1 shows that the gap between closed-book generation and open-book retrieval is still large, especially for the difficult HOVER dataset, where the sophisticated multi-hop retrieval pipeline works much better. GPT-4 achieves the best performance on both datasets among the generative LLMs. Llama2-70B and Falcon-180B have very similar performance. The original language models perform in most cases better than their corresponding instruction-tuned models. This suggests that instruction-tuning can reduce the factuality of the original pre-trained models. The statistics of Llama2 families demonstrate that larger models tend to perform better. Between the two datasets, most generative models perform better on the FEVER dataset than the HOVER dataset. However, the gap is not large. There are many simple claims in the FEVER dataset, e.g., "Ayananka Bose is a person.", "Bradley Cooper refuses to be in films.". There are many ways to support or refute these claims. Often, one of the gold rationales is enough to verify these claims. This can lower the performance of overlap-based metrics. In contrast to FEVER, the claims in HOVER are more complex and need at least two rationales for verification due to the multi-hop characteristic. A typical claim in the HOVER dataset is "Confession of Murder, a 2012 South Korean action thriller film, was turned into the 2014 Indian-Malay melodramatic thriller Angels.". The claims contain more named entities and give LLMs more hints to generate relevant rationales. All these factors lead to comparable performance on overlap-based metrics, although one dataset is more complicated.

Qualitative Analysis We look into the generated rationales in detail. Due to the large number of generated texts, we only inspect the generated rationales with GPT-4, Llama2-70B, and Llama2-70B-Chat. We randomly select 30 examples for each dataset, evenly distributed according to the claim label and number of rationales (or hops). We mainly check the factuality and sufficiency of generated rationales compared to the gold rationales. We query Wikipedia or Google to verify the generated facts that do not exist in the gold rationales. Table 2 summarizes the percentages of hallucination, insufficiency, and correctness of generated rationales for manually inspected claims. We classify the generated text as hallucinated when there are wrong facts or unverifiable information in it. Insufficient means important facts are missing in the generated rationales for claim verification, e.g., in the multi-hop cases. Only when no factual errors exist and all necessary facts exist in the generated rationales, can we say the generation is correct. The results show that hallucination happens more often in HOVER compared to FEVER. Figure 2 demonstrates the gold and generated rationales for a claim from three selected models. The generated rationales by GPT-4 contain factual errors. One hop about the plant Carpentaria is missing in the generation by Llama2-70B. Both hallucination and insufficiency exist in the generation by Llama2-70B-Chat. We observe more hallucinations in the generated texts when the claim is more complex. These factual errors are often minor, e.g., the generated facts about a person are all correct except for the birth date. The insufficiency problem occurs when multi-hop rationales are needed to verify the claims. Multi-hop claims require LLMs' knowledge and reasoning capabilities to generate all necessary rationales. The difficulty for generation increases as the number of hops increases. Compared to the original language models, instruction-tuned models show better reasoning capabilities regarding sufficiency at the cost of factuality.

4.4 Claim Verification

We have three setups for claim verification: claim verification with gold rationales, claim verification with generated rationales, and claim-only verification. We further apply in-context learning, con-

[2] https://www.nltk.org/_modules/nltk/translate/bleu_score.html

[3] https://pypi.org/project/rouge/

	FEVER				HOVER			
Model	Entity	BLEU	Rouge-L	BERT-S	Entity	BLEU	Rouge-L	BERT-S
BLOOM-176B	19.52	6.70	26.94	85.23	18.48	7.27	27.94	85.23
BLOOMZ-176B	17.36	4.42	25.30	85.48	19.23	5.81	26.88	85.80
Falcon-180B	32.50	20.14	39.30	87.51	28.75	17.56	35.85	87.04
Falcon-180B-Chat	33.09	19.08	37.51	87.63	26.35	15.08	32.06	86.66
Llama2-7B	26.76	11.54	32.33	86.48	21.63	10.12	29.60	85.89
Llama2-7B-Chat	23.62	7.73	30.29	86.60	21.87	8.18	28.85	86.64
Llama2-13B	28.93	14.89	35.38	86.84	23.60	11.80	32.76	86.28
Llama2-13B-Chat	27.57	10.13	32.90	87.14	23.88	9.40	30.04	86.71
Llama2-70B	32.71	19.53	38.71	87.58	29.42	17.58	36.26	87.28
Llama2-70B-Chat	28.44	13.32	34.26	87.34	27.28	10.70	31.71	87.19
GPT-4	**40.75**	**25.96**	**45.50**	**88.99**	**35.08**	**22.46**	**38.90**	**88.04**
Open-book Retrieval	53.26	40.62	55.68	89.74	67.85	66.63	64.00	93.32

Table 1: Evaluation of generated rationales with gold rationales as reference. F1 score (%) is used for Entity, Rouge-L, and BERT-S. The numbers in bold indicate the best generation performance. The underlined numbers indicate the best performance among open-source generative LLMs.

Claim: This organism and the genus Ginseng belongs to are both plant genera. The Gulf named after the organism is part of Port McArthur Tidal Wetlands System.
Gold rationales: Ginseng is any one of the species of slow-growing perennial plants with fleshy roots, belonging to the genus "Panax" of the family Araliaceae. The Panax (ginseng) genus belongs to the "Araliaceae" (ivy) family. Carpentaria acuminata (carpentaria palm), the sole species in the genus Carpentaria, is a palm native to tropical coastal regions in the north of Northern Territory, Australia. The Port McArthur Tidal Wetlands System comprises a 994km tract of tidal wetlands on the south-west coast of the Gulf of Carpentaria in the Northern Territory of Australia.

