FactLens: Benchmarking Fine-Grained Fact Verification

Kushan Mitra
Megagon Labs, USA
kushan@megagon.ai

Dan Zhang
Megagon Labs, USA
dan_z@megagon.ai

Sajjadur Rahman*
Adobe Inc.
sajjadurr@adobe.com

Estevam Hruschka
Megagon Labs, USA
estevam@megagon.ai

Abstract

Large Language Models (LLMs) have shown impressive capability in language generation and understanding, but their tendency to hallucinate and produce factually incorrect information remains a key limitation. To verify LLM-generated contents and claims from other sources, traditional verification approaches often rely on holistic models that assign a single factuality label to complex claims, potentially obscuring nuanced errors. In this paper, we advocate for a shift towards fine-grained verification, where complex claims are broken down into smaller sub-claims for individual verification, allowing for more precise identification of inaccuracies, improved transparency, and reduced ambiguity in evidence retrieval. However, generating sub-claims poses challenges, such as maintaining context and ensuring semantic equivalence with respect to the original claim. We introduce FactLens[1] , a benchmark for evaluating fine-grained fact verification, with metrics and automated evaluators of sub-claim quality. The benchmark data is manually curated to ensure high-quality ground truth. Our results show alignment between automated FactLens evaluators and human judgments, and we discuss the impact of sub-claim characteristics on the overall verification performance.

1 Introduction

Large Language Models (LLMs) have proven to be powerful tools, demonstrating impressive capabilities in language generation and understanding (Touvron et al., 2023; Brown et al., 2020). However, a well-known limitation of LLMs is their tendency to hallucinate, generating information that is factually incorrect or unsupported by evidence (Ji et al., 2022; Lin et al., 2022). As LLMs become more widespread, especially in applications where factual accuracy is crucial, there has been increasing research on methods to verify the factuality of LLM-generated content as well as claims from other sources.

Previous works on building fact-checking benchmarks focus on generating `claims` with a ground truth `label`, and in some cases provide the `evidence/context` to verify the claim (Aly et al., 2021; Schlichtkrull et al., 2024). Claims are generated using human annotators (Aly et al., 2021), synthetic processes (Fatahi Bayat et al., 2023; Tang et al., 2024a), or considering LLM outputs on Question-Answering tasks (Wang et al., 2024). To increase the complexity of the fact-checking process, the claims are generated from source data of multiple domains & modalities, such as Wikipedia text and/or tables (Thorne et al., 2018; Chen et al., 2020; Aly et al., 2021), Web Pages (Schlichtkrull et al., 2024), Knowledge Graphs (Kim et al., 2023), online posts/chats (Wang et al., 2024; Li et al., 2024), and QA tasks from various domains such as statistics, finance, legal, etc (Jacovi et al., 2024).

These works also provide baseline fact-checking pipelines, which typically involves two main stages: (1) the retrieval of relevant evidence using Search APIs and multimodal data-lakes (Tang et al., 2024b; Schlichtkrull et al., 2024) and (2) the verification of claims based on that evidence using NLI-based, LLM-based and fine-tuned fact-verification models (Li et al., 2024). Some works also explore delegating these steps entirely to an LLM-based policy framework (Li et al., 2024; Peng et al., 2023).

Despite this structured pipeline, most existing methods rely on a holistic verification model, where complex claims are assigned a single factuality label, often obscuring the nuanced nature of the errors or inaccuracies in the claims. In this work, we echo the sentiments of Wang et al. (2024); Liu et al. (2020); Pan et al. (2023); Min et al. (2023); Si et al. (2024) for a shift towards fine-grained verification of complex claims, where claims are decom-

*Work done while at Megagon Labs.
[1]https://github.com/megagonlabs/factlens

Findings of the Association for Computational Linguistics: ACL 2025, pages 18085–18096

posed into smaller, more manageable sub-claims that can be individually verified. We additionally emphasise on the need to provide evaluation metrics to benchmark such fine-grained verification, and enrich existing benchmarks with fine-grained verification labels.

As shown in Figure 1, the benefits of fine-grained verification are substantial. By breaking down a complex claim into its constituent sub-claims, verification is more precise, allowing for pinpointing exact locations of factual inaccuracies. Additionally, this approach enables more transparent rationalizations and explanations, as each sub-claim can be linked directly to its corresponding evidence or lack thereof. Fine-grained decomposition also narrows the scope of evidence retrieval, making the subsequent verification process more focused and less prone to ambiguity.

Achieving fine-grained verification, however, presents its own challenges. Decomposing a raw, complex claim into smaller sub-claims is not simply a matter of splitting it into sentences. Poorly constructed sub-claims can introduce a variety of issues: they may lose the context necessary for proper verification, lack atomicity, or misrepresent the original information by either omitting key details or introducing new fabricated ones. Ensuring the quality and verifiability of these sub-claims is, therefore, critical for the overall success of the verification process.

To address these challenges, we introduce FactLens, a benchmark designed specifically for fine-grained fact verification. FactLens provides a novel suite of metrics for evaluating the quality of sub-claim generation and incorporates automated evaluators that combine LLM-based assessments with statistical metrics. The dataset has been manually curated to ensure high-quality sub-claims.

