Language Models Hallucinate, but May Excel at Fact Verification

Anonymous ACL submission

Abstract

Recent progress in natural language processing (NLP) owes much to remarkable advances in large language models (LLMs). Nevertheless, LLMs frequently "hallucinate," resulting in non-factual outputs. Our carefully-designed human evaluation substantiates the serious hallucination issue, revealing that even GPT-3.5 produces factual outputs less than 25% of the time. This underscores the importance of fact verifiers in order to measure and incentivize progress. Our systematic investigation affirms that LLMs can be repurposed as effective fact verifiers with strong correlations with human judgments. Surprisingly, FLAN-T5$_{11B}$, the least factual generator in our study, performs the best as a fact verifier, even outperforming more capable LLMs like GPT3.5 and ChatGPT. Delving deeper, we analyze the reliance of these LLMs on high-quality evidence, as well as their deficiencies in robustness and generalization ability. Our study presents insights for developing trustworthy generation models.

1 Introduction

LLMs have demonstrated remarkable performance across various natural language generation (NLG) tasks (Brown et al., 2020; OpenAI, 2023; Chowdhery et al., 2022). However, they persistently suffer from the *hallucination* problem (Bang et al., 2023), often generating non-factual and sometimes misleading outputs. This is quantitatively substantiated by the first part of this paper. In our carefully designed human evaluation of several current LLMs, GPT-3.5 only manages to produce factual outputs less than 25% of the time; other models perform even worse. Such underperformance is achieved on Wikipedia, a domain that they have been extensively trained in and intuitively "familiar with." Our findings highlight the serious challenge that the hallucination issue presents, and underscore the crucial importance of developing effective fact verification methods (Vlachos and Riedel, 2014). These methods are central to evaluating and incentivizing progress in improving LLMs' factuality.

In the second part of the paper, we explore the prospect of leveraging instruction-tuned LLMs for fact verification. We hypothesize that, despite struggling to generate factual outputs, they may still be able to judge whether a piece of text is factual—a task that intuitively appears easier. Our systematic investigation affirms this hypothesis, especially when LLMs are augmented with retrieval components. Specifically, given a statement to be verified, we retrieve evidence from an external corpus and re-frame the statement and evidence into a prompt to instruct an LLM to judge the factuality. We then normalize the LLM's generation probabilities of pre-defined answers as the factuality score, which shows stronger correlations with human judgments than previous statistical and model-based methods. Extensive experiments further reveal that FLAN-T5$_{11B}$ (Chung et al., 2022), the least factual generator in our study, even surprisingly outperforms GPT3.5 and ChatGPT for fact verification.

We further analyze the LLMs-based fact verifiers from the following perspectives: **(1) Influence of given evidence:** ChatGPT is susceptible to irrelevant evidence but deals with relevant but counterfactual evidence better than FLAN-T5$_{11B}$. **(2) Robustness:** GPT variants are less robust to different prompts than FLAN-T5$_{11B}$; **(3) Generalization ability:** It is more difficult to evaluate sentences that are from larger generators, dependent on the context or involving numerals. Evaluating paragraphs is also challenging, and can be facilitated by aggregating judgments of individual, de-contextualized sentences rather than evaluating them directly. Our contributions are as follows:

I. Well-designed human evaluation affirms the serious challenges that current LLMs frequently hallucinate, even in their familiar Wikipedia domain.

II. We explore the potential of LLMs to assess factuality on multiple domains and analyze their

reliance on given evidence, robustness, and generalization ability. These findings may inspire the development of trustworthy generation models and fact verification methods in future research. The evaluation suite also serves as a new comprehensive benchmark for hallucination evaluation[1].

III. Based on our study, we recommend the following practices for fact verification: minimizing irrelevant evidence, taking sentences as base units for long paragraph verification, and de-contextualizing context-dependent sentences before verification.

2 Related Work

Hallucination Many metrics have been proposed to measure hallucinations for directed generation tasks such as summarization, including statistical and model-based metrics. Statistical metrics focus on lexical input-output matching (Dhingra et al., 2019; Wang et al., 2020b; Shuster et al., 2021). Model-based ones further capture semantic-level variations, including unsupervised metrics based on information extraction (IE) (Nan et al., 2021), question answering (QA) (Wang et al., 2020a) and natural language inference (NLI) (Laban et al., 2022), and supervised or semi-supervised metrics trained on specific datasets of evaluation-related tasks (Izacard and Grave, 2021a; Kryściński et al., 2020). These metrics can potentially adapt to open-ended generation by measuring mismatching between outputs and retrieved evidence. One additional challenge lies in retrievers possibly producing noisy, redundant, or contradictory evidence.

Fact Verification Lots of datasets have been collected towards fact verification in various domains, e.g., politics (Vlachos and Riedel, 2014), encyclopedia (Thorne et al., 2018, 2021; Eisenschlos et al., 2021), news (Pérez-Rosas et al., 2018), climate (Diggelmann et al., 2020), science (Wadden et al., 2020), and healthcare (Kotonya and Toni, 2020). Honovich et al. (2022) aggregated multiple datasets to assess the ability to measure input-output consistency. Statements in all above datasets usually contain only single sentences and are crafted by either crawling from dedicated websites (Vlachos and Riedel, 2014), manually mutating sentences from factual articles (Thorne et al., 2018) or re-framing QA pairs (Thorne et al., 2021). We further involve model-generated statements and paragraph-level evaluation in our study.

[1]All the data and evaluation scripts will be made public.

LLMs as Evaluators There are many active efforts to use LLMs' generated answers or generation probabilities for NLG evaluation (Yuan et al., 2021; Colombo et al., 2022; Ke et al., 2022). More recent studies show high correlations of ChatGPT with human judgments for evaluating summarization, story generation, etc. (Wang et al., 2023; Luo et al., 2023; Li et al., 2023). SELFCHECKGPT (Manakul et al., 2023) judged the factuality of a model output based on its similarity with other sampled outputs from the same model, and does not apply to non-model-generated statements or model-agnostic generation. FACTSCORE (Min et al., 2023) used LLMs to evaluate people's biography generation through the fraction of atomic facts supported by retrieved evidence. In contrast, we focus on open-ended generation around various entities and analyze LLMs' robustness and generalization ability.

3 Quantifying LLMs' Hallucination

Our first research question is

> *To what extent do current LLMs hallucinate?*

We quantify this through human evaluation, with a specific focus on the Wikipedia domain, which serves as a reliable information source for annotators. While clean and credible resources exist for other domains such as science and finance, it remains challenging for individuals lacking the expertise to evaluate LLMs' outputs in these domains.

Generation Models We consider four representative LLMs spanning various scales, including FLAN-T5$_{11B}$ (Chung et al., 2022), LLama$_{30B}$, LLama$_{65B}$ (Touvron et al., 2023) and GPT3.5[2].

Generation Tasks We design two open-ended generation tasks that simulate realistic interactions between practitioners and LLMs: **(1) Sentence Completion (SENTCOM):** completing a sentence following the first two tokens of a factual claim from the test set of FEVER (Thorne et al., 2018), a fact verification dataset. The training set provides abundant factual and non-factual claims, enabling us to assess whether supervised verifiers can generalize to model outputs in Section 4. **(2) Wiki Paragraph Generation (PARAGEN)**: generating a paragraph of five sentences about a given entity

[2]In this work, InstructGPT/GPT3.5/ChatGPT/ refers to the OpenAI's API "text-davinci-002"/"text-davinci-003"/"gpt-3.5-turbo-0301," respectively.

SENTCOM	PARAGEN
Please complete the sentence following the given beginning:	Please answer the following questions:
Beginning: Swedish Empire Continuation: was ruled by Gustavus Adolphus from 1611 to 1632 . · · · Beginning: {input} Continuation:	Question: Please write five sentences about facts of "Fire and Darkness" Answer: Fire and Darkness is a cancelled three-dimensional real-time strategy video game developed by Singularity Software. The game consists of a player controlling one of two factions · · · · · · Question: Please write five sentences about facts of "{input}" Answer:

Table 1: Prompts for collecting model outputs. {Input} is two tokens/an entity for SENTCOM/PARAGEN.

Outputs	Models	#	Proportion (%)		
			Factual	Unfactual	NE
SENTCOM	**FLAN-T5$_{11B}$**	48	33.3	50.0	16.7
	Llama$_{30B}$	49	75.5	14.3	10.2
	Llama$_{65B}$	47	68.1	19.2	12.8
	GPT3.5$_{175B}$	49	**89.8**	6.1	4.1
PARAGEN (Sent)	**FLAN-T5$_{11B}$**	147	10.2	58.5	31.3
	LLama$_{30B}$	143	29.4	41.3	29.4
	LLama$_{65B}$	139	33.1	36.0	30.9
	GPT3.5$_{175B}$	139	**37.4**	37.4	25.2
PARAGEN (Para)	**FLAN-T5$_{11B}$**	25	0.0	92.0	8.0
	Llama$_{30B}$	25	4.0	80.0	16.0
	Llama$_{65B}$	21	9.5	85.7	4.8
	GPT3.5$_{175B}$	22	**22.7**	68.2	9.1

Table 2: Statistics of model outputs. **Sent/Para:** Sentence/Paragraph; **NE:** No Evidence. **#**: the number of annotated instances. **Bold**/Underlined percentages indicate the **most**/second most factual outputs.

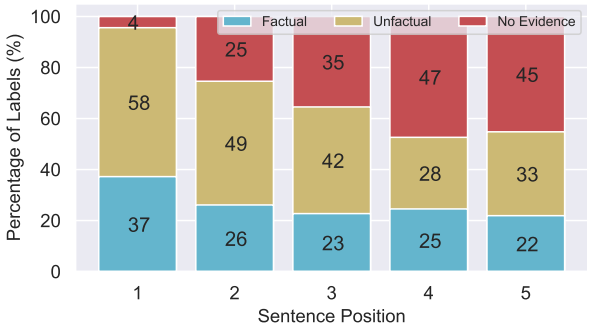

Figure 1: Distribution of different labels arcross sentence positions in PARAGEN (Para).

from Wikipedia. Here the outputs are expected to be longer, and the setting focuses on more long-tailed topics than (1).