GPT-4: The genus Ginseng is a plant genus in the family Araliaceae. [Port McArthur Tidal Wetlands System] The system includes the Gulf of Carpentaria, named after the plant genus Carpentaria, which is a genus of flowering plants in the family Arecaceae.
Llama2-70B:The genus Ginseng belongs to the family Araliaceae. The genus Ginseng is a plant genus. The Gulf of Carpentaria is a large, shallow sea enclosed on three sides by northern Australia and bounded on the north by the Arafura Sea. The Gulf of Carpentaria is part of the Port McArthur Tidal Wetlands System.
Llama2-70B-Chat: Ginseng is a plant genus in the family Araliaceae. The Gulf of Mexico is a body of water named after the organism. The Gulf of Mexico is part of the Port McArthur Tidal Wetlands System.

Figure 4: Example for rationales generation for a supported claim (4-hops) in HOVER. Important entities are marked in blue. Hallucinations are marked in red. Similar examples can be found in Appendix B.1.

Dataset	Model	H	I	C
FEVER	Llama2-70B	23	20	57
	Llama2-70B-Chat	47	17	43
	GPT-4	13	3	83
HOVER	Llama2-70B	57	30	20
	Llama2-70B-Chat	70	23	17
	GPT-4	40	13	47

Table 2: Percentages of generated rationales for the manually inspected 30 claims are hallucinated (H), insufficient (I), and correct (C) for selected models. Since the generated rationales for a claim can be hallucinated and insufficient, the sum of percentages in a row can be larger than 100.

cretely 6-shots learning. The prompt template in Figure 2 is used for claim verification with gold and generated rationales and the template in Figure 3 is used for claim-only verification. Following the original data setup, FEVER has three verification labels, SUPPORTS, REFUTES, NEI (Not Enough Information), and HOVER has two labels, SUPPORTS and REFUTES. For FEVER, we show 2 examples for each label in the prompt, and for HOVER 3 examples. Studies show that few-shot learning can be unstable and the order of example labels in the prompt can impact the accuracy (Zhao et al., 2021). To reduce the impact of the order, we shuffle the examples in the prompt with 3 different seeds. For each claim in the test set, we generate 3 predictions with 3 different prompts, namely 3 different orders of the 6 examples. We show the prompts of one seed in Appendix D.2. We report verification accuracies based on majority voting, by which at least 2 of 3 predictions must be correct. If the verification is undecided, only possible with FEVER, we randomly choose a label. For claim verification with gold rationales, we further fine-tune the ROBERTA-LARGE-MNLI model (Liu et al., 2019) for both datasets with corresponding training data. Fine-tuned models are used as the benchmark for comparison. The fine-tuning details are described in Appendix A.3.

Gold Rationales Given gold rationales, the claim verification mainly relies on the reasoning ability of generative LLMs. Table 3 shows that all instruction-tuned models perform better than their original pre-trained language models. Thus, instruction-tuning has improved the reasoning capability of the original LLMs. After manually inspecting both datasets, we have found 7 labeling errors by FEVER and 8 by HOVER in the 300 test examples. We report the corrected results in Appendix C. Since the claims in FEVER are relatively simple, many generative models perform very well with 6-shot learning, even better than the fine-tuned model, given the gold rationales. Currently, generative LLMs can verify the claims accurately, given the limited number of facts in the premise. However, the verification performance drops significantly in the more complex HOVER dataset. We investigate the verification results according to the number of hops and compare our verification results with ProgmFC (Pan et al., 2023), a few-shot neuro-symbolic multi-hop fact-checking model. The results[4] in Table 4 show that as the number of hops increases, the performance of selected models decreases. However, the performance drop with the fine-tuning model is small. This is partially due to the fact that many refuted claims in HOVER are created by modifying supported claims with word or entity substitution, adding extra unverifiable or wrong information, etc. We find that generative LLMs have difficulty detecting these minor changes, especially when the number of hops increases. In Figure 5, we demonstrate some examples where all our top generative models fail to detect while the fine-tuned model predicts correctly. With fine-tuning, the model can learn these modification patterns from the training data, i.e., substitution, extra non-existing information in the premise, etc. This doesn't necessarily mean that the fine-tuned model has better reasoning capabilities.

Claim: The Swan of Italy was taught by the Italian composer Giovanni Furno.
Gold rationales: Giovanni Furno ... was an Italian composer and famous music teacher. Among his students were Vincenzo Bellini and .... Vincenzo Salvatore Carmelo Francesco Bellini ... for which he was named the Swan of "Catania".
Claim: Vinay Pathak co-hosted the 59th National Film Awards in Bollywood. His co-host currently acts as the lead character in a tv series that premiered on March 02, 2015.
Claim: The son of this director produced the summer 2015 film starring Jesse Eisenberg, Gabriel Byrne, Isabelle Huppert, David Strathairn, and Amy Ryan. This director directed Begynnelsen på en historie.

Figure 5: Wrong verification examples by generative LLMs given gold rationales, which the fine-tuning model verifies correctly. We only show the claims for the second and third examples and leave out the gold rationales here. All facts in both claims are correct except for the red-marked extra information, which doesn't exist in gold rationales. Further examples can be found in Appendix B.2.

Claim-Only & Generated Rationales As of now, the claim verification with these two setups is in the closed-book style, without access to external knowledge bases. Claim-only verification requires both knowledge and reasoning capabilities. Claim verification with gold rationales mainly relies on the reasoning capability of the models. Table 3 shows that most models have improved their claim verification performance with generated rationales compared to the claim-only setup. On FEVER, the separate rationale generation step has improved the average F1 score of all evaluated LLMs by 4.05 percentage points, while HOVER on average 5.12 percentage points. Compared to FEVER, the performance gains on HOVER are bigger for several top models, e.g., Llama2-70B, GPT-4, etc. As shown in Table 5, our top models with generated rationales can outperform ProgramFC's closed-book verification on average.

