COGUMELO at SemEval-2025 Task 3: A Synthetic Approach to Detecting Hallucinations in Language Models based on Named Entity Recognition

Aldan Creo*, **Héctor Cerezo-Costas**†, **Maximiliano Hormazábal Lagos**†, **Pedro Alonso Doval**†

*Independent Author, Dublin, IE
†Fundación Centro Tecnolóxico de Telecomunicacións de Galicia (GRADIANT), Vigo, ES
Correspondence: research@acmc.fyi

Abstract

In this paper, we propose an approach to detecting hallucinations based on a Named Entity Recognition (NER) task. We train a model to identify spans of text likely to contain hallucinations, treating them as a form of named entity. We focus on efficiency, aiming to develop a model that can detect hallucinations without relying on external data sources or expensive computations that involve state-of-the-art large language models with upwards of tens of billions of parameters. We utilize the SQuAD question answering dataset to generate a synthetic version that contains both correct and hallucinated responses and train encoder language models of a moderate size (`RoBERTa` and `FLAN-T5`) to predict spans of text that are highly likely to contain a hallucination. We test our models on a separate dataset of expert-annotated question-answer pairs and find that our approach achieves a Jaccard similarity of up to 0.358 and 0.227 Spearman correlation, which suggests that our models can serve as moderately accurate hallucination detectors, ideally as part of a detection pipeline involving human supervision. We also observe that larger models seem to develop an emergent ability to leverage their background knowledge to make more informed decisions, while smaller models seem to take shortcuts that can lead to a higher number of false positives. We make our data and code publicly accessible, along with an online visualizer. We also release our trained models under an open license.

1 Introduction

Hallucinations in language models (LMs) are a well-known issue that has been studied in the context of text generation tasks (Ye et al., 2023; Huang et al., 2024; Zhang et al., 2023; Rawte et al., 2023), with some authors affirming they are inevitable (Xu et al., 2024; Banerjee et al., 2024). However, despite the open discussion on their avoidability, a community of authors have worked on methods to detect, prevent, or mitigate them (Tonmoy et al., 2024; Mündler et al., 2023; Harrington et al., 2024; Dhuliawala et al., 2023; Manakul et al., 2023, inter alia). Our work contributes to this effort by addressing hallucination detection in instruction-tuned LMs, a shared task proposed in the SemEval 2025 Task 3, Mu-SHROOM (Vázquez et al., 2025). We approach the chanllenge by framing hallucination detection as a Named Entity Recognition (NER) task, leveraging NER's ability to identify specific spans of text.

NER extracts structured information, such as names, dates, or locations, from unstructured text (Nadeau and Sekine, 2007). Traditionally, it has been applied to sequence labeling tasks using rule-based systems or machine learning models trained on annotated datasets (Yang et al., 2024). However, its versatility has led to applications beyond information extraction, including social media analysis, knowledge graph construction, reinforcement learning for entity augmentation, and more (Sufi et al., 2022; Bunescu and Paşca, 2006; Wan et al., 2020; Keraghel et al., 2024). In our approach, we adapt NER to detect hallucinated spans by treating them as a specialized type of named entity, allowing us to efficiently identify incorrect or fabricated text segments without relying on external data sources or computationally expensive large-scale LMs.

The rest of this article is organized as follows: in Section 2, we provide background information on the task setup; in Section 3, we describe our system's approach; in Section 4, we detail our experimental setup and present qualitative evaluations; in Section 5, we report on the quantitative results of our models and discuss their performance; and in Section 6, we conclude our work and suggest future directions. We also address ethical considerations in Section 7 and discuss the limitations of our approach in Section 8. We make our code, dataset and models publicly available at `https://github.com/ACMCMC/hallucinations-ner`.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 2170–2176

2 Background

The shared task that our work is based on, MuSHROOM, involves detecting spans of text that correspond to hallucinations in the outputs of LMs, with the goal of predicting where hallucinations occur in a given text.

For example, given the following example from the validation dataset of the task:

> **Question.** What is the population of the Spanish region of Galicia?
>
> **Tagged answer.** As of 2021, the estimated population in the region is around 1.5 million people.[a]
>
> [a]The opacity of the underlines represents the probability of the character being a hallucination.

The task consists of predicting the spans of text that are more associated with hallucinations, which in this case would be "2021" and "1.5 million".

The authors of the task provide a dataset of expert-annotated question-answer pairs in 14 languages, but we choose to focus on English due to the complexity of generating a synthetic dataset of faithful and hallucinated responses, which we use to train our models (see Section 3).

