Novelty Assessment Report

Paper: Mapping Overlaps in Benchmarks through Perplexity in the Wild

PDF URL: https://openreview.net/pdf?id=QD0cuAmi9z

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-01

Abstract

We construct benchmark signatures that capture the capacity required for strong performance to characterize large language model (LLM) benchmarks and their meaningful overlaps. Formally, we define them as sets of salient tokens drawn from **in-the-wild** corpora whose LLM token perplexity, reflecting training exposure, is highly predictive of benchmark performance. We extract benchmark signatures via stepwise forward selection with linear regression in a large-scale meta-evaluation across 32 LLMs and 89 benchmarks spanning knowledge, coding, logic, instruction following, math, language, reasoning, missing-information detection, and cultural/world modeling. We then analyze how these signatures relate to both the semantic similarity of benchmark questions and the correlation structure of model performance. Performance-level overlaps remain universally high and semantic overlaps stay in a narrow mid-range, but signatures distinguish between benchmarks and illuminate nuanced differences in their capacity demands. For instance, signatures uniquely reveal substantial overlap among knowledge and reasoning benchmarks, whereas humanity- and culture-oriented benchmarks show relatively low similarity, lower even than typical cross-category overlap. Notably, performance-level results are strongly shaped by benchmark-**orthogonal** factors such as question format, whereas benchmark signatures remain robust to such confounds. We further reveal cross-functional overlaps among logic, math, language, instruction following, and cultural/world modeling, with coding emerging as the least overlapping domain, interacting only moderately with the ability of detecting missing information. Qualitative inspection of signatures shows that only the knowledge signature is aligned with actual knowledge, suggesting that LLMs may exhibit a distinctive semantic organization that differs from that of humans. Together, these findings offer insights into benchmark validity, LLM sensitivities, and the broad landscape of interconnected LLM capacities.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Characterizing Benchmark Overlap Through Token-Level Perplexity Patterns**

A total of **9 papers** were analyzed and organized into a taxonomy with **7 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Perplexity-Based Contamination Detection Methods**
- **Benchmark Overlap Characterization via Perplexity Signatures**
- **Benchmark Evaluation Frameworks and Decontamination**
- **Perplexity Applications in Specialized Domains**

Complete Taxonomy Tree

- Characterizing Benchmark Overlap Through Token-Level Perplexity Patterns Survey Taxonomy
- Perplexity-Based Contamination Detection Methods
  - Token-Level Perplexity Analysis for Contamination (3 papers)
  - [1] Investigating data contamination in modern benchmarks for large language models (Chunyuan Deng, 2024) View paper
  - [2] Generalization or memorization: Data contamination and trustworthy evaluation for large language models (Yihong Dong, 2024) View paper
  - [8] Estimating Contamination via Perplexity: Quantifying Memorisation in Language Model Evaluation (Li Yu-Cheng, 2023) View paper
  - Output Distribution and Perplexity Metrics (2 papers)
  - [6] Review of Grey Box/Black Box Data Contamination Metrics on Open and Commercial Models (Casey Casalnuovo, 2025) View paper
  - [9] Generalization or Memorization: Evaluating Data Contamination for Large Language Models (Yihong Dong, n.d.) View paper
- Benchmark Overlap Characterization via Perplexity Signatures
  - Meta-Evaluation of Benchmark Signatures ★ (1 papers)
  - [0] Mapping Overlaps in Benchmarks through Perplexity in the Wild (Anon et al., 2026) View paper
- Benchmark Evaluation Frameworks and Decontamination
  - Perplexity Normalization in Benchmark Design (1 papers)
  - [3] Paloma: A benchmark for evaluating language model fit (Iz Beltagy, 2024) View paper
- Perplexity Applications in Specialized Domains
  - Perplexity for Uncertainty Quantification (1 papers)
  - [7] Measuring large language model uncertainty in women's health using semantic entropy and perplexity: a comparative study (Jahan C. Penny-Dimri, 2025) View paper
  - Perplexity-Based Ranking in Task Automation (1 papers)
  - [4] IoT Rule Generation with Cross-View Contrastive Learning and Perplexity-Based Ranking (Gaetano Cimino, 2025) View paper
  - Prompt Memorization and Leakage Analysis (1 papers)

- [5] Why are my prompts leaked? unraveling prompt extraction threats in customized large language models (Liang Zi, 2024) View paper

Narrative

Core task: characterizing benchmark overlap through token-level perplexity patterns. The field addresses a critical challenge in evaluating large language models: determining whether training data has contaminated evaluation benchmarks, thereby inflating performance metrics. The taxonomy organizes work into several main branches. Perplexity-Based Contamination Detection Methods focus on algorithmic approaches that use perplexity signals to identify potential data leakage, with foundational work like Contamination via Perplexity[8] establishing early techniques. Benchmark Overlap Characterization via Perplexity Signatures examines how perplexity patterns themselves can serve as diagnostic fingerprints of overlap, including meta-evaluation efforts that assess the reliability of these signatures. Benchmark Evaluation Frameworks and Decontamination encompasses broader methodologies for creating clean test sets and validating benchmark integrity, exemplified by efforts like Paloma Benchmark[3]. Finally, Perplexity Applications in Specialized Domains explores how perplexity-based analysis extends beyond contamination detection to domain-specific evaluation challenges.