We generate 50 outputs for two tasks, respectively, using four generation models, leading to $50 \times 2 \times 4 = 400$ statements totally. During generation, we provide five manually selected factual demonstrations to the models, as shown in Tab. 1.

Human Annotation We collect human workers' factuality judgments of LLMs' outputs through Amazon Mechanical Turk (AMT). Each HIT (human intelligence task) contains five statements with the same input—four generated and one gold. Three workers are hired to search for evidence from Wikipedia and annotate the factuality label of each sentence in the statements, including `factual`, `unfactual` and `not sure`[3]. We further instruct workers to choose reasons if they annotate `not sure`, and exclude sentences from our evaluation set that are labeled as `subjective` or `hard to understand` by at least one worker since such sentences can be ambiguous to determine the factuality (Guo et al., 2022). We discard low-quality submissions using well-designed rules and ensure each sentence has three valid annotations. The Fleiss's kappa score (Fleiss and Joseph, 1971) is 0.91/0.74 for SENTCOM/PARAGEN, indicating a substantial inter-annotator agreement. Finally, we use majority voting to obtain sentence-level human judgments. As for paragraph-level judgments for PARAGEN, we first truncate each paragraph to ensure all sentences are not subjective or hard to understand, and then label the paragraph `unfactual` if *any* of the sentences are labeled `unfactual`, `factual` if all sentences are labeled `factual`, and `not sure` otherwise. A paragraph contains 4.3 sentences on average after truncation. Since the only valid reason for `not sure` is `no evidence found`, we call the label `no evidence` onward. Appendix A.2 presents more details.

Tab. 2 summarizes the annotation results: **(1)** Larger models tend to generate more factual outputs. **(2)** GPT3.5, the best-performing generator in this study, yields factual paragraphs less than 25% of the time. **(3)** All models generate notably more factual outputs for SENTCOM than PARAGEN. We conjecture it is because the average frequency of inputs for SENTCOM is ∼335 times more than that of input entities for PARAGEN. We count frequency using the WebText corpus (Radford et al., 2019). **(4)** Fig. 1 shows increasing percentages of `No Evidence` as the generation proceeds to later sentences, which may be caused by irrelevant or spurious information introduced into outputs by the

[3]We only take Wikipedia as the reliable information source since there is much noisy, biased, and non-validated information on the Internet (Shu et al., 2017).

Datasets	WKS			DSS			
	FEVER	BoolQ-FV	FM2	PubMedQA	XsumFaith	SummEval	SciFact
# Examples	1,000	613	1,380	445	853	798	191
# Factual	517	433	681	276	60	719	101
# Unfactal	483	180	699	169	793	79	90
Avg. Len	9.38	9.57	15.34	19.07	25.1	76.53	12.80
Source	Wikipedia Articles	Search-engine Queries	Adversarial Games	PubMed	BBC	CNN/DailyMail	Scientific Papers
Domain	Wikipedia	Wikipedia	Wikipedia	Medicine	News	News	Science

Table 3: Statistics of test sets of three datasets in WKS and four datasets in DSS.

error accumulation inherent in auto-regressive generation (Zhang et al., 2023). Appendix C.6 shows the influence of the context on model generation and fact verification.

The above human evaluation confirms the LLMs' serious hallucination issue, emphasizing the urgent need for effective fact verifiers to measure and incentivize progress in LLMs' factuality. This motivates us to explore LLMs' potential as fact verifiers.

4 Repurposing LLMs as Fact Verifiers

Our second inquiry lies around the question

> *Can LLMs be repurposed as effective fact verifiers?*

We define fact verification as follows: given a statement s and its leading context c, the verifier should give a probability p of s being factual[4]. The context may be absent, and the statement is a sentence or paragraph. Additionally, we retrieve an evidence set, denoted as $E = \{e_1, e_2, \cdots, e_M\}$, from external corpora with the concatenation of c and s as the query. Each piece of evidence is a passage. Next, we first describe the evaluation sets (§4.1), our verification method (§4.2), compared verifiers (§4.3), and then present the experiment results (§4.4).

4.1 Evaluation Sets

We design three evaluation sets across multiple domains and sources: **(1) Wiki-Domain Statements (WKS):** It aggregates three existing fact verification datasets in the Wikipedia domain including FEVER (Petroni et al., 2021), BoolQ-FV (Thorne et al., 2021) and FM2 (Eisenschlos et al., 2021). **(2) Domain-Specific Statements (DSS):** It consists of four existing domain-specific datasets: PubMedQA (Jin et al., 2019), XsumFaith (Maynez et al., 2020), SummEval (Fabbri et al., 2021) and SciFact (Wadden et al., 2020). Statements supported/refuted by gold evidence are labeled `factual`/`unfactual`, and we remove statements without golden evidence since the factuality is unknowable. Tab. 3 summarizes the statistics of WKS and DSS, and Appendix B includes more details. **(3) Model-Generated Statements (MGS):** It includes SENTCOM and PARAGEN statements generated by four LLMs in §3.

[4]We formulate the fact verification task as regression instead of traditional classification because continuous scores are more informative to reflect the nuance of inputs and may give generators fine-grained feedback in future work.

(Task)	Answer the following question:
(Input)	Facts: 1. $\{e_1\}$ 2. $\{e_2\}$ ... M. $\{e_M\}$ Context: $\{c\}$ Statement following the context: $\{s\}$
(Question)	Based on the given facts, is the statement correct? (A) Yes. (B) No. Please answer A or B:

Table 4: An example prompt used to adapt LLMs for fact verification. The prompt may be changed under different settings. For example, when c is empty, we delete "Context: $\{c\}$" and "following the context".

4.2 Verification Method

We transform each input $x = (E, c, s)$ into a prompt, as shown in Tab. 4. The LLM is expected to generate a judgment $W = (w_1, w_2, \cdots, w_T)$ about whether s is factual. We compute the factuality score p by normalizing the LLM's output probabilities of all valid answers (Ke et al., 2022):

$$p = \frac{\sum_{w \in L_A} p_{\text{LLM}}(w \mid x, W_{<t})}{\sum_{w \in L_A \cup L_B} p_{\text{LLM}}(w \mid x, W_{<t})},$$

where p_{LLM} is the LLM's probability distribution over the vocabulary, L_A/L_B is the set of plausible answer words to indicate factual/non-factual, and t is the maximum time step that makes each

Verifiers	FEVER				BoolQ-FV				FM2			
	ECE	ACC	AUR	r	ECE	ACC	AUR	r	ECE	ACC	AUR	r
Constant	*48.3*	*51.7*	*50.0*	*N/A*	*29.4*	*70.6*	*50.0*	*N/A*	*50.6*	*49.4*	*50.0*	*N/A*
Retrieving Evidence from External Corpora												
KF1	51.3	48.3	53.9	9.2	70.3	29.4	44.4	-7.7	48.8	50.6	49.8	-0.6
NLI$_{11B}$	18.3	81.7	83.8	67.4	45.1	54.8	66.8	33.2	34.5	65.5	68.0	38.7
FiD$_{780M}$	**2.9**	**94.6**	**98.2**	**90.5**	13.8	82.7	88.5	62.0	15.5	77.0	85.6	59.1
FLAN-T5$_{11B}$	3.1	93.8	98.2	90.2	**10.2**	85.5	**94.7**	**75.3**	**8.2**	**82.0**	**89.5**	**68.8**
GPT3.5	7.6	91.7	96.6	84.7	17.9	81.7	87.4	61.1	21.0	77.5	82.8	55.9
ChatGPT	8.2	92.8	92.6	84.3	13.1	**87.3**	88.0	70.1	21.4	79.1	78.6	56.7
Not Using any Evidence												
FiD$_{780M}$	**4.2**	77.3	85.7	62.0	**17.4**	64.6	60.8	17.2	21.6	59.4	65.1	26.4
FLAN-T5$_{11B}$	11.1	74.4	**87.0**	62.6	28.6	56.0	65.6	23.6	**15.7**	59.5	66.3	28.2
GPT3.5	25.6	73.8	78.3	52.7	32.8	65.6	67.9	23.2	41.4	57.3	61.5	19.5
ChatGPT	18.1	**82.0**	81.5	**63.9**	27.1	**73.1**	**69.7**	**37.4**	33.5	**66.7**	**66.5**	**34.1**

Table 5: Results on WKS. "Constant" always predicts a factuality score of 1. All instruction-tuned LLMs are under the zero-shot setting. We highlight the best result in **bold** and underline the second best. Note that the results of ChatGPT except for ACC may be underestimated because they are calculated based on generated answers instead of probabilities (details in Appendix C.3). **Takeaway:** In the Wikipedia domain, ChatGPT performs the best when not using any evidence, while FALN-T5$_{11B}$ excels ChatGPT with retrieved evidence.

Verifiers	PubMedQA				XsumFaith				SummEval				SciFact			
	ECE	ACC	AUR	r	ECE	ACC	AUR	r	ECE	ACC	AUR	r	ECE	ACC	AUR	r
FLAN-T5$_{11B}$	**12.6**	78.4	**84.6**	**60.0**	**20.1**	**78.2**	**70.8**	20.0	**3.5**	92.4	**93.2**	**62.0**	**11.1**	83.2	**95.3**	**77.9**
ChatGPT	20.4	**79.6**	76.7	55.7	23.6	76.4	68.1	**21.4**	7.5	**92.5**	71.0	51.4	16.5	**88.0**	86.4	69.5

Table 6: Results on DSS. Both models are under the zero-shot setting and provided with the golden evidence. **Takeaway:** FLAN-T5$_{11B}$ surpasses ChatGPT in terms of most metrics on specific domains.

token in $W_{<t}$ not included in L_A or L_B. We define L_A = {"A", "a", "Yes", "yes", "YES"} and L_B = {"B", "b", "No", "no", "NO"}. If W does not include any valid answer words, we set p to 0.5. In Appendix C.4, we compare other verification methods such as Chain-of-Thought prompting (Wei et al., 2022) and Likert-scale rating.

4.3 Compared Verifiers

We test the following instruction-tuned LLMs for fact verification using the method in §4.2: **(1) FLAN-T5$_{11B}$:** It is tuned from T5$_{11B}$ (Raffel et al., 2020) on 1.8K+ tasks. **(2) GPT3.5:** We approximate p_{LLM} by normalizing the probabilities calculated from top five logits returned by the API. **(3) ChatGPT**: We hard code the score as 1/0 if it returns "A"/"B" using the prompt in Tab. 4 since its output probabilities are unavailable. For all models, the maximum length is set to 4,000 tokens tokenized by the GPT-series BPE tokenizer. Appendix C.3 presents the results of more LLMs.

