

Stat-to-Text

From Tables to Quantified Statements: Evaluating LLM Inference Generation through Executable Verification

Mai Mohamed Eida, Gunjan Anand, Ayush Singh, Aleksandre Maskharashvili

University of Illinois at Urbana-Champaign

{maimm2, gunjana2, ayushs13, am172}@illinois.edu



Research Questions

- 1 Does Stat-to-Text allows to **reliably generate and verify quantified natural language inferences** from statistical tables?
- 2 Does **model family, model scale, and reasoning paradigm** affect the logical **faithfulness** of quantified Stat-to-Text generation?
- 3 How do LLMs differ in the **coverage and diversity** of the **quantified inferences** they generate?

Pipeline Example

Statistical Table

Patient	Diagnosis	Age	Systolic BP	Diastolic BP	Chol.	BMI	Smoker
P1	asthma	46	129	72	192	33.3	no
P2	diabetes	62	143	82	223	22.5	yes
P3	diabetes	68	147	87	198	31.0	no
P4	arthritis	46	135	77	182	27.6	yes
P5	hypertension	76	119	74	245	22.7	no

Quantified Inference

“All individuals have systolic blood pressure between 112 and 156 mmHg.”

LLM Python Checker Code

```
def stmt_1(df: pd.DataFrame):
    df["bp_systolic"] = pd.to_numeric(
        df["bp_systolic"], errors="coerce")
    mask = df["bp_systolic"].between(112, 156)
    truth = bool(mask.all())

    if truth:
        return True, "No violations found."
    return False, (
        f"Found {len(df[~mask])} violating rows.")
```

LLM Python Checker Verification

Decision: True

Explanation: “All 5 individuals have systolic blood pressure between 112 and 156 mmHg.”

Automatic Evaluation

- **Faithfulness:** grounding and logical accuracy
- **Coverage:** table records, variables, and quantifier categories
- **Diversity:** variable and quantifier redundancy

Model	Tables	Faithfulness		Table Coverage			Quantifier Coverage	
		Grounding↑	Accuracy↑	TRC↑	VC↑	VR↓	QC↑	QR↓
LLAMA-8B	86	1	0.487	0.95	0.878	0.149	0.9	0.805
LLAMA-70B	94	0.999	0.713	0.994	0.948	0.168	0.991	0.948
GPT-OSS-20B	78	0.96	0.873	0.929	0.909	0.132	0.754	0.89
GPT-OSS-120B	95	0.995	0.938	1	0.956	0.159	0.956	0.822

Human Evaluation

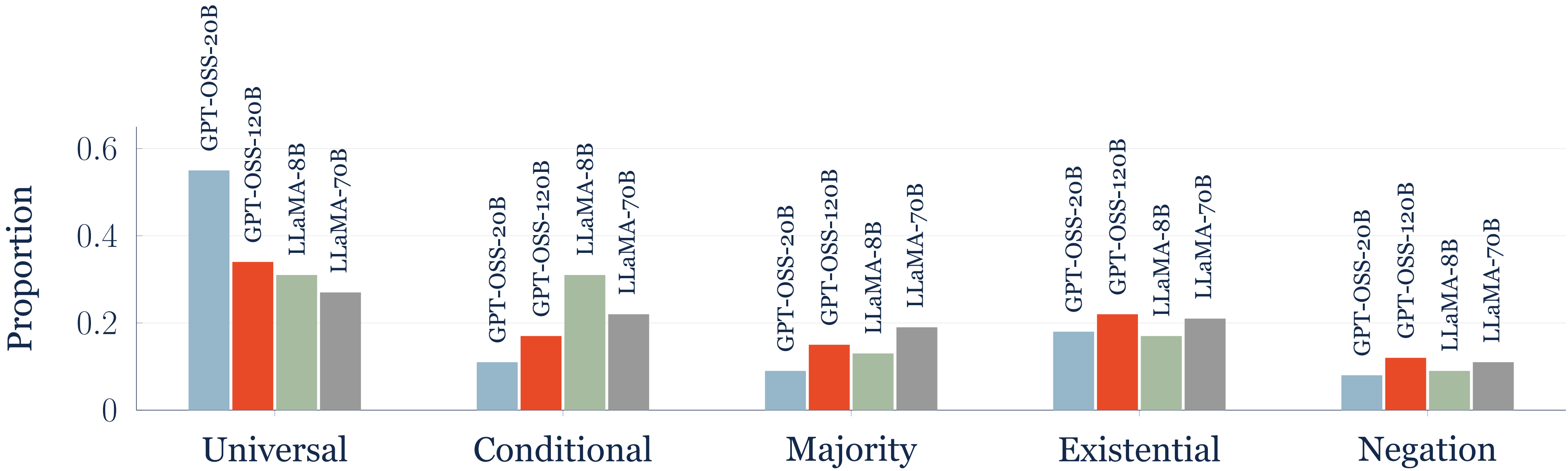
Model	Py. Checker (%)	Human Annot. (%)	Agreement (%)
LLAMA-8B	46.67	56.33	79.67
LLAMA-70B	68.33	75.33	91.67
GPT-OSS-20B	89.00	92.33	96.00
GPT-OSS-120B	94.33	96.50	97.50
Overall	74.58	80.13	91.21

30 randomly sampled tables × 4 models × 20 inferences each.

Discussion & Summary

- Stat-to-Text can reliably **generate and verify quantified inferences**, though performance varies by model.
- GPT-OSS-120B performs best, generating **faithful inferences** without **sacrificing coverage or diversity**.
- Models do not struggle with coverage; however, **diversity of quantifiers** seems to come at a **cost for faithfulness**, especially with smaller models.
- Models each **generate** their own **verification code** so reported errors may also reflect **verifier quality** and not just the quality of generated inferences.
- Future work will extend to real-world, larger, more complex tables and multi-sentence discourse relations.

Quantifier Diversity



Neither model family nor model size are particularly predictive of quantifier diversity. Models with **higher faithfulness** tend to exhibit **lower quantifier diversity** and vice versa—except GPT-OSS-120B.