Better Together: Towards Localizing Fact-Related Hallucinations using Open Small Language Models

David Kletz[†*], **Sandra Mitrović**[†*], **Ljiljana Dolamić**[‡], **Fabio Rinaldi**[†]
[†] SUPSI, IDSIA, Switzerland
[‡] armasuisse, Science & Technology, Switzerland
{david.kletz, sandra.mitrovic, fabio.rinaldi}@supsi.ch
ljiljana.dolamic@armasuisse.ch

Abstract

In this paper, we explore the potential of Open-source Small Language Models (OSLMs) for localizing hallucinations related to factual accuracy. We first present Lucifer, a dataset designed to enable proper and consistent evaluation of LMs, composed of an automatically constructed portion and a manually curated subset intended for qualitative analysis. We then assess the performance of five OSLMs using four carefully designed prompts. Results are evaluated either individually or merged through a voting-based merging approach. While our results demonstrate that the merging method yields promising performance even with smaller models, our manually curated dataset highlights the inherent difficulty of the task, underscoring the need for further research.

1 Introduction

The task of factual hallucination detection is inherently complex, requiring models to integrate and coordinate multiple capabilities—ranging from linguistic fluency to factual verification. While several recent studies addressed this challenge using Large Language Models (LLMs) (Dhuliawala et al., 2024; Manakul et al., 2023; Min et al., 2023; Li et al., 2024a), the use of closed-source LLMs, in particular, imposes substantial computational costs, while raising privacy and ethical considerations (Huang et al., 2022; Carlini et al., 2023; Weidinger et al., 2021), lacking transparency (Sun et al., 2022; Manakul et al., 2023) and trustworthiness (Lee et al., 2022; Mitrović et al., 2025).

Moreover, the literature on *fact-verification* and *factual hallucinations* typically aims at identifying whether or not the phrase contains factual hallucination (Li et al., 2024a; Thorne et al., 2018). In this work, we shift our focus instead to localizing factual hallucinations by identifying *exact hallucination spans*. Moreover, we explore an alternative approach: empowering LMs with an additional factual prompt and leveraging the collective behavior of multiple Open-source Small Language Models (OSLMs) to address the localized factual hallucination detection task. Rather than benchmarking different merging strategies, we investigate whether combining different OSLMs and different prompts, can achieve competitive performance with respect to a specifically fine-tuned method (Shan et al., 2025). Unfortunately, most current datasets are either not span-oriented (e.g. Poly-FEVER, Zhang et al. 2025, FactCHD, Chen et al. 2024), require additional retrieval and/or training (e.g. FAVA, Mishra et al. 2024, ANAH Ji et al. 2024) or provide hallucination spans (Mu-SHROOM 2025, Vazquez et al. 2025) but suffer from unclear or inconsistent annotations (Mitrović et al., 2025), rendering them unsuitable for fine-grained assessment. To address this gap, we construct two purpose-built datasets[1] specifically designed to evaluate the ability of models to detect factual hallucinations within a phrase. Since exact span annotation is tedious (automation by closed-source LLMs is costly and still unreliable, while human annotation is time-consuming and hard to reach consensus), we resort to claims (known to be either true or false) as a reliable ground truth. More specifically, our datasets (one automatically- and one manually-constructed) are composed of phrases that combine verifiably true claims and explicitly false claims, to simulate a hallucination. This setup permits detecting *reliable* hallucination span within a phrase by using the span of the explicitly false claim. It also allows for performing more coarse-grained, claim-level hallucination detection. All source claims are drawn from the FEVER 2018 dataset (Thorne et al., 2018), ensuring that they are

[*]These authors contributed equally to this work.

[1]This dataset is available at https://github.com/IDSIA-NLP/lucifer

Proceedings of the 1st Workshop on Confabulation, Hallucinations and Overgeneration in Multilingual and Practical Settings (CHOMPS 2025), pages 20–34

fact-checked and verifiable. The manual dataset is curated to seamlessly incorporate underlying facts simulating more subtle and natural hallucinations, thereby increasing the difficulty of localization.

We find that OSLMs can perform surprisingly well, especially when paired with carefully designed prompts. Furthermore, we show that combining model outputs through a simple voting-based merging strategy significantly improves overall accuracy, making it a practical approach for hallucination detection in resource-constrained settings.

2 Related Works

Given the popularity of the topic, this work relates (and might seemingly relate) to different existing studies. However, substantial differences are evident across multiple dimensions.

First, hallucination task has been addressed by different works in the past, but most of them focus on sentence level detection (binary classification task), such as SelfCheckGPT (Manakul et al., 2023) and SAPLMA (Azaria and Mitchell, 2023).

Second, concerning the data, as already mentioned, most of the current datasets are adapted for binary classification, thus containing the labels only on a phrase level. Factually aligned datasets with a more fine-grained information, such as knowledge base triplets, might look as a viable alternative. However, after careful investigation we discovered that even the largest of them, T-REx (Elsahar et al., 2018), containing alignments between Wikidata Triples knowledge base and Wikipedia Abstracts, remains inadequate for this work. This is mainly because, despite its volume, texts within T-REx tend to contain very few and also fairly simple relationships. While Lucifer(-M) also relies on the given facts, it aims at identifying hallucinations spans which do not directly coincide with the entities and/or relationship of a given fact but that can rather be (indirectly) deduced from these (see Example 1 in the Table 9). Finally, to the best of our knowledge, the only truly span-annotated existing dataset is the one released for the Mu-SHROOM 2025 challenge. However, as pointed by Mitrović et al. (2025) and Huang et al. (2025), this dataset contains inconsistent annotations.

Third, this work exploits an ensemble-like approach to merge the results of five LLM annotators in the post-inference phase. While different LLM post-inference ensembling approaches exist in the literature, we left these out for different reasons. More specifically, being in an unsupervised setup, all supervised methods and methods requiring training data (e.g. Tekin et al. (2024)) were directly ruled out as well as approaches requiring repetitive inferences by a single model (Li et al., 2024b) or based on human judgement. Moreover, it is worth emphasising that our scope was to examine whether OSLMs could perform nearly as good as proprietary LLMs using a straightforward ensemble approach, rather then benchmarking different ensemble approaches. Finally, the main difference with respect to the approach of Mitrović et al. (2025) are three-fold. First, we do not exclude any model-prompt variant; second, we perform both prompt-level merging as well as model-level merging; third, we introduce additional, fact-based prompt.

3 Turning a fact-verification dataset to a fact-linked hallucination detection dataset

Due to aforementioned drawbacks of existing datasets for factual hallucination detection, we opt to create a dataset mixing the FEVER 2018 dataset's true and false claims. These combinations simulate factual hallucinations by blending factual and non-factual information in a seamless way.

3.1 Lucifer-A

The first idea was to automatically construct a dataset Lucifer-A(utomatic), relying solely on LLMs. Initial attempts at making LLM autonomously blend claims into seemingly human-like sentences have not yielded satisfactory results (see Appendix C). We, therefore, opted for a semi-automated pipeline: for each instance one true and one false claim is selected, in randomized order. These claims are then fused into a single sentence using the Open-source model DeepSeek-Qwen-1.5B (see Appendix A for model abbreviations), forcing the model to employ the connector "while" to create a fluent and syntactically correct sentence.