A particularly active line of investigation centers on developing robust contamination metrics and detection protocols. Works like Data Contamination Investigation[1] and Contamination Trustworthy Evaluation[2] explore the reliability and limitations of various detection strategies, while Contamination Metrics Review[6] synthesizes emerging best practices. The original paper, Perplexity Benchmark Overlaps[0], sits within the meta-evaluation cluster, focusing specifically on how perplexity signatures themselves can be characterized and validated as indicators of benchmark overlap. This positions it closely alongside work examining the trustworthiness of contamination signals, such as Contamination Trustworthy Evaluation[2], but with a distinctive emphasis on the diagnostic properties of token-level perplexity patterns rather than broader evaluation frameworks. The central tension across these branches involves balancing detection sensitivity against false positives, and understanding when perplexity anomalies genuinely indicate memorization versus other statistical artifacts.

Related Works in Same Category

No sibling papers and no sibling subtopics were found under the same parent taxonomy node; the paper appears structurally isolated in the taxonomy.

Contributions Analysis

Overall novelty summary. The paper introduces a framework for characterizing benchmark overlap using 'benchmark signatures'—sets of salient tokens from in-the-wild corpora whose perplexity patterns predict model performance. It sits in the 'Meta-Evaluation of Benchmark Signatures' leaf under 'Benchmark Overlap Characterization via Perplexity Signatures'. This leaf contains only the original paper itself, indicating a sparse research direction. The broader parent category ('Benchmark Overlap Characterization') is also minimally populated, suggesting this approach to using perplexity signatures for overlap analysis represents a relatively unexplored angle within the field.

The taxonomy reveals that most related work clusters in 'Perplexity-Based Contamination Detection Methods', which focuses on identifying training data leakage rather than characterizing benchmark capacity demands. The 'Token-Level Perplexity Analysis for Contamination' leaf contains three papers examining memorization detection, while 'Benchmark Evaluation Frameworks and Decontamination' addresses broader evaluation protocols. The original paper diverges by treating perplexity patterns as diagnostic fingerprints of benchmark overlap rather than contamination signals, positioning it at the intersection of contamination detection methodology and meta-evaluation of benchmark properties. This creates conceptual distance from sibling branches despite shared technical foundations.

Among 20 candidates examined across three contributions, no clearly refuting prior work was identified. The 'Benchmark signatures framework' contribution examined 8 candidates with 0 refutable matches, suggesting limited direct precedent for this specific formulation. The 'Forward selection and regression pipeline' examined only 2 candidates, reflecting a narrower technical scope. The 'Discovery of unexpected cross-functional overlaps' examined 10 candidates with no refutations, indicating that the empirical findings about knowledge-reasoning overlap and culture-oriented benchmark distinctiveness may represent novel observations within this limited search scope. The absence of refutations across all contributions suggests the approach occupies a relatively uncontested niche.

Based on the limited search of 20 candidates, the work appears to introduce a distinctive methodological angle—using perplexity signatures for capacity characterization rather than contamination detection—in a sparsely populated research direction. The taxonomy structure confirms minimal direct competition in this specific framing, though the broader contamination detection literature provides relevant technical context. The analysis cannot rule out related work outside the top-20 semantic matches or in adjacent evaluation methodology domains not captured by the taxonomy.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Benchmark signatures framework for measuring benchmark overlap

Description: The authors propose a novel three-level framework for quantifying overlap among LLM benchmarks. This framework examines benchmarks at the semantic level (question content similarity), performance level (correlated model outcomes), and signature level (perplexity patterns on in-the-wild corpora), providing a more comprehensive characterization than existing approaches.

This contribution was assessed against **8 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. CODEMORPH: Mitigating Data Leakage in Large Language Model Assessment

URL: View paper

Brief Assessment

CodeMorph[24] addresses data leakage in code benchmarks through code perturbation techniques, not benchmark overlap measurement. The original paper's framework uses perplexity patterns from in-the-wild corpora to characterize benchmark relationships, while CodeMorph[24] focuses on generating diverse code variations to prevent contamination.

