EACL 2026
RABAT • MOROCCO
Mars • March 24-29, 2026 • مارس

Quantifying the Impact of Structured Output Format on Large Language Models through Causal Inference

Han Yuan Yue Zhao Li Zhang Wuqiong Luo Zheng Ma
Singapore Decision Science Center of Excellence, Global Decision Science, American Express
{Han.Yuan1, Yue.Zhao, Li.Zhang1, Wuqiong.Luo, Zheng.Ma2}@aexp.com

Motivation

- How effectively do LLMs adhere to structured output formats?
- Does output format influence content quality? Prior studies conclude in a one-sided manner: structured output either improves or reduces quality.
- How to statistically quality the impact? Former explorations apply similar strategy to compare the final aggregate metrics' differences between structured and unstructured output, which is relatively rudimentary.

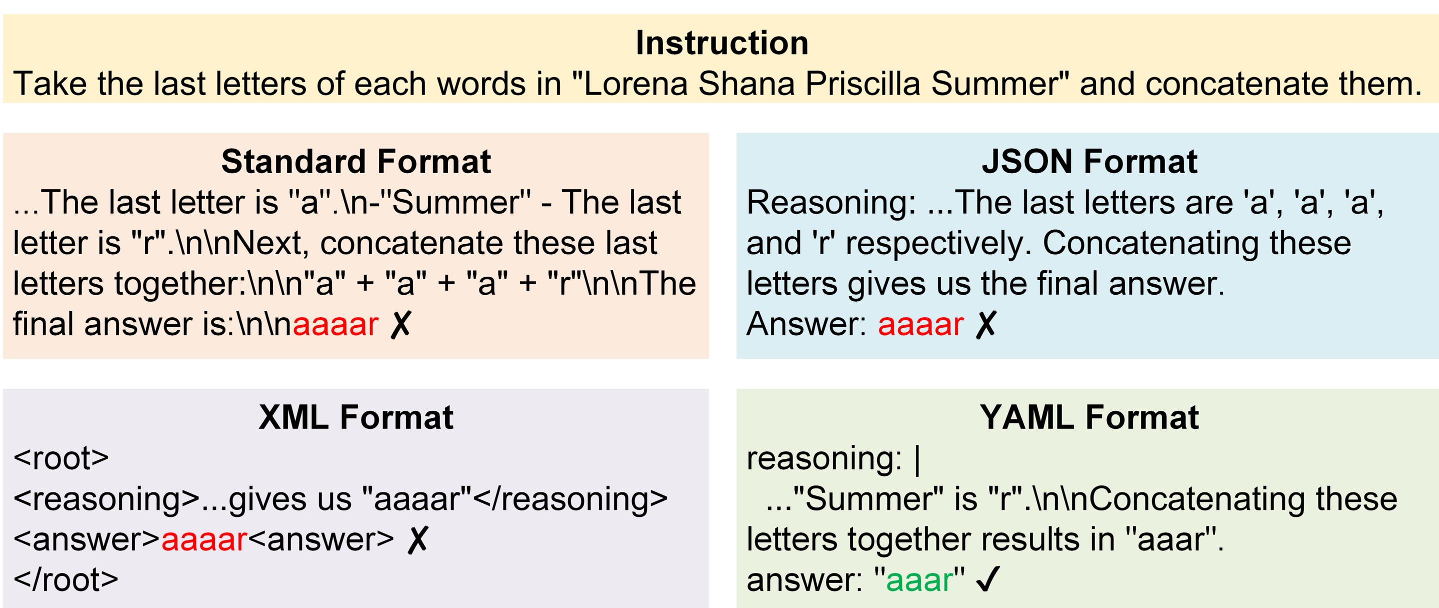

Figure 1: Positive, neutral, and negative effects of structured outputs on LLMs' generation.

Causal inference

- We reduce a large set of directed acyclic graphs (DAGs) to a limited number of candidate structures, enabled by controlled or guaranteed constraints.

