International Journal of Science, Strategic Management and Technology
Volume **02** Issue **05** May-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

A Controlled Study on the Impact of Input-Only vs.Input-Output Contamination on Benchmark Evaluation of Large Language Models

Vansh Deol, Danish Iqbal, Sami Ahmad
Department of Information Technology Noida Institute of Engineering & Technology Greater Noida, India
vanshdeol914@gmail.com, danishiqbal6262@gmail.com, samiahmad2023@gmail.com

https://doi.org/10.55041/ijsmt.v2i5.310

Cite this Article: Deol, V., Ahmad, S. & Iqbal, D. (2026). A Controlled Study on the Impact of Input-Only Vs.Input-Output Contamination on Benchmark Evaluation of Large Language Models. International Journal of Science, Strategic Management and Technology, *02*(05). https://doi.org/10.55041/ijsmt.v2i5.310

***Abstract*—**Benchmark datasets are widely used to evaluate Large Language Models (LLMs), but the inclusion of evaluation data in pretraining corpora raises significant contamination concerns. Such exposure can artificially inflate performance, obscuring true model capabilities. This work conducts a controlled empirical study of benchmark contamination using four open- weight models (TinyLlama, Qwen2.5-1.5B, Phi-2, and Gemma- 2B-it) evaluated on OpenBookQA. We compare two conditions via LoRA finetuning: input-only contamination (exposure to questions) and input-output contamination (exposure to questions and answers).

Beyond standard accuracy, we evaluate contamination using behavioral metrics, including prediction agreement and prediction transitions. Our results demonstrate that input-output contami- nation consistently drives stronger performance gains and larger behavioral shifts. Notably, partial benchmark exposure (input- only) substantially alters benchmark prediction behavior, whereas explicit answer exposure (input-output) primarily promotes memorization. Furthermore, contamination substantially alters prediction behavior even when aggregate benchmark accuracy changes remain small. These findings demonstrate that accuracy alone is an insufficient metric for capturing contamination-induced changes, highlighting the necessity of behavioral evaluation in LLM assessment.

Index Terms—Benchmark contamination, Large Language Models, LLM evaluation, memorization, OpenBookQA, behavioral analysis, LoRA

I. INTRODUCTION

Large Language Models (LLMs) are increasingly evaluated using standardized benchmark datasets to measure reasoning, knowledge retrieval, and general language understanding ca- pabilities. Benchmarks such as OpenBookQA [1], MMLU, and GSM8K have become widely adopted as indicators of model performance and progress. However, the growing scale of pretraining corpora and public benchmark availability have raised concerns regarding benchmark contamination, where evaluation examples or semantically similar content may appear during model training. This contamination can artificially inflate benchmark performance and compromise the validity of reported evaluation results.

Benchmark contamination presents a significant challenge for reliable model assessment because improvements in benchmark accuracy may not necessarily reflect genuine reasoning or generalization capabilities. Instead, performance gains may partially result from memorization or exposure to benchmark- related content during training. Prior work has explored contamination detection and memorization in large language models, but many studies focus primarily on overall accuracy changes while providing limited analysis of how contamination alters prediction behavior across different model families and exposure settings.

In this work, we conduct a controlled empirical study of benchmark contamination using multiple open-weight language models and synthetic contamination conditions. We investigate two distinct contamination scenarios: input-only contamination, where models are exposed only to benchmark questions, and input-output contamination, where models

are exposed to both questions and corresponding answers. Using the OpenBookQA benchmark [1], we evaluate contamination effects across multiple exposure sizes, random seeds, and model architectures, including TinyLlama [2], Qwen2.5-1.5B [3], Phi-2 [4], and Gemma-2B-it [5].

Beyond standard accuracy evaluation, we analyze behavioral changes introduced by contamination through transition-based metrics and prediction agreement analysis. Specifically, we examine wrong-to-correct and correct-to-wrong prediction transitions relative to clean baseline models, allowing us to measure how contamination modifies model behavior rather than only final benchmark scores. Our experiments reveal substantial heterogeneity in contamination effects across models, exposure levels, and contamination conditions. In particular, input-output contamination consistently produces stronger behavioral shifts and larger performance gains than input-only contamination, while weaker models exhibit limited sensitivity to contamination exposure.