We compare the LLMs to baselines widely used to measure hallucination: **(1) knowledgeF1 (KF1)**: It measures the average unigram overlap between the statement and each piece of evidence (Shuster et al., 2021). **(2) NLI**: It is an unsupervised verifier computed as the entailment probability between the evidence and statement. We use the public T5$_{11B}$-based NLI model fine-tuned on a mixture of multiple NLI datasets (Honovich et al., 2022). **(3) FiD**: It is a supervised verifier based on a fine-tuned binary factuality classifier (Izacard and Grave, 2021b; Liu et al., 2022) with the statement and evidence as input. We build FiD based on FLAN-T5$_{780M}$ and fine-tune it on three WKS training sets, respectively, to obtain corresponding verifiers. **(4) FACTSCORE (FAS)**: It first automatically splits a statement into multiple atomic facts and then computes the factual precision as the overall score, i.e., the percentage of atomic facts supported by evidence (Min et al., 2023).

Regarding the retrieval components, we employ the Wikipedia dump from Izacard et al. (2022) as the external corpus for MGS and WKS, with each piece of evidence being a passage of 100 words. Ten pieces of evidence are retrieved for each test sample using Contriever (Izacard and Grave, 2021b). Appendix C.2 presents details about retrievers. For DSS, we use the golden evidence provided by the original papers.

4.4 Results

Results on WKS&DSS Taking human judgments of factual/unfactual statements as 1/0, we use the following metrics to evaluate fact verifiers: **(1) Expected Calibration Error (ECE)** (Guo et al., 2017): It estimates to what extent the predicted score can indicate accuracy. Lower ECE mean better calibration. **(2) Accuracy (ACC)**: It is the fraction of examples that are correctly predicted. **(3) Area Under the ROC Curve (AUR)**: It measures the ability to discriminate factual statements from others. **(4) Pearson's Correlation** (r): It measures the correlation between prediction and human judgments. Appendix C.1 shows more details.

As shown in Tab. 5, on WKS, **when using retrieved evidence**, **(1)** KF1 hardly captures any factuality features; **(2)** NLI_{11B} struggles to generalize to fact verification and significantly underperforms FLAN-T5_{11B}; **(3)** Instruction-tuned LLMs achieve strong performance and are comparable to or better than supervised FiD models; And **(4)** FLAN-T5_{11B} outperforms GPT3.5 in both calibration and discrimination ability. **When not using any evidence**, ChatGPT performs the best, possibly attributed to its superiority in memorizing and utilizing knowledge. The overall inferior performance of not using evidence reveals the importance of retrieval components. Furthermore, the results on DSS in Tab. 6 show that FLAN-T5_{11B} surpasses ChatGPT in terms of most metrics, which also indicates the potential of relatively small-scaled models for hallucination evaluation across various domains. Appendix C.3 shows a similar conclusion on a contemporary dataset FactPrompts (Chern et al., 2023).

Results on MGS Besides metrics used on WKS, we also report **Precision (P)**, **Recall (R)**, and **Area Under the Precision-Recall Curve (AUP)** on the `factual` category for SENTCOM and PARAGEN (Sent), considering the label imbalance. We treat human judgments of factual sentences as 1 and others as 0[5], and calculate the human judgment of a paragraph as the faction of factual sentences. We use FiD trained on FEVER for the experiments.

Tab. 7 shows: **(1)** Despite FiD's best performance on FEVER, it underperforms FLAN-T5_{11B} on SENTCOM, indicating limited generalization of supervised models. **(2)** All verifiers exhibit inferior performance on PARAGEN (Sent) compared to SENTCOM, potentially due to less prevalent entities and more contextual dependencies (e.g., coreference). **(3)** The high recall scores of GPT variants in contrast to FLAN-T5_{11B} show that they prefer to predict `factual`. This may account for the greater disparity in precision and AUP between GPT variants and FLAN-T5_{11B} on PARAGEN that contains more non-factual statements than SENTCOM. **(4)** On PARAGEN (Para), averaging sentence-level scores usually yields better correlations than direct verification. This is attributed to a stronger ability to capture subtle details within a paragraph, facilitated by independent retrieval and verification for every sentence. **(5)** FAS shows comparable or inferior performance in contrast to FLAN-T5_{11B}, suggesting that noise introduced during the generation of atomic facts may impact the final performance.

[5]Since neither "No Evidence" nor "Unfactual" is tolerable for users, we do not distinguish the two categories.

Verifiers	ECE	P	R	AUR	AUP	r
Constant	*33.2*	*66.8*	*100.0*	*50.0*	*66.8*	*N/A*
FiD	11.0	79.0	96.1	88.6	93.7	64.2
FLAN-T5_{11B}	**10.5**	**96.5**	84.5	**93.1**	**96.9**	**74.9**
GPT3.5	17.7	85.2	89.2	87.8	90.8	60.8
ChatGPT	15.0	91.7	85.3	84.8	88.0	67.6
$\text{FAS}_{\text{FLAN-T5}}$	11.7	91.7	77.5	89.8	93.7	69.3
FAS_{CHAT}	18.5	80.4	**92.2**	78.8	83.6	55.0

Verifiers	Sent						Para
	ECE	P	R	AUR	AUP	r	r
Constant	*72.7*	*27.3*	*100.0*	*50.0*	*27.3*	*N/A*	*N/A*
FiD	38.9	35.9	92.9	85.9	76.7	46.5	39.1
FLAN-T5_{11B}	**9.2**	76.0	71.6	**88.0**	**80.9**	**66.8**	45.1 / 79.4
GPT3.5	41.4	38.5	**95.5**	87.7	70.6	39.5	48.6 / 59.7
ChatGPT	25.6	51.7	90.3	79.7	49.6	52.6	52.6 / 68.1
$\text{FAS}_{\text{FLAN-T5}}$	10.6	**77.4**	57.4	84.0	72.6	61.9	**79.7**
FAS_{CHAT}	45.5	35.7	85.2	70.7	42.4	29.4	45.3

Table 7: Results on MGS (**Top:** SENTCOM; **Bottom:** PARAGEN) with retrieved evidence. Two scores **A/B** of LLMs on PARAGEN (Para) mean, **A**: directly evaluating a whole paragraph, **B**: averaging corresponding sentence-level scores. $\text{FAS}_{\text{FLAN-T5/CHAT}}$ refers to FAS using FLAN-T5_{11B}/ChatGPT to judge the factuality of atomic facts. **Takeaway:** FLAN-T5_{11B} exhibits the overall best performance; Taking sentences as base units for verification, as opposed to paragraphs or potentially noisy atomic facts, yields better results.

Verifiers	FEVER	FM2	SENTCOM	PARAGEN	
				Sent	Para
FLAN-T5_{780M}	85.8	61.6	75.4	55.6	71.8
FLAN-T5_{3B}	88.6	66.6	**75.8**	65.8	78.6
FLAN-T5_{11B}	**90.2**	**68.8**	74.9	**66.8**	**79.4**

Table 8: Pearson's correlation scores of FLAN-T5 with different model sizes. We calculate factuality scores on PARAGEN (Para) by averaging sentence-level scores here and below.

Evidence	FEVER		FM2	
	F-T5	ChatGPT	F-T5	ChatGPT
None (0)	74.4	**82.0**	59.5	**66.7**
Golden (1)	93.3	**94.4**	87.3	**88.5**
Random (10)	**64.8**	61.5	**56.7**	55.9
Random+Golden (10)	**92.6**	91.5	**85.8**	83.5
BM25 (10)	**87.5**	87.4	69.1	**69.3**
Contriever (10)	**93.8**	92.8	**82.0**	79.1
Adv (1)	9.4	**24.0**	3.4	**28.2**
Adv+Golden (2)	48.1	**58.9**	29.9	**43.7**

Table 9: Accuracy of FLAN-T5$_{11B}$ (F-T5) and ChatGPT with different evidence. The numbers in the parentheses are the total number of evidence passages. **None/-Golden:** null/golden evidence; **Random:** randomly sampled ten passages from Wikipedia; **Adv**: a sentence adversarial with the statement to be verified, constructed by prompting ChatGPT to convert the statement to its antonym/synonym if the statement is factual/unfactual.

Verifiers	Setting	SENTCOM	PARAGEN	
			Sent	Para
FLAN-T5$_{11B}$	ZS	**74.3**$_{0.8}$	**65.6**$_{1.7}$	**77.7**$_{3.0}$
	FS	73.3$_{0.8}$	60.0$_{0.8}$	67.2$_{2.0}$
GPT3.5	ZS	62.9$_{1.5}$	44.0$_{5.7}$	64.6$_{5.7}$
	FS	68.0$_{1.7}$	52.9$_{6.0}$	71.7$_{4.4}$
ChatGPT	ZS	69.6$_{4.8}$	53.6$_{3.0}$	72.4$_{4.4}$
	FS	70.2$_{1.3}$	51.1$_{2.2}$	66.9$_{3.1}$

Table 10: Pearson's correlations averaged across four prompts. The subscript indicates the standard deviation. ZS/FS means the zero/few-shot setting. Under the few-shot setting, we insert five manually selected demonstrations before each testing example.

Influence of Model Sizes Tab. 8 demonstrates a positive correlation between model sizes and the performance of FLAN-T5. Notably, on more challenging datasets, FM2 and PARAGEN, there exists a relatively larger performance gap between FLAN-T5$_{780M}$ and FLAN-T5$_{3B/11B}$.

5 Analysis on LLM-based Verifiers

This section further analyzes the influence of given evidence on LLM-based verifiers (§5.1), as well as their robustness (§5.2) and generalization ability (§5.3). Appendix C.8 also investigates how LLMs' memorization of inputs may influence their judgments. And Appendix C.9 further shows several representative cases that LLMs fail to judge, to provide more insights.