Our approach is thus based on training a model to predict spans of text that are likely to contain hallucinations, which we model as a Named Entity Recognition (NER) task, under the assumption that hallucinations can be seen as a form of named entity that can be detected by a model trained to recognize them. This assumption may not cover all types of hallucinations, as we discuss in Section 8, but serves as a starting point that can be later expanded upon.

We implement our NER strategy using an IOB (Ramshaw and Marcus, 1995) tagging scheme, which is a common approach in NER tasks that assigns each token in a sequence a label indicating whether it is inside (I), outside (O), or at the beginning (B) of a named entity. In our case, however, instead of named entities, we turn the task into predicting if we are outside or inside a hallucination. Also, while IOB assigns an I label to single-token entities, we slightly alter this approach by assigning a B label to the first token of all entities, including single-token entities, and an I label to all subsequent tokens, often referred to as IOB2 (Sang and Veenstra, 1999).

It is important to note that NER is not limited to IOB tagging; it can be performed using other approaches that may leverage graphs (Muis and Lu, 2017; Wang et al., 2021), neural networks (Sohrab and Miwa, 2018; Wang and Lu, 2019), constituency discriminators (Finkel and Manning, 2009), or translation to an augmented natural language form that can be easily extracted (Paolini et al., 2021), among others. Likewise, the IOB tagging scheme is not exclusive to NER tasks, as it can be applied to any sequence labeling task.

3 System overview

Our goal is to train an encoder model to predict spans of text that are likely to contain hallucinations, so we choose to model the task as a NER problem. We do not employ any of the common NER-specific labels, such as dates or verbs, but rather focus on the general applicability of the IOB tagging scheme to our task.

We first need to have a collection of correct and hallucinated responses to train our model, for which we use the SQuAD dataset (Rajpurkar et al., 2016). SQuAD is primarily intended for question answering based on a given context, but we repurpose it by discarding the context and using exclusively the question and suggested answers to build a synthetic dataset of correct and hallucinated responses.

The first model is tasked with the transformation of the SQuAD answers, which are an extracted span of text from the context, into a full sentence that a human could understand. Then, we use the second model to generate variations of that correct response in a way that it becomes a hallucination.

We assemble a synthetic dataset of question-answer pairs, where the answers are either correct or hallucinated, and we use this dataset to train our models. Figure 1 shows our approach.

4 Experimental Setup

For our training process, we abstain from using the training set provided by the shared task authors, exclusively training on our synthetic dataset, which we split on a 80/10/10 ratio for training, validation, and testing, respectively. We use the validation set to decide when to stop training. We only use the test set to evaluate our models internally; the results shown in Section 5 are based on the test set provided by the shared task authors.

We choose a smaller model, `SmolLM2-360M-Instruct` (Allal et al., 2025)

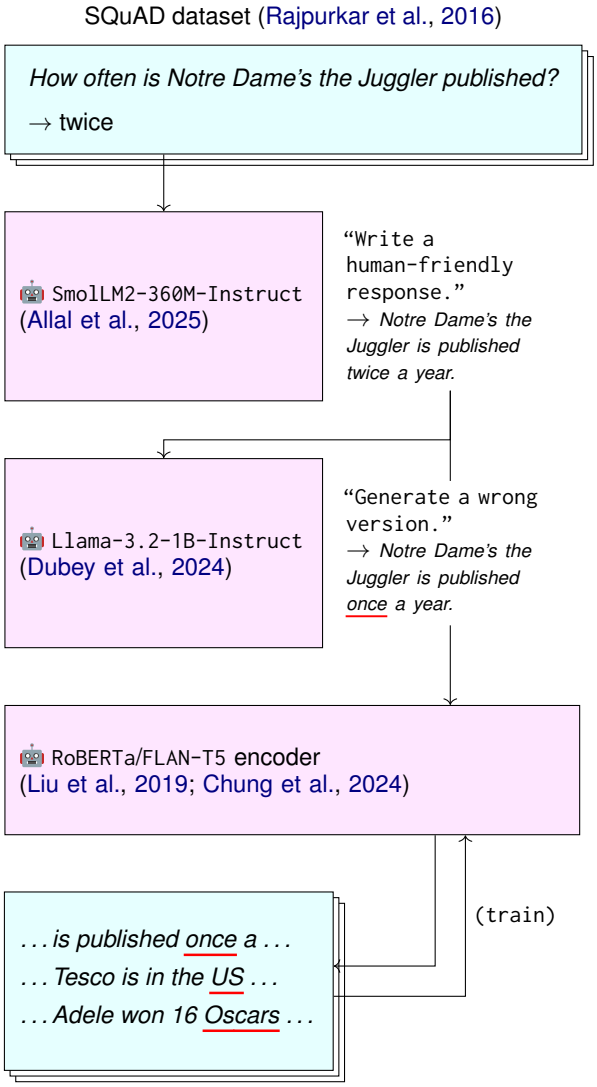


Figure 1: **Our approach.** We employ two instruction-tuned language models to generate synthetic question-answer pairs from the SQuAD dataset, including correct responses and hallucinated variants. These are used to train an encoder model, utilizing a Named Entity Recognition framework, to identify and tag spans of text likely to contain hallucinations.