The contributions of this paper are summarized as follows:

- We introduce a behavioral contamination analysis pipeline using prediction agreement and transition metrics (wrong-to-correct and correct-to-wrong transitions) to evaluate contamination beyond standard benchmark accuracy.
- We present a controlled multi-model study comparing input-only and input-output benchmark contamination in open-weight language models, analyzing sensitivity across varying exposure sizes and multiple random seeds.
- We demonstrate that contamination effects vary substan- tially across model families, with stronger models exhibiting larger contamination-induced behavioral changes than smaller baseline models.

II. Related Work

A. *Benchmark Contamination in Language Models*

Benchmark contamination has become an increasingly im- portant concern in the evaluation of Large Language Models (LLMs). As modern language models are trained on massive internet-scale corpora, benchmark datasets used for evaluation may unintentionally appear within training data. This overlap can artificially inflate benchmark performance and reduce confidence in reported evaluation results. Several studies have shown that benchmark memorization and dataset overlap may significantly affect evaluation reliability, particularly for widely used public benchmarks [6]–[8].

Prior research has explored contamination through methods such as n-gram overlap analysis, dataset reconstruction, bench- mark auditing, and memorization detection [9]. Investigations involving benchmarks such as MMLU, GSM8K, HumanEval, and other standardized evaluation datasets have identified varying levels of overlap between benchmark content and model pretraining corpora. These findings have raised concerns that benchmark accuracy may not always reflect genuine reasoning or generalization capability.

The contamination problem is particularly difficult to address in large pretrained language models because pretraining datasets are often partially undisclosed and extremely large in scale. As a result, identifying direct or indirect benchmark exposure becomes challenging. Existing work has therefore emphasized the need for contamination-aware evaluation protocols and more robust methods for assessing language model reasoning ability.

While prior studies primarily focus on contamination detec- tion and overall benchmark accuracy changes, less attention has been given to how contamination alters prediction be- havior across different model architectures and contamination conditions. In this work, we extend contamination analysis beyond aggregate accuracy by evaluating behavioral changes introduced through controlled contamination exposure.

B. *Memorization and Behavioral Evaluation in Language Models*

Memorization in language models refers to the ability of models to recall or reproduce information encountered during training. Previous studies have demonstrated that transformer- based language models [10] can memorize rare sequences, factual statements, benchmark examples, and other training artifacts under certain conditions. Larger models often exhibit stronger memorization capacity due to increased parameter count and representational complexity.

Several works have investigated the relationship between memorization and benchmark performance. Models exposed to benchmark-related content during training may achieve artificially elevated benchmark accuracy, suggesting that memorization can contribute directly to evaluation performance. However, contamination effects are not always uniform across tasks or architectures, and different models may exhibit substantially different sensitivity to benchmark exposure.

At the same time, recent research in language model evaluation has increasingly emphasized behavioral analysis beyond standard accuracy metrics. Studies examining prediction consistency, calibration, confidence estimation, reasoning be-

havior, and output agreement suggest that aggregate benchmark accuracy alone may not fully capture model behavior under different evaluation conditions.

Motivated by these observations, this work incorporates behavioral contamination analysis through prediction agreement and transition-based metrics. By analyzing wrong-to-correct and correct-to-wrong prediction transitions relative to clean baseline models, our study evaluates how contamination changes model prediction behavior in addition to overall benchmark accuracy. This approach provides a more detailed understanding of contamination-induced behavioral shifts across multiple language model families and exposure settings.

III. METHODOLOGY

A. Research Objective

The objective of this study is to analyze how benchmark contamination affects the evaluation behavior of large language models under controlled experimental conditions. Specifically, the study investigates whether exposure to benchmark-related content during finetuning alters benchmark accuracy and prediction behavior across different contamination settings and model architectures.

The study focuses on two primary research questions:

1) How do different contamination conditions influence benchmark performance across language models?
2) How does benchmark contamination alter model predic- tion behavior beyond overall benchmark accuracy?

To address these questions, controlled contamination datasets were constructed and evaluated using multiple open-weight language models under standardized evaluation settings.

B. Benchmark and Contamination Design

All experiments were conducted using the OpenBookQA benchmark [1], a multiple-choice science reasoning dataset containing question-answer pairs with four candidate answer options labeled A–D. The benchmark was selected because its structured format enables deterministic evaluation and controlled contamination construction.