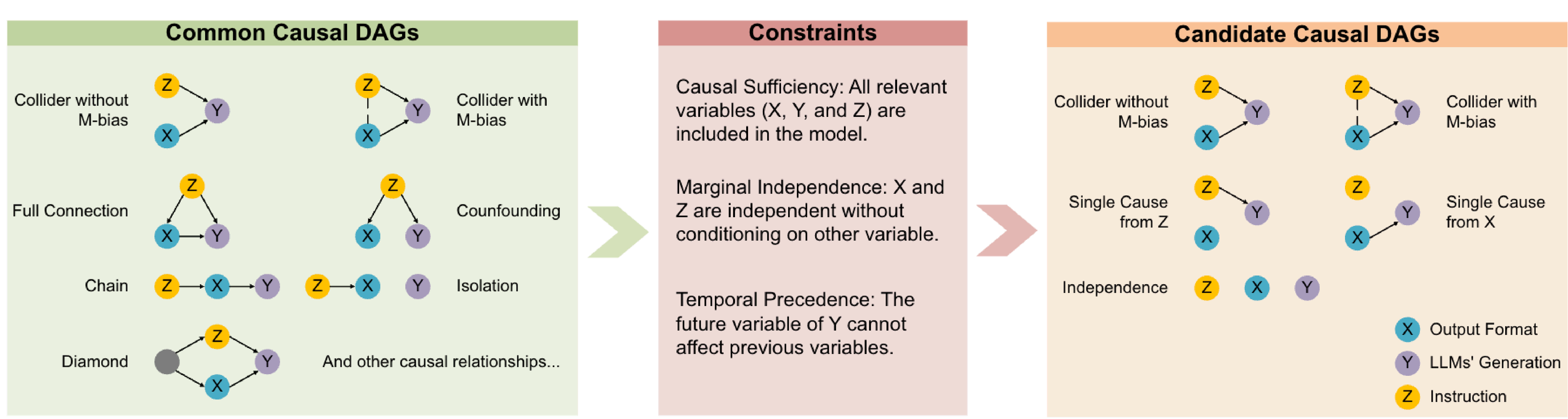

Figure 2: Candidate causal relations isolated from common types.

- We use causal analysis to quantify structured outputs' impact.