The rationale generation example in Figure 4 is an example of showing performance improvement with generated rationales. Initially, the claim is refuted by all three models under the claim-only setup. With generated rationales, GPT-4 and Llama2-70B have verified the claim correctly. One hop information in the Llama2-70B generation is missing: Carpentaria is a plant genus. However, with the insufficient generated rationales, the model can still verify the claim correctly according to its

[4] ProgramFC is evaluated on all dev data of HOVER. Ours is evaluated on random samples from dev data, 100 samples for each hop. The dev data is balanced between two labels. Therefore, we think our results are representative of the dev data.

	FEVER				HOVER			
Model	Gold	Generated	Claim	Δ	Gold	Generated	Claim	Δ
BLOOM-176B	85.89	74.44	70.99	3.45	57.13	47.27	45.52	1.75
BLOOMZ-176B	87.99	69.33	43.78	**25.55**	64.34	54.63	33.33	**21.30**
Falcon-180B	93.64	88.47	87.69	0.78	56.96	56.90	56.13	0.77
Falcon-180B-Chat	**95.66**	89.45	86.57	2.88	68.78	58.96	57.00	1.96
Llama2-7B	84.14	74.16	77.78	-3.62	53.08	49.77	49.94	-0.17
Llama2-7B-Chat	86.01	75.13	71.06	4.07	61.60	56.19	54.29	1.90
Llama2-13B	93.33	82.25	84.64	-2.39	52.90	58.82	49.10	9.72
Llama2-13B-Chat	94.67	82.47	77.61	4.86	69.93	57.99	57.24	0.75
Llama2-70B	95.33	88.00	86.36	1.64	72.11	62.18	54.74	7.44
Llama2-70B-Chat	95.65	87.34	80.73	6.61	72.66	60.09	55.33	4.76
GPT-4	94.95	**92.89**	**92.17**	0.72	**77.98**	**70.29**	**64.13**	6.16
Fine-tuning	95.20	-	-	-	89.00	-	-	-

Table 3: Claim verification results in weighted F1 scores under three setups, given gold rationales as premise, given generated rationales as premise, and claim-only. Δ represents the verification performance improvement with generated rationales compared to the claim-only setup. The numbers in bold indicate the best generative performance. The underlined numbers indicate the best performance among open-source generative LLMs.

Model	2-hop	3-hop	4-hop
Llama2-70B	80.96	68.75	66.43
Llama2-70B-Chat	77.99	70.00	69.89
GPT-4	81.99	77.96	73.96
Fine-tuning	90.99	88.00	88.00
ProgramFC	75.65	68.48	66.75

Table 4: Claim verification results in weighted F1 scores with gold rationales.

Model	2-hop	3-hop	4-hop
Llama2-70B	71.99	63.87	50.40
Llama2-70B-Chat	67.36	62.55	50.40
GPT-4	67.88	73.96	69.00
ProgramFC	54.27	54.18	52.88

Table 5: Claim verification results in weighted F1 scores with generated rationales.

implicit internal knowledge. This kind of implicit knowledge can be a double-edged sword since the outdated or nonfactual knowledge in generative LLMs can negatively impact claim verification.

5 Conclusion

In this paper, we evaluated the knowledge and reasoning capabilities of current generative LLMs with two Wikipedia-based fact-checking datasets. We selected three representative open-source LLM families and measured their performance against GPT-4. For each LLM family, we include original pre-trained and instruction-tuned models. As a standard fact-checking pipeline, we first generated the rationales and then verified the claims. With few-shot learning, the generative LLMs can infer the concept of rationale generation and claim verification tasks. The performance gap between open-book retrieval and closed-book generation is still large, especially for the complex dataset. As claims become complex, it becomes very challenging for LLMs to generate all necessary rationales without factual errors. Minor factual errors often exist in the generated texts. Instruction-tuned models are more likely to hallucinate than the original pre-trained models. In the future, tuning pretrained LLMs without sacrificing factuality will be an interesting research topic. By claim verification, we mainly evaluated the reasoning capability of LLMs with three setups: given gold rationales, generated rationales, and only claims. Given a limited number of facts in the premise, current top generative LLMs can verify the claims very reliably. As the number of facts increases, generative LLMs can't pay attention to every fact and ignore minor factual errors. Most evaluated LLMs have gained performance improvement with generated rationales compared to claim-only verification. The extra generation step has boosted LLMs' verification performance. This suggests the decomposition of complex claims into simpler ones, i.e., more fine-granulated fact-checking, can be a further research direction.

Limitations

The datasets we selected for evaluation only consider Wikipedia-based datasets, which are quite limited. There are other types of fact-checking datasets, e.g., SciFact (Wadden et al., 2020), which is based on abstracts of scientific papers and may also be part of LLM training data. However, we find evaluating generated scientific rationales quite challenging, especially regarding hallucination. Currently, we are not able to detect every factual error in the generated texts, even with Wikipedia-based datasets, since the training data of generative LLMs is beyond Wikipedia corpus and Google search engine. The information generated by LLMs, which can not be found with Google, is not necessarily unfactual.

We have not designed separate prompts for each model family. The prompts used in this paper are first tested with the Llama2 family and further applied to other models. We can imagine that there can be performance improvements of other model families (BLOOM, Falcon) when we customize prompts for them.

Ethical Consideration

Instruction-tuned LLMs have a better understanding of human instructions. Meanwhile, our paper shows that instruction-tuned models are much easier to hallucinate. Given nonfactual claims, instruction-tuned LLMs can generate fluent and convincing rationales. Fact-checking these generated rationales is difficult and time-consuming. We can predict that fact-checking LLM-generated texts will be very challenging in the future.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, and et al. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Miguel Arana-Catania, Elena Kochkina, Arkaitz Zubiaga, Maria Liakata, Robert Procter, and Yulan He. 2022. Natural language inference with self-attention for veracity assessment of pandemic claims. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, United States. Association for Computational Linguistics.

