EuroVis Workshop on Visual Analytics (2025)
H.-J. Schulz and A. Vilanova (Editors)

Supplemental Materials for A Design Space for the Critical Validation of LLM-Generated Tabular Data

Madhav Sachdeva[1], Christopher Narayanan[1], Marvin Wiedenkeller[1], Jana Sedlakova[1,2], and Jürgen Bernard[1,2]

[1]University of Zurich, Switzerland; [2]Digital Society Initiative, Zürich, Switzerland

1. Overview of characterized papers

Paper title	Domain	Type of contribution	Target audience	Granularity	Publisher	Validation	Visualization
Cycles of Thought: Measuring LLM Confidence Through Stable Explanations [BS24]	Machine learning	Algorithm, Evaluation, Theory	Experts	Items	Preprint	Distribution	No
Language Models are Realistic Tabular Data Generators [BSL*22]	Machine learning	Algorithm, Evaluation	Experts	Items, attributes, multiple attributes	ICLR	Align, distribution	No
KnowledgeVis: Interpreting Language Models by Comparing Fill-in-the-Blank Prompts [CE23]	Visualization	System, Methodology, Evaluation	Experts	Items	TVCG	Human Knowledge Externalization, Correlation Analysis	Yes
Zeno: An Interactive Framework for Behavioral Evaluation of Machine Learning [CFB*23]	HCI, Machine learning	System	Experts	Items	CHI	Distribution, counterfactual analysis	Yes
iScore: Visual Analytics for Interpreting How Language Models Automatically Score Summaries [CHM*24]	Visualization	System, methodology, evaluation	Experts	Items, Attribute	IUI	Human Knowledge Externalization, Feature analysis	Yes
Interactive Analysis of LLMs Using Meaningful Counterfactuals [CZC*24]	Machine learning	System, methodology	Experts, non-experts	Items	Preprint TVCG	Counterfactual analysis	Yes
Finding Support for Tabular LLM Outputs [FSM24]	-	Methodology, evaluation	Experts	Multiple attributes	VLDB	Correlation analysis	No
A Rigorous Benchmark of Structured Outputs for Language Models [GCM*25]	NLP	Benchmark, evaluation	Experts	Items	Preprint	Constraint checks	No
TabLLM: Few-Shot Classification of Tabular Data with Large Language Models [HBL*23]	Machine learning	Algorithm, evaluation	Experts	Items	AISTATS	Accuracy analysis	No
EvalLM: Interactive Evaluation of Large Language Model Prompts on User-Defined Criteria [KLS*24]	NLP	System, evaluation	Experts	Items, attribute	CHI	Human Knowledge Externalization, Comparative Evaluation	Yes
PromptAid: Visual Prompt Exploration, Perturbation, Testing, and Iteration for Large Language Models [MDS*25]	Visualization	System, methodology, evaluation	Non-experts	Items	TVCG	Human Knowledge Externalization, What-if analysis	Yes
Human-Centered Design Recommendations for LLM-as-a-Judge [PAD*24]	HCI	Methodology, evaluation	Experts	Items	Preprint	Human Knowledge Externalization	No
Large Language Models with Human-in-the-Loop Validation for Systematic Review Data Extraction [SIZ25]	Machine learning	Methodology	Non-experts	Items	Preprint	Human-in-the-loop	No
SPADE: Synthesizing Data Quality Assertions for Large Language Model Pipelines [SLA*24]	Software engineering	System, algorithm	Experts	Items	Preprint	Constraint checks	No
Interactive and Visual Prompt Engineering for Ad-Hoc Task Adaptation with Large Language Models [SWS*22]	HCI, NLP	System, methodology	Non-experts	Items	TVCG	Human Knowledge Externalization	Yes
The Language Interpretability Tool: Extensible, Interactive Visualizations and Analysis for NLP Models [TWB*20]	NLP	System, methodology	Experts	Items, attribute, multiple Attributes	Preprint	Distribution, feature analysis	Yes
Beyond Yes and No: Improving Zero-shot LLM Rankers via Scoring Fine-grained Relevance Labels [ZQH*23]	Information retrieval	Algorithm, evaluation	Experts	Items	Preprint	Distribution	No
Can ChatGPT Reproduce Human-Generated Labels? A Study of Social Computing Tasks [ZZH*23]	Social computing	Methodology, evaluation	Experts	Items	Preprint	Human Knowledge Externalization	No

References

[BS24] BECKER E., SOATTO S.: Cycles of thought: Measuring llm confidence through stable explanations. *arXiv preprint arXiv:2406.03441* (2024). 2

