From Semantic Alignment to LLM Hallucination Origins: An Algorithmic Approach

Anonymous ACL submission

Abstract

We introduce a new text alignment algorithm that can produce fine-grained alignment between a query document and database documents. Our work explores two under-explored directions: i) alignment granularity at text segment level as opposed to traditional entire document retrieval, and ii) alignment directed by semantic similarity instead of exact matches. We utilize text embeddings produced by Large Language Models (LLM) and perform efficient queries through nearest neighbor data structures. We also introduce an attack strategy exploiting temporal inconsistencies to induce hallucinations in Large Language Models (LLMs) and apply our alignment algorithm to trace these hallucinations back to their possible origins in training data. By creating a database of relevant web documents using keyword filtering on Common Crawl data, our approach demonstrates the effectiveness of identifying candidate origins of LLM hallucinations.

1 Introduction

While document retrieval from a database given a query text is a fundamental Natural Language Processing (NLP) task (Hambarde and Proenca, 2023; Zhu et al., 2021), aligning fine-grained text segments between a query document and database documents at a semantic level remains under-explored. The edit distance algorithm (Levenshtein et al., 1966) is the most widely used text alignment algorithm; it aligns a pair of texts by introducing the minimum number of deletions, insertions, substitutions, and/or transpositions. Pial and Skiena (2023) explored semantic text alignment for pair-wise documents using text embeddings and Dynamic Programming (DP), but the quadratic complexity of pair-wise alignment makes it infeasible to apply these algorithms at database level alignments.

Sequence alignment is a well-studied area in Bioinformatics where fine-grained alignment of DNA residues of different species are usually interpreted to share an evolutionary origin (Katoh et al., 2009). Akin to aligning DNA molecules in Bioinformatics, text alignment involves aligning text units such as n-grams, sentences and paragraphs. Sanchez-Perez et al. (2015) introduced a textual plagiarism detection algorithm paralleling the protein and DNA alignment algorithm BLAST (Altschul et al., 1990), using a seed-and-extend paradigm. First a hashmap of n-grams are created using a database of documents. For a query document q, seed positions (x-th n-gram of document d in database) are identified through matching n-grams in the database hashamp. Next the alignment is extended from the seed position in both directions between q and d as long as they match a predefined heuristic.

As this approach relies on exact n-gram matches or bag-of-words matches for discovering the seed alignment and the subsequent extension, it fails where the wording may differ between paraphrased texts. Taking advantage of recent advancements in embedding texts into vector spaces preserving semantic meaning and efficient Approximate Nearest Neighbor (ANN) search structures, we extend the seed-and-extend paradigm to vector spaces for better semantic alignment.

Large Language Models (LLM) memorize parts of their training data (Carlini et al., 2019, 2021). Memorization becomes a problem when they are regurgitated unprompted or when they are part of a response containing incorrect information i.e. hallucination. We devise an attack strategy exploiting temporal inconsistencies to induce LLM hallucination about historical world events. LLMs are often trained on datasets built on CommonCrawl (Penedo et al., 2024). We use keyword filtering to extract web documents discussing these events from Common Crawl to create a database and apply our alignment algorithm to find parts matching the LLM hallucinated responses.

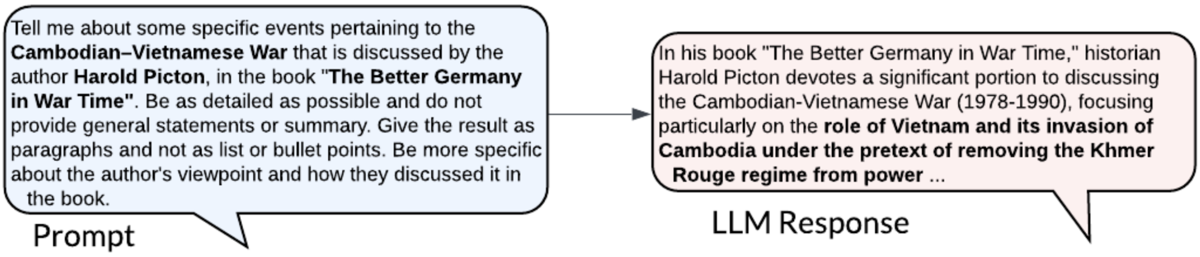

Figure 1: Example attack to generate hallucinations based on temporal inconsistency. *The Cambodian-Vietnamese war* started in 1977 whereas *The Better Germany in War Time* book was published in 1918.