Joris Baan, Nico Daheim, Evgenia Ilia1, Dennis Ulmer, Haau-Sing Li, Raquel Fernández, Barbara Plank, Rico Sennrich, Chrysoula Zerva, and Wilker Aziz. 2023. Uncertainty in natural language generation: From theory to applications.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.*

Mitchell DeHaven and Stephen Scott. 2023. BEVERS: A general, simple, and performant framework for automatic fact verification. In *Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER)*, pages 58–65, Dubrovnik, Croatia. Association for Computational Linguistics.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Llm.int8(): 8-bit matrix multiplication for transformers at scale. In *36th Conference on Neural Information Processing Systems (NeurIPS 2022).*

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, pages 103–166.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. UKP-athene: Multi-sentence textual entailment for claim verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 103–108, Brussels, Belgium. Association for Computational Linguistics.

Casper Hansen, Christian Hansen, and Lucas Chaves Lima. 2021. Automatic fake news detection: Are models learning to reason? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers)*, pages 80–86.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Boyd Adriane. 2020. spacy: Industrial-strength natural language processing in python.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey.

In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *International Conference on Learning Representations*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. Hover: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Omar Khattab, Christopher Potts, and Matei Zaharia. 2021. Baleen: Robust multi-hop reasoning at scale via condensed retrieval. In *Advances in Neural Information Processing Systems*.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA. Association for Computing Machinery.

Nayeon Lee, Belinda Z. Li, Sinong Wang, Wen-tau Yih, Hao Ma, and Madian Khabsa. 2020. Language models as fact checkers? In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, pages 36–41, Online. Association for Computational Linguistics.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. In *Advances in Neural Information Processing Systems*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Alejandro Martín, Javier Huertas-Tato, Álvaro Huertas-García, Guillermo Villar-Rodríguez, and David Camacho. 2022. Facter-check: Semi-automated fact-checking through semantic similarity and natural language inference. *Know.-Based Syst.*, 251(C).

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. MetaICL: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809, Seattle, United States. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and et al. 2022. Training language models to follow instructions with human feedback. In *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004, Toronto, Canada. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, and et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.*, 19(1):22–36.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803.

Amir Soleimani, Christof Monz, and Marcel Worring. 2020. Bert for evidence retrieval and claim verification. In *Advances in Information Retrieval*, pages 359–366, Cham. Springer International Publishing.

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2023. Evaluating the factual consistency of large language models through news summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5220–5255, Toronto, Canada. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and et al. 2023. Llama 2: Open foundation and fine-tuned chat models.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, and et al. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, and Quoc V. Le Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (LongPapers)*, pages 1112–1122. Association for Computational Linguistics.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language modelss. In *Proceedings of the 38th International Conference on Machine Learning*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

A Implementation Details

A.1 LLM Generation

The experiments with open-source LLMs are conducted on a node with 8 NVIDIA-A100-40GB GPUs. Due to memory limitations, we use the quantized 8-bit versions for the Falcon and BLOOM models. According to Dettmers et al. (2022), the inference performance degradation for models with 8-bit quantization is very limited. The experiments with GPT-4 are conducted with the OpenAI API [5], queried at the end of November 2023.

We use the transformers library[6] for generating rationales. Since the goal is to generate relevant facts for the claims, we apply the greedy decoding strategy, which selects the next token with the highest probability. We limit the maximum number of new tokens for each generation to 400. As shown in the rationale template, each example in the prompt ends with "###\n". We utilize the StoppingCriteria class of the transformers library to improve the generation efficiency. As soon as the model generates tokens for newline break "\n", it will stop generation.

A.2 Open-book Retrieval

FEVER The document retrieval has used the document retriever from (Hanselowski et al., 2018)

[5] https://openai.com/blog/openai-api

[6] https://github.com/huggingface/transformers

with 7 search results. From the retrieved documents, we select top-3 sentences based on the classification probabilities of the fine-tuned cross-encoder. Following (Soleimani et al., 2020), we fine-tune the ROBERTA-BASE model with hard negative mining to classify whether the sentences in retrieved documents are relevant to the claim. Claim and candidate sentences are concatenated as input for the cross-encoder. Since there are more negative (irrelevant) sentences than positive ones in the retrieved documents, the imbalance issue exists in the training data. Given batch size n, online hard negative mining selects n negative samples with the highest loss values. We train the cross-encoder 10 epochs and select the epoch with the highest verification accuracy on the dev data. Our training hyper-parameters are batch size 64, learning rata 1e-05, warm-up proportion 0.1, and linear scheduler with weight decay 0.01.

HOVER Baleen is a state-of-the-art retrieval model with a condensed retrieval architecture for multi-hop retrieval (Khattab et al., 2021). We download checkpoints from Baleen's GitHub site[7]. With the checkpoints, we query the relevant sentences for our test samples.

A.3 Fine-tuning for Claim Verification

We use fine-tuned ROBERT-LARGE-MNLI as our base model and fine-tune further with the training data from FEVER and HOVER separately. For FEVER, there are 3 classes of verification results, SUPPORTS, NEI, and REFUTES, as the original setup. For HOVER, the classification head is adapted to 2 classes, SUPPORTS and REFUTES. Our hyper-parameters for the fine-tuning are batch size 64, learning rate 1e-05, warm-up proportion 0.1, and linear scheduler with weight decay 0.01.

B Generation Examples

B.1 Rationale Generation

Figure 6 shows two further examples of rationale generation, where errors exist in the generated texts.

B.2 Claim Verification with Gold Rationales

Figure 7 shows three wrong verification examples of top generative models, where minor changes in the claim have not been detected.