2. Measuring self-deceptive consistency boundaries in large language models through spurious semantic closure networks

URL: View paper

Brief Assessment

Self-Deceptive Consistency[20] focuses on spurious semantic closure networks in LLMs and measuring self-deceptive consistency boundaries. The candidate does not address benchmark overlap measurement, semantic/performance/signature-level analysis frameworks, or perplexity-based characterization of benchmark relationships.

3. Understanding RAG Systems Performance by Profiling Key Factors

URL: View paper

Brief Assessment

RAG Performance Profiling[22] focuses on analyzing retrieval-augmented generation systems through system factors like retrieval recall and document types, not on measuring benchmark overlap through perplexity-based signatures across LLM evaluations.

4. Benchmarking Benchmark Leakage in Large Language Models

URL: View paper

Brief Assessment

Benchmark Leakage[23] focuses on detecting whether benchmark data was used during model training (data leakage detection), not on measuring overlap between different benchmarks to characterize their relationships and capacity demands.

5. Investigating data contamination in modern benchmarks for large language models

URL: View paper

Brief Assessment

Data Contamination Investigation[1] focuses on detecting data contamination through retrieval-based methods and testset slot guessing, rather than proposing a comprehensive framework for measuring benchmark overlap using semantic, performance, and perplexity-based signatures.

6. Studying the Role of Input-Neighbor Overlap in Retrieval-Augmented Language Models Training Efficiency

URL: View paper

Brief Assessment

Input-Neighbor Overlap[21] focuses on retrieval-augmented language models and the role of query-context overlap in training efficiency, not on benchmark evaluation or measuring overlap among LLM benchmarks. The papers address fundamentally different research questions.

7. Estimating Contamination via Perplexity: Quantifying Memorisation in Language Model Evaluation

URL: View paper

Brief Assessment

Contamination via Perplexity[8] focuses on detecting data contamination in evaluation benchmarks by comparing test set perplexity to memorized/clean baselines, not on measuring overlap between different benchmarks or characterizing benchmark relationships across semantic, performance, and signature levels.

8. StyleCloak: Anonymous Source Coding for Textual Attribute Obfuscation with Semantic Fidelity

URL: View paper

Brief Assessment

StyleCloak[25] focuses on textual attribute obfuscation for privacy preservation in communications, not on measuring benchmark overlap or evaluating LLM benchmarks. The paper addresses anonymous source coding design rather than benchmark evaluation methodologies.

Contribution 2: Forward selection and regression pipeline for extracting benchmark signatures

Description: The authors develop a computational method that combines correlation-based screening with forward selection regression to identify salient tokens from large-scale corpora. These tokens form benchmark signatures whose perplexity patterns across models are highly predictive of benchmark performance, enabling systematic characterization of what capacities each benchmark actually measures.

This contribution was assessed against **2 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Natural Language Processing Tools for Reading Level Assessment and Text Simplication for Bilingual Education

URL: View paper

Brief Assessment

Reading Level Assessment[26] focuses on reading level assessment and text simplification for bilingual education, not on extracting benchmark signatures from language model perplexity patterns. The candidate uses forward selection for feature selection in reading level classification, which is a different application domain than characterizing LLM benchmark overlaps through token-level perplexity statistics.

2. Predicting how it sounds: re-ranking dialogue prompts based on TTS quality for adaptive spoken dialogue systems.

URL: View paper

Brief Assessment

TTS Dialogue Prompts[27] applies stepwise regression to select text-to-speech quality features for dialogue prompts, not to extract token-level perplexity statistics from large-scale corpora for benchmark characterization.

Contribution 3: Discovery of unexpected cross-functional benchmark overlaps

Description: Through their signature-based analysis, the authors reveal that many benchmarks claiming to test specific abilities (like logic) actually measure different or overlapping capabilities (like instruction-following) in practice. This finding exposes potential misalignments between benchmark design intentions and what they actually evaluate, highlighting issues in current benchmark validity and the interconnected nature of LLM capabilities.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Craftext benchmark: Advancing instruction following in complex multimodal open-ended world

URL: View paper

Brief Assessment

Craftext Benchmark[16] focuses on instruction-following in dynamic multimodal environments with diverse linguistic constructions, not on analyzing benchmark overlaps or revealing misalignments between intended and actual benchmark measurements across different capability domains.

2. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization

URL: View paper

Brief Assessment

Pandalm[11] focuses on evaluating instruction-tuned LLMs through subjective metrics (conciseness, clarity, adherence to instructions) rather than analyzing benchmark design validity or cross-functional capability overlaps. It does not examine whether benchmarks measure their intended abilities.

3. Ivebench: Modern benchmark suite for instruction-guided video editing assessment

URL: View paper

Brief Assessment

Ivebench[18] focuses on video editing evaluation with instruction-guided prompts, not on analyzing benchmark overlaps or measuring whether benchmarks test their intended abilities versus other capabilities like instruction-following.