Our contributions include:

- A semantic text alignment algorithm at database level that is orders of magnitude faster than quadratic text alignment algorithms both in practice and expected theoretical complexity.
- An attack strategy to induce LLM hallucinations via temporal inconsistencies accompanied by comparative analysis of different LLMs' robustness against this attack. Examples of LLM hallucinations and aligned source text are presented in Table A.3 and A.4.
- Empirical experiments to trace LLM hallucinations back to training data using our algorithm.

2 Related Works

Dynamic Programming has been used for pairwise text alignment in various NLP tasks, such as global alignment of translated books (Thai et al., 2022) and book-to-film script alignment (Pial et al., 2023). In bioinformatics, Smith et al. (1981) and Needleman and Wunsch (1970) are the most popular pairwise sequence alignment algorithms, both based on DP. Our approach is inspired by the database-level sequence alignment algorithm BLAST (Altschul et al., 1990), which leverages the limited vocabulary size of DNA (4) or protein molecules (≈ 100) using hashmaps for k-mers. In NLP, we deal with an infinite number of sentences or paragraphs, so we use embeddings to capture semantic relationships and ANN structures for efficient querying.

We apply our algorithm to discover LLM hallucination origins. Previous works (Biderman et al., 2024; Carlini et al., 2021) have shown how LLMs can emit memorized pretraining data under different attack strategies. The focus has been on discovering exact token sequence matches (typically length 50 (Nasr et al., 2023)) between LLM responses and pretraining data. Lee et al. (2023) introduced the concept of paraphrased plagiarism through memorization. We hypothesize that during hallucinations, LLMs paraphrase pretraining data, and our alignment algorithm can find potential hallucination origins.

3 Aligning Query Text with Database

Given a query document Q segmented into text units $Q_1, Q_2, \ldots, Q_{|Q|}$ and a database D of multiple documents, our objective is to align each Q_i with a unit from a document in D or leave it unaligned. Different units from the same query can align with units from different documents. We experiment with both sentences and paragraphs as units.

3.1 ANN Index for Embeddings

We segment all database documents into text units, embedding each into a d-dimensional vector space using the SBERT model (Reimers and Gurevych, 2019). This model produces semantically meaningful embeddings, where similar texts have closer embeddings. We use the same model for query document text units. For efficient approximate nearest neighbor search, we build the ANNOY index (Bernhardsson, 2018) using these embeddings.

3.2 Query: Seed and Extend

To process a query document Q, we first embed its text units. Let x_j be the nearest neighbor of query unit q_i found via the index where x is a database document. If the similarity between q_i and x_j exceeds threshold th_s, q_i is aligned with x_j and is used as a seed position. We then extend the alignment in both directions, aligning $q_{i\pm k}$ with $x_{j\pm k}$, stopping when $\text{sim}(q_{i+k}, x_{j+k}) < th_e$ or the extension exceeds a predefined maximum length. Both

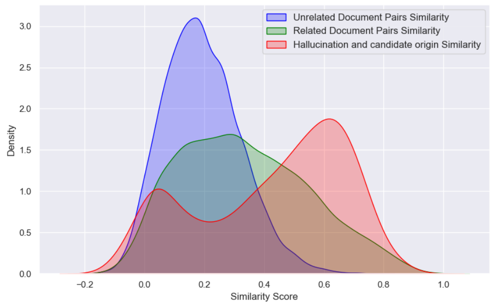

Figure 2: Distribution of cosine similarity between LLM hallucinations and candidate origin discovered by our algorithm is shifted significantly to the right compared to both related and unrelated webpages' similarity distribution.

th_s and th_e are empirically selected thresholds.