C Corrected Results

We show in Table 6 the corrected results of claim verification. Since the correction of the datasets' original labels is subjective, we don't report the corrected results in the main part of the paper.

D Prompt

D.1 Prompt for Rationale Generation

Concretely, for the FEVER dataset, we show 2 examples of claims with 1 rationales, 2 with 2 rationales, and 2 with 3 rationales. In each 2 examples, one claim is supported by the gold rationales, and the other is refuted by the gold rationales. The HOVER dataset is treated accordingly: The examples are selected according to the number of hops, 2 examples for 2 hops, 2 examples for 3 hops, and 2 examples for 4 hops. Due to the multi-hop characteristics of HOVER, we sort the rationales in a more intuitive and logical order instead of a random order. Figure 8 shows the 6 examples for generating rationales on FEVER, and Figure 9 on HOVER.

D.2 Prompt for Claim Verification

We have three setups for claim verification: gold rationales, generated rationales, and claim-only verification. For each claim, we run three predictions and use majority voting to determine the final prediction. The prompts demonstrated here are one of the three seeds. Figure 10 and Figure 11 show the examples used for claim verification with rationales, both gold rationales and generated rationales. Figure 12 and Figure 13 show the examples for the claim-only verification.

[7] https://github.com/stanford-futuredata/Baleen

Claim: Fidel Castro transferred his responsibilities to his brother. (FEVER)
Gold rationales: After his brother Fidel Castro took power, Raúl Castro was one of the most important figures in the party, serving as Minister of the Armed Forces for 49 years, from 1959 to 2008 making him the longest serving minister of the armed forces. [Fidel Castro] In 2006 he transferred his responsibilities to Vice-President Raúl Castro, who was elected to the presidency by the National Assembly in 2008.
GPT-4: On July 31, 2006, Castro, after undergoing intestinal surgery for diverticulitis, transferred his responsibilities to the First Vice-President, his younger brother Raúl Castro.
Llama2-70B: On 31 July 2006, Fidel Castro, due to health complications, transferred his responsibilities to Vice-President Raúl Castro.
Llama2-70B-Chat: On July 31, 2006, Castro announced that he was transferring his responsibilities to his brother Raul Castro, who was officially named his successor.

Claim: Mohammad Akram, is a former Pakistani cricketer (now holding British citizenship) who bowled right arm fast-medium. He is head coach of the undefeated team with the home ground of Arbab Niaz Stadium. (HOVER)
Gold rationales: [Peshawar Zalmi] Darren Sammy is the current captain of the team as of October 2016 and Mohammad Akram is the head coach. Mohammad Akram (born 10 September 1974) is a former Pakistani cricketer (now holding British citizenship) who bowled right arm fast-medium. [Arbab Niaz Stadium] It is the home ground for Peshawar Zalmi.
GPT-4: Mohammad Akram (born 22 June 1974) is a Pakistani cricket coach and former first-class cricketer who played for Pakistan as a right-arm fast-medium bowler. He is now a British citizen. [Mohammad Akram] He was appointed as the head coach of Peshawar Zalmi in 2016. [Peshawar Zalmi] Peshawar Zalmi is a Pakistani franchise T20 cricket team which plays in the Pakistan Super League and represents Peshawar. The team's home ground is Arbab Niaz Stadium. The team's performance record would be needed to verify if they were undefeated during his tenure as head coach.
Llama2-70B: Mohammad Akram (born 12 March 1974) is a former Pakistani cricketer (now holding British citizenship) who bowled right arm fast-medium. He is the head coach of the Peshawar Zalmi team in the Pakistan Super League. [Peshawar Zalmi] The team plays its home games at Arbab Niaz Stadium.
Llama2-70B-Chat: Mohammad Akram (born 13 June 1970) is a former Pakistani cricketer who bowled right-arm fast-medium. He is the head coach of the Multan Sultans, a franchise cricket team that plays in the Pakistan Super League and has its home ground at the Multan Cricket Stadium (previously known as the Arbab Niaz Stadium).

Figure 6: Example for rationales generation for claims in FEVER and HOVER. Important entities are marked in blue. Hallucinations are marked in red.

Claim: Poor old Lu is a band that is considered "one of the most accomplished and creative Christian bands of the '90s". The rap band who created "Pony Express Record" is not.
Goldd Rationales: [Poor Old Lu] The "Encyclopedia of Contemporary Christian Music" calls the band "One of the most accomplished and creative Christian bands of the '90s". Shudder to Think was an American indie rock group. Pony Express Record is a 1994 album by the Washington, D.C.-based post-hardcore group Shudder to Think.
Claim: Elizabeth Appleton is a 1963 novel. The autor of this novel and the person who mentioned Hason Raja in his lectures at the Mahatma Gandhi Institute are not the same nationality.
Gold rationales: John Henry O'Hara (January 31, 1905 – April 11, 1970) was an American writer, best remembered as a keen observer of social status and manners in early to mid 20th century America and pre-eminent among his contemporaries at depicting social realism. Rabindranath Tagore FRAS (;]), also written Ravīndranātha Thākura (7 May 1861 – 7 August 1941), sobriquet Gurudev, was a Bengali polymath who reshaped Bengali literature and music, as well as Indian art with Contextual Modernism in the late 19th and early 20th centuries. Elizabeth Appleton is a novel by John O'Hara first published in 1963. [Hason Raja] He gained international recognition few years after his death, when Nobel laureate, Rabindranath Tagore, mentioned him in his lectures at Oxford University.
Claim: This magazines was co-founded by Joseph J. Thorndike. The magazine and the Knapp Communications magazine founded by Paige Rense are both published in America.
Gold Rationales: American Heritage is a magazine dedicated to covering the history of the United States of America for a mainstream readership. Bon Appétit is an American food and entertaining magazine published monthly by Condé Nast. [Joseph J. Thorndike] He was Managing Editor of "Life" for three years in the late 1940s, and a co-founder of "American Heritage" and "Horizon" magazines. [Paige Rense] Rense founded the cookery magazine "Bon Appétit", was editor in chief of "GEO", and is the author of a mystery novel, "Manor House" (Doubleday, 1997).