Two contamination conditions were evaluated:

a) Input-Only (IO) Contamination: Models were exposed only to benchmark questions during finetuning. Ground-truth answers were intentionally excluded to evaluate whether exposure to benchmark inputs alone influences benchmark prediction behavior.

TABLE I: Evaluated language models.

Model	**Parameters**	**Description**
TinyLlama-1.1B-Chat-v1.0	1.1B	Small instruction-tuned baseline model
Qwen2.5-1.5B-Instruct	1.5B	Strong small-scale instruction-tuned model
Phi-2	2.7B	Compact reasoning-oriented language model
Gemma-2B-it	2B	Instruction-tuned open-weight model

Parameter	Value
Finetuning Method	LoRA
Epochs	1
Batch Size	1
Maximum Sequence Length	256
LoRA Rank (r)	8
LoRA Alpha	16
LoRA Dropout	0.05
Random Seeds	42, 123, 999

TABLE II: LoRA finetuning configuration.

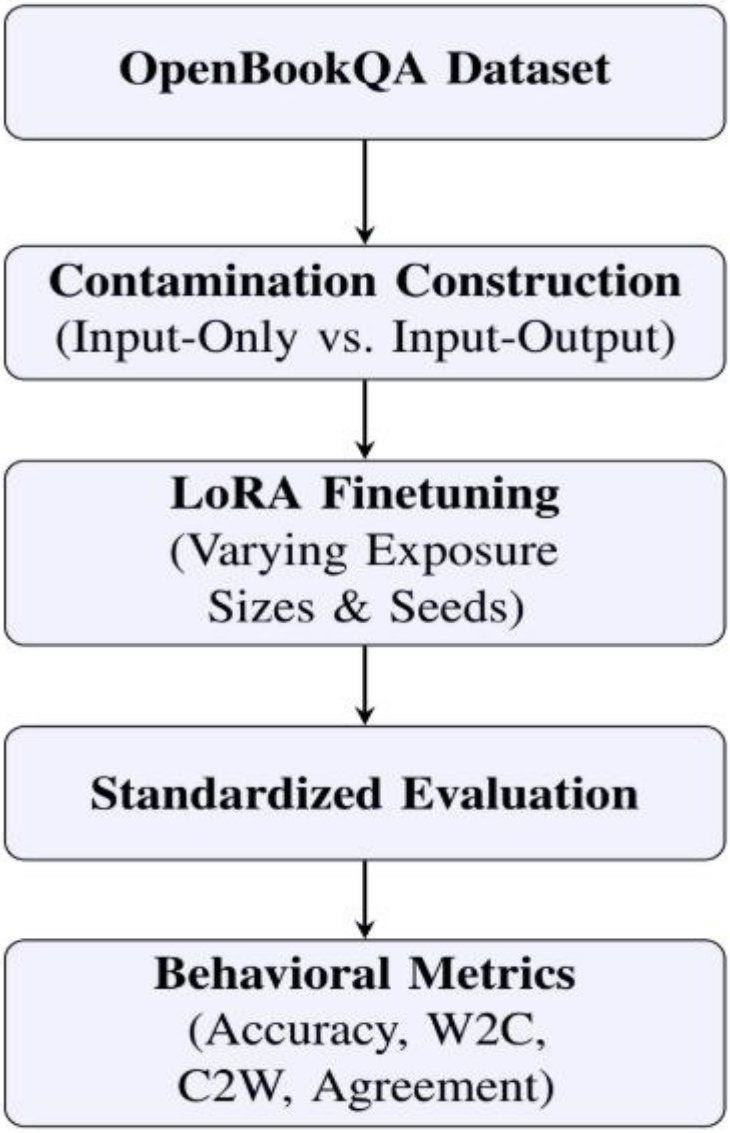

Fig. 1: Methodology pipeline illustrating the experimental workflow from dataset selection to behavioral metric evaluation.

b) Input-Output (I-Out) Contamination: Models were exposed to both benchmark questions and their corresponding correct answers during finetuning. This condition represents direct benchmark leakage.

To analyze contamination scaling effects, experiments were conducted using three contamination exposure sizes:

- 500 examples
- 1000 examples
- 2000 examples

The held-out evaluation split remained fixed across all experi- ments to ensure consistent comparison between contamination conditions and exposure settings.