4 Experiments and Applications

4.1 Matching Translated Books

Pial and Skiena (2023) created the RelBook dataset, comprising 36 pairs of English translations of foreign language books from Project Gutenberg (Gutenberg, n.d.). We augmented this by including one book from each pair and 1000 random books, creating an ANN index with 1.3×10^6 paragraph embeddings. We queried the index using the unselected book from each pair. As shown in Figure 3, we aligned 54% of the query paragraphs to the correct book with a maximum extension length of 5. Nearest neighbor search alone cannot utilize spatial alignment of neighbors, often necessary when consecutive query paragraphs $i - 1$ and i should align with consecutive paragraphs in a database document. Our algorithm addresses this through the seed and extension mechanism.

Without fine-grained gold labels, we defined accuracy by matching the correct book rather than the exact paragraph. To further evaluate, we computed the Pearson correlation of the absolute percentile positions of matched paragraphs, assuming translations maintain event sequence despite variations. Scores ranged from 0.23 to 1.00, averaging 0.80, indicating strong correlation and validating our alignment algorithm though the algorithm does not utilize absolute position information.

4.2 Efficiency-Performance Trade-off

For a query document with n paragraphs, we query the ANN structure n times with expected complexity of $O(\log k)$ where k is the number of database

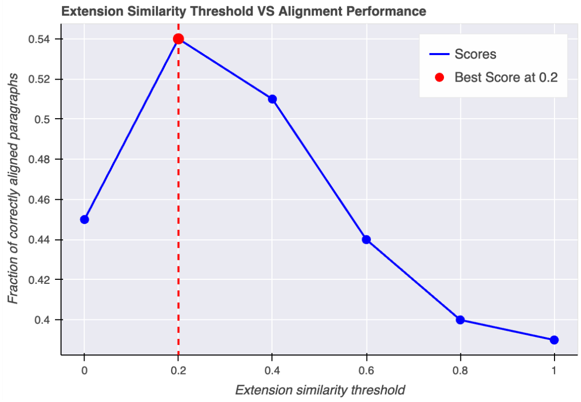

Figure 3: Extension paradigm increases the accuracy of the algorithm. Threshold=1 represents the scenario where no extension is done due to the strict threshold. Without extension, no spatial alignment information is shared between neighboring query document paragraphs. The poor performance on the right shows that alignment results improve when the seed paragraph alignment helps in alignment of neighboring paragraphs through extension. Therefore it is not optimal to always match a paragraph with the nearest neighbor in the ANN structure, rather some paragraphs should be aligned through extension

paragraphs, yielding an expected complexity of $O(n \log k)$. We observed this in practice as we created the ANN structure with 1.6×10^6 paragraphs for the RelBooks experiment of Section 4.1. The query time per paragraph was 0.0002 seconds, with an average of 11.13 seconds per query book.

In comparison, a pair-wise sequence alignment algorithm would require $O(n \times p \times m)$ complexity as it would run p times for p database documents, each with m paragraphs on average. This translates to $O(n \times k)$. To gain the efficiency, our algorithm sacrifices performance since we use heuristics for ANN search. We compare our performance against Pial and Skiena (2023) for the PAN13 plagiarsim detection dataset (Potthast et al., 2013) who use a quadratic DP algorithm. We observe one magnitude of order decrease in performance but more than 3 magnitudes of order increase in efficiency for PAN13 compared to them.

5 LLM Hallucination Origins

5.1 Temporal Attack-Induced Hallucinations

We devise an attack strategy exploiting temporal inconsistencies to induce LLM hallucination. The strategy prompts an LLM to explain what author x discussed about event y in book z, where z was published before y occurred. The correct response

	No Obfuscation			Random Obfuscation			Translation Obfuscation			Summary Obfuscation			Entire Corpus		
	precision	recall	F-1	precision	recall	F-1	precision	recall	F-1	precision	recall	F-1	precision	recall	F-1
Torrejón and Ramos (2013)	.90	.95	**.92**	.91	.63	.74	**.90**	.81	.85	.91	.22	.35	**.89**	.76	.82
Sanchez-Perez et al. (2014)	.83	.98	.90	.91	**.86**	**.88**	.88	**.89**	**.88**	**.99**	.41	.58	.88	**.88**	**.88**
Pial and Skiena (2023)	.91	.79	.84	.89	.43	.58	.86	.66	.75	.93	**.78**	**.85**	.84	.67	.75
BLAST (pairwise)	**.92**	.90	.91	.45	.79	.58	.81	.75	.78	.49	.34	.40	.66	.69	.66