to generate the correct answers since we anticipate that this is a simpler task — however, when it comes to generating hallucinated responses, we want a larger model to generate more diverse and creative hallucinations, so we select `Llama-3.2-1B-Instruct` (Dubey et al., 2024).

Since we wish for the hallucinated responses to be significantly different both from the correct response and among themselves, so we select a generation configuration to encourage this. Specifically, we set the number of beams and beam groups to 3, the diversity penalty to 0.5, the repetition penalty to 1.2, and the temperature to 1.3; all to force diversity in the generated hallucinations.

We take the generated result of the three beams from the hallucination model and add those as hallucinated responses to our dataset. To ensure we have a balance of correct and hallucinated responses, we include the same number of correct and hallucinated responses in our synthetic generation process by upsampling the correct responses to match the number of hallucinated responses.

For our choice of encoders, we selected models of the `BERT` and `T5` families, `RoBERTa-Base` and `RoBERTa-Large` (Liu et al., 2019) and `FLAN-T5-XL` (Chung et al., 2024), respectively. We opted for the `FLAN` variant of `T5` as we hypothesize that its more extensive training on a diverse set of tasks may help it generalize better to our task.

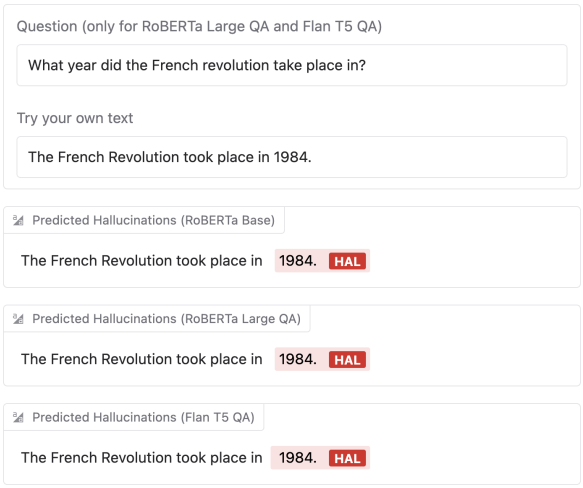
It is important to highlight that we only run the hallucination generation step once, which we acknowledge may lead to a decrease in the representativeness of our synthetic dataset. The test dataset, in fact, can contain more than one hallucination per question-answer pair. This could be addressed by running the hallucination generation step more than once, which we leave for future work.

We only generate hard labels for evaluation (0.0 or 1.0) and do not use the probabilities assigned by the models to each token, which could be a potential improvement to our approach.

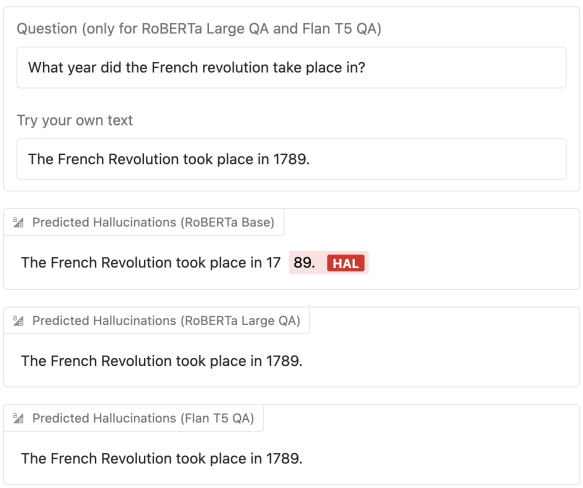
4.1 Qualitative evaluation

We conducted both quantitative (Section 5) and qualitative evaluations, which we describe next.

We developed a visualizer that allows to explore the test split of our synthetic dataset and the predictions we make for each data point, as well as writing any text to explore the predictions of our models. We utilized this tool to qualitatively interpret whether and how our models learned to identify hallucinations. We also make it publicly



(a) **Hallucination.** All models correctly tag "1984".