C. *Behavioral Evaluation Metrics*

In addition to benchmark accuracy, the study incorporates be- havioral evaluation metrics designed to analyze contamination- induced prediction changes.

a) Prediction Agreement: Prediction agreement measures the proportion of predictions that remain identical between contaminated models and their corresponding clean baseline models. Lower agreement indicates stronger behavioral diver- gence caused by contamination exposure.

b) Transition Analysis: Prediction transitions were ana- lyzed relative to clean baseline models using two transition categories:

- **Wrong-to-Correct (W2C):** predictions that were incorrect in the baseline model but became correct after contami- nation finetuning.
- **Correct-to-Wrong (C2W):** predictions that were correct in the baseline model but became incorrect after contami- nation finetuning.

These behavioral metrics provide additional insight into contamination-induced changes that may not be fully captured by aggregate benchmark accuracy alone.

IV. Experimental Setup

A. *Evaluated Language Models*

Four open-weight decoder-only language models were evalu- ated in this study (summarized in Table I). The selected models provide diversity in capability, architecture, and instruction- tuning behavior while remaining computationally feasible for repeated controlled experimentation.

B. *Training Configuration*

All contamination experiments were performed using LoRA (Low-Rank Adaptation) parameter-efficient finetuning [11]. LoRA adapters were attached to transformer attention layers while keeping pretrained model weights frozen. The training configuration used across all experiments is shown in Table II. Separate LoRA adapters were trained independently for each contamination condition, exposure size, random seed, and model configuration.

Parameter	Value
Decoding Strategy	Greedy
do_sample	False
max_new_tokens	5
Temperature	0.0

TABLE III: Decoding configuration.

C. *Evaluation Procedure*

All models were evaluated on the same held-out OpenBookQA test set using deterministic greedy decoding. Each evaluation prompt contained the benchmark question, four answer choices labeled A–D, and a standardized answer generation format. Generation was performed using the config- uration in Table III.

Predicted outputs were parsed to extract the generated answer label, and benchmark accuracy was computed as the proportion of correctly predicted answers.

D. *Hardware and Reproducibility*

Initial development and debugging experiments were con- ducted on an Apple Silicon M-series device using the Metal Performance Shaders (MPS) backend. Large-scale experimental runs were executed on a Windows-based system equipped with an NVIDIA RTX 3050 GPU. The experimental pipeline was implemented using Python, PyTorch, Hugging Face Transformers, Datasets, and the PEFT library.

To ensure reproducibility, automated orchestration scripts were developed to execute experiments across all combinations

of:

- model architectures,
- contamination conditions,
- exposure sizes,
- and random seeds.

Each experimental run stored configuration metadata, prediction outputs, accuracy statistics, transition metrics, and agreement metrics. This structured experimental pipeline ensured repro- ducible evaluation and prevented result overwriting across large experimental grids.a 0.576 baseline. However, input-only contamination causes severe degradation at lower exposure levels (0.284 at 1000 samples). This represents a core finding: partial benchmark exposure substantially alters benchmark prediction behavior, while explicit answer exposure primarily promotes memoriza- tion behavior.

TinyLlama-1.1B demonstrates weak contamination sensitiv- ity across all settings, remaining close to the clean baseline. Gemma-2B-it exhibits a negative contamination profile, where exposure generally lowers benchmark accuracy compared to the clean baseline.

Overall, input-output contamination consistently produces stronger performance shifts than input-only contamination, though the direction of this shift is heavily model-dependent.

B. Behavioral Transition Analysis

Aggregate accuracy obscures internal prediction instability. To analyze these behavioral shifts, we evaluate W2C and C2W transitions relative to clean baseline models. Table V presents the transition counts across exposure sizes.

Qwen2.5-1.5B demonstrates a consistent increase in W2C transitions under input-output contamination, rising from 47.3 to 58.3 as exposure scales. C2W transitions remain low, indicating predictions shift favorably.

Phi-2 exhibits extreme behavioral changes. Under input- output contamination, W2C transitions spike to 85.0. Under input-only contamination, C2W transitions explode to 162.3 at 1000 samples. This massive C2W instability indicates partial benchmark exposure heavily disrupts existing benchmark prediction behavior.