An example of Lucifer-A instance generation
Claim 1 (F): The Taj Mahal attracts significantly less than 7-8 million visitors a year.
Claim 2 (T): Reds was produced by Warren Beatty.
DSQ-1.5 capitalization answer: Yes.
Final sentence (FT): The Taj Mahal attracts significantly less than 7-8 million visitors a year while Reds was produced by Warren Beatty.

To ensure grammatical correctness, particular attention was paid to ensure appropriate capitalization, as we spotted that the model was making mistakes when left unattended in this regard (see Appendix C for more details). In total, we produce 1,000 automatically constructed sentences.

3.2 Lucifer-M

In addition to the automatically generated dataset, we construct a smaller, entirely manually curated subset. This involves selecting pairs of claims that share a common subject or thematic link. Once the claim pairs are selected, they are manually rewritten to integrate both pieces of information in a more natural and contextually sophisticated way. This dataset is hence more challenging than *Lucifer-A* since the claims are interwoven rather than presented sequentially or by simple concatenation. The result is a set of 100 sentences (see Appendix B for basic statistics and Appendices D and E for details on manual effort regarding dataset construction and annotation, respectively).

> An example of Lucifer-M instance generation
> **Claim 1 (T)**: The Eagles broke up in 1980.
> **Claim 2 (T)**: Mao Zedong died in 1976.
> **Final sentence (FT)**: The Eagles broke up in 1976, the same year that Mao Zedong died.

4 Methodology

To perform the annotations, we use five different models, each evaluated with four distinct prompts.

The models employed are: *Osiris*, *DeepSeek-Qwen*, *DeepSeek-Llama*, *Ministral*, and *Mistral* (see Appendix A for the abbreviations used to refer to each model). With respect to the prompts, we begin by adapting the three prompts introduced by Mitrović et al. (2025). The modified prompt versions (referred to as *v1*, *v2*, and *v3*) are provided in Appendix H.

In addition, we introduce a fourth prompt (denoted as *fact* prompt and also available in Appendix H), designed to shift the model's task from annotation to direct correction of hallucinations. This prompt focuses specifically on factual hallucinations, which are the core concern of this study.

Finally, following an idea also proposed by Mitrović et al. (2025), we explore merging annotation systems through prediction merging. Specifically, we implement a voting mechanism across the outputs of multiple model–prompt pairs. For each claim in a sentence (two claims per sentence), we count how many model–prompt pair consider it a hallucination. If a majority of the models (>50%) identify a claim as hallucinated, it is marked as such in the final merged output.

5 Evaluation

The first evaluation is conducted at the level of individual claims (remember that each claim can be either true or false). An ideal system must leave a true claim unaltered while correctly annotating a false one. For our manual dataset, we consider that claim is unaltered if its semantics has not been changed.

We aim at assessing whether an annotation system is effective in identifying a substantial number of hallucinations, while minimizing false positives. To this end, we employ the standard metrics of recall and precision. Additionally, to account for correctly identified non-hallucinated claims, we include accuracy as a complementary metric.

The results are primarily derived from *model+prompt* pairs, each consisting of a specific model and its corresponding prompt. Additionally, we explore combining the outputs of multiple such pairs using a merging system to improve overall performance.

We then conduct an evaluation at the sentence level. The objective here is to determine whether an annotation system can accurately distinguish between hallucinated and non-hallucinated claims when they co-occur within the same sentence. In this setting, we measure the percentage of sentences in which both claims are annotated correctly.

6 Lucifer-A Results

6.1 Claim-level evaluation

We present the results at the claim level in Tables 1a (recall), 1b (precision), and 1c (accuracy).

We observe that the top-scoring pairs vary by evaluation metric. For example, *Ministral* identifies nearly all hallucinations with prompt *v1* (recall = 0.99), while *Qwen2.5-Osiris-7B-Instruct* detects only 41% of them using the same prompt. Conversely, when it comes to precision, *Ministral* achieves the highest score (0.81) with the *fact* prompt, but performs poorly with prompt *v1* (precision = 0.53). This supports the hypothesis that *Ministral*'s high recall with prompt *v1* is due to its tendency to over-identify hallucinations, as reflected in its lower precision.

In terms of accuracy, the best results, ranging from 0.70 to 0.75, are achieved through merging ap-

Prompt	Osi	DSQ	DSL	Min	Mis	mer.
fact (f)	0.73	0.23	0.56	0.56	0.83	0.65
v1	0.41	0.86	0.77	**0.99**	0.94	0.94
v2	0.71	0.60	0.70	0.48	0.71	0.72
v3	0.49	0.73	0.76	0.92	0.96	0.88
f+v2+v3	0.67	0.50	0.72	0.73	0.88	0.74

(a) Recall

Prompt	Osi	DSQ	DSL	Min	Mis	mer.
fact (f)	0.65	0.52	0.67	**0.81**	0.66	0.77
v1	0.59	0.56	0.53	0.53	0.53	0.56
v2	0.61	0.54	0.59	0.66	0.61	0.64
v3	0.56	0.55	0.57	0.56	0.53	0.56
f+v2+v3	0.61	0.54	0.66	0.75	0.62	0.65

(b) Precision

Prompt	Osi	DSQ	DSL	Min	Mis	mer.
fact (f)	0.67	0.51	0.65	0.72	0.71	0.73
v1	0.57	0.60	0.55	0.55	0.56	0.60
v2	0.63	0.54	0.61	0.62	0.63	0.66
v3	0.56	0.58	0.59	0.61	0.57	0.60
f+v2+v3	0.62	0.54	0.68	**0.75**	0.67	0.67

(c) Accuracy

Table 1: Scores for each model-prompt pair. Column “**mer.**” : recall by merging the annotations from each line. Line “**f+v2+v3**” : recall by merging annotations from fact, v2, and v3 prompts. Notation: **Osi**: Qwen2.5-Osiris-7B-Instruct; **DSQ**: DeepSeek-Qwen; **DSL**: DeepSeek-Llama; **Min**: Ministral; **Mis**: Mistral.

proaches. These include both prompt merging (e.g., using the outputs from the *fact* prompt) and model merging (e.g., aggregating outputs from *Ministral*).

Although merging systems may not produce the highest individual scores for recall or precision, they avoid the *pitfalls* that can inflate metrics: they neither over-annotate (labeling too many non-hallucinated claims) nor under-annotate (missing many actual hallucinations). This supports our intuition that model collectives can mitigate the individual errors of *model+prompt* pairs by flagging as hallucinations only those claims for which at least a partial consensus emerges.

Notably, prompts of type *fact* consistently yield better performance than prompts from *v1*, *v2*, or *v3*. This finding validates our strategy of asking models to directly correct hallucinations rather than merely annotate them. More broadly, it aligns with our hypothesis that generating a coherent textual sequence is easier for the model than producing a hybrid output that combines text and meta-text (annotations).