5.1 Influence of Given Evidence

The aforementioned experiments reveal the importance of retrieving external knowledge for fact verification. We further assess how different types of evidence influence performance. Tab. 9 shows: **(1)** Golden evidence is better than retrieved one, both outperforming null and random evidence. **(2)** ChatGPT is more susceptible to irrelevant information in given evidence than FLAN-T5$_{11B}$. For example, ChatGPT is superior with solely golden evidence but underperforms with mixed golden and random evidence. This can potentially elucidate ChatGPT's suboptimal performance when using retrieved evidence that may include irrelevant information. **(3)** ChatGPT performs much better than FLAN-T5$_{11B}$ if given relevant but fake (Adv) or contradictory evidence (Adv+Gold), which may be prevalent cases when retrieving evidence from the Internet (Shu et al., 2017). This reflects ChatGPT's more limited reliance on given evidence and stronger conviction in its internal knowledge. Nevertheless, the overall drop in accuracy compared to using golden evidence confirms the susceptibility of LLMs to false information (Bian et al., 2023).

5.2 Robustness

LLMs are widely observed to be sensitive to synonymous prompts (Jiang et al., 2020). We design three additional prompts by changing the question in Tab. 4 (details in Appendix C.5). Tab. 10 shows the mean and standard deviation on MGS across four prompts. We find (1) FLAN-T5$_{11B}$ and ChatGPT are better under the zero-shot setting, while GPT3.5 is better under the few-shot setting. (2) FLAN-T5$_{11B}$ is more robust to different prompts than GPT3.5 and ChatGPT with lower variance.

5.3 Generalization Ability

It is crucial for fact verifiers to deal with various inputs (Garbacea et al., 2019). We focus on LLMs' generalization to statements generated by different models, depending on the context or not, or involving different types of named entities.

Smaller models struggle to verify outputs from larger models. Fig. 2 shows FLAN-T5$_{11B}$ performs worse at verifying outputs from relatively large models (LLama$_{65B}$ and GPT3.5). ChatGPT is more stable across different generation models.

Verifying context-dependent statements is more challenging and can be enhanced through decontextualization. For each sentence in PARAGEN (Sent), we manually annotate whether it refers to any entities in the context. Tab. 11 shows: **(1)** Performance on context-dependent sentences will notably drop if the context is unseen, indicating the verifiers are indeed utilizing the given context

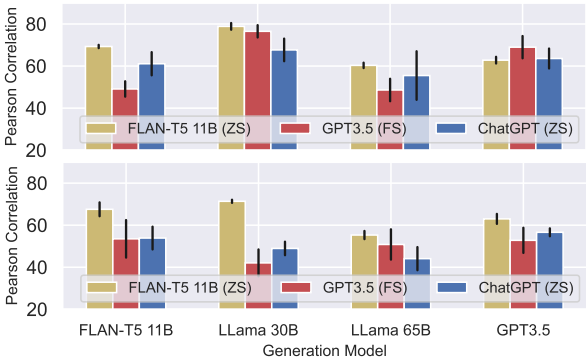

Figure 2: Pearson's correlations when verifying outputs from different generation models. **Top**: SENTCOM; **Bottom**: PARAGEN (Sent). We calculate the standard deviation (black bars) across four instructions in §5.2.

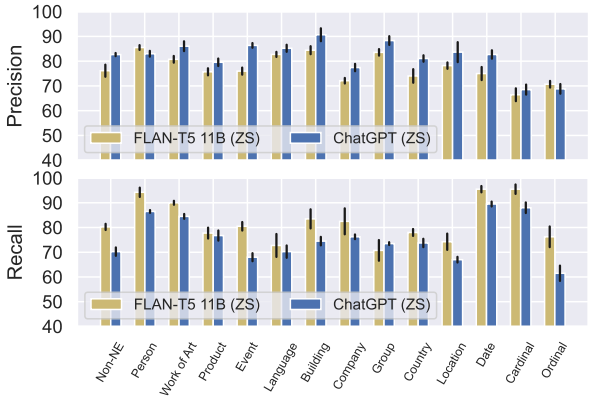

Figure 3: Precision and recall on the `unfactual` category varying with entity types. "Non-NE" means non-named-entity common words (e.g., "computer").

Verifiers	Sent w/ Dependencies			Sent w/o Dependencies			Para
	AUR	AUP	r	AUR	AUP	r	r
FLAN-T5$_{11B}$ (ZS)	87.4	72.6	59.9	88.6	84.7	69.1	79.4
w/o Context	81.0	58.5	49.5	90.2	**85.6**	72.3	82.9
w/ CR	**88.9**	**79.0**	**68.4**	**90.9**	85.2	**72.8**	**86.1**
ChatGPT (ZS)	75.1	38.1	41.1	84.1	62.9	63.4	68.1
w/o Context	72.2	35.0	36.5	84.8	65.9	66.0	74.4
w/ CR	**79.6**	**45.7**	**52.0**	**85.1**	**66.4**	**67.1**	**78.1**

Table 11: Generalization to sentences with or without dependencies on the context. **w/o Context:** not providing the context during verification. **w/ CR:** first performing coreference resolution using ChatGPT to eliminate potential dependencies in a sentence and then verifying it without context.

to judge the factuality. De-contextualization by first performing coreference resolution (CR) can bring substantial improvement. **(2)** Performance on context-independent sentences becomes even better without context, suggesting the context may introduce noise. CR hardly further improves the performance. **(3)** De-contextualization at the sentence level also benefits paragraph-level verification. Appendix C.6 shows more experiment details.

Numeral-involved statements are more difficult to verify. The factuality of a statement mainly manifests in the entities involved and inter-entity relations (Nan et al., 2021). We are curious whether the verifiers perform similarly at verifying statements concerning different types of entities. We experiment on the FaVIQ dataset (Park et al., 2022), which consists of 188k sentences converted from information-seeking QA pairs. The original answer to the question is a word or phrase, and the resulting sentence is factual only if the answer is correct, enabling us to know which part of a sentence is unfactual. We categorize all sentences by the entity types[6] of corresponding answers and randomly sample 100 factual and 100 unfactual sentences in each category to test the verifiers.

Fig. 3 shows: **(1)** LLMs perform better at "Person," "Work of Art," and "Building;" **(2)** The performance is poor at numeral-related types "Cardinal" and "Ordinal." Particularly, LLMs tend to misidentify factual sentences with cardinal numerals as unfactual. The reasons lie in the difficulty of reasoning over inter-numeral logical relations, and retrieving numerals-related evidence. For example, the evidence recall@10 score of the "Cardinal" type is ∼26% lower than that of "Person" (0.43 vs. 0.58). More details are in Appendix C.7.

6 Conclusion

We present a comprehensive study around two research questions: the extent to which current LLMs hallucinate; and their potential as effective fact verifiers. Firstly, we quantitatively affirm the significant hallucination issue of current LLMs through a well-designed human evaluation. This highlights the urgent need for powerful fact verifiers to measure and incentivize progress in LLMs' factuality. To this end, we repurpose the LLMs into fact verifiers, and examine their ability to judge the factuality of model-generated and human-crafted statements. Further analysis shows their heavy reliance on high-quality evidence and discusses their robustness and generalization ability. The evaluation suite and implemented verifiers in this paper can facilitate further research on fact verifiers and trustworthy generation models.

[6]We use spaCy (Honnibal and Montani, 2017) for named entity recognition.

7 Limitations

We summarize our limitations as follows:

I. Regarding the Quantification of Hallucinations

(a) We only focus on quantifying hallucinations for retrieval-free open-ended generation. Although Shuster et al. (2021) showed that retrieval could significantly reduce hallucinations, LLMs are observed to overly depend on retrieved texts for generation (e.g., directly copying from those texts) regardless of fluency and relatedness with input instructions (Liu et al., 2023a), which is out of our scope.

(b) The model outputs are generated under the few-shot setting to ensure that they are as objective and informative as possible, although most generation models are used under the zero-shot setting in reality.

II. Regarding Evaluation Sets

(a) Examples in MGS are limited to the Wikipedia domain due to the difficulty of manually annotating statements involving much professional knowledge in other domains. To mitigate this issue, we include model-generated statements from the news domain (e.g., XsumFaith) in DSS and conduct experiments on ChatGPT-generated statements in the question-answering domain (Appendix C.3). we expect to collect examples in more professional domains (e.g., finance) in future work.

(b) Statements in all three evaluation sets rarely require multi-hop reasoning for judging, considering that outputs of current LLMs are organized as chains of single-hop ones.

III. Regarding Fact Verifiers

(a) Although we endeavor to evaluate LLMs' ability to deal with contradictory or fake evidence in §5.1, the auto-constructed evidence does not occur naturally and may exhibit many artifacts. Contradictory or fake evidence hardly occurs when retrieved from Wikipedia, but it will be much more common if retrieving evidence from the Internet.

(b) We do not perfectly answer why FLAN-T5 is less susceptible to irrelevant information in retrieved evidence than ChatGPT due to the lack of knowledge about the implementation details of OpenAI's GPT series. We conjecture that the reasons possibly lie in that the instruction-tuning data of FLAN-T5 models have included lots of noisy inputs, leading to better robustness to irrelevant information. In contrast, ChatGPT's better knowledge memorization ability may make it less influenced by the relevant but fake information in the evidence. The causal factors underlying these results remain to be investigated in the future work.

8 Ethics Statements

Our experiment results are based on existing public resources (datasets, model checkpoints, and codes). We use widely adopted settings for model generation and evaluation, making our analysis easily replicated. We resorted to Amazon Mechanical Turk (AMT) for human evaluations of model-generated statements. We did not ask about personal privacy or collect any personal information of annotators in the annotation process. We pay each worker \$2.5/\$7.5 for each HIT task of SENT-COM/PARAGEN, respectively, leading to an hourly wage of ∼\$30, which is much higher than the minimum wage of \$7.5/h in the U.S. We decided the payment according to the average length of data examples. We admit that there may still be unpredictable bias in MGS even though we have carefully reviewed all annotated results from an ethical perspective.

References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Ning Bian, Peilin Liu, Xianpei Han, Hongyu Lin, Yaojie Lu, Ben He, and Le Sun. 2023. A drop of ink may make a million think: The spread of false information in large language models. *arXiv preprint arXiv:2305.04812*.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot

learners. *Advances in neural information processing systems*, 33:1877–1901.