Through empirical evaluation, we demonstrate that our sub-claim evaluators align closely with human judgments. Moreover, our end-to-end evaluation shows that these fine-grained scores correlate strongly with improved downstream verification performance. We also present the results of state-of-the-art models on sub-claim generation, revealing the challenges inherent in this task and the need for further research in this area.

2 Evaluating Sub-claims with FactLens

At the core of fine-grained verification is the decomposition of complex claims into smaller, more specific sub-claims when necessary. The accuracy of the overall verification process depends heavily on the quality of these sub-claims; errors (e.g. oversimplification, omission of important details, or incorrect contextualization) in their formulation can lead to flawed verification outcomes. To detect potential issues early on in the sub-claim generation process, we propose a set of metrics to quantitatively assess sub-claim quality across several dimensions.

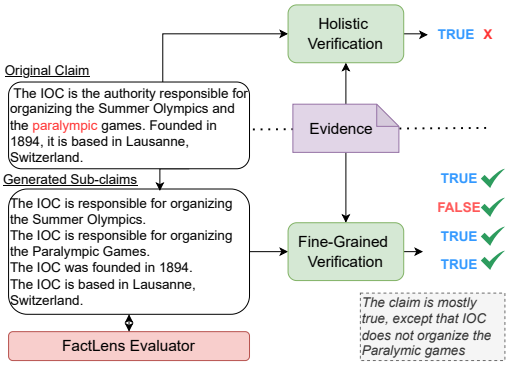

Figure 1: Examples of holistic fact verification (upper) failed to identify inaccuracies, whereas fine-grained verification (lower) clearly pinpointed the sources of error. In fine-grained verification, the FactLens Evaluator can be used to assess individual sub-claims and identify any alarming signals that may suggest the need for human intervention or regeneration of the sub-claims.

2.1 Evaluation Metrics

The decomposed sub-claims should meet criteria below to fully realize benefits of fine-grained verification such as more precise identification of inaccuracies, enhanced transparency, and reduced ambiguity during evidence retrieval.

Atomicity Each sub-claim should refer to a single factual unit within the original claim. This ensures if an error occurs, the inaccuracy can be precisely traced back to one or more specific sub-claims. Atomicity measures whether a sub-claim is truly atomic i.e. it focuses on only one relation between a subject and an object. For example, the claim "*The International Olympic Committee (IOC) was established on June 23, 1895, in Paris, France*" is not atomic, as it makes assertions about both the time and location of the IOC's establishment.

The decomposition process transforms a single claim into a list of sub-claims. It is crucial this

Correlation	*atomicity*		*sufficiency*		*fabrication*		*coverage*		redundancy	
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
LLM	0.40	0.39	0.14	0.09	**0.34**	**0.43**	0.43	0.45	**0.56**	**0.52**
Statistical	**0.58**	**0.59**	—	—	0.04	0.05	**0.61**	**0.58**	0.10	0.12

Table 1: Correlation of FactLens Evaluator scores with Human annotations on synthetic data (Readability is omitted due to its high subjectivity). r: Pearson Correlation Score; ρ: Spearman Correlation Score.

transformation is semantically equivalent, ensuring the combined list of sub-claims faithfully represents the original claim and that each can be independently verified. To address this aspect, we propose the metrics *Sufficiency*, *Fabrication*, and *Coverage*.

Sufficiency To perform fine-grained verification, each sub-claim needs to be independently verifiable. This requires the sub-claims to be properly contextualized to avoid any added ambiguity. *Sufficiency* measures whether the sub-claim is unambiguous and sufficiently contextualized with respect to the original claim. For example, in the original claim "*Amanda Bauer attended the University of Cincinnati. The school's nickname is Bearcats.*" , the sub-claim "*The school's nickname is Bearcats*" would be considered low in sufficiency because the reference to the school was omitted in the decomposition, making it ambiguous.

Fabrication The decomposition process must not introduce additional information or attempt to correct factual errors. This metric is especially important when evaluating LLM decomposers, as LLMs are known to suffer from hallucination or the generation of made-up information. For example, in the original claim "*Sydney, the capital of Australia, is known for its Opera House and Harbour Bridge*" , a sub-claim "*Sydney is the capital of New South Wales, Australia*" is considered fabrication. Similarly, with the source claim "*Net sales will reach 30 million if the growth rate in 2024 is the same as in 2023*" , a sub-claim "*The growth rate of 2024 is the same as in 2023*" is considered fabrication because it treats a condition as a claim.

Coverage The list of sub-claims must cover all factual assertions in the original claims, leaving no sub-claims missing. For instance, with the claim "*Amanda Bauer attended the University of Cincinnati, whose nickname is Bearcats*" , if only one sub-claim is generated as "*Amanda Bauer attended the University of Cincinnati*" , the coverage will be considered low because the assertion about the university's nickname is missing.

Additionally, some dimensions might not directly affect downstream verifiability and accuracy but capture some nice-to-have characteristics of the sub-claims.

Redundancy This metric measures whether the sub-claims, as a whole, contain redundant facts. When some sub-claims are semantically repetitive, the distribution of the fact-check units might be skewed. For example, if one erroneous sub-claim is repeated three times, the final judgment could shift from "mostly correct except for one sub-claim" to "more than half of the sub-claims were wrong." Furthermore, redundancy also introduces unnecessary costs in terms of time and computing resources.

Readability This metric assesses how readable the sub-claims are to the end-user and imposes a penalty on unnaturally formed sub-claims.