Table 1: We report the performance of the top team in the plagiarism detection contest and other more recent results, along with our proposed algorithm on different subsets of the PAN-13 dataset (Potthast et al., 2013). The subsets are created based on how the plagiarism was inserted in the query documents. Despite being multiple magnitudes of order faster than other algorithms, our method demonstrates a competitive performance.

should highlight the temporal inconsistency, but we observed LLMs often hallucinate for lesser-known events, indicating memorization of y from pretraining data. We use keyword filtering to extract web documents from Common Crawl and apply an alignment algorithm to find parts matching LLM hallucinations. Figure 1 has an example prompt and hallucinated output.

We utilized two recent Common Crawl snapshots with minimal overlap. From approximately 670 million web documents, we filtered over 10,000 relevant documents using quality signals and keyword matches pertinent to the events queried in our LLMs. Quality signals included document length, readability, word length, 2-gram frequencies, and relevance to event keywords.

To limit matching documents and reduce ambiguity, we focused on five lesser-known events, selecting 21 books by different authors from Project Gutenberg, generating 105 unique prompts. These prompts were used to query four distinct LLMs, resulting in a dataset of 420 potentially hallucinated text documents. Details of the events are in Table A.2, and Table 2 provides details of the models.

To match segments from hallucinated texts to filtered web document segments, we constructed an Approximate Nearest Neighbors (ANN) data structure for each event, facilitating rapid identification of closest matches based on cosine similarities.

5.2 Analysis of Hallucinations

We observe that all the models hallucinate for majority of the times for all the events, except for the Gemma model and the event *the Football War between El Salvador and Honduras*. Gemma did not know about this event as evident by the vague or historically inaccurate information it provided even when it hallucinated for this event. Sometimes the LLMs discuss the event but correctly identifies that the author did not discuss it in the book - we do not consider these responses as hallucinations.

Model	Cambodian Vietnamese War	Chaco War	Football War	Timur Invasion	Farakka Long March
Mixtral-8x7B	49.54	91.23	64.52	65.71	96.20
Gemma-7b	70.00	82.61	20.00	84.93	63.64
GPT-3.5	100.00	100.00	100.00	100.00	100.00
Mistral-7B	100.00	100.00	100.00	100.00	100.00

Table 2: Percentage of times when the temporal inconsistency attack succeeded in forcing the LLM to hallucinate.

The Mixtral 8x7B model was far superior than GPT 3.5 for refusing to hallucinate. This is in line with the theory that hugely overtrained larger models are prone to emit memorized data more than smaller models (Nasr et al., 2023).

Fig 2 shows that the cosine similarity between hallucinated data and candidate origins follows two different distributions. These distributions originate mostly from scenarios where the LLMs refuse to answer and when they hallucinate. The right part of the distribution crosses the related document similarity distribution, proving these origins have non-trivial semantic similarity with the hallucinated outputs. 3.03% of the alignments we discovered are more than 2 standard deviations higher than the related pairs of documents discussing the same event. The percentage is 40.23% for one standard deviation.

6 Conclusion

We proposed a semantic text alignment algorithm that can align a query document with text segments of a database of millions of documents efficiently. We then applied our method to find candidate origins of LLM hallucinations induced through a temporal inconsistency attack we devised. One future direction involves creating methods for determining statistical significance of alignments of LLM hallucination with potential origins. This can provide concrete p-values and the degree of certainty of alignments. Another interesting direction is creating models that can predict if a LLM is hallucinating. Our attack strategy can create important training data for this task.

Limitations

We explored two snapshots of CommonCrawl for hallucination candidate generation due to resource constraints. There are more than 150 snapshots, all of which potentially have been used for pretraining LLMs. Even though the snapshots have overlaps, they can be explored to discover more and stronger candidate origins with more resources. LLMs also sometimes paraphrase from multiple sources in the same paragraph, but our method does not take into account how two different sources may have contributed to one LLM output paragraph.

References

Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. 1990. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410.

Erik Bernhardsson. 2018. *Annoy*. GitHub. https://github.com/spotify/annoy.