TinyLlama reveals an essential behavioral pattern: despite remarkably stable benchmark accuracy (Table IV), transition counts fluctuate wildly (e.g., up to 69.3 C2W transitions under input-only). This strongly supports our core claim: accuracy alone hides severe contamination-driven prediction instability. Gemma-2B-it similarly exhibits high C2W transitions across all conditions, explaining its overall benchmark degradation.

V. RESULTS

A. Accuracy Trends Under Contamination

We first analyze how contamination affects OpenBookQA accuracy across exposure sizes. Table IV presents benchmark accuracy for all evaluated models under input-only and input- output contamination.

Contamination effects vary substantially across model fami- lies. Qwen2.5-1.5B-Instruct demonstrates clear contamination sensitivity. Under input-output contamination, benchmark accu- racy increases consistently with exposure size, improving from 0.762 at 500 samples to 0.778 at 2000 samples. Conversely, input-only contamination produces only modest gains.

Phi-2 exhibits extreme contamination asymmetry. Input- output contamination substantially improves benchmark ac- curacy, reaching 0.699 at 2000 exposure samples compared to

C. Prediction Agreement Analysis

We measure prediction agreement to quantify the raw propor- tion of predictions remaining unchanged after contamination finetuning. Table VI presents these rates.

Qwen2.5-1.5B demonstrates progressively decreasing agree- ment under input-output exposure (down to 0.776), signifying a steady behavioral drift away from the baseline. Phi-2 shows severe instability under input-only exposure, dropping to a low 0.438 agreement rate at 1000 samples.

Even models with flat accuracy trajectories, like TinyLlama and Gemma-2B-it, exhibit noticeable shifts in raw agree-

ment. Across all evaluations, agreement rates indicate that contamination consistently changes prediction behavior, an observation entirely masked when only reviewing aggregate accuracy metrics.TABLE IV: Main benchmark accuracy results averaged across three random seeds.

Model	Exposure	Base Accuracy	Input-Only Accuracy	Input-Output Accuracy
Qwen2.5-1.5B-Instruct	500	0.720 ± 0.000	0.746 ± 0.005	**0.762** ± 0.003
Qwen2.5-1.5B-Instruct	1000	0.720 ± 0.000	0.752 ± 0.000	**0.780** ± 0.004
Qwen2.5-1.5B-Instruct	2000	0.720 ± 0.000	0.759 ± 0.019	**0.778** ± 0.045
TinyLlama-1.1B-Chat-v1.0	500	**0.284** ± 0.000	0.271 ± 0.017	0.276 ± 0.000
TinyLlama-1.1B-Chat-v1.0	1000	**0.284** ± 0.000	0.270 ± 0.011	0.276 ± 0.000
TinyLlama-1.1B-Chat-v1.0	2000	0.284 ± 0.000	**0.289** ± 0.008	0.278 ± 0.005
gemma-2b-it	500	**0.463** ± 0.001	0.378 ± 0.007	0.374 ± 0.009
gemma-2b-it	1000	**0.463** ± 0.001	0.374 ± 0.010	0.377 ± 0.015
gemma-2b-it	2000	**0.463** ± 0.001	0.383 ± 0.013	0.389 ± 0.013
phi-2	500	0.576 ± 0.000	0.318 ± 0.112	**0.594** ± 0.059
phi-2	1000	0.576 ± 0.000	0.284 ± 0.201	**0.668** ± 0.014
phi-2	2000	0.576 ± 0.000	0.500 ± 0.124	**0.699** ± 0.005

TABLE V: Transition statistics averaged across three random seeds.