I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. 2023. Factool: Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Pierre Jean A Colombo, Chloé Clavel, and Pablo Piantanida. 2022. Infolm: A new metric to evaluate summarization & data2text generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10554–10562.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4884–4895.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*.

Julian Eisenschlos, Bhuwan Dhingra, Jannis Bulian, Benjamin Börschinger, and Jordan Boyd-Graber. 2021. Fool me twice: Entailment from wikipedia gamification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 352–365.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Fleiss and L. Joseph. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt. *arXiv preprint arXiv:2304.02554*.

Cristina Garbacea, Samuel Carton, Shiyan Yan, and Qiaozhu Mei. 2019. Judge the judges: A large-scale evaluation study of neural language models for online review generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3959–3972.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

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

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. True: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning.

Gautier Izacard and Édouard Grave. 2021a. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880.

Gautier Izacard and Edouard Grave. 2021b. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

Pei Ke, Hao Zhou, Yankai Lin, Peng Li, Jie Zhou, Xiaoyan Zhu, and Minlie Huang. 2022. Ctrleval: An unsupervised reference-free metric for evaluating controlled text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2306–2319.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346.

Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. *arXiv e-prints*, pages arXiv–2305.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.

Nelson F Liu, Tianyi Zhang, and Percy Liang. 2023a. Evaluating verifiability in generative search engines. *arXiv preprint arXiv:2304.09848*.

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and William B Dolan. 2022. A token-level reference-free hallucination detection benchmark for free-form text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6723–6737.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for abstractive text summarization. *arXiv preprint arXiv:2303.15621*.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.

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. *arXiv preprint arXiv:2305.14251*.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen Mckeown, and Bing Xiang. 2021. Entity-level factual consistency of abstractive text summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2727–2733.

OpenAI. 2023. Gpt-4 technical report.

Jungsoo Park, Sewon Min, Jaewoo Kang, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. Faviq: Fact verification from information-seeking questions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5154–5166.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. 2021. Kilt: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2020. Superglue:

Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947.

Aalok Sathe, Salar Ather, Tuan Manh Le, Nathan Perry, and Joonsuk Park. 2020. Automated fact-checking of claims from wikipedia. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6874–6882.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin c! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3419–3425, Hong Kong, China. Association for Computational Linguistics.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 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.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

James Thorne, Max Glockner, Gisela Vallejo, Andreas Vlachos, and Iryna Gurevych. 2021. Evidence-based verification for real world information needs. *arXiv preprint arXiv:2104.00640*.

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, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA. Association for Computational Linguistics.

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.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020a. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Zhenyi Wang, Xiaoyang Wang, Bang An, Dong Yu, and Changyou Chen. 2020b. Towards faithful neural table-to-text generation with content-matching constraints. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1072–1086.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

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 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. 2023. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. Paws: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308.

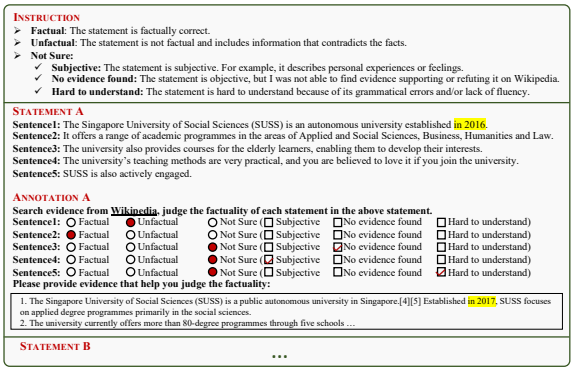

Figure 4: The annotation interface for PARAGEN; that for SENTCOM is similar.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. *arXiv preprint arXiv:2302.10198*.

A Quantifying LLMs' Hallucination

A.1 Generating Statements

The statements in MGS are generated under the few-shot setting with manually selected in-context demonstrations. For SENTCOM, the demonstrations are selected directly from the FEVER training set. And for PARAGEN, the demonstrations are selected by sampling an entity from all Wikipedia titles as the input (not overlapping the inputs for generation) and taking the first five sentences of the corresponding introduction section as the golden output. Tab. 14 shows the demonstrations. We ensure that all words in each input entity of demonstrations and generation outputs for PARAGEN appear more than 100 times in the WebText corpus to avoid rare words.

A.2 Human Evaluation

Fig. 4 shows the annotation interface. To control the annotation quality, we discard those submissions where workers (1) do not provide evidence for sentences annotated as factual or unfactual; (2) do not annotate sentences in golden statements as factual; (3) annotate more than 90% sentences as factual since manual inspection finds that the faction of factual sentences is much less than 90%. When there are sentences assigned with different labels by three workers in a statement, we collect two additional annotations for it and retain three annotations with the highest agreement for each sentence. We obtain the final human judgments by repeating the above steps until each sentence has three valid annotations.

Verifiers	SENTCOM			PARAGEN (Sent)		
	AUR	AUP	r	AUR	AUP	r
Retrieving Evidence from External Corpora						
FLAN-T5$_{11B}$	93.1	96.9	74.9	88.0	80.9	66.8
Using Golden Evidence						
FLAN-T5$_{11B}$	96.0	98.2	80.6	90.7	84.4	72.0
Not Using any Evidence						
FLAN-T5$_{11B}$	77.1	87.0	43.2	70.9	49.7	32.5

Table 12: Results on MGS under different settings.

Verifiers	SENTCOM			PARAGEN (Sent)		
	AUR	AUP	r	AUR	AUP	r
Distinguishing Factual Statements from Unfactual Ones						
FLAN-T5$_{11B}$	93.0	97.7	70.7	90.2	88.3	71.4
ChatGPT	83.3	90.5	62.2	82.8	66.4	63.6
Distinguishing Factual Statements from No-Evidence Ones						
FLAN-T5$_{11B}$	93.3	99.0	62.7	84.7	87.2	62.1
ChatGPT	87.9	96.4	60.4	75.0	65.7	52.3

Table 13: Results on SENTCOM and PARAGEN (Sent) to distinguish factual statements from unfactual/no-evidence ones. The number of factual/unfactual/no-evidence is 129/43/21 in SENTCOM, and is 155/247/166 in PARAGEN (Sent).

We ask workers to provide evidence mainly in order to force workers to search for related information to help them judge the factuality. Tab. 12 shows that LLMs get better results using golden evidence than retrieved evidence on MGS, indicating the high quality of collected evidence.

A.3 Human Judgements for No-Evidence Statements

There are three ground-truth labels for each statement in MGS, i.e., `Factual`, `Unfactual`, and `No Evidence`. To convert human judgments to numerical scores, we regard `Factual` as 1 and both `Unfactual` and `No Evidence` as 0, following prior work that converted the 3-way classification to the 2-way classification (Sarlin et al., 2020; Sathe et al., 2020). We would still like to investigate whether LLMs perform differently to verify unfactual statements and no-evidence ones. As shown in Tab. 13, the conclusion is similar under both settings: FLAN-T5$_{11B}$ outperforms ChatGPT under both settings. This indicates that it is reasonable to merge unfactual and no-evidence statements as one category.

B Collecting WKS

WKS aggregates seven existing fact verification datasets, including:

SENTCOM demonstrations:
1. **Swedish Empire** was ruled by Gustavus Adolphus from 1611 to 1632.
2. **The Boston** Celtics play their home games at TD Garden.
3. **Chris Hemsworth** appeared in A Perfect Getaway.
4. **History of** art includes architecture, dance, sculpture, music, painting, poetry literature, theatre, narrative, film, photography and graphic arts.
5. **Nikolaj Coster-Waldau** worked with the Fox Broadcasting Company.

PARAGEN demonstrations:
1. **Fire and Darkness** is a cancelled three-dimensional real-time strategy video game developed by Singularity Software. The game consists of a player controlling one of two factions, and their main mission is to defeat the enemy faction to secure the planet's resources. Its development started in 1996 and lasted for three years, with developers working mostly on summer. Although the project was incomplete, it became the first game to win the Seumas McNally Grand Prize at the Independent Games Festival of 1999. The development team invested time, but no money into the project.
2. **Patrick Sharp** (born December 27, 1981) is a Canadian former professional ice hockey player who played 15 seasons in the National Hockey League (NHL) for the Philadelphia Flyers, Chicago Blackhawks, and Dallas Stars. Sharp played collegiate hockey at the University of Vermont before he was drafted by the Flyers in 2001. He began his NHL career with the Flyers organization, but was traded to the Blackhawks in 2005. He became a three-time Stanley Cup champion with the Blackhawks in 2010, 2013, and 2015. Sharp was later dealt to the Stars in 2015, where he spent two seasons before returning to the Blackhawks in 2017.
3. The **Cleveland East Ohio Gas Explosion** occurred on the afternoon of Friday, October 20, 1944. The resulting gas leak, explosion and fires killed 130 people and destroyed a one square mile area on Cleveland, Ohio's east side. At 2:30 p.m. on the afternoon of Friday, October 20, 1944, above ground storage tank number 4, holding liquefied natural gas in the East Ohio Gas Company's tank farm, began to emit a vapor that poured from a seam on the side of the tank. The tank was located near Lake Erie on East 61st Street, and winds from the lake pushed the vapor into a mixed use section of Cleveland, where it dropped into the sewer lines via the catch basins located in the street gutters. As the gas mixture flowed and mixed with air and sewer gas, the mixture ignited.
4. **Broke Sky** is a 2007 neo-noir 35 millimeter film, and the directorial debut of cinematographer Thomas L. Callaway. The film stars Will Wallace, Joe Unger, Bruce Glover, Duane Whitaker and Barbara Chisholm, and has earned comparisons to the work of the Cohen Brothers. Bucky and Earl are the two man team that collect and dispose of road kill for the county. A new, specially designed carcass removal truck forces them to choose which one of them gets to keep his job and who is let go. Earl comes up with a plan so they can both keep their jobs, but it means working at night.
5. **Design technology**, or D.T., is the study, design, development, application, implementation, support and management of computer and non-computer based technologies for the express purpose of communicating product design intent and constructability. Design technology can be applied to the problems encountered in construction, operation and maintenance of a product. At times there is cross-over between D.T. and Information Technology, whereas I.T. is primarily focused on overall network infrastructure, hardware & software requirements, and implementation, D.T.

Table 14: Demonstrations for generating statements in MGS. The **red** words correspond to {input} in Tab. 1.