4. Evaluating large language models at evaluating instruction following

URL: View paper

Brief Assessment

Evaluating Instruction Following[14] focuses on meta-evaluation of LLM evaluators for instruction-following tasks, not on analyzing benchmark overlaps or revealing that benchmarks measure different abilities than intended.

5. Hq-edit: A high-quality dataset for instruction-based image editing

URL: View paper

Brief Assessment

HQ-Edit[15] focuses on instruction-based image editing datasets and quality metrics for image-text alignment, not on LLM benchmark analysis or cross-functional capability measurement. The domains are entirely different.

6. IFIR: A comprehensive benchmark for evaluating instruction-following in expert-domain information retrieval

URL: View paper

Brief Assessment

IFIR Benchmark[13] focuses on instruction-following capabilities in expert-domain information retrieval (finance, law, healthcare, scientific literature), not on analyzing cross-functional overlaps between benchmarks measuring different abilities like logic versus instruction-following.

7. Magicbrush: A manually annotated dataset for instruction-guided image editing

URL: View paper

Brief Assessment

Magicbrush[12] focuses on instruction-guided image editing datasets and does not address benchmark evaluation, LLM capabilities, or cross-functional overlaps in language model benchmarks. The domains are entirely distinct.

8. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use

URL: View paper

Brief Assessment

Visit-Bench[17] focuses on evaluating vision-language instruction-following models for real-world use cases, not on analyzing benchmark overlaps or revealing misalignments between benchmark design intentions and actual measurements.

9. When thinking fails: The pitfalls of reasoning for instruction-following in llms

URL: View paper

Brief Assessment

Reasoning Pitfalls[19] focuses on how chain-of-thought reasoning degrades instruction-following performance in LLMs, not on benchmark overlap analysis or misalignment between benchmark design intentions and actual capabilities measured.

10. Infobench: Evaluating instruction following ability in large language models

URL: View paper

Brief Assessment

Infobench[10] focuses on evaluating instruction-following ability through decomposed requirements, not on analyzing cross-functional benchmark overlaps or revealing misalignments between benchmark design intentions and actual measurements.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Mapping Overlaps in Benchmarks through Perplexity in the Wild View paper
- [1] Investigating data contamination in modern benchmarks for large language models View paper
- [2] Generalization or memorization: Data contamination and trustworthy evaluation for large language models View paper
- [3] Paloma: A benchmark for evaluating language model fit View paper
- [4] IoT Rule Generation with Cross-View Contrastive Learning and Perplexity-Based Ranking View paper
- [5] Why are my prompts leaked? unraveling prompt extraction threats in customized large language models View paper
- [6] Review of Grey Box/Black Box Data Contamination Metrics on Open and Commercial Models View paper
- [7] Measuring large language model uncertainty in women's health using semantic entropy and perplexity: a comparative study View paper
- [8] Estimating Contamination via Perplexity: Quantifying Memorisation in Language Model Evaluation View paper
- [9] Generalization or Memorization: Evaluating Data Contamination for Large Language Models View paper
- [10] Infobench: Evaluating instruction following ability in large language models View paper
- [11] Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization View paper
- [12] Magicbrush: A manually annotated dataset for instruction-guided image editing View paper
- [13] IFIR: A comprehensive benchmark for evaluating instruction-following in expert-domain information retrieval View paper
- [14] Evaluating large language models at evaluating instruction following View paper
- [15] Hq-edit: A high-quality dataset for instruction-based image editing View paper

- [16] Craftext benchmark: Advancing instruction following in complex multimodal open-ended world View paper
- [17] Visit-bench: A benchmark for vision-language instruction following inspired by real-world use View paper
- [18] Ivebench: Modern benchmark suite for instruction-guided video editing assessment View paper
- [19] When thinking fails: The pitfalls of reasoning for instruction-following in llms View paper
- [20] Measuring self-deceptive consistency boundaries in large language models through spurious semantic closure networks View paper
- [21] Studying the Role of Input-Neighbor Overlap in Retrieval-Augmented Language Models Training Efficiency View paper
- [22] Understanding RAG Systems Performance by Profiling Key Factors View paper
- [23] Benchmarking Benchmark Leakage in Large Language Models View paper
- [24] CODEMORPH: Mitigating Data Leakage in Large Language Model Assessment View paper
- [25] StyleCloak: Anonymous Source Coding for Textual Attribute Obfuscation with Semantic Fidelity View paper
- [26] Natural Language Processing Tools for Reading Level Assessment and Text Simplication for Bilingual Education View paper
- [27] Predicting how it sounds: re-ranking dialogue prompts based on TTS quality for adaptive spoken dialogue systems. View paper