International Platform of Registered Systematic Review and Meta-analysis Protocols

INPLASY

# Data Contamination in LLMs: A Scoping Literature Review

INPLASY2025110050
doi: 10.37766/inplasy2025.11.0050
Received: 17 November 2025
Published: 17 November 2025

Tudose, M; Ruseti, S; Dascalu, M; Caragea, C.

**Corresponding author:**
Mihai-Cristian Tudose

mihai.tudose1912@stud.acs.upb.ro

**Author Affiliation:**
National University of Science & Technology POLITEHNICA Bucharest.

**ADMINISTRATIVE INFORMATION**

**Support -** NA.

**Review Stage at time of this submission -** Completed but not published.

**Conflicts of interest -** None declared.

**INPLASY registration number:** INPLASY2025110050

**Amendments -** This protocol was registered with the International Platform of Registered Systematic Review and Meta-Analysis Protocols (INPLASY) on 17 November 2025 and was last updated on 17 November 2025.

## INTRODUCTION

**Review question / Objective** Our study presents recent research advances in data contamination across LLMs, followed by practical methods for detecting it. We used the PRISMA for Scoping Reviews (PRISMA-ScR) methodology to extract relevant studies and report our findings. Furthermore, we discuss and analyse the limitations and challenges faced by the methods, including issues related to data and algorithms. We address the following research questions:

RQ1: What types of data contamination exist, and how can these different types be distinguished from one another?
RQ2: What are the primary sources of data contamination, including the contexts and datasets from which they are derived?
RQ3: What are the strategies for detecting data contamination in various LLMs, and in what ways do these strategies vary across different contexts?
RQ4: What is the performance of the existing methods in identifying data contamination?

**Rationale** Data contamination poses a serious challenge because it undermines the ability to accurately measure an LLM's true performance. As a result, this leads to artificially inflated performance scores, giving researchers and developers a false sense of the model's capabilities.

**Condition being studied** NA.

## METHODS

**Search strategy** We conducted the search using 3 databases: Web of Science, arXiv.org, and Science Direct.
The search was limited to publications that were written in English and are subject to the "Computer Science" category published between 2023 and 2025.

The search query targeted the title and abstract using the following Boolean logic:

TS = (contamination OR leakage)
AND TS = ("large language models" OR LLMs)
AND TS = (data OR benchmark OR dataset)
AND WC = ("Computer Science")
AND PY = (2023-2025)
AND DT = ("article" OR "proceedings paper").

**Participant or population** This does not apply to our review.

**Intervention** This does not apply to our review.

**Comparator** This does not apply to our review.

**Study designs to be included** This does not apply to our review.

**Eligibility criteria** Inclusion criteria: (1) published between 2023 and 2025, (2) articles published in English, and (3) were subject to the field of Computer Science.

**Information sources** The Web of Science, Science Direct, and arXiv.org databases, as they existed on November 3, 2025, in their state from 3 November 2025.

**Main outcome(s)** This study employed the PRISMA methodology to systematically retrieve recent papers on data contamination from arXiv.org, Web of Science, and ScienceDirect. The selected literature was analysed for publication trends (per year), institutional contributions, and benchmark frequency. To answer our research question, we then synthesised the findings from the selected literature.

**Additional outcome(s)** The performance of LLMs is continuously and negatively impacted by data contamination. Therefore, benchmark results fail to represent their genuine reasoning power, as the models are merely memorising training examples instead of learning underlying patterns.

**Data management** This does not apply to our review.

**Quality assessment / Risk of bias analysis** This does not apply to our review.

**Strategy of data synthesis** This does not apply to our review.

**Subgroup analysis** This does not apply to our review.

**Sensitivity analysis** This does not apply to our review.

**Language restriction** English.

**Country(ies) involved** Romania.

**Other relevant information** Although prior surveys on this topic exist, our study is the first to apply the PRISMA methodology.

**Keywords** Scoping Data contamination; Large Language Models (LLM); Benchmark leakage; Natural Language Processing.

**Dissemination plans** Publication in Journal..

**Contributions of each author**

Author 1 - Mihai-Cristian Tudose - Conceptualization, methodology, validation, investigation, resources, writing—original draft preparation.
Email: mihai.tudose1912@stud.acs.upb.ro
Author 2 - Ruseti Stefan - Validation, writing—review and editing.
Email: stefan.ruseti@upb.ro
Author 3 - Mihai Dascalu - Conceptualization, methodology, validation, investigation, resources, writing—review and editing, supervision.
Email: mihai.dascalu@upb.ro
Author 4 - Cornelia Caragea - Validation, writing—review and editing.
Email: cornelia@uic.edu