Stella Biderman, USVSN PRASHANTH, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. 2024. Emergent and predictable memorization in large language models. *Advances in Neural Information Processing Systems*, 36.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pages 267–284.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Project Gutenberg. n.d. www.gutenberg.org. Accessed: April 2022.

Kailash A Hambarde and Hugo Proenca. 2023. Information retrieval: recent advances and beyond. *IEEE Access*.

Kazutaka Katoh, George Asimenos, and Hiroyuki Toh. 2009. Multiple alignment of dna sequences with mafft. *Bioinformatics for DNA sequence analysis*, pages 39–64.

Jooyoung Lee, Thai Le, Jinghui Chen, and Dongwon Lee. 2023. Do language models plagiarize? In *Proceedings of the ACM Web Conference 2023*, pages 3637–3647.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*.

Saul B Needleman and Christian D Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2024. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data only. *Advances in Neural Information Processing Systems*, 36.

Tanzir Pial, Shahreen Aunti, Charuta Pethe, Allen Kim, and Steven Skiena. 2023. Analyzing film adaptation through narrative alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15560–15579, Singapore. Association for Computational Linguistics.

Tanzir Pial and Steven Skiena. 2023. GNAT: A general narrative alignment tool. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14636–14652, Singapore. Association for Computational Linguistics.

Martin Potthast, Matthias Hagen, Tim Gollub, Martin Tippmann, Johannes Kiesel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. 2013. Overview of the 5th international competition on plagiarism detection. In *CLEF Conference on Multilingual and Multimodal Information Access Evaluation*, pages 301–331. CELCT.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Miguel A Sanchez-Perez, Alexander Gelbukh, and Grigori Sidorov. 2015. Adaptive algorithm for plagiarism detection: The best-performing approach at pan 2014 text alignment competition. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 402–413. Springer.

Miguel A Sanchez-Perez, Grigori Sidorov, and Alexander F Gelbukh. 2014. A winning approach to text alignment for text reuse detection at pan 2014. *CLEF (Working Notes)*, 2014:1004–1011.

Temple F Smith, Michael S Waterman, et al. 1981. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197.

Katherine Thai, Yapei Chang, Kalpesh Krishna, and Mohit Iyyer. 2022. RELiC: Retrieving evidence for literary claims. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7500–7518, Dublin, Ireland. Association for Computational Linguistics.

Diego A Rodríguez Torrejón and José Manuel Martín Ramos. 2013. Text alignment module in coremo 2.1 plagiarism detector. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'13)*.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv:2101.00774*.

Appendix

	# of documents	Total # of words	# of embeddings in ANN
Pan13	5057	4946819	39007
RelBooks	1085	10009301	1330384
CommonCrawl Snapshots	$\approx 670 \times 10^6$	-	-
Filtered CommonCrawl	10309	52054950	638237

Table A.1: Statistics of the datasets employed in this study.

Event	Year	Description
Cambodian–Vietnamese War	1978-1989	The Cambodian–Vietnamese War was an armed conflict between Democratic Kampuchea, controlled by Pol Pot's Khmer Rouge, and the Socialist Republic of Vietnam.
Chaco War	1932-1935	The Chaco War was fought from 1932 to 1935 between Bolivia and Paraguay, over the control of the northern part of the Gran Chaco region of South America, which was thought to be rich in oil. The war is also referred to as La Guerra de la Sed in literary circles since it was fought in the semi-arid Chaco.
Football War	1969	The Football War, also known as the Soccer War or the 100 Hour War, was a brief military conflict fought between El Salvador and Honduras in 1969. Existing tensions between the two countries coincided with rioting during a 1970 FIFA World Cup qualifier.
Indonesian invasion of East Timor	1975-1999	The Indonesian invasion of East Timor, known in Indonesia as Operation Lotus, began on 7 December 1975 when the Indonesian military invaded East Timor under the pretext of anti-colonialism and anti-communism to overthrow the Fretilin regime that had emerged in 1974.
Farakka Long March	1976	The Farakka Long March occurred in May 1976, and was led by Maulana Abdul Hamid Khan Bhashani, demanding demolition of the Farakka Barrage constructed by India to divert flow of Ganges waters inside its territory, triggering the drying up of river Padma and desertification of Bangladesh.

