Detecting Data Contamination in DeepSeek

Shaowei Zhang and Zhaokun Wang

Keywords: Data Contamination, Large Language Models, Detection Methods, DeepSeek V3, DeepSeek R1

Abstract.

This report examines potential data contamination in DeepSeek V3 and R1 language models. Findings show both models demonstrate small but noticeable data leakage in specific test datasets. When comparing BLEURT and ROUGE-L scores, contamination differences appear based on measurement tools used. This reveals the need for using multiple evaluation methods. Furthermore, DeepSeek-R1, which undergoes additional Reinforcement Learning (RL) and Supervised Fine-Tuning (SFT) stages, displays higher contamination signals than DeepSeek-V3. These findings underscore the need for robust detection techniques tailored to mixture-of-experts architectures and models with complex training pipelines.

1 Background

As LLMs are gaining reputation and attention for their performances and abliities in reasoning, understanding human instructions and generalizing in various tasks[1, 2], they are being trained with massive webtexts collecting from the Internet like the Common Crawl dataset[3] or OpenWebText dataset[4], which has brought the risk for data contamination in the pre-training corpus[5, 6, 7]. Trained on contaminated data, LLMs may achieve higher scores during testing than their actual capabilities. This is a damaging effect on the evaluation of LLMs' abilities. Therefore, it is necessary to detect data contamination in LLMs. [1]

2 Algorithm

In this project, we will employ the algorithm presented in [6]. Based on two assumptions: (1) The pre-training data and weights of LLMs are not easy to directly access. (2) Computational resources are limited, an inexpensive and robust method is proposed, which leverages "guided instruction" detect contamination from instance-level to partition-level.

A "guided instruction" is a prompt instructs the model to finish a sample from a dataset, which contains the name of the dataset and from which split (train, test or validation) the sample is collected. And a "general instruction" doesn't contain the information of the dataset. An example is shown in 1.

For instance-level detection, two methods are introduced. First, BLEURT[8] and ROUGE-L[9] are used to measure the semantic and lexical similarity between the completion from LLMs and the reference. Once the LLMs perform better on guided instructions than general instructions, the sample is detected as an instance-level contamination. Second, the completion from LLMs and the reference will be inputted to GPT-4 to determine if they are exact/near-match leverage the ICL(In-Context-Learning) ability[1] instead of human judgements.

For partition-level detection, two algorithms are developed. First, given ten instances, if average overlap score (measured by BLEURT and ROUGE-L) on completions based on guided instructions is statistically significantly better those based on general instructions under a non-parametric bootstrap resampling test, the data partition will be labeled as contaminated. Second,

[1] See the report's repository at `https://github.com/BufferHund/time-travel-in-llms`.

given ten instances, the data partition will be labeled as contaminated, when at least one completion from LLMs is detected as exact-match or at least two are detected as near-match.

Instruction: You are provided with Sentence 1 from the validation split of the WNLI dataset. Finish Sentence 2 as appeared in the dataset. Sentence 2 must exactly match the instance in the dataset.

Sentence 1: The dog chased the cat, which ran up a tree. It waited at the top.

Label: 1 (entailment)

Sentence 2:

`The cat waited at the top.`

Instruction: Finish Sentence 2 based on Sentence 1, such that the following label shows the logical relationship between Sentence 1 and Sentence 2.

Sentence 1: The dog chased the cat, which ran up a tree. It waited at the top.

Label: 1 (entailment)

Sentence 2:

`The cat was at the top of the tree after being chased by the dog.`

Figure 1: An example of a guided instruction (left) and general instruction (right). LLMs prompted with a guided instruction is believed to perform better.[6]

3 Model

The models to detect are DeepSeek-R1[10] and DeepSeek-V3[11]. DeepSeek-V3 is a powerful Mixture-of-Experts (MoE) language model with a total of 671 billion parameters, of which 37 billion are activated for each token. It inherits the Multi-head Latent Attention (MLA) and DeepSeekMoE[12] architectures from DeepSeek-V2[13] and introduces an auxiliary-loss-free load balancing strategy and a multi-token prediction training objective to enhance performance.