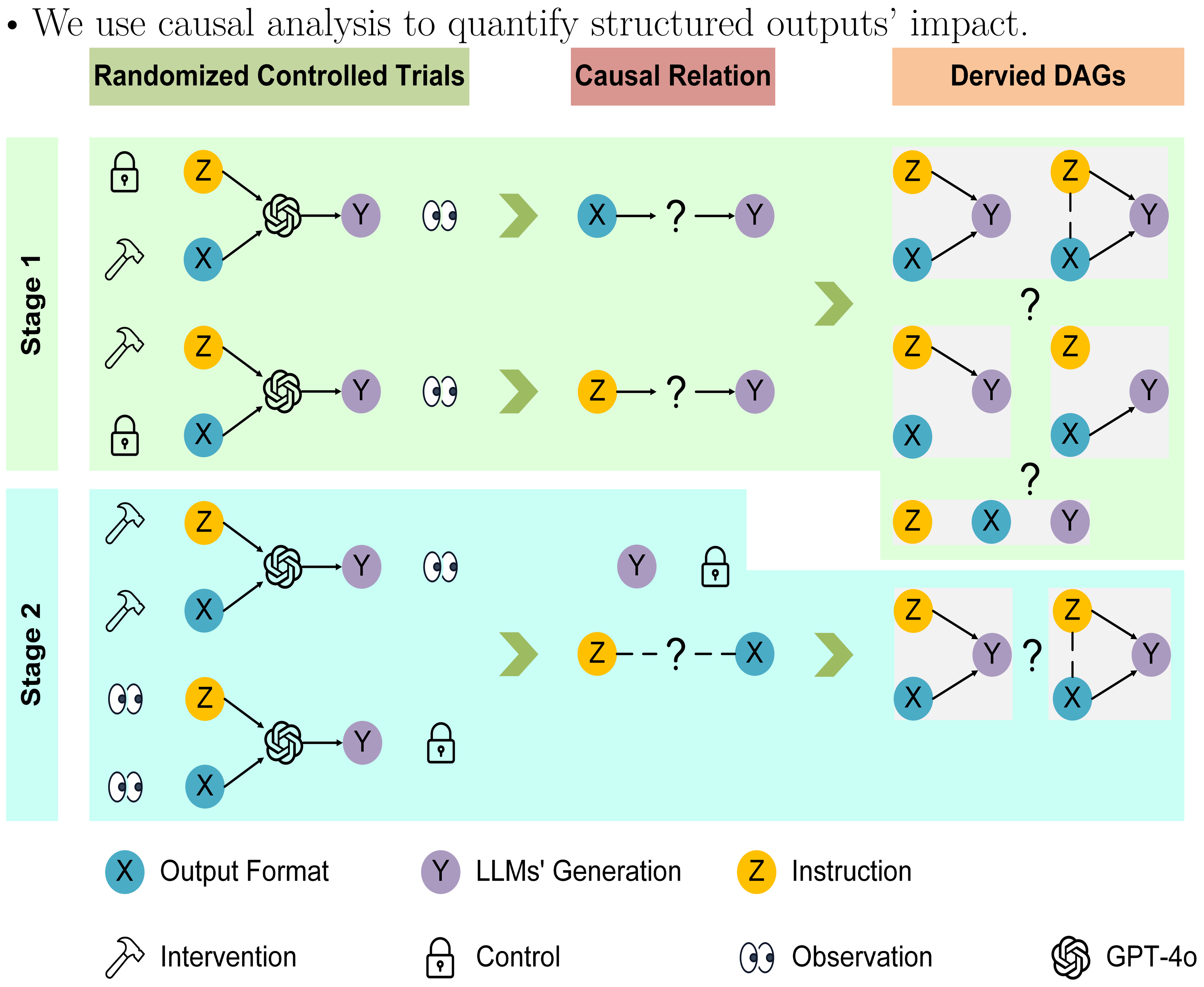

Figure 3: Randomized controlled trials for identifying causal structures.

Results

- GPT-4o demonstrates robust insensitivity to structured output formats. In the exceptions, both instruction and format causally influence its quality.
- GPT-4.1 follows trends similar to GPT-4o, performing robustly in most settings but still exhibiting vulnerability on the SOT task.
- OpenAI-o3 is consistently resilient to structured outputs, showing an underappreciated strength of reasoning models.
- Function calling is a promising approach for obtaining structured outputs: Independence remains the predominant DAGs, while the generation quality outperform that from format-restricting instructions.
- Derived DAGs are robust under additional interventions and API calls

LLM	GPT-4o			GPT-4.1			OpenAI-o3		
Task	SOT	OpsEval	XCodeEval	SOT	OpsEval	XCodeEval	SOT	OpsEval	XCodeEval
JSON format	CwoM	IND	IND	CwoM	IND	IND	IND	IND	IND
XML format	IND	IND	IND	INS	IND	IND	IND	IND	IND
YAML format	IND	IND	IND	CwoM	IND	IND	IND	IND	IND

Table 1: Discovery of DAGs based on structured output by format-restricting instruction. CwoM: collider without m-bias; INS: single cause from instruction; IND: independence.

- SLMs are more sensitive to output formats while GPT-oss-20B demonstrates great robustness to output format interventions.