(b) **Correct answer.** The smaller model, `RoBERTa-Base`, incorrectly tags "89" (part of the year entity) as a hallucination. The two larger models are correct in not finding any.

Figure 2: **Emergent abilities.** Qualitative analysis seems to indicate that larger models go beyond tagging spans of text likely to contain hallucinations, a shortcut that the smaller model seems to take. These models may be learning to extract their background pretrained knowledge to make more informed decisions.

available at https://huggingface.co/spaces/shroom-semeval25/cogumelo-visualizer.

We observe that our models do not exclusively tag spans of text that present a higher possibility of having been hallucinated, such as figures and names of named entities. For instance, when given the question "What year did the French Revolution take place in?", the answer "The French Revolution took place in 1984" gets the correct hallucination tag "1984", while a correct answer (1789) is not tagged in the case of larger models (Figure 2).

This points to the intuition that smaller and larger models are learning in different ways. It appears that the smaller model learns to identify what spans of text usually contain hallucinations (*e.g.*, 1984 or 1789), which is a shortcut that serves to identify some hallucinations — but can also lead to a higher number of false positives. On the other hand, the larger models seem to avoid running into this shortcut and instead seem to be learning to leverage the knowledge that they acquired during their pretraining to identify when a specific figure or claim contradicts such background knowledge. Nevertheless, this observation cannot be generalized, and further investigation is needed to understand the underlying mechanisms that allow our models to make these decisions.

Architecture	**IoU**	**Sp. Corr.**
Neural baseline	0.031	0.119
`RoBERTa-Base`	0.191	0.129
`RoBERTa-Large` (QA)	0.219	0.153
`FLAN-T5-XL` (QA)	0.358	0.227

Table 1: **Scores** obtained on the Mu-SHROOM English test set for the three architectures considered. QA indicates that the model was trained with the question prepended to the answer. The neural baseline is based on `XLM-R` (Conneau et al., 2020).

5 Results

Table 1 shows the scores for our three models, along with a baseline based on `XLM-R` (Conneau et al., 2020). We report on the two metrics official to the Mu-SHROOM shared task: the Intersection over Union (IoU) of characters marked as hallucinations in the gold reference vs. predicted as such, and the Spearman correlation (Sp. Corr.) of the probability assigned by the system that a character is part of a hallucination with the empirical probabilities observed in the annotations.

The results show that larger model sizes seem to correlate with better performance. Model architecture also seems to play a role, with the `FLAN-T5-XL` model outperforming the `RoBERTa` models in both metrics.

In general, while above a baseline model, our results tend to be lower than some other participants in the shared task, which we attribute to the fact that the models we utilize may not have enough capacity to learn the task effectively; the synthetic data we generate, which is not fully aligned with the hallucinations found in the evaluation dataset; and the fact that other techniques such as RAG (Lewis et al., 2020) can retrieve ground truths that greatly improve the performance of the models, since the types of questions asked in the evaluation dataset are at times very specific and we do not expect that such niche knowledge is present in the background knowledge of our models.

It should also be noted that the Spearman correlation is generally lower than the IoU, which is very dependent on the threshold used to determine if a character is part of a hallucination or not. As seen in Figure 2b, the models may sometimes tag just subparts of a hallucination, which we expect will particularly lower the Spearman correlation. Additionally, since we make our models generate hard labels exclusively, we expect that the Spearman correlation would also be lower than if we had used soft labels.

6 Conclusion

In this article, we present an approach to detecting hallucinations in the output of language models based on a named entity recognition task. We train moderate-size encoding models on a synthetic dataset generated from the SQuAD question-answering dataset, which we use to predict text segments that are likely to contain hallucinations.

Our models achieve a Jaccard similarity of up to 0.358 and a Spearman correlation of up to 0.227, suggesting that our models can serve as moderately accurate hallucination detectors, although our scores are lower than some other participants in the shared task. We also observe an interesting pattern in the behavior of our models, where larger models seem to develop an emergent ability to use their background knowledge to make more informed decisions, while smaller models seem to take shortcuts that can lead to a higher number of false positives. We publicly release a synthetic dataset, open-source code and models, along with an interactive visualizer to facilitate further research. Future work could explore enhancing this NER-based approach by incorporating diverse hallucination types and multilingual data to improve detection accuracy.

7 Ethical considerations

Our work aims to contribute to the development of better systems to detect hallucinations, which has important implications for the development of more reliable and trustworthy language models. However, we acknowledge that our models are not perfect and that they may make mistakes. We hope that our approach can be used as part of a pipeline involving human supervision to ensure that the decisions made by the models are correct and that the models do not make decisions which could have negative consequences.