Based on Deepseek-V3, Deepseek-R1 is a model trained with GRPO[14], a Reinforcement Learning (RL) framework to enhance strong reasoning capabilities. During the training procedure, the model first undergoes a cold-start stage with high-quality Supervised Fine-Tuning (SFT) data, followed by a reasoning-oriented RL stage to improve performance on reasoning tasks. When the reasoning-oriented RL converges, the model is used to collect SFT data for subsequent stages. Then the model is trained in a SFT objective with both reasoning and non-reasoning data. Finally, RL is applied again to optimize general abilities, including helpfulness and harmlessness.

The impressive performance of DeepSeek-V3 and DeepSeek-R1 on various benchmarks has led us to question whether data contamination might have occurred during their training processes. Therefore, in this project, we will conduct separate evaluations to detect data contamination of these two models.

4 Datasets and Experimental Setup

4.1 Dataset Selection

For our contamination analysis, we selected a diverse range of benchmark datasets commonly used for LLM evaluation:

- Natural Language Understanding Tasks:
  - WNLI (Winograd Natural Language Inference)
  - RTE (Recognizing Textual Entailment)
- Sentiment Analysis Tasks:
  - IMDB (Internet Movie Database)
  - Yelp (User Reviews)

Table 1: Summary of Datasets and Their Characteristics

Dataset	Task Type	Size (Train/Test)	Key Characteristics
AG News	Text Classification	120K/7.6K	News article classification
IMDB	Sentiment Analysis	85K movies	Movie metadata with ratings/reviews
RTE	Textual Entailment	2.5K/3K	Combined from RTE1-5
SAMSum	Dialogue Summarization	16K	conversations with summaries
WNLI	Natural Language Inference	634/146	GLUE benchmark sentence pairs
XSum	Extreme Summarization	204K/11K	Single-sentence news summaries
Yelp	Sentiment Analysis	6.9M reviews	User reviews with star ratings

- Text Classification Tasks:
  - AG News (News Article Classification)
- Dialogue Summarization Tasks:
  - SAMSum (Summarizing Messenger-style Conversations)
- Generation Tasks:
  - XSum (Extreme Summarization)

Each dataset includes both training and test/validation partitions to enable comprehensive contamination detection across different data splits.

4.2 Experimental Procedure

We implemented the contamination detection methodology from Golchin and Surdeanu with the following workflow:

1. For each dataset and partition, we randomly sampled 10 instances for testing.
2. We used the crafted guided and general instructions for each instance.
3. We queried both DeepSeek-V3 and DeepSeek-R1 with these instructions.
4. We analyzed the responses using both semantic similarity metrics (BLEURT, ROUGE-L) and GPT-4-based classification.
5. We applied both instance-level and partition-level detection algorithms to identify potential contamination.

5 Results and Analysis

5.1 Contamination Detection Findings

The evaluation employed two primary approaches from Algorithm 1—semantic similarity scoring with BLEURT and lexical overlap measurement with ROUGE-L—alongside Algorithm 2's GPT-4-based In-Context Learning (ICL) classification for exact or near-exact matches. Additionally, human evaluations were conducted to provide a complementary perspective.

Table 2 summarizes the overall accuracy of our methods across 28 configurations, encompassing two models evaluated on 14 dataset partitions from seven different benchmarks. Consistent with the original study, GPT-4 ICL demonstrated significantly high accuracy. However, the accuracy observed in our experiments was lower than previously reported values. A plausible

Table 2: The overall accuracy of Deepseek and Deepseek-r1 in identifying contamination.

