Enhancing Knowledge Graph Consistency Through Open Large Language Models: A Case Study

Ankur Padia, Francis Ferraro and Tim Finin

University of Maryland, Baltimore County

pankur1@umbc.edu

Textual Inconsistency in Knowledge Graph

- Information Extraction (IE) system convert text into a knowledge graph and associate provenance sentences as evidence.
- Information Extraction System are not perfect and makes mistake
- One of the error type is **textual inconsistency** which we refer to as Knowledge Graph Consistency

Extracted Fact: ***Mauritania***; org:alternate names; ***CPPCC***
Provenance Text: China thanked **Mauritania** for supporting China on Taiwan, Tibet, Xinjiang, human rights and other issues concerning the country's core interests, Yu said. Yu said the **CPPCC** would like to work with the ESC of Mauritania to carry out exchanges and promote bilateral relations.

Table 1: An example of extracted fact with provenance text

Research Questions

Q1: **Modeling**: How can a Large Language Model help identify inconsistencies in a knowledge graph?

Q2: **Fine-tuning**: Do generic models outperform fine-tuned open models?

Q3: **Size**: Does the size of the language model matter?

Q4: **Domain**: Do language models perform well across different types of relations?

Q5: **Entities**: How does the number of entities affect language models?

Q6: **Number of examples:** How does the number of training examples affect performance?

Convert Knowledge Graph Extraction as a Multi-choice Question Answer Prompt

Extracted Fact:

Mauritania; org:alternate names; *CPPCC*

Provenance Text:

China thanked **Mauritania** for supporting China on Taiwan, Tibet, Xinjiang, human rights and other issues concerning the country's core interests, Yu said.

Yu said the **CPPCC** would like to work with the ESC of Mauritania to carry out exchanges and promote bilateral relations.

→

Context: China thanked **Mauritania** for supporting China on Taiwan, Tibet, Xinjiang, human rights and other issues concerning the country's core interests, Yu said.

Yu said that the **CPPCC** would like to work with the ESC of Mauritania to carry out exchanges and promote bilateral relations.

Question: Which of the following answers is most applicable for "**Mauritania**;org:alternate names;**CPPCC**" (a) True, or (b) False?

Expected Response from LLM: (b), "b", False

Approaches

- **Zero shot**
  - Convert each knowledge graph extracted fact with provenance as input with *no demonstration example* and collect response from LLM
- **Few-shot In Context Learning (ICL)**
  - Convert each knowledge graph extracted fact with provenance as input along with *two demonstration example* as input and collect response form LLM
- **Few-shot Fine-Tuning**
  - Convert each knowledge graph extracted fact with provenance as input with no demonstration example but *fine-tune the model parameters*

Datasets

- **Two datasets:** TAC 2015, TAC 2017

  TAC is the annual Text Analysis Conference held by NIST since 2008
- **Five example per relation** i.e., 9% of training data to fine-tune LLM models

	Train	Validation	Test
TAC-2015	626	6859	6856
TAC-2017	552	5734	5729

Performance

Learning Approach	Model	TAC 2017			TAC 2015		
Baseline (Padia, Ferraro, and Finin 2022)		48.1	**98.0**	63.2	**50.8**	65.2	57.1
Zero Shot	GPT 3.5	41.6	46.7	43.9	40.4	41.6	41.0
	Flan-T5 (large)	50.9	37.4	43.1	63.0	29.0	39.7
In Context Learning	Flan-T5 (large)	39.3	64.8	48.9	41.2	44.9	43.0
Fine tuned Decoder Models	Galactica	34.8	40.2	37.3	29.1	64.0	40.0
	OPT	37.3	45.7	41.1	31.7	61.4	41.8
	Vicuna	35.9	95.1	52.2	27.0	**83.3**	40.8
Fine Tuned Encoder-Decoder Models	BART	34.3	65.1	44.9	29.9	79.9	43.6
	Flan-T5 (large)	**65.3**	66.5	**65.9**	49.5	77.5	**60.5**