Model	Exposure	Input-Only (W2C)	Input-Output (W2C)	Input-Only (C2W)	Input-Output (C2W)
Qwen2.5-1.5B-Instruct	500	38.3 ± 1.5	**47.3** ± 1.2	25.3 ± 2.5	**26.3** ± 0.6
Qwen2.5-1.5B-Instruct	1000	45.3 ± 1.2	**53.0** ± 0.0	**29.3** ± 1.2	23.0 ± 2.0
Qwen2.5-1.5B-Instruct	2000	48.3 ± 1.5	**58.3** ± 3.8	29.0 ± 7.9	**29.3** ± 18.8
TinyLlama-1.1B-Chat-v1.0	500	**49.0** ± 16.5	21.0 ± 0.0	**55.7** ± 20.2	25.0 ± 0.0
TinyLlama-1.1B-Chat-v1.0	1000	**62.3** ± 35.9	21.0 ± 0.0	**69.3** ± 39.3	25.0 ± 0.0
TinyLlama-1.1B-Chat-v1.0	2000	**38.7** ± 13.9	23.0 ± 2.6	**36.3** ± 11.8	26.0 ± 1.0
gemma-2b-it	500	**20.7** ± 6.0	20.0 ± 7.0	63.3 ± 4.6	**64.7** ± 5.5
gemma-2b-it	1000	**34.3** ± 8.0	22.3 ± 5.0	**79.0** ± 6.1	65.7 ± 3.2
gemma-2b-it	2000	**31.0** ± 10.6	23.3 ± 11.0	**71.0** ± 4.0	60.3 ± 7.6
phi-2	500	15.0 ± 4.6	**55.3** ± 2.9	**144.0** ± 51.6	46.3 ± 26.9
phi-2	1000	16.3 ± 13.5	**72.0** ± 7.2	**162.3** ± 88.0	26.0 ± 3.0
phi-2	2000	46.7 ± 12.3	**85.0** ± 3.0	**84.7** ± 50.1	23.7 ± 5.1

TABLE VI: Prediction agreement rates averaged across three random seeds.

Model	Exposure	Input-Only Agreement		Input-Output Agreement	
Qwen2.5-1.5B-Instruct	500	**0.847**	± 0.007	0.826	± 0.005
Qwen2.5-1.5B-Instruct	1000	**0.821**	± 0.004	0.817	± 0.008
Qwen2.5-1.5B-Instruct	2000	**0.803**	± 0.019	0.776	± 0.043
TinyLlama-1.1B-Chat-v1.0	500	0.627	± 0.146	**0.827**	± 0.001
TinyLlama-1.1B-Chat-v1.0	1000	0.545	± 0.280	**0.827**	± 0.001
TinyLlama-1.1B-Chat-v1.0	2000	0.725	± 0.096	**0.819**	± 0.007
gemma-2b-it	500	**0.743**	± 0.025	**0.743**	± 0.025
gemma-2b-it	1000	0.639	± 0.065	**0.730**	± 0.016
gemma-2b-it	2000	0.661	± 0.058	**0.745**	± 0.049
phi-2	500	0.495	± 0.155	**0.686**	± 0.054
phi-2	1000	0.438	± 0.223	**0.701**	± 0.023
phi-2	2000	0.609	± 0.120	**0.695**	± 0.020

VI. Discussion

The results of this study demonstrate that benchmark con- tamination substantially influences model prediction behavior,though these effects are highly dependent on model architecture, contamination type, and exposure size.

Consistently, input-output contamination produces stronger

effects than input-only contamination, driving larger accuracy gains, increased wrong-to-correct transitions, and lower baseline agreement. This implies that direct exposure to benchmark answers enables stronger memorization of benchmark-specific information compared to exposure to questions alone. Further- more, models like Qwen2.5-1.5B-Instruct exhibit a clear dose- response relationship, where increased exposure progressively shifts model behavior and improves aggregate performance.

However, contamination sensitivity varies significantly. While TinyLlama demonstrates minimal behavioral shifts, stronger models like Phi-2 exhibit extreme asymmetry. Under input-only contamination, Phi-2 suffers substantial benchmark degradation, suggesting that incomplete exposure substantially alters benchmark prediction behavior. Conversely, input-output contamination promotes strong memorization effects. Gemma- 2B-it further illustrates that contamination can interfere with stable evaluation, occasionally lowering benchmark accuracy relative to clean baselines.

Crucially, transition and agreement analyses reveal contamination-induced changes that remain hidden by aggre- gate performance metrics. Models frequently exhibit substantial prediction divergence despite relatively stable benchmark accuracy, demonstrating that contamination substantially alters prediction behavior. As benchmark datasets become increas- ingly integrated into training corpora, these findings underscore the necessity of adopting contamination-aware, behavioral evaluation practices to accurately assess language model capabilities.

VII. Limitations

Although this study provides a controlled analysis of bench- mark contamination behavior in open-weight language models, several limitations should be considered when interpreting the results.