	Deepseek-R1		GPT-3.5	
Method	**Success Rate**	**Accuracy**	**Success Rate**	**Accuracy**
Algorithm 1: BLEURT	9/14	64.29%	11/14	78.57%
Algorithm 1: ROUGE-L	7/14	50.00%	9/14	64.29%
Algorithm 2: GPT-4 ICL	10/14	71.43%	13/14	92.86%

explanation for this discrepancy is that DeepSeek models, beyond providing direct answers, frequently incorporate additional reasoning or justifications in their outputs. This supplementary information may inadvertently confound GPT-4 during evaluation, leading to a relative reduction in accuracy compared to prior findings.

Our experiment results detailed in Table 3 revealed that both DeepSeek-V3 and DeepSeek-R1 models show some evidence of contamination across certain datasets, although the overall level of contamination appears to be low.

Interestingly, the evaluation using BLEU and ROUGE-L metrics yielded differing results. Specifically, BLEU scores showed no statistically significant differences between model completions generated from guided instructions compared to general instructions. However, the ROUGE-L metric indicated the opposite, revealing statistically significant differences across several datasets.

The GPT-4 ICL method identified some contamination, particularly in DeepSeek-V3's test/validation split, where multiple datasets received at least one exact (✓) or near-exact match (✓✓). However, the majority of dataset partitions were marked with (×), indicating no strong evidence of direct contamination. These findings are consistent with human evaluations. AG News and WNLI show the most notable evidence of leakage, while datasets like IMDB and RTE are more affected in training.

These findings highlight the complexity of contamination assessment—while traditional similarity metrics may overestimate contamination, human evaluations and GPT-4 ICL provide a more accurate view.

5.2 Analysis

These results suggest that while the contamination is not as severe as observed in some other large language models (like GPT-4), the DeepSeek models are not entirely free from contamination. Particularly, datasets for natural language inference tasks such as WNLI and RTE show consistent signs of contamination across both models.

The variations in contamination levels may reflect differences in task types, dataset characteristics, and the impact of model architectures and training methods. For instance, DeepSeek-V3's mixture-of-experts architecture might influence its memorization and generalization patterns, while DeepSeek-R1's reinforcement learning training process could alter how pre-training information is stored and retrieved.

Notably, the DeepSeek-V3 model appears to exhibit less pronounced data leakage on specific datasets, such as the WNLI training set, XSum test set, and RTE test set, compared to DeepSeek-R1, where leakage was more readily detected. This could be attributed to the Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL) processes enhancing the model's capabilities. Alternatively, data leakage might have occurred during the RL or SFT stages, thus influencing the results.

Despite these findings, most datasets still show low levels of contamination, suggesting that the DeepSeek team may have implemented effective measures to reduce data contamination

Table 3: We evaluate Algorithm 1 using BLEURT and ROUGE-L, as well as Algorithm 2 which relies on GPT-4 decisions via few-shot ICL prompting. The analysis covers 10 randomly selected samples from each dataset segment, focusing on two specific language models. Contamination at the segment level is represented with the following symbols: (1) Stars (*) indicate significant differences between outputs from detailed versus broad instructions, as determined by the similarity metrics, while underlined numbers show setups aligning with human judgments. (2) A single check mark (✓) signifies at least one exact match, a double check mark (✓✓) denotes two or more near-exact matches, and a cross (×) means neither of these conditions is met.