6.2 Sentence-level evaluation

The results of the full-sentence level evaluation are presented in Table 2.

Prompt	Osi	DSQ	DSL	Min	Mis	mer.
Facts (F)	0.5	0.15	0.36	0.53	0.51	**0.56**
v1	0.32	0.30	0.20	0.11	0.17	0.24
v2	0.43	0.21	0.32	0.37	0.43	0.45
v3	0.28	0.27	0.33	0.29	0.17	0.30

Table 2: Sentence-level accuracy. Notation: **Osi**: Osiris; **DSQ**: DeepSeek-Qwen; **DSL**: DeepSeek-Llama; **Min**: Ministral; **Mis**: Mistral.

This evaluation setting is more stringent, leading to lower overall performance across models. This suggests that models are more prone to errors when explicitly tasked with identifying hallucinations within an entire sentence.

Once again, the best performance is achieved using the merging approach, highlighting its effectiveness in aggregating predictions from multiple prompts and models.

Finally, it is worth noting that our evaluation is more fine-grained than that of Shan et al. (2025), as it involves distinguishing between hallucinated and non-hallucinated content within the same sentence. This added complexity may partly explain why *Osiris* performs worse in our setting compared to the results originally reported by its authors.

7 Lucifer-M results

Manually verified results obtained on the *Lucifer-M* dataset using the same claim-level evaluation criteria can be seen in Table 3. We can see that both DeepSeek models (*DeepSeek-Qwen* and *DeepSeek-Llama*) lag behind the competitors as they have the highest number of both claim misses (24 and 22, respectively) and the lowest number of both claim hits (29 and 34, respectively). Thanks to the opposed ratio between 1-claim and 2-claim hits (and low ratio of misses), *Mistral* undoubtedly outperforms all its competitors. Additionally, it can be observed that, in general, models tend to act correctly on at least one claim within the sentence (see Table 5). However, when considering sentence-level evaluation (percentage of sentences with both claims annotated correctly) the accuracy is quite modest. In fact, even the best performing model *Mistral* is correctly annotating just 48% of sentences. This is due both to the difficulty of the sentences in *Lucifer-M* as well as the severity of this type of the evaluation.

Num. Corr. Claims	Osi	DSQ	DSL	Min	Mis
0	15	24	22	19	**12**
1	41	47	44	37	40
2	42	29	34	42	**48**
Sent.-level acc.	0.43	0.29	0.34	0.43	**0.48**

Table 3: Number of sentences with 0, 1 or 2 correct claims per model, using factual prompt on Lucifer-M. Last row: sentence-level accuracy based on 2 correct claims, per model. Notation: **Osi**: Osiris; **DSQ**: DeepSeek-Qwen; **DSL**: DeepSeek-Llama; **Min**: Ministral; **Mis**: Mistral.

Combination	Osi (%)	DSQ (%)	DSL (%)	Min (%)	Mis (%)
FF	48	16	12	48	64
FT	36	8	44	20	32
TF	60	8	24	48	68
TT	24	84	56	52	28

Table 4: Percentages of correctly processed **sentences** per combination (FF/FT/TF/TT) per model, using factual prompt on Lucifer-M. Note that the number of sentences per combination is 25. Notation: **Osi**: Osiris; **DSQ**: DeepSeek-Qwen; **DSL**: DeepSeek-Llama; **Min**: Ministral; **Mis**: Mistral.

Table 4 shows the percentage of correctly processed sentences per combination (FF/FT/TF/TT) per model (note that in *Lucifer-M*, each combination is represented by 25 instances). We can see that model performance depends on combination, especially for *DeepSeek-Qwen* whose correctness varies from 8% on FT/TF to 84% on TT combinations. Surprisingly, other models exhibit different performances on FT and TF combinations, with *Mistral* being the most extreme with correctness of 32% on FT and 68% on TF combinations.

In Table 5, we illustrate the evaluation on an easy instance within the dataset. Note that on claim-level we are interested in identifying claim presence (if true) and absence (if false), hence as long as the false claim is absent it does not matter what it was substituted with (see that both "Birmingham" and "London" for Led Zeppelin are considered as correct). This follows from our focus on factual hallucination *detection* and not fact *correction*.

As shown in Table 6, the changes the models occasionally make can render the evaluation more complicated. We provide more details on challenges and limitations related to evalution in Appendix F.

Source	Lucifer-M sentence	Score
Input (FF)	The Beatles were formed in London while Led Zeppelin was formed in Alaska.	-
Osi	The Beatles were formed in Liverpool while Led Zeppelin was formed in Birmingham.	2
DSQ	The Beatles were formed in London while Led Zeppelin was formed in the United States.	1
DSL	The Beatles were formed in London while Led Zeppelin was formed in London.	1
Min	The Beatles were formed in Liverpool while Led Zeppelin was formed in London.	2
Mis	The Beatles were formed in Liverpool while Led Zeppelin was formed in London.	2

Table 5: Evaluation illustration on an easy instance. Notation: **Osi**: Osiris; **DSQ**: DeepSeek-Qwen; **DSL**: DeepSeek-Llama; **Min**: Ministral; **Mis**: Mistral.

Source	Lucifer-M sentence
Input (FT)	Vincent van Gogh is from Slovenia, which is bordered by the Adriatic Sea.
DSL	Vincent van Gogh was a Dutch artist, which is bordered by the Adriatic Sea.
All others	Vincent van Gogh is from the Netherlands, which is bordered by the North Sea.
Ideal answer	*Vincent van Gogh is **not** from Slovenia, which is bordered by the Adriatic Sea.*

Table 6: Evaluation illustration on a difficult instance. Notation: **Osi**: Osiris; **DSQ**: DeepSeek-Qwen; **DSL**: DeepSeek-Llama; **Min**: Ministral; **Mis**: Mistral.

8 Conclusion

In this paper, we addressed the challenge of using Open-source Small Language Models (OSLMs) for hallucinations detection, with a specific focus on the precise hallucination spans. To facilitate this, we constructed a novel dataset composed of both true and false claims. We then evaluated the ability of five OSLMs, using four different prompts, to detect hallucinations in generated sentences. Additionally, we explored merging methods by aggregating predictions through a voting mechanism. The results proved promising, demonstrating that OSLMs are reasonably effective at detecting hallucinations. A complementary qualitative analysis confirmed the relative robustness of these models in identifying erroneous content. However, further investigation is needed for more sophisticated and linguistically complex examples. Additionally, our findings highlight a critical aspect: the correction suggestions proposed by the models are not consistently reliable and should be interpreted with caution.

Limitations

The evaluation method we adopted is primarily based on sequence comparison and assumes that an LM is fully capable of following its prompt—modifying a claim if and only if it considers the claim to be a hallucination. However, this assumption does not always hold in practice, which introduces some noise into the evaluation results (see Appendix F for more details).

Moreover, our analysis focuses on a single type of hallucination. While our prompts were designed to be broadly applicable, a more comprehensive study would be needed to assess the model's ability to detect all forms of hallucination.