Figure 7: Wrong verification examples by generative LLMs given gold rationales, which the fine-tuning model verifies correctly.

To verify the factuality of the claim <Jack Falahee was born in March.>, following factual evidence is needed: Jack Ryan Falahee (born February 20 , 1989) is an American actor .###

To verify the factuality of the claim <Sayyeshaa acts only on stage.>, following factual evidence is needed: Sayyeshaa is an Indian film actress who appears in Hindi , Tamil and Telugu films . After starring a Telugu film Akhil (2015) , she made her Bollywood debut in Ajay Devgn 's Shivaay (2016) .###

To verify the factuality of the claim <You Only Live Twice had a male British director.>, following factual evidence is needed: You Only Live Twice is the first Bond film to be directed by Lewis Gilbert , who later directed the 1977 film The Spy Who Loved Me and the 1979 film Moonraker , both starring Roger Moore . Lewis Gilbert , (born 6 March 1920) is a British film director , producer and screenwriter , who has directed more than 40 films during six decades ; among them such varied titles as Reach for the Sky (1956) , Sink the Bismarck !###

To verify the factuality of the claim <Raja Hindustani is only a 1993 drama romance film.>, following factual evidence is needed: Raja Hindustani (translation : Indian King) is a 1996 Indian blockbuster Hindi-language drama romance film directed by Dharmesh Darshan . Released on 15 November 1996 , it is a remake of the 1965 Hindi film Jab Jab Phool Khile starring Shashi Kapoor and Nanda . The film was remade in Kannada as Naanu Naane in 2002 starring Upendra and Sakshi Shivanand in lead roles .###

To verify the factuality of the claim <Off the Wall led to someone winning an award.>, following factual evidence is needed: [Off the Wall] The record gained critical acclaim and recognition , and won the singer his first Grammy Award . Jackson received positive reviews for his vocal performance on the record . Jackson wrote three of the songs himself , including the number-one Grammy Award-winning single " Do n't Stop 'Til You Get Enough " .###

To verify the factuality of the claim <Pink is an actress.>, following factual evidence is needed: Alecia Beth Moore (born September 8 , 1979) , known professionally as Pink (frequently stylized as) , is an American singer , songwriter , dancer , and actress .###

Figure 8: Example prompt for FEVER rationale generation.

To verify the factuality of the claim <Henry Tudor was the mother of the man who was King of England from 1483 to 1485.>, following factual evidence is needed: Richard III (2 October 1452 – 22 August 1485) was King of England from 1483 until his death in 1485, at the age of 32, in the Battle of Bosworth Field. Cecily Neville, the mother of the Kings Edward IV and Richard III, was born here.###

To verify the factuality of the claim <The Managing Director of Escorts Group is the son of the English insurance advisor in life insurance.>, following factual evidence is needed: Escorts Group's management team includes Rajan Nanda as the Chairman and Managing Director and Nikhil Nanda as the Managing Director. Part of the Kapoor family, he is the son of insurance agent Ritu Nanda and industrialist Rajan Nanda, and the grandson of actorfilmmaker Raj Kapoor. Ritu Nanda (born Ritu Kapoor; 30 October 1948) is a prominent insurance advisor associated chiefly with the life insurance business.###

To verify the factuality of the claim <This American bass player joined Steve Vai on the Where the Wild Things Are album. He is known for his work in a virtual death metal band featured in "Metalocalypse".>, following factual evidence is needed: Steve Vai is joined on stage by Alex DePue (violin and keyboards), Ann Marie Calhoun (violin and keyboards), Bryan Beller (bass), Jeremy Colson (drums), Dave Weiner (guitar and sitar) and Zack Wiesinger (lap steel). Bryan Beller (born May 6, 1971) is an American bass guitarist known for his work with Joe Satriani, The Aristocrats, Dethklok, Mike Keneally, Steve Vai, James LaBrie of Dream Theater and Dweezil Zappa. Dethklok is a virtual death metal band featured in the Adult Swim animated television series "Metalocalypse".###

To verify the factuality of the claim <A battle came before the fight where Agat Invasion Beach was a landing site. Hoffman Farm was used as a hospital and airfield during this battle.>, following factual evidence is needed: The beaches of Agat were one of the landing sites of American forces in the 1944 Battle of Guam, in which the island was retaken from occupying Japanese forces. The Second Battle of Guam (21 July – 10 August 1944) was the American recapture of the Japanese-held island of Guam, a U.S. territory in the Mariana Islands captured by the Japanese from the U.S. in the 1941 First Battle of Guam during the Pacific campaign of World War II. [Hoffman Farm] The farm buildings were used as a hospital during the American Civil War in Battle of Antietam from the day of the battle on September 17, 1862, and through the following month. The Battle of Antietam , also known as the Battle of Sharpsburg, particularly in the South, was fought on September 17, 1862, near Sharpsburg, Maryland and Antietam Creek as part of the Maryland Campaign.###

To verify the factuality of the claim <A company manages the maximum-security penitentiary where Marco Allen Chapman was executed. This company is headquartered along the Kentucky River.>, following factual evidence is needed: Thirty-seven-year-old Marco Allen Chapman was executed on November 21, 2008 at 8:34 p.m. EST on a Friday by lethal injection in a special chamber at the Kentucky State Penitentiary in Eddyville, Kentucky. [Kentucky State Penitentiary] It is managed by the Kentucky Department of Corrections. [Kentucky Department of Corrections] The agency is headquartered in the Health Services Building in Frankfort. Located along the Kentucky River, Frankfort is the principal city of the Frankfort, Kentucky Micropolitan Statistical Area, which includes all of Franklin and Anderson counties.###

To verify the factuality of the claim <The author of Anastasia on Her Own won the 2002 Rhode Island Children's Book Award.>, following factual evidence is needed: Anastasia on Her Own (1985) is a young-adult novel by Lois Lowry. [Lois Lowry] Her book "Gooney Bird Greene" won the 2002 Rhode Island Children's Book Award.###

Figure 9: Example prompt for HOVER rationale generation.