				Datasets						
Model	**Method**	**Split**	**Instruct.**	**IMDB**	**AG News**	**Yelp**	**RTE**	**WNLI**	**SAMSum**	**XSum**
Deepseek-R1	Alg. 1: BLEURT	Train	General	0.609	0.551	0.579	0.433	0.51	0.66	0.545
			Guided	0.622	0.508	0.592	0.425	0.504	0.465	0.417
		Test/Valid	General	0.609	0.61	0.517	0.466	0.496	0.697	0.56
			Guided	0.616	0.519	0.534	0.449	0.526	0.499	0.438
	Alg. 1: ROUGE-L	Train	General	0.106	0.071	0.117	0.117	0.135	0.073	0.119
			Guided	*0.162	*0.241	*0.148	*0.266	*0.322	0.102	0.188
		Test/Valid	General	0.119	0.077	0.11	0.084	0.126	0.062	0.133
			Guided	0.129	*0.212	*0.136	*0.217	*0.318	*0.139	*0.224
	Alg. 2: GPT-4 ICL	Train	Guided	×	×	×	×	✓	×	×
		Test/Valid	Guided	×	×	×	×	×	×	×
	Human Evaluation	Train	Guided	×	×	×	✓	✓✓	×	×
		Test/Valid	Guided	×	×	×	✓	✓	×	✓✓
Deepseek-V3	Alg. 1: BLEURT	Train	General	0.638	0.589	0.646	0.449	0.526	0.558	0.583
			Guided	0.658	0.59	0.646	0.531	0.573	0.571	0.542
		Test/Valid	General	0.658	0.583	0.593	0.491	0.524	0.566	0.654
			Guided	0.659	0.603	0.597	0.458	0.595	0.578	0.612
	Alg. 1: ROUGE-L	Train	General	0.137	0.116	0.139	0.152	0.122	0.107	0.168
			Guided	0.141	*0.151	0.152	*0.337	*0.55	*0.174	0.223
		Test/Valid	General	0.131	0.12	0.135	0.109	0.154	0.093	0.215
			Guided	0.14	*0.176	0.151	*0.212	*0.568	*0.142	0.281
	Alg. 2: GPT-4 ICL	Train	Guided	×	×	×	✓	×	×	×
		Test/Valid	Guided	×	×	×	×	✓	×	×
	Human Evaluation	Train	Guided	×	×	×	✓✓	✓✓	×	×
		Test/Valid	Guided	×	×	×	×	✓	×	×

during the training process. However, these results also highlight the importance of using multiple detection methods, as different approaches may reveal different patterns of contamination.

6 Conclusion

For this project, we conducted the investigation into data contamination within the DeepSeek-V3 and DeepSeek-R1 models by adopting the ”Time Travel in LLMs” method.Our findings indicate that both DeepSeek-V3 and DeepSeek-R1 exhibit low but detectable levels of data leakage, particularly on datasets like WNLI and RTE, with DeepSeek-R1 showing slightly elevated contamination signals, possibly attributable to its additional Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL) stages. The differing contamination patterns observed between BLEURT and ROUGE-L show the need for multiple evaluation metrics to fully capture the complex nature of data contamination. Although the contamination levels in DeepSeek models are less severe compared to some other LLMs, our results suggest that their unique mixture-of-experts architecture and intricate training pipeline may either effectively reduce contamination or mask it in ways that challenge conventional detection methods like ”Time Travel in LLMs”. These findings emphasize the importance of refining and tailoring detection techniques to accommodate advanced model architectures and training paradigms, ensuring more accurate assessments of data integrity in future LLM evaluations.

7 Contributions

We have contributed to different parts of this project. Zhaokun Wang modified, debugged, and ran the code to obtain experimental results. Shaowei Zhang analyzed most of the experimental results. Together, we collaborated on writing the paper.

References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[2] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

[3] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023.

[4] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners, 2019.

[5] Yonatan Oren, Nicole Meister, Niladri Chatterji, Faisal Ladhak, and Tatsunori B. Hashimoto. Proving test set contamination in black box language models, 2023.

[6] Shahriar Golchin and Mihai Surdeanu. Time travel in llms: Tracing data contamination in large language models, 2024.

[7] Inbal Magar and Roy Schwartz. Data contamination: From memorization to exploitation, 2022.

[8] Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. Bleurt: Learning robust metrics for text generation, 2020.

[9] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[10] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.

[11] DeepSeek-AI. Deepseek-v3 technical report, 2025.

[12] Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y. K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models, 2024.

[13] DeepSeek-AI. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model, 2024.

[14] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024.