(1) **FEVER:** It contains crowd-sourced statements by altering a word or negating sentences from Wikipedia (Thorne et al., 2018), thereby leading to strong artifacts (Schuster et al., 2019). We use the KILT version of FEVER (Petroni et al., 2021). Since the official test set is hidden, we sample 1K examples from the validation set for testing and use the rest for validation.

(2) **BoolQ-FV** (Thorne et al., 2021): It consists of more realistic claims than FEVER, which are derived from real-world information needs by rewriting users' search-engine queries and verifying them against evidence from Wikipedia.

(3) **FM2** (Eisenschlos et al., 2021): It is collected by gamification. Players write challenging statements supported or refuted by evidence from Wikipedia and spot refuted claims written by others.

(4) **PubMedQA** (Jin et al., 2019): It is initially a question-answering dataset specifically designed for the biomedical domain based on the PubMed database[7], which is a comprehensive collection of biomedical literature. Each PubMedQA example consists of a yes-no question and the abstract of the corresponding background paper. We prompt ChatGPT to convert the question into a declarative sentence as the statement to be judged with the abstract as the golden evidence.

(5) **XsumFaith** (Maynez et al., 2020) and **SummEval** (Fabbri et al., 2021): Each example consists of a summary and the corresponding source document. The summaries are generated by various models and paired with human-annotated binary faithfulness labels to the source documents. We regard the summary as a statement to be judged and the source document as the golden evidence.

(6) **SciFact** (Wadden et al., 2020): It contains scientific claims against a corpus of scientific

[7] https://pubmed.ncbi.nlm.nih.gov/

papers. The claims are re-formulated from naturally occurring citation sentences. We use the dataset to test the generalization of LLMs to more professional domains than Wikipedia.

We do not include FaVIQ used in §5.3 in WKS since (1) most statements of FaVIQ are transformed from information-seeking QA pairs, which is similar to BoolQ-FV; and (2) FaVIQ focuses on token-level factual errors, which have been covered by FEVER. To make WKS less redundant, we do not include FaVIQ in WKS.

C Experiments

C.1 Evaluation Metrics

We use multiple metrics to evaluate the verifiers, including ECE, ACC, AUR, AUP, and r, etc. These metrics are focusing on different aspects. **(1)** ECE is the weighted average absolute difference between metric scores and accuracy in each bin of $[0, 1]$ as follows:

$$\text{ECE} = \sum_{b=1}^{B} \frac{n_b}{N} |\text{acc}(b) - \text{conf}(b)|,$$

where acc(b) is the accuracy, conf(b) is the average metric scores (confidence), and n_b is the number of examples in b-th bin. We set $B = 20$ in our experiments. A well-calibrated verifier can serve as an estimation of probabilities of making mistakes, which is more informative than only predicting "factual" or not. Note that lower ECE does not mean better discrimination ability. Supposing that a model always predicts a score of 0.5 on a perfectly balanced dataset, ECE will be 0, while the model is useless. **(2)** ACC is the fraction of examples that are correctly predicted, which means both human judgments and metric scores are either larger or smaller than 0.5. **(3)** AUR is one of the commonest metrics to evaluate binary classifiers. Although both ACC and AUR test the discrimination ability, AUR does not assume a pre-defined threshold, so it is more important on imbalanced datasets (e.g., BoolQ-FV and PARAGEN). **(4)** AUP is also widely used for the evaluation of imbalanced datasets. For comparison, AUR/AUP is more sensitive to the discrimination ability of verifiers on the `factual`/`non-factual` category. Therefore, AUP is a better metric when most statements are non-factual (e.g., PARAGEN). **(5)** All the above metrics require human judgments to be binary, while Pearson's correlation r can be used between two continuous variables (e.g., PARAGEN (Para)). In summary, we recommend comprehensively considering different metrics from multiple perspectives in future research and realistic application.

Retrievers	FEVER	BoolQ-FV	FM2	SciFact
BM25	88.09	63.88	75.50	55.71
Contriever	94.95	88.63	87.80	79.00

Table 15: Token recall scores (%) of different retrievers.

# Retrieved Passages	1	3	5	7	10
FLAN-T5$_{11B}$	54.2	64.0	67.4	68.4	68.8
ChatGPT	47.9	58.2	58.4	57.9	56.7

Table 16: Pearson's correlation on FM2 varying with the number of retrieved passages on the verification performance.

C.2 Retriever

We compare BM25 and Contriever (Izacard et al., 2022) in terms of recall of golden evidence on WKS, i.e., the ratio of tokens in golden evidence that also appear in the top 10 pieces of retrieved evidence. Tab. 15 shows the much higher recall scores of Contriever, so we use it in our experiments.

Furthermore, Tab. 16 shows the Pearson's correlation scores of FLAN-T5$_{11B}$ and ChatGPT varying with the number of retrieved passages (each passage contains 150 tokens) on the FM2 dataset. We see that ChatGPT reaches saturation faster than FLAN-T5$_{11B}$ with increased retrieved passages, although ChatGPT supports a maximum length of 4,096 tokens. This indicates the length extrapolation ability of FLAN-T5 to some extent[8]. We agree that the training-testing length bias may potentially impact the performance of FLAN-T5, although we have not explicitly observed such an impact. Actually, when the length of retrieved passages is much larger than 512 in our paper, FLAN-T5$_{11B}$ always performs the best. We will further investigate the influence of the length bias in our future work.

C.3 Verifiers

NLI$_{11B}$ The corresponding model card on HuggingFace is "google/t5_xxl_true_nli_mixture". It is trained on a mixture of SNLI (Bowman

[8]The encoders of FLAN-T5 models are trained with a maximum length of 512 tokens. Thanks to the relative positional encoding mechanism adopted by FLAN-T5 (all relative distances are truncated into 512), we can test the performance of FLAN-T5 with much longer inputs.

Training Sets	Test Sets			
	FEVER	BoolQ-FV	FM2	SciFact
FEVER	**94.60**	77.65	70.00	75.92
BoolQ-FV	83.10	**82.71**	61.30	70.68
FM2	89.40	74.88	**76.96**	70.16
SciFact	65.80	71.94	52.61	**78.53**

Table 17: Accuracy of the FiD metric which are trained on one dataset and then used for another one.

Datasets	FEVER			BoolQ-FV			FM2		
	ECE	AUR	r	ECE	AUR	r	ECE	AUR	r
Soft	**3.1**	**98.2**	**90.2**	**10.2**	**94.7**	**75.3**	**8.2**	**89.5**	**68.8**
Hard	6.2	93.9	87.7	14.5	87.6	70.0	18.0	81.9	65.0

Table 18: Results of hard coding and soft coding based on FLAN-T5$_{11B}$.

et al., 2015), MNLI (Williams et al., 2018), FEVER (Thorne et al., 2018), Scitail (Khot et al., 2018), PAWS (Zhang et al., 2019) and VitamInC (Schuster et al., 2021). Although FEVER is included in the training set of the model, we still regard this metric as unsupervised since they use golden evidence for training while we mainly use retrieved evidence for testing.

FiD Tab. 17 shows the dataset transfer results of the supervised verifier FiD on WKS, illustrating the poor generalization ability of supervised verifiers with a significant drop in accuracy when transferring it to a different dataset. On the other hand, when applying FiD on PARAGEN (Sent), we do not provide the context during verification since FiD has not been trained to utilize the context.

FLAN-T5 FLNA-T5 series have been trained on SciFact with golden evidence. Therefore, the results of FLAN-T5 on SciFact in Tab. 6 may be over-estimated. We ensure that FLAN-T5 series have not been trained on the other three datasets of WKS in the Wikipedia domain, including FEVER, Boolq-FV, and FM2.

ChatGPT We hard code the predicted factuality score of ChatGPT as 1/0 if it returns "A"/"B" using the prompt in Tab. 4 because its output probabilities are unavailable. Tab. 18 compares the results of hard coding (i.e., using the *generated answers*) and soft coding (i.e., using the *generation probabilities*) based on FLAN-T5$_{11B}$, indicating that the performance of ChatGPT may be underestimated in terms of metrics that depend on probabilities, including ECE, AUR, AUP, and r in our experiments.

Verifiers	FEVER			FM2		
	ECE	ACC	AUR	ECE	ACC	AUR
Constant	48.3	51.7	50.0	50.6	49.4	50.0
Retrieving Evidence from External Corpora						
LLama$_{7B}$	45.9	51.6	47.8	47.7	49.4	45.0
Alpaca$_{7B}$	43.9	51.3	56.3	45.3	49.3	46.2
FLAN-T5$_{11B}$	**3.1**	**93.8**	**98.2**	**8.2**	**82.0**	**89.5**
InstructGPT	6.4	88.8	96.1	15.0	71.5	81.5
GPT3.5	7.6	91.7	96.6	21.0	77.5	82.8
ChatGPT	8.2	92.8	92.6	21.4	79.1	78.6
Using Golden Evidence						
LLama$_{7B}$	6.3	59.6	64.8	**6.4**	54.6	55.8
Alpaca$_{7B}$	43.8	52.1	86.4	46.7	49.2	73.9
FLAN-T5$_{11B}$	5.4	93.3	**98.6**	9.7	87.3	**96.1**
InstructGPT	**3.9**	93.9	98.0	8.0	85.4	92.6
GPT3.5	6.1	93.8	96.6	12.1	87.1	93.2
ChatGPT	5.8	**94.4**	94.3	11.6	**88.5**	88.4
Not Using any Evidence						
LLama$_{7B}$	14.7	48.5	67.5	**6.7**	50.5	49.4
Alpaca$_{7B}$	**5.4**	74.9	80.5	19.9	51.8	54.4
FLAN-T5$_{11B}$	11.1	74.4	**87.0**	15.7	59.5	66.3
InstructGPT	14.4	75.3	83.4	26.4	58.4	62.9
GPT3.5	25.6	73.8	78.3	41.4	57.3	61.5
ChatGPT	18.1	**82.0**	81.5	33.5	**66.7**	**66.5**

Table 19: Performance of different instruction-tuned LLMs on FEVER and FM2. All models are under the zero-shot setting.