[BSL*22] BORISOV V., SESSLER K., LEEMANN T., PAWELCZYK M., KASNECI G.: Language models are realistic tabular data generators. *arXiv preprint arXiv:2210.06280* (2022). 2

[CE23] COSCIA A., ENDERT A.: Knowledgevis: Interpreting language models by comparing fill-in-the-blank prompts. *IEEE Transactions on Visualization and Computer Graphics 30*, 9 (2023), 6520–6532. 2

[CFB*23] CABRERA Á. A., FU E., BERTUCCI D., HOLSTEIN K., TALWALKAR A., HONG J. I., PERER A.: Zeno: An interactive framework for behavioral evaluation of machine learning. In *CHI Conference on Human Factors in Computing Systems* (2023), pp. 1–14. 2

[CHM*24] COSCIA A., HOLMES L., MORRIS W., CHOI J. S., CROSSLEY S., ENDERT A.: iscore: Visual analytics for interpreting how language models automatically score summaries. In *International Conference on Intelligent User Interfaces* (2024), pp. 787–802. 2

[CZC*24] CHENG F., ZOUHAR V., CHAN R. S. M., FÜRST D., STROBELT H., EL-ASSADY M.: Interactive analysis of llms using meaningful counterfactuals. *arXiv preprint arXiv:2405.00708* (2024). 2

[FSM24] FAN G., SHRAGA R., MILLER R. J.: Finding support for tabular llm outputs. *VLDB Endowment. ISSN 2150* (2024), 8097. 2

[GCM*25] GENG S., COOPER H., MOSKAL M., JENKINS S., BERMAN J., RANCHIN N., WEST R., HORVITZ E., NORI H.: Generating structured outputs from language models: Benchmark and studies. *arXiv preprint arXiv:2501.10868* (2025). 2

[HBL*23] HEGSELMANN S., BUENDIA A., LANG H., AGRAWAL M., JIANG X., SONTAG D.: Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics* (2023), PMLR, pp. 5549–5581. 2

[KLS*24] KIM T. S., LEE Y., SHIN J., KIM Y.-H., KIM J.: Evallm: Interactive evaluation of large language model prompts on user-defined criteria. In *CHI Conference on Human Factors in Computing Systems* (2024), pp. 1–21. 2

[MDS*25] MISHRA A., DANZY B., SONI U., ARUNKUMAR A., HUANG J., KWON B. C., BRYAN C.: Promptaid: Visual prompt exploration, perturbation, testing and iteration for large language models. *IEEE Transactions on Visualization and Computer Graphics* (2025). 2

[PAD*24] PAN Q., ASHKTORAB Z., DESMOND M., COOPER M. S., JOHNSON J., NAIR R., DALY E., GEYER W.: Human-centered design recommendations for llm-as-a-judge. *arXiv preprint arXiv:2407.03479* (2024). 2

[SJZ25] SCHROEDER N. L., JALDI C. D., ZHANG S.: Large language models with human-in-the-loop validation for systematic review data extraction. *arXiv preprint arXiv:2501.11840* (2025). 2

[SLA*24] SHANKAR S., LI H., ASAWA P., HULSEBOS M., LIN Y., ZAMFIRESCU-PEREIRA J., CHASE H., FU-HINTHORN W., PARAMESWARAN A. G., WU E.: Spade: Synthesizing data quality assertions for large language model pipelines. *arXiv preprint arXiv:2401.03038* (2024). 2

[SWS*22] STROBELT H., WEBSON A., SANH V., HOOVER B., BEYER J., PFISTER H., RUSH A. M.: Interactive and visual prompt engineering for ad-hoc task adaptation with large language models. *IEEE transactions on visualization and computer graphics 29*, 1 (2022), 1146–1156. 2

[TWB*20] TENNEY I., WEXLER J., BASTINGS J., BOLUKBASI T., COENEN A., GEHRMANN S., JIANG E., PUSHKARNA M., RADEBAUGH C., REIF E., ET AL.: The language interpretability tool: Extensible, interactive visualizations and analysis for nlp models. *arXiv preprint arXiv:2008.05122* (2020). 2

[ZQH*23] ZHUANG H., QIN Z., HUI K., WU J., YAN L., WANG X., BENDERSKY M.: Beyond yes and no: Improving zero-shot llm rankers via scoring fine-grained relevance labels. *arXiv preprint arXiv:2310.14122* (2023). 2

[ZZH*23] ZHU Y., ZHANG P., HAQ E.-U., HUI P., TYSON G.: Can chatgpt reproduce human-generated labels? a study of social computing tasks. *arXiv preprint arXiv:2304.10145* (2023). 2