Generic Models do not Outperform Fine-tuned Models to Identify Inconsistencies

Learning Approach	Model	TAC 2017			TAC 2015		
Baseline (Padia, Ferraro, and Finin 2022)		48.1	**98.0**	63.2	**50.8**	65.2	57.1
Zero Shot	GPT 3.5	41.6	46.7	43.9	40.4	41.6	41.0
	Flan-T5 (large)	50.9	37.4	43.1	63.0	29.0	39.7
In Context Learning	Flan-T5 (large)	39.3	64.8	48.9	41.2	44.9	43.0
Fine tuned Decoder Models	Galactica	34.8	40.2	37.3	29.1	64.0	40.0
	OPT	37.3	45.7	41.1	31.7	61.4	41.8
	Vicuna	35.9	95.1	52.2	27.0	**83.3**	40.8
	BART	34.3	65.1	44.9	29.9	79.9	43.6
Fine Tuned Encoder-Decoder Models	Flan-T5 (large)	**65.3**	66.5	**65.9**	49.5	77.5	**60.5**

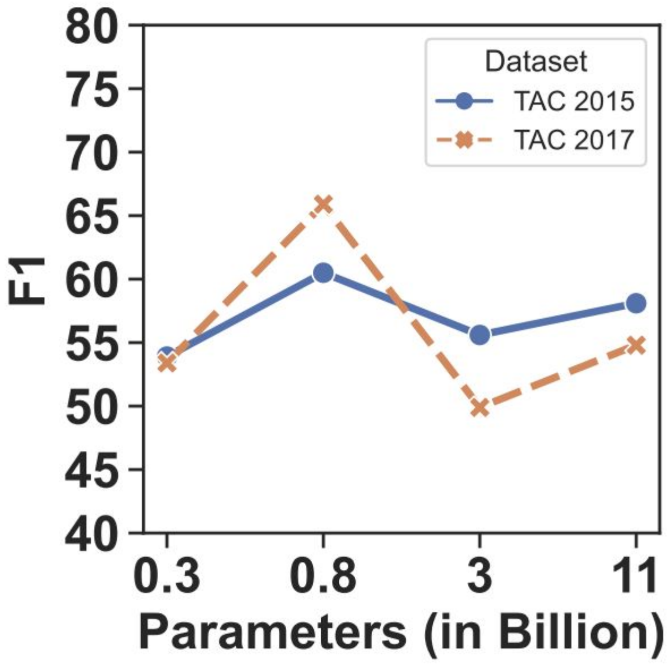
Generic model

Improvement due to demo examples

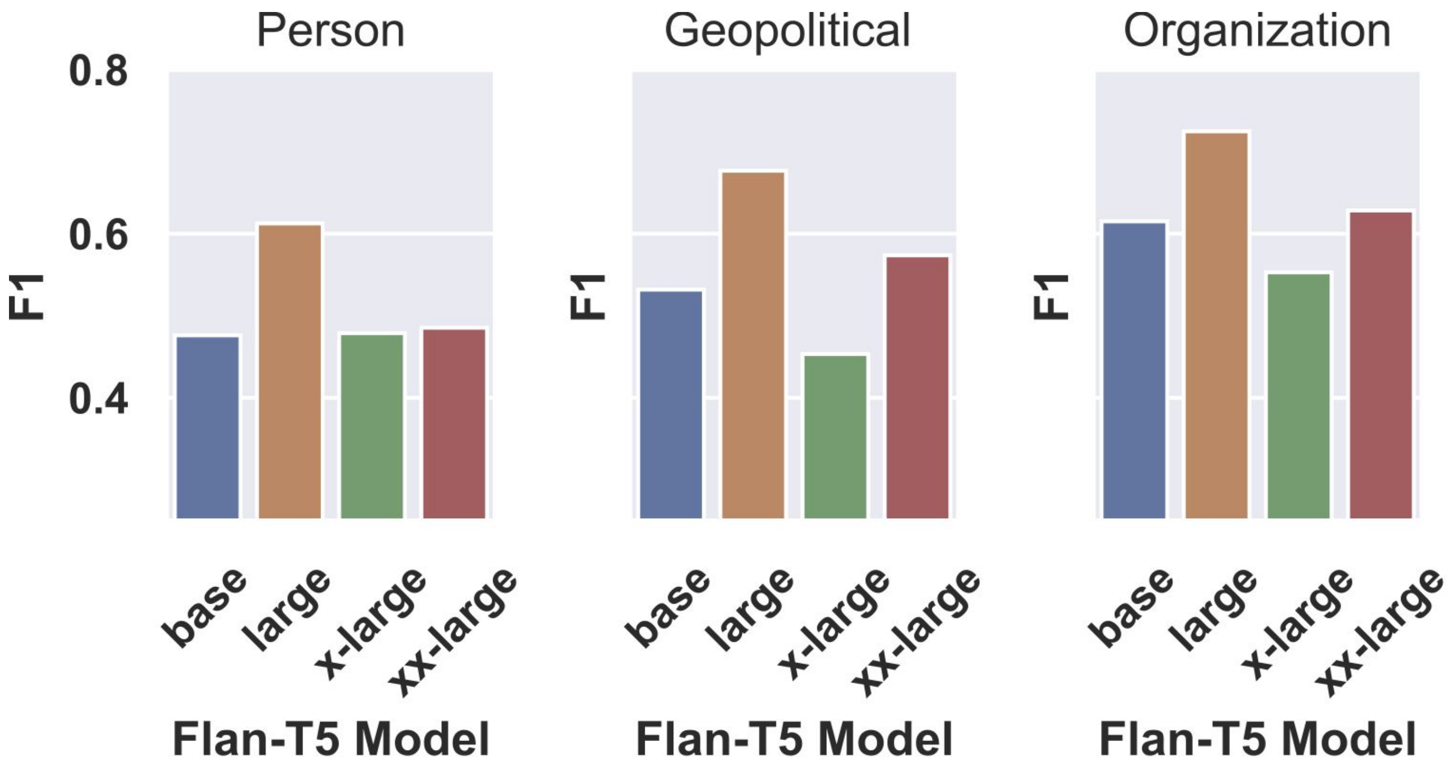
Improvement due to fine-tuning on training dataset

Increasing Model Size does not Increase KG Consistency

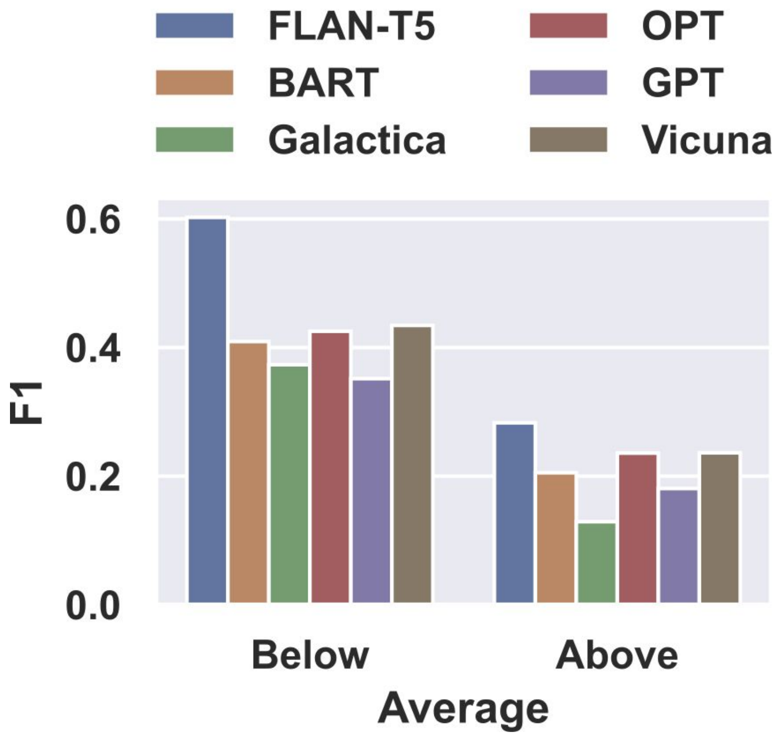
- Performance changes when changing the size of the model.
- Increasing model size does not increase Knowledge Graph Consistency
- Lower performance can be due to
    - **Quantization (8 bits)**
    - **Type of prompt** used for fine-tuning the model
    - **Number of training** examples