More LLMs In addition to FLAN-T5 series, GPT3.5 and ChatGPT, we also test the performance of LLama$_{7B}$, Alpaca$_{7B}$ (Taori et al., 2023) and InstructGPT on WKS. Tab. 19 shows the results under different settings. We observe that:

(1) InstructGPT is better calibrated than GPT3.5 with similar discrimination performance under all settings. Compared with CPT3.5, InstructGPT has not been trained through RLHF (reinforcement learning from human feedback), indicating that RLHF may impact calibration, which is accordant with GPT4's technical report (OpenAI, 2023).

(2) LLama$_{7B}$ performs like random guessing on FM2 under all settings. When using retrieved evidence, its performance is also at the random level on FEVER, which can be improved using golden evidence, suggesting its poor robustness to noise in the given evidence. Surprisingly, when not using any evidence, the AUR score of LLama$_{7B}$ on FEVER instead increases compared with using golden evidence, despite the worse ECE and ACC scores. This means LLama$_{7B}$ is seriously biased toward a certain category when not using evidence, and providing golden evidence can alleviate the bias but impact the discrimination ability.

(3) Alpaca$_{7B}$ is comparable to or better than LLama$_{7B}$ on the whole. One contrary phenomenon to LLama$_{7B}$ is that Alpaca$_{7B}$ is better calibrated with a higher ACC score when not using any evidence than using golden evidence on FEVER, indicating that the instruction tuning fashion of Alpaca$_{7B}$ increases the bias toward a certain category when using golden evidence, despite a better ability to utilize internal and external knowledge.

Results on FactPrompts The dataset is released by a contemporary work (Chern et al., 2023), which comprises real-world prompts from various sources, such as Quora and TruthfulQA (Lin et al., 2022), along with corresponding responses generated by ChatGPT. We regard the response as the statement to be judged and retrieve 10 passages from C4 (Raffel et al., 2020) using an off-the-shelf tool[9] as the evidence. FactPrompts include 177 factual statements and 56 unfactual ones. All these statements are model-generated, so we believe that they should be unseen during the training stages of LLMs. Nevertheless, we acknowledge that texts with similar distributions to these statements might still be seen.

As shown in Tab. 22, FLAN-T5$_{11B}$ outperforms ChatGPT with or without external evidence. ChatGPT without any evidence has a random performance on FactPrompts (AUR≈50.0). This might be because the statements in the dataset are generated by ChatGPT itself, so ChatGPT easily regards them as factual statements.

C.4 Verification Method

Besides using generation probabilities of certain answer words (e.g., "A," and "B") as factuality scores described in Section 4.2, we also try several other methods widely used for NLG evaluation as follows:

- **Chain-of-Thought Prompting (CoT)** was proposed to elicit LLMs' reasoning ability (Wei et al., 2022), and recently was also used for improving NLG evaluators (Liu et al., 2023b).
- **Likert-Scale Rating** is the most common method for human and automatic evaluation (Gao et al., 2023). We prompt LLMs to rate statements from 1 to 5 and re-scale the ratings into the range from 0 to 1 as the final factuality scores.

[9] https://c4-search.apps.allenai.org/

The task is to judge whether the statements are factually correct based on the corresponding facts. Two examples are as follows:
Example 1.
Facts:
0. Baseball Hall of Fame in 1988 · · ·
1. He was a carpenter and construction worker by trade. · · ·
2. players, and 599 of them loved Willie Stargell · · ·

Statement: Willie Stargell lived with his father after his mother and father separated.

Output: Willie Stargell lived with his mother in Florida after his parents' divorce, and later returned to live with her in California. There is no mention of him living with his father after the separation. So the statement is incorrect; the answer is No.

Example 2.
Facts:
0. Alfredo Stroessner Alfredo Stroessner Matiauda · · ·
1. in most other Latin American countries · · ·
2. led to the liberalization of Paraguay · · ·
3. system. On 3 February 1989, Stroessner was · · ·
4. some of Stroessner's policies· · ·

Statement: After the overthrow of Stroessner, Paraguay had its first non-military head of state in four decades.

Output: Stroessner, a military officer, had been in power for 35 years, from 1954 to 1989, and had been overthrown in a military coup led by General Andrés Rodríguez. After Stroessner's ousting, Rodríguez orchestrated a political campaign with the Colorado Party and won the presidency in a multiparty election held in May 1989. This marked the first time in four decades that Paraguay had a non-military head of state. So the statement is correct; the answer is Yes.

Now please judge the following statement:

The task is to rate the factuality level for the statement based on the fraction of details in the statement that can be supported by the facts.
You should output only one single number from 1, 2, 3, 4, and 5, indicating the factuality score of the statement, where a higher score means that more details of the statement are factual.

1 - Almost all details in the statement are not supported by the facts.
2 - Some details in the statement are supported by the facts, but the majority are not.
3 - About half of the details in the statement are supported by the facts.
4 - Most of the details in the statement are supported by the facts.
5 - All details in the statement are supported by the facts.

Please output the score of the following statement:

Table 20: Input Prompt for CoT prompting (Top) and Likert-scale rating (Bottom).

Tab. 20 shows the prompt to instruct LLMs to judge statements using the above two verification methods. We compare different verification methods

Methods	FEVER				FM2			
	ACC	P	R	r	ACC	P	R	r
Direct	**92.8**	92.8	92.8	**92.8**	**79.1**	80.1	75.2	**56.7**
CoT	87.5	**93.2**	81.8	75.7	70.7	**83.1**	51.1	44.5
Likert-Scale	89.6	86.9	**93.4**	79.6	76.1	74.6	**76.1**	53.5

Table 21: Performance of different verification methods. **Direct** means directly predicting "A" or "B" as described in Section 4.2.

Models	ACC	AUR	r
Retrieving Evidence from External Corpora			
FLAN-T5$_{11B}$	64.4	**68.1**	**26.5**
ChatGPT	**71.7**	59.4	19.5
Not Using any Evidence			
FLAN-T5$_{11B}$	63.5	**67.6**	**26.0**
ChatGPT	**72.1**	50.5	1.6

Table 22: Results on FactPrompts.

based on ChatGPT and show the results in Tab. 21:

(1) Directly predicting "A" or "B" is the simplest but best verification method.

(2) Using CoT prompting will lead to a significant performance drop. We attribute the drop to the impact of generated non-factual rationales, which misleads ChatGPT to mistake factual statements for non-factual ones, as indicated by the similar precision but lower recall than direct prediction.

(3) When using Likert-scale rating, LLMs tend to give higher scores than direct prediction, with a lower precision but higher recall. The reason may be that the ground-truth labels of the statements are collected more extremely than Likert-scale rating, which means only very minor factual errors can make a statement labeled `unfactual`. For example, a statement with a ground-truth label of 0 can be rated 4 (0.75 after re-scaling) under the Likert-scale rating. Therefore, Likert-scale rating may also be a good choice for fact verification if users expect the score to be less strict and more informative. In future research, we recommend focusing on standardizing Likert-scale protocols for factuality rating and designing efficient and effective methods for detecting fine-grained factual errors (e.g., at the token level).

C.5 Prompts for Robustness Assessment

We design different prompts to assess the robustness of LLMs in Section 5.2 by changing the "Question" part in Tab. 4 to:

- "Is the statement entailed by the given facts? (A) Yes. (B) No. Please answer A or B:"
- "Based on the given facts, judge whether the statement is factually correct. Please answer Yes or No:"
- "Can the given facts support the statement? Please answer Yes or No:"

Proportion (%)	Factual	Unfactual	No Evidence
Factual Context	45.0	45.5	9.5
Non-Factual Context	15.1	42.1	42.7

Table 23: Label distribution of sentences in PARAGEN (Sent) with factual or non-factual context. Here, factual context means the context of the sentence includes only factual sentences, and non-factual context means the opposite.

C.6 Influence of Non-Factual Context

Influence on Model Generation In Tab. 1, we observe the increase of no-evidence sentences as the generation proceeds to later sentences. We conjecture that it is because of the error accumulation during generation. A contemporary work (Zhang et al., 2023) also highlights the phenomenon and empirically attributes it to error propagation. We show our finding in Tab. 23: There is a lower proportion of no-evidence sentences when the context is factual than non-factual. Here, factual context means the context of the sentence includes only factual sentences, and non-factual context means the opposite.

Intuitively, more errors in the context (unfactual or no-evidence) correlate to more no-evidence sentences. However, it may be difficult to affirm the causality between them. The potential insight is that when generating open-ended long texts using auto-regressive models, it is necessary to dynamically involve external feedback, e.g., from humans, retrievers, planners, etc., to avoid the impact of non-factual contexts.

Dependency Annotation For each sentence in PARAGEN (Sent), we manually annotate whether it involves dependencies on the context, i.e., including references to any entities in the context. We label a sentence `independent` if

(1) It is the first one in the paragraph (137 sentences totally).

(2) It does not contain nouns or pronouns that refer specifically to entities in the context.

Task: Given a statement following its context, find all references to the context in the statement, and replace them with words that they refer to. Be careful not to change the original word order as much as possible.
For example, if the context is "Mary loves pizzas.", and the statement following the context is "She eats them every day.", you should output "Mary eats pizzas every day."
Now finish the following task:
Context: $\{c\}$
Statement following the context: $\{s\}$
Output:

Table 24: Prompts to instruct ChatGPT to perform coreference resolution.

Verifiers	Factual Cont (104/127)			Non-Factual Cont (51/286)		
	AUR	AUP	r	AUR	AUP	r
FLAN-T5$_{11B}$ (ZS)	91.6	91.8	74.2	84.1	49.8	46.4
w/ CR	93.1	92.5	76.0	85.8	62.6	58.4
ChatGPT (ZS)	83.3	72.4	66.1	76.3	30.5	38.3
w/ CR	82.1	71.9	64.0	82.5	40.4	52.1

Table 25: Generalization to sentences whose context is factual or non-factual. **Cont** is short for **context**. Two digits in each parenthesis are the number of factual/non-factual statements. **Factual Context** means all sentences in the context of a statement are factual, or the context is empty. **Non-Factual Context** means there exists at least one non-factual sentence in the context of the statement.

(3) It contains nouns that refer to entities in the context but can be understood solely. For example, in the paragraph "Russian wine is produced in several regions. $\cdots$ Russian wines have won numerous international awards.", the second "Russian wine" in the last sentence refers to the same entity as the context, but the last sentence is not dependent on the context.