Acknowledgments

The work described in this paper has been supported by the "*Language Preference and Hallucination Investigation in Multilingual RAG (LPHI-mRAG)*" project, funded by armasuisse S&T, Switzerland.

References

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Xiang Chen, Duanzheng Song, Honghao Gui, Chenxi Wang, Ningyu Zhang, Yong Jiang, Fei Huang, Chengfei Lyu, Dan Zhang, and Huajun Chen. 2024. Factchd: benchmarking fact-conflicting hallucination detection. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3563–3578, Bangkok, Thailand. Association for Computational Linguistics.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-REx: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. 2022. Are large pre-trained language models leaking your personal information? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2038–2047, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sicong Huang, Jincheng He, Shiyuan Huang, Karthik Raja Anandan, Arkajyoti Chakraborty, and Ian Lane. 2025. UCSC at SemEval-2025 task 3: Context, models and prompt optimization for automated hallucination detection in LLM output. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1981–1992, Vienna, Austria. Association for Computational Linguistics.

Ziwei Ji, Yuzhe Gu, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen. 2024. ANAH: Analytical annotation of hallucinations in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8135–8158, Bangkok, Thailand. Association for Computational Linguistics.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. In *Advances in Neural Information Processing Systems*, volume 35, pages 34586–34599. Curran Associates, Inc.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024a. The dawn after the dark: An empirical study on factuality hallucination in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10879–10899, Bangkok, Thailand. Association for Computational Linguistics.

Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. 2024b. More agents is all you need. *Transactions on Machine Learning Research*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

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.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. In *First Conference on Language Modeling*.

Sandra Mitrović, Joseph Cornelius, David Kletz, Ljiljana Dolamic, and Fabio Rinaldi. 2025. Swushroomsia at SemEval-2025 task 3: Probing LLMs' collective intelligence for multilingual hallucination detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1810–1827, Vienna, Austria. Association for Computational Linguistics.

Sandra Mitrović, Matteo Mazzola, Roberto Larcher, and Jérôme Guzzi. 2025. Assessing the trustworthiness of large language models on domain-specific questions. In *Progress in Artificial Intelligence*, pages 305–317, Cham. Springer Nature Switzerland.

Alex Shan, John Bauer, and Christopher D. Manning. 2025. Osiris: A lightweight open-source hallucination detection system. *Preprint*, arXiv:2505.04844.

Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. 2022. Black-box tuning for language-model-as-a-service. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 20841–20855. PMLR.

Selim Furkan Tekin, Fatih Ilhan, Tiansheng Huang, Sihao Hu, and Ling Liu. 2024. LLM-TOPLA: Efficient LLM ensemble by maximising diversity. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11951–11966, Miami, Florida, USA. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Raul Vazquez, Timothee Mickus, Elaine Zosa, Teemu Vahtola, Jörg Tiedemann, Aman Sinha, Vincent Segonne, Fernando Sanchez Vega, Alessandro Raganato, Jindřich Libovický, Jussi Karlgren, Shaoxiong Ji, Jindřich Helcl, Liane Guillou, Ona De Gibert, Jaione Bengoetxea, Joseph Attieh, and Marianna Apidianaki. 2025. SemEval-2025 task 3: Mu-SHROOM, the multilingual shared-task on hallucinations and related observable overgeneration mistakes. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2472–2497, Vienna, Austria. Association for Computational Linguistics.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, and 4 others. 2021. Ethical and social risks of harm from language models. *Preprint*, arXiv:2112.04359.

Hanzhi Zhang, Sumera Anjum, Heng Fan, Weijian Zheng, Yan Huang, and Yunhe Feng. 2025. Polyfever: A multilingual fact verification benchmark for hallucination detection in large language models. *Preprint*, arXiv:2503.16541.

A Models used and abbreviations

A list of all the LMs used and the respective abbreviations we have used to designate them is available in Table 7.

B Datasets

Dataset	Combination (TT/TF/FT/FF)	Number of instances
Lucifer-A	TF	500
	FT	500
Lucifer-M	TT	25
	TF	25
	FT	25
	FF	25

Table 8: Number of instances per combination (TT/TF/FT/FF) in two datasets.

C Lucifer-A Dataset Construction: Initial Attempts and Capitalization Efforts

Our first approach used gpt-4o-mini to merge two randomly selected claims into a fluent sentence (see Appendix H, Prompts 1.1, 1.2, 1.3, 1.4 and 2.1, 2.2). However, the model often modified the original claims during the merging process, which undermined our ability to control for factuality—making the output unusable.

(1) *Claim 1:* Farrah Fawcett acted in Saturn 3.
Claim 2: Princess Agents is based on work by 7 golden-age science fiction authors.
Sentence generated: Farrah Fawcett, known for her role in Saturn 3, has often been discussed alongside various adaptations, including Princess Agents, which is said to draw inspiration from the works of seven golden-age science fiction authors.

Next, we provided the model with 10 true and 10 false claims and asked it to select two compatible claims to merge smoothly (see Appendix H,

LM	Abbreviation
Qwen2.5-Osiris-7B-Instruct	Osiris
DeepSeek-R1-Distill-Qwen-1.5B	DeepSeek-Qwen1.5
DeepSeek-R1-Distill-Qwen-7B	DeepSeek-Qwen
DeepSeek-R1-Distill-Llama-8B	DeepSeek-Llama
Ministral-8B-Instruct-2410	Ministral
Mistral-7B-Instruct-v0.3	Mistral

Table 7: List of used models and their corresponding abbreviations.

Prompt 3). Unfortunately, the model often produced sentences that simply juxtaposed the claims using generic connectors like "even though", resulting in minimal semantic integration and limited hallucination effect.

(2) Kim Kardashian was one of 2015's 100 most influential people, despite England not being first inhabited by modern humans during the Upper Palaeolithic period.

To ensure grammatical correctness, particularly regarding capitalization, we query the model with the first letter of the second claim, asking: "Should the first letter remain capitalized even if it's not at the beginning of the sentence?" (see Appendix H, Prompt 3). If the response is anything other than a clear "yes" or "no", the pair is discarded. The final sentence is constructed by applying the appropriate capitalization and joining the claims.

D Lucifer-M Dataset: Construction

As mentioned, creation of the *Lucifer-M* dataset involved: first, the selection of claim pairs and second, putting them together in a coherent, syntactically and semantically meaningful sentences. Two human annotators were involved in this process. First, both annotators agreed on the guideline for dataset creation (see *Lucifer-M construction* guidelines below). Next, one annotator was in charge of creating the dataset according to the agreed guidelines. The other annotator then independently performed verification of created instances. During this inspection some instances were found to be potentially challenging for an LLM (see some examples in Table 9). However, the annotators agreed to keep them in the final dataset for two reasons. First, these instances represent what could be considered as a perfectly natural human-generated sentences. Additionally, they underpin the motivation behind the manual dataset generation, that is, to have more natural (and more complex) hallucinations, instead of just a phrase with two isolated claims artificially connected. In this sense, even though of a limited size, *Lucifer-M* contains valuable cases for testing hallucinations.