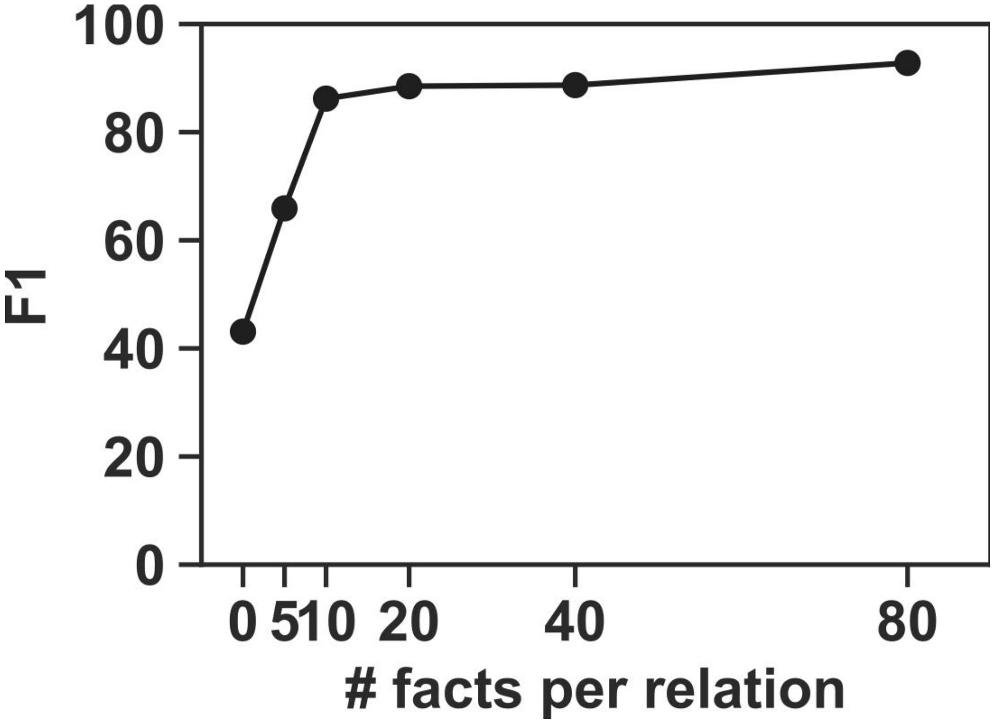
Large Model Variants Perform Differently Based on the Relation Domain



All Models are Sensitive to Number of Entities



LLMs Initially Learn Faster with More Data points, then Slower



Conclusion

- Explored limitations and capabilities of LLMs (BART, Flan-T5, Vicuna, OPT, Galactica, GPT 3.5) on Knowledge Graph consistency task
- Findings:
    - **LLM architecture**: Encoder-Decoder based model Flan-T5 performs better
    - **Size of LLM**: <1 billion parameters models are sufficient for the task
    - **Named entities**: More named entities confuses Large Language Models
    - **Training Examples**: Five to ten training examples are enough to identify knowledge graph inconsistencies
    - **In context learning**: Adding demonstration examples improves performance
    - **Fine-tuning**: Fine-tuning the model with few example performs better than incontext learning and zero-shot approach.