Extracting Semantic Entity Triplets by Leveraging LLMs

Alexander Sternfeld, Andrei Kucharavy, Dimitri Percia David, Julian Jang-Jaccard and Alain Mermoud

September 16, 2024

Extracting Semantic Entity Triplets by Leveraging LLMs

Abstract

As Large Language Models (LLMs) become increasingly powerful and accessible, there is a rise in concerns regarding the automatic generation of academic papers. Several instances of undeniable usage of LLMs in reputable journals have been reported. Probably significantly more articles were partially or entirely written by LLMs but have not yet been detected, posing a threat to the veracity of academic journals. The current consensus among researchers is that detecting LLM-generated text is ineffective or easy to evade in a general setting. Therefore, we explore an alternative approach, targeting the stochastic nature of LLMs by extracting semantic entity triplets. Such triplets can be used to assess a text's factual accuracy and filter the publication corpus accordingly. However, such extraction is all but trivial, and prior approaches have reported poor suitability of both LLMs and embedding-based methods. Here, we show that these issues can be alleviated by few-shot prompting on recent LLMs, notably the `Meta-Llama-3-8B-Instruct`. We show that extracted triplets are more specific, and hallucinations are undetectable in our setting.

1 Introduction

In 2023, Generative Large Language Models (LLMs) took the world by storm with their capability of generating complex, consistent natural language text from a short prompt. Their wide availability has led to the emergence of LLM-generated text in many disciplines, including the scientific community. When searching Google Scholar for the phrases *"As of my last knowledge update"* and *"As an AI language model"*, one can retrieve hundreds of papers with AI-generated content (Maiberg, 2024).

This problem will likely only become more severe due to the increasing power and accessibility of LLMs. Unfortunately, this problem cannot be trivially mitigated by detecting LLM-generated texts. The current research consensus is that LLM detectors do not achieve a satisfying performance and are susceptible to widespread evasion techniques (Henrique et al., 2023; Chen and Shu, 2023). Therefore, we focus instead on using semantic entity triplets to assess factual consistency between and within papers. As LLMs are stochastic text generators, hallucinations in long texts are a persistent problem, and the generated output regularly contains counterfactual components (Li et al., 2024).

Sternfeld et al. (2024) took a first step in this direction by extracting triplets through spaCy, building on the work of Würsch et al. (2023). The limitation of this method is that the extracted triplets are too general, to the point of being domain-agnostic. We improve upon this method by leveraging recently released LLMs for triplet extraction and successfully extracting more specific triplets. The code, annotated data, few-shot examples, and parameter settings are available in a public repository [1].

2 Data and methodology

To fine-tune the LLMs for triplet entity extraction, we compose a training dataset based on the paper *A Survey of Large Language Models* (Zhao et al., 2023). In total, we manually annotated 547 triplets in 100 paragraphs. To evaluate the performance of the fine-tuned LLMs, we use a holdout set of 20% of the annotated paragraphs. Furthermore, we consider the performance of the best-performing LLM on a larger dataset: the arXiv papers from the categories `cs.AI`, `cs.CL` and `cs.LG` in December 2023. We choose these categories as we have in-house expertise in this domain.

We fine-tune pre-trained LLMs with a small domain-specific dataset using parameter-efficient fine-tuning (PEFT). Specifically, we use Low-Rank Adaptation of Large Language Models (LoRA) (Hu et al., 2021). We consider three state-of-the-art LLMs: `Mistral-7B-Instruct-v0.2`, `Meta-Llama-3-8B-Instruct` and `Starling-LM-7B-beta` (Jiang et al., 2023; Zhu et al., 2023). We do not consider larger LLMs, as we require a scalable triplet extraction method with limited resources.

3 Results

Table 1 shows that the spaCy-based extraction method retrieves, on average, 3.9 triplets per paragraph of the reference text. In contrast, on average, the LLMs extract up to 10 triplets per paragraph. When

[1] https://github.com/fully-anonymized-submission/triplet-extraction-llms/tree/main

Table 1: Results from the test set of the manually annotated paragraphs.