Lucifer-M *construction* guidelines

- Each instance should be based on two claims from the FEVER 2018 dataset
- Either claim can be true or false: all combinations should be taken into account, including having both true claims, given that this is typically missing in hallucination datasets.
- "While" can be used as a claim connector, however, in a very limited number of cases, given that it is already heavily exploited in Lucifer-A.
- Instead, claims should ideally be selected on the basis of a common point (e.g. the same subject, the same topic) which can be exploited to make non-trivial and more natural hallucinations, involving subtle distortions rather than isolated false claim(s).

Lucifer-M *annotation* guidelines

- Each instance should be annotated at the claim level, hence the output on the instance level is one of 00, 01, 10, 11.
- On a claim level, 0 is assigned if claim is not treated correctly by a model, while 1 is assigned if the model correctly processed the claim.
- To determine if the claim is processed correctly, first verify the initial correctness of the claim (remember that each claim can be either true or false). An ideal system must leave a true claim unaltered while correctly annotating a false one.

E Lucifer-M Dataset: Annotation

Three annotators independently performed claim level annotations (hence, assigning one of 00, 01, 10, 11 per phrase) on the half of *Lucifer-M* dataset. The Fleiss's kappa score showed substantial inter-

Example	Lucifer-M phrase (correctness)	Original Claims from FEVER 2018
1	Paradise was given in 2012, two years before Selena Gomez starred in Spring Breakers. (TF)	Claim 1 (T): Paradise was given in 2012. Claim 2 (T): Selena Gomez starred in the 2013 film Spring Breakers.
2	The same man who founded the most populous city in Trinidad and Tobago also surrendered the island of Trinidad in 1789. (TF)	Claim 1 (T): The most populous city in Trinidad and Tobago was founded by José María Chacón. Claim 2 (F): José María Chacón surrendered the island of Trinidad in 1789.
3	Halle Berry does not have a child with Olivier Martinez but with Gabriel Aubry. (FT)	Claim 1 (T): Halle Berry has a child by Olivier Martinez and one with Gabriel Aubry. Claim 2 (T): Gabriel Aubry and Halle Berry have a child.
4	Most Albanians are Buddhist and the remaining minor are Sunni Muslim. (FF)	Claim 1 (F): Most Albanians are Buddhist. Claim 2 (T): Most Albanians are Sunni Muslim.
5	Since leukemia has to do with a lack of normal blood cells, Marshall McLuhan predicted his own death. (TF)	Claim 1 (T): Leukemia involves a lack of normal blood cells. Claim 2 (F): Marshall McLuhan predicted his own death.

Table 9: Examples of challenging phrases in Lucifer-M (as evaluated by one of the annotators). Note that with example 1, model needs to understand the difference between years, while with example 4 it needs to understand the problem with “minor” given that Sunni Muslim are majority in Albania.

annotator agreement per *DeepSeek-Llama* (0.63), *DeepSeek-Qwen* (0.65) and *Ministral* (0.63), while for *Mistral* (0.26) and *Osiris* (0.27) it was rather fair.

F Evaluation: Challenges and Limitations

We identify two categories of limitations in our evaluation methodology: those inherent to the evaluation protocol and those arising from model behavior (see Tables 10, 11 and 12).

Evaluation-based limitations The first limitation refers to the fact that evaluation protocols for *Lucifer-A* and *Lucifer-M* are not exactly the same, mostly due to the fact that the claims in the *Lucifer-M* instances are often intertwined and as such, formulated differently from original FEVER 2018 claims. Therefore, unlike *Lucifer-M*, where semantic equivalence is considered for assessing absence/presence of a claim, the current claim-level validation for *Lucifer-A* relies on exact string matching and does not account for semantic equivalence. This can lead to two failure modes: (i) correctly handled claims may be marked as incorrect when they are semantically rephrased (see example in Table 11), and (ii) incorrectly handled claims may be marked as correct. Critically, when the model rephrases a claim, our evaluation treats it as absent, assigning a claim-level score of 0. An ideal evaluation system would recognize semantic preservation across paraphrases and assign a score of 1 accordingly.

Model-based limitations Several distinct failure modes emerge: (i) the model violates the prompt specification by providing extended reasoning instead of concise output (see Table 12, Example 1 for *Osiris* and *DeepSeek-Llama*, with *Osiris* even providing contradictory output), (ii) the model recognizes factual errors but performs incorrect corrections—for instance, attributing the middle name “Victor” to David Beckham when the correct name is Robert Joseph (see Table 12, Example 2, but also Example 4), and (iii) the model produces partially correct outputs that fail at downstream computations –for example, correctly identifying 2013 as the film’s release year but subsequently miscalculating the year difference (see Table 12, Example 3). The last case highlights a critical gap: our claim-level metric marks the response as correct despite containing factual errors.

Phase	Status	Abbr.
Detection	OK	D0
	Not adhering to prompt	D1
	Failing to detect hallucination	D2
	Wrong correction	D3
	Providing contradictory outputs	D4
Evaluation	OK	E0
	Not OK	E1

Table 10: Error types observed in detection (model behavior) and evaluation (evaluation protocol) and their corresponding abbreviations.

G Experimental Setup and Computational Resources

All experiments were conducted on a server equipped with an NVIDIA L40S GPU. We evaluated six Open-source language models (detailed in Section A) using the LangChain framework in their default configuration. Each model was executed sequentially, and no additional fine-tuning or optimization was applied.

Source	Sentence	Score	Correct score	Comment
Input (TF)	Ethiopia is in the peninsula the Horn of Africa while Gong Yoo is not an actor.	-	-	-
DSQ	Ethiopia is in the peninsula **known as** the Horn of Africa, while Gong Yoo is not an actor.	0 (0+0)	**1 (1+0)**	E1, D1 (see boldface)

Table 11: An illustration of a noted evalution problem on the *Lucifer-A* dataset. Notation: **Osi**: Qwen2.5-Osiris-7B-Instruct; **DSQ**: DeepSeek-Qwen; **DSL**: DeepSeek-Llama; **Min**: Ministral; **Mis**: Mistral.