	FEVER				HOVER			
Model	Gold	Generated	Claim	Δ	Gold	Generated	Claim	Δ
BLOOM-176B	87.57	76.13	72.40	3.73	57.83	48.68	45.48	3.20
BLOOMZ-176B	88.32	69.00	44.91	**24.09**	66.41	53.99	36.34	**17.65**
Falcon-180B	95.32	89.47	87.70	1.77	58.51	56.93	56.18	0.75
Falcon-180B-Chat	**97.66**	90.45	86.58	3.87	70.84	58.99	57.69	1.30
Llama2-7B	84.48	75.18	78.78	-3.60	55.21	49.77	49.31	0.46
Llama2-7B-Chat	86.35	76.14	72.81	3.33	61.62	55.52	54.32	1.20
Llama2-13B	93.67	83.92	84.30	-0.38	55.80	59.52	50.48	9.04
Llama2-13B-Chat	94.33	83.47	78.63	4.84	69.96	57.36	55.27	2.09
Llama2-70B	97.00	90.33	87.37	2.96	72.80	62.88	52.73	10.15
Llama2-70B-Chat	97.66	87.68	81.09	6.59	74.01	59.45	55.36	4.09
GPT-4	96.97	**95.26**	**92.85**	2.41	**80.66**	**70.98**	**65.01**	5.97
Fine-tuning	95.86	-	-	-	88.34	-	-	-

Table 6: Corrected claim verification results in weighted F1 scores under three setups, given gold rationales as premise, given generated rationales as premise, and claim-only. Δ represents the verification performance improvement with generated rationales compared to the claim-only setup. The numbers in bold indicate the best generative performance. The underlined numbers indicate the best performance among open-source generative LLMs. The average performance improvements with rationale generation on FEVER reaches 4.51 percentage points and on HOVER 5.08.

Given the premise <[Raja Hindustani] Raja Hindustani (translation : Indian King) is a 1996 Indian blockbuster Hindi-language drama romance film directed by Dharmesh Darshan . [Raja Hindustani] Released on 15 November 1996 , it is a remake of the 1965 Hindi film Jab Jab Phool Khile starring Shashi Kapoor and Nanda . [Raja Hindustani] The film was remade in Kannada as Naanu Naane in 2002 starring Upendra and Sakshi Shivanand in lead roles .>, is the hypothesis <Raja Hindustani is only a 1993 drama romance film.> true? No.###
Given the premise <[System of a Down] System of a Down (also known as SOAD or simply System) is an American heavy metal band formed in Glendale, California, in 1994. [System of a Down] Since 1997, the band has consisted of Serj Tankian (lead vocals, keyboards); Daron Malakian (guitar, vocals); Shavo Odadjian (bass, backing vocals); and John Dolmayan (drums), who replaced original drummer Andy Khachaturian. [System of a Down] The band went on hiatus in 2016 and reunited in 2010.>, is the hypothesis <System of a Down briefly disbanded in limbo> true? Not enough information.###
Given the premise <[Off the Wall] The record gained critical acclaim and recognition , and won the singer his first Grammy Award . [Off the Wall] Jackson received positive reviews for his vocal performance on the record . [Off the Wall] Jackson wrote three of the songs himself , including the number-one Grammy Award-winning single “ Do n’t Stop ’Til You Get Enough ” .>, is the hypothesis <Off the Wall led to someone winning an award.> true? Yes.###
Given the premise <[Kushan Empire] The Kushan Empire was a syncretic empire, formed by the Yuezhi, in the Bactrian territories in the early 1st century. [Kushan Empire] It spread to encompass much of what is now Uzbekistan, Afghanistan, Pakistan and Northern India, at least as far as Saketa and Sarnath near Varanasi (Benares), where inscriptions have been found dating to the era of the Kushan Emperor Kanishka the Great. [Kushan Empire] The Kushans were most probably one of five branches of the Yuezhi confederation, an Indo-European nomadic people of possible Tocharian origin, who migrated from northwestern China (Xinjiang and Gansu) and settled in ancient Bactria.>, is the hypothesis <Afghanistan is the source of the Kushan dynasty.> true? Not enough information.###
Given the premise <[Sayyeshaa] Sayyeshaa is an Indian film actress who appears in Hindi , Tamil and Telugu films . [Sayyeshaa] After starring a Telugu film Akhil (2015) , she made her Bollywood debut in Ajay Devgn ’s Shivaay (2016) .>, is the hypothesis <Sayyeshaa acts only on stage.> true? No.###
Given the premise <[You Only Live Twice (film)] You Only Live Twice is the first Bond film to be directed by Lewis Gilbert , who later directed the 1977 film The Spy Who Loved Me and the 1979 film Moonraker , both starring Roger Moore . [Lewis Gilbert] Lewis Gilbert , (born 6 March 1920) is a British film director , producer and screenwriter , who has directed more than 40 films during six decades ; among them such varied titles as Reach for the Sky (1956) , Sink the Bismarck !>, is the hypothesis <You Only Live Twice had a male British director.> true? Yes.###

Figure 10: Example prompt for FEVER claim verification with rationales. For both verifications with gold rationales and generated rationales, the 6 examples are the same. In the test sample, rationals are differentiated between gold and generated rationales.