8 Limitations

English-centricness. Our models were trained on English data only, which may limit their performance on other languages. Further work is needed to investigate how our models generalize to other languages, and to develop localized versions of our synthetic datasets to train models that can detect hallucinations in other languages.

Alternative architectures. We only considered models from the `RoBERTa` and `T5` families, but it is up for debate whether other encoder architectures may be more suitable for the task. Further experimentation should be conducted to determine the best architecture for the task, which may involve architectures of different families and sizes, not necessarily transformer-based.

Hallucination types. In our approach, we primarily generate *factual* hallucinations, but the evaluation datasets may contain other types of hallucinations, such as logical, context or instruction inconsistencies (Huang et al., 2023). For instance, in the evaluation dataset, we may find question-answer pairs like "What is the capital of France?" with the answer "France is a country in Europe.", which is an instruction inconsistency. We expect this to limit the generalization capabilities of our models, since they have been trained on a narrower set of hallucinations than those found in the evaluation datasets.

Knowledge cutoff date. Our synthetic dataset is derived from the SQuAD dataset, which dates back to 2016 (Rajpurkar et al., 2016) and may not be representative of the current state of knowledge. This may deteriorate the performance of our models.

References

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. 2025. Smollm2: When smol goes big – data-centric training of a small language model. *Preprint*, arXiv:2502.02737.

Sourav Banerjee, Ayushi Agarwal, and Saloni Singla. 2024. Llms will always hallucinate, and we need to live with this. *arXiv preprint arXiv:2409.05746*.

Razvan Bunescu and Marius Paşca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 9–16, Trento, Italy. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jenny Rose Finkel and Christopher D. Manning. 2009. Nested named entity recognition. In *Conference on Empirical Methods in Natural Language Processing*.

Fiona Harrington, Elliot Rosenthal, and Miles Swinburne. 2024. Mitigating hallucinations in large language models with sliding generation and self-checks. *Authorea Preprints*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ArXiv*, abs/2311.05232.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2024. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

Imed Keraghel, Stanislas Morbieu, and Mohamed Nadif. 2024. Recent advances in named entity recognition: A comprehensive survey and comparative study. *Preprint*, arXiv:2401.10825.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

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

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

Aldrian Obaja Muis and Wei Lu. 2017. Labeling gaps between words: Recognizing overlapping mentions with mention separators. *ArXiv*, abs/1810.09073.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *arXiv preprint arXiv:2305.15852*.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cícero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. *ArXiv*, abs/2101.05779.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, SM Tonmoy, Aman Chadha, Amit P Sheth, and Amitava Das. 2023. The troubling emergence of hallucination in large language models–an extensive definition, quantification, and prescriptive remediations. *arXiv preprint arXiv:2310.04988*.

E. Tjong Kim Sang and Jorn Veenstra. 1999. Representing text chunks. *ArXiv*, cs.CL/9907006.

Mohammad Golam Sohrab and Makoto Miwa. 2018. Deep exhaustive model for nested named entity recognition. In *Conference on Empirical Methods in Natural Language Processing*.

Fahim K. Sufi, Imran Razzak, and Ibrahim Khalil. 2022. Tracking anti-vax social movement using ai-based social media monitoring. *IEEE Transactions on Technology and Society*, 3(4):290–299.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*.

Raúl Vázquez, Timothee Mickus, Elaine Zosa, Teemu Vahtola, Jörg Tiedemann, Aman Sinha, Vincent Segonne, Fernando Sánchez-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.

Jing Wan, Haoming Li, Lei Hou, and Juaizi Li. 2020. Reinforcement learning for named entity recognition from noisy data. In *Natural Language Processing and Chinese Computing*, pages 333–345, Cham. Springer International Publishing.

Bailin Wang and Wei Lu. 2019. Combining spans into entities: A neural two-stage approach for recognizing discontiguous entities. In *Conference on Empirical Methods in Natural Language Processing*.

Yucheng Wang, Yu Bowen, Hongsong Zhu, Tingwen Liu, Nan Yu, and Limin Sun. 2021. Discontinuous named entity recognition as maximal clique discovery. In *Annual Meeting of the Association for Computational Linguistics*.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

Jun Yang, Taihua Zhang, Chieh-Yuan Tsai, Yao Lu, and Liguo Yao. 2024. Evolution and emerging trends of named entity recognition: Bibliometric analysis from 2000 to 2023. *Heliyon*.

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2023. Cognitive mirage: A review of hallucinations in large language models. *arXiv preprint arXiv:2309.06794*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.