Otherwise, the sentence is labeled `dependent` on the context. We asked two graduates for annotation and found they annotate the same labels for more than 90% sentences. Finally, we obtain 189 context-dependent sentences (61 factual and 128 non-factual), and 279 context-independent ones (94 factual and 185 non-factual).

De-Contextualization We use ChatGPT for coreference resolution (CR) using the prompt in Tab. 24. To assess the accuracy of the CR method, we inspect twenty randomly sampled CR results and find ChatGPT can complete the task perfectly without any errors. This may be because all sentences in a paragraph of PARAGEN are almost centered on the same entity so that it is easy to resolve coreferences.

Influence on Fact Verification Tab. 11 in the main paper has shown greater difficulty in verifying context-dependent statements than context-independent ones. We are also curious whether non-factual context will impact the performance of the fact verifiers. Tab. 25 answers the question in the affirmative. Performing CR will alleviate the disparity by improving the performance on statements with non-factual context.

C.7 Entity Type

Tab. 26 shows details of entity types used in Fig. 3. We see that "Person," "Work of Art," and "Building" have relatively lower $\frac{3}{4}$ percentile. This means these types of entities are distributed more sharply, which may account for better performance on the statements corresponding to these types than other types.

C.8 Influence of LLMs' Memorization

We are curious about whether there is a degree of memorization when LLMs make judgments about factuality. For example, they may inherently tend to judge those texts that are more probable under them as hallucinatory regardless of the given evidence. To investigate this correlation, we take perplexity as a proxy to measure the memorization degree of an LLM for any one statement. Then, we compute Pearson's correlation between the LLM's perplexity and predicted/golden factuality score on the FEVER and FM2 test sets. It is intractable to compute the perplexity under GPT models since their APIs do not return the logits of any specified texts. Therefore, we only compute the correlation based on the FLAN-T5 models, and use its decoder to compute the perplexity. Tab. 27 shows the correlation results.

We see that the perplexity of FLAN-T5 correlates significantly with the predicted factuality score: When the LLM memorizes a statement better (indicated by a lower perplexity score), it is more likely to regard the statement as factual. However, the perplexity score does not correlate with the golden label. The results suggest that memorization can influence LLMs' judgments about factuality to some extent, but may not be the main contributor to the superior fact verification performance of LLMs.

C.9 Case Study

Fig. 5 plots precision and recall to better understand the preference of different LLMs. On SciFact,

Entity Types	spaCy Label	Description	$\frac{3}{4}$ Percentile
Non-NE	N/A	Common words that are not named entities	9,526. 5
Person	PERSON	People, including fictional	420. 0
Work of Art	WORK_OF_ART	Titles of books, songs, etc.	2,166. 0
Product	PRODUCT	Objects, vehicles, foods, etc. (not services)	3,456. 5
Event	EVENT	Named hurricanes, battles, wars, sports events, etc.	8,559. 0
Language	LANGUAGE	Any named language	245,337. 0
Building	FAC	Buildings, airports, highways, bridges, etc.	603. 0
Company	ORG	Companies, agencies, institutions, etc.	18,557. 5
Group	NORP	Nationalities or religious or political groups	21,983. 5
Country	GPE	Countries, cities, states	205,839. 0
Location	LOC	Non-GPE locations, mountain ranges, bodies of water	8,624.75
Date	DATE	Absolute or relative dates or periods	372,927. 0
Cardinal	CARDINAL	Numerals that do not fall under another type	3,572,566. 0
Ordinal	ORDINAL	"first", "second", etc.	50,076. 0

Table 26: Entity Types and corresponding labels and descriptions from spaCy. $\frac{3}{4}$ percentile means $\frac{3}{4}$ of entities of some entity type appear in fewer Wikipedia passages than that number.

Correlation between Perplexity and Predicted Factuality Score		
Models	**FEVER**	**FM2**
FLAN-T5$_{780M}$	-16.4**	-11.6**
FLAN-T5$_{3B}$	-11.0**	-15.2**
FLAN-T5$_{11B}$	-10.3**	-12.2**
Correlation between Perplexity and Golden Factuality Score		
Models	**FEVER**	**FM2**
FLAN-T5$_{780M}$	-5.7**	3.4**
FLAN-T5$_{3B}$	-2.6**	2.5**
FLAN-T5$_{11B}$	-3.6**	3.7**

Table 27: Correlation scores between FLAN-T5's perplexity and predicted/golden factuality scores on FEVER and FM2. ** means p-value<0.01.

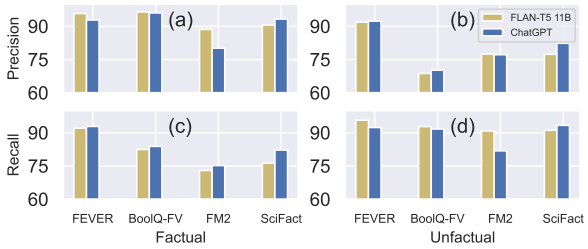

Figure 5: Precision and recall scores with retrieved evidence in WKS (**Left:** Factual **Right:** Unfactual). **Takeaway:** ChatGPT prefers to predict `factual`.

ChatGPT surpasses FLAN-T5$_{11B}$ with higher precision and recall. In the Wikipedia domain, ChatGPT exhibits a preference for predicting `factual` in contrast to FLAN-T5$_{11B}$, resulting in higher recall on factual statements (Fig. 5 c). Such a preference also leads ChatGPT to make more mistakes when predicting `factual` (Fig. 5 a) and misidentify unfactual statements as factual ones (Fig. 5 d). Manual inspection reveals that ChatGPT easily builds spurious connections between related yet distinct concepts, as indicated in the following case study. On the other hand, Zhong et al. (2023) also emphasized the inadequacy of ChatGPT in assessing inter-sentence similarity compared with BERT-based models, which further supports our observation.

Tab. 29 shows a case that ChatGPT fails to judge while FLAN-T5$_{11B}$ does not, suggesting that ChatGPT may build spurious connections between related yet distinct concepts (e.g., "math teacher" and "study mathematics").

Through manual inspection of test examples that both ChatGPT and FLAN-T5$_{11B}$ make wrong judgments, we summarize several typical types of errors. To illustrate these error types, Tab. 28 shows several representative cases:

(1) False Numerical and Logical Reasoning: In Example 1 and 2, despite correct retrieval, ChatGPT mistakes "36 years old" for "40s", and uses "Continuity is not sufficient for differentiability" to refute "differentiability is sufficient for continuity," indicating the weakness of LLMs in understanding numerical and logical relations.

(2) Misled Retriever: In Example 3, the retriever is misled by the non-factual information (i.e., "the United States") in the statement, thus leading to useless evidence and wrong predictions. When giving correct evidence to LLMs (i.e.,

ID	Statement	Retrieved Most-Related Evidence	Label	Pred	ChatGPT Explanation
1	Georges Bizet died in his early **40s**. (*from* FM2)	✔ (Georges Bizet) Georges Bizet (25 October 1838 – 3 June 1875), registered at birth as Alexandre César Léopold Bizet, was a French composer of the Romantic era.	0	1	Georges Bizet was born in 1838 and died in 1875, which means he was 36 years old when he died. This is considered to be in his early 40s.
2	A function is continuous when it is differentiable at that point. (*from* BOOLQ-FV)	✔ (Differentiable function) If f is differentiable at a point x0, then f must also be continuous at x0. In particular, any differentiable function must be continuous at every point in its domain.	1	0	A function can be continuous at a point but not differentiable at that point.
3	**The United States** was the allied nation that was the first to reach the German capital of Berlin. (*from* FaVIQ)	✖ (2nd Armored Division (United States)) After a rest in February, the division, now commanded by Major General Isaac D. White, drove on across the Rhine on 27 March, and was the first American division to reach the Elbe at Schonebeck on 11 April.	0	1	The U.S. Army's 2nd Armored Division was the first American division to reach the Elbe at Schonebeck on 11 April and entered Berlin in July, making the statement correct.
4	Institute for Employment Studies is a **government-linked** think tank on **British** employment policy and practice. (*from* PARAGEN)	✔ (Institute for Employment Studies) The Institute for Employment Studies (IES) is a British centre of research and consultancy on human resources and employment.	0	1	The Institute for Employment Studies is a British think tank that focuses on research and consultancy related to human resources and employment.

Table 28: Several representative cases where both FLAN-T5$_{11B}$ and ChatGPT makes mistakes. We highlight non-factual words in the statements in **bold** and indicate the source datasets after the statements. "Label" or "Pred" refer to ground-truth labels or model predictions, and 1 or 0 means factual or not. ChatGPT explanations are generated by prompting ChatGPT to explain its predictions. **Most-related evidence** means the evidence that is the most related to the ChatGPT explanation, which is manually selected from the top ten pieces of evidence from the retriever. For the evidence, we indicate the corresponding Wikipedia page titles before them and use ✔and ✖ to illustrate whether the evidence can help judge the factuality of the statements.

Statement: Pope Gregory XIII was a math teacher before becoming pope.

Evidence: ··· after studying mathematics, Latin, law, and philosophy, initially at the university of Bologna, and then in Padua in 1545 and becoming a notary. His interests successively extended to philosophy and logic, which he combined with the study of medicine. He obtained a degree in medicine and philosophy ··· He later taught jurisprudence for some years ···

ChatGPT Explanation: The facts state that Pope Gregory XIII studied mathematics and graduated in law, later teaching jurisprudence for some years. He also studied philosophy and logic, which he combined with the study of medicine, obtaining a degree in medicine and philosophy. Therefore, it can be inferred that he had a background in mathematics before becoming pope.

Table 29: An unfactual statement that ChatGPT fails to judge while FLAN-T5$_{11B}$ succeeds. We obtain the explanation by prompting ChatGPT to explain its prediction. Red words are copied from the evidence, and cyan and blue words are conflicting.

the Wikipedia page "Race to Berlin"), we find they can make the right prediction.

(3) Specious Inter-Entity Relations: Statements generated by larger models (such as GPT3.5) tend to seem more fluent and factual with interrelated entities, although the relations between them may be non-factual. In Example 4, the evidence provides useful information about "IES," but it does not mention whether "IES" is linked to the government or its research is only limited to the British. ChatGPT ignores such specious relations and hence makes wrong predictions.