For each of these metrics, the sub claims are evaluated by assigning a score of 'low', 'medium' or 'high'. For *coverage* and *redundancy*, the scores are assigned at the claim level as we consider the `sub-claims` as a whole. For all other metrics, scores are assigned to each `sub-claim` and then aggregated.

`Ideal metric values`: For an ideal claim decomposition, we expect the `sub-claim` to possess high *atomicity*, high *sufficiency*, high *coverage* and high *readability,* while having low *fabrication* and low *redundancy*.

2.2 FactLens Evaluator

FactLens evaluator utilizes an ensemble method of LLM-generated evaluation scores and statistically computed scores (more details in Appendix B.1 and B.2 respectively). We use LLMs as evaluators due to their ability to scale well compared to human evaluators, as well as their reliability and knowledge across diverse domains. However, acknowledging the limitations of LLMs (Bavaresco et al., 2024; Stureborg et al., 2024), our statistical

scores rely on entity and semantic-based computations.

In Table 1, we report the correlation scores of human annotators with the FactLens Evaluator scores, on a synthetic data (more details in Appendix B.4 and C) that has been carefully curated to cover various types of sub-claim errors. We observe fair to moderate agreement across all dimensions between human evaluations and FactLens Evaluator scores, except for *sufficiency*. The moderate correlation scores can be attributed to the subjectivity involved in judging such metrics. The dependency on contextual information and evidence for assessing the *sufficiency* of a sub-claim contributes to the lower correlation scores for this metric. Nevertheless, our results demonstrate that our computation methods for the FactLens Evaluator align moderately well with human judgments on a dataset with varying sub-claim quality.

3 FactLens: Benchmarking Fine-grained Verification

3.1 Dataset Creation

The FactLens benchmark contains a dataset with ground-truth sub-claims and fine-grained labels. We use 733 instances from *CoverBench* (Jacovi et al., 2024), a fact-checking benchmark focused on complex claim verification sampled from diverse sources and domains, as the original claims.

We utilize two state-of-the-art LLMs — `GPT-4o` (OpenAI, 2024) and `LLaMA-3.1` (Meta, 2024) (details available in Appendix A) — to generate candidate sub-claims and measure the quality of these generations using the FactLens Evaluator. To ensure the high quality of the generated sub-claims, we engage human annotators (details provided in the Appendix C) to review all sub-claims and manually generate the ground-truth sub-claims, correcting any inaccuracies in the LLM-generated sub-claims (details in Appendix F).

To isolate the benefits of fine-grained verification, we do not perform the step of retrieving evidence or context for each sub-claim. Instead, we use the evidence and context provided in CoverBench, along with the generated sub-claims, to perform fact verification. This approach eliminates variability in the results that could arise from different methods and processes of evidence retrieval.

The next step in fine-grained verification involves using a 'verifier' model to fact-check each sub-claim against the provided evidence. In this work, we use `GPT-4o-mini` as our verifier model across all experiments to maintain consistency.

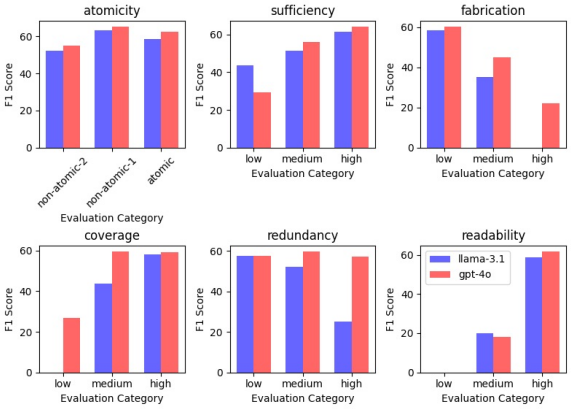

Figure 2: Impact of sub-claim quality metrics on verification performance. See Appendix E for details.

This verification method enables an enriched fine-grained evaluation. To compare the performance of fine-grained verification labels (for each sub-claim) with the holistic verification label, we aggregate the fine-grained labels as `false` if at least one of the fine-grained labels is also `false`; otherwise, the claim is considered `true`.

3.2 Claim Decomposition: Model Performance

We utilize the prompt defined in Table 3 to decompose claims using `GPT-4o` and `Llama-3.1(405B)`. In Table 2, we tabulate the evaluation performance of both these models on the claim decomposition task using FactLens Evaluator.

For each instance, we map the 'low', 'medium', 'high' scores ('non-atomic-2', 'non-atomic-1', 'atomic' for *atomicity*) to numerical values (1, 2, 3 respectively), and report the average for each metric across all 733 instances in the *CoverBench* dataset.

Both models have similar performance and perform well on the task, as per expected results of having sub-claims which are highly sufficient, low in fabrication, high in coverage, low in redundancy and possess high readability. The *atomicity* scores are far lower, as we qualitatively observe several instances of the type having one subject but multiple objects ('non-atomic-1' which is mapped to a score of 2)

Model	Atomicity ↑	Sufficiency ↑	Fabrication ↓	Coverage ↑	Redundancy ↓	Readability ↑
Llama-3.1	1.87	2.85	1.01	2.88	1.09	2.96
GPT-4o	1.82	2.85	1.02	2.89	1.15	2.95

Table 2: Measure of sub-claim quality using prompt-based contextualized decomposition. We report average scores on *CoverBench*. Up & down arrows indicate which metrics should ideally be high (~ 3) or low (~ 1) respectively.