Table A.2: Overview of the five historical events we use in our experiments. The descriptions are taken from Wikipedia

Model	Hallucination
GPT-3.5-turbo	The Football War, also known as the 100-Hour War, was a brief conflict between El Salvador and Honduras that took place in July 1969. The catalyst for the war was a series of tense soccer matches between the two countries during the qualifying rounds of the 1970 FIFA World Cup. The matches led to violent incidents both on and off the field, fueling existing tensions between the two nations. Author Stanley Lane-Poole discusses the events leading up to the war in his book, The Story of the Barbary Corsairs, highlighting the role of the soccer matches as a trigger for the conflict. Lane-Poole delves into the long-standing territorial disputes and economic grievances that underpinned the hostility between El Salvador and Honduras, emphasizing how these factors were exacerbated by the soccer matches. According to Lane-Poole, the situation escalated rapidly after the second game between the two countries, with riots breaking out in both nations and leading to a breakdown in diplomatic relations. Honduras accused El Salvador of supporting insurgents within its territory, while El Salvador claimed that it was defending the rights of Salvadoran immigrants in Honduras. The author describes the outbreak of full-scale hostilities in July 1969, as El Salvador launched a military offensive against Honduras. The conflict lasted for just 100 hours but resulted in significant casualties and widespread destruction, with both countries suffering heavy losses. The Organization of American States (OAS) intervened to broker a ceasefire and peace agreement, bringing an end to the war. Overall, Lane-Poole provides a detailed account of the Football War in his book, portraying it as a tragic and senseless conflict that highlighted the dangers of nationalist fervor and territorial disputes. Through his analysis, the author sheds light on the complex historical and political dynamics that contributed to the outbreak of the war, offering valuable insights into the consequences of sports-related violence and interstate tensions.
Mixtral-8x7B-Instruct-v0.1	In her book "Men, Women, and Gods; and Other Lectures," Helen H. Gardener discusses several specific events related to the Cambodian-Vietnamese war, with a particular focus on the impact of these events on civilians and women. One event she highlights is the forced migration of millions of people during the war. According to Gardener, this displacement was often accompanied by violence, looting, and rape at the hands of soldiers from both sides. She notes that many women were left alone to care for their families after their husbands had been killed or conscripted into military service. These women faced significant challenges, including poverty, hunger, and lack of access to healthcare and education. Gardener also discusses the use of landmines during the conflict, which caused widespread devastation long after the fighting had ended. She cites statistics indicating that unexploded ordnance continued to kill or injure hundreds of people each year in Cambodia, even decades after the end of the war. Many of those affected were children who stumbled upon mines while playing outside. The author also focuses on the Khmer Rouge regime, which came to power in Cambodia following the Vietnamese invasion. During its rule, the Khmer Rouge committed numerous atrocities against the civilian population, including mass executions, forced labor, and torture. Gardener argues that the suffering inflicted by the Khmer Rouge was largely ignored by the international community due to Cold War politics. Furthermore, Gardner criticizes the role of foreign powers, particularly the United States, in exacerbating tensions between Cambodia and Vietnam. She argues that US involvement in Southeast Asia created conditions that allowed extremist groups like the Khmer Rouge to gain power. Additionally, she suggests that American support for anti-communist regimes in the region contributed to the destabilization of Cambodia and led to increased human rights violations. Overall, Helen H. Gardener presents a critical perspective on the Cambodian-Vietnamese war and its impact on civilians, particularly women. She emphasizes the importance of addressing the root causes of conflict and holding perpetrators accountable for their actions. Her analysis underscores the need for greater attention to be paid to the experiences of ordinary people caught up in the midst of violent political struggles.

Table A.3: Model hallucinations

Source Text	LLM Hallucination
Although the Bolivians were better equipped than their adversaries, the latter were accustomed to the roughness of the climate and had a better knowledge of the ground.	In the aftermath of the battle, Scott expresses his disappointment and frustration with the performance of the Bolivian army. He notes that while they possessed superior weapons and equipment, their lack of discipline, poor training, and disorganization proved costly.

Table A.4: Aligned text with LLM Hallucination output