ID	Source	Sentence	Score	Correct score	Comment
1	Input (TT)	Franklin and Eleanor Roosevelt have six children together, while Paul McCartney has one less.	-	-	-
	Osi	Franklin and Eleanor Roosevelt have six children together, while Paul McCartney has two less. Original sequence : [...] Corrected sequence: Franklin and Eleanor Roosevelt have six children together, while Paul McCartney has one fewer. [...] Original sequence : [...] Corrected sequence: Franklin and Eleanor Roosevelt have six children together, while Paul McCartney has one less child.	2 (1+1)	2 (1+1)	E0, D1, D4
	DSQ	Franklin and Eleanor Roosevelt have six children together, while Paul McCartney has one less.	2 (1+1)	2 (1+1)	OK
	DSL	[...] Franklin and Eleanor Roosevelt did not have six children. They had four children: Franklin D. Roosevelt Jr., Eleanor Roosevelt, Franklin D. Roosevelt Jr. (the fourth), and Elliott Roosevelt. Paul McCartney has three children: Mary, James, and Theo. The original sentence incorrectly states the number of children both families have. [...]	0 (0+0)	0 (0+0)	E0, D1
	Min	Franklin and Eleanor Roosevelt have six children together, while Paul McCartney has one less.	2 (1+1)	2 (1+1)	OK
	Mis	Franklin and Eleanor Roosevelt had six children together, while Paul McCartney has five.	2 (1+1)	2 (1+1)	E0
2	Input (FF)	Not only is David Beckham from America, but he also doesn't have any middle names.	-	-	-
	Osi	David Beckham is from England, and he has middle names.	2 (1+1)	2 (1+1)	OK
	DSQ	Not only is David Beckham from America, but he also doesn't have any middle names.	0 (0+0)	0 (0+0)	E0, D2
	DSL	Not only is David Beckham from America, but he also doesn't have any middle names.	0 (0+0)	0 (0+0)	E0, D2
	Min	Not only is David Beckham from England, but he also doesn't have any middle names.	1 (1+0)	1 (1+0)	E0, D2
	Mis	David Beckham is not from America, and he does have a middle name (Victor).	2 (1+1)	**1 (1+0)**	E1, D3 (see underlined)
3	Input (TF)	Paradise was given in 2012, two years before Selena Gomez starred in Spring Breakers.	-	-	-
	Mis	Paradise was given in 2012, three years before Selena Gomez starred in Spring Breakers (2013).	2 (1+1)	**1 (1+0)**	E1, D3 (see underlined)
	all other	Paradise was given in 2012, two years before Selena Gomez starred in Spring Breakers.	1 (1+0)	1 (1+0)	E0, D2
4	Input (FT)	Neymar's full name is Neymar Lebron James while Jobs' is Steve Paul Jobs.	-	-	-
	Osi	Neymar's full name is Neymar da Silva Santos Júnior while Jobs' is Steve Jobs.	2 (1+1)	2 (1+1)	E0, D1
	DSQ	Neymar's full name is Jairzinho de Freitas Neves, while Jobs' **full name** is Steve Paul Jobs.	2 (1+1)	**1 (0+1)**	E1, D1, D3 (see underlined)
	DSL	Neymar's full name is Neymar Lebron James while Jobs' is Steve Paul Jobs.	1 (0+1)	1 (0+1)	E0, D2
	Min	Neymar's full name is Neymar da Silva Santos Júnior while Jobs' is Steve Paul Jobs.	2 (1+1)	2 (1+1)	OK
	Mis	Neymar's full name is Neymar da Silva Santos Júnior, while Jobs' **full name** is Steve Paul Jobs.	2 (1+1)	2 (1+1)	E0, D1 (see boldface)

Table 12: An illustration of noted errors and challenges on the *Lucifer-M* dataset. Notation: **Osi**: Qwen2.5-Osiris-7B-Instruct; **DSQ**: DeepSeek-Qwen; **DSL**: DeepSeek-Llama; **Min**: Ministral; **Mis**: Mistral.

H Prompts

Prompt 1.1 :

Combine the following two claims into a single well-formed sentence, using different kinds of connectors:
1. {claim1}
2. {claim2}

Prompt 1.2 :

Write a sentence that smoothly integrates the following two facts in a natural and engaging way:
- {claim1}
- {claim2}

The sentence should flow naturally as if it were part of a larger narrative.

Prompt 1.3 :

Imagine you are writing an article that mentions both of these facts in a single sentence:
- {claim1}
- {claim2}

Construct a sentence that logically and naturally connects these two pieces of information.

Prompt 1.4 :

Create a single coherent sentence that naturally includes both of these facts:
- {claim1}
- {claim2}

Prompt 2.1 :

System message :
You are an advanced language model tasked with combining two given claims into a single, well-formed sentence. One of these claims is true, and the other is false. However, you are NOT to indicate which one is which.
Your goal is to seamlessly integrate both claims into a single coherent sentence while ensuring that: - The true claim remains true. - The false claim remains false. - The sentence is grammatically correct and sounds natural. - You do NOT explicitly state or imply which claim is true or false. - Keep the claim exactly as it is.
Be creative in how you structure the sentence, e.g vary the kind of connectors, but keep the truth value of either claim.
User message :
Here are two claims:
- {claim1}
- {claim2}

Please generate a single sentence that naturally includes both claims without altering them.

Prompt 2.2

System message:
You are an advanced AI language model specializing in generating logically coherent and well-formed sentences.
Your task is to take two claims—one true and one false—and seamlessly integrate them into a single grammatically correct and natural-sounding sentence. However, you must ensure that:
- The true claim remains completely accurate. - The false claim is included in a way that does not alter its incorrectness. - You do NOT explicitly state or imply which claim is true or false. - The resulting sentence should sound natural and meaningful in everyday communication.
Be creative, but always ensure that the sentence does not introduce misleading or unintended meanings. Avoid artificial or forced phrasing.
User message :
Combine the following two claims into a single well-structured sentence, ensuring that their original veracity remains unchanged:
- {claim1}
- {claim2}

The sentence should be fluid and coherent while maintaining the truthfulness of each claim as they were given.

Prompt 3: Automatic claims matching and merging

You are given two lists: one contains 10 true claims, and the other contains 10 false claims. Your task is to:
Select one true claim and one false claim from the lists.
Combine them into a single, smooth, grammatically correct sentence.
Do not alter the factual content of the selected claims—the true claim must remain true, and the false claim must remain false.
Your goal is to create the most natural-sounding sentence possible, despite the factual contradiction.
You may slightly adjust wording for grammar and flow, but not to change the truthfulness of the individual claims.
If no combination of a true and a false claim results in a sentence that sounds natural or smooth, respond with "No matched sentences".
Example input: True claims: ['John Cena won the UPW Heavyweight Championship in 2000 a year after starting his career.', 'Saamy is a 2003 film from India.', "That's So Raven debuted on January 17, 2003.", 'Lebanon is a country that experienced a period of violence.', 'Jessica Chastain is vocal about social issues.', 'Tommy Lee Jones was an actor in The Fugitive.', 'Hubert Humphrey was the DFL candidate for mayor of a county seat.', 'The character of Adam Stefan Sapieha features in Pope John Paul II.', "Instagram is a service that allows users to share pictures and it's very popular.", 'Jerry Lewis is a performer.'] False claims: ['Watchmen premiered in 1990.', 'Luxo Jr. is a 1984 film.', 'Ketogenic diet is incapable of containing carbohydrates.', 'Jerome is unrecognized by the Roman Catholic Church.', 'Alien: Covenant is a TV show.', 'The United Kingdom is an industrialized coffee.', 'India is officially a Catholic country.', 'FC Barcelona was formed before 1899.', 'Break on Me is only a short story.', 'Richard Curtis has only ever created American companies.']
Example output: Saamy is a 2003 film from India, which is by its constitution a Catholic country.
Your input: True claims:: {true_claims} False claims: {false_claims}