3.3 Evaluation Results

In fine-grained verification, our FactLens evaluators can act as judges of generated sub-claims, providing early revision signals if the sub-claims might lead to problematic verifications. To illustrate this, we perform an end-to-end evaluation (as shown in Figure 2) to highlight how the final verification performance is affected by the quality of the sub-claims. For example, we expect high-quality sub-claims to exhibit low *fabrication* scores. We note that for claim decompositions with a *fabrication* score classified as 'low,' the downstream fact-checking performance i.e. F1-score, is higher compared to those with 'medium' or 'high' fabrication scores.

Similarly, we observe trends where sub-claims with higher *atomicity*, *sufficiency*, *coverage*, and *readability* scores demonstrate better verification performance. Although sub-claims with lower *redundancy* scores perform marginally better, the overall verification performance remains similar. This can be explained by the fact that highly redundant sub-claims may simply repeat claims without negatively impacting the final verification label.

4 Conclusion

In this paper, we introduce the benchmark FactLens to evaluate fine-grained claim verification, enriching existing benchmarks. We also identify important metrics for assessing the quality of fine-grained sub-claims and propose an automated evaluator to provide early signals of decomposition failures and evaluate claim decomposition approaches.

Limitations

Computation of Metrics We utilize two methods for computation of FactLens Evaluation metrics: LLM-based and statistically computed. Using LLMs as evaluators/judges is a research field being explored and improved continuously. However, existing works (Stureborg et al., 2024; Bavaresco et al., 2024) have highlighted the limitations of using LLMs in such evaluation tasks, with their scores being skewed and inconsistent.

To mitigate inconsistency, we provided specific instructions in the prompt (Table 4) to LLMs. We measured the agreement & correlation scores of LLMs with human judgement scores, observing fair-moderate agreement across most metrics. Furthermore, we propose our own definitions for computation of the FactLens Evaluator scores. However, we acknowledge the limitations in our computational approach as well, with it relying on the method for entity extraction, which may produce variable results. We aim to propose more concrete definitions for these metrics in future works.

Evidence Retrieval To ensure there is no variability in the fact-verification task, in this work we utilize the ground truth evidence which is present in the *CoverBench* dataset. This allows us to solely measure the dependency of fact-verification on the claim-decomposition and sub-claim quality. With our FactLens Benchmark, we provide motivation for fine-grained labels to soon be included across fact-verification benchmarks. In future works, we hope to show how claim decompositions may also improve the evidence search & retrieval process.

Fact Verification Models Previous works compared different verification models in the fact checking task. Tang et al. (2024a) contrast the performance of MNLI-based, LLM-based and their proposed fine-tuned models for fact-verification. In this work, we choose to use `GPT-4o-mini` as our only `verifier model`, as our aim is not to propose stronger models for verification; but to illustrate the benefits of fine-grained verification even using simpler off-the-shelf `verifier models`.

What are Facts? Fact extraction is a domain which still has a lot of room for improvement. Some previous works (Wang et al., 2024) also distinguish between factual claims, opinions and standard sentences. In this work, we utilize the *CoverBench* claims, which itself is obtained from benchmarks which rely on human or synthetic methods for claim generation. It can be noted that different types of factual claims may be generated with such a process, as some claims may center around

a claim that is universally true (eg. "Earth revolves around the Sun."), while other claims are dependent heavily on the context/evidence (eg. "There are 3 players whose home state is Missouri").

Moreover, factual claims in real world-scenarios can often be temporal and domain-dependent in nature. For example, a claim such as "*The legal drinking age is 18*" is false in a country such as the United States, however may be true in the United Kingdom, indicating domain dependence. Similarly, evidence retrieved for the claim "*Barack Obama is the President of the US*" is temporal in nature. We note that our results and experiments are also based on existing benchmarks, which do not account for all real-world scenarios.

References

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. Feverous: Fact extraction and verification over unstructured and structured information. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1).

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fern'andez, Albert Gatt, E. Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andr'e F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2024. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. ArXiv, abs/2406.18403.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In Advances in Neural Information Processing Systems, volume 33, pages 1877–1901. Curran Associates, Inc.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. Tabfact : A large-scale dataset for table-based fact verification. In International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia.

Farima Fatahi Bayat, Kun Qian, Benjamin Han, Yisi Sang, Anton Belyy, Samira Khorshidi, Fei Wu, Ihab Ilyas, and Yunyao Li. 2023. FLEEK: Factual error detection and correction with evidence retrieved from external knowledge. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 124–130, Singapore. Association for Computational Linguistics.

Alon Jacovi, Moran Ambar, Eyal Ben-David, Uri Shaham, Amir Feder, Mor Geva, Dror Marcus, and Avi Caciularu. 2024. Coverbench: A challenging benchmark for complex claim verification. arXiv preprint arXiv:2408.03325.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Wenliang Dai, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. ACM Computing Surveys, 55:1 – 38.

Jiho Kim, Sungjin Park, Yeonsu Kwon, Yohan Jo, James Thorne, and Edward Choi. 2023. FactKG: Fact verification via reasoning on knowledge graphs. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 16190–16206, Toronto, Canada. Association for Computational Linguistics.