First, the experiments were conducted using relatively small open-weight language models (1.1B to 2.7B parameters). While these models enable reproducible experimentation, larger frontier-scale language models may exhibit different contamination dynamics due to differences in pretraining scale, architecture, and memorization capacity. Consequently, the findings presented in this work may not fully generalize to substantially larger commercial models.

Second, the study focuses exclusively on the OpenBookQA benchmark. Contamination behavior may vary across different benchmark categories such as mathematical reasoning, coding tasks, or long-form generation. Additional evaluation across multiple benchmark families would provide a broader under- standing of contamination sensitivity.

Third, the contamination setup used in this work is synthetic and explicitly controlled. Models were intentionally exposed

to benchmark-related content during finetuning under predefined contamination conditions. Real-world contamination in large- scale pretraining corpora is typically indirect, noisy, and difficult to measure precisely. As a result, the experimental conditions in this study may not fully capture the complexity of naturally occurring benchmark contamination in large pretrained language models.

Finally, while the behavioral metrics introduced in this work provide insight beyond aggregate benchmark accuracy, they do not directly measure internal memorization mechanisms or causal reasoning processes. The observed behavioral shifts indicate contamination-induced prediction changes, but they do not fully explain the underlying representational or optimization dynamics responsible for these effects.

VIII. Conclusion

This work presented a controlled empirical study of bench- mark contamination in open-weight language models. By evaluating input-only and input-output contamination across multiple architectures and exposure sizes, we demonstrated that contamination substantially influences model prediction dy- namics. Specifically, partial benchmark exposure substantially alters benchmark prediction behavior, while explicit answer exposure primarily promotes rigid memorization.

Crucially, our behavioral evaluation metrics—prediction agreement and transition analysis—revealed that contamination significantly alters model behavior even when aggregate bench- mark accuracy remains relatively stable. This demonstrates that accuracy alone is an insufficient metric for robust contamination analysis.

We also observed substantial heterogeneity across model families; contamination susceptibility is not a uniform phe- nomenon, but depends heavily on architecture, contamination type, and baseline capabilities. As benchmark data increasingly permeates pretraining corpora, future research must prioritize comprehensive, behavioral evaluation methodologies to ensure the reliability of language model assessment.

References

[1] T. Mihaylov, P. Clark, T. Khot, and A. Sabharwal, "Can a suit of armor conduct electricity? a new dataset for open book question answering," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018. [Online]. Available: https://arxiv.org/abs/1809.02789

[2] TinyLlama Team, "Tinyllama: An open-source small language model," Tech. Rep., 2023, technical Report. [Online]. Available: https://arxiv.org/abs/2401.02385

[3] Qwen Team, "Qwen2.5 technical report," 2024, technical Report. [Online]. Available: https://qwenlm.github.io/blog/qwen2.5/

[4] M. Javaheripi, S. Bubeck, M. Abdin, J. Aneja, C. C. T. Mendes,W. Chen, A. Del Giorno, R. Eldan, S. Gopi, and S. Gunasekar, "Phi-2: The surprising power of small language models," Microsoft Research, Tech. Rep., 2023. [Online]. Available: https://www.microsoft.com/en-us/ research/blog/phi-2-the-surprising-power-of-small-language-models/

[5] Gemma Team, "Gemma: Open models based on gemini research and technology," *arXiv preprint*, 2024. [Online]. Available: https://arxiv.org/abs/2403.08295

[6] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal,
A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal,
A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin,S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford,
I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020. [Online]. Available: https://arxiv.org/abs/2005.14165

[7] OpenAI, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023. [Online]. Available: https://arxiv.org/abs/2303.08774

[8] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," OpenAI, Technical Report, February 2019. [Online]. Available: https://cdn.openai.com/better-language-models/language_ models_are_unsupervised_multitask_learners.pdf

[9] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee,A. Roberts, D. Song, U. Erlingsson, A. Oprea, and C. Raffel, "Extracting training data from large language models," in *USENIX Security Symposium*, 2021. [Online]. Available: https://arxiv.org/abs/2012.07805

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. [Online]. Available: https://arxiv.org/abs/1706.03762

[10] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," in *International Conference on Learning Representations (ICLR)*, 2022. [Online]. Available: https://arxiv.org/abs/2106.09685