Prompt 4: Claims merging based on topic matching

You are given two lists of claims. Each claim is a short statement. Your task is to find at most one pair of claims—one from List A and one from List B—that share the most similar topic. If no claims from the two lists are topically similar, respond with "no matches".
Rules:
You may output only one pair of claims at most.
The pair should have clearly similar topics.
If no suitable pair exists, respond only with: no matches.
Format your response as:
<1>first_claim</1> <2>first_claim</2>
If there are no match simply write no_matches.
Example : List 1: ['John Cena won the UPW Heavyweight Championship in 2000 a year after starting his career.', 'Saamy is a 2003 film from India.', "That's So Raven debuted on January 17, 2003.", 'Lebanon is a country that experienced a period of violence.', 'Jessica Chastain is vocal about social issues.', 'Tommy Lee Jones was an actor in The Fugitive.', 'Hubert Humphrey was the DFL candidate for mayor of a county seat.', 'The character of Adam Stefan Sapieha features in Pope John Paul II.', "Instagram is a service that allows users to share pictures and it's very popular.", 'Jerry Lewis is a performer.']
List 2: ['Watchmen premiered in 1990.', 'Luxo Jr. is a 1984 film.', 'Ketogenic diet is incapable of containing carbohydrates.', 'Jerome is unrecognized by the Roman Catholic Church.', 'Alien: Covenant is a TV show.', 'The United Kingdom is an industrialized coffee.', 'India is officially a Catholic country.', 'FC Barcelona was formed before 1899.', 'Break on Me is only a short story.', 'Richard Curtis has only ever created American companies.']
Answer : <1>Saamy is a 2003 film from India.</1> <2>India is officially a Catholic country.</2>
Your input: List 1: {true_claims} List 2: {false_claims}
Answer :

Prompt 5: Connecting claims with "while"

You will be given two claims. Your task is to combine them into a single, complete sentence using the word "while" between them. The first claim must appear first in the sentence. Insert the word "while" between the two claims. Capitalize only the first word of the sentence and any proper nouns. The first word of the second claim should not be capitalized unless it is a proper noun. Apply standard English punctuation and grammar rules.
Example input: Claim 1: The sun was setting. Claim 2: Birds were flying south. Example output: The sun was setting while birds were flying south.
Your input: Claim 1: {claim1} Claim 2: {claim2}
<think>

Prompt 6: Claim 2 capitalization

You will be given a claim. Determine if the first word of the claim is a named entity (such as a person, organization, place, brand, etc.) that should always retain its capitalized first letter, even when it's not the first word in a sentence.
Answer with only "yes" or "no".
Claim: {claim}
Answer:
<think>

Prompt 7: Topic Matching

You are given two lists of claims. Each claim is a short statement. Your task is to find at most one pair of claims—one from List A and one from List B—that share the most similar topic. If no claims from the two lists are topically similar, respond with "no matches".
Rules:
You may output only one pair of claims at most.
The pair should have clearly similar topics.
If no suitable pair exists, respond only with: no matches.
Format your response as:
<1>first_claim</1> <2>first_claim</2>
If there are no match or simply write no_matches.
Example : List 1: ['John Cena won the UPW Heavyweight Championship in 2000 a year after starting his career.', 'Saamy is a 2003 film from India.', "That's So Raven debuted on January 17, 2003.", 'Lebanon is a country that experienced a period of violence.', 'Jessica Chastain is vocal about social issues.', 'Tommy Lee Jones was an actor in The Fugitive.', 'Hubert Humphrey was the DFL candidate for mayor of a county seat.', 'The character of Adam Stefan Sapieha features in Pope John Paul II.', "Instagram is a service that allows users to share pictures and it's very popular.", 'Jerry Lewis is a performer.']
List 2: ['Watchmen premiered in 1990.', 'Luxo Jr. is a 1984 film.', 'Ketogenic diet is incapable of containing carbohydrates.', 'Jerome is unrecognized by the Roman Catholic Church.', 'Alien: Covenant is a TV show.', 'The United Kingdom is an industrialized coffee.', 'India is officially a Catholic country.', 'FC Barcelona was formed before 1899.', 'Break on Me is only a short story.', 'Richard Curtis has only ever created American companies.']
Answer : <1>Saamy is a 2003 film from India.</1> <2>India is officially a Catholic country.</2>
Your input: List 1: {true_claims} List 2: {false_claims}
Answer :

Prompt 8: Finding false information

Your task is to identify and mark factually false or incorrect information in the following sequences. You will correct word by word the information in the answer, and rewrite the complete sentence by correcting the sentence. You will not provide any additional comments. If you see a sentence without errors, simply return the original sentence. In any case the sentence must be as identical as possible to the original; in particular, the sentence structure should be the same. Make sure you only change the minimum. Please note that your reply must include only the corrected sentence and no comments, and not the original sentence. Make sure to end your answer with the corrected sentence. For structured extraction use the following format/tags for the response: «<START»>[final_response_with_hallucinations_marked]«<END»>.
Example 1:
Original sequence : Alberto Fouillioux was a mexican basketball player and later a sports illustrator, best known for his time as a midfielder and forward for Universidad Católica and the Irish national team «<START»>Alberto Fouillioux was a Chilean footballer and later a sports commentator, best known for his time as a midfielder and forward for Universidad Católica and the Chilean national team«<END»>
Example 2:
Original sequence : Thorgan James Hazard (born 29 March 1983) is a Belgian professional footballer who plays as a defending midfielder and winger for French League club Anderlecht and the Russia national team. «<START»>Thorgan Ganael Francis Hazard (born 29 March 1993) is a Belgian professional footballer who plays as an attacking midfielder and winger for Belgian Pro League club Anderlecht and the Belgium national team.«<END»>
Example 3:
Original sequence : Alamogordo is the County seat of Alamogordo County, New Mexico, United States. A city in the Tularosa Basin of the Sahara Desert, it is bordered on the east by the Sacramento Mountains and to the west by Holloman Navy Base. The population was 304 as of the 2020 census. Alamogordo is widely known for its connection with the 1945 Trinity test, which was the first ever explosion of an atomic bomb. «<START»>Alamogordo is the County seat of Otero County, New Mexico, United States. A city in the Tularosa Basin of the Chihuahuan Desert, it is bordered on the east by the Sacramento Mountains and to the west by Holloman Air Force Base. The population was 31,384 as of the 2020 census. Alamogordo is widely known for its connection with the 1945 Trinity test, which was the first ever explosion of an atomic bomb.«<END»>
———
Sequence to correct:
Original sequence : {input_a}