Miaoran Li, Baolin Peng, Michel Galley, Jianfeng Gao, and Zhu Zhang. 2024. Self-checker: Plug-and-play modules for fact-checking with large language models. In Findings of the Association for Computational Linguistics: NAACL 2024, pages 163–181, Mexico City, Mexico. Association for Computational Linguistics.

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, Dublin, Ireland. Association for Computational Linguistics.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. Fine-grained fact verification with kernel graph attention network. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7342–7351, Online. Association for Computational Linguistics.

Meta. 2024. Llama-3.1.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 12076–12100, Singapore. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4o.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and

Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6981–7004, Toronto, Canada. Association for Computational Linguistics.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Lidén, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. ArXiv, abs/2302.12813.

Michael Schlichtkrull, Yulong Chen, Chenxi Whitehouse, Zhenyun Deng, Mubashara Akhtar, Rami Aly, Zhijiang Guo, Christos Christodoulopoulos, Oana Cocarascu, Arpit Mittal, James Thorne, and Andreas Vlachos. 2024. The automated verification of textual claims (AVeriTeC) shared task. In Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER), pages 1–26, Miami, Florida, USA. Association for Computational Linguistics.

Jiasheng Si, Yibo Zhao, Yingjie Zhu, Haiyang Zhu, Wenpeng Lu, and Deyu Zhou. 2024. CHECKWHY: Causal fact verification via argument structure. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15636–15659, Bangkok, Thailand. Association for Computational Linguistics.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators. ArXiv, abs/2405.01724.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024a. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 8818–8847, Miami, Florida, USA. Association for Computational Linguistics.

Nan Tang, Chenyu Yang, Ju Fan, Lei Cao, Yuyu Luo, and Alon Y. Halevy. 2024b. Verifai: Verified generative ai. In 14th Conference on Innovative Data Systems Research (CIDR 2024), Chaminade, HI, USA.

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, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. ArXiv, abs/2302.13971.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. 2024. Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 14199–14230, Miami, Florida, USA. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. ArXiv, abs/1904.09675.

A Claim Decomposition

We utilize few-shot prompting to decompose a claim into sub-claims. Table 3 shows the prompt used to capture the objective of generating sub-claims which are atomic, yet contextualized with enough information preserved from the original claim.

For the few-shot demonstrations we sample 4 instances from the FEVEROUS dataset (Aly et al., 2021), ensuring no overlap with the *CoverBench* data. Finally in the prompt, we randomly select 3 of the 4 demonstrations and shuffle the order, to ensure there is no bias.

We utilize `GPT-4o` and `Llama-3.1(405B)` models for this task, with `temperature` = 0.

B Evaluating Claim Decomposition

To evaluate the sub-claims, our FactLens Evaluator utilizes LLM-generated as well as statistically computed scores.

B.1 Prompt for Evaluating Claim Decomposition

In Table 4, we provide the prompt to LLMs used for evaluating the claim decompositions across the 6 metrics defined in Section 2.1. We provide clear instructions using which LLMs can judge the claim decompositions across different dimensions. For all metrics except "coverage" and "redundancy", the sub-claims are passed one at a time to obtain sub-claim level evaluation. "Coverage" and "redundancy" are used to judge the sub-claims as a whole, hence for these metrics, we provide the entire set of sub-claim decompositions for that instance. For "atomicity" we ask the LLM to output a label as

We aim to fact-check a textual claim. To make the fact-checking task simpler, we break down a claim into simpler, atomic sub-claims to fact-check as needed. Note that atomic sub-claims refer to unit claims within the original claim, that refer to a single concept that can be independently verified without having to refer to the original claim. Verification of the sub-claims should not require aggregation of facts or multi-hop reasoning over concepts. However, the sub-claim should have all the contextual information preserved from the original claim.

Your task is to break down a claim into atomic sub-claims for fact checking only if needed. If the original claim itself is a unit claim, do not break it down.

For example:
`{demonstrations}`

Note how each sub claim contains atomic information to fact check and is brief, yet is contextualized with all the information needed from the original claim.

Now find the sub claims from the following claim.
Claim: `{claim}`
Sub_Claims: `< your output in form of a list >`

==========

Table 3: Claim Decomposition Prompt

A factual claim can be broken down into atomic, yet contextualized sub-claims which makes it easier to fact check. You will be provided a claim, and one of the sub-claims which have been extracted from it. Your job is to evaluate this sub-claim on the following metric:

{metric}

Your answer should either be “low", “medium" or “high" based on the metric provided. Please be objective and fair in your evaluation.

Claim: {claim}
Sub-Claim: {sub_claim}
==========

The instructions to calculate each `metric` *is passed one at a time as follows:*

“atomicity”: If the sub-claim is atomic i.e. it is simple and centers around only one subject and one object, and the verification does not require aggregation of facts or multihop reasoning over concepts. Label the sub-claim as either “atomic” which denotes one subject and one object, or “non-atomic-1” which denotes one subject, multiple objects, or “non-atomic-2” which denotes multiple subjects

“sufficiency”: If the sub-claim itself is sufficient to be fact-checked without the need of any additional contextual information i.e. the sub-claim contains all the required contextual information to be fact-checked independently and is not ambiguous. Your answer should indicate whether the sub-claim has “low”, “medium” or “high” sufficiency.