	Avg. number of triplets	Percentage with correct format	Percentage inconsistent with Levenshtein 2	Percentage inconsistent with Levenshtein 3	Avg. time / line (s)
Annotated triplets	8.1	100%	7.2%	2.6%	-
spaCy extraction	3.9	100%	0%	0%	0.013
Starling - base	-	0%	-	-	0.206
Mistral - base	-	0%	-	-	0.221
Llama - base	4.0	10%	0%	0%	0.097
Starling - base + few-shot	7.6	100%	13.2%	4.6%	0.217
Mistral - base + few-shot	3.1	60%	28%	16%	0.243
Llama - base + few-shot	8.9	100%	7.9%	0.6%	0.104
Starling - fine-tuned + few-shot	6.3	100%	40.8%	24.8%	0.220
Mistral - fine-tuned + few-shot	9.9	75%	50.4%	41.1%	0.244
Llama - fine-tuned + few-shot	7.8	30%	29.8%	21.3%	0.102

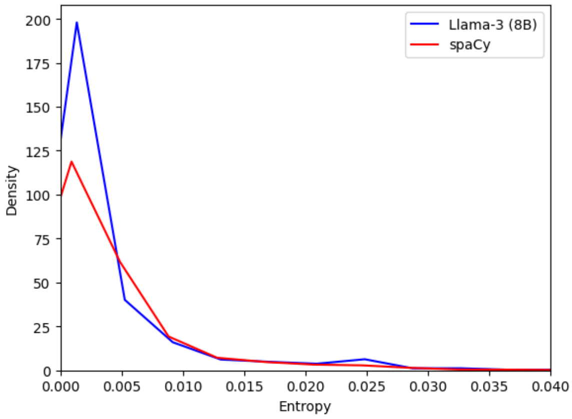

Figure 1: Entropy of the subjects and objects for CS papers December 2023.

considering the format of the extracted triplets, the results show that few-shot prompting is essential to obtain correctly formatted results. Furthermore, fine-tuning the LLM degrades the percentage of correctly formatted generations for Mistral and Llama. We hypothesize this is due to them being extensively fine-tuned and being on a Pareto frontier, right before degeneration kicks in (Bai et al., 2022).

To assess whether the LLMs hallucinate, we investigate whether the subjects and objects from the triplets are present in the original text. We consider the Levenshtein distance, which is defined as the number of single-character edits to change one word into the other. Table 1 shows that even in the human-annotated triplets there are inconsistencies, manual inspection shows that these instances are caused by the stemming of verbs or nouns. Among the LLMs with 100% correct formatting, we find that Llama-3-8B has the least instances of inconsistencies, at the same level as human annotations. Therefore, we choose Llama-3-8B for the extraction of triplets.

We conclude by assessing the specificity of the extracted nouns through the inter-categorical word entropy based on all arXiv papers from October, November, and December 2023. Figure 1 shows that the triplets extracted by Llama-3 have a lower entropy, indicating that the subject and object are more category-specific and thus carry more information.

4 Conclusion

This study considers using fine-tuned LLMs to extract semantic entity triplets. We are able to extract a relatively large number of high-quality triplets by leveraging `Meta-Llama-3-8B-Instruct`. We find that the entropy of the words in the triplets is lower compared to the triplets extracted through spaCy. Moreover, we extract more than twice as many triplets using Llama.

Given that informative semantic triplets extraction has until now been the limiting step in logical consistency analysis for scientific texts, our findings open the direct path toward systematic assessment of factual consistency within and between scientific papers.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback.

Canyu Chen and Kai Shu. 2023. Can llm-generated misinformation be detected? *CoRR*, abs/2309.13788.

Da Silva Gameiro Henrique, Andrei Kucharavy, and Rachid Guerraoui. 2023. Stochastic parrots looking for stochastic parrots: Llms are easy to fine-tune and hard to detect with other llms. *CoRR*, abs/2304.08968.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models.

Emanuel Maiberg. 2024. Scientific journals are publishing papers with ai-generated text.

Alexander Sternfeld, Andrei Kucharavy, Dimitri Percia David, Alain Mermoud, and Julian Jang-Jaccard. 2024. Llm-resilient bibliometrics: Factual consistency through entity triplet extraction.

Maxime Würsch, Andrei Kucharavy, Dimitri Percia David, and Alain Mermoud. 2023. Llm-based entity extraction is not for cybersecurity. In *Proceedings of Joint Workshop of the 4th Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2023) and the 3rd AI + Informetrics (AII2023) co-located with the JCDL 2023, Santa Fe, New Mexico, USA and Online, 26 June, 2023*, volume 3451 of *CEUR Workshop Proceedings*, pages 26–32. CEUR-WS.org.

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

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, Karthik Ganesan, Wei-Lin Chiang, Jian Zhang, and Jiantao Jiao. 2023. Starling-7b: Improving llm helpfulness & harmlessness with rlaif.