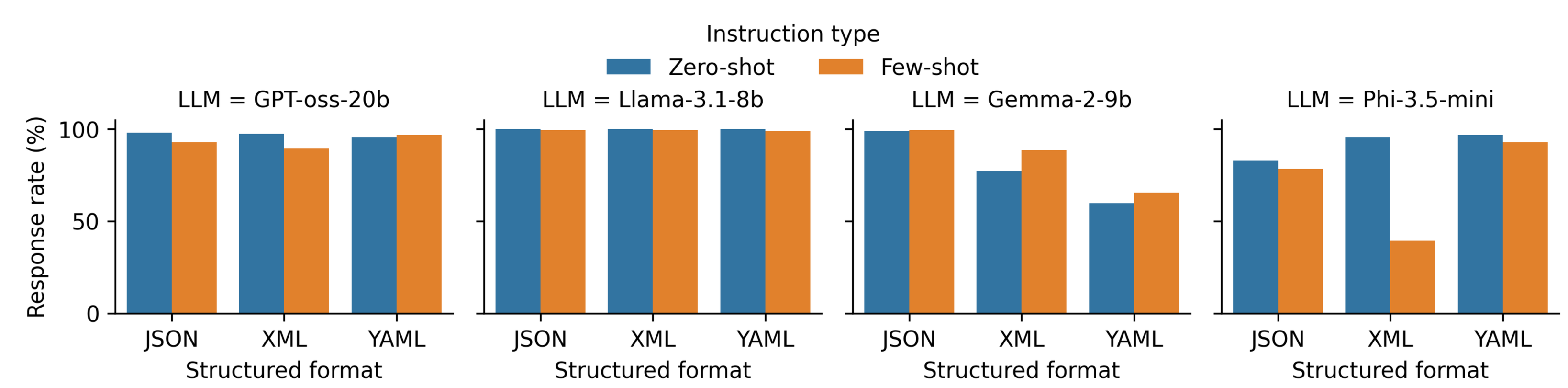

Figure 4: Successful response rate of different SLMs under zero-shot and few-shot Prompts.

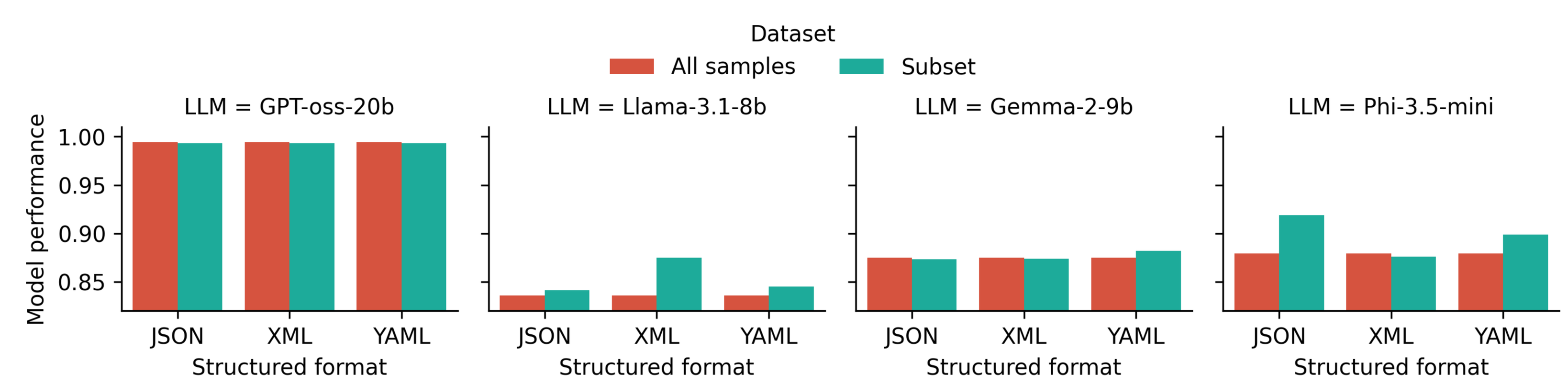

Figure 5: Performance comparison on entire dataset and subsets of successful responses.

- Extended Last Letter Concatenation (ELLC) dataset introduces additional complexity in both symbolic and linguistic reasoning.

Extraction accuracy	Question type	Sample type	Rearrangement accuracy
0.334	Single	Overall	0.286
		Single	0.286
		Multiple	0.304
	Multiple	Overall	0.185
		Single	0.189
		Multiple	0.082

Table 2: GPT-4o performance on ELLC tasks with 6 letters and middle position.

Disclaimer

Reference

- Quantifying the Impact of Structured Output Format on Large Language Models through Causal Inference. EACL. 2026.
- Let Me Speak Freely? A Study On The Impact Of Format Restrictions On Large Language Model Performance. EMNLP. 2024.
- How Likely Do LLMs with CoT Mimic Human Reasoning? COLING. 2025.