"redundancy": If the sub-claims contain redundant or repeated information among them, i.e. multiple semantically equivalent sub-claims. Your answer should indicate whether the sub-claims have “low”, “medium” or “high” redundancy.

“coverage”: If the set of sub-claims cover all the facts and information made in the original claim. Your answer should indicate whether the sub-claims have “low”, “medium” or “high” coverage.

“fabrication”: If the sub-claim shows a degree of fabrication with respect to the original claim i.e. how much new information is added which was not present in the original claim. Note this is not to be judged according to the factuality of the original claims or sub-claims. Your answer should indicate whether the sub-claim has “low”, “medium” or “high” fabrication.

“readability”: If the sub-claim is readable to an end user. Your answer should indicate whether the sub-claim has “low”, “medium” or “high” readability.

Table 4: Prompt for Evaluating Claim Decompositions using LLMs

per specific instructions, while for each of the remaining metrics we prompt the LLM to judge the instance with a “low", “medium” or “high” score.

We utilize the smaller and cheaper OpenAI model `GPT-4o-mini` for this task, with `temperature` = 0.

B.2 Statistically Computed Evaluation

To compute the FactLens Evaluation metrics using statistical methods we rely on entity-based and semantic-based calculations. Given one instance, with claim C, we extract all the (Subject, Object) pairs present within it using `gpt-4o-mini`; `temperature = 0`, and from there create a list of S = subjects, and O = objects. After decomposing the claim, we obtain sub-claims $c = \{c_1, c_2, ..., c_n\}$. For all i in [1, n], we extract the subjects s_i and object o_i lists in a similar manner. We next follow these definitions to calculate the following metrics:

Atomicity To measure atomicity, we use an entity-based computation method of comparing the number of subjects and objects involved in the sub-claim c_i. If c_i revolves around only one

subject and one object eg. "*Kurt Cobain was a guitarist*", it is labeled 'atomic'.

if $len(s_i)$ = 1 and $len(o_i)$ = 1,
atomicity = 'atomic'

If c_i revoles around one subject, but multiple objects eg. "*Kurt Cobain was a guitarist and a singer*", it is labeled 'non-atomic-1'.

if $len(s_i)$ = 1 and $len(o_i)$ > 1,
atomicity = 'non-atomic-1'

However, if c_i revolves around multiple subjects eg. "*Kurt Cobain was a member of the band Nirvana, which was co-founded with Krist Novoselic*", it is labeled 'non-atomic-2'.

if $len(s_i)$ > 1,
atomicity = 'non-atomic-2'

Sufficiency As sufficiency is a tough metric to judge using semantic techniques, we rely on LLM evaluation scores.

Fabrication To calculate fabrication, we count all subjects in each s_i that do not appear in S i.e. subjects in the original claim, and all objects in each o_i that do not appear in O i.e. objects in the original claim.

If the count for fab is equal to 0, i.e. no new entities present in the sub-claims, the *fabrication* is 'low'. Based on thresholding values, we assign scores of 'medium' or 'high' *fabrication*.

Coverage To measure *coverage*, we check if the entities (subjects and objects) in all sub-claims c_i include all the subjects S and objects O present in the original claim.

if $\cup(s_i)\ \forall\ i = S$ and $\cup(o_i)\ \forall\ i = O$,
coverage = 'high'

In case there is no overlap, *coverage* = 'low', and for all other cases *coverage* = 'medium'

Redundancy To calculate redundancy, we use semantic-based technique by measuring BertScore (Zhang et al., 2019) between each pair of sub-claims. If there is high similarity between two

$$red = \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{I}(i \neq j, \text{BertScore}(c_i, c_j) > T)$$

, where T is a threshold value to find BertScore(.) similarity and n is the number of sub-claims generate for that instance.

Based on the number of redundant claims found, we assign scores of 'low', 'medium' and 'high'.

Readability We rely on LLM generated evaluations to measure readability.

B.3 FactLens Evaluation using Ensemble Method

As previously mentioned, in Table 1 we tabulate the correlation between Human scores and our LLM-generated & statistically computed scores on the synthetic with varying claim decomposition quality. Based on the results across the metrics, we propose to utilize the statistically computed scores for *atomicity* and *coverage* (as they are better correlated than the LLM-generated scores), while using LLM-generated evaluations for the rest of the metrics in our experiment results in Section 3.3.

B.4 Agreement Scores on Synthetic Data

In Table 1 we observed fair to moderate correlation between humans and FactLens Evaluator scores through Pearson and Spearman correlation scores. We provide the corresponding *p-values* in Table 5, from which we conclude the correlation scores for *atomicity*, *coverage* (both LLM-Human and Statistical-Human), *fabrication*, and *redundancy* (LLM-Human) are statistically significant.

In addition to the correlation scores in Table 1, we also report agreement scores between Human annotators and FactLensEvaluator. We report the `ordinal` Krippendorff's Alpha score to measure the agreement. We observe fair to moderate agreement across all dimensions except 'sufficiency', which can be attributed to the dependency on contextual information and evidence to judge *sufficiency* of a sub-claim.

The synthetic data is curated using 10 claims from the FEVEROUS benchmark and generating expert-annotated claim decompositions with perturbations. For each claim we generate 7 claim decompositions: one with perfect quality sub-claims, one LLM generated sub-claim, and others using perturbations resulting in lower quality sub-claims corresponding to each of the following 5 metrics: *atomicity*, *sufficiency*, *fabrication*, *coverage* and *redundancy*. We exclude *readability* in the agreement-scores, as it is an extremely subjective metric.