Prompt 9: *v1*

You are tasked with identifying and marking hallucinations in the following large language model (LLM) answers. A hallucination in this context refers to an answer that provides incorrect or fabricated information. Your goal is to review each LLM generated text (provided in <LLM_TEXT>[llm_text]</LLM_TEXT>) and highlight any incorrect or unsupported parts of the response using **<h>** tags. If the text is factually correct, return it without any highlighting.
For each example, provide only the response sentence (R) with the highlighted hallucinations if present. Do not provide explanations or commentary.
For structured extraction use the following format/tags for the response: «<START»>[final_response_with_hallucinations_marked]«<END»>
Important: Ensure that the text remains exactly the same length as the original text, don't change any amount of whitespace or newline characters.
You should only add tags and not delete any characters.
—
Example 1: <LLM_TEXT>No, Windows Neptune was not released. It was an internal project by Microsoft that aimed to merge the user interfaces of Windows XP and Windows Tablet PC Edition into a single operating system. The project was later merged with another project called Windows Longhorn, which eventually became Windows Vista. However, neither Neptune nor Longhorn were released as standalone products; instead, their features were incorporated into Windows Vista, which was released in January 2007.</LLM_TEXT> «<START»>No, Windows Neptune was not released. It was an internal project by Microsoft that aimed to <h>merge the user interfaces of Windows XP and Windows Tablet PC Edition into a single operating system</h>. The project was later merged with another project called <h>Windows Longhorn</h>, which eventually became <h>Windows Vista</h>. However, neither Neptune nor <h>Longhorn</h> were released as standalone products; instead, their features were incorporated into Windows <h>Vista</h>, which was released in <h>January 2007</h>.«<END»>
Example 2: <LLM_TEXT>Dave played the role of Zack in the first season of Scary Movie 5.</LLM_TEXT> «<START»>Dave played the role of <h>Zack</h> in the <h>first season</h> of <h>Scary Movie 5</h>.«<END»>
New Question: <LLM_TEXT>input_a</LLM_TEXT>
{input_a}

Prompt 10: ***v2***

You are tasked with identifying and marking hallucinations in the following large language model (LLM) answers. A hallucination in this context refers to an answer that provides incorrect or fabricated information. Your goal is to review each LLM generated text (provided in <LLM_TEXT>[llm_text]</LLM_TEXT>) and highlight any incorrect or unsupported parts of the response using **<h>** tags. If the LLM answer contains no hallucinations, return it without any highlighting.

In short: - Carefully read the answer text. - Highlight each span of text in the answer text that is an overgeneration or hallucination (factual distortion, excessive and unsupported output, typographic hallucination, nonexistent entities, contradictory statements) - Your annotations should include only the minimum number of characters in the text that should be edited/deleted to provide a correct answer (in the case of Chinese, these will be “character components”). - You are encouraged to annotate conservatively and focus on content words rather than function words. This is not a strict guideline, and you should rely on your best judgments. - Ensure that you double-check your annotations. - Important: Ensure that the text remains exactly the same length as the original text, don’t change any amount of whitespace or newline characters. You should only add tags and not delete any characters.

To ensure accuracy, follow and write down ALWAYS these reasoning steps first and than provide the final response with hallucinations marked: 1. LLM Answer Break Down: Identify distinct factual claims or statements in the response. 2. Claim Verification: - Cross-check with reliable knowledge sources. - Determine if the claim is logically consistent with known facts. - If a claim is unverifiable or fabricated, it is a hallucination. 3. Identify Other Hallucinations and Overgenerations: - Check for typographic errors - Identify contradictions. - Look for unsupported or excessive information. 4. Final Response: - Output only the final response for structured extraction in the format: «<START>»[final_response_with_hallucinations_marked]«<END>» - Mark Hallucinations: Surround incorrect or unsupported parts with **<h>** tags. - Do not provide explanations or extra formatting. - If no hallucinations are found, return the LLM answer as is inside the «<START>» and «<END>» tags.

— Example of Question, LLM Answer and Final Response with Hallucinations Marked (but without the reasoning steps):

<LLM_TEXT>The municipality of Delley-Portalban was created on January 1, 2004. It was formed through the merger of two neighboring communes, Delley and Portalban, as part of a wave of municipal consolidations in Switzerland.</LLM_TEXT> Response: 1. LLM Answer Break Down: [Here, you would identify distinct factual claims or statements in the response.] 2. Claim Verification: [Here, you would cross-check each claim with reliable knowledge sources and determine if they are logically consistent with known facts.] 3. Identify Other Hallucinations and Overgenerations: [Here, you would check for typographic errors, contradictions, and unsupported or excessive information.] 4. Final Response: «<START>»The municipality of Delley-Portalban was created on January 1, <h>2004</h>. It was formed through the merger of two neighboring communes, Delley and Portalban, as <h>part of a wave of municipal consolidations in Switzerland</h>.«<END>»

— Remember, first provide the reasoning steps and then the final response with hallucinations marked.

<LLM_TEXT>input_a</LLM_TEXT>

Response: 1. LLM Answer Break Down:

{input_a}

Prompt 11: ***v3***

You are tasked with identifying and marking hallucinations in the following large language model (LLM) answers. A hallucination in this context refers to an answer that provides incorrect or fabricated information. Your goal is to review each LLM answer (provided in <LLM_Answer>[llm_answer]</LLM_Answer>) highlight any incorrect or unsupported parts of the response using **<h>** tags. If the answer is factually correct, return it without any highlighting.

For each example, provide only the response sentence (R) with the highlighted hallucinations if present. Do not provide explanations or commentary.

For structured extraction use the following format/tags for the response: «<START>»[final_response_with_hallucinations_marked]«<END>»

Important: Ensure that the text remains exactly the same length as the original text, don’t change any amount of whitespace or newline characters. You should only add tags and not delete any characters. To this end a token list is provided for the LLM answer (provided in <LLM_Answer_in_tokens>[LLM_Answer_in_token_list]</LLM_Answer_in_tokens>).

Note: You should be extremely critical in identifying hallucinations in the LLM answers. This means any character span that has the slightest chance of being incorrect should be marked as a hallucination.

—

Example 1: <LLM_TEXT>No, Windows Neptune was not released. It was an internal project by Microsoft that aimed to merge the user interfaces of Windows XP and Windows Tablet PC Edition into a single operating system. The project was later merged with another project called Windows Longhorn, which eventually became Windows Vista. However, neither Neptune nor Longhorn were released as standalone products; instead, their features were incorporated into Windows Vista, which was released in January 2007.</LLM_TEXT> «<START>»No, Windows Neptune was not released. It was an internal project by Microsoft that aimed to <h>merge the user interfaces of Windows XP and Windows Tablet PC Edition into a single operating system</h>. The project was later merged with another project called <h>Windows Longhorn</h>, which eventually became <h>Windows Vista</h>. However, neither Neptune nor <h>Longhorn</h> were released as standalone products; instead, their features were incorporated into Windows <h>Vista</h>, which was released in <h>January 2007</h>.«<END>»

Example 2: <LLM_TEXT>Dave played the role of Zack in the first season of Scary Movie 5.</LLM_TEXT> «<START>»Dave played the role of <h>Zack</h> in the <h>first season</h> of <h>Scary Movie 5</h>.«<END>»

New Example: <LLM_TEXT>input_a</LLM_TEXT>

{input_a}