# Can evaluation metrics designed for tabular data generation predict the alignment performance of LLMs fine-tun

Assignee Research

June 10, 2026

**Abstract**

Language models such as GPT and Llama have shown remarkable ability on diverse natural language tasks, yet their performance on complex table tasks (e.g., NL-to-Code and data cleaning) remains sub-optimal. Improving performance typically requires task-specific fine-tuning, which depends on expensive human labeling and is prone to overfitting. In this work, we propose Table-LLM-Specialist, a self-trained fine-tuning paradigm designed for table tasks. Our key insight is that many table tasks admit two dual formulations: a generative version and a classification version. Leveraging this duality,

## 1 Introduction

This paper examines: Table-LLM-Specialist: Language Model Specialists for Tables using Iterative Generator-Validator Fine-tuning. Research question: Can evaluation metrics designed for tabular data generation predict the alignment performance of LLMs fine-tuned on synthetic instruction datasets derived from benchmark tables?.

## 2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

## 3 Results

15 papers retrieved. 11 claims extracted; 0 independently verified. Quality review score: 3.7/10.

## 4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

## 5 Extracted Claims

| Claim | Verified | Confidence |
|---|---|---|
| Fine-tuning GPT-3.5 on the HXD dataset improves F1 score on the HXD test split. | × | 0.04 |
| Fine-tuning GPT-3.5 on the HXD dataset decreases F1 score on the Wikidata test split compared to vanilla GPT-3.5. | × | 0.05 |
| Fine-tuning GPT-3.5 on the WikiSQL dataset improves execution accuracy on the WikiSQL test split. | × | 0.04 |
| Fine-tuning GPT-3.5 on the WikiSQL dataset decreases execution accuracy on the Text2Analysis test split compared to vani | × | 0.05 |
| GPT-3.5 fine-tuned on HXD achieves an F1 score of approximately 0.95 on the HXD test split. | × | 0.05 |
| GPT-3.5 fine-tuned on HXD achieves an F1 score of approximately 0.85 on the Wikidata test split. | × | 0.05 |
| GPT-3.5 fine-tuned on WikiSQL achieves an execution accuracy of approximately 0.80 on the WikiSQL test split. | × | 0.05 |
| GPT-3.5 fine-tuned on WikiSQL achieves an execution accuracy of approximately 0.40 on the Text2Analysis test split. | × | 0.05 |
| Table-Specialist fine-tuning achieves an average F1 score of 0.938 for Schema Matching tasks across DeepM, WikiData, and | × | 0.10 |
| Table-Specialist fine-tuning achieves an average execution accuracy of 0.616 for NL-to-SQL tasks across WikiSQL, WikiTQ, | × | 0.11 |
| Table-Generalist fine-tuning achieves an average execution accuracy of 0.576 for NL-to-SQL tasks across WikiSQL, WikiTQ, | × | 0.06 |

## References


- http://arxiv.org/abs/2410.12164v2
- http://arxiv.org/abs/2304.12244v3
- http://arxiv.org/abs/2402.11651v2