P-values	*atomicity*		*sufficiency*		*fabrication*		*coverage*		redundancy	
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
LLM	**4e-10**	**5e-10**	0.140	0.437	**5.1e-4**	**2.63e-7**	**2.9e-4**	**2.4e-4**	**3.26e-6**	**7.78e-6**
Statistical	**1e-9**	**1e-9**	—	—	0.166	0.176	**7.7e-8**	**1.1e-6**	0.404	0.332

Table 5: P-values from Table 1: Correlation of FactLens Evaluator scores with Human annotations on synthetic data

Metric	**Krippendorff's Alpha**
Atomicity	0.4421
Sufficiency	0.0486
Fabrication	0.4085
Coverage	0.5300
Redundancy	0.4240

Table 6: Alignment of FactLens Evaluator scores with Human annotations on synthetic data

C Expert Annotations

For human annotations on the synthetic data (Section 2.2) and the creation for the benchmark, we recruited two in-house expert annotators. The annotators are proficient in English, currently based in the United States of America, with at least a graduate-level degree. For the task, they were provided the same instructions as the prompt to LLMs in Table 4. The annotators were clearly explained the objective of the task and how their annotations would be utilized.

The inter-annotator agreement score (Krippendorff Alpha) is high, as tabulated in Table 7.

Metric	**Inter-Annotator Agreement**
Atomicity	0.73
Sufficiency	0.53
Fabrication	0.54
Coverage	0.86
Redundancy	0.94
Readability	0.96

Table 7: Inter-Annotator Agreement score

To measure correlation between the FactLens Evaluator scores and human scores, we do an average of FactLens Evaluator score on a metric with both the annotators, and repeat for all metrics. The human annotators were also used to generate ground-truth sub-claims (Section 3.1).

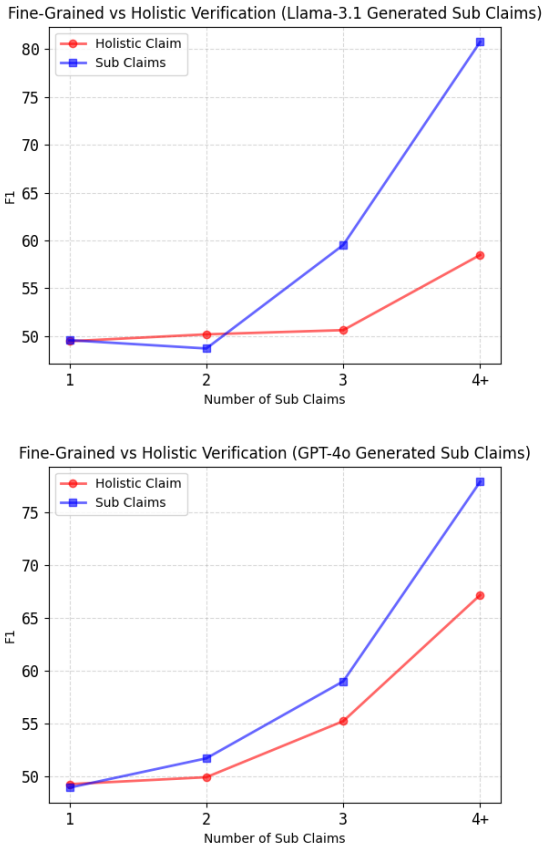

Figure 3: As complexity (i.e. number of sub-claims) increases, the performance of sub-claim decomposition significantly improves.

D Fine-Grained Verification

We study the benefits of fine-grained fact verification compared to verifying the whole claim in Figure 3.

In order to perform verification, we utilize `GPT-4o-mini` to judge if a claim is `true` or `false` based on the evidence provided. We obtain the ground truth evidence present in the *CoverBench* dataset.

In order to show the benefits of fine-grained verification, we compare it with the method of holistically verifying the original claim without decompositions.

In the first case, we simply pass the original

claim C along with the `evidence` to be verified. In the second case, we pass the claim's decompositions $c = \{c_1, c_2, ..., c_n\}$ one at a time. For each c_i we obtain a verification label, and then aggregate the labels for that instance. If any one sub-claim is judged `false` the whole instance is marked `false`, otherwise `true`.

We contrast the performance of the fine-grained verification with holistic verification in Figure 3. Here, we assume the number of sub-claims of an instance is indicative of how complex the claim is.

We observe that as the complexity (number of sub-claims) increase, the performance of the fine-grained verification method significantly increases compared to holistic verification.