Given the premise <[Raby Castle] Cecily Neville, the mother of the Kings Edward IV and Richard III, was born here. [Richard III of England] Richard III (2 October 1452 – 22 August 1485) was King of England from 1483 until his death in 1485, at the age of 32, in the Battle of Bosworth Field.>, is the hypothesis <Henry Tudor was the mother of the man who was King of England from 1483 to 1485.> true? No.###

Given the premise <[Nikhil Nanda] Part of the Kapoor family, he is the son of insurance agent Ritu Nanda and industrialist Rajan Nanda, and the grandson of actorfilmmaker Raj Kapoor. [Ritu Nanda] Ritu Nanda (born Ritu Kapoor; 30 October 1948) is a prominent insurance advisor associated chiefly with the life insurance business. [Escorts Group] Escorts Group's management team includes Rajan Nanda as the Chairman and Managing Director and Nikhil Nanda as the Managing Director.>, is the hypothesis <The Managing Director of Escorts Group is the son of the English insurance advisor in life insurance.> true? No.###

Given the premise <[Bryan Beller] Bryan Beller (born May 6, 1971) is an American bass guitarist known for his work with Joe Satriani, The Aristocrats, Dethklok, Mike Keneally, Steve Vai, James LaBrie of Dream Theater and Dweezil Zappa. [Dethklok] Dethklok is a virtual death metal band featured in the Adult Swim animated television series "Metalocalypse". [Where the Wild Things Are (Steve Vai album)] Steve Vai is joined on stage by Alex DePue (violin and keyboards), Ann Marie Calhoun (violin and keyboards), Bryan Beller (bass), Jeremy Colson (drums), Dave Weiner (guitar and sitar) and Zack Wiesinger (lap steel).>, is the hypothesis <This American bass player joined Steve Vai on the Where the Wild Things Are album. He is known for his work in a virtual death metal band featured in "Metalocalypse".> true? Yes.###

Given the premise <[Battle of Guam (1944)] The Second Battle of Guam (21 July – 10 August 1944) was the American recapture of the Japanese-held island of Guam, a U.S. territory in the Mariana Islands captured by the Japanese from the U.S. in the 1941 First Battle of Guam during the Pacific campaign of World War II. [Battle of Antietam] The Battle of Antietam , also known as the Battle of Sharpsburg, particularly in the South, was fought on September 17, 1862, near Sharpsburg, Maryland and Antietam Creek as part of the Maryland Campaign. [Agat Invasion Beach] The beaches of Agat were one of the landing sites of American forces in the 1944 Battle of Guam, in which the island was retaken from occupying Japanese forces. [Hoffman Farm] The farm buildings were used as a hospital during the American Civil War in Battle of Antietam from the day of the battle on September 17, 1862, and through the following month.>, is the hypothesis <A battle came before the fight where Agat Invasion Beach was a landing site. Hoffman Farm was used as a hospital and airfield during this battle.> true? No.###

Given the premise <[Kentucky Department of Corrections] The agency is headquartered in the Health Services Building in Frankfort. [Frankfort, Kentucky] Located along the Kentucky River, Frankfort is the principal city of the Frankfort, Kentucky Micropolitan Statistical Area, which includes all of Franklin and Anderson counties. [Kentucky State Penitentiary] It is managed by the Kentucky Department of Corrections. [Marco Allen Chapman] Thirty-seven-year-old Marco Allen Chapman was executed on November 21, 2008 at 8:34 p.m. EST on a Friday by lethal injection in a special chamber at the Kentucky State Penitentiary in Eddyville, Kentucky.>, is the hypothesis <A company manages the maximum-security penitentiary where Marco Allen Chapman was executed. This company is headquartered along the Kentucky River.> true? Yes.###

Given the premise <[Anastasia on Her Own] Anastasia on Her Own (1985) is a young-adult novel by Lois Lowry. [Lois Lowry] Her book "Gooney Bird Greene" won the 2002 Rhode Island Childrenś Book Award.>, is the hypothesis <The author of Anastasia on Her Own won the 2002 Rhode Island Childrenś Book Award.> true? Yes.###

Figure 11: Example prompt for HOVER claim verification with rationales.

Is the following claim <Raja Hindustani is only a 1993 drama romance film.> true? No.###
Is the following claim <System of a Down briefly disbanded in limbo> true? Not enough information.###
Is the following claim <Off the Wall led to someone winning an award.> true? Yes.###
Is the following claim <Afghanistan is the source of the Kushan dynasty.> true? Not enough information.###
Is the following claim <Sayyeshaa acts only on stage.> true? No.###
Is the following claim <You Only Live Twice had a male British director.> true? Yes.###

Figure 12: Example prompt for FEVER claim-only verification.

Is the following claim <Henry Tudor was the mother of the man who was King of England from 1483 to 1485.> true? No.###

Is the following claim <The Managing Director of Escorts Group is the son of the English insurance advisor in life insurance.> true? No.###

Is the following claim <This American bass player joined Steve Vai on the Where the Wild Things Are album. He is known for his work in a virtual death metal band featured in "Metalocalypse".> true? Yes.###

Is the following claim <A battle came before the fight where Agat Invasion Beach was a landing site. Hoffman Farm was used as a hospital and airfield during this battle.> true? No.###

Is the following claim <A company manages the maximum-security penitentiary where Marco Allen Chapman was executed. This company is headquartered along the Kentucky River.> true? Yes.###

Is the following claim <The author of Anastasia on Her Own won the 2002 Rhode Island Children's Book Award.> true? Yes.###

Figure 13: Example prompt for HOVER claim-only verification.