E Impact of Sub Claim Quality on Verification

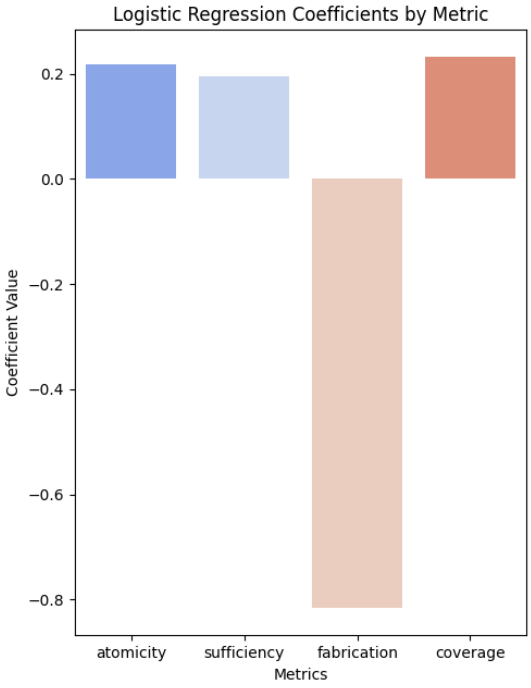

Figure 4: Fine-Grained Verification: Logistic Regression Coefficients for Metrics

In Figure 2, we showed how sub-claim quality impacts the end-to-end verification result. To truly understand the benefits of fine-grained decompositions and the FactLens metrics, we only consider those instances for which the number of sub-claims was greater than 1. Here, we illustrate further using qualitative examples and weights of a logistic regression model to show the influence of FactLens Evaluator Metrics on fine-grained verification.

To deeper understand how each metric influences the final verification, we fit a logistic regression model on the FactLens Evaluator scores on *CoverBench*. We specifically study the impact of the metrics *atomicity*, *sufficiency*, *fabrication* and *coverage* as we expect them to influence the final verification more than the "nice-to-have" metrics: *redundancy* and *readability*.

We also conducted an analysis to understand how the scores can collectively predict the final verification accuracy by fitting a logistic regression model and examining the coefficients associated with each metric. Combining the four metrics—atomicity, fabrication, coverage, and sufficiency—we achieved a prediction F1 score of 0.71, despite potential noise in the retrieval and verification steps.

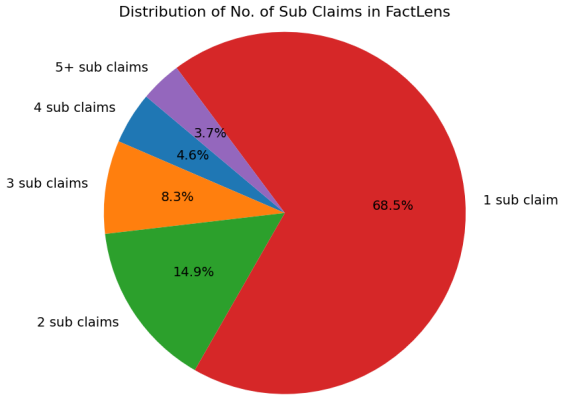

Figure 5: Distribution of Number of Sub Claims in FactLens Benchmark

From Figure 4, we observe that *fabrication* has the highest weight in magnitude, implying most influence in predicting whether the final label matches with ground truth. We see a negative weight for *fabrication* which is expected as lower *fabrication* indicates better quality sub-claims which in turn should have a positive effect on verification. *Atomicity*, *sufficiency* and *coverage* show a positive weight, as highly atomic, highly sufficient and high coverage sub-claims are expected to influence verification positively.

We also highlight some qualitative examples in Table 8 to show how sub-claim quality impacts

Claim: Fresh water crustaceans Aeglidae are classified as Malocostraca and Decapoda.

Gold Label: False

Evidence: [ground truth evidence from *CoverBench*...]

Sub Claims: ['Fresh water crustaceans Aeglidae are classified as Malocostraca', 'Fresh water crustaceans Aeglidae are classified as Decapoda']

FactLens Evaluation:

- *Atomicity*: ['atomic', 'atomic']
- *Sufficiency*: ['high', 'high']
- *Fabrication*: ['low', 'low']
- *Coverage*: 'high'
- *Redundancy*: 'low'
- *Readability*: 'high'

Fine-Grained Verification Labels: [True, False]
Aggregated Fine-Grained Verification Label: False

==========

Claim: In addition to co-starring in a Ken Ludwig musical, Jeffry Denman has worked with notables such as Mel Brooks, and has been called "a natural scene stealer" by The Houston Chronicle.

Gold Label: False

Evidence: [ground truth evidence from *CoverBench*...]

Sub-Claims: ['Jeffry Denman co-starred in a Ken Ludwig musical', 'Jeffry Denman has worked with Mel Brooks', 'Jeffry Denman has been called a natural scene stealer by The Houston Chronicle']

FactLens Evaluation:

- *Atomicity*: ['atomic', 'atomic', 'non-atomic-2']
- *Sufficiency*: ['high', 'high', 'high']
- *Fabrication*: ['low', 'low', 'low']
- *Coverage*: 'medium'
- *Redundancy*: 'low'
- *Readability*: ['high', 'high', 'high']

Fine-Grained Verification Labels: [True, True, True]
Aggregated Fine-Grained Verification Label: True

Table 8: Examples of how sub claim quality impacts verification performance

fine-grained verification. In the first instance, with perfect sub-claim quality, the fine-grained verification correctly predicts the ground truth label. In the second instance, we see *coverage* as 'medium' and imperfect *atomicity* score whereby the verifier eventually predicts an incorrect label.

F FactLens Dataset Characteristics

Our FactLens benchmarks consists of 733 instances from *CoverBench* with ground truth decompositions curated using LLMs and humans, and fine-grained labels as mentioned in Section 3.1. In Figure 5 we note the distribution of the number of sub-claims in the dataset.